Deriving Indistinguishability from Unpredictability: Tools and Applications in Pseudorandomness

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Deriving Indistinguishability from Unpredictability: Tools and Applications in Pseudorandomness"

presented by: Rohit Agrawal

candidate for the degree of Doctor of Philosophy and here by certify that it is worthy of acceptance.

Signature _____
Typed name: Professor S. Vadhan

Signature _____
Typed name: Professor B. Barak

Signature _____
Typed name: Professor F. du Pin Calmon

Signature _____
Typed name: Professor C. Dwork

June 18, 2020

Deriving Indistinguishability from Unpredictability: Tools and Applications in Pseudorandomness

a dissertation presented

by

Rohit Agrawal

to

The Department of Computer Science

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

June 2020

Dissertation Advisor: Professor Salil P. Vadhan Rohit Agrawal

Deriving Indistinguishability from Unpredictability: Tools and Applications in Pseudorandomness

Abstract

Proving that a distribution P is "close to uniform" is an integral part of many problems in pseudorandomness, and is often defined either in terms of *indistinguishability*—no algorithm (possibly required to be efficient) should be able to distinguish between P and the uniform distribution, or *unpredictability*—the distribution P should have high entropy, or be unpredictable by any (efficient) algorithm. In most cases, the application will require the former type of guarantee, although the latter can sometimes be easier to reason about. In this thesis, we develop tools to relate these notions, and apply information-theoretic reasoning to problems in complexity theory and cryptography:

- We extend the definition of randomness extractors to allow the error to be measured in terms of an arbitrary distance measure, and extend the connection between extractors and averaging samplers (Zuckerman, Rand. Struct. Alg.'97) to an arbitrary family $\mathcal{F}$ of test functions and the integral probability metric defined by $\mathcal{F}$. Using this connection, we show that extractors for the Kullback–Leibler (KL) divergence are subgaussian samplers as defined by Błasiok (SODA'18). By showing that KL extractors exist with essentially the same parameters as standard extractors (explicitly and non-explicitly), we construct the first explicit subgaussian samplers matching the best known constructions of averaging samplers for $[0, 1]$-bounded functions in the parameter regime where the approximation error ε and failure probability δ are subconstant.

- We introduce *hardness in relative entropy*, a new notion of hardness for search problems which

on the one hand is satisfied by all one-way functions and on the other hand implies both *next-block pseudoentropy* and *inaccessible entropy*, two forms of computational entropy used in recent constructions of pseudorandom generators and statistically hiding commitment schemes, respectively, thereby shedding light on the apparent "duality" between them.

- We show that the moment generating function of the KL divergence between the empirical distribution of n independent samples from a distribution P over a finite alphabet of size k (i.e. a multinomial distribution) and P itself is no more than that of a gamma distribution with shape $k-1$ and rate n. The resulting exponential concentration inequality becomes meaningful (less than 1) when the divergence ε is larger than $(k-1)/n$, whereas the standard method of types bound requires $\varepsilon > \frac{1}{n} \cdot \log \binom{n+k-1}{k-1} \geq (k-1)/n \cdot \log(1+n/(k-1))$, thus saving a factor of order $\log(n/k)$ in the standard regime of parameters where $n \gg k$.

- We systematically study the relationship between f-divergences and integral probability metrics (IPMs) from the perspective of convex duality. Starting from a tight variational representation of the f-divergence, we derive a generalization of the moment generating function, which we show exactly characterizes the best lower bound of the f-divergence as a function of a given IPM. Using this characterization, we obtain new bounds while also recovering in a unified manner well-known results, such as Hoeffding's lemma, Pinsker's inequality and its extension to subgaussian functions, and the Hammersley–Chapman–Robbins bound. The variational representation also allows us to prove new results on topological properties of the divergence which may be of independent interest.

Contents

Acknowledgments

Most importantly, I would like to thank my advisor Salil Vadhan for his support and mentorship throughout my graduate studies. From the moment I took his pseudorandomness class my first semester and got hooked, Salil has been a fountain of knowledge and enthusiasm for the subject, as well as computer science and mathematics more broadly. I'm also particularly grateful that he gave me the opportunity to investigate any rabbit-hole I found interesting and to develop my own taste for research, while always being there with a helpful and clarifying suggestion or idea.

Along this journey, I was very fortunate to have the opportunity to collaborate with and learn from Yi-Hsiu Chen, Chi-Ning Chou, Thibaut Horel, Christina Ilvento, Preetum Nakkiran, Vladimir Sotirov, and Salil Vadhan. I was also lucky enough to have conversations with Jarosław Błasiok and Muthuramakrishnan Venkitasubramaniam which led to several of the chapters in this thesis. The presentation of the works on which this thesis is based have been greatly improved by the suggestions of Jarosław Błasiok, Flavio du Pin Calmon, Ben Edelman, Julien Fageot, Salil Vadhan, and the anonymous referees who reviewed preliminary versions. I am also grateful to the Harvard computer science community for making my time here so much more rewarding, and especially to the faculty who took the time and effort to serve on my thesis and/or qualifying committees: Boaz Barak, Flavio du Pin Calmon, Yiling Chen, Cynthia Dwork, Salil Vadhan, and Leslie Valiant. I would also like to thank

Allison Choat for making administrative tasks as smooth as possible.

Finally, I want to thank my family for their constant support and encouragement.

The research in this thesis was funded by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, and the work on which Chapter 4 is based was also supported in part by NSF grant CCF-1763299 to Salil Vadhan.

Chapter 1

Introduction

1.1 Background: measuring the distance between distributions

A basic question in statistics is to understand what it means for two distributions to be similar, with the special case of one of the distributions being uniform over a finite set holding particular importance in discrete settings. Perhaps the most common answer to this question, especially in computer science, defines distance in terms of *distinguishability by test functions*, meaning that two distributions P and Q on a finite set Ω are said to have distance

$$d_{\mathcal{F}}(P, Q) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \right|$$

for some class of *test functions* $\mathcal{F}$ from Ω to the real numbers (indeed, usually to $[0, 1]$). Such distance measures, a special case of the *integral probability metrics (IPMs)* [Mül97; Zol84], are extremely operational in nature: they quantify the worst-possible deviation of running a procedure (a function $f \in \mathcal{F}$) on a distribution P instead of a "true" distribution Q. Particular examples include the *total-variation distance* (also called the *statistical distance*), obtained by taking $\mathcal{F}$ to be all functions into the set $\{0, 1\}$

(equivalently the set $[0,1]$) and corresponds to all possible decision procedures, and the standard definition of computational indistinguishability [GM84; Yao82] which takes $\mathcal{F}$ to be the set of all functions computable by polynomial-sized[1] circuits. Unfortunately, the operational nature of IPMs comes at a cost: though they capture well the desired behavior of the *output* of an algorithm or protocol, they are often unsuitable for use in the intermediate analysis of a protocol:

Example 1. Suppose there is a secret random string $X \in \{0,1\}^n$, and the value $f(X)$ is "leaked" for a function $f : \{0,1\}^n \to \{0,1\}$ with $\Pr[f(U_n) = 1] = 1/2$—intuitively the secret X is still mostly random even given the value of $f(X)$, but from the perspective of total variation it is far from uniform, as

$$d_{\mathrm{TV}}\big((f(X), X), (f(X), U_n)\big) = \frac{1}{2}\,. \tag{1.1}$$

Note that a distance of $1/2$ is within a constant factor of the maximal possible distance 1 and that cryptographic protocols usually require error at most $n^{-\omega(1)}$, so that the bound of Eq. (1.1) is almost meaningless.

To better capture the intuition that X is still close to uniform given $f(X)$ in the above example, one can use an alternative measure of the randomness of a random variable based on *unpredictability*, most notably the *(Shannon) entropy* [Sha48] foundational to information theory:

$$\mathrm{H}(P) = \sum_{x \in \Omega} P(x) \lg \frac{1}{P(x)}\,.$$

Recall that the entropy of a random variable over a finite set achieves its maximal possible value of[2] $\lg|\Omega|$ if and only if it is uniformly distributed, and its minimum possible value of 0 if and only if it takes

[1] Here we employ the standard computer science abuse of notation, where asymptotic notation in the context of a single object implicitly refers to a family of objects parametrized by the natural parameter, in this case $\lg|\Omega|$.

[2] To avoid confusion, in this work we avoid the notation log and instead write lg and ln for the binary and natural logarithms respectively.

exactly one value (i.e. is deterministic), so that the entropy can also be thought of as a measure of distance to uniform—explicitly, the quantity $\lg|\Omega| - \mathrm{H}(P)$ is the special case of the Kullback–Leibler (KL) divergence [KL51]

$$\mathrm{KL}(P \parallel Q) = \sum_{x \in \Omega} P(x) \lg \frac{P(x)}{Q(x)}$$

where Q is the uniform distribution.

Entropy of course plays a central role in information theory and coding theory, but it also has found many uses in complexity theory and cryptography. For instance, in the stylized setting of Example 1, the (conditional) entropy of X given $f(X)$ is

$$\mathrm{H}(X \mid f(X)) = \mathrm{H}(X) - \mathrm{H}(f(X)) = n - 1,$$

or equivalently, the (conditional) KL divergence is

$$\mathrm{KL}\big(X \mid f(X) \parallel U_n \mid f(X)\big) = n - (n-1) = 1,$$

which is much smaller than the maximal possible value n, thus capturing the fact that X remains unknown given $f(X)$. As a result, the *leftover hash lemma* [McI87; BBR88; ILL89] in cryptography implies[3] that applying a universal hash function $h : \{0,1\}^n \to \{0,1\}^{n-\omega(\lg n)}$ to X gives output statistically indistinguishable from uniform, i.e. with

$$d_{\mathrm{TV}}\Big(\big(h, f(X), h(X)\big), \big(h, f(X), U_{n-\omega(\lg n)}\big)\Big) \leq n^{-\omega(1)}\,. \tag{1.2}$$

Thus, by arguing via an entropy-type notion, we were able to obtain $n^{-\omega(1)}$ closeness in total variation distance in Eq. (1.2), whereas the direct analysis in Eq. (1.1) led to the much weaker bound of $1/2$. In particular, we see that entropy-type notions are useful in arguing about random variables, even if the

[3] Technically, this requires that the conditional *Rényi (collision) entropy* is $n - 1$, which it is in this case.

desired end result requires a guarantee formulated in terms of an IPM.

1.2 This work

In this work, we extend the reach of the information-theoretic world of entropy and Kullback–Leibler divergence, reducing the need to argue about IPMs directly for problems in pseudorandomness. We give applications to complexity theory and cryptography, and develop statistical tools to further this connection.

1.2.1 Applications

Complexity Theory (Chapter 2, [Agr19])

Averaging samplers, introduced by Bellare and Rompel [BR94] and an important object of study in pseudorandomness, are algorithms which use a short random seed to produce correlated samples from a universe $\{0,1\}^m$ with the property that for every $f : \{0,1\}^m \to [0,1]$, the mean of f on the samples is with high probability close to the true mean of f on all of $\{0,1\}^m$. Such samplers are usually used in the context of randomness-efficient error reduction for algorithms or protocols, where the function f encodes the acceptance or success probability. However, for an application in streaming algorithms, Błasiok [Bła19] recently introduced the notion of a *subgaussian sampler*, defined as an averaging sampler for approximating the mean of functions $f : \{0,1\}^m \to \mathbb{R}$ such that $f(U_m)$ has subgaussian tails, and asked for explicit constructions.

In Chapter 2, based on [Agr19], we give the first explicit constructions of subgaussian samplers (and in fact averaging samplers for the broader class of subexponential functions) that match the best known constructions of averaging samplers for $[0,1]$-bounded functions in the regime of parameters where

the approximation error ε and failure probability δ are subconstant. Our constructions are established via an extension of the standard notion of *randomness extractor* [NZ96], where the error is measured by an arbitrary distance measure rather than total variation distance, and a generalization of Zuckerman's equivalence [Zuc97] between extractors and samplers to arbitrary integral probability metrics and their defining family $\mathcal{F}$. Once recast in the extractor language, we use a result of Boucheron, Lugosi, and Massart [BLM13, §4.9], which shows that the indistinguishability notion of "subgaussian distance" is controlled by a much simpler unpredictability-type notion, the Kullback–Leibler divergence. We thus further develop a framework of KL-extractors, which are stronger than both standard extractors and subgaussian samplers, but we show that they exist with essentially the same parameters as standard extractors, with regards to both explicit constructions and (using a result from Chapter 4) optimal non-explicit constructions.

Cryptography (Chapter 3, [ACHV19])

Since one-way functions (OWFs) [DH76] are the minimal assumption for complexity-based cryptography [IL89], it is of interest to simplify and improve the parameters of the constructions of basic cryptographic primitives from OWFs, such as pseudorandom generators (PRGs) [BM82; Yao82; HILL99], universal-one way hash functions (UOWHFs) [NY89; Rom90], and statistically hiding commitments (SHCs) [BCC88; HNORV09]. A recent line of work [HRVW09; HRV13; HHRVW10; VZ12] has improved all of these constructions using computational versions of entropy: the first step of each of these works is to argue that every one-way function $f : \{0,1\}^n \to \{0,1\}^n$ has a gap between the true entropy of the sequence $(f(X)_1, \ldots, f(X)_n, X_1, \ldots, X_n)$, which is simply n, and some form of computational entropy. Specifically, for PRGs, one argues that the *next-block pseudoentropy* [HRV13; VZ12] is at least $n + \omega(\lg n)$, and for SHCs one argues that the *next-block accessible entropy* [HRVW09]

is at most $n - \omega(\lg n)$.[4] Perhaps surprisingly, once this first step has been accomplished, the remaining parts of the constructions of PRGs, SHCs, and UOWHFs follow a very similar sequence of steps, suggesting the intriguing possibility that the constructions or proofs might be unified.

In Chapter 3, based on joint work with Yi-Hsiu Chen, Thibaut Horel, and Salil Vadhan [ACHV19], we make partial progress towards a unified proof by introducing *hardness in relative entropy*, a new unpredictability-type notion of hardness for search problems which on the one hand is easily shown to be satisfied by all one-way functions, and on the other hand implies gaps between the true entropy and both next-block pseudoentropy and next-block accessible entropy. Furthermore, the proof that it yields a gap in next-block accessible entropy, similar in structure to the proof that one-way functions imply a gap in next-block pseudoentropy [VZ12], is primarily information-theoretic and isolates the part of the argument involving bounded adversaries, thereby simplifying and slightly strengthening the original proof of [HHRVW10], which used a more indistinguishability-style proof directly relating bounded and unbounded adversaries.

1.2.2 Tools

Concentration Inequalities (Chapter 4, [Agr20])

Understanding the rate of convergence of an empirical distribution obtained from samples from an unknown distribution P to the true distribution is a basic problem in statistics and learning. Furthermore, computer science applications, it is often important to have good finite-sample (rather than asymptotic) bounds, as for example is needed in establishing the existence of randomness extractors with optimal parameters (as described in the complexity theory section above). In the case of a finite alphabet of

[4] For UOWHFs, a different notion of accessible entropy is used [HHRVW10], and integrating it with the others is an interesting open problem.

size k and number of samples n, this is equivalent to asking for tail bounds on the random variable

$$V_d = d\left(\left(\frac{X_1}{n}, \dots, \frac{X_k}{n}\right), (p_1, \dots, p_k)\right)$$

where d is some distance measure, and the randomness is taken over the draw of $(X_1, \dots, X_k)$ from the multinomial distribution with n samples and probabilities $(p_1, \dots, p_k)$.

When d is the total variation distance, this is well-understood: we have $d_{\mathrm{TV}}(P, Q)^2 \lesssim \chi^2(P \parallel Q)$ by Jensen's inequality, so that $\mathbb{E}\left[V_{d_{\mathrm{TV}}}\right] \lesssim \sqrt{\mathbb{E}\left[V_{\chi^2}\right]} = \sqrt{(k-1)/n}$, and since the total variation is the IPM defined by the 2^k functions $f : [k] \to \{0, 1\}$, the Chernoff and union bounds imply the exponential concentration inequality $\Pr\left[V_{d_{\mathrm{TV}}} \geq \varepsilon\right] \leq 2^{k - 2n\varepsilon^2/\ln 2}$ which decays exponentially in $n\varepsilon^2$ once $\varepsilon \gtrsim \sqrt{k/n}$, which is of the same order as the expectation[5].

However, when d is the Kullback–Leibler divergence, our understanding is more rudimentary. As with the total variation distance, the expectation can be bounded since $\mathrm{KL}(P \parallel Q) \leq \lg\left(1 + \chi^2(P \parallel Q)\right)$, and thus $\mathbb{E}[V_{\mathrm{KL}}] \leq \lg\left(1 + \frac{k-1}{n}\right) \lesssim (k-1)/n$ [Pan03]. Since Pinsker's inequality implies that $d_{\mathrm{TV}}(P, Q) \lesssim \sqrt{\mathrm{KL}(P \parallel Q)}$, one would hope that V_{KL} concentrates like $V_{d_{\mathrm{TV}}}^2$, that is, that the tail probability of V_{KL} decays exponentially in $n\varepsilon$ for $\varepsilon \gtrsim (k-1)/n$. However, the standard tail bound, based on the method of types [Csi98], states that $\Pr[V_{\mathrm{KL}} \geq \varepsilon] \leq \binom{n+k-1}{k-1} \cdot 2^{-n\varepsilon}$, which only falls below 1 once $\varepsilon \gtrsim \frac{1}{n} \cdot \lg\binom{n+k-1}{k-1} \gtrsim (k-1)/n \cdot \lg(1 + n/(k-1))$, which is off from the expectation by a factor $O(\lg(n/k))$. Though the method of types bound has since been improved by Mardia et al. [MJTNW19], the bound still only becomes meaningful once $\varepsilon \gtrsim (k-1)/n \cdot \lg(1 + n/(k-1))$.

In Chapter 4, based on [Agr20], we give the first tail bound on V_{KL} that decays exponentially in $n\varepsilon$ once $\varepsilon \gtrsim (k-1)/n$, by showing that the moment generating function of V_{KL} is no more than that of

[5] In fact, by McDiarmid's inequality we have $\Pr\left[V_{d_{\mathrm{TV}}} \geq \mathbb{E}\left[V_{d_{\mathrm{TV}}}\right] + \varepsilon\right] \leq 2^{-2n\varepsilon^2/\ln 2}$.

a gamma distribution with shape $k-1$ and rate n.[6] As discussed earlier, we use these results in our analysis of non-explicit constructions of KL extractors in Chapter 2, and in fact this analysis via the KL-divergence is the only method we are aware of to obtain tight bounds on the concentration of the empirical subgaussian distance V_{d_g}, thus again showing the utility of arguing about indistinguishability via unpredictability. As a further consequence, we also obtain finite-sample bounds on all the moments of the empirical divergence which are within constant factors (depending on the moment) of their asymptotic values.

Optimal bounds between f-divergences and IPMs (Chapter 5, [AH20])

Recall that in Chapter 2, we use a generalization of Pinsker's inequality [BLM13, §4.9] which states that the unpredictability-type notion of Kullback–Leibler divergence controls the indistinguishability notion of subgaussian distance defined by test functions g which need not be bounded in $[0,1]$ (as in total variation distance and the standard Pinsker's inequality) if they satisfy appropriate subgaussian tail bounds. This result is based on the Donsker–Varadhan *variational representation* of the KL divergence [DV76, Theorem 5.2], which expresses the divergence between two probability distributions P and Q on the finite set[7] Ω as

$$\begin{aligned}\mathrm{KL}(P \parallel Q) &= \sup_{g:\Omega\to\mathbb{R}} \mathbb{E}[g(P)] - \lg \mathbb{E}\left[2^{g(Q)}\right] \\ &= \sup_{g:\Omega\to\mathbb{R}} \mathbb{E}[g(P)] - \mathbb{E}[g(Q)] - \lg \mathbb{E}\left[2^{g(Q)-\mathbb{E}[g(Q)]}\right]. \end{aligned} \tag{1.3}$$

[6] Technically the rate is n only when the KL is defined in the natural base (as it will be in Chapter 4), in base 2 the rate is $n/\ln 2$.

[7] For consistency, in this introduction we use finite sets and the binary logarithm, but in Chapter 5 we will work with arbitrary probability spaces and the (technically more convenient) natural logarithm.

Equation (1.3) makes clear the connection between such a representation of the divergence and IPMs, and it is natural to ask whether this connection can be generalized beyond the specific case of the KL divergence and subgaussian functions.

The Donsker–Varadhan representation is a specific instance of the *convex conjugate* in convex analysis, and so a natural generalization is to consider the class of *f-divergences* [Csi63; Mor63; Csi67a], also called *Ali–Silvey distances* in statistics [AS66], which are a family of convex divergences generalizing the KL divergence. However, it has been observed by [RRGP12] that the standard variational representation of f-divergences, which is valid for all finite measures, is not optimal for the case of probability distributions. In Chapter 5, based on joint work with Thibaut Horel [AH20], we derive a tight variational representation of the f-divergence, and use it to define a generalization of the moment generating function that we show exactly characterizes the best lower bound of an arbitrary f-divergence as a function of an arbitrary given IPM. Using this characterization, we obtain new bounds while also recovering in a unified manner well-known results, such as Hoeffding's lemma, Pinsker's inequality and its extension to subgaussian functions, and the Hammersley–Chapman–Robbins bound. The variational representation also allows us to prove new results on topological properties of the divergence which may be of independent interest.

Chapter 2

Samplers and Extractors for Unbounded Functions

This chapter is based on [Agr19].

2.1 Introduction

2.1.1 Averaging samplers

Averaging (or oblivious) samplers, introduced by Bellare and Rompel [BR94], are one of the main objects of study in pseudorandomness. Used to approximate the mean of a $[0,1]$-valued function with minimal randomness and queries, an averaging sampler takes a short random string and produces a small set of correlated points such that any given $[0,1]$-valued function will (with high probability) take approximately the same mean on these points as on the entire space. Formally,

Definition 2.1.1 ([BR94])**.** A function $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ is a (δ, ε) *averaging sampler* if for

all $f : \{0,1\}^m \to [0,1]$, it holds that

$$\Pr_{x \sim U_n}\left[\left|\frac{1}{D}\sum_{i=1}^{D} f(\mathrm{Samp}(x)_i) - \mathbb{E}[f(U_m)]\right| > \varepsilon\right] \leq \delta,$$

where U_n is the uniform distribution on $\{0,1\}^n$. The number n is the *randomness complexity* of the sampler, and D is the *sample complexity*. A sampler is *explicit* if $\mathrm{Samp}(x)_i$ can be computed in time $\mathrm{poly}(n, m, \lg D)$.

Traditionally, averaging samplers have been used in the context of randomness-efficient error reduction for algorithms and protocols, where the function f is the indicator of a set ($\{0,1\}$-valued), or more generally the acceptance probability of an algorithm or protocol ($[0,1]$-valued). There has been significant effort in the literature to establish optimal explicit and non-explicit constructions of samplers, which we summarize in Table 2.1. We recommend the survey of Goldreich [Gol11a] for more details, especially regarding non-averaging samplers[1].

However, averaging samplers can also have uses beyond bounded functions: Błasiok [Bła19], motivated by an application in streaming algorithms, introduced the notion of a *subgaussian sampler*, which he defined as an averaging sampler for functions $f : \{0,1\}^m \to \mathbb{R}$ such that $f(U_m)$ is a subgaussian random variable. Since subgaussian random variables have strong tail bounds, subgaussian functions from $\{0,1\}^m$ have a range contained in an interval of length $O(\sqrt{m})$, and thus one can construct a subgaussian sampler from a $[0,1]$-sampler by simply scaling the error ε by a factor of $O(\sqrt{m})$. Unfortunately, looking at Table 2.1 one sees that this induces a multiplicative dependence on m in the sample complexity, and for the expander walk sampler induces a dependence of $m \lg(1/\delta)$ in the randomness complexity. This loss can be avoided for some samplers, such as the sampler of Chor and Goldreich

[1] A non-averaging sampler is an algorithm Samp which makes oracle queries to f and outputs an estimate of its average which is good with high probability, but need not simply output the average of f's values on the queried points.

Table 2.1: Best known constructions of averaging samplers for $[0,1]$-valued functions

Key Idea	Randomness complexity n	Sample complexity D	Best regime
Pairwise-independent Expander Neighbors [GW97]	$m + O(\lg(1/\delta) + \lg(1/\varepsilon))$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Ramanujan Expander Neighbors[a] [KPS85; GW97]	m	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Extractors [Zuc97; GW97; RVW00; GUV09]	$m + (1+\alpha)\cdot\lg(1/\delta)$ any constant $\alpha > 0$	$\mathrm{poly}(\lg(1/\delta), 1/\varepsilon)$	$\varepsilon, \delta = o(1)$
Expander Walk Chernoff [Gil98]	$m + O(\lg(1/\delta)/\varepsilon^2)$	$O\left(\frac{\lg(1/\delta)}{\varepsilon^2}\right)$	$\varepsilon = \Omega(1)$
Pairwise Independence [CG89]	$O(m)$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	None, but simple
Non-Explicit [Zuc97]	$m + \lg(1/\delta) - \lg\lg(1/\delta) + O(1)$	$O\left(\frac{\lg(1/\delta)}{\varepsilon^2}\right)$	All
Lower Bound [CEG95; Zuc97; RT00]	$m + \lg(1/\delta) + \lg(1/\varepsilon) - \lg(D) - O(1)$	$\Omega\left(\frac{\lg(1/\delta)}{\varepsilon^2}\right)$	N/A

[a] Requires explicit constructions of Ramanujan graphs.

[CG89] based on pairwise independence (as its analysis requires only bounded variance) and (as we will show) the Ramanujan Expander Neighbor sampler of [KPS85; GW97], but Błasiok showed [Bła18] that the expander-walk sampler does not in general act as a subgaussian sampler without reducing the error to $o(1)$. We remark briefly that the median-of-averages sampler of Bellare, Goldreich, and Goldwasser [BGG93] still works and is optimal up to constant factors in the subgaussian setting (since the underlying pairwise independent sampler works), but it is not an averaging sampler[1, pg. 11], and matching its parameters with an averaging sampler remains open in general even for $[0,1]$-valued functions.

One of the contributions of this chapter is to give explicit averaging samplers for subgaussian

functions (in fact even for *subexponential* functions that satisfy weaker tail bounds) matching the extractor-based samplers for $[0,1]$-valued functions in Table 2.1 (up to the hidden polynomial in the sample complexity). This achieves the best parameters currently known in the regime of parameters where ε and δ are both subconstant, and in particular has no dependence on m in the sample complexity. We also show non-constructively that subexponentially samplers exist with essentially the same parameters as $[0,1]$-valued samplers.

Theorem 2.1.2 (Informal version of Theorem 2.6.1 and Corollary 2.6.7)**.** *For every integer $m \in \mathbb{N}$, $1 > \delta, \varepsilon > 0$, and $\alpha > 0$, there is a function* $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ *that is:*

- *an explicit subgaussian (in fact subexponential) sampler with randomness complexity $n = m + (1+\alpha) \cdot \lg(1/\delta)$ and sample complexity $D = \mathrm{poly}(\lg(1/\delta), 1/\varepsilon)$ (see Theorem 2.6.1)*
- *a non-constructive subexponential sampler with randomness complexity $n = m + \lg(1/\delta) - \lg\lg(1/\delta) + O(1)$ and sample complexity $D = O(\lg(1/\delta)/\varepsilon^2)$ (see Corollary 2.6.7).*

2.1.2 Randomness extractors

To prove Theorem 2.1.2, we develop a corresponding theory of generalized *randomness extractors* which we believe is of independent interest. For bounded functions, Zuckerman [Zuc97] showed that averaging samplers are essentially equivalent to randomness extractors, and in fact several of the best-known constructions of such samplers arose as extractor constructions. Formally, a randomness extractor is defined as follows:

Definition 2.1.3 (Nisan and Zuckerman [NZ96])**.** A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k,ε) *extractor* if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x \in \{0,1\}^n} \Pr[X = x] \leq 2^{-k}$, the distributions $\mathrm{Ext}(X, U_d)$ and U_m are ε-close in total variation distance. Equivalently, for all

$f : \{0,1\}^m \to [0,1]$ it holds that $\mathbb{E}[f(\mathrm{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$. The number d is called the *seed length*, and m the *output length*.

The formulation of Definition 2.1.3 in terms of $[0,1]$-valued functions implies that extractors produce an output distribution that is indistinguishable from uniform by all bounded functions f. It is therefore natural to consider a variant of this definition for a different set $\mathcal{F}$ of test functions $f : \{0,1\}^m \to \mathbb{R}$ which need not be bounded.

Definition 2.1.4 (Special case of Definition 2.3.1 using Definition 2.2.5)**.** A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε) extractor for a set of real-valued functions $\mathcal{F}$ from $\{0,1\}^m$ if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x \in \{0,1\}^n} \Pr[X = x] \leq 2^{-k}$ and every $f \in \mathcal{F}$, it holds that $\mathbb{E}[f(\mathrm{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$.

We show that much of the theory of extractors and samplers carries over to this more general setting. In particular, we generalize the connection of Zuckerman [Zuc97] to show that extractors for a class of functions of $\mathcal{F}$ are also samplers for that class, along with the converse (though as for total variation distance, there is some loss of parameters in this direction). Thus, to construct a subgaussian sampler it suffices (and is preferable) to construct a corresponding extractor for subgaussian test functions, which is how we prove Theorem 2.1.2.

Unfortunately, the distance induced by subgaussian test functions is not particularly pleasant to work with: for example the point masses on 0 and 1 in $\{0,1\}$ are $O(1)$ apart, but embedding them in the larger universe $\{0,1\}^m$ leads to distributions which are $\Theta(\sqrt{m})$ apart. We solve this problem by constructing extractors for a stronger notion, the *Kullback–Leibler (KL) divergence*, equivalently, extractors whose output is required to have very high Shannon entropy.

Definition 2.1.5 (Special case of Definition 2.3.1 using KL divergence)**.** A function $\mathrm{Ext} : \{0,1\}^n \times$

$\{0,1\}^d \to \{0,1\}^m$ is said to be a (k,ε) KL-extractor if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x\in\{0,1\}^n} \Pr[X = x] \leq 2^{-k}$ it holds that $\mathrm{KL}(\mathrm{Ext}(X, U_d) \parallel U_m) \leq \varepsilon$, or equivalently that $\mathrm{H}(\mathrm{Ext}(X, U_d)) \geq m - \varepsilon$.

A strong form of Pinsker's inequality (e.g. [BLM13, Lemma 4.18]) implies that a (k,ε^2) KL-extractor is also a (k,ε) extractor for subgaussian test functions. The KL divergence has the advantage that is nonincreasing under the application of functions (the famous *data-processing inequality*), and although it does not satisfy a traditional triangle inequality, it does satisfy a similar inequality when one of the segments satisfies stronger ℓ_2 bounds. These properties allow us to show that the zig-zag product for extractors of Reingold, Vadhan, and Wigderson [RVW00] also works for KL-extractors, and therefore to construct KL-extractors with seed length depending on n and k only through the *entropy deficiency* $n - k$ of X rather than n itself, which in the sampler perspective corresponds to a sampler with sample complexity depending on the failure probability δ rather than the universe size 2^m. Hence, we prove Theorem 2.1.2 by constructing corresponding KL-extractors.

Theorem 2.1.6 (Informal version of Theorem 2.6.2)**.** *For all integers m, $1 > \delta, \varepsilon > 0$, and $\alpha > 0$ there is an explicit (k,ε) KL-extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $n = m + (1+\alpha)\cdot \lg(1/\delta)$, $k = n - \lg(1/\delta)$, and $d = O(\lg\lg(1/\delta) + \lg(1/\varepsilon))$.*

Though the above theorem is most interesting in the high min-entropy regime where $n - k = o(n)$, we also show the existence of KL-extractors matching most of the existing constructions of total variation extractors. In particular, we note that extractors for ℓ_2 are immediately KL-extractors without loss of parameters, and also that any extractor can be made a KL-extractor by taking slightly smaller error, so that the extractors of Guruswami, Umans, and Vadhan [GUV09] can be taken to be KL-extractors with essentially the same parameters.

Furthermore, in addition to our explicit constructions, we also show non-constructively that KL-extractors (and hence subgaussian extractors) exist with very good parameters:

Theorem 2.1.7 (Informal version of Theorem 2.5.30)**.** *For any integers $k < n \in \mathbb{N}$ and $1 > \varepsilon > 0$ there is a (k, ε) KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $d = \lg(n-k) + \lg(1/\varepsilon) + O(1)$ and $m = k + d - \lg(1/\varepsilon) - O(1)$.*

One key thing to note about the nonconstructive KL-extractors of the above theorem is that they incur an entropy loss of only $1 \cdot \lg(1/\varepsilon)$, whereas total variation extractors necessarily incur entropy loss $2 \cdot \lg(1/\varepsilon)$ by the lower bound of Radhakrishnan and Ta-Shma [RT00]. In particular, by Pinsker's inequality, (k, ε^2) KL-extractors with the above parameters are also optimal (k, ε) standard (total variation) extractors [RT00], so that one does not lose anything by constructing a KL-extractor rather than a total variation extractor. We also remark that the above theorem gives subgaussian samplers with better parameters than a naive argument that a random function should directly be a subgaussian sampler, as it avoids the need to take a union bound over $O(M^M) = O(2^{M \lg M})$ test functions (for $M = 2^m$) which results in additional additive $\lg \lg$ factors in the randomness complexity.

In the total variation setting, there are only a couple of methods known to explicitly achieve optimal entropy loss $2 \cdot \lg(1/\varepsilon)$, the easiest of which is to use an extractor which natively has this sort of loss, of which only three are known: An extractor from random walks over Ramanujan Graphs due to Goldreich and Wigderson [GW97], the Leftover Hash Lemma due to Impagliazzo, Levin, and Luby [ILL89] (see also [McI87; BBR88]), and the extractor based on almost-universal hashing of Srinivasan and Zuckerman [SZ99]. Unfortunately, all of these are ℓ_2 extractors and so must have seed length linear in $\min(n-k, m)$ (cf. [Vad12, Problem 6.4]), rather than logarithmic in $n-k$ as known non-constructively. The other alternative is to use the generic reduction of Raz, Reingold, and Vadhan [RRV02] which

turns any extractor Ext with entropy loss Δ into one with entropy loss $2 \cdot \lg(1/\varepsilon) + O(1)$ by paying an additive $O(\Delta + \lg(n/\varepsilon))$ in seed length. We show that all of these ℓ_2 extractors and the [RRV02] transformation also work to give KL-extractors with entropy loss $1 \cdot \lg(1/\varepsilon) + O(1)$, so that applications which require minimal entropy loss can also use explicit constructions of KL-extractors.

2.1.3 Future directions

Broadly speaking, we hope that the perspective of KL-extractors will bring new tools (perhaps from information theory) to the construction of extractors and samplers. For example, since KL-extractors can have seed length with dependence on ε of only $1 \cdot \lg(1/\varepsilon)$, trying to explicitly construct a KL-extractor with seed length $1 \cdot \lg(1/\varepsilon) + o(\min(n-k, k))$ may also shed light on how to achieve optimal dependence on ε in the total variation setting.

In the regime of constant $\varepsilon = \Omega(1)$, we do not have explicit constructions of subgaussian samplers matching the expander-walk sampler of Gillman [Gil98] for $[0,1]$-valued functions, which achieves randomness complexity $m + O(\lg(1/\delta))$ and sample complexity $O(\lg(1/\delta))$, as asked for by Błasiok [Bła19]. From the extractor point-of-view, it would suffice (by the reduction of [GW97; RVW00] that we analyze for KL-extractors) to construct explicit *linear degree* KL-extractors with parameters matching the linear degree extractor of Zuckerman [Zuc07], i.e. with seed length $d = \lg(n) + O(1)$ and $m = \Omega(k)$ for $\varepsilon = \Omega(1)$. A potentially easier problem, since the Zuckerman linear degree extractor is itself based on the expander-walk sampler, could be to instead match the parameters of the near-linear degree extractors of Ta-Shma, Zuckerman, and Safra [TZS06] based on Reed–Muller codes, thereby achieving sample complexity $O(\lg(1/\delta) \cdot \operatorname{poly} \lg \lg(1/\delta))$.

Finally, we hope that KL-extractors can also find uses beyond being subgaussian samplers and total variation extractors: for example it seems likely that there are applications (perhaps in coding or

cryptography, cf. [Bar+11]) where it is more important to have high Shannon entropy in the output than small total variation distance to uniform, in which case one may be able to use (k, ε) KL-extractors with entropy loss only $1 \cdot \lg(1/\varepsilon)$ directly, rather than a total variation extractor or (k, ε^2) KL-extractor with entropy loss $2 \cdot \lg(1/\varepsilon)$.

2.2 Preliminaries

2.2.1 (Weak) statistical divergences and metrics

Our results in general will require very few assumptions on notions of "distance" between probability distributions, so we will give a general definition and indicate in our theorems when we need which assumptions.

Definition 2.2.1. A *weak statistical divergence* (or simply *weak divergence*) on a finite set $\mathcal{X}$ is a function D from pairs of probability distributions over X to $\mathbb{R} \cup \{\pm\infty\}$. We write $\mathrm{D}(P \parallel Q)$ for the value of D on distributions P and Q. Furthermore

1. If $\mathrm{D}(P \parallel Q) \geq 0$ with equality iff $P = Q$, then D is *positive-definite*, and we simply call D a *divergence*.

2. If $\mathrm{D}(P \parallel Q) = \mathrm{D}(Q \parallel P)$, then D is *symmetric*.

3. If $\mathrm{D}(P \parallel R) \leq \mathrm{D}(P \parallel Q) + \mathrm{D}(Q \parallel R)$, then D satisfies the *triangle inequality*.

4. If $\mathrm{D}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda \,\mathrm{D}(P_1 \parallel Q_1) + (1-\lambda)\,\mathrm{D}(P_2 \parallel Q_2)$ for all $\lambda \in [0,1]$, then D is *jointly convex*. If this holds only when $Q_1 = Q_2$ then D is *convex in its first argument*.

5. If D is defined on all finite sets $\mathcal{Y}$ and for all functions $f : \mathcal{X} \to \mathcal{Y}$ the divergence is nonincreasing under f, that is $\mathrm{D}(f(P) \parallel f(Q)) \leq \mathrm{D}(P \parallel Q)$, then D satisfies the *data-processing inequality*.

If D is positive-definite, symmetric, and satisfies the triangle inequality, then it is called a *metric*.

Example 2.2.2. The ℓ_p *distance* for $p \geq 1$ between probability distributions over $\mathcal{X}$ is

$$d_{\ell_p}(P, Q) \stackrel{\text{def}}{=} \left(\sum_{x \in \mathcal{X}} |P_x - Q_x|^p \right)^{1/p}$$

is a jointly-convex metric. Furthermore, the ℓ_p distance is nonincreasing in p, and when $p = 1$ it satisfies the data-processing inequality.

Example 2.2.3. The *total variation distance* is

$$d_{\mathrm{TV}}(P, Q) \stackrel{\text{def}}{=} \frac{1}{2} d_{\ell_1}(P, Q) = \sup_{S \subseteq \mathcal{X}} \left| \Pr[P \in S] - \Pr[Q \in S] \right| = \sup_{f \in [0,1]^{\mathcal{X}}} \Big(\mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \Big)$$

and is a jointly convex metric that satisfies the data-processing inequality.

Example 2.2.4 (Rényi Divergences [Rén61])**.** For two probability distributions P and Q over a finite set $\mathcal{X}$, the *Rényi α-divergence* or *Rényi divergence of order α* is defined for real $0 < \alpha \neq 1$ by

$$\mathrm{D}_\alpha(P \parallel Q) \stackrel{\text{def}}{=} \frac{1}{\alpha - 1} \lg \left(\sum_{x \in \mathcal{X}} \frac{P_x^\alpha}{Q_x^{\alpha - 1}} \right)$$

where the logarithm is in base 2 (as are all logarithms in this chapter unless noted otherwise). The Rényi divergence is continuous in α and so is defined by taking limits for $\alpha \in \{0, 1, \infty\}$, giving for $\alpha = 0$ the divergence $\mathrm{D}_0(P \parallel Q) \stackrel{\text{def}}{=} \lg\big(1/\Pr_{x \sim Q}[P_x \neq 0]\big)$, for $\alpha = 1$ the *Kullback–Leibler (or KL) divergence*

$$\mathrm{KL}(P \parallel Q) \stackrel{\text{def}}{=} \mathrm{D}_1(P \parallel Q) = \sum_{x \in X} P_x \lg \frac{P_x}{Q_x},$$

and for $\alpha = \infty$ the *max-divergence* $\mathrm{D}_\infty(P \parallel Q) \stackrel{\text{def}}{=} \max_{x \in X} \lg \frac{P_x}{Q_x}$. The Rényi divergence is nondecreasing in α. Furthermore, when $\alpha \leq 1$ the Rényi divergence is jointly convex, and for all α the Rényi divergence satisfies the data-processing inequality [vEH14].

The *Rényi α-entropy* of P is defined as $\mathrm{H}_\alpha(P) \stackrel{\text{def}}{=} \lg |\mathcal{X}| - \mathrm{D}_\alpha(P \parallel U_{\mathcal{X}})$ for $Q = U_{\mathcal{X}}$ the uniform

distribution over the set $\mathcal{X}$, and satisfies $0 \le \mathrm{H}_\alpha(P) \le \lg|\mathcal{X}|$. For $\alpha = 0$, $\mathrm{H}_0(P) = \lg|\mathrm{Supp}(P)|$ is the *max-entropy of P*, for $\alpha = 1$, $\mathrm{H}_1(P) = \sum_{x\in\mathcal{X}} P_x \lg(1/P_x)$ is the *Shannon entropy of P*, and for $\alpha = \infty$, $\mathrm{H}_\infty(P) = \min_{x\in\mathcal{X}} \lg(1/P_x)$ is the *min-entropy of P*.

For $\alpha = 2$, the Rényi 2-entropy can be expressed in terms of the ℓ_2-distance to uniform:

$$\lg|\mathcal{X}| - \mathrm{H}_2(P) = \mathrm{D}_2(P \parallel U_\mathcal{X}) = \lg\bigl(1 + |\mathcal{X}| \cdot d_{\ell_2}(P, U_\mathcal{X})^2\bigr),$$

and also in terms of the *collision probability*

$$\mathrm{H}_2(P) = \lg \frac{1}{\Pr_{x,x'\sim P}[x = x']} = \lg \frac{1}{\sum_{x\in X} P_x^2}.$$

2.2.2 Integral Probability Metrics, or weak divergences from test functions

Zuckerman's connection [Zuc97] between samplers for bounded functions and extractors for total variation distance is based on the following standard characterization of total variation distance as the maximum distinguishing advantage achieved by bounded functions,

$$d_{\mathrm{TV}}(P, Q) = \sup_{f\in[0,1]^\mathcal{X}} \mathbb{E}[f(P)] - \mathbb{E}[f(Q)].$$

By considering an arbitrary class of functions in the supremum, we get the following weak divergence, a special case of what Müller [Mül97] called *integral probability metrics (IPMs)*:

Definition 2.2.5. Given a finite $\mathcal{X}$ and a set of real-valued functions $\mathcal{F} \subseteq \mathbb{R}^\mathcal{X}$, the *$\mathcal{F}$-distance on $\mathcal{X}$* between probability distributions on $\mathcal{X}$ is denoted by $\mathrm{D}^\mathcal{F}$ and is defined as

$$\mathrm{D}^\mathcal{F}(P \parallel Q) \stackrel{\text{def}}{=} \sup_{f\in\mathcal{F}}\Bigl(\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]\Bigr) = \sup_{f\in\mathcal{F}} \mathrm{D}^{\{f\}}(P \parallel Q).$$

We call the set of functions $\mathcal{F}$ *symmetric* if for all $f \in \mathcal{F}$ there is $c \in \mathbb{R}$ and $g \in \mathcal{F}$ such that $g = c - f$, and *distinguishing* if for all $P \ne Q$ there exists $f \in \mathcal{F}$ with $\mathrm{D}^{\{f\}}(P \parallel Q) > 0$.

Remark 2.2.6. For simplicity, all our probabilistic distributions are given only for random variables and distributions over finite sets as this is all we need for our application in this chapter. We examine the more general case of integral probability metrics [Mül97] on arbitrary probability spaces in Chapter 5.

Example 2.2.7. If $\mathcal{F} = \{0,1\}^{\mathcal{X}}$ or $\mathcal{F} = [0,1]^{\mathcal{X}}$, then $\mathrm{D}^{\mathcal{F}}$ is exactly the total variation distance. Equivalently, the ℓ_1 distance is the $\mathcal{F}$-distance for $\mathcal{F} = \{-1,1\}^{\mathcal{X}}$ or $\mathcal{F} = [-1,1]^{\mathcal{X}}$.

Example 2.2.8. Families of $\mathcal{F}$-distances for $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ have a long history in computer science: for example, by taking $\mathcal{F}$ to be the set of functions computable by efficient algorithms (under some formalization), one recovers the standard notions of pseudorandomness and computational indistinguishability in computer science dating back to Goldwasser and Micali [GM84] and Yao [Yao82], and the work of Reingold et al. [RTTV08b; RTTV08a] has more examples of this definition in computer science.

Example 2.2.9. Generalizing Example 2.2.7, for every $1 \leq p, q \leq \infty$ such that $1/p + 1/q = 1$, we have that that $d_{\ell_p} = |\mathcal{X}|^{-1/q} \cdot d_{\mathcal{M}_q}$ where $\mathcal{M}_q$ is the family of real-valued functions with bounded q-moments

$$\mathcal{M}_q \stackrel{\text{def}}{=} \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f(U_{\mathcal{X}})\|_q \stackrel{\text{def}}{=} \mathbb{E}\left[|f(U_{\mathcal{X}})|^q\right]^{1/q} \leq 1 \right\}.$$

This is a special case of the duality of the Lebesgue spaces L^p and L^q (see for example Dunford and Schwartz [DS58, Theorem IV.3.9]).

Remark 2.2.10. An equivalent definition of $\mathcal{F}$ being symmetric is that for all $f \in \mathcal{F}$ there exists $g \in \mathcal{F}$ with $\mathrm{D}^{\{g\}}(P \parallel Q) = -D^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$ for all distributions P and Q. Hence, one might also consider a weaker notion of symmetry that reverses quantifiers, where $\mathcal{F}$ is "weakly-symmetric" if for all $f \in \mathcal{F}$ and distributions P and Q there exists $g \in \mathcal{F}$ such that $\mathrm{D}^{\{g\}}(P \parallel Q) = -\mathrm{D}^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$. However, such a class $\mathcal{F}$ gives exactly the same weak divergence $\mathrm{D}^{\mathcal{F}}$ as its "symmetrization"

$\overline{\mathcal{F}} = \mathcal{F} \cup \{-f \mid f \in \mathcal{F}\}$, so we do not need to introduce this more complex notion.

Remark 2.2.11. We use a superscript in the notation $\mathrm{D}^{\mathcal{F}}$ to avoid confusion with the Csiszár f-divergences [Csi63], also known as Ali–Silvey distances [AS66], which are a family of divergences parametrized by a convex function f and are commonly denoted D_f.

Remark 2.2.12. By identifying distributions with their probability mass functions, one can realize $\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]$ as an inner product $\langle P - Q, f \rangle$, so that $\mathrm{D}^{\mathcal{F}}(P \parallel Q)$ is the supremum of linear functions on the space of real-valued functions on $\mathcal{X}$.

We now establish some basic properties of $\mathrm{D}^{\mathcal{F}}$.

Lemma 2.2.13. *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of real-valued functions over a finite set $\mathcal{X}$. Then $\mathrm{D}^{\mathcal{F}}$ satisfies the triangle inequality and is jointly convex, and*

1. *if $\mathcal{F}$ is symmetric then $\mathrm{D}^{\mathcal{F}}$ is symmetric and*

$$\mathrm{D}^{\mathcal{F}}(P \parallel Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \right| \geq 0,$$

2. *if $\mathcal{F}$ is distinguishing then $\mathrm{D}^{\mathcal{F}}$ is positive-definite,*

so that if $\mathcal{F}$ is both symmetric and distinguishing then $\mathrm{D}^{\mathcal{F}}$ is a jointly convex metric on probability distributions over $\mathcal{X}$, in which case we also use the notation $d_{\mathcal{F}}(P, Q) \stackrel{\text{def}}{=} \mathrm{D}^{\mathcal{F}}(P \parallel Q)$.

Proof. The triangle inequality and joint convexity both follow from the linearity of each $\mathrm{D}^{\{f\}}$, as by linearity of expectation, for all $f : \mathcal{X} \to \mathbb{R}$ it holds that

$$\mathrm{D}^{\{f\}}(P \parallel R) = \mathrm{D}^{\{f\}}(P \parallel Q) + \mathrm{D}^{\{f\}}(Q \parallel R)$$

$$\mathrm{D}^{\{f\}}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) = \lambda \, \mathrm{D}^{\{f\}}(P_1 \parallel Q_1) + (1-\lambda) \, \mathrm{D}^{\{f\}}(P_2 \parallel Q_2).$$

Upper bounding the terms on the right-hand side by $\mathrm{D}^{\mathcal{F}}$ and taking the supremum of the left hand side over $f \in \mathcal{F}$ then gives the claims. The symmetry and positive-definite claims are immediate from the definitions. □

2.3 Extractors for weak divergences and connections to samplers

2.3.1 Definitions

We now use this machinery to extend the notion of an extractor due to Nisan and Zuckerman [NZ96] and the average-case variant of Dodis et al. [DORS08].

Definition 2.3.1 (Extends Definition 2.1.4)**.** Let D be a weak divergence on the set $\{0,1\}^m$, and $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$. Then if for all distributions X over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$ it holds that

1. $\mathrm{D}(\mathrm{Ext}(X, U_d) \parallel U_m) \leq \varepsilon$, then Ext is said to be a (k,ε) *extractor for* D, or a (k,ε) D*-extractor*.

2. $\mathbb{E}_{s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X,s) \parallel U_m)] \leq \varepsilon$, then Ext is said to be a (k,ε) *strong extractor for* D, or a (k,ε) *strong* D*-extractor*.

Furthermore, if for all joint distributions (Z,X) where X is distributed over $\{0,1\}^n$ with $\widetilde{\mathrm{H}}_\infty(X|Z) \stackrel{\text{def}}{=} \lg\left(1/\mathbb{E}_{z\sim Z}\left[2^{-\mathrm{H}_\infty(X|_{Z=z})}\right]\right) \geq k$, it holds that

3. $\mathbb{E}_{z\sim Z}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m) \leq \varepsilon]$, then Ext is said to be a (k,ε) *average-case extractor for* D, or a (k,ε) *average-case* D*-extractor*.

4. $\mathbb{E}_{z\sim Z, s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)] \leq \varepsilon$, then Ext is said to be a (k,ε) *average-case strong extractor for* D, or a (k,ε) *average-case strong* D*-extractor*.

Remark 2.3.2. By taking D to be the total variation distance we recover the standard definitions of extractor and strong extractor due to [NZ96] and the definition of average-case extractor due to [DORS08].

However, our definitions are phrased slightly differently for strong and average-case extractors as an expectation rather than a joint distance, that is, for strong average-case extractors we require a bound on the expectation $\mathbb{E}_{z\sim Z, s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)]$ rather than a bound on the joint distance $\mathrm{D}(Z, U_d, \mathrm{Ext}(X, U_d) \parallel Z, U_d, U_m)$. In our setting, the weak divergence D need not be defined over the larger joint universe, but it is defined for all random variables over $\{0,1\}^m$. In the case of d_{TV} and KL divergence, both definitions are equivalent (for KL divergence, this is an instance of the *chain rule*).

Remark 2.3.3. The strong variants of Definition 2.3.1 are also non-strong extractors assuming the weak divergence D is convex in its first argument, as it is for most weak divergences of interest, including the ℓ_p norms for $p \geq 1$, all $\mathrm{D}^{\mathcal{F}}$ defined by test functions, the KL divergence, Rényi divergences for $\alpha \leq 1$, and all Csiszár–Morimoto–Ali–Silvey f-divergences. The average-case variants are always non-average-case extractors by taking Z to be independent of X.

Remark 2.3.4. We gave Definition 2.3.1 for general weak divergences which need not be symmetric, and made the particular choice that the output of the extractor was on the left-hand side of the weak divergence and that the uniform distribution was on the right-hand side. This is motivated by the standard information-theoretic divergences such as KL divergence, which require the left-hand distribution to have support contained in the support of the right-hand distribution, and putting the uniform distribution on the right ensures this is always the case. Furthermore, the KL divergence to uniform has a natural interpretation as an entropy difference, $\mathrm{KL}(P \parallel U_m) = m - \mathrm{H}(P)$ for H the Shannon entropy, so that in particular a KL-extractor with error ε requires the output to have Shannon entropy at least $m - \varepsilon$. If for a weak divergence D the other direction is more natural, one can always reverse the sides

by considering the weak divergence $\mathrm{D}'(Q \parallel P) = \mathrm{D}(P \parallel Q)$.

Remark 2.3.5. Definition 2.3.1 does not technically need even a weak divergence, as it suffices to simply have a measure of distance to uniform. However, since weak divergences have minimal constraints, one can define a weak divergence from any distance to uniform by ignoring the second component (or setting it to be infinite for non-uniform distributions).

We also give the natural definition of averaging samplers for arbitrary classes of functions $\mathcal{F}$ extending Definition 2.1.1, along with the strong variant of Zuckerman [Zuc97].

Definition 2.3.6. Given a class of functions $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$, a function $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ is said to be a (δ, ε) *strong averaging sampler for* $\mathcal{F}$ or a (δ, ε) *strong averaging* $\mathcal{F}$*-sampler* if for all $f_1, \ldots, f_D \in \mathcal{F}$, it holds that

$$\Pr_{x \sim U_n}\left[\mathop{\mathbb{E}}_{i \sim U_{[D]}}\left[f_i\big(\mathrm{Samp}(x)_i\big) - \mathbb{E}[f_i(U_m)]\right] > \varepsilon\right] \leq \delta$$

where $[D] = \{1, \ldots, D\}$. If this holds only when $f_1 = \cdots = f_D$, then it is called a *(non-strong)* (δ, ε) *averaging sampler for* $\mathcal{F}$ or (δ, ε) *averaging* $\mathcal{F}$*-sampler*. We say that Samp is a (δ, ε) *strong absolute averaging sampler for* $\mathcal{F}$ if it also holds that

$$\Pr_{x \sim U_n}\left[\left|\mathop{\mathbb{E}}_{i \sim U_{[D]}}\left[f_i\big(\mathrm{Samp}(x)_i\big) - \mathbb{E}[f_i(U_m)]\right]\right| > \varepsilon\right] \leq \delta.$$

with the analogous definition for non-strong samplers.

Remark 2.3.7. We separated a single-sided version of the error bound in Definition 2.3.6 as in [Vad12], as it makes the connection between extractors and samplers cleaner and allows us to be specific about what assumptions are needed. Note that if $\mathcal{F}$ is symmetric then every (δ, ε) (strong) sampler for $\mathcal{F}$ is a $(2\delta, \varepsilon)$ (strong) absolute sampler for $\mathcal{F}$, recovering the standard notion up to a factor of 2 in δ.

2.3.2 Equivalence of extractors and samplers

We now show that Zuckerman's connection [Zuc97] does indeed generalize to this broader setting as promised.

Theorem 2.3.8. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be an* $(n - \lg(1/\delta), \varepsilon)$*-extractor (respectively strong extractor) for the weak divergence* $\mathrm{D}^{\mathcal{F}}$ *defined by a class of test functions* $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$ *as in Definition 2.2.5. Then the function* $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ *for* $D = 2^d$ *defined by* $\mathrm{Samp}(x)_i = \mathrm{Ext}(x,i)$ *is a* (δ, ε)*-sampler (respectively strong sampler) for* $\mathcal{F}$*.*

Proof. The proof is essentially the same as that of Zuckerman [Zuc97, Lemmas 2.6 and 2.14, Propositions 2.7 and 2.15].

Fix a collection of test functions $f_1, \ldots, f_D \in \mathcal{F}$, where if Ext is not strong we restrict to $f_1 = \cdots = f_D$, and let $B_{f_1,\ldots,f_D} \subseteq \{0,1\}^n$ be defined as

$$\begin{aligned} B_{f_1,\ldots,f_D} &\stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_{[D]}} \Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)] \Big] > \varepsilon \right\} \\ &= \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_{[D]}} \Big[\mathrm{D}^{\{f_i\}}\big(U_{\{\mathrm{Ext}(x,i)\}} \,\big\|\, U_m \big) \Big] > \varepsilon \right\}, \end{aligned}$$

where $U_{\{z\}}$ is the point mass on z. Then if X is uniform over $B_{f_1,\ldots,f_D}$, we have

$$\begin{aligned} \varepsilon &< \mathop{\mathbb{E}}_{x \sim X} \Big[\mathop{\mathbb{E}}_{i \sim U_{[D]}} \Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)] \Big] \Big] \\ &= \mathop{\mathbb{E}}_{i \sim U_{[D]}} \Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i) \parallel U_m) \Big] \\ &= \begin{cases} \mathrm{D}^{\{f_1\}}(\mathrm{Ext}(X, U_d) \parallel U_m) & \text{if } f_1 = \cdots = f_D \\ \mathbb{E}_{i \sim U_{[D]}} \Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i) \parallel U_m) \Big] & \text{always} \end{cases} \end{aligned}$$

$$\leq \begin{cases} \mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X, U_d) \parallel U_m) & \text{if } f_1 = \cdots = f_D \\ \mathbb{E}_{i \sim U_{[D]}}\left[\mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X, i) \parallel U_m)\right] & \text{always} \end{cases}$$

Since Ext is an $(n - \lg(1/\delta), \varepsilon)$-extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$ we must have $\mathrm{H}_\infty(X) < n - \lg(1/\delta)$. But $\mathrm{H}_\infty(X) = \lg|B_{f_1,\ldots,f_D}|$ by definition, so we have $|B_{f_1,\ldots,f_D}| < \delta 2^n$. Hence, the probability that a random $x \in \{0,1\}^n$ lands in $B_{f_1,\ldots,f_D}$ is less than δ, and since $B_{f_1,\ldots,f_D}$ is exactly the set of coin tosses which are bad for Samp, this concludes the proof. □

Remark 2.3.9. Hölder's inequality implies that an extractor for ℓ_p with error $\varepsilon \cdot 2^{-m(p-1)/p}$ is also an ℓ_1 extractor and thus $[-1,1]$-averaging sampler with error ε. Example 2.2.9 and Theorem 2.3.8 show that they are in fact samplers for the much larger class of functions $\mathcal{M}_{p/(p-1)}$ with bounded $p/(p-1)$ moments (rather than just ∞ moments), also with error ε.

Furthermore, if all the functions in $\mathcal{F}$ have bounded deviation from their mean (for example, subgaussian functions from $f : \{0,1\}^m \to \mathbb{R}$ have such a bound of $O(\sqrt{m})$ by the tail bounds from Lemma 2.4.3), then we also have a partial converse that recovers the standard converse in the case of total variation distance.

Theorem 2.3.10. *Let $\mathcal{F}$ be a class of functions $\mathcal{F} \subset \{0,1\}^m \to \mathbb{R}$ with finite* maximum deviation from the mean*, meaning* $\max \mathrm{dev}(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} \max_{x \in \{0,1\}^n}\big(f(x) - \mathbb{E}[f(U_m)]\big) < \infty$. *Then given a (δ, ε) $\mathcal{F}$-sampler (respectively (δ, ε) strong $\mathcal{F}$-sampler)* $\mathrm{Samp} : \{0,1\}^n \to \big(\{0,1\}^m\big)^D$*, the function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for $d = \lg D$ defined by* $\mathrm{Ext}(x,i) = \mathrm{Samp}(x)_i$ *is a $\big(k, \varepsilon + \delta \cdot 2^{n-k} \cdot \max \mathrm{dev}(\mathcal{F})\big)$ $\mathrm{D}^{\mathcal{F}}$-extractor (respectively strong $\mathrm{D}^{\mathcal{F}}$-extractor) for every $0 \leq k \leq n$.*

In particular, Ext *is an $\big(n - \lg(1/\delta) + \lg(1/\eta), \varepsilon + \eta \cdot \max \mathrm{dev}(\mathcal{F})\big)$ average-case $\mathrm{D}^{\mathcal{F}}$-extractor (respectively strong average-case $\mathrm{D}^{\mathcal{F}}$-extractor) for every $\delta \leq \eta \leq 1$.*

Proof. Again the proof is analogous to the one in Zuckerman [Zuc97, Propositions 2.8, 2.9, and 2.16].

Fix a distribution X over $\{0,1\}^m$ with $\mathrm{H}_\infty(X) \geq k$ and a collection of test functions $f_1, \ldots, f_D \in \mathcal{F}$, where if Samp is not strong we restrict to $f_1 = \cdots = f_D$. Then since Samp is a (δ, ε) $\mathcal{F}$-sampler, we know that the set of coin tosses for which the sampler is bad must be small. Formally, the set

$$\begin{aligned} B_{f_1,\ldots,f_D} &\stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_d}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)] \Big] > \varepsilon \right\} \\ &= \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_d}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)] \Big] > \varepsilon \right\} \end{aligned}$$

has size $|B_{f_1,\ldots,f_D}| \leq \delta 2^n$. Thus, since X has min-entropy at least k we know $\Pr\big[X \in B_{f_1,\ldots,f_D}\big] \leq \Big(\max_{x \in B_{f_1,\ldots,f_D}} \Pr[X = x]\Big) \cdot |B_{f_1,\ldots,f_D}| \leq 2^{-k} \cdot \delta 2^n$, so we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{i \sim U_d}\Big[\mathbb{E}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \Big]\Big] \\ &\qquad = \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i \sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \Big]\Big] \\ &\qquad = \Pr\big[X \in B_{f_1,\ldots,f_D}\big] \cdot \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i \sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \Big] \;\Big|\; X \in B_{f_1,\ldots,f_D} \Big] \\ &\qquad\quad + \Pr\big[X \notin B_{f_1,\ldots,f_D}\big] \cdot \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i \sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \Big] \;\Big|\; X \notin B_{f_1,\ldots,f_D} \Big] \\ &\qquad \leq \Pr\big[X \in B_{f_1,\ldots,f_D}\big] \cdot \max \mathrm{dev}(\mathcal{F}) + \Pr\big[X \notin B_{f_1,\ldots,f_D}\big] \cdot \varepsilon \\ &\qquad \leq 2^{-k} \cdot \delta 2^n \cdot \max \mathrm{dev}(\mathcal{F}) + \varepsilon \end{aligned}$$

completing the proof of the main claim. The "in particular" statement follows since if (Z, X) are jointly distributed with $\widetilde{\mathrm{H}}_\infty(X|Z) \geq n - \lg(1/\delta) + \lg(1/\eta)$ we have

$$\mathop{\mathbb{E}}_{z \sim Z}\big[\varepsilon + \delta \cdot 2^{n - \mathrm{H}_\infty(X|Z=z)} \cdot \max \mathrm{dev}(\mathcal{F}) \big] = \varepsilon + \delta \cdot 2^{n - \widetilde{\mathrm{H}}_\infty(X|Z)} \cdot \max \mathrm{dev}(\mathcal{F}) \leq \varepsilon + \eta \cdot \max \mathrm{dev}(\mathcal{F})$$

by definition of conditional min-entropy. □

2.3.3 All extractors are average-case

Under a similar boundedness condition for general weak divergences, we can recover the standard fact that all extractors are average-case extractors under a slight loss of parameters (the same loss as achieved by Dodis et al. [DORS08] for the case of total variation distance). More interestingly, if the weak divergence is given by $\mathrm{D}^{\mathcal{F}}$ for a symmetric class of (possibly unbounded) functions $\mathcal{F}$, we can also generalize and recover the result of Vadhan [Vad12, Problem 6.8] that shows that a (k, ε) extractor (for total variation) is a $(k, 3\varepsilon)$ average-case extractor without any other loss.

Theorem 2.3.11. *Let* D *be a bounded weak divergence over* $\{0,1\}^m$*, meaning that*

$$0 \le \|\mathrm{D}\|_\infty \stackrel{\text{def}}{=} \sup_{P \text{ on } \{0,1\}^m} \mathrm{D}(P \parallel U_m) < \infty.$$

Then a (k, ε)*-extractor for* D *(respectively strong extractor)* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is also a* $(k + \lg(1/\eta), \varepsilon + \eta \cdot \|\mathrm{D}\|_\infty)$ *average-case-extractor for* D *(respectively strong average-case-extractor) for any* $0 < \eta \le 1$.

Proof. The proof is analogous to that of [DORS08]. We prove it only for non-strong extractors, the proof for strong extractors is completely analogous by adding more expectations.

For jointly distributed random variables (Z, X) such that $\widetilde{\mathrm{H}}_\infty(X|Z) \ge k + \lg(1/\eta)$, we have by [DORS08, Lemma 2.2] that the probability that $\Pr_{z \sim Z}[\mathrm{H}_\infty(X|_{Z=z}) < k] \le \eta$. Thus

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z \sim Z}\left[\mathrm{D}\left(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m\right)\right] \\
&\quad = \Pr_{z \sim Z}[\mathrm{H}_\infty(X|_{Z=z}) < k] \cdot \mathop{\mathbb{E}}_{z \sim Z}\left[\mathrm{D}\left(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m\right) \middle| \mathrm{H}_\infty(X|_{Z=z}) < k\right] \\
&\qquad + \Pr_{z \sim Z}[\mathrm{H}_\infty(X|_{Z=z}) \ge k] \cdot \mathop{\mathbb{E}}_{z \sim Z}\left[\mathrm{D}\left(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m\right) \middle| \mathrm{H}_\infty(X|_{Z=z}) \ge k\right] \\
&\quad \le \eta \cdot \|D\|_\infty + 1 \cdot \varepsilon
\end{aligned}$$

□

Theorem 2.3.12. *Let $\mathcal{F}$ be a symmetric class of test functions and* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a (k, ε) extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$, where k is at most $n - 1$. Then* Ext *is an $(k, 3\varepsilon)$ average-case extractor (respectively strong average-case extractor) for $\mathrm{D}^{\mathcal{F}}$.*

Remark 2.3.13. Theorem 2.3.12 also applies to extractors for the ℓ_p norms via Example 2.2.9.

The proof of Theorem 2.3.12 follows the strategy outlined by Vadhan [Vad12, Problem 6.8]. We first isolate the following key lemma which shows that any extractor with error that gracefully decays with lower min-entropy is average-case with minimal loss of parameters, as opposed to Theorem 2.3.11 which used a worst-case error bound when the min-entropy is low.

Lemma 2.3.14. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a (k, ε) extractor (respectively strong extractor) for* D *such that for every $0 \le t \le k$,* Ext *is also a $(k - t, 2^{t+1} \cdot \varepsilon)$ extractor (respectively strong extractor) for* D. *Then* Ext *is a $(k, 3\varepsilon)$ average-case extractor (respectively strong average-case extractor) for* D.

