Proceedings of Machine Learning Research vol 336:1–31, 2026 39th Annual Conference on Learning Theory

Partition Function Estimation under Bounded f-Divergence

Adam Block ADAM.BLOCK@COLUMBIA.EDU
Departments of Computer Science and Electrical Engineering, Columbia University

Abhishek Shetty SHETTY@MIT.EDU
MIT and Georgia Institute of Technology

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the statistical complexity of estimating partition functions given sample access to a proposal distribution and an unnormalized density ratio for a target distribution. While partition function estimation is a classical problem, existing guarantees typically rely on structural assumptions about the domain or model geometry. We instead provide a general, information-theoretic characterization that depends only on the relationship between the proposal and target distributions. Our analysis introduces the integrated coverage profile, a functional that quantifies how much target mass lies in regions where the density ratio is large. We show that integrated coverage tightly characterizes the sample complexity of multiplicative partition function estimation and provide matching lower bounds. We further express these bounds in terms of f-divergences, yielding sharp phase transitions depending on the growth rate of f and recovering classical results as a special case while extending to heavy-tailed regimes. Matching lower bounds establish tightness in all regimes. As applications, we derive improved finite-sample guarantees for importance sampling and self-normalized importance sampling, and we show a strict separation between the complexity of approximate sampling and counting under the same divergence constraints. Our results unify and generalize prior analyses of importance sampling, rejection sampling, and heavy-tailed mean estimation, providing a minimal-assumption theory of partition function estimation. Along the way we introduce new technical tools including new connections between coverage and f-divergences as well as a generalization of the classical Paley-Zygmund inequality.

Keywords: partition function estimation, approximate sampling, importance sampling

1. Introduction

There has been significant attention paid to the problem of partition function estimation in a number of specialized regimes, from classical models arising from statistical mechanics and combinatorics like the Ising model (Jerrum, 2003; Vigoda, 2024) to more general settings in scientific applications with physical structure imposed on λ (Chipot and Pohorille, 2007). The goal, to estimate the normalizing constant of an unnormalized density has numerous applications throughout statistics, machine learning, and computer science, including Bayesian inference (Geweke, 1989; Kass and Raftery, 1995), graphical models (Wainwright and Jordan, 2008), energy based models (LeCun et al., 2006), statistical physics and chemistry (Chipot and Pohorille, 2007; Gelman and Meng, 1998), and reinforcement learning for language model post-training (Chen et al., 2025; Brantley et al., 2025), among many others. Indeed, due to the significant interest, many prior works have been devoted to designing and analyzing algorithms to accomplish this important task. While these works have provided numerous theoretical and practical insights in the problem domains studied, they often rely on structural assumptions on the structure of λ (for example, through assumptions that arise in the context of learning on graphs) or the geometry of $\mathcal{X}$ (for example, via smoothness or other regularity conditions on Euclidean space) that limit their applicability to more general settings. Indeed, a surprising lacuna exists in the current literature: despite the fundamental nature of the partition function estimation problem, there is a dearth of general results that characterize the statistical complexity of partition function estimation in terms of natural and information theoretic properties of the underlying distributions μ and ν. This is especially important given modern applications like language modeling, where the domain is unstructured and λ often corresponds to complex learned models and reward functions (Rafailov et al., 2023; Xie et al., 2024).

In this work, we aim to fill this gap by asking: *How many samples from a base distribution μ are required to estimate the partition function of a target distribution ν to a desired accuracy, as a function of natural information theoretic quantities between μ and ν?* We provide a complete answer to this question by providing tight bounds on the sample complexity n required to achieve this estimate in terms of the coverage profile (Definition 4) between the target distribution ν and the proposal distribution μ, as considered by (Chen et al., 2025; Chatterjee and Diaconis, 2018). The coverage profile quantitatively measures how the mass that ν places on regions where the density ratio ${}^{d\nu}/_{d\mu}$ is large, and thus captures the tail behavior of the density ratio. In particular, we capture this decay in terms of a discrepancy measure between ν and μ we introduce and term *integrated coverage*, $\mathrm{ICov}_M(\nu\|\mu)$ (Definition 4). We work in a setting where we have sample access to μ and can evaluate the *unnormalized* density ratio λ, where $\lambda = Z \cdot {}^{d\nu}/_{d\mu}$ for some unknown normalizing constant $Z = \int \lambda(x) d\mu(x)$. We show that the integrated coverage precisely characterizes the sample complexity of partition function estimation as follows:

Theorem 1 (Informal version of Theorems 6 and 9) *Let μ, ν be two probability measures on $\mathcal{X}$. Then, let M be such that $M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu) \leq \varepsilon$, then $n = \Theta\left(M \cdot \varepsilon^{-1}\right)$ samples are necessary and sufficient to estimate Z to multiplicative accuracy $(1 \pm \varepsilon)$.*

This result not only provides a sharp characterization of the sample complexity of partition function estimation in terms of natural information theoretic quantities, but also unifies, generalizes, and applies several prior results on importance sampling and mean estimation (Chatterjee and Diaconis, 2018; Devroye et al., 2016).

The sharpness of our bound can be further interpreted in terms of more familiar notions of discrepancy between ν and μ, *f-divergences* (Csiszár and Shields, 2004; Polyanskiy and Wu, 2022). Recall that f-divergence (Definition 5) is defined as the expectation of a convex function f of the density ratio, and generalizes notions such as total variation, KL-divergence, and Renyi divergences. Indeed, an informal version of our main result stated in terms of f-divergences is as follows:

Theorem 2 (Informal statement of Theorems 7 and 11) *Given two probability measures μ, ν on $\mathcal{X}$ and an f-divergence $\mathrm{D}_f(\nu\|\mu)$ between them, there is a function γ_f depending on f such that*

$$n = \Theta\left(\left[\gamma_f\left(\Theta(1) \cdot \varepsilon^{-1} \cdot \mathrm{D}_f(\nu\|\mu)\right)\right] \vee \left[\mathrm{D}_{\chi^2}(\nu\|\mu) \cdot \varepsilon^{-2}\right]\right)$$

samples are necessary and sufficient to estimate Z to multiplicative accuracy $(1 \pm \varepsilon)$ with constant probability, where $\mathrm{D}_{\chi^2}(\cdot\|\cdot)$ is the χ^2-divergence.

As an application of these results, we provide a sharper finite-sample analysis of importance sampling and self-normalized importance sampling (SNIS) estimators in terms of the target function and the distributions μ and ν (Theorems 14 and 15) and provides a unified perspective on prior results (Owen, 2013; Chatterjee and Diaconis, 2018). In particular, our results could help inform the design of proposal distributions for importance sampling that minimize the required sample complexity for a given set of target functions that is more flexible than the classical variance-minimizing proposal distribution (Owen, 2013; Llorente and Martino, 2025).

We also make an interesting connection to the sample complexity of sampling from ν given samples from μ given access to unnormalized density ratios λ. Generalizing results by Block and Polyanskiy (2023); Flamich and Wells (2024), we show that we provide a tight characterization of the sample complexity of sampling in terms of coverage and f-divergence as follows:

Theorem 3 (Informal version of Proposition 13) *Let μ, ν be probability measures on $\mathcal{X}$. Let M be such that $\mathrm{Cov}_M(\nu \| \mu)\, \varepsilon$. Then, $n = \tilde{\Theta}(M \cdot \log(1/\varepsilon))$ are necessary and sufficient to produce ε-approximate samples from ν in total variation distance given samples from μ and access to an unnormalized density ratio. In particular, this holds for $n \gtrsim \log(1/\varepsilon) \cdot \gamma_f\left(\Theta(1) \cdot \mathrm{D}_f(\nu\|\mu)/\varepsilon\right)$.*

This result shows that sampling is strictly easier than estimation, in conceptual constrast to settings (such as self-reducibility) where sampling is often approximately has the same complexity as estimation (Jerrum, 2003; Vigoda, 2024). We provide a detailed comparison of the sample complexity of sampling and estimation in Section 3.3.

We begin in the next section by formally setting up the problem of interest and introducing and defining the key quantitites used throughout the paper. We then proceed to state our main results in Section 3, including both upper and lower bounds on the sample complexity of partition function estimation, as well as bounds on approximate sampling. We proceed in Section 4 to discuss applications of our results to importance sampling and self-normalized importance sampling. We then provide a high-level overview of the proof techniques used to establish our main upper bounds in Section 5, deferring full proofs of all results to the appendix. We conclude in Section 6 with a discussion of open questions and future directions. Our proofs introduce several novel technical tools of independent interest, including a connection between f-divergences and integrated coverage, a variance bound for truncated density ratios (Lemma 20), and a strong generalization of Paley-Zygmund (Lemma 16).

2. Problem Setup and Preliminaries

We consider a measurable space $\mathcal{X}$ and base probability measure μ over $\mathcal{X}$. We will also consider a target distribution ν over $\mathcal{X}$ from which we would either like to produce a sample or to estimate its normalizing constant given access to samples from μ and an unnormalized density ratio $\lambda : \mathcal{X} \to \mathbb{R}_{\geq 0}$ defined using the Radon-Nikodym derivative of ν with respect to μ as $\lambda(x) = Z \cdot {}^{d\nu}/_{d\mu}(x)$ for some unknown normalizing constant Z. We are primarily interested in understanding the difficulty of obtaining an estimate $\widehat{Z}$ of the normalizing constant Z; more precisely, we ask the following question: *Given access to i.i.d. samples $X_1, \ldots, X_n \sim \mu$ and the ability to evaluate an unnormalized density ratio λ on each sample, how large must n as a function of μ, ν, ε and δ be to ensure that there exists an estimator $\widehat{Z} = \widehat{Z}(X_1, \ldots, X_n, \lambda(X_1), \ldots, \lambda(X_n))$ satisfies $(1-\varepsilon)Z \leq \widehat{Z} \leq (1+\varepsilon)Z$ with probability at least $1-\delta$?*

The rate at which such an estimate can be obtained depends on the similarity between the two distributions μ and ν. For example, if μ and ν had disjoint supports, then no finite number of samples from μ would suffice to estimate Z. The main focus of this work is provide a sharp characterization of the sample complexity in terms of quantitative measures of similarity between μ and ν.

The key notion measuring the divergence between μ and ν is that of *coverage*, considered by Chatterjee and Diaconis (2018); Chen et al. (2025). Conceptually, coverage measures the mass that ν places on regions where the density ratio $\frac{d\nu}{d\mu}$ is large and thus good coverage implies that μ places sufficient mass in high-density regions of ν. Formally, coverage is defined as follows.

Definition 4 (Coverage and Integrated Coverage) *For two probability measures μ, ν on $\mathcal{X}$ and $M > 0$, the* coverage *function at M is defined to be*

$$\mathrm{Cov}_M\left(\nu \| \mu\right) = \nu\left(\{x \in \mathcal{X} : {}^{d\nu}/_{d\mu}(x) \geq M\}\right),$$

where ${}^{d\nu}/_{d\mu}$ is the Radon-Nikodym derivative of ν with respect to μ. Further, we introduce a notion that we term the integrated coverage profile as

$$\mathrm{ICov}_M\left(\nu \| \mu\right) = \int_0^M \mathrm{Cov}_t\left(\nu \| \mu\right) dt.$$

We first note several basic properties of coverage and integrated coverage. First, it is immediate that $\mathrm{Cov}_M(\nu \| \mu) \leq 1$ for all $M > 0$, and thus $\mathrm{ICov}_M(\nu \| \mu) \leq M$; moreover, $\mathrm{Cov}_M(\nu \| \mu)$ is clearly non-increasing in M. Furthermore, an easy calculation implies that $M \mapsto {}^{\mathrm{ICov}_M(\nu\|\mu)}/_M$ is also non-increasing in M and whenever $\nu \ll \mu$, this quantity tends to zero as $M \to \infty$. We will see that this latter map is fundamental to characterizing the sample complexity of partition function estimation.

