Convergence Rates for Non-Log-Concave Sampling and Log-Partition Estimation

David Holzmüller, Francis Bach

CONVERGENCE RATES FOR NON-LOG-CONCAVE SAMPLING AND LOG-PARTITION ESTIMATION

A PREPRINT

David Holzmüller*
University of Stuttgart
Faculty of Mathematics and Physics
Institute for Stochastics and Applications

Francis Bach
INRIA
Ecole Normale Supérieure
PSL Research University

March 6, 2023

ABSTRACT

Sampling from Gibbs distributions $p(x) \propto \exp(-V(x)/\varepsilon)$ and computing their log-partition function are fundamental tasks in statistics, machine learning, and statistical physics. However, while efficient algorithms are known for convex potentials V, the situation is much more difficult in the non-convex case, where algorithms necessarily suffer from the curse of dimensionality in the worst case. For optimization, which can be seen as a low-temperature limit of sampling, it is known that smooth functions V allow faster convergence rates. Specifically, for m-times differentiable functions in d dimensions, the optimal rate for algorithms with n function evaluations is known to be $O(n^{-m/d})$, where the constant can potentially depend on m, d and the function to be optimized. Hence, the curse of dimensionality can be alleviated for smooth functions at least in terms of the convergence rate. Recently, it has been shown that similarly fast rates can also be achieved with polynomial runtime $O(n^{3.5})$, where the exponent 3.5 is independent of m or d. Hence, it is natural to ask whether similar rates for sampling and log-partition computation are possible, and whether they can be realized in polynomial time with an exponent independent of m and d. We show that the optimal rates for sampling and log-partition computation are sometimes equal and sometimes faster than for optimization. We then analyze various polynomial-time sampling algorithms, including an extension of a recent promising optimization approach, and find that they sometimes exhibit interesting behavior but no near-optimal rates. Our results also give further insights on the relation between sampling, log-partition, and optimization problems.

1 Introduction

The tasks of sampling from a Gibbs distribution with density $p(x) \propto \exp(-V(x)/\varepsilon)$ and computing the corresponding normalization constant are important problems in many computational fields such as machine learning, (Bayesian) statistics and statistical physics. Specifically, we are interested in the following setting:

Definition 1 (Sampling and log-partition problems)**.** Let $d \geq 1$, let $\mathcal{X} = [0,1]^d$ and let $f : \mathcal{X} \to \mathbb{R}$ be bounded and measurable.[2] The *sampling problem* is to draw samples from the distribution P_f on $\mathcal{X}$ with density

$$p_f(x) \coloneqq \frac{\exp(f(x))}{Z_f} ,$$

where $Z_f \coloneqq \int_{\mathcal{X}} \exp(f(x)) \, \mathrm{d}x$ is the normalization constant or *partition function*. The *log-partition problem* is to compute the *log-partition function*

$$L_f \coloneqq \log Z_f = \log \left(\int_{\mathcal{X}} \exp(f(x)) \, \mathrm{d}x \right) . \qquad \blacktriangleleft$$

*Work done partially while visiting INRIA. E-mail: david (dot) holzmueller (at) mathematik.uni-stuttgart.de

[2]We choose $\mathcal{X}$ as the unit cube for convenience: It is compact, has unit volume, does not have too sharp corners, there are well-studied approximation results, and it allows to investigate algorithms for periodic functions. However, many of our results could be generalized to other domains.

Distributions of the form P_f are known as *Gibbs distributions*, *Gibbs measures*, or *Boltzmann distributions*. They arise for example in statistical physics with $f(x) = -V(x)/\varepsilon$, where $V(x)$ denotes the (potential) energy of state x, and ε is (proportional to) the temperature of the system.[3] Instead of the temperature, sometimes the inverse temperature (or coldness / thermodynamic beta) $\beta = 1/\varepsilon$ is used. The log-partition problem is also related to the computation of the free energy $-\varepsilon L_{-V/\varepsilon}$. In energy-based models in ML, f could be learned. In a Bayesian statistical or ML model with parameters θ and data D, we could set $f(\theta) \coloneqq \log p(D|\theta) + \log p(\theta)$ to sample from the posterior distribution $p_f(\theta) = p(\theta|D)$ or compute the log-evidence $L_f = \log p(D) = \log\left(\int_{\mathcal{X}} p(D|\theta)p(\theta)\,\mathrm{d}\theta\right)$. In some contexts, L_f is also called the log-marginal likelihood, which is useful for model selection (Robert, 2007).

While exact sampling and log-partition computation are possible for some simple functions f like linear functions, we can usually only expect algorithms to obtain approximations within a limited runtime. Therefore, we are interested in how fast this approximation converges to the true P_f or L_f in terms of the number n of times that the algorithm is allowed to evaluate f. To study this, we need to make some assumptions on f. While efficient sampling algorithms for suitable classes of concave f are known, at least with access to gradients of f (Dwivedi et al., 2018; Mangoubi and Vishnoi, 2018; Chewi et al., 2021; Altschuler and Talwar, 2022), we are interested in larger classes of non-concave functions, which are defined in the following:

Definition 2 (Further notation)**.** For measurable functions $f : \mathcal{X} \to \mathbb{R}$, we use the notation $\|f\|_\infty \coloneqq \operatorname{ess\,sup}_{x\in\mathcal{X}} |f(x)|$. We define function spaces of m-times continuously differentiable functions[4] whose derivatives are bounded by some constant $B \geq 0$:

$$\mathcal{F}_{d,m,B} \coloneqq \{f \in C^m(\mathcal{X}), \|f\|_{C^m} \leq B\}\,, \qquad \|f\|_{C^m} \coloneqq \sup_{\alpha\in\mathbb{N}_0^d : |\alpha|_1 \leq m} \|\partial_\alpha f\|_\infty\,.$$

Here, we use the notation $|\alpha|_1 \coloneqq \alpha_1 + \cdots + \alpha_d$ and $\partial_\alpha f = \frac{\partial^{|\alpha|_1} f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$. Moreover, if f is Lipschitz, we denote its minimal Lipschitz constant by $|f|_1$. If f is bounded, we denote its maximum by M_f. We define $\bar{f} \coloneqq f - L_f$, such that $L_{\bar{f}} = 0$ and $P_{\bar{f}} = P_f$. Finally, we denote the uniform distribution on $\mathcal{X}$ by $\mathcal{U}(\mathcal{X})$. ◀

We study the worst-case error of algorithms over the function class $\mathcal{F}_{d,m,B}$, which is formally defined in Section 2. This error depends on the variables (B, n, d, m). We study the asymptotic behavior in terms of n and B while ignoring constants depending only on m and d for simplicity. Depending on the definition of the norm, such constants are often necessarily exponential in d and represent the part of the curse of dimensionality that cannot be overcome in this setting (Novak and Woźniakowski, 2009). For example, typical convergence rates for function approximation are of the form $O_{m,d}(Bn^{-m/d})$, which we sometimes also write as $O_{m,d}(\|f\|_{C^m} n^{-m/d})$ (Novak, 1988; Wendland, 2004). As we will see later, the dependence on B is not always linear, and tracking the dependence on B is important since the function f appears inside an exponential. When using asymptotic notation like $O_{m,d}$, we mean that the corresponding inequality should hold for *all* values of $n \in \mathbb{N}_{\geq 1}$ and $B > 0$, not only large enough values.[5]

We express bounds on the error E achieved for n function evaluations, such as $E = O_{m,d}(Bn^{-m/d})$. Some authors prefer to express rates in terms of the number of function evaluations needed to reach an error E or lower, which would then for example be $n = O_{m,d}((B/E)^{d/m})$.

Sometimes, we explicitly include a temperature $\varepsilon > 0$ and formulate our theorems in terms of f/ε instead of f. It is well-known that in the limit of low temperatures ($\varepsilon \searrow 0$), sampling becomes essentially equivalent to optimization. Here, we give a quantitative version of this statement:

Lemma 3 (Optimization limit)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz-continuous with Lipschitz constant $|f|_1 < \infty$. Then, for any temperature $\varepsilon > 0$, the maximum $M_f = \max_{x\in\mathcal{X}} f(x)$ satisfies*

$$|M_f - \varepsilon L_{f/\varepsilon}| \leq \varepsilon d \log(1 + 3d^{-1/2}|f|_1/\varepsilon) \longrightarrow 0 \quad \textit{for } \varepsilon \searrow 0\,. \tag{1}$$

Moreover, for any bounded and measurable $f : \mathcal{X} \to \mathbb{R}$ and any $\delta \in (0, 1]$, we have

$$P_{f/\varepsilon}(\{x \in \mathcal{X} \mid f(x) < \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)\}) \leq \delta\,.$$

Lemma 3 is proven in Appendix B and can be related to our function classes by using $|f|_1 \leq d^{1/2}\|f\|_{C^1}$, cf. Lemma B.1. Note that such a convergence result does not hold for general bounded functions, as can be seen for the characteristic

[3] Technically, $\varepsilon = k_B T$, where k_B is Boltzmann's constant and T is the temperature.

[4] It would also be possible to replace $C^m(\mathcal{X})$ with the slightly larger Sobolev space $W^{m,\infty}(\mathcal{X})$.

[5] Specifically, we use the notation $g(B, n, d, m) \leq O_{m,d}(h(B, n, d, m))$ to mean

$$\forall d, m \in \mathbb{N}_{\geq 1} \exists C_{m,d} > 0 \forall B > 0, n \in \mathbb{N}_{\geq 1} : g(B, n, d, m) \leq C_{m,d} h(B, n, d, m)\,,$$

and we similarly write $g \geq \Omega_{m,d}(h)$ for $h \leq O_{m,d}(g)$, as well as $g = \Theta_{m,d}(h)$ for $g \leq O_{m,d}(h)$ and $h \leq O_{m,d}(g)$.

function $f = \mathbb{1}_{\{0\}}$, which satisfies $M_f = 1$ but $\varepsilon L_{f/\varepsilon} = 0$ for all $\varepsilon > 0$. Eq. (1) is related to Corollary 1 by Ma et al. (2019), which shows that in order to achieve an optimization error of E through sampling, it is necessary that $1/\varepsilon = \widetilde{\Omega}(d/E)$.

Hwang (1980) uses Laplace's method to show that under certain assumptions on the Hessians at the maximizers, $P_{f/\varepsilon}$ converges weakly to a distribution on the maximizers. In contrast to optimization, where it typically does not matter which global maximum is found, the low-temperature limit of sampling often yields a unique distribution on the maximizers. Talwar (2019) uses this to show that optimization can be easier than sampling for very particular classes of functions. It is known that the optimal worst-case convergence rate for optimization on $\mathcal{F}_{d,m,B}$ is the same as for approximation, i.e., $O_{m,d}(Bn^{-m/d})$ (see Novak, 1988, and references therein). Recently, Rudi et al. (2020) and Woodworth et al. (2022) have shown polynomial-time optimization algorithms (in n and d) achieving rates close to the optimal rate under relatively mild additional assumptions.

In the high-temperature limit $\varepsilon \to \infty$, $P_{f/\varepsilon}$ converges to a uniform distribution, and Ma et al. (2019) showed that the Metropolis-adjusted Langevin algorithm (MALA) can achieve exponentially fast convergence rates in n, although with exponential dependence on the Lipschitz constant of ∇f. While MALA theoretically does not fit in our framework since it uses gradient information of f, this could be emulated using numerical differentiation, and we show in Section 4.2.1 that in our setting, if B is known, a fixed-budget version of rejection sampling can also achieve similar rates (see also Talwar, 2019).

Since we know that polynomial-time algorithms with fast convergence rates are possible for the high-temperature case and for optimization, which is essentially the low-temperature limit, this poses the question of whether we can find such algorithms for the general sampling and log-partition problems. Ideally, such an algorithm should have the following properties:

- A convergence rate close to the optimal convergence rate $O_{m,d}(\|f\|_{C^m} n^{-m/d})$ for approximation, at least up to $m \geq \Omega(d)$, such that the exponent in the rate does not approach zero for large d,
- Polynomial runtime $O(n^k)$ for some k independent of d and m. Especially, the runtime should be polynomial in d and m. Moreover, the runtime should not depend on $\|f\|$.
- Adaptivity: The algorithm should achieve these rates without knowing m and $\|f\|$. This is not investigated here, and for sampling and log-partition estimation, m can often be known.

Regarding the implications of achieving near-optimal rates in polynomial time, consider the following example:

Example 4. Consider a Bayesian model where the data set $D = (D_1, \dots, D_N)$ consists of N observed samples that are assumed to be drawn in an i.i.d. fashion. Then, we can again model

$$f(\theta) \coloneqq \log p(\theta, D) = \log p(D_1 \mid \theta) + \cdots + \log p(D_N \mid \theta) + \log p(\theta) \,,$$

which is a sum of $N + 1$ functions. Hence, we would expect that $\|f\|_{C^m}$ scales like $\Theta(N)$.

(a) Suppose that we have a log-partition method with rate $\Theta_{m,d}(\|f\|_{C^m} n^{-m/d})$ and polynomial runtime $O_{m,d}(n^k)$. To achieve an error of $O(1)$ for the log-evidence L_f, this method would need $n = \Theta_{m,d}(N^{d/m})$ function evaluations and hence a runtime of $O_{m,d}(N^{kd/m})$. If $m = d$, the exponent kd/m is independent of the dimension.

(b) Now, suppose instead that the runtime is of the form $O_{m,d}(n^m)$ or the rates are of the form $O_{m,d}(\|f\| n^{-1/d})$ or $O_{m,d}(\|f\|^m n^{-m/d})$. In each case, to achieve an error of $O(1)$ for the log-evidence L_f, the resulting runtime would be polynomial in N, but the exponent would be proportional to d. ◀

1.1 Contribution

Our contributions are as follows:

(1) We analyze the information-based complexity of the sampling and log-partition problems, i.e., the worst-case optimal rates without computational constraints, in Section 2. For algorithms that evaluate f at a deterministic set of points, we show that the optimal rate for the log-partition problem is $\Theta_{m,d}(Bn^{-m/d})$, i.e., the same as for approximation. For the bounded total variation and 1-Wasserstein metrics, the optimal rate for sampling is $\Theta_{m,d}(\min\{1, Bn^{-m/d}\})$. For algorithms that are allowed to evaluate f at a stochastic set of points, we show that the optimal rates are the same in an optimization regime but can be faster in the high-temperature regime.

(2) We show reductions between different problems. For example, we analyze how log-partition algorithms can be employed for sampling and vice versa, and analyze the resulting guarantees for the rates. We also discuss how approximate sampling algorithms can be employed for optimization. Moreover, we show how function approximation yields reductions between different runtime complexities, convergence rates, and from stochastic to deterministic evaluation points.

(3) We analyze bounds on the convergence rates for different algorithms. For example, we show that it is possible to achieve the rate $O_{m,d}(Bn^{-m/d})$, but with runtime $O(n^m)$, which is polynomial but still involves the curse of dimensionality since we need $m = \Theta(d)$ to beat the curse of dimensionality in the convergence rate. We show that other simple and efficient algorithms also fail to achieve the optimal rates in different ways, sometimes with multi-regime behavior. Finally, we study an approach toward the log-partition problem by Bach (2022), whose optimization limit has been used by Woodworth et al. (2022) to obtain near-optimal optimization rates in polynomial-time. We show that all versions of this approach necessarily fail to exceed the rate $O_{m,d}(Bn^{-2/d})$ in an intermediate temperature regime $\varepsilon \sim n^{-2/d}$ (corresponding to $B \sim n^{2/d}$).

1.2 Related Work

The analysis of sampling algorithms has received considerable attention in recent years. In the case where p_f is (strongly) log-concave, that is, if f is (strongly) concave, convergence rates of Markov chain Monte Carlo (MCMC) sampling algorithms have been studied extensively. For example, good convergence rates in terms of the dimension d have been established for versions of the Langevin algorithm (Chewi et al., 2021; Altschuler and Talwar, 2022) and Hamiltonian Monte Carlo (Mangoubi and Vishnoi, 2018). Chewi et al. (2022b) establish an algorithm with optimal convergence rate for the case $d = 1$, while not much is known about algorithm-independent lower bounds in other cases.

For sampling from more general non-log-concave distributions, convergence rates have been established for versions of the Langevin algorithm. Bou-Rabee and Hairer (2013) showed an essentially geometric convergence result in TV distance for a class of non-log-concave Gibbs distributions, but without clear dependence of the constants on f. Mangoubi and Vishnoi (2019) and Zou et al. (2021) prove convergence rates that are polynomial in d but additionally depend on properties of f through the Cheeger constant. The analysis of Ma et al. (2019) and Cheng et al. (2018) is closer to our setting, and their convergence rate is polynomial in d as well, but their rate exhibits an exponential dependence on the Lipschitz constant of ∇f and the radius of the domain where f non-log-convex. Bou-Rabee et al. (2020) obtain similar results for Hamiltonian Monte Carlo. Balasubramanian et al. (2022) show that even for non-log-concave distributions, averaged Langevin Monte Carlo converges quickly to a distribution with low relative Fisher information to the target distribution, although this does not imply that the distribution is close to the target distribution with respect to other measures such as the total variation distance. Chewi et al. (2022a) prove corresponding lower bounds. Woodard et al. (2009) show that the mixing time of parallel and simulated tempering for certain distributions can scale exponentially with d, but in a setting different from ours. Achddou et al. (2019) propose and analyze an adaptive rejection sampling algorithm using a piecewise constant approximation of the density. Their setting is significantly different from ours as well, and they only consider functions of low (Hölder) smoothness and regimes with large n. Marteau-Ferey et al. (2022) propose an approximation-based sampling algorithm with a rate similar to $O_{m,d,B}(n^{-m/d})$ but without analyzing the dependence on B.

Another related line of work studies the relation of sampling to optimization. Through their analysis of Langevin algorithms in the non-log-concave setting, Ma et al. (2019) show that there are settings where sampling is easier than optimization. Talwar (2019) provides a simpler argument and shows that the converse can also occur for special function classes. The relation between sampling and optimization is also exploited in simulated annealing (Kirkpatrick et al., 1983). A different connection between sampling and optimization stems from Jordan et al. (1998), who showed that Langevin-type sampling can be interpreted as a gradient flow over distributions for the Wasserstein metric. For an overview of connections between sampling and optimization, we also refer to Cheng (2020).

The log-partition problem is often addressed through sampling algorithms, for example via thermodynamic integration (Kirkwood, 1935). For an overview of thermodynamic integration and other methods for the log-partition problem, we refer to Gelman and Meng (1998) and Friel and Wyse (2012). Ge et al. (2020) analyze an annealing algorithm combined with multilevel Monte Carlo sampling for the log-partition problem in the log-concave setting, and also give an information-based lower bound on the achievable convergence rate. Another popular approach is the Laplace approximation (Laplace, 1774), whose log-partition function however does not converge to the true log-partition function as $n \to \infty$. Well-tempered metadynamics (Barducci et al., 2008) is a popular approach towards the log-partition problem in molecular dynamics simulations, although it relies on a well-chosen low-dimensional collective variable representation. Recently, Marteau-Ferey et al. (2022) have suggested an approach that performs sampling via estimating the log-partition function. Bach (2023) and Bach (2022) suggest further approaches toward solving the log-partition problem.