Proof. We prove this for strong extractors, the non-strong case is analogous. For every (Z, X) with X distributed on $\{0,1\}^n$ and $\widetilde{\mathrm{H}}_\infty(X|Z) \ge k$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim Z, s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)] &= \mathop{\mathbb{E}}_{z\sim Z}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)]\right] \\
&\le \mathop{\mathbb{E}}_{z\sim Z}\left[\begin{cases} \varepsilon & \text{if } \mathrm{H}_\infty(X|_{Z=z}) \ge k \\ 2^{k-\mathrm{H}_\infty(X|_{Z=z})+1} \cdot \varepsilon & \text{otherwise} \end{cases}\right] \\
&\le \varepsilon \cdot \mathop{\mathbb{E}}_{z\sim Z}\left[1 + 2^{k-\mathrm{H}_\infty(X|_{Z=z})+1}\right] \le 3\varepsilon
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}_{z\sim Z}\left[2^{-\mathrm{H}_\infty(X|_{Z=z})}\right] = 2^{-\widetilde{\mathrm{H}}_\infty(X|Z)}$ by definition of conditional min-entropy. □

Proof of Theorem 2.3.12. By the previous lemma, it suffices to prove that for every $t \ge 0$, Ext is a

$(k-t,(2^{t+1}-1)\cdot\varepsilon)$ extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$. Since $\mathrm{D}^{\mathcal{F}}$ is convex in its first argument by Lemma 2.2.13, following Chor and Goldreich [CG88] it is enough to consider only distributions with min-entropy $k-t$ that are supported on a set of at most 2^{n-1}. Fix such a distribution X and a collection of test functions $f_1,\dots,f_D\in\mathcal{F}$ with $f_1=\cdots=f_D$ if Ext is not strong. Then since X is supported on a set of size at most 2^{n-1}, the distribution Y that is uniform over the complement of $\mathrm{Supp}(X)$ has min-entropy at least $n-1\geq k$, and furthermore the mixture $2^{-t}X+(1-2^{-t})Y$ has min-entropy at least k. Hence, as Ext is a (k,ε) extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$,

$$\begin{aligned}
\varepsilon &\geq \mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}\left(\mathrm{Ext}\left(2^{-t}X+(1-2^{-t})Y,i\right)\parallel U_m\right)\right]\\
&= 2^{-t}\mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i)\parallel U_m)\right]+(1-2^{-t})\mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(Y,i)\parallel U_m)\right]\\
&= 2^{-t}\mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i)\parallel U_m)\right]-(1-2^{-t})\mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{c_i-f_i\}}(\mathrm{Ext}(Y,i)\parallel U_m)\right]\\
&\geq 2^{-t}\mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i)\parallel U_m)\right]-(1-2^{-t})\cdot\varepsilon \qquad (\text{since } \mathrm{H}_\infty(Y)\geq k)\\
\left(2^{t+1}-1\right)\cdot\varepsilon &\geq \mathop{\mathbb{E}}_{i\sim U_{[D]}}\left[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i)\parallel U_m)\right]
\end{aligned}$$

where $c_i\in\mathbb{R}$ is such that $c_i-f_i\in\mathcal{F}$ as guaranteed to exist by the symmetry of $\mathcal{F}$. □

2.4 Subgaussian distance and connections to other notions

Now that we've introduced the general machinery we need, we can go back to our motivation of subgaussian samplers. We will need some standard facts about subgaussian and subexponential random variables, we recommend the book of Vershynin [Ver18] for an introduction.

Definition 2.4.1. A real-valued mean-zero random variable Z is said to be *subgaussian with parameter*

σ if for every $t \in \mathbb{R}$ the moment generating function of Z is bounded as

$$\ln \mathbb{E}\left[e^{tZ}\right] \leq \frac{t^2\sigma^2}{2}.$$

If this is only holds for $|t| \leq b$ then Z is said to be (σ, b)-*subgamma,* and if Z is $(\sigma, 1/\sigma)$-subgamma then Z is said to be *subexponential with parameter* σ.

Remark 2.4.2. There are many definitions of subgaussian (and especially subexponential) random variables in the literature, but they are all equivalent up to constant factors in σ and only affect constants already hidden in big-O's.

Lemma 2.4.3. *Let Z be a real-valued random variable. Then*

1. *(Hoeffding's lemma) If Z is bounded in the interval $[0, 1]$, then $Z - \mathbb{E}[Z]$ is subgaussian with parameter $1/2$.*

2. *If Z is mean-zero, then Z is subgaussian (respectively subexponential) with parameter σ if and only if cZ is subgaussian (respectively subexponential) with parameter $|c|\sigma$ for every $c \neq 0$.*

Furthermore, if Z is mean-zero and subgaussian with parameter σ, then

1. *For all $t > 0$, $\max\big(\Pr[Z > t], \Pr[Z < -t]\big) \leq e^{-t^2/2\sigma^2}$.*

2. *$\|Z\|_p \stackrel{\text{def}}{=} \mathbb{E}\left[|Z|^p\right]^{1/p} \leq 2\sigma\sqrt{p}$ for all $p \geq 1$.*

3. *Z is subexponential with parameter σ.*

We are now in a position to formally define the *subgaussian distance*.

Definition 2.4.4. For every finite set $\mathcal{X}$, we define the set $\mathcal{G}_\mathcal{X}$ of *subgaussian test functions on* $\mathcal{X}$ (respectively the set $\mathcal{E}_\mathcal{X}$ of *subexponential test functions on* $\mathcal{X}$) to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that

the random variable $f(U_{\mathcal{X}})$ is mean-zero and subgaussian (respectively subexponential) with parameter 1/2. Then $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are symmetric and distinguishing, so by Lemma 2.2.13 the respective distances induced by $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are jointly convex metrics called the *subgaussian distance* and *subexponential distance* respectively and are denoted as $d_{\mathcal{G}}(P, Q)$ and $d_{\mathcal{E}}(P, Q)$.

Remark 2.4.5. We choose subgaussian parameter 1/2 in Definition 2.4.4 as by Hoeffding's lemma, all functions $f : \{0,1\}^m \to [0,1]$ have that $f(U_m) - \mathbb{E}[f(U_m)]$ is subgaussian with parameter 1/2, so this choice preserves the same "scale" as total variation distance. However, the choice of parameter is essentially irrelevant by linearity, as different choices of parameter simply scale the metric $d_{\mathcal{G}}$.

Note that absolute averaging samplers for $\mathcal{G}_{\{0,1\}^m}$ from Definition 2.3.6 are exactly subgaussian samplers as defined in the introduction. Thus, by Remark 2.3.7 and Theorem 2.3.8, to construct subgaussian samplers it is enough to construct extractors for the subgaussian distance $d_{\mathcal{G}}$.

2.4.1 Composition

Unfortunately, the subgaussian distance has a major disadvantage compared to total variation distance that complicates extractor construction: it does not satisfy the data-processing inequality, that is, there are probability distributions P and Q over a set A and a function $f : A \to B$ such that

$$d_{\mathcal{G}}(f(P), f(Q)) \not\leq d_{\mathcal{G}}(P, Q).$$

This happens because subgaussian distance is defined by functions which are required to be subgaussian only with respect to the *uniform distribution*. A simple explicit counterexample comes from taking $f : \{0,1\}^1 \to \{0,1\}^m$ defined by $x \mapsto (x, 0^{m-1})$ and taking P to be the point mass on 0 and Q the point mass on 1. Their subgaussian distance in $\{0,1\}^1$ is obviously $O(1)$, but the subgaussian distance of $f(P)$ and $f(Q)$ in $\{0,1\}^m$ is $\Theta(\sqrt{m})$.

The reason this matters is because a standard operation (cf. Nisan and Zuckerman [NZ96], Goldreich and Wigderson [GW97], and Reingold, Vadhan, and Wigderson [RVW00]) in the construction of samplers and extractors for bounded functions is to do the following: given extractors

$$\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$

$$\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d,$$

define $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ by

$$\mathrm{Ext}\big((x,y),s\big) = \mathrm{Ext}_{out}\big(x, \mathrm{Ext}_{in}(y,s)\big).$$

The reason this works for total variation distance is exactly the data-processing inequality: if Y has enough min-entropy given X, then $\mathrm{Ext}_{in}(Y, U_{d'})$ will be close in total variation distance to U_d, and by the data-processing inequality for total variation distance this closeness is not lost under the application of Ext_{out}. The assumption that Y has min-entropy given X means that (X,Y) is a so-called *block-source*, and is implied by (X,Y) having enough min-entropy as a joint distribution. From the sampler perspective, this construction uses the inner sampler Ext_{in} to subsample the outer sampler. On the other hand, for subgaussian distance, the distribution $\mathrm{Ext}_{in}(Y, U_{d'})$ can be ε-close to uniform but still have some element with excess probability mass $\Omega(\varepsilon/\sqrt{d})$, and this element (seed) when mapped by Ext_{out} can retain[2] this excess mass in $\{0,1\}^m$, which results in subgaussian distance $\Theta(\varepsilon\sqrt{m/d}) \gg \varepsilon$. Similarly, from the sampler perspective, even when the outer sampler Ext_{out} is a good subgaussian sampler for $\{0,1\}^m$, there is no reason that a good subgaussian sampler Ext_{in} for $\{0,1\}^d$ the seeds of Ext_{out} will preserve the larger sampler property when $m \gg d$.

[2] Given a subgaussian extractor Ext with $d \geq \lg(m/\varepsilon)$, adding a single extra seed $*$ to Ext such that $\mathrm{Ext}(x,*) = 0^m$ results in a subgaussian extractor with error at most $2^{-d} \cdot \sqrt{2m} + \varepsilon \leq 3\varepsilon$ by convexity of $d_{\mathcal{G}}$ and the fact that $\left\| d_{\mathcal{G}_{\{0,1\}^m}} \right\|_\infty < \sqrt{2m}$.

Thus, since this composition operation is used in all existing constructions of high-min entropy extractors for total variation distance with the desired seed length, to construct such extractors for subgaussian distance we need to bypass this barrier. The natural approach is to construct extractors for a better-behaved weak divergence that bounds the subgaussian distance.

Remark 2.4.6. Similar reasoning shows that if Ext is a strong (k, ε) subgaussian extractor, then it is not necessarily the case that the function $(x, s) \mapsto (s, \mathrm{Ext}(x, s))$ that prepends the seed to the output is a (non-strong) (k, ε) subgaussian extractor (in contrast to extractors for total variation distance), though the converse does hold.

2.4.2 Connections to other weak divergences

Therefore, to aid in extractor construction, we show how $d_{\mathcal{G}}$ relates to other statistical weak divergences.

Most basically, the subgaussian distance over $\{0,1\}^m$ differs from total variation distance up to a factor of $O(\sqrt{m})$.

Lemma 2.4.7. *Let P and Q be distributions on $\{0,1\}^m$. Then*

$$d_{\mathrm{TV}}(P, Q) \leq d_{\mathcal{G}}(P, Q) \leq \sqrt{2 \ln 2 \cdot m} \cdot d_{\mathrm{TV}}(P, Q)$$

Proof. That $d_{\mathrm{TV}} \leq d_{\mathcal{G}}$ is immediate from Hoeffding's lemma and the discussion in Remark 2.4.5. The reverse bound holds since any subgaussian function takes values at most $\sqrt{\ln 2/2 \cdot m}$ away from the mean by the tail bounds from part 3 of Lemma 2.4.3, and so any subgaussian test function f has the property that $1/2 + f/\sqrt{2 \ln 2 \cdot m}$ is $[0,1]$-valued and thus lower bounds the total variation distance. □

While this allows constructing subgaussian extractors and samplers from total variation extractors,

as discussed in the introduction the fact that the upper bound depends on m leads to suboptimal bounds. By starting with a stronger measure of error, we pay a much smaller penalty.

Lemma 2.4.8. *Let P and Q be distributions on $\{0,1\}^m$. Then for every $\alpha > 0$*

$$2d_{\mathrm{TV}}(P,Q) = d_{\ell_1}(P,Q) \leq 2^{m\alpha/(1+\alpha)} \cdot d_{\ell_{1+\alpha}}(P,Q)$$
$$d_{\mathcal{G}}(P,Q) \leq 2^{m\alpha/(1+\alpha)} \sqrt{1+\frac{1}{\alpha}} \cdot d_{\ell_{1+\alpha}}(P,Q)$$

In particular, that there is only an additional $\sqrt{1+1/\alpha}$ factor when moving to subgaussian distance compared to total variation, which in particular does not depend on m and is constant for constant α.

Proof. By Example 2.2.9, for any function $f : \{0,1\}^m \to \mathbb{R}$ it holds that

$$\mathrm{D}^{\{f\}}(P \parallel Q) \leq \|f(U_m)\|_{1+\frac{1}{\alpha}} \cdot d_{\mathcal{M}_{1+\frac{1}{\alpha}}}(P,Q) = \|f(U_m)\|_{1+\frac{1}{\alpha}} \cdot 2^{m\alpha/(1+\alpha)} \cdot d_{\ell_{1+\alpha}}(P,Q).$$

The result follows since $[-1,1]$-valued functions f satisfy moment bounds $\|f(U_m)\|_q \leq 1$ for all $q \geq 1$, and functions f which are subgaussian satisfy moment bounds $\|f(U_m)\|_q \leq \sqrt{q}$ by Lemma 2.4.3. □

One downside of starting with bounds on $\ell_{1+\alpha}$ is that, extending a well-known linear seed length linear bound for ℓ_2-extractors (e.g. [Vad12, Problem 6.4]), we show in Corollary 2.5.29 that for every $1 > \alpha > 0$, there is a constant $c_\alpha > 0$ such any $\ell_{1+\alpha}$ extractor with error smaller than $c_\alpha \cdot 2^{-m\alpha/(1+\alpha)}$ requires seed length linear in $\alpha \cdot \min(n-k, m)$, for $n-k$ the entropy deficiency and m the output length. One might hope that sending α to 0 would eliminate this linear lower bound but still bound the subgaussian distance, but phrased this way sending α to 0 just results in a total variation extractor.

However, with a shift in perspective essentially the same approach works: by Example 2.2.4, $d_{\ell_2}(P, U_m) \leq \varepsilon \cdot 2^{-m/2}$ implies $\mathrm{D}_2(P \parallel U_m) \leq \varepsilon^2/\ln 2$, and there is an analogous linear seed length lower bound on constant error $\mathrm{D}_{1+\alpha}$ extractors for every $\alpha > 0$. In this case, however, sending α to 0

results in the *KL divergence*, which does upper bound the subgaussian distance, and in fact with the same parameters as for total variation distance.

Lemma 2.4.9. *Let P be a distribution on $\{0,1\}^m$. Then*

$$d_{\mathcal{G}}(P, U_m) \le \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)}$$

$$d_{\mathcal{E}}(P, U_m) \le \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \le \frac{1}{2\ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2\ln 2} \end{cases}$$

where these bounds are concave in $\mathrm{KL}(P \parallel U_m)$. In the reverse direction, it holds that

$$\mathrm{KL}(P \parallel U_m) \le m \cdot d_{\mathrm{TV}}(P, U_m)$$

Proof. The upper bounds follow from a general form of Pinsker's inequality as in [BLM13, Lemma 4.18], but for completeness we include its proof here in these special cases, based on the Donsker–Varadhan "variational" formulation of KL divergence [DV76, Theorem 5.2]:

$$\mathrm{KL}(P \parallel U_m) = \frac{1}{\ln 2} \cdot \sup_{g:\{0,1\}^m \to \mathbb{R}} \Big(\mathbb{E}[g(P)] - \ln \mathbb{E}\big[e^{g(U_m)}\big] \Big).$$

Now if $f : \{0,1\}^m \to \mathbb{R}$ satisfies $\mathbb{E}[f(U_m)] = 0$, then by letting $g(x) = t \cdot f(x)$, this implies

$$\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] = \frac{1}{t} \cdot \mathbb{E}[g(P)] \le \frac{\ln 2 \cdot \mathrm{KL}(P \parallel U_m) + \ln \mathbb{E}\big[e^{t \cdot f(U_m)}\big]}{t}$$

for all $t > 0$. Thus, when $\ln \mathbb{E}\big[e^{t \cdot f(U_m)}\big] \le t^2/8$, we have $\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] \le \ln 2 \cdot \mathrm{KL}(P \parallel U_m)/t + t/8$.

Then since subgaussian random variables satisfy such a bound for all t, we can make the optimal choice $t = \sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}$ to get the claimed bound on $d_{\mathcal{G}}$. For subexponential random variables,

which satisfy such a bound only for $|t| \leq 2$, we choose $t = \min(\sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}, 2)$, which gives

$$d_{\mathcal{E}}(P, U_m) \leq \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \leq \frac{1}{2 \ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2 \ln 2} \end{cases}$$

as desired. The concavity of this bound follows by noting that it has a continuous and nonincreasing derivative.

The reverse inequality is a special case of the Reverse Pinsker's inequality of Verdú [Ver14, Theorem 7]. □

Remark 2.4.10. In Chapter 5 we give a general framework that provides a more principled and intuitive derivation of the relationship between the subgaussian distance and the Kullback–Leibler divergence.

2.5 Extractors for KL divergence

By Lemma 2.4.9, the subgaussian distance can be bounded in terms of the KL divergence to uniform, so by the following easy lemma to construct subgaussian extractors it suffices to construct extractors for KL divergence.

Lemma 2.5.1. *Let V_1 and V_2 be weak divergences on the set $\{0,1\}^m$ and $f : \mathbb{R} \to \mathbb{R}$ be a function such that $V_1(P \parallel U_M) \leq f(V_2(P \parallel U_m))$ for all distributions P on $\{0,1\}^m$. Then if f is increasing on $(0, \varepsilon)$, every (k, ε) extractor* Ext *for V_1 is also a $(k, f(\varepsilon))$-extractor for V_2, and if f is also concave, then if* Ext *is strong or average-case as a V_1-extractor, it has the same properties as a $(k, f(\varepsilon))$ extractor for V_2.*

Importantly, the KL divergence does not have the flaws of subgaussian distance discussed in Section 2.4.1. The classic *data-processing inequality* says that KL divergence is non-increasing under

postprocessing by (possibly randomized) functions, and the *chain rule* for KL divergence says that

$$\mathrm{KL}(A, B \parallel X, Y) = \mathrm{KL}(A \parallel X) + \mathop{\mathbb{E}}_{a \sim A}[\mathrm{KL}(B|_{A=a} \parallel Y|_{X=a})]$$

for all distributions A, B, X, and Y, so that in particular

$$\mathop{\mathbb{E}}_{s \sim U_d}[\mathrm{KL}(\mathrm{Ext}(X, s) \parallel U_m)] = \mathrm{KL}(U_d, \mathrm{Ext}(X, U_d) \parallel U_d, U_m)$$

and prepending the seed of a strong KL-extractor does in fact give a non-strong KL-extractor:

Lemma 2.5.2. *A function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k, ε) *strong KL-extractor (respectively strong average-case KL-extractor) if and only if the function* $\mathrm{Ext}' : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{d+m}$ *defined by* $\mathrm{Ext}'(x, s) = (s, \mathrm{Ext}(x, s))$ *is a (non-strong)* (k, ε) *KL-extractor (respectively average-case KL-extractor).*

Furthermore, KL divergence satisfies a type of triangle inequality when combined with higher Rényi divergences:

Lemma 2.5.3. *Let P, Q, and R be distributions over a finite set $\mathcal{X}$. Then for all $\alpha > 0$, it holds that*

$$\mathrm{KL}(P \parallel R) \leq \left(1 + \frac{1}{\alpha}\right) \cdot \mathrm{KL}(P \parallel Q) + \mathrm{D}_{1+\alpha}(Q \parallel R)$$

Proof. This follows from a characterization of Rényi divergence due to van Erven and Harremoës [vErv10, Lemma 6.6] [vEH14, Theorem 30] and Shayevitz [Sha11, Theorem 1], who prove that for for every positive real $\beta \neq 1$ and distributions X and Y that

$$(1 - \beta)\, \mathrm{D}_\beta(X \parallel Y) = \inf_Z \{\beta\, \mathrm{KL}(Z \parallel X) + (1 - \beta)\, \mathrm{KL}(Z \parallel Y)\}.$$

In particular, choosing $\beta = 1 + \alpha$, $X = Q$, and $Y = R$ and upper bounding the infimum by the particular choice of $Z = P$ gives the claim. □

2.5.1 Composition

These properties imply that composition does work as we want (without any loss depending on the output length m) assuming we have extractors for KL and higher divergences.

Theorem 2.5.4 (Composition for high min-entropy Rényi entropy extractors, cf. [GW97])**.** *Suppose*

1. $\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is an* $(n - \lg(1/\delta), \varepsilon_{out})$ *extractor for* $\mathrm{D}_{1+\alpha}$ *with* $\alpha > 0$*,*

2. $\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d$ *is an* $(n' - \lg(1/\delta), \varepsilon_{in})$ *average-case KL-extractor,*

and define $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ *by* $\mathrm{Ext}\big((x,y),s\big) = \mathrm{Ext}_{out}(x, \mathrm{Ext}_{in}(y,s))$*. Then* Ext *is an* $\big(n + n' - \lg(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in}\big)$ *extractor for KL. Furthermore, if* Ext_{in} *is a strong average-case KL-extractor, then* Ext *is a strong KL-extractor, and if* Ext_{out} *is average-case then so is* Ext*.*

Proof. Let (Z, X, Y) be jointly distributed random variables with X distributed over $\{0,1\}^n$ and Y over $\{0,1\}^{n'}$ such that $\widetilde{\mathrm{H}}_\infty(X, Y|Z) \geq n + n' - \lg(1/\delta)$. Let S' be a distribution over $\{0,1\}^{d'}$ which is independent of X, Y, and Z. Then for every $z \in \mathrm{Supp}(Z)$, we have by Lemma 2.5.3 and the data-processing inequality for KL divergence that

$$
\begin{aligned}
&\mathrm{KL}(\mathrm{Ext}((X|_{Z=z}, Y|_{Z=z}), S') \parallel U_m) \\
&\quad = \mathrm{KL}(\mathrm{Ext}_{out}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S')) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(\mathrm{Ext}_{out}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S')) \parallel \mathrm{Ext}_{out}(X|_{Z=z}, U_d)) \\
&\qquad + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S') \parallel X|_{Z=z}, U_d) + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m) \\
&\quad = (1 + 1/\alpha) \cdot \mathop{\mathbb{E}}_{x \sim X|_{Z=z}} \big[\mathrm{KL}\big(\mathrm{Ext}_{in}(Y|_{X=x,Z=z}, S') \parallel U_d\big)\big] + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m)
\end{aligned}
$$

where the last equality follows from the chain rule for KL divergence. Now by standard properties

of conditional min-entropy (see for example [DORS08, Lemma 2.2]), we know that $\widetilde{\mathrm{H}}_\infty(X|Z) \geq \widetilde{\mathrm{H}}_\infty(X,Y|Z) - \lg|\mathrm{Supp}(Y)| \geq n - \lg(1/\delta)$ and $\widetilde{\mathrm{H}}_\infty(Y|X,Z) \geq \widetilde{\mathrm{H}}_\infty(X,Y|Z) - \lg|\mathrm{Supp}(X)| \geq n' - \lg(1/\delta)$.

If Ext_{out} is not average-case, take Z to be a constant independent of X and Y, and if Ext_{out} is average-case then take the average of both sides over Z. The claim for non-strong Ext_{in} then follows by taking $S' = U_d$ which bounds the first term by $(1 + 1/\alpha) \cdot \varepsilon_{in}$ and the second by ε_{out}. The claim for strong Ext_{in} follows by choosing $S' = U_{\{s\}}$ to be the point mass on $s \in \{0,1\}^d$ and then taking the expectation of both sides over a uniform $s \in \{0,1\}^d$. □

Remark 2.5.5. Theorem 2.5.4 in fact a construction of a *block-source* KL-extractor, meaning that the claimed error bounds hold for any joint distributions (X,Y) such that $\mathrm{H}_\infty(Y) \geq n' - \lg(1/\delta)$ and $\widetilde{\mathrm{H}}_\infty(X|Y) \geq n - \lg(1/\delta)$ rather than just those distributions with $\mathrm{H}_\infty(X,Y) \geq n + n' - \lg(1/\delta)$. The extra $\lg(1/\delta)$ entropy loss inherent in the non-block analysis is why Reingold, Vadhan, and Wigderson [RVW00] introduced the zig-zag product for extractors, which we will apply for KL-extractors in Corollary 2.5.19.

2.5.2 Existing explicit constructions

The construction of Theorem 2.5.4 required both a $\mathrm{D}_{1+\alpha}$-extractor and an average-case KL-extractor, so for the result not to be vacuous we need to show the existence of such extractors. Thankfully, Example 2.2.4 implies that extractors for ℓ_2 are also extractors for D_2, so we can use existing ℓ_2 extractors from the literature, such as the Leftover Hash Lemma of Impagliazzo, Levin, and Luby [ILL89] (see also [McI87; BBR88]) and its variant using almost-universal hash functions due to Srinivasan and Zuckerman [SZ99].

Proposition 2.5.6 ([McI87; BBR88; ILL89; IZ89; SZ99; DORS08]). *Let $\mathcal{H}$ be a collection of ε-almost universal hash functions from the set $\{0,1\}^n$ to the set $\{0,1\}^m$, meaning that for all $x \neq y \in \{0,1\}^n$ it holds that $\Pr_{h\sim\mathcal{H}}[h(x) = h(y)] \leq (1+\varepsilon)/2^m$. Then the function $\mathrm{Ext} : \{0,1\}^n \times \mathcal{H} \to \mathcal{H} \times \{0,1\}^m$ defined by $\mathrm{Ext}(x,h) = (h, h(x))$ is an average-case $(m + \lg(1/\varepsilon), 2/\ln 2 \cdot \varepsilon)$ D_2-extractor.*

In particular, for every $k, n \in \mathbb{N}$ and $1 > \varepsilon > 0$ there is an explicit strong average-case (k,ε) extractor for D_2 (and KL) with seed length $d = O(k + \lg(n/\varepsilon))$ and $m = k - \lg(1/\varepsilon) - O(1)$, given by $\mathrm{Ext}'(x,h) = h(x)$ for h drawn from an appropriate almost-universal hash family.

Proof. The D_2 claim is implicit in Rackoff's proof of the Leftover Hash Lemma (see [IZ89]) and Srinivasan and Zuckerman's proof of the claim for total variation [SZ99], which both analyzed the *collision probability* of the output, and the average-case claim was proved by Dodis et al. [DORS08], though we include a proof here for completeness.

Given a joint distribution (Z,X) such that X is distributed over $\{0,1\}^n$ with $\widetilde{\mathrm{H}}_\infty(X|Z) \geq m + \lg(1/\varepsilon)$, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim Z}[\mathrm{D}_2(\mathrm{Ext}(X|_{Z=z}, \mathcal{H}) \parallel \mathcal{H} \times U_m)] \\
&\quad= \mathop{\mathbb{E}}_{z\sim Z}\left[\lg\left(2^m \cdot |\mathcal{H}| \cdot \Pr_{h,h'\sim\mathcal{H}, x, x'\sim X|_{Z=z}}[(h,h(x)) = (h',h'(x'))]\right)\right] \\
&\quad= \mathop{\mathbb{E}}_{z\sim Z}\left[\lg\left(2^m \cdot \Pr_{h\sim\mathcal{H}, x, x'\sim X|_{Z=z}}\left[x = x' \vee \left(x \neq x' \wedge h(x) = h(x')\right)\right]\right)\right] \\
&\quad\leq \mathop{\mathbb{E}}_{z\sim Z}\left[\lg\left(2^m \cdot \left(2^{-\mathrm{H}_\infty(X|_{Z=z})} + \frac{1+\varepsilon}{2^m}\right)\right)\right] \\
&\quad\leq \lg\left(\mathop{\mathbb{E}}_{z\sim Z}\left[2^{m-\mathrm{H}_\infty(X|_{Z=z})}\right] + 1 + \varepsilon\right) \qquad \text{(by Jensen's inequality)} \\
&\quad= \lg\left(2^{m-\widetilde{\mathrm{H}}_\infty(X|Z)} + 1 + \varepsilon\right) \leq \lg(1+2\varepsilon) \leq \frac{2}{\ln 2} \cdot \varepsilon.
\end{aligned}
$$

The in particular statement follows from Lemma 2.5.7 below and from the existence of ε-almost universal hash families with size $\mathrm{poly}(2^k, n, 1/\varepsilon)$ as constructed by [SZ99]. □

To establish the claim about strong extractors, we generalize Lemma 2.5.2 to extractors for $\mathrm{D}_{1+\alpha}$ for $\alpha > 0$:

Lemma 2.5.7. *If* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^d \times \{0,1\}^m$ *is a* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor (respectively average-case* $\mathrm{D}_{1+\alpha}$*-extractor) for* $\alpha > 0$ *such that* $\mathrm{Ext}(x,s) = (s, \mathrm{Ext}'(x,s))$*, then* Ext' *is a strong* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor (respectively strong average-case* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor).*

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{s\sim U_d}\left[\mathrm{D}_{1+\alpha}\left(\mathrm{Ext}'(X,s) \parallel U_m\right)\right] &= \mathop{\mathbb{E}}_{s\sim U_d}\left[\frac{1}{\alpha}\lg\left(2^{m\alpha}\sum_{y\in\{0,1\}^m}\Pr\left[\mathrm{Ext}'(X,s)=y\right]^{1+\alpha}\right)\right] \\
&\le \frac{1}{\alpha}\lg\left(2^{m\alpha}\mathop{\mathbb{E}}_{s\sim U_d}\left[\sum_{y\in\{0,1\}^m}\Pr\left[\mathrm{Ext}'(X,s)=y\right]^{1+\alpha}\right]\right) \\
&= \frac{1}{\alpha}\lg\left(2^{\alpha(m+d)}\sum_{(s,y)\in\{0,1\}^{d+m}}\Pr\left[(U_d,\mathrm{Ext}'(X,U_d))=(s,y)\right]^{1+\alpha}\right) \\
&= \mathrm{D}_{1+\alpha}(\mathrm{Ext}(X,U_d) \parallel U_d, U_m)
\end{aligned}$$

□

Following Vadhan [Vad12], we also note that the extractor based on expander walks due to Goldreich and Wigderson [GW97], which has the nice property that its seed length depends only on $n-k$ the entropy deficiency of the source rather than n itself, is also an ℓ_2 extractor. Before stating the extractor formally, we introduce some notation and terminology we will need.

Definition 2.5.8. Let G be a D-regular graph on $\{0,1\}^n$ with adjacency matrix A_G and transition matrix $M_G = \frac{1}{D}A_G$. Then if M_G has eigenvalues $1 = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{2^n} \ge -1$, the *spectral expansion* of G is $\lambda = \max\{\lambda_2, -\lambda_{2^n}\}$. A function $\Gamma_G : \{0,1\}^n \times [D] \to \{0,1\}^n$ is a *neighbor function of* G if there is some labelling of the edges of G for which $\Gamma_G(v,i)$ is the vertex obtained by following the ith edge out of v in G. Γ_G is *consistently labelled* if for all $v \ne v' \in \{0,1\}^n$ and $i \in [D]$ we have $\Gamma(v,i) \ne \Gamma(v',i)$, that is, at most one incoming edge is labelled by i.

Lemma 2.5.9. *Let* $\Gamma : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^n$ *be the neighbor function of a graph* G *with spectral expansion* λ*. Then for every* $0 \le k \le n$*,* Γ *is a* $\left(k, \lambda\sqrt{2^{-k} - 2^{-n}}\right)$ ℓ_2*-extractor and a* $\left(k, \lg\left(1 + \lambda^2\left(2^{n-k} - 1\right)\right)\right)$ D_2*-extractor. Furthermore, if* Γ_G *is consistently labelled, then the function* $\mathrm{W}(x, s) = s$ *is such that* (Γ_G, W) *is an injection out of* $\{0,1\}^n \times \{0,1\}^d$*.*

In particular, if $\lambda^2 \le \varepsilon \cdot 2^{k-n}$ *then* Ext *is an average-case* $(k, \sqrt{\varepsilon} \cdot 2^{-n/2})$ ℓ_2*-extractor and an average-case* $(k, \varepsilon/\ln 2)$ D_2*-extractor.*

Proof. If X is a distribution over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \ge k$, then $\lg\left(1 + 2^n d_{\ell_2}(X, U_n)\right) = \mathrm{D}_2(X \parallel U_n) \le \mathrm{D}_\infty(X \parallel U_n) \le n - k$ so that $d_{\ell_2}(X, U_n) \le \sqrt{2^{-k} - 2^{-n}}$. Then identifying X and U_n with the column vector representation of their probability mass functions, we have

$$d_{\ell_2}\left(\mathrm{Ext}(X, U_d), U_n\right) = \left\|\Gamma_G X - U_n\right\|_2 = \left\|\Gamma_G(X - U_n)\right\|_2 \le \lambda \left\|X - U_n\right\|_2,$$

where the first equality is by definition of Ext, the second is because Γ_G sends the identity to itself since G is regular, and the inequality is because $X - U_n$ is orthogonal to the all-ones vector and Γ_G shrinks all such vectors by a factor of at least λ by definition. This completes the proof of the ℓ_2 extraction claim, and the D_2-extraction claim follows from the fact that $\mathrm{D}_2(Y \parallel U_n) = \lg\left(1 + 2^n d_{\ell_2}(Y, U_n)^2\right)$ for every distribution Y on $\{0,1\}^n$.

For the furthermore claim, we need to show that $(x, s) \mapsto (\Gamma_G(x, s), s)$ is an injection, or equivalently that given $\Gamma_G(x, s)$ and s, one can recover x. But by definition of consistent labelling, at most one edge into $\Gamma_G(x, s)$ is labelled by s, and so taking this edge from $\Gamma_G(x, s)$ gives x, as desired.

Finally, for the in particular claim, we have for any joint distributions (X, Z) with X distributed over $\{0,1\}^n$, we have

$$\mathop{\mathbb{E}}_{z \sim Z}\left[d_{\ell_2}\left(\mathrm{Ext}(X|_{Z=z}, U_d), U_n\right)\right] \qquad \mathop{\mathbb{E}}_{z \sim Z}\left[\mathrm{D}_2(\mathrm{Ext}(X|_{Z=z}, U_d \parallel U_n)\right]$$

$$\leq \mathop{\mathbb{E}}_{z\sim Z}\left[\lambda\sqrt{2^{-\mathrm{H}_\infty(X|_{Z=z})}-2^{-n}}\right] \qquad \leq \mathop{\mathbb{E}}_{z\sim Z}\left[\lg\left(1+\lambda^2\left(2^{n-\mathrm{H}_\infty(X|_{Z=z})}-1\right)\right)\right]$$

$$\leq \lambda\sqrt{\mathop{\mathbb{E}}_{z\sim Z}\left[2^{-\mathrm{H}_\infty(X|_{Z=z})}-2^{-n}\right]} \qquad \leq \lg\left(1+\lambda^2\left(\mathop{\mathbb{E}}_{z\sim Z}\left[2^{n-\mathrm{H}_\infty(X|_{Z=z})}-1\right]\right)\right)$$

$$= \lambda\sqrt{2^{-\widetilde{\mathrm{H}}_\infty(X|Z)}-2^{-n}} \qquad = \lg\left(1+\lambda^2\left(2^{n-\widetilde{\mathrm{H}}_\infty(X|Z)}-1\right)\right)$$

where the first inequality is by the main claim, the second by Jensen's inequality, and the equality is by definition of conditional min-entropy. □

Remark 2.5.10. The fact that $(s, \mathrm{Ext}(x,s))$ is an injection implies that, unlike for the extractors from hashing of Proposition 2.5.6, the result of prepending the seed to the output of the expander-walk extractor does *not* give a D_2 extractor. However, it will be very useful in concert with Reingold, Vadhan, and Wigderson's zig-zag product for extractors [RVW00] to avoid the entropy loss in Theorem 2.5.4.

Corollary 2.5.11 ([GW97] [Vad12, Discussion after Theorem 6.22]). *There is a universal constant $C \geq 1$ such that for every $1 > \varepsilon > 0$, $\Delta > 0$, and $n \in \mathbb{N}$ there is an explicit $(n-\Delta, \varepsilon/\ln 2)$ average-case D_2-extractor (respectively $(n-\Delta, \sqrt{\varepsilon}\cdot 2^{-n/2})$ average-case ℓ_2-extractor) $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^n$ with $d = \lceil C\cdot(\Delta + \lg(1/\varepsilon))\rceil + O(1)$ such that the function $(x,s)\mapsto(s,\mathrm{Ext}(x,s))$ is an injection.*

Moreover, if there is an explicit construction of consistently labelled neighbor functions for Ramanujan graphs over $\{0,1\}^n$ with degree $D = O(2^\Delta/\varepsilon)$, then one can take $C = 1$.

Proof. By Lemma 2.5.9 it suffices to demonstrate the existence of an explicit D-regular expander graph over $\{0,1\}^n$ with a consistently labelled neighbor function Γ_G, spectral expansion $\lambda^2 \leq \varepsilon\cdot 2^{-\Delta}$, and $D = O\left(\left(2^\Delta/\varepsilon\right)^C\right)$. The claim about Ramanujan graphs is thus immediate since a Ramanujan graph with degree $O(2^\Delta/\varepsilon)$ has $\lambda^2 \leq 4/D \leq \varepsilon\cdot 2^{-\Delta}$.

Without the assumption of good Ramanujan graphs, we can use a power of the the explicit constant degree expander of Margulis–Gabber–Galil [Mar73; GG81] (technically this requires n even, which

following Goldreich [Gol11b] we can fix when n is odd by joining two graphs on $\{0,1\}^{n-1}$ by the canonical perfect matching, and we can add self-loops to ensure the degree is a power of 2). This graph G is consistently labelled with degree $D_{MGG} = O(1)$ and constant spectral expansion $\lambda_{MGG} < 1$. Then the graph G^w on $\{0,1\}^n$ with edges representing w-length paths has spectral expansion λ_{MGG}^w and degree D_{MGG}^w, which for $w = \left\lceil \log_{\lambda_{MGG}}(1/2) \cdot \left(\Delta + \lg(1/\varepsilon)\right) \right\rceil$ gives $\lambda \le \varepsilon \cdot 2^{-\Delta}$ and degree $D = O\left(\left(2^{\Delta}/\varepsilon\right)^C\right)$ for $C \le \lg(D_{MGG}) \cdot \log_{\lambda_{MGG}}(1/2)$ as desired. □

We argued that the above extractors are KL-extractors using the fact they are ℓ_2 (and thus D_2) extractors, but one can also show that any total variation extractor with sufficiently small error is a KL-extractor, albeit with some loss of parameters.

Lemma 2.5.12. *If* Ext $: \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k,ε) *extractor for total variation distance, then* Ext *is also a* $(k, m \cdot \varepsilon)$*-KL-extractor. Furthermore, if* Ext *is strong, average-case, or both as a total variation extractor, then it has the same properties as a KL-extractor.*

In particular, every $(k, \varepsilon/(3m))$ *extractor (respectively strong extractor) is an average-case* (k,ε) *KL-extractor (respectively strong average-case* (k,ε) *KL-extractor).*

Proof. The main claim is an immediate corollary of Lemmas 2.4.9 and 2.5.1, and the in particular then follows from Theorem 2.3.12. □

Remark 2.5.13. Reducing ε by a factor of $3m$ increases the seed length and entropy loss of the input extractor. For the former, this is often (but not always) tolerable since the input extractor may already depend suboptimally on $\lg(n/\varepsilon)$. For the latter, we will show in Corollary 2.5.21 how to use the transform of Raz, Reingold, and Vadhan [RRV02] to recover $O(\lg(m/\varepsilon))$ bits of lost entropy (at least this much must be lost by Radhakrishnan and Ta-Shma [RT00]) at a cost of $O(\lg(n/\varepsilon))$ in the seed length.

Instantiating Lemma 2.5.12 with the Guruswami–Umans–Vadhan [GUV09] extractor for total variation distance, we see that the increased seed length and entropy loss are simply absorbed into the existing hidden constants:

Theorem 2.5.14 (KL-analogue of [GUV09, Theorem 1.5])**.** *For every $n \in \mathbb{N}$, $k \leq n$, and $1 > \alpha, \varepsilon > 0$, there is an explicit average-case (respectively strong average-case) (k, ε) KL-extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d \leq \lg n + O_\alpha(\lg(k/\varepsilon))$ and $m \geq (1-\alpha)k$ (respectively $m \geq (1-\alpha)k - O_\alpha(\lg(n/\varepsilon))$).*

2.5.3 Reducing the entropy loss of KL-extractors

In this section, we show how to avoid the entropy loss inherent in Theorem 2.5.4 using the zig-zag product for extractors, introduced by Reingold, Vadhan, and Wigderson [RVW00]. This product combines a technique of Raz and Reingold [RR99] to preserve entropy and the method of Wigderson and Zuckerman [WZ99] to extract entropy left over in a source after an initial extraction, and we show that these techniques extend to the setting of KL-extractors. Furthermore, these techniques (along with the Leftover Hash Lemma) are also the key to the transformation of Raz, Reingold, and Vadhan [RRV02] to convert an arbitrary extractor into one with optimal entropy loss, so we show that this transformation works for KL-extractors as well.

For all of these results, the key is the following lemma:

Lemma 2.5.15 (Re-extraction from leftovers)**.** *Let*

1. $\mathrm{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *be a (k_1, ε_1) KL-extractor,*
2. $\mathrm{W}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^w$ *be a function such that $(\mathrm{Ext}_1, \mathrm{W}_1) : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1} \times \{0,1\}^w$ is an injective map,*

3. $\mathrm{Ext}_2 : \{0,1\}^w \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ *be a* (k_2, ε_2) *average-case KL-extractor for* $k_2 \le k_1 + d_1 - m_1$.

Then $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d_1+d_2} \to \{0,1\}^{m_1+m_2}$ *defined by*

$$\mathrm{Ext}\big(x,(s,t)\big) = \big(\mathrm{Ext}_1(x,s), \mathrm{Ext}_2\big(\mathrm{W}_1(x,s),t\big)\big)$$

is a $(k_1, \varepsilon_1 + \varepsilon_2)$ *KL-extractor. Furthermore, if* Ext_1 *is average-case then so is* Ext.

Remark 2.5.16. The pair $(\mathrm{Ext}_1, \mathrm{W}_1)$ is a special case of what Raz and Reingold [RR99] called an *extractor-condenser pair*. One can think of W_1 as preserving "leftovers" or "waste," which is then "re-extracted" or "recycled" by Ext_2. The identity function on $\{0,1\}^n \times \{0,1\}^{d_1}$ is a valid choice of W_1, but the advantage of the more general formulation is that w can be much smaller than $n + d_1$, and most known explicit constructions of extractors have seed length depending on the input length of the source.

Proof. Given any joint distribution (Z, X) with X distributed over $\{0,1\}^n$ and $\widetilde{\mathrm{H}}_\infty(X|Z) \ge k_1$, we have for every $z \in \mathrm{Supp}(Z)$ that

$$\begin{aligned}
&\mathrm{KL}\big(\mathrm{Ext}(X|_{Z=z}, (U_{d_1}, U_{d_2})) \parallel U_{m_1+m_2}\big) \\
&\quad = \mathrm{KL}\big(\mathrm{Ext}_1(X|_{Z=z}, U_{d_1}), \mathrm{Ext}_2\big(\mathrm{W}_1(X|_{Z=z}, U_{d_1}), U_{d_2}\big) \parallel U_{m_1}, U_{m_2}\big) \\
&\quad = \mathrm{KL}\big(\mathrm{Ext}_1\big(X|_{Z=z}, U_{d_1}\big) \parallel U_{m_1}\big) \\
&\qquad + \mathop{\mathbb{E}}_{o_1 \sim \mathrm{Ext}_1(X|_{Z=z}, s)} \Big[\mathrm{KL}\Big(\mathrm{Ext}_2\Big(\mathrm{W}_1\big(X, U_{d_1}\big)|_{Z=z, \mathrm{Ext}_1(X, U_{d_1})=o_1}, U_{d_2}\Big) \Big\| U_{m_2}\Big)\Big] \qquad (2.1)
\end{aligned}$$

where the last line follows from the chain rule for KL divergence. Note that

$$\begin{aligned}
&\widetilde{\mathrm{H}}_\infty\Big(\mathrm{W}_1(X, U_{d_1}) \,\Big|\, Z, \mathrm{Ext}_1(X, U_{d_1})\Big) \\
&\qquad = \widetilde{\mathrm{H}}_\infty\Big(\mathrm{Ext}_1(X, U_{d_1}), \mathrm{W}_1(X, U_{d_1}) \,\Big|\, Z, \mathrm{Ext}_1(X, U_{d_1})\Big) \\
&\qquad = \widetilde{\mathrm{H}}_\infty\Big(X, U_{d_1} \,\Big|\, Z, \mathrm{Ext}_1(X, U_{d_1})\Big) \qquad (\text{(}\mathrm{Ext}_1, \mathrm{W}_1\text{) is an injection})
\end{aligned}$$

$$
\begin{aligned}
&\geq \widetilde{\mathrm{H}}_\infty(X, U_{d_1} \mid Z) - \lg|\mathrm{Supp}(\mathrm{Ext}_1(X, U_{d_1}))| && (*)\\
&= \widetilde{\mathrm{H}}_\infty(X \mid Z) + \mathrm{H}_\infty(U_{d_1}) - \lg|\mathrm{Supp}(\mathrm{Ext}_1(X, U_{d_1}))| && \text{(by independence)}\\
&\geq k_1 + d_1 - m_1 \geq k_2
\end{aligned}
$$

where the line (*) follows from standard properties of conditional min-entropy (e.g. [DORS08, Lemma 2.2]). That Ext is a $(k_1, \varepsilon_1 + \varepsilon_2)$ KL-extractor now follows immediately from Eq. (2.1) by taking Z independent of X, and the average-case claim follows from taking expectations over $z \sim Z$. □

Remark 2.5.17. The proof above works for any weak divergence D such that

$$
\mathrm{D}(X, Y \,\|\, U_{m_1}, U_{m_2}) \leq \mathrm{D}(X \,\|\, U_{m_1}) + \mathop{\mathbb{E}}_{x \sim X}\left[\mathrm{D}(Y|_{X=x} \,\|\, U_{m_2})\right]
$$

for all joint distributions (X, Y) independent of (U_{m_1}, U_{m_2}). In particular, the same proof also gives Lemma 2.5.15 for standard (total variation) extractors.

By Lemma 2.5.2, we get an analogous result for strong KL-extractors.

Corollary 2.5.18. *Let*

1. $\mathrm{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *be a strong* (k_1, ε_1) *KL-extractor,*
2. $\mathrm{W}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^w$ *be such that the map* $(x, s) \mapsto (s, \mathrm{Ext}_1(x, s), \mathrm{W}_1(x, s))$ *is an injection,*
3. $\mathrm{Ext}_2 : \{0,1\}^w \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ *be a* (k_2, ε_2) *strong average-case KL-extractor for* $k_2 \leq k_1 - m_1$.

Then $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d_1+d_2} \to \{0,1\}^{m_1+m_2}$ *defined by*

$$
\mathrm{Ext}(x, (s, t)) = \left(\mathrm{Ext}_1(x, s), \mathrm{Ext}_2(\mathrm{W}_1(x, s), t)\right)
$$

is a strong $(k_1, \varepsilon_1 + \varepsilon_2)$ *KL-extractor. Furthermore, if* Ext_1 *is average-case then so is* Ext.

The zig-zag product for extractors due to Reingold, Vadhan, and Wigderson [RVW00] (in the special case of injective (Ext, W)-pairs) is an immediate consequence of Lemma 2.5.15 and Theorem 2.5.4 our basic composition result. Recall that Theorem 2.5.4 was able to combine an "outer" extractor, generally taken to have seed length depending only (but linearly) on $n - k$, with an "inner" extractor to produce seeds for the outer extractor with logarithmic seed length. However, as discussed in Remark 2.5.5 that basic composition necessarily lost $\lg(1/\delta)$ bits of entropy, so the zig-zag product uses Lemma 2.5.15 to recover this entropy, using an (Ext, W)-pair to ensure that the re-extraction adds additional seed length depending logarithmically on $n - k$ rather than n.

Corollary 2.5.19 (Zig-zag product for KL-extractors, analogous to [RVW00, Theorem 3.6])**.** *Let*

1. $\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be an* $(n - \lg(1/\delta), \varepsilon_{out})$ *extractor for* $\mathrm{D}_{1+\alpha}$ *with* $\alpha > 0$*,*
2. $\mathrm{W}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^w$ *be a function such that the pair* $(\mathrm{Ext}_{out}, \mathrm{W}_{out})$ *is an injection from* $\{0,1\}^n \times \{0,1\}^d$*,*
3. $\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d$ *be an* $(n' - \lg(1/\delta), \varepsilon_{in})$ *average-case KL-extractor,*
4. $\mathrm{W}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^{w'}$ *be such that the pair* $(\mathrm{Ext}_{in}, \mathrm{W}_{in})$ *is an injection from* $\{0,1\}^{n'} \times \{0,1\}^{d'}$*,*
5. $\mathrm{Ext}_{waste} : \{0,1\}^{w+w'} \times \{0,1\}^{d''} \to \{0,1\}^{m''}$ *be an average-case* $(n + n' - \lg(1/\delta) - m, \varepsilon_{waste})$ *KL-extractor,*

and define

1. $\mathrm{Ext}_{comp} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ *by* $\mathrm{Ext}_{comp}\big((x,y),s\big) = \mathrm{Ext}_{out}\big(x, \mathrm{Ext}_{in}(y,s)\big)$ *as in Theorem 2.5.4,*

2. $\mathrm{W}_{comp} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^{w+w'}$ *by*

$$\mathrm{W}_{comp}\big((x,y),s\big) = \Big(\mathrm{W}_{out}\big(x, \mathrm{Ext}_{in}(y,s)\big), \mathrm{W}_{in}(y,s)\Big),$$

3. $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'+d''} \to \{0,1\}^{m+m''}$ *by*

$$\mathrm{Ext}\big((x,y),(s,t)\big) = \Big(\mathrm{Ext}_{comp}\big((x,y),s\big), \mathrm{Ext}_{waste}\Big(\mathrm{W}_{comp}\big((x,y),s\big),t\Big)\Big)$$

as in Lemma 2.5.15.

Then Ext *is an* $\big(n + n' - \lg(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in} + \varepsilon_{waste}\big)$*-extractor for KL. Furthermore, if* Ext_{in} *and* Ext_{waste} *are strong average-case KL-extractors, then* Ext *is a strong KL-extractor, and if* Ext_{out} *is average-case then so is* Ext.

Proof. We claim that W_{comp} is such that $(\mathrm{Ext}_{comp}, \mathrm{W}_{comp})$ is an injection: by assumption on the pair $(\mathrm{Ext}_{out}, \mathrm{W}_{out})$ we have that given $\mathrm{Ext}_{out}(x, \mathrm{Ext}_{in}(y,s))$ and $\mathrm{W}_{out}(x, \mathrm{Ext}_{in}(y,s))$ we can recover x and $\mathrm{Ext}_{in}(y,s)$, and by assumption on $(\mathrm{Ext}_{in}, \mathrm{W}_{in})$ given $\mathrm{Ext}_{in}(y,s)$ and $\mathrm{W}_{in}(y,s)$ we can recover (y,s), so that $(\mathrm{Ext}_{comp}, \mathrm{W}_{comp})$ has an inverse and is injective as desired. Therefore, since Theorem 2.5.4 implies Ext_{comp} is an $(n + n' - \lg(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in})$ KL-extractor, the result follows from Lemma 2.5.15. The furthermore claims follow from the corresponding claims of these lemmas (and Corollary 2.5.18 for the strong case). □

Remark 2.5.20. Corollary 2.5.19 was presented by Reingold, Vadhan, and Wigderson [RVW00] as a transformation that combined three extractor-condenser pairs into a new extractor-condenser pair. We do not use this generality, so for simplicity we do not present it here, but both Lemma 2.5.15 and Corollary 2.5.19 can be easily extended in this manner if required.

The Raz–Reingold–Vadhan [RRV02] transformation to avoid entropy loss follows similarly using

the Leftover Hash Lemma (Proposition 2.5.6).

Corollary 2.5.21 (KL-extractor analogue of [RRV02, Lemma 28]). *Let* $\mathrm{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *be a strong* $(k, \varepsilon/2)$ *KL-extractor with entropy loss* Δ_1*, meaning* $m_1 = k - \Delta_1$*. Then for every* $d_{extra} \leq \Delta_1$ *there is an explicit* (k, ε) *strong KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d'} \to \{0,1\}^{m'}$ *with seed length* $d' = d_1 + O(d_{extra} + \lg(n/\varepsilon))$ *and entropy loss* $\Delta_1 - d_{extra} + \lg(1/\varepsilon) - O(1)$*, meaning* $m' = k - (\Delta_1 - d_{extra}) - \lg(1/\varepsilon) + O(1)$*, which is computable in polynomial time making one oracle call to* Ext_1*. Furthermore, if* Ext_1 *is average-case then so is* Ext*.*

In particular, by taking $d_{extra} = \Delta_1$ *we get an extractor with optimal entropy loss* $\lg(1/\varepsilon) + O(1)$ *by paying an additional* $O(\Delta + \lg(n/\varepsilon))$ *in seed length.*

Proof. Let $\mathrm{W}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^n$ be given by $\mathrm{W}_1(x,s) = x$, and let $\mathrm{Ext}_2 : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ be the strong average-case $(d_{extra}, \varepsilon/2)$ KL-extractor of Proposition 2.5.6 using almost-universal hash functions, so that $d_2 = O(d_{extra} + \lg(n/\varepsilon))$ and $m_2 = d_{extra} - \lg(1/\varepsilon) - O(1)$. The result follows from taking Ext to be the extractor of Corollary 2.5.18. □

Remark 2.5.22. An analogous versions of the above claim for non-strong KL-extractors follows by taking $\mathrm{W}_1(x,s) = (x,s)$ and using Lemma 2.5.15.

We can apply Corollary 2.5.21 to Theorem 2.5.14 the KL-extractors from the total variation extractors of Guruswami, Umans, and Vadhan [GUV09], thereby avoiding the extra $O(\lg(n/\varepsilon))$ entropy loss in the strong extractors.

Corollary 2.5.23. *For every* $n \in \mathbb{N}$, $1 > \alpha, \varepsilon > 0$*, and* $k, k' \geq 0$ *with* $k + k' \leq n$*, there is an explicit strong average-case* $(k + k', \varepsilon)$ *KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d \leq O_\alpha(\lg(n/\varepsilon)) + O(k')$ *and* $m \geq (1-\alpha)k + k' - \lg(1/\varepsilon) - O(1)$*.*

2.5.4 Lower bounds

In this section, we give lower bounds on extractors for the Rényi divergences D_β of all orders, including the special case $\beta = 1$ of KL-extractors. A reader primarily interested in explicit constructions of subgaussian samplers can skip to Section 2.6.

For Rényi divergences D_β with $\beta \leq 1$ we reduce to Radhakrishnan and Ta-Shma's [RT00] lower bounds for total variation extractors and *dispersers*, which can be understood as a one-sided relaxation of total variation extractors.

Definition 2.5.24 (Sipser [Sip88] and Cohen and Wigderson [CW89])**.** A function $\mathrm{Disp} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k,ε) *disperser* if for all random variables X over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$, it holds that $|\mathrm{Supp}(\mathrm{Disp}(X, U_d))| \geq (1-\varepsilon)2^m$.

Dispersers are of interest in the context of Rényi extractors because the Rényi 0-entropy of a random variable is the logarithm of its support size (see Example 2.2.4), and hence dispersers are equivalent to D_0-extractors:

Lemma 2.5.25. Disp *is a* (k,ε) *disperser if and only if* Disp *is a* $\big(k, \lg(1/(1-\varepsilon))\big)$ D_0*-extractor.*

Given Lemma 2.5.25, we can use the lower bounds of Radhakrishnan and Ta-Shma [RT00] to give an optimal lower bound on the seed length of D_β-extractors for $\beta \leq 1$ in terms of the error ε, input length n and supported entropy k (we will give a matching non-explicit upper bound in the next section), as well as lower bounds on the entropy loss. For the case $\beta = 1$ of KL-extractors, the non-explicit upper bound (Theorem 2.5.30) also shows that the entropy loss lower bound is optimal.

Theorem 2.5.26. *Let* $0 \leq \beta \leq 1$ *and* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* (k,ε) *extractor for* D_β *with* $k \leq n-2$, $d \leq m-1$, *and* $2^{2-m} < \varepsilon < 1/4$. *Then* $d \geq \lg(n-k) + \lg(1/\varepsilon) - O(1)$ *and* $m \leq k + d - \lg\lg(1/\varepsilon) + O(1)$. *Furthermore, if* ε *is at most* $\beta/(2\ln 2)$ *then* $m \leq k + d - \lg(1/\varepsilon) + \lg(1/\beta) + O(1)$.

Proof. Since D_β is nondecreasing in β we have that Ext is a (k, ε) extractor for D_0, and thus by Lemma 2.5.25 it is a $(k, 1 - 2^{-\varepsilon})$ disperser. Then the disperser seed length lower bound of Radhakrishnan and Ta-Shma [RT00] tells us that $d \geq \lg(n-k)+\lg(1/(1-2^{-\varepsilon}))-O(1) \geq \lg(n-k)+\lg(1/\varepsilon)-O(1)$ and $m \leq k + d - \lg\lg(1/(1 - 2^{-\varepsilon})) + O(1) \leq k + d - \lg\lg(1/\varepsilon) + O(1)$.

For the other entropy loss lower bound, we use Gilardoni's [Gil10] generalization of Pinsker's inequality, which shows in particular that $d_{\mathrm{TV}}(P, U_m) \leq \sqrt{\ln 2/(2\beta) \cdot \mathrm{D}_\beta(P \parallel U_m)}$. Thus, Ext is also a $(k, \sqrt{\varepsilon \cdot \ln 2/(2\beta)})$ total variation extractor, and if $\sqrt{\varepsilon \cdot \ln 2/(2\beta)} \leq 1/2$ (equivalently $\varepsilon \leq \beta/(2 \ln 2)$) then the [RT00] total variation extractor entropy loss lower bound implies that $m \leq k + d - 2\lg(1/\sqrt{\varepsilon \cdot \ln 2/(2\beta)}) + O(1) \leq k + d - \lg(1/\varepsilon) + \lg(1/\beta) + O(1)$. □

Remark 2.5.27. For the case of $0 < \beta < 1$, we do not know whether the entropy loss lower bound of Theorem 2.5.26 is tight.

It is well-known that ℓ_2-extractors (which are equivalent to D_2-extractors by Example 2.2.4) require seed length at least linear in $\min(n - k, m)$ (see e.g. [Vad12, Problem 6.4]). We generalize this to give a linear seed length lower bound on D_β extractors for all $\beta > 1$, in the regime of constant ε, improving on the logarithmic lower bound given by Theorem 2.5.26.

Theorem 2.5.28. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* $(k, 0.99)$ $\mathrm{D}_{1+\alpha}$*-extractor for* $\alpha > 0$*. Then* $d \geq \min\{(n - k - 3) \cdot \alpha, (m - 2) \cdot \alpha/(\alpha + 1)\}$.

Proof. We follow the strategy suggested by Vadhan [Vad12, Problem 6.4], and view Ext as a bipartite graph with $N = \{0,1\}^n$ left-vertices, $M = \{0,1\}^m$ right-vertices, and $D = 2^d$ edges per left-vertex given by $E = \{(x \in \{0,1\}^n, y \in \{0,1\}^m) \mid \exists s \in \{0,1\}^d : \mathrm{Ext}(x, s) = y\}$.

Assume for the sake of contradiction that $d \leq \alpha/(\alpha + 1) \cdot (m - 2)$ and $d \leq \alpha(n - k - 3)$, so that $M \geq 4D^{1+1/\alpha}$ and $N/(8D^{1/\alpha}) \geq K$. Now, we claim there exists a set $T \subseteq \{0,1\}^m$ of size

at most $M/(2D^{1+1/\alpha})$ such that $X = \{x \in \{0,1\}^n \mid \exists s \in \{0,1\}^d \text{ s.t } \mathrm{Ext}(x,s) \in T\}$ has size at least $N/(8D^{1/\alpha}) \geq K$. This follows from the following iterative procedure: until $|X| \geq N/(8D^{1/\alpha})$, choose the vertex $y \in \{0,1\}^m$ of highest degree, add it to T, and remove y and its neighbors from the graph (the neighbors go in X). Then at each step we will add to X a number of vertices at least the average degree

$$\frac{(N - |X|) \cdot D}{M - |T|} \geq \frac{(N - N/(8D^{1/\alpha})) \cdot D}{M} \geq \frac{ND}{2M},$$

so that the size of T will be at most $\lceil N/(8D^{1/\alpha}) \cdot 2M/ND \rceil = \lceil M/(4D^{1+1/\alpha}) \rceil \leq M/(2D^{1+1/\alpha})$ as desired. Now, since X has size at least K and Ext is a $(k, 0.99)$ $\mathrm{D}_{1+\alpha}$-extractor, we have that

$$\begin{aligned}
0.99 &\geq \mathrm{D}_{1+\alpha}(\mathrm{Ext}(U_X, U_d) \parallel U_m) \\
&= \frac{1}{\alpha} \lg\left(\sum_{y \in \{0,1\}^m} \frac{\Pr[\mathrm{Ext}(U_X, U_D) = y]^{1+\alpha}}{2^{-m\alpha}} \right) \\
&\geq \frac{1}{\alpha} \lg\left(M^\alpha \sum_{y \in T} \Pr[\mathrm{Ext}(U_X, U_D) = y]^{1+\alpha} \right) \\
&\geq \frac{1}{\alpha} \lg\left(M^\alpha \cdot |T|^{-\alpha} \cdot \left(\sum_{y \in T} \Pr[\mathrm{Ext}(U_X, U_d) = y] \right)^{1+\alpha} \right) && \text{(By Hölder's inequality)} \\
&\geq \frac{1}{\alpha} \lg\left(M^\alpha \cdot (M/(2D^{1+1/\alpha}))^{-\alpha} \cdot (1/D)^{1+\alpha} \right) = 1 && \text{(By definition of } T\text{)}
\end{aligned}$$

which is a contradiction, as desired. □

We can also use this lower bound to get a similar lower bound for $d_{\ell_{1+\alpha}}$-extractors for all $\alpha > 0$, though in this case the lower bound applies up to an error threshold that depends on α.

Corollary 2.5.29. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* $\left(k, \varepsilon_\alpha \cdot 2^{-m\alpha/(1+\alpha)}\right)$ *extractor for* $d_{\ell_{1+\alpha}}$ *where* $\alpha > 0$ *and* $\varepsilon_\alpha = (2/3) \cdot \alpha/(\alpha+1)$. *Then* $d \geq \min\{(n-k-3) \cdot \alpha, (m-2) \cdot \alpha/(\alpha+1)\}$.

Proof. Note that the proof of Theorem 2.5.28 gave a lower bound on the sum $\sum_{y \in \{0,1\}^m} P_y^{1+\alpha}$ where

$P = \mathrm{Ext}(U_X, U_d)$, whereas $d_{\ell_{1+\alpha}}(P, U_m)^{1+\alpha} = \sum_{y\in\{0,1\}^m} \left|P_y - 2^{-m}\right|^{1+\alpha}$. For ℓ_2 these can be related without any loss, but in general we can use the triangle inequality to get

$$\mathrm{D}_{1+\alpha}(P \parallel U_m) \leq \frac{1}{\alpha} \cdot \lg\left(2^{m\alpha} \cdot \left(d_{\ell_{1+\alpha}}(P, U_m) + 2^{-m\alpha/(\alpha+1)}\right)^{1+\alpha}\right)$$

so that if $d_{\ell_{1+\alpha}}(P, U_m) \leq \varepsilon_\alpha \cdot 2^{-m\alpha/(1+\alpha)}$ where $\varepsilon_\alpha = (2/3) \cdot \alpha/(\alpha+1) \leq 2^{0.99\cdot\alpha/(\alpha+1)} - 1$, then $\mathrm{D}_{1+\alpha}(P \parallel U_m) \leq 0.99$, and we conclude by Lemma 2.5.1 and Theorem 2.5.28. □

2.5.5 Non-explicit construction

In this section, we show non-constructively the existence of KL-extractors matching the lower-bound of Theorem 2.5.26 and in particular implying the optimal parameters of standard extractors for total variation distance. Formally, we will prove:

Theorem 2.5.30. *For every $n \in \mathbb{N}$, $k \leq n$, and $1 > \varepsilon > 0$ there is an average-case (respectively strong average-case) (k, ε) KL-extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with seed length $d = \lg(n-k+1) + \lg(1/\varepsilon) + O(1)$ and output length $m = k + d - \lg(1/\varepsilon) + O(1)$ (respectively $m = k - \lg(1/\varepsilon) - O(1)$).*

Remark 2.5.31. For $\varepsilon \gg 1$ the above parameters are not necessarily optimal, and it would be interested to get matching upper and lower bounds in this regime of parameters.

We will prove Theorem 2.5.30 using the probabilistic method, analogously to Zuckerman [Zuc97] or Radhakrishnan and Ta-Shma [RT00] for total variation extractors. However, rather than using Hoeffding's inequality, we use the following lemma:

Lemma 2.5.32. *Let X be uniform over a subset of $\{0,1\}^n$ of size K. Then if $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a random function, it holds for every $\varepsilon > 0$ that*

$$\Pr_{\mathrm{Ext}}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)] > \varepsilon\right] \leq 2^{MD - KD\varepsilon/2}$$

where $D = 2^d$ *and* $M = 2^m$.

Remark 2.5.33. For total variation extractors, the analogous bound is

$$\Pr_{\text{Ext}}\left[d_{\text{TV}}\big((U_d, \text{Ext}(X, U_d)), (U_d, U_m)\big) > \varepsilon\right] \leq 2^{MD - 2KD\varepsilon^2/\ln 2}.$$

One sees that the bounds are very similar, except the KL divergence version depends on ε rather than ε^2. For the regime where $\varepsilon < 1$ the linear dependence is preferable, and is responsible for the $1 \cdot \lg(1/\varepsilon)$ seed length for KL-extractors compared to the $2 \cdot \lg(1/\varepsilon)$ seed length for total variation extractors.

Proof of Lemma 2.5.32. Note that for each $s \in \{0,1\}^d$ and fixed Ext, the random variable $\text{Ext}(X, s)$ is uniform over the multiset $\{\text{Ext}(x, s) \mid x \in \text{Supp}(X)\}$. Hence, since Ext is a random function, this multiset is distributed exactly as taking K iid uniform samples from $\{0,1\}^m$, so we wish to bound the KL divergence between this empirical distribution and the true distribution. For this, in Chapter 4 we give the moment generating function bound

$$\mathop{\mathbb{E}}_{\text{Ext}}\left[2^{t \cdot \text{KL}(\text{Ext}(X,s) \parallel U_m)}\right] \leq \left(\frac{1}{1 - t/K}\right)^{M-1}$$

for every $0 \leq t < K$, which for $t = K/2$ is at most 2^M. Then since $\text{Ext}(X, s)$ is independent across $s \in \{0,1\}^d$, we have

$$\begin{aligned}\Pr_{\text{Ext}}\left[\mathop{\mathbb{E}}_{s \sim U_d}[\text{KL}(\text{Ext}(X, s) \parallel U_m)] > \varepsilon\right] &= \Pr_{\text{Ext}}\left[2^{K/2 \cdot \sum_{s \in \{0,1\}^d} \text{KL}(\text{Ext}(X,s) \parallel U_m)} > 2^{K/2 \cdot D\varepsilon}\right] \\ &\leq 2^{-KD\varepsilon/2} \cdot \prod_{i=1}^{D} 2^M \qquad \square\end{aligned}$$

We can now prove Theorem 2.5.30:

Proof of Theorem 2.5.30. We will show that a random function $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a strong average-case (k, ε) KL-extractor with positive probability, the non-strong version then follows

from Lemma 2.5.2. By Lemma 2.3.14, it is enough to prove that Ext is a strong $(k-t, 2^{t+1}/3 \cdot \varepsilon)$ KL-extractor for every $t \geq 0$. To reduce the range of t we need to consider, note that it suffices to be a $(\lg\lfloor 2^{k-t} \rfloor, 2^{t+1}/3 \cdot \varepsilon)$ extractor for every $t \geq 0$, so that by rounding down it is enough to be a $(k-t, 2^t/3 \cdot \varepsilon)$ strong KL-extractor for each $t \geq 0$ such that 2^{k-t} is an integer.