The notion of coverage has previously been used to analyze the sample complexity of importance sampling (Chatterjee and Diaconis, 2018), the sample complexity of rejection sampling Block and Polyanskiy (2023) and, interestingly, for understanding the efficacy of pretrained language models Chen et al. (2025); Huang et al. (2025a,b). The notion of integrated coverage, introduced here, is a more refined measure of the relationship between μ and ν that allows for sharper characterizations of sample complexity.

A more standard notion of discrepancy between two measures is the f-divergence (Csiszár and Shields, 2004; Polyanskiy and Wu, 2022), which is defined as follows.

Definition 5 *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a convex function satisfying $f(1) = f'(1) = 0$. For two probability measures μ, ν on $\mathcal{X}$, the f-divergence between ν and μ is defined to be*

$$\mathrm{D}_f\left(\nu \| \mu\right) = \mathbb{E}_{X \sim \mu}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right] + \nu\left(\frac{d\nu}{d\mu} = \infty\right) \cdot f'(\infty). \tag{1}$$

The f-divergences generalize many well-known notions of distance between probability measures, including total variation distance (where $f(t) = {}^1/_2\,|t-1|$), the Kullback-Leibler (KL) divergence (where $f(t) = t \log t$), and (a monotone transformation of) the α-Renyi divergences (where $f(t) = t^\alpha - \alpha t$ for $\alpha > 0$, $\alpha \neq 1$). Of special note is the α-Renyi divergence with $\alpha = 2$, which corresponds to the χ^2 divergence, defined as $\chi^2(\nu \| \mu) = \mathbb{E}_{X \sim \mu}\left[\left({}^{d\nu}/_{d\mu}(X) - 1\right)^2\right]$.

Conceptually, a bounded f-divergence controls the tail behavior of the density ratio $\frac{d\nu}{d\mu}$ with faster growing f leading to stronger control. As we show in Appendix C, this intuition is made precise in several ways, most critically in the observation that $\mathrm{Cov}_M(\nu \| \mu) \leq {}^{M \cdot \mathrm{D}_f(\nu\|\mu)}/_M$ for any f-divergence; thus, bounded f-divergence implies coverage decays quickly. As we shall see, our sample complexity results involving the f-divergences will depend on the growth rate of f through that which we term the γ_f function, defined to be the inverse of the map $t \mapsto f(t)/t$ on $[1, \infty)$, i.e.,

$$\gamma_f(M) = \inf\left\{t \geq 1 : f(t)/t \geq M\right\}. \tag{2}$$

The growth rate of f determines the behavior of γ_f; for example, when f is *superlinear*, i.e., $\lim_{t\to\infty} f(t)/t = \infty$, the function γ_f is well-defined on all of $\mathbb{R}_{\geq 0}$; examples of superlinear f include those corresponding to KL divergence (where $\gamma_f(M) \asymp \exp(M)$) and α-Renyi divergences with $\alpha > 1$ (where $\gamma_f(M) \asymp M^{\frac{1}{\alpha-1}}$). We will also have occasion to distinguish between *superquadratic* f, where $\lim_{t\to\infty} f(t)/t^2 = \infty$ (e.g., α-Renyi divergences with $\alpha > 2$) and those that are not (e.g., KL divergence and α-Renyi divergences with $1 < \alpha \leq 2$ such as χ^2). In the case that f is *linear*, we let $\gamma_f(M) = \infty$ for $M > \sup_t f(t)/t$; note that by the assumption of convexity, f cannot be sublinear. These distinctions will control the precise rates in our sample complexity results.

3. Main Results

We now present our main results regarding the sample complexity of estimating the partition function given access to samples from the base distribution and an unnormalized density ratio.

3.1. Upper Bounds for Estimation

We begin by stating our main upper bound on the sample complexity of estimating the partition function in terms of the integrated coverage profile defined in Definition 4.

Theorem 6 *Let μ, ν be probability measures on $\mathcal{X}$. Suppose that we have access to i.i.d. samples $X_1, \ldots, X_n \sim \mu$ and an unnormalized density ratio λ. Let M_ε be such that $M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu) \leq \varepsilon/4$. Then there is an estimator $\widehat{Z}$ such that as long as*

$$n \gtrsim M_\varepsilon \cdot \log(1/\delta)/\varepsilon, \tag{3}$$

then with probability at least $1 - \delta$ it holds that $(1 - \varepsilon) \cdot Z \leq \widehat{Z} \leq (1 + \varepsilon)Z$.

As we discussed above, as long as $\nu \ll \mu$, such an M_ε always exists for any positive ε; thus the above theorem always provides a finite sample complexity bound for estimating the partition function to $(1 \pm \varepsilon)$ accuracy in the absolutely continuous setting.

In order to appreciate (3), consider the simple case when $\chi^2(\nu||\mu)$ is bounded. In this case, a simple tail bound implies $\mathrm{ICov}_M(\nu\|\mu) \leq \chi^2(\nu||\mu)$ for all M. Therefore, we can set $M_\varepsilon \asymp \chi^2(\nu||\mu)/\varepsilon$ in Theorem 6, and obtain that $n \gtrsim \chi^2(\nu||\mu)/\varepsilon^2$ samples suffice to estimate the partition function to $(1 \pm \varepsilon)$ accuracy with high probability. This recovers standard results bounding the sample complexity of importance sampling by computing the variance (Owen, 2013). It is natural to ask, then, how slowly can M_ε grow as a function of ε? Perhaps one might hope that with improved moment bounds, one could achieve a faster estimation rate via M_ε, but this is not the case. Indeed, this can be seen intuitively by observing that, because $\mathrm{Cov}_0(\nu\|\mu) = 1$ for any two distributions ν, μ, we must have $M_\varepsilon \gtrsim 1/\varepsilon$ as $\varepsilon \downarrow 0$ and Theorem 6 can never achieve a sample complexity scaling better than $\Omega(\varepsilon^{-2})$, as one would expect for estimating a mean to $(1 \pm \varepsilon)$ accuracy due to the central limit theorem. Thus, we should instead view the above theorem as a generalization of the bound to more general heavy-tailed settings where the χ^2 divergence may be infinite.

We may thus ask whether we can provide more explicit bounds on the sample complexity in terms of the more familiar notion of f-divergences, which amount to control on the tails of the density ratio. This is the content of our next theorem.

Theorem 7 *Let μ, ν be probability measures on $\mathcal{X}$ and let f be a convex function as in Definition 5 such that there exists*[1] $c \geq 1$ *such that for $t \geq c$ the function $t \mapsto f(t)/t^2$ is non-increasing. Suppose that we have access to i.i.d. samples $X_1, \ldots, X_n \sim \mu$ and an unnormalized density ratio λ. Then there is an estimator $\widehat{Z}$ such that as long as*

$$n \gtrsim \frac{\gamma_f\left({}^{6 \cdot \mathrm{D}_f(\nu \| \mu)}/_{\varepsilon}\right) \cdot \log\left({}^{1}/_{\delta}\right)}{\varepsilon} \vee \frac{c^2}{\varepsilon^2}, \tag{4}$$

then with probability at least $1 - \delta$ it holds that $(1 - \varepsilon) \cdot Z \leq \widehat{Z} \leq (1 + \varepsilon) Z$.

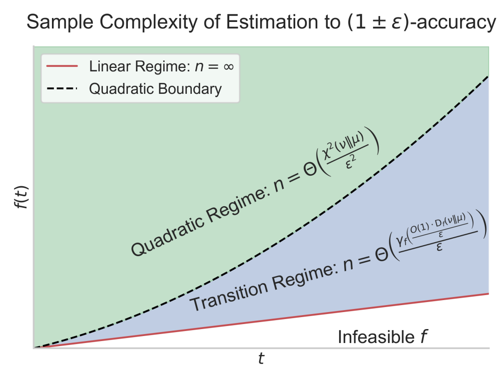

Figure 1: Sample complexity (n) regimes for estimating Z to $(1 \pm \varepsilon)$ accuracy based on the growth rate of the f defining $\mathrm{D}_f(\nu \| \mu)$.

The f-divergence perspective provides a complementary view to the coverage-based bound of Theorem 6. We can separate the sample complexity of estimating the partition function into three regimes (summarized in Figure 1) depending on the growth rate of f: *linear*, *superlinear* but *subquadratic*, and *superquadratic*. Indeed, in the case where f is linear, i.e., $\lim_{t \to \infty} {}^{f(t)}/_{t} < \infty$, the function γ_f is only defined on a bounded domain and thus for sufficiently small ε, the theorem is vaccuous; for example, this is the case for total variation or Hellinger distances. In the superlinear but subquadratic case, where $\lim_{t \to \infty} {}^{f(t)}/_{t^2} = C < \infty$, we see that γ_f is defined on all of $\mathbb{R}_{\geq 0}$ and grows at least as quickly as ε^{-1}, implying that the first term of (4) dominates. This is the case for KL divergence where $n \gtrsim \exp(c \cdot {}^{\mathrm{D_{KL}}(\nu \| \mu)}/_{\varepsilon})/\varepsilon$ samples suffice as well as for α-Renyi divergences with $1 < \alpha \leq 2$, where $n \gtrsim \varepsilon^{-\alpha} \cdot \mathrm{D}_f\left(\nu \| \mu\right)^{1/\alpha - 1}$ samples suffice. Finally, in the superquadratic case where $\lim_{t \to \infty} {}^{f(t)}/_{t^2} = \infty$, we see that γ_f grows too slowly to dominate the sample complexity for small ε and thus the second term of (4) kicks in; this is the case for α-Renyi divergences with $\alpha > 2$, where $n \gtrsim 1/\varepsilon^2$ samples suffice.

Note that our tightest upper bounds can never improve on the ε^{-2} scaling for small ε, as discussed above. We thus conclude the discussion of our upper bounds with an *asymmetric* approach that is capable of improving the sample complexity scaling in ε for the lower tail at the cost of a significantly worse upper tail bound.

Theorem 8 *Let μ, ν be probability measures on $\mathcal{X}$ and suppose $X_1, \ldots, X_n \sim \mu$ are independent with λ an unnormalized density ratio. Then there is an estimator $\widehat{Z}$ such that for any $0 < \varepsilon < 1$, if*

$$n \gtrsim \log\left({}^{1}/_{\delta}\right) \cdot {}^{M}/_{\varepsilon} \quad \textit{where } \operatorname{Cov}_M\left(\nu \| \mu\right) \lesssim \varepsilon, \tag{5}$$

1. The actual constant $c \geq 1$ in Theorem 7 is a technical artifact and should generally thought to be equal to 1. Indeed, this condition is simply ensuring that $f(t)$ is eventually subquadratic, as discussed below.

it holds that with probability at least $1 - \delta$ that $(1 - \varepsilon) \cdot Z \leq \widehat{Z} \leq M \cdot Z$. In particular, if $\mathrm{D}_f(\nu \| \mu) < \infty$, then it suffices to take $M = \gamma_f\left(6 \cdot \mathrm{D}_f(\nu\|\mu)/\varepsilon\right)$.

Note that the M in (5) is no larger than M_ε in Theorem 6, because $\mathrm{ICov}_M(\nu\|\mu) \leq M \cdot \mathrm{Cov}_M(\nu\|\mu)$ for all M by the fact that coverage is decreasing. Thus, in the setting where f is superquadratic, we can achieve a sample complexity of estimating Z from below that is significantly smaller than the ε^{-2} scaling of Theorems 6 and 7.

3.2. Lower Bounds

In order to complement our upper bounds, we provide several lower bounds on the sample complexity of estimating the partition function, both in terms of integrated coverage and in terms of the f-divergences that demonstrate the tightness of our bounds. The lower bound constructions, deferred to Appendix D, are inspired in part by Block and Polyanskiy (2023) and in particular the result of Harremoës and Vajda (2011) that suggests that pairs of Bernoulli random variables extremize the joint range of f-divergences. We begin with the integrated coverage lower bound.