To analyze possible convergence rates for the sampling and log-partition problem without computational constraints, we use the framework of information-based complexity. Here, we refer to Novak (1988) and Traub (2003) for an overview of this topic. In particular, our work is motivated by the works of Rudi et al. (2020) and Woodworth et al. (2022), who

demonstrated that for optimization, convergence rates close to the optimal rates from information-based complexity can be achieved in polynomial time.

The rest of our paper is organized as follows: In Section 2, we study upper and lower bounds for the information-based complexity of different variants of the sampling and log-partition problems. In Section 3, we study relations and reductions between different variants of the sampling, log-partition, and optimization problems. We then study convergence rates of different algorithms in Section 4. We compare some of these algorithms experimentally in Section 5 before concluding in Section 6. All proofs are provided in the appendix, which is structured analogously to the main part of this paper, and whose structure is overviewed in Appendix A.

2 Information-based Complexity

In this section, we look at the log-partition and sampling problems from the viewpoint of (worst-case) *information-based complexity*, where one is interested in what is possible if one is not constrained computationally but only by the number n of function evaluations of the unknown function f. We adopt the general setting of Novak (1988), where one is given a function space $\mathcal{F}$ (such as $\mathcal{F}_{d,m,B}$) of functions $f : \mathcal{X} \to \mathbb{R}$ and wishes to approximate a map $S : \mathcal{F} \to \mathcal{M}$, with the approximation error on $\mathcal{M}$ measured by a metric D. For example, the following problems are considered by Novak (1988):

- *Approximation*: $S_{\mathrm{app}}(f) \coloneqq f$ and $D_\infty(f, g) \coloneqq \|f - g\|_\infty$.
- *Optimization*: $S_{\mathrm{opt}^*}(f) \coloneqq \sup_{x\in\mathcal{X}} f(x)$ and $D_{\mathrm{abs}}(a, b) \coloneqq |a - b|$.
- *Integration*: $S_{\mathrm{int}}(f) \coloneqq \int_{\mathcal{X}} f(x)\,\mathrm{d}x$ and $D_{\mathrm{abs}}(a, b) = |a - b|$.

We can define our sampling and log-partition problems in this context as follows:

- *Log-partition*: $S_L(f) = L_f$ and $D_{\mathrm{abs}}(a, b) = |a - b|$.
- *Sampling*: While a sampling algorithm produces samples, we do not want to compare errors of individual samples but the error of the distribution of the samples. Therefore, we set $S_{\mathrm{samp}}(f) \coloneqq P_f$. For $D(P, Q)$, we can use different metrics or divergences on probability distributions, which will be discussed in Section 2.1.

2.1 Deterministic Evaluation Points

To consider minimax optimal convergence rates, we still need to define a space $\mathcal{A}$ of admissible maps $\tilde{S} : \mathcal{F} \to \mathcal{M}$. Here, we will first consider maps that evaluate functions in a deterministic set of points, before considering stochastic points in Section 2.2. For example, we define

$$\mathcal{A}_n \coloneqq \left\{\tilde{S} = \phi \circ N \mid N(f) = (f(x_1), \dots, f(x_n)) \text{ for some } x_1, \dots, x_n \in \mathcal{X}\right\},$$

the set of maps that only evaluate f in n deterministic and non-adaptive points. We can also allow adaptive points by defining

$$\mathcal{A}_n^{\mathrm{ad}} \coloneqq \left\{\tilde{S} = \phi \circ N \mid N(f) = (f(x_1), f(x_2(f(x_1))), \dots, f(x_n(f(x_1), \dots, f(x_{n-1}))))\right\},$$

where evaluation points may be chosen depending on previous function values. We are interested in the (non-adaptive/adaptive) minimax optimal error

$$e_n(\mathcal{F}, S, D) \coloneqq \inf_{\tilde{S}\in\mathcal{A}_n} \sup_{f\in\mathcal{F}} D(S(f), \tilde{S}(f)), \qquad e_n^{\mathrm{ad}}(\mathcal{F}, S, D) \coloneqq \inf_{\tilde{S}\in\mathcal{A}_n^{\mathrm{ad}}} \sup_{f\in\mathcal{F}} D(S(f), \tilde{S}(f)).$$

The sets $\mathcal{A}_n$ and $\mathcal{A}_n^{\mathrm{ad}}$ can be interpreted as classes of "black-box algorithms" that are only constrained in the number evaluations of f but not in terms of computational efficiency or computability. The minimax-optimal errors e_n and e_n^{ad} thus give lower bounds to what can be achieved by computationally efficient algorithms.

For the case of sampling, maps $\tilde{S} \in \mathcal{A}_n$ (or $\mathcal{A}_n^{\mathrm{ad}}$) produce distributions based on n function evaluations of a function f. They correspond to idealized sampling algorithms in the following sense: We consider an idealized sampling algorithm to take some source of randomness ω sampled from a distribution P_Ω independent of f, and then output a random sample $X_f(\omega) = \tilde{\phi}(N(f), \omega)$. For example, ω could be a sequence of i.i.d. random variables from the uniform distribution $\mathcal{U}([0, 1])$ on the interval $[0, 1]$. The maps $\tilde{S} \in \mathcal{A}_n$ (or $\mathcal{A}_n^{\mathrm{ad}}$) then correspond to the distributions produced by such sampling algorithms, i.e.,

$$\tilde{S}(f) = \text{ distribution of } X_f(\omega) \text{ for } \omega \sim P_\Omega.$$

The following theorem, which is proven in Appendix C.1, adapts known results on minimax optimal rates to our considered function spaces.

Theorem 5 (adapted from Novak (1988))**.** *We have*

$$\begin{aligned} e_n(\mathcal{F}_{d,m,B}, S_{\text{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\text{opt}^*}, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{opt}^*}, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\text{int}}, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{int}}, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}) \,. \end{aligned}$$

Novak (1988) gives these results in a form similar to $e_n(\mathcal{F}_{d,m,1}, S_{\text{app}}, D_\infty) = \Theta_{m,d}(n^{-m/d})$. This implies the rates for general $B \geq 0$ in the theorem above since S_{app} and D_∞ are positively homogeneous, which leads to $D_\infty(S_{\text{app}}(Bf), Bg) = BD_\infty(S_{\text{app}}(f), g)$. The same holds for optimization and integration, but not for log-partition estimation and sampling. Hence, for our considered problems, it is important to explicitly study the dependence on B, since it is not necessarily linear. The optimal rates for approximation can be achieved, for example, using piecewise polynomial interpolation, local polynomial reproductions, or moving least squares (Wendland, 2004), see also Theorem 15. The optimal rates for optimization and integration can be achieved by optimizing or integrating a corresponding approximation.

We use the following distance measures for probability distributions P, Q on $\mathcal{X}$:

- The sup-log distance $D_{\text{sup-log}}(P, Q) \coloneqq \left\| \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \right\|_\infty$, where $\|\cdot\|_\infty$ is taken over $\mathcal{X}$, and $D_{\text{sup-log}}(P, Q) = \infty$ whenever P and Q are not both absolutely continuous with respect to each other. The sup-log distance is a symmetrized version of the max-divergence $D_\infty(P||Q)$, which is used in differential privacy (Dwork et al., 2010) and is a special case of Rényi divergences for $\alpha = \infty$ (cf. Van Erven and Harremos, 2014). The sup-log distance is particularly well-suited to our setting, thanks to its relation to uniform approximation.
- The total variation distance $D_{\text{TV}}(P, Q) \coloneqq \sup_{A \subseteq \mathcal{X} \text{ measurable}} |P(A) - Q(A)|$.
- The 1-Wasserstein distance $W_1(P, Q) \coloneqq \inf_{X \sim P, Y \sim Q} \mathbb{E}\|X - Y\|_2$, also known as Kantorovich–Rubinstein or earth mover distance.

The sup-log, total variation, and 1-Wasserstein distances are metrics.[6] We first show that these quantities can be bounded in terms of the approximation error:

Proposition 6 (Upper bounds via approximation)**.** *For bounded and measurable $f, g : \mathcal{X} \to \mathbb{R}$, we have*

(a) $|L_f - L_g| \leq \|f - g\|_\infty$.
(b) $d^{-1/2} W_1(P_f, P_g) \leq D_{\text{TV}}(P_f, P_g) \leq D_{\text{sup-log}}(P_f, P_g) \leq 2\|f - g\|_\infty$.

Proposition 6 is proven in Appendix C.1. For the KL divergence, which we will not study further, we can leverage the results of Proposition 6 by using the trivial bound $D_{\text{KL}}(P \parallel Q) \leq D_{\text{sup-log}}(P, Q)$ as well as the inequality $D_{\text{KL}}(P \parallel Q) \leq D_{\text{sup-log}}(P, Q)(e^{D_{\text{sup-log}}(P,Q)} - 1)$ from Lemma III.2 of Dwork et al. (2010).

Theorem 5 and Proposition 6 lead to upper bounds on the minimax optimal rates. Combined with the trivial upper bound $D_{\text{TV}}(P, Q) \leq 1$, these are optimal for the deterministic point setting:

Theorem 7 (Information-based complexity of sampling and log-partition with deterministic evaluation points)**.** *We have*

$$\begin{aligned} e_n(\mathcal{F}_{d,m,B}, S_L, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\text{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{TV}}) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{TV}}) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\text{samp}}, W_1) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}), & e_n^{\text{ad}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, W_1) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}). \end{aligned}$$

Theorem 7 is proven in Appendix C.1. The minimax optimal rates for optimization can be related to those for approximation on a very general class of function spaces (Novak, 1988). For sampling, such a general relationship does not hold: For example, the set $\mathcal{F} \coloneqq \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\infty \leq 1, \{x \mid f(x) \neq 0\} \text{ is finite}\}$ satisfies $e_n(\mathcal{F}, S_{\text{app}}, D_\infty) = 1$ for all $n \in \mathbb{N}$, but all functions $f \in \mathcal{F}$ have the same distribution and the same log-partition function. However, our proofs for the lower bounds in Theorem 7 follow the general idea that underlies many lower bounds for Sobolev-type functions: place bumps with small support in regions that the algorithm does not query.

2.2 Stochastic Evaluation Points

We also want to consider methods that are allowed to choose the points x_i stochastically, such as Monte-Carlo type methods (Metropolis and Ulam, 1949; Brooks et al., 2011). For the log-partition problem, we again follow Novak

[6] For the sup-log distance, the triangle inequality follows from $\log\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right) = \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) + \log\left(\frac{\mathrm{d}Q}{\mathrm{d}R}\right)$.

(1988) and define the set ${}^*C(\mathcal{A}_n^{\mathrm{ad}})$ of random variables $\tilde{S} : \Omega \to \mathcal{A}_n^{\mathrm{ad}}$ with given base distribution P_Ω with associated minimax optimal error[7]

$$ {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}, S, D) \coloneqq \inf_{(\tilde{S}, P_\Omega) \in {}^*C(\mathcal{A}_n^{\mathrm{ad}})} \sup_{f \in \mathcal{F}} \mathbb{E}_{\omega \sim P_\Omega} D(S(f), \tilde{S}(\omega)(f)) \, . $$

When applying this definition to sampling, a map $\tilde{S}$ would output a *random* distribution. However, the random samples produced by a sampling algorithm typically still follow a *fixed* distribution, regardless of whether the function f is evaluated in deterministically or randomly chosen points. Hence, the model in Eq. (2) is inadequate for sampling. Instead, we consider again idealized sampling algorithms using some randomness $\omega \sim \Omega$, but this time, we allow the function to be evaluated in randomly and adaptively chosen points, by considering random samples of the form $X_f(\omega) = \tilde{\phi}(N(f, \omega), \omega)$. We then denote the corresponding map from f to P_{X_f} by $\tilde{S}$ and define the set $\mathcal{A}_n^{\text{ad-stoch}}$ of all $\tilde{S}$ that can be realized in this fashion using n function evaluations. We then define

$$ e_n^{\text{ad-stoch}}(\mathcal{F}, S_{\mathrm{samp}}, D) \coloneqq \inf_{\tilde{S} \in \mathcal{A}_n^{\text{ad-stoch}}} \sup_{f \in \mathcal{F}} D(S_{\mathrm{samp}}(f), \tilde{S}(f)) \, . $$

Unlike the deterministic points setting, the stochastic points setting potentially requires to evaluate f at n different points for every generated sample. This has the unintuitive consequence that for a map $\tilde{S} \in \mathcal{A}_n^{\text{ad-stoch}}$, the distribution $\tilde{S}(f)$ typically depends on the values of f at infinitely many points, but a sample from $\tilde{S}(f)$ can be drawn by only evaluating f at n (stochastic) points.

Again, results for approximation, optimization, and integration are known and can be adapted to our function classes:

Theorem 8 (adapted from Novak (1988))**.** *We have*

$$ \begin{aligned} {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}), \\ {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{opt}^*}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), \\ {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{int}}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-1/2-m/d}) \, . \end{aligned} $$

For a proof sketch, we refer to Appendix C.2. The faster rate for integration can be achieved by spending half of the n points for approximating f with g and spending the other half of the points on Monte Carlo quadrature to estimate the error (Novak, 1988)

$$ \int f(x) \, \mathrm{d}x - \int g(x) \, \mathrm{d}x = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{X})}[f(x) - g(x)] \, . $$

For a more practical algorithm, we refer to Chopin and Gerber (2022). For log-partition estimation, we can similarly use an importance sampling formulation

$$ L_f - L_g = \log \left(\mathbb{E}_{x \sim P_g} \left[\exp(f(x) - g(x)) \right] \right) \, . $$

Theorem 9 (Upper bound for stochastic log-partition)**.** *There exists a constant $C_{m,d} > 0$ depending only on m and d such that*

$$ {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) \leq O_{m,d}\left(\min\left\{ Bn^{-m/d}, \exp(C_{m,d} Bn^{-m/d}) Bn^{-1/2-m/d} \right\} \right) \, . $$

The upper bound above, which is proven in Appendix C.2, exhibits a fast transition between the rates $n^{-m/d}$ and $n^{-1/2-m/d}$. This is necessary, as we can exploit the relation of the log-partition problem to optimization to show that the rate $n^{-m/d}$ is optimal in an optimization regime:

Proposition 10 (Lower bound for stochastic log-partition)**.** *For $m \geq 1$, we have*

$$ {}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) \geq \Omega_{m,d}(Bn^{-m/d}) - d \log(1 + 3B) \, . $$

Proposition 10 is proven in Appendix C.2. We leave a lower bound outside of the optimization regime as an open problem, however, we conjecture that the rate $Bn^{-1/2-m/d}$ from the upper bound in Theorem 9 cannot be improved. For a certain class of strongly concave f with Lipschitz gradient, Theorem 5.1 by Ge et al. (2020) contains a lower bound which, in our setting, could be roughly expressed as $\Omega_{d,B}(n^{-1/(2-c/d)})$ for some constant c. A simple Taylor expansion shows $-1/(2 - c/d) \leq -1/2 - (c/4)/d$, hence this rate is compatible with our upper bound for $m = 2$ if $c \geq 8$.

To achieve better rates for sampling in the stochastic points setting, we combine approximation with a budget-limited version of rejection sampling defined in Algorithm 1. If g is shifted appropriately such that it upper-bounds f, we obtain the following convergence rate bound:

[7] Novak (1988) defines further variants, for example with $L_2(P_\Omega)$ instead of $L_1(P_\Omega)$ convergence or more limited stochastic resources, which we will not discuss here for simplicity.

Algorithm 1 Rejection sampling with proposal distribution P_g limited to n function evaluations.

function REJECTIONSAMPLING(f, g, number of steps n)
  for i from 1 to n **do**
    Sample $x \sim P_g$ and $u \sim \mathcal{U}([0,1])$
    Return x if $ue^{g(x)} \leq e^{f(x)}$
  end for
  return Sample from P_g
end function

Lemma 11 (General rejection sampling bound)**.** *Suppose that $f, g : \mathcal{X} \to \mathbb{R}$ are bounded and measurable with $f(x) \leq g(x)$ for all $x \in \mathcal{X}$. Then, the distribution $\tilde{P}_f$ of* REJECTIONSAMPLING(f, g, n) *satisfies*

$$\begin{aligned} \tilde{P}_f &= (1 - p_R)P_f + p_R P_g \\ D_{\text{sup-log}}(P_f, \tilde{P}_f) &\leq \min\left\{D_{\text{sup-log}}(P_f, P_g), p_R(\exp(D_{\text{sup-log}}(P_f, P_g)) - 1)\right\} \\ D_{\text{TV}}(P_f, \tilde{P}_f) &= p_R D_{\text{TV}}(P_f, P_g) \\ W_1(P_f, \tilde{P}_f) &= p_R W_1(P_f, P_g)\,, \end{aligned} \tag{2}$$

where $p_R = (1 - Z_f/Z_g)^n \leq \exp(-nZ_f/Z_g)$ is the probability of overall rejection.

The proof can be found in Appendix C.2. Due to the early stopping after n rejections, rejection sampling may significantly oversample regions where p_f is very small. Since $D_{\text{sup-log}}$ is very sensitive to this behavior, the corresponding bound is worse than for D_{TV} and W_1.

By using half of the n points to create an approximation g and then using a shifted version of g for rejection sampling with the other half of the n points, we obtain the following upper bound on the minimax optimal error:

Theorem 12 (Upper bound for sampling with stochastic evaluation points)**.** *There exists a constant $C_{m,d} > 0$ such that*

$$e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) \leq \begin{cases} O_{m,d}(Bn^{-m/d}) & , C_{m,d}Bn^{-m/d} > 1 \\ O_{m,d}((C_{m,d}Bn^{-m/d})^{n/2+1}) & , C_{m,d}Bn^{-m/d} \leq 1\,. \end{cases}$$

Theorem 12 is proven in Appendix C.2. Combinations of approximation and rejection sampling have also been used, for example, by Achddou et al. (2019) and Chewi et al. (2022b). For $C_{m,d}Bn^{-m/d} \leq 1$, the upper bound above decays faster than exponential in n. The bound is not tight, as the exponent $n/2 + 1$ can at least be improved close to n at the cost of increasing the constant $C_{m,d}$. However, for the optimization regime, the bound is tight:

Theorem 13 (Lower bound for sampling with stochastic evaluation points)**.** *There exists a constant $c_{m,d} > 0$ such that for $B > 0$ and $n \in \mathbb{N}$ with $Bn^{-m/d} \geq c_{m,d}(1 + \log(n))$, we have*

$$\begin{aligned} e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) &\geq \Omega_{m,d}(Bn^{-m/d}) \\ e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{TV}}) &\geq \Omega_{m,d}(1) \\ e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, W_1) &\geq \Omega_{m,d}(1)\,. \end{aligned}$$

The proof of Theorem 13 in Appendix C.2 uses the classical approach of hiding a bump, although explicitly exploiting the relation to optimization via Proposition 20 might also work. Proving lower bounds for sampling with stochastic points outside of the optimization regime seems difficult. Indeed, when restricting the function class a bit further, we can even achieve zero error:

Proposition 14. *Consider the function class $\mathcal{F} \coloneqq \{f \in C(\mathcal{X}) \mid \|f\|_\infty \leq \log(3/2), L_f = 0\}$. Then,*

$$e_n^{\text{ad-stoch}}(\mathcal{F}, S_{\text{samp}}, D_{\text{sup-log}}) = 0$$

for all $n \geq 1$.

