Optimal Transport in Statistical Inference and Computation

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029749">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029749</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Optimal Transport in Statistical Inference and Computation

A DISSERTATION PRESENTED
BY
Espen Bernton
TO
The Department of Statistics
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy IN THE SUBJECT OF Statistics
Harvard University
CAMBRIDGE, MASSACHUSETTS
April 2019
Optimal Transport in Statistical Inference and Computation

Abstract

This manuscript discusses four different topics at the intersection of statistics and optimal transport. Chapters 2 and 3 focus on different features of using the Wasserstein distance as a tool for inference in parametric models. Chapters 4 and 5 apply techniques and algorithms from the optimal transport literature to develop new theoretical results and methods in the field of Monte Carlo simulation. Each chapter is self-contained, and their corresponding abstracts are given below.

Chapter 2: Statistical inference can be performed by minimizing, over the parameter space, the Wasserstein distance between model distributions and the empirical distribution of the data. We study asymptotic properties of such minimum Wasserstein distance estimators, complementing results derived by Bassetti, Bodini and Regazzini in 2006. In particular, our results cover the misspecified setting, in which the data-generating process is not assumed to be part of the family of distributions described by the model. Our results are motivated by recent applications of minimum Wasserstein estimators to complex generative models. We discuss some difficulties arising in the numerical approximation of these estimators. Two of our numerical examples (g-and-k and sum of log-Normals) are taken from the literature on approximate Bayesian computation, and have likelihood functions that are not analytically tractable. Two other examples involve misspecified models.

Chapter 3: A growing number of generative statistical models do not permit the numerical evaluation of their likelihood functions. Approximate Bayesian computation (ABC) has become a popular approach to overcome this issue, in which one simulates synthetic
data sets given parameters and compares summaries of these data sets with the corresponding observed values. We propose to avoid the use of summaries and the ensuing loss of information by instead using the Wasserstein distance between the empirical distributions of the observed and synthetic data. This generalizes the well-known approach of using order statistics within ABC to arbitrary dimensions. We describe how recently developed approximations of the Wasserstein distance allow the method to scale to realistic data sizes, and propose a new distance based on the Hilbert space-filling curve. We provide a theoretical study of the proposed method, describing consistency as the threshold goes to zero while the observations are kept fixed, and concentration properties as the number of observations grows. Various extensions to time series data are discussed. The approach is illustrated on various examples, including univariate and multivariate g-and-k distributions, a toggle switch model from systems biology, a queueing model, and a Lévy-driven stochastic volatility model.

**Chapter 4:** Algorithms based on discretizing Langevin diffusion have become popular tools for sampling from high-dimensional distributions. We develop novel connections between such Monte Carlo algorithms, the theory of Wasserstein gradient flow, and the operator splitting approach to solving PDEs. In particular, we show that a proximal version of the Unadjusted Langevin Algorithm corresponds to a scheme that alternates between solving the Wasserstein gradient flows of the entropy and potential energy functionals on the space of probability measures. Using this composite minimization perspective, we derive some new non-asymptotic results on the convergence properties of this algorithm.

**Chapter 5:** Consider a reference Markov process with initial distribution $\pi_o$ and transition kernels $\{M_t\}_{t \in [0,T]}$, for some $T \in \mathbb{N}$. Assume that you are given distribution $\pi_T$, which
is not equal to the marginal distribution of the reference process at time $T$. In this scenario, Schrödinger addressed the problem of identifying the Markov process with initial distribution $\pi_0$ and terminal distribution equal to $\pi_T$ which is the closest to the reference process in terms of Kullback-Leibler divergence. This special case of the so-called Schrödinger bridge problem can be solved efficiently using the iterative proportional fitting procedure (IPFP), also known as Sinkhorn’s algorithm. We leverage these ideas to develop novel Monte Carlo schemes, termed Schrödinger bridge samplers, to approximate a target distribution $\pi$ on $\mathbb{R}^d$ and to unbiasedly estimate its normalizing constant. This is achieved by iteratively modifying the transition kernels of a reference Markov chain to obtain a process whose marginal distribution at time $T$ becomes closer to $\pi_T = \pi$, via regression-based approximations of the corresponding IPFP recursions. We empirically demonstrate its performance in several applications, and make connections with other problems arising in the optimal transport, optimal control and physics literatures.
## Contents

1 **Introduction**  
   1.1 Historical remarks and literature review  
   1.2 Outline of the manuscript  

2 **On parameter inference with the Wasserstein distance**  
   2.1 Introduction  
   2.2 Theoretical results  
   2.3 Computational aspects  
   2.4 Illustrations  
   2.5 Discussion  

3 **Approximate Bayesian computation with the Wasserstein distance**  
   3.1 Introduction  
   3.2 Wasserstein ABC  
   3.3 Theoretical properties  
   3.4 Time series  
   3.5 Numerical experiments  
   3.6 Discussion  

4 **Langevin Monte Carlo and JKO splitting**  
   4.1 Introduction  
   4.2 Wasserstein gradient flow  

vi
4.3 Operator splitting ................................................. 88
4.4 Proximal Langevin Monte Carlo ......................... 91
4.5 Discussion ......................................................... 95

5 SCHROEDINGER BRIDGE SAMPLERS 96
5.1 Introduction ..................................................... 96
5.2 Schrödinger bridges ............................................. 105
5.3 Sequential Schrödinger bridge samplers .............. 123
5.4 Connections with other problems ..................... 131
5.5 Numerical experiments ..................................... 140
5.6 Discussion ......................................................... 145

6 CONCLUSION ..................................................... 150
6.1 Summary .......................................................... 150
6.2 Future work and open problems .................... 152

A APPENDIX TO CHAPTER 2 154
A.1 Preliminary results .......................................... 154
A.2 Proofs: MWE .................................................. 156
A.3 Proofs: MEWE .................................................. 168
A.4 Checking the assumptions ................................ 175

B APPENDIX TO CHAPTER 3 179
B.1 Proofs .............................................................. 179
B.2 Sequential Monte Carlo sampler for ABC posteriors .. 186
B.3 Explicit rates and checking assumptions ............. 188
B.4 Concentration as \( n \) increases and \( \varepsilon \) decreases of the ABC posterior based
on the Hilbert and swapping distances .................... 195
B.5 Approximation of the Wasserstein distance .......... 202

C APPENDIX TO CHAPTER 4 206
C.1 Proofs .............................................................. 206
C.2 Rates for $\mathcal{W}_1(\rho_o, \pi)$ and $\mathcal{H}(\rho_o|\pi)$ ........................... 215

D Appendix to Chapter 5 ....................................................................................... 219
   D.1 Proofs ........................................................................................................... 219
   D.2 Linear quadratic Gaussian ........................................................................... 220

References ............................................................................................................. 241
Author List

The following authors contributed to Chapter 2:
  Pierre E. Jacob, Mathieu Gerber and Christian P. Robert.

The following authors contributed to Chapter 3:
  Pierre E. Jacob, Mathieu Gerber and Christian P. Robert.

The following authors contributed to Chapter 5:
  Pierre E. Jacob, Jeremy Heng and Arnaud Doucet.
List of Figures

1.1.1 Illustration of Monge’s problem. Figure from Villani (2008). .......................... 2

2.4.1 Estimators in the well-specified g-and-κ model, as described in Section 2.4.1. Figures 2.4.1a and 2.4.1b show the MEWE’s bivariate marginal sampling distributions for (a, b) and (g, κ) respectively, as n ranges from 50 to $10^4$ (colors from red to white to blue as n increases). For each n, we plot $M = 1,000$ estimators based on independent data sets. Each estimator was computed with $p = 1$, $m = 10^4$, $k = 20$, and one iteration of MCEM. Note that for small data sizes ($n = 50$ and $n = 100$), the estimator occasionally appears to be on the boundary of the parameter space, which could mean that the optimization procedure failed to converge. The intersections of the black lines indicate data-generating parameters. Figure 2.4.1c shows the MEWE’s marginal distribution for κ for the different levels of n, centered and rescaled by $\sqrt{n}$, illustrating the rate of convergence anticipated by Theorem 2.2.3. Figure 2.4.1d is a histogram of a data set generated with $\theta_* = (3, 1, 2, 0.5)$ and $n = 1,000$. .............................. 31
2.4.2 Estimators in the g-and-κ model with dependent data, as described in Section 2.4.1. Figures 2.4.2a and 2.4.2b show the MEWE's bivariate marginal sampling distributions for \((a, b)\) and \((g, \kappa)\) respectively, as \(n\) ranges from \(500\) to \(10^5\) (colors from red to white to blue as \(n\) increases). Note that the sample sizes here are 10 times larger than in the plots for the i.i.d. setting. For each \(n\), we plot \(M = 1,000\) estimators based on independent data sets. Each estimator was computed with \(p = 1, m = 10^4, k = 20\), and one iteration of MCEM. The intersections of the black lines indicate data-generating parameters. Figure 2.4.2c shows the MEWE's marginal distribution for \(\kappa\) for the different levels of \(n\), centered and rescaled by \(\sqrt{n}\), illustrating the rate of convergence anticipated by Theorem 2.2.3, but that the asymptotic variance is larger than in the i.i.d. case. Figure 2.4.2d shows the autocorrelation function of a data set generated with \(\theta_* = (3, 1, 2, 0.5), \rho = 0.75\), and \(n = 1,000\). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
2.4.5 Gamma data with \( n = 100 \), fitted with a Normal model, as described in Section 2.4.3. MEWEs are obtained for different values of \( m \) (from 10 to 10,000) and \( k \) (from 1 to 1,000), using one iteration of MCEM, \( M = 500 \) times independently. The intersections of the black lines represent the location of the “exact” MWE computed with \( n = m = 10^8 \).

2.4.6 Cauchy data fitted with a Normal model, as described in Section 2.4.4. Marginals distributions of the MEWE of \( \gamma \) and \( \sigma \), centered by \( \theta^* \) itself computed with \( n = m = 10^8 \), and rescaled by \( \sqrt{n} \), are shown in Figures 2.4.6a and 2.4.6b. Figures 2.4.6c and 2.4.6d show the distributions of the MEWE for \( n = 5,000 \) (full lines), along with the distribution of an estimator obtained by minimizing an estimate of the Kullback–Leibler divergence (dashed lines).

3.2.1 ABC in the bivariate Normal location model of Section 3.2.2. ABC approximations of the posterior after \( 10^6 \) model simulations (left and middle), overlaid the actual posterior. On the right, the Wasserstein distance between ABC posterior samples and exact posterior samples is plotted against the number of model simulations (in log-log scale). In principle, these ABC approximations converge to the posterior as \( \varepsilon \to 0 \). Yet, for a given number of model simulations, the quality of the ABC approximation is sensitive to the choice of distance and sampling algorithm.

3.4.1 ABC posterior samples in the cosine model of Section 3.4.1, using either the Euclidean distance or curve matching with the exact Wasserstein distance and \( \lambda = 1 \), after \( 10^6 \) model simulations. We compare to the posterior distribution, obtained using the 50,000 last samples in a Metropolis–Hastings chain of length 100,000. The standard deviation of the noise \( \sigma \) is better estimated with curve matching than with the Euclidean distance between time series.
3.4.2 Samples from the WABC posteriors of \((\varphi, \log(\sigma))\) in the AR(1) model of Section 3.4.2, as \(c\) decreases over the steps of the SMC sampler (colors from red to white to blue). On the left, using the marginal empirical distribution of the series, the WABC posteriors concentrate around a ridge of values such that \(\sigma^2/(1 - \varphi^2)\) is constant. On the right, using delay reconstruction with lag \(k = 1\), the WABC posteriors concentrate around the data-generating parameters, \(\varphi = 0.7\), \(\log(\sigma) = 0.9\), indicated by full lines. Both methods had a total budget of \(10^5\) model simulations.

3.5.1 Posterior marginals in the univariate g-and-k example of Section 3.5.1 (obtained via MCMC), approximations by Wasserstein ABC and semi-automatic ABC, each with a budget of \(2.4 \times 10^6\) model simulations. Data-generating values are indicated by vertical lines.

3.5.2 3.5.2a-3.5.2d: Posterior marginals in the univariate g-and-k example of Section 3.5.1 (grey, obtained via MCMC) and approximations by Wasserstein ABC from step 20 to step 57 of the SMC algorithm. The color of the WABC approximation changes from red to blue as more steps of the SMC sampler are performed, decreasing the threshold \(c\). For the densities plotted here, the threshold reduces from \(c = 0.20\) to \(c = 0.07\). The range of the plot for \(g\) has been truncated to \((1, 6)\) to highlight the region of interest, despite the densities from the earlier steps of the SMC having support outside this region. Data-generating values are indicated by vertical lines. Figure 3.5.2e shows the number of simulations from the model used in each step of the SMC algorithm (\(y\)-axis in log scale). This number is increasing due to use of the \(r\)-hit kernel within the SMC. Figures 3.5.2f and 3.5.2g show the threshold \(c\) and the \(W_r\)-distance to the posterior respectively, against the number of model simulations (both plots in log-log scale).
3.5.3 3.5.3a shows the number of simulations from the model used in each step of the SMC algorithm for the four distances applied to the bivariate g-and-k model of Section 3.5.1 (y-axis in log scale). This number is increasing due to use of the 𝑟-hit kernel within the SMC. Figure 3.5.3b shows the thresholds ε against the number of model simulations (in log-log scale). Note that the MMD is not on the same scale as the Wasserstein distance and its approximations, and therefore the MMD thresholds are not directly comparable to the those of the other distances. Figure 3.5.3c shows the 𝑉_{ij}-distance between the joint ABC posteriors based on the different distances to the joint true posterior, against the number of model simulations (in log-log scale).

3.5.4 Histogram of observations (3.5.4a), and marginal posteriors based on WABC and the summary statistic from Bonassi et al. (2011); Bonassi and West (2015) in the toggle switch model. The ABC posteriors are computed using the SMC sampler from Section 3.2.1, for a total number of 10^6 model simulations. Data-generating values are indicated by vertical lines.

3.5.5 Posterior marginals in the M/G/1 queueing model of Section 3.5.3 (obtained via particle marginal Metropolis–Hastings), approximations by Wasserstein ABC and semi-automatic ABC, and Wasserstein ABC accounting for the constraint that θ_i has to be less than min_{j∈[n]} y_j, each with a budget of 10^7 model simulations. Data-generating values are indicated by vertical lines.

3.5.6 ABC approximations in the Lévy-driven stochastic volatility model, using the Hilbert distance between delay reconstructions with lag 𝑘 = 1. The plots show samples from the bivariate marginals of (µ, β) (left), (ξ, ω²) (middle), and the marginal distributions of λ (right), as the threshold ε decreases during the steps of the SMC sampler (colors from red to blue). The total budget was 4.2 × 10^5 model simulations. Data-generating parameters are indicated by full lines.
3.5.7 Left: summary, defined as the sum of the first 50 sample autocorrelations of the squared series, against $\lambda$, computed for the output of the WABC algorithm using the Hilbert distance between delay reconstructions, applied to the Lévy-driven stochastic volatility model of Section 3.5.4. Middle and right: approximations of $\omega^2$ and $\lambda$, from the second run of WABC using the Hilbert distance between delay reconstructions combined with the summary for another $6.6 \times 10^5$ model simulations. The colors change from red to blue as more steps of the SMC sampler are performed. The horizontal axis in the right plot is in log-scale, and illustrates the concentration of the ABC posterior towards the data-generating value of $\lambda$. ................................................................. 77

5.1.1 Illustration of the iterative proportional fitting procedure. The blue line represents $P_o(\pi_o) \subset P(E^{T+1})$, denoted $P_o$ in the figure, while the red line represents $P_T(\pi_T) \subset P(E^{T+1})$, denoted $P_T$. The black line illustrates that the KL projections $Q(i) \in P_o(\pi_o)$ and $P(i) \in P_T(\pi_T)$ converge towards the Schrödinger bridge $S$. ................................................................. 103

5.1.2 Illustration of the evolutions of the marginal distributions $q_t$ and $s_t$ of the reference process $Q(dx_{o,T})$ and the associated Schrödinger bridge $S(dx_{o,T})$, respectively. The colors transition from red to green to blue as $t$ increases from 0 to $T = 40$. The Schrödinger bridge interpolates between the initial distribution $\pi_o$ and the target distribution $\pi_T = \pi$, here given in grey, while remaining close to the reference process. ................................................................. 104
5.2.1 Distances between marginals of the Schrödinger bridge \( s_t \) and the marginals of the IPFP iterates \( q_t^{(i)} \), measured as \( \log W_2(s_t, q_t^{(i)}) \), for the LQG setting of Section 5.2.6 with discretized Brownian diffusion reference dynamics. Figure 5.2.1a corresponds to the exact computation of the IPFP iterates for \( i \in [0 : 5] \). The remaining plots correspond to the proposed particle-based approximation of the IPFP iterations using \( N = 1000 \) particles and different values of conditional SMC iterations \( M \) (using \( P = 128 \) cSMC particles). The solid lines correspond to the median value of the log-distance calculated over 100 independent simulations, and corresponding confidence bands represent the 5% and 95% quantiles. Note that the vertical axis of Figure 5.2.1a is on a different scale than those of the other figures, as the exact IPFP iterations yield smaller distances in general.

5.2.2 Distances between marginals of the Schrödinger bridge \( s_t \) and the marginals of the IPFP iterates \( q_t^{(i)} \), measured as \( \log W_2(s_t, q_t^{(i)}) \), for the LQG setting of Section 5.2.6 with discretized Langevin diffusion reference dynamics. Figure 5.2.2a corresponds to the exact computation of the IPFP iterates for \( i \in [0 : 5] \). The remaining plots correspond to the proposed particle-based approximation of the IPFP iterations using \( N = 1000 \) particles and different values of conditional SMC iterations \( M \) (using \( P = 128 \) cSMC particles). The solid lines correspond to the median value of the log-distance calculated over 100 independent simulations, and corresponding confidence bands represent the 5% and 95% quantiles. Note that the vertical axis of Figure 5.2.2a is on a different scale than those of the other figures, as the exact IPFP iterations yield smaller distances in general.
5.3.1 Distances between $\pi_t$ and marginals of the IPFP-based approximations $q_t^{(i)}$ of the multi-marginal Schrödinger bridge, measured as $\log \mathcal{W}_2(\pi_t, q_t^{(i)})$, for the LQG setting of Section 5.3.6 with discretized Langevin diffusion reference dynamics. Figure 5.3.1a correspond to the exact computation of the IPFP iterates for different values of $i$ between 0 and 100. Figure 5.3.1b corresponds to the proposed particle-based approximation using $N = 1,000$ particles and $M = 0$ iterations of cSMC. Figure 5.3.1c illustrates the differences between the approximations obtained with and without the warm start and early stopping schemes discussed in Section 5.3.5. Figure 5.3.1d shows the number of IPFP iterations performed in the early stopping scheme at each time $t \in [1 : T]$, with a lower bound set to 3 iterations. In all the plots, solid lines correspond to median values calculated over 100 independent simulations, and the corresponding confidence bands represent the 5% and 95% quantiles.

5.3.2 On the left, we plot the logarithm of the estimates $\{\hat{Z}_t\}_{t \in [0; T]}$ of the normalizing constants $\{Z_t\}_{t \in [0; T]}$ corresponding to $\{\pi_t\}_{t \in [0; T]}$, based on an SMC sampler using Markov kernels derived from discretizing Langevin diffusion (in red) and the SSB sampler using the same discretized Langevin Markov kernels as the reference process (in blue). The solid lines correspond to median values calculated over 100 independent simulations using $N = 1,000$ particles and $M = 0$ iterations of cSMC, and the corresponding confidence bands represent the 5% and 95% quantiles. Note that the true log normalizing constants (in black) are obscured by the blue line, and that the blue confidence bands are too narrow to be visible. To further illustrate the behavior of the different estimators, we compute the root mean squared error of $\log \hat{Z}_t$, the log of which is shown on the right.
5.4.1 Results of the numerical experiments performed for the LQG model discussed in the particle filtering setting of Section 5.4.4. In Figure 5.4.1a, we perform exact IPFP and plot log \( \mathcal{W}_2(\pi_t, q_i^{(i)}) \) for \( t \in [1 : T] \) and \( i \in [1 : s] \), where \( \pi_t \) was derived using a Kalman filter. In Figure 5.4.1b, we plot the log 2-Wasserstein distances between \( \{ \pi_t \}_t \in [1 : T] \) and their Gaussian approximations based on the Schrödinger bridge and bootstrap particle filters. In Figures 5.4.1c and 5.4.1d we plot the marginal likelihood estimates and effective sample sizes as a percentage of \( N \) for the two particle filters. In each plot based on particle approximations, solid lines correspond to median values calculated over 100 independent simulations, and the corresponding confidence bands represent the 5% and 95% quantiles.

5.5.1 Part 1/3. Distances between \( \pi_t \) and marginals of the IPFP-based approximations \( q_i^{(i)} \) of the multi-marginal Schrödinger bridge, measured as log \( \mathcal{W}_2(\pi_t, q_i^{(i)}) \), for the LQG setting of Section 5.5.1 with discretized Langevin diffusion reference dynamics (left column), and log-RMSE of the log-normalizing constant estimates log \( \tilde{Z}_t \) (right column). The red lines correspond to the reference process and the estimators obtained with the corresponding SMC sampler. The blue and purple lines correspond to the SSB sampler with exact twisting with and without refreshment steps, respectively. The green line correspond to the SSB sampler using Euler-Maruyama approximate twisting and refreshment steps. Figure continued on the next page.

B.3.1 Numerical experiments with a Gamma data-generating distribution \( \mu_* \) and a Normal location model \( \mu_\theta \), described in Section B.3.1, illustrating that the 2-Wasserstein distance between \( \mu_\theta \) and \( \mu_* \) satisfies Assumption B.1.5. On the left, \( \mathcal{W}_2(\mu_*, \mu_\theta) \) is approximated with 500,000 samples and plotted against \( \theta \). On the right \( \sqrt{\mathcal{W}_2(\mu_*, \mu_\theta) - \varepsilon_*} \) can be seen to be linearly associated with \( |\theta - \theta_*| \), for \( \theta \) around \( \theta_* \) (the best linear fit is indicated by a blue line).

B.3.2 Distance \( \rho_{\mathcal{W}}(\theta, \theta_*) \) against \( \mathcal{W}_2(y_{1:n}, z_{1:n}) \), where \( y_{1:n} \sim \mu_{\theta_*} \) and \( z_{1:n} \sim \mu_\theta \), when varying each component of \( \theta \) in a neighborhood of \( \theta_* \), in the g-and-k example as described in Section B.3.3.
B.3.3 Distance $\rho_H(\theta, \theta_*)$ against $\mathcal{W}_2(y_{1:n}, z_{1:n})$, where $y_{1:n} \sim \mu_{\theta_*}$ and $z_{1:n} \sim \mu_{\theta}$ when varying all components of $\theta$ in a neighborhood of $\theta_*$, in the g-and-k example as described in Section B.3.3.

B.5.1 Hilbert, swapping and Wasserstein distances between data generated from multivariate Normal distributions with variance $\theta$ times the identity matrix, versus $\theta_*=4$, with 500 independent draws. Each panel corresponds to a different dimension. Details are provided in Section B.5.
Acknowledgments

First, I want to thank my advisor, Pierre E. Jacob, for his invaluable mentoring throughout my PhD studies. I will always aspire to be as generous with my time and ideas as he has been with me. Second, I would like to thank the other members of my dissertation committee: Neil Shephard, whose advice and guidance I value second to none, and Marco Cuturi, who, together with Nicolas Chopin, was a gracious host during my semester in Paris and has been crucial in developing my interest for optimal transport. Third, I would like to thank my coauthors, Mathieu Gerber, Christian P. Robert, Jeremy Heng, and Arnaud Doucet, for their hard work and dedication. I would also like to thank the rest of the faculty at Harvard for teaching me statistics, and Gabriel Peyré and Arnak Dalalyan for many interesting discussions.

I have been lucky to make many friends during my doctoral studies, and would thank them all for their support. In particular, I owe thanks to my fellow PhD students at Harvard and CREST, and to Christopher Frye and Jérémy L’Hour for being my trustworthy workout partners. I have been equally lucky to maintain the support and patience of my family, despite infrequent visits and phone calls home. Lastly, I would like to thank my girlfriend Audrey, who has been nothing but a distraction over the last four years, but without whom I would be a much less happy man.
1

Introduction

1.1 Historical remarks and literature review

The history of optimal transport began with Gaspard Monge’s famous work Mémoire sur la théorie des déblais et des remblais (Monge, 1781), in which he considered the problem of transporting a certain amount of mass extracted from a quarry (“déblais”) to locations where that material is to be processed (“remblais”). In light of the cost associated with moving a unit of mass from one point to another, Monge was interested finding the location assignments that minimize the total cost of the transportation. In modern terms, Monge’s problem can be stated as follows. Given two distributions of mass $\mu$ and $\nu$ on a measurable space $(X, \mathcal{X})$ and a cost function $c : X \times X \to \mathbb{R}_+$, find a map $T : X \to X$ that pushes $\mu$ onto $\nu$, i.e. $\nu(A) = \mu(T^{-1}(A))$ for every $A \subset \mathcal{X}$, and that minimizes the quantity

$$\int_X c(x, T(x))\,d\mu(x). \quad (1.1)$$
Figure 1.1.1: Illustration of Monge’s problem. Figure from Villani (2008).

Guided by geometric intuition, Monge made deep insights into the nature of the optimal transport problem and its solution in the case where $X = \mathbb{R}^3$ and $c(x, y) = \|x - y\|_2$. However, fundamental questions about the existence of an optimal map went unanswered for the better part of 200 years. It was not until after Leonid V. Kantorovich’s rediscovery of the problem in the 1940s that significant progress was made (Kantorovich, 1942). In fact, Kantorovich considered a relaxation of Monge’s problem in which the deterministic location assignment $T(x)$ is replaced by a probabilistic assignment. A priori, this allows the mass at point $x$ to be moved to multiple locations, a concept also known as mass splitting. Formally, Kantorovich’s problem can be stated as follows. Over couplings $\gamma \in C(\mu, \nu)$, i.e. distributions on $X \times X$ such that $\gamma(A \times X) = \mu(A)$ and $\gamma(X \times B) = \nu(B)$ for all sets $A, B \subset \mathcal{X}$, find the minimum

$$
\min_{\gamma \in C(\mu, \nu)} \int_{X \times X} c(x, y) \, d\gamma(x, y).
$$

Recognizing that problems of this kind were representative of a larger class of questions of optimal allocation of economic resources, Kantorovich was inspired to develop the theory of linear programming, for which he shared the 1975 Nobel Prize in Economics with Tjalling Koopmans.

In the time that has passed since Kantorovich’s discoveries, the field of optimal transport and the span of its applications has grown dramatically. Optimal couplings have proven useful theoretical tools in fluid mechanics (e.g. Brenier, 1991), kinetic theory (e.g. Sznitman, 1991), diffusion equations (e.g. Jordan et al., 1998), probability theory (e.g. Rachev
and Rüschendorf, 1998), and many other topics. This usefulness stems in part from optimal transport’s ability to lift geometric structure captured by the cost \( c \) on the base space \( X \) up to the space of probability measures over \( X \), denoted \( \mathcal{P}(X) \). For instance, if \( c = \rho^p \) for some \( p \geq 1 \) and metric \( \rho \) on \( X \), the corresponding optimal transport problem

\[
\mathcal{W}_p(\mu, \nu)^p = \min_{\gamma \in \mathcal{C}(\mu, \nu)} \int_{X \times X} \rho(x, y)^p d\gamma(x, y) \tag{1.3}
\]

defines the \( p \)-Wasserstein\(^1 \) distance \( \mathcal{W}_p \) on \( \mathcal{P}_p(X) \subset \mathcal{P}(X) \), the space of probability measures with finite \( p \)th moments.

The rich geometric structure captured by the Wasserstein distance gives rise powerful generalizations of concepts like convexity and gradient flows of functionals defined on \( \mathcal{P}(X) \) (McCann, 1997; Otto, 2001; Ambrosio et al., 2005). Inheriting geometric properties from the base metric \( \rho \) also gives the Wasserstein distance important statistical properties, like the fact that it metrizes weak convergence in \( \mathcal{P}_p(X) \) (Villani, 2008, Theorem 6.9). Both of these features will be central to the developments in this manuscript.

Despite its many successful theoretical applications, the role of optimal transport in inference and data analysis was for long time inhibited by high computational costs: Given two empirical distributions \( \hat{\mu}_n \) and \( \hat{\nu}_m \), each consisting of \( n \) atoms, (1.2) reduces to an assignment problem that can be solved using dual ascent methods with computational complexity on the order of \( n^3 \) (Bertsimas and Tsitsiklis, 1997). Recently, Cuturi (2013) recognized that adding an entropic penalty to the objective in (1.2) enables the application of Sinkhorn’s algorithm\(^2 \) (Sinkhorn and Knopp, 1967), which has complexity of \( O(n^3) \) per iteration, at the expense of only yielding an approximate solution. Since then, there has been rapid growth in the development of other fast approximate solvers; see e.g. Benamou et al. (2015); Genevay et al. (2016); Li et al. (2018); Altschuler et al. (2018), and the recent book by Peyré and Cuturi (2019).

In turn, the availability of efficient computational methods has made optimal transport applicable in many data-driven fields, including computer graphics (e.g. Solomon et al., 2015), image processing and generation (e.g. Kolouri et al., 2017; Genevay et al., 2018b),

\(^1\)Despite its name, the Wasserstein distance was first formulated by Kantorovich and Rubinstein (1958).

\(^2\)In statistics, Sinkhorn’s algorithm is better known as the iterative proportional fitting procedure.
signal processing (e.g. Thorpe et al., 2017), and various classical problems in statistics and machine learning (e.g. Froger, et al., 2015; Ramdas et al., 2017; Sommerfeld and Munk, 2018). This practice has lead to the need for a solid theoretical foundation of transport-based inference methods. There have been several recent advances in this direction, including Singh and Póczos (2018); del Barrio and Loubes (2019); Weed and Bach (2019); Weed and Berthet (2019); see also the review of Panaretos and Zemel (2019).

In conjunction with its growing prominence as an inferential tool, optimal transport has also become a popular technical tool in the Monte Carlo literature. For instance, the Wasserstein distance has been fruitfully applied to capture convergence and contraction properties of many stochastic processes and sampling schemes (e.g. Dalalyan, 2017; Bubeck et al., 2018; Rudolf and Schweizer, 2018; Bou-Rabee et al., 2018; Majka et al., 2018). Various constructions of transport maps and couplings have also been useful in developing new algorithms to solve problems in filtering (e.g. Reich, 2013; Taghvaei and Mehta, 2016) and Bayesian computation (e.g. Marzouk et al., 2017; Parno and Marzouk, 2018; Wang et al., 2019).

1.2 Outline of the Manuscript

The contributions of this manuscript follow in the traditions outlined the two paragraphs above: The first half (Chapters 2 and 3) is devoted to studying problems that arise in the estimation of parametric models using the Wasserstein distance. The second half (Chapters 4 and 5) applies tools and algorithms from the optimal transport literature to problems in Monte Carlo simulation. Each chapter is self-contained and reviews the relevant parts of the optimal transport literature.

Chapter 2 is concerned with the statistical properties of frequentist estimators derived by minimizing the Wasserstein distance between a model and an empirical distribution, generalizing previous results by Bassetti et al. (2006); Bassetti and Regazzini (2006). This work is motivated by recent applications of such estimators in complex generative models (e.g. Arjovsky et al., 2017; Genevay et al., 2018b), in which the associated likelihood is not tractable and the supports of the model and data-generating process do not overlap. This prevents the use of standard inference procedures based on the likelihood function or
entropy-based notions of discrepancy. On the other hand, the Wasserstein distance still defines a useful loss function in such settings. Because the models involved are inherently misspecified, the results of Bassetti and coauthors do not apply and need to be extended. We establish basic properties such as existence, measurability and consistency of the estimators in the misspecified settings, and derive the rate of convergence and asymptotic distribution of the minimum Wasserstein estimator with $p = 1$ for one dimensional data.

Chapter 3 is also concerned with likelihood-free parameter inference in generative models, but takes a Bayesian approach. In particular, we use the Wasserstein distance within approximate Bayesian computation (ABC) to estimate the posterior distribution. Existing ABC methods typically capture differences between data sets by computing distances between summary statistics, but given the complex nature of the models to which ABC is typically applied, this can often lead to a systematic loss of information. We instead propose to use the Wasserstein distance as a discrepancy between data sets, and show that this generalizes the common practice of computing distances between order statistics to higher dimensions. The Wasserstein ABC method can under some assumptions be shown to yield consistent approximations of the posterior as the ABC threshold converges to zero. We also investigate some of its other theoretical properties.

In Chapter 4, we show that certain Markov chain Monte Carlo algorithms based on discretizing Langevin diffusion, often called Langevin Monte Carlo methods, can be viewed as approximations of the Wasserstein gradient flow of the Kullback-Leibler divergence with respect to the target distribution. In other words, optimal transport allows us to view a random sampling process on $X = \mathbb{R}^d$ as a deterministic optimization algorithm in $\mathcal{P}(X)$. In turn, this allows us to make use of tools from the optimization literature to provide new non-asymptotic results about the convergence of a proximal version of the LMC algorithm that capture its dependence on $d$. We also point to how this perspective can potentially be used to derive new sampling algorithms by mirroring developments that have been made in optimization and optimal transport.

In Chapter 5, we investigate a regularization of the optimal transport problem known as the Schrödinger bridge problem, and use it to develop an efficient Sequential Monte Carlo (SMC) sampler. Given an easy-to-sample initial distribution $\pi_0 \in \mathcal{P}(X)$ and a sequence of $T \in \mathbb{N}$ Markov kernels that when composed target a distribution $\pi \in \mathcal{P}(X)$, we ob-
serve that the optimal auxiliary target distribution on \( P(X^{T+1}) \) derived by Del Moral et al. (2006) can be seen as the solution of a so-called forward Schrödinger half-bridge problem. By composing the solutions of such forward problems with their backward counterparts, we obtain an instance of the iterative proportional fitting procedure, which is known to converge to the Schrödinger bridge between \( \pi_0 \) and \( \pi \). From the Monte Carlo standpoint, this is desirable because it provides a means to draw samples from the target distribution \( \pi \) by initializing particles from \( \pi_0 \) and propagating them through the bridge. We develop an algorithm to approximate the Schrödinger bridge based on approximate dynamic programming, utilizing an equivalent formulation in terms of optimal control, that only requires pointwise evaluations of the target density up to a normalizing constant. We also study some of the Schrödinger bridge problem’s connections to flow transport and shortcuts to adiabaticity in physics.
2

On parameter inference with the Wasserstein distance

2.1 INTRODUCTION

We consider a statistical estimation approach for parametric models that is based on minimizing the Wasserstein distance between the empirical distribution of the data and the model distributions (Belili et al., 1999; Bassetti et al., 2006). We study two different point estimators, where the first, called the minimum Wasserstein estimator (MWE), arises as the most important special case of the estimator introduced by Bassetti et al. (2006). The second, which we term the minimum expected Wasserstein estimator (MEWE), is better suited to numerical approximations.

We derive theoretical properties of the estimators, such as existence, measurability, and consistency, in the misspecified setting. That is, we do not assume that the observations are
generated from the working model. For one-dimensional data, we also study the convergence rate and asymptotic distribution of the minimum Wasserstein estimator of order 1, extending the work of Bassetti and Regazzini (2006) on location-scale models. Our proofs are based on epi-convergence (Rockafellar and Wets, 2009) and general results on minimum distance estimation (Pollard, 1980), and are as such different from those presented by Bassetti and coauthors.

There are two main motivations for developing these results. Firstly, recent advances in computational optimal transport have led to the application of minimum Wasserstein distance estimators in increasingly complicated settings, where the models are likely to be misspecified. For instance, Genevay et al. (2018b) apply the MEWE in the tuning of image generation models, and Genevay et al. (2017) show that a version of the MEWE also appears in the popular Wasserstein GAN method (Arjovsky et al., 2017). This development has been driven by the advent of efficient numerical algorithms to approximate the Wasserstein distance (see e.g. Peyré and Cuturi, 2019; Cuturi, 2013; Benamou et al., 2015; Genevay et al., 2016; Li et al., 2018; Altschuler et al., 2018).

Secondly, minimum Wasserstein distance estimators, which are particular instances of minimum distance estimators (Basu et al., 2011), appear to be practical and robust alternatives to likelihood-based estimation in the setting of generative models. In these models, synthetic observations can be generated given a parameter, but the likelihood function and associated maximum likelihood estimators might be intractable (Gouriéroux et al., 1993; Marin et al., 2012; Bernton et al., 2019). Some comments on the comparison between the Wasserstein distance and other distances commonly used in minimum distance estimation are provided.

The rest of this chapter is organized as follows: we review the definitions of minimum distance estimation, of the Wasserstein distance, and of the estimators of interest in the rest of this section. Theoretical results, whose proofs can be found in Appendix A, and some open questions are stated in Section 2.2. We briefly review computational strategies to compute the Wasserstein distance and the estimators in Section 2.3, before illustrating their behavior on various examples in Section 2.4. We conclude in Section 2.5. Code to reproduce the numerical results can be found at https://github.com/pierrejacob/winference.
2.1.1 Notation

Throughout this chapter we consider a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), with associated expectation operator \(\mathbb{E}\), on which all the random variables are defined. The set of probability measures on a space \(\mathcal{X}\) is denoted by \(\mathcal{P}(\mathcal{X})\). The data take values in \(\mathcal{Y}\), a subset of \(\mathbb{R}^d\) for some \(d \in \mathbb{N}\), and is endowed with the Borel \(\sigma\)-algebra. We observe \(n \in \mathbb{N}\) data points, \(y_{1:n} = y_1, \ldots, y_n\), that are distributed according to \(\mu^{(n)} \in \mathcal{P}(\mathcal{Y})\). Let \(\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i}\), where \(\delta_y\) is the Dirac distribution with mass on \(y \in \mathcal{Y}\). We refer to \(\hat{\mu}_n\) as the empirical distribution of \(y_{1:n}\), even in settings where the observations are not i.i.d.

A model refers to a collection of distributions on \(\mathcal{Y}^n\), denoted by

\[
\mathcal{M}^{(n)} = \{\hat{\mu}^{(n)}_\theta : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y}^n),
\]

where \(\mathcal{H} \subset \mathbb{R}^{d_\theta}\) is the parameter space, endowed with a distance \(d_\mathcal{H}\) and of dimension \(d_\theta \in \mathbb{N}\). However, we will often assume that the sequence of models \((\mathcal{M}^{(n)})_{n \geq 1}\) is such that, for every \(\theta \in \mathcal{H}\), the sequence \((\hat{\mu}^{(n)}_{\theta,n})_{n \geq 1}\) of random probability measures on \(\mathcal{Y}\) converges (in some sense) to a distribution \(\hat{\mu}_\theta \in \mathcal{P}(\mathcal{Y})\), where \(\hat{\mu}_{\theta,n} = n^{-1} \sum_{i=1}^n \delta_{z_i}\) with \(z_{1:n} \sim \mu^{(n)}_\theta\). Similarly, we will often assume that \(\hat{\mu}_\theta^*\) converges to some distribution \(\mu^* \in \mathcal{P}(\mathcal{Y})\) as \(n \to \infty\). Whenever the notation \(\mu^*_{\theta}^*\) and \(\mu_{\theta}^*\) is used, it is implicitly assumed that these objects exist. In such cases, we instead refer to \(\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y})\) as the model.

We say that it is well-specified if there exists \(\theta^* \in \mathcal{H}\) such that \(\mu^* = \mu_{\theta^*}\); otherwise it is misspecified. Parameters are identifiable if \(\theta = \theta'\) is implied by \(\mu_\theta = \mu_{\theta'}\). The weak convergence of a sequence of measures \(\mu_n\) to \(\mu\) is denoted by \(\mu_n \Rightarrow \mu\). The Kullback-Leibler (KL) divergence between \(\mu\) and \(\nu\) is defined as

\[
KL(\mu \| \nu) = \int \log \frac{d\mu}{d\nu} d\mu
\]

if \(\mu\) is absolutely continuous with respect to \(\nu\), and \(+\infty\) otherwise.
2.1.2 Minimum distance estimation

Minimum distance estimation refers to the minimization, over the parameter $\theta \in \mathcal{H}$, of a distance between the empirical distribution $\hat{\mu}_n$ and the model distribution $\mu_\theta$ (Wolfowitz, 1957; Basu et al., 2011). More formally, denoting by $\mathcal{D}$ a distance or divergence on $\mathcal{P}(\mathcal{Y})$, the associated minimum distance estimator (MDE) can be defined as

$$\hat{\theta}_n = \arg\min_{\theta \in \mathcal{H}} \mathcal{D}(\hat{\mu}_n, \mu_\theta).$$ \hspace{1cm} (2.1)

In broad terms, the minimum distance estimation principle captures the idea of many statistical paradigms. For instance, the generalized method of moments (Hansen, 1982) consists in minimizing a discrepancy $\mathcal{D}$ defined as the weighted Euclidean distance between moments of $\hat{\mu}_n$ and $\mu_\theta$. In the empirical likelihood method (Owen, 2001), $\mathcal{D}$ is taken to be the KL divergence, and the model is supported strictly on the set of observed data and subject to moment conditions. The maximum likelihood estimator minimizes the KL divergence between $\mu_\ast$ and $\mu_\theta$ in the limit of the number of observations going to infinity.

However, it is worth noting that the definition in (2.1) precludes the naive application of some discrepancy measures. For instance, one could not directly choose $\mathcal{D}$ to be the KL divergence or the total variation distance, since for any model distribution $\mu_\theta$ not supported solely on the observed data, they would evaluate to $+\infty$ and 1 respectively. To apply discrepancies of this kind, one would first need to build sample-based estimators of the underlying population quantity $\mathcal{D}(\mu_\ast, \mu_\theta)$, assuming it is well-defined. Many such approaches have been studied in detail by Basu et al. (2011).

The computation of the minimum distance estimator might be intractable, especially in settings where it is assumed that one can simulate data from the model distribution but not evaluate its density. For such generative models, the following minimum expected distance estimator might be more computationally convenient:

$$\hat{\theta}_{n,m} = \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_n \mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m}),$$ \hspace{1cm} (2.2)
where the expectation $\mathbb{E}_m$ is taken over the distribution of the sample $z_{1:m} \sim \mu^{(m)}_\theta$ giving rise to $\hat{\mu}_{\theta, m} = \frac{1}{m} \sum_{i=1}^{m} \delta_{z_i}$. When $n$ is fixed and $m$ is large, or when $n = m$ and $n$ is large, one might hope that the expectation is close to $\mathcal{D}(\hat{\mu}_n; \mu_\theta)$, and that the estimators $\hat{\theta}_n$ and $\hat{\theta}_{n,m}$ have similar properties. Inference techniques such as the method of simulated moments (McFadden, 1989) and indirect inference (Gouriéroux et al., 1993) often (implicitly) use estimators of this form, in which $\mathcal{D}$ defined as the weighted Euclidean distance between sample moments or summary statistics of $y_{1:n}$ and $z_{1:m}$, and the expectation in (2.2) is replaced with a Monte Carlo approximation.

### 2.1.3 Minimum Wasserstein estimation

In this work, we focus on minimum distance estimation with the Wasserstein distance. Let $\rho$ be a distance on the observation space $\mathcal{Y}$, and let $\mathcal{P}_p(\mathcal{Y})$ with $p \geq 1$ (e.g. $p = 1$ or 2) be the set of distributions $\mu \in \mathcal{P}(\mathcal{Y})$ with finite $p$-th moment, i.e. there exists $y_o \in \mathcal{Y}$ such that $\int_{\mathcal{Y}} \rho(y, y_o)^p d\mu(y) < \infty$. The $p$-Wasserstein distance, also called the Monge-Kantorovich, Mallows, or Gini distance, is a finite metric on $\mathcal{P}_p(\mathcal{Y})$, defined by the optimal transport problem

$$
\mathcal{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{Y} \times \mathcal{Y}} \rho(x, y)^p d\gamma(x, y),
$$

(2.3)

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathcal{Y} \times \mathcal{Y}$ with marginals $\mu$ and $\nu$ respectively; see Chapter 6 of Villani (2008) for a brief history of this distance and its central role in optimal transport.

A useful property of the Wasserstein distance is that it is well-defined for distributions with non-overlapping supports. This allows us to define the minimum Wasserstein estimator (MWE) of order $p$, denoted $\hat{\theta}_n$, by simply plugging $\mathcal{W}_p$ into (2.1) in place of $\mathcal{D}$. Some properties of the MWE have been studied in Bassetti et al. (2006), for well-specified models and i.i.d. data; we derive new results in Section 2.2.1 under weaker assumptions. We also propose the minimum expected Wasserstein estimator (MEWE), obtained by replacing $\mathcal{D}$ with $\mathcal{W}_p$ in (2.2) and denoted $\hat{\theta}_{n,m}$. We describe some of its theoretical properties in Section 2.2.2.
Variations of these estimators have recently been applied by for instance Arjovsky et al. (2017) and Genevay et al. (2018b). In the settings they consider, the models are likely to be misspecified, and are supported on low-dimensional manifolds that might not overlap with the support of the data-generating mechanism. While the Wasserstein distance is well-defined in that case, the KL divergence or the total variation are not. This motivates the study of minimum Wasserstein estimators for these settings.

2.2 Theoretical results

We prove the existence, measurability, and consistency of the MWE and MEWE under weak assumptions, allowing the model to be misspecified and to produce data with certain types of dependencies. Under stronger assumptions, we study the rate of convergence and the asymptotic distribution of the MWE when $d = 1$ and $p = 1$. Throughout, we compare our results to those of Bassetti et al. (2006) and Bassetti and Regazzini (2006).

Informally, the consistency of the MWE and MEWE can be understood as follows. Under some conditions, we expect $\hat{\mu}_n$ to converge to $\mu_*$, in the sense that $\mathcal{W}_p(\hat{\mu}_n; \mu_*) \to 0$ as $n \to \infty$. Consequently, the minimum of $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n; \mu_\theta)$ might converge to the minimum of $\theta \mapsto \mathcal{W}_p(\mu_*; \mu_\theta)$, denoted by $\theta_*$, assuming its existence and unicity. The same can be said for the minimum of $\theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,m})$, provided $m \to \infty$ also. The parameter $\theta_*$ is thus the limiting object of interest, also termed the estimand. Beyond its interpretation as the minimizer of $\theta \mapsto \mathcal{W}_p(\mu_*; \mu_\theta)$, this parameter would coincide to the data-generating parameter if we assume that the data are generated from the model. In the misspecified case, note that $\theta_*$ is not necessarily the parameter that minimizes $\text{KL}(\mu_* | \mu_\theta)$, which is the limit of the maximum likelihood estimator under standard regularity conditions.

2.2.1 Minimum Wasserstein estimator

Existence, measurability, and consistency

We first list assumptions on the data-generating process and on the model that are sufficient for the existence, measurability, and consistency for the MWE.
**Assumption 2.2.1.** The data-generating process is such that $\mathcal{W}_p(\hat{\mu}_n; \mu_*) \rightarrow O$, $\mathbb{P}$-almost surely as $n \rightarrow \infty$.

**Assumption 2.2.2.** The map $\theta \mapsto \mu_\theta$ is continuous in the sense that $p_\mathcal{H}(\theta_n, \theta) \rightarrow O$ implies $\mu_{\theta_n} \Rightarrow \mu_\theta$ as $n \rightarrow \infty$.

**Assumption 2.2.3.** For some $\varepsilon > 0$, the set $B_\varepsilon(\varepsilon) = \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_*, \mu_\theta) \leq \varepsilon_* + \varepsilon \}$ is bounded, where $\varepsilon_* = \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta)$.

**Theorem 2.2.1** (Existence and consistency of the MWE). Under Assumptions 2.2.1-2.2.3, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, 

$$\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \rightarrow \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta),$$

and there exists $n(\omega)$ such that for $n \geq n(\omega)$, the sets $\arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ are non-empty and form a bounded sequence with

$$\limsup_{n \rightarrow \infty} \arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta).$$

For a generic function $f$, let $\varepsilon-\arg \min_x f = \{ x : f(x) \leq \varepsilon + \inf_x f \}$. Theorem 2.2.1 also holds if one replaces $\arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ with $\varepsilon_n-\arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$, for any sequence $\varepsilon_n$ converging to zero. If $\hat{\theta}_* = \arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta)$ is unique, the result can be rephrased as $\hat{\theta}_n \rightarrow \hat{\theta}_*$ $\mathbb{P}$-almost surely.

The following theorem derives from a general result by Brown and Purves (1973) on the measurability of estimators defined as minimizers.

**Theorem 2.2.2** (Measurability of the MWE). Suppose that $\mathcal{H}$ is a $\sigma$-compact Borel measurable subset of $\mathbb{R}^d$. Under Assumption 2.2.2, for any $n \geq 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_n : \Omega \rightarrow \mathcal{H}$ that satisfies

$$\hat{\theta}_n(\omega) \in \begin{cases} 
\arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{if this set is non-empty,} \\
\varepsilon-\arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{otherwise.}
\end{cases}$$
Theorem 2.2.1 generalizes the results of Basset et al. (2006), where the model is assumed to be well-specified in the sense that $\mu_* \in \mathcal{M}$. Moreover, Theorem 2.2.1 allows for data-generating processes which do not produce independent data points. For instance, if the data form a stationary and ergodic time series whose marginal distribution has finite $p$-th moments, then Assumption 2.2.1 still holds. These and other sufficient conditions for Assumption 2.2.1 to be satisfied are elaborated upon in Appendix A. Theorem 2.2.2 is only a minor generalization of the result in Basset et al. (2006), where it is assumed that for each $n \geq 1$, $\arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is non-empty for almost every $\omega \in \Omega$. In the next section, this small modification also enables the direct application of results by Pollard (1980).

**Rate of Convergence and Asymptotic Distribution**

Under conditions guaranteeing the consistency of the minimum Wasserstein estimator, we study its rate of convergence and asymptotic distribution in the case where $p = 1$, $\mathcal{Y} = \mathbb{R}$, $\rho(x, y) = |x - y|$. Under this setup, it can be shown that $\mathcal{W}_p(\mu, \nu) = \int_0^1 |F^{-1}_\mu(s) - F^{-1}_\nu(s)| ds = \int_0^1 |F_\mu'(t) - F_\nu'(t)| dt$, where $F_\mu$ and $F_\nu$ denote the cumulative distribution functions (CDFs) of $\mu$ and $\nu$ respectively (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). Additionally, assume that $\mathcal{H}$ is endowed with a norm: $\rho_{\mathcal{H}}(\theta, \theta') = \|\theta - \theta'\|_{\mathcal{H}}$. We also require that $\theta_*$ is “well-separated”:

**Assumption 2.2.4.** For all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\inf_{\theta \in \mathcal{H} : \|\theta - \theta_*\|_{\mathcal{H}} \geq \varepsilon} \mathcal{W}_p(\mu_{\theta_*}, \mu_{\theta}) > \delta.$$

This assumption is commonly made in the asymptotic study of M-estimators; see e.g. Chapter 5 of Van der Vaart (2000) and Appendix A. We focus on the setting in which the model is well-specified, but also discuss some extensions to the misspecified setting in Section 2.2.1.

Our approach to derive asymptotic distributions follows Pollard (1980). Let $F_\theta$, $F_*$ and $F_n$ denote the CDFs of $\mu_\theta$, $\mu_*$ and $\hat{\mu}_n$ respectively. Informally speaking, we show that
\[ \sqrt{n} W_n(\hat{\mu}_n, \mu_\theta) \] can be approximated by
\[
\int_{\mathbb{R}} |\sqrt{n}(F_n(t) - F_\star(t)) - \langle \sqrt{n}(\theta - \theta_\star), D_{\theta_\star}(t) \rangle| \, dt
\]

near \( \theta_\star \) for some \( D_{\theta_\star} \in (L_1(\mathbb{R}))^d \), with \( \langle \theta, u \rangle = \sum_{i=1}^d \theta_i u_i \). Results by del Barrio et al. (1999) and Dede (2009) give conditions under which \( \sqrt{n}(F_n - F_\star) \) converges to a zero mean Gaussian process \( G_\star \) with known covariance structure, for both independent and certain classes of dependent data. Heuristically, the distribution of \( \sqrt{n}(\hat{\theta}_n - \theta_\star) \) is then close to that of \( \operatorname{argmin}_{u \in \mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t) \rangle| \, dt \). The required form of \( D_{\theta_\star} \) is given in the following assumption:

**Assumption 2.2.5.** There exists a non-singular \( D_{\theta_\star} \in (L_1(\mathbb{R}))^d \) such that
\[
\int_{\mathbb{R}} |F_\theta(t) - F_{\theta_\star}(t) - \langle \theta - \theta_\star, D_{\theta_\star}(t) \rangle| \, dt = o(||\theta - \theta_\star||_\mathcal{H}),
\]
as \( ||\theta - \theta_\star||_\mathcal{H} \to 0 \).

To provide some intuition into the nature of the “derivative” \( D_{\theta_\star} \), we consider the following simple example. Let \( \mu_\theta = \mathcal{N}(\theta, I) \) for \( \theta \in \mathbb{R} \), and \( \mu_\star = \mu_{\theta_\star} \) for some \( \theta_\star \). By Taylor expanding \( F_\theta(t) = \Phi(t - \theta) \) around \( \theta_\star \) (for fixed \( t \)), Assumption 2.2.5 can be shown to hold with \( D_{\theta_\star}(t) = -\phi(t - \theta_\star) \), where \( \Phi \) and \( \phi \) denote the CDF and density of a standard Gaussian variable, respectively. Next, we state a result that holds for a well-specified model producing i.i.d. data, and analogous results for misspecified models and certain types of dependent processes can be found in Appendix A.

**Theorem 2.2.3.** Suppose that \( Y_i \sim \mu_\star = \mu_{\theta_\star} \) i.i.d. with \( \theta_\star \) in the interior of \( \mathcal{H} \), and that \( \int_0^\infty \sqrt{\mathbb{P}(|Y_o| > t)} \, dt \leq \infty \). Suppose that Assumptions 2.2.1-2.2.5 hold and that the minimum of \( u \mapsto \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t) \rangle| \, dt \) is almost surely unique. Then, the MWE with \( p = 1 \) satisfies
\[ \sqrt{n}(\hat{\theta}_n - \theta_\star) \Rightarrow \operatorname{argmin}_{u \in \mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t) \rangle| \, dt, \]
as \( n \to \infty \), where \( G_* \) is a zero mean Gaussian process with

\[
\mathbb{E} G_*(s) G_*(t) = \min \{ F_*(s), F_*(t) \} - F_*(s) F_*(t).
\]

A similar statement for \( p = 2 \) can potentially be derived by considering the results of del Barrio et al. (2005). The condition \( \int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)} dt < \infty \) implies the existence of second moments, and is itself implied by the existence of moments of order \( 2 + \varepsilon \) for some \( \varepsilon > 0 \) (see e.g. Section 2.9 in Wellner and van der Vaart, 1996). The uniqueness assumption on the argmin in the limit can be relaxed by considering convergence to the entire set of minimizing values, as in Section 7 of Pollard (1980). Still, uniqueness can sometimes be established, using e.g. the results of Cheney and Wulbert (1969). This approach is taken by Bassetti and Regazzini (2006), who directly show that Theorem 2.2.3 holds when \( \mathcal{M} \) is a location-scale family supported on a bounded open interval. The existence and form of \( D_{\theta^*} \) can in many cases be derived if the model is differentiable in quadratic mean (Le Cam, 1970), which is elaborated upon in Appendix A. There, one can also find results to verify Assumptions 2.2.1 and 2.2.4. It can in some cases potentially be easier to verify the assumptions for a reparameterization of \( \theta \), say \( \phi = r(\theta) \). Provided that the theorem holds for \( \hat{\phi}_n \) and that the inverse map \( r^{-1} \) is differentiable, the limiting distribution of \( \hat{\theta}_n \) can be derived using a delta method argument.

Computing confidence intervals using the asymptotic distribution provided by Theorem 2.2.3 is hard, due in part to its dependence on unknown quantities. However, the existence of the limiting distribution is in itself sufficient to guarantee the asymptotic validity of appropriately constructed subsampling confidence intervals (Politis et al., 1999, Theorem 2.2.1). This also generalizes to settings with certain kinds of dependent data. Under slightly stronger assumptions, the closely related \( m \) out of \( n \) bootstrap produces asymptotically valid confidence intervals as well; see Bickel and Sakov (2008) and references therein. In the numerical experiments of Section 2.4, we find that the standard bootstrap (Efron and Tibshirani, 1994) works well in practice.

Theorem 2.2.3 also holds for approximations of the MWE, say \( \hat{\theta}_n \), provided that \( \hat{\theta}_n = \hat{\theta}_n + o_\mathbb{P}(1/\sqrt{n}) \), as can be seen from its proof. In light of the convergence of the MEWE to the MWE as \( m \to \infty \) established in Section 2.2.2, there exists a sequence \( m(n) \) (depending
on $\omega$) such that the associated MEWE $\hat{\theta}_{n,m(n)}$ satisfies the conclusion of Theorem 2.2.3.

EXTENSIONS

Under slightly stronger assumptions, Theorem 2.2.3 can be extended to the misspecified setting. In particular, suppose that there exists a neighborhood $N$ of $\theta_*$ and a constant $c > 0$ such that for any $\theta \in N$,

$$W_1(\mu_0, \mu_*) \geq W_1(\mu_{\hat{\theta}}_n, \mu_*) + c\|\theta - \theta_*\|_{\mathcal{H}}.$$ 

In the well-specified case, this property is implied by Assumption 2.2.5. Then, as elaborated upon in Appendix A, the minimum of $\theta \mapsto W_1(\hat{\mu}_n, \mu_0)$ is attained on the set $\mathcal{S}_n = \{\theta : \|\theta - \theta_*\|_{\mathcal{H}} \leq 4W_1(\hat{\mu}_n, \mu_*)/c\}$ with probability going to one. Since the conditions of Theorem 2.2.3 imply that $W_1(\hat{\mu}_n, \mu_*) = O_P(1/\sqrt{n})$, this in turn implies that $\|\hat{\theta}_n - \theta_*\|_{\mathcal{H}} = O_P(1/\sqrt{n})$ also. In other words, the minimum Wasserstein estimator retains its rate of convergence in the misspecified case.

To find its asymptotic distribution, one can observe that with probability going to one, the map $\theta \mapsto \sqrt{n}W_1(\hat{\mu}_n, \mu_\theta)$ can be approximated uniformly well over $\mathcal{S}_n$ by the map $\theta \mapsto \sqrt{n}\int_{\mathbb{R}}|F_n(t) - F_{\theta_*}(t) - \langle \theta - \theta_* , D_{\theta_*}(t) \rangle|dt$, which similarly achieves its minimum on $\mathcal{S}_n$. Therefore, as $n$ gets large, $\sqrt{n}(\hat{\theta}_n - \theta_*)$ behaves like a minimum of

$$u \mapsto \int_{\mathbb{R}}\sqrt{n}(F_n(t) - F_\theta(t)) + \sqrt{n}(F_\theta(t) - F_{\theta_*}(t)) - \langle u, D_{\theta_*}(t) \rangle|dt.$$ 

Under the conditions of Theorem 2.2.3, $\sqrt{n}(F_n - F_\theta)$ converges to $G_\theta$ in the sense of del Barrio et al. (1999). In turn, $\sqrt{n}(\hat{\theta}_n - \theta_*)$ should be distributed as the minimizer(s) of $u \mapsto \int_{\mathbb{R}}|G_\theta(t) + \sqrt{n}(F_\theta(t) - F_{\theta_*}(t)) - \langle u, D_{\theta_*}(t) \rangle|dt$ as $n$ grows. A technical complication arises since this function converges pointwise to infinity, and we therefore leave formal statements for Appendix A.

Extensions to cases with multivariate data are left for future research. It is unclear whether convergence to $\theta_*$ will occur at the same $\sqrt{n}$ rate in higher dimensions. This is because $E W_1(\hat{\mu}_n, \mu_*)$ is on the order of $n^{-1/d}$ whenever $\mu_*$ is absolutely continuous with respect to the Lebesgue measure and $d > 2p$ (see e.g. Weed and Bach, 2019, and references therein).
On the other hand, del Barrio and Loubes (2019) show, under some assumptions, that the 2-Wasserstein distance satisfies the following CLT:

$$\sqrt{n} \left( \mathcal{W}_2^2(\hat{\mu}_n; \mu_\theta) - \mathbb{E} \mathcal{W}_2^2(\tilde{\mu}_n; \mu_\theta) \right) \Rightarrow \mathcal{N} \left( 0, \sigma^2(\mu_*, \mu_\theta) \right),$$

where $\sigma^2(\mu_*, \mu_\theta)$ has a known form and the expectation is taken with respect to the observations $y_{i,n} \sim \mu_*^{(n)}$. Similar results are expected to hold for other $p$ also. It therefore seems likely that the distance(s) between the MWE and the minimizer(s) of $\theta \mapsto \mathbb{E} \mathcal{W}_2^2(\tilde{\mu}_n; \mu_\theta)$ converges to zero at the standard $\sqrt{n}$ rate. If these speculations hold true, one could interpret them in terms of a bias-variance trade-off: the bias would appear to be on the order of $n^{-1/4}$, whereas the variance is on the order of $n^{-1/2}$. However, note that the function $\theta \mapsto \mathbb{E} \mathcal{W}_2^2(\tilde{\mu}_n; \mu_\theta)$ depends only on population properties of $\mu_*^{(n)}$. As such, it is a reasonable alternative to the objective function $\theta \mapsto \mathcal{W}_2^2(\mu_*, \mu_\theta)$, and might still yield reasonable identification of the parameters. For instance, if the model is well-specified and Gaussian with $\theta$ being a location parameter, it seems likely that $\theta \mapsto \mathbb{E} \mathcal{W}_2^2(\tilde{\mu}_n; \mu_\theta)$ is minimized at $\theta_*$ for any $n$. It is therefore unclear whether the slow convergence rate of the bias would always be of practical concern.

### 2.2.2 Minimum expected Wasserstein estimator

**Existence, measurability, and consistency**

In order to show similar results for the MEWE as for the MWE, we introduce the following additional assumptions.

**Assumption 2.2.6.** For any $m \geq 1$, if $\rho_{\mathcal{H}}(\theta_n, \theta) \to o$, then $\mu_{\theta_n}^{(m)} \Rightarrow \mu_\theta^{(m)}$ as $n \to \infty$.

**Assumption 2.2.7.** If $\rho_{\mathcal{H}}(\theta_n, \theta) \to o$, then $\mathbb{E}_n \mathcal{W}_p^2(\mu_{\theta_n}; \hat{\mu}_{\theta,n}) \to o$ as $n \to \infty$.

Assumption 2.2.6 is slightly stronger than Assumption 2.2.2, stating that we not only need weak convergence of the “model” distributions $\mu_{\theta_n}$, but also of the sample distributions $\hat{\mu}_{\theta,n}^{(m)}$ for any $m \geq 1$. Assumption 2.2.7 is implied by $\sup_{\theta \in \mathcal{H}} \mathbb{E}_n \mathcal{W}_p^2(\mu_{\theta}; \hat{\mu}_{\theta,n}) \to o$, which in turn might hold when $\mathcal{H}$ is compact and the inequalities in Fournier and Guillin (2015) hold.
In the next result, we prove an analogous version of Theorem 2.2.1 for the MEWE as \( \min\{n, m\} \rightarrow \infty \). For simplicity, we write \( m \) as a function of \( n \) and require that \( m(n) \rightarrow \infty \) as \( n \rightarrow \infty \).

**Theorem 2.2.4 (Existence and consistency of the MEWE).** Under Assumptions 2.2.1-2.2.3 and 2.2.6-2.2.7, there exists a set \( E \subset \Omega \) with \( \mathbb{P}(E) = 1 \) such that, for all \( \omega \in E \),

\[
\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta, m(n)}) \rightarrow \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta),
\]

and there exists \( n(\omega) \) such that, for all \( n \geq n(\omega) \), the sets \( \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta, m(n)}) \) are non-empty and form a bounded sequence with

\[
\limsup_{n \rightarrow \infty} \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta, m(n)}) \subset \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta).
\]

**Theorem 2.2.5 (Measurability of the MEWE).** Suppose that \( \mathcal{H} \) is a \( \sigma \)-compact Borel measurable subset of \( \mathbb{R}^d \). Under Assumption 2.2.6, for any \( n \geq 1 \) and \( m \geq 1 \) and \( \varepsilon > 0 \), there exists a Borel measurable function \( \hat{\theta}_{n,m} : \Omega \rightarrow \mathcal{H} \) that satisfies

\[
\hat{\theta}_{n,m}(\omega) \in \begin{cases} 
\arg\min_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta, m}), & \text{if this set is non-empty,} \\
\varepsilon\text{-argmin}_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta, m}), & \text{otherwise.}
\end{cases}
\]

The results above appear to be the first of their kind for the MEWE.

**Convergence to the MWE**

The next result considers the case where the data are fixed, while \( m \rightarrow \infty \). It shows that the MEWE converges to the MWE, assuming the latter exists. Using the results of del Barrio and Loubes (2019) and references therein, one could potentially derive the rate of this convergence, which we leave for future work. We formulate the following additional assumption, in which the observed empirical distribution is kept fixed and \( \varepsilon_n = \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \).

**Assumption 2.2.8.** For some \( \varepsilon > 0 \), the set \( B_n(\varepsilon) = \{ \theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \leq \varepsilon_n + \varepsilon \} \) is bounded.
**Theorem 2.2.6** (MEWE converges to MWE as \( m \to \infty \)). Under Assumptions 2.2.2 and 2.2.6-2.2.8, then

\[
\inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n; \hat{\theta}_m) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n; \hat{\theta}_0),
\]

and there exists an \( \hat{m} \) such that, for all \( m \geq \hat{m} \), the sets \( \arg \min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n; \hat{\theta}_m) \) are non-empty and form a bounded sequence with

\[
\limsup_{m \to \infty} \arg \min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n; \hat{\theta}_m) \subset \arg \min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n; \hat{\theta}_0).
\]

### 2.3 Computational aspects

#### 2.3.1 Computing the Wasserstein distance

We recall some strategies to calculate or approximate the Wasserstein distance between empirical distributions. In the case where \( \mathcal{Y} \subset \mathbb{R} \), the exact computation is cheap, as the main computational task reduces to sorting the samples. However, in dimensions \( d > 1 \), the cost is in general expensive, which has motivated a rich literature on fast approximations (Peyré and Cuturi, 2019). We will write \( \mathcal{W}_p(\gamma_{1:n}, \gamma_{1:m}) \) for \( \mathcal{W}_p(\hat{\mu}_n, \hat{\nu}_m) \), where \( \hat{\mu}_n \) and \( \hat{\nu}_m \) stand for the empirical distributions \( n^{-1} \sum_{i=1}^n \delta_{y_i} \) and \( m^{-1} \sum_{j=1}^m \delta_{z_j} \). The Wasserstein distance then takes the form

\[
\mathcal{W}_p(\gamma_{1:n}, \gamma_{1:m}) = \inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^n \sum_{j=1}^m \rho(y_i, z_j) \gamma_{i,j}
\]

(2.4)

where \( \Gamma_{n,m} \) is the set of \( n \times m \) matrices with non-negative entries, columns and rows resp. summing to \( m^{-1} \) and \( n^{-1} \).