Theorem 9 *For any non-atomic μ and any $\varepsilon \leq 1/2$, there exists a family of distributions ν such that any estimator $\widehat{Z}$ that achieves $(1 \pm \varepsilon)$ multiplicative accuracy with probability at least $2/3$ requires at least $\varepsilon^{-1} \cdot M_\varepsilon$ samples, where M_ε satisfies*

$$\mathrm{ICov}_{M_\varepsilon}(\nu\|\mu) \leq M_\varepsilon \cdot \varepsilon.$$

In particular, Theorem 9 shows that the sample complexity bound of Theorem 6 is tight in its dependence on integrated coverage and ε. Unfortunately, Theorem 9 does not directly translate to lower bounds in terms of f-divergences; while it is indeed the case that integrated coverage can be tightly controlled via f-divergence, it is not clear that this tight control is realized by the lower bound construction of Theorem 9. We thus complement Theorem 7 with separate lower bounds for each of the three regimes discussed above, beginning with the linear case.

Proposition 10 *Let f be a convex function as in Definition 5 such that $t \mapsto f(t)/t$ is bounded. For any $0 < C < f'(\infty)$, there exist measures μ, ν such that $\mathrm{D}_f(\nu\|\mu) \leq C$ such that no estimator given access to finitely many samples from μ is capable of producing a $1 \pm \varepsilon$ approximation of Z with probability at least $3/4$ for any $\varepsilon \leq f'(\infty)/(f'(\infty) - C)$. Moreover, if $C \geq f'(\infty)$, then there exist measures μ, ν such that $\mathrm{D}_f(\nu\|\mu) \leq C$ and no such estimator can produce a $1 \pm \varepsilon$ approximation of Z with probability at least $3/4$ for any $\varepsilon > 0$.*

Proposition 10 shows that when f is linear, no finite number of samples suffices to estimate the partition function to within any nontrivial multiplicative accuracy. This fact is unsurprising given that f-divergences with linear f (such as total variation or Hellinger) do not control the tails of the density ratio $\frac{d\nu}{d\mu}$ in any meaningful way and indeed allow for the possibility that ν is singular with respect to μ, where clearly estimation of Z is impossible. We now proceed to the second case, where f is superlinear but subquadratic.

Theorem 11 *For any $C > 2$ and $\varepsilon \in (0, \frac{1}{4})$, there exists a probability measure μ on $\mathcal{X}$ and class of probability measures $\mathcal{V}$ on $\mathcal{X}$ such that for all $\nu \in \mathcal{V}$, it holds that $\mathrm{D}_f(\nu\|\mu) \leq C + f(1-\varepsilon)$ and any estimator $\widehat{Z}$ that satisfies $(1-\varepsilon)Z \leq \widehat{Z} \leq (1+\varepsilon)Z$ with probability at least $\frac{2}{3}$ for all $\nu \in \mathcal{V}$ must use at least $n \gtrsim \gamma_f(\frac{C}{2 \cdot \varepsilon})/\varepsilon$ samples.*

While this lower bound holds for arbitrary f-divergences, it is only tight in the superlinear/subquadratic regime, where $\gamma_f(1/\varepsilon) \cdot \varepsilon^{-1} \gg \varepsilon^{-2}$; this applies to KL-divergence and Renyi divergences for $1 < \alpha \leq 2$. For the superquadratic case, we have the following lower bound.

Proposition 12 *Fix $0 < \varepsilon < 1/8$. Then there exist a family of probability measures $\mathcal{U}$ on $\mathcal{X}$ and probability measure ν on $\mathcal{X}$ such that for all $\mu \in \mathcal{U}$, it holds that $\|d\nu/d\mu\|_\infty \leq 4$ and any estimator $\widehat{Z}$ that for all $\nu \in \mathcal{V}$ with probability at least $2/3$ is a $1 \pm \varepsilon$ approximation of Z must use at least $n \gtrsim \frac{1}{\varepsilon^2}$ samples.*

Taken together, the lower bounds presented in this section demonstrate that each of our upper bounds are tight in their respective regimes both in terms of integrated coverage and f-divergences.

3.3. Comparison to Sampling

While our primary focus is on partition function estimation, it is instructive to compare our results to those for the related problem of *sampling* from the target distribution ν given access to samples from the base distribution μ and an unnormalized density ratio λ. In Block and Polyanskiy (2023), the authors examined the question of *approximate rejection sampling*, where the learner is given access to n i.i.d. samples from μ and the *normalized* density ratio $\frac{d\nu}{d\mu}(\cdot)$ and must produce a sample from a distribution ν_n that is close to ν in total variation distance under the assumption of bounded f-divergence between ν and μ. We generalize their result, as well as that of Flamich and Wells (2024) to the setting of *unnormalized* density ratios and arbitrary f-divergences.

Proposition 13 *Let μ, ν be probability measures on $\mathcal{X}$ and let f be a convex function as in Definition 5. Suppose that we have access to i.i.d. samples $X_1, \dots, X_n \sim \mu$ and an unnormalized density ratio λ. Then there is an algorithm that produces a sample $X_{\widehat{j}}$ from a distribution ν_n such that $\mathrm{TV}(\nu_n, \nu) \leq \varepsilon$ as long as*

$$n \gtrsim M \cdot \log(1/\varepsilon) \quad \textit{where } \operatorname{Cov}_M(\nu\|\mu) \lesssim \varepsilon. \tag{6}$$

In particular, this holds for

$$n \gtrsim \log(1/\varepsilon) \cdot \gamma_f\left(\Theta(1) \cdot \mathrm{D}_f(\nu\|\mu)/\varepsilon\right). \tag{7}$$

We observe that the lower bounds in Block and Polyanskiy (2023) for approximate rejection sampling already demonstrate that (7) is tight up to logarithmic factors as the regime in that work is strictly stronger than ours (i.e., access to the normalized density ratio rather than the unnormalized one). We defer a proof of Proposition 13 to Appendix F, where we generalize analysis of the A^*-sampling algorithm in Li and El Gamal (2018); Flamich and Wells (2024).

It is instructive to note that the sample complexity of sampling is strictly smaller than that of partition function estimation. To see this note that $\operatorname{Cov}_M(\nu\|\mu) \leq M^{-1} \cdot \operatorname{ICov}_M(\nu\|\mu)$ for all M, since $\operatorname{Cov}_M(\nu\|\mu)$ is a decreasing function of M. Thus it follows immediately that we have at least an ε^{-1} factor improvement in sample complexity when comparing (6) to (4) in Theorem 6. The improvement can be even more stark, however, depending on the behavior of the coverage profile. Indeed, we saw that $M^{-1} \cdot \operatorname{ICov}_M(\nu\|\mu)$ can never shrink more quickly than M^{-1} as M increases; thus, in the superquadratic regime, where $\mathrm{D}_f(\nu\|\mu)$ is bounded for quickly-growing f, we can have almost a quadratic separation between the sample complexity of sampling and that of partition function estimation. Indeed, in the extreme case where the density ratio is uniformly bounded, i.e., $\|{}^{d\nu}/_{d\mu}\|_\infty \le M$, sampling requires only $n \gtrsim \log(1/\varepsilon)$ samples to produce an ε-approximate sample from ν in total variation distance, while partition function estimation requires $n \gtrsim \varepsilon^{-2}$ samples. One way to interpret this is that estimation depends on the entire coverage profile, while sampling only depends on the coverage at a single value of M.

In this way we see that, in contradistinction to the well-known class of 'self-reducible' problems, the problem of 'counting' (i.e., partition function estimation) is strictly harder than that of 'sampling' under general f-divergence or coverage constraints.

4. Application: Improved Finite Sample Bounds for Importance Sampling

Importance sampling is a fundamental technique in statistics and machine learning which serves as a primitive in many domains, including causal inference, reinforcement learning (Precup et al., 2000; Thomas et al., 2015), variational inference (Burda et al., 2016), and probabilistic programming (Wingate et al., 2011). It serves as a tool for estimating expectations of a function g (we focus on the bounded, positive case) under a target distribution ν using samples from a proposal distribution μ. Given i.i.d. samples $X_1, \dots, X_n \sim \mu$, the importance sampling estimator for $\nu_g = \mathbb{E}_\nu[g(X)]$ is

$$\hat{g}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d\nu}{d\mu}(X_i) g(X_i).$$

This estimator is unbiased, i.e., $\mathbb{E}\left[\hat{g}_{\mathrm{IS}}\right] = \nu_g$ and standard results on importance sampling estimators bound the variance of $\hat{g}_{\mathrm{IS}}$ which can be bounded rely on the χ^2-divergence between the target and proposal distributions, since the variance is bounded by $\chi^2(\nu\|\mu)$.

Towards obtaining sharper bounds, we consider the target distribution weighted by the function g, i.e., the measure $\nu \cdot g$ defined with density given by ${}^{g(x)}/_{\nu_g} \cdot {}^{d\nu}/_{d\mu}(x)$ with respect to μ. This allows us to define the coverage between the weighted target distribution $\nu \cdot g$ and the proposal distribution μ, which in this context maybe expressed as

$$\mathrm{Cov}_M\left(\nu \cdot g \| \mu\right) = \frac{1}{\nu_g} \int g(x)\frac{d\nu}{d\mu}(x)\mathbb{I}\left[g(x)\frac{d\nu}{d\mu}(x) \ge M \cdot \nu_g\right] d\mu(x)$$

Further this allows us to define the integrated coverage and the f-divergence $\mathrm{D}_f\left(\nu \cdot g\|\mu\right)$ between the weighted target distribution $\nu \cdot g$ and the proposal distribution μ in the natural way.

Applying Theorem 6, we obtain finite-sample bounds for importance sampling estimators in terms of general f-divergences.

Theorem 14 (Importance Sampling) *Let μ, ν be measures and g be a bounded, positive function. Then the importance sampling estimator $\hat{g}_{\mathrm{IS}}$ satisfies $|\hat{g}_{\mathrm{IS}} - \nu_g| \le \varepsilon \cdot \nu_g$ with probability at least $1-\delta$, as long as $n \gtrsim {}^{M_{\varepsilon,\delta}}/_{\varepsilon}$. where $M_{\varepsilon,\delta}$ satisfies $\mathrm{ICov}_{M_{\varepsilon,\delta}}\left(g\nu\|\mu\right) \le {}^{\varepsilon\delta}/_{6}$.*

This generalizes the standard variance bounds for importance sampling, which correspond to the case of the χ^2-divergence. The dependence on probability parameter δ can be improved to $\log(1/\delta)$ using the median of means technique as in the proof of Theorem 6 but state it in this form to maintain the form of the estimator that is commonly used in practice.

More importantly, our results bound the error as jointly as function of g, ν and μ, rather than separately e.g. in terms of $D_f(\nu\|\mu)$ and $\|g\|$ Chatterjee and Diaconis (2018). This is particularly useful in settings when we get to design the proposal distribution μ based on the target functions g and distribution ν, bound above can be used to optimize the proposal distribution μ to minimize the required sample complexity for a given set of target functions g. In particular, the optimal proposal distribution is a well studied problem in importance sampling (See (Owen, 2013; Llorente and Martino, 2025) for a survey-level treatment) and is typically chosen to minimize the variance of the importance sampling estimator. Our results suggest that a more refined objective is to minimize the integrated coverage between the weighted target distribution.

Our techniques also extend to self-normalized importance sampling (SNIS) Owen (2013), which estimates expectations $\mathbb{E}_\nu[g(X)]$ when we can only sample from a proposal distribution μ and evaluate the unnormalized density ratio $\lambda(x) = \frac{d\nu}{d\mu}(x) \cdot Z$ for some unknown normalizing constant $Z = \int \lambda(x) d\mu(x)$. Given i.i.d. samples $X_1, \ldots, X_n \sim \mu$, the SNIS estimator is $\hat{g}_{\text{SNIS}} = \sum_{i=1}^n \lambda(X_i) g(X_i) / \sum_{i=1}^n \lambda(X_i)$. Under the assumption that μ is absolutely continuous with respect to ν, the SNIS estimator is asymptotically unbiased (Owen, 2013, Theorem 9.2). The asymptotic variance is $\mathbb{E}_{X\sim\mu}\left[(g(X) - \mathbb{E}_{X\sim\nu}[g(X)])^2 \cdot \left(\frac{d\nu}{d\mu}(X)\right)^2\right]$ which for bounded functions g is at most $\|g\|_\infty^2 \cdot \chi^2(\nu\|\mu)$. These asymptotic bounds yield finite-sample sample complexity of $n \geq \frac{4\|g\|_\infty^2 \chi^2(\nu\|\mu)}{\delta\varepsilon^2}$ (Metelli et al., 2020, Proposition 9).