Now, consider a fixed $t \geq 0$ such that 2^{k-t} is an integer. Since the KL divergence is convex in its first argument and all distributions of min-entropy at least $k-t$ are convex combinations of "flat" distributions which are uniform over a set of size 2^{k-t} (Chor and Goldreich [CG88]), it suffices to analyze the behavior of Ext on such distributions. Then for every subset $X \subseteq \{0,1\}^n$ of size 2^{k-t}, Lemma 2.5.32 tells us that

$$\Pr_{\text{Ext}}\left[\mathop{\mathbb{E}}_{s \sim U_d}[\mathrm{KL}(\mathrm{Ext}(U_X, s) \parallel U_m)] > 2^t/3 \cdot \varepsilon\right] \leq 2^{MD - 2^{k-t} \cdot D \cdot (2^t/3 \cdot \varepsilon)/2} = 2^{MD - KD\varepsilon/6}$$

where $M = 2^m$, $D = 2^d$, and $K = 2^k$. There are $\sum_{j=0}^{K} \binom{N}{j}$ such subsets X of $\{0,1\}^n$ for which we simultaneously need to establish that $\mathbb{E}_{s \sim U_d}[\mathrm{KL}(\mathrm{Ext}(U_X, s) \parallel U_m)] \leq 2^t/3 \cdot \varepsilon$, so we have by a union bound that the probability that Ext is not a strong average-case (k, ε) KL-extractor is at most

$$2^{MD - KD\varepsilon/6} \cdot \sum_{j=0}^{K} \binom{N}{j} \leq 2^{MD - KD\varepsilon/6} \cdot \left(\frac{Ne}{K}\right)^K = 2^{MD + K\lg(Ne/K) - KD\varepsilon/6}.$$

Hence, as long as

$$MD < \frac{KD\varepsilon}{12} \qquad K\lg\left(\frac{Ne}{K}\right) < \frac{KD\varepsilon}{12}$$

$$m \leq k - \lg(1/\varepsilon) - O(1) \qquad d \geq \lg(n-k+1) + \lg(1/\varepsilon) + O(1)$$

we know that a random function is a strong average-case (k, ε) KL-extractor with positive probability as desired. □

2.6 Constructions of subgaussian samplers

2.6.1 Subconstant ε and δ

The goal of this section is to establish the following theorem, which is our explicit construction of subgaussian samplers with sample complexity having no dependence on m, and with randomness complexity and sample complexity matching the best-known $[0,1]$-valued sampler when ε and δ are subconstant (up to the hidden polynomial in the sample complexity).

Theorem 2.6.1. *For all $m \in \mathbb{N}$, $1 > \varepsilon, \delta > 0$, and $\alpha > 0$ there exists an explicit (δ, ε) absolute averaging sampler (respectively strong absolute averaging sampler) for subgaussian and subexponential functions* $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ *with sample complexity $D = \mathrm{poly}(\lg(1/\delta), 1/\varepsilon)$ and randomness complexity $n = m + (1+\alpha)\cdot\lg(1/\delta)$ (respectively $n = m + (1+\alpha)\cdot\lg(1/\delta) + 2\lg(1/\varepsilon) + O(1)$).*

We will use essentially the same construction used for bounded samplers in this regime, namely applying the Reingold, Vadhan, and Wigderson [RVW00] zig-zag product for extractors to combine the expander extractor of Goldreich and Wigderson [GW97] and an extractor with logarithmic seed length. However, as described in detail in Section 2.4.1, even the basic composition used in this construction does not work for general subgaussian extractors, so we will instead use the zig-zag product for KL-extractors (Corollary 2.5.19) combining extractors for Rényi divergences, specifically the D_2-extractor from Corollary 2.5.11 and the KL-extractor from Corollary 2.5.23, to get the following high-entropy KL-extractor:

Theorem 2.6.2. *For all integers m and $1 > \alpha, \delta, \varepsilon > 0$ there is an explicit average-case (respectively strong average-case) (k, ε) KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $n = m + (1+\alpha)\lg(1/\delta) - O(1)$ (respectively $n = m + (1+\alpha)\cdot\lg(1/\delta) + \lg(1/\varepsilon) + O(1)$), $k = n - \lg(1/\delta)$, and $d = O_\alpha(\lg(\lg(1/\delta)/\varepsilon))$.*

Proof. We prove the claim for strong extractors, for the non-strong claim one can simply define $\mathrm{Ext}(x,(s,t)) = \mathrm{Ext}_{strong}((x,t),s)$ where t has length $\lg(1/\varepsilon) + O(1)$.

By Corollary 2.5.11, there is a universal constant $C > 0$ such that for $d_{out} = \lceil C\lg(1/(\delta\varepsilon))\rceil \leq C\lg(1/\delta) + C\lg(1/\varepsilon) + 1$ there is an explicit average-case $(n_{out} - \lg(1/\delta), \varepsilon/4)$ D_2-extractor $\mathrm{Ext}_{out} : \{0,1\}^{n_{out}} \times \{0,1\}^{d_{out}} \to \{0,1\}^{n_{out}}$ with $n_{out} = m - d_{out}$. Furthermore, Ext_{out} has the property that the function $\mathrm{W}_{out}(x,s) = s$ is such that $(\mathrm{Ext}_{out}, \mathrm{W}_{out})$ is an injection.

Let $k'_{in} = C\lg(1/\delta)/(1-\beta)$, $k''_{in} = (C+1)\lg(1/\varepsilon) + O(1)$, and $k_{in} = k'_{in} + k''_{in}$ for $0 < \beta < 1$ some parameter to be chosen later. Then by Corollary 2.5.23, there is an explicit $(k_{in}, \varepsilon/4)$ strong average-case KL-extractor $\mathrm{Ext}_{in} : \{0,1\}^{n_{in}} \times \{0,1\}^{d_{in}} \to \{0,1\}^{m_{in}}$ with $n_{in} = k_{in} + \lg(1/\delta)$, $d_{in} = O_\beta(\lg(n_{in}/\varepsilon)) + O(k''_{in}) = O_\beta(\lg(\lg(1/\delta)/\varepsilon))$, and $m_{in} = (1-\beta)k'_{in} + k''_{in} - \lg(1/\varepsilon) - O(1) = d_{out}$. Furthermore, the function $\mathrm{W}_{in}(x,s) = (x,s)$ is an injection.

Furthermore, for $k_{waste} = (n_{out} + n_{in} - \lg(1/\delta)) - n_{out} = n_{in} - \lg(1/\delta) = k_{in} = k'_{in} + k''_{in}$, by Corollary 2.5.23 there is also an explicit $(k_{waste}, \varepsilon/4)$ strong average-case KL-extractor $\mathrm{Ext}_{waste} : \{0,1\}^{d_{out}+n_{in}+d_{in}} \times \{0,1\}^{d_{waste}} \to \{0,1\}^{m_{waste}}$ such that $m_{waste} = d_{out}$ and

$$d_{waste} = O_\beta\big(\lg((d_{out} + n_{in} + d_{in})/\varepsilon)\big) + O(k''_{in}) = O_\beta\big(\lg(\lg(1/\delta)/\varepsilon)\big).$$

Then by the zig-zag product for KL-extractors (Corollary 2.5.19), there is an explicit $(n_{out} + n_{in} - \lg(1/\delta), \varepsilon)$ strong average-case KL-extractor $\mathrm{Ext} : \{0,1\}^{n_{out}+n_{in}} \times \{0,1\}^{d_{in}+d_{waste}} \to \{0,1\}^{n_{out}+m_{waste}}$, where we have

$$\begin{aligned} n_{out} + n_{in} &= (m - d_{out}) + \Big(\big(C\lg(1/\delta)/(1-\beta) + (C+1)\lg(1/\varepsilon) + O(1)\big) + \lg(1/\delta)\Big) \\ &\leq m + \lg(1/\delta) + \lg(1/\varepsilon) + \lg(1/\delta)\cdot C\cdot\big(1/(1-\beta) - 1\big) + O(1) \\ d_{in} + d_{waste} &= O_\beta\big(\lg(\lg(1/\delta)/\varepsilon)\big) \end{aligned}$$

$$n_{out} + m_{waste} = (m - d_{out}) + d_{out} = m.$$

Choosing $\beta = \alpha/(\alpha + C)$ so that $C \cdot \big(1/(1-\beta) - 1\big) \leq \alpha$ gives the claim. □

We can now prove Theorem 2.6.1.

Proof of Theorem 2.6.1. Let $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be the explicit $\big(n - \lg\big(1/(\delta/2)\big), \varepsilon^2\big)$ KL-extractor (respectively strong KL-extractor) of Theorem 2.6.2, so that $d = O_\alpha\big(\lg\lg(1/\delta)/\varepsilon\big)$ and $n = m + (1+\alpha)\lg(1/\delta)$ (respectively $n = m + (1+\alpha)\lg(1/\delta) + 2\lg(1/\varepsilon) + O(1)$).

Then by Lemmas 2.4.9 and 2.5.1, Ext is also an $\big(n - \lg\big(1/(\delta/2)\big), \varepsilon\big)\, d_{\mathcal{E}}$-extractor (respectively strong $d_{\mathcal{E}}$-extractor), so by Theorem 2.3.8 the function $\mathrm{Samp} : \{0,1\}^n \times \big(\{0,1\}^m\big)^D$ given by $\mathrm{Samp}(x)_i = \mathrm{Ext}(x,i)$ is an explicit $(\delta/2, \varepsilon)$ sampler for $\mathcal{E}$ (respectively strong sampler for $\mathcal{E}$), and thus by symmetry of $\mathcal{E}$ an explicit (δ, ε) absolute subexponential sampler (respectively absolute strong subexponential sampler) as desired. □

2.6.2 Constant δ

We recall from the introduction that the pairwise independent sampler of Chor and Goldreich [CG89] works for subgaussian functions, and in fact the more general class of functions with bounded variance. The sampler has exponentially worse dependence on δ than is necessary for subgaussian samplers, but is very simple and has randomness complexity optimal up to constant factors.

Theorem 2.6.3 ([CG89])**.** *For all $m \in \mathbb{N}$ and $1 > \varepsilon, \delta > 0$ with $1/(\delta\varepsilon^2) < 2^m$, there is an explicit strong sampler $\mathrm{Samp} : \{0,1\}^n \to \big(\{0,1\}^m\big)^D$ for functions with bounded variance $\mathcal{M}_2$, with randomness complexity $n = O(m)$ and sample complexity $D = O\Big(\frac{1}{\varepsilon^2\delta}\Big)$ defined as $\mathrm{Samp}(h)_d = h(d)$ where h is drawn at random from a size 2^n pairwise-independent hash family $\mathcal{H}$ of functions from $[D] \to \{0,1\}^m$.*

Proof. The fact that pairwise independence gives rise to a strong bounded-variance sampler is immediate by Chebyshev's inequality. The existence of a pairwise indepenent hash family with the claimed parameters is due to Chor and Goldreich [CG89], with similar constructions in the probability literature due to Joffe [Jof71]. □

We also show that the Expander Neighbor sampler of [KPS85; GW97] is a bounded-variance sampler.

Theorem 2.6.4. *There is a universal constant $C \geq 1$ such that for all $m \in \mathbb{N}$ and $1 > \varepsilon, \delta > 0$ there is an explicit sampler* $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ *for functions with bounded variance $\mathcal{M}_2$, with randomness complexity $n = m$ and sample complexity $D = O\left(\left(\frac{1}{\varepsilon^2\delta}\right)^C\right)$. Moreover, if the algorithm is given access to a consistently labelled neighbor function of a Ramanujan graph over $\{0,1\}^n$ of degree $O(1/(\delta\varepsilon^2))$, then one can take $C = 1$.*

Proof. By Corollary 2.5.11, there is an explicit $(n - \lg(1/\delta), \varepsilon \cdot 2^{-m/2})$ ℓ_2-extractor $\mathrm{Ext} : \{0,1\}^m \times \{0,1\}^d \to \{0,1\}^m$ with $d = \left\lceil C(\lg(1/\delta) + 2\lg(1/\varepsilon))\right\rceil + O(1)$, where one can take $C = 1$ given the assumed Ramanujan graph. Then by Example 2.2.9 Ext is also an $(n - \lg(1/\delta), \varepsilon)$ $\mathcal{M}_2$-extractor, so we conclude by Theorem 2.3.8. □

Remark 2.6.5. Note that given explicit constructions of Ramanujan graphs, Theorem 2.6.4 has the same sample complexity but better randomness complexity than the sampler of Theorem 2.6.3.

2.6.3 Non-explicit construction

Applying Lemmas 2.4.9 and 2.5.1 to Theorem 2.5.30 our non-explicit construction of KL-extractors gives:

Corollary 2.6.6. *For every $n \in \mathbb{N}$, $k \leq n$, and $1 > \varepsilon > 0$ there is an average-case (respectively strong average-case) (k, ε) $d_{\mathcal{E}}$-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d = \lg(n-k+1) + 2\lg(1/\varepsilon) + O(1)$ *and* $m \geq k + d - 2\lg(1/\varepsilon) - O(1)$ *(respectively* $m \geq k - 2\lg(1/\varepsilon) - O(1)$*)*

Since $d_{\mathcal{E}}$-extractors are also total variation extractors, Corollary 2.6.6 is optimal up to additive constants by the lower bound of Radhakrishnan and Ta-Shma [RT00].

Using the fact that extractors are samplers (Theorem 2.3.8), we get

Corollary 2.6.7. *For every integer m and $1 > \delta, \varepsilon > 0$ there is a (δ, ε) sampler (respectively strong sampler)* $\mathrm{Samp} : \{0,1\}^n \to \left(\{0,1\}^m\right)^D$ *for subgaussian and subexponential functions with sample complexity* $D = O\left(\frac{\lg 1/\delta}{\varepsilon^2}\right)$ *and randomness complexity* $n = m + \lg(1/\delta) - \lg\lg(1/\delta) + O(1)$ *(respectively* $n = m + \lg(1/\delta) + 2\lg(1/\varepsilon) + O(1)$*).*

Note that this matches the best-known (non-explicit) parameters of averaging samplers for $[0,1]$-valued functions due to Zuckerman [Zuc97].

Chapter 3

Unifying Computational Entropies via Kullback–Leibler Divergence

This chapter is based on joint work with Yi-Hsiu Chen, Thibaut Horel, and Salil Vadhan [ACHV19].

3.1 Introduction

3.1.1 One-way functions and computational entropy

One-way functions [DH76] are on one hand the minimal assumption for complexity-based cryptography [IL89], but on the other hand can be used to construct a remarkable array of cryptographic primitives, including such powerful objects as CCA-secure symmetric encryption, zero-knowledge proofs and statistical zero-knowledge arguments for all of **NP**, and secure multiparty computation with an honest majority [GGM86; GMW91; GMW87; HILL99; Rom90; Nao91; HNORV09]. All of these constructions begin by converting the "raw hardness" of a one-way function (OWF) to one of the following more structured cryptographic primitives: a pseudorandom generator (PRG) [BM82;

Yao82], a universal one-way hash function (UOWHF) [NY89], or a statistically hiding commitment scheme (SHC) [BCC88].

The original constructions of these three primitives from arbitrary one-way functions [HILL99; Rom90; HNORV09] were all very complicated and inefficient. Over the past decade, there has been a series of simplifications and efficiency improvements to these constructions [HRVW09; HRV13; HHRVW10; VZ12], leading to a situation where the constructions of two of these primitives — PRGs and SHCs — share a very similar structure and seem "dual" to each other. Specifically, these constructions proceed as follows:

1. Show that every OWF $f : \{0,1\}^n \to \{0,1\}^n$ has a gap between its "real entropy" and an appropriate form of "computational entropy". Specifically, for constructing PRGs, it is shown that the function $G(x) = (f(x), x_1, x_2, \ldots, x_n)$ has "next-block pseudoentropy" at least $n + \omega(\lg n)$ while its real entropy is $H(G(U_n)) = n$ [VZ12] where $H(\cdot)$ denotes Shannon entropy. For constructing SHCs, it is shown that the function $G(x) = (f(x)_1, \ldots, f(x)_n, x)$ has "next-block accessible entropy" at most $n - \omega(\lg n)$ while its real entropy is again $H(G(U_n)) = n$ [HRVW09]. Note that the differences between the two cases are whether we break x or $f(x)$ into individual bits (which matters because the "next-block" notions of computational entropy depend on the block structure) and whether the form of computational entropy is larger or smaller than the real entropy.

2. An "entropy equalization" step that converts G into a similar generator where the real entropy in each block conditioned on the prefix before it is known. This step is exactly the same in both constructions.

3. A "flattening" step that converts the (real and computational) Shannon entropy guarantees of

the generator into ones on (smoothed) min-entropy and max-entropy. This step is again exactly the same in both constructions.

4. A "hashing" step where high (real or computational) min-entropy is converted to uniform (pseudo)randomness and low (real or computational) max-entropy is converted to a small-support or disjointness property. For PRGs, this step only requires randomness extractors [NZ96; HILL99], while for SHCs it requires (information-theoretic) interactive hashing [NOVY98; DHRS04]. (Constructing full-fledged SHCs in this step also utilizes UOWHFs, which can be constructed from one-way functions [Rom90]. Without UOWHFs, we obtain a weaker binding property, which nevertheless suffices for constructing statistical zero-knowledge arguments for all of **NP**.)

This common construction template came about through a back-and-forth exchange of ideas between the two lines of work. Indeed, the uses of computational entropy notions, flattening, and hashing originate with PRGs [HILL99], whereas the ideas of using next-block notions, obtaining them from breaking $(f(x), x)$ into short blocks, and entropy equalization originate with SHCs [HRVW09]. All this leads to a feeling that the two constructions, and their underlying computational entropy notions, are "dual" to each other and should be connected at a formal level.

In this paper, we make progress on this project of unifying the notions of computational entropy, by introducing a new computational entropy notion that yields both next-block pseudoentropy and next-block accessible entropy in a clean and modular fashion. It is inspired by the proof of [VZ12] that $(f(x), x_1, \ldots, x_n)$ has next-block pseudoentropy $n + \omega(\lg n)$, which we will describe now.

3.1.2 Next-block pseudoentropy via relative pseudoentropy

We recall the definition of next-block pseudoentropy, and the result of [VZ12] relating it to one-wayness.

Definition 3.1.1 (next-block pseudoentropy [HRV10], informal)**.** Let n be a security parameter, and $X = (X_1, \ldots, X_m)$ be a random variable distributed on strings of length poly(n). We say that X has *next-block pseudoentropy* at least k if there is a random variable $Z = (Z_1, \ldots, Z_m)$, jointly distributed with X, such that:

1. For all $i = 1, \ldots, m$, $(X_1, \ldots, X_{i-1}, X_i)$ is computationally indistinguishable from $(X_1, \ldots, X_{i-1}, Z_i)$.
2. $\sum_{i=1}^{m} \mathrm{H}(Z_i | X_1, \ldots, X_{i-1}) \geq k$.

Equivalently, for I uniformly distributed in $[m]$, X_I has *conditional pseudoentropy* at least k/m given $(X_1, \ldots, X_{i-1})$.

It was conjectured in [HRV10] that next-block pseudoentropy could be obtained from any OWF by breaking its input into bits, and this conjecture was proven in [VZ12]:

Theorem 3.1.2 ([VZ12], informal)**.** *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function, let X be uniformly distributed in $\{0,1\}^n$, and let $X = (X_1, \ldots, X_m)$ be a partition of X into blocks of length $O(\lg n)$. Then $(f(X), X_1, \ldots, X_m)$ has next-block pseudoentropy at least $n + \omega(\lg n)$.*

The intuition behind Theorem 3.1.2 is that since X is hard to sample given $f(X)$, then it should have some extra computational entropy given $f(X)$. This intuition is formalized using the following notion of "relative pseudoentropy," which is a renaming of [VZ12]'s notion of "KL-hard for sampling," to better unify the terminology with the notions introduced in this work.

Definition 3.1.3 (relative pseudoentropy [VZ12])**.** Let n be a security parameter, and (X, Y) be a pair of random variables, jointly distributed over strings of length poly(n). We say that X has *relative*

pseudoentropy at least Δ given Y if for all probabilistic polynomial-time S, we have

$$\mathrm{KL}(X, Y \parallel \mathsf{S}(Y), Y) \geq \Delta,$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the relative entropy (a.k.a. Kullback–Leibler divergence).[1]

That is, it is hard for any efficient adversary S to sample the conditional distribution of X given Y, even approximately.

The first step of the proof of Theorem 3.1.2 is to show that one-wayness implies relative pseudoentropy (which can be done with a one-line calculation):

Lemma 3.1.4. *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function and let X be uniformly distributed in $\{0,1\}^n$. Then X has relative pseudoentropy at least $\omega(\lg n)$ given $f(X)$.*

Next, we break X into short blocks, and show that the relative pseudoentropy is preserved:

Lemma 3.1.5. *Let n be a security parameter, let (X, Y) be random variables distributed on strings of length $\mathrm{poly}(n)$, let $X = (X_1, \ldots, X_m)$ be a partition of X into blocks, and let I be uniformly distributed in $[m]$. If X has relative pseudoentropy at least Δ given Y, then X_I has relative pseudoentropy at least Δ/m given $(Y, X_1, \ldots, X_{I-1})$.*

Finally, the main part of the proof is to show that, once we have short blocks, relative pseudoentropy is *equivalent* to a gap between conditional pseudoentropy and real conditional entropy.

Lemma 3.1.6. *Let n be a security parameter, Y be a random variable distributed on strings of length $\mathrm{poly}(n)$, and X a random variable distributed on strings of length $O(\lg n)$. Then X has relative pseudoentropy at least Δ given Y iff X has conditional pseudoentropy at least $\mathrm{H}(X|Y) + \Delta$ given Y.*

[1] Recall that for random variables A and B with $\mathrm{Supp}(A) \subseteq \mathrm{Supp}(B)$, the relative entropy is defined by $\mathrm{KL}(A \parallel B) = \mathbb{E}_{a \leftarrow A}[\lg(\Pr[A = a]/\Pr[B = a])]$.

Putting these three lemmas together, we see that when f is a one-way function, and we break X into blocks of length $O(\lg n)$ to obtain $(f(X), X_1, \ldots, X_m)$, on average, the conditional pseudoentropy of X_I given $(f(X), X_1, \ldots, X_{I-1})$ is larger than its real conditional entropy by $\omega(\lg n)/m$. This tells us that the next-block pseudoentropy of $(f(X), X_1, \ldots, X_m)$ is larger than its real entropy by $\omega(\lg n)$, as claimed in Theorem 3.1.2.

We remark that Lemma 3.1.6 explains why we need to break the input of the one-way function into short blocks: it is false when X is long. Indeed, if f is a one-way function, then we have already seen that X has $\omega(\lg n)$ relative pseudoentropy given $f(X)$ (Lemma 3.1.4), but it does not have conditional pseudoentropy noticeably larger than $\mathrm{H}(X|f(X))$ given $f(X)$ (as correct preimages can be efficiently distinguished from incorrect ones using f).

3.1.3 Inaccessible entropy

As mentioned above, for constructing SHCs from one-way functions, the notion of next-block pseudoentropy is replaced with next-block accessible entropy:

Definition 3.1.7 (next-block accessible entropy [HRVW09], informal)**.** Let n be a security parameter, and $Y = (Y_1, \ldots, Y_m)$ be a random variable distributed on strings of length $\mathrm{poly}(n)$. We say that Y has *next-block accessible entropy* at most k if the following holds.

Let $\widetilde{G}$ be any probabilistic $\mathrm{poly}(n)$-time algorithm that takes a sequence of uniformly random strings $\widetilde{R} = (\widetilde{R}_1, \ldots, \widetilde{R}_m)$ and outputs a sequence $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_m)$ in an "online fashion" by which we mean that $\widetilde{Y}_i = \widetilde{G}(\widetilde{R}_1, \ldots, \widetilde{R}_i)$ depends on only the first i random strings of $\widetilde{G}$ for $i = 1, \ldots, m$. Suppose further that $\mathrm{Supp}(\widetilde{Y}) \subseteq \mathrm{Supp}(Y)$.

Then we require:

$$\sum_{i=1}^{m} \mathrm{H}(\widetilde{Y}_i | \widetilde{R}_1, \ldots, \widetilde{R}_{i-1}) \leq k.$$

(Next-block) accessible entropy differs from (next-block) pseudoentropy in two ways:

1. Accessible entropy is useful as an *upper* bound on computational entropy, and is interesting when it is *smaller* than the real entropy $\mathrm{H}(Y)$. We refer to the gap $\mathrm{H}(Y) - k$ as the *next-block inaccessible entropy* of Y.

2. The accessible entropy adversary $\widetilde{G}$ is trying to *generate* the random variables Y_i conditioned on the history rather than recognize them. Note that we take the "history" to not only be the previous blocks $(\widetilde{Y}_1, \ldots, \widetilde{Y}_{i-1})$, but the coin tosses $(\widetilde{R}_1, \ldots, \widetilde{R}_{i-1})$ used to generate those blocks.

Note that one unsatisfactory aspect of the definition is that when the random variable Y is not *flat* (i.e. uniform on its support), then there can be an adversary $\widetilde{G}$ achieving accessible entropy even *larger* than $\mathrm{H}(Y)$, for example by making $\widetilde{Y}$ uniform on $\mathrm{Supp}(Y)$.

Similarly to (and predating) Theorem 3.1.2, it is known that one-wayness implies next-block inaccessible entropy.

Theorem 3.1.8 ([HRVW09]). *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function, let X be uniformly distributed in $\{0,1\}^n$, and let $(Y_1, \ldots, Y_m)$ be a partition of $Y = f(X)$ into blocks of length $O(\lg n)$. Then $(Y_1, \ldots, Y_m, X)$ has next-block accessible entropy at most $n - \omega(\lg n)$.*

Unfortunately, however, the existing proof of Theorem 3.1.8 is not modular like that of Theorem 3.1.2. In particular, it does not isolate the step of relating one-wayness to entropy-theoretic measures (like Lemma 3.1.4 does) or the significance of having short blocks (like Lemma 3.1.6 does).

3.1.4 Our results

We remedy the above state of affairs by providing a new, more general notion of hardness in relative entropy that allows us to obtain next-block inaccessible entropy in a modular way while also encompassing

what is needed for next-block pseudoentropy.

Like in relative pseudoentropy, we will consider a pair of jointly distributed random variables (Y, X). Following the spirit of accessible entropy, the adversary $\widetilde{G}$ for our new notion will try to *generate* Y together with X, rather than taking Y as input. That is, $\widetilde{G}$ will take randomness $\widetilde{R}$ and output a pair $(\widetilde{Y}, \widetilde{X}) = \widetilde{G}(\widetilde{R}) = (\widetilde{G}_1(\widetilde{R}), \widetilde{G}_2(\widetilde{R}))$, which we require to be always within the support of (Y, X). Note that $\widetilde{G}$ need not be an online generator; it can generate both $\widetilde{Y}$ and $\widetilde{X}$ using the same randomness $\widetilde{R}$. Of course, if (Y, X) is efficiently samplable (as it would be in most cryptographic applications), $\widetilde{G}$ could generate $(\widetilde{Y}, \widetilde{X})$ identically distributed to (Y, X) by just using the "honest" sampler G for (Y, X). So, in addition, we require that the adversary $\widetilde{G}$ also come with a *simulator* S, that can simulate its coin tosses given only $\widetilde{Y}$. The goal of the adversary is to minimize the relative entropy

$$\mathrm{KL}\big(\widetilde{R}, \widetilde{Y} \parallel \mathsf{S}(Y), Y\big)$$

for a uniformly random $\widetilde{R}$. This divergence measures both how well $\widetilde{G}_1$ approximates the distribution of Y as well as how well S simulates the corresponding coin tosses of $\widetilde{G}_1$. Note that when $\widetilde{G}$ is the honest sampler G, the task of S is exactly to sample from the conditional distribution of $\widetilde{R}$ given $\mathsf{G}_1(\widetilde{R}) = Y$. However, the adversary may reduce the divergence by instead designing the sampler $\widetilde{G}$ and simulator S to work in concert, potentially trading off how well $\widetilde{G}(\widetilde{R})$ approximates Y in exchange for easier simulation by S. Explicitly, the definition is as follows.

Definition 3.1.9 (hardness in relative entropy, informal version of Definition 3.3.4)**.** Let n be a security parameter, and (Y, X) be a pair of random variables jointly distributed over strings of length $\mathrm{poly}(n)$. We say that (Y, X) has *hardness at least* Δ *in relative entropy* if the following holds.

Let $\widetilde{G} = (\widetilde{G}_1, \widetilde{G}_2)$ and S be probabilistic $\mathrm{poly}(n)$-time algorithms such that $\mathrm{Supp}(\widetilde{G}(\widetilde{R}))$ is contained

in $\mathrm{Supp}((Y,X))$, where $\widetilde{R}$ is uniformly distributed. Then writing $\widetilde{Y} = \widetilde{G}_1(\widetilde{R})$, we require that

$$\mathrm{KL}\big(\widetilde{R}, \widetilde{Y} \,\big\|\, \mathsf{S}(Y), Y\big) \geq \Delta.$$

Similarly to Lemma 3.1.4, we can show that one-way functions achieve this notion of hardness in relative entropy.

Lemma 3.1.10. *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function and let X be uniformly distributed in $\{0,1\}^n$. Then $(f(X), X)$ has hardness $\omega(\lg n)$ in relative entropy.*

Note that this lemma implies Lemma 3.1.4. If we take $\widetilde{G}$ to be the "honest" sampler $\widetilde{G}(x) = (f(x), x)$, then we have:

$$\mathrm{KL}(X, f(X) \,\|\, \mathsf{S}(Y), Y) = \mathrm{KL}\big(\widetilde{R}, \widetilde{Y} \,\big\|\, \mathsf{S}(Y), Y\big),$$

which is is $\omega(\lg n)$ by Lemma 3.1.10. That is, relative pseudoentropy (as in Definition 3.1.3 and Lemma 3.1.4) is obtained by fixing $\widetilde{G}$ and focusing on the hardness for the simulator S, i.e. the divergence $\mathrm{KL}(X, Y \,\|\, S(Y), Y)$. Furthermore, the step of breaking into short blocks (Lemma 3.1.5) is equivalent to requiring the simulator be *online* and showing that relative pseudoentropy implies the following notion of *next-block relative pseudoentropy*:

Definition 3.1.11 (next-block relative pseudoentropy, informal)**.** Let n be a security parameter, (X, Y) be jointly distributed random variables over strings of length $\mathrm{poly}(n)$, and let $X = (X_1, \ldots, X_m)$ be a partition of X into blocks. We say that X has *next-block relative pseudoentropy at least Δ given Y* if for all probabilistic polynomial-time S, we have

$$\sum_{i=1}^{m} \mathrm{KL}(X_i | X_{<i}, Y \,\|\, \mathsf{S}(X_{<i}, Y) | X_{<i}, Y) \geq \Delta,$$

where we use the notation $z_{<i} \stackrel{\text{def}}{=} (z_1, \ldots, z_{i-1})$.

Here, the simulator S is required to be "online" in the sense that it cannot simulate $(X_1, \ldots, X_m)$ at once, but must simulate X_i only as a function of $X_{<i}$ and Y.

In particular, Lemma 3.1.6 is thus equivalent to the statement that having next-block relative pseudoentropy at least Δ for blocks of length $O(\lg n)$ is equivalent to having next-block pseudoentropy at least $\Delta + \sum_{i=1}^{m} \mathrm{H}(X_i|X_{<i}, Y)$ in the sense of Definition 3.1.1.

Conversely, we show that inaccessible entropy arises from hardness in relative entropy by first requiring the *generator* G to be online and breaking the relative entropy into blocks to obtain the following next-block hardness property.

Definition 3.1.12 (next-block hardness in relative entropy, informal)**.** Let n be a security parameter, and $Y = (Y_1, \ldots, Y_m)$ be a random variable distributed on strings of length $\mathrm{poly}(n)$. We say that Y has *next-block hardness at least* Δ *in relative entropy* if the following holds.

Let $\widetilde{G}$ be any probabilistic $\mathrm{poly}(n)$-time algorithm that takes a sequence of uniformly random strings $\widetilde{R} = (\widetilde{R}_1, \ldots, \widetilde{R}_m)$ and outputs a sequence $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_m)$ in an "online fashion" by which we mean that $\widetilde{Y}_i = \widetilde{G}(\widetilde{R}_1, \ldots, \widetilde{R}_i)$ depends on only the first i random strings of $\widetilde{G}$ for $i = 1, \ldots, m$. Suppose further that $\mathrm{Supp}(\widetilde{Y}) \subseteq \mathrm{Supp}(Y)$. Additionally, let S be a probabilistic $\mathrm{poly}(n)$-time algorithm such for all $i = 1, \ldots, m$, S takes as input $\widehat{R}_1, \ldots, \widehat{R}_{i-1}$ and Y_i and outputs $\widehat{R}_i$, where $\widehat{R}_j$ has the same length as $\widetilde{R}_j$. Then we require that for all such $(\widetilde{G}, \mathsf{S})$, we have:

$$\sum_{i=1}^{m} \mathrm{KL}\big(\widetilde{R}_i, \widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \,\big\|\, \widehat{R}_i, Y_i|\widehat{R}_{<i}, Y_{<i}\big) \geq \Delta.$$

Observe that hardness in relative entropy can be seen as the specific case of next-block hardness in relative entropy when there is only one block (*i.e.,* setting $m = 1$ in the previous definition).

Next, we fix the *simulator*, analogously to how relative pseudoentropy was obtained by fixing the generator, and obtain *next-block inaccessible relative entropy*:

Definition 3.1.13 (next-block inaccessible relative entropy, informal)**.** Let n be a security parameter, and $Y = (Y_1, \ldots, Y_m)$ be a random variable distributed on strings of length $\mathrm{poly}(n)$. We say that Y has *next-block inaccessible relative entropy at least* Δ if the following holds.

Let $\widetilde{G}$ be any probabilistic $\mathrm{poly}(n)$-time algorithm that takes a sequence of uniformly random strings $\widetilde{R} = (\widetilde{R}_1, \ldots, \widetilde{R}_m)$ and outputs a sequence $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_m)$ in an online fashion, and such that $\mathrm{Supp}(\widetilde{Y}) \subseteq \mathrm{Supp}(Y)$. Then we require that for all such $\widetilde{G}$, we have:

$$\sum_{i=1}^{m} \mathrm{KL}\big(\widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i | R_{<i}, Y_{<i}\big) \geq \Delta,$$

where $R = (R_1, \ldots, R_m)$ is a dummy random variable independent of Y.

That is, the goal of the online generator $\widetilde{G}$ is to generate $\widetilde{Y}_i$ given the history of coin tosses $\widetilde{R}_{<i}$ with the same conditional distribution as Y_i given $Y_{<i}$. As promised, there is no explicit simulator in the definition of next-block inaccessible relative entropy, as we essentially dropped all $\widehat{R}$ variables from the definition of next-block hardness in relative entropy. Nevertheless we can obtain it from hardness in relative entropy by using sufficiently short blocks:

Lemma 3.1.14. *Let n be a security parameter, let Y be a random variable distributed on strings of length* $\mathrm{poly}(n)$*, and let $Y = (Y_1, \ldots, Y_m)$ be a partition of Y into blocks of length $O(\lg n)$.*

If $(Y_1, \ldots, Y_m)$ has next-block hardness at least Δ in relative entropy, then $(Y_1, \ldots, Y_m)$ has next-block inaccessible relative entropy at least $\Delta - \mathrm{negl}(n)$.

An intuition for the proof is that since the blocks are of logarithmic length, given Y_i we can simulate the corresponding coin tosses of $\widetilde{R}_i$ of $\widetilde{G}$ by rejection sampling and succeed with high probability in $\mathrm{poly}(n)$ tries.

A nice feature of the definition of next-block inaccessible relative entropy compared to inaccessible

entropy is that it is meaningful even for non-flat random variables, as the Kullback–Leibler divergence is always nonnegative. Moreover, for flat random variables, it equals the inaccessible entropy:

Lemma 3.1.15. *Suppose $Y = (Y_1, \ldots, Y_m)$ is a flat random variable. Then Y has next-block inaccessible relative entropy at least Δ if and only if Y has accessible entropy at most $\mathrm{H}(Y) - \Delta$.*

Intuitively, this lemma comes from the identity that if Y is a flat random variable and $\mathrm{Supp}(\widetilde{Y}) \subseteq \mathrm{Supp}(Y)$, then $\mathrm{H}(\widetilde{Y}) = \mathrm{H}(Y) - \mathrm{KL}(\widetilde{Y} \parallel Y)$. We stress that we do not require the individual blocks Y_i have flat distributions, only that the random variable Y as a whole is flat. For example, if f is a function and X is uniform, then $(f(X), X)$ is flat even though $f(X)$ itself may be far from flat.

Putting together Lemmas 3.1.10, 3.1.14, and 3.1.15, we obtain a new, more modular (and slightly tighter) proof of Theorem 3.1.8. The reduction implicit in the combination of these lemmas is the same as the one in [HRVW09], but the analysis is different. (In particular, [HRVW09] makes no use of KL divergence.) Like the existing proof of Theorem 3.1.2, this proof separates the move from one-wayness to a form of hardness involving relative entropies, the role of short blocks, and the move from hardness in relative entropy to computational entropy, as summarized in Figure 3.1. Moreover, this further illumination of and toolkit for notions of computational entropy may open the door to other applications in cryptography.

We remark that another interesting direction for future work is to find a construction of universal one-way hash functions (UOWHFs) from one-way functions that follows a similar template to the above constructions of PRGs and SHCs. There is now a construction of UOWHFs based on a variant of inaccessible entropy [HHRVW10], but it remains more complex and inefficient than those of PRGs and SHCs.

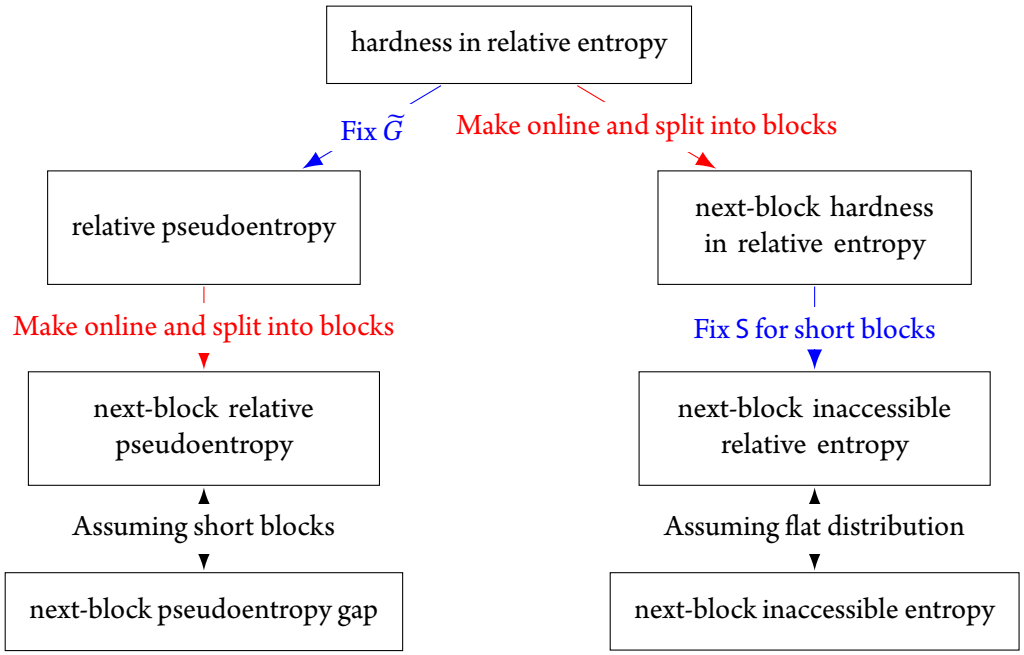

Figure 3.1: Relationships between hardness notions.

3.2 Preliminaries

Notations. For a tuple $x = (x_1, \dots, x_n)$, we write $x_{\leq i}$ for $(x_1, \dots, x_i)$, and $x_{<i}$ for $(x_1, \dots, x_{i-1})$.

poly denotes the set of polynomial functions and negl the set of all negligible functions: $\varepsilon \in \text{negl}$ if for all $p \in \text{poly}$ and large enough $n \in \mathbb{N}$, $\varepsilon(n) \leq 1/p(n)$. We will sometimes abuse notations and write $\text{poly}(n)$ to mean $p(n)$ for some $p \in \text{poly}$ and similarly for $\text{negl}(n)$.

PPT stands for probabilistic polynomial time and can be either in the uniform or non-uniform model of computation. All our results are stated as uniform polynomial time oracle reductions and are thus meaningful in both models.

For a random variable X over $\mathcal{X}$, $\text{Supp}(X) \stackrel{\text{def}}{=} \{x \in \mathcal{X} \ : \ \Pr[X = x] > 0\}$ denotes the support of X. A random variable is *flat* if it is uniform over its support. Random variables will be written with uppercase letters and the associated lowercase letter represents a generic element from its support.

Information theory.

Definition 3.2.1 (Entropy)**.** For a random variable X and $x \in \mathrm{Supp}(X)$, the *sample entropy* (also called surprise) of x is $\mathrm{H}^*_x(X) \stackrel{\text{def}}{=} \lg(1/\Pr[X = x])$. The *entropy* $\mathrm{H}(X)$ of X is the expected sample entropy: $\mathrm{H}(X) \stackrel{\text{def}}{=} \mathbb{E}_{x \leftarrow X}[\mathrm{H}^*_x(X)]$.

Definition 3.2.2 (Conditional entropy)**.** Let (A, X) be a pair of random variables and consider $(a, x) \in \mathrm{Supp}(A, X)$, the *conditional sample entropy* of (a, x) is $\mathrm{H}^*_{a|x}(A|X) \stackrel{\text{def}}{=} \lg(1/\Pr[A = a \mid X = x])$ and the *conditional entropy* of A given X is the expected conditional sample entropy:

$$\mathrm{H}(A|X) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(a,x)\leftarrow(A,X)}\left[\lg \frac{1}{\Pr[A = a \mid X = x]}\right].$$

Proposition 3.2.3 (Chain rule for entropy)**.** *Let (A, X) be a pair of random variables, then* $\mathrm{H}(A, X) = \mathrm{H}(A|X) + \mathrm{H}(X)$ *and for* $(a, x) \in \mathrm{Supp}(A, X)$, $\mathrm{H}^*_{a,x}(A, X) = \mathrm{H}^*_{a|x}(A|X) + \mathrm{H}^*_x(X)$.

Definition 3.2.4 (Relative entropy[2])**.** For a pair (A, B) of random variables and $(a, b) \in \mathrm{Supp}(A, B)$ the *sample relative entropy* (lg-probability ratio) is:

$$\mathrm{KL}^*_a(A \parallel B) \stackrel{\text{def}}{=} \lg \frac{\Pr[A = a]}{\Pr[B = a]},$$

and the *relative entropy* of A with respect to B is the expected sample relative entropy:

$$\mathrm{KL}(A \parallel B) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{a\leftarrow A}\left[\lg \frac{\Pr[A = a]}{\Pr[B = a]}\right].$$

Definition 3.2.5 (Conditional relative entropy)**.** For pairs of random variables (A, X) and (B, Y), and

[2] Relative entropy is another name for the Kullback–Leibler divergence, but in this chapter we prefer the relative entropy terminology to better match existing cryptographic notions of entropy and for uniformity across the notions discussed in this chapter.

$(a, x) \in \mathrm{Supp}(A, X)$, the *conditional sample relative entropy* is:

$$\mathrm{KL}^*_{a|x}(A|X \,||\, B|Y) \stackrel{\text{def}}{=} \lg \frac{\Pr[A = a|X = x]}{\Pr[B = a|Y = x]},$$

and the *conditional relative entropy* is:

$$\mathrm{KL}(A|X \,||\, B|Y) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(a,x)\leftarrow(A,X)}\left[\lg \frac{\Pr[A = a|X = x]}{\Pr[B = a|Y = x]}\right].$$

Proposition 3.2.6 (Chain rule for relative entropy)**.** *For pairs of random variables* (X, A) *and* (Y, B)*:*

$$\mathrm{KL}(A, X \,||\, B, Y) = \mathrm{KL}(A|X \,||\, B|Y) + \mathrm{KL}(X \,||\, Y),$$

and for $(a, x) \in \mathrm{Supp}(A, X)$*:*

$$\mathrm{KL}^*_{a,x}(A, X \,||\, B, Y) = \mathrm{KL}^*_{a|x}(A|X \,||\, B|Y) + \mathrm{KL}^*_x(X \,||\, Y).$$

Proposition 3.2.7 (Data-processing inequality)**.** *Let* (X, Y) *be a pair of random variables and let* f *be a function defined on* $\mathrm{Supp}(Y)$*, then:*

$$\mathrm{KL}(X \,||\, Y) \geq \mathrm{KL}(f(X) \,||\, f(Y)).$$

Definition 3.2.8 (min relative entropy)**.** Let (X, Y) be a pair of random variables and $\delta \in [0, 1]$. We define $\mathrm{KL}^{\delta}_{\min}(X \,||\, Y)$ to be the quantile of level δ of $\mathrm{KL}^*_x(X \,||\, Y)$, equivalently it is the smallest $\Delta \in \mathbb{R}$ satisfying:

$$\Pr_{x\leftarrow X}[\mathrm{KL}^*_x(X \,||\, Y) \leq \Delta] \geq \delta,$$

and it is characterized by the following equivalence:

$$\mathrm{KL}^{\delta}_{\min}(X \,||\, Y) > \Delta \iff \Pr_{x\leftarrow X}[\mathrm{KL}^*_x(X \,||\, Y) \leq \Delta] < \delta.$$

Block generators

Definition 3.2.9 (Block generator)**.** An *m-block generator* is a function $\mathsf{G} : \{0,1\}^s \to \prod_{i=1}^m \{0,1\}^{\ell_i}$. $\mathsf{G}_i(r)$ denotes the i-th block of G on input r and $|\mathsf{G}_i| = \ell_i$ denotes the bit length of the i-th block.

Definition 3.2.10 (Online generator)**.** An *online m-block generator* is a function $\widetilde{G} : \prod_{i=1}^m \{0,1\}^{s_i} \to \prod_{i=1}^m \{0,1\}^{\ell_i}$ such that for all $i \in [m]$ and $r \in \prod_{i=1}^m \{0,1\}^{s_i}$, $\widetilde{G}_i(r)$ only depends on $r_{\leq i}$. We sometimes write $\widetilde{G}_i(r_{\leq i})$ when the input blocks $i+1, \ldots, m$ are unspecified.

Definition 3.2.11 (Support)**.** The *support* of a generator G is the support of the random variable $\mathrm{Supp}\big(\mathsf{G}(R)\big)$ for uniform input R. If G is an $(m+1)$-block generator, and Π is a binary relation, we say that G is *supported on* Π if $\mathrm{Supp}\big(\mathsf{G}_{\leq m}(R), \mathsf{G}_{m+1}(R)\big) \subseteq \Pi$.

When G is an $(m+1)$-block generator supported on a binary relation Π, we will often use the notation $\mathsf{G}_{\mathsf{w}} \stackrel{\text{def}}{=} \mathsf{G}_{m+1}$ to emphasize that the last block corresponds to a witness for the first m blocks.

Cryptography.

Definition 3.2.12 (One-way Function)**.** Let n be a security parameter, $t = t(n)$ and $\varepsilon = \varepsilon(n)$. A function $f : \{0,1\}^n \to \{0,1\}^n$ is a (t,ε)-*one-way function* if:

1. For all time t randomized algorithm A: $\Pr_{x \leftarrow U_n}\big[\mathsf{A}\big(f(x)\big) \in f^{-1}\big(f(x)\big)\big] \leq \varepsilon$, where U_n is uniform over $\{0,1\}^n$.

2. There exists a polynomial time algorithm B such that $\mathsf{B}(x) = f(x)$ for all $x \in \{0,1\}^n$.

If f is $(n^c, 1/n^c)$-one-way for every $c \in \mathbb{N}$, we say that f is *(strongly) one-way*.

3.3 Search Problems and Hardness in Relative Entropy

In this section, we first present the classical notion of hard-on-average search problems and introduce the new notion of hardness in relative entropy. We then relate the two notions by proving that average-case hardness implies hardness in relative entropy.

3.3.1 Search problems

For a binary relation $\Pi \subseteq \{0,1\}^* \times \{0,1\}^*$, we write $\Pi(y,w)$ for the predicate that is true iff $(y,w) \in \Pi$ and say that w is a *witness* for the *instance* y[3]. To each relation Π, we naturally associate (1) a *search problem*: given y, find w such that $\Pi(y,w)$ or state that no such w exist and (2) the *decision problem* defined by the language $L_\Pi \stackrel{\text{def}}{=} \{y \in \{0,1\}^* \ : \ \exists w \in \{0,1\}^*, \ \Pi(y,w)\}$. **FNP** denotes the set of all relations Π computable by a polynomial time algorithm and such that there exists a polynomial p such that $\Pi(y,w) \Rightarrow |w| \leq p(|y|)$. Whenever $\Pi \in \mathbf{FNP}$, the associated decision problem L_Π is in **NP**. We now define average-case hardness.

Definition 3.3.1 (distributional search problem)**.** A *distributional search problem* is a pair (Π, Y) where $\Pi \subseteq \{0,1\}^* \times \{0,1\}^*$ is a binary relation and Y is a random variable supported on L_Π.

The problem (Π, Y) is *(t, ε)-hard* if $\Pr\left[\Pi\big(Y, \mathsf{A}(Y)\big)\right] \leq \varepsilon$ for all time t randomized algorithm A, where the probability is over the distribution of Y and the randomness of A.

Example 3.3.2. For $f \ : \ \{0,1\}^n \to \{0,1\}^n$, the problem of inverting f is the search problem associated with the relation $\Pi^f \stackrel{\text{def}}{=} \{(f(x),x) \ : \ x \in \{0,1\}^n\}$. If f is a (t,ε)-one-way function, then the distributional search problem $\big(\Pi^f, f(X)\big)$ of inverting f on a uniform random input $X \in \{0,1\}^n$ is

[3]We used the unconventional notation y for the instance (instead of x) because our relations will often be of the form Π^f for some function f; in this case an instance is some y in the range of f and a witness for y is any preimage $x \in f^{-1}(y)$.

(t, ε)-hard.

Remark 3.3.3. Consider a distributional search problem (Π, Y). Without loss of generality, there exists a (possibly inefficient) two-block generator $\mathsf{G} = (\mathsf{G}_1, \mathsf{G}_\mathsf{w})$ supported on Π such that $\mathsf{G}_1(R) = Y$ for uniform input R. If G_w is polynomial-time computable, it is easy to see that the search problem $\left(\Pi^{\mathsf{G}_1}, \mathsf{G}_1(R)\right)$ is at least as hard as (Π, Y). The advantage of writing the problem in this "functional" form is that the distribution $(\mathsf{G}_1(R), R)$ over (instance, witness) pairs is flat, which is a necessary condition to relate hardness to inaccessible entropy (see Theorem 3.4.12).

Furthermore, if G_1 is also polynomial-time computable and (Π, Y) is $(\mathrm{poly}(n), \mathrm{negl}(n))$-hard, then $R \mapsto \mathsf{G}_1(R)$ is a one-way function. Combined with the previous example, we see that the existence of one-way functions is equivalent to the existence of $(\mathrm{poly}(n), \mathrm{negl}(n))$-hard search problems for which (instance, witness) pairs can be efficiently sampled.

3.3.2 Hardness in relative entropy

Instead of considering an adversary directly attempting to solve a search problem (Π, Y), the adversary in the definition of hardness in relative entropy comprises a pair of algorithm $(\widetilde{G}, \mathsf{S})$ where $\widetilde{G}$ is a two-block generator outputting valid (instance, witness) pairs for Π and S is a *simulator* for $\widetilde{G}$: given an instance y, the goal of S is to output randomness r for $\widetilde{G}$ such that $\widetilde{G}_1(r) = y$. Formally, the definition is as follows.

Definition 3.3.4 (hardness in relative entropy)**.** Let (Π, Y) be a distributional search problem. We say that (Π, Y) has *hardness* (t, Δ) *in relative entropy* if:

$$\mathrm{KL}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}(Y), Y\big) > \Delta \,,$$

for all pairs $(\widetilde{G}, \mathsf{S})$ of time t algorithms where $\widetilde{G}$ is a two-block generator supported on Π and $\widetilde{R}$ is

uniform randomness for $\widetilde{G}_1$. Similarly, for $\delta \in [0,1]$, (Π, Y) has *hardness* (t, Δ) *in* δ*-min relative entropy* if for all such pairs:

$$\mathrm{KL}^{\delta}_{\min}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \parallel \mathsf{S}(Y), Y\big) > \Delta \,.$$

Note that a pair $(\widetilde{G}, \mathsf{S})$ achieves a relative entropy of zero in Definition 3.3.4 if $\widetilde{G}_1(R)$ has the same distribution as Y and if $\widetilde{G}_1\big(\mathsf{S}(y)\big) = y$ for all $y \in \mathrm{Supp}(Y)$. In this case, writing $\widetilde{G}_{\mathsf{w}} \stackrel{\text{def}}{=} \widetilde{G}_2$, we have that $\widetilde{G}_{\mathsf{w}}\big(\mathsf{S}(Y)\big)$ is a valid witness for Y since $\widetilde{G}$ is supported on Π.

More generally, the composition $\widetilde{G}_{\mathsf{w}} \circ \mathsf{S}$ solves the search problem (Π, Y) whenever $\widetilde{G}_1\big(\mathsf{S}(Y)\big) = Y$. When the relative entropies in Definition 3.3.4 are upper-bounded, we can lower bound the probability of the search problem being solved (Lemma 3.3.7). This immediately implies that hard search problems are also hard in relative entropy.

Theorem 3.3.5. *Let* (Π, Y) *be a distributional search problem. If* (Π, Y) *is* (t, ε)*-hard, then it has hardness* (t', Δ') *in relative entropy and* (t', Δ'') *in* δ*-min relative entropy for every* $\delta \in [0,1]$ *where* $t' = \Omega(t)$,[4] $\Delta' = \lg(1/\varepsilon)$ *and* $\Delta'' = \lg(1/\varepsilon) - \lg(1/\delta)$.

Remark 3.3.6. As we see, a "good" simulator S for a generator $\widetilde{G}$ is one for which $\widetilde{G}_1\big(\mathsf{S}(Y)\big) = Y$ holds often. It will be useful in Section 3.4 to consider simulators S which are allowed to fail by outputting a failure string $r \notin \mathrm{Supp}(\widetilde{R})$, (e.g. $r = \perp$) and adopt the convention that $\widetilde{G}_1(r) = \perp$ whenever $r \notin \mathrm{Supp}(\widetilde{R})$. With this convention, we can without loss of generality add the requirement that $\widetilde{G}_1\big(\mathsf{S}(Y)\big) = Y$ whenever $\mathsf{S}(Y) \in \mathrm{Supp}(\widetilde{R})$: indeed, S can always check that it is the case and if not output a failure symbol. For such a simulator S, observe that for all $r \in \mathrm{Supp}(\widetilde{R})$, the second variable on both sides of the relative entropy in Definition 3.3.4 is obtained by applying $\widetilde{G}_1$ on the first variable and can thus be dropped, leading to a simpler definition of hardness in relative entropy: $\mathrm{KL}\big(\widetilde{R} \parallel \mathsf{S}(Y)\big) > \Delta$.

[4] For the theorems in this paper that relate two notions of hardness, the notation $t' = \Omega(t)$ means that there exists a constant C depending *only* on the computational model such that $t' \geq C \cdot t$.

Theorem 3.3.5 is an immediate consequence of the following lemma.

Lemma 3.3.7. *Let (Π, Y) be a distributional search problem and $(\widetilde{G}, \mathsf{S})$ be a pair of algorithms with $\widetilde{G} = (\widetilde{G}_1, \widetilde{G}_w)$ a two-block generator supported on Π. Define the linear-time oracle algorithm $\mathsf{A}^{\widetilde{G}_w,\mathsf{S}}(y) \stackrel{\text{def}}{=} \widetilde{G}_w(\mathsf{S}(y))$. For $\Delta \in \mathbb{R}^+$ and $\delta \in [0,1]$:*

1. *If $\mathrm{KL}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}(Y), Y\big) \leq \Delta$ then $\Pr\Big[\Pi(Y, \mathsf{A}^{\widetilde{G}_w,\mathsf{S}}(Y))\Big] \geq 1/2^{\Delta}$.*

2. *If $\mathrm{KL}^{\delta}_{\min}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}(Y), Y\big) \leq \Delta$ then $\Pr\Big[\Pi(Y, \mathsf{A}^{\widetilde{G}_w,\mathsf{S}}(Y))\Big] \geq \delta/2^{\Delta}$.*

Proof. We have:

$$\begin{aligned}
\Pr\Big[\Pi\big(Y, \mathsf{A}^{\widetilde{G}_w,\mathsf{S}}(Y)\big)\Big] &= \Pr\big[\Pi(Y, \widetilde{G}_w(\mathsf{S}(Y)))\big] \\
&\geq \Pr\big[\widetilde{G}_1(\mathsf{S}(Y)) = Y\big] && (\widetilde{G} \text{ is supported on } \Pi) \\
&= \sum_{r \in \mathrm{Supp}(\widetilde{R})} \Pr\big[\mathsf{S}(Y) = r \wedge Y = \widetilde{G}_1(r)\big] \\
&= \mathop{\mathbb{E}}_{r \leftarrow \widetilde{R}}\left[\frac{\Pr\big[\mathsf{S}(Y) = r \wedge Y = \widetilde{G}_1(r)\big]}{\Pr\big[\widetilde{R} = r\big]}\right] \\
&= \mathop{\mathbb{E}}_{\substack{r \leftarrow \widetilde{R} \\ y \leftarrow \widetilde{G}_1(r)}}\left[2^{-\mathrm{KL}^*_{r,y}\left(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\|\, \mathsf{S}(Y), Y\right)}\right].
\end{aligned}$$

Now, the first claim follows by Jensen's inequality (since $x \mapsto 2^{-x}$ is convex) and the second claim follows by Markov' inequality when considering the event that the sample relative entropy is smaller than Δ (which occurs with probability at least δ by assumption). □

Relation to relative pseudoentropy. In [VZ12], the authors introduced the notion of relative pseudoentropy[5]: for jointly distributed variables (Y, W), W has relative pseudoentropy given Y if it is

[5] As already mentioned in the introduction, this notion was in fact called "KL-hardness for sampling" in [VZ12] but we rename it here to unify the terminology between the various notions discussed here.

hard for a polynomial time adversary to approximate—measured in relative entropy—the conditional distribution W given Y. Formally:

Definition 3.3.8 (relative pseudoentropy, Def. 3.4 in [VZ12])**.** Let (Y, W) be a pair of random variables, we say that W has *relative pseudoentropy* (t, Δ) *given* Y if for all time t randomized algorithm S, we have:

$$\mathrm{KL}(Y, W \parallel Y, \mathsf{S}(Y)) > \Delta\,.$$

As discussed in Section 3.1.2, it was shown in [VZ12] that if $f : \{0,1\}^n \to \{0,1\}^n$ is a one-way function, then $\big(f(X), X_1, \ldots, X_n\big)$ has next-bit pseudoentropy for uniform $X \in \{0,1\}^n$ (see Theorem 3.1.2). The first step in proving this result was to prove that X has relative pseudoentropy given $f(X)$ (see Lemma 3.1.4).

We observe that when (Y, W) is of the form $(f(X), X)$ for some function $f : \{0,1\}^n \to \{0,1\}^n$ and variable X over $\{0,1\}^n$, then relative pseudoentropy is implied by hardness in relative entropy by simply fixing $\widetilde{G}$ to be the "honest sampler" $\widetilde{G}(X) = (f(X), X)$. Indeed, in this case we have:

$$\mathrm{KL}\big(X, \widetilde{G}_1(X) \parallel \mathsf{S}(Y), Y\big) = \mathrm{KL}(X, f(X) \parallel \mathsf{S}(Y), Y)\,.$$

We can thus recover Lemma 3.1.4 as a direct corollary of Theorem 3.3.5.

Corollary 3.3.9. *Consider a function* $f : \{0,1\}^n \to \{0,1\}^n$ *and define* $\Pi^f \stackrel{\text{def}}{=} \{(f(x), x) : x \in \{0,1\}^n\}$ *and* $Y \stackrel{\text{def}}{=} f(X)$ *for* X *uniform over* $\{0,1\}^n$*. If* f *is* (t, ε)*-one-way, then* (Π^f, Y) *has hardness* $\big(t', \lg(1/\varepsilon)\big)$ *in relative entropy and* X *has relative pseudoentropy* $\big(t', \lg(1/\varepsilon)\big)$ *given* Y *with* $t' = \Omega(t)$*.*

Witness hardness in relative entropy. We also introduce a relaxed notion of hardness in relative entropy called witness hardness in relative entropy. In this notion, we further require $(\widetilde{G}, \mathsf{S})$ to approximate the joint distribution of (instance, witness) pairs rather than only instances. For example, the

problem of inverting a function f over a random input X is naturally associated with the distribution $\big(f(X), X\big)$. The relaxation in this case is analogous to the notion of *distributional one-way function* for which the adversary is required to approximate the uniform distribution over preimages.

Definition 3.3.10 (witness hardness in relative entropy)**.** Let Π be a binary relation and (Y, W) be a pair of random variables supported on Π. We say that (Π, Y, W) has *witness hardness* (t, Δ) *in relative entropy* if for all pairs of time t algorithms $(\widetilde{G}, \mathsf{S})$ where $\widetilde{G}$ is a two-block generator supported on Π, for uniform $\widetilde{R}$:

$$\mathrm{KL}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}), \widetilde{G}_{\mathsf{w}}(\widetilde{R}) \,\big\|\, \mathsf{S}(Y), Y, W\big) > \Delta \,.$$

Similarly, for $\delta \in [0,1]$, (Π, Y, W) has *witness hardness* (t, Δ) *in* δ*-min relative entropy*, if for all such pairs:

$$\mathrm{KL}^{\delta}_{\min}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}), \widetilde{G}_{\mathsf{w}}(\widetilde{R}) \,\big\|\, \mathsf{S}(Y), Y, W\big) > \Delta \,.$$

We introduced hardness in relative entropy first, since it is the notion which is most directly obtained from the hardness of distributional search problems. Observe that by the data processing inequality for relative entropy (Proposition 3.2.7), dropping $\widetilde{G}_{\mathsf{w}}(\widetilde{R})$ and W in the relative entropies in Definition 3.3.10 only decreases them. Hence, hardness in relative entropy (as in Theorem 3.3.5) implies witness hardness (as in Theorem 3.3.11 below). As we will see in Section 3.4 witness hardness in relative entropy is the "correct" notion to obtain inaccessible entropy from: it is in fact equal to inaccessible entropy up to 1/ poly losses.

Theorem 3.3.11. *Let* Π *be a binary relation and* (Y, W) *be a pair of random variables supported on* Π*. If* (Π, Y) *is* (t, ε)*-hard, then* (Π, Y, W) *has witness hardness* (t', Δ') *in relative entropy and* (t', Δ'') *in* δ*-min relative entropy for every* $\delta \in [0,1]$ *where* $t' = \Omega(t)$*,* $\Delta' = \lg(1/\varepsilon)$ *and* $\Delta'' = \lg(1/\varepsilon) - \lg(1/\delta)$*.*

Remark 3.3.12. The data processing inequality does not hold exactly for $\mathrm{KL}_{\min}$, hence the statement

about δ-min relative entropy in Theorem 3.3.11 does not follow with the claimed parameters in a black-box manner from Theorem 3.3.5. However, an essentially identical proof yields the result.

3.4 Inaccessible Entropy and Hardness in Relative Entropy

In this section, we relate our notion of witness hardness in relative entropy to the inaccessible entropy definition of [HRVW16]. Roughly speaking, we "split" the relative entropy into blocks and obtain the intermediate notion of next-block inaccessible relative entropy (Section 3.4.1) which we then relate to inaccessible entropy (Section 3.4.2). Together, these results show that if f is a one-way function, the generator $\mathsf{G}^f(X) = \big(f(X)_1, \ldots, f(X)_n, X\big)$ has superlogarithmic inaccessible entropy.

3.4.1 Next-block hardness and rejection sampling

For an online (adversarial) generator $\widetilde{G}$, it is natural to consider simulators S that also operate in an online fashion. That is:

Definition 3.4.1 (online simulator)**.** Let $\widetilde{G} : \prod_{i=1}^{m} \{0,1\}^{s_i} \to \prod_{i=1}^{m} \{0,1\}^{\ell_i}$ be an online m-block generator. An *online simulator* for $\widetilde{G}$ is a PPT algorithm S such that for all $y = (y_1, \ldots, y_m) \in \prod_{i=1}^{m} \{0,1\}^{\ell_i}$, defining inductively $\hat{r}_i \stackrel{\text{def}}{=} \mathsf{S}(\hat{r}_{<i}, y_i) \in \{0,1\}^{s_i}$, we have for all $i \in [m]$:

$$\widetilde{G}_i(\hat{r}_{\le i}) = y_i \quad \text{or} \quad \hat{r}_i = \bot\,.$$

The *running time* of S is the total amount of time required to compute $\hat{r}_1, \ldots, \hat{r}_m$.

The goal of such an online simulator S is to ensure that the distribution of $\widehat{R}_i = \mathsf{S}(\hat{r}_{<i}, y_i)$ is close to that of $\widetilde{R}_i | (\widetilde{R}_{<i} = \hat{r}_{<i}, \widetilde{Y}_i = y_i)$ where $(\widetilde{Y}_1, \ldots, \widetilde{Y}_m) \stackrel{\text{def}}{=} \widetilde{G}(\widetilde{R}_{\le m})$ for uniformly random $(\widetilde{R}_1, \ldots, \widetilde{R}_m)$. Equivalently, $\widehat{R}_i$ should be close to uniform on $\{\hat{r}_i : \widetilde{G}_i(\hat{r}_{\le i}) = y_i\}$. Measuring closeness with relative

entropy, we have:

Definition 3.4.2 (next-block hardness in relative entropy)**.** The joint distribution $Y = (Y_1, \dots, Y_m)$ has *next-block hardness* (t, Δ) *in relative entropy* if the following holds for every time t online m-block generator $\widetilde{G}$ and every time t online simulator S for $\widetilde{G}$.