#### Exact computation

The formulation in (2.4) is a linear program, and can be solved with generic linear program solvers. However, specialized approaches can be more efficient. In the univariate case with \( \rho(x, y) = |x - y| \), the optimal transport coupling can be found by sorting the vectors \( y_{1:n} \) and \( z_{1:m} \) to get the collections of order statistics \( \{y(i)\}_{i=1}^n \) and \( \{z(i)\}_{j=1}^m \). Suppose that
\[ m = \ell n \text{ for some } \ell \geq 1. \] Then, the \( p \)-Wasserstein distance in (2.4) can be expressed as
\[
\mathcal{W}_p^n(y_{i:n}, z_{i:m}) = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{\ell} |y(i) - z(\ell(i-1)+j)|^p,
\]
which can be seen from the representation \( \mathcal{W}_p^n(\tilde{\mu}_n, \tilde{\nu}_m) = \int_{0}^{1} |F_{\mu}^{-1}(s) - F_{\nu}^{-1}(s)|^p \, ds \) (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). The cost of the Wasserstein distance computation is thus of order \( m \log m \) in the univariate setting. Note that, in some cases, the generation of \( m \) sorted observations can be done directly for a cost of order \( m \), for instance by generating already-sorted uniforms and applying a quantile function (Devroye, 1985). It should also be noted that the expression \( \mathcal{W}_p^n(\mu, \nu) = \int_{0}^{1} |F_{\mu}^{-1}(s) - F_{\nu}^{-1}(s)|^p \, ds \), in combination with a numerical integrator, could be used whenever the quantile functions of \( \mu \) and \( \nu \) are known (as in the g-and-k example of Section 2.4.1). In that case one can directly target the MWE with a numerical optimizer, as an alternative to computing the MEWE. The same is true if the CDFs are available, using the expression \( \mathcal{W}_c^n(\mu, \nu) = \int_{\mathbb{R}}|F_{\mu}(t) - F_{\nu}(t)| \, dt \) given in Section 2.2.1.

In multivariate settings, one can solve the problem in (2.4) using dual ascent methods (see e.g. Bertsimas and Tsitsiklis, 1997). This includes the Hungarian algorithm, applicable in the setting where \( m = n \), at a cost of order \( n^3 \). Other algorithms have a cost of order \( n^{3.5} \log(n \, C_n) \), with \( C_n = \max_{1 \leq i \leq n} \rho(y_i, z_j) \), and can therefore be more efficient when \( C_n \) is small (Burkard et al., 2009, Section 4.1.3). A practical alternative is the short-list method, derived from the network simplex algorithm, presented by Gottschlich and Schuhmacher (2014) and implemented in the transport R package (Schuhmacher et al., 2017). In general, simplex algorithms come without guarantees of polynomial running times, but Gottschlich and Schuhmacher (2014) show empirically that their method tends to have sub-cubic cost. When the cost of computing the Wasserstein distance exactly gets prohibitively large, we can resort to various approximations.

**Approximations**

In parallel with its increasing popularity as an inferential tool in statistics and machine learning, there has been fast growth in the number of algorithms that approximate the
Wasserstein distance at reduced computational costs. The book of Peyré and Cuturi (2019) provides an overview of many such methods. In particular, they provide a thorough discussion of the method introduced by Cuturi (2013), which regularizes the optimization problem in (2.4) using an entropic constraint. Specifically, the regularized version of (2.4) reads:

$$\gamma^\xi = \arg\min_{\gamma \in \Gamma_{s,m}} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(y_i, z_j)^p \gamma_{ij} + \xi \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} \log \gamma_{ij},$$

which includes a penalty on the entropy of $\gamma$. The regularized problem can be solved iteratively by Sinkhorn's algorithm (Cuturi, 2013) or iterative Bregman projections (Benamou et al., 2015) for a total cost of order $nm$. Define the dual-Sinkhorn divergence $S^\xi_p(y_{i:n}, z_{j:m}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(y_i, z_j)^p \gamma_{ij}^\xi$. If $\xi$ goes to zero, the dual-Sinkhorn divergence goes to the Wasserstein distance. If $\xi$ goes to infinity, it converges to the energy distance (Ramdas et al., 2017). Other fast approximations of the Wasserstein distance include Ye et al. (2017); Altschuler et al. (2017, 2018); Li et al. (2018).

In the case where $n = m$, computing the Wasserstein distance can be viewed as an assignment problem, which leads to other specialized approaches. For instance, Puccetti (2017) proposes a greedy algorithm based on swaps in the assignment, for a cost of $n^2$ per iteration. When a cost of order $nm$ or $n^2$ is too large, Bernton et al. (2019) propose a new distance generalizing the idea of sorting when $d > 1$. It consists in sorting samples according to their projection via the Hilbert space-filling curve and computing a distance analogous to the one in (2.5), for a computational cost of the order of $m \log m$. A similar idea underlies the sliced Wasserstein distance (Rabin et al., 2011; Bonneel et al., 2015), which can be estimated by projecting the data onto $L$ random lines, and by averaging the Wasserstein distances computed in the associated one-dimensional spaces, for a total cost on the order of $Lm \log m$.

2.3.2 Computing the estimators

The exact computation of the MWE and MEWE is in general intractable. This is also true when $W_p$ is substituted for any of its approximations mentioned above. However, we can envision various schemes to numerically approximate the estimators.
The calculation of the MEWE can be based on Monte Carlo approximation of the function \( \theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \tilde{\mu}_{\theta, m}) \) using synthetic samples generated given \( \theta \). Assume that a data set \( z_{i:m} \) can be sampled from \( \mu^{(m)}_\theta \) by setting \( z_{i:m} = g_m(u, \theta) \), where \( g_m \) is a deterministic function of the parameter \( \theta \) and \( u \) a random variable independent of \( \theta \). Then, the mean \( k^{-1} \sum_{i=1}^{k} \mathcal{W}_p(y_{i:n}; g_m(u^{(i)})) \), where the \( u^{(i)} \) are i.i.d. is a natural estimate of \( \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta, m}) \). In the limit \( k \to \infty \), \( k^{-1} \sum_{i=1}^{k} \mathcal{W}_p(y_{i:n}; g_m(u^{(i)})) \to \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta, m}) \) almost surely. Since this estimator is an average of i.i.d. random variables, the CLT indicates that the rate of convergence is \( \sqrt{k} \). Moreover, this approximation is a deterministic function of \( \theta \), which can be optimized with standard methods. In turn, this optimization step can be placed within a Monte Carlo Expectation-Maximization (MCEM) algorithm (Wei and Tanner, 1990), which would alternate between optimization of \( \theta \) and resampling of \( u^{(i)} \). Convergence results for such algorithms, as both the number of iterations and \( k \) go to infinity, are reviewed in Neath et al. (2013).

In practice, we are naturally constrained to finite values of \( m \) and \( k \). The incremental cost of increasing \( k \) is typically lower than that of increasing \( m \), due in part to the potential for parallelization when calculating the distances \( \mathcal{W}_p(y_{i:n}; g_m(\theta, u^{(i)})) \) for a given \( \theta \), and in part to the algorithmic complexity in \( m \), which is super-linear as described in the previous section. In the numerical experiments of Section 2.4, we found that \( m = 10^4 \) and \( k = 20 \) within a single iteration of MCEM yielded accurate estimators. That is, we draw \( u^{(i)} \) for \( i = 1, \ldots, k \) once and for all, and optimize over \( \theta \). We illustrate the effect of choosing different \( m \) and \( k \) in Section 2.4.3.

Several alternatives to the MCEM approach exist. An approach to computing the MEWE was proposed in Genevay et al. (2018b) based on the Sinkhorn divergence approximation to the Wasserstein distance. They derive gradients of \( S_p^k(y_{i:n}, g_m(u, \theta)) \) with respect to \( \theta \) while \( u \) is fixed, allowing for the application of stochastic gradient descent. In practice, the gradients can be computed with auto-differentiation. A method for computing the MWE was proposed by Chen and Li (2018), in which they pull back the 2-Wasserstein metric tensor in \( \mathcal{P}_2(\mathcal{Y}) \) to \( \mathcal{H} \), under which \( \mathcal{H} \) becomes a Riemannian manifold. In turn, this structure allows them to derive a novel gradient descent algorithm. Alternatively, in the spirit of Monte Carlo optimization, one can modify the sampling algorithms used for the approximate Bayesian computation (ABC) approach described by Bernton et al. (2019).
to approximate the MEWE. This has the benefit of not requiring the synthetic data to be
generated via a deterministic function $g_m$ with fixed-dimensional arguments. Related dis-
cussions can be found in Wood (2010); Rubio et al. (2013).

2.4 Illustrations

In Sections 2.4.1 and 2.4.2, we compute the MEWE in two well-specified models with in-
tractable likelihoods that produce i.i.d. data, taken from the ABC literature. We empirically
estimate the coverage of bootstrap confidence intervals for the data-generating parameter.
In Section 2.4.1, we also compute the MEWE in a setting where the data-generating pro-
cess produces a time series. In Section 2.4.3, we compare the distribution of the MEWE
with that of the maximum likelihood estimator (MLE) in a simple misspecified setting.
We also investigate the effect of $k$ and $m$ on the distribution of the approximate MEWE.
In Section 2.4.4, we highlight the robustness of this choice by considering a heavy-tailed
data-generating process for which the MLE is not consistent. Throughout the numerical
experiments, we have chosen $p = 1$, as this imposes minimal assumptions on the existence
of moments of both the data-generating process and the model.

2.4.1 Quantile "g-and-k" distribution

Independent data

The g-and-κ distribution (Tukey, 1977; Jorge and Boris, 1984) is defined in terms of its
quantile function:

$$
r \in (0, 1) \mapsto a + b \left( 1 + 0.8 \frac{1 - \exp(-gz(r))}{1 + \exp(-gz(r))} \right) (1 + z(r)^\kappa) z(r),
$$

(2.7)

where $z(r)$ refers to the $r$-th quantile of the standard Normal distribution. The model
is indexed by the parameter $\theta = (a, b, g, \kappa) \in [0, 10]^4$, and we take $\mu_* = \mu_{\theta_*}$ with
$\theta_* = (3, 1, 2, 0.5)$. The probability density function, and therefore the likelihood of the
model, is analytically intractable; thus the model has become a standard benchmark for
ABC methods (Sisson et al., 2018). Though, the likelihood can be estimated by numeri-
cally inverting and then differentiating the quantile function, as described in Rayner and MacGillivray (2002); Bernton et al. (2019).

Sampling i.i.d. variables from the g-and-κ distribution can be achieved straightforwardly by plugging independent standard Normals into (2.7) in place of $z(r)$. Therefore, the MEWE with large $m$ can be computed to high precision. In Figure 2.4.1, we show the behavior of the MEWE with $p = 1$ and $m = 10^4$ for different numbers of observed data, and illustrate its concentration around the data-generating parameter $\theta_*$. In computing the MEWE, we used $k = 20$ and only one iteration of MCEM. That is, we approximate the MEWE by sampling $k = 20$ independent $u^{(i)}$ random variables and minimize $\theta \mapsto k^{-1} \sum_{i=1}^{k} \mathcal{W}_p(y_i, g_m(u^{(i)}(\theta)))$ to form the estimator, using the opt.im function in R (R Core Team, 2015).

We check the coverage of bootstrap confidence intervals calculated for $\theta_* = (3, 1, 2, 0.5)$. We use the percentile bootstrap (Efron and Tibshirani, 1994) for data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each of these. The observed coverage rates of the resulting 0.95 confidence intervals were 0.928 for $a$, 0.945 for $b$, 0.960 for $g$, and 0.938 for $\kappa$. The coverage rates should approach 0.95 as $n \to \infty$, $m \to \infty$, and $k \to \infty$ within the MCEM algorithm. After a Bonferroni correction, the observed coverage of the confidence sets for $\theta_*$ was 0.935.

As mentioned in Section 2.3.1, since the g-and-κ distribution has an explicit quantile function (insofar as the Normal quantile function can be considered explicit), one could instead directly estimate the Wasserstein distance between the g-and-k distribution and some empirical distribution using a representation of the distance in terms of an integral of the difference of quantile functions, combined with a numerical integrator.

Dependent data

To illustrate the behavior of the estimator when the data-generating process produces dependent data, we also generated g-and-κ variables using Normals from an AR(1) process. Specifically, we let $x_0 \sim \mathcal{N}(0, 1)$ and $x_t = \rho x_{t-1} + \eta_t$ for $t \geq 1$, where $\eta_t \sim \mathcal{N}(0, 1 - \rho^2)$ independently, and $\rho = 0.75$. Hence, these variables are marginally distributed as $\mathcal{N}(0, 1)$,
but are positively correlated. To produce the observation $y_i$ for each $t$, we plugged $x_t$ into (2.7) in place of $z(r)$, using the same $\theta_*$ as in the independent setting. The marginal distribution of the data are therefore the same as before, but the sequence of observations now forms a stationary and ergodic time series. This setting is covered by the theoretical results of Section 2.2; Assumption 2.2.1 holds with $\mu_* = \mu_{\scriptscriptstyle \hat{\theta}_*}$. The model, as before, is taken to generate i.i.d. data.

To approximate the MEWE, we used the same computational approach as in the i.i.d. setting, with $p = 1$, $m = 10^4$, and $k = 20$. In Figure 2.4.2, we show that the MEWE appears to concentrate around $\theta_*$ at the same rate as in the i.i.d. setting, but that its asymptotic distribution has higher variance. Note that in Figure 2.4.2, the data sizes are 10 times larger than in the plots for the i.i.d. setting (Figure 2.4.1), as the correlation between the samples effectively reduces the sample size and makes the estimators poorly behaved when $n$ is small.

### 2.4.2 Sum of log-Normal random variables

The distribution of the sum of log-Normal random variables appears in various settings (Fenton, 1960; Rodrigues et al., 2018), but no analytical formula is available for its probability density function, and thus the associated likelihood function is intractable. For a given positive integer $L$, $\gamma \in \mathbb{R}$ and $\sigma > 0$, the model generates an observation $y \in \mathbb{R}$ by sampling $x_1, \ldots, x_L \sim \mathcal{N}(\gamma, \sigma^2)$ independently, and defining $y = \sum_{i=1}^{L} \exp(x_i)$. Thus, sampling synthetic observations from the model is simple. We consider the task of estimating $\theta = (\gamma, \sigma)$ from data, fixing $L$ to 10, and using the MEWE. We generate $n$ observations independently using $\theta_* = (0, 1)$.

In Figure 2.4.3, we illustrate the behavior of the MEWE with $p = 1$ and $m = 10^4$ for different sizes of observed data $n$. The sampling distribution of the MEWE appears to concentrate around the data-generating parameter $\theta_*$ at the $\sqrt{n}$ rate as $n$ increases. In computing the MEWE, we used $k = 20$ and one iteration of MCEM as in the previous section.

We estimate the coverage of bootstrap confidence intervals calculated for $\theta_* = (0, 1)$. As before, we use the percentile bootstrap (Efron and Tibshirani, 1994) for data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. 26
We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each. The observed coverage rates were 0.945 and 0.940 for $\gamma_*$ and $\sigma_*$ respectively, which are close to the limiting 0.95 coverage rates. After a Bonferroni correction, the observed coverage of the confidence sets for $\theta_*$ was 0.960.

2.4.3 Gamma data fitted with a Normal model

We now consider a mis-specified setting. Let $\mu_*$ be a Gamma(10, 5) distribution (parametrized by shape and rate) and $\mathcal{M} = \{N(\gamma, \sigma^2) : \gamma \in \mathbb{R}, \sigma > 0\}$. The Normal location-scale model is very simple, yet it is widely used in practice in the form of regression models. Figure 2.4.4 compares the sampling distributions of the maximum likelihood estimator and approximations of the MEWE of order 1, over $M = 1,000$ experiments, for different values of $n$. The MEWE converges at the same $\sqrt{n}$ rate as the MLE, albeit to a distribution that is centered at a different location. Therefore, despite both estimation techniques leading to similar values for $\gamma$ and $\sigma$, the distributions of the estimators have very little overlap for large $n$, as observed in Figures 2.4.4c and 2.4.4d. For the MEWE, we have again used $m = 10^4$, $k = 20$, and one iteration of MCEM.

In Figure 2.4.5, we fix an observed data set of size $n = 100$, and compute $M = 500$ instances of the approximate MEWE for 8 different values of $k$ and $m$, ranging from 1 to 1,000 and 10 to 10,000 respectively. In Figure 2.4.5a, we plot the estimators obtained for all the levels of $k$, given 4 different values of $m$. In Figure 2.4.5b, we plot the estimators obtained for all the levels of $m$, given 4 different values of $k$. The axis scales are different for each subplot. In both figures, black points correspond to the “true” MWE, calculated using a very large value of $m$ ($m = 10^8$). For low values of $m$, the estimators might be significantly different from the MWE, as can be seen from the lower-right sub-plots of Figure 2.4.5b. When $m$ increases, the estimators converge to the MWE. Increasing $k$ reduces variation in the estimator. The changes in $k$ and $m$ had no significant impact on the number of evaluations of the objective required to locate the maximum using the `optim` function in R (R Core Team, 2015), which uses the Nelder–Mead simplex method (Nelder and Mead, 1965).
We check the coverage of bootstrap confidence intervals calculated for $\theta_*$ (itself calculated using $n = m = 10^8$ and $k = 1$). As before, we use the percentile bootstrap (Efron and Tibshirani, 1994) for data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each of these. The observed coverage rates of the resulting 0.95 confidence intervals were 0.960 and 0.953 for $\gamma_*$ and $\sigma_*$ respectively. After a Bonferroni correction, the observed coverage rate of the confidence sets for $\theta_* = (\gamma_*, \sigma_*)$ was 0.955.

2.4.4 Cauchy data fitted with a Normal model

Let $\mu_*$ be Cauchy with median zero and scale one, and consider the model $\mathcal{M} = \{N(\gamma, \sigma^2) : \gamma \in \mathbb{R}, \sigma > 0\}$. We explore the behavior of the MEWE of order 1, over $M = 1,000$ repeated experiments. Figure 2.4.6 shows its sampling distributions, for $n$ ranging from 50 to $10^4$. The marginal distribution of the estimator of $\gamma$ concentrates around 0, the median of $\mu_*$. The marginal distribution of the estimator of $\sigma$ also concentrates to a value close to 2.2. The concentration appears to occur at rate $\sqrt{n}$, as shown by the marginal densities of the rescaled estimators of $\gamma$ and $\sigma$ in Figures 2.4.6a and 2.4.6b.

In this setting the maximum likelihood estimator would not converge as $n \to \infty$, as the maximum likelihood estimator for $\gamma$ is the sample average, and the sample average of independent Cauchy variables is also Cauchy, with the same location and scale. As an alternative, we consider an estimator defined by minimizing a sample based estimator of the Kullback-Leibler divergence between $\mu_0$ and $\mu_*$. For the KL approximation we use the function KL.divergence in the FNN package (Beygelzimer et al., 2013), which approximates the KL divergence using $\ell$-nearest neighbor estimates described in Boltz et al. (2009) (and using the default parameter $\ell = 5$). The resulting estimator is termed the minimum KL estimator (MKLE), and is a variation of the MDEs discussed by Basu et al. (2011). We compute it using the same approach as for the MEWE, using $k = 20$, $m = 10^4$, and one iteration of MCEM. For $n = 5,000$ the distributions of MEWEs and MKLEs are plotted in Figures 2.4.6c and 2.4.6d. Both estimators appear to be robust in the sense that they converge to well-defined limits, unlike the MLE approach. The estimates of $\gamma$ are concen-
trated around 0, but the estimators of $\sigma$ are concentrated around two different values: the MEWEs seem to concentrate around 2.15 and the MKLEs around 1.65. The marginal distributions of the MEWE appear to have slightly smaller variance than those of the MKLE.

Note that this example is not covered by the theoretical results of Section 2.2 since the Cauchy distribution does not have a finite first moment. Robustness properties of general minimum distance estimators are discussed in Parr and Schucany (1980), and of the MWE in location models in Bassetti and Regazzini (2006). In the location-scale model considered here, if the approximation of the MEWE is computed with $k = 1$ and $m = \ell n$ for some $\ell \geq 1$, it can be written

$$\arg\min_{\gamma, \sigma} \sum_{i=1}^{n} \sum_{j=1}^{\ell} |y(i) - (\sigma x_{(\ell(i-1)+j)} + \gamma)|.$$  

As such, the approximate MEWE can be seen as the coefficients in a median regression (Koenker and Hallock, 2001) of a vector $\tilde{Y}$ on a vector $\tilde{X}$, where $\tilde{Y}_{\ell(i-1)+j} = y(i)$ for each $i = 1, \ldots, n$, and $\tilde{X}$ contains the order statistics of an $m$-sample of $\mathcal{N}(0, 1)$ random variables. Quantile regression is often presented as a robust alternative to linear regression in the presence of outliers, and further connections might explain the observed robustness of the MEWE with $p = 1$ in this example.

2.5 Discussion

The minimum Wasserstein (or Kantorovich) estimation approach (Bassetti et al., 2006) has received a renewed attention, due to recent advances in the field of computational optimal transport (Peyré and Cuturi, 2019), along with various applications in machine learning. In the broad context of generative models, these estimators present various appeals compared to maximum likelihood estimators. For instance, in Sections 2.4.1 and 2.4.2, we have observed the satisfactory behavior of minimum expected Wasserstein estimators in models where the likelihood function is not analytically available. In Sections 2.4.3 and 2.4.4 we have observed similarities and differences between MEWE and MLE in misspecified settings, illustrating some robustness properties of minimum Wasserstein estimation.
Minimum distance estimators were originally developed for obtaining almost surely convergent estimators (Wolfowitz, 1957), and we have showed that both the MWE and MEWE have this strong consistency property under mild conditions. We have also proved that the MWE converges to \( \theta_0 \), at the optimal \( \sqrt{n} \) convergence rate when the observations are univariate, and have derived its asymptotic distribution. The generalization of this result to multivariate data is left for future research. Interestingly, given the known convergence properties of the Wasserstein distance, it seems reasonable to conjecture that the rate of the MWE depends (negatively) on the dimension of the observation space rather than that of the parameter space. Other topics for future research include a more general derivation of the limiting distributions of the estimators, whose existence is needed to justify the asymptotic coverage of subsampling confidence intervals, as well as the development of a better understanding of their robustness properties.
Figure 2.4.1: Estimators in the well-specified g-and-κ model, as described in Section 2.4.1. Figures 2.4.1a and 2.4.1b show the MEWE’s bivariate marginal sampling distributions for \((a, b)\) and \((g, \kappa)\) respectively, as \(n\) ranges from 50 to 10^4 (colors from red to white to blue as \(n\) increases). For each \(n\), we plot \(M = 1,000\) estimators based on independent data sets. Each estimator was computed with \(p = 1\), \(m = 10^4\), \(k = 20\), and one iteration of MCEM. Note that for small data sizes (\(n = 50\) and \(n = 100\)), the estimator occasionally appears to be on the boundary of the parameter space, which could mean that the optimization procedure failed to converge. The intersections of the black lines indicate data-generating parameters. Figure 2.4.1c shows the MEWE’s marginal distribution for \(\kappa\) for the different levels of \(n\), centered and rescaled by \(\sqrt{n}\), illustrating the rate of convergence anticipated by Theorem 2.2.3. Figure 2.4.1d is a histogram of a data set generated with \(\theta_* = (3, 1, 2, 0.5)\) and \(n = 1,000\).
Figure 2.4.2: Estimators in the g-and-κ model with dependent data, as described in Section 2.4.1. Figures 2.4.2a and 2.4.2b show the MEWE’s bivariate marginal sampling distributions for \( (a, b) \) and \( (g, \kappa) \) respectively, as \( n \) ranges from \( 500 \) to \( 10^5 \) (colors from red to white to blue as \( n \) increases). Note that the sample sizes here are 10 times larger than in the plots for the i.i.d. setting. For each \( n \), we plot \( M = 1,000 \) estimators based on independent data sets. Each estimator was computed with \( p = 1, m = 10^4, k = 20 \), and one iteration of MCEM. The intersections of the black lines indicate data-generating parameters. Figure 2.4.2c shows the MEWE’s marginal distribution for \( \kappa \) for the different levels of \( n \), centered and rescaled by \( \sqrt{n} \), illustrating the rate of convergence anticipated by Theorem 2.2.3, but that the asymptotic variance is larger than in the i.i.d. case. Figure 2.4.2d shows the autocorrelation function of a data set generated with \( \theta_* = (3, 1, 2, 0.5) \), \( \rho = 0.75 \), and \( n = 1,000 \).
Figure 2.4.3: Estimators in the well-specified sum of log-Normals model, as described in Section 2.4.2. Figure 2.4.3a shows the sampling distributions of the MEWE, as \( n \) ranges from 50 to 10,000 (colors from red to white to blue as \( n \) increases). For each \( n \), we plot \( M = 1,000 \) estimators based on independent data sets. Each estimator was computed with \( p = 1 \), \( m = 10^4 \), \( k = 20 \), and one iteration of MCEM. The intersections of the black lines indicate data-generating parameters. Figures 2.4.3b and 2.4.3c show the MEWE’s marginal distributions for the different levels of \( n \), centered and rescaled by \( \sqrt{n} \), illustrating the rate of convergence anticipated by Theorem 2.2.3. Figure 2.4.3d is a histogram of a data set generated with \( \theta_* = (0, 1) \) and \( n = 1,000 \).
**Figure 2.4.4:** Gamma data fitted with a Normal model, as described in Section 2.4.3. Figures 2.4.4a and 2.4.4b show the sampling distributions of the MLE and MEWE of order 1 respectively, as \( n \) ranges from \( 50 \) to \( 10^4 \) (colors from red to white to blue). Figures 2.4.4c and 2.4.4d show the marginal densities of the estimators of \( \gamma \) and \( \sigma \) respectively, for \( n = 10^4 \); the MLEs are shown in dashed lines and the MEWE in full lines. For the MEWE, we have used \( m = 10^4 \), \( k = 20 \) and one iteration of MCEM.
(a) Approximate MEWE for increasing $k$ (colors from red to white to blue as $k$ increases), for different values of $m$.

(b) Approximate MEWE for increasing $m$ (colors from red to white to blue as $m$ increases), for different values of $k$.

Figure 2.4.5: Gamma data with $n = 100$, fitted with a Normal model, as described in Section 2.4.3. MEWEs are obtained for different values of $m$ (from 10 to 10,000) and $k$ (from 1 to 1,000), using one iteration of MCEM, $M = 500$ times independently. The intersections of the black lines represent the location of the “exact” MWE computed with $n = m = 10^8$. 
Figure 2.4.6: Cauchy data fitted with a Normal model, as described in Section 2.4.4. Marginals distributions of the MEWE of $\gamma$ and $\sigma$, centered by $\theta^*$ itself computed with $n = m = 10^5$, and rescaled by $\sqrt{n}$, are shown in Figures 2.4.6a and 2.4.6b. Figures 2.4.6c and 2.4.6d show the distributions of the MEWE for $n = 5,000$ (full lines), along with the distribution of an estimator obtained by minimizing an estimate of the Kullback–Leibler divergence (dashed lines).
3

Approximate Bayesian computation with the Wasserstein distance

3.1 Introduction

The likelihood function plays a central role in modern statistics. However, for many models of interest, the likelihood cannot be numerically evaluated. It might still be possible to simulate synthetic data sets from the model given parameters. A popular approach to Bayesian inference in such generative models is approximate Bayesian computation (ABC, Beaumont et al., 2002; Marin et al., 2012). ABC constructs an approximation of the posterior distribution by simulating parameters and synthetic data sets, and retaining the parameters such that the associated data sets are similar enough to the observed data set. Measures of similarity between data sets are often based on summary statistics, such as sample moments. In other words, data sets are considered close if some distance between their
summaries is small. The resulting ABC approximations have proven extremely useful, but can lead to a systematic loss of information compared to the original posterior distribution.

We propose here to instead view data sets as empirical distributions and to rely on the Wasserstein distance between synthetic and observed data sets. The Wasserstein distance, also called the Gini, Mallows, or Kantorovich distance, defines a metric on the space of probability distributions, and has become increasingly popular in statistics and machine learning, due to its appealing computational and statistical properties (e.g. Cuturi, 2013; Srivastava et al., 2015; Sommerfeld and Munk, 2018; Panaretos and Zemel, 2019). We will show that the resulting ABC posterior, which we term the Wasserstein ABC (WABC) posterior, can approximate the posterior distribution arbitrarily well in the limit of the threshold $\varepsilon$ going to zero, while bypassing the choice of summaries. Furthermore, we derive asymptotic settings under which the WABC posterior behaves differently from the posterior, illustrating the potential impact of model misspecification and the effect of the dimension of the observation space, by providing upper bounds on concentration rates as the number of observations goes to infinity. The WABC posterior is a particular case of coarsened posterior, and our results are complementary to those of Miller and Dunson (2018).

We further develop two strategies to deal with the specific case of time series. The challenge is that the marginal empirical distributions of time series might not contain enough information to identify all model parameters. In the first approach, which we term curve matching, each data point is augmented with the time at which it was observed. A new ground metric is defined on this extended observation space, which in turn allows for the definition of a Wasserstein distance between time series, with connections to Thorpe et al. (2017). A tuning parameter $\lambda > 0$ allows the proposed distance to approximate the Euclidean distance as $\lambda \to \infty$, and the Wasserstein distance between the marginal distributions as $\lambda \to 0$. The second approach involves transforming the time series such that its empirical distribution contains enough information for parameter estimation. We refer to such transformations as reconstructions and discuss delay reconstructions, as studied in dynamical systems (Stark et al., 2003), and residual reconstructions, as already used in ABC settings (Mengersen et al., 2013).

The calculation of Wasserstein distances is fast for empirical distributions in one dimen-
tion, as the main computational task reduces to sorting. For multivariate data sets, we can leverage the rich literature on the computation of Wasserstein distances and approximations thereof (Peyré and Cuturi, 2019). We also propose a new distance utilizing the idea of sorting, termed the Hilbert distance, based on the Hilbert space-filling curve (Sagan, 1994; Gerber and Chopin, 2015). The proposed distance approximates the Wasserstein distance well in low dimensions, but can be computed faster than the exact distance. We also shed light on some theoretical properties of the resulting ABC posterior.

In the following subsections we set up the problem we consider in this work, and briefly introduce ABC and the Wasserstein distance; we refer to Marin et al. (2012) and to Villani (2008) and Santambrogio (2015) for more detailed presentations of ABC and the Wasserstein distance, respectively.

3.1.1 Set-up, notation, and generative models

Throughout this chapter we consider a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), with associated expectation operator \(\mathbb{E}\), on which all the random variables are defined. The set of probability measures on a space \(\mathcal{X}\) is denoted by \(\mathcal{P}(\mathcal{X})\). The data take values in \(\mathcal{Y}\), a subset of \(\mathbb{R}^d\) for \(d \in \mathbb{N}\). We observe \(n \in \mathbb{N}\) data points, \(y_{1:n} = y_1, \ldots, y_n\), that are distributed according to \(\mu_s^{(n)} \in \mathcal{P}(\mathcal{Y}^n)\). Let \(\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i}\), where \(\delta_y\) is the Dirac distribution with mass on \(y \in \mathcal{Y}\). With a slight abuse of language, we refer below to \(\hat{\mu}_n\) as the empirical distribution of \(y_{1:n}\) even in the presence of non i.i.d. observations.

Formally, a model refers to a collection of distributions on \(\mathcal{Y}^n\), denoted by \(\mathcal{M}^{(n)} = \{\mu_{\theta}^{(n)} : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y}^n)\), where \(\mathcal{H} \subset \mathbb{R}^d\) is the parameter space, endowed with a distance \(\rho_{\mathcal{H}}\) and of dimension \(d_{\theta} \in \mathbb{N}\). However, we will often assume that the sequence of models \((\mathcal{M}^{(n)})_{n \geq 1}\) is such that, for every \(\theta \in \mathcal{H}\), the sequence \((\hat{\mu}_{\theta,n})_{n \geq 1}\) of random probability measures on \(\mathcal{Y}\) converges (in some sense) to a distribution \(\mu_{\theta} \in \mathcal{P}(\mathcal{Y})\), where \(\hat{\mu}_{\theta,n} = n^{-1} \sum_{i=1}^n \delta_{z_i}\), with \(z_{1:n} \sim \mu_{\theta}^{(n)}\). Similarly, we will often assume that \(\hat{\mu}_n\) converges to some distribution \(\mu_* \in \mathcal{P}(\mathcal{Y})\) as \(n \to \infty\). Whenever the notation \(\mu_*\) and \(\mu_{\theta}\) is used, it is implicitly assumed that these objects exist. In such cases, we instead refer to \(\mathcal{M} = \{\mu_{\theta} : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y})\) as the model. We say that it is well-specified if there exists \(\theta_* \in \mathcal{H}\) such that \(\mu_* = \mu_{\theta_*}\); otherwise it is misspecified. Parameters are identifiable if \(\theta = \theta'\) is implied
by $\mu_{\theta} = \mu_{\theta'}$.

We consider parameter inference for purely generative models: it is possible to generate observations $z_{1:n}$ from $\mu_{\theta}^{(n)}$, for all $\theta \in \mathcal{H}$, but it is not possible to numerically evaluate the associated likelihood. In some cases, observations from the model are obtained as $z_{1:n} = g_n(u, \theta)$, where $g_n$ is a known deterministic function and $u$ some known fixed-dimensional random variable independent of $\theta$. Some methods require access to $g_n$ and $u$ (Prangle et al., 2016; Graham and Storkey, 2017); by contrast, here we do not place assumptions on how data sets are generated from the model.

### 3.1.2 Approximate Bayesian Computation

Let $\pi$ be a prior distribution on the parameter $\theta$. Consider the following algorithm, where $\varepsilon > 0$ is referred to as the threshold, and $\mathcal{D}$ denotes a discrepancy measure between two data sets $y_{1:n}$ and $z_{1:n}$, taking non-negative values.

1. Draw a parameter $\theta$ from the prior distribution $\pi$, and a synthetic dataset $z_{1:n} \sim \mu_{\theta}^{(n)}$.

2. If $\mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon$, keep $\theta$, otherwise reject it.

The accepted samples are drawn from the ABC posterior distribution

$$
\pi^{(n)}_{\varepsilon} (d\theta) = \frac{\pi(d\theta) \int_{\mathcal{Y}} 1(\mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon) \mu_{\theta}^{(n)} (dz_{1:n})}{\int_{\mathcal{H}} \pi(d\theta) \int_{\mathcal{Y}} 1(\mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon) \mu_{\theta}^{(n)} (dz_{1:n})},
$$

(3.1)

where $1$ is the indicator function. A more sophisticated algorithm to approximate ABC posteriors, which we will apply in our numerical experiments, is described in Section 3.2.1.

Let $\rho$ be a distance on the observation space $\mathcal{Y}$, referred to as the ground distance. Suppose that $\mathcal{D}$ is chosen as

$$
\mathcal{D}(y_{1:n}, z_{1:n}) \rho = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, z_i)^p.
$$

(3.2)

Then, the resulting ABC posterior can be shown to have the desirable theoretical property of converging to the standard posterior as $\varepsilon \to 0$ (Prangle et al., 2016, see also Proposition

40
3.3.1. In the case where \( p = 2 \), \( Y \subset \mathbb{R} \), and \( \rho(y_i, z_i) = |y_i - z_i| \), \( D \) is a scaled version of the Euclidean distance between the vectors \( y_{1:n} \) and \( z_{1:n} \).

However, this approach is in most cases impractical due to the large variation of \( D(y_{1:n}, z_{1:n}) \) over repeated samples from \( \mu^{(n)}_\theta \). A rare example of practical use of ABC with the Euclidean distance is given in Sousa et al. (2009). A large proportion of the ABC literature is devoted to studying ABC posteriors in the setting where \( D \) is the Euclidean distance between summaries, i.e. \( D(y_{1:n}, z_{1:n}) = ||\eta(y_{1:n}) - \eta(z_{1:n})||_\eta \), where \( \eta : \mathbb{R}^n \rightarrow \mathbb{R}^{d_\eta} \) for some small \( d_\eta \).

Using summaries can lead to a loss of information: the resulting ABC posterior converges, at best, to the conditional distribution of \( \theta \) given \( \eta(y_{1:n}) \), as \( \varepsilon \rightarrow 0 \). A trade-off ensues, where using more summaries reduces the information loss, but increases the variation in the distance over repeated model simulations (Fearnhead and Prangle, 2012).

3.1.3 Wasserstein distance

A natural approach to reducing the variance of the distance defined in (3.2), while hoping to avoid the loss of information incurred by the use of summary statistics, is to instead consider the distance

\[
W_p(y_{1:n}, z_{1:n}) = \inf_{\pi \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \rho(y_{\pi(i)}, z_{\pi(i)})^p, \tag{3.3}
\]

where \( \mathcal{S}_n \) is the set of permutations of \( \{1, \ldots, n\} \). Indeed, when the observations are univariate and \( \rho(y_i, z_i) = |y_i - z_i| \), the above infimum is achieved by sorting \( y_{1:n} \) and \( z_{1:n} \) in increasing order and matching the order statistics. Using order statistics as a choice of summary within ABC has been suggested multiple times in the literature, see e.g. Sousa et al. (2009); Fearnhead and Prangle (2012). It turns out that \( W_p(y_{1:n}, z_{1:n}) \) is the \( p \)-Wasserstein distance between the empirical distributions supported on the data sets \( y_{1:n} \) and \( z_{1:n} \). From this perspective, our proposal of using the Wasserstein distance between empirical distributions can be thought of as generalizing the use of order statistics within ABC to arbitrary dimensions.

More formally, let \( \mathcal{P}_p(Y) \) with \( p \geq 1 \) (e.g. \( p = 1 \) or 2) be the set of distributions \( \mu \in \mathcal{P}(Y) \) with finite \( p \)-th moment: there exists \( y_\circ \in Y \) such that \( \int_Y \rho(y, y_\circ)^p d\mu(y) < \infty \). The
space $\mathcal{P}_p(\mathcal{Y})$ is referred to as the $p$-Wasserstein space of distributions on $\mathcal{Y}$ (Villani, 2008). The $p$-Wasserstein distance is a finite metric on $\mathcal{P}_p(\mathcal{Y})$, defined by the transport problem

$$\mathcal{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{Y} \times \mathcal{Y}} \rho(x, y)^p \, d\gamma(x, y),$$

(3.4)

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathcal{Y} \times \mathcal{Y}$ with marginals $\mu$ and $\nu$ respectively; see the notes in Chapter 6 of Villani (2008) for a brief history of this distance and its central role in optimal transport.

As in (3.3), we also write $\mathcal{W}_p(y_{i:n}, z_{i:m})$ for $\mathcal{W}_p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n$ and $\hat{\nu}_m$ stand for the empirical distributions $n^{-1} \sum_{i=1}^n \delta_{y_i}$ and $m^{-1} \sum_{j=1}^m \delta_{z_j}$, respectively. In particular, the Wasserstein distance between two empirical distributions with unweighted atoms takes the form

$$\mathcal{W}_p(y_{i:n}, z_{i:m})^p = \inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^n \sum_{j=1}^m \rho(y_i, z_j)^p \gamma_{ij},$$

(3.5)

where $\Gamma_{n,m}$ is the set of $n \times m$ matrices with non-negative entries, columns summing to $m^{-1}$, and rows summing to $n^{-1}$. We focus on the case $n = m$, for which it is known that the solution to the optimization problem, $\gamma^*$, corresponds to an assignment matrix with only one non-zero entry per row and column, equal to $n^{-1}$ (see e.g. the introductory chapter in Villani, 2003). In this special case, the Wasserstein distance can thus be represented as in (3.3). Computing the Wasserstein distance between two samples of the same size can therefore also be thought of as a matching problem; see Section 3.2.3.

### 3.1.4 Related works and plan

The minimum Wasserstein estimator (MWE), defined as $\hat{\theta}_n = \text{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta)$, was first studied by Bassetti et al. (2006) and is an example of a minimum distance estimator (Basu et al., 2011). To make this approach applicable in generative models, Bernton et al. (2017) introduce the minimum expected Wasserstein estimator (MEWE), defined as $\hat{\theta}_{n,m} = \text{argmin}_{\theta \in \mathcal{H}} \mathbb{E} \left[ \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta,m}) \right]$, where the expectation refers to the distribution of $z_{i:m} \sim \mu^{(m)}_\theta$. General results on both the MWE and MEWE are obtained in the technical report of Bernton et al. (2017); see also Chapter 2 of this manuscript. Another method
related to our approach was proposed by Park et al. (2016), who bypass the choice of summary statistics in the definition of the ABC posterior in (3.1) by using a discrepancy measure \( D \) such that \( D(y_{i:n}, z_{i:n}) \) is an estimate of the maximum mean discrepancy (MMD) between \( \hat{\mu}_n \) and \( \mu_0 \).

Our contributions are structured as follows: the proposed approach to Bayesian inference in generative models using the Wasserstein distance is described in Section 3.2, some theoretical properties of the Wasserstein ABC posterior is detailed in Section 3.3, methods to handle time series are proposed in Section 3.4, and numerical illustrations in Section 3.5, where in each example we make comparisons to existing methods, such as semi-automatic ABC (Fearnhead and Prangle, 2012). The code is available on GitHub at github.com/pierrejacob/winference. Appendix B contains additional theoretical results and details on computational aspects, as referenced in the present chapter.

3.2 Wasserstein ABC

The distribution \( \pi_{y_{i:n}}^\varepsilon (d\theta) \) of (3.1), with \( D \) replaced by \( \mathcal{W}_p \) for some choice of \( p \geq 1 \), is referred to as the Wasserstein ABC (WABC) posterior; that is, the WABC posterior is defined by

\[
\pi_{y_{i:n}}^\varepsilon (d\theta) = \frac{\pi(d\theta) \int_{\mathcal{Y}_n} 1 \left( \mathcal{W}_p(y_{i:n}, z_{i:n}) \leq \varepsilon \right) \mu_\theta^{(n)}(dz_{i:n})}{\int_{\mathcal{H}} \pi(d\theta) \int_{\mathcal{Y}_n} 1 \left( \mathcal{W}_p(y_{i:n}, z_{i:n}) \leq \varepsilon \right) \mu_\theta^{(n)}(dz_{i:n})}
\]

(3.6)

with \( \mathcal{W}_p(y_{i:n}, z_{i:n}) \) defined in (3.3). Throughout the experiments of this article we set \( p = 1 \), which makes minimal assumptions on the existence of moments of the data-generating process.

As mentioned in the introductory section, the motivation for choosing \( D \) to be the Wasserstein distance is to have a discrepancy measure \( D(y_{i:n}, z_{i:n}) \) that has both a small variance and results in an ABC posterior that has satisfactory theoretical properties. In particular, we show in Section 3.3 that, as per ABC based on the Euclidean distance, the WABC posterior converges to the true posterior distribution as \( \varepsilon \to 0 \). In that section, we also provide a result showing that, as \( n \to \infty \) and the threshold \( \varepsilon \) converges slowly enough to some minimal value \( \varepsilon_* \geq 0 \), the WABC posterior concentrates around \( \theta_* := \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*; \mu_\theta) \). In the well-specified case, \( \theta_* \) coincides with the data-
generating parameter. In the misspecified case, $\theta_*$ is typically different from where the actual posterior concentrates, which is around the minimizer of $\theta \mapsto KL(\mu^*|\mu_0)$, where KL refers to the Kullback–Leibler divergence. The experiments in Section 3.5 contain examples where the WABC posterior provides a practical and accurate approximation of the standard posterior, and examples where it does not, partly because of the computational difficulty of sampling from the WABC posterior when $\varepsilon$ is small.

### 3.2.1 Sampling sequentially from the WABC posterior

Instead of the rejection sampler of Section 3.1.2, we will target the WABC and other ABC posteriors using a sequential Monte Carlo (SMC) approach, with $N$ particles exploring the parameter space (Del Moral et al., 2012). The algorithm starts with a threshold $\varepsilon_0 = +\infty$, for which the WABC posterior is the prior. Given the Monte Carlo approximation of the WABC posterior for $\varepsilon_{t-1}$, the next value $\varepsilon_t$ is chosen so as to maintain a number of unique particles of at least $aN$, with $a \in (0, 1]$. Upon choosing $\varepsilon_t$, resampling and rejuvenation steps are triggered and the algorithm proceeds. In the experiments, we will run the algorithm until a fixed budget of model simulations is reached. At the end of the run, the algorithm provides $N$ parameter samples and synthetic data sets, associated with a threshold $\varepsilon_T$.

The algorithm is parallelizable over the $N$ particles, and thus over equally many model simulations and distance calculations. Any choice of MCMC kernel can be used within the rejuvenation steps. In particular, we use the r-hit kernel of Lee (2012), shown to be advantageous compared to standard ABC-MCMC kernels in Lee and Łatuszyński (2014). We choose the number of hits to be 2 by default. For the proposals of the MCMC steps, we use a mixture of multivariate Normal distributions, with 5 components by default. We set $N$ to be 2, 048 and $a$ to be 50%. These default tuning parameters are used throughout all the numerical experiments of Section 3.5, unless otherwise specified. Full details on the SMC algorithm are given in Appendix B.
3.2.2 Illustration on a Normal location model

Consider 100 i.i.d. observations generated from a bivariate Normal distribution. The mean components are drawn from a standard Normal distribution, and the generated values are approximately $-0.71$ and $0.09$. The covariance is equal to 1 on the diagonal and 0.5 off the diagonal. The parameter $\theta$ is the mean vector, and is assigned a centered Normal prior with variance 25 on each component.

We compare WABC with two other methods: ABC using the Euclidean distance between the data sets, and ABC using the Euclidean distance between sample means, which for this model are sufficient summary statistics. All three ABC posteriors are approximated using the SMC sampler described in Section 3.2.1. The summary-based ABC posterior is also approximated using the simple rejection sampler given in Section 3.1.2 to illustrate the benefit of the SMC approach. All methods are run for a budget of $10^6$ model simulations, using $N = 2,048$ particles in the SMC sampler. The rejection sampler accepted only the 2,048 draws yielding the smallest distances. Approximations of the marginal posterior distributions of the parameters are given in Figures 3.2.1a and 3.2.1b, illustrating that the SMC-based ABC methods with the Wasserstein distance and with sufficient statistics both approximate the posterior accurately.

To quantify the difference between the obtained ABC samples and the posterior, we again use the Wasserstein distance. Specifically, we independently draw 2,048 samples from the posterior distribution, and compute the Wasserstein distance between these samples and the $N = 2,048$ ABC samples produced by the SMC algorithm. We plot the resulting distances against the number of model simulations in Figure 3.2.1c, in log-log scale. As expected, ABC with sufficient statistics converges fastest to the posterior. It should be noted that sufficient statistics are almost never available in realistic applications of ABC. The proposed WABC approach performs almost as well, but requires more model simulations to yield comparable results. In contrast, the ABC approach with the Euclidean distance struggles to approximate the posterior accurately. Extrapolating from the plot, it would seemingly take billions of model simulations for the latter ABC approach to approximate the posterior as accurately as the other two methods. Similarly, despite being based on the sufficient statistic, the rejection sampler does not adequately estimate the posterior.
distribution for the given sample budged. The estimated 1-Wasserstein distance between the 2, 048 accepted samples and the posterior was 0.63.

In terms of computing time, based on our R implementation on an Intel Core i5 (2.5GHz), simulating a data set took on average 4.0 \times 10^{-5}s. Computing the discrepancy between data sets took on average 6.4 \times 10^{-5}s for the summary-based distance, 3.8 \times 10^{-4}s for the Euclidean distance, and 1.2 \times 10^{-5}s for the Wasserstein distance; see Section 3.2.3 for fast approximations of the Wasserstein distance. The SMC sampler is algorithmically more involved than the rejection sampler, and one could ask whether the added computational effort is justified. In this example, the total time required by the SMC algorithm using the summary statistic was 169s, whereas the analogous rejection sampler took 141s. This illustrates that even when one can very cheaply simulate data and compute distances, the added costs associated with an SMC sampler are relatively small; see Del Moral et al. (2012); Filippini et al. (2013); Sisson et al. (2018) for more details on SMC samplers for ABC purposes.

![Figure 3.2.1](image-url)

**Figure 3.2.1:** ABC in the bivariate Normal location model of Section 3.2.2. ABC approximations of the posterior after 10^6 model simulations (left and middle), overlaid the actual posterior. On the right, the Wasserstein distance between ABC posterior samples and exact posterior samples is plotted against the number of model simulations (in log-log scale). In principle, these ABC approximations converge to the posterior as \( \epsilon \to 0 \). Yet, for a given number of model simulations, the quality of the ABC approximation is sensitive to the choice of distance and sampling algorithm.
3.2.3 Computing and approximating the Wasserstein distance

Computing the Wasserstein distance between the distributions \( \hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i} \) and \( \hat{\nu}_n = n^{-1} \sum_{i=1}^n \delta_{z_i} \) reduces to a linear sum assignment problem, as in (3.3). In the univariate case, finding the optimal permutation can be done by sorting the vectors \( y_{1:n} \) and \( z_{1:n} \) in increasing order, obtaining the orders \( \sigma_y(i) \) and \( \sigma_z(i) \) for \( i \in \{1, \ldots, n\} \). Then, one associates each \( y_i \) with \( z_{\sigma(i)} \) where \( \sigma(i) = \sigma_z \circ \sigma_y^{-1}(i) \). The cost of the Wasserstein distance computation is thus of order \( n \log n \) for distributions on one-dimensional spaces.

In multivariate settings, (3.3) can be solved by the Hungarian algorithm for a cost of order \( n^3 \). Other algorithms have a cost of order \( n^{3.5} \log(n C_n) \), with \( C_n = \max_{1 \leq i \leq j \leq n} \rho(y_i, z_j) \), and can therefore be more efficient when \( C_n \) is small (Burkard et al., 2009, Section 4.1.3). In our numerical experiments, we use the short-list method presented in Gottschlich and Schuhmacher (2014) and implemented in Schuhmacher et al. (2017). This method is derived from the simplex algorithm and comes without guarantees of polynomial running times, but Gottschlich and Schuhmacher (2014) show empirically that their method tends to have sub-cubic cost.

The cubic cost of computing Wasserstein distances in the multivariate setting can be prohibitive for large data sets. However, many applications of ABC involve relatively small numbers of observations from complex models which are expensive to simulate. In these settings, the cost of simulating synthetic data sets might dominate the model-free cost of computing distances. Note also that the dimension \( d_y \) of the observation space only enters the ground distance \( \rho \), and thus the cost of computing the Wasserstein distance under a Euclidean ground metric is linear in \( d_y \).

Fast approximations

In conjunction with its increasing popularity as a tool for inference in statistics and machine learning, there has been a rapid growth in the number of algorithms that approximate the Wasserstein distance at reduced computational costs; see Peyré and Cuturi (2019). In particular, they provide an in-depth discussion of the popular method proposed by Cuturi (2013), in which the optimization problem in (3.5) is regularized using an entropic
constraint on the joint distribution $\gamma$. Consider

$$
\gamma^\xi = \arg\min_{\gamma \in \Gamma_\alpha} \sum_{i,j=1}^{n} \rho(y_i, z_j)^x \gamma_{ij} + \xi \sum_{i,j=1}^{n} \gamma_{ij} \log \gamma_{ij},
$$

which includes a negative penalty on the entropy of $\gamma$, and define the dual-Sinkhorn divergence $S_{p}^{\xi}(y_{1:n}, z_{1:n})^x = \sum_{i,j=1}^{n} \rho(y_i, z_j)^x \gamma_{ij}^\xi$. The regularized problem can be solved iteratively by Sinkhorn’s algorithm, which involves matrix-vector multiplications resulting in a cost of order $n^2$ per iteration. If $\xi \to 0$, the dual-Sinkhorn divergence converges to the Wasserstein distance, whereas if $\xi \to \infty$ it converges to the maximum mean discrepancy (Ramdas et al., 2017). It can therefore be seen as an interpolation between optimal transport and kernel-based distances. Further properties of the dual-Sinkhorn divergence and other algorithms to approximate it are discussed in Peyré and Cuturi (2019).

Unlike the optimal coupling that yields the exact Wasserstein distance, the coupling obtained in the regularized problem is typically not an assignment matrix. In the following subsections, we discuss two simple approaches with different computational complexities that yield couplings that are assignments. This has the benefit of aiding the theoretical analysis in Section 3.3.

**Hilbert distance**

The assignment problem in (3.3) can be solved in $n \log n$ in the univariate case by sorting the samples. We propose a new distance generalizing this idea when $d_y > 1$, by sorting samples according to their projection via the Hilbert space-filling curve. As shown in Gerber and Chopin (2015) and Schretter et al. (2016), transformations through the Hilbert space-filling curve and its inverse preserve a notion of distance between probability measures. The Hilbert curve $H : [0, 1] \to [0, 1]^{d_y}$ is a Hölder continuous mapping from $[0, 1]$ into $[0, 1]^{d_y}$. One can define a measurable pseudo-inverse $h : [0, 1]^{d_y} \to [0, 1]$ verifying $h(H(x)) = x$ for all $x \in [0, 1]$ (Gerber et al., 2019). We assume in this subsection that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ is such that there exists a mapping $\psi : \mathcal{Y} \to (0, 1)^{d_y}$ verifying, for $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, $\psi(y) = \left(\psi_1(y_1), \ldots, \psi_{d_y}(y_{d_y})\right)$ where the $\psi_i$’s are continuous and strictly monotone. For instance, if $\mathcal{Y} = \mathbb{R}^{d_y}$, one can take $\psi$ to be the component-wise
logistic transformation; see Gerber and Chopin (2015) for more details. By construction, the mapping \( h_Y := h \circ \psi : \mathcal{Y} \rightarrow (0, 1) \) is one-to-one. For two vectors \( y_{1:n} \) and \( z_{1:n} \), denote by \( \sigma_y \) and \( \sigma_z \) the permutations obtained by mapping the vectors through \( h_Y \) and sorting the resulting univariate vectors in increasing order. We define the Hilbert distance \( \mathcal{H}_p \) between the empirical distributions of \( y_{1:n} \) and \( z_{1:n} \) by

\[
\mathcal{H}_p(y_{1:n}, z_{1:n})^p = \frac{1}{n} \sum_{i=1}^{n} \rho(y_{i}; z_{\sigma(i)})^p,
\]

where \( \sigma(i) = \sigma_z \circ \sigma_y^{-1}(i) \) for all \( i \in \{1, \ldots, n\} \).

**Proposition 3.2.1.** For any integer \( n \geq 1 \) and real number \( p \geq 1 \), \( \mathcal{H}_p \) defines a distance on the space of empirical distributions of size \( n \).

The Hilbert distance can be computed at a cost in the order of \( n \log n \) and an implementation is provided by the function hilbert_sort in The Computational Geometry Algorithms Library (2016). From a practical point of view, this implementation has the attractive property of not having to map the samples to \((0, 1)^d_y\) and hence having to choose a specific mapping \( \psi \). Instead, this function directly constructs the Hilbert curve around the input point set.

Despite not being defined in terms of a transport problem, the Hilbert distance yields approximations of the Wasserstein distance that are accurate for small \( d_y \), as illustrated in Appendix B. More importantly for its use within ABC, the level sets of the (random) map \( \theta \mapsto \mathcal{H}_p(\mu_n, \hat{\mu}_{\theta,n}) \) appear to be close to those of the analogous Wasserstein distance. The two distances therefore discriminate between parameters in similar fashions. However, this behavior tends to deteriorate as the dimension \( d_y \) grows.

The coupling produced by Hilbert sorting is feasible for the assignment problem in (3.3). Therefore, it is always greater than the Wasserstein distance, which minimizes the objective therein. This property plays an important role in showing that the ABC posterior based on the Hilbert distance concentrates on \( \theta \), as \( n \rightarrow \infty \) and the threshold \( \varepsilon \) decreases sufficiently slowly. In Appendix B, we provide such a result under the assumption that the model is well-specified, but leave further theoretical analysis under milder conditions for future research. Other one-dimensional projections of multivariate samples, followed by
Wasserstein distance computation using the projected samples, have been proposed in the computational optimal transport literature (Rabin et al., 2011; Bonneel et al., 2015), also leading to computational costs in $n \log n$.

**Swapping distance**

Viewing the Wasserstein distance calculation as the assignment problem in (3.3), Puccetti (2017) proposed a greedy swapping algorithm to approximate the optimal assignment. Consider an arbitrary permutation $\sigma$ of $\{1, \ldots, n\}$, and the associated transport cost $\sum_{i=1}^{n} \rho(y_i, z_{\sigma(i)})^p$. The swapping algorithm consists in checking, for all $1 \leq i < j \leq n$, whether $\rho(y_i, z_{\sigma(i)})^p + \rho(y_j, z_{\sigma(j)})^p$ is less or greater than $\rho(y_i, z_{\sigma(j)})^p + \rho(y_j, z_{\sigma(i)})^p$. If it is greater, then one swaps $\sigma(i)$ and $\sigma(j)$, resulting in a decrease of the transport cost. One can repeat these sweeps over $1 \leq i < j \leq n$, until the assignment is left unchanged, and denote it by $\tilde{\sigma}$. Each sweep has a cost of order $n^2$ operations. There is no guarantee that the resulting assignment $\tilde{\sigma}$ corresponds to the optimal one. Note that we initialize the algorithm with the assignment obtained by Hilbert sorting for a negligible cost of $n \log n$. We refer to the resulting distance $\left(n^{-1} \sum_{i=1}^{n} \rho(y_i, z_{\tilde{\sigma}(i)})^p\right)^{1/p}$ as the swapping distance.

The swapping distance between $y_{i:n}$ and $z_{i:n}$ takes values that are, by construction, between the Wasserstein distance $\mathcal{W}_p(y_{i:n}, z_{i:n})$ and the Hilbert distance $\mathcal{H}_p(y_{i:n}, z_{i:n})$. Thanks to this property, we show in Appendix B that the associated ABC posterior concentrates on $\theta_*$ as $n \to \infty$ and the threshold $\varepsilon$ decreases sufficiently slowly. As with the Hilbert distance, this result is obtained under the assumption that the model is well-specified and leave further theoretical analysis under milder conditions for future research. In Appendix B, we also observe that the swapping distance can approximate the Wasserstein distance more accurately than the Hilbert distance as the dimension $d_y$ grows.

**Sub-sampling**

Any of the aforementioned distances can be computed faster by first sub-sampling $m < n$ points from $y_{i:n}$ and $z_{i:n}$, and then computing the distance between the resulting distributions. This increases the variance of the calculated distances, introducing a trade-off with computation time. In the case of the Wasserstein distance, this approach could be studied
formally using the results of Sommerfeld and Munk (2018). Other multiscale approaches can also be used to accelerate computation (Mérigot, 2011). We remark that computing the distance between vectors containing subsets of order statistics (Fearnhead and Prangle, 2012) can be viewed as an example of a multiscale approach to approximating the Wasserstein distance.

**Combining distances**

It might be useful to combine distances. For instance, one might want to start exploring the parameter space with a cheap approximation, and switch to the exact Wasserstein distance in a region of interest; or use the cheap approximation to save computations in a delayed acceptance scheme. One might also combine a transport distance with a distance between summaries. We can combine distances in the ABC framework by introducing a threshold for each distance, and define the ABC posterior as in (3.1), with a product of indicators corresponding to each distance. We explore the combination of distances in the numerical experiments of Section 3.5.4.

**3.3 Theoretical properties**

We study the behavior of the Wasserstein ABC posterior under different asymptotic regimes. First, we give conditions on a discrepancy measure for the associated ABC posterior to converge to the posterior as the threshold $\varepsilon$ goes to zero, while keeping the observed data fixed. We then discuss the behavior of the WABC posterior as $n \to \infty$ for fixed $\varepsilon > 0$. Finally, we establish bounds on the rates of concentration of the WABC posterior as the data size $n$ grows and the threshold $\varepsilon$ shrinks sufficiently slowly at a rate dependent on $n$, similar to Frazier et al. (2018) in the case of summary-based ABC. Proofs are deferred to the appendix.

We remark that the assumptions underlying our results are typically hard to check in practice, due to the complexity and intractable likelihoods of the models to which ABC methods are applied. This is also true for state-of-the-art asymptotic results about summary-based ABC methods, which, for example, require injectivity and growth conditions on the “binding function” to which the summary statistics converge (Frazier et al., 2018).
Nonetheless, we believe that our results provide insight into the statistical properties of the WABC posterior. For instance, in Corollary 3.3.1 we give conditions under which the WABC posterior concentrates around $\theta_* = \arg\min_{\theta \in \mathcal{H}} W_p(\mu_\theta, \mu_*)$ as $n$ grows. When the model is misspecified, this is in contrast with the posterior, which is known to concentrate around $\arg\min_{\theta \in \mathcal{H}} \text{KL}(\mu_\theta, \mu_*)$ (see e.g. Müller, 2013).

3.3.1 Behavior as $\varepsilon \to 0$ for fixed observations

The following result establishes conditions under which a non-negative measure of discrepancy between data sets $\mathcal{D}$ yields an ABC posterior that converges to the true posterior as $\varepsilon \to 0$, while the observations are kept fixed.

**Proposition 3.3.1.** Suppose that $\mu^{(n)}_\theta$ has a continuous density $f^{(n)}_\theta$ and that

$$
\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_\mathcal{H}} f^{(n)}_\theta(y_{1:n}) < \infty,
$$

where $\mathcal{N}_\mathcal{H}$ is a set such that $\pi(\mathcal{N}_\mathcal{H}) = 0$. Suppose that there exists $\varepsilon > 0$ such that

$$
\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_\mathcal{H}} \sup_{z_{1:n} \in \mathcal{A}_\varepsilon} f^{(n)}_\theta(z_{1:n}) < \infty,
$$

where $\mathcal{A}_\varepsilon = \{z_{1:n} : \mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\}$. Suppose also that $\mathcal{D}$ is continuous in the sense that $\mathcal{D}(y_{1:n}, z_{1:n}) \to \mathcal{D}(y_{1:n}, x_{1:n})$ whenever $z_{1:n} \to x_{1:n}$ component-wise in the metric $\rho$. If either

1. $f^{(n)}_\theta$ is n-exchangeable, such that $f^{(n)}_\theta(y_{1:n}) = f^{(n)}_\theta(y_{\sigma(1:n)})$ for any $\sigma \in \mathcal{S}_n$, and $\mathcal{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, or

2. $\mathcal{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{1:n}$

then, keeping $y_{1:n}$ fixed, the ABC posterior converges strongly to the posterior as $\varepsilon \to 0$.

The Wasserstein distance applied to unmodified data satisfies $\mathcal{W}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, making condition (1) of Proposition 3.3.1 applicable. In Section 3.4, we will discuss two methods applicable to time series that lead to discrepancies for which condition (2) holds. Note that this result does not guarantee that
the Monte Carlo algorithm employed to sample the ABC posterior distribution, with an adaptive mechanism to decrease the threshold, will be successful at reaching low thresholds in a reasonable time.

3.3.2 Behavior as $n \to \infty$ for fixed $\varepsilon$

Under weak conditions, the WABC posterior distribution $\pi_{Y^n}(d\theta)$ in (3.1) converges to $\pi(d\theta | \mathcal{W}_p(\mu_\theta, \mu_\star) < \varepsilon)$ as $n \to \infty$ for a fixed threshold $\varepsilon$, following the reasoning in Miller and Dunson (2018) for general weakly-continuous distances, which include the Wasserstein distance. Therefore, the WABC distribution with a fixed $\varepsilon$ does not converge to a Dirac mass, contrarily to the standard posterior. As argued in Miller and Dunson (2018), this can have some benefit in case of model misspecification: the WABC posterior is less sensitive to perturbations of the data-generating process than the standard posterior.

3.3.3 Concentration as $n$ increases and $\varepsilon$ decreases

A sequence of distributions $\pi_{Y^n}$ on $\mathcal{H}$, depending on the data $Y^n$, is consistent at $\theta_\star$ if, for any $\delta > 0$, $E[\pi_{Y^n}(\{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_\star) > \delta\})] \to o$ as $n \to \infty$, where the expectation is taken with respect to $\mu_\star^{(n)}$. Finding rates of concentration for $\pi_{Y^n}$ involves finding the fastest decaying sequence $\delta_n > 0$ such that the limit above holds. More precisely, we say that the rate of concentration of $\pi_{Y^n}$ is bounded above by the sequence $\delta_n$ if $E[\pi_{Y^n}(\{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_\star) > \delta_n\})] \to o$.

We establish upper bounds on the rates of concentration of the sequence of WABC posteriors around $\theta_\star = \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\theta, \mu_\star)$, as the data size $n$ grows and the threshold shrinks slowly towards $\varepsilon_\star = \mathcal{W}_p(\mu_{\hat{\theta}_n}, \mu_\star)$ at a rate dependent on $n$. Although we focus on the Wasserstein distance in this section, the reasoning also holds for other metrics on $\mathcal{P}(\mathcal{Y})$; see Section 3.2.3 and Appendix B.

Our first assumption is on the convergence of the empirical distribution of the data.

**Assumption 3.3.1.** The data-generating process is such that $\mathcal{W}_p(\hat{\mu}_n, \mu_\star) \to o$, in $\mathbb{P}$-probability, as $n \to \infty$.

In Appendix A, we derive a few different conditions under which Assumption 1 holds for i.i.d. data and certain classes of dependent processes. Additionally, the moment and
concentration inequalities of Fournier and Guillin (2015); Weed and Bach (2019) can also be used to verify both this and the next assumption.

**Assumption 3.3.2.** For any \( \varepsilon > 0 \), \( \mu_{\vartheta}^{(n)}(W_p(\mu_\vartheta, \hat{\mu}_{\vartheta,n}) > \varepsilon) \leq c(\vartheta)f_n(\varepsilon) \), where \( f_n(\varepsilon) \) is a sequence of functions that are strictly decreasing in \( \varepsilon \) for fixed \( n \) and \( f_n(\varepsilon) \to 0 \) for fixed \( \varepsilon \) as \( n \to \infty \). The function \( c : \mathcal{H} \to \mathbb{R}^+ \) is \( \pi \)-integrable, and satisfies \( c(\vartheta) \leq c_o \) for some \( c_o > 0 \), for all \( \vartheta \) such that, for some \( \delta_o > 0 \), \( W_p(\mu_\vartheta, \mu_{\vartheta_0}) \leq \delta_o + \varepsilon_\star \).

For well-specified models, note that Assumption 3.3.2 implies Assumption 3.3.1. The next assumption states that the prior distribution puts enough mass on the sets of parameters \( \vartheta \) that yield distributions \( \mu_\vartheta \) close to \( \mu_\star \) in the Wasserstein distance.

**Assumption 3.3.3.** There exist \( L > 0 \) and \( c_\varpi > 0 \) such that, for all \( \varepsilon \) small enough,

\[
\pi \left( \{ \vartheta \in \mathcal{H} : W_p(\mu_\star, \mu_{\vartheta}) \leq \varepsilon + \varepsilon_\star \} \right) \geq c_\varpi \varepsilon^{L}.
\]

The main result of this subsection is on the concentration of the WABC posteriors on the aforementioned sets.

**Proposition 3.3.2.** Under Assumptions 3.3.1-3.3.3, consider a sequence \( (\varepsilon_n)_{n \geq 0} \) such that, as \( n \to \infty \), \( \varepsilon_n \to 0 \), \( f_n(\varepsilon_n) \to 0 \), and \( P(W_p(\mu_\star, \mu_{\star}) \leq \varepsilon_n) \to 1 \). Then, the WABC posterior with threshold \( \varepsilon_n + \varepsilon_\star \) satisfies, for some \( 0 < C < \infty \) and any \( 0 < R < \infty \),

\[
\pi_{\varepsilon_n + \varepsilon_\star} \left( \{ \vartheta \in \mathcal{H} : W_p(\mu_\star, \mu_{\vartheta}) > \varepsilon_n + 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n/R) \} \right) \leq \frac{C}{R},
\]

with \( P \)-probability going to 1 as \( n \to \infty \).