The proof idea, executed in Appendix C.2, is simple: We use REJECTIONSAMPLING$(\tilde{f}, g, 1)$, where $\tilde{f}(x) \coloneqq \log(2\exp(f(x)) - 1)$ and $g(x) = \log(2)$ are constructed such that the resulting distribution is exactly P_f. The assumption that L_f is known is necessary to exactly control the acceptance probability in the rejection sampling step.

3 Relations Between Different Problems

In this section, we study how different problems such as sampling, log-partition estimation, and optimization are related, especially also in terms of reductions between algorithms, their runtime complexities, and their convergence rates. Again, certain bounds can be established via the connection to function approximation. For this, we need an efficient approximation method achieving optimal convergence rates while producing a smooth approximation. This is possible using the moving least squares method (Lancaster and Salkauskas, 1981), which produces an approximant $g(x) = g_x(x)$, where g_x is a local polynomial regression function fitted using a smooth local weight function $w(x_i, x)$. The following theorem shows that the moving least squares method achieves the desired properties:

Theorem 15 (adapted from Li (2016) and Mirzaei (2015))**.** *Let $m, d \in \mathbb{N}_{\geq 1}$. Using the moving least squares method, it is possible to construct an approximation f_n of $f \in C^m(\mathcal{X})$ using n deterministic non-adaptive function evaluations such that*

(a) $\|f - f_n\|_{C^k} \leq O_{m,d}(\|f\|_{C^m} n^{-(m-k)/d})$ for $k \in \{0, 1, \dots, m\}$,
(b) the runtime of constructing f_n is zero (construction takes place on-the-fly during evaluation), and
(c) the runtime of evaluating f_n at a point $x \in \mathcal{X}$ is $O_{m,d}(1)$.

We prove Theorem 15 in Appendix D.

3.1 Runtime-Accuracy Trade-off

When investigating sampling and log-partition algorithms, we study their convergence rate and their runtime complexity both in terms of the number n of required function evaluations. Here, we show that these two quantities can be traded off against each other to some extent. Improving the computational complexity at the cost of worse convergence rates is easy by increasing n without using the additional function values:

Example 16 (Trading convergence rates for better runtime complexity)**.** Suppose we have an algorithm A for the sampling or log-partition problems with convergence rate $\Theta_{m,d}(\|f\|_{C^m} n^{-\alpha_{m,d}})$ and runtime $\Theta_{m,d}(n^{\beta_{m,d}})$. We can then evaluate f in n points but only use $N \leq n$ of these points for A. If $N = \Theta_{m,d}(n^\gamma)$, $\gamma \in (0, 1]$, we obtain a convergence rate of $\Theta_{m,d}(\|f\|_{C^m} N^{-\alpha_{m,d}}) = \Theta_{m,d}(\|f\|_{C^m} n^{-\gamma\alpha_{m,d}})$ and a runtime of $\Theta_{m,d}(n + N^{\beta_{m,d}}) = \Theta_{m,d}(n^{\max\{1, \gamma\beta_{m,d}\}})$. A similar construction could be used to move constants $C_{m,d} \geq 1$ or potential factors $\|f\|_{C^m}^k \geq 1$ from the runtime to the convergence rate. ◀

Of course, the construction in Example 16 does not improve the runtime needed to reach a desired error level, but it shows that some combinations of runtime complexity and convergence rates are not better than others. To trade runtime complexity for better convergence rates, an analogous construction is not possible, since it would need to use $N > n$ function evaluations, which would contradict the definition of n. However, we can instead use N evaluations of an interpolant created using n function evaluations:

Example 17 (Trading runtime complexity for better convergence rates)**.** Suppose again that we have an algorithm A for the sampling or log-partition problems with convergence rate $\Theta_{m,d}(\|f\|_{C^m} n^{-\alpha_{m,d}})$ and runtime $\Theta_{m,d}(n^{\beta_{m,d}})$. We consider an algorithm resulting from the following construction:

(1) Use an approximation algorithm as in Theorem 15 to create an approximation f_n of f using n (deterministic) function evaluations.
(2) Run algorithm A on $N = \Theta_{m,d}(n^\gamma)$ function evaluations of f_n, $\gamma \in (0, \infty)$.

By Theorem 15, we have $\|f_n - f\|_\infty \leq O_{m,d}(\|f\|_{C^m} n^{-m/d})$, and by Proposition 6, this rate also applies to the considered distances of L_{f_n} to L_f or P_{f_n} to P_f. By the triangle inequality, the resulting algorithm has a convergence rate of

$$O_{m,d}(\|f\|_{C^m} n^{-m/d} + \|f_n\|_{C^m} N^{-\alpha_{m,d}}) = O_{m,d}(\|f\|_{C^m} n^{-\min\{m/d, \gamma\alpha_{m,d}\}}) ,$$

where we used $\|f_n\|_{C^m} \leq O_{m,d}(\|f\|_{C^m})$ due to Theorem 15 (a) with $k = m$, and runtime complexity

$$O_{m,d}(n^{\gamma\beta_{m,d}}) . \qquad \blacktriangleleft$$

While the construction in Example 17 also does not improve the runtime complexity needed to reach a desired error level, it can still be useful if evaluations of the approximant (or surrogate model) f_n are much cheaper than evaluations of f. This principle is used for example in computational chemistry, where expensive direct simulations f are approximated with machine-learned interatomic potentials f_n (Deringer et al., 2019).

3.2 Relation between stochastic and deterministic evaluation points

When the construction in Example 17 is applied to a sampling algorithm with stochastic evaluation points, it yields a sampling algorithm with deterministic evaluation points. This can be advantageous since the latter only needs n function evaluations to draw an arbitrary number of samples, while the former may require n new function evaluations for every drawn sample. On the other hand, this construction limits the convergence rate of the sampling algorithm to $\Omega_{m,d}(Bn^{-m/d})$, a rate which can be improved by sampling algorithms with stochastic evaluation points outside of the optimization regime (cf. Theorem 12).

Applying the construction in Example 17 to a log-partition algorithm with stochastic evaluation points yields a stochastic log-partition algorithm with deterministic evaluation points. We did not consider such algorithms separately in Section 2.2. However, such an algorithm is never better than its median or expected output, which is a deterministic log-partition method with deterministic evaluation points. Hence, it follows from Theorem 7 that the convergence rate of the construction in Example 17 is limited to $\Omega_{m,d}(Bn^{-m/d})$, and this rate can be improved by log-partition algorithms with stochastic evaluation points outside of the optimization regime (cf. Theorem 9).

3.3 Relation Between Sampling and Log-partition Estimation

A natural question is whether efficient sampling algorithms can be used to obtain efficient log-partition estimators and vice versa. We study both of these directions in the following. In fact, sampling algorithms are frequently employed for log-partition estimation in computational statistical physics and other fields (Frenkel and Smit, 2001; Friel and Wyse, 2012). One method to achieve this is thermodynamic integration (Kirkwood, 1935), of which we present a particularly simple version here. By integrating the derivative of $L(\beta) \coloneqq L_{\beta f}$, it is possible to derive the following formula (Gelman and Meng, 1998; Friel and Wyse, 2012):

$$L_f = \int_0^1 \mathbb{E}_{x\sim P_{\beta f}}[f(x)]\,\mathrm{d}\beta = \mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta f}} f(x)\,.$$

Thermodynamic integration can be used more generally to estimate a difference $L_f - L_g$ by integrating along a path between f and g. In practice, the inner expectation is typically evaluated by Monte Carlo methods using sampling algorithms to sample from $P_{\beta f}$, while the outer integral is typically approximated with a suitable (deterministic) quadrature rule. For convenience of analysis, we will consider the case where both expectations are approximated using Monte Carlo quadrature:

Theorem 18 (Convergence of thermodynamic integration)**.** *Given $N \in \mathbb{N}_{\geq 1}$ and a sampling algorithm producing samples from approximate distributions $\tilde{P}_{\beta f}$, consider the following algorithm:*

- *Sample $\beta_1, \dots, \beta_N \sim \mathcal{U}([0,1])$ independently.*
- *Draw $X_i \sim \tilde{P}_{\beta_i f}$ independently.*
- *Output $\tilde{L}_f \coloneqq \frac{1}{N}\sum_{i=1}^N f(X_i)$.*

Then, for $\delta > 0$, we have

$$|\tilde{L}_f - L_f| \leq |L_f - \mathbb{E}\tilde{L}_f| + 2\|f\|_\infty \sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability $\geq 1-\delta$, where

$$|L_f - \mathbb{E}\tilde{L}_f| \leq 2\|f\|_\infty \sup_{\beta\in[0,1]} D_{\mathrm{TV}}(P_{\beta f}, \tilde{P}_{\beta f}),$$
$$|L_f - \mathbb{E}\tilde{L}_f| \leq |f|_1 \sup_{\beta\in[0,1]} W_1(P_{\beta f}, \tilde{P}_{\beta f}).$$

Theorem 18 is proven in Appendix D.1. In the upper bounds above, we obtain additional factors $\|f\|_\infty$ or $|f|_1$, which deteriorate the convergence rate. While it appears that these factors are in general necessary for the TV and 1-Wasserstein distances, we explain in Remark D.2 that better bounds in terms of $D_{\text{sup-log}}$ seem plausible but appear to be more difficult to prove. When considering the runtime complexity and convergence rate of the construction in Theorem 18, it is important to set them in relation to the total number n of function evaluations used. For example, if sampling from $P_{\beta_i f}$ uses $\tilde{n}$ function evaluations, then in general $n = (\tilde{n}+1)N$. If the employed sampling algorithm is non-adaptive with deterministic evaluation points, we only need $n = \tilde{n} + N$ function evaluations. Still, due to the Monte Carlo nature of thermodynamic integration, the convergence rate is at least limited to $\Omega_{m,d,f}(n^{-1/2})$, which is not optimal as we showed in Theorem 9. Of course, thermodynamic integration can be performed on top of an approximation of f instead, similar to Example 17.

Algorithm 2 Bisection sampling algorithm using a log-partition algorithm $\tilde{L}$.

function BISECTIONSAMPLING(Function $f : \mathcal{X} \to \mathbb{R}^d$, Log-partition algorithm $\tilde{L}$, Number $M \in \mathbb{N}_0$ of bisection steps per dimension)
  For a hyperrectangle $\mathcal{Z} = \bigtimes_{i=1}^{d} [z_i, z_i + h_i]$, define $f_{\mathcal{Z}} : \mathcal{X} \to \mathbb{R}$ by $f_{\mathcal{Z}}(x) \coloneqq f(z_1 + h_1 x_1, \ldots, z_d + h_d x_d)$
  $\mathcal{Z} \leftarrow \mathcal{X}$
  for i from 1 to M **do**
    for j from 1 to d **do**
      Split $\mathcal{Z}$ along dimension j into two equal-sized hyperrectangles $\mathcal{Z}_1$ and $\mathcal{Z}_2$
      Compute $p_1 \coloneqq \sigma(\tilde{L}_{f_{\mathcal{Z}_1}} - \tilde{L}_{f_{\mathcal{Z}_2}})$, where $\sigma(u) = (1 + \exp(-u))^{-1}$ is the sigmoid function
      Sample $k = 1$ with probability p_1 and $k = 2$ otherwise
      $\mathcal{Z} \leftarrow \mathcal{Z}_k$
    end for
  end for
  return sample from the uniform distribution $\mathcal{U}(\mathcal{Z})$
end function

Now, we ask the converse question: Can an efficient log-partition algorithm be used for efficient sampling? To achieve such a reduction, we note that we can apply a log-partition algorithm not only to the target function f but also for example to multiple shifted and rescaled versions of f, which amounts to computing the log-partition function on subsets of the cube $\mathcal{X}$. This is exploited in Algorithm 2, which we refer to as bisection sampling. Bisection sampling has been studied for example by Marteau-Ferey et al. (2022). We give an upper bound on its error in the sup-log distance:

Theorem 19 (Convergence of bisection sampling)**.** *Let $m \geq 1, B \geq 0$ and $M \in \mathbb{N}_0$. Let $f \in \mathcal{F}_{d,m,B}$ and let $\tilde{L}$ be a log-partition estimator with worst-case error $E \geq 0$ on $\mathcal{F}_{d,m,B}$. Let $f \in C^m(\mathcal{X})$ and let $\tilde{P}_f$ be the distribution of samples produced by* BISECTIONSAMPLING$(f, \tilde{L}, M)$ *in Algorithm 2. Then,*

$$D_{\text{sup-log}}(\tilde{P}_f, P_f) \leq 2MdE + 2^{-M} d \|f\|_{C^1} \, .$$

Of course, Theorem 19, which is proven in Appendix D.1, also implies bounds on the TV and 1-Wasserstein distances using Proposition 6. The first term in the upper bound grows with M, which stems from the possibility to make an error of order $2E$ per loop iteration. However, when the resulting error decays quickly enough in the loop, it is possible to make the first term independent of M. For example, this could arise because the log-partition algorithm achieves smaller errors for smoother functions. It is also possible if we consider the 1-Wasserstein distance, which provides better error bounds on smaller hyperrectangles.

In order to analyze the resulting convergence rates, suppose that the log-partition algorithm $\tilde{L}$ uses N evaluation points. Ignoring rounding issues, we can set $M = \log_2(N^{m/d})$ and obtain the rate

$$D_{\text{sup-log}}(\tilde{P}_f, P_f) \leq O_{m,d}(E \log(N) + \|f\|_{C^1} N^{-m/d}) \, ,$$

where BISECTIONSAMPLING uses up to $n \coloneqq 2MdN = O_{m,d}(N \log(N))$ function evaluations. Hence, we typically only lose polylogarithmic terms in the convergence rate, unlike for thermodynamic integration. Even for log-partition algorithms with deterministic evaluation points, the resulting sampling algorithm uses stochastic evaluation points. Again, bisection sampling can be performed on top of an approximation of f instead, similar to Example 17.

3.4 Relation to Optimization

Due to the relationship between sampling and optimization, a natural question is in which sense approximate sampling algorithms can perform approximate optimization. Actually, we can consider two kinds of optimization problems, similar to Novak (1988):

- (OPT) The problem of outputting $x \in \mathcal{X}$ such that $|M_f - f(x)|$ is small can be seen as the low-temperature limit of the sampling problem.
- (OPT*) The problem of outputting an estimate $\tilde{M}_f$ such that $|M_f - \tilde{M}_f|$ is small can be seen as the low-temperature limit of the log-partition problem.

As special cases of the reductions between sampling and log-partition estimation in Section 3.3, we can obtain reductions between (OPT) and (OPT*): For (OPT*), we can simply evaluate f at the estimate x obtained from (OPT), which can be seen as a simple special case of thermodynamic integration. On the other hand, for (OPT), we can recursively use

(OPT*) to see whether the optimum is contained in a subdomain of $\mathcal{X}$, which corresponds to the low-temperature limit of bisection sampling.

To obtain a bound for approximate (OPT*) via approximate log-partition estimation $\tilde{L}$, we note that Lemma 3 directly yields

$$|M_f - \varepsilon\tilde{L}_{f/\varepsilon}| \leq |M_f - \varepsilon L_{f/\varepsilon}| + \varepsilon|L_{f/\varepsilon} - \tilde{L}_{f/\varepsilon}| \leq \varepsilon d \log(1 + 3d^{-1/2}\varepsilon|f|_1) + \varepsilon|L_{f/\varepsilon} - \tilde{L}_{f/\varepsilon}|$$

for temperatures $\varepsilon > 0$.

When performing approximate (OPT) via sampling from an approximate distribution $Q = \tilde{P}_f$, the result depends on the employed distance metric. Since the result of sampling is of course stochastic, we will upper-bound probabilities of the form $Q(\{x \in \mathcal{X} \mid f(x) \leq \alpha\})$ of obtaining a function value $f(x) \leq \alpha$ when drawing x from Q.

Proposition 20 (Optimization by approximate sampling)**.** *Let Q be a probability distribution on $\mathcal{X}$. Then, for any $\delta \in (0, 1]$ and $\varepsilon > 0$,*

(a) $Q(\{x \in \mathcal{X} \mid f(x) \leq \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta) - \varepsilon D_{\text{sup-log}}(P_{f/\varepsilon}, Q)\}) \leq \delta$,
(b) $Q(\{x \in \mathcal{X} \mid f(x) \leq \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)\}) \leq \delta + D_{\text{TV}}(P_{f/\varepsilon}, Q)$,
(c) $Q(\{x \in \mathcal{X} \mid f(x) < \varepsilon L_{f/\varepsilon} - \varepsilon \log(2/\delta) - 2\delta^{-1}|f|_1 W_1(P_{f/\varepsilon}, Q)\}) \leq \delta$.

Proposition 20 is proven in Appendix D.2. If $Q = P_{g/\varepsilon}$ for some bounded $g : \mathcal{X} \to \mathbb{R}$, the bound in (a) recovers the known optimization bound $f(\operatorname{argmax} g) \geq M_f - 2\|f - g\|_\infty$ in the limit $\varepsilon \searrow 0$ using Lemma 3 and Proposition 6. We can deduce from (a) that a sampling algorithm achieving the optimal rate $O_{m,d}(\|f\|_{C^m} n^{-m/d})$ in terms of $D_{\text{sup-log}}$ can be used (with sufficiently small ε) to achieve the optimal rate for (OPT) as well. On the other hand, the bounds (b) and (c) are much weaker. For example, (a) still gives a good bound for $D_{\text{sup-log}}(P_{f/\varepsilon}, Q) = 1/2$, but (b) only gives a low-probability bound for $D_{\text{TV}}(P_{f/\varepsilon}, Q) = 1/2$ and (c) is trivial for $W_1(P_{f/\varepsilon}, Q) = 1/2$. Note that an argument analogous to (b) has been used in Corollary 1 by Ma et al. (2019) to analyze the convergence of Langevin algorithms for approximate optimization.

4 Algorithms

In this section, we investigate the convergence rates of different algorithmic approaches toward the sampling and log-partition problems.

4.1 Approximation-based Algorithms

First, we study approximation-based algorithms. In Section 2.1, we have seen that in principle, approximation-based methods can achieve the optimal rates for the sampling and log-partition problems with deterministic points. However, for most approximations g, it is unclear how to sample from P_g or compute L_g. In the following, we will consider a few cases where this is possible:

4.1.1 Piecewise Constant Approximation

A very simple approximation method is piecewise constant approximation. Here, we study the convenient setting where $n = N^d$ for some $N \in \mathbb{N}$:

- Divide $\mathcal{X}$ into N^d equally-sized cubes $\mathcal{X}_1, \ldots, \mathcal{X}_n$ by dividing $[0, 1]$ into N intervals.
- Output the function g_n that is piecewise constant on each cube and interpolates f at the center $x^{(i)}$ of the cube $\mathcal{X}_i$. Boundary points can be assigned to an arbitrary adjacent cube.