Write $\widetilde{Y}_{\leq m} \stackrel{\text{def}}{=} \widetilde{G}(\widetilde{R}_{\leq m})$ for uniform $\widetilde{R}_{\leq m}$, and define inductively $\widehat{R}_i \stackrel{\text{def}}{=} \mathsf{S}(\widehat{R}_{<i}, Y_i)$. Then we require:

$$\sum_{i=1}^{m} \mathrm{KL}\Big(\widetilde{R}_i, \widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \;\Big\|\; \widehat{R}_i, Y_i | \widehat{R}_{<i}, Y_{<i}\Big) > \Delta\,.$$

Similarly, for $\delta \in [0,1]$, we say that $(Y_1, \dots, Y_m)$ has *next-block hardness* (t, Δ) *in* δ*-min relative entropy* if, with the same notations as above:

$$\Pr_{\substack{r_{\leq m} \leftarrow \widetilde{R}_{\leq m} \\ y_{\leq m} \leftarrow \widetilde{G}(r_{\leq m})}} \left[\sum_{i=1}^{m} \mathrm{KL}^{*}_{r_i, y_i | r_{<i}, y_{<i}}\Big(\widetilde{R}_i, \widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \;\Big\|\; \widehat{R}_i, Y_i | \widehat{R}_{<i}, Y_{<i}\Big) \leq \Delta \right] < \delta\,.$$

Observe that using the chain rule for relative entropy, the sum of relative entropies appearing in Definition 3.4.2 is exactly equal to the relative entropies appearing in Definition 3.3.4. Since, furthermore considering an online generator $\widetilde{G}$ and online simulator S is only less general than arbitrary pairs $(\widetilde{G}, \mathsf{S})$, we immediately obtain the following theorem.

Theorem 3.4.3. *Let* (Π, Y) *be a distributional search problem. If* (Π, Y) *has hardness* (t, Δ) *in relative entropy then* $(Y_1, \dots, Y_m)$ *has next-block hardness* (t, Δ) *in relative entropy.*

Similarly, for any $\delta \in [0,1]$*, if* (Π, Y) *has hardness* (t, Δ) *in* δ*-min relative entropy then* $(Y_1, \dots, Y_m)$ *has next-block hardness* (t, Δ) *in* δ*-min relative entropy.*

Proof. Immediate using the chain rule for relative (sample) entropy. □

The next step is to obtain a notion of hardness that makes no reference to simulators by considering, for an online block generator $\widetilde{G}$, a specific simulator $\mathsf{Sim}^{\widetilde{G}, T}$ which on input $(\hat{r}_{<i}, y_i)$, generates $\widehat{R}_i$ using

rejection sampling until $\widetilde{G}_i(\hat{r}_{<i}, \widehat{R}_i) = y_i$. The superscript T is the maximum number of attempts after which $\mathsf{Sim}^{\widetilde{G},T}$ gives up and outputs $\perp$. The formal definition of $\mathsf{Sim}^{\widetilde{G},T}$ is given in Algorithm 1.

Algorithm 1 Rejection sampling simulator $\mathsf{Sim}^{\widetilde{G},T}$ for $1 \le i \le m$

Input: $y_i \in \{0,1\}^*, \hat{r}_{<i} \in (\{0,1\}^v \cup \{\perp\})^{i-1}$
Output: $\hat{r}_i \in \{0,1\}^v \cup \{\perp\}$
  if $\hat{r}_{i-1} = \perp$ **then**
    $\hat{r}_i \leftarrow \perp$; **return**
  end if
  repeat
    sample $\hat{r}_i \leftarrow \{0,1\}^v$
  until $\widetilde{G}_i(\hat{r}_{\le i}) = y_i$ or $\ge T$ attempts
  if $\widetilde{G}_i(\hat{r}_{\le i}) \ne y_i$ **then**
    $\hat{r}_i \leftarrow \perp$
  end if

For the rejection sampling simulator $\mathsf{Sim}^{\widetilde{G},T}$, we will show in Lemma 3.4.6 that the next-block hardness in relative entropy in Definition 3.4.2 decomposes as the sum of two terms:

1. A term measuring how well $\widetilde{G}_{\le m}$ approximates the distribution Y in an online manner, without any reference to a simulator.

2. An error term measuring the failure probability of the rejection sampling procedure due to having a finite time bound T.

As we show in Lemma 3.4.7, the error term can be made arbitrarily small by setting the number of trials T in $\mathsf{Sim}^{\widetilde{G},T}$ to be a large enough multiple of $m \cdot 2^{\ell}$ where ℓ is the length of the blocks of $\widetilde{G}_{\le m}$. This leads to a poly(m) time algorithm whenever ℓ is logarithmic in m. That is, given an online block generator $\widetilde{G}$ for which $\widetilde{G}_{\le m}$ has short blocks, we obtain a corresponding simulator "for free". Thus, considering only the first term leads to the following clean definition of next-block inaccessible relative entropy that makes no reference to simulators.

Definition 3.4.4 (next-block inaccessible relative entropy)**.** The joint distribution $(Y_1, \dots, Y_m)$ has *next-block inaccessible relative entropy* (t, Δ), if for every time t online m-block generator $\widetilde{G}$ supported on $Y_{\le m}$, writing $\widetilde{Y}_{\le m} \stackrel{\text{def}}{=} \widetilde{G}(\widetilde{R}_{\le m})$ for uniform $\widetilde{R}_{\le m}$, we have:

$$\sum_{i=1}^{m} \mathrm{KL}\big(\widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i | R_{<i}, Y_{<i}\big) > \Delta\,,$$

where R_i is a "dummy" random variable over the domain of $\widetilde{G}_i$ and independent of $Y_{\le m+1}$. Similarly, for $\delta \in [0,1]$, we say that $(Y_1, \dots, Y_{m+1})$ has *next-block inaccessible δ-min relative entropy* (t, Δ) if for every $\widetilde{G}$ as above:

$$\Pr_{\substack{r_{\le m} \leftarrow \widetilde{R}_{\le m} \\ y_{\le m} \leftarrow \widetilde{G}(r_{\le m})}} \left[\sum_{i=1}^{m} \mathrm{KL}^{*}_{y_i | r_{<i}, y_{<i}}\big(\widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i | R_{<i}, Y_{<i}\big) \le \Delta \right] < \delta\,,$$

where $(\widetilde{Y}_{\le m}, \widetilde{R}_{\le m})$ are defined as above.

Remark 3.4.5. Since $\widetilde{Y}_{<i}$ is a function of $\widetilde{R}_{<i}$, the first conditional distribution in the KL is effectively $\widetilde{Y}_i | \widetilde{R}_{<i}$. Similarly the second distribution is effectively $Y_i | Y_{<i}$. The extra random variables are there for syntactic consistency.

With this definition in hand, we can make formal the claim that, even as sample notions, the next-block hardness in relative entropy decomposes as next-block inaccessible relative entropy plus an error term.

Lemma 3.4.6. *For a joint distribution $(Y_1, \dots, Y_m)$, let $\widetilde{G}$ be an online m-block generator supported on $Y_{\le m}$. Define $(\widetilde{Y}_1, \dots, \widetilde{Y}_m) \stackrel{\text{def}}{=} \widetilde{G}(\widetilde{R})$ for uniform random variable $\widetilde{R} = (\widetilde{R}_1, \dots, \widetilde{R}_m)$ and let R_i be a "dummy" random variable over the domain of $\widetilde{G}_i$ and independent of $Y_{\le m}$. We also define $\widehat{R}_i \stackrel{\text{def}}{=} \mathsf{Sim}^{\widetilde{G}, T}(\widehat{R}_{<i}, Y_i)$ and $\widehat{Y}_i = \widetilde{G}(\widehat{R}_{\le i})$. Then, for all $r \in \mathrm{Supp}(\widetilde{R})$ and $y \stackrel{\text{def}}{=} \widetilde{G}(r)$:*

$$\sum_{i=1}^{m} \mathrm{KL}^{*}_{r_i, y_i | r_{<i}, y_{<i}}\Big(\widetilde{R}_i, \widetilde{Y}_i | \widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel \widehat{R}_i, Y_i | \widehat{R}_{<i}, Y_{<i}\Big)$$

$$= \sum_{i=1}^{m} \mathrm{KL}^{*}_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i|R_{<i}, Y_{<i}\big) + \sum_{i=1}^{m} \lg\left(\frac{1}{\Pr\big[\widehat{Y}_i = y_i | Y_i = y_i, \widehat{R}_{<i} = r_{<i}\big]}\right).$$

Moreover, the running time of $\mathsf{Sim}^{\widetilde{G},T}$ *on input* $\widehat{R}_{<i}, Y_i$ *is* $O(|r_i| \cdot T)$*, with at most* T *oracle calls to* $\widetilde{G}$*.*

Proof. Consider $r \in \mathrm{Supp}(\widetilde{R})$ and $y \stackrel{\text{def}}{=} \widetilde{G}(r)$. Then:

$$\begin{aligned}
&\sum_{i=1}^{m} \mathrm{KL}^{*}_{r_i,y_i|r_{<i},y_{<i}}\big(\widetilde{R}_i, \widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel \widehat{R}_i, Y_i|\widehat{R}_{<i}, Y_{<i}\big) \\
&\quad = \sum_{i=1}^{m} \mathrm{KL}^{*}_{r_i,y_i|r_{<i},y_{<i}}\big(\widetilde{R}_i, \widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel \widehat{R}_i, \widehat{Y}_i|\widehat{R}_{<i}, \widehat{Y}_{<i}\big) \\
&\quad = \sum_{i=1}^{m} \Big(\mathrm{KL}^{*}_{r_i|r_{<i},y_{\le i}}\big(\widetilde{R}_i|\widetilde{R}_{<i}, \widetilde{Y}_{\le i} \parallel \widehat{R}_i|\widehat{R}_{<i}, \widehat{Y}_{\le i}\big) + \mathrm{KL}^{*}_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel \widehat{Y}_i|\widehat{R}_{<i}, \widehat{Y}_{<i}\big)\Big) \\
&\quad = \sum_{i=1}^{m} \mathrm{KL}^{*}_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel \widehat{Y}_i|\widehat{R}_{<i}, \widehat{Y}_{<i}\big) = \sum_{i=1}^{m} \mathrm{KL}^{*}_{y_i|r_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i} \parallel \widehat{Y}_i|\widehat{R}_{<i}\big).
\end{aligned}$$

The first equality is because $Y_i = \widehat{Y}_i$ since we are only considering non-failure cases ($r_i \neq \bot$). The second equality is the chain rule. The penultimate equality is by definition of rejection sampling: $\widetilde{R}_i|\widetilde{R}_{<i}, \widetilde{Y}_{\le i}$ and $\widehat{R}_i|\widehat{R}_{<i}, \widehat{Y}_{\le i}$ are identical on $\mathrm{Supp}(\widetilde{R}_i)$ since conditioning on $\widehat{Y}_i = y$ implies that only non-failure cases ($r_i \neq \bot$) are considered. The last equality is because $\widetilde{Y}_{<i}$ (resp. $\widehat{Y}_{<i}$) is a deterministic function of $\widetilde{R}_{<i}$ (resp. $\widehat{R}_{<i}$).

We now relate $\widehat{Y}_i|\widehat{R}_{<i}$ to $Y_i|Y_{<i}$:

$$\begin{aligned}
\Pr\big[\widehat{Y}_i = y_i|\widehat{R}_{<i} = r_{<i}\big] &= \Pr\big[\widehat{Y}_i = y_i, Y_i = y_i|\widehat{R}_{<i} = r_{<i}\big] && (\widehat{Y}_i = y_i \Leftrightarrow \widehat{Y}_i = y_i \wedge Y_i = y_i) \\
&= \Pr\big[\widehat{Y}_i = y_i|Y_i = y_i, \widehat{R}_{<i} = r_{<i}\big] \cdot \Pr\big[Y_i = y_i|\widehat{R}_{<i} = r_{<i}\big] && \text{(Bayes' Rule)} \\
&= \Pr\big[\widehat{Y}_i = y_i|Y_i = y_i, \widehat{R}_{<i} = r_{<i}\big] \cdot \Pr[Y_i = y_i|Y_{<i} = y_{<i}],
\end{aligned}$$

where the last equality is because when $r \in \mathrm{Supp}(\widetilde{R})$, $\widehat{R}_{<i} = r_{<i} \Rightarrow Y_{<i} = y_{<i}$ and because Y_i is independent of $\widehat{R}_{<i}$ given $Y_{<i}$ (as $\widehat{R}_{<i}$ is simply a randomized function of $Y_{<i}$). The conclusion of the lemma follows by combining the previous two derivations. □

Observe that taking expectations with respect to a uniform $\widetilde{R}$ on both sides in the conclusion of Lemma 3.4.6, we get that next-block hardness in relative entropy is equal to the sum of next-block inaccessible relative entropy and the expectation of the error term coming from the rejection sampling procedure. The following lemma upper bounds this expectation.

Lemma 3.4.7. *Let $\widetilde{G}$ be an online m-block generator, and let $L_i \stackrel{\text{def}}{=} 2^{|\widetilde{G}_i|}$ be the size of the codomain of $\widetilde{G}_i$, $i \in [m]$. Then for all $i \in [m]$, $r_{<i} \in \mathrm{Supp}(\widetilde{R}_{<i})$ and uniform $\widetilde{R}_i$:*

$$\mathop{\mathbb{E}}_{y_i \leftarrow \widetilde{G}_i(r_{<i}, \widetilde{R}_i)} \left[\lg \frac{1}{\Pr\left[\widehat{Y}_i = y_i | Y_i = y_i, \widehat{R}_{<i} = r_{<i}\right]} \right] \leq \lg\left(1 + \frac{L_i - 1}{T}\right) .$$

Proof of Lemma 3.4.7. By definition of $\mathsf{Sim}^{\widetilde{G},T}$, we have:

$$\Pr\left[\widehat{Y}_i = y_i | Y_i = y_i, \widehat{R}_{<i} = r_{<i}\right] = 1 - \left(1 - \Pr\left[\widetilde{G}_i(r_{<i}, \widetilde{R}_i) = y_i\right]\right)^T .$$

Applying Jensen's inequality, we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_i \leftarrow \widetilde{G}_i(r_{<i}, \widetilde{R}_i)} \left[\lg \left(\frac{1}{\Pr\left[\widehat{Y}_i = y_i | Y_i = y_i, \widehat{R}_{<i} = r_{<i}\right]} \right) \right] \\
&\qquad \leq \lg \mathop{\mathbb{E}}_{y_i \leftarrow \widetilde{G}_i(r_{<i}, \widetilde{R}_i)} \left[\frac{1}{\Pr\left[\widehat{Y}_i = y_i | Y_i = y_i, \widehat{R}_{<i} = r_{<i}\right]} \right] \\
&\qquad = \lg \left(\sum_{y \in \mathrm{Im}(\widetilde{G}_i(r_{<i}, \cdot))} \frac{p_y}{1 - (1 - p_y)^T} \right)
\end{aligned}$$

where $p_y = \Pr\left[\widetilde{G}_i(r_{<i}, \widetilde{R}_i) = y\right]$. Since the function $x / \left(1 - (1 - x)^T\right)$ is convex (see Lemma 3.4.8 below), the maximum of the expression inside the logarithm over probability distributions $\{p_y\}$ is achieved at the extremal points of the standard probability simplex. Namely, when all but one $p_y \to 0$ and the other one is 1. Since $\lim_{x \to 0} x / \left(1 - (1 - x)^T\right) = 1/T$:

$$\lg \left(\sum_{y \in \mathrm{Im}(\widetilde{G}_i)} \frac{p_y}{1 - (1 - p_y)^T} \right) \leq \lg\left(1 + (L_i - 1) \cdot \frac{1}{T}\right) . \qquad \square$$

Lemma 3.4.8. *For all $t \geq 1$, $f : x \mapsto \frac{x}{1-(1-x)^t}$ is convex over $[0,1]$.*

Proof. We instead show convexity of $\tilde{f} : x \mapsto f(1-x)$. A straightforward computation gives:

$$\tilde{f}''(x) = \frac{x^{t-2}t\big(t(1-x)(x^t+1) - (1+x)(1-x^t)\big)}{(1-x^t)^3}$$

so that it suffices to show the non-negativity of $g(x) = t(1-x)(x^t+1) - (1+x)(1-x^t)$ over $[0,1]$. The function g has second derivative $t(1-x)(t^2-1)x^{t-2}$, which is non-negative when $x \in [0,1]$, and thus the first derivative g' is non-decreasing. Also, the first derivative at 1 is equal to zero, so that g' is non-positive over $[0,1]$ and hence g is non-increasing over this interval. Since $g(1) = 0$, this implies that g is non-negative over $[0,1]$ and f is convex as desired. □

By combining Lemmas 3.4.6 and 3.4.7, we are now ready to state the main result of this section, relating witness hardness in relative entropy to next-block inaccessible relative entropy.

Theorem 3.4.9. *Let Π be a binary relation and let (Y, W) be a pair of random variables supported on Π. Let $Y = (Y_1, \ldots, Y_m)$ be a partition of Y into blocks of at most ℓ bits. Then we have:*

1. *if (Π, Y, W) has witness hardness (t, Δ) in relative entropy, then for every $0 < \Delta' \leq \Delta$, $(Y_1, \ldots, Y_m, W)$ has next-block inaccessible relative entropy $(t', \Delta - \Delta')$ where $t' = \Omega(t\Delta'/(m^2 2^\ell))$.*

2. *if (Π, Y, W) has witness hardness (t, Δ) in δ-min relative entropy then for every $0 < \Delta' \leq \Delta$ and $0 \leq \delta' \leq 1 - \delta$, we have that $(Y_1, \ldots, Y_m, W)$ has next-block inaccessible $(\delta + \delta')$-min relative entropy $(t', \Delta - \Delta')$ where $t' = \Omega(t\delta'\Delta'/(m^2 2^\ell))$.*

Proof. We consider an online generator $\widetilde{G}$ supported on $(Y_1, \ldots, Y_m, W)$ and the simulator $\mathsf{Sim}^{\widetilde{G},T}$. For convenience, we sometimes write Y_{m+1} for W. Define $\widetilde{R} \stackrel{\text{def}}{=} \widetilde{R}_{\leq m}$ where $\widetilde{R}_{\leq m}$ is a sequence of independent and uniformly random variables, $\widetilde{Y}_{\leq m+1} \stackrel{\text{def}}{=} \widetilde{G}(\widetilde{R})$, $\widetilde{G}_1(\widetilde{R}) \stackrel{\text{def}}{=} \widetilde{Y}_{\leq m}$ and $\widetilde{G}_{\mathsf{w}}(\widetilde{R}) \stackrel{\text{def}}{=} \widetilde{Y}_{m+1}$. We also write for $1 \leq i \leq m$, $\widehat{R}_i \stackrel{\text{def}}{=} \mathsf{Sim}^{\widetilde{G},T}(\widehat{R}_{<i}, Y_i)$, $\widehat{Y}_i \stackrel{\text{def}}{=} \widetilde{G}(\widehat{R}_{\leq i})_i$. Finally we define $\mathsf{S}^{\widetilde{G},T}(Y) \stackrel{\text{def}}{=} \widehat{R}_{\leq m}$.

Observe that $(\widetilde{G}_1, \widetilde{G}_{\mathsf{w}})$ is a two-block generator supported on Π, so the pair $(\widetilde{G}, \mathsf{S}^{\widetilde{G},T})$ forms a pair a algorithms as in the definition of witness hardness in relative entropy (Definition 3.3.10). We focus on sample notions first, and consider $r \in \mathrm{Supp}(\widetilde{R})$, $y \in \mathrm{Supp}(\widetilde{Y}_{\leq m})$ and $w \in \mathrm{Supp}(\widetilde{Y}_{m+1})$. First we use the chain rule to isolate the witness block:

$$\begin{aligned}
&\mathrm{KL}^*_{r,y,w}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}), \widetilde{G}_{\mathsf{w}}(\widetilde{R}) \,\big\|\, \mathsf{S}^{\widetilde{G},T}(Y), Y, W\big) \\
&\quad = \mathrm{KL}^*_{w|r,y}\big(\widetilde{G}_{\mathsf{w}}(\widetilde{R})|\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, W|\mathsf{S}^{\widetilde{G},T}(Y), Y\big) + \mathrm{KL}^*_{r,y}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}^{\widetilde{G},T}(Y), Y\big) \\
&\quad = \mathrm{KL}^*_{y_{m+1}|r_{\leq m}, y_{\leq m}}\big(\widetilde{Y}_{m+1}|\widetilde{R}_{\leq m}, \widetilde{Y}_{\leq m} \,\big\|\, Y_{m+1}|R_{\leq m}, Y_{\leq m}\big) + \mathrm{KL}^*_{r,y}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}^{\widetilde{G},T}(Y), Y\big).
\end{aligned}$$

Next, as in Theorem 3.4.3 we apply the chain rule to decompose the second term on the right-hand side and obtain next-block hardness in relative entropy:

$$\mathrm{KL}^*_{r,y}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}) \,\big\|\, \mathsf{S}^{\widetilde{G},T}(Y), Y\big) = \sum_{i=1}^{m} \mathrm{KL}^*_{r_i,y_i|r_{<i},y_{<i}}\big(\widetilde{R}_i, \widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \,\big\|\, \widehat{R}_i, Y_i|\widehat{R}_{<i}, Y_{<i}\big).$$

Finally, we use Lemma 3.4.6 to further decompose the right-hand side term into inaccessible relative entropy and the rejection sampling error:

$$\begin{aligned}
&\sum_{i=1}^{m} \mathrm{KL}^*_{r_i,y_i|r_{<i},y_{<i}}\big(\widetilde{R}_i, \widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \,\big\|\, \widehat{R}_i, Y_i|\widehat{R}_{<i}, Y_{<i}\big) \\
&\quad = \sum_{i=1}^{m} \mathrm{KL}^*_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \,\big\|\, Y_i|R_{<i}, Y_{<i}\big) + \sum_{i=1}^{m} \lg\left(\frac{1}{\Pr\big[\widehat{Y}_i = y_i|Y_i = y_i, \widehat{R}_{<i} = r_{<i}\big]}\right).
\end{aligned}$$

Combining the previous derivations, we obtain:

$$\begin{aligned}
&\mathrm{KL}^*_{r,y,w}\big(\widetilde{R}, \widetilde{G}_1(\widetilde{R}), \widetilde{G}_{\mathsf{w}}(\widetilde{R}) \,\big\|\, \mathsf{S}^{\widetilde{G},T}(Y), Y, W\big) \\
&\quad = \sum_{i=1}^{m+1} \mathrm{KL}^*_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \,\big\|\, Y_i|R_{<i}, Y_{<i}\big) + \sum_{i=1}^{m} \lg\left(\frac{1}{\Pr\big[\widehat{Y}_i = y_i|Y_i = y_i, \widehat{R}_{<i} = r_{<i}\big]}\right).
\end{aligned}$$

Now, the first claim of the theorem follows by taking expectations on both sides and observing that

when $T = m \cdot 2^\ell/(\Delta' \ln 2)$, Lemma 3.4.7 implies that the expected value of the rejection sampling error is smaller than Δ'.

For the second claim, we first establish using Lemma 3.4.7 and Markov's inequality that:

$$\Pr_{\substack{y_{\le m+1} \leftarrow \widetilde{Y}_{\le m+1} \\ r \leftarrow \widetilde{R}}}\left[\sum_{i=1}^{m} \lg\left(\frac{1}{\Pr\left[\widehat{Y}_i = y_i | \widehat{R}_{<i} = r_{<i}, \widehat{Y}_{<i} = y_{<i}\right]}\right) \ge \frac{m \cdot 2^\ell}{T\delta' \ln 2}\right] \le \delta'$$

and we reach a similar conclusion by setting $T = m \cdot 2^\ell/(\delta'\Delta' \ln 2)$. □

Remark 3.4.10. For fixed distribution and generators, in the limit where T grows to infinity, the error term caused by the failure of rejection sampling in time T vanishes. In this case, hardness in relative entropy implies next-block inaccessible relative entropy without any loss in the relative entropy parameters.

3.4.2 Next-block inaccessible relative entropy and inaccessible entropy

We first recall the definition from [HRVW16], slightly adapted to our notations.

Definition 3.4.11 (Inaccessible Entropy)**.** Let $(Y_1, \ldots, Y_{m+1})$ be a joint distribution.[6] We say that $(Y_1, \ldots, Y_{m+1})$ has *inaccessible entropy* (t, Δ) if for all $(m+1)$-block online generators $\widetilde{G}$ running in time t and consistent with $(Y_1, \ldots, Y_{m+1})$:

$$\sum_{i=1}^{m+1} \left(\mathrm{H}(Y_i|Y_{<i}) - \mathrm{H}(\widetilde{Y}_i|\widetilde{R}_{<i})\right) > \Delta \,.$$

where $(\widetilde{Y}_1, \ldots, \widetilde{Y}_{m+1}) = \widetilde{G}(\widetilde{R}_1, \ldots, \widetilde{R}_{m+1})$ for a uniform $\widetilde{R}_{\le m+1}$.

Similarly $(Y_1, \ldots, Y_{m+1})$ has *inaccessible δ-max entropy* (t, Δ) if for all $(m+1)$-block online generators

[6] We write $m+1$ the total number of blocks, since in this section we will think of Y_{m+1} (also written as W) as the witness of a distributional search problem and $(Y_1, \ldots, Y_m)$ are the blocks of the instance as in the previous section.

$\widetilde{G}$ running in time t and consistent with $(Y_1, \ldots, Y_{m+1})$:

$$\Pr_{\substack{r_{\leq m+1} \leftarrow \widetilde{R}_{\leq m+1} \\ y_{\leq m+1} \leftarrow \widetilde{G}(r_{\leq m+1})}} \left[\sum_{i=1}^{m+1} \left(\mathrm{H}^*_{y_i|y_{<i}}(Y_i|Y_{<i}) - \mathrm{H}^*_{y_i|r_{<i}}(\widetilde{Y}_i|\widetilde{R}_{<i}) \right) \leq \Delta \right] < \delta \,.$$

Unfortunately, one unsatisfactory aspect of Definition 3.4.11 is that inaccessible entropy can be negative since the generator $\widetilde{G}$ could have more entropy than $(Y_1, \ldots, Y_{m+1})$: if all the Y_i are independent biased random bits, then a generator $\widetilde{G}$ outputting unbiased random bits will have negative inaccessible entropy. On the other hand, next-block inaccessible relative entropy (Definition 3.4.4) does not suffer from this drawback.

Moreover, in the specific case where $(Y_1, \ldots, Y_{m+1})$ is a flat distribution[7], then no distribution with the same support can have higher entropy and in this case Definitions 3.4.4 and 3.4.11 coincide as stated in the following theorem.

Theorem 3.4.12. *Let $(Y_1, , \ldots, Y_{m+1})$ be a flat distribution and $\widetilde{G}$ be an $(m+1)$-block generator consistent with $Y_{\leq m+1}$. Then for $\widetilde{Y}_{\leq m+1} = \widetilde{G}(\widetilde{R}_{\leq m+1})$ for uniform $\widetilde{R}_{\leq m+1}$:*

1. *For every $y_{\leq m+1}, r_{\leq m+1} \in \mathrm{Supp}(\widetilde{Y}_{\leq m+1}, \widetilde{R}_{\leq m+1})$, it holds that*

$$\sum_{i=1}^{m+1} \left(\mathrm{H}^*_{y_i|y_{<i}}(Y_i|Y_{<i}) - \mathrm{H}^*_{y_i|r_{<i}}(\widetilde{Y}_i|\widetilde{R}_{<i}) \right) = \sum_{i=1}^{m+1} \mathrm{KL}^*_{y_i|r_{<i},y_{<i}}(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i|R_{<i}, Y_{<i})$$

   In particular, $(Y_1, \ldots, Y_{m+1})$ has next-block inaccessible δ-min relative entropy (t, Δ) if and only if it has inaccessible δ-max entropy (t, Δ).

[7] For example, the distribution $(Y_{\leq m}, Y_{m+1}) = (f(U), U)$ for a function f and uniform input U is always a flat distribution even if f itself is not regular.

2. *Furthermore,*

$$\sum_{i=1}^{m+1} \left(\mathrm{H}(Y_i|Y_{<i}) - \mathrm{H}\big(\widetilde{Y}_i|\widetilde{R}_{<i}\big)\right) = \sum_{i=1}^{m+1} \mathrm{KL}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i|R_{<i}, Y_{<i}\big),$$

so in particular, $(Y_1, \ldots, Y_{m+1})$ *has next-block inaccessible relative entropy* (t, Δ) *if and only if it has inaccessible entropy* (t, Δ)*.*

Proof. For the sample notions, the chain rule (Proposition 3.2.6) gives:

$$\sum_{i=1}^{m+1} \mathrm{H}^*_{y_i|y_{<i}}(Y_i|Y_{<i}) = \mathrm{H}^*_y\big(Y_{\le m+1}\big) = \lg|\operatorname{Supp}(Y_{\le m+1})|$$

for all y since Y is flat. Hence:

$$\begin{aligned}\lg|\operatorname{Supp}(Y_{\le m+1})| - \sum_{i=1}^{m+1} \mathrm{H}^*_{y_i|r_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}\big) &= \sum_{i=1}^{m+1} \left(\mathrm{H}^*_{y_i|y_{<i}}(Y_i|Y_{<i}) - \mathrm{H}^*_{y_i|r_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}\big)\right) \\ &= \sum_{i=1}^{m+1} \mathrm{KL}^*_{y_i|r_{<i},y_{<i}}\big(\widetilde{Y}_i|\widetilde{R}_{<i}, \widetilde{Y}_{<i} \parallel Y_i|R_{<i}, Y_{<i}\big)\,,\end{aligned}$$

so the second claim follows by taking the expectation over $(\widetilde{Y}_{\le m+1}, \widetilde{R}_{\le m+1})$ on both sides. □

By chaining the reductions between the different notions of hardness considered in this work (hardness in relative entropy, next-block inaccessible relative entropy and inaccessible entropy), we obtain a more modular proof of the theorem of Haitner, Reingold, Vadhan, and Wee [HRVW16], obtaining inaccessible entropy from any one-way function.

Theorem 3.4.13. *Let* n *be a security parameter,* $f : \{0,1\}^n \to \{0,1\}^n$ *be a* (t, ε)*-one-way function, and* X *be uniform over* $\{0,1\}^n$*. For* $\ell \in \{1, \ldots, n\}$*, decompose* $f(X) \stackrel{\text{def}}{=} (Y_1, \ldots, Y_{n/\ell})$ *into blocks of length* ℓ*. Then:*

1. *For every* $0 \le \Delta \le \lg(1/\varepsilon)$*,* $(Y_1, \ldots, Y_{n/\ell}, X)$ *has inaccessible entropy* $\big(t', \lg(1/\varepsilon) - \Delta\big)$ *for* $t' = \Omega\left(t \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right)$*.*

2. *For every* $0 < \delta \leq 1$ *and* $0 \leq \Delta \leq \lg(1/\varepsilon) - \lg(2/\delta)$, $(Y_1, \ldots, Y_{n/\ell}, X)$ *has inaccessible* δ*-max entropy* $\left(t', \lg(1/\varepsilon) - \lg(2/\delta) - \Delta\right)$ *for* $t' = \Omega\left(t \cdot \delta \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right)$.

Proof. Since f is (t, ε)-one-way, the distributional search problem $\left(\Pi^f, f(X)\right)$ where $\Pi^f = \{(f(x), x) : x \in \{0,1\}^n\}$ is (t, ε)-hard. Clearly, $(f(X), X)$ is supported on Π^f, so by applying Theorem 3.3.11, we have that $(\Pi^f, f(X), X)$ has witness hardness $(\Omega(t), \lg(1/\varepsilon))$ in relative entropy and $(\Omega(t), \lg(1/\varepsilon) - \lg(2/\delta))$ in $\delta/2$-min relative entropy. Thus, by Theorem 3.4.9 we have that $(Y_1, \ldots, Y_{n/\ell}, X)$ has next-block inaccessible relative entropy $\left(\Omega\left(t \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \lg(1/\varepsilon) - \Delta\right)$ and next-block inaccessible δ-min relative entropy $\left(\Omega\left(t \cdot \delta \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \lg(1/\varepsilon) - \lg(2/\delta) - \Delta\right)$, and we conclude by Theorem 3.4.12. □

Remark 3.4.14. For comparison, the original proof of [HRVW16] shows that for every $0 < \delta \leq 1$, the joint distribution $(Y_1, \ldots, Y_{n/\ell}, X)$ has inaccessible δ-max entropy $\left(t', \lg(1/\varepsilon) - 2\lg(1/\delta) - O(1)\right)$ for $t' = \tilde{\Omega}\left(t \cdot \delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right)$, which in particular for fixed t' has quadratically worse dependence on δ in terms of the achieved inaccessible entropy: $\lg(1/\varepsilon) - 2 \cdot \lg(1/\delta) - O(1)$ rather than our $\lg(1/\varepsilon) - 1 \cdot \lg(1/\delta) - O(1)$.

Corollary 3.4.15 (Theorem 4.2 in [HRVW16])**.** *Let* n *be a security parameter,* $f : \{0,1\}^n \to \{0,1\}^n$ *be a strong one-way function, and* X *be uniform over* $\{0,1\}^n$*. Then for every* $\ell = O(\lg n)$*, the joint distribution* $(f(X)_{1\ldots\ell}, \ldots, f(X)_{n-\ell+1\ldots n}, X)$ *has inaccessible entropy* $\left(n^{\omega(1)}, \omega(\lg n)\right)$ *and inaccessible* $1/n^{\omega(1)}$*-max entropy* $\left(n^{\omega(1)}, \omega(\lg n)\right)$.

Chapter 4

Finite-Sample Concentration of the Multinomial in Relative Entropy

This chapter is based on [Agr20].

4.1 Introduction

A key problem in statistics is to understand the rate of convergence of an empirical distribution of independent samples to the true underlying distribution. Indeed, this convergence is the basis of hypothesis testing and statistical inference in general [Pit79]. For the case of discrete distributions over a finite alphabet, the Neyman–Pearson lemma [NP33] shows that for optimal hypothesis testing it is important to consider the *likelihood-ratio statistic*, or equivalently [HT12], the Kullback–Leibler divergence (relative entropy) from the true distribution to the empirical distribution, as formally defined in Definition 4.1.1:

Definition 4.1.1. Let $X = (X_1, \dots, X_k)$ be distributed according to a multinomial distribution with n

samples and probabilities $P = (p_1, \dots, p_k)$, and define

$$V_{n,k,P} \stackrel{\text{def}}{=} \mathrm{KL}\Big((X_1/n, \dots, X_k/n) \,\Big\|\, (p_1, \dots, p_k)\Big)$$

where

$$\mathrm{KL}\Big((q_1, \dots, q_k) \,\Big\|\, (p_1, \dots, p_k)\Big) \stackrel{\text{def}}{=} \sum_{i=1}^{k} q_i \ln \frac{q_i}{p_i}$$

is the Kullback–Leibler divergence between two probability distributions on a finite set $\{1, \dots, k\}$ (represented as probability mass functions), and ln is the logarithm in the natural base (as are all logarithms and exponentials in this chapter). The likelihood-ratio statistic is $2nV_{n,k,P}$ [HT12].

In this language, the Neyman–Pearson lemma states that the uniformly most powerful hypothesis test for significance α rejects a hypothesis $P = (p_1, \dots, p_k)$ if and only if $V_{n,k,P}$ is at least ε_α, where ε_α is such that $\Pr\big[V_{n,k,P} \geq \varepsilon_\alpha\big] \leq \alpha$. To apply this test in practice an upper bound on ε_α is needed, so to maximize the power of a provably correct finite-sample test we seek upper bounds on $\Pr[V \geq \varepsilon]$ which are meaningful (less than 1) for ε as small as possible. Equivalently, tight control on ε reduces the number of samples needed to obtain a given level of significance, which is of importance in areas as disparate as high-dimensional statistics [Wai19], combinatorial constructions in complexity theory (Chapter 2), learning theory [Can20], and private machine learning [DWCS20].

In this chapter, we focus on tail bounds for $\Pr\big[V_{n,k,P} \geq \varepsilon\big]$ which decay exponentially for small ε, ideally when $\varepsilon \approx \mathbb{E}\big[V_{n,k,P}\big]$. Paninski [Pan03] showed that $\mathbb{E}\big[V_{n,k,P}\big] \leq \ln\Big(1 + \frac{k-1}{n}\Big) \leq \frac{k-1}{n}$, and conversely Jiao et al. [JVHW17] showed that for P the uniform distribution and large enough n that $\mathbb{E}\big[V_{n,k,U_k}\big] \geq \frac{k-1}{n} \cdot \frac{1}{2}$, so in general the smallest ε for which one can expect a meaningful bound is of order $(k-1)/n$. In this chapter, we derive the first tail bound decaying exponentially in ε for ε as small as $(k-1)/n$, whereas existing bounds either require ε to be at least order $(k-1)/n \cdot \ln(n/k)$ when

$k < n$ ([Csi98; MJTNW19]) or work only for the uniform distribution and decay exponentially in ε^2 ([AK01]), which when $\varepsilon < 1$ is significantly weaker than decay in ε^1. Formally, our result is as follows:

Theorem 4.1.2. *Let $V_{n,k,P}$ be as in Definition 4.1.1. Then for all $\varepsilon > \frac{k-1}{n}$, it holds that*

$$\Pr\left[V_{n,k,P} \geq \varepsilon\right] \leq e^{-n\varepsilon} \cdot \left(\frac{e\varepsilon n}{k-1}\right)^{k-1}.$$

Theorem 4.1.2 is in fact an immediate corollary of our main technical result, which is a bound on the moment generating function of $V_{n,k,P}$.

Theorem 4.1.3. *Let $V_{n,k,P}$ be as in Definition 4.1.1. Then for all $0 \leq t < n$ it holds that*

$$\mathbb{E}\left[\exp\left(t \cdot V_{n,k,P}\right)\right] \leq \left(\frac{1}{1-t/n}\right)^{k-1}.$$

Note that this is also the moment generating function of a gamma distribution with shape $k-1$ and rate n. Bounding the moment generating function is a standard technique to obtain concentration bounds (see e.g. [BLM13]), but to the best of our knowledge Theorem 4.1.3 is the first to give a finite bound on $\mathbb{E}\left[\exp\left(s \cdot 2nV_{n,k,P}\right)\right]$ independent of n for any constant $s > 0$. As a consequence, we are able to give the first (to the best of our knowledge) upper bounds on the m'th moments of $2nV_{n,k,P}$ which do not depend on n for all $m > 2$. Using Wilks' theorem [Wil38] on the asymptotic distribution of the likelihood-ratio statistic, we are then able to compute the asymptotic moments of $2nV_{n,k,P}$ for fixed k and P as n goes to infinity. Furthermore, our finite sample bounds on the m'th non-central moment are within constant factors (with the constant depending on m) of the asymptotic value.

The rest of this chapter is organized as follows. In Section 4.2 we prove Theorems 4.1.2 and 4.1.3, with the proof divided into two parts: in Section 4.2.1 we show Theorem 4.1.3 can be derived from bounds for the special case of a binary alphabet ($k = 2$), e.g. a binomial distribution, and in Section 4.2.2 we give a bound for this simpler case. In Section 4.3 we use Theorem 4.1.3 to derive moment bounds

and asymptotic results. Finally, in Section 4.4 we compare our bounds to existing results in the literature and suggest possible directions for future research, and in particular conjecture an improvement to Theorem 4.1.3 which would nearly close the quadratic gap between our finite-sample bound and the bound of Wilks' theorem on the asymptotic distribution of likelihood-ratio statistic (which does not hold in general for finite n).

4.2 Proof of Finite-Sample Bounds

In this section we prove our main technical result, the moment generating function bound of Theorem 4.1.3, and use it to derive our new tail bound Theorem 4.1.2.

4.2.1 Reducing the Multinomial to the Binomial

We first show that the moment generating function of the empirical relative entropy for arbitrary finite alphabets of size k can be bounded in terms of the special case $k = 2$. Formally, this requires the bound to be of a particular form:

Definition 4.2.1. A function $f : [0, 1) \to \mathbb{R}$ is a *sample-independent MGF bound for the binomial KL* if for every positive integer n, real $t \in [0, n)$, and $p \in [0, 1]$ it holds that

$$\mathbb{E}\left[\exp\left(t \cdot V_{n,2,(p,1-p)}\right)\right] \leq f(t/n).$$

Remark 4.2.2. Recalling that $2nV_{n,k,P}$ is the likelihood-ratio statistic, Definition 4.2.1 is equivalent to requiring bounds on the moment generating function $\mathbb{E}\left[\exp\left(s \cdot 2nV_{n,2,(p,1-p)}\right)\right]$ for $0 \leq s < 1/2$ which do not depend on n or p.

We can now state our reduction.

Proposition 4.2.3. *Let $P = (p_1, \dots, p_k)$ be a distribution on a set of size k for $k \geq 2$. Then for every sample-independent MGF bound for the binomial KL $f : [0,1) \to \mathbb{R}$ and $0 \leq t < n$, the moment generating function of $V_{n,k,P}$ satisfies*

$$\mathbb{E}\big[\exp\big(t \cdot V_{n,k,P}\big)\big] \leq f(t/n)^{k-1}.$$

Proof. This is a simple induction on k. The base case $k = 2$ holds by definition of sample-independent MGF bound for the binomial KL.

For the inductive step, we compute conditioned on the value of X_k. Note that if $p_k = 1$ then the inductive step is trivial since $V_{n,k,P} = 0$ with probability 1, so assume that $p_k < 1$. For each $i \in \{1, \dots, k-1\}$ define $p'_i = p_i/(1-p_k)$, so that conditioned on $X_k = m$, the variables $(X_1, \dots, X_{k-1})$ are distributed multinomially with $n - m$ samples and probabilities $P' = (p'_1, \dots, p'_{k-1})$. Simple rearranging (using the chain rule) implies that

$$\begin{aligned} V_{n,k,P} &= \mathrm{KL}((X_1/n, \dots, X_k/n) \,\|\, (p_1, \dots, p_n)) \\ &= \mathrm{KL}\big((X_k/n, 1 - X_k/n) \,\|\, (p_k, 1-p_k)\big) + \frac{n - X_k}{n} \cdot V_{n-X_k,k-1,P'} \end{aligned} \tag{4.1}$$

where

$$V_{n-X_k,k-1,P'} = \mathrm{KL}\left(\left(\frac{X_1}{n-X_k}, \dots, \frac{X_{k-1}}{n-X_k}\right) \middle\| (p'_1, \dots, p'_{k-1})\right)$$

and where we treat the second term of Eq. (4.1) as 0 if $X_k = n$. Now for every $0 \leq t < n$ we have

$$\begin{aligned} \mathbb{E}\Big[\exp\big(t \cdot V_{n,k,P}\big)\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\exp\big(t \cdot V_{n,k,P}\big) \,\Big|\, X_k\Big]\Big] \\ &= \mathbb{E}\Big[\exp\Big(t \cdot \mathrm{KL}\big((X_k/n, 1 - X_k/n) \,\|\, (p_k, 1-p_k)\big)\Big) \\ &\qquad \cdot \mathbb{E}\Big[\exp\Big(t \cdot \frac{n - X_k}{n} \cdot V_{n-X_k,k-1,P'}\Big) \,\Big|\, X_k\Big]\Big]. \end{aligned}$$

Since $0 \le t \cdot \frac{n-X_k}{n} < n - X_k$, the inductive hypothesis for $V_{n-X_k,k-1,P'}$ implies the upper bound

$$\le \mathbb{E}\left[\exp\Big(t \cdot \mathrm{KL}\big((X_k/n, 1-X_k/n) \,\|\, (p_k, 1-p_k)\big)\Big) \cdot f\left(\frac{t(n-X_k)/n}{n-X_k}\right)^{k-2}\right]$$

$$= f(t/n)^{k-2} \cdot \mathbb{E}\Big[\exp\Big(t \cdot \mathrm{KL}\big((X_k/n, 1-X_k/n) \,\|\, (p_k, 1-p_k)\big)\Big)\Big].$$

By definition of a sample-independent MGF bound for the binomial KL, the second term is at most $f(t/n)$, so we get a bound of $f(t/n)^{k-1}$ as desired. □

Remark 4.2.4. Mardia et al. [MJTNW19] use the same chain rule decomposition of the multinomial KL to inductively bound the (non-exponential) moments.

4.2.2 Bounding the Binomial

It remains to give a sample-independent MGF bound for the binomial KL:

Proposition 4.2.5. *The function*

$$f(x) = \frac{1}{1-x}$$

is a sample-independent MGF bound for the binomial KL.

Remark 4.2.6. Hoeffding's inequality [Hoe63] can be used to give a simple proof of the weaker claim that $2^x/(1-x)$ is a sample-independent MGF bound for the binomial KL.

Proof. Let $B_{n,p}$ denote a random variable with Binomial(n, p) distribution. Using the fact that

$$\exp\Big(n \cdot \mathrm{KL}\big((i/n, 1-i/n) \,\|\, (p, 1-p)\big)\Big) = \frac{\Pr[B_{n,i/n} = i]}{\Pr[B_{n,p} = i]}$$

for any integers $0 \le i \le n$, we can expand the moment generating function as

$$\mathbb{E}\left[\exp\left(nx \cdot \mathrm{KL}\left(\left(\frac{B_{n,p}}{n}, 1-\frac{B_{n,p}}{n}\right) \,\middle\|\, (p, 1-p)\right)\right)\right] = \sum_{i=0}^{n} \Pr[B_{n,p} = i]^{1-x} \Pr[B_{n,i/n} = i]^{x}.$$

For every n and i, the function $q \mapsto \Pr\big[B_{n,q} = i\big] = \binom{n}{i} q^i (1-q)^{n-i}$ is easily seen to be log-concave over $[0,1]$, so we can upper bound the moment generating function by

$$G_n(p,x) \stackrel{\text{def}}{=} \sum_{i=0}^{n} \Pr\big[B_{n,(1-x)p+ix/n} = i\big] = \sum_{i=0}^{n} \binom{n}{i} \big((1-x)p + ix/n\big)^i \big(1 - \big((1-x)p + ix/n\big)\big)^{n-i}$$

It turns out G_n does not depend on p and can be simplified significantly, which we prove in the following two lemmas.

Lemma 4.2.7. *For all non-negative integers n and real numbers x and p we have $G_n(p,x) = G_n(0,x)$.*

Proof. Define $R_n(q,x) = \sum_{i=0}^{n} \binom{n}{i} (q + ix/n)^i (1 - q - ix/n)^{n-i}$ (where when $i = n = 0$ we treat $0/0 = 1$) so that $G_n(p,x) = R_n((1-x)p, x)$ and it suffices to prove that $R_n(q,x) = R_n(0,x)$. We prove this by induction on n: the base case of $n = 0$ holds since $R_n(q,x) = 1$ always, and for the inductive step we have

$$\begin{aligned}
&\frac{\partial}{\partial q} R_n(q,x) \\
&\quad = \sum_{i=0}^{n} \binom{n}{i} \frac{\partial}{\partial q} \big((q + ix/n)^i (1 - q - ix/n)^{n-i}\big) \\
&\quad = \sum_{i=0}^{n} \binom{n}{i} \big(i(q + ix/n)^{i-1}(1 - q - ix/n)^{n-i} - (n-i)(q + ix/n)^i (1 - q - ix/n)^{n-i-1}\big) \\
&\quad = n \sum_{i=1}^{n} \binom{n-1}{i-1} \left(q + x/n + \frac{i-1}{n-1} \cdot \frac{x(n-1)}{n}\right)^{i-1} \left(1 - q - x/n - \frac{i-1}{n-1} \cdot \frac{x(n-1)}{n}\right)^{n-1-(i-1)} \\
&\qquad - n \sum_{i=0}^{n-1} \binom{n-1}{i} \left(q + \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{i} \left(1 - q - \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{n-1-i} \\
&\quad = n \sum_{i=0}^{n-1} \binom{n-1}{i} \left(q + x/n + \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{i} \left(1 - q - x/n - \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{n-1-i} \\
&\qquad - n \sum_{i=0}^{n-1} \binom{n-1}{i} \left(q + \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{i} \left(1 - q - \frac{i}{n-1} \cdot \frac{x(n-1)}{n}\right)^{n-1-i} \\
&\quad = n\left(R_{n-1}\left(q + \frac{x}{n}, \frac{x(n-1)}{n}\right) - R_{n-1}\left(q, \frac{x(n-1)}{n}\right)\right) \\
&\quad = n\big(R_{n-1}(0, x(n-1)/n) - R_{n-1}(0, x(n-1)/n)\big) = 0
\end{aligned}$$

where the last line is by the inductive hypothesis. □

Lemma 4.2.8. *For all non-negative integers n we have $G_n(p,x) = \sum_{i=0}^{n} \frac{n!}{n^i(n-i)!} \cdot x^i$.*

Proof. By Lemma 4.2.7 we have that $G_n(p,x) = G_n(0,x) = \sum_{i=0}^{n} \left(\frac{ix}{n}\right)^i \left(1 - \frac{ix}{n}\right)^{n-i}$ is a polynomial in x of degree at most n. For any non-negative integer $i \leq n$ we can compute the coefficient of x^i in $G_n(0,x)$ by summing over the power of x contributed by the $(jx/n)^j$ term for each j:

$$\begin{aligned}\sum_{j=0}^{i} \binom{n}{j}\left(\frac{j}{n}\right)^j \cdot \binom{n-j}{i-j}\left(-\frac{j}{n}\right)^{i-j} &= \sum_{j=0}^{i} \frac{n!}{j!(n-j)!} \cdot \frac{(n-j)!}{(i-j)!(n-i)!} \cdot \left(\frac{j}{n}\right)^i (-1)^{i-j} \\ &= \frac{n!}{n^i(n-i)!} \cdot \frac{1}{i!} \sum_{j=0}^{i} \binom{i}{j} j^i (-1)^{i-j}\end{aligned}$$

where $\frac{1}{i!}\sum_{j=0}^{i}\binom{i}{j}j^i(-1)^{i-j}$ is by definition the Stirling number of the second kind $\left\{{i \atop i}\right\}$ and is equal to 1 (see e.g. [GKP94, Chapter 6.1]), so that we can simplify this to

$$\frac{n!}{n^i(n-i)!}$$

as desired. □

Putting together Lemma 4.2.7 and Lemma 4.2.8, we have that the moment generating function is at most $G_n(p,x) = \sum_{i=0}^{n} \frac{n!}{n^i(n-i)!}x^i$, where $\frac{n!}{n^i(n-i)!} = \prod_{j=0}^{i-1}(1-j/n) \leq 1$ and thus for each $x \in [0,1)$ we have $G_n(p,x) \leq \sum_{i=0}^{n} x^i \leq \sum_{i=0}^{\infty} x^i = 1/(1-x)$. □

Together, Propositions 4.2.3 and 4.2.5 imply our moment generating function bound (Theorem 4.1.3), and thus a Chernoff bound implies our tail bound:

Proof of Theorem 4.1.2. By Theorem 4.1.3, we know for every $t \in [0,n)$ that $\mathbb{E}\left[\exp\left(t \cdot V_{n,k,P}\right)\right] \leq \left(\frac{1}{1-t/n}\right)^{k-1}$, so by a Chernoff bound

$$\Pr\left[V_{n,k,P} \geq \varepsilon\right] \leq \inf_{t \in [0,n)} \exp(-t\varepsilon) \cdot \left(\frac{1}{1-t/n}\right)^{k-1}.$$

The result follows by making the optimal choice $t/n = 1 - (k-1)/(\varepsilon n)$ when $\varepsilon > (k-1)/n$. □

4.3 Moment and Asymptotic Bounds

In this section we use Theorem 4.1.3 to give finite-sample and asymptotic bounds on the moments of $V_{n,k,P}$. We will need some basic facts about subexponential random variables, for which we follow the textbook of Vershynin [Ver18].

Lemma 4.3.1 ([Ver18, Definition 2.7.5, Proposition 2.7.1])**.** *There is a universal constant $C > 0$ such that every real-valued random variable X with finite* subexponential norm $\|X\|_{\psi_1} \stackrel{\text{def}}{=} \inf\left\{t > 0 : \mathbb{E}\left[\exp(|X|/t)\right] \le 2\right\}$ *satisfies $\mathbb{E}\left[|X|^m\right]^{1/m} \le Cm\|X\|_{\psi_1}$ for all $m \ge 1$.*

Lemma 4.3.1 allows us to bound the moments of $2nV_{n,k,P}$ uniformly for all n.

Theorem 4.3.2. *For every n, k, and P, it holds that $\left\|2nV_{n,k,P}\right\|_{\psi_1} \le 4(k-1)$ and that $\left\|2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right\|_{\psi_1} \le 8(k-1)$. In particular, there exist universal constants $C_1, C_2 > 0$ such that for all n, k, P and $m \ge 1$*

$$\mathbb{E}\left[\left(2nV_{n,k,P}\right)^m\right] \le (C_1 m(k-1))^m \qquad \mathbb{E}\left[\left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)^m\right] \le (C_2 m(k-1))^m$$

Proof. Theorem 4.1.3 implies for all n, k, and P that

$$\mathbb{E}\left[\exp\left(\frac{1}{4(k-1)} \cdot 2nV_{n,k,P}\right)\right] \le \left(1 - \frac{1}{2(k-1)}\right)^{-(k-1)} \le 2,$$

so by Lemma 4.3.1 we have that $\left\|2nV_{n,k,P}\right\|_{\psi_1} \le 4(k-1)$. By the triangle inequality and convexity of norms, this lets us bound the norm of the centered random variable as $\left\|2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right\|_{\psi_1} \le 2\left\|2nV_{n,k,P}\right\|_{\psi_1} \le 8(k-1)$. □

Our asymptotic results rely on Wilks' theorem [Wil38] on the asymptotic behavior of the likelihood

ratio test, which for fixed k and P implies that the random variable $2nV_{n,k,P}$ converges in distribution to the chi-squared distribution with $k-1$ degrees of freedom as n goes to infinity (see also [CS05, Theorem 4.2]). Though in general convergence in distribution does not imply convergence of moments or of the moment generating function [Bil99], it turns out that the bounds from Theorem 4.3.2 are strong enough for convergence in distribution to imply convergence of the moments.

Theorem 4.3.3. *Let $k \geq 2$ be an integer and $P = (p_1, \ldots, p_k)$ be a probability distribution over a finite alphabet of size k with $p_i \neq 0$ for every $i \in \{1, \ldots, k\}$. Then for every $m \geq 1$ we have*

$$\lim_{n\to\infty} \mathbb{E}\left[(2nV_{n,k,P})^m\right] = \mathbb{E}\left[\left(\chi^2_{k-1}\right)^m\right] = 2^m \frac{\Gamma\left(m + \frac{k-1}{2}\right)}{\Gamma\left(\frac{k-1}{2}\right)}$$

$$\lim_{n\to\infty} \mathbb{E}\left[\left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)^m\right] = \mathbb{E}\left[\left(\chi^2_{k-1} - \mathbb{E}\left[\chi^2_{k-1}\right]\right)^m\right]$$

and for every $s \in [0, 1/2)$ we have

$$\lim_{n\to\infty} \mathbb{E}\left[\exp\left(s \cdot 2nV_{n,k,P}\right)\right] = \mathbb{E}\left[\exp\left(s \cdot \chi^2_{k-1}\right)\right] = (1-2s)^{-(k-1)/2}$$

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\exp\left(s \cdot \left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)\right)\right] &= \mathbb{E}\left[\exp\left(s \cdot \left(\chi^2_{k-1} - \mathbb{E}\left[\chi^2_{k-1}\right]\right)\right)\right] \\ &= e^{-(k-1)s}(1-2s)^{-(k-1)/2}\end{aligned}$$

Remark 4.3.4. [MJTNW19] prove the one-sided lower bound that $\liminf_{n\to\infty} \mathrm{Var}\left(2nV_{n,k,P}\right) \geq \mathrm{Var}\left(\chi^2_{k-1}\right)$, which is a special case of the second equality above.

Proof. Given a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ which convergence in distribution to a random variable X, a sufficient condition for $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ is that $\sup_n \mathbb{E}\left[|X_n|^{1+\alpha}\right] < \infty$ for some $\alpha > 0$ (see e.g. [Bil99]).

Wilks' theorem [Wil38] shows that $2nV_{n,k,P}$ converges in distribution to χ^2_{k-1}, and thus the continuous mapping theorem implies that $\left(2nV_{n,k,P}\right)^m$ converges in distribution to $\left(\chi^2_{k-1}\right)^m$ for every $m \geq 1$.

Theorem 4.3.2 implies $\sup_n \mathbb{E}\left[\left|\left(2nV_{n,k,P}\right)^m\right|^2\right] \leq (Cm(k-1))^{2m} < \infty$, which establishes the first claim. In particular, for $m = 1$ we have $\lim_{n\to\infty} \mathbb{E}\left[2nV_{n,k,P}\right] = \mathbb{E}\left[\chi^2_{k-1}\right]$, so Slutsky's theorem implies that $2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]$ converges in distribution to $\chi^2_{k-1} - \mathbb{E}\left[\chi^2_{k-1}\right]$. Again by the continuous mapping theorem we thus have that $\left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)^m$ converges in distribution to $\left(\chi^2_{k-1} - \mathbb{E}\left[\chi^2_{k-1}\right]\right)^m$, so since Theorem 4.3.2 implies $\sup_n \mathbb{E}\left[\left|\left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)^m\right|^2\right] \leq (Cm(k-1))^{2m} < \infty$, we also get the second claim.

For the moment generating function claims, first note that they are trivial for $s = 0$, as both sides are always 1, and for $s \in (0, 1/2)$ we have $1/2 > 1/4 + s/2 > s$. Now, since the continuous mapping theorem implies $\exp(s \cdot 2nV_{n,k,P})$ converges in distribution to $\exp\left(s \cdot \chi^2_{k-1}\right)$, and Theorem 4.1.3 implies $\sup_n \mathbb{E}\left[\left|\exp\left((1/4 + s/2) \cdot 2nV_{n,k,P}\right)\right|\right] \leq (1/2 - s)^{k-1} < \infty$, we get the third claim. Finally, for the last claim, we again have that $\exp\left(s \cdot \left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)\right)$ converges in distribution to $\exp\left(s \cdot \left(\chi^2_{k-1} - \mathbb{E}\left[\chi^2_{k-1}\right]\right)\right)$ by the continuous mapping theorem, and since $V_{n,k,P} \geq 0$ we have

$$\exp\left((1/4 + s/2) \cdot \left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)\right) \leq \exp\left((1/4 + s/2) \cdot 2nV_{n,k,P}\right)$$

and we conclude as for the third claim. □

4.4 Discussion

In this section we compare our bounds to existing results in the literature and discuss possible directions for future work.

4.4.1 Moment generating function bounds

To the best of our knowledge, this work is the first to explicitly consider the moment generating function of the empirical divergence, and existing tail bounds do not give finite bounds on the quantity

$\sup_n \mathbb{E}\left[\exp\left(x \cdot nV_{n,k,P}\right)\right] = \sup_n \int_0^\infty \Pr\left[nV_{n,k,P} > \frac{\ln t}{x}\right] dt$ for any $k \geq 3$ or constant $x > 0$. Thus, we focus on comparing our finite sample bound (Theorem 4.1.3) to the asymptotic one (Theorem 4.3.3).

In Theorem 4.3.3 we showed for all $x \in [0, 1)$ that $\lim_{n\to\infty} \mathbb{E}\left[\exp\left(x \cdot nV_{n,k,P}\right)\right] = (1-x)^{-(k-1)/2}$, whereas our finite sample bound of Theorem 4.1.3 instead gave the upper bound $\mathbb{E}\left[\exp\left(x \cdot nV_{n,k,P}\right)\right] \leq (1-x)^{-(k-1)}$, which is quadratically worse. This loss arises from our binomial bound from Proposition 4.2.5 of $(1-x)^{-1}$ for the case $k = 2$, where the correct asymptotic bound is $(1-x)^{-1/2}$. Unfortunately, it is *not* the case that this latter asymptotic bound holds for all n, p, and $0 \leq x < 1$: indeed, this is violated even for $(n, p, x) = (2, 1/2, 1/2)$. Nevertheless, we conjecture that Proposition 4.2.5 can be improved to something closer to the asymptotic bound:

Conjecture 4.4.1. *The function*

$$f(x) = \frac{2}{\sqrt{1-x}} - 1$$

is a sample-independent MGF bound for the binomial KL.

Remark 4.4.2. $1/\sqrt{1-x} \leq 2/\sqrt{1-x} - 1 \leq 1/(1-x)$ for all $x \in [0, 1)$.

Conjecture 4.4.1 would follow from the following more natural conjecture, which looks at a single branch of the KL divergence and is supported by numerical evidence:

Conjecture 4.4.3. *Letting*

$$\mathrm{KL}_>(p \parallel q) \stackrel{\text{def}}{=} \begin{cases} 0 & p \leq q \\ \mathrm{KL}\Big((p, 1-p) \,\Big\|\, (q, 1-q)\Big) & p > q \end{cases}$$

it holds for every positive integer n, real $t \in [0, n)$, and $p \in [0, 1]$ that

$$\mathbb{E}\left[\exp\left(t \cdot \mathrm{KL}_>(B/n \parallel p)\right)\right] \leq \frac{1}{\sqrt{1 - t/n}}$$

where $B \sim \text{Binomial}(n, p)$.

Remark 4.4.4. We believe the results (or techniques) of Zubkov and Serov [ZS13] and Harremoës [Har17] strengthening Hoeffding's inequality may be of use in proving these conjectures.

Proof of Conjecture 4.4.1 given Conjecture 4.4.3. We have that

$$\mathrm{KL}\Big((p, 1-p) \,\Big\|\, (q, 1-q)\Big) = \mathrm{KL}_{>}(p \,\|\, q) + \mathrm{KL}_{>}(1-p \,\|\, 1-q)$$

so for every $i \in \{0, 1, \dots, n\}$

$$\exp\Big(t \cdot \mathrm{KL}\Big((i/n, 1-i/n) \,\Big\|\, (p, 1-p)\Big)\Big) = \exp\Big(t \cdot \mathrm{KL}_{>}(i/n \,\|\, p)\Big) \cdot \exp\Big(t \cdot \mathrm{KL}_{>}(1-i/n \,\|\, 1-p)\Big).$$

Letting $x = \exp\Big(t \cdot \mathrm{KL}_{>}(i/n \,\|\, p)\Big)$ and $y = \exp\Big(t \cdot \mathrm{KL}_{>}(1-i/n \,\|\, 1-p)\Big)$, we have that at least one of x and y is equal to 1, so that

$$xy = \big(1 + (x-1)\big)\big(1 + (y-1)\big) = 1 + (x-1) + (y-1) + (x-1)(y-1) = x + y - 1,$$

and thus by taking expectations over $i = B$ for $B \sim \text{Binomial}(n, p)$, we get

$$\mathbb{E}\big[\exp(t \cdot \mathrm{KL}(B/n \,\|\, p))\big] = \mathbb{E}\big[\exp(t \cdot \mathrm{KL}_{>}(B/n \,\|\, p))\big] + \mathbb{E}\big[\exp(t \cdot \mathrm{KL}_{>}(1-B/n \,\|\, 1-p))\big] - 1.$$

We conclude by bounding both terms using Conjecture 4.4.3, since $n - B \sim \text{Binomial}(n, 1-p)$. □

It would also be interesting to improve the bounds in the regime of parameters where k is comparable to or larger than n, since [Pan03] show $\mathbb{E}\big[V_{n,k,P}\big] \le \ln\Big(1 + \frac{k-1}{n}\Big)$ but Theorem 4.1.3 implies only the weaker $\mathbb{E}\big[V_{n,k,P}\big] \le \frac{k-1}{n}$. After the initial dissemination of this work, Guo and Richardson [GR20] extend the technique used in Proposition 4.2.5 to bound the moment generating function of $V_{n,k,P}$ directly for arbitrary $k \ge 2$, thereby strengthening Theorem 4.1.3 and furthermore obtaining the correct

order of growth when $k \gg n$.

Another approach to this problem would be to try to bound the centered moment generating function of $V_{n,k,P}$, that is, the moment generating function of $V_{n,k,P} - \mathbb{E}\left[V_{n,k,P}\right]$. Numerical evidence suggests the following strengthening of Theorem 4.1.3, which asserts that $V_{n,k,P}$, after centering, has moment generating function dominated by that of a centered gamma-distributed random variable with shape $k-1$ and rate n.

Conjecture 4.4.5. *Letting $V_{n,k,P}$ as in Definition 4.1.1, for all $0 \leq t < n$ it holds that*

$$\mathbb{E}\left[\exp\left(t \cdot \left(V_{n,k,P} - \mathbb{E}\left[V_{n,k,P}\right]\right)\right)\right] \leq \left(\frac{e^{-t/n}}{1-t/n}\right)^{k-1}.$$

We remark that Conjecture 4.4.5 would imply concentration of $V_{n,k,P}$ around its expectation as conjectured in [MJTNW19, Conjecture 2].

4.4.2 Moment bounds

The moments of $V_{n,k,P}$ have seen some study in the literature. Most notably, Paninski [Pan03] showed by comparison to the χ^2-statistic that $\mathbb{E}\left[V_{n,k,P}\right] \leq \ln\left(1+\frac{k-1}{n}\right) \leq \frac{k-1}{n}$. In the reverse direction, [JVHW17] showed that if $n \geq 15k$ then for the uniform distribution it holds that $\mathbb{E}\left[V_{n,k,U_k}\right] \geq \frac{k-1}{2n}$, complementing the asymptotic result that $\lim_{n\to\infty} \mathbb{E}\left[nV_{n,k,U_k}\right] = \frac{k-1}{2}$, which follows from Theorem 4.3.3 (and can also be derived from [MJTNW19]). For higher moments, [MJTNW19] showed that $\mathrm{Var}(V_{n,k,P}) \leq Ck/n^2$ for some constant C, and asymptotically that $\liminf_{n\to\infty} \mathrm{Var}(2nV_{n,k,P}) \geq \mathrm{Var}(\chi^2_{k-1}) = 2(k-1)$. To the best of our knowledge, no bounds on the higher moments have appeared in the literature.

In Theorem 4.3.2 we showed for every $m \geq 1$ that $\mathbb{E}\left[\left(2nV_{n,k,P}\right)^m\right] \leq (Cm(k-1))^m$ for some universal constant $C > 0$, and we showed in Theorem 4.3.3 the asymptotic equality $\lim_{n\to\infty} \mathbb{E}\left[\left(2nV_{n,k,P}\right)^m\right]$

$= 2^m \frac{\Gamma\left(m+\frac{k-1}{2}\right)}{\Gamma\left(\frac{k-1}{2}\right)} = (C'm(k-1))^m$ where C' is bounded in a constant range. Thus, our finite-sample bound is asymptotically optimal up to the universal constant C.

However, the situation is different for the central moments $\mathbb{E}\left[\left(2nV_{n,k,P} - \mathbb{E}\left[2nV_{n,k,P}\right]\right)^m\right]$, where we again showed the finite sample bound $(Cm(k-1))^m$, but asymptotically from Theorem 4.3.3 the bound is $(C'm(k-1))^{\lfloor m/2 \rfloor}$ for $m \geq 2$ and some C' in a constant range. For $m = 2$, [MJTNW19] were able to achieve this bound up to constant factors, but it is an intriguing open question to get finite sample central moment bounds with the asymptotically correct power for $m > 2$, with one possible approach being bounding the centered moment generating function as suggested in Conjecture 4.4.5.