The assumptions that \( f_n(\varepsilon_n) \to 0 \) and that \( P(W_p(\mu_\star, \mu_{\star}) \leq \varepsilon_n) \to 1 \) imply that \( \varepsilon_n \) has to be the slowest of the two convergence rates: that of \( \hat{\mu}_n \) to \( \mu_\star \) and that of \( \hat{\mu}_{\vartheta,n} \) to \( \mu_\vartheta \). We can further relate concentration on the sets \( \{ \vartheta : W_p(\mu_\star, \mu_{\vartheta}) < \delta + \varepsilon_\star \} \), for some \( \delta > 0 \), to concentration on the sets \( \{ \vartheta : \rho_\mathcal{H}(\vartheta, \vartheta_\star) < \delta \} \), for some \( \delta > 0 \), assuming the parameter \( \vartheta_\star = \text{argmin}_{\vartheta \in \mathcal{H}} W_p(\mu_\star, \mu_{\vartheta}) \) is well-defined. In turn, this leads to concentration rates of the WABC posteriors. To that end, consider the following assumptions.
**Assumption 3.3.4.** The parameter $\theta_* = \text{argmin }_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*; \mu_\theta)$ exists, and is well-separated in the sense that, for all $\delta > 0$, there exists $\delta' > 0$ such that

$$
\inf_{\{\theta \in \mathcal{H} : p(\theta; \theta_*) > \delta\}} \mathcal{W}_p(\mu_\theta, \mu_*) > \mathcal{W}_p(\mu_{\theta_*}, \mu_*) + \delta'.
$$

This assumption is akin to those made in the study of the asymptotic properties of the maximum likelihood estimator under misspecification, where $\theta_*$ is defined in terms of the Kullback–Leibler divergence. In Appendix A, we give a proposition establishing conditions under which Assumption 3.3.4 holds.

Under Assumption 3.3.4, note that the last part of Assumption 3.3.2 is implied by $c(\theta) \leq c_0$ for all $\theta$ with $\mathcal{W}_p(\mu_\theta, \mu_\theta) \leq \delta_0$, for some $\delta_0 > 0$. Indeed, $\mathcal{W}_p(\mu_\theta, \mu_\theta) \leq \delta_0$ implies that $\mathcal{W}_p(\mu_\theta, \mu_*) - \mathcal{W}_p(\mu_\theta, \mu_\theta) \leq \delta_0$. Since $\varepsilon_* = \mathcal{W}_p(\mu_*, \mu_\theta)$, the argument follows. By the same reasoning, Assumption 3.3.3 is implied by

$$
\pi\left(\{\theta \in \mathcal{H} : \mathcal{W}_p(\mu_{\theta_*}, \mu_\theta) \leq \varepsilon\}\right) \geq c_\pi \varepsilon^L,
$$

for some $c_\pi > 0$ and $L > 0$.

**Assumption 3.3.5.** The parameters are identifiable, and there exist $K > 0$, $a > 0$ and an open neighborhood $U \subset \mathcal{H}$ of $\theta_*$, such that, for all $\theta \in U$,

$$
\rho_{\mathcal{H}}(\theta, \theta_*) \leq K(\mathcal{W}_p(\mu_\theta, \mu_*) - \varepsilon_*)^a.
$$

**Corollary 3.3.1.** Under Assumptions 3.3.1–3.3.5, consider a sequence $(\varepsilon_n)_{n \geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to 0, f_n(\varepsilon_n) \to 0, f_n^{-1}(\varepsilon_n^L) \to 0$ and $\mathbb{P}(\mathcal{W}_p(\hat{\mu}_n, \mu_* \leq \varepsilon_n) \to 1$. Then the WABC posterior with threshold $\varepsilon_* + \varepsilon_*$ satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,

$$
\pi_{y_{n+}}^{\varepsilon_* + \varepsilon_*}\left(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_*) > K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^a\}\right) \leq \frac{C}{R},
$$

with $\mathbb{P}$-probability going to 1.

This result bounds the concentration rate from above through the expression $\delta_n = K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^a$, but we remark that it is not clear whether this bound is optimal.
in any sense. Explicit upper bounds for certain classes of models and data-generating processes, such as location-scale models and the AR(1) model in Section 3.4.2, are given in Appendix B. Important aspects of the method that appear in these bounds include the dimension of the observation space \( \mathcal{Y} \), the order \( p \) of the Wasserstein distance, and model misspecification, through the exponent \( a \) in Assumption 3.3.5.

The result provides some insight into the behavior of the method when \( \varepsilon_n \) converges \textit{slowly} to \( \varepsilon_* \). However, it is unclear what happens when \( \varepsilon_n \) decays to a value smaller than \( \varepsilon_* \) at a rate faster than that prescribed by Corollary 3.3.1. As shown in Proposition 3.3.1, the WABC posterior converges to the true posterior when \( \varepsilon \rightarrow o \) for fixed observations. The posterior itself is known to concentrate around the point in \( \mathcal{H} \) minimizing the KL divergence between \( \mu_* \) and \( \mu_{\theta} \) when \( n \rightarrow \infty \) (see e.g. Müller, 2013), and it might be that the WABC posterior inherits similar properties for faster decaying thresholds.

In high dimensions, the rate of convergence of the Wasserstein distance between empirical measures is known to be slow (Talagrand, 1994). On the other hand, recent results establish that it concentrates quickly around its expectation: For instance, del Barrio and Loubes (2019) show that regardless of dimension, \( \mathcal{W}^2_z(\hat{\mu}_n; \hat{\mu}_{\theta,n}) = \mathcal{W}^2_z(\hat{\mu}_n; \hat{\mu}_{\theta,n}) \) converges weakly at the \( \sqrt{n} \) rate to a centered Gaussian random variable with known (finite) variance \( \sigma^2(\mu_*, \mu_{\theta}) \). If the map \( \theta \mapsto \mathbb{E}\mathcal{W}^2_z(\hat{\mu}_n; \hat{\mu}_{\theta,n}) \) offers discrimination between the parameters that is similar to \( \theta \mapsto \mathcal{W}^2_z(\mu_*, \mu_{\theta}) \), it is not clear how the Wasserstein distance’s convergence rate would impact the WABC posterior. Detailed analysis of WABC’s dependence on dimension is an interesting avenue of future research.

3.4 Time series

Viewing data sets as empirical distributions requires some additional care in the case of dependent data, which are common in settings where ABC methods are applied. A naïve approach consists in ignoring dependencies, which might be enough to estimate all parameters in some cases, as illustrated in Section 3.5.3. However, in general, ignoring dependencies might prevent some parameters from being identifiable, as illustrated in the examples of this section. We propose two main approaches to extend the WABC methodology to time series.
3.4.1 Curve matching

Visually, we might consider two time series to be similar if their curves are similar, in a trace plot of the series in the vertical axis against the time indices on the horizontal axis. The Euclidean vector distance between curves sums the vertical differences between pairs of points with identical time indices. We can instead introduce the points $\tilde{y}_t = (t, y_t)$ and $\tilde{z}_t = (t, z_t)$ for all $t \in 1 : n$, viewing the trace plot as a scatter plot. The distance between two points, $(t, y_t)$ and $(s, z_s)$, can be measured by a weighted distance $\rho_\lambda ( (t, y_t), (s, z_s) ) = \|y_t - z_s\| + \lambda |t - s|$, where $\lambda$ is a non-negative weight, and $\|y - z\|$ refers to the Euclidean distance between $y$ and $z$. Intuitively, the distance $\rho_\lambda$ takes into account both vertical and horizontal differences between points of the curves, $\lambda$ tuning the importance of horizontal differences relative to vertical differences. We can then define the Wasserstein distance between two empirical measures supported by $\tilde{y}_{1:n}$ and $\tilde{z}_{1:n}$, with $\rho_\lambda$ as a ground distance on the observation space $\{1, \ldots, n\} \times \mathcal{Y}$. Since computing the Wasserstein distance can be thought of as solving an assignment problem, a large value of $\lambda$ implies that $y_t$ will be assigned to $z_s$, for all $t$. The transport cost will then be $n^{-1} \sum_{t=1}^n \|y_t - z_t\|$, corresponding to the Euclidean distance (up to a scaling factor). If $\lambda$ is smaller, $(t, y_t)$ is assigned to some $(s, z_s)$, for some $s$ possibly different than $t$. If $\lambda$ goes to zero, the distance coincides with the Wasserstein distance between the marginal empirical distributions of $y_{1:n}$ and $z_{1:n}$, where the time element is entirely ignored. Thus curve matching provides a compromise between the Euclidean distance between the series seen as vectors, and the Wasserstein distance between marginal empirical distributions.

For any $\lambda > 0$, the curve matching distance satisfies condition (2) of Proposition 3.3.1, implying that the resulting WABC posterior converges to the standard posterior distribution as $\varepsilon \to 0$. To estimate the WABC posterior, we can utilize any of the methods for computing and approximating the Wasserstein distance discussed in Section 3.2.3 in combination with the SMC algorithm of Section 3.2.1. In Section 3.4.1, we use the exact Wasserstein curve matching distance to infer parameters in a cosine model. The choice of $\lambda$ is open, but a simple heuristic for univariate time series goes as follows. Consider the aspect ratio of the trace plot of the time series $(y_t)$, with horizontal axis spanning from 1 to $n$, and vertical axis from $\min_{t \in 1:n} y_t$ to $\max_{t \in 1:n} y_t$. For an aspect ratio of $H : V$, one can
choose \( \lambda \) as \((\max_{t \in T} y_t - \min_{t \in T} y_t)/V \times (H/n)\). For this choice \( \rho_3 \) corresponds to the Euclidean distance in a rectangular plot with the given aspect ratio.

Generalizations of the curve matching distance have been proposed independently by Thorpe et al. (2017) under the name “transportation \( L_p \) distances”. In that paper, the properties of the curve matching distance are studied in detail, and compared to and combined with the related notion of dynamic time warping (Berndt and Clifford, 1994). Other related distances between time series include the Skorokhod distance between curves (Majumdar and Prabhu, 2015) and the Fréchet distance between polygons (Buchin et al., 2008), in which \( y_t \) would be compared to \( z_{r(t)} \), where \( r \) is a retiming function to be optimized.

**Example: Cosine model**

Consider a cosine model where \( y_t = A \cos(2\pi \omega t + \phi) + \sigma \epsilon_t \), where \( \epsilon_t \sim \mathcal{N}(0, 1) \), for all \( t \geq 1 \), are independent. Information about \( \omega \) and \( \phi \) is mostly lost when considering the marginal empirical distribution of \( y_{1:n} \). In Figure 3.4.1, we compare the ABC posteriors obtained either with the Euclidean distance between the series, or with curve matching, with an aspect ratio of one; in both cases the algorithm is run for \( 10^6 \) model simulations. The figure also shows an approximation of the exact posterior distribution, obtained via Metropolis–Hastings. The prior distributions are uniform on \([0, 1/10]\) and \([0, 2\pi]\) for \( \omega \) and \( \phi \) respectively, and standard Normal on \( \log(\sigma) \) and \( \log(A) \). The data are generated using \( \omega = 1/80, \phi = \pi/4, \log(\sigma) = 0 \) and \( \log(A) = \log(2) \), with \( n = 100 \). We see that curve matching yields a more satisfactory estimation of \( \sigma \) in Figure 3.4.1c, and a similar approximation for the other parameters. By contrast, an ABC approach based on the marginal distribution of \( y_{1:n} \) would fail to identify \( \phi \).

### 3.4.2 Reconstructions

Our second approach consists in transforming the time series to define an empirical distribution \( \tilde{\mu}_n \) from which parameters can be estimated.
Figure 3.4.1: ABC posterior samples in the cosine model of Section 3.4.1, using either the Euclidean distance or curve matching with the exact Wasserstein distance and $\lambda = 1$, after $10^6$ model simulations. We compare to the posterior distribution, obtained using the 50,000 last samples in a Metropolis–Hastings chain of length 100,000. The standard deviation of the noise $\sigma$ is better estimated with curve matching than with the Euclidean distance between time series.

Delay reconstruction

In time series analysis, the lag-plot is a scatter plot of the pairs $(y_{t}, y_{t-k})_{t=k+1}^{n}$ for some lag $k \in \mathbb{N}$, from which one can inspect the dependencies between lagged values of the series. In ABC applied to time series models, lag-$k$ autocovariances, defined as the sample covariance of $(y_{t}, y_{t-k})_{t=k+1}^{n}$, are commonly used statistics to summarize these dependencies (Marin et al., 2012; Mengersen et al., 2013; Li and Fearnhead, 2018). Here, we also propose to use the joint samples $(y_{t}, y_{t-k})_{t=k+1}^{n}$ but bypass their summarization into sample covariances. In particular, we define delay reconstructions as $\tilde{y}_{t} = (y_{t}, y_{t-\tau_1}, \ldots, y_{t-\tau_k})$ for some integers $\tau_1, \ldots, \tau_k$. The sequence, denoted $\tilde{y}_{1:n}$ after relabelling and redefining $n$,
inherits many properties from the original series, such as stationarity. Therefore, the empirical distribution of $\tilde{y}_{i,n}$ denoted by $\tilde{\mu}_{\tilde{\theta}}$, might converge to a limit $\tilde{\mu}_*$. In turn, $\tilde{\mu}_*$ is likely to capture more features of the dependency structure than $\mu_*$, and the resulting procedure might provide more accurate inference on the model parameters than if we were to compare the lag-$k$ autocovariances alone.

Delay reconstructions (or embeddings) play a central role in dynamical systems (Kantz and Schreiber, 2004), for instance in Takens’ theorem and variants thereof (Stark et al., 2003). The Wasserstein distance between the empirical distributions of delay reconstructions has previously been proposed as a way of measuring distance between time series (Moeckel and Murray, 1997; Muskulus and Verduyn-Lunel, 2011), but not as a device for parameter inference. In the ABC setting, we propose to construct the delay reconstructions of each synthetic time series, and to compute the Wasserstein distance between their empirical distribution and the empirical distribution of $\tilde{y}_{i,n}$. We refer to this approach as WABC with delay reconstruction.

Denote by $\tilde{\mu}_{\tilde{\theta},n}$ the empirical distribution of the delay reconstructed series $\tilde{z}_{i,n}$, formed from $z_{i,n} \sim \tilde{\mu}_{\tilde{\theta}}(n)$. Then, provided that $\tilde{\mu}_{\tilde{\theta},n}$ converges to an identifiable distribution $\tilde{\mu}_{\tilde{\theta}}$ as $n \to \infty$, we are back in a setting where we can study the concentration behavior of the WABC posterior around $\tilde{\theta}_* = \text{argmin}_{\tilde{\theta} \in \mathcal{H}} \mathcal{W}_P(\tilde{\mu}_*, \tilde{\mu}_{\tilde{\theta}})$, assuming its existence and uniqueness (see Section 3.3.3). In well-specified settings, $\tilde{\theta}_*$ must correspond to the data-generating parameters.

When the entries of the vectors $y_{i,n}$ and $z_{i,n}$ are all unique, which happens with probability one when $\mu_*(n)$ and $\mu_{\tilde{\theta}}(n)$ are continuous distributions, then $\mathcal{W}_P(\tilde{y}_{i,n}, \tilde{z}_{i,n}) = 0$ if and only if $y_{i,n} = z_{i,n}$. To see this, consider the setting where $\tilde{y}_t = (y_{t-1}, y_{t-1})$, and $\tilde{z}_t = (z_{t-1}, z_{t-1})$. For the empirical distributions of $\tilde{y}_{i,n}$ and $\tilde{z}_{i,n}$ to be equal, we require that for every $t$ there exists a unique $s$ such that $\tilde{y}_t = \tilde{z}_s$. However, since the values in $y_{i,n}$ and $z_{i,n}$ are unique, the values $y_t$ and $z_t$ appear only as the second coordinates of $\tilde{y}_t$ and $\tilde{z}_s$ respectively. It therefore has to be that $y_{t} = z_t$, and $\tilde{y}_t = \tilde{z}_t$. In turn, this implies that $y_{t} = z_t$, and inductively, $y_t = z_t$ for all $t \in 1:n$. A similar reasoning can be done for any $k \geq 2$ and $1 \leq \tau_t < \ldots < \tau_k$. This property can be used to establish the convergence of the WABC posterior based on delay reconstruction to the posterior, as $\epsilon \to 0$, which can be deduced using condition (2) of Proposition 3.3.1.
In practice, for a non-zero value of $\varepsilon$, the obtained ABC posteriors might be different from the posterior, but still identify the parameters with a reasonable accuracy, as illustrated in Section 3.4.2. The quality of the approximation will depend on the choice of lags $\tau_1, \ldots, \tau_k$. Data-driven ways of making such choices are discussed by Kantz and Schreiber (2004). Still, since the order of the original data is only partly reflected in delay reconstructions, some model parameters might be difficult to estimate with delay reconstruction, such as the phase shift $\varphi$ in Section 3.4.1.

**Example: AR(1) model**

Consider an autoregressive process of order 1, written AR(1), where $y_i \sim \mathcal{N}(0, \sigma^2/(1 - \varphi^2))$, for some $\sigma > 0$ and $\varphi \in (-1, 1)$. For each $t \geq 2$, let $y_t = \varphi y_{t-1} + \sigma w_t$, where $w_t \sim \mathcal{N}(0, 1)$ are independent. The marginal distribution of each $y_i$ is $\mathcal{N}(0, \sigma^2/(1 - \varphi^2))$. Furthermore, by an ergodic theorem, the empirical distribution $\hat{\mu}_n$ of the time series converges to this marginal distribution. The two parameters ($\varphi, \sigma^2$) are not identifiable from the limit $\mathcal{N}(0, \sigma^2/(1 - \varphi^2))$. Figure 3.4.2a shows WABC posterior samples derived while ignoring time dependence, obtained for decreasing values of $\varepsilon$, using a budget of $10^5$ model simulations. The prior is uniform on $[-1, 1]$ for $\varphi$, and standard Normal on $\log(\sigma)$. The data are generated using $\varphi = 0.7$, $\log(\sigma) = 0.9$ and $n = 1,000$. The WABC posteriors concentrate on a ridge of values with constant $\sigma^2/(1 - \varphi^2)$.

Using $k = 1$, we consider $\tilde{y}_t = (y_t, y_{t-1})$ for $t \geq 2$. The reconstructions are then sub-sampled to 500 values, $\tilde{y}_2 = (y_2, y_1), \tilde{y}_4 = (y_4, y_3), \ldots, \tilde{y}_{1000} = (y_{1000}, y_{999})$; similar results were obtained with the 999 reconstructed values, but sub-sampling leads to computational gains in the exact Wasserstein distance calculations; see Section 3.2.3. The stationary distribution of $\tilde{y}_t$ is given by

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\sigma^2}{1 - \varphi^2} \begin{pmatrix} 1 & \varphi \\ \varphi & 1 \end{pmatrix}\right).$$

Both parameters $\sigma^2$ and $\varphi$ can be identified from a sample approximating the above distribution. Figure 3.4.2b shows the WABC posteriors obtained with delay reconstruction and a budget of $10^5$ model simulations concentrating around the data-generating values as $\varepsilon$
(a) WABC using $\mathcal{W}_i$ between marginal distributions.  
(b) WABC using $\mathcal{W}_i$ between delay reconstructions.

**Figure 3.4.2:** Samples from the WABC posteriors of $(\varphi, \log(\sigma))$ in the AR(1) model of Section 3.4.2, as $\varepsilon$ decreases over the steps of the SMC sampler (colors from red to white to blue). On the left, using the marginal empirical distribution of the series, the WABC posteriors concentrate around a ridge of values such that $\sigma^2/(1 - \varphi^2)$ is constant. On the right, using delay reconstruction with $\text{lag } k = 1$, the WABC posteriors concentrate around the data-generating parameters, $\varphi = 0.7, \log(\sigma) = 0.9$, indicated by full lines. Both methods had a total budget of $10^5$ model simulations.

decreases.

**Residual reconstruction**

Another approach to handle dependent data is advocated in Mengersen et al. (2013), in the context of ABC via empirical likelihood. In various time series models, the observations are modeled as transformations of some parameter $\theta$ and residual variables $w_1, \ldots, w_n$. Then, given a parameter $\theta$, one might be able to reconstruct the residuals corresponding to the observations. In Section 3.4.1, one can define $w_i = (y_i - A \cos(2\pi \omega t + \varphi))/\sigma$. In Section 3.4.2, one can define $w_i = (y_i - \varphi y_{i-1})/\sigma$; other examples are given in Mengersen et al. (2013). Once the residuals have been reconstructed, their empirical distribution can be compared to the distribution that they would follow under the model, e.g. a standard Normal in Sections 3.4.2 and 3.4.1.
3.5 Numerical experiments

We illustrate the proposed approach and make comparisons to existing methods in various models taken from the literature. In each example, we approximate the WABC posterior using the SMC algorithm and default parameters outlined in Section 3.2.1.

3.5.1 Quantile “g-and-k” distribution

We first consider an example where the likelihood can be approximated to high precision, which allows comparisons between the standard posterior and WABC approximations. We observe that the WABC posterior converges to the true posterior in the univariate g-and-k example, as suggested by Proposition 3.3.1. We also compare WABC to a method developed by Fearnhead and Prangle (2012) that uses a semi-automatic construction of summary statistics. Lastly, we compare the use of the Wasserstein distance with other distances described in Section 3.2.3 on a bivariate version of the g-and-k distribution.

Univariate “g-and-k”

A classical example in the ABC literature (see e.g. Fearnhead and Prangle, 2012; Mengersen et al., 2013), the univariate g-and-k distribution is defined in terms of its quantile function:

\[ r \in (0, 1) \mapsto a + b \left( 1 + o.8 \frac{1 - \exp(-gz(r))}{1 + \exp(-gz(r))} \right) (1 + z(r)^2) z(r), \]  

(3.9)

where \( z(r) \) refers to the \( r \)-th quantile of the standard Normal distribution.

Sampling from the g-and-k distribution can be done by plugging standard Normal variables into (3.9) in place of \( z(r) \). The probability density function is intractable, but can be numerically calculated with high precision since it only involves one-dimensional inversions and differentiations of the quantile function in (3.9), as described in Rayner and MacGillivray (2002). Therefore, Bayesian inference can be carried out with e.g. Markov chain Monte Carlo.

We generate \( n = 250 \) observations from the model using \( a = 3, b = 1, g = 2, k = 0.5 \), and the parameters are assigned a uniform prior on \( [0, 10]^4 \). We estimate the posterior distribution by running 5 Metropolis–Hastings chains for 75,000 iterations, and discard the
first 50,000 as burn-in. For the WABC approximation, we use the SMC sampler outlined in Section 3.2.1 with $N = 2,048$ particles, for a total of $2.4 \times 10^6$ simulations from the model. The resulting marginal WABC posteriors are also compared to the marginal posteriors obtained with the semi-automatic ABC approach of Fearnhead and Prangle (2012). We used the rejection sampler in the abctools package (Nunes and Prangle, 2015), also for a total of $2.4 \times 10^6$ model simulations, of which $N = 2,048$ draws from the prior are accepted. We observed no benefit to accepting fewer draws. The semi-automatic approach requires the user to specify a set of initial summary statistics, for which we used every 25th order statistic as well as the minimum (that is, $y_{(1)}; y_{(2)}; y_{(25)}; \ldots; y_{(250)}$) and their powers up to fourth order, following guidance in Fearnhead and Prangle (2012).

Figure 3.5.1 shows the marginal posterior distributions and their approximations obtained with WABC and semi-automatic ABC. The plots show that the WABC posteriors appear to be closer to the target distributions, especially on the $a$, $b$ and $k$ parameters. Neither method captures the marginal posterior of $g$ well, though the WABC posterior appears more concentrated in the region of interest on that parameter as well.

In both of the ABC approaches, the main computational costs stem from simulating from the model and sorting the resulting data sets. Over 1,000 repetitions, the average wall-clock time to simulate a data set was $7.7 \times 10^{-3}s$ on an Intel Core i5 (2.5 GHz). The average time to sort the resulting data sets was $8.9 \times 10^{-4}s$, and computing the Wasserstein distance to the observed data set was negligibly different from this. In semi-automatic ABC, one additionally has to perform a regression step. The rejection sampler in semi-automatic ABC is easier to parallelize than our SMC approach, but on the other hand requires more memory due to the regression used in constructing the summary statistic. This makes the method hard to scale up beyond the number of model simulations considered here, without using specialized tools for large scale regression.

This problem does not arise in the WABC sequential Monte Carlo sampler, and Figure 3.5.2 illustrates the behavior of the marginal WABC posteriors as more steps of the SMC sampler are performed. In particular, we can see that the WABC approximations for $a$, $b$ and $k$ converge to the corresponding posteriors (up to some noise). The approximations for $g$ also shows convergence towards the posterior, but have not yet reached the target distribution at the stage when the sampler was terminated. The convergence is further il-
illustrated in Figure 3.5.2g, where the $W_i$- distance between the joint WABC posterior and joint posterior is plotted as a function of the number of simulations from the model. The plot shows that the Wasserstein distance between the distributions decreases from around 10 to around 0.06 over the course of the SMC run. The distances are approximated by thinning the MCMC samples left after burn-in down to 2,048 samples, and computing the Wasserstein between the corresponding empirical distribution and the empirical distributions supported at the $N = 2,048$ SMC particles at each step.

Figure 3.5.2f shows the development of the threshold as a function of the number of model simulations, showing that $\varepsilon$ decreases to 0.07 over the course of the SMC. The threshold decreases at a slower rate as it approaches zero, suggesting that the underlying sampling problem becomes harder as $\varepsilon$ becomes smaller. This is also illustrated by Figure 3.5.2e, which shows the number of model simulations performed at each step of the SMC algorithm. This number is increasing throughout the run of the algorithm, as the $r$-hit kernel requires more and more attempts before it reaches the desired number of hits.

**Bivariate “g-and-k”**

We also consider the bivariate extension of the g-and-k distribution (Drovandi and Pettitt, 2011), where one generates bivariate Normals with mean zero, variance one, and correlation $\rho$, and substitutes $z(r)$ with them in (3.9), with parameters $(a_i, b_i, g_i, k_i)$ for each component $i \in \{1, 2\}$. Since the model generates bivariate data, we can no longer rely on simple sorting to calculate the Wasserstein distance. We compare the exact Wasserstein distance to the approximations discussed in Section 3.2.3, as well as the maximum mean discrepancy, whose use within ABC was proposed by Park et al. (2016) (see Section 3.1.4).

We generate $n = 500$ observations from the model using $a_1 = 3, b_1 = 1, g_1 = 1, k_1 = 0.5, a_2 = 4, b_2 = 0.5, g_2 = 2, k_2 = 0.4, \rho = 0.6$, as in Section 5.2 of Drovandi and Pettitt (2011). The parameters $(a_i, b_i, g_i, k_i)$ are assigned a uniform prior on $[0, 10]^4$, independently for $i \in \{1, 2\}$, and $\rho$ a uniform prior on $[-1, 1]$. We estimate the posterior distribution by running 8 Metropolis–Hastings chains for 150,000 iterations, and discard the first 50,000 as burn-in. For each of the ABC approximations, we run the SMC sampler
Figure 3.5.1: Posterior marginals in the univariate g-and-k example of Section 3.5.1 (obtained via MCMC), approximations by Wasserstein ABC and semi-automatic ABC, each with a budget of $2.4 \times 10^6$ model simulations. Data-generating values are indicated by vertical lines.
Figure 3.5.2: 3.5.2a-3.5.2d: Posterior marginals in the univariate g-and-k example of Section 3.5.1 (grey, obtained via MCMC) and approximations by Wasserstein ABC from step 20 to step 57 of the SMC algorithm. The color of the WABC approximation changes from red to blue as more steps of the SMC sampler are performed, decreasing the threshold $\varepsilon$. For the densities plotted here, the threshold reduces from $\varepsilon = 0.20$ to $\varepsilon = 0.07$. The range of the plot for $g$ has been truncated to $(1, 6)$ to highlight the region of interest, despite the densities from the earlier steps of the SMC having support outside this region. Data-generating values are indicated by vertical lines. Figure 3.5.2e shows the number of simulations from the model used in each step of the SMC algorithm ($y$-axis in log scale). This number is increasing due to use of the $r$-hit kernel within the SMC. Figures 3.5.2f and 3.5.2g show the threshold $\varepsilon$ and the $\mathcal{W}_1$-distance to the posterior respectively, against the number of model simulations (both plots in log-log scale).
outlined in Section 3.2.1 for a total of $2 \times 10^6$ simulations from the model. For the MMD, we use the estimator

$$
\text{MMD}^2(y_{i:n}, z_{1:n}) = \frac{1}{n^2} \sum_{i,j=1}^n k(y_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(z_i, z_j) - \frac{2}{n^2} \sum_{i,j=1}^n k(y_i, z_j),
$$

(3.10)

with the kernel $k(x, x') = \exp(-\|x - x'\|^2/2h^2)$. The bandwidth $h$ was fixed to be the median of the set $\{\|y_i - y_j\| : i, j = 1, \ldots, n\}$, following guidance in Park et al. (2016).

Figure 3.5.3c shows the $\mathcal{W}_1$-distance between the $N = 2, 048$ ABC posterior samples, obtained with the different distances, and a sample of $2, 048$ points thinned out from the Markov chains targeting the posterior. This distance is plotted against the number of model simulations. It shows that all distances yield ABC posteriors that get closer to the actual posterior. On the other hand, for this number of model simulations, all of the ABC posteriors are significantly different from the actual posterior. For comparison, the $\mathcal{W}_1$-distance between two samples of size $2, 048$ thinned out from the Markov chains is on average about 0.07. In this example, it appears that the MMD leads to ABC posteriors that are not as close to the posterior as the other distances given the same budget of model simulations. The Hilbert distance provides a particularly cheap and efficient alternative to the Wasserstein distance in this bivariate case, providing a very similar approximation to the posterior as both the exact Wasserstein and swapping distances.

Figure 3.5.3b shows the development of the threshold as a function of the number of model simulations for the different distances. Note that the MMD is not on the same scale as the Wasserstein distance and its approximations, and therefore the MMD thresholds are not directly comparable to the those of the other distances. As for the univariate g-and-k distribution, the thresholds decrease at a slower rate as they become smaller for each of the distances, suggesting that the underlying sampling problem becomes harder as $\varepsilon$ becomes smaller. The behaviors of the thresholds based on the exact Wasserstein, swapping, and Hilbert distances appear negligibly different. Figure 3.5.3a shows the number of model simulations performed at each step of the SMC algorithm for each of the distances. As before, these numbers are increasing throughout the run of the algorithm, as the $r$-hit kernel requires more and more attempts before it reaches the desired number of hits.
An important distinction between the distances is the time they take to compute. For data sets of size \( n = 500 \) simulated using the data-generating parameter, the average wall-clock times to compute distances between simulated and observed data, on an Intel Core i7-5820K (3.30 GHz), are as follows: 0.002s for the Hilbert distance, 0.015s for the MMD, 0.03s for the swapping distance, and 0.22s for the exact Wasserstein distance; these average times were computed on 1,000 independent data sets. In this example, simulating from the model takes a negligible amount of time, even compared to the Hilbert distance. Calculating the likelihood over 1,000 parameters drawn from the prior, we find an average computing time of 0.05s. In combination with the information conveyed by Figure 3.5.3, the Hilbert and swapping-based ABC posteriors provide good approximations of the exact Wasserstein-based ABC posteriors in only a fraction of the time the latter takes to compute.

### 3.5.2 Toggle switch model

We borrow the system biology “toggle switch” model used in Bonassi et al. (2011); Bonassi and West (2015), inspired by studies of dynamic cellular networks. This provides an example where the design of specialized summaries can be replaced by the Wasserstein distance between empirical distributions. For \( i \in \{1, \ldots, n\} \) and \( t \in \{1, \ldots, T\} \), let \((u_{i,t}, v_{i,t})\) denote the expression levels of two genes in cell \( i \) at time \( t \). Starting from \((u_{i,0}, v_{i,0}) = (10, 10)\), the evolution of \((u_{i,t}, v_{i,t})\) is given by

\[
\begin{align*}
u_{i,t+1} & = v_{i,t} + a_2/(1 + u_{i,t}^{\beta_2}) - (1 + 0.03u_{i,t}) + 0.55^{\xi_{i,t,1}}, \\
u_{i,t+1} & = v_{i,t} + a_1/(1 + u_{i,t}^{\beta_1}) - (1 + 0.03v_{i,t}) + 0.55^{\xi_{i,t,2}},
\end{align*}
\]

where \( a_1, a_2, \beta_1, \beta_2 \) are parameters, and \( \xi \)'s are standard Normal variables, truncated so that \((u_{i,t}, v_{i,t})\) only takes non-negative values. For each cell \( i \), we only observe a noisy measurement of the terminal expression level \( u_{i,T} \). Specifically, the observations \( y_i \) are assumed to be independently distributed as \( \mathcal{N}(\mu + u_{i,T}, \mu^2\sigma^2/u_{i,T}^{2\gamma}) \) random variables truncated to be non-negative, where \( \mu, \sigma, \gamma \) are parameters. We generate \( n = 2,000 \) observations using \( a_1 = 22, a_2 = 12, \beta_1 = 4, \beta_2 = 4.5, \mu = 325, \sigma = 0.25, \gamma = 0.15 \). A histogram of the data is shown in Figure 3.5.4a.
Figure 3.5.3: 3.5.3a shows the number of simulations from the model used in each step of the SMC algorithm for the four distances applied to the bivariate g-and-k model of Section 3.5.1 (y-axis in log scale). This number is increasing due to use of the r-hit kernel within the SMC. Figure 3.5.3b shows the thresholds $\varepsilon$ against the number of model simulations (in log-log scale). Note that the MMD is not on the same scale as the Wasserstein distance and its approximations, and therefore the MMD thresholds are not directly comparable to those of the other distances. Figure 3.5.3c shows the $W_1$-distance between the joint ABC posteriors based on the different distances to the joint true posterior, against the number of model simulations (in log-log scale).
We consider the task of estimating the data-generating values, using uniform prior distributions on \([0, 50]\) for \(a_1, a_2\), on \([0, 5]\) for \(\beta_1, \beta_2\), on \([250, 450]\) for \(\mu\), on \([0, 0.5]\) for \(\sigma\) and on \([0, 0.4]\) for \(\gamma\). These ranges are derived from Figure 5 in Bonassi and West (2015). We compare our method using \(p = 1\) with a summary-based approach using the 11-dimensional tailor-made summary statistic from Bonassi et al. (2011); Bonassi and West (2015). Since the data are one-dimensional, the Wasserstein distance between data sets can be computed quickly via sorting. For both methods, we use the SMC sampler outlined in Section 3.2.1, for a total number of \(10^6\) model simulations.

The seven marginal ABC posterior distributions obtained in the final step of the SMC sampler are shown in Figure 3.5.4. We find that the marginal WABC and summary-based posteriors concentrate to the same distributions for the \(a_2, \beta_1\) and \(\beta_2\) parameters. On the remaining parameters, the marginal WABC posteriors show stronger concentration around the data-generating parameters than the summary-based approach. Comparing the results, we see that the design of a custom summary can be bypassed using the Wasserstein distance between empirical distributions: the resulting posterior approximations appear to be more concentrated around the data-generating parameters, and our proposed approach is fully black-box. The time to simulate data from the model does not seem to depend noticeably on the parameter, and the average wall-clock time to simulate a data set over 1,000 repetitions was 0.523s on an Intel Core i5 (2.5GHz). The average time compute the Wasserstein distance to the observed data set was 0.0002s, whereas the average time to compute the summary statistic was 0.176s.

### 3.5.3 Queueing Model

We turn to the \(M/G/1\) queueing model, which has appeared frequently as a test case in the ABC literature, see e.g. Fearnhead and Prangle (2012). It provides an example where the observations are dependent, but where the parameters can be identified from the marginal distribution of the data. In the model, customers arrive at a server with independent interarrival times \(w_i\) exponentially distributed with rate \(\theta_1\). Each customer is served with independent service times \(u_i\) taken to be uniformly distributed on \([\theta_1, \theta_2]\). We observe
Figure 3.5.4: Histogram of observations (3.5.4a), and marginal posteriors based on WABC and the summary statistic from Bonassi et al. (2011); Bonassi and West (2015) in the toggle switch model. The ABC posteriors are computed using the SMC sampler from Section 3.2.1, for a total number of $10^8$ model simulations. Data-generating values are indicated by vertical lines.
only the interdeparture times \( y_i \) given by the process

\[
y_i = u_i + \max \left\{ 0, \sum_{j=1}^{i} w_j - \sum_{j=1}^{i-1} y_j \right\}.
\]

(3.13)

The prior on \((\theta_1, \theta_2 - \theta_1, \theta_3)\) is Uniform on \([0, 10]^3 \times [0, 1/3]\).

We use the data set given in Shestopaloff and Neal (2014), which was generated using the parameters \((\theta_1, \theta_2 - \theta_1, \theta_3) = (4, 3, 0.15)\) and \(n = 50\). The WABC posterior based on the empirical distribution of \(y_{i,n}\) ignoring dependencies, is approximated using the SMC algorithm of Section 3.2.1, with a budget of \(10^7\) model simulations. We compare with the semi-automatic ABC approach of Fearnhead and Prangle (2012) with the same budget of model simulations, using a subset of 20 evenly spaced order statistics as the initial summary statistics in that method. The semi-automatic ABC posteriors are computed using the rejection sampler in the \texttt{abctools} package (Nunes and Prangle, 2015), accepting the 100 best samples. The actual posterior distribution is approximated with a particle marginal Metropolis–Hastings (PMMH) run (Andrieu et al., 2010), using 4,096 particles and \(10^5\) iterations. The use of PMMH was suggested in Shestopaloff and Neal (2014) as an alternative to the model-specific Markov chain Monte Carlo algorithm they propose.

Upon observing \(y_{i,n}, \theta_i\) has to be less than \(\min_{i \in [n]} y_i\), which is implicitly encoded in the likelihood, but not in an ABC procedure. One can add this constraint explicitly, rejecting parameters that violate it, which is equivalent to redefining the prior on \(\theta_i\) to be uniform on \([0, \min_{i \in [n]} y_i]\). Figure 3.5.5 shows the marginal distributions of the parameters obtained with PMMH, semi-automatic ABC, and WABC, with or without the additional constraint.

Overall, the WABC approximations are close to the posterior, in comparison to the relatively vague prior distribution on \((\theta_1, \theta_2 - \theta_1)\). Furthermore, we see that incorporating the constraint leads to marginal WABC approximations that are closer to the marginal posteriors. Both variations of WABC appear to perform better than semi-automatic ABC, except on \(\theta_i\), where the semi-automatic ABC approximation is closer to the posterior than the unconstrained WABC approximation. We observed no significant difference in the semi-automatic ABC posterior when incorporating the constraint on \(\theta_i\), and hence only show the approximated posterior for the unconstrained approach.
As in the univariate g-and-k model of Section 3.5.1, the computation costs for the WABC and semi-automatic approaches are similar, as they both rely on simulating from the model and sorting the resulting data. Over 1,000 repetitions, the average wall-clock time to simulate a data set was $7.5 \times 10^{-5}$ s on an Intel Core i5 (2.5GHz). Sorting a data set took on average $7.7 \times 10^{-5}$ s, and computing the Wasserstein distance was negligibly different from this. For the semi-automatic ABC approach, one additionally has to perform the regression step. The model simulations in semi-automatic ABC are easier to parallelize, but the method is hard to scale up without specialized tools for large-scale regression, due to memory requirements of the regression used to construct the summary statistics.

![Posterior marginals](image)

**Figure 3.5.5:** Posterior marginals in the M/G/1 queueing model of Section 3.5.3 (obtained via particle marginal Metropolis–Hastings), approximations by Wasserstein ABC and semi-automatic ABC, and Wasserstein ABC accounting for the constraint that $\theta_1$ has to be less than $\min_{t \in T} y_t$, each with a budget of $10^7$ model simulations. Data-generating values are indicated by vertical lines.

### 3.5.4 Lévy-driven stochastic volatility model

We consider a Lévy-driven stochastic volatility model (e.g. Barndorff-Nielsen and Shephard, 2002), used in Chopin et al. (2013) as a challenging example of parameter inference in state space models. We demonstrate how ABC with transport distances can identify some of the parameters in a black-box fashion, and can be combined with summaries to identify the remaining parameters. The observation $y_t$ at time $t$ is the log-return of a fi-
nancial asset, assumed Normal with mean $\mu + \beta \nu_t$ and variance $\nu_t$, where $\nu_t$ is the actual volatility. Together with the spot volatility $z_t$, the pair $(\nu_t, z_t)$ constitutes a latent Markov chain, assumed to follow a Lévy process. Starting with $z_0 \sim \Gamma(\xi^2/\omega^2, \xi/\omega^2)$ (where the second parameter is the rate), and an arbitrary $\nu_0$, the evolution of the process goes as follows:

$$
k \sim \mathcal{P}oisson\left(\lambda \xi^2/\omega^2\right), \quad c_{t,k} \overset{i.i.d.}{\sim} \mathcal{U}(t, t+1), \quad e_{t,k} \overset{i.i.d.}{\sim} \mathcal{E}xp\left(\xi/\omega^2\right),
$$

$$
z_{t+1} = e^{-\lambda}z_t + \sum_{j=1}^{k} e^{-\lambda(t+1-t_j)}e_j, \quad \nu_{t+1} = \frac{1}{\lambda}[z_t - z_{t+1} + \sum_{j=1}^{k} e_j]. \quad (3.14)
$$

The random variables $(k, c_{t,k}, e_{t,k})$ are generated independently for each time period, and $1 : k$ is the empty set when $k = 0$. The parameters are $(\mu, \beta, \xi, \omega^2, \lambda)$. We specify the prior as Normal with mean zero and variance 2 for $\mu$ and $\beta$, Exponential with rate 0.2 for $\xi$ and $\omega^2$, and Exponential with rate 1 for $\lambda$.

We generate synthetic data with $\mu = 0, \beta = 0, \xi = 0.5, \omega^2 = 0.0625, \lambda = 0.01$, which were used also in the simulation study of Barndorff-Nielsen and Shephard (2002); Chopin et al. (2013), of length $n = 10,000$. We use delay reconstruction with a lag $k = 1$, and the Hilbert distance $\mathcal{H}_p$ of Section 3.2.3 with $p = 1$. Given the length of the time series, the cost of computing the Hilbert distance is much smaller than that of the other distances discussed in Section 3.2.3. We ran the SMC algorithm outlined in Section 3.2.1 until a total of $4.2 \times 10^5$ data sets had been simulated. Figure 3.5.6 shows the resulting quasi-posterior marginals for $(\mu, \beta)$, $\xi, \omega^2)$, and $\lambda$. The parameters $(\mu, \beta, \xi, \omega^2)$ are accurately identified, from a vague prior to a region close to the data-generating values. On the other hand, the approximation of $\lambda$ is barely different from the prior distribution. Indeed, the parameter $\lambda$ represents a discount rate which impacts the long-range dependencies of the process, and is thus not captured by the bivariate marginal distribution of $(y_1, y_{n-1})$.

Hoping to capture long-range dependencies in the series, we define a summary $\eta(y_{1:n})$ as the sum of the first 50 sample autocorrelations among the squared observations. For each of the parameters obtained with the first run of WABC described above, we compute the summary of the associated synthetic data set. We plot the summaries against $\lambda$ in Fig-
ure 3.5.7a. The dashed line indicates the value of the summary calculated on the observed data. The plot shows that the summaries closest to the observed summary are those obtained with the smallest values of $\lambda$. Therefore, we might be able to learn more about $\lambda$ by combining the previous Hilbert distance with a distance between summaries.

Denote by $\tilde{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n})$ the Hilbert distance between delay reconstructions, and by $\varepsilon_h$ the threshold obtained after the first run of the algorithm. A new distance between data sets is defined as $|\eta(y_{1:n}) - \eta(z_{1:n})| \text{ if } \tilde{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n}) < \varepsilon_h$, and $+\infty$ otherwise. We then run the SMC sampler of Section 3.2.1, initializing with the results of the first run, but using the new distance. In this second run, a new threshold is introduced and adaptively decreased, keeping the first threshold $\varepsilon_h$ fixed. One could also decrease both thresholds together or alternate between decreasing either. Note that the Hilbert distance and the summaries could have been combined in other ways, for instance in a weighted average.

We ran the algorithm with the new distance for an extra $6.6 \times 10^5$ model simulations. Figures 3.5.7b and 3.5.7c show the evolution of the WABC posterior distributions of $\omega^2$ and $\lambda$ during the second run. The WABC posteriors concentrate closer to the data-generating values, particularly for $\lambda$; for $(\mu, \beta, \xi)$, the effect is minimal and not shown. In terms of computing time, it took on average $1.3 \times 10^{-1}$s to generate time series given the data-generating parameter, $2.4 \times 10^{-2}$s to compute the Hilbert distance, and $1.5 \times 10^{-3}$s to compute the summary statistic, on an Intel Core i5 (2.5 GHz). Thus, most of the time consumed by the algorithm was spent generating data.

The WABC posterior could then be used to initialize a particle MCMC algorithm (Andrieu et al., 2010) targeting the posterior. The computational budget of roughly $1.1 \times 10^6$ model simulations, as performed in total by the WABC procedure in this section, would be equivalent to relatively few iterations of particle MCMC in terms of number of model transitions. Therefore, the cost of initializing a particle MCMC algorithm with the proposed ABC approach is likely to be negligible. The approach could be valuable in settings where it is difficult to initialize particle MCMC algorithms, for instance due to the large variance of the likelihood estimator for parameters located away from the posterior mode, as illustrated in Figure 2 (c) of Murray et al. (2013).
(a) Posterior of \((\mu, \beta)\). (b) Posterior of \((\xi, \omega^2)\) (in log-log scale). (c) Posterior of \(\lambda\) (in log-log scale).

**Figure 3.5.6:** ABC approximations in the Lévy-driven stochastic volatility model, using the Hilbert distance between delay reconstructions with lag \(k = 1\). The plots show samples from the bivariate marginals of \((\mu, \beta)\) (left), \((\xi, \omega^2)\) (middle), and the marginal distributions of \(\lambda\) (right), as the threshold \(\varepsilon\) decreases during the steps of the SMC sampler (colors from red to blue). The total budget was \(4.2 \times 10^5\) model simulations. Data-generating parameters are indicated by full lines.

(a) Summary against \(\lambda\). (b) Distributions of \(\omega^2\). (c) Distributions of \(\lambda\) (log-scale).

**Figure 3.5.7:** Left: summary, defined as the sum of the first 50 sample autocorrelations of the squared series, against \(\lambda\), computed for the output of the WABC algorithm using the Hilbert distance between delay reconstructions, applied to the Lévy-driven stochastic volatility model of Section 3.5.4. Middle and right: approximations of \(\omega^2\) and \(\lambda\), from the second run of WABC using the Hilbert distance between delay reconstructions combined with the summary for another \(6.6 \times 10^5\) model simulations. The colors change from red to blue as more steps of the SMC sampler are performed. The horizontal axis in the right plot is in log-scale, and illustrates the concentration of the ABC posterior towards the data-generating value of \(\lambda\).
3.6 Discussion

Using the Wasserstein distance in approximate Bayesian computation leads to a new way of inferring parameters in generative models, bypassing the choice of summaries. The approach can also be readily used for deterministic models. We have demonstrated how the proposed approach can identify high posterior density regions, in settings of both i.i.d. (Section 3.5.1) and dependent data (Section 3.5.3). In some examples the proposed approximations appear to be at least as close to the posterior distribution as those produced by state-of-the-art summary-based ABC. For instance, in the toggle switch model of Section 3.5.2, our black-box method obtained posterior approximations that are more concentrated on the data-generating parameters than those obtained with sophisticated, case-specific summaries, while being computationally cheaper. Furthermore, we have shown how summaries and transport distances can be fruitfully combined in Section 3.5.4. There are various ways of combining distances in the ABC approach, which could deserve more research.

We have proposed multiple ways of defining empirical distributions of time series data, in order to identify model parameters. The proposed approaches have tuning parameters, such as $\lambda$ in the curve matching approach of Section 3.4.1 or the lags in delay reconstruction in Section 3.4.2. The choice of these parameters has been commented on by Thorpe et al. (2017) in the case of curve matching, and by Muskulus and Verduyn-Lunel (2011); Stark et al. (2003); Kantz and Schreiber (2004) in the case of delay reconstructions. Making efficient choices of these parameters might be easier than choosing summary statistics. Further research could leverage e.g. the literature on Skorokhod distances for $\lambda$ (Majumdar and Prabhu, 2015). The investigation of similar methods for the setting of spatial data would also be interesting.

We have established some theoretical properties of the WABC distribution, adding to the existing literature on asymptotic properties of ABC posteriors (Frazier et al., 2018; Li and Fearnhead, 2018). In particular, we have considered settings where the threshold $\varepsilon$ goes to zero for a fixed set of observations, and where the number of observations $n$ goes to infinity with a slowly decreasing threshold sequence $\varepsilon_n$. In the first case, we establish conditions under which the WABC posterior converges to the posterior, as illustrated em-
pirically in Section 3.5.1. In the second case, our results show that under certain conditions, the WABC posterior can concentrate in different regions of the parameter space compared to the posterior. We also derive upper bounds on the concentration rates, which highlight the potential impact of the order \( p \) of the Wasserstein distance, of the dimension of the observation space, and of model misspecification. The dependence on dimension of the observation space could be a particularly interesting avenue of future research.

In comparison with the asymptotic regime, less is known about the properties of ABC posteriors for fixed \( \varepsilon \). Viewing the WABC posterior as a coarsened posterior (Miller and Dunson, 2018), one can justify its use in terms of robustness to model misspecification. On the other hand, ABC posteriors in general do not yield conservative statements about the posterior for a fixed threshold \( \varepsilon \) and data set \( y_{\cdot,n} \). For instance, Figure 3.4.1c shows that ABC posteriors can have little overlap with the posterior, despite having shown signs of concentration away from the prior distribution.

As Wasserstein distance calculations scale super-quadratically with the number of observations \( n \), we have introduced a new distance based on the Hilbert space-filling curve, computable in order \( n \log n \), which can be used to initialize a swapping distance with a cost of order \( n^2 \). We have derived some posterior concentration results for the ABC posterior distributions using the Hilbert and swapping distances, similarly to Proposition 3.3.2 obtained for the Wasserstein distance. Many other distances related to optimal transport could be used; we mentioned Park et al. (2016) who used the maximum mean discrepancy, and recently Genevay et al. (2018b) consider Sinkhorn divergences, and Jiang et al. (2018) consider the Kullback–Leibler divergence. A thorough comparison between these different distances, none of which involve summary statistics, could be of interest to ABC practitioners.
4

Langevin Monte Carlo and JKO splitting

4.1 INTRODUCTION

In this chapter, we shed new light on Langevin-based Monte Carlo algorithms by drawing connections to the Wasserstein gradient flow literature and the operator splitting approach to solving PDEs. In a seminal paper, Jordan et al. (1998) expressed the solution of the Fokker–Planck equation as the gradient flow of the relative entropy functional (otherwise known as the KL-divergence) with respect to the 2-Wasserstein distance. Their constructive proof used a time discretization approach that has since become known as the JKO scheme. We show that applying the JKO scheme in conjunction with a splitting approach to solving the Fokker–Planck equation reduces to a proximal version of the Unadjusted Langevin Algorithm. Our proofs rely heavily on the theory developed by Ambrosio et al. (2005), and have the benefit of holding for potentials that are not necessarily differentiable. In turn, this allows us to provide some new results regarding the convergence of the algo-
algorithm. Our work is related to Durmus et al. (2016), and we will make comparisons to their theoretical results.

To motivate the use of Langevin-based Monte Carlo algorithms, consider a log-concave target distribution \( \pi \), given in terms of the Lebesgue density \( \pi(x) = Z^{-1}e^{-V(x)} \), where \( V: \mathbb{R}^d \to \mathbb{R} \) is a convex function, \( d \in \mathbb{N} \) is an integer, and \( Z \) is the normalizing constant. In the case where \( V \) is differentiable, we can associate with it the Langevin diffusion, given in terms of the Itô stochastic differential equation

\[
    dX(t) = -\nabla V(X(t))dt + \sqrt{2}dW(t), \quad X(0) = X_0 \sim \rho_0. \tag{4.1}
\]

It represents the position \( X(t) \in \mathbb{R}^d \) of a particle at time \( t > 0 \), initialized at the random location \( X_0 \sim \rho_0 \), with drift according to the gradient of the potential \( V \) and subject to random perturbations \( dW(t) \). The process \( W(t) \) is the standard Wiener process. The density of \( X(t) \) at time \( t \), written \( \rho(t) \), satisfies the linear Fokker–Planck equation:

\[
    \frac{d\rho}{dt} = \text{div}(\rho \nabla V) + \Delta \rho, \quad \rho(0) = \rho_0. \tag{4.2}
\]

A classical result says that under quite weak convexity and smoothness conditions on \( V \), the unique stationary solution of (4.2) is equal to \( \pi \), and that convergence to \( \pi \) is exponentially fast (see for example Pavliotis, 2014, Chapter 4). These attractive properties have spawned a range of sampling algorithms targeting \( \pi \) based on time discretizations of the process in (4.1). Notably, the Unadjusted Langevin Algorithm (ULA) and its Metropolis adjusted counterpart MALA have received much attention.

The Unadjusted Langevin Algorithm is simply an explicit Euler discretization of (4.1): for a time-step \( h > 0 \) and for \( k \geq 0 \),

\[
    X_{h}^{k+1} = X_{h}^{k} - h\nabla V(X_{h}^{k}) + \sqrt{2h}\eta^{k+1}, \quad X_{0}^{0} = X_0, \tag{4.3}
\]

where \( (\eta^k)_{k \geq 1} \) is a sequence of independent \( \mathcal{N}(0, I_d) \) random variables and \( I_d \) is the \( d \)-dimensional identity matrix. In MALA, \( X_{h}^{k+1} \) is either accepted or rejected in a Metropolis step with the purpose of removing the asymptotic bias of ULA stemming from discretization error.
Originating with Roberts and Tweedie (1996), there has been a lot of interest in quantifying the performance of these algorithms, with early work primarily focusing on MALA (see e.g. Jarner and Hansen, 2000; Roberts and Stramer, 2002; Pillai et al., 2012; Xifara et al., 2014). It was not until Dalalyan (2014), who gave precise bounds for the total variation distance between the law of $X^k_h$ and $\pi$ in terms of $d$, $k$, and $h$, that ULA garnered similar attention. His results were further improved and extended to other metrics and discrepancies by Durmus and Moulines (2016b, 2017); Cheng and Bartlett (2017); Dalalyan (2017). For instance, Dalalyan and Karagulyan (2017) show that if $V$ is strongly convex and has Lipschitz continuous gradient, then $\Omega(d/\varepsilon^2)$ iterations are sufficient for ULA to achieve an error of $\varepsilon$ in the 2-Wasserstein distance. Similar results also hold in situations where only a (sufficiently regular) approximation of the gradient is available.

In what follows, we will view Langevin-based Monte Carlo through the lens of Wasserstein gradient flow, and show that this perspective can lead to interesting results on the computational complexity of such algorithms. Wasserstein gradient flow was also used by Cheng and Bartlett (2017) as a theoretical tool to study ULA, but our approach makes closer connections to the operator splitting literature, and as such leads to different results. We hope that further connections can have methodological implications in these fields, by considering the wide variety of JKO schemes, splitting schemes, and Langevin Monte Carlo algorithms that exist.

The rest of this chapter is structured as follows. Section 4.1.1 defines the notation and states some important definitions, Section 4.2 reviews some concepts from the Wasserstein gradient flow literature, Section 4.3 briefly discusses the operator splitting approach to solving PDEs, Section 4.4 establishes connections between Wasserstein gradient flow, operator splitting and Langevin Monte Carlo and includes some convergence results on the proximal version of the ULA algorithm, and Section 4.5 concludes. Proofs are given in Appendix C.

### 4.1.1 Notation and definitions

Let $\| \cdot \|_p$ be the $\ell_p$-norm on $\mathbb{R}^d$, unless $p = 2$, in which case it reduces to the Euclidean distance and is denoted by $\| \cdot \|$. Define $\mathcal{P}_2(\mathbb{R}^d)$ to be the set of probability measures on
\( \mathbb{R}^d \) with finite second moments with respect to the Euclidean distance. The 2-Wasserstein distance is a metric on \( \mathcal{P}_1(\mathbb{R}^d) \), and is for any \( \mu, v \in \mathcal{P}_1(\mathbb{R}^d) \) defined by

\[
\mathcal{W}_2(\mu, v) = \left( \inf_{\gamma \in \Gamma(\mu, v)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \gamma(x, y) \right)^{\frac{1}{2}},
\]

where \( \Gamma(\mu, v) \) is the set of all joint distributions with marginals \( \mu \) and \( v \). A desirable feature of the 2-Wasserstein distance is that \( \mathcal{W}_2(\mu_n, \mu) \to 0 \) as \( n \to \infty \) if and only if \( \mu_n \) converges weakly to \( \mu \) and the corresponding sequence of second moments also converges (Villani, 2008, Theorem 6.9).

The entropy and potential energy functionals, \( \rho \mapsto \mathcal{H}(\rho) \) and \( \rho \mapsto \mathcal{V}(\rho) \) respectively, are given by

\[
\mathcal{H}(\rho) = \begin{cases} 
\int \log \rho \, d\rho & \text{for } \rho \ll \mu_{\text{Leb}}, \\
+\infty & \text{otherwise},
\end{cases}
\]

where \( \mu_{\text{Leb}} \) denotes the Lebesgue measure on \( \mathbb{R}^d \), and

\[
\mathcal{V}(\rho) = \int V d\rho.
\]

The relative energy functional \( \rho \mapsto \mathcal{H}(\rho|\pi) \), also called the KL-divergence, is given by

\[
\mathcal{H}(\rho|\pi) = \mathcal{H}(\rho) + \mathcal{V}(\rho) + \log Z.
\]

An important concept in optimal transport, which will play a significant role later, is the notion of displacement convexity. A functional \( \rho \mapsto \mathcal{F}(\rho) \) is said to be \( \lambda \)-displacement convex for some \( \lambda \in \mathbb{R} \) if, for all \( t \in [0, 1] \),

\[
\mathcal{F}(\mu_t) \leq (1 - t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\lambda}{2}t(1 - t)\mathcal{W}_2^2(\mu_0, \mu_1)
\]

for any constant speed geodesic \( \mu : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d) \). A curve \( \mu : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d) \) is a constant speed geodesic if, for any \( 0 \leq s \leq t \leq 1 \), we have that

\[
\mathcal{W}_2(\mu_s, \mu_t) = (t - s)\mathcal{W}_2(\mu_0, \mu_1).
\]
We use the following notation for the density of a Gaussian distribution with zero mean and covariance matrix $2tI_d$:

$$
\phi_t(x) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{\|x\|^2}{4t}\right). \tag{4.8}
$$

By a Markov operator, we mean a linear functional $R$ that maps the set of non-negative Lebesgue integrable functions into itself. A family of Markov operators $(R_t)_{t \geq 0}$ is called a Markov semigroup if $R_0$ is the identity map, $R_{t+s} = R_t R_s$ for any $s, t \geq 0$, and the map $t \mapsto R_t f$ is continuous for any non-negative and Lebesgue integrable $f$.

### 4.2 Wasserstein gradient flow

The theory of gradient flows in the space of probability measures was pioneered by Ambrosio, Gigli and Savaré in their book Ambrosio et al. (2005), generalizing the variational structure Jordan et al. (1998) had used to describe the diffusion and Fokker–Planck equations. With Langevin Monte Carlo in mind, we provide only a brief introduction to this theory, and refer to the aforementioned references and the accessible review of Santambrogio (2016) for further details.

We first consider continuous time flows, which will lead to a useful perspective on generalizations of the continuous time processes in (4.1) and (4.2). Secondly, we consider the time discretizations through which the existence and uniqueness of gradient flows are typically established. Although they were originally introduced as theoretical tools in the literature, it will later become clear that Langevin Monte Carlo in fact numerically approximates such a time discretization.

#### 4.2.1 Continuous time flows

In Euclidean space, a curve $x: [0, \infty) \to \mathbb{R}^d$ is the gradient flow, or steepest descent, of a differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ if

$$
\frac{dx}{dt} = -\nabla f(x), \quad x(0) = x_0. \tag{4.9}
$$
By analogy, one can interpret the gradient flow of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ to be a curve $\rho : [0, \infty) \to \mathcal{P}_2(\mathbb{R}^d)$ that satisfies
\[
\frac{d\rho}{dt} = -\nabla_{W_2} F(\rho), \quad \rho(0) = \rho_o,
\] (4.10)
for some generalized notion of gradient $\nabla_{W_2}$ in terms of the $W_2$ metric. For sufficiently regular $\rho$ and $F$, $\nabla_{W_2} F(\rho)$ corresponds to $-\text{div}(\rho \nabla \frac{\delta F}{\delta \rho})$, where $\frac{\delta F}{\delta \rho}$ is the first variation of $F$. Applied to the functional of interest, namely $F(\rho) = \mathcal{H}(\rho|\pi)$, one has that $\frac{\delta F}{\delta \rho} = V + \log \rho + 1$. Thus, if $V$ is differentiable one recovers (4.2); see e.g. Ambrosio et al. (2005, Lemma 10.4.1).

Due to the technically challenging nature of defining Wasserstein gradients this way when $V$ is not differentiable, we instead adopt the definition given in Ambrosio et al. (2009), inspired by the characterization of gradient flows in terms of evolution variational inequalities (EVIs) shown in Ambrosio et al. (2005, Theorem 11.1.4). In particular, we say that a continuous curve $\rho : (0, +\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is a gradient flow of a $\lambda$-displacement convex functional $F$ if
\[
\frac{d}{dt} \mathcal{W}_2^2(\rho(t), v) + \frac{\lambda}{2} \mathcal{W}_2^2(\rho(t), v) + F(\rho(t)) \leq F(v),
\] (4.11)
holds in the sense of distributions, for all $v \in D(F) = \{ \mu \in \mathcal{P}_2(\mathbb{R}^d) : F(\mu) < +\infty \}$. The flow is said to start from $\rho_o$ if $\mathcal{W}_2(\rho(t), \rho_o) \to 0$ as $t \to 0$. Here, “in the sense of distributions” means that for all infinitely differentiable and compactly supported test functions, denoted $f \in C_c^\infty((0, \infty); \mathbb{R})$, such that $f \geq 0$, we have
\[
-\frac{1}{2} \int_0^\infty \mathcal{W}_2^2(\rho(t), v)f(t)dt \leq \int_0^\infty \left[ F(v) - F(\rho(t)) - \frac{\lambda}{2} \mathcal{W}_2^2(\rho(t), v) \right] f(t)dt.
\] (4.12)

The connection between (4.11) and (4.12) can be seen by imagining the left hand side of (4.12) being integrated by parts.

One of the most attractive features of gradient flows are their convergence properties. For any $\lambda$-displacement convex functional $F$ with $\lambda > 0$, the map $\rho \mapsto F(\rho)$ has a unique minimum $\bar{\rho}$, and Theorem 11.2.1 of Ambrosio et al. (2005) states that there exists a unique
gradient flow $t \mapsto \rho(t)$, which satisfies
\[
\mathcal{W}_2(\rho(t), \bar{\rho}) \leq \mathcal{W}_2(\rho_o, \bar{\rho}) e^{-\lambda t} \quad \text{and} \quad \mathcal{F}(\rho(t)) - \mathcal{F}(\bar{\rho}) \leq \left[ \mathcal{F}(\rho_o) - \mathcal{F}(\bar{\rho}) \right] e^{-\lambda t},
\]
or any $t \geq 0$. Convergence results also exist in the case where $\lambda = 0$, but do not yield the exponential convergence observed above.

This result can be applied to the relative entropy by making the following observations: when $V$ is $\lambda$-strongly convex with $\lambda > 0$, it follows that $\rho \mapsto \mathcal{V}(\rho)$ is $\lambda$-displacement convex (Ambrosio et al., 2005, Proposition 9.3.2). In turn, this implies that $\rho \mapsto \mathcal{H}(\rho|\pi)$ is $\lambda$-displacement convex. Recall that $\mathcal{H}(\rho|\pi) \geq 0$ for any $\rho$, and that $\rho \mapsto \mathcal{H}(\rho|\pi)$ is uniquely minimized at $\pi$ due to the strict convexity of the function $x \mapsto x \log x$ for $x > 0$ appearing in $\mathcal{H}(\rho)$, and Jensen’s inequality. The result in (4.13) can then be formulated as
\[
\mathcal{W}_2(\rho(t), \pi) \leq \mathcal{W}_2(\rho_o, \pi) e^{-\lambda t} \quad \text{and} \quad \mathcal{H}(\rho(t)|\pi) \leq \mathcal{H}(\rho_o|\pi) e^{-\lambda t}. \tag{4.14}
\]

This is a more general statement of the exponential convergence to $\pi$ of the solution to the Fokker–Planck equation mentioned in the introduction, and is as such one of the main motivations for studying Langevin Monte Carlo algorithms.

### 4.2.2 Time discretized flows

An important theoretical tool in establishing the existence of gradient flows is the minimizing movement scheme, often also called the JKO scheme. For a time-step $h > 0$, $k \geq 0$, and $\rho_h^0 = \rho_o$, consider the iterated minimization problems
\[
\rho_h^{k+1} = \arg\min_{\rho \in \mathcal{P}_1(\mathbb{R}^d)} \mathcal{F}(\rho) + \frac{1}{2h} \mathcal{W}_2^2(\rho, \rho_h^k). \tag{4.15}
\]

Such minimizers exist and are unique under weak assumptions, such as lower semi-continuity and (strong) displacement convexity of $\mathcal{F}$ (see e.g. Ambrosio et al., 2009, Proposition 4.2). Both of these conditions hold for the relative entropy functional $\rho \mapsto \mathcal{H}(\rho|\pi)$ when $V$ is convex: the first property holds in more generality and is well-known, whereas the second was proved in McCann (1997).
In the Euclidean setting, the sequence \((x^k_h)_{k \geq 0}\) is an implicit Euler discretization with step-size \(h\) of the gradient flow of \(f : \mathbb{R}^d \to \mathbb{R}\) given in (4.9) with initial condition \(x^0_h = x_0\) if
\[
x^{k+1}_h = \arg\min_{y \in \mathbb{R}^d} f(y) + \frac{1}{2h}||x^k_h - y||^2.
\] (4.16)

The map defined by the right hand side of (4.16) is often written \(\text{prox}^h_f(x^k_h)\) in the optimization literature, and is referred to as the proximal operator (see e.g. Parikh and Boyd, 2014).

By analogy, the JKO scheme (4.15) can be seen as an implicit Euler discretization of the flow in (4.10). It was this time discretization scheme applied to the functional \(\rho \mapsto \mathcal{H}(\rho | \pi)\) that Jordan et al. (1998) employed, showing that the interpolation
\[
\rho^h(t) = \rho^{k+1}_h \quad \text{for } t \in (kh, (k + 1)h]
\] (4.17)
converges (in some formal sense) to the solution of the Fokker–Planck equation as \(h \to 0\), in the case where \(V\) is smooth and satisfies certain growth conditions.

Building on results by Cépa (1998), Ambrosio et al. (2009) used a minimizing movement scheme to show existence and uniqueness of the gradient flow of the relative entropy functional given any convex \(V\). In particular, they show that there exists a semigroup \((P_t)_{t \geq 0}\) and a unique Markov family \(\{\mathbb{P}_x : x \in \mathbb{R}^d\}\) of probability measures on \((\mathbb{R}^d)^{[0, +\infty)}\) such that \(\mathbb{E}_df(X_t) = P_t f(x)\) for all bounded Borel functions \(f\) and all \(x \in \mathbb{R}^d\). Moreover, it is shown that \(\{\mathbb{P}_x : x \in \mathbb{R}^d\}\) is reversible with respect to \(\pi\), and that \(\pi\) is uniquely invariant for \((P_t)_{t \geq 0}\). Restricting \((P_t)_{t \geq 0}\) to indicator functions of Borel sets \(B \in \mathcal{B}(\mathbb{R}^d)\), we define \((R_t)_{t \geq 0}\) by \(R_t \rho_\circ(B) = \int P_t 1_B d\rho_\circ\). The process \(\rho(t) = R_t \rho_\circ\) then uniquely satisfies (4.11) and the associated properties outlined in the previous section.

After originally being introduced as a theoretical tool, there has recently been interest in developing numerical implementations of the JKO scheme for solving PDEs. Several Eulerian grid-based approaches exist, see e.g. Burger et al. (2012); Carrillo et al. (2015a); Peyré (2015). By virtue of being grid-based, these have limited application in the high-dimensional sampling setting.