Given a piecewise constant function g_n, we can easily compute $L_{g_n} = \log\left(\frac{1}{n}\sum_{i=1}^n e^{f(x^{(i)})}\right)$ in time $O_{m,d}(n)$. Similarly, we can sample from g_n in time $O_{m,d}(n)$ by first sampling a subcube $\mathcal{X}_i$ with probability $p_i = e^{f(x^{(i)}) - L_{g_n}}$ and then drawing a uniform random sample from $\mathcal{X}_i$.[8] However, the convergence rate is bad, as we prove in Appendix E.1.1:

Theorem 21 (Convergence rate of piecewise constant approximation)**.** *Let $m \geq 1$ and $n = N^d$ as above. If g_n is a piecewise constant interpolant as above, we have*

$$\sup_{f \in \mathcal{F}_{d,m,B}} |L_f - L_{g_n}| = \begin{cases} \Theta_{m,d}(Bn^{-1/d}) & \text{, if } m = 1 \text{ or } Bn^{-1/d} > 1 \\ \Theta_{m,d}(\max\{B, B^2\}n^{-2/d}) & \text{, otherwise.} \end{cases}$$

[8] This could be improved to $O_{m,d}(\log n)$ by precomputing partial sums of the p_i once in $O(n)$. Then, for drawing a sample, we can sample $u \sim \mathcal{U}(\{1, \ldots, n\})$ and use binary search to find the correct "bucket" k with $\sum_{i=1}^{k-1} p_i \leq u < \sum_{i=1}^k p_i$.

$$\sup_{f \in \mathcal{F}_{d,m,B}} D_{\text{sup-log}}(P_f, P_{g_n}) = \Theta_{m,d}(Bn^{-1/d}) \,.$$

The rates of piecewise constant approximation are thus optimal for $m = 1$, but not for $m > 1$. For $m > 1$, using a combination with higher-order function approximation as in Example 17, it is possible to achieve the rate $O_{m,d}(Bn^{-m/d})$ with runtime $O_{m,d}(n^m)$. The result above also shows that the piecewise constant log-partition method can achieve the faster rate $O(n^{-2/d})$ of midpoint quadrature only outside of the optimization regime. We leave it as an open problem whether such faster rates are also achieved for sampling with D_{TV} or W_1. Achddou et al. (2019) analyze a combination of piecewise constant approximation with rejection sampling, but in a setting incomparable to ours. They also note that piecewise constant approximation achieves optimal rates for Hölder classes of functions.

Beyond piecewise constant approximations, piecewise linear approximations also allow for efficient sampling and log-partition estimation, and they should allow to achieve convergence rates of $O_{m,d}(Bn^{-2/d})$. We leave a precise analysis of this approach as an open problem.

4.1.2 Density-based Approximation

Another option to obtain tractable sampling and log-partition algorithms is to directly approximate the unnormalized density $p(x) = e^{f(x)}$. Since probability distributions are normalized, approximating λp with λq yields the same sampling and log-partition errors as approximating p with q, but the approximation error $\|\lambda p - \lambda q\|_\infty$ depends on $\lambda > 0$. To obtain a scale-invariant bound for the sampling and log-partition errors, we need to divide the approximation bound by a normalization constant:

Proposition 22 (Density approximation bounds)**.** *Let $p, q : \mathbb{R} \to [0, \infty)$ be bounded and measurable such that $I_p, I_q > 0$, where $I_p \coloneqq \int_{\mathcal{X}} p(x) \,\mathrm{d}x$. Define probability distributions P, Q with densities p/I_p and q/I_q, respectively. Then,*

$$|\log I_p - \log I_q| \leq \log\left(\frac{1}{1 - \|p - q\|_\infty / I_p}\right) \quad \text{if } \|p - q\|_\infty < I_p,$$
$$D_{\text{TV}}(P, Q) \leq \frac{3\|p - q\|_\infty}{\max\{I_p, I_q\}} \leq \frac{3\|p - q\|_\infty}{I_p} \,.$$

We prove Proposition 22 in Appendix E.1.2. While $\|p\|_\infty = \|e^f\|_\infty$ can be exponential in $\|f\|_\infty$, Proposition 22 demonstrates that we need to incorporate the normalization constant to obtain reasonable estimates for sampling and log-partition computation. After incorporating the normalization constant, we arrive at terms of the form $\|p\|/I_p = \|e^f\|/Z_f = \|e^f/Z_f\| = \|p_f\|$. Hence, the norm of the (normalized) density plays an important role for convergence rates of density-based approximation approaches. As it turns out, $\|p_f\|$ does not scale exponentially in $\|f\|_\infty$, but still badly:

Theorem 23 (Density norm)**.** *For $m \geq 1$, we have*

$$\sup_{f \in \mathcal{F}_{d,m,B}} \|p_f\|_{C^m} = \Theta_{m,d}\left(\max\{1, B\}^{m+d}\right)$$

and this asymptotic rate is attained by $f_{d,m,B}(x) = Bd^{-1}(x_1 + \cdots + x_d)$.

Theorem 23 is proven in Appendix E.1.2. Suppose that $f \in C^m(\mathcal{X})$ and we can approximate $p(x) = e^{f(x)}$ with a non-negative function q with optimal[9] rate $O_{m,d}(\|p\|_{C^m} n^{-m/d})$. By combining Proposition 22 and Theorem 23, the distribution Q associated with the unnormalized density q then satisfies

$$D_{\text{TV}}(P_f, Q) \leq O_{m,d}(\max\{1, \|f\|_{C^m}\}^{m+d} n^{-m/d}) \,. \tag{3}$$

Although this rate is optimal in terms of n for deterministic evaluation points, it is bad in terms of $\|f\|_{C^m}$, cf. also Example 4.

Marteau-Ferey et al. (2022) propose a sampling algorithm based on approximating the density with a (non-negative) sum-of-squares model. Specifically, for a Gibbs distribution, they suggest approximating $\sqrt{p}$ with q and then using q^2 as an unnormalized density. They achieve a rate of $O_{m,d,f}(n^{-m/d})$ in polynomial time without explicitly stating the dependence on $\|f\|$, but we conjecture that the dependence on $\|f\|$ is similar to Eq. (3).

[9]This rate is worst-case optimal if we only know $\|p\|_{C^m}$. However, it might not be optimal over the smaller class of functions of the form e^f with small $\|f\|_{C^m}$.

4.2 Simple Stochastic Algorithms

We now analyze the convergence rates for some simple stochastic algorithms.

4.2.1 Rejection Sampling With Uniform Proposal Distribution

A simple stochastic algorithm is rejection sampling with a uniform proposal distribution. The following proposition shows that this can achieve better rates in terms of the TV distance than the density-based approximation rates in Eq. (3) if the maximum M_f of f is known:

Proposition 24 (Convergence of rejection sampling)**.** *Let $m \geq 1$ and let $f \in C^1(\mathcal{X})$. Then, the distribution $\tilde{P}_f$ produced by* RejectionSampling*(f, M_f, n) (see Algorithm 1) satisfies*

$$D_{\text{sup-log}}(P_f, \tilde{P}_f) \leq \min\{2\|f\|_\infty, \exp(2\|f\|_\infty - n/\|p_f\|_\infty)\}$$
$$D_{\text{TV}}(P_f, \tilde{P}_f) \leq \min\{1, 2\|f\|_\infty\} \exp(-n/\|p_f\|_\infty) \leq O_{m,d}(\min\{1, \|f\|_\infty\} \max\{1, \|f\|_{C^1}\}^m n^{-m/d}) .$$

A proof can be found in Appendix E.2.1. Lower bounds for the convergence of rejection sampling could be obtained using Lemma 11, but the resulting formula would not be easy to interpret. In any case, an argument similar to the one in Section 4.2.3 and Appendix E.2.3 can be made to show that RejectionSampling(f, M_f, n) cannot achieve the rate $O_{m,d}(Bn^{-m/d})$. We leave it as an open question whether similar rates to Proposition 24 can be achieved when M_f is approximately known or when a guess for M_f is used that slowly increases with n. Note that Talwar (2019) studies a similar setting where rejection sampling is not stopped after n rejections.

4.2.2 Monte Carlo Log-partition

Since the log-partition problem involves an integral, it is natural to approximate the integral by Monte Carlo (MC) quadrature. The following theorem gives an upper bound on the convergence rate:

Theorem 25 (Upper bounds for MC log-partition)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz, let $X_1, \dots, X_n \sim \mathcal{U}(\mathcal{X})$ be independent and let*

$$\tilde{L}_n \coloneqq \log S_n, \qquad S_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} \exp(f(X_i)).$$

Then, for any $\delta \in (0, 1]$, the following convergence rates hold:

(a) ***Optimization regime:*** *If $n \leq 4\log(2/\delta)(1 + 3d^{-1/2}|f|_1)^d$, we have*

$$|\tilde{L}_n - L_f| \leq d^{1/2}(\log(1/\delta))^{1/d}|f|_1 n^{-1/d} + \log(4\log(2/\delta)) + d\log(1 + 3d^{-1/2}|f|_1)$$

with probability $\geq 1 - \delta$.

(b) ***Quadrature regime:*** *If $n \geq 4\log(2/\delta)(1 + 3d^{-1/2}|f|_1)^d$, we have*

$$|\tilde{L}_n - L_f| \leq 4\log(2/\delta)^{1/2}(1 + 3d^{-1/2}|f|_1)^{d/2} n^{-1/2}$$

with probability $\geq 1 - \delta$.

We prove Theorem 25 in Appendix E.2.2. Roughly speaking, the rates in the theorem above behave like $|f|_1 n^{-1/d}$ until an error of $O(1)$ is reached, and then they change to $|f|_1^{d/2} n^{-1/2}$. Figure 1 and our experiments later in Figure 2 show that this reflects the qualitative behavior of the error on linear f in practice.

4.2.3 Monte Carlo Sampling

We can also consider a sampling version of the Monte Carlo log-partition method considered in Section 4.2.2. The following theorem shows that it cannot achieve good rates in the optimization regime either:

Theorem 26 (Lower bound for MC sampling)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be bounded and measurable. Let $X_1, \dots, X_n \sim \mathcal{U}(\mathcal{X})$ and let the random index $I \in \{1, \dots, n\}$ be distributed as*

$$P(I = i) = \frac{\exp(f(X_i))}{\sum_{j=1}^{n} \exp(f(X_j))}.$$

Consider the distribution $\tilde{P}_f$ of the random sample X_I. Then, for all $B > 0$ and $n \geq 1$ with $Bn^{-1/d} \geq 4d\log(4d)$,

$$\sup_{f \in \mathcal{F}_{d,m,B}} D_{\text{TV}}(P_f, \tilde{P}_f) \geq \frac{1}{2} .$$

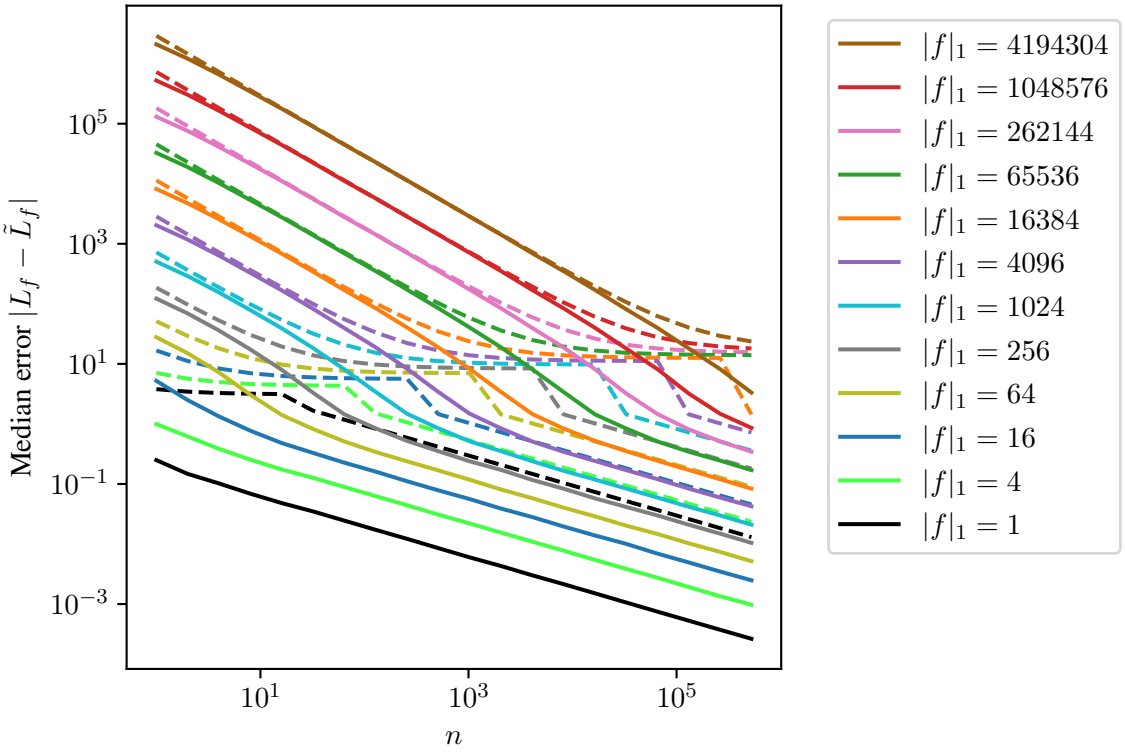

Figure 1: Median error of MC log-partition for the function $f : [0, 1] \to \mathbb{R}, x \mapsto \beta x$ (with $d = 1$), for varying numbers of points n and values of $\beta = |f|_1 > 0$. Medians were computed out of 10001 repetitions. The dashed lines show the corresponding upper bounds from Theorem 25 for the median ($\delta = 1/2$).

The lower bounds in Theorem 26, proven in Appendix E.2.3, show that $n \geq \Omega_{m,d}(B^d)$ points are required to achieve an error below $O(1)$, which is significantly worse than the around $\Theta_{m,d}(B^{d/m})$ points required by a method with the optimal rate for deterministic evaluation points. The proof only uses that the density $\tilde{p}_f$ is upper-bounded by n, and would apply analogously (using $n + 1$ instead of n) to rejection sampling with uniform proposal distribution as considered in Section 4.2.1.

4.3 Markov Chain Monte Carlo Algorithms

Markov Chain Monte Carlo (MCMC) methods are a very popular class of sampling algorithms. In particular, gradient-based MCMC algorithms such as versions of Langevin MCMC and Hamiltonian Monte Carlo (Duane et al., 1987) have been studied intensively in recent years. While most theoretical guarantees only consider the case of concave f, there have been a few extensions where f is allowed to be non-concave in a compact region of the domain. For example, Ma et al. (2019) study a certain class of functions whose gradient is L-Lipschitz and that are non-concave in a region with radius R but α-strongly convex outside of it. For the Metropolis-adjusted Langevin algorithm (MALA) to reach a TV distance error $E > 0$, they obtain the mixing time bound

$$n \leq O\left(\frac{e^{40LR^2}}{\alpha}(L/\alpha)^{3/2} d^{1/2} (d \ln(L/\alpha) + \ln(1/E))^{3/2}\right). \tag{4}$$

We have used n here for the mixing time since it corresponds to the number of gradient evaluations, which are potentially more informative than the function evaluations normally allowed in our setting but can be approximated using $d + 1$ function evaluations. The dependence of the upper bound in Section 4.3 on L, which is related to $\|f\|_{C^2}$ in our setting, is exponential. We are not aware of a lower bound, but conjecture that a tight lower bound will also have an exponential dependence on $\|f\|$ in some fashion. This indicates that Langevin MCMC could perform worse than rejection sampling in our setting.

Beyond Langevin MCMC, there are many other popular MCMC methods, for example, variants of Hamiltonian Monte Carlo, parallel tempering (or replica exchange MCMC), and simulated tempering. Obtaining convergence rates for

these methods on function classes like $\mathcal{F}_{d,m,B}$ is an interesting problem but left open in this paper. While Woodard et al. (2009) prove torpid (slow) mixing for parallel and simulated tempering in some settings, they show an exponential dependency on d for certain mixtures of Gaussians, which does not appear to imply suboptimal rates in our setting.

4.4 Variational Formulation for Log-Partition Estimation

In the following, we will introduce the variational approach to the log-partition problem by Bach (2022). We will first start with the simpler optimization setting. Let $\mathcal{P}(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. We start with the formulation

$$M_f = \max_{x \in \mathcal{X}} f(x) = \sup_{P \in \mathcal{P}(\mathcal{X})} \int f(x) \, \mathrm{d}P(x) \,, \tag{5}$$

which converts a finite-dimensional non-concave optimization problem into an infinite-dimensional convex optimization problem. To apply the approach by Bach (2022), we need to approximate the function f by a model of the form

$$g(x) = \varphi(x)^* H \varphi(x),$$

where H is a Hermitian matrix and $\varphi : \mathcal{X} \to \mathbb{C}^N$ is a suitable feature map. For example, for $d = 1$, if f is periodic and we use Fourier features $\varphi(x) = (1, e^{ix}, \dots, e^{(N-1)ix})^\top$, then H can be determined by trigonometric interpolation, see also Woodworth et al. (2022).

For a probability distribution $P \in \mathcal{P}(\mathcal{X})$, we define the moment matrix

$$\Sigma_P \coloneqq \int_{\mathcal{X}} \varphi(x) \varphi(x)^* \, \mathrm{d}P(x) \,.$$

Because of

$$\int_{\mathcal{X}} g(x) \, \mathrm{d}P(x) = \int_{\mathcal{X}} \operatorname{tr}[\varphi(x)^* H \varphi(x)] \, \mathrm{d}x = \int_{\mathcal{X}} \operatorname{tr}[H \varphi(x) \varphi(x)^*] \, \mathrm{d}x = \operatorname{tr}[H \Sigma_P] \,,$$

we then obtain

$$M_g = \sup_{P \in \mathcal{P}(\mathcal{X})} \operatorname{tr}[H \Sigma_P] = \sup_{\Sigma \in \mathcal{K}} \operatorname{tr}[H \Sigma] \,,$$

where $\mathcal{K}$ is the (convex) set of all possible values of Σ_P. This reduces the infinite-dimensional convex optimization problem in Eq. (5) to a finite-dimensional convex optimization problem, and at least for certain feature maps, the set $\mathcal{K}$ has a sufficiently nice structure for optimization.