4.4.3 Tail bound

To understand our tail bound (Theorem 4.1.2), we compare our result to existing bounds in the literature. Antos and Kontoyiannis [AK01] used McDiarmid's bounded differences inequality [McD89] to give a concentration bound for the empirical entropy, which in the case of the uniform distribution implies the bound

$$\Pr\left[\left|V_{n,k,U_k} - \mathbb{E}\left[V_{n,k,U_k}\right]\right| \geq \varepsilon\right] \leq 2e^{-n\varepsilon^2/(2\ln^2 n)}.$$

This bound has the advantage of providing subgaussian concentration around the expectation, but for the case of small $\varepsilon < 1$ it is preferable to have a bound with linear dependence on ε. Unfortunately, existing tail bounds which decay like $e^{-n\varepsilon}$ are not, in the common regime of parameters where $n \gg k$, meaningful for ε close to $\mathbb{E}\left[V_{n,k,P}\right] \leq (k-1)/n$. For example, the method of types [Csi98] is used to prove the standard bound

$$\Pr\left[V_{n,k,P} > \varepsilon\right] \leq e^{-n\varepsilon} \cdot \binom{n+k-1}{k-1}, \tag{4.2}$$

which is commonly used in proofs of Sanov's theorem (see e.g. [CT06]). However, this bound is meaningful only for $\varepsilon > \frac{1}{n} \cdot \ln \binom{n+k-1}{k-1} \geq \frac{k-1}{n} \cdot \ln\left(1 + \frac{n}{k-1}\right)$, which is off by a factor of order $\ln\left(1 + \frac{n}{k-1}\right)$. A recent bound due to Mardia et al. [MJTNW19] improved on the method of types bound for all settings of k and n, but for $3 \leq k \leq \frac{e^2}{2\pi} \cdot n$ still requires $\varepsilon > \frac{k}{n} \cdot \ln\left(\sqrt{\frac{e^3 n}{2\pi k}}\right) > \frac{k-1}{n} \cdot \ln\left(1 + \frac{n-1}{k}\right)/2$, which again has dependence on $\ln\left(1 + \frac{n-1}{k}\right)$.

Thus, if $k \leq n$, then our bound is meaningful for ε smaller than what is needed for the method of types bound or the bound of [MJTNW19] by a factor of order $\ln(n/k)$, which for k as large as $n^{0.99}$ is still $\ln(n)$, and for k as large as $n/\ln n$ is of order $\ln \ln n$. However, Theorem 4.1.2 has slightly worse dependence on ε than the other bounds, so for example it is better than the method of types bound if and only if

$$\frac{k-1}{n} < \varepsilon < \frac{k-1}{n} \cdot \left(\frac{1}{e} \sqrt[k-1]{\binom{n+k-1}{k-1}}\right). \tag{4.3}$$

In particular, when $n \geq e(k-1)$, our bound is better for ε up to order $\frac{n}{k-1}$ times larger than $\frac{k-1}{n}$. However, we can also see that our bound can be better only when $\sqrt[k-1]{\binom{n+k-1}{k-1}} \geq e$, which asymptotically is equivalent to $k-1 \leq Cn$, where $C \approx 1.84$ is the solution to the equation $(1+C)/C \cdot H(C/(1+C)) = 1$ for H the binary entropy function in nats. From a finite-sample perspective, note that the condition is always satisfied in the standard setting of parameters where $n \geq k$, that is, the number of samples is larger than the size of the alphabet. In this regime, we can also compare to the "interpretable" upper bound of [MJTNW19, Theorem 3], to see that Theorem 4.1.2 is better if

$$\frac{k-1}{n} < \varepsilon < \frac{k-1}{n} \cdot \frac{1}{e}\left(\frac{6e^2}{\pi^{3/2}} \sqrt{\frac{e^3 n}{2\pi k}}^{k}\right)^{1/(k-1)},$$

so that in particular our bound is better for ε up to order $\sqrt{\frac{n}{k}}^{1+1/(k-1)} \geq \sqrt{\frac{n}{k}}$ times larger than $\frac{k-1}{n}$.

Chapter 5

Optimal Bounds between f-Divergences and Integral Probability Metrics

This chapter is based on joint work with Thibaut Horel [AH20].

5.1 Introduction

Quantifying the extent to which two probability distributions differ from one another is central in most, if not all, problems and methods in machine learning and statistics. In a line of research going back at least to the work of Kullback [Kul59], information theoretic measures of dissimilarity between probability distributions have provided a fruitful and unifying perspective on a wide range of statistical procedures. A prototypical example of this perspective is the interpretation of maximum likelihood estimation as minimizing the Kullback–Leibler divergence between the empirical distribution—or the ground truth distribution in the limit of infinitely large sample—and a distribution chosen from a parametric family.

A natural and important generalization of the Kullback–Leibler divergence is provided by the family of ϕ-divergences[1] [Csi63; Csi67a] also known in statistics as Ali–Silvey distances [AS66]. Informally, a ϕ-divergence quantifies the divergence between two distributions μ and ν as an average cost of the likelihood ratio, that is, $D_\phi(\mu \parallel \nu) \stackrel{\text{def}}{=} \int \phi(d\mu/d\nu)\, d\nu$ for a convex cost function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Notable examples of ϕ-divergences include the Hellinger distance, the α-divergences (a convex transformation of the Rényi divergences), and the χ^2-divergence.

Crucial in applications of ϕ-divergences are their so-called *variational representations*. For example, the Donsker–Varadhan representation [DV76, Theorem 5.2] expresses the Kullback–Leibler divergence $\mathrm{KL}(\mu \parallel \nu)$ between probability distributions μ and ν as

$$\mathrm{KL}(\mu \parallel \nu) = \sup_{g \in \mathcal{L}^b} \left\{ \int g\, d\mu - \ln \int e^g\, d\nu \right\}, \tag{5.1}$$

where $\mathcal{L}^b$ is the space of bounded measurable functions. Similar variational representations were for example used by [NWJ08; NWJ10; RRGP12; Bel+18] to construct estimates of ϕ-divergences by restricting the optimization problem in (5.1) to a class of functions $\mathcal{G} \subseteq \mathcal{L}^b$ for which the problem becomes tractable (for example when $\mathcal{G}$ is a RKHS or representable by a given neural network architecture). In recent work, [NCT16; NCMQW17] conceptualized an extension of generative adversarial networks (GANs) in which the problem of minimizing a ϕ-divergence is expressed via variational representations such as (5.1) as a two-player game involving two neural networks, one minimizing over probability distributions μ, the other maximizing over g as in (5.1).

Another important class of distances between probability distributions is given by Integral Proba-

[1] In the rest of this chapter, we use ϕ-divergence instead of f-divergence and reserve the letter f for a generic function.

bility Metrics (IPMs) defined by [Mül97] and taking the form

$$d_{\mathcal{G}}(\mu,\nu) = \sup_{g\in\mathcal{G}}\left\{\left|\int g\,d\mu - \int g\,d\nu\right|\right\}, \tag{5.2}$$

where $\mathcal{G}$ is a class of functions parametrizing the distance. Notable examples include the total variation distance ($\mathcal{G}$ is the class of all functions taking value in $[0,1]$), the Wasserstein metric ($\mathcal{G}$ is a class of Lipschitz functions) and Maximum Mean Discrepancies ($\mathcal{G}$ is the unit ball of a RKHS). Being already expressed as a variational problem, IPMs are amenable to estimation, as was exploited by [SFGSL12; GBRSS12]. MMDs have also been used in lieu of ϕ-divergences to train GANs as was first done by [DRG15].

Rewriting the optimization problem (5.1) as

$$\sup_{g\in\mathcal{L}^b}\left\{\int g\,d\mu - \int g\,d\nu - \ln\int e^{(g-\nu(g))}\,d\nu\right\} \tag{5.3}$$

reveals an important connection between ϕ-divergences and IPMs. Indeed, (5.3) expresses the divergence as the solution to a regularized optimization problem in which one attempts to maximize the mean deviation $\int g\,d\mu - \int g\,d\nu$, as in (5.2), while also penalizing functions g which are too "complex" as measured by the centered log moment-generating function of g. In this chapter, we further explore the connection between ϕ-divergences and IPMs, guided by the following question:

what is the best lower bound of a given ϕ-divergence

as a function of a given integral probability metric?

Some specific instances of this question are already well understood. For example, the best lower bound of the Kullback–Leibler divergence by a quadratic function of the total variation distance is known as Pinsker's inequality. More generally, describing the best lower bound of a ϕ-divergence as a function of the total variation distance (without being restricted to being a quadratic), is known as Vajda's problem,

to which an answer was given by [FHT03] for the Kullback–Leibler divergence and by [Gil06] for an arbitrary ϕ-divergence.

Beyond the total variation distance—in particular, when the class $\mathcal{G}$ in (5.2) contains unbounded functions—few results are known. Using (5.3), [BLM13, §4.10] shows that Pinsker's inequality holds as long as the log moment-generating function grows at most quadratically. Since this is the case for bounded functions (via Hoeffding's lemma), this recovers Pinsker's inequality and extends it to the class of so-called *subgaussian* functions. This was recently used by [RZ20] to control bias in adaptive data analysis.

In this chapter, we systematize the convex analytic perspective underlying many of these results, thereby developing the necessary tools to resolve the above guiding question. As an application, we recover in a unified manner the known bounds between ϕ-divergences and IPMs, and extend them along several dimensions. Specifically, starting from the observation of [RRGP12] that the variational representation of ϕ-divergences commonly used in the literature is not "tight" for probability measures (in a sense which will be made formal in the chapter), we make the following contributions:

- we derive a tight representation of ϕ-divergences for probability measures, exactly generalizing the Donsker–Varadhan representation of the Kullback–Leibler divergence.
- we define a generalization of the log moment-generating function and show that it exactly characterizes the best lower bound of a ϕ-divergence by a given IPM. As an application, we show that this function grows quadratically if and only if the ϕ-divergence can be lower bounded by a quadratic function of the given IPM and recover in a unified manner the extension of Pinsker's inequality to subgaussian functions and the Hammersley–Chapman–Robbins bound.
- we characterize the existence of *any* non-trivial lower bound on an IPM in terms of the general-

ized log moment-generating function, and give implications for topological properties of the divergence, for example regarding compactness of sets of measures with bounded ϕ-divergence and the relationship between convergence in ϕ-divergence and weak convergence.

- the answer to Vajda's problem for bounded functions is re-derived in a principled manner, providing a new geometric interpretation on the optimal lower bound of the ϕ-divergence by the total variation distance. From this, we derive a refinement of Hoeffding's lemma and generalizations of Pinsker's inequality to a large class of ϕ-divergences.

The rest of this chapter is organized as follows: Section 5.2 discusses related work, Section 5.3 gives a brief overview of concepts and tools used in this chapter, Section 5.4 derives the tight variational representation of the ϕ-divergence, Section 5.5 focuses on the case of an IPM given by a single function g with respect to a reference measure ν, deriving the optimal bound in this case and discussing topological applications, and Section 5.6 extends this to arbitrary IPMs and sets of measures, with applications to subgaussian functions and Vajda's problem.

5.2 Related work

The question studied in this chapter is an instance of the broader problem of the constrained minimization of a ϕ-divergence, which has been extensively studied in works spanning information theory, statistics and convex analysis.

Kullback–Leibler divergence. The problem of minimizing the Kullback–Leibler divergence [KL51] subject to a convex constraint can be traced back at least to [San57] in the context of large deviation theory and to [Kul59] for the purpose of formulating an information theoretic approach to statistics.

In information theory, this problem is known as an I-projection [Csi75; CM03]. The case where the convex set is defined by finitely many affine equality constraints, which is closest to our work in this chapter, was specifically studied in [BC77; BC79] via a convex duality approach. This special case is of particular relevance to the field of statistics, since the exponential family arises as the optimizer of this problem.

Convex integral functionals and general ϕ. With the advent of the theory of convex integral functionals, initiated in convex analysis by [Roc66; Roc68], the problem is generalized to arbitrary ϕ-divergences, sometimes referred to as ϕ-entropies, especially when seen as functionals over spaces of functions, and increasingly studied via a systematic application of convex duality [TV93]. In the case of affine constraints, the main technical challenge is to identify constraint qualifications guaranteeing that strong duality holds: [BL91; BL93; BK06] investigate the notion of quasi-relative interior for this purpose, and [Léo01a; Léo01b] consider integrability conditions on the functions defining the affine constraints. A comprehensive account of this case can be found in [CM12]. We also note the work [AS06], which shows a duality between *approximate* divergence minimization—where the affine constraints are only required to hold up to a certain accuracy—and maximum a posteriori estimation in statistics.

At a high level, in this chapter we show in Section 5.6 that one can essentially reduce the problem of minimizing the divergence on probability measures subject to a constraint on an IPM to the problem of minimizing the divergence on finite measures subject to two affine constraints: the first restricting to probability measures, and the second constraining the mean deviation of a single function in the class defining the IPM. For the restriction to probability measures, we prove that constraint qualification always holds, a fact which was not observed in the aforecited works, to the best of our knowledge. For

the second constraint, we show in Section 5.5.3 that by focusing on a single function, we can relate strong duality of the minimization problem to compactness properties of the divergence. In particular, we obtain strong duality under similar assumptions as those considered in [Léo01a], even when the usual interiority conditions for constraint qualification do not hold.

Relationship between ϕ-divergences. A specific case of the minimization question which has seen significant work is when the feasible set is defined by other ϕ-divergences, and most notably is a level set the total variation distance. The best-known result in this line is Pinsker's inequality, first proved in a weaker form in [Пин60] and then strengthened independently in [Kul67; Kem69; Csi67a], which gives the best possible quadratic lower bound on the Kullback–Leibler divergence by the total variation distance. More recently, for ϕ-divergences other than the Kullback–Leibler divergence, [Gil10] identified conditions on ϕ under which quadratic "Pinsker-type" lower bounds can be obtained.

More generally, the problem of finding the best lower bound of the Kullback–Leibler divergence as a (possibly non-quadratic) function of the total variation distance was introduced by Vajda in [Vaj70] and generalized to arbitrary ϕ-divergences in [Vaj72], and is therefore sometimes referred to as *Vajda's problem*. Approximations of the best lower bound were obtained in [BH79; Vaj70] for the Kullback–Leibler divergence and in [Vaj72; Gil08; Gil10] for ϕ-divergences under various assumptions on ϕ. The optimal lower bound was derived in [FHT03] for the Kullback–Leibler divergence and in [Gil06] for any ϕ-divergence. As an example application of Section 5.6, in Section 5.6.3 we rederive the optimal lower bound as well as its quadratic relaxations in a unified manner.

In [RW09; RW11], the authors consider the generalization of Vajda's problem of obtaining a tight lower bound on an arbitrary ϕ-divergence given multiple values of *generalized total variation distances*; their result contains [Gil06] as a special case. Beyond the total variation distance, [HV11] introduced

the general question of studying the *joint range* of values taken by an arbitrary pair of ϕ-divergences, which has its boundary given by the best lower bounds of one divergence as a function of the other. [GSS14] generalize this further and consider the general problem of understanding the joint range of multiple ϕ-divergences, i.e. minimizing a ϕ-divergence subject to a finite number of constraints on other ϕ-divergences. A key conceptual contribution in this line of work is to show that these optimization problems, which are defined over (infinitely dimensional) spaces of measures, can be reduced to finite dimensional optimization problems.

Our work in the present chapter differs from results of this type since we are primarily concerned with IPMs other than the total variation distance, and in particular with those containing unbounded functions. It was shown in [KFG06; KFG07; SGFSL09; SFGSL12] that the class of ϕ-divergences and the class of pseudometrics (including IPMs) intersect *only* at the total variation distance. As such, the problem studied in this chapter cannot be phrased as the one of a joint range between two ϕ-divergences, and to the best of our knowledge cannot be handled by the techniques used in studying the joint range.

Transport inequalities. Starting with the work of Marton [Mar86], transportation inequalities upper bounding the Wasserstein distance by a function of the relative entropy have been instrumental in the study of the concentration of measure phenomenon (see e.g. [GL10] for a survey). These inequalities are related to the question studied in this chapter since the 1-Wasserstein distance is an IPM when the probability space is a Polish space and coincides with the total variation distance when the probability space is discrete and endowed with the discrete metric. In an influential paper, Bobkov and Götze [BG99] proved that upper bounding the 1-Wasserstein distance by a square root of the relative entropy is equivalent to upper bounding the log moment-generating function of all 1-Lipschitz functions by a quadratic function. The extension of Pinsker's inequality in [BLM13, §4.10], which was inspired by

[BG99], is also based on quadratic upper bounds of the log moment-generating function and we in turn follow similar ideas in Sections 5.4.3 and 5.5.1 of this chapter.

5.3 Preliminaries

5.3.1 Measure Theory

Notation. Unless otherwise noted, all the probability measures in this chapter are defined on a common measurable space $(\Omega, \mathcal{A})$, which we assume is non-trivial in the sense that $\{\varnothing, \Omega\} \subsetneq \mathcal{A}$, as otherwise all questions considered in this chapter become trivial. We denote by $\mathcal{M}(\Omega, \mathcal{A})$, $\mathcal{M}^+(\Omega, \mathcal{A})$ and $\mathcal{M}^1(\Omega, \mathcal{A})$ the sets of finite signed measures, finite non-negative measures, and probability measures respectively. $\mathcal{L}^0(\Omega, \mathcal{A})$ denotes the space of all measurable functions from Ω to $\mathbb{R}$, and $\mathcal{L}^b(\Omega, \mathcal{A}) \subseteq \mathcal{L}^0(\Omega, \mathcal{A})$ is the set of all bounded measurable functions. For $\nu \in \mathcal{M}(\Omega, \mathcal{A})$, and $1 \leq p \leq \infty$, $\mathcal{L}^p(\nu, \Omega, \mathcal{A})$ denotes the space of measurable functions with finite p-norm with respect to ν, and $L^p(\nu, \Omega, \mathcal{A})$ denotes the space obtained by taking the quotient with respect to the space of functions which are 0 ν-almost everywhere. Similarly, $L^0(\nu, \Omega, \mathcal{A})$ is the space of all measurable functions Ω to $\mathbb{R}$ up to equality ν-almost everywhere. When there is no ambiguity, we drop the indication $(\Omega, \mathcal{A})$.

For two measures $\mu, \nu \in \mathcal{M}$, $\mu \ll \nu$ (resp. $\mu \perp \nu$) denotes that μ is absolutely continuous (resp. singular) with respect to ν and we define $\mathcal{M}_c(\nu) \stackrel{\text{def}}{=} \{\mu \in \mathcal{M} \mid \mu \ll \nu\}$ and $\mathcal{M}_s(\nu) \stackrel{\text{def}}{=} \{\mu \in \mathcal{M} \mid \mu \perp \nu\}$, so that by the Lebesgue decomposition theorem we have the direct sum $\mathcal{M} = \mathcal{M}_c(\nu) \oplus \mathcal{M}_s(\nu)$. For $\mu \in \mathcal{M}_c(\nu)$, $\frac{d\mu}{d\nu} \in L^1(\nu)$ denotes the Radon–Nikodym derivative of μ with respect to ν. For a signed measure $\nu \in \mathcal{M}$, we write the Hahn–Jordan decomposition $\nu = \nu^+ - \nu^-$ where $\nu^+, \nu^- \in \mathcal{M}^+$, and denote by $|\nu| = \nu^+ + \nu^-$ the total variation measure.

For a measurable function $f \in \mathcal{L}^0$ and measure $\mu \in \mathcal{M}$, $\mu(f) \stackrel{\text{def}}{=} \int f \, d\mu$ denotes the integral

of f with respect to μ, and $\operatorname{ess\,sup}_\nu f$ and $\operatorname{ess\,inf}_\nu f$ denote the ν-essential supremum and infimum respectively.

Finally, for brevity, we define for a subspace $X \subseteq \mathcal{M}$ of finite signed measures the subsets $X^+ \stackrel{\text{def}}{=} X \cap \mathcal{M}^+$ and $X^1 \stackrel{\text{def}}{=} X \cap \mathcal{M}^1$, and for $\nu \in \mathcal{M}$ we also define $X_c(\nu) \stackrel{\text{def}}{=} X \cap \mathcal{M}_c(\nu)$ and $X_s(\nu) \stackrel{\text{def}}{=} X \cap \mathcal{M}_s(\nu)$.

Integral Probability Metrics.

Definition 5.3.1. For a non-empty set of measurable functions $\mathcal{G} \subseteq \mathcal{L}^0$, the *integral probability metric* associated with $\mathcal{G}$ is defined by

$$d_{\mathcal{G}}(\mu, \nu) \stackrel{\text{def}}{=} \sup_{g \in \mathcal{G}} \left\{ \left| \int g \, d\mu - \int g \, d\nu \right| \right\},$$

for all pairs of measures $(\mu, \nu) \in \mathcal{M}^2$ such that all functions in $\mathcal{G}$ are absolutely μ- and ν-integrable. We extend this definition to all pairs of measures $(\mu, \nu) \in \mathcal{M}^2$ by $d_{\mathcal{G}}(\mu, \nu) = +\infty$ in cases where there exists a function in $\mathcal{G}$ which is not μ- or ν- integrable.

Remark 5.3.2. When the class $\mathcal{G}$ is closed under negation, one can drop the absolute value in the definition.

Example 5.3.3. The total variation distance $d_{\mathrm{TV}}(\mu, \nu)$ is obtained when $\mathcal{G}$ is the class of measurable functions taking values in $[0, 1]$, and the related L^1 distance is obtained by taking the class $\mathcal{B}$ of measurable functions taking values in $[-1, 1]$.

Example 5.3.4. Note that the integrals $\int g \, d\mu$ and $\int g \, d\nu$ depend only on the pushforward measures $g_*\mu$ and $g_*\nu$ on $\mathbb{R}$. Equivalently, when μ and ν are the probability distributions of random variables X and Y taking values in Ω, we have that $\int g \, d\mu = \int \mathrm{Id}_{\mathbb{R}} \, dg_*\mu = \mathbb{E}[g(X)]$, the expectation of the random variable $g(X)$, and similarly $\int g \, d\nu = \mathbb{E}[g(Y)]$. The integral probability metric $d_{\mathcal{G}}$ thus defines

the distance between random variables X and Y as the largest difference in expectation achievable by "observing" X and Y through a function from the class $\mathcal{G}$.

5.3.2 Convex analysis

Most of the convex functions considered in this chapter will be defined over spaces of measures or functions. Consequently, we will apply tools from convex analysis in its general formulation for locally convex topological vector spaces. References on this subject include [BCR84] and [Bou87, II. and IV.§1] for the topological background, and [ET99, Part I] and [Zălo2, Chapters 1 & 2] for convex analysis. We now briefly review the main concepts appearing in this chapter.

Definition 5.3.5 (Dual pair)**.** A *dual pair* is a triplet $(X, Y, \langle \cdot\,, \cdot \rangle)$ where X and Y are real vector spaces, and $\langle \cdot\,, \cdot \rangle : X \times Y \to \mathbb{R}$ is a bilinear form satisfying the following properties:

(i) for every $x \in X \setminus \{0\}$, there exists $y \in Y$ such that $\langle x, y \rangle \neq 0$.

(ii) for every $y \in Y \setminus \{0\}$, there exists $x \in X$ such that $\langle x, y \rangle \neq 0$.

We say that the *pairing* $\langle \cdot\,, \cdot \rangle$ *puts* X *and* Y *in (separating) duality*. Furthermore, a topology τ on X is said to be *compatible* with the pairing if it is locally convex and if the topological dual $X^\star$ of X with respect to τ is isomorphic to Y. Topologies on Y compatible with the pairing are defined similarly.

Example 5.3.6. For an arbitrary dual pair $(X, Y, \langle \cdot\,, \cdot \rangle)$, the *weak topology* $\sigma(X, Y)$ induced by Y on X is defined to be the coarsest topology such that for each $y \in Y$, $x \mapsto \langle x, y \rangle$ is a continuous linear form on X. It is a locally convex Hausdorff topology induced by the family of seminorms $p_y : x \mapsto |\langle x, y \rangle|$ for $y \in Y$ and is thereby compatible with the duality between X and Y.

Note that in finite dimension, all Hausdorff vector space topologies coincide with the standard topology.

In the remainder of this section, we fix a dual pair $(X, Y, \langle \cdot\,, \cdot \rangle)$ and endow X and Y with topologies compatible with the pairing. As is customary in convex analysis, convex functions take values in the set of extended reals $\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{-\infty, +\infty\}$ to which the addition over $\mathbb{R}$ is extended using the usual conventions, including $(+\infty) + (-\infty) = +\infty$. In this manner, convex functions can always be extended to be defined on the entirety of their domain by assuming the value $+\infty$ when they are not defined. For a convex function $f : X \to \overline{\mathbb{R}}$, $\operatorname{dom} f \stackrel{\text{def}}{=} \{x \in X \mid f(x) < +\infty\}$ is the *effective domain* of f and $\partial f(x) \stackrel{\text{def}}{=} \{y \in Y \mid \forall x' \in X, f(x') \geq f(x) + \langle x' - x, y \rangle\}$ denotes its subdifferential at $x \in X$.

Definition 5.3.7 (Lower semicontinuity, inf-compactness)**.** The function $f : X \to \overline{\mathbb{R}}$ is *lower semicontinuous* (resp. *inf-compact*) if for every $t \in \mathbb{R}$ the sublevel set $f^{-1}(-\infty, t] \stackrel{\text{def}}{=} \{x \in X \mid f(x) \leq t\}$ is closed (resp. compact).

Lemma 5.3.8. *If $f : X \times C \to \overline{\mathbb{R}}$ is a convex function for C a convex subset of some linear space, then $g : X \to \overline{\mathbb{R}}$ defined as $g(x) \stackrel{\text{def}}{=} \inf_{c \in C} f(x, c)$ is convex. Furthermore, if for some topology on C the function f is inf-compact with respect to the product topology, then g is also inf-compact.*

Definition 5.3.9 (Properness)**.** A convex function $f : X \to \overline{\mathbb{R}}$ is *proper* if $\operatorname{dom} f \neq \varnothing$ and $f(x) > -\infty$ for all $x \in X$.

Definition 5.3.10 (Convex conjugate)**.** The *convex conjugate* (also called Fenchel dual or Fenchel–Legendre transform) of $f : X \to \overline{\mathbb{R}}$ is the function $f^\star : Y \to \overline{\mathbb{R}}$ defined for $y \in Y$ by

$$f^\star(y) \stackrel{\text{def}}{=} \sup_{x \in X} \big\{ \langle x, y \rangle - f(x) \big\}.$$

For a set $C \subseteq X$, $\delta_C : X \to \overline{\mathbb{R}}_{\geq 0}$ denotes the characteristic function of C, that is $\delta_C(x)$ is 0 if $x \in C$ and $+\infty$ elsewhere. The support function of C is $h_C : Y \to \mathbb{R} \cup \{+\infty\}$ defined by $h_C(y) = \sup_{x \in C} \langle x, y \rangle$. If C is closed and convex then (δ_C, h_C) form a pair of convex conjugate functions.

Proposition 5.3.11. *Let $f : X \to \overline{\mathbb{R}}$ be a function. Then:*

1. *$f^\star : Y \to \overline{\mathbb{R}}$ is convex and lower semicontinuous.*
2. *for all $x \in X$ and $y \in Y$, $f(x) + f^\star(y) \geq \langle x, y \rangle$ with equality iff $y \in \partial f(x)$.*
3. *$f^{\star\star} \leq f$ with equality iff f is proper convex lower semicontinuous, $f \equiv +\infty$ or $f \equiv -\infty$.*
4. *if $f \leq g$ for some $g : X \to \overline{\mathbb{R}}$, then $g^\star \geq f^\star$.*

Remark 5.3.12. In Proposition 5.3.11, Item 2 is known as the Fenchel–Young inequality and Item 3 as the Fenchel–Moreau theorem.

In the special case of $X = \mathbb{R} = Y$ and a proper convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, we can be more explicit about the domain of $f^\star$.

Definition 5.3.13. For $f : \mathbb{R} \to \overline{\mathbb{R}}$ a proper convex function, we define for $\ell \in \{-\infty, +\infty\}$ the quantity $f'(\ell) \stackrel{\text{def}}{=} \lim_{x \to \ell} f(x)/x \in \mathbb{R} \cup \{+\infty\}$.

Remark 5.3.14. The limit is always well-defined in $\mathbb{R} \cup \{+\infty\}$ for proper convex functions. The name $f'(\ell)$ is motivated by the fact that when f is differentiable, we have $f'(\ell) = \lim_{x \to \ell} f'(x)$.

Lemma 5.3.15. *If $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a proper convex function, then the domain of $f^\star : \mathbb{R} \to \overline{\mathbb{R}}$ satisfies* $\operatorname{int}(\operatorname{dom} f^\star) = \big(f'(-\infty), f'(+\infty)\big)$.

Fenchel duality theorem is arguably the most fundamental result in convex analysis, and we will use it in this chapter to compute the convex conjugate and minimum of a convex function subject to a linear constraint. The following proposition summarizes the conclusions obtained by instantiating the duality theorem to this specific case.

Proposition 5.3.16. *Let $f : X \to (-\infty, +\infty]$ be a convex function. For $y \in Y$ and $\varepsilon \in \mathbb{R}$, define $f_{y,\varepsilon} : X \to (-\infty, +\infty]$ by*

$$f_{y,\varepsilon}(x) \stackrel{\text{def}}{=} f(x) + \delta_{\{\varepsilon\}}\big(\langle x, y\rangle\big) = \begin{cases} f(x) & \text{if } \langle x, y\rangle = \varepsilon \\ +\infty & \text{otherwise} \end{cases}$$

for all $x \in X$.

1. *Assume that f is lower semicontinuous and define $\langle \operatorname{dom} f, y\rangle \stackrel{\text{def}}{=} \{\langle x, y\rangle \mid x \in \operatorname{dom} f\}$. If $\varepsilon \in \operatorname{int}\big(\langle \operatorname{dom} f, y\rangle\big)$, then $f^\star_{y,\varepsilon}(x^\star) = \inf_{\lambda\in\mathbb{R}} f^\star(x^\star + \lambda y) - \lambda\cdot\varepsilon$ for all $x^\star \in Y$, where the infimum is reached whenever $f^\star_{y,\varepsilon}(x^\star)$ is finite.*

2. *Assume that f is non-negative and satisfies $f(0) = 0$. Define the* marginal value function

$$\mathscr{L}_{y,f}(\varepsilon) \stackrel{\text{def}}{=} \inf_{x\in X} f_{y,\varepsilon}(x) = \inf\{f(x) \mid x \in X \wedge \langle x, y\rangle = \varepsilon\}. \tag{5.4}$$

Then $\mathscr{L}_{y,f}$ is a non-negative convex function satisfying $\mathscr{L}_{y,f}(0) = 0$ and its convex conjugate is given by $\mathscr{L}^\star_{y,f}(t) = f^\star(ty)$. Furthermore, $\mathscr{L}_{y,f}$ is lower semicontinuous at ε, that is $\mathscr{L}_{y,f}(\varepsilon) = \mathscr{L}^{\star\star}_{y,f}(\varepsilon)$, if and only if strong duality holds for problem (5.4), i.e. if and only if

$$\inf\{f(x) \mid x \in X \wedge \langle x, y\rangle = \varepsilon\} = \sup\big\{t\cdot\varepsilon - f^\star(t\cdot y) \mid t \in \mathbb{R}\big\}.$$

Proof. 1. This follows from a direct application of Fenchel's duality theorem (see e.g. [Zălo2, Corollary 2.6.4, Theorem 2.8.1]).

2. Define the *perturbation function* $F : X \times \mathbb{R} \to \overline{\mathbb{R}}$ by $F(x,\varepsilon) \stackrel{\text{def}}{=} f_{y,\varepsilon}(x) = f(x) + \delta_{\{0\}}\big(\langle x, y\rangle - \varepsilon\big)$ so that $\mathscr{L}_{y,f}(\varepsilon) = \inf_{x\in X} F(x,\varepsilon)$. Since F is non-negative, jointly convex over the convex set $X \times \mathbb{R}$ and $F(0,0) = 0$, we get that $\mathscr{L}_{y,f}$ is itself convex, non-negative, and satisfies $\mathscr{L}_{y,f}(0) = 0$.

Furthermore, $F^\star(x^\star, t) = f^\star(x^\star + ty)$ and $\mathscr{L}^\star_{y,f}(t) = F^\star(0,t) = f^\star(ty)$ by e.g. [Zălo2, Theorem 2.6.1, Corollary 2.6.4]. □

Finally, we will use the following result giving a sufficient condition for a convex function to be bounded below. Most such results in convex analysis assume that the function is either lower semicontinuous or bounded above on an open set. In contrast, the following lemma assumes that the function is upper bounded on a closed, convex, bounded set of a Banach space, or more generally on a *cs-compact* subset of a real Hausdorff topological vector space.

Lemma 5.3.17 (cf. [Kön86, Example 1.6(o), Remark 1.9]). *Let C be a cs-compact subset of a real Hausdorff topological vector space. If $f : C \to \mathbb{R}$ is a convex function such that $\sup_{x\in C} f(x) < +\infty$, then $\inf_{x\in C} f(x) > -\infty$. In particular, if $f : C \to \mathbb{R}$ is linear, then $\sup_{x\in C} f(x) < +\infty$ if and only if $\inf_{x\in C} f(x) > -\infty$.*

The notion of cs-compactness (called σ-convexity in [Kön86]) was introduced and defined by Jameson in [Jam72], and Proposition 2 of the same paper states that closed, convex, bounded sets of Banach spaces are cs-compact.

5.3.3 Orlicz spaces

We will use elementary facts from the theory of Orlicz spaces which we now briefly review (see for example [Léo07] for a concise exposition or [RR91] for a more complete reference). A function $\theta : \mathbb{R} \to [0, +\infty]$ is a *Young function* if it is a convex, lower semicontinuous, and even function with $\theta(0) = 0$ and $0 < \theta(s) < +\infty$ for some $s > 0$. Then writing $I_{\theta,\nu} : f \mapsto \int \theta(f)\, d\nu$ for $\nu \in \mathcal{M}$, one defines[2] two spaces associated with θ:

[2] The definition and theory of Orlicz spaces holds more generally for σ-finite measures. The case of finite measures already covers all the applications considered in this chapter whose focus is primarily on probability measures.

- the Orlicz space $L^\theta(\nu) \stackrel{\text{def}}{=} \{f \in L^0(\nu) \mid \exists \alpha > 0, I_{\theta,\nu}(\alpha f) < \infty\}$,
- the Orlicz heart [ES89] $L^\theta_\heartsuit(\nu) \stackrel{\text{def}}{=} \{f \in L^0(\nu) \mid \forall \alpha > 0, I_{\theta,\nu}(\alpha f) < \infty\}$, also known as the Morse–Transue space [MT50],

which are both Banach spaces when equipped with the Luxemburg norm $\|f\|_\theta \stackrel{\text{def}}{=} \inf\{t > 0 \,|\, I_{\theta,\nu}(f/t) \leq 1\}$. Furthermore, $L^\theta_\heartsuit(\nu) \subseteq L^\theta(\nu) \subseteq L^1(\nu)$ and $L^\infty(\nu) \subseteq L^\theta(\nu)$ for all θ, and $L^\infty(\nu) \subseteq L^\theta_\heartsuit(\nu)$ when $\operatorname{dom}\theta = \mathbb{R}$. If $\theta^\star$ is the convex conjugate of θ, we have the following analogue of Hölder's inequality: $\int f_1 f_2 \, d\nu \leq 2\|f_1\|_\theta \|f_2\|_{\theta^\star}$, for all $f_1 \in L^\theta(\nu)$ and $f_2 \in L^{\theta^\star}(\nu)$, implying that $(L^\theta, L^{\theta^\star})$ are in dual pairing. Furthermore, if $\operatorname{dom}\theta = \mathbb{R}$, we have that the dual Banach space $(L^\theta_\heartsuit, \|\cdot\|_\theta)^\star$ is isomorphic to $(L^{\theta^\star}, \|\cdot\|_{\theta^\star})$.

5.4 Variational representations of ϕ-divergences

In the rest of this chapter, we fix a convex and lower semicontinuous function $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that $\phi(1) = 0$. After defining ϕ-divergences in Section 5.4.1, we start with the usual variational representation of the ϕ-divergence in Section 5.4.2, which we then strengthen in the case of probability measures in Section 5.4.3. A reader interested primarily in optimal bounds between ϕ-divergences and IPMs can skip Sections 5.4.2 and 5.4.3 at a first reading.

5.4.1 Convex integral functionals and ϕ-divergences

The notion of a ϕ-divergence is closely related to the one of a convex integral functional that we define first.

Definition 5.4.1 (Integral functional)**.** For $\nu \in \mathcal{M}^+$ and $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ a proper convex function, the convex integral functional associated with f and ν is the function $I_{f,\nu} : L^1(\nu) \to \mathbb{R} \cup \{\infty\}$ defined

for $g \in L^1(\nu)$ by

$$I_{f,\nu}(g) = \int f \circ g \, d\nu \,.$$

The systematic study of convex integral functionals from the perspective of convex analysis was initiated by Rockafellar in [Roc68; Roc71], who considered more generally functionals of the form $g \mapsto \int f(\omega, g(\omega))\, d\nu$ for $g : \Omega \to \mathbb{R}^n$ and $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$ such that $f(\omega, \cdot)$ is convex ν-almost everywhere. A good introduction to the theory of such functionals can be found in [Roc76; RW98]. The specific case of Definition 5.4.1 is known as an *autonomous* integral functional, but we drop this qualifier since it applies to all functionals studied in this chapter.

Definition 5.4.2 (ϕ-divergence)**.** For $\mu \in \mathcal{M}$ and $\nu \in \mathcal{M}^+$, write $\mu = \mu_c + \mu_s$ with $\mu_c \ll \nu$ and $\mu_s \perp \nu$, the Lebesgue decomposition of μ with respect to ν, and $\mu_s = \mu_s^+ - \mu_s^-$ with $\mu_s^+, \mu_s^- \in \mathcal{M}^+$, the Hahn–Jordan decomposition of μ_s. The *ϕ-divergence* of μ with respect to ν is the quantity $\mathrm{D}_\phi(\mu \parallel \nu) \in \mathbb{R} \cup \{\infty\}$ defined by

$$\mathrm{D}_\phi(\mu \parallel \nu) \stackrel{\text{def}}{=} \int \phi\left(\frac{d\mu_c}{d\nu}\right) d\nu + \mu_s^+(\Omega) \cdot \phi'(\infty) - \mu_s^-(\Omega) \cdot \phi'(-\infty)\,,$$

with the convention $0 \cdot (\pm\infty) = 0$.

Remark 5.4.3. An equivalent definition of $\mathrm{D}_\phi(\mu \parallel \nu)$ which does not require decomposing μ is obtained by choosing $\lambda \in \mathcal{M}^+$ dominating both μ and ν (e.g. $\lambda = |\mu| + \nu$) and defining

$$\mathrm{D}_\phi(\mu \parallel \nu) = \int \frac{d\nu}{d\lambda} \cdot \phi\left(\frac{d\mu/d\lambda}{d\nu/d\lambda}\right) d\lambda,$$

with the conventions coming from continuous extension that $0 \cdot \phi(a/0) = a \cdot \phi'(\infty)$ if $a \geq 0$ and $0 \cdot \phi(a/0) = a \cdot \phi'(-\infty)$ if $a \leq 0$ (see Definition 5.3.13). It is easy to check that this definition does not depend on the choice of λ and coincides with Definition 5.4.2.

The notion of ϕ-divergence between probability measures was introduced by Csiszár in [Csi63;

Table 5.1: Common ϕ-divergences (see e.g. [SV16])

Name	ϕ	$\phi'(\infty) < \infty$?	$\phi(0) < \infty$?	Notes
α-divergences	$\frac{x^\alpha - 1}{\alpha(\alpha-1)}$	when $\alpha < 1$	when $\alpha > 0$	$\phi_\alpha^\dagger = \phi_{1-\alpha}$
KL	$x \ln x$	No	Yes	Limit of $\alpha \to 1^-$
reverse KL	$-\ln x$	Yes	No	Limit of $\alpha \to 0^+$
squared Hellinger	$(\sqrt{x} - 1)^2$	Yes	Yes	Scaling of $\alpha = \frac{1}{2}$
χ^2-divergence	$(x-1)^2$	No	Yes	Scaling of $\alpha = 2$
Jeffreys	$(x-1)\ln x$	No	No	KL + reverse KL
χ^α-divergences	$\lvert x-1\rvert^\alpha$	when $\alpha = 1$	Yes	For $\alpha \geq 1$ [Vaj73]
Total variation	$\frac{1}{2}\lvert x-1\rvert$	Yes	Yes	Scaling of χ^1-divergence
Jensen–Shannon	$x \ln x - (1+x)\ln\left(\frac{1+x}{2}\right)$	Yes	Yes	a.k.a. total divergence to the average
Triangular discrimination	$\frac{(x-1)^2}{x+1}$	Yes	Yes	a.k.a. Vincze–Le Cam distance

Csi67a] in information theory and independently by Ali and Silvey [AS66] in statistics. The generalization to finite signed measures is from [CGG99]. Some useful properties of the ϕ-divergence include: it is jointly convex in both its arguments, if $\mu(\Omega) = \nu(\Omega)$ then $D_\phi(\mu \parallel \nu) \geq 0$, with equality if and only if $\mu = \nu$ assuming that ϕ is strictly convex at 1.

Remark 5.4.4. If $\mu \ll \nu$, the definition simplifies to $D_\phi(\mu \parallel \nu) = \nu\left(\phi \circ \frac{d\mu}{d\nu}\right)$. Furthermore, if $\phi'(\pm\infty) = \pm\infty$, then $D_\phi(\mu \parallel \nu) = +\infty$ whenever $\mu \not\ll \nu$. When either $\phi'(+\infty)$ or $\phi'(-\infty)$ is finite, some authors implicitly or explicitly redefine $D_\phi(\mu \parallel \nu)$ to be $+\infty$ whenever $\mu \not\ll \nu$, thus departing from Definition 5.4.2. This effectively defines $D_\phi(\cdot \parallel \nu)$ as the integral functional $I_{\phi,\nu}$ and the rich theory of convex integral functionals can be readily applied. As we will see in this chapter, this change of definition is unnecessary and the difficulties arising from the case $\mu \not\ll \nu$ in Definition 5.4.2 can be addressed by separately treating the component of μ singular with respect to ν.

An important reason to prefer the general definition is the equality $D_\phi(\nu \parallel \mu) = D_{\phi^\dagger}(\mu \parallel \nu)$ where $\phi^\dagger : x \mapsto x\phi(1/x)$ is the Csiszár dual of ϕ, which identifies the *reverse* ϕ-divergence—where the

arguments are swapped—with the divergence associated with $\phi^\dagger$. Consequently, any result obtained for the partial function $\mu \mapsto \mathrm{D}_\phi(\mu \,\|\, \nu)$ can be translated into results for the partial function $\nu \mapsto \mathrm{D}_\phi(\mu \,\|\, \nu)$ by swapping the role of μ and ν and replacing ϕ with $\phi^\dagger$. Note that $(\phi^\dagger)'(\infty) = \lim_{x\to 0+} \phi(x)$ and $(\phi^\dagger)'(-\infty) = \lim_{x\to 0-} \phi(x)$, and for many divergences of interest (including the Kullback–Leibler divergence) at least one of $\phi'(\infty)$ and $\phi(0)$ is finite. See Table 5.1 for some examples.

5.4.2 Variational representations: general measures

In this section, we fix a finite non-negative measure $\nu \in \mathcal{M}^+ \setminus \{0\}$ and study the convex functional $\mathrm{D}_{\phi,\nu} : \mu \mapsto \mathrm{D}_\phi(\mu \,\|\, \nu)$ on a space of finite measures put in dual pairing with a space of functions via $\langle \mu, g \rangle = \mu(g)$ for a measure μ and function g.

We consider two types of dual pairs depending on whether the space of measures is contained in $\mathcal{M}_c(\nu)$. Generic instances of these two types of duals pairs will be denoted by (X, Y) and $(\mathcal{X}, \mathcal{Y})$ respectively, and we assume that the spaces constituting those pairs are endowed with topologies compatible with the pairing (for example the weak topologies as in Example 5.3.6). Furthermore, we require that the spaces considered contain a large enough class of "elementary" measures or functions as defined next.

Assumption 5.4.5. The pairs (X, Y) and $(\mathcal{X}, \mathcal{Y})$ are in separating duality and satisfy:

1. $\left\{\mu \in \mathcal{M}_c(\nu) \,\middle|\, \frac{d\mu}{d\nu} \in L^\infty(\nu)\right\} \subseteq X \subseteq \mathcal{M}_c(\nu)$ and $L^\infty(\nu) \subseteq Y \subseteq L^0(\nu)$.
2. $\left\{\mu \in \mathcal{M}_c(\nu) \,\middle|\, \frac{d\mu}{d\nu} \in L^\infty(\nu)\right\} \subseteq \mathcal{X} \subseteq \mathcal{M}$ and $\mathcal{L}^b(\Omega) \subseteq \mathcal{Y} \subseteq \mathcal{L}^0(\Omega)$. Furthermore for any $A \in \mathcal{A} \setminus \{\emptyset\}$ with $\nu(A) = 0$, there exists $\mu \in \mathcal{X}^+ \setminus \{0\}$ such that $\mu(\Omega \setminus A) = 0$.

Note that in Assumption 5.4.5 1., since $X \subseteq \mathcal{M}_c(\nu)$ cannot distinguish two functions equal on a ν-null set, we require that $Y \subseteq L^0(\nu)$. For Assumption 5.4.5 2., since not all measures are continuous

with respect to ν, we require $Y \subseteq \mathcal{L}^0(\Omega)$ in order for integration to be well-defined. Examples of such dual pairs are $(X, Y) = \big(\mathcal{M}_c(\nu), L^\infty(\nu)\big)$ and $(\mathcal{X}, \mathcal{Y}) = \big(\mathcal{M}, \mathcal{L}^b(\Omega)\big)$. Another example is given by the following definition constructing a dual pair tailored to a specific class of function $\mathcal{G}$, as will be useful when considering IPMs.

Definition 5.4.6. For a (possibly empty) set $\mathcal{G} \subseteq \mathcal{L}^0(\Omega, \mathcal{A})$, define the space $\mathcal{X}_\mathcal{G} \subseteq \mathcal{M}$ as the set of all μ integrating every $g \in \mathcal{G}$ (i.e. such that $|\mu|(|g|) < \infty$), and define the space $\mathcal{Y}_\mathcal{G} \subseteq \mathcal{L}^0(\Omega, \mathcal{A})$ as the set of all h such that h is μ-integrable for every $\mu \in \mathcal{X}_\mathcal{G}$.

Similarly, for $\nu \in \mathcal{M}^+$ and a (possibly empty) set $\mathcal{G} \subseteq L^0(\nu)$, define $X_\mathcal{G}(\nu) \stackrel{\text{def}}{=} \mathcal{X}_\mathcal{G} \cap \mathcal{M}_c(\nu)$, and define the space $Y_\mathcal{G}(\nu) \subseteq L^0(\nu)$ as the set of all h such that h is μ-integrable for every $\mu \in X_\mathcal{G}(\nu)$.

For brevity, if $\mathcal{G} = \{g\}$ is a singleton, we write $\mathcal{X}_g$ and $X_g(\nu)$ for $\mathcal{X}_{\{g\}}$ and $X_{\{g\}}(\nu)$ respectively.

Lemma 5.4.7. *The pairs $(\mathcal{X}_\mathcal{G}, \mathcal{Y}_\mathcal{G})$ and $\big(X_\mathcal{G}(\nu), Y_\mathcal{G}(\nu)\big)$ from Definition 5.4.6 satisfy*

(i) For every $\mu \in \mathcal{X}_\mathcal{G}$ (resp. $\mu \in X_\mathcal{G}(\nu)$), $\left\{\mu' \in \mathcal{M}_c(\mu) \;\middle|\; \frac{d\mu'}{d\mu} \in L^\infty(\mu)\right\}$ is contained in $\mathcal{X}_\mathcal{G}$ (resp. $X_\mathcal{G}(\nu)$). Furthermore, $\mathcal{X}_\mathcal{G}$ contains all Dirac measures.

(ii) $\mathcal{Y}_\mathcal{G}$ contains $\mathcal{G} \cup \mathcal{L}^b(\Omega, \mathcal{A})$, and $Y_\mathcal{G}(\nu)$ contains $\mathcal{G} \cup L^\infty(\nu)$.

(iii) $(\mathcal{X}_\mathcal{G}, \mathcal{Y}_\mathcal{G})$ and $\big(X_\mathcal{G}(\nu), Y_\mathcal{G}(\nu)\big)$ are in dual pairing via $(\mu, h) \mapsto \mu(h)$.

In particular, $\big(X_\mathcal{G}(\nu), Y_\mathcal{G}(\nu)\big)$ satisfies Assumption 5.4.5 1., and $(\mathcal{X}_\mathcal{G}, \mathcal{Y}_\mathcal{G})$ satisfies Assumption 5.4.5 2.

Proof. (i) For $\mu \in \mathcal{X}_\mathcal{G}$ (resp. $\mu \in X_\mathcal{G}(\nu)$) and functions $\frac{d\mu'}{d\mu} \in L^\infty(\mu)$ and $g \in \mathcal{G}$, we have

$$|\mu'|(|g|) = \int \left|\frac{d\mu'}{d\mu}\right| |g| \, d|\mu| \leq \left\|\frac{d\mu'}{d\mu}\right\|_\infty \cdot \; |\mu|(|g|) < \infty \,.$$

$\mathcal{X}_\mathcal{G}$ contains all Dirac measures since for every $g \in \mathcal{G}$ and $\omega \in \Omega$, we have $|\delta_\omega|(|g|) = |g(\omega)| < \infty$.

(ii) All finite signed measures integrate all bounded measurable functions, and the containment of $\mathcal{G}$ is by definition.

(iii) Bilinearity and well-definedness of $(\mu, h) \mapsto \mu(h)$ is by definition, and separation is by (i) and (ii). □

The following proposition gives an explicit formula for the convex conjugate $\mathrm{D}^\star_{\phi,\nu}$ defined for $g \in Y$ by

$$\mathrm{D}^\star_{\phi,\nu}(g) = \sup_{\mu \in X} \{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} \tag{5.5}$$

and states that $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous. Using the identity $\mathrm{D}_{\phi,\nu} = \mathrm{D}^{\star\star}_{\phi,\nu}$ we thus obtain a *variational representation* of $\mathrm{D}_\phi(\mu \parallel \nu)$, expressing it as the solution of an optimization problem over Y. Since $X \subseteq \mathcal{M}_c(\nu)$, $\mathrm{D}_{\phi,\nu}$ coincides with the integral functional $I_{\phi,\nu}$. This lets us exploit the well-known fact that under mild assumptions, $(I_{\phi,\nu}, I_{\phi^\star,\nu})$ form a pair of convex conjugate functionals. This fact was first observed in [LZ56] in the context of Orlicz spaces, and then generalized in [Roc68; Roc71].

Proposition 5.4.8. *Let (X, Y) be a dual pair as in Assumption 5.4.5 1. Then the functional $\mathrm{D}_{\phi,\nu}$ over X has convex conjugate $\mathrm{D}^\star_{\phi,\nu}$ given for all $g \in Y$ by*

$$\mathrm{D}^\star_{\phi,\nu}(g) = I_{\phi^\star,\nu}(g) = \int \phi^\star \circ g \, d\nu.$$

Furthermore $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous, therefore for all $\mu \in X$

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup_{g \in Y} \left\{ \int g \, d\mu - \int \phi^\star \circ g \, d\nu \right\}. \tag{5.6}$$

Proof. Since $\nu \in X$ by assumption, the function $\mathrm{D}_{\phi,\nu}$ is proper and convex over X. The proposition is then immediate consequence of [Roc76, Theorem 3C] after identifying $\mathcal{M}_c(\nu)$ with $L^1(\nu)$ by the Radon–Nikodym theorem and noting that X and Y are decomposable [Roc76, Section 3] by Assumption 5.4.5.

□

Example 5.4.9. Consider the case of the Kullback–Leibler divergence, corresponding to the function $\phi : x \mapsto x \ln x$. A simple computation gives $\phi^\star(x) = e^{x-1}$ and (5.6) yields as a variational representation, for all $\mu \in X$

$$\mathrm{KL}(\mu \parallel \nu) = \sup_{g \in Y} \left\{ \mu(g) - \int e^{g-1} \, d\nu \right\}, \tag{5.7}$$

Note that this representation differs from the Donsker–Varadhan representation (5.1) discussed in the introduction. This discrepancy will be explained in the next section.

The variational representation of the ϕ-divergence in Proposition 5.4.8 is well-known (see e.g. [RRGP12]). If $\phi'(\pm\infty) \neq \pm\infty$, it is also of interest to consider the case of a space $\mathcal{X}$ containing measures μ such that $\mu \not\ll \nu$, which has comparatively been less studied in the literature. The following proposition gives an expression for $\mathrm{D}^\star_{\phi,\nu}$ in this case, which is new to the best of our knowledge.

Proposition 5.4.10. *Let $(\mathcal{X}, \mathcal{Y})$ be a dual pair as in Assumption 5.4.5 2. Then the functional $\mathrm{D}_{\phi,\nu}$ over $\mathcal{X}$ has convex conjugate $\mathrm{D}^\star_{\phi,\nu}$ given for all $g \in \mathcal{Y}$ by*

$$\mathrm{D}^\star_{\phi,\nu}(g) = \begin{cases} I_{\phi^\star,\nu}(g) & \text{if } g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)] \\ +\infty & \text{otherwise} \end{cases}. \tag{5.8}$$

Proof. For $g \in \mathcal{Y}$, let $C(g)$ be the right-hand side of Eq. (5.8), our claimed expression for $\mathrm{D}^\star_{\phi,\nu}(g)$.

First, we show the upper bound $\sup_{\mu \in \mathcal{X}} \{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} \leq C(g)$. We assume that $g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]$, otherwise $C(g) = +\infty$ and there is nothing to prove. For $\mu \in \mathcal{X}$, write $\mu = \mu_c + \mu_s^+ - \mu_s^-$ with $\mu_c \in \mathcal{M}_c(\nu)$ and $\mu_s^+, \mu_s^- \in \mathcal{M}_s^+(\nu)$, so that

$$\mu(g) - \mathrm{D}_{\phi,\nu}(\mu) = \mu_c(g) - I_{\phi,\nu}\left(\frac{d\mu_c}{d\nu}\right) + \mu_s^+(g) - \mu_s^+(\Omega) \cdot \phi'(\infty) - \mu_s^-(g) + \mu_s^-(\Omega) \cdot \phi'(-\infty). \tag{5.9}$$

Observe that $\mu_c(g) - I_{\phi,\nu}\left(\frac{d\mu_c}{d\nu}\right) = \nu\left(\frac{d\mu_c}{d\nu} \cdot g - \phi \circ \frac{d\mu_c}{d\nu}\right) \leq \nu(\phi^\star \circ g) = I_{\phi^\star,\nu}(g)$, by the Fenchel–Young inequality applied to ϕ and monotonicity of the integral. Since $\sup g(\Omega) \leq \phi'(\infty)$, we have $\mu_s^+(g) - \mu_s^+(\Omega) \cdot \phi'(\infty) = \mu_s^+(g - \phi'(\infty)) \leq 0$. Similarly $\mu_s^-(\Omega) \cdot \phi'(-\infty) - \mu_s^-(g) \leq 0$. Using these bounds in (5.9) yields $\mu(g) - \mathrm{D}_{\phi,\nu}(\mu) \leq C(g)$ as desired.

Next, we show that $\sup_{\mu \in \mathcal{X}}\{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} \geq C(g)$. Observe that

$$\sup_{\mu \in \mathcal{X}}\{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} \geq \sup_{\mu \in \mathcal{X}_c(\nu)}\{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} = I_{\phi^\star,\nu}(g)\,, \tag{5.10}$$

where the equality follows from Proposition 5.4.8 applied to $X = \mathcal{X}_c(\nu)$ and $Y = \mathcal{Y}/\sim_\nu$ where $\sim_\nu$ is the equivalence relation of being equal ν-almost everywhere. If $g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]$, then $I_{\phi^\star,\nu}(g) = C(g)$ and (5.10) gives the desired conclusion. If $\sup g(\Omega) > \phi'(\infty)$, let $\alpha \in \mathbb{R}$ such that $\phi'(\infty) < \alpha < \sup g(\Omega)$. Then $A = \{\omega \in \Omega | g(\omega) > \alpha\}$ is a non-empty measurable set. If $\nu(A) > 0$, then $I_{\phi^\star,\nu}(g) = \infty = C(g)$, since $\operatorname{dom} \phi^\star \subseteq [\phi'(-\infty), \phi'(\infty)]$ and (5.10) again gives the desired conclusion. If $\nu(A) = 0$, then by Assumption 5.4.5, there exists $\mu_A \in \mathcal{X}^+ \setminus \{0\}$ such that $\mu_A(\Omega \setminus A) = 0$. But then

$$\begin{aligned}
&\sup_{\mu \in \mathcal{X}}\{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\} \\
&\quad \geq \sup_{c>0}\{(\nu + c\mu_A)(g) - \mathrm{D}_{\phi,\nu}(\nu + c\mu_A)\} = \nu(g) + \sup_{c>0}\{c\mu_A(g) - c\mu_A(\Omega) \cdot \phi'(\infty)\} \\
&\quad \geq \nu(g) + \sup_{c>0}\{c\mu_A(\Omega) \cdot (\alpha - \phi'(\infty))\} = +\infty = C(g)\,,
\end{aligned}$$

where the first equality is because $I_{\phi,\nu}\left(\frac{d\nu}{d\nu}\right) = \phi(1) = 0$ and $\mu_A \in \mathcal{X}_s^+(\nu)$, and the second is because $\mu_A(\Omega) > 0$ and $\alpha > \phi'(\infty)$. The case $\inf g(\Omega) < \phi'(-\infty)$ is analogous. □

Remark 5.4.11. Compared to the expression for $\mathrm{D}^\star_{\phi,\nu}$ obtained in Proposition 5.4.8, the expression in Proposition 5.4.10 explicitly constrains the range of g to be contained in $[\phi'(-\infty), \phi'(\infty)]$. Note however, that there is an implicit constraint on the *essential* range of g in Proposition 5.4.8. Indeed,

unless it is contained in $\overline{\operatorname{dom}\phi^\star} = [\phi'(-\infty), \phi'(\infty)]$, the integral functional $I_{\phi^\star,\nu}(g)$ is infinite. As such, Proposition 5.4.10 simply extends this implicit constraint to the entire range of g to account for the measures $\mu \in \mathcal{X}$ with $\mu \not\ll \nu$.

Finally, we prove that $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous over $\mathcal{X}$, yielding a variational representation of $\mathrm{D}_\phi(\mu \parallel \nu)$, even when $\mu \not\ll \nu$.

Proposition 5.4.12. *Let $(\mathcal{X}, \mathcal{Y})$ be a dual pair as in Assumption 5.4.5 2. Then $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous, equivalently, we have for all $\mu \in \mathcal{X}$ the biconjugate representation*

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup\left\{\mu(g) - I_{\phi^\star,\nu}(g) \,\middle|\, g \in \mathcal{Y} \wedge g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]\right\}.$$

Proof. Since $\mathrm{D}_{\phi,\nu}$ is proper, by the Fenchel–Moreau theorem it suffices to show that $\mathrm{D}^{\star\star}_{\phi,\nu} \geq \mathrm{D}_{\phi,\nu}$. For $\mu \in X$, write $\mu = \mu_c + \mu_s^+ - \mu_s^-$ with $\mu_c \in \mathcal{M}_c(\nu)$, and $\mu_s^+, \mu_s^- \in \mathcal{M}_s^+(\nu)$ by the Lebesgue and Hahn–Jordan decompositions. Furthermore, let $(C, P, N) \in \mathcal{A}^3$ be a partition of Ω such that $|\mu_c|(\Omega \setminus C) = \nu(\Omega \setminus C) = 0, \mu_s^+(\Omega \setminus P) = 0$ and $\mu_s^-(\Omega \setminus N) = 0$. By Proposition 5.4.10,

$$\mathrm{D}^{\star\star}_{\phi,\nu}(\mu) = \sup\left\{\mu_c(g) - I_{\phi^\star,\nu}(g) + \mu_s^+(g) - \mu_s^-(g) \,\middle|\, g \in \mathcal{Y} \wedge g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]\right\}. \quad (5.11)$$

Let $\alpha \in \mathbb{R}$ such that $\alpha < I_{\phi,\nu}\left(\frac{d\mu_c}{d\nu}\right)$. Applying Proposition 5.4.8 with $X = \mathcal{M}_c(\nu)$ and $Y = L^\infty(\nu)$, we get the existence of $g_c \in L^\infty(\nu)$ such that $\mu_c(g_c) - I_{\phi^\star,\nu}(g_c) > \alpha$. Furthermore, since $\operatorname{dom}\phi^\star \subseteq [\phi'(-\infty), \phi'(\infty)]$, we have that $g_c \in [\phi'(-\infty), \phi'(\infty)]$ ν-almost everywhere. Consequently, there exists a representative $\tilde{g}_c \in \mathcal{L}^b(\Omega)$ of g_c such that for every $\omega \in \Omega$, $\phi'(-\infty) \leq \operatorname{ess\,inf}_\nu g_c \leq \tilde{g}_c(\omega) \leq \operatorname{ess\,sup}_\nu g_c \leq \phi'(\infty)$.

For $\beta, \gamma \in \mathbb{R} \cap [\phi'(-\infty), \phi'(\infty)] \supseteq \operatorname{dom}\phi^\star$ (nonempty since ϕ is convex and proper), define

$g : \Omega \to \mathbb{R}$ by

$$g(\omega) = \begin{cases} \tilde{g}_c(\omega) & \text{if } \omega \in C \\ \beta & \text{if } \omega \in P \\ \gamma & \text{if } \omega \in N \end{cases} .$$

By construction $g \in \mathcal{L}^b(\Omega)$ and thus $g \in \mathcal{Y}$ by Assumption 5.4.5. Furthermore, $\mu_c(g) - I_{\phi^\star,\nu}(g) = \mu_c(\tilde{g}_c) - I_{\phi^\star,\nu}(\tilde{g}_c) = \mu_c(g_c) - I_{\phi^\star,\nu}(g_c) > \alpha$, $\mu_s^+(g) = \mu_s^+(\Omega) \cdot \beta$, and $\mu_s^-(g) = \mu_s^-(\Omega) \cdot \gamma$. Since $g(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]$ by construction, for this choice of $g \in Y$, the optimand in (5.11) is at least $\alpha + \mu_s^+(\Omega) \cdot \beta - \mu_s^-(\Omega) \cdot \gamma$, which concludes the proof since α, β, γ can be made arbitrarily close to $I_{\phi,\nu}\left(\frac{d\mu_c}{d\nu}\right)$, $\phi'(\infty)$, and $\phi'(-\infty)$ respectively. □

5.4.3 Variational representations: probability measures

When applied to *probability measures*, which are the main focus of this chapter, the variational representations provided by Propositions 5.4.8 and 5.4.12 are loose. This fact was first explicitly mentioned in [RRGP12], where the authors also suggested that tighter representations could be obtained by specializing the derivation to probability measures.

Specifically, given a dual pair $(\mathcal{X}, \mathcal{Y})$ as in Section 5.4.2, we restrict $D_{\phi,\nu}$ to probability measures by defining $\widetilde{D}_{\phi,\nu} : \mu \mapsto D_{\phi,\nu}(\mu) + \delta_{\mathcal{M}^1}(\mu)$ for $\mu \in \mathcal{X}$. For $g \in \mathcal{Y}$ we get

$$\widetilde{D}^\star_{\phi,\nu}(g) = \sup_{\mu \in \mathcal{X}} \{\mu(g) - \widetilde{D}_{\phi,\nu}(\mu)\} = \sup_{\mu \in \mathcal{X}^1} \{\mu(g) - D_{\phi,\nu}(\mu)\}. \tag{5.12}$$

Observe that compared to (5.5), the supremum is now taken over the smaller set $\mathcal{X}^1 = \mathcal{X} \cap \mathcal{M}^1$, and thus $\widetilde{D}^\star_{\phi,\nu} \leq D^\star_{\phi,\nu}$. When $\widetilde{D}_{\phi,\nu}$ is lower semicontinuous we then get for $\mu \in \mathcal{X}^1$

$$D_\phi(\mu \,\|\, \nu) = \widetilde{D}_{\phi,\nu}(\mu) = \widetilde{D}^{\star\star}_{\phi,\nu}(\mu) = \sup_{g \in \mathcal{Y}} \{\mu(g) - \widetilde{D}^\star_{\phi,\nu}(g)\}. \tag{5.13}$$

This representation should be contrasted with the one obtained in Section 5.4.2,

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup_{g \in \mathcal{Y}} \{\mu(g) - \mathrm{D}^\star_{\phi,\nu}(g)\},$$

which holds for any $\mu \in X$ and in which the optimand is smaller than in (5.13) for all $g \in \mathcal{Y}$ (see also Examples 5.4.16 and 5.4.18 below for an illustration).

In the rest of this section, we carry out the above program by giving an explicit expression for $\widetilde{\mathrm{D}}^\star_{\phi,\nu}$ defined in (5.12) and showing that $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semi-continuous. We will assume in the rest of this chapter that $\operatorname{dom}\phi$ contains a neighborhood of 1, as otherwise the ϕ-divergence on probability measures becomes the discrete divergence $\mathrm{D}_\phi(\mu \parallel \nu) = \delta_{\{\nu\}}(\mu)$ which is only finite when $\mu = \nu$ and for which the questions studied in this work are trivial. We start with the following lemma giving a simpler expression for $\widetilde{\mathrm{D}}_{\phi,\nu}$.

Lemma 5.4.13. *Define $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$ for $x \in \mathbb{R}$. Then for all $\mu \in \mathcal{X}$*

$$\widetilde{\mathrm{D}}_{\phi,\nu}(\mu) = \mathrm{D}_{\phi_+,\nu}(\mu) + \delta_{\{1\}}\big(\mu(\Omega)\big).$$

Proof. Using the same notations as in Definition 5.4.2, and since $\phi'_+(-\infty) = -\infty$, it is easy to see that $\mathrm{D}_{\phi_+,\nu}(\mu)$ equals $+\infty$ whenever $\mu_s^- \neq 0$ or $\nu\big(\{\omega \in \Omega \,|\, \frac{d\mu_c}{d\nu}(\omega) < 0\}\big) \neq 0$ and equals $\mathrm{D}_{\phi,\nu}(\mu)$ otherwise. In other words, $\mathrm{D}_{\phi_+,\nu}(\mu) = \mathrm{D}_{\phi,\nu}(\mu) + \delta_{\mathcal{M}^+}(\mu)$. This concludes the proof since $\delta_{\mathcal{M}^+}(\mu) + \delta_{\{1\}}\big(\mu(\Omega)\big) = \delta_{\mathcal{M}^1}(\mu)$. □

In the expression of $\widetilde{\mathrm{D}}_{\phi,\nu}$ given by Lemma 5.4.13, the non-negativity constraint on μ is "encoded" directly in the definition of ϕ_+ (cf. [BL91]), only leaving the constraint $\mu(\Omega) = 1$ explicit. Since $\mu(\Omega) = \int \mathbf{1}_\Omega \, d\mu$, this is an affine constraint which is well-suited to a convex duality treatment. In particular, we can use Proposition 5.3.16 to compute $\widetilde{\mathrm{D}}^\star_{\phi,\nu}$.