It will later be seen that Langevin-based Monte Carlo can be considered a Lagrangian
scheme using a particle approximation to the gradient flow. Other Lagrangian approaches have been considered by e.g. Carrillo et al. (2015b); Benamou et al. (2016); Carrillo et al. (2017). These methods are typically adapted to accurately solving PDEs in two or three dimensions, and do not scale well with $d$. For instance, Carrillo et al. (2017) used the modified relative entropy functional

$$
\mathcal{F}_\gamma(\rho) = \int \log(\varphi_\gamma * \rho) d\rho + \int V d\rho + \log Z,
$$

where $\varphi_\gamma = \gamma^{-d} \phi(x/\gamma)$ denotes a mollifier, typically a Gaussian kernel with standard deviation $\gamma > 0$. This modification makes the functional well-behaved when evaluated at an empirical measure, with the first term providing a kernel-based estimate of the entropy of the underlying distribution. For small time steps $h$, their algorithm reduces to solving a system of ODEs to evolve the particles in the empirical measure. The application of this approach to the high-dimensional setting is limited by the kernel-based estimate of entropy.

### 4.3 Operator Splitting

In the previous section, we alluded to the idea that Langevin Monte Carlo numerically approximates the time discretizations used to theoretically study Wasserstein gradient flows. Before making this connection clear, we first need to introduce the concept of operator splitting.

Consider the generic Cauchy problem

$$
\frac{df}{dt} = \mathcal{A}(f), \quad f(0) = f_0, \quad (4.18)
$$

with solution given by $f(t) = Sf_0$ in semigroup notation. In many situations, the operator $\mathcal{A}$ can be split into the sum of two simpler operators: $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$. Let $f_j(t) = S^j f_0$ for $j = 1, 2$ denote the solutions to the problems

$$
\frac{df_j}{dt} = \mathcal{A}_j(f_j), \quad f_j(0) = f_0. \quad (4.19)
$$
One can hope to estimate the solution $f(t)$ of (4.18) via $f(t) \approx (S_{t/n}^2 S_{t/n}^1)^n f_0$ for some large positive integer $n$, which can be justified if a Lie–Trotter–Kato product formula of the form

$$f(t) = \lim_{n \to +\infty} (S_{t/n}^2 S_{t/n}^1)^n f_0$$  \hspace{1cm} (4.20)


Returning to the Fokker–Planck equation (4.2), there is a natural split between the transport part of the equation:

$$\frac{d\rho}{dt} = \text{div}(\rho \nabla V), \quad \rho(0) = \rho_o,$$  \hspace{1cm} (4.21)

and the diffusion part:

$$\frac{d\rho}{dt} = \Delta \rho, \quad \rho(0) = \rho_o.$$  \hspace{1cm} (4.22)

In his PhD thesis, Stojković (2011) considers such a split for the Fokker–Planck equation with smooth drift satisfying a monotonicity property, but which is not necessarily a gradient. Bowles and Agueh (2015) also consider this split for the fractional Fokker–Planck equation, where the Laplacian in the diffusion equation (4.22) is substituted for a fractional Laplacian. In both of these works, operator splitting is introduced as a theoretical tool to establish the existence of solutions to generalized Fokker–Planck equations, but they do not consider numerical aspects nor the general case of convex $V$.

The splitting interpretation carries over to the Wasserstein gradient flow formulation, where the transport equation (4.21) can be interpreted as the gradient flow of the potential energy functional $\rho \mapsto V(\rho)$, and the diffusion equation (4.22) can be interpreted as the gradient flow of the entropy functional $\rho \mapsto \mathcal{H}(\rho)$. We now take a brief closer look at these two gradient flows.
4.3.1 The transport equation

In addition to the formulation in (4.11), the gradient flow of \( \rho \mapsto V(\rho) \) can be characterized by the semigroup \((T_t)_{t \geq 0}\), induced by the differential inclusion

\[
\frac{d}{dt} T_t(x) \in -\partial V(T_t(x)), \quad T_0(x) = x \quad \text{for all } x \text{ s.t. } V(x) < +\infty. \tag{4.23}
\]

According to Theorem 11.2.3 of Ambrosio et al. (2005), there exists a unique gradient flow of \( \rho \mapsto V(\rho) \) and solution to (4.23). This gradient flow satisfies \( \rho(t) = (T_t)_# \rho_0 \), where \((T_t)_# \) denotes the push-forward map associated with \( T_t \).

The corresponding JKO scheme performs minimizations of the form

\[
\rho_h^{k+1} = \arg\min_{\rho \in \mathcal{P}_1(\mathbb{R}^d)} \mathcal{V}(\rho) + \frac{1}{2h} \mathcal{W}_2(\rho, \rho_h^k).
\]

By the proof of Proposition 10.4.2 in Ambrosio et al. (2005), it is clear that these steps are well-defined. Moreover, the map \( T_h(x) = \prox_{\mathcal{V}}^h(x) \) is such that \( \rho_h^{k+1} = (T_h)_# \rho_h^k \). Since the proximal operator satisfies \( y = \prox_{\mathcal{V}}^h(x) \iff (x - y) / h \in \partial \mathcal{V}(x) \) (see e.g. Parikh and Boyd, 2014), this can be seen as an implicit Euler step for the evolution of \( T_t \) given in (4.23).

4.3.2 The diffusion equation

The classical diffusion equation (4.22), also known as the heat equation, was first described as the gradient flow of the entropy functional \( \rho \mapsto \mathcal{H}(\rho) \) on the set of densities in \( \mathcal{P}_2(\mathbb{R}^d) \) by Jordan et al. (1998). Note that \( \mathcal{H}(\rho) \) is the negative Gibbs–Boltzmann entropy of \( \rho \). As pointed out in the aforementioned paper, the interpretation of the diffusion equation as the gradient flow of \( \mathcal{H} \) therefore provides a natural interpretation of diffusion as the tendency of a system to maximize entropy.

Unlike the other gradient flows we have discussed, the flow of \( \rho \mapsto \mathcal{H}(\rho) \) is known in closed form: it is well-known that the solution of the diffusion equation (4.22) is given by the density \( \rho(t) = \varphi_t * \rho_0 \), where \( \varphi_t \) is the Gaussian kernel defined in (4.8).
4.4 Proximal Langevin Monte Carlo

We are now ready to describe connections between JKO discretized gradient flows, operator splitting, and Langevin-based Monte Carlo algorithms. For a time-step $h > 0$ and for $k \geq 0$, consider the iterative scheme

\[ \rho_h^{k+1/2} = (\mathcal{T}_h)^* \rho_h^k, \quad \rho_h^{k+1} = \varphi_h * \rho_h^{k+1/2}, \]  

(4.24)

which can be seen as alternating between performing a JKO step for the gradient flow of $\rho \mapsto \mathcal{V}(\rho)$ and solving the exact gradient flow of $\rho \mapsto \mathcal{H}(\rho)$. Taking instead the particle perspective, let $X_h^k \sim \rho_o$ and perform

\[ X_h^{k+1/2} = \mathcal{T}_h(X_h^k) = \text{prox}_\nu^k(X_h^k), \quad X_h^{k+1} = X_h^{k+1/2} + \sqrt{2h}\eta_h^{k+1}, \]  

(4.25)

where $(\eta_h^k)_{k \geq 1}$ is a sequence of independent $\mathcal{N}(0, I_d)$ random variables. For each $k$, the laws of $X_h^{k+1/2}$ and $X_h^{k+1}$ are equal to $\rho_h^{k+1/2}$ and $\rho_h^{k+1}$ respectively. A generalization of this algorithm was proposed by Pereyra (2016) and studied further in Durmus et al. (2016).

Note that $\text{prox}_\nu^k(x) = x - h\nabla M^h_V(x)$, where

\[ M^h_V(x) = \inf_{y \in \mathbb{R}^d} \left\{ \mathcal{V}(y) + \frac{1}{2h} ||x - y||^2 \right\} \]

is the Moreau–Yosida regularization of $V$. Moreover, in the case where $V$ is twice differentiable with positive definite Hessian $D^2V(x)$ for every $x \in \mathbb{R}^d$, it is known that $\text{prox}_\nu^h(x) = x - h\nabla V(x) + o(h)$ as $h \to 0$ (see e.g. Parikh and Boyd, 2014, Section 3.3). Hence, for small $h$, the steps in (4.25) can be thought of as approximating the Unadjusted Langevin Algorithm.

4.4.1 Convergence analysis

We follow the approach of Clément and Maas (2011), which itself is an adaptation of the methods in Ambrosio et al. (2005, Chapter 4), to establish that the scheme in (4.24) satisfies a Lie–Trotter–Kato formula. We will also derive an upper bound on the 2-Wasserstein
distance between the interpolation \( \rho_h^k(t) = \rho_h^{k+1} \) for \( t \in (kh, (k+1)h) \) and the gradient flow \( \rho(t) \) of \( \rho \mapsto \mathcal{H}(\rho|\pi) \). In turn, this allows us to bound the quantity of interest, \( \mathcal{W}_2(\rho_h^k(t), \pi) \). Before stating the main results, we introduce some notation.

For any \( n \geq 1 \) and any \( 0 \leq k \leq n-1 \), define the quantities

\[
\delta_h^{k+1} = \mathcal{V}(\rho_h^{k+1}) - \mathcal{V}(\rho_h^{k+1/2}), \quad \Delta_h^{k+1} = \sum_{j=1}^{k+1} \delta_h^j. \tag{4.26}
\]

Note that \( \delta_h^{k+1} \) can also be expressed

\[
\delta_h^{k+1} = \mathbb{E}V(X + \eta) - \mathbb{E}V(X), \tag{4.27}
\]

where \( X \sim \rho_h^{k+1/2} \) and \( \eta \sim \mathcal{N}(0, 2hI_4) \) independently. By convexity of \( V \) and Jensen's inequality, it is clear that \( \delta_h^{k+1} \geq \mathbb{E}V(\mathbb{E}(X + \eta|X)) - \mathbb{E}V(X) \geq 0 \). The next results show that controlling these quantities is sufficient to establish convergence. We also remark that if one has access to independent runs of the algorithm given in (4.25), one can estimate \( \delta_h^{k+1} \) by averaging \( V(X_h^{k+1}) - V(X_h^{k}) \) across those runs.

**Theorem 4.4.1.** Let \( (\rho_{hn}^m(t))_{m \geq 1} \) be a sequence of discrete solutions generated from \( \rho_o \), such that \( h_m \Delta_m \to 0 \) and \( h_m m \to T \) for some \( T > 0 \), as \( m \to \infty \). Then, \( \rho_{hn}^m(t) \) converges uniformly on \([0, T]\) to \( \rho(t) \), the gradient flow of \( \rho \mapsto \mathcal{H}(\rho|\pi) \) started from \( \rho_o \). Moreover, if \( h > 0 \) and \( n \geq 1 \) are such that \( hn \leq T \), then for any \( t \in [0, hn] \),

\[
\mathcal{W}_2(\rho_t^h(t), \rho(t)) \leq \sqrt{6h \left( \mathcal{H}(\rho_o|\pi) + \Delta_h^n \right)}. \tag{4.28}
\]

The corollary below follows from combining (4.14) and (4.28) via the triangle inequality.

**Corollary 4.4.1.** Suppose \( V \) is \( \lambda \)-strongly convex. Then, under the assumptions of Theorem 4.4.1, we have

\[
\mathcal{W}_2(\rho_t^h(t), \pi) \leq \sqrt{6h \left( \mathcal{H}(\rho_o|\pi) + \Delta_h^n \right)} + \mathcal{W}_2(\rho_o, \pi)e^{-\lambda t}, \tag{4.29}
\]

for any \( t \in [0, hn] \), where \( h > 0 \) and \( n \geq 1 \).
4.4.2 Explicit rates

It is clear that the rate at which \( h \Delta_h^n \to 0 \) as \( h \to 0 \) is crucial in determining the quality of the approximation \( \rho^h(t) \). Under some assumptions on \( \rho \) and \( V \), we can obtain explicit bounds on \( \Delta_h^n \) in terms of \( h, n, \text{ and } d \), as will be seen below.

Suppose \( V = f + g \), where \( f \) is \( \lambda \)-strongly convex and has Lipschitz continuous gradient, and \( g \) is convex and Lipschitz. That is, assume that there exist \( M(d) \) and \( L(d) \) such that for all \( x, y \in \mathbb{R}^d \),

\[
\| \nabla f(x) - \nabla f(y) \| \leq M(d) \| x - y \| \tag{4.30}
\]

\[
|g(x) - g(y)| \leq L(d) \| x - y \|, \tag{4.31}
\]

where the notation \( M(d) \) and \( L(d) \) reflects potential dependence of the Lipschitz constants on dimension. Under this assumption, we can bound \( \delta_h^{k+1} \) as follows:

\[
\mathbb{E} V(X + \eta) - \mathbb{E} V(X) = \mathbb{E} [f(X + \eta) - f(X)] + \mathbb{E} [g(X + \eta) - g(X)] \leq \mathbb{E} \left[ \nabla f(X)^\top \eta + \frac{M(d)}{2} \| \eta \|^2 \right] + L(d) \mathbb{E} \| \eta \| \tag{4.32}
\]

\[
\leq M(d) \| x \| + \frac{M(d)}{2} \| x \|^2 + L(d) \| x \| \tag{4.33}
\]

\[
\leq M(d) h \| x \| + L(d) \sqrt{2hd}, \tag{4.34}
\]

where (4.33) follows from the basic property that

\[
f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{M(d)}{2} \| x - y \|^2, \tag{4.35}
\]

for all \( x, y \in \mathbb{R}^d \), see for example Nesterov (2013). Then, \( h \Delta_h^n \leq M(d) h d \cdot h n + L(d) \sqrt{2hd} \cdot hn \). Hence, for any \( T > 0 \) we could take \( h_m = T / m \) and satisfy the conditions of Corollary 4.4.1.

Next, we can use these bounds to derive explicit rates for \( n \) and \( h \) that yield a desired approximation error. When selecting the initial distribution, it is not unreasonable to assume that one can choose \( \rho_0 \) such that \( \mathcal{W}_2(\rho_0, \pi) = \mathcal{O}(\sqrt{d}) \) and \( \mathcal{H}(\rho_0 | \pi) = \mathcal{O}(d) \). See Appendix C for justifications and an explicit example where these assumptions hold.

Now, if we want \( \mathcal{W}_2(\rho^h(hn), \pi) = \mathcal{O}(\epsilon) \) for a threshold \( \epsilon > 0 \), we could require that
both $h\mathcal{H}(\rho_0|\pi) + h\Delta^n_h = \mathcal{O}(\varepsilon^3)$ and $W_2(\rho_0, \pi)e^{-\lambda_h n} = \mathcal{O}(\varepsilon)$. Under the assumptions above, to ensure $W_2(\rho_0, \pi)e^{-\lambda_h n} = \mathcal{O}(\varepsilon)$, it is sufficient to take $hn = \Omega(\log(d/\varepsilon^2))$. To get $h\mathcal{H}(\rho_0|\pi) = \mathcal{O}(\varepsilon^2)$, one can require that $h = \mathcal{O}(\varepsilon^2/d)$. Lastly, to get $h\Delta^n_h = \mathcal{O}(\varepsilon^2)$, one can in turn require that both $M(d)h\log(\sqrt{d}/\varepsilon) = \mathcal{O}(\varepsilon^4)$ and $L(d)\sqrt{2hd}\log(\sqrt{d}/\varepsilon) = \mathcal{O}(\varepsilon^4)$. The former can be achieved if

$$n = \Omega\left(\frac{dM(d)\log(\sqrt{d}/\varepsilon)^2}{\varepsilon^4}\right) \quad \text{and} \quad h = \mathcal{O}\left(\frac{\varepsilon^4}{dM(d)\log(\sqrt{d}/\varepsilon)}\right), \quad (4.36)$$

while maintaining $hn = \Omega(\log(d/\varepsilon^2))$. Similarly, the latter can be achieved if

$$n = \Omega\left(\frac{dL(d)^2\log(\sqrt{d}/\varepsilon)^3}{\varepsilon^4}\right) \quad \text{and} \quad h = \mathcal{O}\left(\frac{\varepsilon^4}{dL(d)^2\log(\sqrt{d}/\varepsilon)^3}\right), \quad (4.37)$$

still keeping $hn = \Omega(\log(d/\varepsilon^2))$.

In the case where $g = o$ (or equivalently $L(d) = o$) and $M(d) = \mathcal{O}(1)$, we recover the assumptions on $V$ that were made in e.g. Dalalyan (2017); Dalalyan and Karagulyan (2017). Using (4.36), we see that $n = \Omega(d\varepsilon^{-2}\log(d\varepsilon^{-2}))$ iterations with a step-size of $h = \log(d/\varepsilon^2)/n$ are sufficient to achieve a 2-Wasserstein error of $\mathcal{O}(\varepsilon)$. Up to log-terms, this is the same rate as those derived for ULA in the aforementioned papers.

In the case where $g(x) \propto ||x||$, so that $L(d) = \mathcal{O}(\sqrt{d})$, we get that $n = \Omega(d^2/\varepsilon^4)$ iterations are sufficient (ignoring the log-terms). This improves upon the recent results of Grappin (2018), who showed that if additionally $f$ is quadratic, then $n = \Omega(d^4/\varepsilon^4)$ iterations are sufficient to yield a 2-Wasserstein error of $\mathcal{O}(\varepsilon)$. Comparing to the remark accompanying Theorem 3 of Durmus et al. (2016), our results appears less sharp than the TV bounds they derive, in which $n$ depends linearly on $d$ (up to log-terms) whenever $V$ is strongly convex. As can be seen in Appendix C, this likely stems from not optimally accounting for $\lambda$-displacement convexity in Lemma C.1.4 of Appendix C.
4.5 Discussion

In this chapter, we have developed novel connections between the fields of Wasserstein gradient flow, operator splitting, and Langevin Monte Carlo. We have demonstrated that the gradient flow perspective allows us to derive new convergence results about a proximal version of the Unadjusted Langevin Algorithm. Under certain assumptions on the potential $V$, we derive results that are on par with the contemporary literature on ULA. However, we point out that there is room for improvement in our current proofs. In particular, they could be improved by better accounting for the condition that $V$ is $\lambda$-strongly convex, allowing us to obtain sharper bounds when that assumption is present. On the other hand, the proof of Theorem 4.4.1 generalizes to any convex $V$. Hence, to obtain control over the proximal ULA algorithm in such a case, one would only need to formulate conditions under which one can still derive a rate of convergence of the exact gradient flow to $\pi$, though one should no longer expect this convergence to be exponentially fast. Some recent progress in this direction based on Lojasiewicz inequalities was made by Blanchet and Bolte (2016).

We also hope that these connections can have implications on methodology. The many other splitting schemes discussed by Holden et al. (2010) and in the optimization literature can potentially lead to new sampling algorithms. The same holds for other numerical schemes, such as the alternative JKO algorithm developed by Legendre and Turinici (2017). For the Fokker–Planck equation, they show that their new scheme is second-order convergent, improving the original JKO scheme’s first-order convergence. Recently, Plazzotta (2018) developed a variational formulation of the BDF2 scheme applicable to the estimation of gradient flows. It is also likely that the growing literature on Langevin Monte Carlo and its variations can lead to new time discretization schemes that are of both practical and theoretical interest to the gradient flow community.
5

Schrödinger bridge samplers

5.1 Introduction

Let \( \pi \) be a distribution which admits a density, with respect to some dominating measure on a measurable space \( (E, \mathcal{E}) \), that can only be evaluated pointwise up to a normalizing constant \( Z \). We are interested in approximating expectations with respect to \( \pi \) as well as the value of \( Z \). State-of-the-art Monte Carlo methods to address this problem include Annealed Importance Sampling (AIS; Crooks, 1998; Neal, 2001) and Sequential Monte Carlo (SMC; Del Moral et al., 2006). The basis of these methods is to simulate \( N \) independent non-homogeneous Markov chains\(^1\) with initial distribution \( \pi_o \) and transition kernels \( \{M_t\}_{t \in [0:T]} \), designed such that the marginal distribution of each Markov chain at time \( T \) is approximately equal to \( \pi \). However, this marginal distribution is typically not analytically available, prohibiting its direct application as a proposal distribution within importance

\(^1\)We can also use an interacting particle system (Del Moral et al., 2006).
sampling. In AIS and SMC, this intractability is circumvented by introducing an appropriate auxiliary target distribution on the path space $E^{T+1}$ whose marginal at time $T$ coincides with $\pi$ and with respect to which importance weights can be calculated. The resulting schemes also provide estimates of the normalizing constant $Z$.

These methods have found numerous applications in physics and statistics, but can perform poorly when the marginal distribution of the samples at time $T$ differs significantly from $\pi$, resulting in importance weights with high (or potentially infinite) variance. Building upon previous contributions for inference in partially observed diffusions and state-space models (Richard and Zhang, 2007; Kappen and Ruiz, 2016; Guarniero et al., 2017), the controlled SMC sampler methodology of Heng et al. (2017) uses ideas from optimal control to iteratively modify the initial distribution and transition kernels of the reference Markov process to reduce the Kullback-Leibler (KL) divergence between the induced path distribution and the auxiliary target distribution on $E^{T+1}$. When applicable, controlled SMC samplers demonstrate clear improvements over AIS and SMC. However, a significant limitation of this approach is that one must be able to sample from a modified initial distribution. Practically, this means that $\pi_0$ must be a mixture of Gaussians and that the value functions of the underlying optimal control problem need to be approximated by quadratic forms.

We propose here an alternative approach which is more widely applicable. First, we only modify the transition kernels and not the initial distribution. Second, instead of minimizing the KL divergence on path space $E^{T+1}$ with respect to a fixed auxiliary target distribution, the auxiliary target is itself being optimized across iterations. We describe our algorithm as an approximation of the iterative proportional fitting procedure (IPFP), which is known to converge to the solution of the Schrödinger bridge problem (Léonard, 2014). In finite state-spaces, the steps of IPFP, which is also known as Sinkhorn’s algorithm, can be computed exactly (Deming and Stephan, 1940; Sinkhorn, 1967; Kullback, 1968). Although the algorithm can be formally extended to continuous state-spaces (Kullback, 1968; Beghi, 1996; Rüschendorf, 1995), the corresponding iterations are intractable in all but trivial scenarios. Recent computational approaches that have been proposed to address similar problems to the one we consider here either rely on deterministic (Chen et al., 2016) or stochastic (Reich, 2018) discretization of the space using $N$ atoms, and then fall
back on the finite state-space IPFP algorithm. Such procedures therefore have computational complexity of $O(N^2)$ at each iteration. Our algorithm instead relies on approximations of IPFP through regression techniques in the spirit of Heng et al. (2017), and admits a linear complexity in the number of Monte Carlo samples.

The rest of this chapter is organized as follows. In the remaining part of Section 5.1, we define our notation, formalize the problem statement, review Sequential Monte Carlo samplers and their limitations, and give a brief overview of the proposed method. In Section 5.2, we discuss Schrödinger bridges and their various formulations, and introduce numerical algorithms to approximate them. In Section 5.3, we discuss our main computational contribution, which we term the sequential Schrödinger bridge sampler. In Section 5.4, we discuss connections between the Schrödinger bridge problem and various other topics. Section 5.5 contains numerical experiments and Section 5.6 concludes.

5.1.1 Notation

Given integers $n \leq m$ and a sequence $\{x_i\}_{i \in \mathbb{N}}$, we define the set $[n : m] = \{n, \ldots, m\}$ and write the subsequence $x_{n:m} = (x_n, x_{n+1}, \ldots, x_m)$. Let $(E, \mathcal{E})$ be an arbitrary measurable space. Given $\mu, \nu \in \mathcal{P}(E)$, we write $\mu \ll \nu$ if $\mu$ is absolutely continuous with respect to $\nu$, and denote the corresponding Radon-Nikodym derivative as $d\mu/d\nu$. The Kullback-Leibler (KL) divergence from $\nu \in \mathcal{P}(E)$ to $\mu \in \mathcal{P}(E)$ is defined as

$$\text{KL}(\mu|\nu) = \int_E \log \frac{d\mu}{d\nu}(x) \mu(dx)$$

if the integral is finite and $\mu \ll \nu$, and $\text{KL}(\mu|\nu) = \infty$ otherwise. The set of all real-valued, $\mathcal{E}$-measurable and bounded functions on $E$ is denoted by $\mathcal{B}(E)$. Given $\mu \in \mathcal{P}(E)$, $M \in \mathcal{M}(E)$ and $\phi \in \mathcal{B}(E)$, we define the integral $\mu(\phi) = \int_E \phi(x) \mu(dx)$ and the function $M(\phi)(\cdot) = \int_E \phi(y) M(\cdot, dy) \in \mathcal{B}(E)$. For ease of presentation, we will often assume that measures and transition kernels admit densities with respect to a $\sigma$-finite dominating measure $dx$, in which case we write the densities of $\mu \in \mathcal{P}(E)$ and $M \in \mathcal{M}(E)$ as $\mu(dx) = \mu(x)dx$ and $M(x, dy) = M(x, y)dy$, respectively.
5.1.2 Problem formulation and SMC samplers

In the rest of this chapter, we will restrict ourselves to $E := \mathbb{R}^d$, with $\mathcal{E}$ being the corresponding Borel $\sigma$-algebra. We are interested in sampling from a target distribution $\pi$ on $E$ which admits a density $\pi(x) = Z^{-1} \gamma(x) \in \mathcal{P}(E)$ with respect to the Lebesgue measure $dx$, assuming that we can evaluate $\gamma(x)$ pointwise. We are also interested in estimating its normalizing constant $Z = \gamma(E)$. A standard strategy is to introduce a collection of auxiliary probability measures $\{\pi_t\}_{t \in [0, T]} \subset \mathcal{P}(E)$ that interpolate between an easy-to-sample distribution $\pi_o \in \mathcal{P}(E)$ and the target distribution $\pi_T = \pi$, for some $T \in \mathbb{N}$. A typical example is the geometric interpolation where the auxiliary distributions admit densities of the following form:

$$
\gamma_t(x_t) := \pi_o(x_t)^{1-\lambda_t} \gamma(x_t)^{\lambda_t}, \quad \pi_t(x_t) := \gamma_t(x_t)/Z_t, \quad t \in [0 : T],
$$

where $Z_t = \gamma_t(E)$ and $\{\lambda_t\}_{t \in [0, T]} \subset [0, 1]$ is an increasing sequence satisfying $\lambda_0 = 0$ and $\lambda_T = 1$.

The rationale for introducing the sequence $\{\pi_t\}_{t \in [0, T]}$ is that if neighboring distributions $\pi_{t-1}$ and $\pi_t$ are not too different, it might be possible to construct forward Markov transition kernels $\{M_t\}_{t \in [0, T]} \subset \mathcal{M}(E)$ such that samples from $\pi_{t-1}$ are approximately distributed as $\pi_t$ when moved with $M_t$. This sampling strategy results in a non-homogeneous Markov chain with initial distribution $\pi_o$ and transition kernels $\{M_t\}_{t \in [0, T]}$, giving rise to the path measure

$$
\mathbb{Q}(dx_{o:T}) = \pi_o(dx_o) \prod_{t=1}^T M_t(x_{t-1}, dx_t).
$$

The distribution $\mathbb{Q}$ is such that the marginal distribution $q_T$ of $X_T$ is an approximation of the target $\pi_T$. This idea motivates using $q_T$ as a proposal distribution targeting $\pi_T$ in importance sampling, but the corresponding Radon-Nikodym derivative $d\pi_T/dq_T$ cannot be computed up to a normalizing constant, as $q_T$ is typically intractable.

Sequential Monte Carlo samplers (Del Moral et al., 2006) avoid this intractability by instead performing importance sampling on path space $(E^{T+1}, \mathcal{E}^{T+1})$ by defining the ex-
tended target distribution

\[ \mathbb{P}(dx_{0:T}) = \pi_T(dx_T) \prod_{t=1}^{T} L_{t-1}(x_t, dx_{t-1}), \quad (5.3) \]

where \( \{L_t\}_{t \in [0:T-1]} \) is a sequence of arbitrary backward Markov transition kernels, selected such that \( \mathbb{P} \ll \mathbb{Q} \) and \( \frac{d\mathbb{P}}{d\mathbb{Q}} \) can be evaluated pointwise up to a normalizing constant. As the distribution \( p_T \) of \( X_T \) under \( \mathbb{P} \) is such that \( p_T = \pi_T \), an importance sampling approximation of \( \mathbb{P} \) provides directly an approximation of \( \pi_T \) and an unbiased Monte Carlo estimate of \( Z_T/Z_o \) using samples from \( \mathbb{Q} \), thanks to the identity

\[ \frac{Z_T}{Z_o} = \frac{Z_T}{Z_o} \mathbb{E}_\mathbb{Q} \left[ \frac{d\mathbb{P}}{d\mathbb{Q}}(X_{0:T}) \right] = \mathbb{E}_\mathbb{Q} \left[ \prod_{t=1}^{T} w_t(x_{t-1}, x_t) \right], \quad (5.4) \]

where the incremental importance weights are given by

\[ w_t(x_{t-1}, x_t) := \frac{dL_{t-1} \otimes \gamma_t}{d\gamma_{t-1} \otimes M_t}(x_{t-1}, x_t), \quad (5.5) \]

and

\[ L_{t-1} \otimes \gamma_t(dx_{t-1}, dx_t) := \gamma_t(dx_t)L_{t-1}(x_t, dx_{t-1}), \]

\[ \gamma_{t-1} \otimes M_t(dx_{t-1}, dx_t) := \gamma_{t-1}(dx_{t-1})M_t(x_{t-1}, dx_t). \]

However, the performance of any importance sampling scheme depends on the discrepancy between the target and the proposal distributions, which can only increase when extending the domain of integration from \( E \) to \( E^{T+1} \). This can be seen from the decomposition

\[ \text{KL}(\mathbb{P}||\mathbb{Q}) = \text{KL}(\pi_T||q_T) + \int_E \text{KL}(\mathbb{P}(dx_{0:T-1}|x_T)||\mathbb{Q}(dx_{0:T-1}|x_T)) \pi_T(dx_T), \quad (5.6) \]

in which the first term captures the discrepancy between \( q_T \) to \( \pi_T \), and the second captures the additional discrepancy arising from the introduction of the backward kernels.
\( \{ L_t \}_{t \in [0; T-1]} \). In other words, making the importance weights tractable comes at a cost.

The formula in (5.6) also shows that KL(\( \mathbb{P} \| \mathbb{Q} \)) is minimized by choosing the backward kernels to make the equality

\[
\mathbb{P}(dx_{o:T}) = \frac{d\pi_T}{dq_T}(x_T)\mathbb{Q}(dx_{o:T})
\]  

(5.7)

hold, as this is equivalent to \( \mathbb{P}(dx_{o:T-1}|x_T) = \mathbb{Q}(dx_{o:T-1}|x_T) \) for \( \pi_T \)-almost every \( x_T \). The optimal backward kernels \( \{ L_i^{\text{opt}} \}_{i \in [0; T-1]} \) must therefore satisfy the forward-backward relation

\[
\pi_o(dx_o) \prod_{t=1}^{T} M_t(x_{t-1}, dx_t) = q_T(dx_T) \prod_{t=1}^{T} L_i^{\text{opt}}(x_t, dx_{t-1}),
\]  

(5.8)

where \( q_i(dx)L_i^{\text{opt}}(x, dx') = q_{i-1}(dx')M_i(x', dx) \), as noticed by Del Moral et al. (2006). They also showed that these kernels minimize the variance of the weights \( \prod_{t=1}^{T} w_t(x_{t-1}, x_t) \) under \( X_{0:T} \sim \mathbb{Q} \). Note that this choice remains intractable, and one must typically resort to approximations of the optimal backward kernels in practice.

In Crooks (1998) and Neal (2001), the authors restrict themselves to scenarios where \( M_t \) is selected to be \( \pi_t \)-invariant and \( L_{t-1} \) is the reversal of \( M_t^2 \). Under these conditions, (5.8) corresponds to a discrete-time version of the celebrated Jarzynski’s identity (Jarzynski, 1997). The generalized framework of Del Moral et al. (2006) presented here is key to the developments in this chapter, as we will use forward transition kernels that are not invariant with respect to the measures \( \pi_t \).

Suppose for a moment that we flip the roles of \( \pi_o \) and \( \pi_T \) and that we can sample from the path measure \( \mathbb{P}^{(i)} := \mathbb{P} \) defined in (5.7) to target \( \pi_o \). Following the reasoning above, but applied backwards in time, the corresponding optimal forward kernels would give rise to a path measure \( \mathbb{Q}^{(i)} \) such that \( \mathbb{Q}^{(i)}(dx_{o:T} | x_o) = \mathbb{P}^{(i)}(dx_{o:T} | x_o) \). In turn, the path measure \( \mathbb{Q}^{(i)} \) could hypothetically be used as a path space proposal targeting \( \pi_T \). Furthermore, this process could be iterated to construct forward and backward kernels \( \{ M_i^{(i)} \}_{i \in [1; T]} \) and \( \{ L_t^{(i)} \}_{t \in [0; T-1]} \) and corresponding path measures \( \mathbb{Q}^{(i)} \) and \( \mathbb{P}^{(i)} \). In what follows, we will describe this algorithm as the iterative proportional fitting procedure (IPFP). Under certain

\(^2\)The kernel \( L_{t-1} \) is the reversal of the \( \pi_t \)-invariant kernel \( M_t \) if \( \pi_t(dx)M_t(x, dx') = \pi_t(dx')L_{t-1}(x', dx) \). In particular, \( L_{t-1} = M_t \) if \( M_t \) is \( \pi_t \)-reversible.
regularity conditions, the measures $Q^{(i)}$ and $P^{(i)}$ both convergence to the solution of the so-called Schrödinger bridge problem. By numerically approximating the steps of the IPFP algorithm, we will leverage the reduction in discrepancy between $Q^{(i)}$ and $P^{(i)}$ to yield efficient SMC sampling from the target distribution $\pi_T$.

5.1.3 Outline of the proposed method

The formulas in equations (5.6) and (5.7) are at the core of our approach. In Section 5.2, we will describe (5.7) as providing the solution to the forward Schrödinger half-bridge problem

$$P^{(i)}(dx_{o:T}) = \arg\min_{H \in \mathcal{P}_t(\pi_t)} KL(H|Q^{(i-1)}), \quad (5.9)$$

where for any $t \in [0 : T]$, $\mathcal{P}_t(\pi_t)$ denotes the set of path measures on $(E^{T+1}, \mathcal{E}^{T+1})$ that have $\pi_t$ as their $t$-th marginal distribution. Approximating its solution allows us to implicitly estimate the optimal backward kernels for a given set of forward kernels. By alternating this step with solving the analogous backward Schrödinger half-bridge problem

$$Q^{(i)}(dx_{o:T}) = \arg\min_{H \in \mathcal{P}_o(\pi_o)} KL(H|P^{(i)}), \quad (5.10)$$

we will also make refinements on the forward kernels.

Iterating between the forward and backward half-bridge problems can be seen as an instance of the iterative proportional fitting procedure, introduced in various forms by Deming and Stephan (1940); Ireland and Kullback (1968); Kullback (1968). In discrete spaces, IPFP is also known as Sinkhorn’s algorithm (Sinkhorn and Knopp, 1967), and has recently gained attention as a computational tool in the field of optimal transport (Cuturi, 2013; Peyré and Cuturi, 2019). Under the weak regularity conditions detailed by Rüschendorf (1995), the iterates $P^{(i)}$ and $Q^{(i)}$, initialized at $Q^{(0)} = Q$, are known to converge as $i \to \infty$ to the solution of the Schrödinger bridge problem

$$S(dx_{o:T}) = \arg\min_{H \in \mathcal{P}_{o,T}(\pi_o, \pi_T)} KL(H|Q), \quad (5.11)$$
Figure 5.1.1: Illustration of the iterative proportional fitting procedure. The blue line represents $\mathcal{P}_o(\pi_o) \subset \mathcal{P}(E^{T+1})$, denoted $\mathcal{P}_o$ in the figure, while the red line represents $\mathcal{P}_T(\pi_T) \subset \mathcal{P}(E^{T+1})$, denoted $\mathcal{P}_T$. The black line illustrates that the KL projections $Q^{(i)} \in \mathcal{P}_o(\pi_o)$ and $F^{(i)} \in \mathcal{P}_T(\pi_T)$ converge towards the Schrödinger bridge $S$.

where $\mathcal{P}_{o,T}(\pi_{o,T}) = \mathcal{P}_o(\pi_o) \cap \mathcal{P}_T(\pi_T)$. The IPFP algorithm is illustrated in Figure 5.1.1.

The Schrödinger bridge $S$ is by definition the path measure with marginals that interpolate between $\pi_o$ and $\pi_T$ which is the closest to the reference dynamics $Q$ in terms of KL divergence. Once $S$ has been computed, one could generate samples from $\pi_o$ and propagate them through the forward Markov kernels induced by $S$ to obtain samples from $\pi_T$ at time $T$. Figure 5.1.2 illustrates a reference process and the associated Schrödinger bridge.

The Schrödinger bridge problem has a long history in physics, and was first studied by its namesake Schrödinger (1931, 1932) because of its connections to quantum mechanics. It was later rediscovered and posed in its modern formulation (5.11) in probability (e.g. Dawson and Gärtner, 1987; Föllmer, 1988) and control theory (e.g. Mikami, 1990; Dai Pra, 1991). We will elaborate on some of these connections throughout this chapter, and refer to Léonard (2014) for a more thorough overview.

Unlike in discrete spaces, where the iterations (5.9) and (5.10) can be implemented exactly, we need to approximate them in continuous spaces. In such scenarios it will be typically easier to approximate the Schrödinger bridge when $q_T$ is not too different from $\pi_T$ to start with. This motivates composing the solutions of a sequence of intermediate
(a) Marginal distributions $q_t$ of a reference process $Q(dx_{0:T})$.

(b) Marginal distributions $s_t$ of the Schrödinger bridge $S(dx_{0:T})$.

**Figure 5.1.2:** Illustration of the evolutions of the marginal distributions $q_t$ and $s_t$ of the reference process $Q(dx_{0:T})$ and the associated Schrödinger bridge $S(dx_{0:T})$, respectively. The colors transition from red to green to blue as $t$ increases from $0$ to $T = 40$. The Schrödinger bridge interpolates between the initial distribution $\pi_0$ and the target distribution $\pi_T = \pi$, here given in grey, while remaining close to the reference process.

Schrödinger bridge problems, for instance between adjacent distributions in the geometric interpolation (§1), which underpins the sequential Schrödinger bridge (SSB) sampler we propose in Section 5.3.

Of central importance in our method will be the class of path measures that arise as $\psi$-twisted versions of $Q$. That is, path measures of the form

$$Q^\psi(dx_{0:T}) = \pi_0^\psi(dx_0) \prod_{t=1}^{T} M_t^\psi(x_{t-1}, dx_t), \quad (5.12)$$

where $\psi = \{\psi_t\}_{t \in [0:T]}$ is a collection of strictly positive functions on $\mathcal{E}$ referred to as a
policy (we will sometimes also refer to the collection \([- \log \psi_t]_{t \in [0 : T]}\) as the policy), and
\[
\pi^*_t(dx) = \frac{\pi_o(dx_o)\psi_o(x_0)}{\pi_o(\psi_o)}, \quad M^*_t(x_{t-1}, dx_t) = \frac{M_t(x_{t-1}, dx_t)\psi_t(x_t)}{M_t(\psi_t)(x_{t-1})}, \quad t \in [1 : T].
\]

In particular, we show that \(S = Q^{\psi^*}\) for some policy \(\psi^*\) in Section 5.2. Moreover, formulating the half-bridge problems (5.9) and (5.10) as policy refinements, we can approximate their solutions with parametric policy classes and approximate dynamic programming, utilizing a method in the spirit of Heng et al. (2017).

In our implementation of the SSB sampler, the forward kernels of the initial Markov process \(Q\) will often be obtained by discretizing the continuous-time Langevin dynamics
\[
dX_t = \frac{1}{2} \nabla \log \pi_t(X_t)ds + dW_t, \quad s \in [0, \tau],
\]
where \(W_t\) denotes the standard Brownian motion, and \((\pi_t)_{t \in [0, \tau]}\) is a smooth curve such that \(\pi_\tau = \pi\), e.g. a continuous version of the geometric interpolation in (5.1). There are two main motivations for this choice. Firstly, if \(\tau\) is large enough, the distribution of \(X_t\) in (5.14) is close to \(\pi_t\) for any \(s\); see e.g. Chiang et al. (1987). Provided the discretization of \([0, \tau]\) is fine enough, we expect the marginal distribution of the discretized process to be close to \(\pi_t\), also. Such kernels are therefore likely to provide us with an adequate starting point. Secondly, this perspective will allow us to more readily estimate the optimal backward kernels, and to make use of flexible function estimation methods to approximate the policies.

### 5.2 Schrödinger Bridges

We review different formulations of the Schrödinger bridge problem introduced in (5.11) and the iterative proportional fitting procedure to find its solution. We then reformulate IPFP as an algorithm that acts in the space of policies. Parameterizing this space allows us to develop a numerical algorithm to approximate the Schrödinger bridge based on approximate dynamic programming. In practice we often work with kernels that arise from discretizing continuous-time processes, and thus also review some properties of the
continuous-time formulation of the Schrödinger bridge problem.

5.2.1 Dynamic and static formulations

Recall that for a reference path measure \( \mathcal{Q} \) and initial and terminal constraints \( \pi_o \) and \( \pi_T \) respectively, the dynamic Schrödinger bridge problem takes the form

\[
S(dx_{0:T}) = \arg \min_{\mathbb{H} \in \mathcal{P}(E^{T+1})} KL(\mathbb{H}|\mathcal{Q}). \tag{5.15}
\]

It can equivalently be expressed in its static form as follows: Notice that for any \( \mathbb{H} \in \mathcal{P}(E^{T+1}) \) we have the decomposition

\[
KL(\mathbb{H}|\mathcal{Q}) = KL(h_{0,T}|q_{0,T}) + \int_{E^t} KL(\mathbb{H}(dx_{t-1}|x_0,x_T)|\mathcal{Q}(dx_{t-1}|x_0,x_T)) h_{0,T}(dx_0,dx_T), \tag{5.16}
\]

where for any \( 0 \leq s < t \leq T \), \( h_{s,t} \) and \( q_{s,t} \) denote the two-time marginals of \( \mathbb{H} \) and \( \mathcal{Q} \). Since the constraint in (5.15) applies only to the first term in the sum in (5.16), the second term can be minimized by setting \( \mathbb{H}(dx_{t-1}|x_0,x_T) = \mathcal{Q}(dx_{t-1}|x_0,x_T) \) for \( h_{s,t} \)-almost every \( (x_0,x_T) \). Hence, (5.15) can be reduced to the static problem

\[
s_{0,T}(dx_0,dx_T) = \arg \min_{h_{0,T} \in \mathcal{C}(\pi_0,\pi_T)} KL(h_{0,T}|q_{0,T}), \tag{5.17}
\]

where \( \mathcal{C}(\pi_0,\pi_T) \subset \mathcal{P}(E^t) \) is the set of couplings on \( E^t \) with marginal distributions \( \pi_o \) and \( \pi_T = \pi \) respectively.

If \( q_{0,T} \ll q_o \otimes q_T \) and there exists \( h_{0,T} \in \mathcal{C}(\pi_0,\pi_T) \) such that \( KL(h_{0,T}|q_{0,T}) \leq \infty \), Rüschendorf and Thomsen (1993) showed that the minimum in (5.17) exists, is unique, and of the form

\[
s_{0,T}(dx_0,dx_T) = \phi^*(x_0)q_{0,T}(dx_0,dx_T)\phi^*(x_T), \tag{5.18}
\]

for two functions \( \phi^*, \phi^*: E \to \mathbb{R}_+ \). These functions are often called potentials, and are themselves unique up to a multiplicative constant. From (5.16) and (5.18), we obtain

\[
S(dx_{0:T}) = s_{0,T}(dx_0,dx_T)\mathcal{Q}(dx_{t-1}|x_0,x_T) = \phi^*(x_0)\mathcal{Q}(dx_{0:T})\phi^*(x_T). \tag{5.19}
\]
The marginal constraints on $\mathcal{S}(dx_{o:T})$ give rise to the so-called Schrödinger equations

$$
\pi_o(dx_o) = \phi^o(x_o)q_o(dx_o)\int_E \mathbb{Q}(dx_T|x_o)\phi^*(x_T),
$$

(5.20)

$$
\pi_T(dx_T) = \phi^*(x_T)q_T(dx_T)\int_E \mathbb{Q}(dx_o|x_T)\phi^o(x_o),
$$

(5.21)

where we recall that, from (5.2), $\pi_o = q_o$ in our context. Finding the solution to the Schrödinger bridge problem can therefore be reduced to finding the potentials $\phi^o$ and $\phi^*$ that solve the Schrödinger equations (5.20) and (5.21).

To succinctly express the transition probabilities and marginal distributions under $\mathcal{S}$, we define the harmonic functions

$$
\psi^*(x_T) = \phi^*(x_T),
$$

(5.22)

$$
\psi^0(x_t) = \int_E \psi^*(x_T)\mathbb{Q}(dx_T|x_t), \quad t \in [o : T - 1],
$$

(5.23)

and the co-harmonic functions

$$
\psi^o(x_o) = \phi^o(x_o),
$$

(5.24)

$$
\psi^0(x_t) = \int_E \psi^o(x_o)\mathbb{Q}(dx_o|x_t), \quad t \in [1 : T].
$$

(5.25)

Using this notation, the one- and two-time marginals of $\mathcal{S}$ can be expressed as

$$
s_t(dx_t) = \psi^0(x_t)\psi^*(x_t)q_t(dx_t) \quad \text{and} \quad s_{s,t}(dx_s, dx_t) = \psi^o(x_s)\psi^*(x_t)q_{s,t}(dx_s, dx_t)
$$

any $t \in [o : T]$ and any $o \leq s < t \leq T$, respectively. Thus, the transition probabilities under $\mathcal{S}$ take the form

$$
\mathcal{S}(dx_t|x_s) = \frac{\mathbb{Q}(dx_t|x_s)\psi^*(x_t)}{\psi^0(x_t)}, \quad o \leq s < t \leq T.
$$

(5.26)

From this representation we can see that if $\mathbb{Q}$ is Markovian, then so is $\mathcal{S}$. In this case, we also have that $\psi^*(x_s) = \int_E \psi^*(x_t)\mathbb{Q}(dx_t|x_s)$. In particular, if $\mathbb{Q}$ is of the form (5.2), then
\[ Q(dx_t|x_{t-1}) = M_t(x_{t-1}, dx_t) \text{ and } \psi^*_t(x_{t-1}) = M_t(\psi^*_t)(x_{t-1}). \] In other words, \( S \) is the \( \psi^* \)-twisted version of \( Q \), where \( \psi^* = \{ \psi^*_t \}_{t \in [0:T]} \) is the collection of harmonic functions.

### 5.2.2 Iterative Proportional Fitting

The iterative proportional fitting procedure (IPFP) was introduced by Deming and Stephan (1940) to estimate cell probabilities in contingency tables subject to certain marginal constraints, and later generalized to continuous state spaces by Ireland and Kullback (1968); Kullback (1968). Notably, it was also studied as a matrix scaling algorithm by Sinkhorn and Knopp (1967) and is often referred to as Sinkhorn’s algorithm in computer science and machine learning (Cuturi, 2013; Peyré and Cuturi, 2019).

In our notation, the steps of IPFP initialized at \( Q^{(0)} = Q \) can be expressed

\[ P^{(i)}(dx_{0:T}) = \arg\min_{H \in P_{\pi_T}} \text{KL}(H|Q^{(i-1)}), \tag{5.27} \]
\[ Q^{(i)}(dx_{0:T}) = \arg\min_{H \in P_{\pi_0}} \text{KL}(H|P^{(i)}), \tag{5.28} \]

for \( i \geq 1 \), and are sometimes called Schrödinger half-bridge problems; in each KL projection, only one of the two marginal constraints are enforced. Let \( S^{(i+1)} = P^{(i+1)} \) and \( S^{(i)} = Q^{(i)} \) for any \( i \geq 0 \). Then, if the minimizer in (5.17) exists, Rüschendorf (1995, Proposition 2.1) shows that the initial and terminal marginals of \( S^{(i)} \), denoted \( s_0^{(i)} \) and \( s_T^{(i)} \) respectively, converge to \( \pi_0 \) and \( \pi_T = \pi \) in both KL divergence and Total Variation (TV) as \( i \to \infty \). In the following proposition, which is inspired by Altschuler et al. (2017) and whose proof can be found in Appendix D, we provide further insight into the convergence properties of IPFP.

**Proposition 5.2.1.** For any \( \varepsilon > 0 \), the iterative proportional fitting procedure returns a distribution \( S^{(i)} \) satisfying

\[ \text{KL}(\pi_0|s_0^{(i)}) + \text{KL}(\pi_T|s_T^{(i)}) \leq \varepsilon \text{ in fewer than } \left\lceil \frac{\text{KL}(S|Q)}{\varepsilon} \right \rceil \text{ iterations.} \]

In particular, this result illustrates the benefit of starting with an initial distribution \( Q \) that is close to \( S \). Under the additional assumption that \( Q(dx_T|x_0) \geq c \pi_T(dx_T) \) for \( q_0 \)-almost every \( x_0 \) and some \( c > 0 \), the sequence \( S^{(i)} \) converges to \( S \) in both KL and TV as well (Rüschendorf, 1995, Theorem 3.5).
Unlike the problem in (5.15), the half-bridge problems have explicit solutions: Using
the KL decomposition in (5.6), we know that

$$\mathbb{P}^{(i)}(dx_{0:T}) = \frac{d\pi_T}{dq_{i-1}^T}(x_T)Q^{(i-1)}(dx_{0:T}). \quad (5.29)$$

By analogous reasoning, we also have that

$$\mathbb{Q}^{(i)}(dx_{0:T}) = \frac{d\pi_o}{dp_{i-1}^o}(x_o)\mathbb{P}^{(i)}(dx_{0:T}). \quad (5.30)$$

Next, we show that these iteration can be expressed as policy refinements. To set up an
inductive argument, assume that $\mathbb{Q}^{(i-1)} = \mathbb{Q}^{\psi^{(i-1)}}$ for some policy $\psi^{(i-1)} = \{\psi_t^{(i-1)}\}_{t \in [0:T]}$
and $i \geq 1$, i.e. $\mathbb{Q}^{(i-1)}$ is the $\psi^{(i-1)}$-twisted version of $\mathbb{Q}$ in (5.2). Note that this is trivially
true for $i = 1$, with $\psi_t^{(1)} = 1$ for all $t \in [0 : T]$. By the representation (5.29) and Heng et al.
(2017, Proposition 1), we have that $\mathbb{P}^{(i)} = \mathbb{Q}^{\psi^{(i-1)} \cdot \varphi^{(i)}}$, where we have used the notation
$\psi \cdot \varphi = \{\psi_t \cdot \varphi_t\}_{t \in [0:T]}$ and $\varphi^{(i)}$ satisfies the backward recursion (in $t$):

$$\varphi^{(i)}_T(x_T) = \frac{d\pi_T}{dq_{i-1}^T}(x_T), \quad (5.31)$$
$$\varphi^{(i)}_t(x_t) = M_{i+1}^{\psi^{(i-1)} \cdot \varphi^{(i)}_t}(x_t), \quad t \in [0 : T). \quad (5.32)$$

In other words, we have the representation

$$\mathbb{P}^{(i)}(dx_{0:T}) = \pi_o^{\psi^{(i-1)} \cdot \varphi^{(i)}}(dx_{0}) \prod_{t=1}^{T} M_t^{\psi^{(i-1)} \cdot \varphi^{(i)}}(x_{t-1}, dx_t). \quad (5.33)$$

With $\mathbb{P}^{(i)}$ expressed this way, the solution of the second half-bridge problem (5.30) can be
written

$$\mathbb{Q}^{(i)}(dx_{0:T}) = \pi_o(dx_o) \prod_{t=1}^{T} M_t^{\psi^{(i-1)} \cdot \varphi^{(i)}}(x_{t-1}, dx_t). \quad (5.34)$$

Hence, $\mathbb{Q}^{(i)} = \mathbb{Q}^{\psi^{(i)}}$, where $\psi_0^{(i)} = 1$ and $\psi_t^{(i)} = \psi_t^{(i-1)} \cdot \varphi_t^{(i)}$ for $t \in [1 : T]$.

Using the policy refinement perspective, the convergence of the IPFP algorithm can be
stated informally as follows. As \( i \to \infty \), the sequence of policies \( \psi^{(i)} \) converges to the policy \( \psi^* \) defined by the harmonic functions (5.22) and (5.23). Alternatively, the convergence can also be expressed in terms of the convergence of a fixed point method to solve (5.20) and (5.21) and obtain the corresponding Schrödinger potentials: For \( i \geq 1 \), we have

\[
a^{(i)}(x_o) = a^{(i-1)}(x_o) \frac{d\pi_o}{dp^{(i)}_o}(x_o) = a^{(i-1)}(x_o) \left[ \frac{\psi^{(i)}_o(x_o)}{\pi_o(\psi^{(i)}_o)} \right]^{-1}, \quad (5.35)
\]

\[
\beta^{(i)}(x_T) = \beta^{(i-1)}(x_T) \frac{d\pi_T}{dq^{(i)}_T}(x_T) = \beta^{(i-1)}(x_T) \phi^{(i)}_T(x_T), \quad (5.36)
\]

initialized at \( a^{(0)} = 1 \) and \( \beta^{(0)} = d\pi_T/dq_T \). The convergence of IPFP to the Schrödinger bridge is then equivalent to \( a^{(i)} \) and \( \beta^{(i)} \) converging to \( \phi^o \) and \( \phi^* \) as \( i \to \infty \), respectively.

In all but simple cases, such as the Gaussian setting treated in Section 5.2.6, the policies \( \psi^{(i)} \) are not available in closed form due to the intractability of the recursion (5.31). In what follows, we utilize the connection to optimal control that underlies the framework developed by Heng et al. (2017) to build numerical approximations of the policies.

5.2.3 Approximate iterative proportional fitting

Assuming we can evaluate the Radon-Nikodym derivative \( R^{\psi^{(i)}} := d\pi_T/dq^{\psi^{(i)}}_T \) pointwise, we can use approximate dynamic programming to approximate the recursion (5.31). However, \( q^{\psi^{(i)}}_T \) is typically intractable, making \( R^{\psi^{(i)}} \) intractable also. To circumvent this difficulty, we derive estimators of \( R^{\psi^{(i)}} \) to use in its place. A high-level view of our approximate IPFP is summarized in Algorithm 1.

Approximate dynamic programming

Our implementation of the approximate dynamic programming (ADP) algorithm estimates the intractable policy functions \( \phi^{(i)} = \{\phi^{(i)}_t\}_{t \in [0:T]} \) by least squares projections (in log-scale) onto a collection of function classes \( \{F_t\}_{t \in [0:T]} \). Specifically, for an estimate \( \psi^{(i-1)} \) of \( \psi^{(i-1)} \) and trajectories \( \{X^n\}_{n \in [1:N]} \) such that \( X^n \sim Q^{\psi^{(i-1)}} \) for \( n \in [1 : N] \), ADP
performs the backward recursion

\[ \hat{\psi}^{(i)}_T = \arg\min_{f \in F_T} \sum_{n=1}^{N} \left| \log f(X^n_T) - \log R^{\hat{\psi}^{(i-1)}_T}(X^n_T) \right|^2, \quad (5.37) \]

\[ \hat{\psi}^{(i)}_t = \arg\min_{f \in F_t} \sum_{n=1}^{N} \left| \log f(X^n_t) - \log M^{\hat{\psi}^{(i-1)}_{t+1}}_t(\hat{\psi}^{(i)}_t)(X^n_t) \right|^2, \quad t \in [0 : T - 1], \quad (5.38) \]

and returns the output \( \hat{\psi}^{(i-1)} \cdot \hat{\psi}^{(i)} \) and the approximation \( Q^{\hat{\psi}^{(i-1)}_T \cdot \hat{\psi}^{(i)}_T} \) of \( P(i) \). To form \( \hat{\psi}^{(i)} \)
and the corresponding approximation \( Q^{\hat{\psi}^{(i)}_T} \) of \( Q(i) \), we set \( \hat{\psi}^{(0)}_o = 1 \) and \( \hat{\psi}^{(i)}_t = \hat{\psi}^{(i-1)}_t \cdot \hat{\psi}^{(i)}_t \)
for \( t \in [1 : T] \). We refer the reader to Heng et al. (2017, Section 4) for analysis of the error
and large \( N \) behavior of this scheme.

The requirement that we can sample trajectories \( X^n \sim Q^{\hat{\psi}^{(i-1)}_T} \) for any \( i \geq 1 \), i.e. that we

can sample \( X^n_0 \sim \pi_o(dx_o) \) and \( X^n_t \sim M^{\hat{\psi}^{(i-1)}_T}_t(X^n_{t-1}, dx_t) \) for \( t \in [1 : T] \), places restrictions on

the initial kernels \( \{M_t\}_{t \in [1 : T]} \) and the function classes \( \{F_t\}_{t \in [0 : T]} \).

So does the requirement

that we are able to evaluate \( \log M^{\hat{\psi}^{(i-1)}_T}_t(\hat{\psi}^{(i)}_t)(X^n_t) \). In many cases, these conditions can only

be shown to hold when the class \( F_t \) is conjugate with the kernel \( M_t \) for each \( t \). An important

special case where these conditions are satisfied is when the kernels are Gaussian and the

function classes contain only quadratic forms, so that the twisted kernels remain Gaussian.

We elaborate on this point in Section 5.2.4.

**Estimating the Radon-Nikodym derivative**

The ADP implementation in the previous section requires the evaluation of \( R^{\hat{\psi}^{(i-1)}_T} \) =

\( d\pi_T/dQ^{\psi}_T \) at the locations \( \{X^n_T\}_{n \in [1 : N]} \), where \( X^n \sim Q^{\psi}_n \) for \( n \in [1 : N] \). Here,

we discuss here how to obtain unbiased and asymptotically consistent estimates of these

Radon-Nikodym derivatives. For simplicity, we denote by \( X_{o:T} \) a generic trajectory from \( Q^{\psi} \)

and suppress the notational dependence on \( n \) and \( i \). Similarly, we express \( R^{\psi}_T \) in terms of

the normalized measure \( \pi_T \) for simplicity, but in practice only need access to \( \gamma_T \).

Our approach makes use of the identity

\[ \frac{d\pi_T}{dQ_T}(x_T) = \int_{E_T} \frac{dQ_T}{dQ^{\psi}_T}(x_{o:T})Q^{\psi}_T(dx_{o:T-1}|x_T), \quad (5.39) \]
which holds for any policy $\psi$ and any $\mathbb{H} \in \mathcal{P}_T(\pi_T)$ such that $\mathbb{H} \ll \mathbb{Q}^\psi$. As in (5.3), we consider path measures $\mathbb{H}$ defined in terms of backward Markov transition kernels:

$$
\mathbb{H}(dx_{0:T}) = \pi_T(dx_T) \prod_{t=1}^{T} L_{t-1}(x_t, dx_{t-1}). 
$$

(5.40)

If we can generate $M + 1$ samples $\{X^{(m)}_{o:T-1}\}_{m=0}^{M}$ from the distribution $\mathbb{Q}^\psi(dx_{o:T-1}|x_T)$, an unbiased estimator of $R^\psi(x_T)$ is given by

$$
\hat{R}_M^\psi(x_T) = \frac{1}{M + 1} \sum_{m=0}^{M} \frac{d\mathbb{H}}{d\mathbb{Q}^\psi}(X^{(m)}_{o:T-1}, x_T). 
$$

(5.41)

Assuming the realization $X_T = x_T$, then one such sample is already available, since in this case $X^{(0)}_{o:T-1} \sim \mathbb{Q}^\psi(dx_{o:T-1}|x_T)$. In the setting where $\{L_t\}_{t \in [0:T-1]}$ corresponds to the optimal choice of backward kernels, i.e. such that

$$
\prod_{t=1}^{T} L_{t-1}(x_t, dx_{t-1}) = \mathbb{Q}^\psi(dx_{o:T-1}|x_T), 
$$

(5.42)

then the single sample $X^{(0)}_{o:T-1}$ would be sufficient, as in this case $\hat{R}_M^\psi(x_T)$ is a zero-variance unbiased estimator of $R^\psi(x_T)$ for any $M \geq 0$. This choice is intractable, and we discuss methods for constructing backward kernels that are useful in practice in Section 5.2.4.

In cases where $M = 0$ is not sufficient, we use conditional SMC to produce more samples (Andrieu et al., 2010). Specifically, we derive a Markov kernel $K_{x_T} \in \mathcal{M}(E^T)$ whose invariant distribution is $\mathbb{Q}^\psi(dx_{o:T-1}|x_T)$. By setting $X^{(0)}_{o:T-1} = X^{(0)}_{o:T-1}$, and iterating

$$
X^{(m)}_{o:T-1} \sim K_{x_T}(X^{(m-1)}_{o:T-1}, dx_{o:T-1}), \quad m \in [1 : M], 
$$

(5.43)

we construct samples that are correlated, but have exactly $\mathbb{Q}^\psi(dx_{o:T-1}|x_T)$ as their marginal distribution. The resulting estimator $\hat{R}_M^\psi(x_T)$ therefore stays unbiased, and converges to $R^\psi(x_T)$ at the rate of $\sqrt{M}$.

Next, we explain how the kernel $K_{x_T}$ is built. For simplicity we present the construc-
tion without resampling, but the details of its resampling counterpart are discussed by
Andrieu et al. (2010). Given the current state of the Markov chain \( X^{(m-1)}_{\omega : T-1} \), we first sam-
pie \( P - 1 \in \mathbb{N} \) independent trajectories \( \{Y^p_{\omega : T-1}\}_{p \in [1 : P-1]} \) from the conditional distribution \( \mathbb{H}(dx_{\omega : T-1}|x_T) \). It follows from (5.40) that this can be done by sampling \( Y^p_{T-1} \sim L_{T-1}(x_T, dx_{T-1}) \) and \( Y^p_t \sim L_t(Y^p_{t+1}, dx_t) \) for \( t \in [0 : T-2] \), and setting \( Y^p_{\omega : T-1} = (Y^p_0, \ldots, Y^p_{T-1}) \) for \( p \in [1 : P-1] \). In addition, we set \( Y^p_{\omega : T-1} = X^{(m-1)}_{\omega : T-1} \) as a reference trajectory. Now, sample an index \( A \) from the categorical distribution on \( [1 : P] \) with associated probabilities

\[
\frac{d \mathbb{Q}^\psi}{d \mathbb{H}}(Y^A_{\omega : T-1}, x_T) \quad \sum_{q=1}^{P} \frac{d \mathbb{Q}^\psi}{d \mathbb{H}}(Y^q_{\omega : T-1}, x_T), \quad p \in [1 : P], \tag{5.44}
\]

and set \( X^{(m)}_{\omega : T-1} = Y^A_{\omega : T-1} \).

The efficiency of this approach and the required number of samples \( M+1 \) depend heavily
on the design of the backward kernels, which remains a challenging task. In fact, approxi-
mating the optimal backward kernels for \( \mathbb{Q}^\psi(\cdot) \) may be harder than for the initial measure \( \mathbb{Q} \), as \( M^\psi(\cdot) \) no longer necessarily targets \( \pi \). In the next section, we illustrate this point for specific choices of initial kernels, derived from discretizing continuous-time processes.

5.2.4 Continuous-time Schrödinger bridges

In applications, we often choose the initial kernels \( \{M_t\}_{t \in [0 : T]} \) to be the Euler-Maruyama
discretization of some continuous-time dynamics, e.g. the Langevin dynamics in (5.14).
By approximating the solution of the corresponding Schrödinger bridge problem, we pro-
pose a version of the IPFP algorithm that allows flexible function classes to be used. First,
we review some of the continuous-time problem’s features, restricting ourselves to refer-
ence processes of the kind

\[
dX_s = b_s(X_s)ds + dW_s, \quad s \in [0, \tau], \tag{5.45}
\]
Algorithm 1 Approximate IPFP

Input: Initial kernels \( \{M_t\}_{t \in [0:T]} \), function classes \( \{F_t\}_{t \in [0:T]} \), number of particles \( N \in \mathbb{N} \), number of IPFP iterations \( I \in \mathbb{N} \).

1. Initialize: Set \( \hat{\psi}_t^{(o)} = 1 \) for \( t \in \{0 : T\} \).

2. For \( 1 \leq i \leq I \):

   (a) Sample trajectories \( \{X^n_t\}_{n \in [0:N]} \) from \( Q^{(i-1)} \): For each \( n \in [1 : N] \), sample \( X^n_0 \sim \pi_0(dx_0) \) and \( X^n_t \sim M^n_t(X^n_{t-1}, dx_t) \) for \( t \in [1 : T] \).

   (b) For each \( n \in [1 : N] \), compute an estimator \( \hat{R}^{(i-1)}(X^n_T) \) of \( R^{(i-1)}(X^n_T) = d\pi_T/dq_T^{(i-1)}(X^n_T) \) using the strategies discussed in Section 5.2.3.

   (c) Approximate dynamic programming: perform the recursion

   \[
   \hat{\phi}_T^{(i)} = \arg\min_{f \in F_T} \sum_{n=1}^{N} \left| \log f(X^n_T) - \log \hat{R}^{(i-1)}(X^n_T) \right|^2,
   \]

   \[
   \hat{\phi}_t^{(i)} = \arg\min_{f \in F_t} \sum_{n=1}^{N} \left| \log f(X^n_t) - \log M^n_{t+1}(\hat{\phi}^{(i)}_{t+1}(X^n_{t+1})) \right|^2, \quad t \in [1 : T - 1].
   \]

   (d) Set \( \hat{\psi}_0^{(i)} = 1 \) and \( \hat{\psi}_t^{(i)} = \hat{\psi}_{t-1}^{(i-1)} \cdot \hat{\phi}_t^{(i)} \) for \( t \in [1 : T] \).

Output: Policy \( \hat{\psi}^{(i)} \).
where $W_t$ denotes the standard Brownian motion, the initial value $X_0$ is distributed according to some measure, and $\tau > 0$. Let $Q$ denote the law of this process on the space of continuous $\mathbb{E}$-valued paths over $[0, \tau]$. Important special cases include the reversible Brownian motion, for which $b_s = 0$ for all $s \in [0, \tau]$ and $X_0$ is distributed according to the Lebesgue measure, as well as the Langevin dynamics with $X_0 \sim \pi_0$ which we consider in the next section.

As in the discrete-time setting, the continuous-time Schrödinger bridge problem with respect to $Q$ can be expressed as the minimization of KL divergence over processes $H$ that satisfy the marginal constraints $H_0 = \pi_0$ and $H_\tau = \pi$ (see e.g. Léonard, 2014). Following Dai Pra (1991), it can equivalently be described as a stochastic control problem: Find the change in drift $s \mapsto u_s$ such that the process

$$dY_s = \{b_s(Y_s) + u_s(Y_s)\} \, ds + dW_s, \quad s \in [0, \tau], \quad (5.46)$$

satisfies $Y_0 \sim \pi_0$ and $Y_\tau \sim \pi$, and $s \mapsto u_s$ minimizes the cost function $\mathbb{E} \int_0^\tau \|u_s(Y_s)\|^2 \, ds$. Using this perspective, Dai Pra (1991) shows, under some assumptions, that the optimal controller $s \mapsto u^*_s$ is of the form $u^*_s = \nabla \log \psi^*_s$ and gives an explicit form of the smooth map $s \mapsto \psi^*_s$ in terms of Doob’s $h$-transform. This map is the continuous-time analog of the policy defined by (5.22) and (5.23).