Our results yield an improved finite-sample analysis of the SNIS estimator in terms of general f-divergences, applicable even when the χ^2-divergence is infinite.

Theorem 15 (Self-Normalized Importance Sampling) *Let μ, ν be measures and g be a bounded, positive function. Then the self-normalized importance sampling estimator $\hat{g}_{\text{SNIS}}$ satisfies $|\hat{g}_{\text{SNIS}} - \nu_g| \leq \varepsilon \cdot \nu_g$ with probability at least $1 - \delta$, as long as $n \gtrsim M_{\varepsilon,\delta}/\varepsilon$, where $M_{\varepsilon,\delta}$ satisfies $\mathrm{ICov}_{M_{\varepsilon,\delta}}(g\nu\|\mu) \leq \frac{\varepsilon\delta}{6}$ and $\mathrm{ICov}_{M_{\varepsilon,\delta}}(\nu\|\mu) \leq \frac{\varepsilon\delta}{6}$.*

Note that the requirement on the integrated coverage of both the weighted target distribution $g\nu$ and the target distribution ν is necessary since the SNIS estimator is a ratio of two importance sampling estimators where the denominator estimates the normalizing constant Z.

5. Proof Techniques for Upper Bounds

In this section, we outline some of the key proof techniques used to establish our main upper bounds, with a focus on technical results that may be of independent interest. We defer full proofs of all results to the appendix.

5.1. Upper Bounds on Sample Complexity

Our upper bound relies on the same general proof strategy and the same estimator, namely the *median-of-means* (Lugosi and Mendelson, 2019; Alon et al., 1996) Formally, given n samples from μ, we partition them into k groups of size $m = \lfloor n/k \rfloor$ and compute the sample mean within each group, i.e., for each group $j \in [k]$, we let

$$\widehat{Z}_j = \frac{1}{m} \sum_{i=1}^{m} \lambda(X_i^{(j)}) \quad \text{and} \quad \widehat{Z} = \mathrm{Med}\left(\widehat{Z}_1, \ldots, \widehat{Z}_k\right), \tag{8}$$

where $\mathrm{Med}(\cdot)$ denotes the median operator. Using standard analysis (cf. Lemma 19), it suffices to show that each group mean $\widehat{Z}_j$ is within a constant factor of the true partition function Z with constant probability, say $2/3$. Thus, the main technical challenge is to analyze the concentration of the sample mean $\widehat{Z}_j$ around Z. While we defer full proofs of Theorems 6 and 7 to Appendix B and section C.2, we sketch the main ideas below.

Proof Sketch of Theorem 6. The first natural approach to analyzing the concentration of $\widehat{Z}_j$ would be to apply standard concentration inequalities such as Chernoff bounds. Unformately as discussed earlier, this approach fails because the density ratio $d\nu/d\mu$ may be heavy-tailed, and thus the variance of the summands may be infinite. In order to overcome this challenge, we look at the concentration of truncated versions of the density ratio; note that this truncation purely an analytical tool and the estimator $\widehat{Z}_j$ does not involve any truncation. We thus analyze the 'bulk' of the density ratios and the 'tail' separately. To lower bound $\widehat{Z}_j$, we can ignore the tail entirely, while to upper bound the estimate, we apply Markov's inequality to the definition of coverage.

Thus, much of the effort arises in controlling the average of the truncated density ratios, whose expectation can be appropriately controlled via the definition of coverage. One might at first like to apply a standard Chernoff bound, because truncated at M the summands are bounded, but this could at best achieve a sample complexity scaling as M/ε^2 where M is such that $\mathrm{Cov}_M(\nu\|\mu) \leq \varepsilon$, which is loose. Instead, we apply Bernstein's inequality and bound the *variance* of the truncated density ratio, which we show can be controlled by the integrated coverage itself (Lemma 20). This interesting *self-normalization* type property that relates the variance of the truncated density ratio to the bias introduced by truncation is the key technical tool that allows us to improve the dependence on ε in the sample complexity. ■

We can then use this result to prove Theorem 7 by relating the integrated coverage to the f-divergence, as we sketch below.

Proof Sketch of Theorem 7. In order to get a sample complexity that depends on the f-divergence, we need to relate the integrated coverage to the f-divergence. At first glance, this seems simple since an application of Markov's inequality directly relates the two quantities (Lemma 22). Unfortunately, this direct relationship is not sufficient to get us the desired bound on the integrated coverage. Towards this end, we need to use the growth properties of the function f. In particular, since we are interested primarily in the subquadratic regime, we use the assumption that $t^{-2} \cdot f(t)$ is decreasing for $t \geq c$ to relate the integrated coverage to the f-divergence as shown in Lemma 24. ■

5.2. Asymmetric Estimation via Coverage

In order to obtain improved dependence on ε, in Theorem 8 we relaxed the symmetric estimation requirement to only tightly controlling the lower tail while allowing a looser constant factor upper tail. The technical approach is built on the following lemma which can be read as a generalization of the Paley-Zygmund inequality to f-divergences, whose proof can be found in Section E.1.

Lemma 16 *Let μ, ν be probability measures on $\mathcal{X}$ and let f be a convex function as in Definition 5. Then for any $0 < \varepsilon, u < 1$, it holds that*

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq 1-\varepsilon\right) \geq {}^{(1-u)\varepsilon}/_{M} \quad \textit{where } \mathrm{Cov}_M(\nu\|\mu) \leq u\cdot\varepsilon.$$

In particular, $\mathbb{P}_{X\sim\mu}\left(d\nu/d\mu(X) \geq 1-\varepsilon\right) \geq \sup_{0<u<1} {}^{(1-u)\cdot\varepsilon}/_{\gamma_f(\mathrm{D}_f(\nu\|\mu)/u\cdot\varepsilon)}$.

While Lemma 16 is phrased in terms of f-divergences and likelihood ratios, it could otherwise be stated for arbitrary nonnegative random variables, providing a lower bound on the probability that such a variable exceeds a $(1-\varepsilon)$ fraction of its mean. Indeed, taking $f(t) = (t-1)^2$ and letting $Y \geq 0$, we see that we recover Paley-Zygmund up to a factor of 4. We similarly recover the standard generalization of Paley-Zygmund to higher moments by taking $f(t) = t^p - 1$ for $p > 1$. In the case that we only assume a $|X|\log(|X|)$ moment, we derive a bound that, to the best of our knowledge, is novel, showing that a nonnegative random variable exceeds $(1-\varepsilon)$ times its mean with probability at least $\varepsilon \cdot \exp\left(-O(1) \cdot \mathbb{E}[X\log(X)]/\varepsilon\cdot\mathbb{E}[X]\right)$. We now sketch the proof of Theorem 8 using Lemma 16; full details can be found in Appendix E.

Proof Sketch of Theorem 8. Unlike the previous upper bounds, we do not use a median-of-means estimator here, instead simply taking the $(1-\alpha)$-quantile of the samples for appropriately chosen α. The key idea is to use Lemma 16 to lower bound the probability that a single sample exceeds $(1-\varepsilon)Z$, which then allows us to control the lower tail of the quantile estimator via a Chernoff bound. We then control the upper tail of the estimator via the definition of coverage to ensure that the quantile does not exceed $M \cdot Z$ with high probability. ■

6. Discussion and Related Work

In this work, we have provided a complete characterization of the sample complexity of partition function estimation in terms of integrated coverage and f-divergences between the target and proposal distributions. We now discuss related work.

Partition Function Estimation. Partition function estimation is a classical problem with roots dating back to Gibbs and Boltzman in statistical mechanics, with many classical techniques developed and analyzed in the intervening years (Neal, 2001; Gelman and Meng, 1998; Jerrum, 2003; Owen, 2013). One common line of work in theoretical computer science and statistical mechanics assumes some kind of discrete structure and constructs sophisticated annealing schedules and samplers in order to estimate partition functions of complex models such as the Ising model or counting problems such as the permanent of a matrix (Jerrum, 2003; Bezáková et al., 2008; Huber, 2015; Kolmogorov, 2018; Štefankovič et al., 2009). Another line of work, especially in scientific and Bayesian applications, focuses on continuous distributions in Euclidean space and makes use of strong regularity assumptions in order to ensure quality (Chehab et al., 2023; Grosse et al., 2013; Liu et al., 2015). In contradistinction to those works, we focus on a minimal assumption setting where we characterize the difficulty of partition function estimation in terms of purely information theoretic quantities, without making any structural or regularity assumptions.

Sampling and Importance Sampling. In parallel to partition function estimation, the closely related problem of approximate sampling from a target distribution ν given sample access to a proposal distribution μ and the ability to evaluate the density ratio $d\nu/d\mu$ in order to construct mean estimates under target distributions has also been extensively studied (Knuth, 1976; Cappé et al., 2005; Owen, 2013). Most relevant to our work is those of Devroye et al. (2016); Chatterjee and Diaconis (2018). The former, while not explicitly about partition function estimation, considers heavy-tailed mean estimation under polynomial moment constraints; our f-divergence recover their results when considering Renyi divergences. On the other hand, Chatterjee and Diaconis (2018) consider importance sampling and examine the sample complexity under KL constraints, proving that exponentially many samples in the KL divergence are necessary and sufficient for importance sampling to succeed, but do not consider more general f-divergences or coverage. Our results generalize and sharpen theirs, recovering their bounds as a special case and unifying these paradigms. More recently, Block and Polyanskiy (2023) studied the problem of approximate rejection sampling, which is closely related to our results, except they assume *normalized* density ratio access. On the other hand, Li and El Gamal (2018) consider sampling with unnormalized density ratios but assume bounded density ratios, whereas Flamich and Wells (2024) generalizes this to a bounded KL constraint. We apply the same algorithm, introduced in Maddison et al. (2014), and generalize these prior results to arbitrary f-divergences and coverage constraints.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the 28th ACM Symposium on Theory of Computing*, pages 20–29, 1996.

Ivona Bezáková, Daniel Štefankovič, Vijay V Vazirani, and Eric Vigoda. Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM Journal on Computing*, 37(5):1429–1454, 2008.

Adam Block and Yury Polyanskiy. The sample complexity of approximate rejection sampling with applications to smoothed online learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 228–273. PMLR, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating rl for llm reasoning with optimal advantage regression. *arXiv preprint arXiv:2505.20686*, 2025.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer, 2005.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Omar Chehab, Aapo Hyvarinen, and Andrej Risteski. Provable benefits of annealing for estimating normalizing constants: Importance sampling, noise-contrastive estimation, and beyond. *Advances in Neural Information Processing Systems*, 36:45945–45970, 2023.

Fan Chen, Audrey Huang, Noah Golowich, Sadhika Malladi, Adam Block, Jordan T Ash, Akshay Krishnamurthy, and Dylan J Foster. The coverage principle: How pre-training enables post-training. *arXiv preprint arXiv:2510.15020*, 2025.

Christophe Chipot and Andrew Pohorille. Free energy calculations: Theory and applications in chemistry and biology. *Springer Series in Chemical Physics*, 86, 2007.

Imre Csiszár and Paul C Shields. *Information Theory and Statistics: A Tutorial*. Now Publishers Inc, 2004.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Gergely Flamich and Lennie Wells. Some notes on the sample complexity of approximate channel simulation. In *2024 IEEE International Symposium on Information Theory Workshops (ISIT-W)*, pages 1–6. IEEE, 2024.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

Roger B Grosse, Chris J Maddison, and Russ R Salakhutdinov. Annealing between distributions by averaging moments. *Advances in Neural Information Processing Systems*, 26, 2013.

Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, 2011.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *Forty-second International Conference on Machine Learning*, 2025b.

Mark Huber. Approximation algorithms for the normalizing constant of gibbs distributions. 2015.

Mark Jerrum. *Counting, Sampling and Integrating: Algorithms and Complexity*. Birkhäuser, Basel, 2003.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Donald E Knuth. Mathematics and computer science: Coping with finiteness: Advances in our ability to compute are bringing us substantially closer to ultimate limitations. *Science*, 194(4271): 1235–1242, 1976.

Vladimir Kolmogorov. A faster approximation algorithm for the gibbs partition function. In *Conference On Learning Theory*, pages 228–249. PMLR, 2018.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Qiang Liu, Jian Peng, Alexander Ihler, and John Fisher III. Estimating the partition function by discriminance sampling. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 514–522, 2015.

Fernando Llorente and Luca Martino. Optimality in importance sampling: a gentle survey, 2025. URL https://arxiv.org/abs/2502.07396.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Art B Owen. *Monte Carlo theory, methods and examples*. Self-published, 2013.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning. https://people.lids.mit.edu/yp/homepage/data/itbook.pdf, 2022. Draft book.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766. Morgan Kaufmann, 2000.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Daniel Štefankovič, Santosh Vempala, and Eric Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM*, 56(3):1–36, 2009.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3000–3006. AAAI Press, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Eric Vigoda. Sampling and counting. In *Handbook on Approximation Algorithms and Metaheuristics*. CRC Press, 2nd edition, 2024. See also earlier 2007 edition.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Martin J. Wainwright and Michael I. Jordan. *Graphical models, exponential families, and variational inference*, volume 1. Now Publishers, 2008.

David Wingate, Andreas Stuhlmüller, and Noah D. Goodman. Lightweight implementations of probabilistic programming languages via transformational compilation. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 770–778. JMLR Workshop and Conference Proceedings, 2011.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Appendix A. Technical Lemmata

In this section, we collect some technical lemmata used in the proofs of our main results. We begin with two concentration inequalities, before proving Lemma 22 relating coverage to f-divergences.

A.1. Concentration Inequalities

We first state the standard Chernoff bound (cf. for example (Boucheron et al., 2013)).

Lemma 17 (Chernoff Bound) *Let $X_1, \ldots, X_m$ be independent random variables such that for each $i \in [m]$, it holds that $X_i \in [0, M]$ almost surely and $\mu = \mathbb{E}[X_i]$. Then for any $1 \geq t > 0$ if holds that*

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i \leq (1-t)\mu\right) \leq e^{-\frac{m\mu t^2}{2M}}.$$

Moreover, for any $t > 0$, it holds that

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i \geq (1+t)\mu\right) \leq \exp\left(-{}^{m\mu}/_{M} \cdot ((1+t)\log(1+t) - t)\right).$$

Because we are generally considering $t \lesssim \varepsilon \ll 1$, we will often simplify the upper tail bound using the fact that $(1+t)\log(1+t) - t \geq {}^{t^2}/_{3}$ for $t \in (0,1)$ so that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq t\mu\right) \leq 2\exp\left(-{}^{m\mu t^2}/_{3M}\right)$$

for $t \in (0,1)$.

We also make use of the classical Bernstein's inequality (cf. for example (Boucheron et al., 2013)), which we state here for completeness.

Lemma 18 (Bernstein's Inequality) *Let $\xi_1, \ldots, \xi_m$ be independent random variables such that for each $i \in [m]$, the following properties hold:*

1. *The variables are centred, i.e. $\mathbb{E}[\xi_i] = 0$.*
2. *The variables are bounded almost surely, i.e. $|\xi_i| \leq M$.*
3. *The variables have variance at most σ^2, i.e. $\mathbb{E}[\xi_i^2] - \mathbb{E}[\xi_i]^2 \leq \sigma^2$.*

Then it holds for any $0 < \delta < 1$ that with probability at least $1 - \delta$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{m}} + \frac{2M \log(2/\delta)}{3m}.$$

Appendix B. Proof of Theorem 6

B.1. Choice of Estimator

In this section, we define the estimator we will use to prove our upper bounds. Recall that the median of k samples $Y_1, \ldots, Y_k$ is defined to be $y = \mathrm{Med}(\{Y_1, \ldots, Y_k\})$ such that $y \in \{Y_1, \ldots, Y_k\}$ satisfying

$$|\{j \in [k] | Y_j \leq y\}| \geq \frac{k}{2} \quad \text{and} \quad |\{j \in [k] | Y_j \geq y\}| \geq \frac{k}{2}.$$

We will use a median of means estimator to prove the upper bound. Let $k = \lceil 8 \log(1/\delta) \rceil$ and partition the n samples into k groups of size $m = \lfloor n/k \rfloor$. Recall that we define our estimator such that for each group $j \in [k]$, we let

$$\widehat{Z}_j = \frac{1}{m} \sum_{i=1}^{m} \lambda(X_i^{(j)}) \quad \text{and} \quad \widehat{Z} = \mathrm{Med}\left(\widehat{Z}_1, \ldots, \widehat{Z}_k\right).$$

The key fact about medians we will use is the following standard lemma, which allows us to convert constant probability bounds into high probability bounds using the median of means technique. This lemma is typically used in the context of estimation with bounded variance, but we state it here in a more general form for completeness.

Lemma 19 *Let $Y_1, \ldots, Y_k$ be independent random variables such that for each $i \in [k]$, it holds that $\mathbb{P}(Y_i \in (a, b)) \geq 1/2 + c$ for some $c > 0$. Then it holds that*

$$\mathbb{P}(\mathrm{Med}(Y_1, \ldots, Y_k) \in (a, b)) \geq 1 - \exp\left(-2kc^2\right).$$

Proof Let $y = \mathrm{Med}(\{Y_1, \ldots, Y_k\})$ and let $X_i = \mathbb{I}[Y_i \in (a, b)]$. Note that $\mathbb{E}[X_i] \geq 1/2 + c$ for all $i \in [k]$ and let $S = \sum_{i=1}^{k} X_i$. Note that if $y \notin (a, b)$, then it must hold that $S \leq k/2$. Thus we have

$$\mathbb{P}(y \notin (a, b)) \leq \mathbb{P}(S \leq k/2) \leq \mathbb{P}(\mathrm{Bin}(k/2 + c, k) \leq k/2) \leq e^{-2kc^2}$$

where the last inequality follows from a standard Chernoff bound (Lemma 17). ■

B.2. Analysis of Sample Mean

Our upper bound proofs will proceed by showing that $\widehat{Z}_j$ is ε-close to Z with constant probability for each $j \in [k]$, and then conclude by applying Lemma 19. Further, since our estimator is scale invariant and the guarantees are multiplicative, we can analyze the case where $Z = 1$ without loss of generality.

The first key lemma we need involves controlling the variance of truncated versions of the density ratio $\frac{d\nu}{d\mu}$ in terms of integrated coverage.

Lemma 20 *For any two probability measures μ, ν, we have that*

$$\mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X)\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right] \leq \mathrm{ICov}_M\left(\nu \| \mu\right)$$

Proof We use a standard tail integration argument. Indeed, we have

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X)\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right] &= \mathbb{E}_{X\sim\nu}\left[\frac{d\nu}{d\mu}(Y)\mathbb{I}\left[\frac{d\nu}{d\mu}(Y) \leq M\right]\right] \\
&= \int_0^\infty \mathbb{P}_{Y\sim\nu}\left(\frac{d\nu}{d\mu}(Y) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(Y) \leq M\right] \geq t\right) dt \\
&= \int_0^M \mathbb{P}_{Y\sim\nu}\left(t \leq \frac{d\nu}{d\mu}(Y) \leq M\right) dt \\
&= \int_0^M \left(\mathrm{Cov}_t\left(\nu \| \mu\right) - \mathrm{Cov}_M\left(\nu \| \mu\right)\right) dt \\
&= \int_0^M \mathrm{Cov}_t\left(\nu \| \mu\right) dt - M \cdot \mathrm{Cov}_M\left(\nu \| \mu\right).
\end{aligned}$$

The result follows.

■

Next, we prove a concentration result for the empirical average of the density ratio in terms of integrated coverage. The key idea is to note that even though the density ratio may have unbounded variance, we can analyze truncated versions of the density ratio to control the variance.

Lemma 21 (Sample Mean Concentration via Integrated Coverage) *For any two probability measures μ, ν, let $X_1, \ldots, X_m$ be i.i.d. samples from μ. Suppose $M_{\varepsilon,\delta} > 0$ is defined to satisfy*

$$\mathrm{ICov}_{M_{\varepsilon,\delta}}\left(\nu \| \mu\right) \leq \varepsilon\delta M_{\varepsilon,\delta},$$

with the convention that $M_{\varepsilon,\delta} = \infty$ if no such finite value exists. Then for any $0 < \varepsilon, \delta < 1$, it holds that as long as

$$m \geq \frac{M_{\varepsilon,\delta} \cdot \log(1/\delta)}{\varepsilon},$$

then with probability at least $1 - \delta$ it holds that

$$1 - \varepsilon \leq \frac{1}{m}\sum_{i=1}^m \frac{d\nu}{d\mu}(X_i) \leq 1 + \varepsilon.$$

Proof Fix an M to be chosen later and let $\eta_i = \frac{d\nu}{d\mu}(X_i) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) \leq M\right]$ for $i \in [m]$. It is immediate that $\mathbb{E}\left[\eta_i\right] \leq 1$. Further, from Definition 4, we have that

$$\mathbb{E}\left[\eta_i\right] \geq 1 - \mathrm{Cov}_M\left(\nu \| \mu\right) \tag{9}$$

By Lemma 20, it holds that

$$\mathbb{E}\left[\eta_i^2\right] \leq \mathrm{ICov}_M\left(\nu \| \mu\right)$$

and clearly $0 \leq \eta_i \leq M$ almost surely. Thus by Lemma 18, it holds that with probability at least $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^{m}\eta_i - \mathbb{E}\left[\eta_i\right]\right| \leq \sqrt{\frac{2\mathrm{ICov}_M\left(\nu\|\mu\right)\log\left(2/\delta\right)}{m}} + \frac{2M\log\left(2/\delta\right)}{3m}. \tag{10}$$

We now analyze the upper and lower tails separately.