Proposition 5.4.14. *Assume that $\nu \in \mathcal{M}^1$ and define $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$. Then,*

1. *the convex conjugate of $\widetilde{\mathrm{D}}_{\phi,\nu}$ with respect to a dual pair (X, Y) satisfying Assumption 5.4.5 1. is, for all $g \in Y$,*

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}}\left\{\int \phi_+^\star(g+\lambda)\,d\nu - \lambda\right\}. \tag{5.14}$$

2. *the convex conjugate of $\widetilde{\mathrm{D}}_{\phi,\nu}$ with respect to a dual pair $(\mathcal{X}, \mathcal{Y})$ satisfying Assumption 5.4.5 2. is, for all $g \in \mathcal{Y}$,*

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf\left\{\int \phi_+^\star(g+\lambda)\,d\nu - \lambda \;\middle|\; \lambda + \sup g(\Omega) \leq \phi'(\infty)\right\}. \tag{5.15}$$

In (5.14) and (5.15) the infimum is reached whenever it is finite. In particular, this is the case in (5.14) whenever $g \in L^\infty(\nu)$ and in (5.15) whenever $g \in \mathcal{L}^b(\Omega)$.

Proof. The first part of the proof is identical for both statements of Proposition 5.4.14. We use Lemma 5.4.13 and apply Proposition 5.3.16 with $f = \mathrm{D}_{\phi_+,\nu}$, $y = \mathbf{1}_\Omega$ and $\varepsilon = 1$. We need to verify that $1 \in \operatorname{int}(\{\mu(\mathbf{1}_\Omega) \mid \mu \in \operatorname{dom}\mathrm{D}_{\phi_+,\nu}\})$. This is immediate since $(1 \pm \alpha)\nu \in \operatorname{dom}\mathrm{D}_{\phi_+,\nu}$ for sufficiently small α by the assumption that $1 \in \operatorname{int}\operatorname{dom}\phi$.

Thus, by Proposition 5.3.16, for all $g \in Y$ (resp. $g \in \mathcal{Y}$)

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}}\left\{\mathrm{D}^\star_{\phi_+,\nu}(g+\lambda) - \lambda\right\},$$

where the infimum is reached whenever it is finite.

1. Equation (5.14) follows immediately since in this case, $\mathrm{D}^\star_{\phi_+,\nu}(g) = I_{\phi_+^\star,\nu}(g)$ for all $g \in Y$ by Proposition 5.4.8.

2. Similarly, Equation (5.15) follows by using the expression of $\mathrm{D}^\star_{\phi_+,\nu}(g)$ given by Proposition 5.4.10 after observing that $\phi'_+(\infty) = \phi'(\infty)$ and $\phi'_+(-\infty) = -\infty$.

It remains to verify the claims about finiteness of $\widetilde{\mathrm{D}}^{\star}_{\phi,\nu}(g)$. For $g \in L^{\infty}(\nu)$, write $M \stackrel{\text{def}}{=} \operatorname{ess\,sup}_{\nu} g$. Since $\operatorname{int}(\operatorname{dom} \phi_+^{\star}) = \big(-\infty, \phi'(\infty)\big)$, for any $A < \phi'(\infty)$, the choice of $\lambda = A - M$ makes the optimand in (5.14) finite. Similarly the choice $\lambda = A - \sup g(\Omega)$ for $A < \phi'(\infty)$ makes the optimand in (5.15) finite when $g \in \mathcal{L}^b(\Omega)$. □

Remark 5.4.15. Since $\overline{\operatorname{dom} \phi_+^{\star}} = \big(-\infty, \phi'(\infty)\big]$, the optimization variable λ in (5.14) is in fact implicitly constrained to $\lambda + \operatorname{ess\,sup}_{\nu} g \le \phi'(\infty)$. In (5.15) this constraint is extended to the (true) supremum of g to account for singular measures in $\mathcal{X}$ (see also Remark 5.4.11 above).

Example 5.4.16. The effect of the restriction to probability measures is particularly pronounced for the L^1 distance, which is the ϕ-divergence for $\phi(x) = |x - 1|$. In the unrestricted case, a simple calculation shows ϕ has convex conjugate $\phi^{\star}(x) = x$ for $|x| \le 1$ and $\phi^{\star}(x) = +\infty$ when $|x| > 1$, so that the conjugate of the unrestricted divergence $\mathrm{D}^{\star}_{\phi,\nu}(g)$ is $+\infty$ unless $g(\Omega) \subseteq [-1, 1]$. In the case of probability measures, the restriction of ϕ to the non-negative reals $\phi_+(x)$ has conjugate $\phi_+^{\star}(x) = x$ when $|x| \le 1$, $\phi_+^{\star}(x) = +\infty$ when $x > 1$, but $\phi_+^{\star}(x) = -1$ when $x < -1$. Thus, $\mathrm{D}^{\star}_{\phi_+,\nu}(g) < +\infty$ whenever $g(\Omega) \subseteq (-\infty, 1]$. Furthermore, because of the additive λ shift in Eq. (5.14), we have $\widetilde{\mathrm{D}}^{\star}_{\phi,\nu}(g) < +\infty$ whenever $\sup g(\Omega) < +\infty$, in particular whenever $g \in \mathcal{L}^b(\Omega)$.

As a corollary, we obtain a different variational representation of the ϕ-divergence, valid for probability measures and containing as a special case the Donsker–Varadhan representation of the Kullback–Leibler divergence.

Corollary 5.4.17. *Assume that $\nu \in \mathcal{M}^1$ and define $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\ge 0}}(x)$. Then,*

1. *For a dual pair (X, Y) satisfying Assumption 5.4.5 1., $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semicontinuous over X. In particular*

for all probability measures $\mu \in X^1 = X \cap \mathcal{M}^1$

$$D_\phi(\mu \parallel \nu) = \sup_{g \in Y} \{\mu(g) - \inf_{\lambda \in \mathbb{R}} \{I_{\phi_+^\star, \nu}(g + \lambda) - \lambda\}\}.$$

2. *For a dual pair* $(\mathcal{X}, \mathcal{Y})$ *satisfying Assumption* 5.4.5 2., $\widetilde{D}_{\phi,\nu}$ *is lower semicontinuous over* $\mathcal{X}$. *In particular for all probability measures* $\mu \in \mathcal{X}^1 = \mathcal{X} \cap \mathcal{M}^1$

$$D_\phi(\mu \parallel \nu) = \sup_{g \in \mathcal{Y}} \{\mu(g) - \inf\{I_{\phi_+^\star, \nu}(g + \lambda) - \lambda \mid \lambda + \sup g(\Omega) \leq \phi'(\infty)\}\}.$$

Proof. Since $\mathbf{1}_\Omega \in Y$ (resp. $\mathbf{1}_\Omega \in \mathcal{Y}$), the linear form $\mu \mapsto \mu(\mathbf{1}_\Omega)$ is continuous for any topology compatible with the dual pair (X, Y) (resp. $(\mathcal{X}, \mathcal{Y})$). Consequently, the function $\mu \mapsto \delta_{\{1\}}(\mu(\Omega))$ is lower semicontinuous as the composition of the lower semicontinuous function $\delta_{\{1\}}$ with a continuous function. Finally, $D_{\phi_+,\nu}$ is lower semicontinuous by Propositions 5.4.8 and 5.4.12. Hence $\widetilde{D}_{\phi,\nu}$ is lower semicontinuous as the sum of two lower semicontinuous functions, by using the expression in Lemma 5.4.13. The variational representation immediately follows by expressing $\widetilde{D}_{\phi,\nu}$ as its biconjugate.

□

Example 5.4.18. As in Example 5.4.9, we consider the case of the Kullback–Leibler divergence, given by $\phi(x) = \phi_+(x) = x \ln x$. For a dual pair (X, Y) satisfying Assumption 5.4.5 1., since $\phi^\star(x) = e^{x-1}$, Proposition 5.4.14 implies that for $\nu \in \mathcal{M}^1$ and $g \in Y$ we have

$$\widetilde{D}^\star_{\phi,\nu}(g) = \inf_{\lambda \in \mathbb{R}} \int e^{g+\lambda-1} \, d\nu - \lambda = \ln \int e^g \, d\nu \,,$$

where the last equality comes from the optimal choice of $\lambda = -\ln \int e^{g-1} \, d\nu$. Using Corollary 5.4.17 we obtain for all probability measure $\mu \in X^1$

$$\mathrm{KL}(\mu \parallel \nu) = \sup_{g \in Y} \left\{\mu(g) - \ln \int e^g \, d\nu\right\}$$

$$= \sup_{g \in Y} \left\{ \mu(g) - \nu(g) - \ln \int e^{\left(g - \nu(g)\right)} \, d\nu \right\},$$

which is the Donsker–Varadhan representation of the Kullback–Leibler divergence [DV76]. For $\mu \in X^1$, the variational representation obtained in (5.7) can be equivalently written

$$\mathrm{KL}(\mu \parallel \nu) = \sup_{g \in Y} \left\{ 1 + \mu(g) - \int e^g \, d\nu \right\}.$$

Using the inequality $\ln(x) \leq x - 1$ for $x > 0$, we see that the optimand in the previous supremum is smaller than the optimand in the Donsker–Varadhan representation for all $g \in Y$. We thus obtained a "tighter" variational representation by restricting the divergence to probability measures.

Example 5.4.19. Consider the family of divergences $\phi(x) = |x - 1|^\alpha / \alpha$ for $\alpha \geq 1$. A simple computation gives $\phi^\star(y) = y + |y|^\beta / \beta$ where $\beta \geq 1$ is such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. In [JHW17], the authors used the variational representation given by Proposition 5.4.8, that is $\mathrm{D}_\phi(\mu \parallel \nu) = \sup_g \mu(g) - \nu\big(\phi^\star(g)\big)$. However, Corollary 5.4.17 shows that the tight representation uses $\phi_+^\star(y)$ which has the piecewise definition $y + |y|^\beta / \beta$ when $y \geq -1$ and the constant $-1/\alpha$ when $y \leq -1$, and writes $\mathrm{D}_\phi(\mu \parallel \nu) = \sup_g \mu(g) - \inf_\lambda \nu\big(\phi_+^\star(g + \lambda)\big)$. Note that the additive λ shift, in e.g. the case $\alpha = 2$, reduces the second term from the raw second moment $\nu\big(g^2\big)$ to something no larger than the variance $\nu\big((g - \nu(g))^2\big)$, which is potentially much smaller.

5.5 Optimal bounds for a single function and reference measure

As a first step to understand the relationship between a ϕ-divergence and an IPM, we consider the case of a single fixed probability measure $\nu \in \mathcal{M}^1$ and measurable function $g \in \mathcal{L}^0$, and study the optimal lower bound of $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms on the *mean deviation* $\mu(g) - \nu(g)$. We characterize this optimal lower bound and its convex conjugate in Section 5.5.1 and then present implications for topological

question regarding the divergence itself in subsequent sections.

In the remainder of this chapter, since we are interested in probability measures, which are in particular non-negative, we assume without loss of generality that ϕ is infinite on the negative reals, that is $\phi(x) = \phi_+(x) = \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$. As per the discussion in Section 5.4.3 (see in particular Lemma 5.4.13), this does not change the value of the divergence on non-negative measures, that is $\mathrm{D}_\phi(\mu \parallel \nu) = \mathrm{D}_{\phi_+}(\mu \parallel \nu)$ for $\mu \in \mathcal{M}^+$, but yields a tighter variational representation since $\phi_+^\star \leq \phi^\star$.

Furthermore, since for probability measures $\mathrm{D}_\phi(\mu \parallel \nu)$ is invariant to affine shifts of the form $\widetilde{\phi}(x) = \phi(x) + c \cdot (x - 1)$ for $c \in \mathbb{R}$, it will be convenient to assume that $0 \in \partial\phi(1)$ (e.g. $\phi'(1) = 0$), equivalently that ϕ is non-negative and has global minimum at $\phi(1) = 0$. This can always be achieved by an appropriate choice of c and is therefore without loss of generality. As an example, we now write for the Kullback–Leibler divergence $\phi(x) = x \ln x - x + 1$ which is non-negative with $\phi'(1) = 0$, and equivalent to the standard definition $\phi(x) = x \ln x$.

5.5.1 Derivation of the bound

We first define the optimal lower bound function, which comes in two flavors depending on whether the mean deviation or the absolute mean deviation is considered.

Definition 5.5.1. For a probability measure $\nu \in \mathcal{M}^1$, a function $\mathrm{g} \in \mathcal{L}^1(\nu)$, and set of probability measures M integrating g, the *optimal lower bound on* $\mathrm{D}_\phi(\mu \parallel \nu)$ *in terms of the mean deviation* is the function $\mathscr{L}_{\mathrm{g},\nu,M}$ defined for $\varepsilon \in \mathbb{R}$ by:

$$\begin{aligned} \mathscr{L}_{\mathrm{g},\nu,M}(\varepsilon) &\stackrel{\text{def}}{=} \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \,\Big|\, \mu \in M \wedge \mu(\mathrm{g}) - \nu(\mathrm{g}) = \varepsilon\Big\} \\ &= \inf_{\mu \in M}\big\{\mathrm{D}_\phi(\mu \parallel \nu) + \delta_{\{0\}}(\mu(\mathrm{g}) - \nu(\mathrm{g}) - \varepsilon)\big\} \qquad (5.16) \\ \mathscr{L}_{\{\pm\mathrm{g}\},\nu,M}(\varepsilon) &\stackrel{\text{def}}{=} \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \,\Big|\, \mu \in M \wedge |\mu(\mathrm{g}) - \nu(\mathrm{g})| = \varepsilon\Big\} \end{aligned}$$

$$= \min\{\mathscr{L}_{g,\nu,M}(\varepsilon), \mathscr{L}_{g,\nu,M}(-\varepsilon)\} \tag{5.17}$$

where we follow the standard convention that the infimum of the empty set is $+\infty$. For the common case where $M = X_g^1(\nu)$ (see Definition 5.4.6) we drop the M subscript and simply write $\mathscr{L}_{g,\nu}$, and similarly the case $M = \mathcal{X}_g^1$ is abbreviated as $\mathscr{L}_{g,\nu,\perp}$.

Lemma 5.5.2. *For every $\nu \in \mathcal{M}^1$, $g \in \mathcal{L}^1(\nu)$, and convex set M of probability measures integrating g, the function $\mathscr{L}_{g,\nu,M}$ is convex and non-negative. Furthermore, $\mathscr{L}_{g,\nu,M}(0) = 0$ whenever $\nu \in M$, and if $\phi'(\infty) = \infty$ then $\mathscr{L}_{g,\nu,M} = \mathscr{L}_{g,\nu,M\cap\mathcal{M}_c(\nu)}$.*

Proof. Convexity is immediate from Lemma 5.3.8 applied to Eq. (5.16), non-negativity follows from non-negativity of $\mathrm{D}_\phi(\cdot \,\|\, \nu)$, the choice $\mu = \nu$ implies $\mathscr{L}_{g,\nu,M}(0) = 0$ when $\nu \in M$, and if $\phi'(\infty) = \infty$ then $\mathrm{D}_\phi(\mu \,\|\, \nu) = +\infty$ when $\mu \in M \setminus \mathcal{M}_c(\nu)$. □

We compute the convex conjugate of $\mathscr{L}_{g,\nu}$ by applying Fenchel duality to Eq. (5.16).

Proposition 5.5.3. *Let $\nu \in \mathcal{M}^1$. Then for $g \in L^1(\nu)$ and $t \in \mathbb{R}$,*

$$\mathscr{L}^\star_{g,\nu}(t) = \inf\left\{\int \phi^\star(tg + \lambda)\, d\nu - t \cdot \nu(g) - \lambda \;\middle|\; \lambda \in \mathbb{R}\right\}, \tag{5.18}$$

and similarly for $g \in \mathcal{L}^1(\nu)$ and $t \in \mathbb{R}$,

$$\mathscr{L}^\star_{g,\nu,\perp}(t) = \inf\left\{\int \phi^\star(tg + \lambda)\, d\nu - t \cdot \nu(g) - \lambda \;\middle|\; \lambda + \sup\big(t \cdot g(\Omega)\big) \le \phi'(\infty)\right\}. \tag{5.19}$$

Furthermore, we have for $M \in \{X_g^1(\nu), \mathcal{X}_g^1\}$ and $\varepsilon \in \mathbb{R}$ that $\mathscr{L}_{g,\nu,M}(\varepsilon) = \mathscr{L}^{\star\star}_{g,\nu,M}(\varepsilon)$ if and only if strong duality holds in the optimization problem Eq. (5.16).

Remark 5.5.4. Proposition 5.5.3 holds more generally for $\mathscr{L}_{g,\nu,X^1}$ (resp. $\mathscr{L}_{g,\nu,\mathcal{X}^1}$) where (X, Y) (resp. $(\mathcal{X}, \mathcal{Y})$) is a dual pair satisfying Assumption 5.4.5 1. (resp. Assumption 5.4.5 2.) and $g \in Y$ (resp.

$g \in \mathcal{Y}$). This shows in particular that the convex conjugate of $\mathscr{L}_{g,\nu,X^1}$ does not depend on X as long as X is sufficiently large. Thus, in the rest of this section we focus on the largest possible choice, that is $X = X_g(\nu)$ and $X = \mathcal{X}_g$ (see Definition 5.4.6).

Remark 5.5.5. Note that if $\phi'(\infty) = \infty$ then the infimum in Eq. (5.19) is taken over all $\lambda \in \mathbb{R}$ and so Eq. (5.19) and Eq. (5.18) coincide, which is consistent with the fact that, in this case, $\mathrm{D}_{\phi,\nu}$ is infinite on singular measures. More generally, since $\operatorname{dom} \phi^\star \subseteq [-\infty, \phi'(\infty)]$, Eq. (5.18) is equivalent to optimizing over λ such that $\lambda + \operatorname{ess\,sup}_\nu tg \leq \phi'(\infty)$, where in Eq. (5.19) the essential supremum is replaced by the true supremum.

Proof. For $X \in \{X_g(\nu), \mathcal{X}_g\}$, let $\Phi : X \to \overline{\mathbb{R}}$ defined by $\Phi(x) = \widetilde{\mathrm{D}}_{\phi,\nu}(x + \nu)$ so that Φ is convex, lsc, non-negative, and 0 at 0. Furthermore, $\Phi^\star(h) = \widetilde{\mathrm{D}}^\star_{\phi,\nu}(h) - \nu(h)$ for $h \in Y$, and $\mathscr{L}_{g,\nu,X^1}(\varepsilon) = \inf\{\Phi(x) \mid x \in X \wedge \langle x, g\rangle = \varepsilon\}$. The result then follows by Propositions 5.3.16 and 5.4.14. □

Remark 5.5.6. Unlike in Proposition 5.4.14, it is not always true that the interiority constraint qualification conditions hold, and indeed strong duality does not always hold for the optimization problem (5.16). For example, for $\Omega = (-1/2, 1/2)$, ν the Lebesgue measure, g the canonical injection into $\mathbb{R}$, and $\phi : x \mapsto \frac{1}{2}|x - 1|$ corresponding to the total variation distance, we have $\mathscr{L}_{g,\nu}(\pm 1/2) = \mathscr{L}_{g,\nu,\perp}(\pm 1/2) = \infty$ but $\mathscr{L}_{g,\nu,\perp}(x) \leq \mathscr{L}_{g,\nu}(x) \leq 1$ for $|x| < 1/2$. However, as noted in Theorem 5.5.12 below, this generally does not matter since it only affects the boundary of the domain of $\mathscr{L}_{g,\nu}$ or $\mathscr{L}_{g,\nu,\perp}$, which contains at most two points. Furthermore, we will show in Corollary 5.5.35 via a compactness argument that when $\phi'(\infty) = \infty$ and $\operatorname{dom} \mathscr{L}^\star_{g,\nu} = \mathbb{R}$—e.g. when $g \in L^\infty(\nu)$—strong duality holds in (5.16).

We can simplify the expressions in Proposition 5.5.3 by introducing the function $\psi : x \mapsto \phi(x+1)$. We state some useful properties of its conjugate $\psi^\star$ below, which follow immediately from basic results

in convex analysis on $\mathbb{R}$ (recall that at the beginning of Section 5.5 we assumed, without loss of generality, that $0 \in \partial\phi(1)$ and $\operatorname{dom}\phi \subseteq \mathbb{R}_{\geq 0}$, which is necessary for Lemma 5.5.7 to hold).

Lemma 5.5.7. *The function $\psi^\star : x \mapsto \phi^\star(x) - x$ is non-negative, convex, and inf-compact. Furthermore, it satisfies $\psi^\star(0) = 0$, $\psi^\star(x) \leq -x$ when $x \leq 0$, and $\operatorname{int}(\operatorname{dom}\psi^\star) = \big(-\infty, \phi'(\infty)\big)$.*

The right-hand side of Eq. (5.18), expressed in terms of $\psi^\star$, will be central to our theory, so we give it a name in the following definition.

Definition 5.5.8 (Cumulant generating function)**.** For $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$, we define the (ϕ, ν)-*cumulant generating function* $K_{g,\nu} : \mathbb{R} \to \overline{\mathbb{R}}$ for $t \in \mathbb{R}$ by

$$K_{g,\nu}(t) \stackrel{\text{def}}{=} \inf_{\lambda \in \mathbb{R}} \int \psi^\star(tg + \lambda)\, d\nu\,, \tag{5.20}$$

where $\psi : x \mapsto \phi(x+1)$. Similarly, given $g \in \mathcal{L}^0(\Omega)$, the $(\phi, \nu, \perp)$-*cumulant generating function* is given, for $t \in \mathbb{R}$, by

$$K_{g,\nu,\perp}(t) \stackrel{\text{def}}{=} \inf\left\{ \int \psi^\star(tg + \lambda)\, d\nu \;\middle|\; \lambda + \sup\big(t \cdot g(\Omega)\big) \leq \phi'(\infty) \right\}.$$

In particular $K_{g,\nu,\perp} \geq K_{g,\nu}$, and $K_{g,\nu,\perp} = K_{g,\nu}$ if $\phi'(\infty) = \infty$. Note also that $\sup\big(t \cdot g(\Omega)\big)$ is the piecewise-linear function

$$\sup\big(t \cdot g(\Omega)\big) = \begin{cases} t \cdot \sup g(\Omega) & t \geq 0 \\ t \cdot \inf g(\Omega) & t \leq 0 \end{cases}.$$

Example 5.5.9. For the Kullback–Leibler divergence, we see by Example 5.4.18 that $K_{g,\nu}(t) = \ln \nu\big(e^{t(g-\nu(g))}\big)$, which is the standard (centered) cumulant generating function, thereby justifying the name.

Note that the (ϕ, ν)-cumulant generating function $K_{g,\nu}$ depends only on the pushforward measure $g_*\nu$ of ν through g. In particular, when ν is the probability distribution of a random variable X, as in Example 5.3.4, $K_{g,\nu}(t)$ can be equivalently written as

$$K_{g,\nu}(t) = \inf_{\lambda \in \mathbb{R}} \mathbb{E}[\psi^\star(t \cdot g(X) + \lambda)], \tag{5.21}$$

highlighting the fact that $K_{g,\nu}$ only depends on $g(X)$. This contrasts with the $(\phi, \nu, \perp)$-cumulant generating function, for which the constraint on the minimization parameter λ depends on the range $g(\Omega)$, which is not a property of the random variable $g(X)$ since it depends on the value of g on ν-null sets.

Furthermore, since for $t \in \mathbb{R}$, the function $\lambda \mapsto I_{\psi^\star,\nu}(tg + \lambda)$ is convex in λ, the (ϕ, ν)-cumulant generating function is defined by a single-dimensional convex optimization problem whose objective function is expressed as an integral with respect to a probability measure (5.20, 5.21). Hence, the rich spectrum of stochastic approximation methods, such as stochastic gradient descent, can be readily applied, leading to efficient numerical procedures to evaluate $K_{g,\nu}(t)$, as long as the pushforward measure $g_*\nu$ is efficiently samplable.

Remark 5.5.10. Since the mean deviation, and thus the optimal bound $\mathscr{L}_{g,\nu}$ is invariant to shifting g by a constant, we are in fact implicitly working in the quotient space $L^1(\nu)/\mathbb{R}\mathbf{1}_\Omega$. As such, $g \mapsto \inf_{\lambda \in \mathbb{R}} I_{\psi^\star,\nu}(g + \lambda)$ can be interpreted as the integral functional induced by $I_{\psi^\star,\nu}$ on this quotient space, by considering its infimum over all representatives of a given equivalence class. This is analogous to the definition of a norm on a quotient space.

The following proposition states some basic properties of the cumulant generating function.

Proposition 5.5.11. *For every $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$, $K_{g,\nu} : \mathbb{R} \to \overline{\mathbb{R}}$ is non-negative, convex, lower semicontinuous, and satisfies $K_{g,\nu}(0) = 0$. Furthermore, it is inf-compact unless there is $c \in \mathbb{R}$ such that $g = c$ ν-a.s., in which case $K_{g,\nu} \equiv 0$.*

Proof. Define $f(t,\lambda) \stackrel{\text{def}}{=} \int \psi^\star(tg+\lambda)\,d\nu$, so that $K_{g,\nu}(t) = \inf_{\lambda\in\mathbb{R}} f(t,\lambda)$. Then since $\psi^\star$ is a non-negative proper convex function which is 0 at 0, we have that $f : \mathbb{R}^2 \to \overline{\mathbb{R}}$ is as well, so that $K_{g,\nu}(t)$ is non-negative, convex by Lemma 5.3.8, and satisfies $K_{g,\nu}(0) = 0$. Furthermore, we get that if $g = c$ holds ν-almost surely for $c \in \mathbb{R}$, then $0 \le K_{g,\nu}(t) \le f(t,-tc) = \int \psi^\star(tg-tc)\,d\nu = \psi^\star(0) = 0$ so that $K_{g,\nu}(t) = 0$ for all t, and the constant 0 function is lsc but not inf-compact.

Now, assume g is not almost-surely constant, so that we wish to show that $K_{g,\nu}$ is inf-compact, for which it suffices by Lemma 5.3.8 to show that the function $f(t,\lambda)$ is inf-compact as a function on $\mathbb{R}^2$. For lower semicontinuity, given a sequence $(t_n,\lambda_n) \to (t,\lambda)$, we have by Fatou's lemma and the non-negativity and lower semicontinuity of $\psi^\star$ that

$$\begin{aligned}\liminf_{n\to\infty} f(t_n,\lambda_n) = \liminf_{n\to\infty} \int \psi^\star(t_n g+\lambda_n)\,d\nu \\ \ge \int \liminf_{n\to\infty} \psi^\star(t_n g+\lambda_n)\,d\nu \ge \int \psi^\star(tg+\lambda)\,d\nu = f(t,\lambda)\,,\end{aligned}$$

so since $\mathbb{R}^2$ is a metric space the sets $\left\{(t,\lambda)\in\mathbb{R}^2 \mid f(t,\lambda)\le\alpha\right\}$ are closed, and it remains only to show that they are bounded. Since g is not almost surely constant, there exists $p > 0$ and disjoint compact intervals $[a_1,b_1]$ and $[a_2,b_2]$ such that $\Pr_\nu\big(g \in [a_i,b_i]\big) \ge p$ for $i \in \{1,2\}$. Furthermore, since $\psi^\star$ is inf-compact, there exists $\tau \ge 0$ such that $|x| \ge \tau$ implies $\psi^\star(x) \ge 2(1+\alpha)/p$. Thus, by Markov's inequality we have that $f(t,\lambda) \le \alpha$ only if $\Pr_\nu\big(|tg+\lambda| \ge \tau\big) \le p/2$, which in particular requires that there exist $x_i \in [a_i,b_i]$ such that $|tx_i+\lambda| \le \tau$ for $i \in \{1,2\}$. But since $[a_1,b_1]$ and $[a_2,b_2]$ are disjoint, there is a $t_0 > 0$ such that for $|t| > t_0$, the intervals $[ta_1,tb_1]$ and $[ta_2,tb_2]$ are at distance at least 4τ so no λ satisfies the constraint, and then for smaller $|t| \le t_0$ we have that $|tx_i+\lambda| \le \tau$ implies $|\lambda| \le \tau + |tx_i| \le \tau + t_0 \cdot \sup_{x_i\in[a_i,b_i]}|x_i|$ so that the set of (t,λ) with $|tx_i+\lambda| < \tau$ is uniformly bounded. □

With these definitions, we can state the main result of this section giving an expression for the optimal lower bound function.

Theorem 5.5.12. *Let $\nu \in \mathcal{M}^1$. Then for every $g \in L^1(\nu)$ and $\varepsilon \in \operatorname{int}(\operatorname{dom} \mathscr{L}_{g,\nu})$,*

$$\mathscr{L}_{g,\nu}(\varepsilon) = K^\star_{g,\nu}(\varepsilon)\,. \tag{5.22}$$

Similarly, for every $g \in \mathcal{L}^1(\nu)$ and $\varepsilon \in \operatorname{int}(\operatorname{dom} \mathscr{L}_{g,\nu,\perp})$,

$$\mathscr{L}_{g,\nu,\perp}(\varepsilon) = K^\star_{g,\nu,\perp}(\varepsilon)\,, \tag{5.23}$$

and if $\phi'(\infty) = \infty$ then $\mathscr{L}_{g,\nu,\perp} = \mathscr{L}_{g,\nu}$.

Furthermore, if $\mathscr{L}$ is lower semi-continuous, equivalently if strong duality holds in (5.16), then (5.22) and (5.23) hold for all $\varepsilon \in \mathbb{R}$.

Remark 5.5.13. As in Remark 5.5.4, the above theorem holds more generally for $\mathscr{L}_{g,\nu,X^1}$ and $\mathscr{L}_{g,\nu,\mathcal{X}^1}$ for any spaces X and $\mathcal{X}$ satisfying Assumption 5.4.5.

Proof. Lemma 5.5.2 implies that if $\phi'(\infty) = \infty$ that $\mathscr{L}_{g,\nu} = \mathscr{L}_{g,\nu,\perp}$, and more generally that $\mathscr{L}$ is proper and convex for $\mathscr{L} \in \{\mathscr{L}_{g,\nu}, \mathscr{L}_{g,\nu,\perp}\}$. Thus, by the Fenchel–Moreau theorem, we have $\mathscr{L} = \mathscr{L}^{\star\star}$ except possibly at the boundary of its domain, so this is simply a restatement of Proposition 5.5.3 using the terminology from Definition 5.5.8. □

The following easy corollary is an "operational" restatement of Theorem 5.5.12 highlighting the duality between upper bounding the cumulant generating function and lower bounding the ϕ-divergence by a convex lower semicontinuous function of the mean deviation.

Corollary 5.5.14. *Let $\nu \in \mathcal{M}^1$ and $g \in L^1(\nu)$ (resp. $g \in \mathcal{L}^1(\nu)$). Then for every convex lower semicontinuous function $L : \mathbb{R} \to \overline{\mathbb{R}}_{\geq 0}$, the following are equivalent:*

(i) $D_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ *for every* $\mu \in \mathcal{M}_c^1(\nu)$ *(resp.* $\mu \in \mathcal{M}^1$*) integrating* g*.*

(ii) $K_{g,\nu} \leq L^\star$ *(resp.* $K_{g,\nu,\perp} \leq L^\star$*).*

Example 5.5.15. The Hammersley–Chapman–Robbins bound in statistics is an immediate corollary of Corollary 5.5.14 applied to the χ^2-divergence given by $\phi(x) = (x-1)^2 + \delta_{\mathbb{R}_{\geq 0}}(x)$: The convex conjugate of $\psi(x) = x^2 + \delta_{[-1,\infty)}(x)$ is

$$\psi^\star(x) = \begin{cases} x^2/4 & x \geq -2 \\ -1 - x & x < -2 \end{cases}$$

and satisfies in particular $\psi^\star(x) \leq x^2/4$, so that $K_{g,\nu}(t) \leq \inf_\lambda \int (tg + \lambda)^2/4 \, d\nu = t^2 \operatorname{Var}_\nu(g)/4$. Since the convex conjugate of $t \mapsto t^2 \operatorname{Var}_\nu(g)/4$ is $t \mapsto t^2/\operatorname{Var}_\nu(g)$, we obtain for all $\mu, \nu \in \mathcal{M}^1$ and $g \in L^1(\nu)$ that $\chi^2(\mu \parallel \nu) \geq (\mu(g) - \nu(g))^2/\operatorname{Var}_\nu(g)$.

Theorem 5.5.12 also gives a useful characterization of the existence of a non-trivial lower bound by the *absolute* mean deviation.

Corollary 5.5.16. *For* $\nu \in \mathcal{M}^1$ *and* $g \in L^1(\nu)$*, the optimal lower bound* $\mathscr{L}_{\{\pm g\},\nu}$ *is non-zero if and only if* $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$*. Similarly, for* $g \in \mathcal{L}^1(\nu)$ *the optimal lower bound* $\mathscr{L}_{\{\pm g\},\nu,\perp}$ *is non-zero if and only if* $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu,\perp})$*. In other words, the following are equivalent*

(i) there exists a non-zero function $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}_{\geq 0}$ *such that* $D_\phi(\mu \parallel \nu) \geq L\big(|\mu(g) - \nu(g)|\big)$ *for every* $\mu \in \mathcal{M}_c^1(\nu)$ *(resp.* $\mu \in \mathcal{M}^1$*) integrating* g*.*

(ii) the function $K_{g,\nu}$ *(resp.* $K_{g,\nu,\perp}$*) is finite on an open interval around* 0*.*

Proof. For $M \in \{X_g^1(\nu), \mathcal{X}_g^1\}$ we have by Eq. (5.17) that the function $\mathscr{L}_{\{\pm g\},\nu,M}$ is non-zero if and only if there exists $\varepsilon > 0$ such that $\mathscr{L}_{g,\nu,M}(\varepsilon) \neq 0 \neq \mathscr{L}_{g,\nu,M}(-\varepsilon)$. Since $\mathscr{L}_{g,\nu,M}$ is convex, non-

negative, and 0 at 0 by Lemma 5.5.2, such an ε exists if and only if 0 is contained in the interval $\left(\mathscr{L}'_{g,\nu,M}(-\infty), \mathscr{L}'_{g,\nu,M}(\infty)\right)$, the interior of the domain of $\mathscr{L}^{\star}_{g,\nu,M}$. □

Remark 5.5.17. We will see in Theorem 5.6.15 that when we consider a true IPM where we require the bound L to hold for all ν, the distinction between the absolutely continuous case $\mathcal{M}^1_c(\nu)$ and the general case $\mathcal{M}^1$ disappears.

5.5.2 Subexponential functions and connections to Orlicz spaces

In Sections 5.5.2 to 5.5.4, we explore properties of the set of functions satisfying the conditions of Corollary 5.5.16, i.e. for which there is a non-trivial lower bound of the ϕ-divergence in terms of the absolute mean deviation, and show its relation to topological properties of the divergence. A reader primarily interested in quantitative bounds for IPMs can skip to Section 5.6.

In light of Corollary 5.5.16, we need to consider the set of functions g such that $\operatorname{dom} K_{g,\nu}$ or $\operatorname{dom} K_{g,\nu,\perp}$ contains a neighborhood of zero. The following lemma shows that this is the case for bounded functions, and that furthermore, when $\phi'(\infty) < \infty$, boundedness is necessary. In other words, when $\phi'(\infty) < \infty$, the ϕ-divergence cannot upper bound the absolute mean deviation of an unbounded function. This is in sharp contrast with the KL divergence (satisfying $\phi'(\infty) = \infty$), for which such upper bounds exist as long as the function satisfies Gaussian-type tail bounds [BLM13, §4.10].

Lemma 5.5.18. *Let $\nu \in \mathcal{M}^1$ be a probability measure. If $g \in L^\infty(\nu)$ (resp. $g \in \mathcal{L}^b(\Omega)$) then* $\operatorname{dom} K_{g,\nu}$ *(resp.* $\operatorname{dom} K_{g,\nu,\perp}$*) is all of $\mathbb{R}$, and in particular contains a neighborhood of zero. Furthermore, when $\phi'(\infty) < \infty$, we have conversely that if* 0 *is in the interior of the domain of $K_{g,\nu}$ (resp.* $\operatorname{dom} K_{g,\nu,\perp}$*), then $g \in L^\infty(\nu)$ (resp. $g \in \mathcal{L}^b(\Omega)$), in which case $K_{g,\nu}(t) = K_{g,\nu,\perp}(t)$ whenever $|t| \cdot \left(\sup g(\Omega) - \inf g(\Omega)\right) \le \phi'(\infty)$.*

Remark 5.5.19. As already discussed, Lemma 5.5.18 implies that when $\phi'(\infty) < \infty$, boundedness of g is necessary for the existence of a non-trivial lower bound on $D_\phi(\mu \parallel \nu)$ in terms of the $|\mu(g) - \nu(g)|$. Moreover, we can deduce from Lemma 5.5.18 that in this case, any non-trivial lower bound must depend on $\|g\|_\infty$ and cannot depend only on properties of g such as its ν-variance. In particular, any non-trivial lower bound must converge to 0 as $\|g\|_\infty$ converges to $+\infty$, for if it were not the case, one could obtain a non-trivial lower bound for an unbounded function g by approximating it with bounded functions $g \cdot \mathbf{1}\{|g| \le n\}$.

Proof. Recall that $(-\infty, 0] \subseteq \operatorname{dom} \psi^\star$ and that $\psi^\star(x) \le -x$ for $x \le 0$ by Lemma 5.5.7. For $g \in L^\infty(\nu)$ (resp. $g \in \mathcal{L}^b(\nu)$), write B for $\operatorname{ess\,sup}_\nu |g|$ (resp. $\sup_{\omega\in\Omega} |g(\omega)|$), and for $t \in \mathbb{R}$, write $\lambda \stackrel{\text{def}}{=} -|t| \cdot B$. Then we have that $-2|t|B \le t \cdot g(\omega) + \lambda \le 0 \le \phi'(\infty)$ holds ν-a.s. (resp. for all $\omega \in \Omega$), and thus also $\psi^\star(tg(\omega) + \lambda) \le 2|t| \cdot B$ holds ν-a.s. Thus $K_{g,\nu}(t)$ (resp. $K_{g,\nu,\perp}(t)$) is at most $2|t| \cdot B < \infty$ by definition, and since t is arbitrary, we get $\operatorname{dom} K_{g,\nu} = \mathbb{R}$ (resp. $\operatorname{dom} K_{g,\nu,\perp} = \mathbb{R}$).

We now assume $\phi'(\infty) < \infty$ and prove the converse claim. If $K_{\nu,g}(t)$ (resp. $K_{\nu,g,\perp}(t)$) is finite for some $t \in \mathbb{R}$, then $tg + \lambda \le \phi'(\infty)$ holds ν-a.s. (resp. for all $\omega \in \Omega$). In particular, if it holds for some $t > 0$, then $\operatorname{ess\,sup}_\nu g$ (resp. $\sup g(\Omega)$) is finite, and if it holds for some $t < 0$, then $\operatorname{ess\,inf}_\nu g$ (resp. $\inf g(\Omega)$) is finite.

For the remaining claim, since $\psi^\star$ is non-decreasing on the non-negative reals we have that $K_{g,\nu}(t) = \inf\{I_{\psi^\star,\nu}(tg+\lambda) \,|\, \lambda \in \mathbb{R}\} = \inf\{I_{\psi^\star,\nu}(tg+\lambda) \,|\, \inf(\lambda + t \cdot g(\Omega)) \le 0\}$. But if $\sup(t \cdot g(\Omega)) - \inf(t \cdot g(\Omega)) \le \phi'(\infty)$, then $\inf(\lambda + t \cdot g(\Omega)) \le 0$ implies $\lambda + \sup(t \cdot g(\Omega)) \le \phi'(\infty)$ and $K_{g,\nu}(t) \ge K_{g,\nu,\perp}(t) \ge K_{g,\nu}(t)$. □

Since Lemma 5.5.18 completely characterizes the existence of a non-trivial lower bound when $\phi'(\infty) < \infty$, we focus on the case $\phi'(\infty) = \infty$ in the remainder of this section. Recall that $K_{g,\nu} = K_{g,\nu,\perp}$

in this case, so we only need to consider $K_{g,\nu}$ in the following definition.

Definition 5.5.20 ((ϕ,ν)-subexponential functions)**.** Let $\nu \in \mathcal{M}^1$ be a probability measure. We say that the function $g \in L^0(\nu)$ is (ϕ,ν)-*subexponential* if $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$ and we denote by $S^\phi(\nu)$ the space of all such functions. We further say that $g \in L^0(\nu)$ is *strongly* (ϕ,ν)-*subexponential* if $\operatorname{dom} K_{g,\nu} = \mathbb{R}$ and denote by $S^\phi_\star(\nu)$ the space of all such functions.

Example 5.5.21. For the case of the KL-divergence, if the pushforward $g_*\nu$ of ν induced by g on $\mathbb{R}$ is the Gaussian distribution (respectively the gamma distribution), then g is strongly subexponential (respectively subexponential). Furthermore, it follows from Example 5.5.9 that $g \in S^\phi(\nu)$ iff the moment-generating function of g is finite on a neighborhood of 0, which is the standard definition of subexponential functions (see e.g. [Ver18, §2.7]) and thus justifies our terminology.

Example 5.5.22. Lemma 5.5.18 shows that $L^\infty(\nu) \subseteq S^\phi_\star(\nu)$ and that furthermore, if $\phi'(\infty) < \infty$, then $L^\infty(\nu) = S^\phi_\star(\nu) = S^\phi(\nu)$.

We start with the following key lemma allowing us to relate the finiteness of $K_{g,\nu}$ to the finiteness of the function $t \mapsto I_{\psi^\star,\nu}(tg)$.

Lemma 5.5.23. *For $\nu \in \mathcal{M}^1$, $g \in L^0(\nu)$, and $t \in \operatorname{dom} K_{g,\nu}$, we have that if $\phi'(\infty) = \infty$ (resp. $\phi'(\infty) > 0$) then $\alpha t g \in \operatorname{dom} I_{\psi^\star,\nu}$ for all $\alpha \in (0,1)$ (resp. for sufficiently small $\alpha > 0$).*

Proof. Let $\lambda \in \mathbb{R}$ be such that $\int \psi^\star(tg+\lambda)\,d\nu < \infty$ (such a λ exists since $t \in \operatorname{dom} K_{g,\nu}$). Using the convexity of $\psi^\star$, we get for any $\alpha \in (0,1)$

$$\begin{aligned}\int \psi^\star(\alpha t g)\,d\nu &= \int \psi^\star\left(\alpha(tg+\lambda) + (1-\alpha)\frac{-\alpha\lambda}{1-\alpha}\right)d\nu \\ &\leq \alpha \int \psi^\star(tg+\lambda)\,d\nu + (1-\alpha)\psi^\star\left(\frac{-\alpha\lambda}{1-\alpha}\right).\end{aligned}$$

The first summand is finite by definition, and if $-\alpha\lambda/(1-\alpha) \in \operatorname{dom}\psi^\star \supseteq (-\infty, \phi'(\infty))$ then so is the second summand. If $\phi'(\infty) = \infty$ this holds for all $\alpha \in (0,1)$, and if $\phi'(\infty) > 0$ it holds for sufficiently small $\alpha > 0$. □

Remark 5.5.24. When $\phi'(\infty) < \infty$, it is not necessarily true that any $\alpha \in (0,1)$ can be used in Lemma 5.5.23. For example, Lemma 5.5.18 implies that $\operatorname{dom} K_{g,\nu} = \mathbb{R}$ for all $g \in L^\infty(\nu)$, but since $\operatorname{dom}\psi^\star \subseteq \left(-\infty, \phi'(\infty)\right]$ we have $I_{\psi^\star,\nu}(tg) = \infty$ for sufficiently large (possibly only positive or negative) t, unless g is zero ν-a.s.

The following proposition gives useful characterizations of subexponential functions in terms of the finiteness of different integral functionals of g.

Proposition 5.5.25. *Suppose that $\phi'(\infty) = \infty$ and fix $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$. Then the following are equivalent:*

(i) *g is (ϕ, ν)-subexponential*

(ii) *$K_{|g|,\nu}(t) < \infty$ for some $t > 0$*

(iii) *$g \in L^\theta(\nu)$ for $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$ (here $L^\theta(\nu)$ is the Orlicz space defined in Section 5.3.3)*

Proof. *(i)* $\implies$ *(ii)* If $\operatorname{dom} K_{g,\nu}$ contains an open interval around 0, Lemma 5.5.23 and the convexity of $\operatorname{dom} I_{\psi^\star,\nu}$ imply that there exists $s > 0$ such that $\int \psi^\star(tg)\, d\nu < \infty$ for all $|t| < s$. By non-negativity of $\psi^\star$, $\int \psi^\star(t|g|)\, d\nu \le \int \psi^\star(tg) + \psi^\star(-tg)\, d\nu < \infty$ for all $t \in (-s, s)$, which in turns implies $(-s, s) \subseteq \operatorname{dom} K_{|g|,\nu}$.

(ii) $\implies$ *(iii)* Define $\eta(x) \stackrel{\text{def}}{=} \psi^\star(|x|)$. Since $\psi^\star(x) \le -x$ for $x \le 0$ by Lemma 5.5.7, we have that $\eta(x) \le \theta(x) \le \eta(x) + |x|$ for all $x \in \mathbb{R}$. Since we also have $L^\eta(\nu) \subseteq L^1(\nu)$, this implies that

$g \in L^\eta(\nu)$ if and only if $L^\theta(\nu)$. We conclude after observing that $K_{|g|,\nu}(t) < \infty$ for some $t > 0$ implies that $g \in L^\eta(\nu)$ by Lemma 5.5.23.

$(iii) \implies (i)$ Observe that for all $t \in \mathbb{R}$,

$$\max\{K_{g,\nu}(t), K_{g,\nu}(-t)\} \le \max\left\{\int \psi^\star(tg)\,d\nu, \int \psi^\star(-tg)\,d\nu\right\} \le \int \theta(tg)\,d\nu\,, \tag{5.24}$$

where the first inequality is by definition of $K_{g,\nu}$ and the second inequality is by monotonicity of the integral and the definition of θ. Since $\operatorname{dom} K_{g,\nu}$ is convex, if there exists $t > 0$ such that $I_{\theta,\nu}(tg) < \infty$, then (5.24) implies that $[-t, t] \subseteq \operatorname{dom} K_{g,\nu}$ and g is (ϕ, ν)-subexponential. □

Remark 5.5.26. Though Proposition 5.5.25 implies that the set of (ϕ, ν)-subexponential functions is the same as the set $L^\theta(\nu)$ for $\theta(x) = \max\{\psi^\star(x), \psi^\star(-x)\}$, we emphasize that the Luxemburg norm $\|\cdot\|_\theta$ does *not* capture the relationship between $\mathrm{D}_\phi(\mu \,\|\, \nu)$ and the absolute mean deviation $|\mu(g) - \nu(g)|$. First, the function θ, being a symmetrization of $\psi^\star$, induces integral functionals which are potentially much larger than those defined by $\psi^\star$, in particular it is possible to have $\max\{K_{g,\nu}(t), K_{g,\nu}(-t)\} < \inf_{\lambda\in\mathbb{R}} I_{\theta,\nu}(tg + \lambda) < I_{\theta,\nu}(tg)$. Furthermore, the Luxemburg norm summarizes the growth of $t \mapsto I_{\theta,\nu}(tg)$ with a single number (specifically its inverse at 1), whereas Theorem 5.5.12 shows that the relationship with the mean deviation is controlled by $K^\star_{g,\nu}$, which depends on the growth of $K_{g,\nu}(t)$ with t.

We are now ready to prove the main result of this section, which is that the space $S^\phi(\nu)$ of (ϕ, ν)-subexponential functions is the largest space of functions which can be put in dual pairing with (the span of) all measures $\mu \in \mathcal{M}_c(\nu)$ such that $\mathrm{D}_\phi(\mu \,\|\, \nu) < \infty$, i.e. $\operatorname{dom} I_{\phi,\nu}$.

Theorem 5.5.27. *For $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$, the following are equivalent:*

(i) g is (ϕ, ν)-subexponential, i.e. $g \in S^\phi(\nu)$.

(ii) *g is μ-integrable for every $\mu \in \mathcal{M}_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$.*

(iii) *g is μ-integrable for every $\mu \in \mathcal{M}_c^1(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$.*

Proof. $(i) \implies (ii)$ If $\phi'(\infty) < \infty$ this follows since $L^\infty(\nu) = S^\phi(\nu)$, so assume that $\phi'(\infty) = \infty$. If $g \in S^\phi(\nu)$ then $g \in L^\theta(\nu)$ for $\theta(x) = \max\{\psi^\star(x), \psi^\star(-x)\}$ by Proposition 5.5.25. Since $\theta \geq \psi^\star$ we have $\theta^\star \leq \psi$, and thus for $\mu \in \mathcal{M}_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$,

$$I_{\theta^\star,\nu}\left(\frac{d\mu}{d\nu} - 1\right) \leq I_{\psi,\nu}\left(\frac{d\mu}{d\nu} - 1\right) = \mathrm{D}_\phi(\mu \parallel \nu) < \infty\,,$$

implying that $\frac{d\mu}{d\nu} - 1 \in L^{\theta^\star}(\nu)$. Furthermore, since $1 \in L^\infty(\nu) \subseteq L^{\theta^\star}(\nu)$ we get that $\frac{d\mu}{d\nu} \in L^{\theta^\star}(\nu)$. Property 2. then follows from the fact that $(L^{\theta^\star}, L^\theta)$ form a dual pair.

$(ii) \implies (iii)$ Immediate.

$(iii) \implies (i)$ Define $C \stackrel{\text{def}}{=} \left\{\mu \in \mathcal{M}_c^1(\nu) \,\middle|\, \mathrm{D}_\phi(\mu \parallel \nu) \leq 1\right\}$, which is closed and convex as a sublevel set of the convex lower semicontinuous functional $\widetilde{\mathrm{D}}_{\phi,\nu}$ on the Banach space $\mathcal{M}_c(\nu)$ with the total variation norm (recall that this space is isomorphic to $L^1(\nu)$ by the Radon–Nikodym theorem). Since furthermore $C \subseteq \mathcal{M}^1$, it is bounded in $\mathcal{M}_c(\nu)$ and so is cs-compact [Jam72, Proposition 2]. Then by assumption, the linear function $\mu \mapsto \mu(|g|)$ is well-defined and bounded below by 0 on C, so Lemma 5.3.17 implies that there exists $B \in \mathbb{R}$ such that $\mu(|g|) \leq B$ for all $\mu \in C$. Thus, we get that for all $\mu \in C$, $|\mu(g) - \nu(g)| \leq \mu(|g|) + \nu(|g|) \leq B + \nu(|g|)$. In particular, if $|\mu(g) - \nu(g)| > B + \nu(|g|)$ then $\mathrm{D}_\phi(\mu \parallel \nu) > 1$, proving the existence of a non-zero function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(|\mu(g) - \nu(g)|\big)$. This implies that $g \in S^\phi(\nu)$ by Corollary 5.5.16. □

We have the following characterization of the space $S_\star^\phi(\nu)$ of strongly subexponential functions. In particular $S_\star^\phi(\nu)$ can be identified as a set with $L^\infty(\nu)$ or the Orlicz heart $L_\heartsuit^\theta(\nu)$ depending on whether $\phi'(\infty)$ is finite or infinite (with the finite case from Lemma 5.5.18).

Proposition 5.5.28. *Suppose that $\phi'(\infty) = \infty$ and fix $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$. Then the following are equivalent:*

(i) *g is strongly (ϕ, ν)-subexponential, i.e. $g \in S^\phi_\star(\nu)$.*

(ii) *$K_{|g|,\nu}(t) < \infty$ for all $t > 0$.*

(iii) *$g \in L^\theta_\heartsuit(\nu)$ for $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$.*

Proof. $(i) \implies (ii)$ Since $\phi'(\infty) = \infty$, Lemma 5.5.23 implies that $tg \in \operatorname{dom} I_{\psi^\star,\nu}$ for all $t \in \mathbb{R}$, and since $\psi^\star$ is non-negative we have for each $t > 0$ that $K_{|g|,\nu}(t) \leq \int \psi^\star(t|g|)\, d\nu \leq \int \psi^\star(tg) + \psi^\star(-tg)\, d\nu < \infty$.

$(ii) \implies (iii)$ Define $\eta : x \mapsto \psi^\star(|x|)$, so that by Lemma 5.5.23 we have $\int \eta(tg)\, d\nu = \int \psi^\star(t|g|)\, d\nu < \infty$ for all $t > 0$, and hence Property 2. implies $g \in L^\eta_\heartsuit(\nu)$. As in the proof of Proposition 5.5.25, $\eta(x) \leq \theta(x) \leq \eta(x) + |x|$ for all $x \in \mathbb{R}$ and since $L^\eta_\heartsuit(\nu) \subseteq L^1(\nu)$, we have that $g \in L^\eta_\heartsuit(\nu)$ iff $g \in L^\theta_\heartsuit(\nu)$.

$(iii) \implies (i)$ Immediate since for $t \in \mathbb{R}$, $K_{g,\nu}(t) \leq \int \psi^\star(tg)\, d\nu \leq \int \theta(tg)\, d\nu < \infty$. □

Finally, we collect several statements from this section and express them in a form which will be convenient for subsequent sections.

Corollary 5.5.29. *Define $\theta(x) \stackrel{\text{def}}{=} \max\{\psi^\star(x), \psi^\star(-x)\}$. Then we have $S^\phi_\star(\nu) \subseteq S^\phi(\nu) \subseteq L^1(\nu)$ and $\operatorname{dom} I_{\phi,\nu} \subseteq L^{\theta^\star}(\nu) \subseteq L^1(\nu)$. Furthermore, $L^{\theta^\star}(\nu)$ is in dual pairing with both $S^\phi(\nu)$ and $S^\phi_\star(\nu)$, and when $\phi'(\infty) = \infty$ the topology induced by $\|\cdot\|_\theta$ on $S^\phi_\star(\nu)$ is complete and compatible with the pairing.*

Proof. The containment $S^\phi(\nu) \subseteq L^1(\nu)$ is because $S^\phi(\nu)$ is equal as a set to the Orlicz space $L^\theta(\nu)$ by Proposition 5.5.25, and the containment $\operatorname{dom} I_{\phi,\nu} \subseteq L^{\theta^\star}(\nu)$ can be found in the proof of $(i) \implies (ii)$ of Theorem 5.5.27. The fact that $\left(L^{\theta^\star}(\nu), S^\phi(\nu)\right)$ form a dual pair is also immediate from the identification

of $S^\phi(\nu)$ with $L^\theta(\nu)$ as a set. Finally, the last claim follows from the identification of $S^\phi_\star(\nu)$ with $L^\theta_\heartsuit(\nu)$ as a set and the fact that when $\phi'(\infty) = \infty$, then $\operatorname{dom}\theta = \mathbb{R}$ implying that the topological dual of the Banach space $(L^\theta_\heartsuit(\nu), \|\cdot\|_\theta)$ is isomorphic to $(L^{\theta^\star}(\nu), \|\cdot\|_{\theta^\star})$. □

5.5.3 Inf-compactness of divergences and connections to strong duality

In this section, we study the question of inf-compactness of the functional $\mathrm{D}_{\phi,\nu}$ and that of its restriction $\widetilde{\mathrm{D}}_{\phi,\nu}$ to probability measures. Specifically, we wish to understand under which topology the information "ball" $\mathcal{B}_{\phi,\nu}(\tau) \stackrel{\text{def}}{=} \{\mu \in \mathcal{M} \mid \mathrm{D}_\phi(\mu \,\|\, \nu) \le \tau\}$ is compact. Beyond being a natural topological question, it also has implications for strong duality in Theorem 5.5.12, since the following lemma shows that compactness of the ball under suitable topologies implies strong duality.

Lemma 5.5.30. *For every g, ν, and M as in Definition 5.5.1, if $\mu \mapsto \mathrm{D}_\phi(\mu \,\|\, \nu)$ is inf-compact (or even countably inf-compact) with respect to a topology on M such that $\mu \mapsto \mu(\mathrm{g})$ is continuous, then $\mathscr{L}_{\mathrm{g},\nu,M}$ is inf-compact (and in particular lower semicontinuous), so that strong duality holds in Theorem 5.5.12.*

Proof. Recall from Eq. (5.16) that

$$\mathscr{L}_{\mathrm{g},\nu,M}(\varepsilon) = \inf_{\mu\in M} \mathrm{D}_\phi(\mu \,\|\, \nu) + \delta_{\{0\}}(\mu(\mathrm{g}) - \nu(\mathrm{g}) - \varepsilon)$$

where $f(\varepsilon,\mu) = \mathrm{D}_\phi(\mu \,\|\, \nu) + \delta_{\{0\}}(\mu(\mathrm{g}) - \nu(\mathrm{g}) - \varepsilon)$ is convex. Furthermore, under the stated assumption, we have that f is also inf-compact so that Lemma 5.3.8 gives the claim. □

Throughout this section, we assume that $\phi'(\infty) = \infty$,[3] which implies that $\operatorname{dom}\psi^\star = \mathbb{R}$ by Lemma 5.5.7, and furthermore that $\mu \in \mathcal{M}_c(\nu)$ whenever $\mathrm{D}_\phi(\mu \,\|\, \nu) < \infty$ and hence $\mathrm{D}_{\phi,\nu} = I_{\phi,\nu}$ and $\mathcal{B}_{\phi,\nu}(\tau) \subset \mathcal{M}_c(\nu)$ for all $\tau \ge 0$. It is well known that in this case, $\mathcal{B}_{\phi,\nu}(\tau)$ is compact in the weak

[3] When $\phi'(\infty) < \infty$, compactness of information balls is very dependent on the specific measure space $(\Omega, \mathcal{A}, \nu)$, and in this chapter we avoid such conditions.

topology $\sigma\big(L^1(\nu), L^\infty(\nu)\big)$ (e.g. [Roc71, Corollary 2B] or [TV93]). This fact can be derived as a simple consequence of the Dunford–Pettis theorem since $\mathcal{B}_{\phi,\nu}(\tau)$ is uniformly integrable by the de la Vallée-Poussin theorem (see e.g. [Val70, pages 67–68]). In light of Lemma 5.5.30, it is however useful to understand whether $\mathcal{B}_{\phi,\nu}(\tau)$ is compact under topologies for which $\mu \mapsto \mu(g)$ is continuous, where g could be unbounded. Léonard [Léo01a, Theorem 3.4] showed, in the context of convex integral functionals on Orlicz spaces, that strong duality holds when $g \in S^\phi_\star(\nu)$, and in this section we reprove this result in the language of ϕ-divergences by noting (as is implicit in [Léo01a, Lemma 3.1]) that $\mathcal{B}_{\phi,\nu}(\tau)$ is compact for the initial topology induced by the maps of the form $\mu \mapsto \mu(g)$ for all strongly subexponential function $g \in S^\phi_\star(\nu)$.

Proposition 5.5.31. *Fix $\nu \in \mathcal{M}^1$ and define $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$ as in Proposition 5.5.28. If $\phi'(\infty) = \infty$, then the functional $I_{\phi,\nu}$ is $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\star(\nu)\big)$ inf-compact.*

Proof. By Corollary 5.5.29, we know that $\big(S^\phi_\star(\nu), \|\cdot\|_\theta\big)$ is a Banach space in dual pairing with $L^{\theta^\star}(\nu)$. Thus, from Proposition 5.4.8, the integral functional $I_{\phi^\star,\nu}$ defined on $S^\phi_\star(\nu)$ is convex, lower semicontinuous, and has conjugate $I_{\phi^{\star\star},\nu} = I_{\phi,\nu}$ on $L^{\theta^\star}(\nu)$. Furthermore, from Lemma 5.5.23 we know for every $g \in S^\phi_\star(\nu)$ that $I_{\phi^\star,\nu}(g) < \infty$, so $I_{\phi^\star,\nu}$ is convex, lsc, and finite everywhere on a Banach space, and thus continuous everywhere by [Brø64, p. 2.10]. Finally, [Mor64, Proposition 1] implies that its conjugate $I_{\phi,\nu}$ is inf-compact on $L^{\theta^\star}(\nu)$ with respect to the weak topology $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\star(\nu)\big)$. □

Remark 5.5.32. This result generalizes [Roc71, Corollary 2B] since $L^\infty(\nu) \subseteq S^\phi_\star(\nu)$ whenever $\phi'(\infty) = \infty$ (see Example 5.5.22).

Corollary 5.5.33. *Under the same assumptions and notations as Proposition 5.5.31, the functional $\widetilde{\mathrm{D}}_{\phi,\nu}$ is $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\star(\nu)\big)$ inf-compact.*

Proof. Observe that since $\phi(x) = \infty$ for $x < 0$, we have for every $\tau \in \mathbb{R}$ that $\{\mu \in L^{\theta^\star}(\nu) \mid \widetilde{\mathrm{D}}_{\phi,\nu}(\mu) \leq \tau\} = \{\mu \in \mathcal{M}^1 \cap L^{\theta^\star}(\nu) \mid I_{\phi,\nu}(\mu) \leq \tau\} = \{\mu \in L^{\theta^\star}(\nu) \mid I_{\phi,\nu}(\mu) \leq \tau\} \cap f^{-1}(1)$ where $f : \mu \to \mu(\mathbf{1}_\Omega)$ is continuous in the weak topology $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\star(\nu)\big)$ since $L^\infty(\nu) \subseteq S^\phi_\star(\nu)$ by Lemma 5.5.18. Hence, $\mathcal{M}^1 \cap \mathcal{B}_{\phi,\nu}(\tau)$ is compact as a closed subset of a compact set. □

Corollary 5.5.34. *If $\phi'(\infty) = \infty$, then for every $\tau \in \mathbb{R}$ the sets $\mathcal{B}_{\phi,\nu}(\tau)$ and $\mathcal{M}^1 \cap \mathcal{B}_{\phi,\nu}(\tau)$ are compact in the initial topology induced by $\{\mu \mapsto \mu(g) \mid g \in S^\phi_\star(\nu)\}$.*

Proof. Immediate from Proposition 5.5.31 and Corollary 5.5.33. □

Corollary 5.5.35. *Let $\nu \in \mathcal{M}^1$ be a probability measure and assume that $\phi'(\infty) = \infty$. If $g \in L^0(\nu)$ is strongly (ϕ, ν)-subexponential and $M \subseteq \mathcal{M}^1_c(\nu)$ is a convex set of probability measures containing every $\mu \in \mathcal{M}^1_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$, then the function $\mathscr{L}_{g,\nu,M}$ is lower semicontinuous.*

Proof. Follows from Lemma 5.5.30 and Corollary 5.5.34. □

Remark 5.5.36. Corollary 5.5.35 does not apply when $\phi'(\infty) < \infty$ or $g \in S^\phi(\nu) \setminus S^\phi_\star(\nu)$ (e.g. when the pushforward measure $g_*\nu$ is gamma-distributed in the case of the KL divergence), and it would be interesting to identify conditions other than inf-compactness of $\mathrm{D}_{\phi,\nu}$ under which $\mathscr{L}_{g,\nu}$ is lower semicontinuous.

5.5.4 Convergence in ϕ-divergence and weak convergence

Our goal in this section is to relate two notions of convergence for a sequence of probability measures $(\nu_n)_{n\in\mathbb{N}}$ and $\nu \in \mathcal{M}^1$: (i) $\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$,[4] and (ii) $|\nu_n(g) - \nu(g)| \to 0$ for $g \in \mathcal{L}^0(\Omega)$. Specifically, we would like to identify the largest class of functions $g \in \mathcal{L}^0(\Omega)$ such that *convergence in ϕ-divergence* (i)

[4] Throughout this section, we restrict our attention to ϕ which are not the constant 0 on a neighborhood of 1, i.e. such that $1 \notin \operatorname{int}\operatorname{dom}\{x \in \mathbb{R} \mid \phi(x) = 0\}$, as otherwise it is easy construct probability measures $\mu \neq \nu$ such that $\mathrm{D}_\phi(\mu \parallel \nu) = 0$, hence $\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$ does not define a meaningful convergence notion.

implies (ii). In other words, we would like to identify the finest initial topology induced by linear forms $\mu \mapsto \mu(g)$ for which (sequential) convergence is implied by (sequential) convergence in ϕ-divergence[5]. This question is less quantitative than computing the best lower bound of the ϕ-divergence in terms of the absolute mean deviation, since it only characterizes when $|\nu_n(g) - \nu(g)|$ converges to 0, whereas the optimal lower bound quantifies the *rate of convergence* to 0 when it occurs.

This has been studied in the specific case of the Kullback–Leibler divergence by Harremoës, who showed [Haro7, Theorem 25] that $\mathrm{KL}(\nu_n \parallel \nu) \to 0$ implies $|\nu_n(g) - \nu(g)| \to 0$ for every non-negative function g whose moment generating function is finite at some positive real (in fact, the converse was also shown in the same paper under a so-called *power-dominance* condition on ν). In this section, we generalize this to an arbitrary ϕ-divergence and show that convergence in ϕ-divergence implies $\nu_n(g) \to \nu(g)$ if and only if g is (ϕ, ν)-subexponential.

This question is also closely related the one of understanding the relationship between weak convergence and *modular convergence* in Orlicz spaces (e.g. [Nak50] or [Mus83]). Although convergence in ϕ-divergence as defined above only formally coincides with the notion of modular convergence when ϕ is symmetric about 1 (though this can sometimes be relaxed [Her67]) and satisfies the so-called Δ_2 growth condition, it is possible that this line of work could be adapted to the question studied in this section.

We start with the following proposition, showing that this question is equivalent to the differentiability of $\mathscr{L}^\star_{g,\nu}$ at 0.

Proposition 5.5.37. *Let $\nu \in \mathcal{M}^1$, $g \in \mathcal{L}^1(\nu)$, and $M \subseteq \mathcal{M}^1$ be a convex set of measures integrating g and*

[5] The natural notion of convergence in ϕ-divergence defines a topology on the space of probability measures for which continuity and sequential continuity coincide (see e.g. [Kis60; Dud64; Haro7]), so it is without loss of generality that we consider only sequences rather than nets in the rest of this section. Note that the information balls $\{\mu \in \mathcal{M}^1 \mid \mathrm{D}_\phi(\mu \parallel \nu) < \tau\}$ for $\tau > 0$ need not be neighborhoods of ν in this topology, and the information balls do not in general define a basis of neighborhoods for a topology on the space of probability measures [Csi62; Csi64; Csi67b; Dud98].

containing ν. Then the following are equivalent:

(i) $\lim_{n\to\infty} \nu_n(g) = \nu(g)$ *for all* $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ *such that* $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$.

(ii) $\mathscr{L}_{g,\nu,M}$ *is strictly convex at* 0*, that is* $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ *if and only if* $\varepsilon = 0$.

(iii) $\partial\mathscr{L}^\star_{g,\nu,M}(0) = \{0\}$*, that is* $\mathscr{L}^\star_{g,\nu,M}$ *is differentiable at* 0 *and* $\mathscr{L}^{\star\prime}_{g,\nu,M}(0) = 0$.