**Schrödinger bridges for discretized Langevin dynamics**

Consider the case where $b_s = \frac{4}{3} \nabla \log \pi_s$ for some smooth curve of distributions $(\pi_s)_{s \in [0, \tau]}$ such that $\lim_{s \to 0} = \pi_0$ and $\lim_{s \to \tau} \pi_s = \pi$. Examples of such curves include continuous versions of the geometric interpolation in (5.1). Provided the drift function $s \mapsto \nabla \log \pi_s$ changes adiabatically, or infinitely slowly in time, the distribution of the particle $X_t$ is equal to $\pi$, for all $s \in [0, \tau]$ (Chiang et al., 1987; Patra and Jarzynski, 2017). In all but trivial cases, this effectively requires that $\tau \to \infty$. However, for $\tau$ large enough, one expects the distribution of $X_t$ to be close to $\pi$.

From the sampling perspective, this property motivates using the kernels $\{M_t\}_{t \in [0, \tau]}$ de-
defined by the Euler-Maruyama discretization

\[ M_t(x_{t-1}, x_t) = \mathcal{N}\left(x_t; x_{t-1} + \frac{h}{2} \nabla \log \pi_t(x_{t-1}), h\mathcal{I}_d\right), \quad t \in [1 : T]. \tag{5.47} \]

to initialize the path measure \( \mathbb{Q}_s \), where \( \{\pi_t\}_{t \in [0, T]} \) is a discretization of the curve \( (\pi_s)_{s \in [0, T]} \).

If the step-size \( h = \tau / T \) is small and \( \tau \) is large, we expect \( q_T \) to be close to \( \pi_T = \pi \), providing us a good starting point. This can also be seen by noting that if one takes \( L_{t-1}(x, x') = M_t(x, x') \), the resulting SMC sampler would approximate AIS, as \( \pi_{t-1} \approx \pi_t \) in the small \( h \) setting.

Consider also the Euler-Maruyama discretization of (5.46) in the case where \( b_s = \frac{\partial}{\partial x} \log \pi_s \) and \( u_s = \nabla \log \psi_s \) for some smooth curve of functions \( (\psi_s)_{s \in [0, T]} \), and define

\[ M_t^\psi(x_{t-1}, x_t) = \mathcal{N}\left(x_t; x_{t-1} + \frac{h}{2} \nabla \log \pi_t(x_{t-1}) + h \nabla \log \psi_t(x_{t-1}), h\mathcal{I}_d\right), \tag{5.48} \]

for \( t \in [1 : T] \). These kernels can also be expressed

\[ M_t^\psi(x_{t-1}, x_t) = \frac{M_t(x_{t-1}, x_t)\psi_t(x_t)}{M_t(\psi_t(x_{t-1}))}, \tag{5.49} \]

where \( \log \psi_t(x_t) = \log \psi_t(x_{t-1}) + \nabla \log \psi_t(x_{t-1})^T(x_t - x_{t-1}) \) and \( \log M_t(\psi_t(x_{t-1})) = \log \psi_t(x_{t-1}) + \frac{h}{2} \nabla \log \psi_t(x_{t-1})^T \left[ \nabla \log \pi_t(x_{t-1}) + \nabla \log \psi_t(x_{t-1}) \right] \). Provided the functions \( \{\psi_t\}_{t \in [0, T]} \) are appropriately smooth, the Taylor expansion of \( \log \psi_t \) around \( x_{t-1} \), can be written

\[ \log \psi_t(x_t) = \log \psi_t(x_{t-1}) + \mathcal{O}(\|x_t - x_{t-1}\|^2), \tag{5.50} \]

as \( \|x_t - x_{t-1}\| \to 0 \). In the small \( h \) regime, we expect \( \|x_t - x_{t-1}\| \to 0 \) to be of size \( \mathcal{O}(h) \) under \( M_t(x_{t-1}, dx_t) \). Hence, \( \{M_t^\psi\}_{t \in [0, T]} \) are likely to provide good approximations of the kernels \( \{M_t\}_{t \in [0, T]} \) for sufficiently small step sizes.

The benefit of using the policy \( \tilde{\psi} \) in place of \( \psi \) is that sampling from \( M_t^\psi(x_{t-1}, x_t) \) and evaluating the integrals \( \log M_t(\tilde{\psi}_t(x_{t-1})) \) only require access to the pointwise evaluation of the functions \( \log \psi_t \) and \( \nabla \log \psi_t \). Thus, using these approximations we could feasibly run the ADP algorithm without requiring the function classes and kernels to be conjugate. This
would allow us to make use of flexible function estimation methods such as deep neural networks within the approximate IPFP algorithm. A version of approximate IPFP using Euler-Maruyama kernels is given in Algorithm 2.

The continuous-time perspective suggests a natural choice of backward kernels. For diffusion processes of the form (5.46), Haussmann and Pardoux (1986) show that the time reversed process satisfies

\[
d\tilde{Y}_s = \left\{ \nabla \log \rho_s(\tilde{Y}_s) - b_s(\tilde{Y}_s) - u_s(\tilde{Y}_s) \right\} ds + d\tilde{W}_s, \quad s \in [0, \tau], \tag{5.51}
\]

where \( \rho_s \) is the marginal density of the process (5.46) at time \( s \), and \( \tilde{W}_s \) is another Brownian motion. The backward kernels could therefore be sensibly chosen to be the Euler-Maruyama discretization of (5.51), but the term \( \nabla \log \rho_s \) is typically intractable. If \( b_s = \frac{1}{2} \nabla \log \pi_s \) and \( u_s = \nabla \log \psi_s \) is an approximation of the optimal forward drift \( \nabla \log \psi_s^* \), a generic choice could be to replace \( \rho_s \) by \( \pi_s \). In this case, the backward kernels would amount to

\[
L^\psi_{t-1}(x_t, x_{t-1}) = \mathcal{N} \left( x_{t-1}; x_t + \frac{h}{2} \nabla \log \pi_{t-1}(x_t) - h \nabla \log \psi_t(x_t), hI_d \right), \tag{5.52}
\]

for \( t \in [1 : T] \). However, the usefulness of this choice relies on the assumption that the marginal of \( Q \) at time \( s \) is close to \( \pi_s \), which might not be the case. In Section 5.3.2, we circumvent the intractability of \( \rho_s \) by designing a process in which the marginal distributions can be consistently approximated by \( \pi_s \), making the kernel in (5.52) more suitable.

5.2.5 Schrödinger bridge sampling

Given the policy \( \hat{\psi}^{(I)} \) after \( I \geq 1 \) iterations of the approximate IPFP algorithm, one can proceed to use the proposal \( Q^{\hat{\psi}^{(I)}} \) within importance sampling or SMC on path space. The terminal marginal \( q_T^{\hat{\psi}^{(I)}} \) is likely to be closer to the target distribution \( \pi_T = \pi \) than the initial \( q_T \), but this condition is not in itself enough to guarantee efficient sampling. In particular, as mentioned in the previous sections, constructing backward kernels \( \{L_t\}_{t \in [0: T-1]} \) to define the target distribution on path space remains challenging. Moreover, the incre-
Algorithm 2 Approximate IPFP for discretized Langevin dynamics

**Input:** Number of time-steps $T \in \mathbb{N}$, terminal time $\tau > 0$, function classes $\{F_t\}_{t \in [0,T]}$, number of particles $N \in \mathbb{N}$, number of iterations $I \in \mathbb{N}$.

1. Initialize: Set $h = \tau/T$, $\hat{\psi}_t^{(0)} = 1$ for $t \in [0 : T]$, and

   $$ M_t(x_{t-1}, x_t) = \mathcal{N}\left(x_t; x_{t-1} + \frac{h}{2} \nabla \log \pi_t(x_{t-1}), hI_d\right), \quad t \in [1 : T]. $$

2. For $1 \leq i \leq I$:
   (a) For each $t \in [0 \ldots T]$, define
   $$ \log \tilde{\psi}_t^{(i-1)} = \log \hat{\psi}_t^{(i-1)}(x_{t-1}) + \nabla \log \hat{\psi}_t^{(i-1)}(x_{t-1})^T(x_t - x_{t-1}). $$
   (b) Sample trajectories $\{X^n_t\}_{n \in [1 : N]}$ from $Q^{\tilde{\psi}^{(i-1)}}$: For each $n \in [1 : N]$, sample $X^n_0 \sim \pi_o(dx_o)$ and
   $$ X^n_t \sim \mathcal{N}\left(x_t; x_{t-1} + \frac{h}{2} \nabla \log \pi_t(x_{t-1}) + h \nabla \log \psi_t(x_{t-1}), hI_d\right), \quad t \in [1 : T]. $$
   (c) For each $n \in [1 : N]$, compute an estimator $\hat{R}^{\tilde{\psi}^{(i-1)}}(X^n_T) = (d\pi_T/d\hat{q}^{(i-1)})(X^n_T)$ using the strategies discussed in Section 2.3, e.g. using the backward kernels
   $$ L_t^{\tilde{\psi}^{(i-1)}}(x_{t-1}, x_t) = \mathcal{N}\left(x_t; x_{t-1} + \frac{h}{2} \nabla \log \pi_{t-1}(x_t) - h \nabla \log \psi_t(x_t), hI_d\right). $$
   (d) Approximate dynamic programming: perform the recursion
   $$ \bar{\phi}_T^{(i)} = \text{argmin}_{f \in F_T} \sum_{n=1}^N \left| \log f(X^n_T) - \log \hat{R}^{\tilde{\psi}^{(i-1)}}(X^n_T) \right|^2, $$
   $$ \bar{\phi}_t^{(i)} = \text{argmin}_{f \in F_t} \sum_{n=1}^N \left| \log f(X^n_t) - \log M_t(x_t; \bar{\phi}_t^{(i-1)}(X^n_t)) \right|^2, $$
   for $t \in [1 : T - 1]$, where $\bar{\phi}^{(i)}$ is defined analogously to $\bar{\psi}^{(i-1)}$.

   (e) Set $\bar{\psi}^{(i)}_0 = 1$ and $\bar{\psi}^{(i)}_t = \bar{\psi}^{(i-1)} \cdot \bar{\phi}^{(i)}$ for $t \in [1 : T]$. 

**Output:** Policy $\bar{\psi}^{(i)}$. 

118
mental weights within an importance sampling or SMC scheme would more naturally be based on the sequence of Schrödinger bridge marginals \( \{ s_i \}_{i \in [0:T]} \) rather than \( \{ \pi_i \}_{i \in [0:T]} \), but these distributions are intractable (even up to normalizing constants). In Section 5.3, we propose a scheme based on the multi-marginal Schrödinger bridge problem which helps mitigate these issues.

5.2.6 Example: Linear quadratic Gaussian

We illustrate the computation of Schrödinger bridges on a simple example in which the initial, target, and transitions are all Gaussian: Let \( \pi_o (dx_o) = N(x_o; \mu_o, \Sigma_o) dx_o \) and \( \pi_T (dx_T) = N(x_T; \mu_T, \Sigma_T) dx_T \) for some \( \mu_o, \mu_T \in \mathbb{R}^d \) and \( \Sigma_o, \Sigma_T \in \mathbb{R}^{d \times d} \), and for each \( t \in [1 : T] \), let \( M_t(x_{t-1}, dx_t) = N(x_t; K_t x_{t-1} + r_t, H_t) dx_t \) for some \( r_t \in \mathbb{R}^d \) and \( K_t, H_t \in \mathbb{R}^{d \times d} \). By conjugacy, it can be shown that the exact IPFP iterations can be written in terms of policies of the form \( -\log \psi_t^{(i)} = x_t^T A_t^{(i)} x_t + x_t^T b_t^{(i)} + c_t^{(i)} \) for every \( i \geq 1 \) and \( t \in [0 : T] \). We derive expressions for the coefficients \( A_t^{(i)}, b_t^{(i)}, c_t^{(i)} \) and the resulting twisted Markov kernels in Appendix D. We compare the exact potentials \( \psi^{(i)} \) and the corresponding path measure \( \mathbb{Q}^{\psi^{(i)}} \) with their approximations \( \tilde{\psi}^{(i)} \) and \( \mathbb{Q}^{\tilde{\psi}^{(i)}} \) constructed with ADP. We choose the function classes \( F = \{ F_t \}_{t \in [0:T]} \) to contain only quadratic forms. This choice is well-specified in the sense that \( \psi^{(i)} \in F \) for every \( i \geq 1 \).

In particular, we consider a prior distribution \( \pi_o \) with \( \mu_o = 0 \) and \( \Sigma_o = I_d \) and a log-likelihood defined by \( \ell(x) = - (y - x)^T R^{-1} (y - x) / 2 \) for some \( y \in \mathbb{R}^d \) and symmetric positive definite \( R \in \mathbb{R}^{d \times d} \), giving rise to a posterior distribution \( \pi = \pi_T \) with \( \Sigma_T = (\Sigma_o^{-1} + R^{-1})^{-1} \) and \( \mu_T = \Sigma_T (\Sigma_o^{-1} \mu_o + R^{-1} y) \). By conjugacy, the distributions defining the geometric path (5.1) are Gaussian: \( \pi_t (dx_t) = N(x_t; \mu_t, \Sigma_t) dx_t \) with \( \Sigma_t = (\Sigma_o^{-1} + \lambda_t R^{-1})^{-1} \) and \( \mu_t = \Sigma_t (\Sigma_o^{-1} \mu_o + \lambda_t R^{-1} y) \). Let \( y = (\xi, \ldots, \xi)^T \) and \( R \) have 1’s on the diagonal and \( \rho \)’s on the off-diagonal. Here, we take \( d = 2, \xi = 8 \) and \( \rho = 0.8 \).

We consider two different settings of the reference process parameters \( \{ K_t \}_{t \in [1:T]} \), \( \{ r_t \}_{t \in [1:T]} \) and \( \{ H_t \}_{t \in [1:T]} \). In the first, they are constructed from discretizing Brownian motion over the time-interval \( [0, \tau] \) with \( \tau = 2 \) and step size \( h = 1/20 \). That is, we take \( K_t = I_d, r_t = 0 \) and \( H_t = h I_d \) for every \( t \in [1 : T] \), with \( T = 4.0 \). In Section 5.4, we discuss the connection between Schrödinger bridges with Brownian motion reference process and the optimal

119
transport problem with quadratic cost. In the second setting, we consider the discretization of the Langevin dynamics discussed in Section 5.2.4, i.e. \( \lambda_i = h \tau, K_i = I_d - h \Sigma_i^{-1} / 2, r_i = h \Sigma_i^{-1} \mu_i / 2 \) and \( H_i = h I_d \) for \( t \in [1 : T] \), with \( T \) and \( h \) as in the Brownian setting. In both cases, we take the backward kernels to be (5.52).

To measure the discrepancy between the marginals of the Schrödinger bridge \( s_t \) and the marginals obtained with \( i \) iterations of IPFP \( q_i^{(i)} \), we compute the corresponding 2-Wasserstein distance. Since the distributions involved are Gaussian, the 2-Wasserstein distance between them is known in closed form (see e.g. Peyré and Cuturi, 2019, Chapter 2.6): For Gaussian distributions \( \nu_i \) with means \( \gamma_i \) and covariances \( \Gamma_i \) for \( i = 1, 2 \), we have

\[
\mathcal{W}_2^2(\nu_1, \nu_2) = \| \gamma_1 - \gamma_2 \|^2 + B(\Gamma_1, \Gamma_2)^{\frac{1}{2}},
\]

where \( B \) is the Bures metric between positive definite matrices (Bures, 1969), defined

\[
B(\Gamma_1, \Gamma_2)^{\frac{1}{2}} = \text{Tr} \left( \Gamma_1 + \Gamma_2 - 2 \left( \Gamma_1^{\frac{1}{2}} \Gamma_2 \Gamma_1^{\frac{1}{2}} \right)^{\frac{1}{2}} \right).
\]

In Figures 5.2.1 and 5.2.2 we plot \( \log \mathcal{W}_2(\nu_1, q_i^{(i)}) \) for \( t \in [1 : T] \), \( i \in [0 : \tilde{s}] \), and exact and approximate IPFP for different number of conditional SMC iterations \( \tilde{M} \) (using \( P = 128 \) cSMC particles). In both Figures 5.2.1a and 5.2.2a, we observe that the marginals given by the exact IPFP iterations appear to converge exponentially fast. For the approximate schemes, we observe that performing only one or two iteration of IPFP leads to large reductions in the distance to the Schrödinger bridge, but that the benefit of performing further IPFP iterations depends on the variance of the Radon-Nikodym estimates via the value of \( \tilde{M} \). For less variable estimates, the benefit of further IPFP iterations is larger. This also illustrates that the choice of backward kernels (5.52) is suboptimal in these settings, which motivates the method proposed in Section 5.3.2. Increasing the number of particles \( N \) did not lead to qualitatively different behavior across the IPFP iterations, but reduced the width of the confidence bands (not illustrated).
(a) Exact IPFP.  (b) Approx. IPFP, $M = 0$.  (c) Approx. IPFP, $M = 5$.  
(d) Approx. IPFP, $M = 10$.  (e) Approx. IPFP, $M = 50$.  (f) Approx. IPFP, $M = 100$.  

**Figure 5.2.1:** Distances between marginals of the Schrödinger bridge $s_t$ and the marginals of the IPFP iterates $q^{(i)}_t$, measured as $\log W_2(s_t, q^{(i)}_t)$, for the LQG setting of Section 5.2.6 with discretized Brownian diffusion reference dynamics. Figure 5.2.1a corresponds to the exact computation of the IPFP iterates for $i \in [0 : 5]$. The remaining plots correspond to the proposed particle-based approximation of the IPFP iterations using $N = 1,000$ particles and different values of conditional SMC iterations $M$ (using $P = 128$ CSMC particles). The solid lines correspond to the median value of the log-distance calculated over 100 independent simulations, and corresponding confidence bands represent the 5% and 95% quantiles. Note that the vertical axis of Figure 5.2.1a is on a different scale than those of the other figures, as the exact IPFP iterations yield smaller distances in general.
(a) Exact IPFP.  
(b) Approx. IPFP, \( M = 0 \).  
(c) Approx. IPFP, \( M = 5 \).

(d) Approx. IPFP, \( M = 10 \).  
(e) Approx. IPFP, \( M = 50 \).  
(f) Approx. IPFP, \( M = 100 \).

**Figure 5.2.2:** Distances between marginals of the Schrödinger bridge \( s_t \) and the marginals of the IPFP iterates \( q^{(i)}_t \), measured as \( \log \mathcal{W}_2(s_t, q^{(i)}_t) \), for the LQG setting of Section 5.2.6 with discretized Langevin diffusion reference dynamics. Figure 5.2.2a corresponds to the exact computation of the IPFP iterates for \( i \in [0 : 5] \). The remaining plots correspond to the proposed particle-based approximation of the IPFP iterations using \( N = 1,000 \) particles and different values of conditional SMC iterations \( M \) (using \( P = 128 \) cSMC particles). The solid lines correspond to the median value of the log-distance calculated over 100 independent simulations, and corresponding confidence bands represent the 5\% and 95\% quantiles. Note that the vertical axis of Figure 5.2.2a is on a different scale than those of the other figures, as the exact IPFP iterations yield smaller distances in general.
5.3 Sequential Schrödinger Bridge Samplers

The Monte Carlo approximation of the Schrödinger bridge proposed in Section 5.2 has some limitations. In particular, choosing a useful set of backward kernels is difficult, in part because the optimal choice depends on the intractable marginal distributions of the forward process. In this section, we construct forward kernels such that the corresponding marginal distributions are approximately equal to the sequence $\{ \pi_t \}_{t \in [0:T]}$, and therefore circumvent this issue. The scheme we introduce is based on estimating and composing a sequence of intermediate Schrödinger bridges, which will be seen to approximate the solution of the multi-marginal Schrödinger bridge problem.

5.3.1 Multi-marginal Schrödinger bridges

The multi-marginal Schrödinger bridge problem is defined by

$$S_K(dx_{o:T}) = \arg\min_{H \in \mathcal{P}_K(\pi_K)} \text{KL}(H \| Q),$$

where $\mathcal{K} = \{ t_k \}_{k \in [i:K]} \subset [o : T]$ with $t_i = o$, $t_K = T$ and $t_k < t_{k+1}$ for each $k$, and $\mathcal{P}_K(\pi_K) = \bigcap_{k \in [i:K]} \mathcal{P}_{t_k}(\pi_{t_k})$. In other words, the set of admissible path measures have $\pi_{t_k}$ as their $t_k$-marginal for every $k \in [i : K]$. An important special case is where $\mathcal{K} = [o : T]$. Similar to the two-marginal Schrödinger bridge considered earlier (in which $\mathcal{K} = \{ o, T \}$), the multi-marginal Schrödinger bridge can be written

$$S_K(dx_{o:T}) = Q(dx_{o:T}) \prod_{k=1}^K \phi_{t_k}(x_{t_k}),$$

where the potentials $\{ \phi_{t_k} \}_{k \in [i:K]}$ are unique up to a multiplicative constant and solve the $K$ Schrödinger equations: For each $k \in [1 : K]$,

$$\pi_{t_k}(x_{t_k}) = \int_{\mathcal{E}} \left[ \prod_{\ell=1}^K \phi_{t_{\ell}}(x_{t_{\ell}}) Q(x_{o:T}) \right] dx_{-t_k},$$
where we have used the notation $x_{-t} = (x_0, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T)$. Furthermore, the solution of the multi-marginal problem can be approximated using a variation of the IPFP algorithm. In particular, Kullback (1968) introduced a scheme which systematically cycles through KL projections onto $\mathcal{P}_{h_k}(\pi_k)$ for each $k \in [1 : K]$. For $\mathcal{K} = \{0, T\}$, this scheme reduces to the iterations in (5.27).

However, this version of IPFP is challenging to implement for several reasons. For instance, we only assume being able to initialize samples from $\pi_o$, which means that before each projection onto $\mathcal{P}_{h_k}(\pi_k)$ for $k \in [2 : K]$, we would need to first project onto $\mathcal{P}_o(\pi_0)$. Even so, the scheme would suffer from the same difficulties as discussed for the two-marginal problem, as we would still be required to construct a full set of $T$ backward kernels for each iteration of IPFP.

### 5.3.2 Sequential Schrödinger bridges

Instead of solving the multi-marginal Schrödinger bridge problem using the cyclic IPFP scheme considered by Kullback (1968), we instead introduce a sequential approach. In particular, we sequentially solve the intermediate two-marginal Schrödinger bridge problems

$$S_k(\text{d}x_{k,i:k+1}) = \arg\min_{H \in \mathcal{P}_{h_k}(\pi_{k,i:k+1})} \text{KL}(H \| Q_{k,i:k+1}), \quad k \in [1 : K]. \quad (5.57)$$

where for each $0 \leq s < u \leq T$, we define $Q_{s,u} = \pi_s(\text{d}x_s) \prod_{t=s+1}^u M_t(x_{t-1}, \text{d}x_t)$. As shown in Section 5.2, we know that for each $k \in [1 : K]$ we can write

$$S_k(\text{d}x_{k,i:k+1}) = \pi_k(\text{d}x_k) \prod_{t=t_{k+1}}^{t_{k+1}} M_t^{\psi_k}(x_{t-1}, \text{d}x_t) \quad (5.58)$$

where $\psi_k = \{\psi_k^{\star}\}_{t \in [k:k+1]}$ denotes the corresponding harmonic functions.

By the Markov property of the initial path measure $Q_o$, solving the set of two-marginal Schrödinger bridge problems is equivalent to solving the multi-marginal problem. Their solutions can be related explicitly in the following way. Let $\phi_{k,i}(x_k) = \left[\psi_{k,i}^{\star}(x_k)\right]^{-1}$ and

124
\( \phi_{t_k}(x_{t_k}) = \psi_{K-t_k}^{*}(x_{t_k}), \) and for any \( k \in [2 : K - 1] \) let

\[
\phi_{t_k}(x_{t_k}) = \frac{\psi_{K-t_{k+1}}^{*}(x_{t_k})}{\psi_{K-t_k}^{*}(x_{t_k})}.
\] (5.59)

For \( \{\phi_{t_k}\}_{k \in [1 : K]} \) defined this way, we have that

\[
Q(dx_{o:T}) \prod_{k=1}^{K} \phi_{t_k}(x_{t_k}) = \pi_o(dx_o) \prod_{k=1}^{K} \prod_{i=t_k+1}^{t_k} M_i^{\psi_k^*}(x_{t_{i-1}}, x_t) \, dx_i.
\] (5.60)

By construction of the policies \( \{\psi_k^*\}_{k \in [1 : K]} \) via the two-marginal problems, the \( t_k \)-marginal of this path measure is equal to \( \pi_{t_k} \). Hence, the potentials \( \{\phi_{t_k}\}_{k \in [1 : K]} \) solve the Schrödinger equations (5.56), and we can write the solution to the multi-marginal problem as

\[
S_K(dx_{o:T}) = \pi_o(dx_o) \prod_{k=1}^{K} \prod_{i=t_k+1}^{t_k} M_i^{\psi_k^*}(x_{t_{i-1}}, x_t) \, dx_i.
\] (5.61)

The reason for introducing the intermediate problems is that approximating the Schrödinger bridge between nearby distributions \( \pi_{t_k} \) and \( \pi_{t_{k+1}} \) with reference process \( Q_{t_k : t_{k+1}} \) is typically easier than estimating the full Schrödinger bridge between \( \pi_o \) and \( \pi_T \) with reference process \( Q \). However, applying approximate IPFP to find \( S_k \) requires being able to initialize particles from \( \pi_{t_k} \). This can be done approximately by initializing particles from \( \pi_o \) and propagating them through the approximations of the bridges \( S_1, \ldots, S_{k-1} \), suggesting a sequential approach to estimating the multi-marginal bridge \( S_K \). A high-level summary of this scheme is given in Algorithm 3.

In Section 5.3.3, we approximate the multi-marginal Schrödinger bridge in the case where the reference process \( Q \) is the Euler-Maruyama discretization of the Langevin dynamics introduced in Section 5.2.4. In Section 5.3.4, we discuss how to use the sequential estimation approach for sampling, and in Section 5.3.5 develop an adaptive construction of the set \( \mathcal{K} \).
Algorithm 3 Sequential Schrödinger bridge approximation

Input: Initial kernels \( \{M_t\}_{t \in [0, T]} \), number of particles \( N \in \mathbb{N} \), set of indices \( \{0, T\} \subset \mathcal{K} \subset [0 : T] \) with \( \mathcal{K} = |\mathcal{K}| \), number of iterations \( \mathcal{I}_k \in \mathbb{N} \) for each \( k \in [1 : K] \).

1. Initialize: For each \( n \in [1 : N] \), sample \( X^n_0 \sim \pi_0 = \pi_k \).

2. For \( k \in [1 : K] \):
   
   (a) Perform \( \mathcal{I}_k \) iterations of approximate IPFP to estimate \( \hat{S}_{k,n} \), initialized with \( \hat{\psi}_{k,t}^{(0)} = 1 \) for \( t \in [t_k : t_{k+1}] \), using the samples \( \{X^n_t\}_{n \in [1 : N]} \) as approximate draws from \( \pi_k \). Output the policy \( \hat{\psi}_{k,t}^{(k)} \).
   
   (b) Propagate the samples \( \{X^n_{t_k}\}_{n \in [1 : N]} \) using \( \{M_t^{\hat{\psi}_{k,t}^{(k)}}\}_{t \in [t_{k+1} : t_{k+1}]} \) to produce the approximate draws \( \{X^n_{t_{k+1}}\}_{n \in [1 : N]} \) from \( \pi_{t_{k+1}} \).

Output: Policies \( \{\hat{\psi}_{k,t}^{(k)}\}_{k \in [1 : K]} \).

5.3.3 Sequential Schrödinger bridges for discretized Langevin dynamics

We illustrate the SSB approach in the case where the initial forward kernels correspond to the Euler-Maruyama discretization of the continuous-time Langevin dynamics in (5.14). The main contrast with the corresponding two-marginal problem discussed in Section 5.2.4 is that the backward kernels introduced in (5.52) are likely to be more efficient within the SSB methodology. To illustrate this point, we consider the setting where \( \mathcal{K} = [0 : T] \).

The resulting multi-marginal problem can be seen as a discretization of the following control problem: Find the control \( s \mapsto u_s = \nabla \log \psi_s \) such that the process defined in (5.46) with \( b_s = \frac{1}{2} \nabla \log \pi_s \) satisfies \( Y_s \sim \pi_s \) for every \( s \in [0, \tau] \), and \( s \mapsto \nabla \log \psi_s \) minimizes the cost function \( \mathbb{E} \int_0^\tau \| \nabla \log \psi_s (Y_s) \|^2 ds \). This problem arises in different literatures, e.g. in physics, where any feasible potential is said to yield a shortcut to adiabaticity (Patra and Jarzynski, 2017). We elaborate on these connections in Section 5.4.

As noted in Section 5.2.4, the time-reversed version of the controlled continuous-time process (5.46) can be expressed as (5.51). However, in the setting we consider here, the marginal distribution \( \rho_{s} \) of the forward process is no longer intractable, as it is forced to equal \( \pi_s \). Hence, the backward kernels (5.52) that arise as the Euler-Maruyama discretiza-
tion of the time reversed process are likely to become increasingly efficient as our policy approximation improves, provided the discretization of $[0, \tau]$ is fine enough.

5.3.4 Sequential Schrödinger bridge sampling

Given the policy $\hat{\psi} = \{\hat{\psi}^{(k)}_k\}_{k \in [i:K]}$ produced by the sequential application of approximate IPFP, one can proceed to use the proposal $Q^{\hat{\psi}}$ within importance sampling or SMC on path space. We note that the calculation of the incremental importance weights (and, if applicable, the corresponding resampling step) can be integrated into the forward IPFP sweep. Similar to other SMC algorithms, one can also incorporate rejuvenation steps into the SSB algorithm. In particular, for each iteration of IPFP targeting $\mathcal{F}_k$ one can move the particles $\{X^n_k\}_{n \in [i:N]}$ approximating $\pi_k$ using Markov kernel that is invariant to $\pi_k$. This serves two purposes: First, these moves can improve the particle approximation of the $\pi_k$. Second, “refreshing” the particles can prevent overfitting the estimated policies to the fixed set $\{X^n_k\}_{n \in [i:N]}$.

In Algorithm 4, we present a version of the SSB sampler for the important special case where $\mathcal{K} = [0 : T]$. For ease of presentation, we use $M = o$ in the estimation of the Radon-Nikodym derivatives (see Section 5.2.3), the same number of IPFP iterations $I$ to estimate each intermediate bridge, no resampling, and no rejuvenation steps. Other sets $\mathcal{K}$ and varying numbers of IPFP iterations will be considered in Section 5.3.5, and versions using $M > o$ and resampling are useful in practice. The potential benefit of refreshing the particles is illustrated in the high-dimensional Gaussian setting of Section 5.5.1.

Each step of Algorithm 4 has a computational complexity of $O(IN)$. In Section 5.3.5, we discuss methods for warm starting the IPFP algorithm and adapting the number iterations to the difficulty of the policy estimation problems, to further reduce the computational cost. Since the SSB sampler is fundamentally an SMC algorithm, many of the theoretical results that have been derived for such methods are applicable (see e.g. Del Moral, 2013). More specifically, the SSB sampler is an instance of an adaptive SMC sampler, some of the properties of which were discussed by Beskos et al. (2016).
Algorithm 4 Sequential Schrödinger bridge sampler with index set $\mathcal{K} = [0 : T], M = 0$
cSMC iterations, no resampling, and no rejuvenation

**Input:** Initial kernels $\{M_t\}_{t \in [0 : T]}$, function classes $\{F_t\}_{t \in [0 : T]}$, number of particles $N \in \mathbb{N}$,
number of IPFP iterations $I \in \mathbb{N}$.

1. Initialize: Sample $X^n_0 \sim \pi_0(dx_0)$ for $n \in [1 : N]$.
2. For $t \in [1 : T]$,
   (a) Set $\hat{\psi}^{(o)}_t = 1$.
   (b) For $i \in [1 : I]$,
      i. Sample $Y^n_t \sim M^{\hat{\phi}^{(i-1)}_t}_t(X^n_{t-1}, dx_t)$ for $n \in [1 : N]$.
      ii. Construct a backward kernel $L^{\hat{\phi}^{(i-1)}_t}_{t-1}$, and for each $n \in [1 : N]$ compute
          \[
          \hat{R}^{\hat{\phi}^{(i-1)}_t}_t(Y^n_t) = \frac{\gamma_t(Y^n_t)L^{\hat{\phi}^{(i-1)}_t}_{t-1}(Y^n_t, X^n_{t-1})}{\gamma_{t-1}(X^n_{t-1})M^{\hat{\phi}^{(i-1)}_t}_{t-1}(X^n_{t-1}, Y^n_t)}.
          \]
      iii. Fit $\hat{\psi}^{(i)}_t = \arg\min_{\psi_t \in F_t} \sum_{n=1}^N \log f(Y^n_t) - \log \hat{R}^{\hat{\phi}^{(i-1)}_t}_t(Y^n_t) \|^{1/2}$, and set
           \[
           \Psi_t^{(i)} = \Psi_t^{(i-1)} \cdot \hat{\psi}^{(i)}_t.
           \]
   (c) Sample $X^n_t \sim M^{\psi^{(i)}_t}_t(X^n_{t-1}, dx_t)$ for $n \in [1 : N]$, and compute
       \[
       w^n_t = \frac{\gamma_t(X^n_t)L^{\psi^{(i)}_t}_{t-1}(X^n_t, X^n_{t-1})}{\gamma_{t-1}(X^n_{t-1})M^{\psi^{(i)}_t}_{t-1}(X^n_{t-1}, X^n_t)}.
       \]
3. Compute $w^n_{0:T} = \prod_{t=1}^T w^n_t$.

**Output:** Trajectories $\{X^n_{0:T}\}_{n \in [1 : N]}$ and importance weights $\{w^n_{0:T}\}_{n \in [1 : N]}$. 
5.3.5 Sequential Schrödinger bridge sampling with adaptive IPFP

In this section, we discuss different methods to reduce the computational cost of the SSB sampler. The first approach aims to reduce the number of IPFP iterations by using warm starts and stopping IPFP when the policy approximations appear to have converged. The second approach adaptively constructs the set $\mathcal{K}$ by triggering IPFP only when the effective sample size (ESS) of the particle system falls below a given threshold.

IPFP with early stopping and warm starts

Over the course of the IPFP iterations, we can monitor changes in the policy approximations and stop when they appear to have converged. Since the policies are estimated using a finite number of random particles, the measure by which we evaluate convergence needs to be able to account for the resulting noise in the approximations. Hence, we perform hypothesis tests to check whether the estimated policies have reached their stationary distribution under the approximate IPFP scheme.

In practice, we use parametric function classes for policy approximations, which means that monitoring the convergence of the policies can be reduced to tracking the evolution of the corresponding parameters. In the numerical experiments of Sections 5.3.6 and 5.5, we consider, at IPFP iteration $i$, a window of the differences between the parameters at iterations $j$ and $j - 1$ for the last $J$ iterations of IPFP (i.e. $j = i - J, \ldots, j = i$). For each parameter, we perform a $t$-test of whether the mean of these differences are equal to zero. The IPFP iterations are stopped when none of the tests are significant, controlling for false discoveries using e.g. the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), or when a prescribed number of iterations is reached. If the IPFP iterations are stopped early, we can reduce the variance in the policy approximations by averaging each parameter over the window for which the test was performed.

In the important special case where $\mathcal{K} = [0 : T]$, in which we approximate the solution of the two-marginal Schrödinger bridge between $\pi_{t-1}$ and $\pi_t$ for each $t \in [1 : T]$, we can also warm start the IPFP iterations. Since in the analogous continuous-time problem we expect the curve of policies $(\psi_s^*)_{s \in [0,T]}$ to be smooth, we can hope that an extrapolation of the approximations of $\{\psi_r^*\}_{r \in [1:T]}$ creates a good starting point for the approximation of
\( \psi_{t \pm t'} \) provided the discretization of \([0, \tau]\) is fine enough. In practice, we often use a finite difference of the form \( \hat{\psi}_t + (\psi_t - \hat{\psi}_{t-1}) \), or simply \( \hat{\psi}_t \) itself. In Section 5.3.6, we illustrate the benefits of warm starts, and how the combination of warm starts and early stopping can yield large reductions in computational cost without increasing the approximation error.

**Adaptive construction of \( \mathcal{K} \)**

Instead of a priori defining the set \( \mathcal{K} \), we can sequentially add indices to it by triggering IPFP steps only when the effective sample size (ESS) of the particle system falls below a given threshold \( 0 \leq \varepsilon \leq N \). That is, starting from \( t_k \), propagate the particles using the untwisted Markov kernels and define \( t_{k+1} \) to be the first \( t > t_k \) for which \((\sum_{n=1}^{N} (W^n_{t_{k+1}})^2)^{-1} < \varepsilon\), where

\[
W^n_{t_{k+1}} = \frac{w^n_{t_{k+1}}}{\sum_{i=1}^{N} w^n_{t_{k+1}}} , \quad w^n_{t_{k+1}} = \prod_{s=t_k}^{t} w^n_s .
\]

**5.3.6 Example: Linear Quadratic Gaussian (continued)**

Recall the linear quadratic Gaussian setting described in Section 5.2.6. Here, we estimate the solution to the corresponding multi-marginal Schrödinger bridge problem with discretized Langevin dynamics reference process and \( \mathcal{K} = [0 : T] \). In Figures 5.3.1a and 5.3.1b we plot \( \log \mathcal{W}_2(\pi_t, q_t^{(l)}) \) for \( t \in [1 : T] \) and various values of \( l \) between 0 and 100, calculated using exact and approximate IPFP with \( M = 0 \) iterations of conditional SMC. As for the two-marginal Schrödinger bridge problem, the marginals constructed in the exact IPFP iterations appear to converge exponentially fast, albeit at a slower rate. On the other hand, the approximate scheme appears to behave more similarly to the exact algorithm than in the two-marginal problem. This is likely because our construction of the backward kernels are better adapted to the multi-marginal problem.

Figure 5.3.1c plots \( \log \mathcal{W}_2(\pi_t, q_t^{(l)}) \) for marginals \( q_t^{(l)} \) obtained with and without the warm starts and early stopping schemes discussed in Section 5.3.5, with the non-adaptive algorithm using \( I = 100 \) IPFP iterations, and the adaptive one performing at least 3 iterations for each value of \( t \in [1 : T] \). To monitor the convergence of the policies, we use a window of size \( J = \min\{15, i\} \) at iteration \( i \geq 3 \). The plot illustrates the benefit of warm
starts, and that the combination of warm starts and early stopping appears to yield at least as good approximations as the non-adaptive scheme. The reduction in computational cost is dramatic, and the number of IPFP iterations performed at each \( t \in [1 : T] \) using the early stopping criterion is illustrated in Figure 5.3.1d. In terms of wall-clock time, the non-adaptive algorithm took on average 8.68s, whereas the adaptive one took 0.906s on an Intel Core i5 (2.5 GHz).

To assess the SSB algorithm from the sampling perspective, we compute the estimates \( \{ \hat{Z}_t \}_{t \in [1 : T]} \) of the normalizing constants \( \{ Z_t \}_{t \in [1 : T]} \) corresponding to \( \{ \pi_t \}_{t \in [0 : T]} \). In Figure 5.3.2, we compare an SMC sampler with discretized Langevin dynamics Markov kernels with the SSB sampler using warm starts and early stopping. In Figure 5.3.2b, we plot the root mean squared error (RMSE) of \( \log \hat{Z}_n \) for the two different methods. At time \( t = T \), the RMSE of the estimator obtained with standard SMC was 86 times higher than the estimator obtained with the SSB sampler, but took only 0.122s to compute on average. In other words, the RMSE of \( \log \hat{Z}_T \) was reduced by a factor of 86 at the cost of 7.4 times greater running time. In comparison, increasing the number of particles (and hence running time) of the standard SMC algorithm by a factor of 7.4 is only expected to reduce the RMSE by a factor of \( \sqrt{7.4} \).

5.4 Connections with other problems

In Sections 5.2 and 5.3.2, we have discussed the two- and multi-marginal Schrödinger bridge problems in their minimum KL and stochastic control formulations. Here, we elaborate on a few other perspectives that arise in different literatures. In particular, we review the two-marginal Schrödinger bridge problem as a regularization of an optimal transport problem, and discuss how our method might be used to approximate the \( 2 \)-Wasserstein distance between \( \pi_o \) and \( \pi_T \). Similarly, by analogy with the continuous-time formulation, the multi-marginal Schrödinger bridge can be viewed as an approximation of a solution of the flow transport problem. In the physics literature, both the two- and multi-marginal problems with Langevin diffusion reference dynamics can be said to yield shortcuts to adiabaticity. Recently, Chen et al. (2019) also develop connections between the multi-marginal prob-
Figure 5.3.1: Distances between $\pi_t$ and marginals of the IPFP-based approximations $q_t^{(i)}$ of the multi-marginal Schrödinger bridge, measured as $\log J_{\mathcal{V}}(\pi_t, q_t^{(i)})$, for the LQG setting of Section 5.3.6 with discretized Langevin diffusion reference dynamics. Figure 5.3.1a correspond to the exact computation of the IPFP iterates for different values of $i$ between 0 and 100. Figure 5.3.1b corresponds to the proposed particle-based approximation using $N = 1,000$ particles and $M = 0$ iterations of cSMC. Figure 5.3.1c illustrates the differences between the approximations obtained with and without the warm start and early stopping schemes discussed in Section 5.3.5. Figure 5.3.1d shows the number of IPFP iterations performed in the early stopping scheme at each time $t \in [1 : T]$, with a lower bound set to 3 iterations. In all the plots, solid lines correspond to median values calculated over 100 independent simulations, and the corresponding confidence bands represent the 5% and 95% quantiles.
Figure 5.3.2: On the left, we plot the logarithm of the estimates \( \{ \hat{Z}_t \}_{t \in [0, T]} \) of the normalizing constants \( \{ Z_t \}_{t \in [0, T]} \) corresponding to \( \{ \pi_t \}_{t \in [0, T]} \), based on an SMC sampler using Markov kernels derived from discretizing Langevin diffusion (in red) and the SSB sampler using the same discretized Langevin Markov kernels as the reference process (in blue). The solid lines correspond to median values calculated over 100 independent simulations using \( N = 1,000 \) particles and \( M = 0 \) iterations of cSMC, and the corresponding confidence bands represent the 5% and 95% quantiles. Note that the true log normalizing constants (in black) are obscured by the blue line, and that the blue confidence bands are too narrow to be visible. To further illustrate the behavior of the different estimators, we compute the root mean squared error of \( \log \hat{Z}_t \), the log of which is shown on the right.
lem and measure-valued splines. We also briefly discuss Schrödinger bridge-based particle filtering and some difficulties associated with applying the methodology in this setting.

### 5.4.1 Optimal transport

The first formal connections between the Schrödinger bridge problem and optimal transport were developed by Mikami (2004), who considered the control problem (5.46) for reference dynamics of the form $dX_s = \sigma dW_s$ on $[0, \tau]$. If $\mathcal{S}^e_{o,r}$ denotes the joint distribution of the optimally controlled process at times $s = o$ and $s = \tau$, he showed that as $\sigma \to o$, $\mathcal{S}^e_{o,r}$ converges to a solution of the quadratic Monge-Kantorovich optimal transport problem:

$$
\mathcal{W}^2_2(\pi_o, \pi_T) = \min_{h_{o,T} \in C(x_o, \pi_T)} \int_{E \times E} \|x_o - x_T\|^2 h_{o,T}(dx_o, dx_T),
$$

where the notation $\mathcal{W}^2_2(\pi_o, \pi_T)$ reflects that this defines the 2-Wasserstein distance. Additionally, when the reference path measure $Q^r$ is the $\sigma$-scaled reversible Brownian motion, he showed that $\sigma^2 \text{KL}(\mathcal{S}_o^e | Q_o^r) \to \mathcal{W}^2_2(\pi_o, \pi_T)$. Recently, Léonard (2012) has extended these results to more general Markov reference processes.

A useful perspective for making these connections is a formulation of the Schrödinger bridge problem developed by Léonard (2014), derived by considering the Fokker-Planck equation associated with (5.46); see also Chen et al. (2017). In particular, the problem studied by Mikami (2004) can be expressed

$$
\min_{\rho, \psi} \int_0^\tau \int_E \| \nabla \log \psi_s(x) \|^2 \rho_s(dx) ds,
$$

$$
\frac{d\rho_s}{ds} = -\text{div}(\rho_s \nabla \log \psi_s) + \frac{\sigma^2}{2} \Delta \rho_s,
$$

$$
\rho_o(dx) = \pi_o(dx), \quad \rho_\tau(dx) = \pi_T(dx).
$$

When $\sigma = o$, in which (5.64) becomes the continuity equation (Ambrosio et al., 2005, p.169), the above problem reduces to the Benamou-Brenier fluid mechanics formulation of the quadratic optimal transport problem (Benamou and Brenier, 2000).

Recently, these connections have been utilized to create fast approximate solvers of optimal transport problems with general cost functions. In particular, we have seen that the
original Schrödinger bridge problem (5.15) can be reduced to the static problem (5.17), given by
\[ s_{o,T}(dx_o, dx_T) = \arg \min_{h_{o,T} \in C(\pi_o, \pi_T)} \text{KL}(h_{o,T} | q_{o,T}). \] (5.66)
If we can write \( q_{o,T}^* (dx_o, dx_T) = Z_{\sigma}^{-1} \exp \left\{ -c(x_0, x_T)/\sigma^2 \right\} \lambda(dx_o, dx_T) \) for some cost function \( c : E \times E \to \mathbb{R}_+ \) and a \( \sigma \)-finite measure \( \lambda \) on \( E \times E \), then\(^3\)
\[ \text{KL}(h_{o,T} | q_{o,T}^*) = \text{KL}(h_{o,T} | \lambda) + \frac{1}{\sigma^2} \int_{E \times E} c(x_o, x_T) h_{o,T}(dx_o, dx_T) + \log Z_{\sigma}. \] (5.67)

The corresponding Schrödinger bridge can then be written
\[ s_{o,T}^* (dx_o, dx_T) = \arg \min_{h_{o,T} \in C(\pi_o, \pi_T)} \int_{E \times E} c(x_o, x_T) h_{o,T}(dx_o, dx_T) + \sigma^2 \text{KL}(h_{o,T} | \lambda). \] (5.68)
If \( \lambda \) is equal to the Lebesgue measure and \( c(x, y) = \|x - y\|^2 \), the above setting reduces to the one considered by Mikami (2004). When \( \lambda \) is taken to be \( \pi_o \otimes \pi_T \), the minimization in (5.68) is often called the entropically regularized optimal transport problem, and its objective function evaluated at the minimizer \( s_{o,T}^* \), called the Sinkhorn divergence (Cuturi, 2013; Peyré and Cuturi, 2019).

The output of Algorithm 1 applied to the Schrödinger bridge problem with discretized Brownian reference dynamics can be used to approximate the quadratic optimal transport cost in several ways. The simplest is the estimator \( N^{-1} \sum_{n=1}^N \| \hat{X}_0^n - \hat{X}_T^n \|^2 \), where \( \{ \hat{X}_n \}_{n=1}^N \) are the particle trajectories obtained in the last IPFP iteration. Since the obtained coupling will in general be sub-optimal for the transport problem, this estimator provides upper bound of \( \mathcal{W}_2^2(\pi_o, \pi_T) \) for any non-zero \( \sigma \) (up to noise and approximation errors). For the 100 independent simulations in performed in the LQG example of Section 5.2.6 with \( \sigma = 1 \) and \( M = 10 \), the average estimated value of \( \mathcal{W}_2(\pi_o, \pi_T) \) was 4.27 with a standard deviation of 0.033, whereas the exact distance is equal to 4.09. The discrepancy stems from the large value of \( \sigma \).

Alternatively, one can use the approximated Schrödinger potentials together with the

\(^3\)See e.g. Léonard (2014) for a discussion of the definition of the KL divergence with respect to an unbounded measure.
identity \(KL(s_o,T|q_{o,T}) = \int_{E \times E} \{\log \phi^o(x_o) + \log \phi^*(x_T)\} s_{o,T}(dx_o, dx_T)\) to construct a particle-based estimate. Thirdly, by analogy with the continuous-time problem, one can approximate (5.63) using the estimated policies and the associated particle trajectories.

5.4.2 Flow transport

Recall the control problem stated in Section 5.3.3, in which want to find \(s \mapsto u^*_s = \nabla \log \psi^*_s\) such that the process defined in (5.46) with \(b_t = \frac{1}{\tau} \nabla \log \pi_t\) satisfies \(Y_t \sim \pi_t\) for every \(s \in [0, \tau]\), and \(s \mapsto \nabla \log \psi^*_s\) minimizes the cost function \(E \int_0^\tau \|\nabla \log \psi^*_s(Y_t)\|^2 ds\). Re-expressing this problem in the language of (5.63)-(5.65), we can write

\[
\begin{align*}
\min \psi & \int_0^\tau \int_E \|\nabla \log \psi^*_s(x)\|^2 \pi_s(dx) ds, \quad (5.69) \\
\frac{d\pi_s}{ds} & = -\text{div} \left( \pi_s \left\{ \nabla \log \psi^*_s + \frac{1}{2} \nabla \log \pi_s \right\} \right) + \frac{1}{2} \Delta \pi_s. \quad (5.70)
\end{align*}
\]

Note, however, that the Fokker-Planck equation (5.70) reduces to the continuity equation:

\[
\frac{d\pi_s}{ds} = -\text{div} \left( \pi_s \nabla \log \psi^*_s \right). \quad (5.71)
\]

For any \(s \mapsto \nabla \log \psi^*_s\) that solves (5.71), we have that if \(x'_s = \nabla \log \psi^*_s(x_s)\) subject to \(x_o \sim \pi_o\), then \(x_s \sim \pi_t\) for any \(s \in [0, \tau]\). Finding such a policy is often called the flow transport problem. Simultaneously solving (5.69) yields the minimum kinetic energy solution among all solutions of the flow transport problem, and has been considered by e.g. Reich (2011, 2012).

In the linear quadratic Gaussian case, the minimal kinetic energy solution of the associated continuous time flow transport problem is known exactly (Bergemann and Reich, 2012). The optimal policy is given by

\[
s \mapsto \nabla \log \psi^*_s(x_s) = -\frac{1}{2 \tau} \Sigma_s \Sigma^{-1}(x_s + \mu_s - 2y),
\]

where \(\Sigma_s\) and \(\mu_s\) are defined as in the discrete setting of Section 5.2.6, but with \(\lambda_s = s/\tau\). As illustrated in Section 5.3.6, the sequential Schrödinger bridge algorithm yields policies that in turn give define marginal distributions \(q_t^{(1)}\) that are very close to \(\pi_t\). Here, we can also
numerically approximate the optimal cost and compare it to the cost estimates produced by the SSB sampling algorithm.

Using the same discretization of \([0, \tau]\) as in Section 5.3.6, we numerically solve \(\chi' = \nabla \log \psi^*_s(x_s)\) for 100,000 particles initialized from \(\pi_o\). These trajectories were then used to approximate the cost \(\mathbb{E} \int_0^\tau \|\nabla \log \psi^*_s(Y_s)\|^2 \, ds\), which was estimated to be 7.69. Over the 100 independent runs of the SSB algorithm with \(N = 1,000\) particles, the associated cost was estimated to be 8.50 with a standard deviation of 0.036. In other words, the policies approximated with the SSB algorithm yield intermediate distributions that are close to the targets, as shown in Section 5.3.6, but appear to have not completely converged to optimality in terms of cost.

5.4.3 Shortcuts to adiabaticity

In the thermodynamics literature, it is well known that it takes an infinitely long time to transition between equilibrium states of a system that stays in equilibrium with the thermal reservoir. This is the physical intuition for why we require \(\tau \to \infty\) to satisfy \(X_s \sim \pi\) for all \(s\) in (5.45) with \(b_s = \frac{1}{\tau} \nabla \log \pi_s\). An interesting question is whether such transitions can be realized in a finite time \(\tau\) by allowing the system to not be in equilibrium with the thermal reservoir. A process that achieves such a transition is said to be a shortcut to adiabaticity (see e.g. Patra and Jarzynski, 2017).

Solutions of both the two-marginal Schrödinger bridge problem with Langevin dynamics reference and the flow transport problem yield such shortcuts. Maps \(s \mapsto -\log \psi_s\) that are feasible for the flow transport problem are often called counterdiabatic potentials, as the system follows the adiabatic evolution \(s \mapsto \pi_s\). On the other hand, maps that solve the two-marginal Schrödinger bridge problem are often called fast-forward potentials, as they allow the intermediate distributions to deviate from \(\pi_s\), but return to the adiabatic evolution as \(s \to \tau\). In addition to the stochastic setting considered here, there has been growing interest in defining protocols that achieve shortcuts to adiabaticity in both quantum and classical Hamiltonian systems; see e.g. the recent survey of del Campo and Kim (2019). We hope that the methods developed in this chapter can potentially be useful in these fields.
5.4.4 Particle filtering

Consider a hidden Markov chain \( \{X_t\}_{t \in [0:T]} \) with distribution

\[
\mathbb{Q}(dx_{0:T}) = \pi_0(dx_0) \prod_{t=1}^{T} f_t(x_{t-1}, dx_t),
\]

and a sequence of observations \( \{Y_t\}_{t \in [0:T]} \) assumed to be conditionally independent given \( \{X_t\}_{t \in [0:T]} \) and distributed with densities \( g_t(x_t, \cdot) \). The goal in particle filtering is develop online approximations the sequence of filtering distributions \( \pi_t \), defined as the marginal laws of the hidden states \( X_t \) given a realization of the observations \( y_{1:t} \). Given \( \pi_{t-1} \), the filtering distribution at time \( t \) can be written

\[
\pi_t(dx_t) = \int_{E} g_t(x_t, y_t) f_t(x_{t-1}, dx_t) \pi_{t-1}(dx_{t-1}). \quad (5.72)
\]

We could envision solving the multi-marginal Schrödinger bridge problem associated with the reference measure \( \mathbb{Q} \) and marginal constraints \( \{\pi_t\}_{t \in [0:T]} \) using the sequential algorithm developed in Section 5.3.2. Recall that such an approach would require estimates of pointwise evaluations of Radon-Nikodym derivatives of the form \( d\pi_t/dq_t \). This estimation is harder in the filtering setting than in the SMC sampling setting considered earlier, in part due to the intractability of pointwise evaluations of \( \pi_t \) and any of its unnormalized versions.

Instead of relying on exact evaluations of (unnormalized) filtering distributions, we can use sample-based estimates, at the cost of \( N^2 \) density evaluations per iteration of IPPF. Assuming \( \{X^n_{t-1}\}_{n \in [1:N]} \) are distributed according to \( \pi_{t-1} \) and that \( X^n_t \sim f_t(X^n_{t-1}, \cdot) \) for each \( n \in [1 : N] \), an estimate of \( d\pi_t/dq_t \) evaluated at \( X^n_t \) is given by

\[
g_t(X^n_t, y_t) \sum_{k=1}^{N} f_t(X^k_{t-1}, X^n_t) \over \sum_{k=1}^{N} f_t(X^k_{t-1}, X^n_t).
\]

Note that these estimates are biased, but can be made unbiased in the following way. First,

\[\text{Note that resulting multi-marginal Schrödinger bridge would be different from the smoothing distribution (that is, the law of } X_{0:T} \text{ given } y_{1:T} \text{), as this does not have marginals given by } \{\pi_t\}_{t \in [0:T]} \text{.} \]
assuming that \( N \) is even, we can split the particles into the sets \( \{ (X_{t-1}^n, X_{t}^n) \}_{n \in [1: N/2]} \) and \( \{ (X_{t-1}^n, X_{t}^n) \}_{n \in [(N/2+1):N]} \) and estimate the numerator and denominator independently. Second, we can recognize that for any \( n \in [1 : N/2] \), \( \left( \sum_{k=1}^{N/2} f_t (X_{t-1}^k, X_{t}^k) \right)^{-1} \) is an unbiased estimator of \( \left( q_{t}^y (X_{t}^n) \right)^{-1} = (\int_E f_t (x_{t-1}, X_{t}^n) \pi_{t-1}(dx_{t-1}))^{-1} \), which follows from an auxiliary variable argument of the particle MCMC type (Andrieu et al., 2010).

We illustrate this approach on a simple linear quadratic Gaussian model, in which the hidden states \( \{X_t\}_{t \in [0:T]} \) arise as a discretization of \( dX_t = AX_t ds + dW_t \) over the interval [0, \( \tau \)], and where \( W_t \) denotes a Brownian motion and \( \pi_0 (dx_0) = \mathcal{N}(x_0; 0, \mathcal{I}_d)dx_0 \). In other words, \( f_t (x_{t-1}, dx_t) = \mathcal{N}(x_t; x_{t-1} + hAx_{t-1}, h\mathcal{I}_d)dx_t \) for a step size \( h > 0 \) such that \( h\tau = \tau \). The matrix \( A \in \mathbb{R}^{d \times d} \) is defined by \( A_{ij} = A_{ij} = a_{i-j+1} \). The observation densities are given by \( g_t (x_t, dy_t) = \mathcal{N}(y_t; x_t, \sigma^2 I_d)dy_t \). In this setting, the filtering distributions can be calculated exactly using a Kalman filter, to which we compare the Schrödinger bridge particle filter discussed above. We also make comparisons with the classical bootstrap particle filter (Gordon et al., 1993).

In Figure 5.4.1 we illustrate numerical results in the setting where \( d = 2 \), \( a = 0.1 \), \( \sigma = 0.1 \), \( \tau = 2 \) and \( T = 80 \). In Figure 5.4.1a, we use the exact expressions for \( \{ \pi_t \}_{t \in [0 : T]} \) derived with a Kalman filter to perform exact IPFP, and plot log \( \mathcal{W}_2 (\pi_t, q_{t}^{(i)}) \) for \( t \in [1 : T] \) and \( i \in [1 : 5] \). For the particle based methods, we took the number of particles to be \( N = 200 \) for the Schrödinger bridge-based method and \( N = 150,000 \) for the bootstrap particle filter, leading to average running times of 9.09s and 9.74s respectively over 100 independent runs on an Intel Core i5 (2.5 GHz). For the Schrödinger bridge particle filter, the we used IPFP with early stopping. In Figure 5.4.1b, we plot the log 2-Wasserstein distance between \( \pi_t \) and Gaussian distributions with means and covariance matrices estimated using the particle systems, illustrating that the Schrödinger bridge particle filter tends to yields approximations closer than the bootstrap particle filter, with the exception of certain times \( t \in [1 : T] \). In Figures 5.4.1c and 5.4.1d we also plot the normalizing constant (or marginal likelihood) estimates and effective sample sizes as percentage of the sample size, again illustrating improvements of the Schrödinger bridge scheme relative to the bootstrap particle filter.

However, we do not anticipate that the Schrödinger bridge particle filter using the Radon-Nikodym derivative estimates discussed here will scale well with dimension of the state.
space. This is because larger dimension $d$ would require larger sample sizes $N$ to yield estimates with sufficiently low variance, but the cost of increasing $N$ is quadratic. An alternative filtering scheme based on sample-based approximations of $\pi_t$ and Schrödinger bridges have recently been proposed by Reich (2018), but does not reduce to a standard particle filter and therefore does not benefit from theoretical results developed for such methods, unlike the approach presented here. We intend to return to the problem of applying Schrödinger bridges within state space modeling in future work.

5.5 Numerical experiments

In this section, we illustrate the proposed method in two different settings. In the first, we compare the performance of the SSB sampler to a standard SMC sampler in a high-dimensional Gaussian model. In the second, we apply the SSB sampler to a Bayesian logistic regression model with an instance of the weakly informative priors recommended by Gelman et al. (2008). Because these priors are non-Gaussian and consequently not conjugate with the Gaussian transition kernels, they represent a setting in which controlled SMC (Heng et al., 2017) is not directly applicable.

5.5.1 High-dimensional LQG

In this section, we apply the SSB sampler to a high-dimensional version of the Gaussian model discussed in Sections 5.2.6 and 5.3.6. That is, we consider a prior distribution $\pi_o$ with $\mu_o = 0$ and $\Sigma_o = \mathcal{I}_d$ and a log-likelihood defined by $\ell(x) = -(y - x)^\top R^{-1}(y - x)/2$ for some $y \in \mathbb{R}^d$ and symmetric positive definite $R \in \mathbb{R}^{d \times d}$. This defines a Gaussian posterior distribution $\pi = \pi_T$ with covariance matrix $\Sigma_T = (\Sigma_o^{-1} + R^{-1})^{-1}$ and mean $\mu_T = \Sigma_T (\Sigma_o^{-1} \mu_o + R^{-1} y)$. The intermediate distributions on the geometric path (5.1) are also Gaussian: $\pi_t(dx_t) = \mathcal{N}(x_t; \mu_t, \Sigma_t)dx_t$ with $\Sigma_t = (\Sigma_o^{-1} + \lambda_t R^{-1})^{-1}$ and $\mu_t = \Sigma_t (\Sigma_o^{-1} \mu_o + \lambda_t R^{-1} y)$. We consider the discretized Langevin reference process defined by the kernels (5.47) with $h = \tau/T$ and $\lambda_t = ht/T$ and set $\tau = 2$ and $T = 40$. We let $y = (\xi, \ldots, \xi)^\top$ and $R$ have 1’s on the diagonal and $\rho$’s on the off-diagonal. Throughout the experiments, we fix $\xi = 25$ and $\rho = 0.8$, and investigate the effect of the dimension of
Figure 5.4.1: Results of the numerical experiments performed for the LQG model discussed in the particle filtering setting of Section 5.4.4. In Figure 5.4.1a, we perform exact IPFP and plot \( \log \mathcal{W}_2(\pi_t, q_t^{(i)}) \) for \( t \in [1 : T] \) and \( i \in [1 : S] \), where \( \pi_t \) was derived using a Kalman filter. In Figure 5.4.1b, we plot the log 2-Wasserstein distances between \( \{\pi_t\}_{t \in [1 : T]} \) and their Gaussian approximations based on the Schrödinger bridge and bootstrap particle filters. In Figures 5.4.1c and 5.4.1d we plot the marginal likelihood estimates and effective sample sizes as a percentage of \( N \) for the two particle filters. In each plot based on particle approximations, solid lines correspond to median values calculated over 100 independent simulations, and the corresponding confidence bands represent the 5\% and 95\% quantiles.
We compare the standard SMC sampler defined by the reference process with three
versions of the SSB sampler. The first method corresponds to the same algorithm
applied in the two-dimensional LQG setting of Section 5.3.6. The second method uses the
modification of this algorithm discussed in Section 5.3.2, in which we refresh the particles
\( \{X_{t,n}^n\}_{n \in [1:N]} \) using a Markov kernel that is invariant to \( \pi_{t-1} \), for each iteration of IPFP, in
order to prevent the regression-based approximation of \( \psi_t \) from overfitting. In particular,
we used one step of the Metropolis-adjusted Langevin algorithm (Roberts and Tweedie,
1996) scaled with a diagonal covariance matrix estimated using \( \{X_{t,n}^n\}_{n \in [1:N]} \) and step-size
chosen to be \( 3/d^{n/3} \). The third method uses the Euler-Maruyama approximation to the
exact Markov kernel twisting discussed in Section 5.2.4 in combination with refreshment
steps. In each method we used \( M = 0 \) cSMC iterations and resampled the particles ac-
gording to the incremental importance weights at each time \( t \).

In \( d \)-dimensional space, parameterizing the quadratic function class containing policies
of the kind
\[
- \log \psi_t = x_t^T A_t x_t + x_t^T b_t + c_t \]
requires a total of \( d(d+1)/2 + d + 1 \) parameters. By restricting \( A_t \) to be diagonal, this number is reduced to \( 2d + 1 \). In the following simulations, we use the fully parameterized form of \( A_t \) when \( d = 4, 8 \) and the diagonal form of \( A_t \) when \( d = 32, 64 \). We compare impact of the two different parameterization in the \( d = 16 \) case.

Of the three SSB samplers, the one using refreshment steps and exact kernel twisting
was the most computationally expensive. For this method, we set the number of particles
to be \( N = 1,000 \) for \( d = 4, 8, 16 \), and \( N = 2,000 \) for \( d = 32, 64 \). The sample sizes of the
other schemes were tuned so that the four methods ran in approximately the same wall-
clock time; for each simulation setting, the most time consuming method among the four
took no more than 10% longer than the least time consuming (averaged over 100 runs).
For each SSB method, we used IPFP with warm starts, early stopping and a maximum of
\( I = 100 \) iterations, while setting \( \mathcal{K} = [0 : T] \).

In Figure 5.5.1, we plot the \( 2 \)-Wasserstein distance between the target sequence \( \{ \pi_t \}_{t \in [0:T]} \)
and the marginal distributions induced by the different approximate IPFP schemes (left
column), and the RMSE of the corresponding log-normalizing constant estimators \( \log \tilde{Z}_t \)
(right column). Across all dimensions, we observe that the SSB samplers outperform the
standard SMC sampler in estimating the log-normalizing constant. For \( d = 4, 8 \), there ap-
pears to be only limit benefit in including refreshment steps, but seems to become increasingly useful as the dimension increases. Moreover, the Euler-Maruyama approximation provides a reasonable alternative to exact twizzling in all the examples. For instance, in the $d = 64$ setting, the RMSE of the SMC log-normalizing constant estimator was 21 and 17 times higher than the SSB estimators using exact twizzling and Euler-Maruyama twizzling, respectively.

On the other hand, the policies obtained with the different SSB schemes appear to yield marginal distributions $q_t^{(i)}$ that are of similar 2-Wasserstein distance from the targets, at least for $t$ close to $T$. This illustrates that the quality of the marginal approximations alone do not explain difference in performance of the log-normalizing constant estimators. This is further illustrated in the $d = 16$ case, in which the distances between the marginals and the targets are observed to increase when moving from the full parameterization of $A_t$ to the diagonal parameterization, while the performance of the corresponding SSB samplers do not appear to deteriorate. On the contrary, the SSB sampler without rejuvenation looks to benefit from using the diagonal representation. A possible explanation for this behavior is that the relatively high number of parameters in the full parameterization leads the regression within the IPFP steps to overfit, and that using the diagonal parameterization acts as a regularization. The fact that there is only marginal difference between the performance of the SSB samplers with rejuvenation between the full and diagonal parameterizations suggests that refreshing the samples can help mitigate such overfitting. Other approaches to prevent overfitting, such as estimating the policies with penalized regression, would also be worth investigating.

5.5.2 Bayesian logistic regression

We fit a Bayesian logistic regression model to the Cleveland heart disease database in the UCI machine learning repository\(^5\), which contains information about the presence of heart disease in 303 patients as well as measurements of 13 other predictors. We removed 6 en-

---

tries with missing covariates, and turned categorical variables with more than two categories into corresponding binary vectors, leaving us with a data set with $M = 297$ individuals, each with $d = 20$ binary or continuous covariates. Additionally, we made the response variable binary by only indicating the presence or absence of heart disease even though the data set contained four different categories of disease presence.