Lower Tail. Combining (10) and (9), we have that with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i) &\geq \frac{1}{m}\sum_{i=1}^{m}\eta_i \\ &\geq 1 - \mathrm{Cov}_M\left(\nu\|\mu\right) - \sqrt{\frac{2\mathrm{ICov}_M\left(\nu\|\mu\right)\log\left(2/\delta\right)}{m}} - \frac{2M\log\left(2/\delta\right)}{3m}.\end{aligned}$$

Applying Young's inequality, which says that for $a, b, \lambda \geq 0$ it holds that $\sqrt{ab} \leq a/\lambda + \lambda b$, we see that

$$\sqrt{\frac{2\mathrm{ICov}_M\left(\nu\|\mu\right)\log\left(2/\delta\right)}{m}} \leq \frac{\mathrm{ICov}_M\left(\nu\|\mu\right)}{M} + \frac{M\log\left(2/\delta\right)}{2m}.$$

This gives that with probability at least $1-\delta$,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i) \geq 1 - \mathrm{Cov}_M\left(\nu\|\mu\right) - \frac{\mathrm{ICov}_M\left(\nu\|\mu\right)}{M} - \frac{2M\log\left(2/\delta\right)}{3m}.$$

Setting, M such that $\mathrm{Cov}_M\left(\nu\|\mu\right) \leq \varepsilon/3$, $\mathrm{ICov}_M\left(\nu\|\mu\right) \leq \varepsilon M/3$ and m such that $\frac{2M\log(2/\delta)}{3m} \leq \varepsilon/3$, we have that with probability at least $1-\delta$,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i) \geq 1-\varepsilon.$$

Upper Tail. We separate the sample mean as

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i) = \frac{1}{m}\sum_{i=1}^{m}\eta_i + \frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) > M\right]$$

and analyze each term separately. By Markov's inequality, it holds that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) > M\right] > t\right) &\leq \frac{\mathbb{E}\left[\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) > M\right]\right]}{t} \\ &\leq \frac{\mathrm{Cov}_M\left(\nu\|\mu\right)}{t}.\end{aligned}$$

Thus, as long as $\mathrm{Cov}_M(\nu\|\mu) \le \frac{\varepsilon\delta}{4}$ it holds that with probability at least $1 - \delta/2$, that

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) > M\right] \le \frac{\varepsilon}{2}.$$

We apply Lemma 18 again to control the first term. Indeed, it holds that with probability at least $1 - \delta/3$ that

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i \le 1 - \mathrm{Cov}_M(\nu\|\mu) + \sqrt{\frac{2\mathrm{ICov}_M(\nu\|\mu)\log(2/\delta)}{m}} + \frac{2M\log(2/\delta)}{3m}$$

As before, applying Young's inequality gives that with probability at least $1 - \delta/3$,

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i \le 1 - \mathrm{Cov}_M(\nu\|\mu) + \frac{\mathrm{ICov}_M(\nu\|\mu)}{M} + \frac{7M\log(3/\delta)}{6m}.$$

Setting, M such that $\mathrm{Cov}_M(\nu\|\mu) \le \frac{\varepsilon}{2}$, $\mathrm{ICov}_M(\nu\|\mu) \le \frac{\varepsilon M}{2}$ and m such that $\frac{7M\log(3/\delta)}{6m} \le \frac{\varepsilon}{2}$, we have that with probability at least $1 - \delta/3$,

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i \le 1 + \varepsilon.$$

Combining the two parts of the upper tail and applying a union bound, we get the result. ■

Proof [Proof of Theorem 6] By the previous lemma, we have that for each $j \in [k]$, as long as

$$m \ge \frac{M_{\varepsilon,1/4}\cdot\log(4)}{\varepsilon/2},$$

it holds that with probability at least $3/4$,

$$1 - \varepsilon \le \widehat{Z}_j \le 1 + \varepsilon.$$

Applying Lemma 19 with $c = 1/4$ and $k = \lceil 8\log(1/\delta)\rceil$ gives that with probability at least $1 - \delta$,

$$1 - \varepsilon \le \widehat{Z} \le 1 + \varepsilon$$

as required. ■

Appendix C. Relationship between Coverage and f-Divergences and the Proof of Theorem 7

In this appendix we establish some relationships between coverage and f-divergences, which we then use to prove Theorem 7 in Section C.2. We begin with these relationships, culminating in a key relationship (Lemma 24) used in the proof of Lemma 25, which is in turn necessary to prove our upper bound.

C.1. Coverage and f-Divergences

We begin by stating a fundamental relationship between coverage and f-divergences, which will be useful in proving upper bounds on the sample complexity in terms of f-divergences.

Lemma 22 *For any two probability measures μ, ν on $\mathcal{X}$ and any convex function f as in Definition 5, it holds for all $M > 1$ that*

$$\mathrm{Cov}_M(\nu \| \mu) \leq \frac{M \cdot \mathrm{D}_f(\nu \| \mu)}{f(M)}.$$

Thus if $M \geq \gamma_f(\mathrm{D}_f(\nu\|\mu)/\varepsilon)$ for some $\varepsilon \in (0,1)$, then $\mathrm{Cov}_M(\nu\|\mu) \leq \varepsilon$.

Proof By definition of coverage, we have that

$$\begin{aligned}
\mathrm{Cov}_M(\nu\|\mu) &= \mathbb{P}_{X\sim\nu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) = \mathbb{E}_{X\sim\mu}\left[\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \geq M\right]\right] \\
&\leq \mathbb{E}\left[\left(1 + \frac{M-1}{f(M)} \cdot f\left(\frac{d\nu}{d\mu}(X)\right)\right) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \geq M\right]\right] \\
&\leq \mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) + \frac{M-1}{f(M)} \cdot \mathrm{D}_f(\nu\|\mu),
\end{aligned}$$

where the first inequality follows from Lemma 23 with $x = \frac{d\nu}{d\mu}(X)$ for $X \sim \mu$ and the second inequality follows from the definition of f-divergence in (1) and the fact that $f \geq 0$. To conclude, we observe that for $M \geq 1$, it holds that f is monotone increasing and thus we may apply Markov's inequality to obtain

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) \leq \mathbb{P}_{X\sim\mu}\left(f\left(\frac{d\nu}{d\mu}(X)\right) \geq f(M)\right) \leq \frac{\mathbb{E}_{X\sim\mu}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right]}{f(M)} \leq \frac{\mathrm{D}_f(\nu\|\mu)}{f(M)}.$$

Adding this to the previous bound yields the result. ■

Lemma 23 *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a convex function such that $f(1) = f'(1) = 0$ and let $M \geq 1$. Then it holds that for any $x \geq M$,*

$$x \leq 1 + \frac{M-1}{f(M)} \cdot f(x).$$

Proof Let $t = \frac{M-1}{x-1}$; by the assumption that $1 < M \leq x$ we have that $t \in [0,1)$. Observe that $M = t \cdot 1 + (1-t) \cdot x$ and so by convexity of f, it holds that

$$f(M) \leq t \cdot f(1) + (1-t) \cdot f(x) = (1-t) \cdot f(x) = \frac{M-1}{x-1} \cdot f(x).$$

The result follows by rearranging the above inequality. ■

We now prove a key relationship between the second moment of the truncated density ratio and the f-divergence between the two measures.

Lemma 24 *Let f be a convex function as in Definition 5 such that there exists some $c > 1$ such that for all $M \geq t \geq c$ it holds that*

$$\frac{f(M)}{M^2} \leq \frac{f(t)}{t^2}.$$

Then for any probability measures μ, ν on $\mathcal{X}$ and any $M \geq c$, it holds that

$$\frac{\mathrm{ICov}_M\left(\nu \| \mu\right)}{M} \leq \frac{c^2}{M} + \frac{M \cdot \mathrm{D}_f\left(\nu \| \mu\right)}{f(M)}.$$

Proof Note first that if ${f(M)}/{M^2} \leq {f(t)}/{t^2}$ then

$$t^2 \leq \frac{M^2 \cdot f(t)}{f(M)}. \tag{11}$$

We compute:

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2\right] &= \mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq c\right]\right] \\
&\quad + \mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2 \mathbb{I}\left[M \geq \frac{d\nu}{d\mu}(X) > c\right]\right] \\
&\leq c^2 + \mathbb{E}_{X\sim\mu}\left[\frac{M^2 \cdot f\left(\frac{d\nu}{d\mu}(X)\right)}{f(M)} \mathbb{I}\left[M \geq \frac{d\nu}{d\mu}(X) > c\right]\right] \\
&\leq c^2 + \frac{2M^2 \cdot \mathrm{D}_f\left(\nu \| \mu\right)}{f(M)},
\end{aligned}$$

where the first inequality follows from (11) and the second inequality follows the nonnegativity of the integrand. ∎

Note that this lemma connnects the key quantity $\mathrm{ICov}_M\left(\nu \| \mu\right)/M$ used in our upper bounds to the f-divergence between the two measures. In particular, it shows that if we choose M sufficiently large so that ${M \cdot \mathrm{D}_f(\nu\|\mu)}/{f(M)}$ is small, then ${\mathrm{ICov}_M(\nu\|\mu)}/{M}$ is also small. We are now ready to prove our main upper bound in terms of f-divergences.

C.2. Proof of Theorem 7

As before, we will use the median-of-means estimator defined in (8) and thus it suffices to understand the behavior of the sample mean estimators in terms of f-divergences.

Lemma 25 (Sample Mean Concentration via f-divergence) *For any two probability measures μ, ν, let $X_1, \ldots, X_m$ be i.i.d. samples from μ. Let f be a convex function as in Definition 5 such that there exists some $c > 1$ such that for all $M \geq t \geq c$ it holds that*

$$\frac{f(M)}{M^2} \leq \frac{f(t)}{t^2}.$$

Then for any $0 < \varepsilon, \delta < 1$, *as long as*

$$m \geq \frac{\gamma_f\left(\mathrm{D}_f(\nu\|\mu)/\varepsilon\delta\right) \cdot \log(1/\delta)}{\varepsilon} \vee \frac{c^2 \cdot \log(1/\delta)}{\varepsilon^2\delta},$$

then with probability at least $1 - \delta$ *it holds that*

$$1 - \varepsilon \leq \frac{1}{m}\sum_{i=1}^{m} \frac{d\nu}{d\mu}(X_i) \leq 1 + \varepsilon.$$

Proof The first step is the following lemma relating coverage to f-divergences, the proof of which we defer to Appendix C.

To apply Lemma 21, we need to chose M such that the integrated coverage $\mathrm{ICov}_M(\nu\|\mu) \leq \varepsilon\delta M$. From Lemma 24, we have that we need $c^2/M + M \cdot \mathrm{D}_f(\nu\|\mu)/f(M) \leq \varepsilon\delta$. Recall from the definition of γ_f in (2), we have if we set $M = \gamma_f(\mathrm{D}_f(\nu\|\mu)/\varepsilon\delta)$, then by definition of γ_f it holds that $\frac{M \cdot \mathrm{D}_f(\nu\|\mu)}{f(M)} \leq \varepsilon\delta$. Thus, we setting $M \geq \max\left\{c^2/\varepsilon\delta, \gamma_f\left(\frac{\mathrm{D}_f(\nu\|\mu)}{\varepsilon\delta}\right)\right\}$ and applying Lemma 21 completes the proof. ■

Appendix D. Proofs of Lower Bounds

In this appendix, we provide complete proofs of our lower bound results. We first prove the most interesting case, the general lower bound in Theorem 11 (Section D.1) for superlinear but subquadratic f-divergences. We proceed with the lower bound in terms of integrated coverage (Theorem 9) in Section D.2 and then the linear lower bound in Proposition 10 (Section D.3); finally we prove the superquadratic lower bound in Proposition 12 (Section D.4).

D.1. Proof of Theorem 11

We first outline the construction and then prove the proposition. Let μ be any measure and consider a family of distributions $\nu_{p,\varepsilon}$ parameterized by $p \in [0, 1]$ and $\varepsilon \in (0, \frac{1}{4})$ defined such that

$$\frac{d\nu_{p,\varepsilon}}{d\mu}(X) = \begin{cases} 1 - \varepsilon & \text{w.p. } 1 - p, \\ 1 + \varepsilon\left(\frac{1}{p} - 1\right) & \text{w.p. } p. \end{cases} \tag{12}$$

We will let $\mathcal{V} \subset \{\nu_{p,\varepsilon}\}$ such that for all $\nu \in \mathcal{V}$, it holds that $\mathrm{D}_f(\nu\|\mu) \leq C + f(1 - \varepsilon)$. To help with this objective, we first compute the f-divergence between $\nu_{p,\varepsilon}$ and μ.

Lemma 26 *Suppose that*

$$\gamma_f\left(\frac{C}{2\varepsilon}\right) \geq 2.$$

Then it holds that $\mathrm{D}_f(\nu_{p,\varepsilon}\|\mu) \leq C + f(1 - \varepsilon)$ *whenever*

$$p \geq \frac{2 \cdot \varepsilon}{\gamma_f\left(\frac{C}{2\cdot\varepsilon}\right)}.$$

Proof By definition of the f-divergence and (12), we have that

$$\begin{aligned}
\mathrm{D}_f(\nu\|\mu) &= (1-p)\cdot f(1-\varepsilon) + p\cdot f\left(1+\varepsilon\left(\frac{1}{p}-1\right)\right) \\
&\leq f(1-\varepsilon) + p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right).
\end{aligned}$$

Thus as long as

$$p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right) \leq C, \tag{13}$$

the f-divergence condition will be satisfied.