Proof. $(i) \implies (ii)$ Assume for the sake of contradiction that $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ for some $\varepsilon \neq 0$. Then by definition of $\mathscr{L}_{g,\nu,M}$, there exists a sequence $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ such that for all $n \in \mathbb{N}$, $\mathrm{D}_\phi(\nu_n \parallel \nu) \leq 1/n$ and $\nu_n(g) - \nu(g) = \varepsilon$, thus contradicting (i). Hence, $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ if and only if $\varepsilon = 0$, which is equivalent to strict convexity at 0 since $\mathscr{L}_{g,\nu,M}$ is convex with global minum $\mathscr{L}_{g,\nu,M}(0) = 0$ by Lemma 5.5.2.

$(ii) \implies (i)$ Let $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ be a sequence such that $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$. By definition of $\mathscr{L}_{g,\nu,M}$, we have that $\mathrm{D}_\phi(\nu_n \parallel \nu) \geq \mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) \geq 0$ for all $n \in \mathbb{N}$, and in particular $\lim_{n\to\infty} \mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) = 0$. Assume for the sake of contradiction that $\nu_n(g)$ does not converge to $\nu(g)$. This implies the existence of $\varepsilon > 0$ such that $|\nu_n(g) - \nu(g)| \geq \varepsilon$ for infinitely many $n \in \mathbb{N}$. But then $\mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) \geq \min\{\mathscr{L}_{g,\nu,M}(\varepsilon), \mathscr{L}_{g,\nu,M}(-\varepsilon)\} > 0$ for infinitely many $n \in \mathbb{N}$, a contradiction.

$(ii) \iff (iii)$ By a standard characterization of the subdifferential (see e.g. [Zălo2, Theorem 2.4.2(iii)]), we have that $\partial\mathscr{L}^\star_{g,\nu,M}(0) = \{x \in \mathbb{R} \mid \mathscr{L}^\star_{g,\nu,M}(0) + \mathscr{L}^{\star\star}_{g,\nu,M}(x) = 0 \cdot x\} = \{x \in \mathbb{R} \mid \mathscr{L}^{\star\star}_{g,\nu,M}(x) = 0\}$. Since $\mathscr{L}_{g,\nu,M}$ is convex, non-negative, and 0 at 0, this subdifferential contains $\varepsilon \neq 0$ if and only if there exists $\varepsilon \neq 0$ with $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$. □

The above proposition characterizes continuity in terms of the differentiability at 0 of the conjugate of the optimal lower bound function, or equivalently by Proposition 5.5.3, differentiability of the

functions $K_{g,\nu}$ and $K_{g,\nu,\perp}$. In the previous section we investigated in detail the finiteness (or equivalently by convexity, the continuity) of these functions around 0; in this section we show that continuity at 0 is equivalent to differentiability at 0 assuming that ϕ is not the constant 0 on a neighborhood of 1.

Proposition 5.5.38. *Assume that* $1 \notin \operatorname{int}\{x \in \mathbb{R} \mid \phi(x) = 0\}$. *Then for* $\nu \in \mathcal{M}^1$ *and* $g \in \mathcal{L}^0(\Omega)$, *we have that* $0 \in \operatorname{int} \operatorname{dom} K_{g,\nu}$ *(resp.* $0 \in \operatorname{int} \operatorname{dom} K_{g,\nu,\perp}$*) if and only if* $K'_{g,\nu}(0) = 0$ *(resp.* $K'_{g,\nu,\perp}(0) = 0$*).*

Proof. The if direction is immediate, since differentiability at 0 implies continuity at 0. Thus, for the remainder of the proof we assume that $K_{g,\nu}$ (resp. $K_{g,\nu,\perp}$) is finite on a neighborhood of 0.

We first consider the case $\phi'(\infty) < \infty$, where Lemma 5.5.18 implies $g \in L^\infty(\nu)$ (resp. $g \in \mathcal{L}^b(\Omega)$). Define $B \stackrel{\text{def}}{=} \operatorname{ess\,sup}_\nu |g|$ (resp. $B \stackrel{\text{def}}{=} \sup |g|(\Omega)$), and let $\sigma \in \{-1, 1\}$ be such that $\phi(1 + \sigma x) > 0$ for all $x > 0$ as exists by assumption on ϕ. Since ψ is non-negative and 0 at 0, a standard characterization of the subdifferential (e.g. [Zălo2, Theorem 2.4.2(iii)]) implies that the function $t \mapsto \psi^\star(\sigma|t|)$ has derivative 0 at 0. Then for all $t \in \mathbb{R}$, by considering $\lambda = \sigma t B$ in (5.20), we obtain $K_{g,\nu}(t)$ (resp. $K_{g,\nu,\perp}(t)$) is at most $\nu\big(\psi^\star(tg + \sigma t B)\big) + \delta_{[-\infty,\phi'(\infty)]}(2\sigma|t|B) \le \psi^\star(2\sigma|t|B) + \delta_{[-\infty,\phi'(\infty)]}(2\sigma|t|B)$. Now, if $\sigma = -1$ then $2\sigma|t|B \le 0 \le \phi'(\infty)$ for all t, and if $\sigma = 1$ then necessarily $\phi'(\infty) > 0$ and so $2\sigma|t|B \le \phi'(\infty)$ for sufficiently small $|t|$. Thus, we have for sufficiently small $|t|$ that $K_{g,\nu}(t)$ (resp. $K_{g,\nu,\perp}(t)$) is between 0 and $\psi^\star(2\sigma|t|B)$, both of which are 0 with derivative 0 at 0, completing the proof in this case.

Now, assume that $\phi'(\infty) = \infty$, so that we have $K_{g,\nu,\perp} = K_{g,\nu} = \inf_{\lambda\in\mathbb{R}} f(\cdot, \lambda)$ for $f(t, \lambda) \stackrel{\text{def}}{=} \nu\big(\psi^\star(tg + \lambda)\big)$. Note that $\psi \ge 0$ implies $f \ge 0$, so since $K_{g,\nu}(0) = f(0, 0) = 0$ we have by standard results in convex analysis (e.g. [Zălo2, Theorem 2.6.1(ii)]) that $\partial K_{g,\nu}(0) = \{t^\star \mid (t^\star, 0) \in \partial f(0, 0)\}$. Furthermore, by assumption $K_{g,\nu}$ is finite on a neighborhood of 0, so since $K_{g,\nu} = K_{g+c,\nu}$ for all $c \in \mathbb{R}$, Lemma 5.5.23 implies $\operatorname{int}\big(\operatorname{dom} K_{g,\nu}\big) \times \mathbb{R} \subseteq \operatorname{dom} f$ and in particular $(0, 0) \in \operatorname{int} \operatorname{dom} f$. Thus, defining for each $\omega \in \Omega$ the function $f_\omega(t, \lambda) \stackrel{\text{def}}{=} \psi^\star(t \cdot g(\omega) + \lambda)$, standard results on convex integral

functionals (e.g. [Lev68, Theorem 1] or [IT69, Formula (7)]) imply that $(t^\star, \lambda^\star) \in \partial f(0,0)$ if and only $(t^\star, \lambda^\star) = \big(\nu(t^\star_\omega), \nu(\lambda^\star_\omega)\big)$ for measurable functions $t^\star_\omega, \lambda^\star_\omega : \Omega \to \mathbb{R}$ such that $(t^\star_\omega, \lambda^\star_\omega) \in \partial f_\omega(0,0)$ holds ν-a.s.

Now, for each $\omega \in \Omega$, we have that $(t^\star_\omega, \lambda^\star_\omega) \in \partial f_\omega(0,0)$ if and only if $\psi^\star(t \cdot g(\omega) + \lambda) \geq t^\star_\omega \cdot t + \lambda^\star_\omega \cdot \lambda$ for all $(t, \lambda) \in \mathbb{R}^2$. By considering $t = 0$, this implies that $\lambda^\star_\omega \in \partial\psi^\star(0) = \{x \in \mathbb{R} \mid \psi(x) = 0\}$, which is contained in either $\mathbb{R}_{\geq 0}$ or $\mathbb{R}_{\leq 0}$ since ψ is not 0 on a neighborhood of 0. Then since the integral of a function of constant sign is zero if and only if it is zero almost surely, we have that $(t^\star, 0) = \big(\nu(t^\star_\omega), \nu(\lambda^\star_\omega)\big)$ if and only if $\lambda^\star_\omega = 0$ holds ν-a.s. But $(t^\star_\omega, 0) \in \partial f_\omega(0,0)$ if and only if for all $t \in \mathbb{R}$ we have $t^\star_\omega \cdot t \leq \inf_\lambda \psi^\star(t \cdot g(\omega) + \lambda) = \psi^\star(0) = 0$, i.e. if and only if $t^\star_\omega = 0$.

Putting this together, we get that $\partial K_{g,\nu}(0) = \{t^\star \mid (t^\star, 0) \in \partial f(0,0)\} = \{\nu(t^\star_\omega) \mid (t^\star_\omega, 0) \in \partial f_\omega(0,0)\ \nu\text{-a.s.}\} = \{\nu(t^\star_\omega) \mid t^\star_\omega = 0\ \nu\text{-a.s.}\} = \{0\}$ and $K'_{g,\nu}(0) = 0$ as desired. □

Remark 5.5.39. If ϕ is 0 on a neighborhood of 1, then it is easy to show that $K_{g,\nu}$ is not differentiable at 0 unless g is ν-essentially constant. Thus, the above proposition shows that the following are equivalent: (i) $1 \notin \operatorname{int}\operatorname{dom}\{x \in \mathbb{R} \mid \phi(x) = 0\}$, (ii) for every g, continuity of $K_{g,\nu}$ at 0 implies differentiability at 0, (iii) $\mathrm{D}_\phi(\mu \parallel \nu) = 0$ for probability measures μ and ν if and only if $\mu = \nu$.

A similar (but simpler) proof shows that the following are equivalent: (i) ϕ strictly convex at 1, (ii) for every g, continuity of $t \mapsto I_{\psi^\star,\nu}(tg)$ at 0 implies differentiability at 0, and (iii) $\mathrm{D}_\phi(\mu \parallel \nu) = 0$ for finite measures μ and ν if and only if $\mu = \nu$. The similarity of the statements in both cases suggest there may be a common proof of the equivalences using more general techniques in convex analysis.

Thus, combining the previous two propositions and Proposition 5.5.3 computing the convex conjugate of the optimal lower bound function, we obtain the following theorem in the case of absolutely continuous measures.

Theorem 5.5.40. *Assume that* $1 \notin \operatorname{int}(\{x \in \mathbb{R} \mid \phi(x) = 0\})$. *Then for* $\nu \in \mathcal{M}^1$, $g \in L^1(\nu)$, *and* $M = X_g^1(\nu)$, *the following are equivalent:*

(i) *for all* $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$, $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \,\|\, \nu) = 0$ *implies* $\lim_{n\to\infty} \nu_n(g) = \nu(g)$.

(ii) $\partial K_{g,\nu}(0) = \{0\}$, *i.e.* $K_{g,\nu}$ *is differentiable at* 0 *and* $K'_{g,\nu}(0) = 0$.

(iii) g *is* (ϕ, ν)*-subexponential, i.e.* $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$.

Recall that measures which are not absolutely continuous are only interesting when $\phi'(\infty) < \infty$, as otherwise the ϕ-divergence is infinite on such measures. Since in this case, Lemma 5.5.18 shows that the space of subexponential functions coincides with the space of bounded functions, we also obtain the following theorem.

Theorem 5.5.41. *Assume that* $\phi'(\infty) < \infty$ *and that* $1 \notin \operatorname{int}(\{x \in \mathbb{R} \mid \phi(x) = 0\})$. *Then for* $\nu \in \mathcal{M}^1$, $g \in \mathcal{L}^1(\nu)$, *and* $M = \mathcal{X}_g^1$, *the following are equivalent:*

(i) *for all* $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$, $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \,\|\, \nu) = 0$ *implies* $\lim_{n\to\infty} \nu_n(g) = \nu(g)$.

(ii) $\{0\} = \partial K_{g,\nu,\perp}(0)$, *i.e.* $K_{g,\nu,\perp}$ *is differentiable at* 0 *with* $K'_{g,\nu,\perp}(0) = 0$.

(iii) $g \in \mathcal{L}^b(\Omega)$.

5.6 Optimal bounds relating ϕ-divergences and IPMs

In this section we generalize Theorem 5.5.12 on the optimal lower bound function for a single measure and function to the case of sets of measures and measurable functions.

5.6.1 On the choice of definitions

When considering a class of functions $\mathcal{G}$, there are several ways to define a lower bound of the divergence in terms of the mean deviation of functions in $\mathcal{G}$. The first one is to consider the IPM $d_{\mathcal{G}}$ induced by $\mathcal{G}$ and to ask for a function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(d_{\mathcal{G}}(\mu,\nu)\big)$ for all probability measures μ and ν, leading to the following definition of the optimal bound.

Definition 5.6.1. Let $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ be a non-empty set of measurable functions and let $N, M \subseteq \mathcal{M}^1$ be two sets of probability measures such that $\mathcal{G} \subseteq L^1(\nu)$ for every $\nu \in N \cup M$. The optimal lower bound function $\mathscr{L}_{\mathcal{G},N,M} : \mathbb{R} \to \overline{\mathbb{R}}_{\geq 0}$ is defined by

$$\mathscr{L}_{\mathcal{G},N,M}(\varepsilon) \stackrel{\text{def}}{=} \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; (\nu,\mu) \in N \times M \wedge \sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = \varepsilon\Big\}.$$

The definition generalizes to the case where $M : N \to 2^{\mathcal{M}^1}$ is a set-valued function by ranging over all pairs (μ,ν) with $\nu \in N$ and $\mu \in M(\nu)$.

Remark 5.6.2. Note that when $\mathcal{G}$ is closed under negation, then $\sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = d_{\mathcal{G}}(\mu,\nu)$ and $\mathscr{L}_{\mathcal{G},N,M}$ exactly quantifies the smallest value taken by the ϕ-divergence given a constraint on the IPM defined by $\mathcal{G}$.

An alternative definition, using the notations of Definition 5.6.1, is to consider the largest function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ for all $(\nu,\mu) \in N \times M$ and $g \in \mathcal{G}$. It is easy to see that this function can simply be expressed as

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon) = \inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} \mathscr{L}_{g,\nu,M}(\varepsilon) = \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; (\nu,\mu,g) \in N \times M \times \mathcal{G} \wedge \mu(g) - \nu(g) = \varepsilon\Big\}.$$

Observe that $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M} = \mathscr{L}_{\mathcal{G},N,M}$ when $\mathcal{G} = \{g\}$ or $\mathcal{G} = \{-g, g\}$. More generally, the goal of this section is to explore the relationship between $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$. In particular, we will

show that assuming a basic convexity condition on the set of measures M, both of these functions are non-decreasing on the non-negative reals, and can differ only on their (at most countably many) discontinuity points.

Lemma 5.6.3. *Let $N, M \subseteq \mathcal{M}^1$ be two sets of probability measures with $N \subseteq M$ and M convex. Then the functions $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ are non-negative and non-decreasing on the non-negative reals. The result holds more generally for $N \subseteq \mathcal{M}^1$ and $M : N \to 2^{\mathcal{M}^1}$ a set-valued function such that $\nu \in M(\nu)$ and $M(\nu)$ is convex for all $\nu \in N$.*

Proof. Let N and M be as in the lemma statement. It is sufficient to prove the result for $\mathscr{L}_{\mathcal{G},N,M}$, since the result for $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ follows from the fact that taking infima preserves sign and monotonicity.

Fix $0 \leq x \leq y$ and consider $\alpha > \mathscr{L}_{\mathcal{G},N,M}(y)$, so that by definition there exist $\mu \in M$ and $\nu \in N$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \alpha$ and $\sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = y$. Define $\mu' = x/y \cdot \mu + (1 - x/y) \cdot \nu$, which is a probability measure in M since $\nu \in N \subseteq M$ and M is convex. Then we have for every $g \in \mathcal{G}$ that $\mu'(g) - \nu(g) = x/y \cdot \big(\mu(g) - \nu(g)\big)$, and thus $\sup_{g\in\mathcal{G}}\big(\mu'(g) - \nu(g)\big) = x$. Furthermore, by convexity of $\mathrm{D}_{\phi,\nu}$ we have $\mathrm{D}_\phi(\mu' \parallel \nu) \leq x/y \cdot \mathrm{D}_\phi(\mu \parallel \nu) + (1 - x/y) \cdot \mathrm{D}_\phi(\nu \parallel \nu) < x/y \cdot \alpha \leq \alpha$ since $x/y \leq 1$. This implies that $\mathscr{L}_{\mathcal{G},N,M}(x) < \alpha$ and since α can be made arbitrarily close to $\mathscr{L}_{\mathcal{G},N,M}(y)$ that $\mathscr{L}_{\mathcal{G},N,M}(x) \leq \mathscr{L}_{\mathcal{G},N,M}(y)$. □

Remark 5.6.4. For convex sets of measures M and N and a single function $g \in L^1(\nu)$, a simple adaptation of Lemma 5.5.2 shows that $\mathscr{L}_{g,N,M}$ is convex, non-decreasing, and non-negative on the non-negative reals. Lemma 5.6.3 extends the latter two properties to the case of $\mathscr{L}_{\mathcal{G},N,M}$ for a set of functions $\mathcal{G}$, and in fact its proof shows that $\mathscr{L}_{\mathcal{G},N,M}(y)/y$ is non-decreasing, which is necessary for convexity. It would be interesting to characterize the set of $\mathcal{G}$, N, and M for which $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ are in fact convex.

Proposition 5.6.5. *Under the assumptions of Lemma 5.6.3, we have for every $\varepsilon > 0$ that*

$$\lim_{\varepsilon' \to \varepsilon^-} \inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') \leq \mathscr{L}_{\mathcal{G},N,M}(\varepsilon) \leq \inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon),$$

with equality if $\inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}$ is lower semicontinuous (equivalently left-continuous) at ε or if $\mathcal{G}$ is compact in the initial topology on $\mathcal{L}^0$ induced by the maps $\langle \mu - \nu, \cdot \rangle$ for $\mu \in M$ and $\nu \in N$.

Proof. Under the assumptions of Lemma 5.6.3 we have $\inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}$ and $\mathscr{L}_{\mathcal{G},N,M}$ are non-decreasing on the positive reals. Thus, we have

$$\begin{aligned}
&\inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon) \\
&\quad = \inf\left\{ D_\phi(\mu \parallel \nu) \,\middle|\, (\nu, \mu) \in N \times M \wedge \exists g \in \mathcal{G}, \mu(g) - \nu(g) = \varepsilon \right\} \\
&\quad \geq \inf\left\{ D_\phi(\mu \parallel \nu) \,\middle|\, (\nu, \mu) \in N \times M \wedge \sup_{g \in \mathcal{G}} \big(\mu(g) - \nu(g)\big) \geq \varepsilon \right\} && (5.25) \\
&\quad = \inf_{\varepsilon' \geq \varepsilon} \mathscr{L}_{\mathcal{G},N,M}(\varepsilon') = \mathscr{L}_{\mathcal{G},N,M}(\varepsilon) && (5.26) \\
&\quad = \inf\left\{ D_\phi(\mu \parallel \nu) \,\middle|\, (\nu, \mu) \in N \times M \wedge \forall \varepsilon' < \varepsilon \; \exists g \in \mathcal{G},\; \mu(g) - \nu(g) \geq \varepsilon' \right\} \\
&\quad \geq \sup_{\varepsilon' < \varepsilon} \inf\left\{ D_\phi(\mu \parallel \nu) \,\middle|\, (\nu, \mu) \in N \times M \wedge \exists g \in \mathcal{G}, \mu(g) - \nu(g) \geq \varepsilon' \right\} \\
&\quad = \sup_{\varepsilon' < \varepsilon} \inf\left\{ D_\phi(\mu \parallel \nu) \,\middle|\, (\nu, \mu, g) \in N \times M \times \mathcal{G} \wedge \mu(g) - \nu(g) \geq \varepsilon' \right\} \\
&\quad = \sup_{\varepsilon' < \varepsilon} \inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') = \lim_{\varepsilon' \to \varepsilon^-} \inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') && (5.27)
\end{aligned}$$

where Eq. (5.25) is since if there is $g \in \mathcal{G}$ with $\mu(g) - \nu(g) = \varepsilon$ then $\sup_{g \in \mathcal{G}} \mu(g) - \nu(g) \geq \varepsilon$, Eq. (5.26) is because $\mathscr{L}_{\mathcal{G},N,M}$ is non-decreasing, and Eq. (5.27) is because $\inf_{g \in G} \mathscr{L}_{g,N,M}$ is non-decreasing.

For the equality claims, since $\inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}$ is non-decreasing, lower semicontinuity at ε is equivalent to left-continuity, and $\lim_{\varepsilon' \to \varepsilon^-} \inf_{g \in \mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') = \inf_{g \in G} \mathscr{L}_{g,N,M}(\varepsilon)$ in this case. If $\mathcal{G}$ is compact in the claimed topology, then $\sup_{g \in \mathcal{G}} \big(\mu(g) - \nu(g)\big)$ is the supremum of the continuous function $\langle \mu - \nu, \cdot \rangle$

on the compact set $\mathcal{G}$, so that $\sup_{g\in\mathcal{G}}\big(\mu(g)-\nu(g)\big) = \max_{g\in\mathcal{G}}\big(\mu(g)-\nu(g)\big)$ and thus Eq. (5.25) is an equality. □

Thus, the functions $\inf_{g\in G}\mathscr{L}_{g,N,M}$ and $\mathscr{L}_{\mathcal{G},N,M}$ differ only at the (at most countably many) discontinuity points of the non-decreasing function $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$, where at those points we have $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M} \geq \mathscr{L}_{\mathcal{G},N,M}$. In particular, these two functions have the same lsc regularization and thus the same convex conjugate and biconjugate (recall that the lsc regularization of a function is the largest lsc function which lower bounds it pointwise). This will be useful in the next section where the optimal lower bound function will be described via its convex conjugate, which is easier to compute from the expression $\inf_{g\in G}\mathscr{L}_{g,N,M}$.

Another consequence of Proposition 5.6.5 is that $\inf_{g\in G}\mathscr{L}_{g,N,M}$ and $\mathscr{L}_{\mathcal{G},N,M}$ have the same generalized inverse. This generalized inverse is simply the optimal *upper bound* function, that is the smallest function U such that $\mu(g)-\nu(g) \leq U\big(\mathrm{D}_\phi(\mu \parallel \nu)\big)$ for all $(\mu,\nu,g)\in M\times N\times\mathcal{G}$, or equivalently such that $d_{\mathcal{G}}(\mu,\nu) \leq U\big(\mathrm{D}_\phi(\mu \parallel \nu)\big)$ for all $(\mu,\nu)\in M\times N$. In this language, any discontinuity of $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ corresponds to an interval on which U is constant, i.e. in which changing the value of the divergence does not change the largest possible value of $d_{\mathcal{G}}(\mu,\nu)$.

We conclude this section with two lemmas showing how the lower bound is preserved under natural transformations of the sets of functions $\mathcal{G}$ or measures M, N.

Lemma 5.6.6. *For every set $\mathcal{G}\subseteq\mathcal{L}^0(\Omega)$ and pair of measures $\mu,\nu\in\mathcal{X}_{\mathcal{G}}$, we have that*

$$\sup_{g\in\mathcal{G}}\big(\mu(g)-\nu(g)\big) = \sup_{g\in\overline{\mathrm{co}}\,\mathcal{G}}\big(\mu(g)-\nu(g)\big)$$

where $\overline{\mathrm{co}}\,\mathcal{G}$ is the $\sigma(\mathcal{Y}_{\mathcal{G}},\mathcal{X}_{\mathcal{G}})$-closed convex hull of $\mathcal{G}$.

Proof. We have $\mathcal{G}\subseteq\overline{\mathrm{co}}\,\mathcal{G}$, and furthermore since $\langle\mu-\nu,\ \cdot\ \rangle$ is a $\sigma(\mathcal{Y}_{\mathcal{G}},\mathcal{X}_{\mathcal{G}})$-continuous linear function

we have that the set $\{h \in \mathcal{Y}_{\mathcal{G}} \mid \langle \mu - \nu, h\rangle \leq \sup_{g\in\mathcal{G}}(\mu(g) - \nu(g))\}$ is convex, $\sigma(\mathcal{Y}_{\mathcal{G}}, \mathcal{X}_{\mathcal{G}})$-closed, and contains $\mathcal{G}$, and so also contains $\overline{\operatorname{co}}\,\mathcal{G}$. □

Lemma 5.6.7. *For every $g \in \mathcal{L}^0(\Omega)$, we have $\mathscr{L}_{g,\mathcal{X}_g^1,\mathcal{X}_g^1} = \mathscr{L}_{\mathrm{Id}_{\mathbb{R}},g_*\mathcal{X}_g^1,g_*\mathcal{X}_g^1}$ where $g_*\mathcal{X}_g^1 = \{g_*\nu \mid \nu \in \mathcal{X}_g^1\}$ is the set of probability measures on $\mathbb{R}$ obtained by pushing forward through g the probability measures $\nu \in \mathcal{M}^1(\Omega)$ integrating g. Furthermore, for every $\nu \in \mathcal{M}^1$ and $g \in L^1(\nu)$ we have that $\mathscr{L}_{g,\nu} = \mathscr{L}_{\mathrm{Id}_{\mathbb{R}},g_*\nu,g_*X_g^1(\nu)}$.*

Proof. We first prove the main claim. As in Example 5.3.4, we have for every $\mu, \nu \in \mathcal{X}_g^1$ that $\mu(g)-\nu(g) = \int \mathrm{Id}_{\mathbb{R}}\, dg_*\mu - \int \mathrm{Id}_{\mathbb{R}}\, dg_*\nu$, so it suffices to show for every $\mu_0, \nu_0 \in \mathcal{X}_g^1$ the existence of $\mu, \nu \in \mathcal{X}_g^1$ with $g_*\mu = g_*\mu_0$, $g_*\nu = g_*\nu_0$, and $\mathrm{D}_\phi(g_*\mu_0 \parallel g_*\nu_0) = \mathrm{D}_\phi(\mu \parallel \nu) \leq \mathrm{D}_\phi(\mu_0 \parallel \nu_0)$.

For this, write $\xi = \frac{1}{2}(\mu_0+\nu_0)$ so that $\mu_0, \nu_0 \ll \xi$ and $\xi \in \mathcal{X}_g^1$, and define the measures $\mu, \nu \in \mathcal{M}_c^1(\xi)$ by $\frac{d\mu}{d\xi} = \frac{dg_*\mu_0}{dg_*\xi} \circ g$ and $\frac{d\nu}{d\xi} = \frac{dg_*\nu_0}{dg_*\xi} \circ g$ (note that these are just the conditional expectations of $\frac{d\mu_0}{d\xi}$ and $\frac{d\nu_0}{d\xi}$ with respect to g). It remains to show that μ and ν have the desired properties, for which we first note that for every (Borel) measurable function $h : \mathbb{R}^3 \to \mathbb{R} \cup \{+\infty\}$ we have

$$\begin{aligned}\int h\Big(\frac{d\mu}{d\xi}, \frac{d\nu}{d\xi}, g\Big)\, d\xi &= \int h\Big(\frac{dg_*\mu_0}{dg_*\xi} \circ g, \frac{dg_*\nu_0}{dg_*\xi} \circ g, g\Big)\, d\xi \\ &= \int h\Big(\frac{dg_*\mu_0}{dg_*\xi}, \frac{dg_*\nu_0}{dg_*\xi}, \mathrm{Id}_{\mathbb{R}}\Big)\, dg_*\xi\,.\end{aligned}$$

Then taking $h(x, y, z) = x$ we get $\mu(\Omega) = \mu(\mathbf{1}_\Omega) = g_*\mu_0(\mathbf{1}_{\mathbb{R}}) = \mu_0(\mathbf{1}_\Omega) = 1$, and similarly by taking $h(x, y, z) = y$ we get $\nu(\Omega) = 1$. Taking $h(x, y, z) = x \cdot |z|$ we get $\mu(|g|) = \mu_0(|g|) < \infty$ so that $\mu \in \mathcal{X}_g^1$, and similarly by taking $h(x, y, z) = y \cdot |z|$ we get $\nu(|g|) = \nu_0(|g|) < \infty$ and $\nu \in \mathcal{X}_g^1$. Finally, as in Remark 5.4.3, taking $h(x, y, z) = y \cdot \phi(x/y)$ if $y \neq 0$ and $h(x, y, z) = x \cdot \phi'(\infty)$ if $y = 0$ gives $\mathrm{D}_\phi(\mu \parallel \nu) = \mathrm{D}_\phi(g_*\mu_0 \parallel g_*\nu_0)$, and furthermore Jensen's inequality implies that $\mathrm{D}_\phi(\mu \parallel \nu) \leq \mathrm{D}_\phi(\mu_0 \parallel \nu_0)$ since h is convex.

The furthermore claim is analogous after noting that since we have $\mu \ll \nu$ for every $\mu \in X_g^1(\nu)$ we

can take $\xi = \nu_0 = \nu$. □

5.6.2 Derivation of the bound

In this section we give our main results computing optimal lower bounds on a ϕ-divergence given an integral probability metric. Note that from Section 5.6.1, the optimal lower bound is simply the infimum of the optimal lower bound $\mathscr{L}_{g,\nu}$ for each $g \in \mathcal{G}$ and $\nu \in N$. Since $\mathscr{L}^\star_{g,\nu} = K_{g,\nu}$ by Proposition 5.5.3, and given the order-reversing property of convex conjugacy, it is natural to consider the the best *upper bound* on $K_{g,\nu}$ which holds *uniformly* over all $g \in \mathcal{G}$ and $\nu \in N$. Formally, we have the following definition.

Definition 5.6.8. For a set of measurable functions $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ and a set of measures $N \subseteq \mathcal{M}^1$, we write $K_{\mathcal{G},N}(t) \stackrel{\text{def}}{=} \sup\{K_{g,\nu}(t) \mid (g,\nu) \in \mathcal{G} \times N\}$, and $K_{\mathcal{G},N,\perp}(t) \stackrel{\text{def}}{=} \sup\{K_{g,\nu,\perp}(t) \mid (g,\nu) \in \mathcal{G} \times N\}$.

Note that $K_{\mathcal{G},N}$ and $K_{\mathcal{G},N,\perp}$ are convex and lower semicontinuous as suprema of convex and lower semicontinuous functions. Furthermore, as alluded to before Definition 5.6.8, we expect $K_{\mathcal{G},N}$ to be equal to the conjugate of the optimal lower bound functions. This is stated formally in the following theorem which also gives a sufficient condition under which the optimal lower bound functions are convex and lower semicontinuous (see also Remark 5.6.10 below).

Theorem 5.6.9. *Let $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ be a non-empty set of functions and $N \subseteq \mathcal{X}^1_{\mathcal{G}}$ be a non-empty set of probability measures integrating all functions in $\mathcal{G}$. Then, we have for all $\varepsilon \geq 0$ that*

$$\mathscr{L}^\star_{\mathcal{G},N,\mathcal{X}^1_{\mathcal{G}}}(\varepsilon) = \left(\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,\mathcal{X}^1_{\mathcal{G}}}\right)^\star(\varepsilon) = K_{\mathcal{G},N,\perp}(\varepsilon)\,, \tag{5.28}$$

and similarly for absolutely continuous measures,

$$\mathscr{L}^\star_{\mathcal{G},N,X^1_{\mathcal{G}}(\cdot)}(\varepsilon) = \left(\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1_{\mathcal{G}}(\cdot)}\right)^\star(\varepsilon) = K_{\mathcal{G},N}(\varepsilon)\,. \tag{5.29}$$

Proof. It follows from Proposition 5.6.5 and the discussion following it that the convex conjugates of $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,\mathcal{X}_{\mathcal{G}}^1}$ and $\mathscr{L}_{\mathcal{G},N,\mathcal{X}_{\mathcal{G}}^1}$ coincide on the non-negative reals, which justifies the first equality in (5.28). For the second equality,

$$\left(\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,\mathcal{X}_{\mathcal{G}}^1}\right)^\star(\varepsilon) = \left(\inf_{(g,\nu)\in\mathcal{G}\times N} \mathscr{L}_{g,\nu,\mathcal{X}_{\mathcal{G}}^1}\right)^\star(\varepsilon) = \sup_{\substack{g\in G\\ \nu\in N}} \mathscr{L}^\star_{g,\nu,\mathcal{X}_{\mathcal{G}}^1}(\varepsilon) = \sup_{\substack{g\in G\\ \nu\in N}} K_{g,\nu,\perp}(\varepsilon) = K_{\mathcal{G},N,\perp}(\varepsilon)\,,$$

where we used successively the definition of $\mathscr{L}_{g,N,\mathcal{X}_{\mathcal{G}}^1}$, the fact that $(\inf_{\alpha\in A} f_\alpha)^\star = \sup_{\alpha\in A} f_\alpha^\star$ for any collection $(f_\alpha)_{\alpha\in A}$ of functions, Proposition 5.5.3 and Remark 5.5.4, and Definition 5.6.8. □

Remark 5.6.10. Theorem 5.6.9 computes the conjugate of the optimal lower bound functions, but if this function is not convex or lsc, it is also useful to discuss what we can say about $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N}$ itself. First, if $\mathscr{L}_{g,\nu}$ (resp. $\mathscr{L}_{g,\nu,\perp}$) is lower semicontinuous for each $g \in \mathcal{G}$ and $\nu \in N$ (e.g. when $\phi'(\infty) = \infty$ and $\mathcal{G} \subseteq S_\star^\phi(\nu)$ for all $\nu \in N$ by Corollary 5.5.35), then

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,\mathcal{X}_{\mathcal{G}}^1}(\varepsilon) = \inf_{(g,\nu)\in\mathcal{G}\times N} \mathscr{L}_{g,\nu,\perp}(\varepsilon) = \inf_{(g,\nu)\in\mathcal{G}\times N} K^\star_{g,\nu,\perp}(\varepsilon)\,,$$

and similarly for absolutely continuous measures. Furthermore, if we also know that the function $\inf_{(g,\nu)\in\mathcal{G}\times N} K^\star_{g,\nu,\perp}$ is itself convex and lsc, then

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,\mathcal{X}_{\mathcal{G}}^1}(\varepsilon) = \mathscr{L}_{\mathcal{G},N,\mathcal{X}_{\mathcal{G}}^1}(\varepsilon) = K^\star_{\mathcal{G},N,\perp}(\varepsilon)\,,$$

and similarly for absolutely continuous measures.

Similarly to Corollary 5.5.14, we give in the following corollary an "operational" restatement of Theorem 5.6.9 emphasizing the duality between upper bounds on $K_{\mathcal{G},N}$ and lower bounds on $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of $d_{\mathcal{G}}(\mu,\nu)$.

Corollary 5.6.11. *Let $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ be a non-empty set of measurable functions and let $N \subseteq \mathcal{X}_{\mathcal{G}}^1$ be a non-empty set of probability measures. Then for every convex and lower semicontinuous function $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$,*

the following are equivalent:

(i) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$ *for all* $\nu \in N$ *and* $\mu \in \mathcal{M}^1$ *(resp.* $\mathcal{M}_c^1(\nu)$*) integrating* $\mathcal{G}$*.*

(ii) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(|\mu(g) - \nu(g)|)$ *for all* $g \in \mathcal{G}$*,* $\nu \in N$*,* $\mu \in \mathcal{M}^1$ *(resp.* $\mathcal{M}_c^1(\nu)$*) integrating* $\mathcal{G}$*.*

(iii) $K_{g,\nu,\perp}(t) \leq L^\star(|t|)$ *(resp.* $K_{g,\nu}(t) \leq L^\star(|t|)$*) for all* $t \in \mathbb{R}$*,* $g \in \mathcal{G}$*, and* $\nu \in N$*.*

Proof. If $\mathcal{G}$ is closed under negation then the result is immediate from Theorem 5.6.9 since then $\sup_{g\in G} \mu(g) - \nu(g) = d_{\mathcal{G}}(\mu,\nu)$ and $K_{\mathcal{G},N}(t) = K_{\mathcal{G},N}(-t)$. For the general case, the result follows by applying Theorem 5.6.9 to $\mathcal{G}' \stackrel{\text{def}}{=} \mathcal{G} \cup -\mathcal{G}$ where $-\mathcal{G} \stackrel{\text{def}}{=} \{-g \mid g \in \mathcal{G}\}$, for which $K_{\mathcal{G}',N}(t) = \max\{K_{\mathcal{G},N}(t), K_{\mathcal{G},N}(-t)\}$. □

Example 5.6.12 (Subgaussian functions)**.** For the Kullback–Leibler divergence, [BLM13, Lemma 4.18] shows that $\mathrm{KL}(\mu \parallel \nu) \geq \frac{1}{2} d_{\mathcal{G}}(\mu,\nu)^2$ for all $\mu \in \mathcal{M}^1$ if and only if $\ln \int e^{t(g-\nu(g))}\, d\nu \leq t^2/2$ for all $g \in \mathcal{G}$ and $t \in \mathbb{R}$. Such a quadratic upper bound on the log moment-generating function is one of the characterizations of the so-called *subgaussian* functions, which contain as a special case the class of bounded functions by Hoeffding's lemma [Hoe63] (see also Example 5.6.31). Corollary 5.6.11 recovers this result by considering the (self-conjugate) function $L : t \mapsto t^2/2$, thus showing that Pinsker's inequality generalize to all subgaussian functions.

Theorem 5.6.9 generalizes this further to an arbitrary ϕ-divergence, showing that a subset $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ of measurable functions satisfies $\mathrm{D}_\phi(\mu \parallel \nu) \geq \frac{1}{2} d_{\mathcal{G}}(\mu,\nu)^2$ for all $\mu \in \mathcal{M}^1$ if and only if $K_{g,\nu}(t) \leq t^2/2$ for all $g \in \mathcal{G}$ and $t \in \mathbb{R}$. By analogy, we refer to functions whose cumulant generating function admits such a quadratic upper bound as *ϕ-subgaussian* functions.

Example 5.6.13. Recall from Example 5.5.15 that the χ^2-divergence given by $\phi(x) = (x-1)^2 + \delta_{\mathbb{R}_{\geq 0}}(x)$

satisfies

$$\psi^\star(x) = \begin{cases} x^2/4 & x \geq -2 \\ -1 - x & x < -2 \end{cases}$$

and $K_{g,\nu}(t) \leq \inf_\lambda \int (tg + \lambda)^2/4 \, d\nu = t^2 \operatorname{Var}_\nu(g)/4$, showing that the class of χ^2-subgaussian functions (see Example 5.6.12) includes all those with bounded variance.

Example 5.6.14. As a step towards understanding the Wasserstein distance, Bolley and Villani [BV05] define a "weighted total variation distance" between probability measures μ and ν as $\int g \, d|\mu - \nu|$ for some non-negative measurable function $g \in \mathcal{L}^0(\Omega)$, and their main result [BV05, Theorem 2.1] bounds this weighted total variation in terms of the KL divergence.

We rederive their result by noting that the g-weighted total variation is $d_{g\mathcal{B}}(\mu, \nu)$ for $g\mathcal{B} = \{g \cdot b \mid b \in \mathcal{B}\}$ where $\mathcal{B}$ is the set of measurable functions taking values in $[-1, 1]$, so that it suffices by Theorem 5.6.9 to upper bound $K_{g \cdot b, \nu}(t)$ for each $b \in \mathcal{B}$ in terms of $\ln \int e^g \, d\nu$ or $\ln \int e^{g^2} \, d\nu$. But since $g \geq 0$, we have $g \cdot b \leq |g| = g$ and we conclude by using the fact that finiteness of $\ln \int e^h \, d\nu$ (resp. $\ln \int e^{h^2} \, d\nu$) implies a quadratic upper bound bound on the centered log-moment generating function $K_{h,\nu}(t)$ for $|t| \leq 1/4$ (resp. all $t \in \mathbb{R}$) for any non-negative function h (see e.g. [Ver18, Propositions 2.5.2 and 2.7.1]).

Finally, we show that when we take $N = \mathcal{X}^1_{\mathcal{G}} = M$, that is, we want a lower bound L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_{\mathcal{G}}(\mu, \nu))$ for all probability measures μ and ν, we no longer need to distinguish between absolutely continuous and non-absolutely continuous measures for the best convex lsc bound. Intuitively, this is because it is always possible to approximate a non-absolutely continuous measure by continuous ones.

Theorem 5.6.15. *Let $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ be a non-empty set of measurable functions. Then*

$$\mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},x^1_{\mathcal{G}}} = \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)} = K^{\star}_{\mathcal{G},x^1_G}(|\cdot|)\,.$$

In other words, the following are equivalent for every convex lsc $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$:

(i) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$ *for all $\mu,\nu \in \mathcal{M}^1$ integrating $\mathcal{G}$.*

(ii) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(|\mu(g) - \nu(g)|)$ *for all $g \in \mathcal{G}$ and $\mu,\nu \in \mathcal{M}^1$ integrating $\mathcal{G}$.*

(iii) $K_{g,\nu}(t) \leq L^\star(|t|)$ *for all $t \in \mathbb{R}$, $g \in \mathcal{G}$, and $\nu \in \mathcal{M}^1$ integrating $\mathcal{G}$.*

Proof. We first show that $\mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},x^1_{\mathcal{G}}} = \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}$. The $\leq$ direction is immediate since $X^1_{\mathcal{G}}(\nu) \subseteq \mathcal{X}^1_{\mathcal{G}}$, so we need to show the other direction. Given any $\mu,\nu \in \mathcal{X}^1_{\mathcal{G}}$ and $\delta \in [0,1]$ let $\nu_\delta = (1-\delta)\cdot\nu + \delta\cdot\mu$ so that $\nu_\delta \in \mathcal{X}^1_{\mathcal{G}}$. Then for each $\delta \in [0,1]$ we have that

$$d_{\mathcal{G}}(\mu,\nu_\delta) = (1-\delta)d_{\mathcal{G}}(\mu,\nu) \quad \text{and} \quad \mathrm{D}_\phi(\mu \parallel \nu_\delta) \leq (1-\delta)\,\mathrm{D}_\phi(\mu \parallel \nu) \leq \mathrm{D}_\phi(\mu \parallel \nu)\,,$$

where the equality is because $\mu(g) - \nu_\delta(g) = (1-\delta)(\mu(g) - \nu(g))$ for all $g \in \mathcal{G}$, and where the inequalities are by convexity and-negativity of $\mathrm{D}_\phi(\mu \parallel \cdot\,)$. Furthermore, for $\delta \in (0,1]$ we have that $\mu \ll \nu_\delta$ and so $\mu \in X^1_{\mathcal{G}}(\nu_\delta)$, and thus for all $\delta \in (0,1]$

$$\mathrm{D}_\phi(\mu \parallel \nu) \geq \mathrm{D}_\phi(\mu \parallel \nu_\delta) \geq \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}(d_{\mathcal{G}}(\mu,\nu_\delta)) = \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}\big((1-\delta)d_{\mathcal{G}}(\mu,\nu)\big).$$

Then by lower semicontinuity of $\mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}$ we get that

$$\mathrm{D}_\phi(\mu \parallel \nu) \geq \lim_{\delta\to 0^+} \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}\big((1-\delta)d_{\mathcal{G}}(\mu,\nu)\big) \geq \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}(d_{\mathcal{G}}(\mu,\nu))\,.$$

In particular, $\mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}$ is a convex lsc lower bound on $d_{\mathcal{G}}(\mu,\nu)$ in terms of $\mathrm{D}_\phi(\mu \parallel \nu)$ for all $\mu,\nu \in \mathcal{X}^1_{\mathcal{G}}$, establishing $\mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},x^1_{\mathcal{G}}} \geq \mathscr{L}^{\star\star}_{\mathcal{G},x^1_{\mathcal{G}},X^1_{\mathcal{G}}(\cdot)}$ as desired. The remaining claims follow from Theorem 5.6.9. □

5.6.3 Application to bounded functions and the total variation

In this section, we consider the problem of lower bounding the ϕ-divergence by a function of the total variation distance. Though it is a well-studied problem and most of the results we derive are already known, we consider this case to demonstrate the applicability of the results obtained in Section 5.6.2. In Section 5.6.3, we study Vajda's problem [Vaj72]: obtaining the best lower bound of the ϕ-divergence by a function of the total variation distance, and in Section 5.6.3 we show how to obtain quadratic relaxations of the best lower bound as in Pinsker's inequality and Hoeffding's lemma. Note that following the conventions of the literature on this problem, we actually formulate everything in terms of the L^1 distance, which is equal to twice the total variation distance as defined in the rest of this thesis.

Vajda's problem

The Vajda problem [Vaj72] is to quantify the optimal relationship between the ϕ-divergence and the total variation, that is to compute the function

$$\begin{aligned}\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) &= \inf\left\{\mathrm{D}_\phi(\mu \parallel \nu) \,\middle|\, (\mu,\nu) \in \mathcal{M}^1 \times \mathcal{M}^1 \wedge 2d_{\mathrm{TV}}(\mu,\nu) = \varepsilon\right\} \\ &= \inf\left\{\mathrm{D}_\phi(\mu \parallel \nu) \,\middle|\, (\mu,\nu) \in \mathcal{M}^1 \times \mathcal{M}^1 \wedge d_{\mathcal{B}}(\mu,\nu) = \varepsilon\right\}\end{aligned}$$

where $\mathcal{B}$ is the set of measurable functions $\Omega \to [-1,1]$. In this section, we use Theorem 5.6.15 to give for an arbitrary ϕ an expression for the Vajda function as the convex conjugate of a natural geometric quantity associated with the function $\psi^\star$, the inverse of its *sublevel set volume function*, which we call the *height-for-width* function.

Definition 5.6.16. The *sublevel set volume* function $\mathrm{sls}_{\psi^\star} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ maps $h \in \mathbb{R}$ to the Lebesgue measure of the sublevel set $\{x \in \mathbb{R} \mid \psi^\star(x) \leq h\}$. Since $\psi^\star$ is convex and inf-compact, the sublevel sets are compact intervals and their Lebesgue measure is simply their length.

The *height-for-width* function $\mathrm{hgt}_{\psi^\star} : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$ is the (right) inverse of the sublevel set volume function given by $\mathrm{hgt}_{\psi^\star}(w) = \inf\left\{h \in \overline{\mathbb{R}} \;\middle|\; \mathrm{sls}_{\psi^\star}(h) \geq w\right\}$.

To understand this definition, note that since $\psi^\star$ is defined on $\mathbb{R}$, the sublevel set volume function can be interpreted as giving for each height h the length of longest horizontal line segment that can be placed in the epigraph of $\psi^\star$ but no higher than h. The inverse, the height-for-width function, asks for the minimal height at which one can place a horizontal line segment of length w in the epigraph of $\psi^\star$. See Fig. 5.1 for an illustration of this in the case of $\psi^\star(x) = e^x - x - 1$, corresponding to the Kullback–Leibler divergence.

The following lemma shows that the height-for-width function can be equivalently formulated as the optimal value of a simple convex optimization problem.

Lemma 5.6.17. *For all* $w \in \mathbb{R}_{\geq 0}$, $\mathrm{hgt}_{\psi^\star}(w) = \inf_{\lambda \in \mathbb{R}} \max\left\{\psi^\star(\lambda + w/2), \psi^\star(\lambda - w/2)\right\}$. *Furthermore, if for* $w > 0$ *there exists* λ_w *such that* $\psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$, *then* $\mathrm{hgt}_{\psi^\star}(w) = \psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$.

Proof. For every $w \geq 0$, define the function $h_w : \lambda \mapsto \max\{\psi^\star(\lambda - w/2), \psi^\star(\lambda + w/2)\}$ which is the supremum of two convex inf-compact functions with overlapping domain, and so is itself proper, convex, and inf-compact. In particular, h_w achieves its global minimum $y_w \in \mathbb{R}$, where by definition and convexity of $\psi^\star$ we have y_w is the smallest number such that there exists an interval $[\lambda - w/2, \lambda + w/2]$ of length w such that $\psi^\star([\lambda - w/2, \lambda + w/2]) \subseteq (-\infty, y_w]$, and thus $y_w = \inf\left\{x \in \overline{\mathbb{R}} \;\middle|\; \mathrm{sls}_{\psi^\star}(x) \geq w\right\} = \mathrm{hgt}_{\psi^\star}(w)$ as desired.

For the remaining claim, consider $w > 0$ for which there is $\lambda_w \in \mathbb{R}$ such that $\psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$. By convexity of $\psi^\star$ we have for every $\lambda < \lambda_w$ that $\psi^\star(\lambda - w/2) \geq \psi^\star(\lambda_w - w/2)$, and analogously for every $\lambda > \lambda_w$ that $\psi^\star(\lambda + w/2) \geq \psi^\star(\lambda_w + w/2)$. Thus for every λ we have $\max\{\psi^\star(\lambda -$

$w/2), \psi^\star(\lambda + w/2)\} \geq \min\{\psi^\star(\lambda_w - w/2), \psi^\star(\lambda_w + w/2)\} = \psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$, so the result follows from the main claim. □

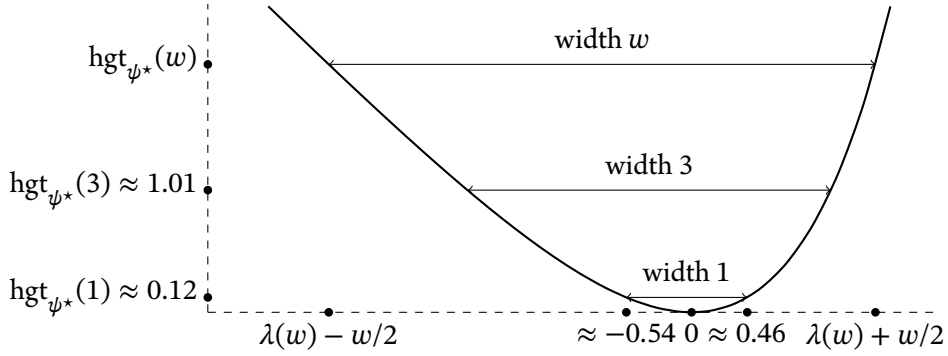

Figure 5.1: Illustration of height-for-width function for $\psi^\star(x) = e^x - x - 1$

Example 5.6.18. For the case of the KL divergence for which $\psi^\star(w) = e^w - w - 1$, one can compute that $\psi^\star(\lambda(w) + w/2) = \psi^\star(\lambda(w) - w/2)$ for $\lambda(w) = -\ln \frac{e^{w/2} - e^{-w/2}}{w} = -\ln \frac{2\sinh(w/2)}{w}$, so that $\mathrm{hgt}_{\psi^\star}(w) = -1 + \frac{w}{2}\coth\frac{w}{2} + \ln\frac{2\sinh(w/2)}{w}$.

The duality result of Theorem 5.6.15 computes the biconjugate of the optimal bound $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}$, so we first prove that this function is convex and lsc.

Lemma 5.6.19. *Let M the set of probability measures on the set $\{-1, 1\}$ with the discrete σ-algebra. Then $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$ is convex and lower semicontinuous. In particular $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) = K^\star_{\mathcal{B},\mathcal{M}^1}(\varepsilon)$ for $\varepsilon \geq 0$.*

Proof. By Theorem 5.6.15 we have that $\mathscr{L}^\star_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = K_{\mathcal{B},\mathcal{M}^1}$, so the in particular statement follows immediately from the main claim. The main claim, that $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$ is convex and lower semicontinuous, is well-known and can easily be derived using the methods of e.g. [Vaj72], but we include a proof here in our language for completeness and to illustrate how it could be generalized beyond the total variation.

Note that the set $\mathcal{B} = [-1,1]^\Omega \cap \mathcal{L}^0(\Omega)$ is convex, and furthermore is $\sigma(\mathcal{L}^b(\Omega), \mathcal{M})$-compact by the Banach–Alaoglu theorem, and so by the Krein–Milman theorem $\mathcal{B}$ is the $\sigma(\mathcal{L}^b(\Omega), \mathcal{M})$-closed convex hull of its extreme points $\mathrm{ext}(\mathcal{B}) = \{-1,1\}^\Omega \cap \mathcal{L}^0(\Omega)$ the set of measurable $\{-1,1\}$-valued functions. Thus, Lemma 5.6.6 implies $d_{\mathcal{B}} = d_{\mathrm{ext}(\mathcal{B})}$, and so $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{ext}(\mathcal{B}),\mathcal{M}^1,\mathcal{M}^1}$.

We now prove that $\inf_{g\in\mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1}$ is convex and lsc, which by Proposition 5.6.5 also implies $\mathscr{L}_{\mathrm{ext}(\mathcal{B}),\mathcal{M}^1,\mathcal{M}^1} = \inf_{g\in\mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1}$ is convex and lsc. By Lemma 5.6.7, for each $g \in \mathrm{ext}(\mathcal{B})$ we have $\mathscr{L}_{g,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M_g,M_g}$ for $M_g = \{g_*\mu \,|\, \mu \in \mathcal{M}^1\}$. In particular, if g is constant this set is the singleton $M_g = \{\delta_{g(\Omega)}\}$, and if g is non-constant then it is exactly the set M of probability measures supported on $\{-1,1\}$. Thus, $\inf_{g\in\mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$.

Note that the set M with the total variation norm is homeomorphic to the unit interval $[0,1]$ via the linear map $p \mapsto p\cdot\delta_{\{1\}} + (1-p)\cdot\delta_{\{-1\}}$. Then the function $f : \mathbb{R}\times M^2 \to \overline{\mathbb{R}}$ given by $f\big(\varepsilon,(\mu,\nu)\big) = \mathrm{D}_\phi(\mu \,\|\, \nu) + \delta_{\{0\}}\big(\mu\big(\mathrm{Id}_{\{-1,1\}}\big) - \nu\big(\mathrm{Id}_{\{-1,1\}}\big) - \varepsilon\big)$ is jointly convex and lower semi-continuous, and hence since M is compact also inf-compact. Thus, by Lemma 5.3.8, the function $\mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M} = \inf_{(\mu,\nu)\in M^2} f\big(\,\cdot\,,(\mu,\nu)\big)$ is convex and inf-compact as desired. □

Lemma 5.6.19 implies that it suffices to compute $K_{\mathcal{B},\mathcal{M}^1}$.

Lemma 5.6.20. *$K_{\mathcal{B},\mathcal{M}^1}(t) = \mathrm{hgt}_{\psi^\star}(2t)$ for every $t \geq 0$.*

Proof. For $M = \{p\cdot\delta_{\{1\}} + (1-p)\cdot\delta_{\{-1\}} \mid p \in [0,1]\}$, we have by Lemma 5.6.19 and Theorem 5.6.15 that $K_{\mathcal{B},\mathcal{M}^1} = \mathscr{L}^\star_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}^\star_{\mathrm{Id}_{\{-1,1\}},M,M} = \sup_{\nu\in M} K_{\mathrm{Id}_{\{-1,1\}},\nu}$. For $p \in [0,1]$ we have $K_{\mathrm{Id}_{\{-1,1\}},p\cdot\delta_{\{1\}}+(1-p)\cdot\delta_{\{-1\}}} = \inf_{\lambda\in\mathbb{R}}\big(p\cdot\psi^\star(t+\lambda) + (1-p)\cdot\psi^\star(-t+\lambda)\big)$, so that

$$K_{\mathcal{B},\mathcal{M}^1}(t) = \sup_{p\in[0,1]} \inf_{\lambda\in\mathbb{R}}\big(p\cdot\psi^\star(\lambda+t) + (1-p)\cdot\psi^\star(\lambda-t)\big). \tag{5.30}$$

This mixed optimization problem is convex in λ for each p and linear in p for each $\lambda \in \mathbb{R}$, and the

interval $[0,1]$ is compact, so by the Sion minimax theorem [Sio58] we can swap the supremum and infimum to get

$$\begin{aligned} K_{\mathcal{B},\mathcal{M}^1}(t) &= \inf_{\lambda\in\mathbb{R}} \sup_{p\in[0,1]} \left(p\cdot\psi^\star(\lambda+t) + (1-p)\cdot\psi^\star(\lambda-t)\right) \\ &= \inf_{\lambda\in\mathbb{R}} \max\{\psi^\star(\lambda+t), \psi^\star(\lambda-t)\} \end{aligned}$$

so the claim follows from Lemma 5.6.17. □

Example 5.6.21. For the Kullback–Leibler divergence, since $K_{g,\nu}(t) = \ln \nu\left(e^{t(g-\nu(g))}\right)$ as in Example 5.4.18, Lemma 5.6.20 and Example 5.6.18 imply that the optimal bound on the cumulant generating function of a random variable g with $\nu(g) = 0$ and $m \le g \le M$ ν-a.s. is $\ln \nu\left(e^{tg}\right) \le \operatorname{hgt}_{\psi^\star}[(M-m)t] = -1 + \frac{M-m}{2}\coth\frac{M-m}{2} + \ln\frac{2\sinh((M-m)t/2)}{t}$. This is a refinement of Hoeffding's lemma, which gives the upper bound of $(M-m)^2t^2/8$, which we will also derive as consequence of a general quadratic relaxation on the height function in Example 5.6.31.

Corollary 5.6.22. *$\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) = \operatorname{hgt}^\star_{\psi^\star}(\varepsilon/2)$ for all $\varepsilon \ge 0$. In particular, if $\operatorname{hgt}_{\psi^\star}$ is differentiable then $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(2\operatorname{hgt}'_{\psi^\star}(x)) = x\operatorname{hgt}'_{\psi^\star}(x) - \operatorname{hgt}_{\psi^\star}(x)$.*

Proof. The main claim is immediate from Lemmas 5.6.19 and 5.6.20, and the supplemental claim follows from the explicit expression for the convex conjugate for differentiable functions. □

Example 5.6.23. For the Kullback–Leibler divergence, using Example 5.6.18, the supplemental claim of Corollary 5.6.22 applied to $x = 2t$ gives $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(V(t)) = \ln\frac{t}{\sinh t} + t\coth t - \frac{t^2}{\sinh^2 t}$ for $V(t) = 2\coth t - \frac{t}{\sinh^2 t} - 1/t$, which is exactly the formula derived by [FHT03].

Remark 5.6.24. Corollary 5.6.22 shows that lower bounds on the ϕ-divergence in terms of the total variation are equivalent to upper bounds on the height-for-width function $\operatorname{hgt}_{\psi^\star}$, equivalently to lower

bounds on the sublevel set volume function of $\psi^\star$. The complementary problem of obtaining upper bounds on the sublevel set volume function is of interest in harmonic analysis due to its connection to studying oscillatory integrals (e.g. [Ste93, Chapter 8, Proposition 2] and [CCW99, §1-2]), and it would be interesting to see if techniques from that literature could be applied in this context.

Remark 5.6.25. Since the L^1 distance $d_{\mathcal{B}}(\mu, \nu)$ is symmetric in terms of μ and ν, the optimal lower bound on $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of $d_{\mathcal{B}}(\mu, \nu)$ is the same as the optimal lower bound on $\mathrm{D}_\phi(\nu \parallel \mu) = \mathrm{D}_{\phi^\dagger}(\mu \parallel \nu)$ for $\phi^\dagger = x\phi(1/x)$. By Corollary 5.6.22, this implies that $\mathrm{hgt}_{\psi^\star} = \mathrm{hgt}_{(\psi^\dagger)^\star}$ (note that this can also be derived directly from the definition).

Application to Pinsker-type inequalities

Corollary 5.6.22 implies that to obtain Pinsker-type inequalities, it suffices to upper bound the height function $\mathrm{hgt}_{\psi^\star}(t)$ by a quadratic function of t. In this section, we show such bounds under mild assumptions on $\psi^\star$, both rederiving optimal Pinsker-type inequalities for the Kullback–Leibler divergence and α-divergences for $-1 \le \alpha \le 2$ due to Gilardoni [Gil10], and deriving new but not necessarily optimal Pinsker-type inequalities for all $\alpha \in \mathbb{R}$. We proceed by giving two arguments approximating the minimizer $\lambda(t)$ in the optimization problem defining the height (Lemma 5.6.17), and an argument that works directly with the optimal $\lambda(t)$.

We begin with the crudest but most widely applicable bound.

Corollary 5.6.26. *If ϕ is twice differentiable on its domain and ϕ'' is monotone, then $\mathrm{hgt}_{\psi^\star}(t) \le t^2/(2\phi''(1))$ for all $t \ge 0$. Equivalently, for such ϕ we have that $\mathrm{D}_\phi(\mu \parallel \nu) \ge \frac{\phi''(1)}{8} \cdot d_{\mathcal{B}}(\mu, \nu)^2$ for all $\mu, \nu \in \mathcal{M}^1$.*

Proof. If $\phi''(1) = 0$, then the claim is trivial, so we assume that $\phi''(1) > 0$. If ϕ'' is non-decreasing, we have by Taylor's theorem that $\phi(x) \ge \frac{\phi''(1)}{2}(x-1)^2$ for $x \ge 1$, equivalently $\psi(x) \ge \frac{\phi''(1)}{2}x^2$ for

$x \geq 0$, so that $\psi^\star(x) \leq \frac{1}{2\phi''(1)} x^2$ for $x \geq 0$. Then $\mathrm{hgt}_{\psi^\star}(t) = \inf_{\lambda \in \mathbb{R}} \max\{\psi^\star(\lambda - t/2), \psi^\star(\lambda + t/2)\} \leq \max\{\psi^\star(0), \psi^\star(t)\} \leq t^2/(2\phi''(1))$. On the other hand, if ϕ'' is non-increasing, then analogously we have $\psi^\star(x) \leq \frac{1}{2\phi''(1)} x^2$ for $x \leq 0$, so that Then $\mathrm{hgt}_{\psi^\star}(t) = \inf_{\lambda \in \mathbb{R}} \max\{\psi^\star(\lambda - t/2), \psi^\star(\lambda + t/2)\} \leq \max\{\psi^\star(0), \psi^\star(-t)\} \leq t^2/(2\phi''(1))$. □

Example 5.6.27. Most of the standard ϕ-divergences satisfy the condition of Corollary 5.6.26, in particular the α-divergences given by $\phi_\alpha = \frac{x^\alpha - \alpha(x-1) - 1}{\alpha(\alpha-1)}$ have $\phi_\alpha''(x) = x^{\alpha-2}$ which is monotone for all α. As a result, we get for all α the (possibly suboptimal) Pinsker inequality $\mathrm{D}_{\phi_\alpha}(\mu \,\|\, \nu) \geq \frac{1}{8} \cdot d_{\mathcal{B}}(\mu, \nu)^2$ for all $\mu, \nu \in \mathcal{M}^1$. Such a bound appears to be new for $\alpha > 2$, but for $\alpha \in [-1, 2]$ Gilardoni [Gil10] established the better bound $\mathrm{D}_{\phi_\alpha}(\mu \,\|\, \nu) \geq \frac{1}{2} \cdot d_{\mathcal{B}}(\mu, \nu)^2$, extending the standard case of the Kullback–Leibler divergence $\alpha = 1$. We rederive this optimal constant for these divergences below, and also give general conditions under which such bounds hold.

Corollary 5.6.26 used the crude linear relaxation $-t/2 \leq \lambda(t) \leq t/2$. In the following Corollary, we derive a tighter Pinsker-type inequality by using a Taylor expansion of $\lambda(t)$.

Corollary 5.6.28. *Suppose that ϕ strictly convex and twice differentiable on its domain, thrice differentiable at 1 and that*

$$\frac{27\phi''(1)}{(3 - z\phi'''(1)/\phi''(1))^3} \leq \phi''(1+z)$$

for all $z \geq -1$. Then $\mathrm{hgt}_{\psi^\star}(t) \leq t^2/(8\phi''(1))$ for all $t \geq 0$, equivalently, for such ϕ we have $\mathrm{D}_\phi(\mu \,\|\, \nu) \geq \frac{\phi''(1)}{2} \cdot d_{\mathcal{B}}(\mu, \nu)^2$ for all $\mu, \nu \in \mathcal{M}^1$.

Remark 5.6.29. The Pinsker constant in Corollary 5.6.28 is best-possible, since if ϕ is twice-differentiable at 1, then Taylor's theorem gives the local expansion $\phi(x) = \phi''(1)/2 \cdot (x-1)^2 + o\big((x-1)^2\big)$, and thus the distributions $\mu_\varepsilon = (1/2 + \varepsilon/2, 1/2 - \varepsilon/2)$ and $\nu = (1/2, 1/2)$ on the set $\{0, 1\}$ have $d_{\mathcal{B}}(\mu_\varepsilon, \nu) = \varepsilon$ and $\mathrm{D}_\phi(\mu_\varepsilon \,\|\, \nu) = \phi''(1)/2 \cdot \varepsilon^2 + o(\varepsilon^2)$.

Proof. Under suitable regularity assumptions on ϕ and $\psi^\star$, one can easily show that the second order expansion of the function $\lambda(t)$ implicitly defined by $\psi^\star(\lambda(t)+t/2) = \psi^\star(\lambda(t)-t/2)$ is $L(t) = -\frac{ct^2}{24}$ for $c = (\psi^\star)'''(0)/(\psi^\star)''(0) = -\phi'''(1)/\phi''(1)^2$. Taking this as given, we show under the stated assumptions of the proposition that for $L(t) = -\frac{ct^2}{24}$ and $c = -\phi'''(1)/\phi''(1)^2$, we have that $\psi^\star(L(t) + st/2) \leq t^2/(8\phi''(1))$ for $s \in \{\pm 1\}$. Since both sides are 0 at 0, it thus suffices to show $\big(L'(t) + s/2\big)(\psi^\star)'(L(t) + st/2) \leq t/(4\phi''(1))$. Now, let $\lesseqgtr$ indicate $\leq$ if $L'(t) + s/2 \geq 0$ and $\geq$ if $L'(t) + s/2 \leq 0$. Since ϕ strictly convex implies $\psi' = ((\psi^\star)')^{-1}$ is strictly increasing, we thus have that this is equivalent to

$$L(t) + st/2 \lesseqgtr \psi'\left(\frac{t/(4\phi''(1))}{L'(t) + s/2}\right) \tag{5.31}$$

Write $z = \frac{t/(4\phi''(1))}{L'(t)+s/2} = \frac{t/(4\phi''(1))}{-ct/12+s/2}$ so that z has the same sign as $L'(t) + s/2$ and $t = \frac{6sz\phi''(1)}{3+cz\phi''(1)}$. Plugging this in and using the fact that $s^2 = 1$, we wish to show that

$$\frac{3z\phi''(1)(6 + cz\phi''(1))}{2(3 + cz\phi''(1))^2} - \psi'(z) \lesseqgtr 0 \tag{5.32}$$

for all z such that $t \geq 0$. The left hand side of Eq. (5.32) is 0 at 0, so since $z > 0$ implies $\lesseqgtr$ is $\leq$ and $z < 0$ implies $\lesseqgtr$ is $\geq$, it suffices to show that the derivative of the left-hand side of Eq. (5.32) with respect to z is non-positive for all z. This derivative is

$$\frac{27\phi''(1)}{(3 + cz\phi''(1))^3} - \psi''(z) = \frac{27\phi''(1)}{(3 - z\phi'''(1)/\phi''(1))^3} - \phi''(1 + z) \tag{5.33}$$

which since $\operatorname{dom}\psi \subseteq [-1, \infty)$ is non-positive for all z if and only if it is non-positive for all $z \geq -1$. □

Example 5.6.30 ([Gil10])**.** For the α-divergences, we have $\phi''_\alpha(x) = x^{\alpha-2}$, and $\phi'''_\alpha(x) = (\alpha-2)x^{\alpha-3}$ so that Corollary 5.6.28 is equivalent to the condition $\frac{27}{(3+(2-\alpha)z)^3} \leq (1+z)^{\alpha-2}$ for $z \geq -1$. Note that this is true for $z = 0$ for all α, and the derivative of $\frac{27(1+z)^{2-\alpha}}{(3+(2-\alpha)z)^3}$ with respect to z is $\frac{27(\alpha-2)(\alpha+1)z(1+z)^{1-\alpha}}{(3+(2-\alpha)z)^4}$. Thus, for $\alpha \in [-1, 2]$ the sign of the derivative is the opposite of the sign of z, and the condition holds

for all $z \geq -1$, recovering the result of Gilardoni [Gil10] as desired.