We use a prior distribution on the $d = 20$ regression coefficients from the class of weakly informative default priors for logistic regression suggested by Gelman et al. (2008). In particular, we first shift all the binary predictors to have mean 0 and to differ by 1 in their lower and upper conditions. Similarly, the non-binary predictors are shifted to have means of 0 and normalized to have standard deviations equal to 0.5. This puts all the variables on the same scale. Next, we place independent $t$-distributions with centers 0, scales 2.5 and four degrees of freedom on each of the coefficients. The log-likelihood is defined via the logistic link function and is given by

$$
\ell(x) = y^T X x - \sum_{m=1}^{M} \log(1 + \exp(x^T X_m)),
$$

where $y \in \{0, 1\}^M$ is the response variable, $X_m \in \mathbb{R}^d$ is the $m$-th row of the covariate matrix $X \in \mathbb{R}^{M \times d}$. The gradient of the log-likelihood is in turn given by

$$
\nabla \ell(x) = X^T y - \sum_{m=1}^{M} (1 + \exp(-x^T X_m))^{-1} X_m.
$$

To draw samples from the resulting posterior distribution, we obtain a reference process by discretizing the Langevin diffusion associated with the interpolation (5.1), setting $\tau = 2$, $T = 40$ and $h = \tau / T$ as in the previous sections. Unlike the earlier sections, however, we choose $\lambda_t = t^2 / T^2$ for each $t \in [0 : T]$. Compared to the linear interpolation, this makes consecutive distributions $\pi_t$ and $\pi_{t+1}$ closer when $t$ is small. Using the quadratic interpolation appeared to make the samplers better behaved, which is possibly because the prior distribution is dispersed compared to the likelihood by virtue of being weakly informative. We used the SSB sampler with potentials estimated within the quadratic function class, which, unlike in the Gaussian setting, is no longer guaranteed to contain the opti-
marginal policy. Additionally, we restricted the matrix $A_t$ to be diagonal. Again, we used IPFP with warm starts and early stopping, allowing a maximum of $I = 20$ IPFP iterations per $t$. For each IPFP iteration, we refresh the particles using the same MALA kernel as in Section 5.5.1 with a step-size equal to $1/d^{1/3}$. The number of particles was chosen to be $N = 2,000$. We compare the SSB sampler with the standard SMC sampler induced by the reference process, for which we tuned the number of particles to be such that the two algorithms ran in the same amount of wall-clock time. Over 100 independent runs, the percentage ESS at time $t = T$ was on average around 85% for both the SSB and SMC samplers. However, the corresponding log-normalizing constant estimators were on average -126.47 and -128.11 with standard deviations of 0.034 and 1.47 respectively, illustrating a large reduction in variance.

5.6 Discussion

In this chapter, we have presented a new Sequential Monte Carlo (SMC) sampling algorithm based on a novel method to sequentially approximate the solution of a multimarginal Schrödinger bridge problem, which we termed the Sequential Schrödinger bridge (SSB) sampler. The algorithm uses a new method to compute two-marginal Schrödinger bridges based on an approximate version of the iterative proportional fitting procedure (IPFP), also known as Sinkhorn’s algorithm. We showed that the first iteration of IPFP can be seen as finding the optimal SMC auxiliary target distribution derived by Del Moral et al. (2006) for a given set of Markov kernels, and that further iterations modify the kernels in such a way that the induced joint distribution converges to the Schrödinger bridge.

By making use of the problem’s equivalent formulation in terms of optimal control, we showed how to formulate the steps of the IPFP algorithm as policy refinements and how to estimate them using the approximate dynamic programming (APD) methodology developed by Heng et al. (2017). Unlike the controlled SMC algorithm they propose, the SSB sampler is applicable when the initial distribution is non-Gaussian. Furthermore, by analogy with the continuous-time formulation of the Schrödinger bridge problem with Langevin diffusion reference process, we proposed an Euler-Maruyama approximation to the APD algorithm that can alleviate the need for using quadratic function classes when
estimating the policies.

We illustrated our approach in various numerical experiments, including high-dimensional linear quadratic Gaussian and Bayesian logistic regression models. While controlling for computational time, we showed that the SSB sampler outperforms a standard SMC algorithm in the estimation of log-normalizing constants, in some cases reducing the RMSE of the estimators by several orders of magnitude. In the Gaussian setting, we also showed that the Euler-Maruyama approximation of the ADP algorithm provided a reasonable alternative to its exact counterpart. However, there are several methodological extensions left to consider. In particular, applying more general and flexible policy approximation methods, such as penalized regression and deep learning, could potentially lead to improvements. There are also several theoretical aspects that are left to consider, such as a more thorough analysis of the asymptotic properties of the Schrödinger bridge approximation as \( N \) and \( I \) (and also \( M, R \) and \( T \)) grows, and how their relative sizes impact the efficiency of the method. For instance, the IPFP algorithm is known to converge exponentially fast in many settings, but has not yet been proven under weak conditions in the continuous state-space setting we consider here. Moreover, we have not formally studied the behavior of the IPFP algorithm when the function classes used within ADP are misspecified in the sense that they do not contain the optimal policy given by the underlying system of Schrödinger equations.

The Schrödinger bridge problem falls at the intersection of many different literatures, including probability theory, optimal transport, control theory, and physics. We have discussed some of the connections between our methodology and computational approaches to approximate the 2-Wasserstein distance between two distributions, the flow transport problem, and the problem of constructing shortcuts to adiabaticity in thermodynamics. We anticipate that further exploration into these and other related problems, such as particle filtering and inference for diffusion processes, would be interesting. The Schrödinger bridge problem is also a special case of a minimum relative entropy principle, in which one minimizes the KL divergence to a reference distribution under constraints that are more general than restricting the marginals. Such problems arise in a variety of different challenging computational settings, e.g. in modeling of protein folding, and is an interesting direction for future work.
Figure 5.5.1: Part 1/3. Distances between \( \pi_t \) and marginals of the IPFP-based approximations \( q_t^{(i)} \) of the multi-marginal Schrödinger bridge, measured as \( \log W_2(\pi_t, q_t^{(i)}) \), for the LQG setting of Section 5.5.1 with discretized Langevin diffusion reference dynamics (left column), and log-RMSE of the log-normalizing constant estimates \( \log \hat{Z}_t \) (right column). The red lines correspond to the reference process and the estimators obtained with the corresponding SMC sampler. The blue and purple lines correspond to the SSB sampler with exact twisting with and without refreshment steps, respectively. The green line correspond to the SSB sampler using Euler-Maruyama approximate twisting and refreshment steps. Figure continued on the next page.
\((e)\) \(d = 16\), \(\log \mathcal{W}_2(\pi_t, q_t^{(i)})\), policies estimated using full \(A_t\).

\((f)\) \(d = 16\), log-RMSE of \(\log \hat{Z}_t\), policies estimated using full \(A_t\).

\((g)\) \(d = 16\), \(\log \mathcal{W}_2(\pi_t, q_t^{(i)})\), policies estimated using diagonal \(A_t\).

\((h)\) \(d = 16\), log-RMSE of \(\log \hat{Z}_t\), policies est. using diagonal \(A_t\).

**Figure 5.5.1: Part 1/3.** Distances between \(\pi_t\) and marginals of the IPFP-based approximations \(q_t^{(i)}\) of the multi-marginal Schrödinger bridge, measured as \(\log \mathcal{W}_2(\pi_t, q_t^{(i)})\), for the LQG setting of Section 5.5.1 with discretized Langevin diffusion reference dynamics (left column), and log-RMSE of the log-normalizing constant estimates \(\log \hat{Z}_t\) (right column). The red lines correspond to the reference process and the estimators obtained with the corresponding SMC sampler. The blue and purple lines correspond to the SSB sampler with exact twisting with and without refreshment steps, respectively. The green line corresponds to the SSB sampler using Euler-Maruyama approximate twisting and refreshment steps. Figure continued on the next page.
Figure 5.5.1: Part 3/3. Distances between $\pi_t$ and marginals of the IPFP-based approximations $q_t^{(l)}$ of the multi-marginal Schrödinger bridge, measured as $\log \mathcal{W}_2(\pi_t, q_t^{(l)})$, for the LQG setting of Section 5.5.1 with discretized Langevin diffusion reference dynamics (left column), and log-RMSE of the log-normalizing constant estimates $\log \tilde{Z}_t$ (right column). The red lines correspond to the reference process and the estimators obtained with the corresponding SMC sampler. The blue and purple lines correspond to the SSB sampler with exact twisting with and without refreshment steps, respectively. The green line correspond to the SSB sampler using Euler-Maruyama approximate twisting and refreshment steps.
6

Conclusion

6.1 Summary

In this manuscript, we have investigated different applications of optimal transport in statistical inference and computation. Here we summarize our main contributions and briefly discuss a few open problems and avenues of future research.

In Chapter 2, we derived some statistical properties of estimators constructed by minimizing the Wasserstein distance between a parametric model and an empirical distribution. Without assuming that the model is well specified, we established consistency of two versions of minimum Wasserstein estimators, one originally proposed by Bassetti et al. (2006) and one that is more amenable to computation in generative models. These results provide theoretical support for recent applications of minimum Wasserstein estimators in complex generative models to which previous results do not apply. By making use of a convenient representation of the Wasserstein distance, we also derive the asymptotic distribu-
tion of the minimum Wasserstein estimator of order 1 when the data are one-dimensional. Although the asymptotic distribution is itself not useful in the construction of confidence intervals due to its dependence on unknown quantities, its existence provides asymptotic validity of confidence sets constructed via subsampling. We also illustrated our theoretical results in several simple examples.

In Chapter 3, we proposed an approach to likelihood-free approximate Bayesian computation (ABC) in generative models using the Wasserstein distance. Our results show that the Wasserstein ABC posterior can, under some conditions, yield consistent approximation of the posterior distribution as the ABC threshold converges to zero. This is in contrast with standard ABC methods based on summary statistics, which can suffer from a systematic loss of information. We also discussed some other theoretical features of the Wasserstein ABC posterior, including the derivation of an asymptotic regime in which the Wasserstein ABC posterior concentrates around the point in the parameter space that minimizes the Wasserstein distance between the model and the data-generating process. The method was illustrated on variety of examples, showing that it in many settings appears to outperform existing ABC approaches.

In Chapter 4, we showed how certain Monte Carlo sampling algorithms based on the discretization of Langevin diffusion can be viewed as deterministic optimization algorithms in the space of probability measures. Langevin diffusion can, via the Fokker-Planck equation, be seen as the Wasserstein gradient flow of KL divergence with respect to the target distribution. In turn, Langevin Monte Carlo can be seen as an approximation of this gradient flow. We characterize the approximation error as stemming from the operator splitting approach to solving the Fokker-Planck equation, which can equivalently been seen as a splitting approach to a composite minimization problem. Using this optimization perspective, we derive some new non-asymptotic results a proximal version of the algorithm and its dependence on dimension.

In Chapter 5, we presented a new Sequential Monte Carlo (SMC) sampling algorithm based on the Schrödinger bridge problem. The Schrödinger bridge is defined as the interpolation between the initial and target distributions that is closest to some reference process in KL divergence, and hence provides a means to sample from the target by propagating particles through it. We approximate the Schrödinger bridge using an iterative
proportional fitting procedure based on approximate dynamic programming, utilizing its equivalent formulation as the solution to an optimal control problem. The Schrödinger bridge can also be viewed as a regularization of the optimal transport coupling, and this connection underpins several recently proposed numerical schemes in computational optimal transport. We discussed some interesting aspects of this connection and presented new links with flow transport and shortcuts to adiabaticity in physics.

6.2 Future work and open problems

Despite many recent methodological and theoretical advances at the intersection of optimal transport and statistics, several important questions remain open. For instance, beyond the one-dimensional setting considered in Chapter 2, the rate of convergence and asymptotic distribution of the minimum Wasserstein estimator is unknown. This is in part due to our still incomplete understanding of empirical processes in the space of probability measures metrized by the Wasserstein distance. This topic is an active area of research in its own right, with recent contributions by e.g. Fournier and Guillin (2015); Weed and Bach (2019); del Barrio and Loubes (2019). What we do know is that the rate of convergence of empirical measures in the Wasserstein distance is typically slow. For instance, if \( \mu_* \ll \mu_{\text{Leb}} \) and \( \hat{\mu}_n \) is the empirical measure of \( n \) independent draws from \( \mu_* \), then Weed and Bach (2019) show that

\[
n^{-1/t} \lesssim \mathbb{E} \mathcal{W}_t(\mu_*, \hat{\mu}_n) \lesssim n^{-1/s},
\]

for any \( t < d < s \) and \( d \geq 4 \). In contrast, we know that the distribution of \( \mathcal{W}_t^2(\mu_*, \hat{\mu}_n) \) concentrates around \( \mathbb{E} \mathcal{W}_t^2(\mu_*, \hat{\mu}_n) \) at the \( \sqrt{n} \) rate for any \( d \) (del Barrio and Loubes, 2019). How these different rates impact the convergence of the minimum Wasserstein estimator and other Wasserstein-based inference procedures remains unclear.

In practice, the Wasserstein distance is often computed using approximate schemes such as the one introduced by Cuturi (2013). The statistical properties of these approximations are mostly unknown, with the exception of recent contributions by Genevay et al. (2018a) and Rigollet and Weed (2018). The former of these showed that the empirical measure
enjoys faster rates of convergence in the entropically regularized Wasserstein distance (or Sinkhorn divergence) than in the Wasserstein distance itself. This has lead to the hypothesis that regularization can also improve the statistical properties of minimum Wasserstein estimators and other inference procedures, but has yet to be investigated formally.

Optimal transport provides a useful framework for studying connections between optimization and sampling, which Chapters 4 and 5 are both examples of. The perspective in Chapter 4 of viewing sampling algorithms as optimization algorithms in the space of probability measures metrized by the Wasserstein distance can be extended to other stochastic processes than the discretized Langevin diffusion considered there. Another interesting feature of this approach is that it can inspire new sampling algorithms by drawing on insights made in the optimization literature. With the exception of recent work by Wibisono (2018) and Ma et al. (2019), this research direction is still largely unexplored.

Despite its long history in physics, probability and control theory, the Schrödinger bridge problem has not yet received much attention from statisticians. Our use of the Schrödinger bridge in Sequential Monte Carlo sampling is only scratching the surface of its potential applications in computational statistics. For instance, we anticipate that further explorations into its use in particle filtering and inference for diffusion processes would likely be fruitful. The Schrödinger bridge problem is also a special case of a minimum relative entropy principle with an even larger span of applications. Instead of minimizing the KL divergence subject to constraints on the marginals of the joint distribution as in the Schrödinger bridge problem, one could instead imagine constraining other features of the joint. Such problems arise in a variety of different settings, e.g. in modeling of protein folding, and is a very interesting avenue of future research.
Appendix to Chapter 2

A.1 Preliminary results

Before proving the results stated in Chapter 2 of the main text, we first provide some preliminary results.

A sequence of probability measures \((\mu_n)_{n \geq 1}\) is said to converge weakly in \(\mathcal{P}_p(\mathcal{Y})\) to \(\mu\) as \(n \to \infty\) if \(\mu_n \Rightarrow \mu\), i.e. converges weakly in the usual sense, and there exists \(y_o \in \mathcal{Y}\) such that \(\int_{\mathcal{Y}} \rho(y, y_o)^p d\mu_n(y) \to \int_{\mathcal{Y}} \rho(y, y_o)^p d\mu(y)\). Recall that \(\mathcal{Y}\) is a subset of \(\mathbb{R}^d\) for \(d \geq 1\).

**Theorem A.1.1.** The \(p\)-Wasserstein distance metrizes weak convergence in \(\mathcal{P}_p(\mathcal{Y})\): a sequence \(\mu_n\) converges weakly to \(\mu\) in \(\mathcal{P}_p(\mathcal{Y})\) if and only if \(\mathcal{W}_p(\mu_n, \mu) \to 0\).

For a proof, see Villani (2008, Theorem 6.9). From this result one can deduce the continuity of the \(p\)-Wasserstein distance. If \(\mu_n\) and \(\nu_n\) converge weakly in \(\mathcal{P}_p(\mathcal{Y})\) to \(\mu\) and \(\nu\), then \(\mathcal{W}_p(\mu_n, \nu_n) \to \mathcal{W}_p(\mu, \nu)\). On the other hand, if \(\mu_n\) and \(\nu_n\) converge weakly in the usual
sense, the Wasserstein distance is only lower semicontinuous:

$$\liminf_{n \to \infty} \mathcal{W}_p(\mu_n, \nu_n) \geq \mathcal{W}_p(\mu, \nu).$$

The following lemma is extended from Bassetti et al. (2006). Its second condition corresponds to Assumption A.2.2, and is implied by the first condition. All limits in the lemma are understood to be as $n \to \infty$.

**Lemma A.1.1.** Let $(\theta_n)_{n \geq 1}$ be a sequence in $\mathcal{H}$ and $\theta \in \mathcal{H}$. Suppose that either of the following conditions holds.

1. $\rho_\mathcal{H}(\theta_n, \theta) \to 0$ implies that $\mathcal{W}_p(\mu_{\theta_n}, \mu_{\theta}) \to 0$.

2. $\rho_\mathcal{H}(\theta_n, \theta) \to 0$ implies that $\mu_{\theta_n} \Rightarrow \mu_{\theta}$.

Then, respectively,

1. $\mathcal{H} \times \mathcal{P}(\mathcal{Y}) \ni (\theta, \mu) \mapsto \mathcal{W}_p(\mu_{\theta}, \mu)$ is continuous.

2. $\mathcal{H} \times \mathcal{P}(\mathcal{Y}) \ni (\theta, \mu) \mapsto \mathcal{W}_p(\mu_{\theta}, \mu)$ is lower semicontinuous.

**Proof.** This follows directly from the two assumptions and the continuity/lower semicontinuity derived from Theorem A.1.1. □

**Lemma A.1.2.** The function $(v, \mu^{(m)}) \mapsto \mathbb{E}\mathcal{W}_p(v, \hat{\mu}_m)$ is lower semicontinuous with respect to weak convergence. Furthermore, if $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ implies that $\mu_{\theta_n}^{(m)} \Rightarrow \mu_{\theta}^{(m)}$, then the map $(v, \theta) \mapsto \mathbb{E}\mathcal{W}_p(v, \hat{\mu}_{\theta,m})$ is lower semicontinuous.

**Proof.** Let $\mu_k^{(m)} \Rightarrow \mu^{(m)}$ and $v_k \Rightarrow v$. Then there exist versions of the corresponding empirical measures such that $\hat{\mu}_{k,m} \Rightarrow \hat{\mu}_m$ almost surely. Indeed, by Skorokhod’s representation theorem, there exists a probability space $(\hat{\mathbb{P}}, \hat{\Omega}, \hat{\mathcal{E}})$ and random variables $\hat{X}_{k,m} \sim \mu_k^{(m)}$ and $\hat{X}_{1:m} \sim \mu^{(m)}$ such that $\hat{X}_{k,m} \Rightarrow \hat{X}_{1:m}$ $\hat{\mathbb{P}}$-almost surely. Let $\hat{\mu}_{k,m}$ and $\hat{\mu}_m$ be the empirical distributions of these samples. By Varadarajan (1958b) and since $\mathcal{Y}$ is separable, there exists a fixed countable subset $C^*$ of continuous
and bounded functions on $\mathcal{Y}$, such that for any sequence of measures $\mu_n \in \mathcal{P}(\mathcal{Y})$, $\mu_n$ converges weakly to $\mu$ if and only if $\int f \, d\mu_n \to \int f \, d\mu$ for all $f \in C^*$. Fix one such $f$. Then,

$$\int f \, d\hat{\mu}_{k,m} = \frac{1}{m} \sum_{i=1}^{m} f(\hat{X}^i_k) \to \frac{1}{m} \sum_{i=1}^{m} f(\hat{X}^i) = \int f \, d\hat{\mu}_m,$$

on a set of $\mathbb{P}$-probability one, by the continuous mapping theorem. Taking the countable intersection of these sets over $f \in C^*$, we get that $\hat{\mu}_{k,m} \Rightarrow \hat{\mu}_m$ $\mathbb{P}$-almost surely.

By the lower semicontinuity of the $p$-Wasserstein distance and Fatou’s lemma,

$$\mathbb{E} \mathcal{W}_p(v, \hat{\mu}_m) \leq \mathbb{E} \liminf_{k \to \infty} \mathcal{W}_p(v_k, \hat{\mu}_{k,m}) \leq \liminf_{k \to \infty} \mathbb{E} \mathcal{W}_p(v_k, \hat{\mu}_{k,m}).$$

The lower semicontinuity of $(v, \theta) \mapsto \mathbb{E} \mathcal{W}_p(v, \hat{\mu}_{\theta,m})$ is proved analogously to Lemma A.1.1. \hfill \square

### A.2 Proofs: MWE

#### A.2.1 Existence, measurability, and consistency

For ease of presentation, we recall the assumptions made in the main text.

**Assumption A.2.1.** The data-generating process is such that $\mathcal{W}_p(\hat{\mu}_n, \mu_\star) \to 0$ $\mathbb{P}$-almost surely as $n \to \infty$.

**Assumption A.2.2.** The map $\theta \mapsto \mu_\theta$ is continuous in the sense that $\rho_\mathcal{H}(\theta_n, \theta) \to 0$ implies $\mu_{\theta_n} \Rightarrow \mu_\theta$ as $n \to \infty$.

For the next assumption, define $\varepsilon_\star = \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$; we will use this definition throughout.

**Assumption A.2.3.** For some $\varepsilon > 0$, the set $B_\star(\varepsilon) = \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon \}$ is bounded.
Theorem A.2.1 (Existence and consistency of the MWE). Under Assumptions A.2.1-A.2.3, there exists a set \( E \subset \Omega \) with \( \mathbb{P}(E) = 1 \) such that, for all \( \omega \in E \),

\[
\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta),
\]

and there exists \( n(\omega) \) such that, for all \( n \geq n(\omega) \), the sets \( \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \) are non-empty and form a bounded sequence with

\[
\limsup_{n \to \infty} \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \subseteq \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta).
\]

Before giving the proof, we recall a definition and a proposition.

Definition A.2.1. A sequence of functions \( f_n : \mathcal{H} \to \mathbb{R} \) is said to epi-converge to \( f : \mathcal{H} \to \mathbb{R} \) if for all \( \theta \in \mathcal{H} \),

\[
\begin{cases}
\liminf_{n \to \infty} f_n(\theta_n) \geq f(\theta) & \text{for every sequence } \theta_n \to \theta, \\
\limsup_{n \to \infty} f_n(\theta_n) \leq f(\theta) & \text{for some sequence } \theta_n \to \theta.
\end{cases}
\]

A useful equivalent formulation of epi-convergence can be found in Proposition 7.29 of Rockafellar and Wets (2009), paraphrased here.

Proposition A.2.1 (Proposition 7.29 of Rockafellar and Wets (2009)). The sequence \( f_n : \mathcal{H} \to \mathbb{R} \) epi-converges to \( f : \mathcal{H} \to \mathbb{R} \) if and only if

\[
\begin{cases}
\liminf_{n \to \infty} \inf_{\theta \in \mathcal{K}} f_n(\theta) \geq \inf_{\theta \in \mathcal{K}} f(\theta) & \text{for every compact set } \mathcal{K} \subset \mathcal{H}, \\
\limsup_{n \to \infty} \inf_{\theta \in \mathcal{O}} f_n(\theta) \leq \inf_{\theta \in \mathcal{O}} f(\theta) & \text{for every open set } \mathcal{O} \subset \mathcal{H}.
\end{cases}
\]

In an colloquial sense, epi-convergence is a weak notion of convergence for which the minimizer of \( f_n \) converges to the minimizer of \( f \). Showing that the function \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) epi-converges to \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \) almost surely is the key step in the proof of Theorem A.2.1.

Proof of Theorem A.2.1. First note that, for any \( \nu \in \mathcal{P}(\mathcal{Y}) \), the lower semicontinuity of the map \( \theta \mapsto \mathcal{W}_p(\nu, \mu_\theta) \) follows from Lemma A.1.1, via Assumption A.2.2. Next, since
\[
\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta) = \varepsilon_*, \text{ the set } B_*(\varepsilon) \text{ with the } \varepsilon \text{ of Assumption A.2.3 is non-empty, by definition of the infimum. Moreover, since } \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \text{ is lower semicontinuous, the set } B_*(\varepsilon) \text{ is closed. By Assumption A.2.3, } B_*(\varepsilon) \text{ is therefore compact. In other words, again by lower semicontinuity, the set } \text{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta) \text{ is non-empty.}
\]

We now show that \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) epi-converges to \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \) \( \mathbb{P} \)-almost surely. Let \( E \) denote the set of probability one from Assumption A.2.1 and let \( \omega \in E \). Fix \( \mathcal{K} \subset \mathcal{H} \) compact. By lower semicontinuity of \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \), we know that

\[
\inf_{\theta \in \mathcal{K}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}),
\]

for some sequence \( \theta_n = \theta_n(\omega) \in \mathcal{K} \). Hence,

\[
\liminf_{n \to \infty} \inf_{\theta \in \mathcal{K}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \liminf_{n \to \infty} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})
\]

\[
= \lim_{k \to \infty} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_{k_n}}) \quad \exists \text{ subsequence converging to the lim inf,}
\]

\[
= \lim_{m \to \infty} \mathcal{W}_p(\hat{\mu}_{n_m}(\omega), \mu_{\theta_{n_m}}) \quad \exists \text{ subsequence } \theta_{n_m} \to \bar{\theta} \in \mathcal{K} \text{ by compactness,}
\]

\[
= \liminf_{m \to \infty} \mathcal{W}_p(\hat{\mu}_{n_m}(\omega), \mu_{\theta_{n_m}})
\]

\[
\geq \mathcal{W}_p(\mu_*, \mu_{\theta}) \quad \text{by l.s.c., Assumptions A.2.1 and A.2.2, } \omega \in E,
\]

\[
\geq \inf_{\theta \in \mathcal{K}} \mathcal{W}_p(\mu_*, \mu_\theta).
\]

Fix \( \mathcal{O} \subset \mathcal{H} \) open. By definition of the infimum, there exists a sequence \( \theta_n \in \mathcal{O} \) such that \( \mathcal{W}_p(\mu_*, \mu_{\theta_n}) \to \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\mu_*, \mu_\theta) \). Now, \( \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}) \).

Hence,

\[
\limsup_{n \to \infty} \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \limsup_{n \to \infty} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})
\]

\[
\leq \limsup_{n \to \infty} \left( \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_*) + \mathcal{W}_p(\mu_*, \mu_{\theta_n}) \right) \quad \text{by the triangle inequality,}
\]

\[
\leq \limsup_{n \to \infty} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_*) + \limsup_{n \to \infty} \mathcal{W}_p(\mu_*, \mu_{\theta_n}) \quad \text{by positivity,}
\]

\[
= \limsup_{n \to \infty} \mathcal{W}_p(\mu_*, \mu_{\theta_n}) \quad \text{by Assumption A.2.1, } \omega \in E,
\]

158
\[
\inf_{\theta \in \mathcal{D}} \mathcal{W}_p(\mu_\ast, \mu_\theta) \quad \text{by definition of } \theta_n.
\]

Using Proposition A.2.1, the sequence of functions \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \) epi-converges to \( \theta \mapsto \mathcal{W}_p(\mu_\ast, \mu_\theta) \).

Theorem 7.29b) of Rockafellar and Wets (2009) implies that

\[
\limsup_{n \to \infty} \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\ast, \mu_\theta) = \varepsilon_\ast.
\]

So, for all \( a > 0 \), there exists \( n_\ast(\omega) \), such that for \( n \geq n_\ast(\omega) \), \( \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\ast + a \). Let \( a \in (0, \varepsilon/2) \). The set

\[
\{ \theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\ast + \varepsilon/2 \}
\]

is non-empty for \( n \geq n_\ast(\omega) \), by definition of the infimum. Let \( \theta \) belong to this set. Then, by the triangle inequality,

\[
\mathcal{W}_p(\mu_\ast, \mu_\theta) \leq \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\ast) + \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta).
\]

By Assumption A.2.1, there exists an \( n_\varepsilon(\omega) \) such that for \( n \geq n_\varepsilon(\omega) \), \( \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\ast) \leq \varepsilon/2 \). So, if \( n \geq \max\{n_\varepsilon(\omega), n_\ast(\omega)\} \), we have that \( \mathcal{W}_p(\mu_\ast, \mu_\theta) \leq \varepsilon_\ast + \varepsilon \). This means that

\[
\{ \theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\ast + \varepsilon/2 \} \subset B_\varepsilon(\varepsilon).
\]

As a consequence, we know that for \( n \geq \max\{n_\varepsilon(\omega), n_\ast(\omega)\} \), \( \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta \in B_\varepsilon(\varepsilon)} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \).

By Theorem 7.31a) of Rockafellar and Wets (2009), this implies

\[
\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\ast, \mu_\theta).
\]

For \( n \geq \max\{n_\varepsilon(\omega), n_\ast(\omega)\} \) and by the same reasoning as for \( \theta \mapsto \mathcal{W}_p(\mu_\ast, \mu_\theta) \), the sets

\[
\arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)
\]

are non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. The same argument holds for \( \varepsilon_n \)-argmin \( \theta \in \mathcal{H} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \).
with $\varepsilon_n \to 0$, since, eventually, $\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) + \varepsilon_n \leq \varepsilon_* + \alpha$.

\begin{proof}

**Theorem A.2.2** (Measurability of the MWE). Suppose that $\mathcal{H}$ is a $\sigma$-compact Borel measurable subset of $\mathbb{R}^d$. Under Assumption A.2.2, for any $n \geq 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_n : \Omega \to \mathcal{H}$ that satisfies

\[
\hat{\theta}_n(\omega) \in \begin{cases} \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta), & \text{if this set is non-empty}, \\ \varepsilon-\arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta), & \text{otherwise}. \end{cases}
\]

Before the proof, we first recall a useful result from Brown and Purves (1973), also used in Bassetti et al. (2006).

**Theorem A.2.3** (Corollary 1 in Brown and Purves (1973)). Let $X, Y$ be Polish, $D \subset Y \times X$ be Borel, and $f : D \to \mathbb{R}$ be Borel measurable. Suppose that for all $y \in \text{proj}(D)$, the section $D_y = \{x : (y, x) \in D\}$ is $\sigma$-compact and that $f_x = f(y, \cdot)$ is lower semicontinuous with respect to the relative topology on $D_y$. Then

1. The sets $G = \text{proj}(D)$ and $I = \{y \in G : \text{for some } x \in D_y, f(y, x) = \inf f_x\}$ are Borel.

2. For each $\varepsilon > 0$, there exists a Borel measurable function $\varphi_\varepsilon$ such that for $y \in G$,

\[
f(y, \varphi_\varepsilon(y)) = \begin{cases} \inf f_y & \text{if } y \in I, \\ \varepsilon + \inf f_y & \text{if } y \notin I, \inf f_y \neq -\infty, \\ -\varepsilon^{-1} & \text{if } y \notin I, \inf f_y = -\infty. \end{cases}
\]

**Proof of Theorem A.2.2.** First note that $\mathcal{Y}^\infty$ endowed with the product topology is Polish since $(\mathcal{Y}, \rho)$ is Polish. Also, $\hat{\mu}_n(\omega)$ depends on $\omega$ only through $y = Y(\omega)$, where $Y = (Y_t)_{t \in \mathcal{T}}$. We can therefore write $\hat{\mu}_n(\omega) = \hat{\mu}_n(y)$, where $y \in \mathcal{Y}^\infty$, and consider the empirical measure a function on $\mathcal{Y}^\infty$. The map $y \mapsto \hat{\mu}_n(y)$ is measurable with respect to the Borel $\sigma$-algebra on $\mathcal{P}_p(\mathcal{Y})$ with respect to weak convergence. Recall also that $(\mathcal{P}_p(\mathcal{Y}), \mathcal{W}_p)$ is Polish since $\mathcal{Y}$ is Polish by Theorem 6.18 of Villani (2008).
Let $\mathcal{D} = \mathcal{Y}^\infty \times \mathcal{H}$. By Lemma A.1.1 and Assumption A.2.2, the map $(\mu, \theta) \mapsto \mathcal{W}_p(\mu, \mu_\theta)$ is lower semicontinuous (and therefore measurable). Hence the map $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is also lower semicontinuous on $\mathcal{H}$ for any $y \in \mathcal{Y}^\infty$. Being the composition of measurable functions, $(y, \theta) \mapsto \mathcal{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is measurable on $\mathcal{D}$. In light of this, the result follows by a direct application of Theorem A.2.3.

\[ \square \]

A.2.2 Asymptotic distribution

Let $p = 1$, $\mathcal{Y} = \mathbb{R}$, and $\rho(x, y) = |x - y|$. In this case we have

$$\mathcal{W}_1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)| ds = \int_\mathbb{R} |F_\mu(t) - F_\nu(t)| dt,$$

where $F_\mu$ and $F_\nu$ denote the cumulative distribution functions (CDFs) of $\mu$ and $\nu$ respectively (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). For this reason, we will occasionally use the notation $\mathcal{W}_1(\mu, \nu) = \|F_\mu - F_\nu\|_1$. Assume also that $\mathcal{H}$ is endowed with a norm: $\rho_{\mathcal{H}}(\theta, \theta') = \|\theta - \theta'\|_{\mathcal{H}}$. We recall results from del Barrio et al. (1999) and Dede (2009), after a few definitions.

Definition A.2.2. Suppose that the sequence $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto X_n(\omega, t)$ for all $n$, and $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto X(\omega, t)$, are stochastic processes with almost all their sample paths in $L_1(\mathbb{R})$. Then $X_n$ is said to converge weakly to $X$ in $L_1(\mathbb{R})$ if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ as $n \to \infty$ for all bounded continuous functions $f : L_1(\mathbb{R}) \to \mathbb{R}$.

Definition A.2.3. The stochastic process $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto G_n(\omega, t)$ is a $\mu$-Brownian bridge if it is a zero mean Gaussian process with covariance function $\mathbb{E}G_n(s)G_n(t) = \min\{F\mu(s), F\mu(t)\} - F\mu(s)F\mu(t)$.

Theorem A.2.4 (Theorem 2.1a in del Barrio et al. (1999)). Suppose that $Y = (Y_i)_{i \in \mathbb{Z}} \sim \mu_*^\infty$, and define $F_n(\omega, t) = \hat{\mu}_n(\omega)(-\infty, t]$ and $F_*(t) = \mu_*(-\infty, t]$. The stochastic process $\sqrt{n}(F_n - F_*)$ converges weakly in $L_1(\mathbb{R})$ to a $\mu_*^\infty$-Brownian bridge $G_*$ if and only if the condition $\int_0^\infty \mathbb{P}(|Y_0| > t) dt < \infty$ is satisfied.

For a stationary sequence, let $\tilde{a}_t = \sup_{u \in \mathbb{R}} \mathbb{E}|\mathbb{P}(Y_t \leq u | F_{-\infty}^\infty) - \mathbb{P}(Y_t \leq u)|$. Note that for stationary sequences, $\tilde{a}$-mixing is weaker than $a$-mixing, as defined later in Section A.4.
Theorem A.2.5 (Proposition 3.5 in Dede (2009)). Suppose that $Y = (Y_t)_{t \in \mathbb{Z}}$ is ergodic and stationary, and that

$$\sum_{k \geq 1} \frac{1}{\sqrt{k}} \int_0^\infty \min \{ \sqrt{\alpha_k}, \sqrt{\mathbb{P}(|Y_0| > t)} \} dt < \infty.$$  

Then $\sqrt{n}(F_n - F_\star)$ converges weakly in $L_1(\mathbb{R})$ to a zero mean Gaussian process $G_\star$ with sample paths in $L_1(\mathbb{R})$ and covariance satisfying: for every $f, g \in L_\infty(\mathbb{R})$,

$$\mathbb{E}f(G_\star)g(G_\star) = \int_{\mathbb{R}^2} f(s)g(u)C(s, u)ds du,$$

where

$$C(s, u) = \sum_{t \in \mathbb{Z}} \{ \mathbb{P}(X_0 \leq s, X_t \leq u) - F_\star(s)F_\star(u) \}.$$  

Dede (2009) also provides other conditions, e.g. on $\varphi$-mixing coefficients, for which the convergence above holds. We first consider the well-specified setting, in which our results follow directly from Pollard (1980).

Well-specified setting

Suppose that $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of $\mathcal{H}$, and consider the following assumptions:

Assumption A.2.4. For all $\varepsilon > \circ$, there exists $\delta > \circ$ such that

$$\inf_{\theta \in \mathcal{H} : \|\theta - \theta_\star\|_H \geq \varepsilon} \mathcal{W}_i(\mu_{\theta_\star}, \mu_\theta) > \delta.$$  

Assumption A.2.5. There exists a non-singular $D_{\theta_\star} \in (L_1(\mathbb{R}))^{d_\theta}$ such that

$$\int_\mathbb{R} |F_\theta(t) - F_{\theta_\star}(t) - \langle \theta - \theta_\star, D_{\theta_\star}(t) \rangle| dt = o(\|\theta - \theta_\star\|_H),$$

as $\|\theta - \theta_\star\|_H \to \circ$.

The following results contain Theorem 2.2.3 of the main text as a special case.
Theorem A.2.6. Suppose that \( \mu_* = \mu_{\theta_*} \) for some \( \theta_* \) in the interior of \( \mathcal{H} \), and that the conditions of either Theorem A.2.4 or Theorem A.2.5 are satisfied. Under Assumptions A.2.1-A.2.5, the goodness-of-fit statistic satisfies
\[
\sqrt{n} \inf_{\theta \in \mathcal{H}} \mathcal{W}_1(\hat{\mu}_n, \mu_\theta) \Rightarrow \inf_{u \in \mathcal{H}} \int_\mathbb{R} |G_*(t) - \langle u, D_{\theta_*}(t) \rangle| dt,
\]
as \( n \to \infty \), where \( G_* \) is given as in Theorem A.2.4 or Theorem A.2.5 respectively.

Theorem A.2.7. Suppose that the conditions in Theorem A.2.6 hold. Suppose also that the random map \( \mathcal{H} \ni u \mapsto \int_\mathbb{R} |G_*(t) - \langle u, D_{\theta_*}(t) \rangle| dt \) has an almost surely unique infimum. Then the MWE of order 1 satisfies
\[
\sqrt{n}(\hat{\theta}_n - \theta_*) \Rightarrow \arg \min_{u \in \mathcal{H}} \int_\mathbb{R} |G_*(t) - \langle u, D_{\theta_*}(t) \rangle| dt,
\]
as \( n \to \infty \), where \( G_* \) is given as in Theorem A.2.4 or Theorem A.2.5.

Proof. The proofs of these two results follow the steps outlined in Pollard (1980)'s Theorems 4.2 and 7.2 respectively, which also generalize to the setting where the map \( \mathcal{H} \ni u \mapsto \int_\mathbb{R} |G_*(t) - \langle u, D_{\theta_*}(t) \rangle| dt \) does not necessarily have a unique minimum (see also Section A.2.2 below). The delta methods employed therein hold for the 1-Wasserstein distance due to the representation \( \mathcal{W}_1(\mu, \nu) = \|F_\mu - F_\nu\|_{L_1} \). Moreover, the well-separation of \( \theta_* \) provided by Assumption A.2.4, the consistency and measurability of the MWE proved earlier, and Theorems A.2.4 and A.2.5 proved in del Barrio et al. (1999) and Dede (2009) respectively, guarantee that Pollard's conditions are satisfied. Note that the measurability concerns outlined in his Section 3 do not apply to here, as \( L_1(\mathbb{R}) \) is separable.

\[ \square \]

Misspecified setting

To study the asymptotic distribution of the MWE in the misspecified setting, we adapt the arguments outlined in Section 7 of Pollard (1980). Define \( f(x, u) = \|x - \langle u, D_{\theta_*} \rangle\|_L \) and \( m(x) = \inf_u f(x, u) \). Let \( \mathcal{K} \) be the class of all compact, convex, non-empty subsets of a set \( L_1(\mathbb{R}) \) equipped with its canonical distance. The corresponding Hausdorff metric on \( \mathcal{K} \) is
defined by $d_H(K_i, K_j) = \inf\{ \delta > 0 : K_i \subset K^\delta, K_j \subset K^\delta \}$, where $K^\delta = \bigcup_{x \in K}\{ z \in M : \|z - x\|_{L_i} \leq \delta \}$. Let $K(x, \beta) = \{ u : f(x, u) \leq m(x) + \beta \}$. The function $x \mapsto K(x, \beta)$ maps into $\mathcal{K}$ and, by Pollard (1980, Lemma 7.1), is measurable. Let also

$$H_n = \sqrt{n}(F_n - F_{\theta_*}) = \sqrt{n}(F_n - F_\star) + \sqrt{n}(F_\star - F_{\theta_*})$$

and $H_n^* = G_\star + \sqrt{n}(F_\star - F_{\theta_*})$. Let

$$M_n = \{ \theta \in \mathcal{H} : \mathcal{W}_i(\mu_\theta, \mu_\star) \leq \inf_\theta \mathcal{W}_i(\mu_\theta, \mu_\star) + n^{-1/2}\eta_n \}$$

where $\eta_n > \theta$ is any sequence such that $\eta_n = o_P(1)$ and $M_n$ is non-empty. That is, $M_n$ is a set of approximate MWEs.

Consider the following assumption:

**Assumption A.2.6.** There exists a neighborhood $N$ of $\theta_\star$ and a constant $c_\star > 0$ such that for any $\theta \in N$,

$$\mathcal{W}_i(\mu_\theta, \mu_\star) \geq \mathcal{W}_i(\mu_\theta, \mu_\star) + c_\star\|\theta - \theta_\star\|_{\mathcal{H}}.$$  

In the well-specified setting, this condition follows from Assumption A.2.5. The next result concerns the distribution of the set $M_n$ as $n$ becomes large.

**Theorem A.2.8.** Suppose Assumptions A.2.1-A.2.6 hold for some $\theta_\star$ in the interior of $\mathcal{H}_i$ and that the conditions of either Theorem A.2.4 or Theorem A.2.5 are satisfied. Then, there exist positive real numbers $\beta_n \to 0$ such that

1. $P_* \left( \{ M_n \subset \theta_\star + n^{-1/2}K(H_n, \beta_n) \} \right) \to 1$ as $n \to \infty$, where $P_*$ denotes inner probability, and

2. if $F_\star$ and $G_\star$ are versions of the processes such that $\sqrt{n}(F_n - F_\star) \to G_\star$ in $L_i(\mathbb{R})$ almost surely, then $d_H(K(H_n^*, o), K(H_n, \beta_n)) = o_P(1)$.

Since $K(H_n^*, o) = \arg\min_u \|G_\star + \sqrt{n}(F_\star - F_{\theta_*}) - \langle u, D_{\theta_*} \rangle\|_{L_i}$, one can interpret this result as saying that the limit of the set of approximate MWEs $M_n$ behaves distributionally like the limit of the sets $\theta_\star + n^{-1/2}\arg\min_u \|G_\star + \sqrt{n}(F_\star - F_{\theta_*}) - \langle u, D_{\theta_*} \rangle\|_{L_i}$ in the Hausdorff metric sense. Note that the latter sequence does not depend on the data. Since
the assumptions guarantee that $\sqrt{n}(F_n - F_\theta) \rightarrow G_\cdot$ weakly in $L_1(\mathbb{R})$, there exist versions of these variables that converge almost surely. For simplicity, we assume without loss of generality that these are the variables we work with. As noted by Pollard (1980), establishing the measurability of the sets \( \{ M_n \subset \theta_* + n^{-1/2} K(H_n, \beta_n) \} \subset \Omega \) is hard, which is why the result is stated in terms of inner probability. See also Pollard (1980, pp. 67) for further comments on the sequence $\beta_n$.

**Proof of Theorem A.2.8.** Let $\theta \in N$, where $N$ is the set from Assumption A.2.6. By Assumption A.2.4 and Pollard (1980, Lemma 4.1) or the proof of Theorem A.2.1, we know that the minimizers of $\|F_n - F_\theta\|_{L_1}$ will be attained in $N$ with probability going to one. For $\theta \in N$, we have that

\[
\|F_n - F_\theta\|_{L_1} \geq \|F_\theta - F_\\theta\|_{L_1} - \|F_n - F_\\theta\|_{L_1} \quad \text{by the triangle inequality,}
\]

\[
\geq \|F_\\theta - F_\theta\|_{L_1} + c_\|\theta - \theta_*\|_H - \|F_n - F_\\theta\|_{L_1} \quad \text{by Assumption A.2.6,}
\]

\[
\geq \|F_n - F_\\theta\|_{L_1} + c_\|\theta - \theta_*\|_H - 2\|F_n - F_\\theta\|_{L_1} \quad \text{by the triangle inequality.}
\]

Let $\xi_n = \sqrt{n}(4\|F_n - F_\\theta\|_{L_1} + 2\eta_n)/c_\|$, and $S_n = \{ \theta : \sqrt{n}\|\theta - \theta_*\|_H \leq \xi_n \}$. Then, by the assumptions on $\eta_n$ and $\sqrt{n}(F_n - F_\\theta)$, we know that $n^{-1/2}\xi_n = o_P(1)$. If $\theta \in N \cap S_n$, then by the inequality derived above, $\|F_n - F_\theta\|_{L_1} > \|F_n - F_\\theta\|_{L_1} + 2(\|F_n - F_\\theta\|_{L_1} + \eta_n)$. Thus, with inner probability going to one, it has to be that $M_n \subset S_n$.

Next, we approximate $\theta \mapsto \sqrt{n}\|F_n - F_\theta\|_{L_1}$ with the convex map $\theta \mapsto \sqrt{n}\|F_n - F_\theta - \langle \theta - \theta_*; D_\theta \rangle\|_{L_1}$ over the set $S_n$. First, note that Assumption A.2.5 implies that the remainder $R_\theta = F_\theta - F_\\theta - \langle \theta - \theta_*; D_\theta \rangle$ satisfies

\[
\|R_\theta\|_{L_1} \leq \|\theta - \theta_*\|_H \cdot \Delta(\|\theta - \theta_*\|_H),
\]

where $\Delta$ is an increasing function such that $\Delta(t) = o(1)$ as $t \rightarrow 0$. Define

\[
\Gamma_n = \sup_{\theta \in S_n} \sqrt{n}\|F_n - F_\theta\|_{L_1} - \sqrt{n}\|F_n - F_\\theta - \langle \theta - \theta_*; D_\theta \rangle\|_{L_1}.
\]

165
We then have that

\[
\Gamma_n = \sup_{\theta \in S_n} \left| \sqrt{n} \| F_n - F_{\theta_*} - \langle \theta - \theta_*, D_{\theta_*} \rangle - R_0 \|_{L_i} - \sqrt{n} \| F_n - F_{\theta_*} - \langle \theta - \theta_*, D_{\theta_*} \rangle \|_{L_i} \right|
\]

\[
\leq \sup_{\theta \in S_n} \sqrt{n} \| R_0 \|_{L_i} \text{ by the triangle inequality,}
\]

\[
\leq \sup_{\theta \in S_n} \sqrt{n} \| \theta - \theta_* \|_{\mathcal{H}} \cdot \Delta(\| \theta - \theta_* \|_{\mathcal{H}}) \text{ by Assumption A.2.5,}
\]

\[
\leq \xi_n \Delta \left( \frac{\xi_n}{\sqrt{n}} \right) = o_P(1) \text{ by the definitions of } S_n \text{ and } \xi_n.
\]

Hence, we have uniform control over the difference between \( \theta \mapsto \sqrt{n} \| F_n - F_{\theta} \|_{L_i} \) and its convex approximation over \( S_n \). Moreover, the map \( \theta \mapsto \sqrt{n} \| F_n - F_{\theta_*} - \langle \theta - \theta_*, D_{\theta_*} \rangle \|_{L_i} \) also attains its minimum on \( S_n \) with probability going to one, since for \( \theta \in N \) such that \( \Delta(\| \theta - \theta_* \|_{\mathcal{H}}) \leq c_*/2 \),

\[
\| F_n - F_{\theta_*} - \langle \theta - \theta_*, D_{\theta_*} \rangle \|_{L_i} = \| F_n - F_{\theta} + R_0 \|_{L_i}
\]

\[
\geq \| F_n - F_{\theta} \|_{L_i} - \| R_0 \|_{L_i} \text{ by the triangle inequality,}
\]

\[
\geq \| F_n - F_{\theta_*} \|_{L_i} + c_* \| \theta - \theta_* \|_{\mathcal{H}} - \frac{1}{2} \| F_n - F_* \|_{L_i}
\]

\[
- \| \theta - \theta_* \|_{\mathcal{H}} \cdot \Delta(\| \theta - \theta_* \|_{\mathcal{H}}) \text{ by Ass. A.2.6, tri. ineq.,}
\]

\[
\geq \| F_n - F_{\theta_*} \|_{L_i} + \frac{1}{2} c_* \| \theta - \theta_* \|_{\mathcal{H}} - \frac{1}{2} \| F_n - F_* \|_{L_i} .
\]

Hence, if \( \theta \in N \cap S_n \) and \( \Delta(\| \theta - \theta_* \|_{\mathcal{H}}) \leq c_*/2 \), then

\[
\| F_n - F_{\theta} \|_{L_i} > \| F_n - F_{\theta_*} \|_{L_i} + \eta_n = \| F_n - F_{\theta_*} - \langle \omega, D_{\theta_*} \rangle \|_{L_i} + \eta_n .
\]

In other words, \( m(H_n) = \inf_{u \in L_n} f(H_n, u) \) with probability going to one, where we have used the reparameterization \( L_n = \{ u : u = \sqrt{n}(\theta - \theta_*), \theta \in S_n \} \), or equivalently \( S_n = \theta_* + n^{-1/2} L_n \).

Now, since \( \Gamma_n = o_P(1) \), we can find a sequence of positive real numbers \( \gamma_n \to 0 \) such that \( P(\Gamma_n \leq \gamma_n) \to 1 \). Similarly, we can find \( \delta_n > 0 \) and \( a_n > 0 \) such that \( P(\eta_n \leq \delta_n) \to 1 \) and \( P(\| H_n - H_* \|_{L_i} \leq a_n) \to 1 \). Define \( \beta_n = \max \{ 2 \gamma_n + \delta_n, 2a_n \} \). Let \( \tau \) be such that \( \tau \in L_n \) and \( \theta_* + n^{-1/2} \tau \in M_n \) and suppose that \( \Gamma_n \leq \gamma_n \) and \( \eta_n \leq \delta_n \). By combining the
approximations developed above, we have that

$$m(H_n) \geq \inf_{u \in L_n} \sqrt{n} \| F_n - F_{\theta_n + n^{-1/3} u} \|_{L_2} - \gamma_n$$

$$\geq \sqrt{n} \| F_n - F_{\theta_n + n^{-1/3} \tau} \|_{L_2} - \gamma_n - \delta_n$$

$$\geq f(H_n, \tau) - 2\gamma_n - \delta_n.$$

Since $2\gamma_n + \delta_n \leq \beta_n$, we have that $\tau \in K(H_n, \beta_n)$. This proves the first part of the theorem, as the events considered above all hold with (inner) probability going to one.

By the triangle inequality, $u \in K(H_n^*, 0)$ implies that $u \in K(H_n, 2\| H_n - H_n^* \|_{L_2})$. Hence, with probability going to one, $K(H_n^*, 0) \subset K(H_n, \beta_n)$. Similarly, $u \in K(H_n, \beta_n)$ implies that $u \in K(H_n^*, \beta_n + 2\| H_n - H_n^* \|_{L_2})$. Recall that $\beta_n + 2\| H_n - H_n^* \|_{L_2} \to 0$ almost surely. Let $E \subset \Omega$ denote the set on which this occurs. Then, for every every $\delta > 0$, there exists $n(\omega)$ such that for $n \geq n(\omega)$, $K(H_n(\omega), \beta_n) \subset K(H_n^*(\omega), \sigma^\delta)$. By the definition of the Hausdorff metric, these set inclusions imply that

$$d_H(K(H_n^*, 0), K(H_n, \beta_n)) = o(1).$$

\[ \square \]

**Differentiability condition**

The condition in Assumption A.2.5 can sometimes be established from more familiar concepts of differentiability, such as differentiability in quadratic mean (Le Cam, 1970). The following proposition gives such a result. Suppose that the model family is absolutely continuous with respect to the Lebesgue measure $\lambda$ on $\mathbb{R}$, and denote the density $d\mu_{\theta}/d\lambda$ of $\mu_{\theta}$ by $f_{\theta}$. Let $\xi_\theta(y) = \sqrt{f_{\theta}(y)}$ for all $y \in \mathbb{R}$. Le Cam (1970) introduced the concept of differentiability in quadratic mean, which we define below.

**Definition A.2.4.** The model $\mathcal{M}$ is differentiable in quadratic mean at $\theta_*$ if there exists $\tilde{\xi}_{\theta_*} \in (L_2(\mathbb{R}))^{d_\theta}$ and $R_{\theta-\theta_*} \in (L_2(\mathbb{R}))^{d_\theta}$ such that $\tilde{\xi}_{\theta} = \tilde{\xi}_{\theta_*} + (\theta - \theta_*, \tilde{\xi}_{\theta_*}) + R_{\theta-\theta_*}$, where $[\int_\mathbb{R} R_{\theta-\theta_*}(y)dy]^{1/2} = o(\|\theta - \theta_*\|_{\mathcal{H}})$ as $\|\theta - \theta_*\|_{\mathcal{H}} \to 0$.

Differentiability in quadratic mean holds for many classical models, such as exponential
families and many location-scale families (see e.g. Section 12.2 in Lehmann and Romano, 2005).

**Proposition A.2.2.** Suppose that the model family is supported on a set $S \subset \mathbb{R}$ of bounded Lebesgue measure, and that it is differentiable in quadratic mean at $\theta_\ast$. Let

$$D_{\theta_\ast}(t) = \int_{-\infty}^{t} 2\xi_{\theta_\ast}(y)\dot{\xi}_{\theta_\ast}(y)dy$$

for $t \in S$ and zero elsewhere. Then, as $\|\theta - \theta_\ast\|_H \to 0$,

$$\int_{\mathbb{R}}|F_\theta(t) - F_{\theta_\ast}(t) - \langle \theta - \theta_\ast, D_{\theta_\ast}(t) \rangle|dt = o(\|\theta - \theta_\ast\|_H).$$

**Proof.** Consider

$$\int_{\mathbb{R}}|F_\theta(t) - F_{\theta_\ast}(t) - \langle \theta - \theta_\ast, D_{\theta_\ast}(t) \rangle|dt$$

$$= \int_{\mathbb{R}}\left|\int_{-\infty}^{t} \xi_{\theta_\ast}^*(y) - \xi_{\theta_\ast}(y) - 2\xi_{\theta_\ast}(y)\langle \theta - \theta_\ast, \dot{\xi}_{\theta_\ast}(y) \rangle dy \right| dt$$

$$\leq \int_{\mathbb{R}}\int_{-\infty}^{t} |\xi_{\theta_\ast}^*(y) - \xi_{\theta_\ast}(y) - 2\xi_{\theta_\ast}(y)\langle \theta - \theta_\ast, \dot{\xi}_{\theta_\ast}(y) \rangle| dy dt$$

$$\leq c \int_{\mathbb{R}}|\theta - \theta_\ast, \dot{\xi}_{\theta_\ast}(y)|^2 + R_{\theta - \theta_\ast}(y) + 2|\xi_{\theta_\ast}(y)R_{\theta - \theta_\ast}(y)| + 2|\theta - \theta_\ast, \dot{\xi}_{\theta_\ast}(y)R_{\theta - \theta_\ast}(y)| dy$$

$$= o(\|\theta - \theta_\ast\|_H),$$

where $c$ is some constant and the last equality follows by applying the Cauchy-Schwarz inequality to the two last terms of the integrand.

**A.3 Proofs: MEWE**

**A.3.1 Existence, measurability, and consistency**

In order to show similar results for the MEWE, we introduce the following assumptions.

**Assumption A.3.1.** For any $m \geq 1$, if $\rho_H(\theta_n, \theta) \to o$, then $\mu^{(m)}_{\hat{\theta}_n} \Rightarrow \mu^{(m)}_{\hat{\theta}}$ as $n \to \infty$. 

168
Assumption A.3.2. If \( \rho_{\mathcal{H}}(\theta_n; \theta) \to 0 \), then \( \mathbb{E}_n \mathcal{W}_p(\mu_{\theta_n}; \hat{\mu}_{\theta,n}) \to 0 \) as \( n \to \infty \).

Assumption A.3.1 is a slightly stronger version of Assumption A.2.2, stating that we not only need weak convergence of the “model” distributions \( \mu_{\theta_n} \) but also of the sample distributions \( \mu^{(m)}_{\theta} \) for any \( m \geq 1 \). Assumption A.3.2 is implied by \( \sup_{\theta \in \mathcal{H}} \mathbb{E}_n \mathcal{W}_p(\mu_{\theta}; \hat{\mu}_{\theta,n}) \to 0 \), which in turn might hold when \( \mathcal{H} \) is compact and the inequalities in Fournier and Guillin (2015) hold. In the next result, we prove an analogous version of Theorem A.2.1 for the MEWE as \( \min \{ n, m \} \to \infty \). For simplicity, we write \( m \) as a function of \( n \) and require that \( m(n) \to \infty \) as \( n \to \infty \).

Theorem A.3.1. Under Assumptions A.2.1-A.2.3 and A.3.1-A.3.2, there exists a set \( E \subseteq \Omega \) with \( \mathbb{P}(E) = 1 \) such that, for all \( \omega \in E \),

\[
\inf_{\theta \in \mathcal{H}} \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\ast; \mu_\theta),
\]

and there exists \( n(\omega) \) such that, for all \( n \geq n(\omega) \), the sets argmin \( \theta \in \mathcal{H} \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta,m(n)}) \) are non-empty and form a bounded sequence with

\[
\lim_{n \to \infty} \sup_{\theta \in \mathcal{H}} \arg\min \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta,m(n)}) \subset \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\ast; \mu_\theta).
\]

Proof of Theorem A.3.1. For any \( \nu \in \mathcal{P}(\mathcal{Y}) \), lower semicontinuity of the map \( \theta \mapsto \mathcal{W}_p(\nu; \mu_\theta) \) follows from Lemma A.1.1, via Assumption A.2.2. Since \( \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\ast; \mu_\theta) = \varepsilon_\ast, B_\ast(\varepsilon) \) with the \( \varepsilon \) of Assumption A.2.3 is non-empty, by definition of the infimum. Moreover, since \( \theta \mapsto \mathcal{W}_p(\mu_\ast; \mu_\theta) \) is lower semicontinuous, the set \( B_\ast(\varepsilon) \) is closed. By Assumption A.2.3, \( B_\ast(\varepsilon) \) is therefore compact. In other words, again by lower semicontinuity, the set argmin \( \theta \in \mathcal{H} \mathcal{W}_p(\mu_\ast; \mu_\theta) \) is non-empty.

We show that \( \theta \mapsto \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,m(n)}) \) epi-converges to \( \theta \mapsto \mathcal{W}_p(\mu_\ast; \mu_\theta) \) \( \mathbb{P} \)-almost surely. Let \( E \) denote the set of probability one from Assumption A.2.1 and let \( \omega \in E \). Fix \( \mathcal{K} \subseteq \mathcal{H} \) compact. By lower semicontinuity of \( \theta \mapsto \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta,m(n)}) \), ensured by Lemma A.1.2 and Assumption A.3.1, we know that

\[
\inf_{\theta \in \mathcal{K}} \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta,m(n)}) = \mathbb{E}_m(n) \mathcal{W}_p(\hat{\mu}_n(\omega); \hat{\mu}_{\theta_n,m(n)}).
\]
for some sequence \( \theta_n = \theta_n(\omega) \in \mathcal{K} \). Hence,

\[
\liminf_{n \to \infty} \inf_{\theta \in \mathcal{K}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}) \\
= \liminf_{n \to \infty} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}) \\
= \lim_{k \to \infty} \mathbb{E}_{m(m_k)} \mathcal{W}_p(\hat{\mu}_{m_k}(\omega), \hat{\mu}_{m_k}(m_k)) \quad \exists \text{ subsequence converging to the lim inf},
\]

\[
= \lim_{\ell \to \infty} \mathbb{E}_{m(m_k)} \mathcal{W}_p(\hat{\mu}_{m_k}(\omega), \hat{\mu}_{m_k}(m_k)) \quad \exists \text{ subseq. } \theta_{m_k} \to \bar{\theta} \in \mathcal{K} \text{ by compactness},
\]

\[
= \liminf_{\ell \to \infty} \mathbb{E}_{m(m_k)} \mathcal{W}_p(\hat{\mu}_{m_k}(\omega), \hat{\mu}_{m_k}(m_k)) \\
\geq \liminf_{\ell \to \infty} \mathcal{W}_p(\hat{\mu}_{m_k}(\omega), \mu_{\theta_{m_k}}) - \mathbb{E}_{m(m_k)} \mathcal{W}_p(\mu_{\theta_{m_k}}, \hat{\mu}_{m_k}(m_k)) \quad \text{by the triangle ineq.},
\]

\[
\geq \liminf_{\ell \to \infty} \mathcal{W}_p(\hat{\mu}_{m_k}(\omega), \mu_{\theta_{m_k}}) - \limsup_{\ell \to \infty} \mathbb{E}_{m(m_k)} \mathcal{W}_p(\mu_{\theta_{m_k}}, \hat{\mu}_{m_k}(m_k)) \\
\geq \mathcal{W}_p(\mu_*, \mu_\theta) \quad \text{by l.s.c., Assumptions A.2.1, A.2.2, A.3.2, and } \omega \in E,
\]

\[
\geq \inf_{\theta \in \mathcal{K}} \mathcal{W}_p(\mu_*, \mu_\theta).
\]

Fix \( \mathcal{O} \subset \mathcal{H} \) open. By definition of the infimum, there exists a sequence \( \theta_n \in \mathcal{O} \) such that \( \mathcal{W}_p(\mu_*, \mu_{\theta_n}) \to \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\mu_*, \mu_\theta) \). Now,

\[
\inf_{\theta \in \mathcal{O}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}) \leq \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}).
\]

Hence,

\[
\limsup_{n \to \infty} \inf_{\theta \in \mathcal{O}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}) \\
\leq \limsup_{n \to \infty} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{m(n)}) \\
\leq \limsup_{n \to \infty} \left[ \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_*) + \mathcal{W}_p(\mu_*, \mu_\theta) + \mathbb{E}_{m(n)} \mathcal{W}_p(\mu_\theta, \hat{\mu}_{m(n)}) \right] \quad \text{by tri. ineq.},
\]

\[
= \limsup_{n \to \infty} \mathcal{W}_p(\mu_*, \mu_\theta) \quad \text{by Assumptions A.2.1 and A.3.2, } \omega \in E,
\]

\[
= \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\mu_*, \mu_\theta) \quad \text{by definition of } \theta_n.
\]

170
Theorem 7.29b) of Rockafellar and Wets (2009) implies that

$$\limsup_{n \to \infty} \left( \inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \right) \leq \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta) = \varepsilon_* .$$

Hence, for all \( \alpha > 0 \), there exists \( n_\alpha(\omega) \), such that \( n \geq n_\alpha(\omega) \) implies that

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \leq \varepsilon_* + \alpha .$$

Let \( \alpha \in (0, \varepsilon/3) \). The set \( \{ \theta \in \mathcal{H} : \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \leq \varepsilon_* + \varepsilon/3 \} \) is non-empty for \( n \geq n_\alpha(\omega) \), by definition of the infimum. Let \( \theta \) belong to this set. Then, by the triangle inequality,

$$\mathcal{W}_p(\mu_*, \mu_\theta) \leq \mathcal{W}_p(\tilde{\mu}_n(\omega), \mu_*) + \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) + \mathbb{E}_{m(n)} \mathcal{W}_p(\mu_\theta, \tilde{\mu}_{\theta,m(n)}) .$$

By Assumption A.2.1, there exists an \( n_\varepsilon(\omega) \) such that for \( n \geq n_\varepsilon(\omega) \), \( \mathcal{W}_p(\tilde{\mu}_n(\omega), \mu_*) \leq \varepsilon/3 \). By Assumption A.3.2, there exists an \( \tilde{n}(\omega) \) such that for any \( n \geq \tilde{n}(\omega) \), we have \( \mathbb{E}_{m(n)} \mathcal{W}_p(\mu_\theta, \tilde{\mu}_{\theta,m(n)}) \leq \varepsilon/3 \). So, if \( n \geq \max\{ n_\alpha(\omega), n_\varepsilon(\omega), \tilde{n}(\omega) \} \), we have \( \mathcal{W}_p(\mu_*, \mu_\theta) \leq \varepsilon_* + \varepsilon \). This means that

$$\{ \theta \in \mathcal{H} : \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \leq \varepsilon_* + \varepsilon/3 \} \subset B_\varepsilon(\varepsilon) .$$

As a consequence, for \( n \geq \max\{ n_\alpha(\omega), n_\varepsilon(\omega), \tilde{n}(\omega) \} \),

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) = \inf_{\theta \in B_\varepsilon(\varepsilon)} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) .$$

By Theorem 7.31a) of Rockafellar and Wets (2009), we have

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \rightarrow \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta) .$$

Also, for \( n \geq \max\{ n_\alpha(\omega), n_\varepsilon(\omega), \tilde{n}(\omega) \} \) and by the same reasoning as for the map \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \), the sets \( \text{argmin}_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\tilde{\mu}_n(\omega), \tilde{\mu}_{\theta,m(n)}) \) are non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. The same argument holds for the sets
\( \varepsilon_n \rightarrow \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\theta}_m(\omega)) \) with \( \varepsilon_n \rightarrow 0 \), since \( \inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\theta}_m(\omega)) \) + \( \varepsilon_n \leq \varepsilon_\ast + a \) eventually.

**A.3.2 Convergence to the MWE**

The next result considers the case where the data and \( n \) is fixed, while \( m \rightarrow \infty \). It shows that the MEWE converges to the MWE, assuming the latter exists. We summarize this condition in the following assumption, in which the observed empirical distribution is kept fixed and \( \varepsilon_n = \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \).

**Assumption A.3.3.** For some \( \varepsilon > 0 \), the set \( B_n(\varepsilon) = \{ \theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \leq \varepsilon_n + \varepsilon \} \) is bounded.

**Theorem A.3.2.** Under Assumptions A.2.2 and A.3.1-A.3.3,

\[
\inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\theta}_m) \rightarrow \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta),
\]

and there exists an \( \hat{m} \) such that, for all \( m \geq \hat{m} \), the sets \( \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\theta}_m) \) are non-empty and form a bounded sequence with

\[
\limsup_{m \rightarrow \infty} \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\theta}_m) \subset \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta).
\]

**Proof of Theorem A.3.2.** Lower semicontinuity of the map \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) follows from Lemma A.1.1, via Assumption A.2.2. Since \( \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) = \varepsilon_n, B_n(\varepsilon) \) with the \( \varepsilon \) of Assumption A.2.3 is non-empty, by definition of the infimum. Moreover, since \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) is lower semicontinuous, the set \( B_n(\varepsilon) \) is closed. By Assumption A.3.3, \( B_n(\varepsilon) \) is therefore compact. In other words, by lower semicontinuity, the set \( \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) is non-empty.