We first claim that if (13) holds for p, then it holds for all $p < q \leq 1$. To see this, observe that for such q,

$$\begin{aligned}
q\cdot f\left(1-\varepsilon+\frac{\varepsilon}{q}\right) &= (q(1-\varepsilon)+\varepsilon)\cdot \frac{f\left(\frac{q(1-\varepsilon)+\varepsilon}{q}\right)}{\frac{q(1-\varepsilon)+\varepsilon}{q}} \\
&\leq (p(1-\varepsilon)+\varepsilon)\cdot \frac{f\left(\frac{p(1-\varepsilon)+\varepsilon}{p}\right)}{\frac{p(1-\varepsilon)+\varepsilon}{p}} \\
&= p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right) \leq C,
\end{aligned}$$

where the first inequality comes from the fact that $t \mapsto f(t)/t$ is non-decreasing for $t \geq 1$.

Now observe that the (13) holds if and only if

$$\frac{f\left(\frac{p(1-\varepsilon)+\varepsilon}{p}\right)}{\frac{p(1-\varepsilon)+\varepsilon}{p}} \leq \frac{C}{p(1-\varepsilon)+\varepsilon}.$$

By definition of the γ_f function in (2), this holds whenever

$$\frac{p(1-\varepsilon)+\varepsilon}{p} \leq \gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right),$$

which, after rearranging, is satisfied whenever

$$p \geq \frac{\varepsilon}{\gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right)-1}.$$

By the assumption that $\gamma_f\left(\frac{C}{2\varepsilon}\right) \geq 2$, this last is implied by $p \geq \frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right)}$. Note that for $p \leq \varepsilon$ it holds that

$$\frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right)} \leq \frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{2\cdot\varepsilon}\right)},$$

and thus the result follows. ■

Thus for fixed ε, let

$$\mathcal{V} = \left\{ \nu_{p,\varepsilon'} \middle| \frac{2\varepsilon'}{\gamma_f\left(\frac{C}{2\varepsilon'}\right)} \leq p \leq 1,\, 0 \leq \varepsilon' \leq \frac{1}{4} \right\} \tag{14}$$

We can now complete the proof of the proposition.

Proof [Proof of Theorem 11] Let μ be any probability measure and let $\mathcal{V}$ be defined as in (14). By Lemma 26, it holds that for all $\nu \in \mathcal{V}$, we have that $\mathrm{D}_f(\nu\|\mu) \leq C + f(1-\varepsilon)$.

Note that for any fixed p, in order to distinguish between distributions $\nu_0 = \mu$ and $\nu_{p,\varepsilon}$, with probability at least $2/3$, one must observe at least one sample from the high-density region of $\nu_{p,\varepsilon}$, which occurs with probability p under μ. On the other hand, the probability of observing only low-density samples after n samples is

$$(1-p)^n \geq e^{-2np}.$$

Thus for $n \leq \log(3/2)/2p$, no estimator can distinguish between ν_0 and $\nu_{p,\varepsilon}$ with probability at least $2/3$. On the other hand, the normalizing constants satisfy $Z = \lambda(X_1)$ in the former case and $Z = \lambda(X_1)/1-\varepsilon$ in the latter case, and thus any estimator that achieves $(1-\varepsilon)Z \leq \widehat{Z} \leq (1+\varepsilon)Z$ with probability at least $2/3$ must be able to distinguish between these two distributions. The result follows by setting $p = \frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{2\cdot\varepsilon}\right)}$. ■

D.2. Proof of Theorem 9

We use the same construction as in the proof of Theorem 11. Indeed, let $\nu_{p,\varepsilon}$ be defined as in (12). Observe that for any $\nu_{p,\varepsilon}$, it holds that

$$\mathrm{Cov}_M(\nu_{p,\varepsilon}\|\mu) = \begin{cases} 1 & M < 1-\varepsilon \\ p & 1-\varepsilon \leq M < 1+\varepsilon\left(\frac{1}{p}-1\right) \\ 0 & M \geq 1+\varepsilon\left(\frac{1}{p}-1\right) \end{cases}$$

Thus it holds that

$$\mathrm{ICov}_M(\nu_{p,\varepsilon}\|\mu) = \begin{cases} M & M < 1-\varepsilon \\ (1-\varepsilon) + p\cdot(M-(1-\varepsilon)) & 1-\varepsilon \leq M < 1+\varepsilon\left(\frac{1}{p}-1\right) \\ 1+\varepsilon\left(\frac{1}{p}-1\right) & M \geq 1+\varepsilon\left(\frac{1}{p}-1\right) \end{cases}.$$

By the identical argument as the proof of Theorem 11, in order to achieve $(1\pm\varepsilon)$ multiplicative accuracy with probability at least $2/3$, one must be able to distinguish $\nu_0 = \mu$ from $\nu_{p,\varepsilon}$ with probability at least $2/3$, which requires at least $\log(3/2)/2p$ samples. Letting $p = \varepsilon^2/1+\varepsilon$, we claim that for $M = \varepsilon/p$, it holds that $\mathrm{ICov}_M(\nu_{p,\varepsilon}\|\mu) \leq M\cdot\varepsilon$. Indeed, observe that for such M, we are clearly are in the intermediate regime above and thus

$$\frac{\mathrm{ICov}_M(\nu\|\mu)}{M} = \frac{(1-\varepsilon)(1-p)}{M} + p \leq \frac{p}{4\cdot\varepsilon} + p \leq \varepsilon.$$

On the other hand, we have established that at least $\log(3/2)/2p = \Theta\left(M\varepsilon/\varepsilon\right)$ samples are required to achieve the desired accuracy. The result follows. ■

D.3. Proof of Proposition 10

We prove a slightly restated version of Proposition 10.

Proposition 27 *Let f be a convex function as in Definition 5 such that the map $t \mapsto f(t)/t$ is uniformly bounded. For any $0 < C < f'(\infty)$, there exist measures μ, ν such that $\mathrm{D}_f(\nu \| \mu) \leq C$ and such that no algorithm given any finite number of independent samples from μ and oracle access to $\frac{d\nu}{d\mu}(\cdot) \cdot Z$ can produce an estimate $\widehat{Z}$ satisfying a $1 \pm \varepsilon$ approximation of Z with probability at least $3/4$ for any $\varepsilon \leq f'(\infty)/(f'(\infty)-C)$.*

Moreover, for any $C \geq f'(\infty)$, there exist measures μ, ν such that $\mathrm{D}_f(\nu \| \mu) \leq C$ and such that no algorithm given any finite number of independent samples from μ and oracle access to $\frac{d\nu}{d\mu}(\cdot) \cdot Z$ can produce an estimate $\widehat{Z}$ satisfying a $1 \pm \varepsilon$ approximation of Z with probability at least $3/4$ for any $\varepsilon > 0$.

Proof Let $\mu = \delta(0)$ and $\nu = \mathrm{Ber}(1-q)$. Then it holds that

$$\mathrm{D}_f(\nu \| \mu) = q f'(\infty).$$

Thus, for any finite $C > 0$, choosing $q \leq C/f'(\infty)$ ensures that $\mathrm{D}_f(\nu \| \mu) \leq C$. However, as μ is a point mass at 0, no finite number of samples from μ reveals any information about q.

In the case where $C \leq f'(\infty)$, we have

$$1 - \frac{C}{f'(\infty)} \leq q \leq 1.$$

and thus the tightest possible constant factor approximation of Z is at least

$$\frac{f'(\infty)}{f'(\infty) - C}.$$

In the case where $C > f'(\infty)$, we may choose any $q \in [0,1]$ and thus no finite number of samples from μ can yield any constant factor approximation of Z.

■

D.4. Proof of Proposition 12

Let $\mathcal{X} = \{0,1\}$ with $\mu_p(0) = 1-p$ and $\mu_p(1) = p$; let $\mathcal{U} = \{\mu_p : 1/4 \leq p \leq 1/2\}$. Let $\nu = \delta_1$ so

$$\frac{d\nu}{d\mu}(X) = \begin{cases} 0 & X = 0 \\ \frac{1}{p} & X = 1 \end{cases}.$$

Thus $\left\| \frac{d\nu}{d\mu} \right\|_\infty \leq 4$ for all $\mu \in \mathcal{U}$ by the lower bound on p. Let $\lambda(X) = \mathbb{I}[X = 1]$ be an unnormalized density ratio for ν with respect to μ_p and observe that estimating Z is equivalent to estimating p. Let $p = 1/2$ and $p' = 1/2 - 2\varepsilon$ and consider the two measures $\mu = \mu_p$ and $\mu' = \mu_{p'}$. In order to estimate Z to within a factor of $(1 \pm \varepsilon)$ with probability at least $2/3$ under both μ and μ', one must distinguish between the two distributions with probability at least $1/3$. By standard hypothesis testing lower bounds such as Le Cam's method (cf. e.g. Wainwright (2019)), the probability of error is lower bounded by $1/2\left(1-\mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right)\right)$. Thus we require $\mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right) \geq 1/3$. On the other hand, by Pinsker's inequality and the chain rule for KL divergence, it holds that

$$\mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right) \leq \sqrt{\frac{n}{2} \cdot \mathrm{D_{KL}}\left(\mu \| \mu'\right)}.$$

Thus the result will follow if we can show that

$$\mathrm{D_{KL}}\left(\mu \| \mu'\right) \lesssim \varepsilon^2.$$

We compute that

$$\begin{aligned} \mathrm{D_{KL}}\left(\mu \| \mu'\right) &= \frac{1}{2} \cdot \left(\log\left(\frac{1/2}{1/2 - 2\varepsilon}\right) + \log\left(\frac{1/2}{1/2 + 2\varepsilon}\right)\right) \\ &= \frac{1}{2} \cdot \log\left(\frac{1}{1 - 16\varepsilon^2}\right) \\ &\leq 8\varepsilon^2, \end{aligned}$$

where the last inequality follows from the fact that $\log\left(1/1-x\right) \leq 2x$ for all $x \in (0, 1/2)$. This completes the proof. ■

Appendix E. Proof of Theorem 8

In this appendix, we prove Theorem 8. We begin by proving the Generalized Paley-Zygmund inequality, before applying it to demonstrate that for appropriately chosen α, the empirical $(1-\alpha)$-quantile of the samples has the stated performance guarantees with high probability.

E.1. Generalized Paley-Zygmund

We begin by stating a slightly stronger result than the Generalized Paley-Zygmund inequality in Lemma 16.

Lemma 28 *Let μ, ν be probability measures on $\mathcal{X}$. Then for any $0 < \varepsilon < 1$, it holds for any $0 < u < 1$ that*

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq 1 - \varepsilon\right) \geq \frac{(1-u)\cdot\varepsilon}{M} \quad \textit{where } \mathrm{Cov}_M\left(\nu \| \mu\right) \leq u \cdot \varepsilon.$$

Proof Let $Y = \frac{d\nu}{d\mu}(X)$ for $X \sim \mu$. By definition, we have that $\mathbb{E}[Y] = 1$ and thus

$$1 = \mathbb{E}\left[Y\right] \leq 1 - \varepsilon + \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq 1 - \varepsilon\right]\right].$$

Rearranging, we have for any M that

$$\begin{aligned} \varepsilon &\leq \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq 1 - \varepsilon\right]\right] \\ &= \mathbb{E}\left[Y \cdot \mathbb{I}\left[1 - \varepsilon \leq Y < M\right]\right] + \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq M\right]\right] \\ &\leq M \cdot \mathbb{P}\left(Y \geq 1 - \varepsilon\right) + \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq M\right]\right] \\ &= M \cdot \mathbb{P}\left(Y \geq 1 - \varepsilon\right) + \mathrm{Cov}_M(\nu \| \mu). \end{aligned}$$

The result follows by rearranging and using the assumption that $\mathrm{Cov}_M(\nu\|\mu) \leq u \cdot \varepsilon$. ■

We now derive Lemma 16 from Lemma 28 by applying Lemma 22.