Example 5.6.31. For the case of the Kullback–Leibler divergence, Example 5.6.30 rederives Pinsker's inequality and Hoeffding's lemma.

Finally, we show that one can also obtain optimal Pinsker-type inequalities while arguing directly about the optimal $\lambda(t)$, for which we need the following lemma.

Lemma 5.6.32. *Suppose that $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a convex function continuously differentiable on (a, b) the interior of its domain with a unique global minimum and such that $\lim_{x\to a^+} f(x) = \infty = \lim_{x\to b^-} f(x)$. Then there is a continuously differentiable function $\lambda : (a-b, b-a) \to \mathbb{R}$ such that $\operatorname{hgt}_f(t) = f(\lambda(t) + t/2) = f(\lambda(t) - t/2)$ and*

$$\lambda'(t) = \frac{f'(\lambda(t) + t/2) + f'(\lambda(t) - t/2)}{2\left(f'(\lambda(t) - t/2) - f'(\lambda(t) + t/2)\right)} \tag{5.34}$$

$$\operatorname{hgt}_f'(t) = \frac{f'(\lambda(t) + t/2) f'(\lambda(t) - t/2)}{f'(\lambda(t) - t/2) - f'(\lambda(t) + t/2)}\,. \tag{5.35}$$

Proof. For each $t \in (a-b, b-a)$, the function $\lambda \mapsto f(\lambda+t/2)-f(\lambda-t/2)$ is continuously differentiable on its domain $(a + \frac{|t|}{2}, b - \frac{|t|}{2})$, with limits $-\infty$ and ∞. Thus, for all such t there exists λ satisfying the implicit equation $f(\lambda(t) + t/2) = f(\lambda(t) - t/2)$, which by Lemma 5.6.17 also defines $\operatorname{hgt}_f(t)$. Furthermore, the fact that f has a unique global minimum implies this function is strictly increasing in λ for each t, and thus the implicit function theorem guarantees the existence of the claimed continuously differentiable $\lambda(t)$.

Given the existence of $\lambda(t)$, we have by its definition that $\frac{d}{dt} f(\lambda(t) + t/2) = \frac{d}{dt} f(\lambda(t) - t/2)$, which implies by the chain rule the claimed value for $\lambda'(t)$, which since $\operatorname{hgt}_f'(t) = \frac{d}{dt} f(\lambda(t) + t/2)$ implies the claimed expressions for the derivative of hgt_f. □

Using the previous lemma, we obtain the same optimal Pinsker-type inequality as in Corollary 5.6.28

under related but incomparable assumptions.

Corollary 5.6.33. *If ϕ is strictly convex, has a positive second derivative on its domain, $1/\phi''$ is concave, and $\lim_{x\to\phi'(\infty)^-}\psi^\star(x) = \infty$ (e.g. if $\phi'(\infty) = \infty$), then $\mathrm{hgt}_{\psi^\star}(t) \le t^2/(8\phi''(1))$ for all $t \ge 0$. Equivalently, for such ϕ we have $\mathrm{D}_\phi(\mu \parallel \nu) \ge \frac{\phi''(1)}{2}\cdot d_{\mathcal{B}}(\mu,\nu)^2$ for all $\mu,\nu \in \mathcal{M}^1$.*

Proof. By standard results in convex analysis, the existence and positivity of ψ'' imply that $\psi^\star$ is itself twice differentiable (e.g. [HL93, Proposition 6.2.5] or [Gor91, Proposition 1.1]). Thus, by Lemma 5.6.32, it suffices to show that $\mathrm{hgt}'_{\psi^\star}(t) \le t/(4\phi''(1))$, or equivalently

$$\frac{(\psi^\star)'(\lambda(t)+t/2)(\psi^\star)'(\lambda(t)-t/2)}{(\psi^\star)'(\lambda(t)-t/2)-(\psi^\star)'(\lambda(t)+t/2)} \le \frac{t}{4\phi''(1)}. \tag{5.36}$$

Since $\psi^\star(\lambda(t)+t/2) = \psi^\star(\lambda(t)-t/2)$ and $\psi^\star$ has global minimum at 0, we have $\lambda(t)-t/2 \le 0$ and $\lambda(t)+t/2 \ge 0$, and $(\psi^\star)'(\lambda(t)-t/2) \le 0$ and $(\psi^\star)'(\lambda(t)+t/2) \ge 0$. Thus, we have that the left-hand side of Eq. (5.36) is half the harmonic mean of $(\psi^\star)'(\lambda(t)+t/2)$ and $-(\psi^\star)'(\lambda(t)-t/2)$, so it suffices by the arithmetic mean–harmonic mean inequality to prove

$$(\psi^\star)'(\lambda(t)+t/2)-(\psi^\star)'(\lambda(t)-t/2) \le \frac{t}{\phi''(1)}. \tag{5.37}$$

Since Eq. (5.37) holds when $t = 0$, it suffices to prove that

$$(1/2+\lambda'(t))\cdot(\psi^\star)''(\lambda(t)+t/2) + (1/2-\lambda'(t))\cdot(\psi^\star)''(\lambda(t)-t/2) \le \frac{1}{\phi''(1)}. \tag{5.38}$$

By the relationship between the second derivative of a function and the one of its conjugate (e.g. [HL93, Proposition 6.2.5]), this is equivalent to

$$\frac{1/2+\lambda'(t)}{\psi''\big((\psi^\star)'(\lambda(t)+t/2)\big)} + \frac{1/2-\lambda'(t)}{\psi''\big((\psi^\star)'(\lambda(t)-t/2)\big)} \le \frac{1}{\phi''(1)}. \tag{5.39}$$

Now, by Eq. (5.34), we have that $\lambda'(t) \in [-1/2, 1/2]$, so that by Jensen's inequality and the concavity

of $1/\psi''$, the left-hand side of Eq. (5.39) is at most

$$1/\psi''\Big((1/2+\lambda'(t))(\psi^\star)'(\lambda(t)+t/2)-(\lambda'(t)-1/2)(\psi^\star)'(\lambda(t)-t/2)\Big). \tag{5.40}$$

Finally, since by definition $\psi^\star(\lambda(t)+t/2)=\psi^\star(\lambda(t)-t/2)$, the term inside $1/\psi''$ in Eq. (5.40) is 0, so since $\psi(x)=\phi(1+x)$ we are done. □

Example 5.6.34. For the α-divergences, we have $1/\phi''_\alpha(x)=x^{2-\alpha}$ which is concave for $\alpha\in[1,2]$, so Corollary 5.6.33 applies for these divergences. Furthermore, by Remark 5.6.25, we can consider the reverse α-divergences with $\phi^\dagger_\alpha(x)=x\phi_\alpha(1/x)$ which has $1/(\phi^\dagger_\alpha)''(x)=x^{1+\alpha}$, which is concave for $\alpha\in[-1,0]$.

5.7 Discussion

Throughout this chapter, the ϕ-cumulant generating function has proved central in explicitating the relationship between ϕ-divergences and integral probability metrics. As a starting point, the identity $K_{g,\nu}=\mathscr{L}^\star_{g,\nu}$ (Theorem 5.5.12) expresses the cumulant generating function as the convex conjugate of the best lower bound of $\mathrm{D}_\phi(\mu\parallel\nu)$ in terms of $\mu(g)-\nu(g)$. This establishes a "correspondence principle" by which properties of the relationship between ϕ-divergences and integral probability metrics translate by duality into properties of the cumulant generating function, and vice versa. An advantage of this correspondence is that the function $K_{g,\nu}$, being expressed as the solution of a single-dimensional convex optimization problem (Definition 5.5.8), is arguably easier to evaluate and analyze than its counterpart $\mathscr{L}_{g,\nu}$, expressed as the solution to an infinite-dimensional optimization problem. Following Theorem 5.5.12, several results from this chapter can be seen as instantiations of this "correspondence principle" and we summarize some of them in Table 5.2.

Table 5.2: Several examples, proved in this chapter, of the dual correspondence between properties of the ϕ-cumulant generating function and properties of the relationship between the ϕ-divergence and mean deviations. Throughout, $\mu \in \mathcal{M}^1$, $g \in L^1(\nu)$, $B : \mathbb{R} \to \mathbb{R}$ is arbitrary, $E : \mathbb{R} \to \mathbb{R}$ is even, and $\mathcal{G} \subseteq \mathcal{L}^0$. Recall that $X_g^1(\nu)$ is the set of all probability measures $\mu \ll \nu$ and integrating g, and that $\mathcal{X}_{\mathcal{G}}^1$ is the set of all probability measures integrating all functions in $\mathcal{G}$.

Ref.	Property of the ϕ-cumulant generating function	Property of the ϕ-divergence
§5.5.1	$K_{g,\nu}(t) \leq B(t)$ for all $t \in \mathbb{R}$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq B^\star\big(\mu(g) - \nu(g)\big)$ for all $\mu \in X_g^1(\nu)$
§5.5.2	$0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(\lvert\mu(g) - \nu(g)\rvert\big)$ for some $L \not\equiv 0$, all $\mu \in X_g^1(\nu)$
§5.5.4	$K_{g,\nu}$ differentiable at 0	$\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$ implies $\nu_n(g) \to \nu(g)$ for all $(\nu_n) \in X_g^1(\nu)^{\mathbb{N}}$
§5.6.2	$K_{g,\nu}(t) \leq E(t)$ for all $t \in \mathbb{R}$, $g \in \mathcal{G}$, $\nu \in \mathcal{X}_{\mathcal{G}}^1$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq E^\star\big(d_{\mathcal{G}}(\mu, \nu)\big)$ for all $\mu, \nu \in \mathcal{X}_{\mathcal{G}}^1$
§5.6.3	$\operatorname{hgt}_{\psi^\star}(2t) \leq B(t)$ for all $t \in \mathbb{R}$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq B^\star\big(d_{\mathcal{B}}(\mu, \nu)\big)$ for all $\mu, \nu \in \mathcal{M}^1$

A limitation of this correspondence is that it only describes the optimal lower bound function $\mathscr{L}_{g,\nu}$ via its convex conjugate. When $\mathscr{L}_{g,\nu}$ is lower semicontinuous, this is without any loss of information by the Fenchel–Moreau theorem, but in general this only provides information about the *biconjugate* $\mathscr{L}_{g,\nu}^{\star\star}$. While $\mathscr{L}_{g,\nu}$ and $\mathscr{L}_{g,\nu}^{\star\star}$ differ in at most two points, as discussed in Section 5.5.1, the difference between the optimal lower bound and its biconjugate is potentially much more important when considering a class of functions $\mathcal{G}$ or a class of measures N as in Section 5.6.1. Some conditions under which this function is necessarily convex and lower semicontinuous were derived in Sections 5.5.3 and 5.6.3, but they do not provide a complete characterization (cf. Remarks 5.5.36 and 5.6.4). We believe that an interesting direction for future work would be to identify natural necessary or sufficient conditions under which $\mathscr{L}_{\mathcal{G},N}$ is convex and lower semicontinuous.

Chapter 6

Conclusion

We have seen throughout this thesis the power of reasoning about random variables via unpredictability-type notions, even when working with computationally bounded algorithms or when the goal is to later bound an integral probability metric. We conclude by giving two directions for future research in this vein which are more general than the specific open problems mentioned in Chapters 2 to 5.

Composition for unpredictability notions. Though (as we have seen) the Kullback–Leibler divergence is useful for analyzing constructions in complexity theory and cryptography, it has a major drawback that complicates reasoning about composition: the KL divergence does not satisfy (even relaxed versions of) the triangle inequality, that is, for all constants $C \in \mathbb{R}$, there are distributions P, Q, and R on the set $\{0,1\}$ such that

$$\mathrm{KL}(P \parallel R) > C\big(\mathrm{KL}(P \parallel Q) + \mathrm{KL}(Q \parallel R)\big).$$

In Chapter 2, to get around this we used a different inequality (Lemma 2.5.3) which replaces the $\mathrm{KL}(Q \parallel R)$ term with a larger Rényi divergence, and in Chapter 3 we worked almost exclusively with

sample notions (equivalently log-probability ratios) and took an expectation only at the end of the analysis, which allowed us to use the triangle *equality* satisfied by log-ratios. It would be interesting to find more general-purpose ways to reason about composition in the absence of the triangle inequality, perhaps by either identifying structural conditions on the distributions P, Q, and R under which similar inequalities hold, or by using a different measure than KL that is better behaved in this sense but still easy to reason about.

Isolating computational reasoning. One way to interpret the role played by the KL divergence in Chapter 2 and especially Chapter 3 is that it allowed us to cleanly delineate the statistical or information-theoretic versus computational aspects of the problems. For example, the fact that one-way function adversaries in cryptography are required to be polynomial-time bounded appears in Chapter 3 only in the first step of establishing hardness in relative entropy from one way functions (Theorem 3.3.5) and in the fact that we explicitly measure the error in rejection sampling a finite number of times (Lemma 3.4.7). It seems likely that using entropy-type notions is not the only way to obtain a similar separation of concerns, and it would be interesting to find other such techniques.

References

[ACHV19] R. Agrawal, Y.-H. Chen, T. Horel, and S. Vadhan. "Unifying Computational Entropies via Kullback–Leibler Divergence". In: *Advances in Cryptology – CRYPTO 2019*. Ed. by A. Boldyreva and D. Micciancio. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 831–858. ISBN: 978-3-030-26951-7. DOI: 10.1007/978-3-030-26951-7_28. arXiv: 1902.11202 (cit. on pp. 5, 6, 64).

[Agr19] R. Agrawal. "Samplers and Extractors for Unbounded Functions". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Ed. by D. Achlioptas and L. A. Végh. Vol. 145. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 59:1–59:21. ISBN: 978-3-95977-125-2. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2019.59. arXiv: 1904.08391 (cit. on pp. 4, 10).

[Agr20] R. Agrawal. "Finite-Sample Concentration of the Multinomial in Relative Entropy". In: *IEEE Transactions on Information Theory* Early Access (May 20, 2020). DOI: 10.1109/TIT.2020.2996134. arXiv: 1904.02291 (cit. on pp. 6, 7, 98).

[AH20] R. Agrawal and T. Horel. "Optimal Bounds between f-Divergences and Integral Probability Metrics". In: *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020. arXiv: 2006.05973 (cit. on pp. 8, 9, 114).

[AK01] A. Antos and I. Kontoyiannis. "Convergence Properties of Functional Estimates for Discrete Distributions". In: *Random Structures & Algorithms* 19.3-4 (2001), pp. 163–193. ISSN: 1098-2418. DOI: 10.1002/rsa.10019 (cit. on pp. 100, 112).

[AS06] Y. Altun and A. Smola. "Unifying Divergence Minimization and Statistical Inference Via Convex Duality". In: *Learning Theory*. Ed. by G. Lugosi and H. U. Simon. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 139–153. ISBN: 978-3-540-35296-9. DOI: 10.1007/11776420_13 (cit. on p. 119).

[AS66] S. M. Ali and S. D. Silvey. "A General Class of Coefficients of Divergence of One Distribution from Another". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 28.1 (1966), pp. 131–142. ISSN: 0035-9246 (cit. on pp. 9, 22, 115, 131).

[Bar+11] B. Barak, Y. Dodis, H. Krawczyk, O. Pereira, K. Pietrzak, F.-X. Standaert, and Y. Yu. "Leftover Hash Lemma, Revisited". In: *Advances in Cryptology – CRYPTO 2011*. Ed. by P. Rogaway. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 1–20. ISBN: 978-3-642-22792-9. DOI: 10.1007/978-3-642-22792-9_1 (cit. on p. 18).

[BBR88] C. H. Bennett, G. Brassard, and J.-M. Robert. "Privacy Amplification by Public Discussion". In: *SIAM Journal on Computing* 17.2 (Apr. 1, 1988), pp. 210–229. ISSN: 0097-5397. DOI: 10.1137/0217014 (cit. on pp. 3, 16, 41, 42).

[BC77] A. Ben-Tal and A. Charnes. *A Dual Optimization Framework for Some Problems of Information Theory and Statistics.* Center for Cynbernetic Studies, University of Texas, Austin, Nov. 1, 1977 (cit. on p. 119).

[BC79] A. Ben-Tal and A. Charnes. "A Dual Optimization Framework for Some Problems of Information Theory and Statistics". In: *Problems of Control and Information Theory. Problemy Upravlenija i Teorii Informacii* 8.5-6 (1979), pp. 387–401. ISSN: 0370-2529 (cit. on p. 119).

[BCC88] G. Brassard, D. Chaum, and C. Crépeau. "Minimum Disclosure Proofs of Knowledge". In: *Journal of Computer and System Sciences* 37.2 (Oct. 1, 1988), pp. 156–189. ISSN: 0022-0000. DOI: 10.1016/0022-0000(88)90005-0 (cit. on pp. 5, 65).

[BCR84] C. Berg, J. P. R. Christensen, and P. Ressel. "Introduction to Locally Convex Topological Vector Spaces and Dual Pairs". In: *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. New York, NY: Springer, 1984, pp. 1–15. ISBN: 978-1-4612-1128-0. DOI: 10.1007/978-1-4612-1128-0_1 (cit. on p. 124).

[Bel+18] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. "Mutual Information Neural Estimation". In: *Proceedings of the 35th International Confer-*

ence on Machine Learning. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, July 10–15, 2018, pp. 531–540 (cit. on p. 115).

[BG99] S. Bobkov and F. Götze. "Exponential Integrability and Transportation Cost Related to Logarithmic Sobolev Inequalities". In: *Journal of Functional Analysis* 163.1 (Apr. 1999), pp. 1–28. ISSN: 00221236. DOI: 10.1006/jfan.1998.3326 (cit. on pp. 121, 122).

[BGG93] M. Bellare, O. Goldreich, and S. Goldwasser. "Randomness in Interactive Proofs". In: *computational complexity* 3.4 (Dec. 1, 1993), pp. 319–354. ISSN: 1420-8954. DOI: 10.1007/BF01275487 (cit. on p. 12).

[BH79] J. Bretagnolle and C. Huber. "Estimation des densités : risque minimax". In : *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 47.2 (Jan. 1979), pp. 119–137. ISSN : 1432-2064. DOI : 10.1007/BF00535278 (cit. on p. 120).

[Bil99] P. Billingsley. *Convergence of Probability Measures*. 2nd ed. Wiley Series in Probability and Statistics. Probability and Statistics Section. New York: Wiley, 1999. 277 pp. ISBN: 978-0-471-19745-4 (cit. on p. 107).

[BK06] M. Broniatowski and A. Keziou. "Minimization of ϕ-Divergences on Sets of Signed Measures". In: *Studia Scientiarum Mathematicarum Hungarica* 43.4 (Dec. 2006), pp. 403–442. ISSN: 0081-6906, 1588-2896. DOI: 10.1556/SScMath.43.2006.4.2 (cit. on p. 119).

[BL91] J. M. Borwein and A. S. Lewis. "Duality Relationships for Entropy-Like Minimization Problems". In: *SIAM Journal on Control and Optimization* 29.2 (Mar. 1, 1991), pp. 325–338. ISSN: 0363-0129. DOI: 10.1137/0329017 (cit. on pp. 119, 139).

[BL93] J. M. Borwein and A. S. Lewis. "Partially-Finite Programming in L_1 and the Existence of Maximum Entropy Estimates". In: *SIAM Journal on Optimization* 3.2 (May 1, 1993), pp. 248–267. ISSN: 1052-6234. DOI: 10.1137/0803012 (cit. on p. 119).

[Bła18] J. Błasiok. Private Communication. Cambridge, MA USA, 2018 (cit. on p. 12).

[Bła19] J. Błasiok. "Optimal Streaming and Tracking Distinct Elements with High Probability". In: *ACM Transactions on Algorithms* 16.1 (Dec. 5, 2019), 3:1–3:28. ISSN: 1549-6325. DOI: 10.1145/3309193 (cit. on pp. 4, 11, 17).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. 1 edition. Oxford University Press, Feb. 7, 2013. 489 pp. DOI: 10.1093/acprof:oso/9780199535255.001.0001 (cit. on pp. 5, 8, 15, 37, 100, 117, 121, 152, 174).

[BM82] M. Blum and S. Micali. "How to Generate Cryptographically Strong Sequences of Pseudo Random Bits". In: *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (Sfcs 1982)*. Nov. 1982, pp. 112–117. DOI: 10.1109/SFCS.1982.72 (cit. on pp. 5, 64).

[Bou87] N. Bourbaki. *Topological Vector Spaces*. Trans. by H. G. Eggleston and S. Madan. Elements of Mathematics. Berlin: Springer-Verlag, 1987. viii+364. ISBN: 978-3-540-13627-9. DOI: 10.1007/978-3-642-61715-7 (cit. on p. 124). Trans. of *Espaces vectoriels topologiques*. Éléments de mathématique. Paris : Masson, 1981. vii+368. ISBN : 978-2-225-68410-4.

[BR94] M. Bellare and J. Rompel. "Randomness-Efficient Oblivious Sampling". In: *Proceedings 35th Annual Symposium on Foundations of Computer Science*. Nov. 1994, pp. 276–287. DOI: 10.1109/SFCS.1994.365687 (cit. on pp. 4, 10).

[Brø64] A. Brøndsted. "Conjugate Convex Functions in Topological Vector Spaces". In: *Matematiskfysiske Meddelelser udgivet af det Kongelige Danske Videnskabernes Selskab* 34.2 (1964), p. 27. ISSN: 0023-3323 (cit. on p. 160).

[BV05] F. Bolley and C. Villani. "Weighted Csiszár-Kullback-Pinsker Inequalities and Applications to Transportation Inequalities". In: *Annales de la Faculté des sciences de Toulouse : Mathématiques*. 6th ser. 14.3 (2005), pp. 331–352 (cit. on p. 175).

[Can20] C. L. Canonne. "A Short Note on Learning Discrete Distributions". Feb. 25, 2020. arXiv: 2002.11457 (cit. on p. 99).

[CCW99] A. Carbery, M. Christ, and J. Wright. "Multidimensional van der Corput and Sublevel Set Estimates". In: *Journal of the American Mathematical Society* 12.4 (1999), pp. 981–1015. ISSN: 1088-6834. DOI: 10.1090/S0894-0347-99-00309-4 (cit. on p. 182).

[CEG95] R. Canetti, G. Even, and O. Goldreich. "Lower Bounds for Sampling Algorithms for Estimating the Average". In: *Information Processing Letters* 53.1 (Jan. 13, 1995), pp. 17–25. ISSN: 0020-0190. DOI: 10.1016/0020-0190(94)00171-T (cit. on p. 12).

[CG88] B. Chor and O. Goldreich. "Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity". In: *SIAM Journal on Computing* 17.2 (Apr. 1, 1988), pp. 230–261. ISSN: 0097-5397. DOI: 10.1137/0217015 (cit. on pp. 31, 58).

[CG89] B. Chor and O. Goldreich. "On the Power of Two-Point Based Sampling". In: *Journal of Complexity* 5.1 (Mar. 1, 1989), pp. 96–106. ISSN: 0885-064X. DOI: 10.1016/0885-064X(89)90015-0 (cit. on pp. 11, 12, 61, 62).

[CGG99] I. Csiszár, F. Gamgoa, and E. Gassiat. "MEM Pixel Correlated Solutions for Generalized Moment and Interpolation Problems". In: *IEEE Transactions on Information Theory* 45.7 (Nov. 1999), pp. 2253–2270. ISSN: 00189448. DOI: 10.1109/18.796367 (cit. on p. 131).

[CM03] I. Csiszár and F. Matúš. "Information Projections Revisited". In: *IEEE Transactions on Information Theory* 49.6 (June 2003), pp. 1474–1490. ISSN: 0018-9448. DOI: 10.1109/TIT.2003.810633 (cit. on p. 119).

[CM12] I. Csiszár and F. Matúš. "Generalized Minimizers of Convex Integral Functionals, Bregman Distance, Pythagorean Identities". In: *Kybernetika* 48.4 (2012), pp. 637–689. ISSN: 0023-5954. arXiv: 1202.0666 (cit. on p. 119).

[CS05] I. Csiszár and P. C. Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory. Hanover, MA: Now Publishers, 2005. 115 pp. ISBN: 978-1-933019-05-5 (cit. on p. 107).

[Csi62] I. Csiszár. "Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen". In: *A Magyar Tudományos Akadémia. Matematikai Kutató Intézetének Közleményei* 7 (1962), pp. 137–158. ISSN: 0541-9514 (cit. on p. 162).

[Csi63] I. Csiszár. "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten". In: *A Magyar Tudományos Akadémia. Matematikai Kutató Intézetének Közleményei* 8 (1963), pp. 85–108 (cit. on pp. 9, 22, 115, 130).

[Csi64] I. Csiszár. "Über topologische und metrische Eigenschaften der relativen Information der Ordnung α". In: *Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, 1962*. Prague: Publishing House of the Czechoslovak Academy of Science, 1964, pp. 63–73 (cit. on p. 162).

[Csi67a] I. Csiszár. "Information-Type Measures of Difference of Probability Distributions and Indirect Observations". In: *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 299–318 (cit. on pp. 9, 115, 120, 131).

[Csi67b] I. Csiszár. "On Topological Properties of f-Divergences". In: *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 329–339 (cit. on p. 162).

[Csi75] I. Csiszár. "I-Divergence Geometry of Probability Distributions and Minimization Problems". In: *The Annals of Probability* 3.1 (Feb. 1975), pp. 146–158. ISSN: 0091-1798. DOI: 10.1214/aop/1176996454 (cit. on p. 119).

[Csi98] I. Csiszár. "The Method of Types". In: *IEEE Transactions on Information Theory* 44.6 (Oct. 1998), pp. 2505–2523. ISSN: 0018-9448. DOI: 10.1109/18.720546 (cit. on pp. 7, 100, 112).

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006. 748 pp. ISBN: 978-0-471-24195-9 (cit. on p. 113).

[CW89] A. Cohen and A. Wigderson. "Dispersers, Deterministic Amplification, and Weak Random Sources". In: *30th Annual Symposium on Foundations of Computer Science*. Oct. 1989, pp. 14–19. DOI: 10.1109/SFCS.1989.63449 (cit. on p. 53).

[DH76] W. Diffie and M. E. Hellman. "New Directions in Cryptography". In: *IEEE Transactions on Information Theory* 22.6 (Nov. 1976), pp. 644–654. ISSN: 1557-9654. DOI: 10.1109/TIT.1976.1055638 (cit. on pp. 5, 64).

[DHRS04] Y. Z. Ding, D. Harnik, A. Rosen, and R. Shaltiel. "Constant-Round Oblivious Transfer in the Bounded Storage Model". In: *Theory of Cryptography, First Theory of Cryptography Conference, TCC 2004*. Ed. by M. Naor. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 446–472. ISBN: 978-3-540-24638-1. DOI: 10.1007/978-3-540-24638-1_25 (cit. on p. 66).

[DORS08] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith. "Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data". In: *SIAM Journal on Computing* 38.1 (Jan. 1, 2008), pp. 97–139. ISSN: 0097-5397. DOI: 10.1137/060651380 (cit. on pp. 23, 24, 29, 41, 42, 49).

[DRG15] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. "Training Generative Neural Networks via Maximum Mean Discrepancy Optimization". In: *Proceedings of the Thirty-First*

Conference on Uncertainty in Artificial Intelligence. UAI'15. Arlington, Virginia, USA: AUAI Press, 2015, pp. 258–267. ISBN: 978-0-9966431-0-8 (cit. on p. 116).

[DS58] N. Dunford and J. T. Schwartz. *Linear Operators. I. General Theory*. In collab. with W. G. Bade and R. G. Bartle. Pure and Applied Mathematics 7. Interscience Publishers, Inc., New York; Interscience Publishers, Ltd., London, 1958. xiv+858. ISBN: 0-470-22605-6 (cit. on p. 21).

[Dud64] R. M. Dudley. "On Sequential Convergence". In: *Transactions of the American Mathematical Society* 112.3 (1964), pp. 483–507. ISSN: 0002-9947, 1088-6850. DOI: 10.1090/S0002-9947-1964-0175081-6 (cit. on p. 162).

[Dud98] R. M. Dudley. "Consistency of M-Estimators and One-Sided Bracketing". In: *High Dimensional Probability*. Ed. by E. Eberlein, M. Hahn, and M. Talagrand. Basel: Birkhäuser Basel, 1998, pp. 33–58. ISBN (print): 978-3-0348-9790-7. ISBN (online): 978-3-0348-8829-5. DOI: 10.1007/978-3-0348-8829-5_3 (cit. on p. 162).

[DV76] M. D. Donsker and S. R. S. Varadhan. "Asymptotic Evaluation of Certain Markov Process Expectations for Large Time—III". In: *Communications on Pure and Applied Mathematics* 29.4 (1976), pp. 389–461. ISSN: 1097-0312. DOI: 10.1002/cpa.3160290405 (cit. on pp. 8, 37, 115, 143).

[DWCS20] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar. "On the Robustness of Information-Theoretic Privacy Measures and Mechanisms". In: *IEEE Transactions on Information Theory* 66.4 (Apr. 2020), pp. 1949–1978. ISSN: 1557-9654. DOI: 10.1109/TIT.2019.2939472 (cit. on p. 99).

[ES89] G. A. Edgar and L. Sucheston. "On Maximal Inequalities in Orlicz Spaces". In: *Contemporary Mathematics*. Ed. by R. D. Mauldin, R. M. Shortt, and C. E. Silva. Vol. 94. Providence, Rhode Island: American Mathematical Society, 1989, pp. 113–129. ISBN (print): 978-0-8218-5099-2. ISBN (online): 978-0-8218-7682-4. DOI: 10.1090/conm/094/1012982 (cit. on p. 129).

[ET99] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, Jan. 1999. ISBN (print): 978-0-89871-450-0. ISBN (online): 978-1-61197-108-8. DOI: 10.1137/1.9781611971088 (cit. on p. 124).

[FHT03] A. A. Fedotov, P. Harremoës, and F. Topsøe. "Refinements of Pinsker's Inequality". In: *IEEE Transactions on Information Theory* 49.6 (June 2003), pp. 1491–1498. ISSN: 0018-9448. DOI: 10.1109/TIT.2003.811927 (cit. on pp. 117, 120, 181).

[GBRSS12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25 (Mar. 2012), pp. 723–773. ISSN: 1532-4435 (cit. on p. 116).

[GG81] O. Gabber and Z. Galil. "Explicit Constructions of Linear-Sized Superconcentrators". In: *Journal of Computer and System Sciences* 22.3 (June 1, 1981), pp. 407–420. ISSN: 0022-0000. DOI: 10.1016/0022-0000(81)90040-4 (cit. on p. 45).

[GGM86] O. Goldreich, S. Goldwasser, and S. Micali. "How to Construct Random Functions". In: *Journal of the ACM* 33.4 (Aug. 10, 1986), pp. 792–807. ISSN: 0004-5411. DOI: 10.1145/6490.6503 (cit. on p. 64).

[Gil06] G. L. Gilardoni. "On the Minimum f-Divergence for given Total Variation". In: *Comptes Rendus Mathematique* 343.11 (Dec. 1, 2006), pp. 763–766. ISSN: 1631-073X. DOI: 10.1016/j.crma.2006.10.027 (cit. on pp. 117, 120).

[Gil08] G. L. Gilardoni. "An Improvement on Vajda's Inequality". In: *In and Out of Equilibrium 2*. Ed. by V. Sidoravicius and M. E. Vares. Progress in Probability. Basel: Birkhäuser Basel, 2008, pp. 299–304. ISBN: 978-3-7643-8786-0. DOI: 10.1007/978-3-7643-8786-0_14 (cit. on p. 120).

[Gil10] G. L. Gilardoni. "On Pinsker's Type Inequalities and Csiszar's f-Divergences". In: *IEEE Transactions on Information Theory* 56.11 (Nov. 2010), pp. 5377–5386. ISSN: 0018-9448. DOI: 10.1109/TIT.2010.2068710 (cit. on pp. 54, 120, 182–185).

[Gil98] D. Gillman. "A Chernoff Bound for Random Walks on Expander Graphs". In: *SIAM Journal on Computing* 27.4 (Aug. 1, 1998), pp. 1203–1220. ISSN: 0097-5397. DOI: 10.1137/S0097539794268765 (cit. on pp. 12, 17).

[GKP94] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science, Second Edition*. Upper Saddle River, NJ: Addison-Wesley, 1994 (cit. on p. 105).

[GL10] N. Gozlan and C. Léonard. "Transport Inequalities. A Survey". In: *Markov Processes and Related Fields* 16.4 (2010), pp. 635–736. ISSN: 1024-2953. arXiv: 1003.3852 (cit. on p. 121).

[GM84] S. Goldwasser and S. Micali. "Probabilistic Encryption". In: *Journal of Computer and System Sciences* 28.2 (Apr. 1984), pp. 270–299. ISSN: 00220000. DOI: 10.1016/0022-0000(84)90070-9 (cit. on pp. 2, 21).

[GMW87] O. Goldreich, S. Micali, and A. Wigderson. "How to Play Any Mental Game or a Completeness Theorem for Protocols with Honest Majority". In: *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*. STOC '87. New York, New York, USA: Association for Computing Machinery, Jan. 1, 1987, pp. 218–229. ISBN: 978-0-89791-221-1. DOI: 10.1145/28395.28420 (cit. on p. 64).

[GMW91] O. Goldreich, S. Micali, and A. Wigderson. "Proofs That Yield Nothing but Their Validity or All Languages in NP Have Zero-Knowledge Proof Systems". In: *Journal of the ACM* 38.3 (July 1, 1991), pp. 691–729. ISSN: 0004-5411. DOI: 10.1145/116825.116852 (cit. on p. 64).

[Gol11a] O. Goldreich. "A Sample of Samplers: A Computational Perspective on Sampling". In: *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*. Ed. by O. Goldreich. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 302–332. ISBN: 978-3-642-22670-0. DOI: 10.1007/978-3-642-22670-0_24 (cit. on p. 11).

[Gol11b] O. Goldreich. "Basic Facts about Expander Graphs". In: *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*. Ed. by O. Goldreich. Vol. 6650. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 451–464. ISBN (print): 978-3-642-22669-4. ISBN (online): 978-3-642-22670-0. DOI: 10.1007/978-3-642-22670-0_30 (cit. on p. 46).

[Gor91] G. Gorni. "Conjugation and Second-Order Properties of Convex Functions". In: *Journal of Mathematical Analysis and Applications* 158.2 (July 1, 1991), pp. 293–315. ISSN: 0022-247X. DOI: 10.1016/0022-247X(91)90237-T (cit. on p. 186).

[GR20] F. R. Guo and T. S. Richardson. "Chernoff-Type Concentration of Empirical Probabilities in Relative Entropy". In: *arXiv e-prints* (Mar. 19, 2020). arXiv: 2003.08614 [math, stat] (cit. on p. 110).

[GSS14] A. Guntuboyina, S. Saha, and G. Schiebinger. "Sharp Inequalities for f-Divergences". In: *IEEE Transactions on Information Theory* 60.1 (Jan. 2014), pp. 104–121. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2013.2288674 (cit. on p. 121).

[GUV09] V. Guruswami, C. Umans, and S. Vadhan. "Unbalanced Expanders and Randomness Extractors from Parvaresh–Vardy Codes". In: *Journal of the ACM* 56.4 (July 2009), 20:1–20:34. ISSN: 0004-5411. DOI: 10.1145/1538902.1538904 (cit. on pp. 12, 15, 47, 52).

[GW97] O. Goldreich and A. Wigderson. "Tiny Families of Functions with Random Properties: A Quality-Size Trade-off for Hashing". In: *Random Structures & Algorithms* 11.4 (1997), pp. 315–343. ISSN: 1098-2418. DOI: 10.1002/(SICI)1098-2418(199712)11:4<315::AID-RSA3>3.0.CO;2-1 (cit. on pp. 12, 16, 17, 34, 40, 43, 45, 59, 62).

[Har07] P. Harremoës. "Information Topologies with Applications". In: *Entropy, Search, Complexity*. Ed. by I. Csiszár, G. O. H. Katona, G. Tardos, and G. Wiener. Bolyai Society Mathematical Studies. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 113–150. ISBN: 978-3-540-32777-6. DOI: 10.1007/978-3-540-32777-6_5 (cit. on p. 162).

[Har17] P. Harremoës. "Bounds on Tail Probabilities for Negative Binomial Distributions". In: *Kybernetika* (Feb. 2, 2017), pp. 943–966. ISSN: 0023-5954, 1805-949X. DOI: 10.14736/kyb-2016-6-0943 (cit. on p. 110).

[Her67] H. H. Herda. "On Non-Symmetric Modular Spaces". In: *Colloquium Mathematicum* 17.2 (1967), pp. 333–346. ISSN: 0010-1354. DOI: 10.4064/cm-17-2-333-346 (cit. on p. 162).

[HHRVW10] I. Haitner, T. Holenstein, O. Reingold, S. Vadhan, and H. Wee. "Universal One-Way Hash Functions via Inaccessible Entropy". In: *Advances in Cryptology – EUROCRYPT 2010*. Ed. by H. Gilbert. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 616–637. ISBN: 978-3-642-13190-5 (cit. on pp. 5, 6, 65, 75).

[HILL99] J. Håstad, R. Impagliazzo, L. Levin, and M. Luby. "A Pseudorandom Generator from Any One-Way Function". In: *SIAM Journal on Computing* 28.4 (Jan. 1, 1999), pp. 1364–1396. ISSN: 0097-5397. DOI: 10.1137/S0097539793244708 (cit. on pp. 5, 64–66).

[HL93] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Red. by M. Artin, S. S. Chern, J. Coates, J. M. Fröhlich, H. Hironaka, F. Hirzebruch, L. Hörmander, C. C. Moore, J. K. Moser, M. Nagata, W. Schmidt, D. S. Scott, Y. G. Sinai, J. Tits, M. Waldschmidt, S. Watanabe, M. Berger, B. Eckmann, and S. R. S. Varadhan. Vol. 305. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993. ISBN (print): 978-3-642-08161-3. ISBN (online): 978-3-662-02796-7. DOI: 10.1007/978-3-662-02796-7 (cit. on p. 186).

[HNORV09] I. Haitner, M. Nguyen, S. Ong, O. Reingold, and S. Vadhan. "Statistically Hiding Commitments and Statistical Zero-Knowledge Arguments from Any One-Way Function". In: *SIAM Journal on Computing* 39.3 (Jan. 1, 2009), pp. 1153–1218. ISSN: 0097-5397. DOI: 10.1137/080725404 (cit. on pp. 5, 64, 65).

[Hoe63] W. Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. ISSN: 0162-1459. DOI: 10.2307/2282952 (cit. on pp. 103, 174).

[HRV10] I. Haitner, O. Reingold, and S. Vadhan. "Efficiency Improvements in Constructing Pseudorandom Generators from One-Way Functions". In: *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC)*. STOC '10. Cambridge, Massachusetts, USA: Association for Computing Machinery, June 5, 2010, pp. 437–446. ISBN: 978-1-4503-0050-6. DOI: 10.1145/1806689.1806750 (cit. on p. 67).

[HRV13] I. Haitner, O. Reingold, and S. Vadhan. "Efficiency Improvements in Constructing Pseudorandom Generators from One-Way Functions". In: *SIAM Journal on Computing* 42.3 (Jan. 1, 2013), pp. 1405–1430. ISSN: 0097-5397. DOI: 10.1137/100814421 (cit. on pp. 5, 65).

[HRVW09] I. Haitner, O. Reingold, S. Vadhan, and H. Wee. "Inaccessible Entropy". In: *Proceedings of the 41st Annual ACM Symposium on Symposium on Theory of Computing - STOC '09*. Bethesda, MD, USA: ACM Press, 2009, p. 611. ISBN: 978-1-60558-506-2. DOI: 10.1145/1536414.1536497 (cit. on pp. 5, 65, 66, 69, 70, 75).

[HRVW16] I. Haitner, O. Reingold, S. P. Vadhan, and H. Wee. "Inaccessible Entropy I: Inaccessible Entropy Generators and Statistically Hiding Commitments from One-Way Functions."

2016. URL: www.cs.tau.ac.il/~iftachh/papers/AccessibleEntropy/IE1.pdf (cit. on pp. 86, 94, 96, 97).

[HT12] P. Harremoës and G. Tusnády. "Information Divergence Is More χ^2-Distributed than the χ^2-Statistics". In: *2012 IEEE International Symposium on Information Theory Proceedings*. July 2012, pp. 533–537. DOI: 10.1109/ISIT.2012.6284247 (cit. on pp. 98, 99).

[HV11] P. Harremoës and I. Vajda. "On Pairs of f-Divergences and Their Joint Range". In: *IEEE Transactions on Information Theory* 57.6 (June 2011), pp. 3230–3235. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2011.2137353. arXiv: 1007.0097 (cit. on p. 120).

[IL89] R. Impagliazzo and M. Luby. "One-Way Functions Are Essential for Complexity Based Cryptography". In: *Proceedings of the 30th Annual Symposium on Foundations of Computer Science (FOCS)*. Oct. 1989, pp. 230–235. DOI: 10.1109/SFCS.1989.63483 (cit. on pp. 5, 64).

[ILL89] R. Impagliazzo, L. A. Levin, and M. Luby. "Pseudo-Random Generation from One-Way Functions". In: *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing* (Seattle, Washington, USA). STOC '89. New York, NY, USA: ACM, 1989, pp. 12–24. ISBN: 978-0-89791-307-2. DOI: 10.1145/73007.73009 (cit. on pp. 3, 16, 41, 42).

[IT69] A. D. Ioffe and V. M. Tikhomirov. "On Minimization of Integral Functionals". In: *Functional Analysis and Its Applications* 3.3 (July 1969), pp. 218–227. ISSN: 0016-2663, 1573-8485. DOI: 10.1007/BF01676623 (cit. on p. 165). Trans. of A. Д. Иоффе and B. M. Тихомиров. "О минимизации интегральных функционалов". In: *Функциональный анализ и его приложения* 3.3 (1969), pp. 61–70.

[IZ89] R. Impagliazzo and D. Zuckerman. "How to Recycle Random Bits". In: *30th Annual Symposium on Foundations of Computer Science*. Oct. 1989, pp. 248–253. DOI: 10.1109/SFCS.1989.63486 (cit. on p. 42).

[Jam72] G. J. O. Jameson. "Convex Series". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 72.1 (July 1972), pp. 37–47. ISSN: 1469-8064, 0305-0041. DOI: 10.1017/S0305004100050933 (cit. on pp. 128, 157).

[JHW17] J. Jiao, Y. Han, and T. Weissman. "Dependence Measures Bounding the Exploration Bias for General Measurements". In: *2017 IEEE International Symposium on Information*

Theory (ISIT). June 2017, pp. 1475–1479. DOI: 10.1109/ISIT.2017.8006774. arXiv: 1612.05845 (cit. on p. 143).

[Jof71] A. Joffe. "On a Sequence of Almost Deterministic Pairwise Independent Random Variables". In: *Proceedings of the American Mathematical Society* 29 (1971), pp. 381–382. ISSN: 0002-9939. DOI: 10.2307/2038147 (cit. on p. 62).

[JVHW17] J. Jiao, K. Venkat, Y. Han, and T. Weissman. "Maximum Likelihood Estimation of Functionals of Discrete Distributions". In: *IEEE Transactions on Information Theory* 63.10 (Oct. 2017), pp. 6774–6798. ISSN: 0018-9448. DOI: 10.1109/TIT.2017.2733537 (cit. on pp. 99, 111).

[Kem69] J. H. B. Kemperman. "On the Optimum Rate of Transmitting Information". In: *Annals of Mathematical Statistics* 40.6 (Dec. 1969), pp. 2156–2177. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177697293 (cit. on p. 120).

[KFG06] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver. "Exceptionality of the Variational Distance". In: *Proceedings of the 2006 IEEE Information Theory Workshop*. Chengdu, China: IEEE, Oct. 2006, pp. 274–276. ISBN (print): 978-1-4244-0067-6. ISBN (online): 978-1-4244-0068-3. DOI: 10.1109/ITW2.2006.323802 (cit. on p. 121).

[KFG07] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver. "Confliction of the Convexity and Metric Properties in f-Divergences". In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E90-A.9 (Sept. 1, 2007), pp. 1848–1853. ISSN: 0916-8508. DOI: 10.1093/ietfec/e90-a.9.1848 (cit. on p. 121).

[Kis60] J. Kisyński. "Convergence du type $\mathcal{L}$". In : *Colloquium Mathematicum* 7.2 (1960), pp. 205–211. ISSN : 0010-1354. DOI : 10.4064/cm-7-2-205-211 (cit. on p. 162).

[KL51] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729694 (cit. on pp. 3, 118).

[Kön86] H. König. "Theory and Applications of Superconvex Spaces". In: *Aspects of Positivity in Functional Analysis (Tübingen, 1985)*. Vol. 122. North-Holland Math. Stud. North-Holland, Amsterdam, 1986, pp. 79–118 (cit. on p. 128).

[KPS85] R. Karp, N. Pippenger, and M. Sipser. "A Time-Randomness Tradeoff". In: Durham, New Hampshire, 1985 (cit. on pp. 12, 62).

[Kul59] S. Kullback. *Information Theory and Statistics*. New York: Wiley, 1959. xvii+395 (cit. on pp. 114, 118).

[Kul67] S. Kullback. "A Lower Bound for Discrimination Information in Terms of Variation". In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), pp. 126–127. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.1967.1053968 (cit. on p. 120).

[Léo01a] C. Léonard. "Minimization of Energy Functionals Applied to Some Inverse Problems". In: *Applied Mathematics and Optimization* 44.3 (Jan. 1, 2001), pp. 273–297. ISSN: 0095-4616, 1432-0606. DOI: 10.1007/s00245-001-0019-5 (cit. on pp. 119, 120, 160).

[Léo01b] C. Léonard. "Minimizers of Energy Functionals". In: *Acta Mathematica Hungarica* 93.4 (2001), pp. 281–325. ISSN: 02365294. DOI: 10.1023/A:1017919422086 (cit. on p. 119).

[Léo07] C. Léonard. *Orlicz Spaces*. Apr. 23, 2007, p. 10. URL: http://leonard.perso.math.cnrs.fr/papers/Leonard-Orlicz%20spaces.pdf (cit. on p. 128).

[Lev68] V. L. Levin. "Some Properties of Support Functionals". In: *Mathematical Notes of the Academy of Sciences of the USSR* 4.6 (Dec. 1968), pp. 900–906. ISSN: 0001-4346, 1573-8876. DOI: 10.1007/BF01110826 (cit. on p. 165). Trans. of В. Л. Левин. "О некоторых свойствах опорных функционалов". In: *Математические заметки* 4.6 (1968), pp. 685–696.

[LZ56] W. Luxemburg and A. Zaanen. "Conjugate Spaces of Orlicz Spaces". In: *Indagationes Mathematicae (Proceedings)* 59 (1956), pp. 217–228. ISSN: 1385-7258. DOI: 10.1016/S1385-7258(56)50029-7 (cit. on p. 134).

[Mar73] G. A. Margulis. "Explicit Constructions of Expanders". In: *Problems of Information Transmission* 9.4 (1973), pp. 325–332. ISSN: 0555-2923 (cit. on p. 45). Trans. of Г. А. Маргулис. "Явные Конструкции Расширителей". In: *Проблемы передачи информации* 9.4 (1973), pp. 71–80.

[Mar86] K. Marton. "A Simple Proof of the Blowing-Up Lemma". In: *IEEE Transactions on Information Theory* 32.3 (May 1986), pp. 445–446. ISSN: 1557-9654. DOI: 10.1109/TIT.1986.1057176 (cit. on p. 121).

[McD89] C. McDiarmid. "On the Method of Bounded Differences". In: *Surveys in Combinatorics, 1989*. Ed. by J. Siemons. Cambridge: Cambridge University Press, 1989, pp. 148–188. ISBN: 978-1-107-35994-9. DOI: 10.1017/CBO9781107359949.008 (cit. on p. 112).

[McI87] J. L. McInnes. *Cryptography Using Weak Sources of Randomness*. Technical Report 194/87. University of Toronto, 1987 (cit. on pp. 3, 16, 41, 42).

[MJTNW19] J. Mardia, J. Jiao, E. Tánczos, R. D. Nowak, and T. Weissman. "Concentration Inequalities for the Empirical Distribution of Discrete Distributions: Beyond the Method of Types". In: *Information and Inference: A Journal of the IMA* (Nov. 18, 2019), iaz025. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaz025. arXiv: 1809.06522 (cit. on pp. 7, 100, 103, 107, 111–113).

[Mor63] T. Morimoto. "Markov Processes and the H-Theorem". In: *Journal of the Physical Society of Japan* 18.3 (Mar. 15, 1963), pp. 328–331. ISSN: 0031-9015. DOI: 10.1143/JPSJ.18.328 (cit. on p. 9).

[Mor64] J. J. Moreau. "Sur la fonction polaire d'une fonction semi-continue supérieurement". In : *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 258 (1964), pp. 1128–1130 (cit. on p. 160).

[MT50] M. Morse and W. Transue. "Functionals F Bilinear Over the Product $A \times B$ of Two Pseudo-Normed Vector Spaces: II. Admissible Spaces A". In: *Annals of Mathematics* 51.3 (1950), pp. 576–614. ISSN: 0003-486X. DOI: 10.2307/1969370 (cit. on p. 129).

[Mül97] A. Müller. "Integral Probability Metrics and Their Generating Classes of Functions". In: *Advances in Applied Probability* 29.2 (1997), pp. 429–443. ISSN: 0001-8678. DOI: 10.2307/1428011 (cit. on pp. 1, 20, 21, 116).

[Mus83] J. Musielak. *Orlicz Spaces and Modular Spaces*. Vol. 1034. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983. ISBN (print): 978-3-540-12706-2. ISBN (online): 978-3-540-38692-6. DOI: 10.1007/BFb0072210 (cit. on p. 162).

[Nak50] H. Nakano. *Modulared Semi-Ordered Linear Spaces*. Vol. 1. Tokyo Mathematical Book Series. Tokyo: Maruzen Co., Ltd., 1950. 288 pp. (cit. on p. 162).

[Nao91] M. Naor. "Bit Commitment Using Pseudorandomness". In: *Journal of Cryptology* 4.2 (Jan. 1991), pp. 151–158. ISSN: 0933-2790, 1432-1378. DOI: 10.1007/BF00196774 (cit. on p. 64).

[NCMQW17] R. Nock, Z. Cranko, A. K. Menon, L. Qu, and R. C. Williamson. "f-GANs in an Information Geometric Nutshell". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 456–464. ISBN: 978-1-5108-6096-4 (cit. on p. 115).

[NCT16] S. Nowozin, B. Cseke, and R. Tomioka. "f-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. NIPS'16. Red Hook, NY, USA: Curran Associates, Inc., 2016, pp. 271–279. ISBN: 978-1-5108-3881-9 (cit. on p. 115).

[NOVY98] M. Naor, R. Ostrovsky, R. Venkatesan, and M. Yung. "Perfect Zero-Knowledge Arguments for NP Using Any One-Way Permutation". In: *Journal of Cryptology* 11.2 (Mar. 1, 1998), pp. 87–108. ISSN: 1432-1378. DOI: 10.1007/s001459900037 (cit. on p. 66).

[NP33] J. Neyman and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231.694-706 (Jan. 1, 1933), pp. 289–337. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.1933.0009 (cit. on p. 98).

[NWJ08] X. Nguyen, M. J. Wainwright, and M. I. Jordan. "Estimating Divergence Functionals and the Likelihood Ratio by Penalized Convex Risk Minimization". In: *Advances in Neural Information Processing Systems 20*. Ed. by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis. Curran Associates, Inc., 2008, pp. 1089–1096 (cit. on p. 115).

[NWJ10] X. Nguyen, M. J. Wainwright, and M. I. Jordan. "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization". In: *IEEE Transactions on Information Theory* 56.11 (Nov. 2010), pp. 5847–5861. ISSN: 0018-9448. DOI: 10.1109/TIT.2010.2068870 (cit. on p. 115).

[NY89] M. Naor and M. Yung. "Universal One-Way Hash Functions and Their Cryptographic Applications". In: *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*. STOC '89. Seattle, Washington, USA: Association for Computing Machinery, Feb. 1, 1989, pp. 33–43. ISBN: 978-0-89791-307-2. DOI: 10.1145/73007.73011 (cit. on pp. 5, 65).

[NZ96] N. Nisan and D. Zuckerman. "Randomness Is Linear in Space". In: *Journal of Computer and System Sciences* 52.1 (Feb. 1, 1996), pp. 43–52. ISSN: 0022-0000. DOI: 10.1006/jcss.1996.0004 (cit. on pp. 5, 13, 23, 24, 34, 66).

[Pan03] L. Paninski. “Estimation of Entropy and Mutual Information”. In: *Neural Computation* 15.6 (June 1, 2003), pp. 1191–1253. ISSN: 0899-7667. DOI: 10.1162/089976603321780272 (cit. on pp. 7, 99, 110, 111).

[Пин60] М. С. Пинскер. *Информация и информационная устойчивость случайных величин и процессов*. Проблемы передачи информации 7. АН СССР, 1960. 203 pp. (cit. on p. 120). M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Trans. by A. Feinstein. Holden-Day, 1964.

[Pit79] E. J. G. Pitman. *Some Basic Theory for Statistical Inference*. Monographs on Applied Probability and Statistics. London : New York: Chapman and Hall ; distributed in the U.S.A. by Halsted Press, 1979. 110 pp. ISBN: 978-0-470-26554-3 (cit. on p. 98).

[Rén61] A. Rényi. “On Measures of Entropy and Information”. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 1961 (cit. on p. 19).

[Roc66] R. T. Rockafellar. “Level Sets and Continuity of Conjugate Convex Functions”. In: *Transactions of the American Mathematical Society* 123.1 (1966), pp. 46–63. ISSN: 0002-9947, 1088-6850. DOI: 10.1090/S0002-9947-1966-0192318-X (cit. on p. 119).

[Roc68] R. T. Rockafellar. “Integrals Which Are Convex Functionals.” In: *Pacific Journal of Mathematics* 24.3 (1968), pp. 525–539. ISSN: 0030-8730 (cit. on pp. 119, 130, 134).

[Roc71] R. T. Rockafellar. “Integrals Which Are Convex Functionals. II.” In: *Pacific Journal of Mathematics* 39.2 (1971), pp. 439–469. ISSN: 0030-8730 (cit. on pp. 130, 134, 160).

[Roc76] R. T. Rockafellar. “Integral Functionals, Normal Integrands and Measurable Selections”. In: *Nonlinear Operators and the Calculus of Variations*. Ed. by J. P. Gossez, E. J. Lami Dozo, J. Mawhin, and L. Waelbroeck. Vol. 543. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1976, pp. 157–207. ISBN (print): 978-3-540-07867-8. ISBN (online): 978-3-540-38075-7. DOI: 10.1007/BFb0079944 (cit. on pp. 130, 134).

[Rom90] J. Rompel. “One-Way Functions Are Necessary and Sufficient for Secure Signatures”. In: *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*. STOC ’90. Baltimore, Maryland, USA: Association for Computing Machinery, Apr. 1, 1990, pp. 387–394. ISBN: 978-0-89791-361-4. DOI: 10.1145/100216.100269 (cit. on pp. 5, 64–66).

[RR91] M. M. Rao and Z. D. Ren. *Theory of Orlicz Spaces*. Monographs and Textbooks in Pure and Applied Mathematics 146. New York: M. Dekker, 1991. 449 pp. ISBN: 978-0-8247-8478-2 (cit. on p. 128).

[RR99] R. Raz and O. Reingold. "On Recycling the Randomness of States in Space Bounded Computation". In: *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing - STOC '99*. Atlanta, Georgia, United States: ACM Press, 1999, pp. 159–168. ISBN: 978-1-58113-067-6. DOI: 10.1145/301250.301294 (cit. on pp. 47, 48).

[RRGP12] A. Ruderman, M. Reid, D. García-García, and J. Petterson. "Tighter Variational Representations of f-Divergences via Restriction to Probability Measures". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (Edinburgh, Scotland). Ed. by J. Langford and J. Pineau. ICML '12. New York, NY, USA: Omnipress, July 2012, pp. 671–678. ISBN: 978-1-4503-1285-1 (cit. on pp. 9, 115, 117, 135, 138).

[RRV02] R. Raz, O. Reingold, and S. Vadhan. "Extracting All the Randomness and Reducing the Error in Trevisan's Extractors". In: *Journal of Computer and System Sciences* 65.1 (Aug. 1, 2002), pp. 97–128. ISSN: 0022-0000. DOI: 10.1006/jcss.2002.1824 (cit. on pp. 16, 17, 46, 47, 51, 52).

[RT00] J. Radhakrishnan and A. Ta-Shma. "Bounds for Dispersers, Extractors, and Depth-Two Superconcentrators". In: *SIAM Journal on Discrete Mathematics* 13.1 (Jan. 1, 2000), pp. 2–24. ISSN: 0895-4801. DOI: 10.1137/S0895480197329508 (cit. on pp. 12, 16, 46, 53, 54, 56, 63).

[RTTV08a] O. Reingold, L. Trevisan, M. Tulsiani, and S. Vadhan. "Dense Subsets of Pseudorandom Sets". In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. Oct. 2008, pp. 76–85. DOI: 10.1109/FOCS.2008.38 (cit. on p. 21).

[RTTV08b] O. Reingold, L. Trevisan, M. Tulsiani, and S. Vadhan. "New Proofs of the Green-Tao-Ziegler Dense Model Theorem: An Exposition". In: *arXiv e-prints* (June 2, 2008). arXiv: 0806.0381 [math] (cit. on p. 21).

[RVW00] O. Reingold, S. Vadhan, and A. Wigderson. "Entropy Waves, the Zig-Zag Graph Product, and New Constant-Degree Expanders and Extractors". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. Nov. 2000, pp. 3–13. DOI: 10.1109/SFCS.2000.892006 (cit. on pp. 12, 15, 17, 34, 41, 45, 47, 50, 51, 59).

[RW09] M. D. Reid and R. C. Williamson. "Generalised Pinsker Inequalities". In: *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*. 2009, p. 10 (cit. on p. 120).

[RW11] M. D. Reid and R. C. Williamson. "Information, Divergence and Risk for Binary Experiments". In: *Journal of Machine Learning Research* 12.22 (2011), pp. 731–817. ISSN: 1533-7928 (cit. on p. 120).

[RW98] R. T. Rockafellar and R. J. B. Wets. "Measurability". In: *Variational Analysis*. Red. by M. Berger, P. de la Harpe, F. Hirzebruch, N. J. Hitchin, L. Hörmander, A. Kupiainen, G. Lebeau, M. Ratner, D. Serre, Y. G. Sinai, N. J. A. Sloane, A. M. Vershik, and M. Waldschmidt. Vol. 317. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. Chap. 14, pp. 642–683. ISBN (print): 978-3-540-62772-2. ISBN (online): 978-3-642-02431-3. DOI: 10.1007/978-3-642-02431-3_14 (cit. on p. 130).

[RZ20] D. Russo and J. Zou. "How Much Does Your Data Exploration Overfit? Controlling Bias via Information Usage". In: *IEEE Transactions on Information Theory* 66.1 (Jan. 2020), pp. 302–323. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2019.2945779 (cit. on p. 117).

[San57] I. N. Sanov. "On the Probability of Large Deviations of Random Variables". In: *Mat. Sb. N. S.* 42.84 (1957), pp. 11–44 (cit. on p. 118).

[SFGSL12] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. "On the Empirical Estimation of Integral Probability Metrics". In: *Electronic Journal of Statistics* 6 (2012), pp. 1550–1599. ISSN: 1935-7524. DOI: 10.1214/12-EJS722 (cit. on pp. 116, 121).

[SGFSL09] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. "A Note on Integral Probability Metrics and ϕ-Divergences". In: *arXiv e-prints* (Jan. 18, 2009). arXiv: 0901.2698v1 [cs, math] (cit. on p. 121).

[Sha11] O. Shayevitz. "On Rényi Measures and Hypothesis Testing". In: *2011 IEEE International Symposium on Information Theory Proceedings*. July 2011, pp. 894–898. DOI: 10.1109/ISIT.2011.6034266 (cit. on p. 39).

[Sha48] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 2).

[Sio58] M. Sion. "On General Minimax Theorems". In: *Pacific Journal of Mathematics* 8.1 (Mar. 1, 1958), pp. 171–176. ISSN: 0030-8730, 0030-8730. DOI: 10.2140/pjm.1958.8.171 (cit. on p. 181).

[Sip88] M. Sipser. "Expanders, Randomness, or Time versus Space". In: *Journal of Computer and System Sciences* 36.3 (June 1, 1988), pp. 379–383. ISSN: 0022-0000. DOI: 10.1016/0022-0000(88)90035-9 (cit. on p. 53).

[Ste93] E. M. Stein. *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*. Princeton Mathematical Series 43. Princeton University Press, Princeton, NJ, 1993. xiv+695. ISBN: 978-0-691-03216-0 (cit. on p. 182).

[SV16] I. Sason and S. Verdú. "f-Divergence Inequalities". In: *IEEE Transactions on Information Theory* 62.11 (Nov. 2016), pp. 5973–6006. DOI: 10.1109/TIT.2016.2603151 (cit. on p. 131).

[SZ99] A. Srinivasan and D. Zuckerman. "Computing with Very Weak Random Sources". In: *SIAM Journal on Computing* 28.4 (Jan. 1, 1999), pp. 1433–1459. ISSN: 0097-5397. DOI: 10.1137/S009753979630091X (cit. on pp. 16, 41, 42).

[TV93] M. Teboulle and I. Vajda. "Convergence of Best ϕ-Entropy Estimates". In: *IEEE Transactions on Information Theory* 39.1 (Jan. 1993), pp. 297–301. ISSN: 00189448. DOI: 10.1109/18.179378 (cit. on pp. 119, 160).

[TZS06] A. Ta-Shma, D. Zuckerman, and S. Safra. "Extractors from Reed–Muller Codes". In: *Journal of Computer and System Sciences*. Special Issue on FOCS 2001 72.5 (Aug. 1, 2006), pp. 786–812. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2005.05.010 (cit. on p. 17).

[Vad12] S. P. Vadhan. *Pseudorandomness*. Boston, Mass.: Now Publishers Inc, Oct. 23, 2012. 352 pp. ISBN: 978-1-60198-594-1 (cit. on pp. 16, 25, 29, 30, 36, 43, 45, 54).

[Vaj70] I. Vajda. "Note on Discrimination Information and Variation". In: *IEEE Transactions on Information Theory* 16.6 (Nov. 1970), pp. 771–773. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.1970.1054557 (cit. on p. 120).

[Vaj72] I. Vajda. "On the f-Divergence and Singularity of Probability Measures". In: *Periodica Mathematica Hungarica* 2.1 (Mar. 1, 1972), pp. 223–234. ISSN: 1588-2829. DOI: 10.1007/BF02018663 (cit. on pp. 120, 177, 179).

[Vaj73] I. Vajda. "χ^α-Divergence and Generalized Fisher's Information". In: *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*. Prague: Academia, 1973, pp. 873–886 (cit. on p. 131).

[Val70] M. Valadier. "Intégration de convexes fermés notamment d'épigraphes. Inf-convolution continue". In : *Revue française d'informatique et de recherche opérationnelle* 4 (Série R-2 1970), pp. 57–73 (cit. on p. 160).

[vEH14] T. van Erven and P. Harremoës. "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7 (July 2014), pp. 3797–3820. ISSN: 0018-9448. DOI: 10.1109/TIT.2014.2320500 (cit. on pp. 19, 39).

[Ver14] S. Verdú. "Total Variation Distance and the Distribution of Relative Information". In: *2014 Information Theory and Applications Workshop (ITA)*. Feb. 2014, pp. 1–3. DOI: 10.1109/ITA.2014.6804281 (cit. on p. 38).

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics 47. Cambridge: Cambridge University Press, 2018. ISBN: 978-1-108-41519-4 (cit. on pp. 31, 106, 154, 175).

[vErv10] T. van Erven. "When Data Compression and Statistics Disagree: Two Frequentist Challenges for the Minimum Description Length Principle". PhD thesis. Leiden University, 2010 (cit. on p. 39).

[VZ12] S. Vadhan and C. J. Zheng. "Characterizing Pseudoentropy and Simplifying Pseudorandom Generator Constructions". In: *Proceedings of the 44th Symposium on Theory of Computing - STOC '12*. New York, New York, USA: ACM Press, 2012, p. 817. ISBN: 978-1-4503-1245-5. DOI: 10.1145/2213977.2214051 (cit. on pp. 5, 6, 65–67, 83, 84).

[Wai19] M. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics 48. Cambridge ; New York, NY: Cambridge University Press, 2019. ISBN: 978-1-108-49802-9 (cit. on p. 99).

[Wil38] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *The Annals of Mathematical Statistics* 9.1 (Mar. 1938), pp. 60–62. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177732360 (cit. on pp. 100, 106, 107).

[WZ99] A. Wigderson and D. Zuckerman. "Expanders That Beat the Eigenvalue Bound: Explicit Construction and Applications". In: *Combinatorica* 19.1 (Jan. 1, 1999), pp. 125–138. ISSN: 1439-6912. DOI: 10.1007/s004930050049 (cit. on p. 47).

[Yao82] A. C.-C. Yao. "Theory and Application of Trapdoor Functions". In: *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (Sfcs 1982)*. Nov. 1982, pp. 80–91. DOI: 10.1109/SFCS.1982.45 (cit. on pp. 2, 5, 21, 65).

[Zăl02] C. Zălinescu. *Convex Analysis in General Vector Spaces*. River Edge, N.J. ; London: World Scientific, 2002. 367 pp. ISBN: 978-981-238-067-8 (cit. on pp. 124, 127, 128, 163, 164).

[Zol84] V. M. Zolotarev. "Probability Metrics". In: *Theory of Probability & Its Applications* 28.2 (Jan. 1, 1984), pp. 278–302. ISSN: 0040-585X. DOI: 10.1137/1128025 (cit. on p. 1).

[ZS13] A. M. Zubkov and A. A. Serov. "A Complete Proof of Universal Inequalities for the Distribution Function of the Binomial Law". In: *Theory of Probability & Its Applications* 57.3 (Jan. 1, 2013), pp. 539–544. ISSN: 0040-585X. DOI: 10.1137/S0040585X97986138 (cit. on p. 110).

[Zuc07] D. Zuckerman. "Linear Degree Extractors and the Inapproximability of Max Clique and Chromatic Number". In: *Theory of Computing* 3.1 (Aug. 6, 2007), pp. 103–128. ISSN: 1557-2862. DOI: 10.4086/toc.2007.v003a006 (cit. on p. 17).

[Zuc97] D. Zuckerman. "Randomness-Optimal Oblivious Sampling". In: *Random Structures & Algorithms* 11.4 (1997), pp. 345–367. ISSN: 1098-2418. DOI: 10.1002/(SICI)1098-2418(199712)11:4<345::AID-RSA4>3.0.CO;2-Z (cit. on pp. 5, 12–14, 20, 25, 26, 28, 56, 63).