We show that \( \theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\theta}_m) \) epi-converges to \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \) as \( m \rightarrow \infty \). Fix \( \mathcal{K} \subset \mathcal{H} \) compact. By lower semicontinuity of \( \theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\theta}_m) \), ensured by Lemma

172
A.1.2 and Assumption A.3.1, we know that
\[
\inf_{\theta \in \mathcal{K}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m}) ,
\]
for some sequence $\theta_m \in \mathcal{K}$. Hence,
\[
\liminf_{m \to \infty} \inf_{\theta \in \mathcal{K}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \\
= \liminf_{m \to \infty} \mathbb{E}_m \mathcal{W}_p(\mu_n, \hat{\mu}_{\theta,m}) \\
= \lim_{k \to \infty} \mathbb{E}_{m_k} \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_{m_k}}) \quad \exists \text{ subsequence converging to the lim inf},
\]
\[
= \lim_{\ell \to \infty} \mathbb{E}_{m_{k\ell}} \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_{m_{k\ell}}}) \\
\geq \lim_{\ell \to \infty} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta_{m_{k\ell}}}) - \mathbb{E}_{m_{k\ell}} \mathcal{W}_p(\mu_{\theta_{m_{k\ell}}}, \hat{\mu}_{\theta_{m_{k\ell}}}) \quad \text{by the triangle ineq.},
\]
\[
\geq \liminf_{\ell \to \infty} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) \quad \text{by l.s.c., Assumptions A.2.2 and A.3.2},
\]
\[
\geq \inf_{\theta \in \mathcal{K}} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) .
\]

Fix $\mathcal{O} \subset \mathcal{H}$ open. By definition of the inf, there exists a sequence $\theta_m \in \mathcal{O}$ such that $\mathcal{W}_p(\hat{\mu}_n, \mu_{\theta_m}) \to \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta})$. Now, $\inf_{\theta \in \mathcal{O}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$. Hence,
\[
\limsup_{m \to \infty} \inf_{\theta \in \mathcal{O}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \limsup_{m \to \infty} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \\
\leq \limsup_{m \to \infty} \left[ \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta_m}) + \mathbb{E}_m \mathcal{W}_p(\mu_{\theta_m}, \hat{\mu}_{\theta,m}) \right] \quad \text{by the triangle inequality},
\]
\[
= \limsup_{m \to \infty} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta_m}) \quad \text{by Assumption A.3.2},
\]
\[
= \inf_{\theta \in \mathcal{O}} \mathcal{W}_p(\mu_n, \mu_{\theta}) \quad \text{by definition of } \theta_m .
\]
Theorem 7.29b) of Rockafellar and Wets (2009) implies that
\[
\limsup_{m \to \infty} \left( \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \right) \leq \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) = \varepsilon_n.
\]
Hence, for all \( a > 0 \), there exists \( m_a \), such that for \( m \geq m_a \), \( \inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \varepsilon_n + a \). Let \( a \in (0, \varepsilon/2) \). The set
\[
\{ \theta \in \mathcal{H} : \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \varepsilon_n + \varepsilon/2 \}
\]
is non-empty for \( m \geq m_a \), by definition of the infimum. Let \( \theta \) belong to this set. Then, by the triangle inequality,
\[
\mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) \leq \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) + \mathbb{E}_m \mathcal{W}_p(\mu_{\theta}, \hat{\mu}_{\theta,m}).
\]
By Assumption A.3.2, there exists an \( \hat{m} \) such that for \( m \geq \hat{m} \), \( \mathbb{E}_m \mathcal{W}_p(\mu_{\theta}, \hat{\mu}_{\theta,m}) \leq \varepsilon/2 \). So, if \( m \geq \max\{m_a, \hat{m}\} \), we have that \( \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) \leq \varepsilon_n + \varepsilon \). This means that
\[
\{ \theta \in \mathcal{H} : \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \varepsilon_n + \varepsilon/2 \} \subset B_n(\varepsilon).
\]
Hence, for \( m \geq \max\{m_a, \hat{m}\} \), \( \inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = \inf_{\theta \in B_n(\varepsilon)} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \).

By Theorem 7.31a) of Rockafellar and Wets (2009), we know that
\[
\inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta})
\]
as \( m \to \infty \). Also, for any \( m \geq \max\{m_a, \hat{m}\} \), and by the same reasoning as for the map \( \theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_{\theta}) \), the set \( \arg\min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \) is non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. \( \square \)

**Theorem A.3.3** (Measurability of the MEWE). Suppose that \( \mathcal{H} \) is a \( \sigma \)-compact Borel measurable subset of \( \mathbb{R}^{d_E} \). Under Assumption A.3.1, for any \( n \geq 1 \) and \( m \geq 1 \) and \( \varepsilon > 0 \), there
exists a Borel measurable function \( \hat{\theta}_{n,m} : \Omega \to \mathcal{H} \) that satisfies

\[
\hat{\theta}_{n,m}(\omega) \in \begin{cases} 
\arg\min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m}), & \text{if this set is non-empty}, \\
\varepsilon\text{-}\arg\min_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m}), & \text{otherwise}.
\end{cases}
\]

Proof. The proof is identical to that of Theorem A.2.2, applying Lemma A.1.2 instead of A.1.1.

\[\square\]

A.4 CHECKING THE ASSUMPTIONS

The following proposition gives three data-generating mechanisms for which \( \mathcal{W}_p(\hat{\mu}_n, \mu_\star) \rightarrow 0 \) \( \mathbb{P} \)-almost surely, which is Assumption A.2.1. The three conditions below are mainly chosen for illustrative purposes, and are by no means exhaustive. We first give definitions that are used in the conditions. We denote by \( \mathcal{F} \) the measurable sets of \( \Omega \).

Definition A.4.1. The stochastic process \( Y = (Y_t)_{t \in \mathbb{Z}} \) is stationary if for any \( k \in \mathbb{N} \) and \( \tau, t_1, \ldots, t_k \in \mathbb{Z} \) we have that \( (Y_{t_1}, \ldots, Y_{t_k}) \) and \( (Y_{t_1+\tau}, \ldots, Y_{t_k+\tau}) \) have the same distribution.

Definition A.4.2. The map \( T : \Omega \to \Omega \) is \( \mathbb{P} \)-measure preserving if \( \mathbb{P}(T^{-1}(A)) = \mathbb{P}(A) \) for all \( A \in \mathcal{F} \).

Definition A.4.3. The map \( T : \Omega \to \Omega \) is \( \mathbb{P} \)-ergodic if it is \( \mathbb{P} \)-measure preserving, and such that for all \( A \in \mathcal{F} \) with \( T^{-1}(A) = A \) we have that \( \mathbb{P}(A) = 0 \) or \( \mathbb{P}(A) = 1 \). The stochastic process \( Y = (Y_t)_{t \in \mathbb{Z}} \) is ergodic if it can be represented by \( Y_t = Y_0 \circ T^t \) for some ergodic \( T \) and some random variable \( Y_0 \).

Definition A.4.4. The stochastic process \( Y = (Y_t)_{t \in \mathbb{Z}} \) is \( a \)-mixing with mixing coefficients

\[
a_t = \sup_{k \in \mathbb{Z}} \sup_{A \in \mathcal{F}_k, B \in \mathcal{F}_k^{\infty}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|,
\]

if \( a_t \to 0 \) as \( t \to \infty \), where \( \mathcal{F}_k = \sigma(Y_i : i \leq k) \) and \( \mathcal{F}_k^{\infty} = \sigma(Y_i : i \geq k) \).

Proposition A.4.1. Suppose that \( Y = (Y_t)_{t \in \mathbb{Z}} \) is a stochastic process such that either
1. \( Y \sim \mu_\infty \) for some \( \mu_\infty \in \mathcal{P}_P(\mathcal{Y}) \), i.e. the observations are i.i.d., or

2. \( (Y_t)_{t \in \mathbb{Z}} \) is ergodic and stationary, represented by \( Y_t = Y_o \circ T^t \), where \( Y_o \sim \mu_\infty \in \mathcal{P}_P(\mathcal{Y}) \) and \( T \) is an ergodic, measure preserving map, or

3. \( (Y_t)_{t \in \mathbb{Z}} \) is a−mixing with mixing coefficients \( a_t \) such that \( \sum_{t=1}^{\infty} a_t^{-1/3} < \infty \), with \( Y_t \sim \mu_t \) such that \( \mu_t \) converges weakly to \( \mu_\infty \in \mathcal{P}_P(\mathcal{Y}) \) and satisfies \( \sup_t \mathbb{E}|Y_t|^3 < \infty \) for some \( 1 \leq \max(r, p) < q < 2r \) (where it is assumed \( \rho(x, y) = \|x - y\|_Y \) for simplicity).

Then there exists a set \( E \in \mathcal{F} \) with \( \mathbb{P}(E) = 1 \) such that, for all \( \omega \in E \), \( \mathcal{W}_P(\hat{\mu}_n(\omega), \mu_\infty) \to 0 \).

Proof. Under condition 1., Theorem 3 in Varadarajan (1958a) establishes that there exists a set \( E_i \) with \( \mathbb{P}(E_i) = 1 \) such that for all \( \omega \in E_i \), \( \hat{\mu}_n(\omega) \) converges weakly to \( \mu_\infty \). By the strong law of large numbers, there exist a set \( E_2 \) with \( \mathbb{P}(E_2) = 1 \) and an \( x_o \in \mathcal{X} \) such that \( \int_{\mathcal{X}} \rho(x, x_o)^p d\hat{\mu}_n(\omega)(x) \to \int_{\mathcal{X}} \rho(x, x_o)^p d\mu_\infty(x) \) for all \( \omega \in E_2 \). Then, in light of Theorem A.1.1, the claim holds on \( E = E_i \cap E_2 \).

Consider condition 2. By Varadarajan (1958b), there exists a fixed countable set \( C^* \) of continuous and bounded functions on \( \mathcal{Y} \), such that for any sequence of measures \( \mu_n \) on \( \mathcal{Y} \), \( \mu_n \) converges weakly to \( \mu \) if and only if \( \int f d\mu_n \to \int f d\mu \) for all \( f \in C^* \). Fix \( f \in C^* \). We know that \( f \circ Y_o \) is measurable and that \( \mathbb{E}|f \circ Y_o| < \infty \) since \( f \) is bounded, so by Birkhoff’s ergodic theorem there exists a set \( E_f \) such that \( \mathbb{P}(E_f) = 1 \) and

\[
\int_{\mathcal{Y}} f d\hat{\mu}_n(\omega) = \frac{1}{n} \sum_{t=1}^{n} f(Y_t(\omega)) = \frac{1}{n} \sum_{t=1}^{n} f \circ Y_o \circ T^t(\omega) \to \int_{\mathcal{Y}} f d\mu_\infty,
\]

for all \( \omega \in E_f \). Moreover, since \( \mu_\infty \in \mathcal{P}_P(\mathcal{Y}) \) we know \( \int_{\mathcal{Y}} \rho(y, x_o)^p d\mu_\infty(y) < \infty \) and that there exists a set \( E_o \) with \( \mathbb{P}(E_o) = 1 \) such that

\[
\int \rho(y, y_o)^p d\hat{\mu}_n(y)(\omega) \to \int \rho(y, y_o)^p d\mu_\infty(y),
\]

for all \( \omega \in E_o \). Since \( C^* \) is countable we know that \( \mathbb{P}(\cap_{f \in C^*} E_f \cap E_o) = 1 \). In other words, this means that \( \mathcal{W}_P(\hat{\mu}_n(\omega), \mu_\infty) \to 0 \) for all \( \omega \in E = \cap_{f \in C^*} E_f \cap E_o \).
Under condition 3., we first note that since \((Y_i)_{i \in \mathbb{Z}}\) is \(a\)–mixing, then so is \((f \circ Y_i)_{i \in \mathbb{Z}}\) for any measurable \(f\), with mixing coefficients bounded above by \(a\), since \(\sigma(f(Y_i) : i \leq k) \subset \sigma(Y_i : i \leq k)\). Also, since \(\mu_i\) converges weakly to \(\mu_*\) in \(P(Y)\) we have that for all \(f \in C^*\),

\[
\frac{1}{n} \sum_{i=1}^{n} \int_{Y} f d\mu_i \to \int_{Y} f d\mu_*,
\]

and

\[
\frac{1}{n} \sum_{i=1}^{n} \int_{Y} \|y\|_{Y}^p d\mu_i(y) \to \int_{Y} \|y\|_{Y}^p d\mu_*(y).
\]

By Hansen (1991) Corollary 4, we know that for all \(f \in C^*\) we have that the zero-mean, \(a\)-mixing sequence \(f(Y_i) - \int_{Y} f d\mu_i\) satisfies

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ f(Y_i) - \int_{Y} f d\mu_i \right\} \to 0 \quad \mathbb{P}\text{-almost surely.}
\]

Similarly,

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \|Y_i\|_{Y}^p - \int_{Y} \|y\|_{Y}^p d\mu_i(y) \right\} \to 0 \quad \mathbb{P}\text{-almost surely.}
\]

Together this gives us that

\[
\int_{Y} f d\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \to \int_{Y} f d\mu_* \quad \mathbb{P}\text{-almost surely.}
\]

and

\[
\int_{Y} \|y\|_{Y}^p d\hat{\mu}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \|Y_i\|_{Y}^p \to \int_{Y} \|y\|_{Y}^p d\mu_*(y) \quad \mathbb{P}\text{-almost surely.}
\]

Then, again by the countability of \(C^*\), we can conclude that \(W_p(\hat{\mu}_n(\omega), \mu_*) \to 0\) for all \(\omega\) in a set \(E\) defined analogously to the one for the second set of conditions.

The following proposition can be used to verify Assumption A.2.4.

**Proposition A.4.2.** Suppose that either of the conditions of Lemma A.1.1 holds. Suppose that there exists a proper, connected and compact subset \(S \subset \mathcal{H}\) with positive Lebesgue measure such
that \( \inf_{\theta \in \mathcal{H} \setminus S} \mathcal{W}_p(\mu_*, \mu_\theta) > \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_*, \mu_\theta) \). Then there exists a \( \theta_* \) attaining the infimum of \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \). If \( \theta_* \) is unique, then it is well-separated.

**Proof.** Since \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \) is continuous/lower semicontinuous, it attains a minimum \( \theta_* \) on \( S \). This is also the global minimum by the assumption on \( S \). If \( \theta_* \) is unique, it is well-separated in the sense of Assumption A.2.4, for all \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\inf_{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon} \mathcal{W}_p(\mu_*, \mu_\theta) > \mathcal{W}_p(\mu_*, \mu_{\theta_*}) + \delta.
\]

Indeed, let \( \varepsilon > 0 \), and consider \( \{ \theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \). Either the set is contained in \( \mathcal{H} \setminus S \), and thus well-separation follows, or, \( \{ \theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \cap S \) is not empty. Then we show that it is compact. Since \( S \) is compact, there exists \( \varepsilon_* \geq \varepsilon \) such that \( S \subset \{ \theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \leq \varepsilon_* \} \). Therefore,

\[
\{ \theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \cap S = \{ \theta \in \mathcal{H} : \varepsilon_* \geq \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \cap S.
\]

Now \( \{ \theta \in \mathcal{H} : \varepsilon_* \geq \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \) is compact. An intersection of compact sets is compact. Therefore, \( \theta \mapsto \mathcal{W}_p(\mu_*, \mu_\theta) \) being continuous/lower semicontinuous, an infimum is attained on \( \{ \theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \geq \varepsilon \} \cap S \), and by uniqueness of \( \theta_* \), well-separation follows. \( \Box \)
Appendix to Chapter 3

B.1 Proofs

For ease of presentation, we restate the theoretical results of Chapter 3 and their assumptions before giving the proofs.

Proposition B.1.1. For any integer $n \geq 1$ and real number $p \geq 1$, $\mathcal{H}_p$ defines a distance on the space of empirical distributions of size $n$.

Proof. Let $x_{i:n}$, $y_{i:n}$ and $z_{i:n}$ be three vectors in $\mathcal{Y}^n$ and denote by $\hat{\mu}_n^x$, $\hat{\mu}_n^y$ and $\hat{\mu}_n^z$ the corresponding empirical distributions of size $n$. Since $\rho$ is a metric on $\mathcal{Y}$,

$$\mathcal{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) \geq 0, \quad \mathcal{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) = \mathcal{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z)$$

and $\mathcal{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) = 0$ if and only if $\hat{\mu}_n^x = \hat{\mu}_n^z$. To conclude the proof it therefore remains to
show that

$$\mathcal{S}_p(\mu_n^x, \hat{\mu}_n^x) \leq \mathcal{S}_p(\mu_n^y, \hat{\mu}_n^y) + \mathcal{S}_p(\mu_n^z, \hat{\mu}_n^z).$$

To this end, we define

$$\rho_{xy} = (\rho(x_{\sigma_x(i)}; y_{\sigma_y(i)}), \ldots, \rho(x_{\sigma_x(n)}; y_{\sigma_y(n)})),$$

$$\rho_{xz} = (\rho(x_{\sigma_x(i)}; z_{\sigma_z(i)}), \ldots, \rho(x_{\sigma_x(n)}; z_{\sigma_z(n)})),$$

$$\rho_{yz} = (\rho(y_{\sigma_y(i)}; z_{\sigma_z(i)}), \ldots, \rho(y_{\sigma_y(n)}; z_{\sigma_z(n)})),$$

and denote by $\| \cdot \|_p$ the $L_p$-norm on $\mathbb{R}^n$. Then,

$$\mathcal{S}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) = n^{-1/p} \| \rho_{xz} \|_p$$

$$\leq n^{-1/p} \| \rho_{xy} \|_p + n^{-1/p} \| \rho_{xz} - \rho_{xy} \|_p$$

$$= \mathcal{S}_p(\hat{\mu}_n^x, \hat{\mu}_n^y) + \mathcal{S}_p(\hat{\mu}_n^z, \hat{\mu}_n^z) + \mathcal{S}_p(\hat{\mu}_n^y, \hat{\mu}_n^y)$$

$$\leq \mathcal{S}_p(\hat{\mu}_n^y, \hat{\mu}_n^y) + n^{-1/p} \left( \sum_{i=1}^n \rho(x_{\sigma_x(i)}; z_{\sigma_z(i)}) - \rho(x_{\sigma_x(i)}; y_{\sigma_y(i)}) \right)^{1/p}$$

$$\leq \mathcal{S}_p(\hat{\mu}_n^y, \hat{\mu}_n^y) + n^{-1/p} \left( \sum_{i=1}^n \rho(y_{\sigma_y(i)}; z_{\sigma_z(i)}) \right)^{1/p}$$

$$\leq \mathcal{S}_p(\hat{\mu}_n^z, \hat{\mu}_n^z) + \mathcal{S}_p(\hat{\mu}_n^y, \hat{\mu}_n^y),$$

where the first inequality uses the triangle inequality, and the last uses the reverse triangle inequality.

\[\square\]

**Proposition B.1.2.** Suppose that $\mu^{(n)}_\theta$ has a continuous density $f^{(n)}_\theta$ and that

$$\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_H} f^{(n)}_\theta(y_{1:n}) < \infty,$$

where $\mathcal{N}_H$ is a set such that $\pi(\mathcal{N}_H) = 0$. Suppose that there exists $\varepsilon > 0$ such that

$$\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_H} \sup_{z_{1:n} \in A^\varepsilon} f^{(n)}_\theta(z_{1:n}) < \infty,$$

where $A^\varepsilon = \{z_{1:n} : D(y_{1:n}, z_{1:n}) \leq \varepsilon\}$. Suppose also that $D$ is continuous in the sense that
\( \mathcal{D}(y_{1:n}, z_{1:n}) \to \mathcal{D}(y_{1:n}, x_{1:n}) \) whenever \( z_{1:n} \to x_{1:n} \) component-wise in the metric \( \rho \). If either

1. \( f_{\theta}^{(n)} \) is \( n \)-exchangeable, such that \( f_{\theta}^{(n)}(y_{1:n}) = f_{\theta}^{(n)}(y_{\sigma(1:n)}) \) for any \( \sigma \in S_n \), and

   \[ \mathcal{D}(y_{1:n}, z_{1:n}) = 0 \text{ if and only if } z_{1:n} = y_{\sigma(1:n)} \text{ for some } \sigma \in S_n, \]

2. \( \mathcal{D}(y_{1:n}, z_{1:n}) = 0 \text{ if and only if } z_{1:n} = y_{1:n} \)

then, keeping \( y_{1:n} \) fixed, the ABC posterior converges strongly to the posterior as \( \varepsilon \to 0 \).

**Proof.** We follow a similar approach to that in Proposition 1 of Rubio and Johansen (2013). Fix \( y_{1:n} \) and let \( \varepsilon \) be as in the statement of our proposition. For any \( 0 < \varepsilon < \bar{\varepsilon} \), let \( q^\varepsilon(\theta) \) denote the normalized quasi-likelihood induced by the ABC procedure, i.e.

\[
q^\varepsilon(\theta) = \frac{\int_{y_{1:n}} \mathbb{1}(\mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n}}{\int_{y_{1:n}} \mathbb{1}(\mathcal{D}(y_{1:n}, z'_{1:n}) \leq \varepsilon) dz'_{1:n}} = \int_{y_{1:n}} K^\varepsilon(y_{1:n}, z_{1:n}) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n},
\]

where \( K^\varepsilon(y_{1:n}, z_{1:n}) \) denotes the uniform density on \( \mathcal{A}^\varepsilon = \{ z_{1:n} : \mathcal{D}(y_{1:n}, z_{1:n}) \leq \varepsilon \} \), evaluated at some \( z_{1:n} \). Note that the sets \( \mathcal{A}^\varepsilon \) are compact, due to the continuity of \( \mathcal{D} \).

Now, for any \( \theta \in \Theta \setminus \mathcal{N}\mathcal{H} \) we have

\[
\left| q^\varepsilon(\theta) - f_{\theta}^{(n)}(y_{1:n}) \right| = \int_{y_{1:n}} K^\varepsilon(y_{1:n}, z_{1:n}) \left| f_{\theta}^{(n)}(z_{1:n}) - f_{\theta}^{(n)}(y_{1:n}) \right| dz_{1:n}
\]

\[
\leq \sup_{z_{1:n} \in \mathcal{A}^\varepsilon} \left| f_{\theta}^{(n)}(z_{1:n}) - f_{\theta}^{(n)}(y_{1:n}) \right|
\]

for some \( z_{1:n}^\varepsilon \in \mathcal{A}^\varepsilon \), where the second inequality holds since \( \int_{y_{1:n}} K^\varepsilon(y_{1:n}, z_{1:n}) dz_{1:n} = 1 \), and the last equality holds by compactness of \( \mathcal{A}^\varepsilon \) and continuity of \( f_{\theta}^{(n)} \). Since for each \( \varepsilon > 0 \), \( z_{1:n}^\varepsilon \in \mathcal{A}^\varepsilon \), we know \( \lim_{\varepsilon \to 0} z_{1:n}^\varepsilon \in \cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^\varepsilon \). Under condition 1, \( \cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^\varepsilon = \{ y_{\sigma(1:n)} : \sigma \in S_n \} \), by continuity of \( \mathcal{D} \). Similarly, under condition 2, \( \cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^\varepsilon = \{ y_{1:n} \} \). In both cases, taking the limit \( \varepsilon \to 0 \) yields \( \left| q^\varepsilon(\theta) - f_{\theta}^{(n)}(y_{1:n}) \right| \to 0 \), due to the continuity of \( f_{\theta}^{(n)} \) (and \( n \)-exchangeability under condition 1).

Let \( \varepsilon \leq \bar{\varepsilon} \), so that

\[
\sup_{\theta \in \Theta \setminus \mathcal{N}\mathcal{H}} q^\varepsilon(\theta) = \sup_{\theta \in \Theta \setminus \mathcal{N}\mathcal{H}} \int_{y_{1:n}} K^\varepsilon(y_{1:n}, z_{1:n}) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n}
\]
\[
\leq \sup_{\theta \in \mathcal{H}} \sup_{z_{i:n} \in A^c} f^{(n)}_{\theta}(z_{i:n}) < M,
\]
for some \( \alpha < M < \infty \). By the bounded convergence theorem, for any measurable \( B \subset \mathcal{H} \) we have that \( \int_{B} \pi(d\theta) q^\varepsilon(\theta) \to \int_{B} \pi(d\theta) f^{(n)}_{\theta}(y_{i:n}) \) as \( \varepsilon \to 0 \). Hence,

\[
\lim_{\varepsilon \to 0} \int_{B} \pi^\varepsilon_{y_{i:n}}(d\theta) = \lim_{\varepsilon \to 0} \frac{\int_{B} \pi(d\theta) q^\varepsilon(\theta)}{\int_{\mathcal{H}} \pi(d\theta) q^\varepsilon(\theta)} = \frac{\int_{B} \pi(d\theta) f^{(n)}_{\theta}(y_{i:n})}{\int_{\mathcal{H}} \pi(d\theta) f^{(n)}_{\theta}(y_{i:n})} = \int_{B} \pi(d\theta|y_{i:n}) .
\]

\( \Box \)

**Assumption B.1.1.** The data-generating process is such that \( \mathcal{W}_p(\hat{\mu}_n; \mu_*) \to \alpha \) in \( \mathbb{P} \)-probability, as \( n \to \infty \).

**Assumption B.1.2.** For any \( \varepsilon > 0 \), \( f^{(n)}_{\theta}(\mathcal{W}_p(\mu_\theta; \hat{\mu}_n, \mu_*) > \varepsilon) \leq c(\theta) f_n(\varepsilon) \), where \( f_n(\varepsilon) \) is a sequence of functions that are strictly decreasing in \( \varepsilon \) for fixed \( n \) and \( f_n(\varepsilon) \to 0 \) for fixed \( \varepsilon \) as \( n \to \infty \). The function \( c : \mathcal{H} \to \mathbb{R}^+ \) is \( \pi \)-integrable, and satisfies \( c(\theta) \leq c_\alpha \) for some \( c_\alpha > \alpha \), for all \( \theta \) such that, for some \( \delta_\alpha > 0 \), \( \mathcal{W}_p(\mu_*, \mu_\theta) \leq \delta_\alpha + \varepsilon_* \).

**Assumption B.1.3.** There exist \( L > 0 \) and \( c_\varepsilon > 0 \) such that, for all \( \varepsilon \) small enough,

\[
\pi \left( \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_*, \mu_\theta) \leq \varepsilon + \varepsilon_* \} \right) \geq c_\varepsilon^L .
\]

**Proposition B.1.3.** Under Assumptions B.1.1-B.1.3, consider a sequence \( (\varepsilon_n)_{n \geq 0} \) such that, as \( n \to \infty \), \( \varepsilon_n \to 0 \), \( f_n(\varepsilon_n) \to 0 \), and \( \mathbb{P}(\mathcal{W}_p(\hat{\mu}_n, \mu_*) \leq \varepsilon_n) \to 1 \). Then, the WABC posterior with threshold \( \varepsilon_n + \varepsilon_* \) satisfies, for some \( \alpha < C < \infty \) and any \( \alpha < R < \infty \),

\[
\pi^{\varepsilon_n + \varepsilon_*}_{\mu_*, \mu_\theta} \left( \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_*, \mu_\theta) > \varepsilon_* + 4\varepsilon_n/3 + f^{-1}_n(\varepsilon_n/R) \} \right) \leq \frac{C}{R},
\]

with \( \mathbb{P} \)-probability going to 1 as \( n \to \infty \).

**Proof.** We first look at the WABC posterior probability of the sets \( \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_*, \mu_\theta) > \delta \} \). Note that, using Bayes’ formula, for all \( \varepsilon, \delta > 0 \),

\[
\pi^{\varepsilon + \varepsilon_*}_{y_{i:n}} \left( \mathcal{W}_p(\mu_*, \mu_\theta) > \delta \right) = \frac{\mathbb{P}_\theta(\mathcal{W}_p(\mu_*, \mu_\theta) > \delta, \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*)}{\mathbb{P}_\theta(\mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*)} ,
\]

182
where \( \mathbb{P}_\theta \) denotes the distribution of \( \theta \sim \pi \) and of the synthetic data \( z_{1:n} \sim \mu_\theta^{(n)} \), keeping the observed data \( y_{1:n} \) and hence \( \hat{\mu}_n \) fixed. We aim to upper bound this expression, and proceed by upper bounding the numerator and lower bounding the denominator.

By the triangle inequality,

\[
\mathcal{W}_p(\mu_*; \mu_\theta) \leq \mathcal{W}_p(\mu_*; \hat{\mu}_n) + \mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,n}) + \mathcal{W}_p(\hat{\mu}_{\theta,n}; \mu_\theta).
\]

On the events \( \{ \mathcal{W}_p(\mu_*; \mu_\theta) > \delta, \, \mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_* \} \), we have

\[
\delta < \mathcal{W}_p(\mu_*; \mu_\theta) \leq \mathcal{W}_p(\mu_*; \hat{\mu}_n) + \mathcal{W}_p(\hat{\mu}_{\theta,n}; \mu_\theta) + \varepsilon + \varepsilon_*.
\]

Let \( A(n, \varepsilon) = \{ y_{1:n} : \mathcal{W}_p(\hat{\mu}_n; \mu_*) \leq \varepsilon / 3 \} \). Assuming \( y_{1:n} \in A(n, \varepsilon) \) implies that

\[
\delta < \mathcal{W}_p(\hat{\mu}_{\theta,n}; \mu_\theta) + \frac{4\varepsilon}{3} + \varepsilon_*.
\]

Using this to bound the numerator, we get by a simple reparametrization that for any \( \xi > 0 \),

\[
\pi_{y_{1:n}}^{\varepsilon + \varepsilon_*} (\mathcal{W}_p(\mu_*; \mu_\theta) > 4\varepsilon / 3 + \varepsilon_* + \xi) \leq \frac{\mathbb{P}_\theta(\mathcal{W}_p(\hat{\mu}_{\theta,n}; \mu_\theta) > \xi)}{\mathbb{P}_\theta(\mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*)}.
\]

The remainder of the proof follows from further bounding this fraction using the assumptions we made on the convergence rate of empirical measures in the Wasserstein distance. Focusing first on the numerator, for any \( \xi > 0 \) we have by Assumption B.1.2 that

\[
\mathbb{P}_\theta(\mathcal{W}_p(\hat{\mu}_{\theta,n}; \mu_\theta) > \xi) = \int_{\mathcal{H}} \mu_\theta^{(n)} (\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \xi) \pi(d\theta) \leq \int_{\mathcal{H}} c(\theta) f_n(\xi) \pi(d\theta) \leq c f_n(\xi),
\]

for some constant \( c_1 \leq +\infty \). For the denominator,

\[
\mathbb{P}_\theta(\mathcal{W}_p(\hat{\mu}_n; \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*) = \int_{\mathcal{H}} \mu_\theta^{(n)} (\mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*) \pi(d\theta) \geq \int_{\mathcal{W}_p(\mu_*; \mu_\theta) \leq \varepsilon / 3 + \varepsilon_*} \mu_\theta^{(n)} (\mathcal{W}_p(\mu_*; \mu_{\theta,n}) \leq \varepsilon + \varepsilon_*) \pi(d\theta) \quad \text{(by non-neg. of integrand)}
\]
\[ \geq \int_{W_p(\mu_*) \leq \varepsilon_{/3 + \varepsilon_*}} \mu_\theta^{(n)} (W_p(\mu_*, \mu_\theta) + W_p(\hat{\mu}_n, \mu_*) + W_p(\mu_\theta, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_*) \pi(d\theta) \]

(by the triangle inequality)

\[ \geq \int_{W_p(\mu_*) \leq \varepsilon_{/3 + \varepsilon_*}} \mu_\theta^{(n)} (W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3}) \pi(d\theta) \]

(since \( W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*} \) and \( W_p(\hat{\mu}_n, \mu_*) \leq \varepsilon_{/3} \))

\[ = \pi(W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*}) - \int_{W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*}} \mu_\theta^{(n)} (W_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon_{/3}) \pi(d\theta) \]

\[ \geq \pi(W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*}) - \int_{W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*}} c(\theta)f_n(\varepsilon_{/3}) \pi(d\theta) \] (by Ass. B.1.2).

We now make more specific choices for \( \varepsilon \) and \( \zeta \), starting with assuming that \( \varepsilon_{/3} \leq \delta_0 \), such that \( c(\theta) \leq c_o \) for some constant \( c_o > 0 \) in the last integrand above, by Assumption B.1.2. The last line above is then greater than or equal to

\[ \pi(W_p(\mu_*, \mu_\theta) \leq \varepsilon_{/3 + \varepsilon_*} (1 - c_0f_n(\varepsilon_{/3})). \]

Replacing \( \varepsilon \) with \( \varepsilon_n \) such that \( f_n(\varepsilon_n/3) \to o \) implies that \( c_0f_n(\varepsilon_n/3) \leq 1/2 \) for sufficiently large \( n \). Hence,

\[ \pi(W_p(\mu_*, \mu_\theta) \leq \varepsilon_n/3 + \varepsilon_*) (1 - c_0f_n(\varepsilon_n/3)) \geq \frac{1}{2} \pi(W_p(\mu_*, \mu_\theta) \leq \varepsilon_n/3 + \varepsilon_*) \geq c_e \varepsilon_n^L, \]

for sufficiently large \( n \), by Assumption B.1.3. We can summarize the bounds derived above as follows,

\[ \pi_{\varepsilon_{n_{/3}} + \varepsilon_*(W_p(\mu_*, \mu_\theta) > 4\varepsilon_n/3 + \varepsilon_*)} \leq C\varepsilon_n^{-L}, \]

where \( C = c_*/c_e. \)

Set some \( R > 0 \) and note that for any \( n \geq 1 \), because the function \( f_n \) is strictly decreasing under Assumption B.1.2, \( f_{n^{-1}}(\varepsilon_{n}/R) \) is well-defined in the sense that \( f_{n^{-1}} \) is defined at \( \varepsilon_{n}/R \). Choosing \( \zeta_n = f_{n^{-1}}(\varepsilon_{n}/R) \) leads to

\[ \pi_{\varepsilon_{n_{/3}} + \varepsilon_*(W_p(\mu_*, \mu_\theta) > 4\varepsilon_n/3 + \varepsilon_*)} \leq \frac{C}{R}. \]

184
Since we assumed that $\mathbb{P}\left(\{\omega : y_{1:n}(\omega) \in A(n, \varepsilon_n)\} \to 1\right)$ as $n \to \infty$, the statement above holds with probability going to one. \hfill \Box

**Assumption B.1.4.** The parameter $\theta_* = \arg\min_{\theta \in \mathcal{H}} \mathcal{W}_*^{P}(\mu_*, \mu_\theta)$ exists, and is well-separated in the sense that, for all $\delta > \alpha$, there exists $\delta' > \alpha$ such that

$$\inf_{\{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) > \delta\}} \mathcal{W}_*^{P}(\mu_\theta, \mu_*) > \mathcal{W}_*^{P}(\mu_\theta, \mu_*) + \delta'.$$

**Assumption B.1.5.** The parameters are identifiable, and there exist $K > \alpha$, $\alpha > \alpha$ and an open neighborhood $U \subset \mathcal{H}$ of $\theta_*$, such that, for all $\theta \in U$,

$$\rho_\mathcal{H}(\theta, \theta_*) \leq K(\mathcal{W}_*^{P}(\mu_\theta, \mu_*) - \varepsilon_*)^\alpha.$$

**Corollary B.1.1.** Under Assumptions B.1.1-B.1.5, consider a sequence $(\varepsilon_n)_{n \geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to \alpha$, $f_n(\varepsilon_n) \to \alpha$, $f_n^{-1}(\varepsilon_n^3) \to 0$, and $\mathbb{P}(\mathcal{W}_*^{P}(\hat{\mu}_n, \mu_*) \leq \varepsilon_n) \to 1$. Then the WABC posterior with threshold $\varepsilon_n + \varepsilon_*$ satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,

$$\pi_{y_{1:n}}^{\varepsilon_n + \varepsilon_*}\left(\{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) > \mathcal{W}_*^{P}(\mu_*, \mu_\theta) - \varepsilon_* \leq 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^3/R)\}^3\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to one.

**Proof.** Let $\delta > \alpha$ be such that $\{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_*) \leq \delta\} \subset U$, where $U$ is the set in Assumption B.1.5. By Assumption B.1.4, there exists $\delta' > \alpha$ such that $\rho_\mathcal{H}(\theta, \theta_*) > \delta$ implies $\mathcal{W}_*^{P}(\mu_\theta, \mu_*) - \varepsilon_* > \delta'$. Let $n$ be large enough such that $4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^3/R) < \delta'$, which implies $\{\theta \in \mathcal{H} : \mathcal{W}_*^{P}(\mu_\theta, \mu_*) - \varepsilon_* \leq 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^3/R)\} \subset U$.

From Proposition B.1.3, we know that

$$\pi_{y_{1:n}}^{\varepsilon_n + \varepsilon_*}\left(\mathcal{W}_*^{P}(\mu_*, \mu_\theta) - \varepsilon_* \leq 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^3/R)\right) \geq 1 - \frac{C}{R},$$

with probability going to one. Applying the inequality in Assumption B.1.5 gives

$$\pi_{y_{1:n}}^{\varepsilon_n + \varepsilon_*}\left(\rho_\mathcal{H}(\theta, \theta_*) \leq K[4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^3/R)]^\alpha\right) \geq 1 - \frac{C}{R},$$

with probability going to one. \hfill \Box

185
B.2 Sequential Monte Carlo sampler for ABC posteriors

To approximate ABC posterior distributions, we apply an SMC algorithm which leads to an automatic choice of decreasing thresholds \((\varepsilon_t)\), is parallelizable over the particles, and can leverage any MCMC kernel within the rejuvenation step. Choices of Markov kernel and adaptation rule for the thresholds are discussed in Sections B.2.1 and B.2.2. For dependent data, the algorithm can be readily modified to include delay and residual reconstructions. We write \(D\) for a generic distance between data sets in this section.

Let \(N\) be the number of particles, and \(\varepsilon_0 = \infty\). The first step goes as follows.

1. Sample \(\theta^0, \ldots, \theta^N\) from the prior \(\pi\), set \(w^k_0 = N^{-1}\) for all \(k \in 1:N\).

2. For each \(k \in 1:N\), sample \(z^k_{1:n}\) from \(\mu_{\theta^k_0}(\cdot)\), and compute the distance \(d^k_0 = D(y_{1:n}, z^k_{1:n})\).

3. Based on \((\theta^k_0)^N_{k=1}\) and \((d^k_0)^N_{k=1}\), compute the next threshold \(\varepsilon_1\).

To approximate \(\pi^{\varepsilon_1}\) based on the samples approximating \(\pi^{\varepsilon_{t-1}}\), perform the following for \(t \geq 1\).

1. Define the weight \(w^1_t \propto \mathbb{1}(d^k_{t-1} \leq \varepsilon_t)\) for all \(k \in 1:N\), and normalize the weights.

2. Sample ancestor indices \(a^1_{t-1}, \ldots, a^N_{t-1}\) such that \(\mathbb{P}(a^j_{t-1} = j) = w^j_t\) for all \(j, k \in 1:N\).

3. Sample \(\theta^k_t\) from \(K^{\varepsilon_t}(\theta^k_{t-1}, \cdot)\), a Markov kernel leaving \(\pi^{\varepsilon_t}\) invariant, and store the associated distance \(d^k_t\) and synthetic data set \(z^k_{1:n}\), for all \(k \in 1:N\); this is the rejuvenation step.

4. Based on \((\theta^k_t)^N_{k=1}\) and \((d^k_t)^N_{k=1}\), compute the next threshold \(\varepsilon_{t+1}\).

We use systematic resampling to obtain \(a^1_{t-1:N}\) at step \(t\), noting that other resampling schemes exist with varying degrees of parallelism (Murray et al., 2016).
B.2.1 Choice of Markov kernels

Any choice of kernel $K^\varepsilon$ leaving $\pi^\varepsilon$ invariant would be valid, the simplest arguably being the kernel of Marjoram et al. (2003), i.e. a standard pseudo-marginal Metropolis–Hastings kernel. This kernel tends to be less and less effective at diversifying the particles as $\varepsilon$ goes to zero, which can be compensated for by iterating the kernel more and more times. Instead, we use the $r$-hit kernel of Lee (2012), and Algorithm 6 of that paper in particular, which automatically adapts the computational budget of the kernel to the value of $\varepsilon$; see Lee and Łatuszyński (2014) for a theoretical comparison of MCMC kernels for ABC-type targets. Starting from a value $\theta$ and an associated distance $d$, the kernel works as follows, based on a desired number of hits $r \geq 2$ and a proposal distribution $g(\cdot|\theta)$ on the parameter space.

1. For $i \geq 1$, sample $\theta^i \sim g(\cdot|\theta)$ and $z_{i,n}^j \sim \mu_{\theta^i}^{(n)}$, until $\sum_{j=1}^i 1(D(y_{i,n}, z_{i,n}^j) \leq \varepsilon) = r$. Let $K' = i$.

2. Sample uniformly $L$ among $\{ j \in \{1, \ldots, K' - 1 \} : 1(D(y_{i,n}, z_{i,n}^j) \leq \varepsilon) \}$.

3. For $i \geq 1$, sample $\theta^j \sim g(\cdot|\theta^{(n)})$ and $x_{i,n}^j \sim \mu_{\theta^j}^{(n)}$, until $\sum_{j=1}^i 1(D(y_{i,n}, x_{i,n}^j) \leq \varepsilon) = r - 1$. Let $K = i$.

4. With probability

$$\min \left( 1, \frac{\pi(\theta^{(n)}) g(\theta|\theta^{(n)})}{\pi(\theta) g(\theta^{(n)}|\theta)} K' - 1 \right),$$

output $\theta^{(n)}$ and the distance $D(y_{i,n}, z_{i,n}^j)$; otherwise output $\theta$ and $d$.

We use $r = 2$ as a default choice. The proposal distribution $g(\cdot|\theta)$ can be adapted based on the particles $(w_{t,k}^{i,k}, \theta_{t,k}^{i,k})_{k=1}^N$ at step $t$, which approximate $\pi^\varepsilon$. By default, we can fit a parametric distribution on the particles, such as a Normal distribution using the empirical mean and covariance matrix of the particles. In the numerical experiments, we define $g(\cdot|\theta)$ as a mixture of $5$ multivariate Normal distributions, fitted on the particles available at step $t$, to partially accommodate non-Gaussian features of the target.
B.2.2 Adaptation of thresholds

The SMC framework allows the automatic adaptation of thresholds (see e.g. Del Moral et al., 2012; Silk et al., 2013). A simple approach consists in choosing $\varepsilon_{t+1}$ such that the effective sample size computed on the resulting weights $(w_{t+1}^k)_{k=1}^N$, defined as $\text{ESS}_{t+1} = (\sum_{k=1}^N (w_{t+1}^k)^2)^{-1}$, is above a certain value, e.g. $N/2$. We use a slightly different approach, motivated by the fact that multiple particles might be identical at any step $t \geq 1$ of the SMC sampler. Indeed, the resampling step leads to duplicate values, which the MCMC steps only partially diversify. Criteria that are based solely on the weights fail to account for this potential lack of diversity.

We define a diversity parameter $a$, set to $0.5$ by default, which indicates the desired minimum proportion of unique particles within our sample, at all steps. The idea is to find $\varepsilon_{t+1}$ such that, upon resampling the particles using weights $w_{t+1}^k \propto 1(d_t^k \leq \varepsilon_{t+1})$, we obtain at least a proportion $a$ of unique particles. This is complicated by the randomness of the resampling step. Therefore, we clamp that randomness by drawing the uniform variables of the resampling step and keeping them fixed during the calculation of $\varepsilon_{t+1}$. The proportion of unique particles after resampling is then a deterministic function of $\varepsilon$, denoted by $f(\varepsilon)$. We could define $\varepsilon_{t+1}$ as the value $\varepsilon$ in the interval $[0, \varepsilon_1]$ such that $\int f(\varepsilon) = a$. Since that equality might not be achievable, we use a numerical optimizer to minimize $|f(\varepsilon) - a|$ over $\varepsilon \in [0, \varepsilon_1]$.

As a result, the sequence of thresholds $(\varepsilon_t)$ decreases, but only if the MCMC steps have managed to diversify the particles. We can let the SMC sampler run indefinitely, saving the particles and distances obtained at every step $t$. We can decide to stop the algorithm if the sequence of thresholds $(\varepsilon_t)$ does not decrease anymore, or if the cost of each step becomes prohibitive. We can otherwise run the algorithm for a given wall-clock time, or a given number of model simulations.

B.3 Explicit rates and checking assumptions

Some of the assumptions can be verified using the results developed in Section 2. We check the assumptions and establish explicit rate bounds in a few special cases.
B.3.1 Independent and identically distributed data

We take \( \rho(x, y) = \|x - y\| \) to be the Euclidean distance and denote by \( \mathcal{E}_{\beta, \gamma}(\mu) = \int_{\mathbb{R}^d} e^{\|x\|^\beta} d\mu \) exponential moments with \( \beta > 0, \gamma > 0 \). We start with the case of i.i.d. observations in \( \mathcal{Y} \subseteq \mathbb{R}^d \). Let \( k = 2p \) if \( d_y/2 < p \) and \( k = d_y \) if \( d_y/2 > p \). Assume that \( \mathcal{E}_{\beta, \gamma}(\mu_0) < \infty \) for some \( \beta > p \) and \( \gamma > 0 \) and for all \( \theta \in \mathcal{H} \); otherwise, assumptions can be put on \( q \)-th moments for some \( q > p \).

By Theorem 2 of Fournier and Guillin (2015), for \( \varepsilon < 1 \), we have \( \mu_\theta^{(n)}(\mathcal{W}_p(\mu_\theta; \hat{\mu}_{\theta,n}) > \varepsilon) \leq C(\theta) \exp\left(-c(\theta)n\varepsilon^k\right) \), for some constants \( C(\theta) \geq 0 \) and \( c(\theta) > 0 \). To proceed, we further assume that the rate of convergence of \( \hat{\mu}_{\theta,n} \) to \( \mu_\theta \) varies with \( \theta \) in such a way that \( C(\theta) \) and \( c(\theta) \) can be replaced by constants: \( \mu_\theta^{(n)}(\mathcal{W}_p(\mu_\theta; \hat{\mu}_{\theta,n}) > \varepsilon) \leq C \exp\left(-cn\varepsilon^k\right) \). This might be achieved by restricting the parameters to compact subsets if necessary. We can then define \( f_n : \varepsilon \mapsto \exp\left(-cn\varepsilon^k\right) \), and consider its inverse \( f_n^{-1} : \varepsilon \mapsto \left(-\frac{\log \varepsilon}{cn}\right)^{1/k} \).

One can check that \( \varepsilon_n = c_\varepsilon (\log(n)/n)^{1/k} \), for some \( c_\varepsilon > 0 \), satisfies \( \varepsilon_n \to 0 \), \( f_n(\varepsilon_n) \to 0 \) and \( f_n^{-1}(\varepsilon_n^k) \to 0 \) for all \( L > 0 \) as \( n \to \infty \). If the rate of convergence of the observed empirical distribution \( \hat{\mu}_n \) toward \( \mu_\star \) is at least as fast at that of \( \hat{\mu}_{\theta,n} \) toward \( \mu_\theta \) for all \( \theta \), as specified by the function \( f_n \), then the conditions on \( \varepsilon_n \) in Proposition B.1.3 and Corollary B.1.1 hold. The resulting concentration rate \( K(4\varepsilon_n^3 + f_n^{-1}(\varepsilon_n^k/R))^a \) behaves asymptotically as \( (\log(n)/n)^{a/k} \), where we recall that \( k = 2p \) when \( p > d_y/2 \) and \( k = d_y \) when \( p < d_y/2 \). The obtained rate thus worsens quickly with the dimension of the observation space and a rate close to \( n^{-1/2} \) is retrieved if \( k/a = 2 \), which is the case if e.g. \( p = 1, d_y = 1 \), and \( a = 1 \). The above computations also suggest that the concentration rate decreases with \( p \) and thus that a small value for this parameter should be chosen.

Well-specified location model

Suppose the distributions \( \mu_\theta \) on \( \mathcal{Y} \) have densities such that \( f_\theta(y) = f_\theta(y - \theta) \) for all \( y \in \mathcal{Y} \) and \( \theta \in \mathcal{H} = \mathcal{Y} = \mathbb{R}^d \). Like \( \rho(x, y) \), let also \( \rho(\mathcal{H}; x, y) = \|x - y\| \) be the Euclidean distance. Then, using the definition of the Wasserstein distance, the convexity of the function \( x \mapsto \|x + \theta - \theta_x\|^\gamma \), and Jensen’s inequality, one gets \( \|\theta - \theta_x\| \leq \mathcal{W}_p(\mu_\theta; \mu_{\theta_x}) \) for all \( p \geq 1 \). Therefore, with \( K = 1 \) and \( a = 1 \), Assumption B.1.5 holds. For location models, the rate of convergence of \( \hat{\mu}_{\theta,n} \) to \( \mu_\theta \) is the same for all \( \theta \), so that we can indeed take \( C(\theta) \)
and \(c(\theta)\) to be constants. The assumption \(\mathcal{E}_{\tilde{\beta}, \gamma}(\mu_\theta) < \infty\) for some \(\beta > p\) and \(\gamma > 0\) is, for instance, satisfied in the Normal case. Note also that

\[
\mathcal{W}_p^p(\mu_\theta, \mu_{\theta_*}) = \inf_{\gamma \in \Gamma(\mu_\theta, \mu_{\theta_*})} \int_{Y \times Y} ||x - y||^p \gamma(x, y)
\]

\[
= \inf_{\gamma \in \Gamma(\mu_\theta, \mu_{\theta_*})} \int_{Y \times Y} ||\theta + x - (\theta_* + y)||^p \gamma(x, y) \leq ||\theta - \theta_*||^p,
\]

where the inequality is obtained by considering the maximal coupling between \(\mu_\theta\) and \(\mu_{\theta_*}\). Combined with the reverse inequality obtained via Jensen, we have \(\mathcal{W}_p^p(\mu_\theta, \mu_{\theta_*}) = ||\theta - \theta_*||\) for all \(p \geq 1\). This formula shows that Assumption B.1.4 is satisfied, and that Assumption B.1.5 holds with \(K = 1\) and \(a = 1\).

To satisfy Assumption B.1.3, we need to lower-bound the prior mass of balls \(\{\theta \in \mathcal{H} : ||\theta - \theta_*|| \leq \varepsilon\}\). These have a volume \(v \varepsilon^{d_j}\) for some \(v > 0\) independent of \(\varepsilon\), thus the assumption holds with \(L = d_j\) for any prior with strictly positive density on \(\mathcal{H} = \mathcal{Y} = \mathbb{R}^{d_j}\).

The dimension \(d_j\) of \(\mathcal{Y}\) has an apparent impact on the concentration rate, as in the expression \((\log(n)/n)^{a/d_j}\) derived above. Beside dimension, the order \(p\) also enters the concentration rate. Model misspecification enters through the value of \(a\), which might have to be small for Assumption B.1.5 to hold, as illustrated in the next section.

**Misspecified location model**

Consider a misspecified location model, in which \(\theta\) denotes the mean of \(\mu_\theta\) and \(\theta_*\) the mean of \(\mu_{\theta_*}\). Assume that both \(\mathcal{E}_{\tilde{\beta}, \gamma}(\mu_\theta) < \infty\) and \(\mathcal{E}_{\tilde{\beta}, \gamma}(\mu_{\theta_*}) < \infty\), and let \(p = 2\). Let \(\varepsilon_n = c_\varepsilon (\log(n)/n)^{i/k}\) for some constant \(c_\varepsilon\), where \(k = 2p\) if \(p > d_j/2\) and \(k = d_j\) if \(p < d_j/2\). In the case \(p = 2\), Bickel and Freedman (1981) showed that \(\mathcal{W}_2^2(\mu_\theta, \mu_{\theta_*}) = ||\theta - \theta_*||^2 + \mathcal{W}_2^2(\mu_\theta, \mu_{\theta_*}^o) = ||\theta - \theta_*||^2 + \varepsilon_n^2\), where \(\mu_{\theta_*}^o\) refers to \(\mu_{\theta_*}\) centered at zero. This formula shows that Assumption B.1.4 holds. Now,

\[
K(\mathcal{W}_2(\mu_\theta, \mu_{\theta_*}) - \varepsilon_n^2) = K(\sqrt{||\theta - \theta_*||^2 + \varepsilon_n^2} - \varepsilon_n^2) = K\left\{\frac{||\theta - \theta_*||^2}{2\varepsilon_n^2} - \frac{||\theta - \theta_*||^4}{8\varepsilon_n^4}\right\}^{a} \quad (\text{for } ||\theta - \theta_*||^2 \leq 8\varepsilon_n^2)
\]
\[
\begin{align*}
\geq K \left\{ \frac{\|\theta - \theta_\star\|^2}{2\varepsilon_\star^2} \left( 1 - \frac{\|\theta - \theta_\star\|^2}{4\varepsilon_\star^2} \right) \right\}^a & \quad \text{(for } \|\theta - \theta_\star\|^2 \leq 4\varepsilon_\star^2) \\
\geq K \left( \frac{\|\theta - \theta_\star\|^2}{4\varepsilon_\star^2} \right)^a & \quad \text{(for } \|\theta - \theta_\star\|^2 \leq 2\varepsilon_\star^2),
\end{align*}
\]

where the first inequality uses the fact that
\[
\sqrt{a^2 + b^2} - a \geq \frac{b^2}{2a} - \frac{b^2}{8a^2}
\]
whenever \(0 \leq b^2 \leq 8a^2\). Hence, Assumption B.1.5 is satisfied with \(a = 1/2\) and \(K = 2\sqrt{\varepsilon_\star}\), leading to concentration rates that are the square root of those derived for the well-specified case. Next, we illustrate that this is not simply due to the lack of sharpness in the above approximation.

For this purpose, we use a Normal location model fitted on Gamma-distributed observations. For \(n = 500,000\), we generate \(y_{i,n}\) independently from a Gamma distribution with parameters \((10, 5)\), with a mean of 2. We compute the Wasserstein distance between \(y_{i,n}\) and \(z_{i,n}\) for \(z_{i,n}\) generated from \(\mathcal{N}(\theta, 1)\), for a range of values of \(\theta\) around \(\theta_\star = 2\). The approximation of \(\theta \mapsto \mathcal{W}_2(\mu_\star, \mu_\theta)\) is shown in Figure B.3.1a. In Figure B.3.1b, \(\sqrt{\mathcal{W}_2(\mu_\star, \mu_\theta)} - \varepsilon_\star\) is plotted against \(|\theta - \theta_\star|\), yielding a straight line. This suggests that the relation in Assumption B.1.5 might be an equality when \(a = 1/2\).

![Figure B.3.1](attachment:figure.png)

**Figure B.3.1:** Numerical experiments with a Gamma data-generating distribution \(\mu_\star\) and a Normal location model \(\mu_\theta\), described in Section B.3.1, illustrating that the 2-Wasserstein distance between \(\mu_\theta\) and \(\mu_\star\) satisfies Assumption B.1.5. On the left, \(\mathcal{W}_2(\mu_\star, \mu_\theta)\) is approximated with 500,000 samples and plotted against \(\theta\). On the right \(\sqrt{\mathcal{W}_2(\mu_\star, \mu_\theta)} - \varepsilon_\star\) can be seen to be linearly associated with \(|\theta - \theta_\star|\), for \(\theta\) around \(\theta_\star\) (the best linear fit is indicated by a blue line).
B.3.2 AR(1) model

As described in Fournier and Guillin (2015), deviation inequalities for the convergence of the empirical distribution in the Wasserstein sense have also been derived for certain classes of dependent data. Their inequalities are in the form of moment inequalities, which we convert to concentration inequalities via Markov’s inequality. This does not yield sharp bounds, and is one reason for why the rates found for this example will be slower than those obtained for i.i.d. data.

A stationary stochastic process $Y = (Y_t)_{t \in \mathbb{Z}}$ with marginal distribution $\mu$ is $\rho$-mixing with mixing coefficients $\rho_t$ for $t \geq 0$ if $\rho_t \to 0$ and for all $f, g \in L_2(\mu)$ and $i, j \geq 1$ we have

$$\text{Cov}(f(Y_i), g(Y_j)) \leq \rho_{|i-j|} \sqrt{\text{Var}(f(X_i)) \text{Var}(g(X_j))}.$$  

From the proof of Theorem 14 of Fournier and Guillin (2015) together with Markov’s inequality, it can be seen that a $\rho$-mixing process with marginal distribution $\mu_*$ and $s = \sum_{t=0}^{\infty} \rho_t < \infty$ satisfies

$$\mathbb{P}(W_p(\mu_*, \tilde{\mu}_n) > \varepsilon) \leq c(s, p, q, d_r) M^{p/q}_q(\mu_*) f_n(\varepsilon) = c(s, p, q, d_r) M^{p/q}_q(\mu) \frac{1}{\varepsilon^p v_n(p, q, d_r)},$$

where $v_n = v_n(p, q, d_r) \to \infty$ at a rate depending of $p, q$ and $d_r$, $c(s, p, q, d_r)$ is function depending on $s, p, q$ and $d_r$, and $M_q(\mu) = \int_{\mathbb{R}} ||x||^q d\mu(x)$.

We now focus on the autoregressive model in Example 22. Recall that we defined the delay reconstructed time series $\tilde{y}_t = (y_t, y_{t-1})$ for $t \geq 2$, which can be written

$$\tilde{y}_{t+1} = \begin{pmatrix} \varphi & 0 \\ 1 & 0 \end{pmatrix} \tilde{y}_t + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \left( \begin{pmatrix} \sigma \omega_t \\ v_t \end{pmatrix}, \right)$$

where $v_t$ is an independent Gaussian noise process. Expressed in this way, the process $\tilde{y}_t$ satisfies the definition of an ARMA process given in Pham and Tran (1985). Their Theorem 3.1 can then be shown to hold, with the conclusion that $\tilde{y}_t$ satisfies the weaker notion of $\alpha$-mixing, with coefficients $a_t = O(|\varphi|^t)$ (for a definition, see e.g. Bradley et al., 2005). However, Kolmogorov and Rozanov (1960) show in their Theorem 2 that stationary Gaus-
sian processes that are a-mixing are also ρ-mixing, with coefficients satisfying \( \rho_t \leq 2\pi a_t \), implying that also \( \rho_t \sim O(\| \varphi \|') \).

In particular, note that \( s \leq 1/(1 - |\varphi|) \) and \( d_s = 2 \). Restricting \( \mathcal{H} \) to be the compact set \( \{ \theta = (\varphi, \sigma) : \theta \in [-1 + \delta_1, 1 - \delta_1] \times [\delta_2, M] \} \), for some small \( \delta_1, \delta_2 > 0 \) and some large \( M < \infty \), implies that both \( c(s, p, q, d_p) \) and \( M(p/q)(\mu_\theta) \) are bounded functions of \( \theta \). Letting \( p = 2 \) and \( q > 4 \), the concentration property can then be written, for all \( \theta \in \mathcal{H} \),

\[
\mathbb{P}(W_n(\mu_\theta; \hat{\mu}_{\theta,n}) > \varepsilon) \leq c \frac{n^{-1/2} + n^{-(q-1)/2}}{\varepsilon^2} \leq 2c' \frac{n^{-1/2}}{\varepsilon^2},
\]

for some constant \( c' > 0 \). Letting \( \varepsilon_n = c_n n^{-1/(sL+4)} \) and doing the same calculation of \( f_n^{-1}(\varepsilon_n^L/R) \) and \( f_n(\varepsilon_n) \) as in the earlier examples, we can verify the conditions of Proposition B.1.3.

We now look for a value of \( a \) such that Assumption B.1.5 is satisfied. Let \( \Sigma \) be the covariance of \( (y_t, y_{t-1}) \) from Section 3.4.2,

\[
\Sigma = \frac{\sigma^2}{1 - \varphi^2} \begin{pmatrix} 1 & \varphi \\ \varphi & 1 \end{pmatrix},
\]

and define \( \Sigma * \) analogously for some \( \theta_* = (\varphi_*, \sigma_*) \) in the interior of \( \mathcal{H} \). Then

\[
\mathcal{W}_2(\mu_\theta; \hat{\mu}_{\theta,n}) = \text{tr}(\Sigma + \Sigma_* - 2(\Sigma^{1/2} \Sigma_\star \Sigma^{1/2})^{1/2}) = \| \Sigma^{1/2} - \Sigma_*^{1/2} \|_F^2,
\]

where \( \| \cdot \|_F \) denotes the Frobenius matrix norm, the first equality follows from the closed form 2-Wasserstein distance between two multivariate Normal distributions, and the second follows because \( \Sigma, \Sigma = \Sigma \Sigma_\star \). This formula shows that Assumption B.1.4 holds. By Cholesky decomposition,

\[
\| \Sigma^{1/2} - \Sigma_*^{1/2} \|_F^2 = \left\| \frac{\sigma}{\sqrt{1 - \varphi^2}} \begin{pmatrix} 1 & \varphi \\ \varphi & 1 \end{pmatrix} - \frac{\sigma_*}{\sqrt{1 - \varphi_*^2}} \begin{pmatrix} 1 & \varphi_* \\ \varphi_* & 1 \end{pmatrix} \right\|_F^2
\]

\[
= \left( \frac{\sigma}{\sqrt{1 - \varphi^2}} - \frac{\sigma_*}{\sqrt{1 - \varphi_*^2}} \right)^2 + \left( \frac{\sigma \varphi}{\sqrt{1 - \varphi^2}} - \frac{\sigma_* \varphi_*}{\sqrt{1 - \varphi_*^2}} \right)^2 + (\sigma - \sigma_*)^2
\]

193
\[
\geq \left( \frac{\sigma \varphi}{\sqrt{1 - \varphi^2}} - \frac{\sigma_* \varphi_*}{\sqrt{1 - \varphi_*^2}} \right)^2 + (\sigma - \sigma^*)^2
\]
\[
= (\eta - \eta^*)^2 + (\sigma - \sigma^*)^2,
\]
where \(\eta = \sigma \varphi / \sqrt{1 - \varphi^2}\). Note that the reparametrization \((\varphi, \sigma) \mapsto (\eta, \sigma)\) is one-to-one. Choosing \(\rho_\mathcal{H}(\theta, \theta^*) = \sqrt{(\eta - \eta^*)^2 + (\sigma - \sigma^*)^2}\) yields \(\rho_\mathcal{H}(\theta, \theta^*) \leq W_2(\mu_\theta, \mu_{\theta^*})\). Therefore, Assumption B.1.5 holds with \(K = 1\) and \(a = 1\). Using Corollary B.1.1, we obtain a concentration rate of \(n^{-1/(2L+4)}\). \(\mathcal{H}\) is compact also in the \(\rho_\mathcal{H}\) distance, and with a uniform prior on \((\eta, \sigma)\) it follows that \(L = 2\). We can therefore only bound the concentration rate by \(n^{-1/8}\).

B.3.3 Checking Assumption B.1.5 for the g-and-k Example

With numerical experiments we illustrate Assumption 3.3.5 of the main document (Assumption B.1.5 in this appendix), on the univariate g-and-k model. Since it generates univariate observations, we can quickly compute the Wasserstein distance for large data sets, and here let \(n = 5 \times 10^5\). We generate data \(y_{i,n}\) using the parameter \(\theta^* = (a^*, b^*, g^*, k^*) = (3, 1, 2, 0.5)\). Then, for each parameter \(\theta_i\), we create a grid of 100 values on the interval \(\theta^*_i \pm 0.3\). For each parameter in the grid, we generate \(z_{i,n}\) and compute the associated Wasserstein distance of order \(p = 2\). We plot the distance \(\rho_\mathcal{H}(\theta, \theta^*)\), here taken as the Euclidean distance between \(\theta\) and \(\theta^*\) against Wasserstein distances in Figure B.3.2. If the assumption holds, it means that we should find the distance \(\rho_\mathcal{H}(\theta, \theta^*)\) to be below a polynomial function of the horizontal axis. Here we see that these distances behave seemingly linearly in the Wasserstein distance, which suggests that the assumption holds when treating the parameter component-wise.

In a second experiment, we generate 1000 draws for \(\theta\), uniformly in a hypercube centered at \(\theta^*\) and with width 0.6 on each component. We generate data sets using these parameters, and compute the associated Wasserstein distances relative to \(y_{1:n}\). The resulting distances between parameters are plotted against Wasserstein distances in Figure B.3.3. The figure suggests that the distances between parameters are upper-bounded by a linear function of
the Wasserstein distances in a neighborhood of $\theta_\ast$.

**B.4 Concentration as $n$ increases and $\varepsilon$ decreases of the ABC posterior based on the Hilbert and swapping distances**

Below we establish versions of Proposition 2.3.2 of the main text that hold for the ABC posteriors based on the Hilbert and swapping distances (Proposition B.4.1 and Corollary B.4.1 of this document, respectively). To simplify the presentation, the consistency results presented below assume that the model is well-specified.

**B.4.1 Notation and assumptions**

Let $H : [0, 1] \to [0, 1]^d$ be the Hilbert space-filling curve such that $H(0) = 0 \in \mathbb{R}^d$. Although we refer to $H$ as the Hilbert curve in what follows, there exist in fact several Hilbert curves. We recall that the mapping $H$ is Hölder continuous: there exists a constant $C_{d_H} < \infty$ such that $\rho(H(x), H(y)) \leq C_{d_H}|x - y|^d$, $\forall x, y \in [0, 1]$. In what follows, let $\varepsilon = C_{d_H}^{-1}e^{d_H}$. The mapping $H$ is also surjective, which implies that $H$ admits a Borel measurable pseudo-inverse $h : [0, 1]^d \to [0, 1]$ such that $H(h(x)) = x$ for all $x \in [0, 1]^d$ (Gerber et al., 2019, Proposition 4).

Below we denote by $\mu^h$ the image of probability measure $\mu \in \mathcal{P}((0, 1)^d)$ by $h$ and by $\mathcal{P}_b(\mathcal{Y}) \subset \mathcal{P}(\mathcal{Y})$ the set of probability measures on $\mathcal{Y}$ that admit a continuous and bounded density w.r.t. $\lambda$, the Lebesgue measure on $\mathbb{R}^d$. Henceforth, we fix $p \geq 1$ and a metric $\rho$ on $(0, 1)^d$ and denote by $W_1^*$ the 1-Wasserstein distance on $\mathcal{P}(\mathbb{R})$ obtained for $\rho(x, y) = |x - y|$.

**Assumption B.4.1.** The model $\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\}$ is well-specified; that is, there exists $\theta_\ast \in \mathcal{H}$ such that $\mu_{\theta_\ast} = \mu_{\ast}$. In addition, $\mathcal{M} \subset \mathcal{P}_p(\mathcal{Y}) \cap \mathcal{P}_b(\mathcal{Y})$.

**Assumption B.4.2.** For any $\varepsilon > 0$, $\mu_{\theta}^{(n)}(W_p(\mu_{\theta}, \hat{\mu}_{\theta,n}) > \varepsilon) \leq c(\theta)f_n(\varepsilon)$, where $(f_n)_{n \geq 1}$, is a sequence of strictly decreasing functions and such that $f_n(\varepsilon) \to 0$ as $n \to \infty$, and the function $c : \mathcal{H} \to \mathbb{R}^+$ is $\pi$-integrable.
Assumption B.4.3. For any \( \varepsilon > 0 \), \( \mu^{(n)}_\theta (\mathcal{W}^*_i (\mu^h_\theta \mu^h_\bar{\theta}, n)) > \varepsilon \) \( \leq \tilde{c} (\theta) \bar{f}_n (\varepsilon) \), where \( \bar{f}_n (\varepsilon) \), is a sequence of strictly decreasing functions such that \( f_n (\varepsilon_n) \to 0 \) for any sequence \((\varepsilon_n)_{n\geq 1}\), that converges to 0 sufficiently slow as \( n \to \infty \). The function \( \tilde{c} : \mathcal{H} \to \mathbb{R}^+ \) satisfies \( \tilde{c} (\theta) \leq \tilde{c}_o \) for some \( \tilde{c}_o > 0 \), for all \( \theta \) such that, for some \( \tilde{\delta}_o > 0 \), \( \mathcal{W}^*_i (\mu^* , \mu^{h}_\theta) \leq \tilde{\delta}_o \).

Note that Assumption B.4.3 tends to hold under weaker conditions than the analogous Assumption B.4.2, since \( \mu^h_\theta \) and \( \mu^h_\bar{\theta} \) are distributions on the bounded interval \([0, 1] \). The moment conditions in Fournier and Guillin (2015) are therefore satisfied. For instance, in the case of i.i.d. data, Assumption B.4.3 is automatically satisfied by their Theorem 1 and Markov’s inequality. However, it may be that other functions \( \tilde{f}_n \) and \( \tilde{c} \) exist to give sharper bounds.

Assumption B.4.4. The function \( r(\varepsilon) := \pi \left( \{ \theta \in \mathcal{H} : \mathcal{W}^*_i (\mu^* , \mu^h_\theta) \leq \varepsilon \} \right) \) satisfies \( r(\varepsilon) > 0 \) for any \( \varepsilon > 0 \).

In Section B.4.4, we illustrate that this assumption holds under relatively weak conditions.