To extend this approach to the log-partition problem, Bach (2022) uses the following variational formulation by Donsker and Varadhan (1983) for general base distributions Q, where $D_{\mathrm{KL}}(P \parallel Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P$ is the KL divergence:

$$\begin{aligned} L_f(Q) &\coloneqq \log \int_{\mathcal{X}} e^{f(x)} \, \mathrm{d}Q(x) \\ &= \sup_{P \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} f(x) \, \mathrm{d}P(x) - D_{\mathrm{KL}}(P \parallel Q) \,. \end{aligned}$$

Again, after approximating f by g, we can replace the integral by $\operatorname{tr}[H \Sigma_P]$. However, to obtain a finite-dimensional optimization problem, we also need to replace the KL divergence with something that only depends on Σ_P instead of P. Since this is not possible exactly, Bach (2022) proposes multiple lower bounds, of which the tightest one (and most difficult to compute) is

$$D_{\mathrm{KL}}^{\mathrm{OPT}}(\Sigma_P \parallel \Sigma_Q) \coloneqq \inf_{\tilde{P}, \tilde{Q} \in \mathcal{P}(\mathcal{X}) : \Sigma_P = \Sigma_{\tilde{P}}, \Sigma_Q = \Sigma_{\tilde{Q}}} D_{\mathrm{KL}}(\tilde{P} \parallel \tilde{Q}) \,.$$

This yields the following upper bound on the log-partition function:

$$\begin{aligned} L_g^{\mathrm{OPT}}(Q) &\coloneqq \sup_{P \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} g(x) \, \mathrm{d}P(x) - D_{\mathrm{KL}}^{\mathrm{OPT}}(\Sigma_P \parallel \Sigma_Q) \\ &= \sup_{P \in \mathcal{P}(\mathcal{X})} \operatorname{tr}[H \Sigma_P] - D_{\mathrm{KL}}^{\mathrm{OPT}}(\Sigma_P \parallel \Sigma_Q) = \sup_{\Sigma \in \mathcal{K}} \operatorname{tr}[H \Sigma] - D_{\mathrm{KL}}^{\mathrm{OPT}}(\Sigma \parallel \Sigma_Q) \geq L_g(Q) \,. \end{aligned} \tag{6}$$

Our investigation begins here: After inserting the definition of $D_{\mathrm{KL}}^{\mathrm{OPT}}$, a simple calculation shows that due to the minus sign, the infimum over $\tilde{P}$ merges with the supremum over P, and the infimum over $\tilde{Q}$ turns into a supremum:

Lemma 27. *For a model of the form $g(x) = \varphi(x)^* H \varphi(x)$ as above, we have*

$$L_g^{\mathrm{OPT}}(Q) = \sup_{\tilde{Q} \in \mathcal{P}(\mathcal{X}) : \Sigma_{\tilde{Q}} = \Sigma_Q} L_g(\tilde{Q}) \, .$$

This formulation allows us to show a lower bound on the achievable convergence rate. The basic idea is as follows: Since Q is only known through finitely many moments Σ_Q, we can find a discrete distribution $\tilde{Q}$ with the same moments. We then choose f such that it attains its maximum at one of the discrete points, and if g approximates f sufficiently well, we show that the discrete distribution places too much weight on the maximum.

Theorem 28 (Lower bound for OPT relaxation)**.** *Let $\varphi : \mathcal{X} \to \mathbb{C}^N$ be continuous. Let*

$$n \coloneqq \dim_{\mathbb{C}} \mathcal{V}_{\mathrm{lin}}, \qquad \mathcal{V}_{\mathrm{lin}} \coloneqq \mathrm{Span}_{\mathbb{C}} \left\{ \varphi(x)\varphi(x)^* \mid x \in \mathcal{X} \right\} \subseteq \mathbb{C}^{N \times N} \, .$$

In other words, n is the number of effective degrees-of-freedom of the model $g(x) = \varphi(x)^* H \varphi(x)$, and hence corresponds to the maximum number of points where such a model can interpolate arbitrary function values. Then, there exists a point $z \in \mathcal{X}$ depending only on φ, such that the periodic and analytic function

$$f : \mathcal{X} \to \mathbb{R}, x \mapsto \sum_{i=1}^{d} \cos(2\pi(x_i - z_i))$$

satisfies

$$|L_g^{\mathrm{OPT}}(\mathcal{U}([0,1])) - L_{\beta f}(\mathcal{U}([0,1]))| \geq \log\left(\frac{\beta^{d/2}}{2n+1}\right) - \|g - \beta f\|_\infty \tag{7}$$

for any model $g(x) = \varphi(x)^* H \varphi(x)$ and any $\beta > 0$.

What are the implications of Theorem 28, which is proven in Appendix E.3, on convergence rates? To answer this question, we need to consider the limit $n \to \infty$, which means that N, φ, f, g in general depend on n, and we will denote them by N_n, φ_n, f_n, g_n, respectively. We also consider an inverse temperature $\beta_n \coloneqq (e(2n+1))^{2/d}$. Since f_n is analytic, an approximation method with optimal rate should achieve the rate $\|g_n - \beta_n f_n\|_\infty \leq O_{m,d}(\|\beta_n f_n\|_{C^m} n^{-m/d}) = O_{m,d}(n^{-(m-2)/d})$ for every $m \in \mathbb{N}$. Suppose that this is at least achieved for $m = 3$, such that $\lim_{n\to\infty} \|g_n - \beta_n f_n\|_\infty = 0$. Then,

$$\begin{aligned} |L_{g_n}^{\mathrm{OPT}}(\mathcal{U}([0,1])) - L_{\beta_n f_n}(\mathcal{U}([0,1]))| &\geq \log\left(\frac{\beta_n^{d/2}}{2n+1}\right) - \|g_n - \beta_n f_n\|_\infty \geq 1 - O_{m,d}(n^{-(m-2)/d}) \\ &= \Omega_{m,d}(\beta_n n^{-2/d}) \text{ for sufficiently large } n. \end{aligned}$$

In other words, the approximation error and the log-partition error of the OPT relaxation in Eq. (6) cannot both achieve a rate strictly better than $O_{m,d}(\|f\| n^{-2/d})$ even for infinitely smooth functions, no matter which (continuous) feature map φ is chosen.

5 Experiments

To further investigate the convergence behavior of some simple algorithms, we study them numerically on functions of the form $f : [0,1]^3 \to \mathbb{R}, x \mapsto \beta(x_1 + x_2 + x_3)$. While these functions are simple (and concave), they pose a challenge to some general algorithms as they have a large range in relation to their Lipschitz constant. The dimension $d = 3$ has been chosen for visualization purposes, to be able to distinguish the convergence rates $n^{-1/d}$ and $n^{-2/d}$ from the typical MC convergence rate of $n^{-1/2}$. Our plots can be reproduced using the code at

`github.com/dholzmueller/sampling_experiments`

5.1 Log-partition Estimation

For the log-partition problem, we consider the following algorithms:

- **PC**: Compute the log-partition function of a piecewise constant approximation as in Section 4.1.1.
- **MC**: Monte carlo log-partition estimation as in Section 4.2.2.
- **PC+MC**: We use importance sampling, specifically MC quadrature on top of a piecewise constant approximation as described in Section 2.2: We use $n/2$ function evaluations to compute a piecewise constant approximation g of f and then use the other $n/2$ function evaluations for an MC approximation of the right-hand side in

$$L_f = L_g + \log\left(\mathbb{E}_{x \sim P_g}\left[\exp(f(x) - g(x))\right]\right) \, .$$

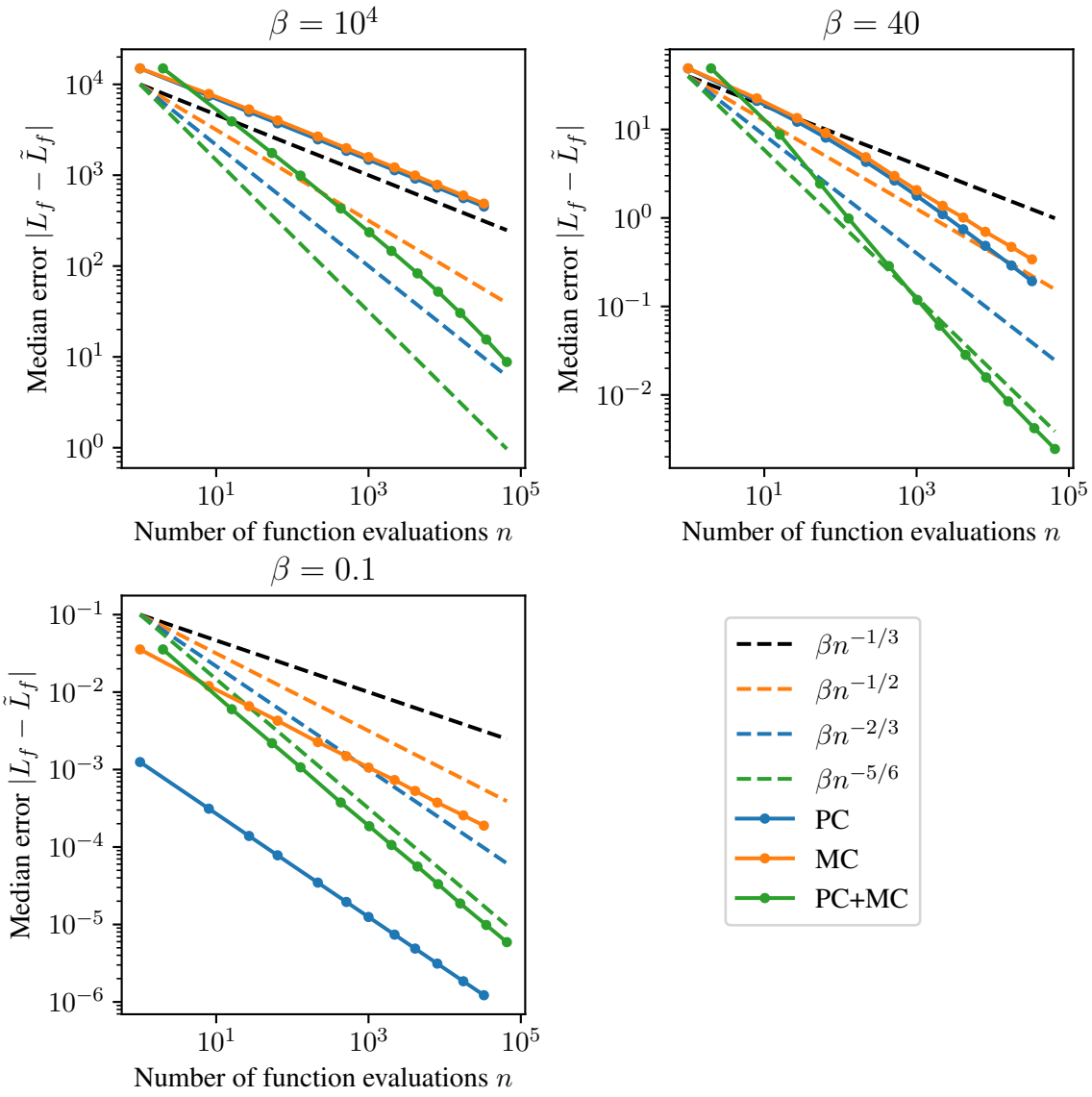

Figure 2: Convergence of the (median) error $|L_f - \tilde{L}_f|$ for different values of $\beta \in \{0.1, 40, 10000\}$. For the stochastic methods MC and PC+MC, the median is taken over 10001 independent runs.

Note that PC and MC run in linear time $O_{m,d}(n)$ while PC+MC can require a runtime of $O_{m,d}(n^2)$ or $O_{m,d}(n \log n)$ depending on the implementation.

Figure 2 shows the convergence of these methods for $\beta \in \{0.1, 40, 10000\}$. For $\beta = 10000$, the methods are in an optimization regime, where PC and MC follow the rate $O(n^{-1/3})$ of the corresponding upper bounds in Theorem 21 and Theorem 25. Meanwhile, PC+MC follows the rate $O(n^{-2/3})$. This can be understood intuitively by noting that due to the linear nature of f, the PC proposal distribution will mostly propose points close to the optimum, such that the MC component can get much closer to the optimum than with a uniform proposal distribution.

For $\beta = 30$, we observe a transition between an optimization regime and a quadrature regime. In the quadrature regime, the convergence rate of MC is the classical MC quadrature rate $O(n^{-1/2})$, matching the upper bound in Theorem 25. Meanwhile, the convergence rate of PC transitions to $O(n^{-2/3})$, matching the worst-case bound in Theorem 21, whose proof uses a linear f for the lower bound. The combination PC+MC approaches a convergence rate around $O(n^{-5/6})$. This can be understood as the MC rate $O(n^{-1/2})$ combined with the *approximation* rate (not log-partition rate) of PC,

which is $O(n^{-1/3})$. The rate $O(n^{-5/6})$ can be proven formally using arguments analogous to the proof of Theorem 9 in Appendix C.2.

For $\beta = 0.1$, we see the same quadrature regime rates as for $\beta = 30$, except that now the constant in the rate for PC is smaller than those of MC and PC+MC. This can be explained by an observation in the proof of Theorem 21 in Appendix E.1: Since PC performs midpoint quadrature, its error depends on the curvature of $\exp(f)$. Since f is linear, the curvature of $\exp(f)$ is significantly smaller than the worst-case curvature when $\beta \ll 1$. On the other hand, the convergence rate of PC+MC depends on the *approximation* rate of PC, which does not exhibit this phenomenon.

5.2 Sampling

To study convergence rates for sampling, we need a way to estimate distances between probability distributions through samples. While this can be achieved for the Wasserstein distance, and more efficiently for the related Sinkhorn distances, an even more efficient and easy-to-compute measure is the energy distance (see e.g. Székely and Rizzo, 2013) given by

$$D_{\text{energy}}(P,Q)^2 = 2\mathbb{E}_{x\sim P,x'\sim Q}\|x-x'\|_2 - \mathbb{E}_{x\sim P,x'\sim P}\|x-x'\|_2 - \mathbb{E}_{x\sim Q,x'\sim Q}\|x-x'\|_2 \,.$$

We estimate the energy distance $D_{\text{energy}}(P_f, \tilde{P}_f)$ by sampling a finite number of samples $x_1,\dots,x_N \sim P_f$ and $\tilde{x}_1,\dots,\tilde{x}_N \sim \tilde{P}_f$) and then computing the energy distance $D_{\text{energy}}(Q,\tilde{Q})$ of the empirical distributions

$$Q \coloneqq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \qquad \tilde{Q} \coloneqq \frac{1}{N}\sum_{i=1}^{N}\delta_{\tilde{x}_i}\,,$$

where δ_x is the Dirac distribution at x. We compare the following sampling algorithms:

- **PC**: Sampling from a piecewise constant approximation as in Section 4.1.1.
- **MC**: Monte carlo sampling as defined in Section 4.2.3.
- **RS**: We return RejectionSampling(f, M_f, n) as defined in Algorithm 1 and investigated in Section 4.2.1. Here, we know M_f explicitly due to the simple nature of f.
- **PC+MC**: Performing MC sampling on top of a piecewise constant proposal distribution: We compute a piecewise constant approximant g of f with $n/2$ points, then draw samples $X_1,\dots,X_{n/2} \sim P_g$ and output X_I, where
$$P(I=i) = \frac{\exp(f(X_i)-g(X_i))}{\sum_{j=1}^{n/2}\exp(f(X_j)-g(X_j))}\,.$$
- **PC+RS**: We use $n/2$ points to compute a piecewise constant approximation g of f and then return RejectionSampling$(f, g+M_{f-g}, n/2)$ as defined in Algorithm 1. Here, we know M_{f-g} explicitly due to the simple nature of f.

Note that PC, MC, and RS run in linear time $O_{m,d}(n)$ while PC+MC and PC+RS can require $O_{m,d}(n^2)$ or $O_{m,d}(n\log n)$ depending on the implementation.

For the sampling algorithms in Figure 3, the behavior in terms of convergence rates is less clear than for the log-partition algorithms. Nonetheless, it is evident that combining approximation-based and stochastic methods performs better than either of the two in isolation. Moreover, it is noticeable that RS, our budget-limited variant of rejection sampling, initially performs poorly while reaching fast convergence for larger values of n, when the probability of overall rejection becomes small.

6 Conclusion

In this paper, we studied the convergence rates of sampling and log-partition estimation methods on classes of m-smooth functions on the d-dimensional unit cube $\mathcal{X} = [0,1]^d$. In Section 2, we showed that without computational constraints, the optimal achievable convergence rates are of the form $O_{m,d}(Bn^{-m/d})$ or even better depending on the setting. We then investigated several computational reductions between problems in Section 3, showing that several problems are similarly hard. In Section 4, we studied convergence rates of specific algorithms, which are however far from being optimal unless one is willing to spend a computational effort on the order of $O(n^m)$, i.e., exponential in the smoothness m for which the optimal rate should be achieved. Our experimental study nonetheless confirms practical differences between the convergence rates of some of the investigated efficient algorithms.

Our work poses the central question of whether near-optimal convergence rates for smooth functions can be achieved with runtimes that are of fixed polynomial order $O_{m,d}(n^k)$, i.e., where k does not depend on m or d. Moreover,

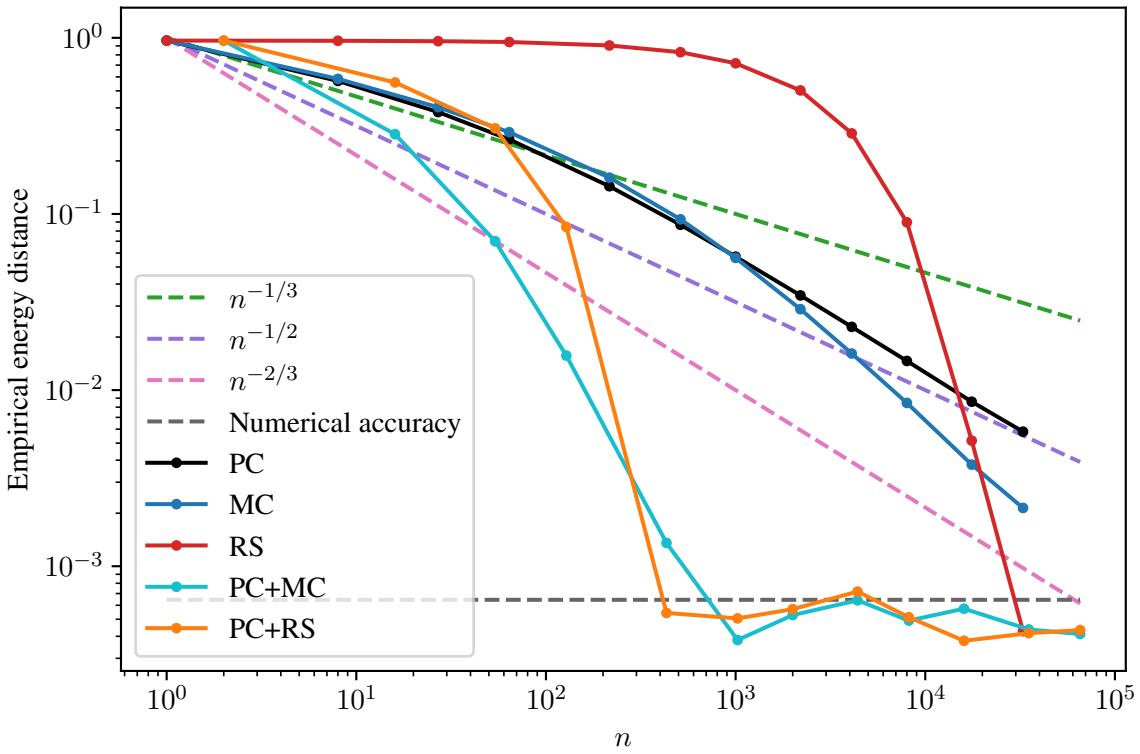

Figure 3: Convergence of different sampling methods in terms of the empirical energy distance, computed using $N = 10^6$ samples for each distribution, to the true distribution P_f for $\beta = 15$. Here, n denotes the number of function evaluations used for drawing a sample, where PC uses the same function evaluations for each sample while MC and RS need new function evaluations for every drawn sample. The gray dashed line corresponds to the maximum empirical energy distance of two sets of $N = 10^6$ samples both drawn from P_f, where the maximum is taken over three random draws.

for many sampling algorithms, it is unclear which convergence rates they can achieve in our setting. For example, variants of parallel tempering are often employed for non-log-concave problems, and diffusion models might prove to be relevant if the score function can be approximated efficiently (Chen et al., 2022). An analysis of (mixtures of) Laplace approximations might also be interesting in this context. Beyond specific algorithms, proving lower bounds outside of the optimization regime is still an open question except for some special cases (Chewi et al., 2022b), and other probability distance measures such as the KL divergence could be considered as well.