Proof [Proof of Lemma 16] The first part of the statement is just Lemma 28. For the second part, by Lemma 22, it holds that for

$$M \geq \gamma_f \left(\frac{\mathrm{D}_f \left(\nu\|\mu\right)}{u \cdot \varepsilon} \right),$$

we have that $\mathrm{Cov}_M(\nu\|\mu) \leq u \cdot \varepsilon$. The result follows by plugging this into Lemma 28 and optimizing over u. ■

E.2. Completing the Proof of Theorem 8

We restate the result with constants included.

Theorem 29 *Let μ, ν be probability measures on $\mathcal{X}$ and suppose $X_1, \ldots, X_n \sim \mu$ are independent with λ an unnormalized density ratio. Then there is an estimator $\widehat{Z}$ such that for any $0 < \varepsilon < 1$, if*

$$n \geq \frac{18M}{\varepsilon} \cdot \log \left({}^{2}/_{\delta} \right) \quad \textit{where } \mathrm{Cov}_M \left(\nu\|\mu\right) \leq {}^{\varepsilon}/_{4},$$

it holds that with probability at least $1 - \delta$ that $(1 - \varepsilon) \cdot Z \leq \widehat{Z} \leq M \cdot Z$.

We first define the estimator, motivated by Lemma 28. Indeed for arbitrary u, t, we will choose some M such that $\mathrm{Cov}_M(\nu\|\mu) < u \cdot \varepsilon$ and let $\alpha = \frac{(1-t)(1-u) \cdot \varepsilon}{M}$. We then define the estimator as the $(1 - \alpha)$-quantile of the samples $\lambda(X_1), \ldots, \lambda(X_n)$; that is, letting $\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(n)}$ be the order statistics of the samples, we define $\widehat{Z} = \lambda_{(\lceil (1-\alpha)n \rceil)}$. We now analyze the quality of this estimator when $t = u = {}^{1}/_{2}$.

Proof [Proof of Theorem 29] We analyze each tail separately and then apply a union bound to conclude.

Lower Tail. Let $\eta_i = \mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) \geq 1 - \varepsilon\right]$ for $i \in [n]$. By Lemma 28 it holds for M such that $\mathrm{Cov}_M(\nu\|\mu) \leq u \cdot \varepsilon$ that

$$\mathbb{E}[\eta_i] = \mathbb{P}_{X \sim \mu} \left(\frac{d\nu}{d\mu}(X) \geq 1 - \varepsilon \right) \geq \frac{(1 - u) \cdot \varepsilon}{M}.$$

Applying a standard Chernoff bound (Lemma 17), we have that

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^{n} \eta_i \leq \frac{(1 - t)(1 - u) \cdot \varepsilon}{M} \right) \leq \exp \left(-\frac{nt^2(1 - u) \cdot \varepsilon}{2M} \right).$$

Now observe that by definition of the quantile,

$$\mathbb{P} \left(\widehat{Z} \leq (1 - \varepsilon)Z \right) \leq \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^{n} \eta_i \leq \alpha \right) \leq \exp \left(-\frac{nt^2(1 - u) \cdot \varepsilon}{2M} \right).$$

Thus by setting

$$n \geq \frac{2M}{t^2(1 - u) \cdot \varepsilon} \cdot \log \left({}^{2}/_{\delta} \right), \tag{15}$$

we have with probability at least $1 - {}^{\delta}/_{2}$ that $\widehat{Z} \geq (1 - \varepsilon)Z$.

Upper Tail. Fix $M' \geq 1$ and let $\xi_i = \mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) \geq M'\right]$ for $i \in [n]$. By Lemma 30, it holds that

$$\mathbb{E}[\xi_i] = \mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M'\right) \leq \frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'}.$$

Thus by a standard Chernoff bound (Lemma 17), we have that for $0 < s \leq 1$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq (1+s)\frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'}\right) \leq \exp\left(-\frac{ns^2 \cdot \mathrm{Cov}_{M'}(\nu\|\mu)}{3M'}\right).$$

It then follows that as long as

$$n \geq \frac{3M'}{s^2 \cdot \mathrm{Cov}_{M'}(\nu\|\mu)} \cdot \log\left({}^{2}/_{\delta}\right) \quad \text{and} \quad (1+s)\frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'} \leq \alpha, \tag{16}$$

we have that with probability at least $1 - {}^{\delta}/_{2}$, that $\widehat{Z} \leq M'Z$.

Concluding the Proof. We now set $t = u = {}^{1}/_{2}$ so that $\alpha = \frac{\varepsilon}{4M}$. Thus (15) implies that as long as $n \geq {}^{16M}/_{\varepsilon} \cdot \log\left({}^{2}/_{\delta}\right)$ with $\mathrm{Cov}_M(\nu\|\mu) \leq {}^{\varepsilon}/_{2}$, we have with probability at least $1 - {}^{\delta}/_{2}$ that $\widehat{Z} \geq (1-\varepsilon)Z$. Moreover, setting $s = {}^{1}/_{2}$, it follows from (16) that as long as

$$n \geq \frac{12}{{}^{2\alpha}/_{3}} \cdot \log\left({}^{2}/_{\delta}\right) = \frac{18M}{\varepsilon} \cdot \log\left({}^{2}/_{\delta}\right)$$

that with probability at least $1-{}^{\delta}/_{2}$, it holds that $\widehat{Z} \leq M' \cdot Z$ for any M' such that $\mathrm{Cov}_{M'}(\nu\|\mu) \leq {}^{\varepsilon}/_{4}$. The result follows by applying a union bound. ∎

We now prove a necessary technical lemma used in the proof of Theorem 29 above.

Lemma 30 *Let μ, ν be probability measures on $\mathcal{X}$ and let $M \geq 1$. Then it holds that*

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) \leq \frac{\mathrm{Cov}_M(\nu\|\mu)}{M}.$$

Proof We compute

$$\begin{aligned}
\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) &= \mathbb{E}_{X\sim\mu}\left[\mathbb{I}\left[\frac{d\nu}{d\mu}(X) \geq M\right]\right] \\
&= \mathbb{E}_{Y\sim\nu}\left[\frac{1}{\frac{d\nu}{d\mu}(Y)} \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(Y) \geq M\right]\right] \\
&\leq \frac{1}{M} \cdot \mathbb{E}_{Y\sim\nu}\left[\frac{d\nu}{d\mu}(Y) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(Y) \geq M\right]\right] \\
&= \frac{\mathrm{Cov}_M(\nu\|\mu)}{M}.
\end{aligned}$$

∎

We now conclude this section by deriving Theorem 8 from Theorem 29.

Proof [Proof of Theorem 8] By Lemma 22, it holds that for

$$M \geq \gamma_f\left(\frac{4 \cdot \mathrm{D}_f(\nu\|\mu)}{\varepsilon}\right),$$

we have that $\mathrm{Cov}_M(\nu\|\mu) \leq {}^{\varepsilon}/_{4}$. The result follows by plugging this into Theorem 29. ∎

Appendix F. Proof of Proposition 13

We use the classical A^* sampling algorithm (Li and El Gamal, 2018) which is known to produce approximate samples from the target distribution in the case where the Radon-Nikodym derivative is uniformly bounded or more generally under finite $\mathrm{D}_{\mathrm{KL}}(\nu \| \mu)$ (Flamich and Wells, 2024). The algorithm proceeds by first sampling $X_1, \ldots, X_n \sim \mu$ and $E_1, \ldots, E_n \sim \operatorname{Exp}(1)$ independently. We let $N_i = \sum_{j=1}^{i} E_j$ be the Poisson arrival times and let

$$S_i = \frac{N_i}{{}^{d\nu}/_{d\mu}(X_i)}.$$

In order to construct a sample, we let $\widehat{j} = \operatorname{argmin}_{1 \le i \le n} S_i$ and return $X_{\widehat{j}}$. We now analyze the total variation distance between the law of $X_{\widehat{j}}$ and ν.

Lemma 31 *Let μ, ν be probability measures on $\mathcal{X}$ and let $X_1, \ldots, X_n \sim \mu$ be independent. Let $\widehat{j}$ be defined as above. Letting ν_n be the law of $X_{\widehat{j}}$, it holds that $\mathrm{TV}(\nu_n, \nu) \le \varepsilon$ as long as*

$$n \ge 2M \cdot \log\left({}^{3}/_{\varepsilon}\right) \quad \textit{where} \quad \operatorname{Cov}_M(\nu \| \mu) \le {}^{\varepsilon}/_{3}.$$

Proof We adopt the proof strategy found in Li and El Gamal (2018, Theorem 1). We form a coupling between the law of $Z_{\widehat{j}}$ and ν as follows. Extend the sequences X_i, E_i to be infinite and let $K = \operatorname{argmin}_{i \ge 1} \frac{N_i}{{}^{d\nu}/_{d\mu}(X_i)}$. By the argument in Li and El Gamal (2018), it holds that $Z_K \sim \nu$. Thus it holds that

$$\mathrm{TV}(\nu_n, \nu) \le \mathbb{P}\left(K \neq \widehat{j}\right) = \mathbb{P}(K > n).$$

Letting $\Theta = \min_{i \ge 1} \frac{N_i}{{}^{d\nu}/_{d\mu}(X_i)}$, we have by Li and El Gamal (2018) the the following hold: (i) $\Theta \sim \operatorname{Exp}(1)$ and (ii) conditioned on Θ and X_K, it holds that $K - 1$ is a Poisson random variable with rate at most $\Theta \frac{d\nu}{d\mu}(X_K)$. Thus it holds that

$$\mathbb{P}(K > n) \le \mathbb{E}_{Y \sim \nu}\left[\int_0^\infty e^{-\theta} \mathbb{P}\left(\operatorname{Pois}\left(\theta \frac{d\nu}{d\mu}(Y)\right) > n\right) d\theta\right].$$

Observe that poisson random variables with larger rate stochastically dominate those with smaller rate; thus it holds for any $M > 1$ that

$$\begin{aligned}
\mathbb{E}_{Y \sim \nu}\left[\int_0^\infty e^{-\theta} \mathbb{P}\left(\operatorname{Pois}\left(\theta \frac{d\nu}{d\mu}(Y)\right) > n\right) d\theta\right] &\le \operatorname{Cov}_M(\nu \| \mu) + \int_0^\infty e^{-\theta} \mathbb{P}(\operatorname{Pois}(\theta M) > n)\, d\theta \\
&\le \operatorname{Cov}_M(\nu \| \mu) + \inf_{t > 0} e^{-t} + \mathbb{P}(\operatorname{Pois}(tM) > n) \\
&\le {}^{2\varepsilon}/_{3} + \mathbb{P}\left(\operatorname{Pois}\left(M \cdot \log\left({}^{3}/_{\varepsilon}\right)\right) > n\right).
\end{aligned}$$

Using standard Poisson tail bounds (Vershynin, 2018), for $\lambda, t > 0$, it holds that

$$\mathbb{P}(\operatorname{Pois}(\lambda) > (1+t)\lambda) \le e^{-\frac{t^2 \lambda}{2(1+t)}}.$$

Thus for $n \ge 2M \log\left({}^{3}/_{\varepsilon}\right)$, we have that $\mathbb{P}\left(\operatorname{Pois}\left(M \cdot \log\left({}^{3}/_{\varepsilon}\right)\right) > n\right) \le {}^{\varepsilon}/_{3}$. The result follows. ■

We can now prove the proposition by appealing to the above lemma and Lemma 22.
Proof [Proof of Proposition 13] By Lemma 22, it holds that

$$\mathrm{Cov}_M\left(\nu \| \mu\right) \le \frac{M \cdot \mathrm{D}_f\left(\nu \| \mu\right)}{f(M)}.$$

Thus by setting M such that $\frac{M \cdot \mathrm{D}_f(\nu \| \mu)}{f(M)} \le \varepsilon/3$, we have that

$$n \gtrsim \log\left(3/\varepsilon\right) \cdot \gamma_f\left(\frac{3 \cdot \mathrm{D}_f\left(\nu \| \mu\right)}{\varepsilon}\right)$$

samples suffice to ensure that $\mathrm{TV}\left(\nu_n, \nu\right) \le \varepsilon$. ■