B.4.2 Concentration of the ABC posterior based on the Hilbert distance

We denote below by \( \pi^c_{\mathcal{H}_p} \) the ABC posterior based on the Hilbert distance \( \mathcal{H}_p \); that is,

\[
\pi^c_{\mathcal{H}_p ; \varepsilon_n} (d\theta) = \frac{\pi (d\theta) \int_{\mathcal{Y}^n} \mathcal{H}^{(n)} (dz_{i:n}) \mathbb{1}_{\{ \delta(p, \mu_{\varepsilon_n, n}) < \varepsilon \}}}{\int_{\mathcal{H}} \pi (d\theta) \int_{\mathcal{Y}^n} \mathcal{H}^{(n)} (dz_{i:n}) \mathbb{1}_{\{ \delta(p, \mu_{\varepsilon_n, n}) < \varepsilon \}}}.
\]

Proposition B.4.1. Let \( \mathcal{Y} = (0, 1)^d \), \( h_\mathcal{Y} = h \), and \( 1 \leq p \leq d_\mathcal{Y} \). Under Assumptions B.4.1-B.4.4, consider a sequence \((\varepsilon_n)_{n \geq 0}\) such that, as \( n \to \infty \), \( \varepsilon_n \to 0 \), \( f_n (\varepsilon_n) \to 0 \), and \( f_n (\varepsilon_n) \to 0 \). Then, the ABC posterior based on the Hilbert distance with threshold \( \varepsilon_n \) satisfies, for some \( 0 < C < \infty \) and any \( 0 < R < \infty \),

\[
\pi^c_{\mathcal{H}_p ; \varepsilon_n} \left( \{ \theta \in \mathcal{H} : \mathcal{W}^*_p (\mu^* , \mu_{\theta}) > 4\varepsilon_n / 3 + f_n (r(\varepsilon_n / 3) / R) \} \right) \leq \frac{C}{R},
\]

with \( \mathbb{P} \)-probability going to 1 as \( n \to \infty \).
Proof. We first look at the ABC posterior probability of the set \( \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu_\star, \mu_\theta) > \delta \} \). Note that, using Bayes’ formula, for all \( \varepsilon, \delta > 0 \),

\[
\pi^\varepsilon \left( \mathcal{W}_p(\mu_\star, \mu_\theta) > \delta \right) = \frac{\mathbb{P}_\theta \left( \mathcal{W}_p(\mu_\star, \mu_\theta) > \delta, \mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \varepsilon \right)}{\mathbb{P}_\theta \left( \mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \varepsilon \right)},
\]

where \( \mathbb{P}_\theta \) denotes the distribution of \( \theta \sim \pi \) and of the synthetic data \( z_{\varepsilon,n} \sim \mu_\theta^{(n)} \), keeping the observed data \( y_{\varepsilon,n} \) and hence \( \hat{\mu}_n \) fixed.

We first study the numerator. Using the fact that for any \( \theta \in \mathcal{H} \),

\[
\mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n), \quad \mu_\theta^{(n)} \text{-almost surely},
\]

it follows that

\[
\mathbb{P}_\theta \left( \mathcal{W}_p(\mu_\star, \mu_\theta) > \delta, \mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \varepsilon \right) \leq \mathbb{P}_\theta \left( \mathcal{W}_p(\mu_\star, \mu_\theta) > \delta, \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \varepsilon \right),
\]

where, under Assumption B.4.2, the right-hand side is bounded as in the proof of Proposition 2.3.2 of the main text. Hence, for any \( \varepsilon > 0 \) and

\[
y_{\varepsilon,n} \in A(n, \varepsilon) := \left\{ y_{\varepsilon,n} : \mathcal{W}_p \left( n^{-1} \sum_{i=1}^{n} \delta_{y_i, \mu_\star} \right) \leq \varepsilon / 3 \right\},
\]

we have

\[
\pi^\varepsilon \left( \mathcal{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon / 3 + \xi \right) \leq \frac{c_{\varepsilon} f_{\varepsilon}(\xi)}{\mathbb{P}_\theta \left( \mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \varepsilon \right)} \quad (B.1)
\]

for a constant \( c_{\varepsilon} < \infty \) and with the sequence of functions \( (f_n)_{n \geq 1} \), as in Assumption B.4.2.

We now lower bound the denominator in (B.1). Let \( y_i^h = h(y_i) \) and \( z_i^h = h(z_i) \), for \( i \in 1 : n \). Then, for any \( \theta \in \mathcal{H} \) and \( \mu_\theta^{(n)} \)-almost surely,

\[
\mathcal{S}_p(\hat{\mu}_n, \hat{\mu}_\theta, n) \leq \mathcal{S}_{d_p}(\hat{\mu}_n, \hat{\mu}_\theta, n),
\]

(since \( \| \cdot \|_p \leq \| \cdot \|_{d_p} \), for any \( x \in \mathbb{R}^n \) and \( 1 \leq p \leq d_p \))

\[
= \left( \frac{1}{n} \sum_{i=1}^{n} \rho(y_{\varepsilon(i), i}, z_{\varepsilon(i), i}) \right)^{1/d_p} \quad \text{(by definition of} \mathcal{S}_{d_p} \text{)}
\]

197
where

$$= \left( \frac{1}{n} \sum_{i=1}^{n} \rho \left( H(y_{(i)}^h), H(z_{(i)}^h) \right) \right)^{1/d_y} \quad \text{(since } H(h(x)) = x, \forall x \in (0, 1)^d)$$

$$\leq \left( \frac{C_{d_y}}{n} \sum_{i=1}^{n} |y_{(i)}^h - z_{(i)}^h| \right)^{1/d_y} \quad \text{(by the Hölder property of } H)$$

$$= (C_{d_y} \mathcal{W}^*_i(y_{(n)}^h, \hat{\theta}_{(n)}))^{1/d_y} \quad \text{(by definition of } \mathcal{W}^*_i),$$

so that

$$\mathbb{P}_\theta(S_{\mathcal{P}}(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon) \geq \mathbb{P}_\theta\left(C_{d_y} \mathcal{W}^*_i(y_{(n)}^h, \hat{\theta}_{(n)})^{1/d_y} \leq \varepsilon\right) = \mathbb{P}_\theta\left(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq C_{d_y}^{-1} \varepsilon^{d_y}\right).$$

To bound the right-hand side, let $\bar{\varepsilon} = C_{d_y}^{-1} \varepsilon^{d_y}$ and assume henceforth that

$$\mathcal{W}^*_i(y_{(n)}^h, \hat{\theta}_{(n)}) \leq \bar{\varepsilon} \subseteq \left\{ y_{(n)}^h : \mathcal{W}^*_i\left(n^{-1} \sum_{i=1}^{n} S_{h(y_{(i)}^h), \mu_i^h}\right) \leq C_{d_y}^{-1} \varepsilon^{d_y}/3 \right\}. \quad \text{(B.2)}$$

Then,

$$\mathbb{P}_\theta(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}) = \int_{\mathcal{H}} \mu_{\theta}^{(n)}(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}) \pi(d\theta) \geq \int_{\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3} \mu_{\theta}^{(n)}(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3) \pi(d\theta) \quad \text{(by non-negativity of integrand)}$$

$$\geq \int_{\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3} \mu_{\theta}^{(n)}(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}) \pi(d\theta) \quad \text{(by the triangle inequality)}$$

$$\geq \pi(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}) - \int_{\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3} \mu_{\theta}^{(n)}(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) > \bar{\varepsilon}/3) \pi(d\theta) \quad \text{(since } \mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3 \text{ and } \mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3)$$

$$\geq \pi(\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}) - \int_{\mathcal{W}^*_i(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \bar{\varepsilon}/3} \varepsilon(\hat{\theta}_{(n)}(\bar{\varepsilon}/3)) \pi(d\theta) \quad \text{(by Assumption B.4.3),}$$

198
where the third inequality uses (B.2). We now make specific choices for \(\varepsilon\) (and thus \(\tilde{\varepsilon}\)) and \(\zeta\), starting by assuming that \(\tilde{\varepsilon}/3 \leq \tilde{\delta}_o\), such that \(\tilde{c}(\theta) \leq \tilde{c}_o\) for some constant \(\tilde{c}_o > 0\) in the last integrand above, by Assumption B.4.3. The last line above is then greater than or equal to \(\pi(\mathcal{W}_1^*(\mu^h, \mu^h) \leq \tilde{\varepsilon}/3)(1 - \tilde{c}_o\tilde{f}_n(\tilde{\varepsilon}/3))\). Replacing \(\tilde{\varepsilon}\) with \(\varepsilon_n\) such that \(\tilde{f}_n(\varepsilon_n/3) \to 0\) implies that \(\tilde{c}_o\tilde{f}_n(\varepsilon_n/3) \leq 1/2\) for sufficiently large \(n\). Hence,

\[
\pi(\mathcal{W}_1^*(\mu^h, \mu^h) \leq \varepsilon_n/3) \left(1 - \tilde{c}_o\tilde{f}_n(\varepsilon_n/3)\right) \geq \frac{1}{2}\pi(\mathcal{W}_1^*(\mu^h, \mu^h) \leq \varepsilon_n/3) = \frac{1}{2}r(\varepsilon_n/3)
\]

for sufficiently large \(n\), by Assumption B.4.4. Summarizing the bounds derived above, we get

\[
\pi_{\mathcal{H}_p}^z(\mathcal{W}_p(\mu, \mu) > 4\varepsilon_n/3 + \zeta|y_{1:n}) \leq \frac{Cf_\varepsilon(\zeta)}{r(\varepsilon_n/3)},
\]

where \(C = 2\zeta_i\). Set some \(R > 0\) and assume that \(f_n^{-1}(r(\varepsilon_n/3)/R)\) is well-defined, in the sense that \(f_n^{-1}\) is defined at \(r(\varepsilon_n/3)/R\), which can be achieved by taking \(n\) large enough. Choosing \(\zeta = f_n^{-1}(r(\varepsilon_n/3)/R)\) leads to

\[
\pi_{\mathcal{H}_p, y_{1:n}}^z(\mathcal{W}_p(\mu, \mu) > 4\varepsilon_n/3 + f_n^{-1}(r(\varepsilon_n/3)/R)) \leq \frac{C}{R}. \quad (B.3)
\]

If the model is well-specified, the convergence \(\mathbb{P}(\{\omega : y_{1:n}(\omega) \in A(n, \varepsilon_n)\}) \to 1\) and \(\mathbb{P}(\{\omega : y_{1:n}(\omega) \in A^h(n, \varepsilon_n)\}) \to 1\) as \(n \to \infty\) is implied by \(f_n(\varepsilon_n/3) \to 0\) and \(\tilde{f}_n(\varepsilon_n/3) \to 0\) respectively. As a consequence, the statement in (B.3) holds with probability going to one.

\[\square\]

**B.4.3 Concentration of the ABC posterior based on the swapping distance**

Let \(\mathcal{G}_p(\hat{\mu}_n, \hat{\theta}_n)\) be the swapping distance. Then, the result for the ABC posterior based on the this distance, defined by

\[
\pi_{\mathcal{G}_p, y_{1:n}}^z(\mathcal{D}) = \frac{\pi(d\theta) \int_{\mathcal{D}} \mu_{\theta}(d\varepsilon_{1:n}) \mathcal{G}_p(\hat{\mu}_n, \hat{\theta}_n) \leq \varepsilon}{\int_{\mathcal{D}} \pi(d\theta) \int_{\mathcal{D}} \mu_{\theta}(d\varepsilon_{1:n}) \mathcal{G}_p(\hat{\mu}_n, \hat{\theta}_n) \leq \varepsilon},
\]

199
is given in the next corollary.

**Corollary B.4.1.** Let \( Y = (0, 1)^d \), \( y = h \), and \( 1 \leq p \leq d_y \). Under Assumptions B.4.1-B.4.4, consider a sequence \( (\varepsilon_n)_{n \geq 0} \) such that, as \( n \to \infty \), \( \varepsilon_n \to 0 \), \( f_n(\varepsilon_n) \to 0 \) and \( f_n(\varepsilon_n) \to 0 \). Then, the ABC posterior based on the swapping distance with threshold \( \varepsilon_n \) satisfies, for some \( 0 < C < \infty \) and any \( 0 < R < \infty \),

\[
\pi_{\varepsilon, y, n}^{\varepsilon_n} \left( \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu^*_{\varepsilon}, \mu_\theta) > 4\varepsilon_n / 3 + f_n^{-1}(r(\varepsilon_n / 3) / R) \} \right) \leq \frac{C}{R},
\]

with \( \mathbb{P} \)-probability going to 1 as \( n \to \infty \).

**Proof.** The result follows from the proof of Proposition B.4.1 and from the fact that, for all \( n \geq 1 \) and \( \theta \in \mathcal{H} \),

\[
\mathcal{W}_p(\mu_{\varepsilon_n}, \mu_{\theta, n}) \leq \mathcal{G}_p(\mu_{\varepsilon_n}, \mu_{\theta, n}) \leq \mathcal{F}_p(\mu_{\varepsilon_n}, \mu_{\theta, n}), \quad \mathbb{P} - \text{a.s.}
\]

\[\Box\]

**B.4.4 Verifying the Assumptions**

The main text and the earlier discussion in this document contain remarks and references to results that are useful for establishing Assumptions B.4.1-B.4.3. In this section, we develop conditions under which Assumption B.4.4 holds. Consider the following assumptions and technical lemmas.

**Assumption B.4.5.** The map \( \theta \mapsto \mu_\theta \) is injective.

**Assumption B.4.6.** For any sequence \( (\theta_n)_{n \geq 1} \) and \( \theta \in \mathcal{H} \) such that \( \rho_\mathcal{H}(\theta_n, \theta) \to 0 \), we have that \( \mathcal{W}_p(\mu_{\theta_n}, \mu_\theta) \to 0 \).

**Lemma B.4.1.** Let \( \mu \in \mathcal{P}_p((0,1)^d) \cap \mathcal{P}_p((0,1)^d) \) and \( (\mu_n)_{n \geq 1} \) be a sequence of probability measures in \( \mathcal{P}_p((0,1)^d) \) such that \( \mu_n \Rightarrow \mu \). Then, as \( n \to \infty \), \( \mathcal{W}_p(\mu_n, \mu^h) \to 0 \).

**Proof.** By Theorem 3, \( \mathcal{W}_p(\mu_n, \mu^h) \to 0 \) if and only if (i) \( \mu_n^h \Rightarrow \mu^h \) and (ii) there exists \( y_0 \in (0, 1) \) such that \( \int_{(y_0,1)} \rho(y, y_0) \mu^h_n(dy) \to \int_{(y_0,1)} \rho(y, y_0) \mu^h(dy) \). By Gerber et al. (2019,
Theorem 9), \( \mu_n^h \Rightarrow \mu^h \) and thus the result follows from the fact that, for any fixed \( y_0 \in (0, 1) \), the mapping \( y \mapsto \rho(y, y_0)^p \) is continuous and bounded on \((0, 1)\). \( \square \)

**Lemma B.4.2.** Let \( \mathcal{Y} = (0, 1)^d \). Under Assumption B.4.5, the map \( \theta \mapsto \mu^h_\theta \) is injective.

**Proof.** Let \( \theta \) and \( \theta' \) be two distinct points in \( \mathcal{H} \). Then, under Assumption B.4.5, there exists a measurable set \( A \subset \mathcal{Y} \) such that \( \mu^h_\theta(A) \neq \mu^h_{\theta'}(A) \). To conclude the proof let \( I_A = h(A) \) and note that, since the mapping \( h : (0, 1)^d \rightarrow [0, 1] \) is one-to-one, we have \( \mu^h_\theta(I_A) = \mu^h_{\theta'}(I_A) \).

**Lemma B.4.3.** Let \( \mathcal{Y} = (0, 1)^d \). Under Assumptions B.4.1 and B.4.6, for any sequence \( (\theta_n)_{n \geq 1} \), and \( \theta \in \mathcal{H} \) such that \( \rho(\mathcal{H}(\theta_n, \theta) \rightarrow 0 \), we have that \( \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_\theta) \rightarrow 0 \).

**Proof.** The result is a direct consequence of Lemma B.4.1 and Assumption B.4.6. \( \square \)

**Lemma B.4.4.** Let \( \mathcal{Y} = (0, 1)^d \). Under Assumptions B.4.1, B.4.5 and B.4.6, the mapping \( \theta \mapsto \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_\theta) \) has a unique minimum on \( \mathcal{H} \) at \( \theta = \theta_\star \). In addition, for any \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that

\[
\inf_{\{\theta \in \mathcal{H} : \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_\theta) \geq \varepsilon\}} \mathcal{W}_p(\mu^h_{\theta_\star}, \mu^h_\theta) > \delta.
\]

**Proof.** Under Assumptions B.4.5 and B.4.1, there exists a unique \( \theta_\star \in \mathcal{H} \) such that \( \mu^h_{\theta_\star} = \mu^h_{\star} \) and thus \( \mu^h_{\theta_\star} = \mu^h_{\star} \). Then, the first part of the result follows from the fact that, under Assumption B.4.5, the map \( \theta \mapsto \mu^h_\theta \) is injective by Lemma B.4.2. The second part of the lemma follows from Proposition A.4.2, using the fact that, under Assumptions B.4.6 and B.4.1, the map \( \theta \mapsto \mathcal{W}_p(\mu^h_\theta, \mu^h_{\theta_\star}) \) is continuous by Lemmas A.1.1 and B.4.3. \( \square \)

**Lemma B.4.5.** Let \( \mathcal{Y} = (0, 1)^d \). Under Assumptions B.4.1, B.4.5 and B.4.6, for any \( \varepsilon > 0 \) the set \( \tilde{B}_\theta(\varepsilon) = \{ \theta \in \mathcal{H} : \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_{\theta_\star}) \leq \varepsilon \} \) is compact and has strictly positive Lebesgue measure.

**Proof.** Under Assumptions B.4.5 and B.4.6, by Lemmas A.1.1 and B.4.3, the map \( \theta \mapsto \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_{\theta_\star}) \) is continuous. Therefore, for any \( \varepsilon > 0 \) the set \( \tilde{B}_\theta(\varepsilon) \) is compact since it is the pre-image of the compact set \([0, \varepsilon]\) by the continuous mapping \( \theta \mapsto \mathcal{W}_p(\mu^h_{\theta_n}, \mu^h_{\theta_\star}) \). Lastly, for any \( \varepsilon > 0 \) the set \( \tilde{B}_\theta(\varepsilon) \) has strictly positive measure since, under Assumption B.4.5, the map \( \theta \mapsto \mu^h_\theta \) is injective by Lemma B.4.2. \( \square \)
**Proposition B.4.2.** Let $\mathcal{Y} = (0,1)^d$, $h_\mathcal{Y} = h$, and $1 \leq p \leq d_p$. Assume that the prior distribution $\pi(d\theta)$ admits a density (with respect to the Lebesgue measure) which is continuous and strictly positive on a neighborhood of $\theta_*$. Then, under Assumptions B.4.1, B.4.5 and B.4.6, Assumption B.4.4 holds.

**Proof.** By Lemma B.4.5, the set $\tilde{B}_{\delta_n}(\varepsilon) := \{\theta \in \mathcal{H} : \mathcal{W}_i^r(\mu_\theta^h, \mu_\theta^h) \leq \varepsilon\}$ is compact and has strictly positive Lebesgue measure for any $\varepsilon > 0$. In addition, because the mapping $\theta \mapsto \mathcal{W}_i^r(\mu_\theta^h, \mu_\theta^h)$ has a unique minimum at $\theta = \theta_*$ (Lemma B.4.4) and is continuous (Lemma B.4.3), it follows that $\cap_{n=1}^{\infty} \tilde{B}_{\delta_n}(\varepsilon_n/3) = \{\theta_*\}$. Since, by assumption, the prior distribution $\pi(d\theta)$ admits a density (w.r.t. Lebesgue measure) which is continuous and strictly positive on a neighborhood of $\theta_*$, it follows that $\pi\left(\{\theta \in \mathcal{H} : \mathcal{W}_i^r(\mu_\theta^h, \mu_\theta^h) \leq \varepsilon\}\right) > 0$ for any $\varepsilon > 0$.

\[\square\]

**B.5 Approximation of the Wasserstein Distance**

We present numerical results on the approximation of Wasserstein distances between empirical distributions by the swapping and Hilbert distances with a simple example. We consider a multivariate Normal model: the observations are assumed i.i.d. $\mathcal{N}(\mu_d, \theta I_d)$, where $\mu_d$ is a vector of $d$ zeros, and $I_d$ is the $d \times d$ identity matrix. The univariate parameter $\theta \geq 0$ scales the covariance matrix. We define the data-generating distribution as $\mathcal{N}(\mu_d, 4 I_d)$, so the inference task is to retrieve $\theta_* = 4$. We generate $y_i$ for $i = 1, \ldots, n$, with $n = 500$, from $\mathcal{N}(\mu_d, 4 I_d)$, and for $d = 2, 3, 4, 5$. Then, over a grid of 100 values of $\theta$ equispaced between 0.1 and 9, we generate $z_i$ for $i = 1, \ldots, n$, from $\mathcal{N}(\mu_d, \theta I_d)$. We then compute the Wasserstein distance of order $p = 2$, $\mathcal{W}_2(y_{1:n}, z_{1:n})$, as well as the Hilbert and the swapping distances of the same order. For each dimension, the distances are plotted against $\theta$, in Figure B.5.1. We observe that for $d = 2$, the shape of the curve corresponding to each distance is very similar. We note that, as expected, the swapping distance is always larger than the Wasserstein distance, and the Hilbert distance is larger than the swapping distance. When the dimension increases, the curve corresponding to the Hilbert distance seems to depart from that of the Wasserstein distance. In particular, an important difference is that the minimizer over $\theta$ is visibly different with the Hilbert distance compared
to the Wasserstein distance, when the dimension increases. We observe that the swapping distance remains closer to the Wasserstein distance, but that the discrepancy also increases with the dimension.

In this example, the computing time is found to be practically insensitive to the dimension $d$. For each dimension, the Hilbert distances were computed in a median time of $2 \times 10^{-3}$ seconds, swapping distances in a median time of $3 \times 10^{-2}$ seconds, and exact Wasserstein distances in a median time of $2 \times 10^{-1}$ seconds. In this instance, generating data sets takes around $10^{-4}$ seconds, thus the cost of computing distances is much larger than the cost of generating data; this would not necessarily be the case in typical ABC settings.
Figure B.3.2: Distance $\rho_H(\theta, \theta_*)$ against $W_2(y_{lin}, z_{lin})$, where $y_{lin} \sim \mu_{\theta_*}$ and $z_{lin} \sim \mu_{\theta}$, when varying each component of $\theta$ in a neighborhood of $\theta_*$, in the g-and-k example as described in Section B.3.3.
**Figure B.3.3:** Distance $\rho_d(\theta, \theta_*)$ against $W_2(y_{1:n}, z_{1:n})$, where $y_{1:n} \sim \mu_{\theta_*}$ and $z_{1:n} \sim \mu_{\theta}$, when varying all components of $\theta$ in a neighborhood of $\theta_*$, in the g-and-k example as described in Section B.3.3.

**Figure B.5.1:** Hilbert, swapping and Wasserstein distances between data generated from multivariate Normal distributions with variance $\theta$ times the identity matrix, versus $\theta_* = 4$, with 500 independent draws. Each panel corresponds to a different dimension. Details are provided in Section B.5.
Appendix to Chapter 4

C.1 Proofs

Closely following Clément and Maas (2011) and Ambrosio et al. (2005), we start by proving a discrete version of the evolution variational inequality used to characterize gradient flows. Using interpolations of the discrete solutions, we use the discrete EVI to build a continuous approximation to the desired EVI. With this approximation, we derive a bound that quantifies the closeness of two discrete solutions. This bound is used to show that under appropriate assumptions on a sequence of discrete solutions, this sequence is Cauchy and therefore has a limit. Lastly, this limit is shown to be the desired gradient flow.

Lemma C.1.1 (Discrete Evolution Variation Inequality). For any $n \geq 1$, $h > 0$, $v \ll \mu_{\text{Leb}}$
and \( k = 0, \ldots, n - 1 \) we have

\[
\frac{1}{2h} \left[ \mathcal{W}_2^z(p_h^{k+1/2}, v) - \mathcal{W}_2^z(p_h^k, v) \right] + \frac{\lambda}{2} \mathcal{W}_2^z(p_h^{k+1/2}, v) \\
\leq \mathcal{H}(v|\pi) - \mathcal{H}(p_h^{k+1}|\pi) - \frac{1}{2h} \mathcal{W}_2^z(p_h^{k+1/2}, p_h^k) + \delta_h^{k+1}.
\]

**Proof.** By Corollary 4.1.3 of Ambrosio et al. (2005) (see also their Lemma 9.2.7), for any \( p_h^k \ll \mu_{\text{Leb}} \), we have

\[
\frac{1}{2h} \left[ \mathcal{W}_2^z(p_h^{k+1/2}, v) - \mathcal{W}_2^z(p_h^k, v) \right] + \frac{\lambda}{2} \mathcal{W}_2^z(p_h^{k+1/2}, v) \\
\leq \mathcal{V}(v) - \mathcal{V}(p_h^{k+1/2}) - \frac{1}{2h} \mathcal{W}_2^z(p_h^{k+1/2}, p_h^k).
\]

(C.1)

Recall that \( t \mapsto \phi_t \ast p_h^{k+1/2} \) is the gradient flow of the \( o \)-displacement convex entropy functional \( \rho \mapsto \mathcal{H}(\rho) \). Therefore,

\[
\frac{d}{dt} \frac{1}{2} \mathcal{W}_2^z(\phi_t \ast p_h^{k+1/2}, v) + \mathcal{H}(\phi_t \ast p_h^{k+1/2}) \leq \mathcal{H}(v),
\]

in the sense of distributions. By Remark 1.2 of Clément and Maas (2011), an equivalent condition is: for all \( 0 < a < b < \infty \),

\[
\frac{1}{2} \left[ \mathcal{W}_2^z(\phi_b \ast p_h^{k+1/2}, v) - \mathcal{W}_2^z(\phi_a \ast p_h^{k+1/2}, v) \right] \\
\leq (b - a) \mathcal{H}(v) - \int_a^b \mathcal{H}(\phi_t \ast p_h^{k+1/2})dt.
\]

Noting that \( t \mapsto \mathcal{H}(\phi_t \ast p_h^{k+1/2}) \) is non-increasing by Theorem 11.2.1 of Ambrosio et al. (2005) (see equation 11.2.4), we have that for all \( 0 < a < b < \infty \),

\[
\frac{1}{2} \left[ \mathcal{W}_2^z(\phi_b \ast p_h^{k+1/2}, v) - \mathcal{W}_2^z(\phi_a \ast p_h^{k+1/2}, v) \right] \\
\leq (b - a) \mathcal{H}(v) - (b - a) \mathcal{H}(\phi_b \ast p_h^{k+1/2}).
\]
Letting $a \to 0$, $b = h$, we have

$$\frac{1}{2h} \left[ W_z^2(\rho_h^{k+1}, \nu) - W_z^2(\rho_h^{k+1/2}, \nu) \right] \leq H(\nu) - H(\rho_h^{k+1}).$$

(C.2)

Adding inequalities (C.1) and (C.2), as well as adding and subtracting $W(\rho_h^{k+1})$ to the right hand side to make $\delta_h^{k+1}$ appear, yields the result. $\square$

It can be deduced from Lemma C.1.1 that

$$\frac{1}{2h} W_z^2(\rho_h^{k+1}, \rho_h^k) \leq H(\rho_h^k|\pi) - H(\rho_h^{k+1}|\pi) - \frac{1 + \lambda h}{2h} W_z^2(\rho_h^{k+1/2}, \rho_h^k) + \delta_h^{k+1},$$

(C.3)

by taking $\nu = \rho_h^k$, so that

$$\sum_{k=0}^{n-1} W_z^2(\rho_h^{k+1}, \rho_h^k) \leq 2h \left[ H(\rho_h^k|\pi) - H(\rho_h^k|\pi) + \Delta_h^n \right],$$

$$\leq 2h \left[ H(\rho_h^k|\pi) + \Delta_h^n \right].$$

Similarly,

$$W_z^2(\rho_h^{k+1/2}, \rho_h^k) \leq \frac{2h}{1 + \lambda h} \left[ H(\rho_h^k|\pi) - H(\rho_h^{k+1}|\pi) + \delta_h^{k+1} \right],$$

so that

$$\sum_{k=0}^{n-1} W_z^2(\rho_h^{k+1/2}, \rho_h^k) \leq \frac{2h}{1 + \lambda h} \left[ H(\rho_h^k|\pi) + \Delta_h^n \right].$$

(C.4)

Before proceeding, we introduce some more notation. Introduce the delayed interpolation $\rho_h(t) = \rho_h^k$ if $t \in [hk, (k + 1)h)$, and note that $\rho^h(t)$ and $\rho_h(t)$ are left and right continuous respectively. Introduce also an interpolation of the half-steps, denoted by $\rho_{h/2}(t) = \rho_h^{k+1/2}$ if $t \in [hk, (k + 1)h)$.

Define the piecewise affine function

$$\ell_h(t) = \frac{t - hk}{h} \quad \text{if} \ t \in [hk, (k + 1)h).$$
and in turn let
\[
\mathcal{W}^z_h(t, v) = (1 - \ell_h(t))\mathcal{W}^z_h(\rho^h(t), v) + \ell_h(t)\mathcal{W}^z(\rho^k(t), v),
\]
\[
\mathcal{H}_h(t) = (1 - \ell_h(t))\mathcal{H}(\rho^h(t)|\pi) + \ell_h(t)\mathcal{H}(\rho^k(t)|\pi).
\]
Let also
\[
R_h(t) = 2(1 - \ell_h(t))\left(\mathcal{H}(\rho^k|\pi) - \mathcal{H}(\rho^{k+1}|\pi) + \delta^k_h\right) + 2\ell_h(t)\delta^k_h
\]
for \( t \in [hk, (k + 1)h) \). By (C.3) and \( \delta^k_h \geq 0 \), it is clear that \( R_h(t) \geq 0 \). The following result is an analog of Theorem 4.1.4 of Ambrosio et al. (2005).

**Lemma C.1.2** (Gradient flow approximation). For any \( n \geq 1 \), \( h > 0 \), \( v \ll \mu_{t, d} \) and \( t \in [o, hn] \setminus \{kh : k = o, \ldots, n\} \), we have
\[
\frac{d}{dt} \mathcal{W}^z_h(t, v) + \frac{\lambda}{2} \mathcal{W}^z_h(\rho^h(t), v) + \mathcal{H}_h(t) - \mathcal{H}(v|\pi) \leq \frac{1}{2} R_h(t),
\]
where \( d/dt \) denotes the pointwise derivative.

**Proof.** If \( t \in (hk, (k + 1)h) \), then
\[
\frac{d}{dt} \mathcal{W}^z_h(t, v) = \frac{1}{2h} \left[ \mathcal{W}^z_h(\rho^{k+1}, v) - \mathcal{W}^z_h(\rho^k, v) \right].
\]
By Lemma C.1.1, this means
\[
\frac{d}{dt} \mathcal{W}^z_h(t, v) + \frac{\lambda}{2} \mathcal{W}^z_h(\rho^h(t), v) + \mathcal{H}_h(t) - \mathcal{H}(v|\pi)
\]
\[
= \frac{1}{2h} \left[ \mathcal{W}^z_h(\rho^{k+1}, v) - \mathcal{W}^z_h(\rho^k, v) \right] + \frac{\lambda}{2} \mathcal{W}^z_h(\rho^h(t), v) + \mathcal{H}_h(t) - \mathcal{H}(v|\pi)
\]
\[
\leq \mathcal{H}_h(t) - \mathcal{H}(\rho^{k+1}|\pi) + \delta^k_h
\]
\[
= (1 - \ell_h(t))\mathcal{H}(\rho^k|\pi) + \ell_h(t)\mathcal{H}(\rho^{k+1}|\pi) - \mathcal{H}(\rho^{k+1}|\pi) + \delta^k_h
\]
\[
= (1 - \ell_h(t))\left(\mathcal{H}(\rho^k|\pi) - \mathcal{H}(\rho^{k+1}|\pi)\right) + \delta^k_h
\]
\[
= \frac{1}{2} R_h(t).
\]
Lemma C.1.3. For any $n \geq 1$, $h > 0$ and $k = 0, \ldots, n - 1$, we have the estimate

$$0 \leq \int_0^{(k+1)h} R_h(t)dt \leq h \left( \mathcal{H}(\rho_h^0|\pi) + 2\Delta_h^n \right).$$

Proof. The lower bound follows from $R_h(t) \geq 0$ for all $t \in [0, hn]$. Observe that

$$\int_{kh}^{(k+1)h} \ell_h(t)dt = \int_{kh}^{(k+1)h} (1 - \ell_h(t))dt = \frac{1}{2},$$

which in turn implies that

$$\int_0^{(k+1)h} R_h(t)dt = \sum_{j=0}^{k-1} \int_{jh}^{(j+1)h} R_h(t)dt$$

$$= \sum_{j=0}^{k-1} h \left( \mathcal{H}(\rho_h^j|\pi) - \mathcal{H}(\rho_h^{j+1}|\pi) + \delta_h^{j+1} \right) + \sum_{j=0}^{k-1} h\delta_h^{j+1}$$

$$\leq h \left( \mathcal{H}(\rho_h^0|\pi) - \mathcal{H}(\rho_h^{k+1}|\pi) + \Delta_h^{k+1} \right) + h\Delta_h^{k+1}$$

$$\leq h \left( \mathcal{H}(\rho_h^0|\pi) + 2\Delta_h^n \right).$$

Let $(\gamma_r^j)_{j=0}^m$ denote a trajectory corresponding to another time-step $r$, and define the quantities $\gamma_r(s)$, $\gamma^r(s)$, $\ell_r(s)$, $\mathcal{H}_r(s)$ and $R_r(s)$ analogously to those defined in terms of $h$. Define

$$W^\pi_{h,r}(t,s) = (1 - \ell_r(s))W^\pi_{h}(t, \gamma_r(s)) + \ell_r(s)W^\pi_{h}(t, \gamma^r(s)),$$

and observe that this function is continuous in $t$ and $s$.

Lemma C.1.4. For any $n, m \geq 1$, $h, r > 0$ and $t \in [0, \min\{hn, rm\}]$,

$$W^{\pi}_{h,r}(t,t) \leq W^{\pi}_{h}(\rho_h^0, \gamma^0_r) + \int_0^t R_h(t) + R_r(t)dt. \quad (C.5)$$
Proof. Let \( s \in \left[ o, rm \right] \) and \( t \in \left[ o, hn \right] \setminus \{kh : k = o, \ldots, n\} \). By Lemma C.1.2,

\[
\frac{\partial}{\partial t} \frac{1}{2} \mathcal{W}_{h,r}^2(t, s) + \mathcal{H}_h(t) - \mathcal{H}_r(s) \leq \frac{1}{2} R_h(t).
\]

Similarly, for \( s \in \left[ o, rm \right] \setminus \{jr : j = o, \ldots, m\} \) and \( t \in \left[ o, hn \right] \setminus \{kh : k = o, \ldots, n\} \),

\[
\frac{\partial}{\partial s} \frac{1}{2} \mathcal{W}_{r,h}^2(s, t) + \mathcal{H}_r(s) - \mathcal{H}_h(t) \leq \frac{1}{2} R_r(s).
\]

Note the symmetry

\[
\mathcal{W}_{h,r}^2(t, s) = \mathcal{W}_{r,h}^2(s, t),
\]

so that for \( s \in \left[ o, rm \right] \setminus \{jr : j = o, \ldots, m\} \) and \( t \in \left[ o, hn \right] \setminus \{kh : k = o, \ldots, n\} \),

\[
\frac{\partial}{\partial t} \mathcal{W}_{h,r}^2(t, s) + \frac{\partial}{\partial s} \mathcal{W}_{r,h}^2(t, s) \leq R_h(t) + R_r(s),
\]

by adding the inequalities above. Setting \( s = t \) and letting \( t \in \left[ o, \min\{hn, rm\} \right] \setminus (\{kh : k = o, \ldots, n\} \cup \{jr : j = o, \ldots, m\}) \),

\[
\frac{d}{dt} \mathcal{W}_{h,r}^2(t, t) \leq R_h(t) + R_r(t).
\]

Since \( t \mapsto \mathcal{W}_{h,r}^2(t, t) \) is continuous and piecewise differentiable, the Fundamental Theorem of Calculus implies that

\[
\mathcal{W}_{h,r}^2(t, t) \leq \mathcal{W}_{h,r}^2(o, o) + \int_o^t R_h(t) + R_r(t)dt
\]

\[
= \mathcal{W}_{h}^2(p_h^o, \gamma_r^o) + \int_o^t R_h(t) + R_r(t)dt.
\]

\[\square\]

**Lemma C.1.5.** For any \( n, m \geq 1, h, r > o \) and \( t \in \left[ o, \min\{hn, rm\} \right] \),

\[
\mathcal{W}_{h}^2(p_h^o(t), \gamma_r(t))
\]

\[
\leq 6 \left[ \mathcal{W}_{h}^2(p_h^o, \gamma_r^o) + h (\mathcal{H}(p_h^o|\pi) + \Delta_h^o) + r (\mathcal{H}(\gamma_r^o|\pi) + \Delta_r^m) \right].
\]

211
Proof. Suppose \( j \) and \( k \) are such that \( t \in [kh, (k + 1)h] \cap [jr, (j + 1)r) \). Then,

\[
\begin{align*}
\mathcal{W}_2^z(\rho^k(t), \gamma^r(t)) &= \mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r) \\
&= (1 - \ell_h(t))(1 - \ell_r(t))\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r) \\
&\quad + (1 - \ell_h(t))\ell_r(t)\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r) \\
&\quad + \ell_h(t)(1 - \ell_r(t))\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r) \\
&\quad + \ell_h(t)\ell_r(t)\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r)
\end{align*}
\]

\[
\leq 3(1 - \ell_h(t))(1 - \ell_r(t))\left[ \mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) + \mathcal{W}_2^z(\rho^{k}_h, \gamma^{r}_r) + \mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r) \right] \\
+ 2(1 - \ell_h(t))\ell_r(t)\left[ \mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) + \mathcal{W}_2^z(\rho^{k}_h, \gamma^{r+1}_r) \right] \\
+ 2\ell_h(t)(1 - \ell_r(t))\left[ \mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r) + \mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{r}_r) \right] \\
+ \ell_h(t)\ell_r(t)\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r)
\]

\[
\leq 3(1 - \ell_h(t))(1 - \ell_r(t))\left[ \mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) + \mathcal{W}_2^z(\rho^{k}_h, \gamma^{r}_r) + \mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r) \right] \\
+ 3(1 - \ell_h(t))\ell_r(t)\left[ \mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) + \mathcal{W}_2^z(\rho^{k}_h, \gamma^{r+1}_r) \right] \\
+ 3\ell_h(t)(1 - \ell_r(t))\left[ \mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r) + \mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{r}_r) \right] \\
+ 3\ell_h(t)\ell_r(t)\mathcal{W}_2^z(\rho^{k+1}_h, \gamma^{j+1}_r)
\]

\[
= 3\mathcal{W}_{h,r}^z(t, t) + 3(1 - \ell_h(t))\mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) + 3(1 - \ell_r(t))\mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r).
\]

Now, by Lemmas C.1.4 and C.1.3,

\[
\mathcal{W}_{h,r}^z(t, t) \leq \mathcal{W}_2^z(\rho^0_h, \gamma^0_r) + \int_0^t R_h(t) + R_r(t)dt \\
\leq \mathcal{W}_2^z(\rho^0_h, \gamma^0_r) + h \left( \mathcal{H}(\rho^0_h|\pi) + 2\Delta^0_h \right) + r \left( \mathcal{H}(\gamma^0_r|\pi) + 2\Delta^m_r \right).
\]

Lastly, we know by Lemma C.1.1 that

\[
\mathcal{W}_2^z(\rho^{k+1}_h, \rho^{k}_h) \leq 2h \left( \mathcal{H}(\rho^0_h|\pi) + \Delta^0_h \right),
\]

\[
\mathcal{W}_2^z(\gamma^{j+1}_r, \gamma^{r}_r) \leq 2r \left( \mathcal{H}(\gamma^0_r|\pi) + \Delta^m_r \right).
\]
In conclusion, and without optimizing the constant, we get

\[
\mathcal{W}_2^n (\rho^h(t), \gamma^r(t)) \
\leq 6 \left[ \mathcal{W}_2^n (\rho_h^o, \gamma_o^r) + h \left( \mathcal{H} (\rho_h^o | \pi) + \Delta_h^n \right) + r \left( \mathcal{H} (\gamma_o^r | \pi) + \Delta_r^n \right) \right].
\]

Before giving its proof, we restate the main theorem of the paper:

**Theorem 1.** Let \((\rho^{hn}(t))_{m \geq 1}\) be a sequence of discrete solutions generated from \(\rho_o\), such that \(h_m \Delta_h^n \to 0\) and \(h_mm \to T\) for some \(T > 0\), as \(m \to \infty\). Then, \(\rho^{hn}(t)\) converges uniformly on \([0, T]\) to \(\rho(t)\), the gradient flow of \(\rho \mapsto \mathcal{H}(\rho | \pi)\) started from \(\rho_o\) as \(m \to \infty\). Moreover, if \(h > 0\) and \(n \geq 1\) are such that \(hn \leq T\), then for any \(t \in [0, hn]\),

\[
\mathcal{W}_2^n (\rho^h(t), \rho(t)) \leq \sqrt{6h \left( \mathcal{H} (\rho_o | \pi) + \Delta_h^n \right)}.
\]

**Proof.** Let the discrete solutions \(\rho^{hn}(t)\) and \(\rho^{hn}(t)\) be members of the sequence. From Lemma C.1.5, we know that \(\mathcal{W}_2^n (\rho^{hn}(t), \rho^{hn}(t)) \to 0\) as \(m, n \to \infty\), for any \(t \in [0, T]\).

This implies that \((\rho^{hn}(t))_{m \geq 1}\) is a Cauchy sequence. Since \((\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)\) is complete, this means that the sequence converges to a function \(\rho(t)\). Since the bound in Lemma C.1.5 does not depend on \(t\), this convergence is uniform on \([0, T]\).

Since the convergence is uniform and \(\rho^{hn}(t)\) is left continuous, then so is the limit \(\rho(t)\).

Moreover, since if \(t \in [kh, (k+1)h)\) for some \(k = 0, \ldots, n - 1\),

\[
\mathcal{W}_2^n (\rho^{hn}(t), \rho^{hn}(t)) \leq \mathcal{W}_2^n (\rho_{hn}^{k+1}, \rho_{hn}^{k}) \leq 2hn \left( \mathcal{H} (\rho_o | \pi) + \Delta_h^n \right) \to 0 \quad \text{as } n \to \infty.
\]

Hence, \(\rho_{hn}(t)\) converges to \(\rho(t)\) in the same manner as \(\rho^{hn}(t)\), meaning that the limit \(\rho(t)\) is right continuous also. Combining these facts, it is clear that \(\rho(t)\) is continuous.

Similarly,

\[
\mathcal{W}_2^n (\rho_{hn}(t), \rho_{hn}^{k+1/2}(t)) \leq \mathcal{W}_2^n (\rho_{hn}^{k+1/2}, \rho_{hn}^{k}) \leq 2hn \left( \mathcal{H} (\rho_o | \pi) + \Delta_h^n \right) \to 0 \quad \text{as } n \to \infty,
\]

by the bound in (C.4). This implies that \(\rho_{hn}^{h_k}(t)\) converges to \(\rho(t)\) in the same manner as \(\rho_{hn}(t)\) and \(\rho^{hn}(t)\).
It remains to show that $\rho(t)$ is the gradient flow of $\rho \mapsto \mathcal{H}(\rho|\pi)$. Indeed, let $f \in C^\infty_\mathbb{R}((0, \infty);\mathbb{R})$ be non-negative and $\nu \ll \mu_{\text{Leb}}$. Note that $\lim_{n \to \infty} \mathcal{W}_h^\nu_n(t, \nu) = \mathcal{W}_h^\nu(t, \nu)$ uniformly on $[0, T]$. Since $t \mapsto \mathcal{W}_h^\nu_n(t, \nu)$ is continuous, so is the limit $\mathcal{W}_h^\nu(t, \nu)$. Thus, $t \mapsto f(t)\mathcal{W}_h^\nu_n(t, \nu)$ is continuous, i.e. integrable, on $[0, T]$. The continuity of $f'$ implies that there exists an $M > 0$ such that $|f'(t)| \leq M$. In combination with the aforementioned uniform convergence, we know that

$$\lim_{n \to \infty} \int_0^T f(t)\mathcal{W}_h^\nu_n(t, \nu)dt = \int_0^T f(t)\mathcal{W}_h^\nu(t, \nu)dt.$$  

By the same reasoning, and the fact that $\lim_{n \to \infty} \mathcal{W}_h^\nu_n(t, \nu) = \mathcal{W}_h^\nu(t, \nu)$ uniformly on $[0, T]$, we have

$$\lim_{n \to \infty} \int_0^T f(t)\mathcal{W}_h^\nu_n(t, \nu)dt = \int_0^T f(t)\mathcal{W}_h^\nu(t, \nu)dt.$$ 

Now, since $f$ and $\mathcal{H}(\cdot|\pi)$ are non-negative, so is the function $t \mapsto f(t)\mathcal{H}_h(t)$. Thus, by Fatou’s lemma,

$$\lim\inf_{n \to \infty} \int_0^T f(t)\mathcal{H}_h(t)dt \geq \int_0^T \lim\inf_{n \to \infty} f(t)\mathcal{H}_h(t)dt.$$  

By Lemma 2.8 of Clément and Maas (2011),

$$\int_0^T \lim\inf_{n \to \infty} f(t)\mathcal{H}_h(t)dt \geq \int_0^T f(t)\mathcal{H}(\rho(t)\pi)dt.$$  

So,

$$\int_0^T \left[ -f(t)\frac{1}{2}\mathcal{W}_h^\nu(t, \nu) + f(t)\frac{\lambda}{2}\mathcal{W}_h^\nu(t, \nu) + f(t)\mathcal{H}(\rho(t)\pi) \right]dt \geq \lim\inf_{n \to \infty} \int_0^T \left[ -f(t)\frac{1}{2}\mathcal{W}_h^\nu_n(t, \nu) + f(t)\frac{\lambda}{2}\mathcal{W}_h^\nu_n(t, \nu) + f(t)\mathcal{H}_h_n(t) \right]dt \geq \lim\inf_{n \to \infty} \int_0^T f(t)\frac{d}{dt}\mathcal{W}_h^\nu_n(t, \nu) + f(t)\frac{\lambda}{2}\mathcal{W}_h^\nu_n(t, \nu) + f(t)\mathcal{H}_h_n(t)dt$$  

214
\begin{align}
\leq \liminf_{n \to \infty} \int_0^T f(t) \left[ \frac{1}{2} R_{h_n}(t) + H(v|\pi) \right] dt
\tag{C.11}
\end{align}

\begin{align}
= \int_0^T f(t) H(v|\pi) dt + \liminf_{n \to \infty} \int_0^T f(t) \frac{1}{2} R_{h_n}(t) dt
\tag{C.12}
\end{align}

\begin{align}
\leq \int_0^T f(t) H(v|\pi) dt + \sup_{t \in [0, T]} \liminf_{n \to \infty} \int_0^T \frac{1}{2} R_{h_n}(t) dt
\tag{C.13}
\end{align}

\begin{align}
\leq \int_0^T f(t) H(v|\pi) dt + \sup_{t \in [0, T]} \liminf_{n \to \infty} \left[ \frac{1}{2} h_n \left( \mathcal{H}(\rho_v|\pi) + 2 \Delta_{h_n}^n \right) \right]
\tag{C.14}
\end{align}

\begin{align}
= \int_0^T f(t) H(v|\pi) dt.
\tag{C.15}
\end{align}

where (C.9) follows from (C.6) and (C.7), (C.10) follows by integration by parts, (C.11) follows by Lemma C.1.2, (C.13) follows by \( f \) being non-negative and continuous, and \( R_{h_n}(t) \geq 0 \), (C.14) follows by Lemma C.1.3, and (C.15) follows by the assumption. This concludes the proof that \( \rho(t) \) is indeed the gradient flow.

Now, fix \( h > 0 \) and \( n \geq 1 \) such that \( hn \leq T \). Then, for any \( m \geq 1 \),

\[ \mathcal{W}_2^n(\rho^h(t), \rho^{hm}(t)) \leq \left[ h (\mathcal{H}(\rho_v|\pi) + \Delta_{h_n}^n) + h_m (\mathcal{H}(\rho_v|\pi) + \Delta_{h_m}^m) \right], \]

for any \( t \in [0, \min\{hn, hm\}] \) by Lemma C.1.5. Taking \( m \to \infty \) yields the conclusion.

\[ \Box \]

C.2 Rates for \( \mathcal{W}_2(\rho_v, \pi) \) and \( \mathcal{H}(\rho_v|\pi) \)

In this section, we provide some heuristic support for the claim that one can often assume that \( \mathcal{H}(\rho_v|\pi) = \mathcal{O}(d) \) and \( \mathcal{W}_2(\rho_v, \pi) = \mathcal{O}(\sqrt{d}) \). These assumptions can also be shown to be hold for more general settings than those we consider below.

Let \( \rho_v(x) = Z^{-1} e^{-V_v(x)} \), and note that

\[ \mathcal{W}_2^n(\rho_v, \pi) = \inf_{\gamma \in \Gamma(\rho_v, \pi)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} ||x - y||^2 d\gamma(x, y) \]

\[ \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} ||x - y||^2 d\pi(x) d\rho_v(y) \]

\[ 215 \]
\[ = \int_{\mathbb{R}^d} \|x - \bar{x}\|^2 d\pi(x) + \int_{\mathbb{R}^d} \|y - \bar{y}\|^2 d\rho_o(y) + \|\bar{x} - \bar{y}\|^2, \]

and where \( \bar{x} \) and \( \bar{y} \) are the means of \( \pi \) and \( \rho_o \) respectively. The third term on the last line safely be assumed to be \( O(d) \). By Theorem 1 of Durmus and Moulines (2016a), the first term can be bounded by \( d/\lambda \) under the \( \lambda \)-strong convexity assumption. Under similar assumptions on \( \rho_o \), or e.g. assuming that \( V_o(x) = \sum_{i=1}^d V^i(x_i) \), one can also defend imposing a bound of \( O(d) \) for second term.

Secondly, one can easily support the assumption \( \mathcal{H}(\rho_o | \pi) = O(d) \) if both \( V_o(x) = \sum_{i=1}^d V^i(x_i) \) and \( V(x) = \sum_{i=1}^d V(x_i) \). A less restrictive condition is to assume that \( 0 \leq V(x) - V_o(x) \leq a\|x\|^2 + b \) for some \( a \geq 0 \) and \( b \in \mathbb{R} \) not dependent on \( d \). The first inequality is analogous to saying that \( \rho_o \) has heavier tails than \( \pi \), whereas the second inequality constrains exactly how much heavier these tails can be. Under this assumption, and using the proof of Lemma 3 of Dalalyan (2014), we can write

\[
\mathcal{H}(\rho_o | \pi) = \int_{\mathbb{R}^d} \log \left( \frac{\rho_o}{\pi} \right) d\rho_o \\
= \int_{\mathbb{R}^d} [V(x) - V_o(x)] d\rho_o + \log \left( \int_{\mathbb{R}^d} e^{V_o(x) - V(x)} d\rho_o \right) \\
\leq \int_{\mathbb{R}^d} (a\|x\|^2 + b) \ d\rho_o,
\]

by noting that \( e^{V_o(x) - V(x)} \leq 1 \) by the assumption. One can then proceed as in the last paragraph.

C.2.1 Gaussian initial distribution

Let \( x^* \) denote the minimum of \( V \), and let \( V_o(x) = \frac{\alpha}{2}\|x - \mu\|^2 + V(x^*) \) with \( \alpha < M(d) \), so that \( \rho_o \) is a Gaussian distribution. We focus on bounding \( \mathcal{H}(\rho_o | \pi) \), as bounding the Wasserstein distance can be done as in the previous section. Then, using strong convexity, (4.30) and (4.31),

\[ V(x) \leq V(x^*) + L(d)\|x - x^*\| + \nabla f(x^*)^\top (x - x^*) + \frac{M(d)}{2} \|x - x^*\|^2, \]
\[ V(x) \leq V(x^*) - L(d)\|x - x^*\| + \nabla f(x^*)^\top (x - x^*) + \frac{\lambda}{2} \|x - x^*\|^2, \]

so that

\[
\int_{\mathbb{R}^d} [V(x) - V_0(x)] \, d\rho_o = \int_{\mathbb{R}^d} \left[ \frac{M(d)}{2} \|x - x^*\|^2 - \frac{a}{2} \|x - \mu\|^2 + L(d)\|x - x^*\| + \nabla f(x^*)^\top (x - x^*) \right] \, d\rho_o \\
\leq \frac{M(d)}{2} \|\mu - x^*\|^2 + \frac{M(d)d}{2a} - \frac{ad}{2a} + \nabla f(x^*)^\top (\mu - x^*) + L(d) \left( \int_{\mathbb{R}^d} \|x - x^*\|^2 \, d\rho_o \right)^{1/2} \\
\leq \frac{M(d)}{2} \|\mu - x^*\|^2 + \frac{(M(d) - a)d}{2a} + \nabla f(x^*)^\top (\mu - x^*) + L(d) \left( \|\mu - x^*\|^2 + \frac{d}{a} \right)^{1/2},
\]

and

\[
\log \int_{\mathbb{R}^d} e^{V_o(x) - V(x)} \, d\rho_o \leq \log \left( \frac{1}{Z_{1/a}} \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x - x^*\|^2 + L(d)\|x - x^*\| - \nabla f(x^*)^\top (x - x^*)} \, dx \right) \\
\leq \log \left( \frac{1}{Z_{1/a}} \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x - x^*\|^2 + (L(d) + \|\nabla f(x^*)\|)\|x - x^*\|} \, dx \right) \\
= \log \left( \frac{1}{Z_{1/a}} \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x - x^*\|^2 + c \|x - x^*\|} \, dx \right),
\]

where \( c = L(d) + \|\nabla f(x^*)\| \) and \( Z_{1/a} = \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x - x^*\|^2} \, dx \). Furthermore,

\[
\log \left( \frac{1}{Z_{1/a}} \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x\|^2 + c \|x\|} \, dx \right) \leq \log \left( \frac{1}{Z_{1/a}} \int_{\mathbb{R}^d} e^{-\frac{d}{2} \|x\|^2 + \frac{d}{2} \|x\|} \, dx \right) \\
= \log \left( \frac{Z_{2/a}}{Z_{1/a}} e^{\frac{d}{2}} \right) \\
= \frac{d}{2} \log(2) + \frac{(L(d) + \|\nabla f(x^*)\|)^2}{\lambda} \\
\leq \frac{d}{2} \log(2) + \frac{(L(d) + M(d)\|x^* - x^d\|)^2}{\lambda},
\]
where $x^\dagger$ is the minimum of $f$. Hence,

\[
\mathcal{H}(\rho_\omega | \pi) \leq \frac{M(d)}{2} \|\mu - x^*\|^2 + \frac{(M(d) - a)d}{2a} + \|x^* - x^\dagger\| \|\mu - x^*\|
\]

\[
+ L(d) \left( \|\mu - x^*\|^2 + \frac{d}{a} \right)^{\frac{1}{2}} + \frac{d}{2} \log(2) + \frac{(L(d) + M(d)\|x^* - x^\dagger\|)^2}{\lambda}.
\]

Take $a = \lambda$ and $\mu$ such that $\|\mu - x^*\|^2 = \mathcal{O}(d)$, and make the safe assumption that $\|x^* - x^\dagger\|^2 = \mathcal{O}(d)$. If $M(d) = \mathcal{O}(1)$ and $L(d) = \sqrt{d}$ like in Section 4.4.2, we get $\mathcal{H}(\rho_\omega | \pi) = \mathcal{O}(d)$. 

---

218
Appendix to Chapter 5

D.1 Proofs

Recall the following proposition from the main text.

**Proposition D.1.1** (Proposition 5.2.1 of the main text). For any $\varepsilon > 0$, the iterative proportional fitting procedure returns a distribution $S^{(i)}$ satisfying $KL(\pi_0|s^{(i)}_0) + KL(\pi_T|s^{(i)}_T) \leq \varepsilon$ in fewer than $\left\lceil \frac{KL(S|Q)}{\varepsilon} \right\rceil$ iterations.

**Proof.** Recall that $S^{(2)} = Q^{(i)} \in \mathcal{P}_o(\pi_o)$ and $S^{(2i+1)} = \mathbb{P}^{(i+1)} \in \mathcal{P}_T(\pi_T)$, and that for any $\mathbb{H} \in \mathcal{P}_T(\pi_T)$,

$$KL(\mathbb{H}|G) = KL(\pi_T|g_T) + \int_E KL(H(dx_{0:T-1}|x_T)|G(dx_{0:T-1}|x_T)) \pi_T(dx_T).$$

Hence, $\mathbb{P}^{(i+1)} = \arg\min_{\mathbb{H} \in \mathcal{P}_T(\pi_T)} KL(\mathbb{H}|Q^{(i)})$ is such that $\mathbb{P}^{(i+1)}(dx_{0:T-1}|x_T) = Q^{(i)}(dx_{0:T-1}|x_T)$.
for $\pi_T$-almost every $x_T$. So, for any $H \in P_T(\pi_T)$,

$$
\text{KL}(H|S^{(2i)}) - \text{KL}(H|S^{(2i+1)}) = \text{KL}(\pi_T|s_T^{(2i)})
= \text{KL}(\pi_o|s_o^{(2i)}) + \text{KL}(\pi_T|s_T^{(2i)}),
$$

since $s_o^{(2i)} = \pi_o$. By analogous reasoning, we have that $Q^{(i)}(dx_{i:T}|x_o) = P^{(i)}(dx_{i:T}|x_o)$ for $\pi_o$-almost every $x_o$ and that for any $H \in P_o(\pi_o)$,

$$
\text{KL}(H|S^{(2i-1)}) - \text{KL}(H|S^{(2i)}) = \text{KL}(\pi_o|s_o^{(2i-1)})
= \text{KL}(\pi_o|s_o^{(2i-1)}) + \text{KL}(\pi_T|s_T^{(2i-1)}).
$$

Hence, for any $H \in P_{o,T}(\pi_o, T)$ we have that

$$
\text{KL}(H|S^{(o)}) - \text{KL}(H|S^{(k)}) = \sum_{i=1}^{k-1} \text{KL}(\pi_o|s_o^{(i)}) + \text{KL}(\pi_T|s_T^{(i)}).
$$

Let $\varepsilon > 0$. If $k^*$ is the first iteration such that $\text{KL}(\pi_o|s_o^{(i)}) + \text{KL}(\pi_T|s_T^{(i)}) \leq \varepsilon$, then

$$
\text{KL}(H|S^{(o)}) \geq \text{KL}(H|S^{(o)}) - \text{KL}(H|S^{(k^*)}) \geq k^*\varepsilon,
$$

from which it follows that

$$
k^* \leq \frac{\text{KL}(H|S^{(o)})}{\varepsilon} = \frac{\text{KL}(H|Q)}{\varepsilon}.
$$

In particular, $k^* \leq \text{KL}(S|Q)/\varepsilon$ since $S \in P_{o,T}(\pi_o, T)$, from which the result follows. \hfill \Box

**D.2 Linear Quadratic Gaussian**

Recall the setting introduced in Section 5.2.6, in which we set

$$
\pi_o(dx_o) = \mathcal{N}(x_o; \mu_o, \Sigma_o)dx_o \quad \text{and} \quad \pi_T(dx_T) = \mathcal{N}(x_T; \mu_T, \Sigma_T)dx_T,
$$
for some $\mu_0, \mu_T \in \mathbb{R}^d$ and $\Sigma_0, \Sigma_T \in \mathbb{R}^{d \times d}$, and for each $t \in [1 : T]$, the kernel
\[ M_t(x_{t-1}, dx_t) = \mathcal{N}(x_t; K_t x_{t-1} + r_t, H_t) dx_t \]
for some $r_t \in \mathbb{R}^d$ and $K_t, H_t \in \mathbb{R}^{d \times d}$. We derive the form of the exact policies $\psi^{(i)}$, which enables the comparison with the approximate method.

To find $\psi^{(i)}$ and the corresponding path measure $\mathbb{Q}_{\psi^{(i)}}$ defined by the IPFP iterations, we proceed by induction. Suppose that
\[ \mathbb{Q}_{\psi^{(i-1)}}(dx_{0:T}) = \pi_0(dx_0) \prod_{t=1}^T M_t^{\psi^{(i-1)}}(x_{t-1}, dx_t), \]
where
\[ M_t^{\psi^{(i-1)}}(x_{t-1}, dx_t) = \mathcal{N}(x_t; K_t^{(i-1)} x_{t-1} + r_t^{(i-1)}, H_t^{(i-1)}) dx_t. \]
The marginal distributions of $\mathbb{Q}_{\psi^{(i-1)}}$ are then Gaussian: For each $t \in [1 : T]$,
\[ q_t^{\psi^{(i-1)}}(dx_t) = \mathcal{N}(x_t; \mu_t^{(i-1)}, \Sigma_t^{(i-1)}) dx_t, \]
where $\mu_t^{(i-1)}$ and $\Sigma_t^{(i-1)}$ are given by the recursions
\[ \mu_t^{(i-1)} = K_t^{(i-1)} \mu_{t-1}^{(i-1)} + r_t^{(i-1)}, \quad \Sigma_t^{(i-1)} = H_t^{(i-1)} + K_t^{(i-1)} \Sigma_{t-1}^{(i-1)} (K_t^{(i-1)})^\top, \quad t \in [1 : T]. \]
In particular, this representation holds for $t = T$, which allows us to express $\psi_T^{(i)} = dx_T/dq_T^{\psi^{(i-1)}}$ on the form $- \log \psi_T^{(i)} = x_T^T A_T^{(i)} x_T + x_T^T b_T^{(i)} + c_T^{(i)}$, where
\[ A_T^{(i)} = \frac{1}{2} \left( \Sigma_T - (\Sigma_T^{(i-1)})^{-1} \right), \quad b_T^{(i)} = (\Sigma_T^{(i-1)})^{-1} \mu_T^{(i-1)} - \Sigma_T^{-1} \mu, \]
\[ c_T^{(i)} = \frac{1}{2} \left[ \mu_T^\top \Sigma_T^{-1} \mu_T - \mu_T^{(i-1)}^\top (\Sigma_T^{(i-1)})^{-1} \mu_T^{(i-1)} + \log \left( \det (\Sigma_T^{(i-1)})^{-1} \right) - \log \left( \det \Sigma_T^{-1} \right) \right]. \]
By induction on $t$ (while keeping $i$ fixed), one can show that $\psi_{t-1}^{(i)}(x_{t-1})$ can be written
\[ - \log \psi_{t-1}^{(i)}(x_{t-1}) = x_{t-1}^T A_{t-1}^{(i)} x_{t-1} + x_{t-1}^T b_{t-1}^{(i)} + c_{t-1}^{(i)}, \]

221
where $A_t^{(i)}, b_t^{(i)}$ and $c_t^{(i)}$ satisfy the backward recursions

$$A_{t-1}^{(i)} = \frac{1}{2} (K_t^{(i-1)})^T \left[ (H_t^{(i-1)})^{-1} - \frac{1}{2} (H_t^{(i-1)})^{-1} \left( A_t^{(i)} + \frac{1}{2} (H_t^{(i-1)})^{-1} \right)^{-1} (H_t^{(i-1)})^{-1} \right] K_t^{(i-1)},$$

$$b_{t-1}^{(i)} = (K_t^{(i-1)})^T (H_t^{(i-1)})^{-1} \left[ r_t^{(i-1)} - \frac{1}{2} (A_t^{(i)} + \frac{1}{2} (H_t^{(i-1)})^{-1})^{-1} \left( (H_t^{(i-1)})^{-1} - r_t^{(i-1)} - b_t^{(i)} \right) \right],$$

$$c_{t-1}^{(i)} = c_t^{(i)} + \frac{1}{2} r_t^{(i-1)} (H_t^{(i-1)})^{-1} r_t^{(i-1)}$$

$$- \frac{1}{4} \left( (H_t^{(i-1)})^{-1} - r_t^{(i-1)} - b_t^{(i)} \right)^T \left( A_t^{(i)} + \frac{1}{2} (H_t^{(i-1)})^{-1} \right)^{-1} \left( (H_t^{(i-1)})^{-1} - r_t^{(i-1)} - b_t^{(i)} \right),$$

for $t \in [1 : T]$, initialized at the $A_T^{(i)}, b_T^{(i)}$ and $c_T^{(i)}$ given above.

The updated Markov kernels are given by

$$M_t^{(i)}(x_{t-1}, dx_t) = \mathcal{N}(x_t; K_t^{(i)} x_{t-1} + r_t^{(i)}, H_t^{(i)}) \, dx_t,$$

where the updated parameters $K_t^{(i)}, H_t^{(i)}$ and $r_t^{(i+1)}$ satisfy

$$H_t^{(i)} = \left( (H_t^{(i-1)})^{-1} + 2 A_t^{(i)} \right)^{-1}, \quad K_t^{(i)} = \left( (H_t^{(i-1)})^{-1} + 2 A_t^{(i)} \right)^{-1} (H_t^{(i-1)})^{-1} K_t^{(i-1)},$$

$$r_t^{(i)} = \left( (H_t^{(i-1)})^{-1} + 2 A_t^{(i)} \right)^{-1} \left( (H_t^{(i-1)})^{-1} - r_t^{(i-1)} - b_t^{(i)} \right).$$

The next IPFP iterates can then be expressed

$$\mathbb{P}^{(i)}(dx_{0:T}) = p^{(i)}_0(dx_0) \prod_{t=1}^T M_t^{(i)}(x_{t-1}, dx_t),$$

$$\mathbb{Q}^{(i)}(dx_{0:T}) = p_0(dx_0) \prod_{t=1}^T M_t^{(i)}(x_{t-1}, dx_t),$$

where $p_0^{(i)}(dx_0) = \mathcal{N}(x_0; \gamma^{(i)}, \Gamma^{(i)}) \, dx_0$, with

$$\Gamma^{(i)} = \left( \Sigma_0^{-1} + 2 A_0^{(i)} \right)^{-1}, \quad \gamma^{(i)} = \left( \Sigma_0^{-1} + 2 A_0^{(i)} \right)^{-1} (\Sigma_0^{-1} \mu_0 - b_0^{(i)}).$$

222
The potential function $a^{(i)}$ defined by (??) can be written

$$- \log a^{(i)}(x_o) = - \log a^{(i-1)}(x_o) + x_o^\top A_a^{(i)} x_o + x_o^\top b_a^{(i)} + c_a^{(i)},$$

where

$$A_a^{(i)} = \frac{1}{2} \left( \Sigma_o^{-1} - (\Gamma^{(i)})^{-1} \right) = -A_o^{(i)}, \quad b_a^{(i)} = (\Gamma^{(i)})^{-1} \gamma^{(i)} - \Sigma_o^{-1} \mu_o = -b_o^{(i)},$$

$$c_a^{(i)} = \frac{1}{2} \left[ \mu_o^\top \Sigma_o^{-1} \mu_o - \gamma^{(i)}^\top (\Gamma^{(i)})^{-1} \gamma^{(i)} + \log \left( \det(\Gamma^{(i)})^{-1} \right) - \log \left( \det \Sigma_o^{-1} \right) \right]$$

In other words, $- \log a^{(i)}(x_o) = x_o^\top F_a^{(i)} x_o + x_o^\top v_a^{(i)} + d_a^{(i)}$, where

$$F_a^{(i)} = \sum_{j=1}^{i} A_a^{(j)}, \quad v_a^{(i)} = \sum_{j=1}^{i} b_a^{(j)}, \quad d_a^{(i)} = \sum_{j=1}^{i} c_a^{(j)}.$$

Similarly, the potential $\beta^{(i)}$ defined by (??) can be written

$$- \log \beta^{(i)}(x_T) = - \log \beta^{(i-1)}(x_T) + x_T^\top A_T^{(i)} x_T + x_T^\top b_T^{(i)} + c_T^{(i)}$$

$$= x_T^\top F_\beta^{(i)} x_T + x_T^\top v_\beta^{(i)} + d_\beta^{(i)},$$

where

$$F_\beta^{(i)} = \sum_{j=1}^{i} A_T^{(j)}, \quad v_\beta^{(i)} = \sum_{j=1}^{i} b_T^{(j)}, \quad d_\beta^{(i)} = \sum_{j=1}^{i} c_T^{(j)}.$$


Bibliography


227


234