Acknowledgements We thank Alessandro Rudi, Viktor Zaverkin, Hans Kersting, Ingo Steinwart, Davoud Mirzaei, and Eric Moulines for helpful discussions. Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting David Holzmüller.

References

Juliette Achddou, Joseph Lam-Weil, Alexandra Carpentier, and Gilles Blanchard. A minimax near-optimal algorithm for adaptive rejection sampling. In *Algorithmic Learning Theory*, 2019.

Jason M. Altschuler and Kunal Talwar. Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling. *arXiv:2210.08448*, 2022.

Herbert Amann and Joachim Escher. *Analysis*. Springer, 2005.

Francis Bach. Sum-of-squares relaxations for information theory and variational inference. *arXiv:2206.13285*, 2022.

Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2023.

Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, 2022.

Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian monte carlo. *The Annals of Applied Probability*, 30(3):1209–1250, 2020.

Susanne C. Brenner, L. Ridgway Scott, and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 3. Springer, 2008.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Xiang Cheng. *The Interplay between Sampling and Optimization*. PhD thesis, University of California, Berkeley, 2020.

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv:1805.01648*, 2018.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory*, 2021.

Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. *arXiv:2210.02482*, 2022a.

Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. In *Conference on Learning Theory*, 2022b.

Nicolas Chopin and Mathieu Gerber. Higher-order stochastic integration through cubic stratification. *arXiv:2210.01554*, 2022.

Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

Monroe D. Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, 2018.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Symposium on Foundations of Computer Science*, 2010.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier, 2001.

Nial Friel and Jason Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

Jean Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi diagrams and Delaunay triangulations. *arXiv:0805.0292*, 2008.

Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: Algorithms and lower bounds. In *Symposium on Theory of Computing*, 2020.

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

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Chii-Ruey Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980. Publisher: JSTOR.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Scott Kirkpatrick, C. Daniel Gelatt Jr, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598): 671–680, 1983.

John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

Peter Lancaster and Kes Salkauskas. Surfaces generated by moving least squares methods. *Mathematics of Computation*, 37(155):141–158, 1981.

Pierre Simon Laplace. Mémoire sur la probabilité de causes par les évènements. *Mémoires de Mathématique et de Physique, Presentés à l'Académie Royale des Sciences, par divers Savants & lus dans ses Assemblées. Tome Sixième*, pages 621–656, 1774.

Xiaolin Li. Error estimates for the moving least-square approximation and the element-free Galerkin method in n-dimensional spaces. *Applied Numerical Mathematics*, 99:77–97, 2016.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Neural Information Processing Systems*, volume 31, 2018.

Oren Mangoubi and Nisheeth K. Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, 2019.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Sampling from arbitrary functions via psd models. In *Artificial Intelligence and Statistics*, 2022.

Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44 (247):335–341, 1949.

Davoud Mirzaei. Analysis of moving least squares approximation revisited. *Journal of Computational and Applied Mathematics*, 282:237–250, 2015.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer, 1988.

Erich Novak and Henryk Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *Journal of Complexity*, 25(4):398–404, 2009.

Christian P. Robert. *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*. Springer, 2007.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *arXiv:2012.11978*, 2020.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Kunal Talwar. Computational separations between sampling and optimization. In *Neural Information Processing Systems*, volume 32, 2019.

Joseph F. Traub. Information-based complexity. In *Encyclopedia of Computer Science*, pages 850–854. John Wiley and Sons Ltd., GBR, 2003. ISBN 0-470-86412-5.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.

Dawn Woodard, Scott Schmidler, and Mark Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804, 2009.

Blake Woodworth, Francis Bach, and Alessandro Rudi. Non-convex optimization with certificates and fast rates through kernel sums of squares. In *Conference on Learning Theory*, 2022.

Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, 2021.

A Overview

The structure of the appendix matches that of the main part of this paper. We give proofs for Section 1 in Appendix B, for Section 2 in Appendix C, for Section 3 in Appendix D, and for Section 4 in Appendix E.

B Proofs for Introduction

Lemma 3 (Optimization limit)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz-continuous with Lipschitz constant $|f|_1 < \infty$. Then, for any temperature $\varepsilon > 0$, the maximum $M_f = \max_{x\in\mathcal{X}} f(x)$ satisfies*

$$|M_f - \varepsilon L_{f/\varepsilon}| \le \varepsilon d \log(1 + 3d^{-1/2}|f|_1/\varepsilon) \longrightarrow 0 \quad \text{for } \varepsilon \searrow 0 \,. \tag{1}$$

Moreover, for any bounded and measurable $f : \mathcal{X} \to \mathbb{R}$ and any $\delta \in (0, 1]$, we have

$$P_{f/\varepsilon}(\{x \in \mathcal{X} \mid f(x) < \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)\}) \le \delta \,.$$

Proof. For the first part of the lemma, let $x^* \in \mathcal{X}$ be a maximizer of f. Without loss of generality, assume that f is shifted such that $f(x^*) = 0$.

Step 1: Upper bound. We have

$$L_f = \log \int_{\mathcal{X}} e^{f(x)} \,\mathrm{d}x \le \log \int_{\mathcal{X}} 1 \,\mathrm{d}x = \log 1 = 0 = M_f \,.$$

Step 2: Lower bound. To show a lower bound on L_f, define the side length $R \coloneqq (\max\{1, d^{-1/2}|f|_1\})^{-1}$. Since $R \le 1$, $\mathcal{X}$ contains an axis-aligned subcube $\tilde{\mathcal{X}}$ of side length R containing x^*. Each point $x \in \tilde{\mathcal{X}}$ has distance at most $\sqrt{d}R$ from x^*, and hence by Lipschitzness, we have

$$f(x) \ge f(0) - |f|_1\sqrt{d}R = -|f|_1\sqrt{d}R \,.$$

We consider two cases:

(a) **Case 1: $d^{-1/2}|f|_1 \le 1$.** In this case, we have $R = 1$ and hence

$$L_f = \log \int_{\mathcal{X}} e^{f(x)} \,\mathrm{d}x \ge \log \int_{\mathcal{X}} e^{-|f|_1\sqrt{d}} \,\mathrm{d}x = -|f|_1\sqrt{d} = -d(d^{-1/2}|f|_1) \,.$$

The function $h(x) \coloneqq \log(1 + 3x) - x$ is concave and $h(0), h(1) \ge 0$, which shows $h(x) \ge 0$ for $x \in [0, 1]$. Hence,

$$L_f \ge -d \log(1 + 3d^{-1/2}|f|_1) \,.$$

(b) **Case 2: $d^{-1/2}|f|_1 > 1$.** In this case, we have $R = (d^{-1/2}|f|_1)^{-1}$ and hence $f(x) \ge -d$ for $x \in \tilde{\mathcal{X}}$. This yields

$$\begin{aligned} L_f &= \log \int_{\mathcal{X}} e^{f(x)} \,\mathrm{d}x \ge \log \int_{\tilde{\mathcal{X}}} e^{f(x)} \,\mathrm{d}x \\ &\ge \log \int_{\tilde{\mathcal{X}}} e^{-d} \,\mathrm{d}x = -d + d\log(R) = -d - d\log(d^{-1/2}|f|_1) \\ &= -d \log(ed^{-1/2}|f|_1) \ge -d \log(1 + 3d^{-1/2}|f|_1) \,. \end{aligned}$$

Step 3: Including the temperature. By replacing f with f/ε, we obtain

$$|M_f - \varepsilon L_{f/\varepsilon}| = |\varepsilon M_{f/\varepsilon} - \varepsilon L_{f/\varepsilon}| = \varepsilon |M_{f/\varepsilon} - L_{f/\varepsilon}| \le \varepsilon d \log(1 + 3d^{-1/2}|f|_1/\varepsilon) \,.$$

Step 4: Probabilistic bound. We have

$$\begin{aligned} P_{f/\varepsilon}(x : f(x) \le \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)) &\le \int_{\{x\in\mathcal{X}: f(x) \le \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)\}} \exp(f(x)/\varepsilon - L_{f/\varepsilon}) \,\mathrm{d}x \\ &\le \int_{\mathcal{X}} \exp(-\log(1/\delta)) \,\mathrm{d}x = \delta \,. \end{aligned}$$

□

The following lemma will be useful to deal with Lipschitz constants:

Lemma B.1. *Let $f \in C^m(\mathcal{X}), m \ge 1$. Then, $|f|_1 \le d^{1/2}\|f\|_{C^m}$.*

Proof. We have

$$|f|_1 = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|_2 \le \sup_{x\in\mathcal{X}} d^{1/2}\|\nabla f(x)\|_\infty \le d^{1/2}\|f\|_{C^1} \le d^{1/2}\|f\|_{C^m} \,.$$

□

C Proofs for Information-based Complexity

Most of our lower bounds rely on the common strategy of hiding smooth functions with small support somewhere in the domain (see e.g. Novak, 1988). We will consider the following bump functions:

Definition C.1 (Bump functions)**.** We define the template one-dimensional bump function

$$\tilde{b} : \mathbb{R} \to \mathbb{R}, x \mapsto \begin{cases} \exp(4 - (1-x)^{-1} - (x+1)^{-1}) & , \text{ if } x \in (-1,1) \\ 0 & , \text{ otherwise} \end{cases}$$

and, for given dimension d, the template multi-dimensional bump function

$$b : \mathbb{R}^d \to \mathbb{R}, x \mapsto \tilde{b}(x_1) \cdots \tilde{b}(x_d) \, .$$

for $z \in \mathbb{R}^d$ and $\delta > 0$, the shifted and scaled bump functions

$$b_{z,\delta} : \mathbb{R}^d \to \mathbb{R}, x \mapsto b\left(\frac{x-z}{\delta}\right) \, .$$

Moreover, we define the open cube

$$B_\infty(x, \delta) \coloneqq \{z \in \mathbb{R}^d \mid \|z - x\|_\infty < \delta\} \, . \qquad \blacktriangleleft$$

The following lemma illustrates some important properties of these bump functions:

Lemma C.2 (Bump functions)**.** *The bump functions $b_{z,\delta}$ from Definition C.1 satisfy*

(a) $b_{z,\delta}$ is zero outside of $B_\infty(z, \delta)$.
(b) $b_{z,\delta}$ is infinitely often continuously differentiable and all of its derivatives are bounded,
(c) there exists a constant $C_{m,d} > 0$ independent of z and δ such that for all $z \in \mathbb{R}^d$ and $\delta > 0$,

$$\|b_{z,\delta}\|_{C^m(\mathbb{R}^d)} \leq C_{m,d} \max\{1, \delta^{-m}\} \, .$$

(d) For $x \in B_\infty(z, \delta/2)$, we have $b_{z,\delta}(x) \geq 1$.

Proof.

(a) This is easy to verify from the definition.

(b) It is well-known, see e.g. Remark 3.4 (d) in Chapter V.3 of Amann and Escher (2005), that the function

$$\hat{b} : \mathbb{R} \to \mathbb{R}, x \mapsto \begin{cases} 0 & , x \leq 0 \\ \exp(-1/x) & , x > 0 \end{cases}$$

is C^∞. Since $\tilde{b}(x) = e^4 \hat{b}(1-x) \cdot \hat{b}(x+1)$, $\tilde{b}$ is also C^∞, and so must be b and $b_{z,\delta}$. Moreover, since $b_{z,\delta}$ has compact support, all the derivatives are bounded.

(c) Let $C_{m,d} \coloneqq \|b\|_{C^m(\mathbb{R}^d)}$. We have

$$\sup_{x \in \mathbb{R}^d} \left| \frac{\partial^\alpha b_{z,\delta}(x)}{\partial x^\alpha} \right| = \delta^{-|\alpha|_1} \sup_{x \in \mathbb{R}^d} \left| \frac{\partial^\alpha b(x)}{\partial x^\alpha} \right| \, .$$

Therefore, by definition of the C^m-norm, we have

$$\|b_{z,\delta}\|_{C^m(\mathbb{R}^d)} \leq C_{m,d} \max\{1, \delta^{-m}\} \, .$$

(d) It is easy to verify that $\tilde{b}(x) \geq 1$ for $|x| \leq 1/2$. For $x \in B_\infty(z, \delta/2)$, we have

$$\|(x-z)/\delta\|_\infty \leq 1/2 \, ,$$

hence

$$b_{z,\delta}(x) = b\left(\frac{x-z}{\delta}\right) \geq 1 \cdots 1 = 1 \, . \qquad \square$$

The following lemma is useful to bound the number of bump functions that we can hide in a domain:

Lemma C.3. *For $k \in \mathbb{N}_{\geq 1}$, a third-slice $\tilde{\mathcal{X}} \coloneqq [0, 1/3] \times [0,1]^{d-1}$ of the cube $\mathcal{X}$ contains at least k disjoint open cubes $B_\infty(z_1, r_k), \ldots, B_\infty(z_k, r_k)$ with radius*

$$r_k = \frac{k^{-1/d}}{12} \, .$$

Proof. Choose $N \coloneqq \lceil k^{1/d} \rceil$. We can divide $\tilde{\mathcal{X}}$ into $N \cdot (3N)^{d-1} \geq N^d \geq k$ cubes of side length $(3N)^{-1}$ and radius

$$r = \frac{1}{6N} \geq \frac{1}{6(k^{1/d}+1)} \geq \frac{1}{12 k^{1/d}} = \frac{k^{-1/d}}{12} = r_k \, . \qquad \square$$

C.1 Deterministic Evaluation Points

We first adapt some results from Novak (1988) to our setting.

Theorem 5 (adapted from Novak (1988))**.** *We have*

$$\begin{aligned} e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{opt}^*}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{opt}^*}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), \\ e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{int}}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{int}}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}) \,. \end{aligned}$$

Proof. **Step 1: Upper bounds.** For $S \in \{S_{\mathrm{app}}, S_{\mathrm{opt}^*}, S_{\mathrm{int}}\}$, Novak (1988) states upper bounds of the form $O_{m,d}(n^{-m/d})$ for bounded classes of functions in the Sobolev space $W_\infty^{m,d}$, which contain $\mathcal{F}_{d,m,B_{m,d}}$ for some $B_{m,d} > 0$ (see Section 1.3.11 and 1.3.12 in Novak (1988)). Hence, for the corresponding metric D, we have

$$e_n(\mathcal{F}_{d,m,B_{m,d}}, S, D) \le O_{m,d}(n^{-m/d}) \,.$$

For another value of B, we can then take a near-optimal $\tilde{S} \in \mathcal{A}_n$ for $\mathcal{F}_{d,m,B_{m,d}}$ and define

$$\hat{\tilde{S}}(f) \coloneqq \frac{B}{B_{m,d}} \tilde{S}\left(\frac{B_{m,d}}{B} f\right) \,,$$

and by positive homogeneity of S and D, this then achieves the rate $O_{m,d}(Bn^{-m/d})$.

Step 2: Lower bounds. For lower bounds, it is again sufficient to consider $\mathcal{F}_{d,m,B_{m,d}}$ for a single $B_{m,d} > 0$. Novak (1988) uses bump functions created by rescaling and shifting the template bump function

$$\Phi(x) = \begin{cases} a \prod_{i=1}^d (1 - x_i^2)^m & , x \in [-1,1]^d \\ 0 & , \text{ otherwise} \end{cases}$$

for some appropriate constant $a > 0$. This function is in $W_\infty^{m,d}$ but not all of its weak m-th derivatives are continuous. Hence, the constructed counterexamples do not directly apply to $\mathcal{F}_{d,m,B_{m,d}}$. However, it is possible to replace Φ by the C^∞ bump function b from Definition C.1 since the norms of the derivatives behave in the same fashion for scaled and shifted versions of b, as shown in Lemma C.2. Hence, the same lower bounds still apply to $\mathcal{F}_{d,m,B_{m,d}}$. □

We can now turn to our upper bounds through approximation:

Proposition 6 (Upper bounds via approximation)**.** *For bounded and measurable $f, g : \mathcal{X} \to \mathbb{R}$, we have*

(a) $|L_f - L_g| \le \|f - g\|_\infty$.
(b) $d^{-1/2} W_1(P_f, P_g) \le D_{\mathrm{TV}}(P_f, P_g) \le D_{\text{sup-log}}(P_f, P_g) \le 2\|f - g\|_\infty$.

Proof. Since L_f and P_f are not influenced by changing f on null sets, we will ignore exceptional null sets in the essential supremum in the definition of $\|\cdot\|_\infty$ in the following.

(a) We have

$$\begin{aligned} L_g &= \log \int_{\mathcal{X}} e^{g(x)} \,\mathrm{d}x \\ &\le \log \int_{\mathcal{X}} e^{f(x) + \|f-g\|_\infty} \,\mathrm{d}x = \log\left(e^{\|f-g\|_\infty} \cdot \int_{\mathcal{X}} e^{f(x)} \,\mathrm{d}x\right) = \left(\log \int_{\mathcal{X}} e^{f(x)} \,\mathrm{d}x\right) + \|f-g\|_\infty \\ &= L_f + \|f-g\|_\infty \,, \end{aligned}$$

and the other inequality follows analogously.

(b) We have $p_f(x) = \exp(f(x) - L_f)$ and $p_g(x) = \exp(g(x) - L_g)$, hence

$$\begin{aligned} D_{\text{sup-log}}(P_f, P_g) &= \left\| \log\left(\frac{p_f}{p_g}\right) \right\|_\infty = \|(f - L_f) - (g - L_g)\|_\infty \le \|f-g\|_\infty + |L_f - L_g| \\ &\overset{(a)}{\le} 2\|f-g\|_\infty \,. \end{aligned}$$

Let $\bar{f} \coloneqq f - L_f$. By a well-known property of the TV distance (see e.g. Gibbs and Su, 2002),

$$D_{\mathrm{TV}}(P_f, P_g) = D_{\mathrm{TV}}(P_{\bar{f}}, P_{\bar{g}}) = \frac{1}{2} \int_{\mathcal{X}} |e^{\bar{f}(x)} - e^{\bar{g}(x)}| \,\mathrm{d}x \,.$$

Now, consider a fixed $x \in \mathcal{X}$. Without loss of generality, assume $\bar{f}(x) \leq \bar{g}(x)$. Then,

$$e^{\bar{g}(x)-\|\bar{f}-\bar{g}\|_\infty} \leq e^{\bar{f}(x)} \leq e^{\bar{g}(x)} ,$$

which yields

$$|e^{\bar{f}(x)} - e^{\bar{g}(x)}| \leq (1 - e^{-\|\bar{f}-\bar{g}\|_\infty})e^{\bar{g}(x)} \leq (1 - e^{-\|\bar{f}-\bar{g}\|_\infty})(e^{\bar{f}(x)} + e^{\bar{g}(x)}) \leq \|\bar{f} - \bar{g}\|_\infty(e^{\bar{f}(x)} + e^{\bar{g}(x)}) .$$

Therefore,

$$D_{\mathrm{TV}}(P_f, P_g) \leq \frac{1}{2}\int_{\mathcal{X}} \|\bar{f} - \bar{g}\|_\infty(e^{\bar{f}(x)} + e^{\bar{g}(x)})\,\mathrm{d}\mu(x) = \|\bar{f} - \bar{g}\|_\infty = D_{\text{sup-log}}(P_f, P_g) .$$

The bound $W_1(P_f, P_g) \leq \mathrm{diam}(\mathcal{X})D_{\mathrm{TV}}(P_f, P_g) = d^{1/2}D_{\mathrm{TV}}(P_f, P_g)$ for the 1-Wasserstein distance, where $\mathrm{diam}(\mathcal{X})$ is the diameter of $\mathcal{X}$, is well-known (see e.g. Gibbs and Su, 2002). □

The following technical lemmas will be used for the lower bound afterward.

Lemma C.4. *Let $a, b > 0$. Then,*

$$\frac{a}{a+b} \geq \frac{1}{2}\min\left\{1, \frac{a}{b}\right\} .$$

Proof. If $a \leq b$, we have

$$\frac{a}{a+b} \geq \frac{a}{2b} \geq \frac{1}{2}\min\left\{1, \frac{a}{b}\right\} .$$

Similarly, if $a \geq b$, we have

$$\frac{a}{a+b} \geq \frac{a}{2a} = \frac{1}{2} \geq \frac{1}{2}\min\left\{1, \frac{a}{b}\right\} .$$ □

Lemma C.5. *Let $c \in (0, 1]$. Then, the function*

$$h : [0, \infty) \to \mathbb{R}, x \mapsto \log(1 + c(e^x - 1))$$

satisfies $h(x) \geq cx$ for all $x \geq 0$.

Proof. For all $x \geq 0$, we have

$$h'(x) = \frac{ce^x}{1 + c(e^x - 1)} = \frac{c}{c + (1 - c)e^{-x}} \geq \frac{c}{c + (1 - c)} = c .$$

Therefore,

$$h(x) = h(0) + \int_0^x h(u)\,\mathrm{d}u \geq \int_0^x c\,\mathrm{d}u = cx .$$ □

Now, we are ready to prove the exact minimax optimal rates. The main technical difficulty is that for the lower bound in the 1-Wasserstein distance, we need to hide many bumps that are far apart, and we need to bound the resulting Wasserstein distance.

Theorem 7 (Information-based complexity of sampling and log-partition with deterministic evaluation points)**.** *We have*

$$\begin{aligned}
e_n(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}),\\
e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, D_{\text{sup-log}}) &= \Theta_{m,d}(Bn^{-m/d}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, D_{\text{sup-log}}) &= \Theta_{m,d}(Bn^{-m/d}),\\
e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, D_{\mathrm{TV}}) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, D_{\mathrm{TV}}) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}),\\
e_n(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, W_1) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}), & e_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{samp}}, W_1) &= \Theta_{m,d}(\min\{1, Bn^{-m/d}\}).
\end{aligned}$$

Proof. **Step 0: Upper bounds.** We know from Theorem 5 that the rate $O_{m,d}(Bn^{-m/d})$ can be achieved for approximation with non-adaptive deterministic evaluation points, and we know from Proposition 6 that this rate can therefore also be achieved for the log-partition problem and the sampling problem with $D_{\text{sup-log}}$, D_{TV}, and W_1. Moreover, since $D_{\mathrm{TV}}(P, Q) \leq 1$ for all distributions P, Q, we obtain an upper bound of $O_{m,d}(\max\{1, Bn^{-m/d}\})$ for D_{TV}. Similarly,

since $\mathcal{X}$ has diameter $d^{1/2}$, W_1 is upper bounded by $d^{1/2} = O_{m,d}(1)$, and hence we also obtain an upper bound of $O_{m,d}(\max\{1, Bn^{-m/d}\})$ for W_1. The upper bounds also hold for the adaptive setting since it is more permissive.

In the following, we will derive matching asymptotic lower bounds for the adaptive setting, which then also hold for the non-adaptive setting. To this end, let $\tilde{S} \in \mathcal{A}_n^{\mathrm{ad}}$ for the log-partition or sampling problem on the function class $\mathcal{F}_{d,m,B}$.

Step 1: Defining grids in the cube. We can cut the cube $\mathcal{X}$ along one axis into three equally shaped slices $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2$:

$$\mathcal{C}_k \coloneqq [k/3, (k+1)/3] \times [0,1]^{d-1}, \qquad k \in \{0,1,2\} .$$

Then, by Lemma C.3, we can find a finite set of points $\mathcal{G}_k \subseteq \mathcal{C}_k$ with $|\mathcal{G}_k| = 2n$ such that the open cubes $B_\infty(x, \delta_n)$ for $x \in \mathcal{G}_k$ and radius

$$\delta_n = \frac{(2n)^{-1/d}}{12} \geq \frac{n^{-1/d}}{24}$$

are contained in $\mathcal{C}_k$ and disjoint.

Step 2: Removing points close to queried points. Let $\mathcal{X}_n$ denote the $\leq n$ points where $\tilde{S}$ queries the zero function. For fixed $k \in \{0,2\}$, the $2n$ cubes $(B_\infty(x, \delta_n))_{x \in \mathcal{G}_k}$ are disjoint. Hence there must be a subset $\tilde{\mathcal{G}}_k \subseteq \mathcal{G}_k$ containing n points whose corresponding cubes do not contain any point from $\mathcal{X}_n$.

Step 3: Two different functions. Now, for $k \in \{0,2\}$ and $C_{m,d}$ as in Lemma C.2, define the functions

$$f_k(x) \coloneqq BC_{m,d}^{-1} \delta_n^m \sum_{z \in \tilde{\mathcal{G}}_k} b_{z,\delta_n}(x) .$$

We have $\delta_n \leq 1$ and hence $\|b_{z,\delta_n}\|_{C^m} \leq C_{m,d} \delta_n^{-m}$ by Lemma C.2. Because the support of the bump functions does not overlap, we have $\|f_k\|_{C^m} \leq B$ by Lemma C.2 and hence $f_k \in \mathcal{F}_{d,m,B}$. By the construction of $\tilde{\mathcal{G}}_k$, f_0 and f_2 are zero on $\mathcal{X}_n$. Hence, even an adaptive $\tilde{S}$ must also query f_k at the points in $\mathcal{X}_n$, and since both are equal at those points, we must have

$$\tilde{S}(f_0) = \tilde{S}(f_2) .$$

Step 4: Wasserstein distance of both functions. Because f_0 and f_2 use the same number of equally wide bump functions whose support is fully contained in $\mathcal{X}$, we have

$$L_{f_0} = L_{f_2} . \tag{8}$$

To lower-bound the 1-Wasserstein distance, we use its dual formulation and choose the 1-Lipschitz function $\varphi(x) \coloneqq x_1 - 1/3$. This yields

$$\begin{aligned} W_1(P_{f_0}, P_{f_2}) &\geq \mathbb{E}_{x \sim P_{f_2}} \varphi(x) - \mathbb{E}_{x \sim P_{f_0}} \varphi(x) = \int_{\mathcal{X}} \varphi(x)(e^{\overline{f}_2(x)} - e^{\overline{f}_0(x)}) \,\mathrm{d}x \\ &\overset{\text{Eq. (8)}}{=} e^{-L_{f_0}} \int_{\mathcal{X}} \varphi(x)(e^{f_2(x)} - e^{f_0(x)}) \,\mathrm{d}x . \end{aligned} \tag{9}$$

Step 5: Lower-bounding the normalization constant. We first define the "bump integral"

$$I_n \coloneqq \int_{B_\infty(z,\delta_n)} (e^{BC_{m,d}^{-1}\delta_n^m b_{z,\delta_n}(x)} - 1) \,\mathrm{d}x ,$$

which is independent of z. Then, we have

$$\begin{aligned} I_n &\overset{\text{Lemma C.2}}{\geq} \int_{B_\infty(z,\delta_n/2)} (e^{BC_{m,d}^{-1}\delta_n^m} - 1) \,\mathrm{d}x \\ &= \delta_n^d (e^{BC_{m,d}^{-1}\delta_n^m} - 1) . \end{aligned} \tag{10}$$

We then obtain

$$\begin{aligned} e^{L_{f_0}} &= \int_{\mathcal{X}} e^0 \,\mathrm{d}x + \sum_{z \in \tilde{\mathcal{G}}_0} \int_{B_\infty(z,\delta_n)} (e^{BC_{m,d}^{-1}\delta_n^m b_{z,\delta_n}(x)} - e^0) \,\mathrm{d}x \\ &= 1 + nI_n . \end{aligned} \tag{11}$$

Step 6: Lower-bounding the integral. By construction of the functions φ, f_0, and f_2, we know that

$$\int_{\mathcal{X}} \varphi(x)(e^{f_2(x)} - e^{f_0(x)})\,\mathrm{d}x \geq \int_{\mathcal{C}_2} \varphi(x)(e^{f_2(x)} - e^{f_0(x)})\,\mathrm{d}x\,. \tag{12}$$

Using $\varphi(x) \geq 1/3$ and $f_2(x) \geq f_0(x)$ for $x \in \mathcal{C}_2$, we can lower-bound the latter integral as

$$\begin{aligned}\int_{\mathcal{C}_2} \varphi(x)(e^{f_2(x)} - e^{f_0(x)})\,\mathrm{d}x &\geq \sum_{z \in \tilde{\mathcal{G}}_2} \int_{B_\infty(z,\delta_n)} \frac{1}{3}(e^{BC_{m,d}^{-1}\delta_n^m b_{z,\delta_n}(x)} - 1)\,\mathrm{d}x \\ &= \frac{1}{3} n I_n\,.\end{aligned}$$

Step 7: Wasserstein distance lower bound. By combining the previous lower bounds with Eq. (12), Eq. (11) and Eq. (9), we arrive at

$$W_1(P_{f_0}, P_{f_2}) \geq \frac{(1/3)nI_n}{1 + nI_n} = \frac{1}{3}\frac{I_n}{I_n + n^{-1}}\,.$$

We can then apply Lemma C.4 and Eq. (10) to obtain, for a suitable constant $c_{m,d} > 0$,

$$\begin{aligned}W_1(P_{f_0}, P_{f_2}) &\geq \frac{1}{6}\min\left\{1, \frac{I_n}{n^{-1}}\right\} \\ &\geq \frac{1}{6}\min\left\{1, \frac{\delta_n^d(e^{BC_{m,d}^{-1}\delta_n^m} - 1)}{n^{-1}}\right\} \\ &\geq \frac{1}{6}\min\left\{1, n\delta_n^d BC_{m,d}^{-1}\delta_n^m\right\} \\ &\geq \frac{1}{6}\min\left\{1, c_{m,d}Bn^{-m/d}\right\}\,.\end{aligned}$$

Step 8: Wasserstein minimax rate lower bound. Suppose that we are considering the sampling problem. As argued before, we have $\tilde{S}(f_0) = \tilde{S}(f_2)$. Hence, by an application of the triangle inequality, we must have $k \in \{0, 2\}$ such that

$$W_1(P_{f_k}, \tilde{S}(f_k)) \geq \frac{1}{12}\min\left\{1, c_{m,d}Bn^{-m/d}\right\}\,.$$

The Wasserstein minimax lower bound then follows by setting $f \coloneqq f_k$.

Step 9: TV distance minimax lower bound. Since $D_{\mathrm{TV}}(P_f, \tilde{S}(f)) \geq d^{-1/2}W_1(P_f, \tilde{S}(f))$ (see e.g. Gibbs and Su, 2002), we obtain the same asymptotic lower bound for the TV distance.

Step 10: Log-sup minimax lower bound. We have

$$D_{\text{sup-log}}(P_{f_0}, P_{f_2}) = \|(f_0 - L_{f_0}) - (f_2 - L_{f_2})\|_\infty = \|f_0 - f_2\|_\infty \geq \Omega_{m,d}(BC_{m,d}^{-1}\delta_n^m) = \Omega_{m,d}(Bn^{-m/d})\,.$$

Since $\tilde{S}(f_0) = \tilde{S}(f_2)$, by the triangle inequality, there must hence exist $k \in \{0, 2\}$ such that

$$D_{\text{sup-log}}(P_{f_k}, \tilde{S}(f_k)) \geq \Omega_{m,d}(Bn^{-m/d})\,.$$

Step 11: Log-partition minimax lower bound. Suppose that we instead consider the log-partition problem. Setting $c_d = 24^{-d}$, we obtain

$$\begin{aligned}L_{f_2} &= \log(1 + nI_n) \geq \log(1 + n\delta_n^d(e^{BC_{m,d}^{-1}\delta_n^m} - 1)) \\ &\geq \log\left(1 + c_d(e^{BC_{m,d}^{-1}\delta_n^m} - 1)\right) \\ &\overset{\text{Lemma C.5}}{\geq} c_d BC_{m,d}^{-1}\delta_n^m \\ &\geq \Omega_{m,d}(Bn^{-m/d})\,.\end{aligned}$$

Since $\tilde{S}$ cannot distinguish the zero function $f \equiv 0$ and f_2, we must have

$$\max\{|L_f - \tilde{S}(f)|, |L_{f_2} - \tilde{S}(f_2)|\} \geq \Omega_{m,d}(Bn^{-m/d})\,. \qquad \square$$

C.2 Stochastic Evaluation Points

Again, we first adapt some related results from Novak (1988) to our setting.

Theorem 8 (adapted from Novak (1988))**.** *We have*

$$\begin{aligned}
{}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{app}}, D_\infty) &= \Theta_{m,d}(Bn^{-m/d}),\\
{}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{opt}^*}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-m/d}),\\
{}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_{\mathrm{int}}, D_{\mathrm{abs}}) &= \Theta_{m,d}(Bn^{-1/2-m/d})\ .
\end{aligned}$$

Proof. Analogous to the proof of Theorem 5 in Appendix C.1, this can be shown using the positive homogeneity of $S \in \{S_{\mathrm{app}}, S_{\mathrm{opt}^*}, S_{\mathrm{int}}\}$ and $D \in \{D_\infty, D_{\mathrm{abs}}\}$, and by replacing the bump functions in the lower bound by the C^∞ bump functions from Definition C.1. □

We now prove our upper bound for log-partition estimation with stochastic evaluation points through approximation and importance sampling:

Theorem 9 (Upper bound for stochastic log-partition)**.** *There exists a constant $C_{m,d} > 0$ depending only on m and d such that*

$${}^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) \le O_{m,d}\left(\min\left\{Bn^{-m/d}, \exp(C_{m,d}Bn^{-m/d})Bn^{-1/2-m/d}\right\}\right)\ .$$

Proof. The bound $O_{m,d}(Bn^{-m/d})$ can be achieved even through methods with deterministic evaluation points, as proven in Theorem 7, hence we only need to show the other bound. Since the bounds agree for $n = 1$, we can in the following assume $n \ge 2$.

Let $\tilde{S}_n \in \mathcal{A}_n$ be a sequence of methods for which the worst-case errors

$$e_n \coloneqq \sup_{f \in \mathcal{F}_{d,m,B}} D_\infty(S_{\mathrm{app}}(f), \tilde{S}_n(f))$$

achieve the optimal rate $O_{m,d}(Bn^{-m/d})$ for the approximation problem on $\mathcal{F}_{d,m,B}$.

Set $N \coloneqq \lfloor n/2 \rfloor$, such that $N = \Omega(n)$ (since we assumed $n \ge 2$) and $2N \le n$. Set $g \coloneqq \tilde{S}_N(f)$. For N i.i.d. random variables $X_1, \dots, X_N \sim P_g$, set

$$\begin{aligned}
\mu_N &\coloneqq \frac{1}{N}\sum_{i=1}^{N} \exp(f(X_i) - g(X_i))\\
\tilde{S}_L(f) &\coloneqq L_g + \log \mu_N\ .
\end{aligned}$$

Then, $\tilde{S}_L$ only uses $2N \le n$ function evalutaions of f, hence $\tilde{S}_L \in {}^*C(\mathcal{A}_n^{\mathrm{ad}})$.

Since $\|f - g\|_\infty \le e_N$ and since exp is $\exp(e_N)$-Lipschitz on $(-\infty, e_N)$, we have

$$1 - e_N \le \exp(-e_N) \le \exp(f(X_i) - g(X_i)) \le \exp(e_N) \le 1 + e_N \exp(e_N)\ .$$

Hence, $|\exp(f(X_i) - g(X_i)) - \mathbb{E}\mu_N| \le (\exp(e_N) + 1)e_N$, which implies

$$\operatorname{Var} \mu_N \le N^{-1}((\exp(e_N) + 1)e_N)^2\ .$$

Additionally,

$$\log \mathbb{E}\mu_N = \log \int_{\mathcal{X}} e^{f(x)-g(x)} e^{g(x)-L_g}\,\mathrm{d}x = L_f - L_g\ .$$

Moreover, we have $e^{f(X_i)-g(X_i)} \in [\exp(-e_N), \exp(e_N)]$ and hence $\mu_N \in [\exp(-e_N), \exp(e_N)]$. Since log is $\exp(e_N)$-Lipschitz on $[\exp(-e_N), \exp(e_N)]$, we obtain

$$\begin{aligned}
\mathbb{E}|L_f - \tilde{S}_L(f)| &= \mathbb{E}|\log(\mathbb{E}\mu_N) - \log(\mu_N)| \le \mathbb{E}\exp(e_N)|\mu_N - \mathbb{E}\mu_N|\\
&\le \exp(e_N)\sqrt{\mathbb{E}[(\mu_N - \mathbb{E}\mu_N)^2]} = \exp(e_N)\sqrt{\operatorname{Var}\mu_N}\\
&\le N^{-1/2}e_N \exp(e_N)(\exp(e_N) + 1) \le 2\exp(2e_N)N^{-1/2}e_N\\
&\le O_{m,d}(\exp(C_{m,d}Bn^{-m/d})Bn^{-1/2-m/d})
\end{aligned}$$

for a suitable constant $C_{m,d} > 0$. □

In the optimization regime, we can directly exploit the relation to optimization to get a lower bound:

Proposition 10 (Lower bound for stochastic log-partition)**.** *For $m \geq 1$, we have*

$$^*\sigma_n^{\mathrm{ad}}(\mathcal{F}_{d,m,B}, S_L, D_{\mathrm{abs}}) \geq \Omega_{m,d}(Bn^{-m/d}) - d\log(1+3B)\ .$$

Proof. Take any stochastic log-partition method $\tilde{S} \in {}^*C(\mathcal{A}_n^{\mathrm{ad}})$. We can also interpret this as a stochastic optimization method. Hence, we know from an adaptation of the corresponding lower bound by Novak (1988) that there exists a constant $c_{m,d} > 0$ and a function $f \in \mathcal{F}_{d,m,B}$ such that $\mathbb{E}D_{\mathrm{abs}}(\tilde{S}(f), S_{\mathrm{opt}^*}(f)) \geq c_{m,d}Bn^{-m/d}$. But then, using $|f|_1 \leq d^{1/2}\|f\|_{C^1} \leq d^{1/2}B$ from Lemma B.1, we obtain

$$\begin{aligned}\mathbb{E}|\tilde{S}(f) - L_f| &\geq \mathbb{E}|\tilde{S}(f) - M_f| - |M_f - L_f| = \mathbb{E}D_{\mathrm{abs}}(\tilde{S}(f), S_{\mathrm{opt}^*}(f)) - |M_f - L_f| \\ &\overset{\text{Lemma 3}}{\geq} c_{m,d}Bn^{-m/d} - d\log(1+3B)\ .\end{aligned}$$

□

The following lemma will be useful to obtain a bound for rejection sampling in the sup-log distance:

Lemma C.6. *Let $p \in [0,1]$ and $c \geq 0$. Then, for any $a \in [-c, c]$, we have*

$$|\log(1 + p(e^a - 1))| \leq \min\{c, p(e^c - 1)\}\ .$$

Proof. For an upper bound, we use $\log(1+x) \leq x$ to obtain

$$\begin{aligned}\log(1 + p(e^a - 1)) &\leq \log(1 + p(e^c - 1)) \leq p(e^c - 1)\ , \\ \log(1 + p(e^a - 1)) &\leq \log(1 + (e^c - 1)) = c\ .\end{aligned}$$

For lower bounds, we note that

$$1 + p(e^a - 1) \geq 1 + p(e^{-c} - 1) \geq 1 + (e^{-c} - 1) = e^{-c}\ .$$

This immediately yields $\log(1 + p(e^a - 1)) \geq -c$. Moreover, because $\log$ is e^c-Lipschitz on $[e^{-c}, \infty)$, we have

$$\begin{aligned}\log(1 + p(e^a - 1)) &\geq \log(1 + p(e^{-c} - 1)) = \log(1 + p(e^{-c} - 1)) - \log(1) \geq -e^c|p(e^{-c} - 1)| \\ &= -p(e^c - 1)\ .\end{aligned}$$

□

Now, we can prove upper bounds for rejection sampling:

Lemma 11 (General rejection sampling bound)**.** *Suppose that $f, g : \mathcal{X} \to \mathbb{R}$ are bounded and measurable with $f(x) \leq g(x)$ for all $x \in \mathcal{X}$. Then, the distribution $\tilde{P}_f$ of* RejectionSampling(f, g, n) *satisfies*

$$\begin{aligned}\tilde{P}_f &= (1 - p_R)P_f + p_R P_g \qquad (2)\\ D_{\text{sup-log}}(P_f, \tilde{P}_f) &\leq \min\left\{D_{\text{sup-log}}(P_f, P_g), p_R(\exp(D_{\text{sup-log}}(P_f, P_g)) - 1)\right\} \\ D_{\mathrm{TV}}(P_f, \tilde{P}_f) &= p_R D_{\mathrm{TV}}(P_f, P_g) \\ W_1(P_f, \tilde{P}_f) &= p_R W_1(P_f, P_g)\ ,\end{aligned}$$

where $p_R = (1 - Z_f/Z_g)^n \leq \exp(-nZ_f/Z_g)$ is the probability of overall rejection.

Proof. **Step 1: Exact distribution.** We prove Eq. (2) via induction on n. For $n = 0$, this is clear. Now, suppose the statement is true for $n \in \mathbb{N}_0$. Denote by A the event that RejectionSampling$(f, g, n+1)$ accepts in the first iteration. Then, we have

$$P(A) = \mathbb{E}_{x\sim P_g}\mathbb{E}_{u\sim\mathcal{U}([0,1])}\mathbb{1}[ue^{g(x)} \leq e^{f(x)}] \overset{f\leq g}{=} \mathbb{E}_{x\sim P_g}\frac{e^{f(x)}}{e^{g(x)}} = \int_{\mathcal{X}} e^{f(x)-g(x)}\frac{e^{g(x)}}{Z_g}\,\mathrm{d}x = \frac{Z_f}{Z_g}\ .$$

The density of $x \in \mathcal{X}$ conditional on acceptance is

$$p(x|A) \propto p(A|x)p(x) = e^{f(x)-g(x)}\frac{e^{g(x)}}{Z_g} \propto p_f(x)\ ,$$

hence $P(x|A) = P_f$. On the other hand, the distribution $P(x|A^c)$, i.e. the distribution of x conditioned on not-acceptance is the distribution for RejectionSampling(f, g, n), which we know from Eq. (2) by the induction hypothesis. Hence, the distribution $\tilde{P}_f$ for RejectionSampling$(f, g, n+1)$ is

$$\tilde{P}_f = P(A)P(\cdot|A) + P(A^c)P(\cdot|A^c) = \frac{Z_f}{Z_g}P_f + \left(1 - \frac{Z_f}{Z_g}\right)\left(P_f + \left(1 - \frac{Z_f}{Z_g}\right)^n(P_g - P_f)\right)$$

$$= P_f + \left(1 - \frac{Z_f}{Z_g}\right)^{n+1} (P_g - P_f) = (1 - p_R)P_f + p_R P_g \ .$$

The argument above also shows that the overall rejection probability is $(1 - P(A))(1 - Z_f/Z_g)^n = (1 - Z_f/Z_g)^{n+1}$. Moreover, the bound $(1 - Z_f/Z_g)^n \leq \exp(-nZ_f/Z_g)$ follows from $1 - x \leq \exp(-x)$ for $x \geq 0$.

Step 2: Sup-log distance. From step 1, we see that

$$\begin{aligned} D_{\text{sup-log}}(P_f, \tilde{P}_f) &= \left\| \log\left((1 - p_R)e^{\bar{f}} + p_R e^{\bar{g}}\right) - \bar{f} \right\|_\infty \\ &= \left\| \log\left((1 - p_R) + p_R e^{\bar{g} - \bar{f}}\right) \right\|_\infty \\ &= \left\| \log\left(1 + p_R(e^{\bar{g} - \bar{f}} - 1)\right) \right\|_\infty \\ &\overset{\text{Lemma C.6}}{\leq} \min\left\{ \|\bar{g} - \bar{f}\|_\infty, p_R\left(e^{\|\bar{g} - \bar{f}\|_\infty} - 1\right) \right\} \\ &= \min\left\{ D_{\text{sup-log}}(P_g, P_f), p_R(\exp(D_{\text{sup-log}}(P_g, P_f)) - 1) \right\} \ . \end{aligned}$$

Step 3: TV distance. Using Eq. (2), we obtain for the TV distance:

$$\begin{aligned} D_{\text{TV}}(P_f, \tilde{P}_f) &= \sup_{A \subseteq \mathcal{X} \text{ measurable}} |P_f(A) - \tilde{P}_f(A)| = \sup_{A \subseteq \mathcal{X} \text{ measurable}} |p_R P_f(A) - p_R P_g(A)| \\ &= p_R D_{\text{TV}}(P_f, P_g) \ . \end{aligned}$$

Step 4: 1-Wasserstein distance. Using Eq. (2), we obtain for the 1-Wasserstein distance:

$$\begin{aligned} W_1(P_f, \tilde{P}_f) &= \sup_{\varphi \text{ 1-Lipschitz}} \left(\int \varphi(x) \, \mathrm{d}P_f(x) - \int \varphi(x) \, \mathrm{d}\tilde{P}_f(x) \right) \\ &= p_R \sup_{\varphi \text{ 1-Lipschitz}} \left(\int \varphi(x) \, \mathrm{d}P_f(x) - \int \varphi(x) \, \mathrm{d}P_g(x) \right) \\ &= p_R W_1(P_f, P_g) \ . \end{aligned}$$

□

With the upper bounds for rejection sampling proven above, we can analyze a combination of approximation and rejection sampling to prove the following upper bound:

Theorem 12 (Upper bound for sampling with stochastic evaluation points)**.** *There exists a constant $C_{m,d} > 0$ such that*

$$e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) \leq \begin{cases} O_{m,d}(Bn^{-m/d}) & , C_{m,d}Bn^{-m/d} > 1 \\ O_{m,d}((C_{m,d}Bn^{-m/d})^{n/2+1}) & , C_{m,d}Bn^{-m/d} \leq 1 \ . \end{cases}$$

Proof. **Step 1: Sampling method definition.** We consider the following sampling method:

(1) Use $\lfloor n/2 \rfloor$ function evaluations to create an approximation g of f, using a near-optimal approximation method such that the worst-case sup-log error is $E_n \leq O_{m,d}(Bn^{-m/d})$.
(2) Return a sample using RejectionSampling$(f, g + e_n, \lceil n/2 \rceil)$.

For step (1) we note that we have $\lfloor n/2 \rfloor \geq \Omega(n)$ except if $n = 1$. However, in the case $n = 1$, we can use the approximation $g = 0$ with worst-case error $E_n = B \leq O_{m,d}(Bn^{-m/d})$. Thus, it is indeed possible to achieve the bound in step (1).

Step 2: Upper bound. Denote by $C_{m,d} > 0$ a constant such that $E_n \leq C_{m,d}Bn^{-m/d}/2$. Moreover, denote by $\tilde{P}_f$ the distribution produced by the sampling method defined in step 1. By Lemma 11, we have for $\tilde{g} \coloneqq g + E_n$:

$$D_{\text{sup-log}}(P_f, \tilde{P}_f) \leq \min\left\{ D_{\text{sup-log}}(P_f, P_g), \left(1 - \frac{Z_f}{Z_{\tilde{g}}}\right)^{\lceil n/2 \rceil} (\exp(D_{\text{sup-log}}(P_f, P_g)) - 1) \right\} \ . \tag{13}$$

The first bound $D_{\text{sup-log}}(P_g, P_f)$ already yields the desired bound for $C_{m,d}Bn^{-m/d} > 1$. Now, consider the case $C_{m,d}Bn^{-m/d} \leq 1$. We have

$$Z_f = \int_{\mathcal{X}} e^{f(x)} \, \mathrm{d}x \geq \int_{\mathcal{X}} e^{\tilde{g}(x) - 2E_n} \, \mathrm{d}x \geq \exp(-C_{m,d}Bn^{-m/d}) Z_{\tilde{g}} \ .$$

Now, the second bound in Eq. (13) yields

$$\begin{aligned} D_{\text{sup-log}}(P_f, \tilde{P}_f) &\leq \left(1 - \exp(-C_{m,d} B n^{-m/d})\right)^{n/2} (\exp(C_{m,d} B n^{-m/d}) - 1) \\ &\leq \exp(C_{m,d} B n^{-m/d}) \left(1 - \exp(-C_{m,d} B n^{-m/d})\right)^{n/2+1} \\ &\leq e \cdot (C_{m,d} B n^{-m/d})^{n/2+1} \leq \Omega_{m,d}((C_{m,d} B n^{-m/d})^{n/2+1}) \,. \end{aligned}$$ □

Next, we prove corresponding lower bounds in the optimization regime, again using bump functions:

Theorem 13 (Lower bound for sampling with stochastic evaluation points)**.** *There exists a constant $c_{m,d} > 0$ such that for $B > 0$ and $n \in \mathbb{N}$ with $Bn^{-m/d} \geq c_{m,d}(1 + \log(n))$, we have*

$$\begin{aligned} e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{sup-log}}) &\geq \Omega_{m,d}(Bn^{-m/d}) \\ e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, D_{\text{TV}}) &\geq \Omega_{m,d}(1) \\ e_n^{\text{ad-stoch}}(\mathcal{F}_{d,m,B}, S_{\text{samp}}, W_1) &\geq \Omega_{m,d}(1) \,. \end{aligned}$$

Proof. We re-use some results from the proof of Theorem 7 in Appendix C.1. We consider again the decomposition of the cube $\mathcal{X}$ into three slices

$$\mathcal{C}_k \coloneqq [k/3, (k+1)/3] \times [0,1]^{d-1}, \qquad k \in \{0,1,2\} \,.$$

Consider a sampling algorithm $\tilde{S} \in \mathcal{A}_n^{\text{ad-stoch}}$ with stochastic evaluation points and consider a corresponding random sample $X_f = \phi(N(f,\omega),\omega)$ as defined in Section 2.2.

Step 1.1: Candidate functions for the sup-log distance. By Lemma C.3, $\mathcal{C}_0$ contains $4(n+1)$ disjoint open balls $B_\infty(z_i, \delta_n)$ with radius

$$\delta_n \coloneqq r_{4(n+1)} = \frac{(4(n+1))^{-1/d}}{12} \geq \frac{n^{-1/d}}{96} \,.$$

Let $f_0 \equiv 0$ be the zero function. Consider the set $\mathcal{Q}(\omega)$ containing the n random points where $N(f_0,\omega)$ queries f_0 and the one random point $\phi(N(f_0,\omega),\omega)$ that the sampling method outputs. We can pick an $i \in \{1,\dots,4(n+1)\}$ such that the cube $B_\infty(z_i,\delta_n)$ contains a point from $\mathcal{Q}(\omega)$ only with probability $\leq 1/4$. With $C_{m,d}$ as in Lemma C.2, we define

$$f_1(x) \coloneqq C_{m,d}^{-1} B \delta_n^{-1} b_{z_i,\delta_n}(x) \,,$$

which satisfies $f_1 \in \mathcal{F}_{d,m,B}$. Using analogous arguments to the proof of Theorem 7 in Appendix C.1, we obtain

$$\begin{aligned} L_{f_1} = \log(1 + I_n) &\geq \log(1 + \delta_n^d (e^{C_{m,d}^{-1} B \delta_n^m} - 1)) \geq C_{m,d}^{-1} B \delta_n^m + \log(\delta_n^d) \\ &\geq \tilde{c}_{m,d} B n^{-m/d} - \log(n) - d\log(96) \end{aligned}$$

for a suitable constant $\tilde{c}_{m,d} > 0$.

Step 1.2: Bounding the distribution on f_1. Now, the probability that $\phi(N(f_0,\omega),\omega) \in B_\infty(z_i,\delta_n)$ is at most $1/4$ by construction. Moreover, the probability of $N(f_0,\omega)$ querying $B_\infty(z_i,\delta_n)$ is also at most $1/4$, hence the probability of $N(f_1,\omega)$ querying $B_\infty(z_i,\delta_n)$ is also at most $1/4$. By the union bound, the probability that $\phi(N(f_1,\omega),\omega) \in B_\infty(z_i,\delta_n)$ is at most $1/2$. Now, to have $D_{\text{sup-log}}(\tilde{S}(f_1), P_{f_1}) < \infty$, $\tilde{S}(f_1)$ must be of the form P_g for some function $g : \mathcal{X} \to \mathbb{R}$. Without loss of generality, we can assume $L_g = 0$. Then, since the set $\hat{\mathcal{X}} \coloneqq \mathcal{X} \setminus B_\infty(z_i,\delta_n)$ satisfies $P_g(\hat{\mathcal{X}}) \geq 1/2$, there exists $x \in \hat{\mathcal{X}}$ with $p_g(x) \geq 1/2$, implying $g(x) \geq \log(1/2)$. But then,

$$\begin{aligned} D_{\text{sup-log}}(\tilde{S}(f_1), P_{f_1}) &\geq |(g(x) - L_g) - (f_1(x) - L_{f_1})| = |g(x) + L_{f_1}| \\ &\geq \tilde{c}_{m,d} B n^{-m/d} - \log(n) - d\log(96) - \log(2) \geq \tilde{c}_{m,d} B n^{-m/d} - \log(n) - 6d \,. \end{aligned}$$

Especially, for $Bn^{-m/d} \geq 12d\tilde{c}_{m,d}^{-1}(1 + \log(n))$, we have

$$D_{\text{sup-log}}(\tilde{S}(f_1), P_{f_1}) \geq \tilde{c}_{m,d} B n^{-m/d} - \frac{1}{2}\tilde{c}_{m,d} B n^{-m/d} = \Omega_{m,d}(Bn^{-m/d}) \,.$$

Step 2.1: Candidate functions for the Wasserstein distance. By Lemma C.3, for $M \in \mathbb{N}$ to be determined later, we can place Mn subcubes each in $\mathcal{C}_0$ and $\mathcal{C}_2$ with radius

$$\delta_n \coloneqq r_{Mn} \coloneqq \frac{(Mn)^{-1/d}}{12} \,.$$

By an analogous argument to Step 1.1, we can find subcubes $B_\infty(z_0, \delta_n)$ and $B_\infty(z_2, \delta_n)$ of $\mathcal{C}_0$ and $\mathcal{C}_2$ such that the probability of one of them being queried for f_0 is at most $2/M$. Following Lemma C.2, we construct the functions

$$f_k(x) \coloneqq C_{m,d}^{-1} B \delta_n^m b_{z_k, \delta_n}(x), \qquad k \in \{0, 2\} ,$$

which are contained in $\mathcal{F}_{d,m,B}$.

Step 2.2: Bounding the Wasserstein distance. We set $M \coloneqq 20d$. Since the two subcubes are only queried with probability at most $2/M$, we know that

$$W_1(\tilde{S}(f_0), \tilde{S}(f_2)) \leq d^{1/2} D_{\mathrm{TV}}(\tilde{S}(f_0), \tilde{S}(f_2)) \leq d^{1/2} \frac{2}{M} \leq \frac{1}{10} .$$

With an argument analogous to the proof of Theorem 7 in Appendix C.1, we obtain

$$W_1(P_{f_0}, P_{f_2}) \geq \frac{1}{3} \frac{I_n}{1 + I_n} \overset{\text{Lemma C.4}}{\geq} \frac{1}{6} \min\{1, I_n\} ,$$

and

$$I_n \geq \delta_n^d (e^{B C_{m,d}^{-1} \delta_n^m} - 1) \geq \tilde{c}_{m,d} n^{-1} (e^{\tilde{c}_{m,d} B n^{-m/d}} - 1)$$

for a suitable constant $\tilde{c}_{m,d} \in (0, 1)$. Now, suppose that

$$B n^{-m/d} \geq \tilde{c}_{m,d}^{-1} (1 + \log(\tilde{c}_{m,d}^{-1}))(1 + \log(n)) .$$

We obtain

$$\tilde{c}_{m,d} B n^{-m/d} \geq 1 + \log(\tilde{c}_{m,d}^{-1}) + \log(n) \geq \log(\tilde{c}_{m,d}^{-1} n + 1)$$

and therefore

$$W_1(P_{f_0}, P_{f_2}) \geq \frac{1}{6} \min\{1, I_n\} \geq \frac{1}{6} \min\{1, 1\} = \frac{1}{6} .$$

Since W_1 satisfies the triangle inequality, there must exist $k \in \{0, 2\}$ with

$$W_1(\tilde{S}(f_k), P_{f_k}) \geq \frac{1}{2} \left(\frac{1}{6} - \frac{1}{10} \right) = \frac{1}{30} = \Omega_{m,d}(1) .$$

Step 3: TV lower bound. The corresponding lower bound for the TV distance follows from the inequality $W_1(P, Q) \leq d^{1/2} D_{\mathrm{TV}}(P, Q)$. □

Finally, we prove our auxiliary result on the complexity of sampling when the log-partition function is known:

Proposition 14. *Consider the function class $\mathcal{F} \coloneqq \{f \in C(\mathcal{X}) \mid \|f\|_\infty \leq \log(3/2), L_f = 0\}$. Then,*

$$e_n^{\text{ad-stoch}}(\mathcal{F}, S_{\mathrm{samp}}, D_{\text{sup-log}}) = 0$$

for all $n \geq 1$.

Proof. For $f \in \mathcal{F}$, define $\tilde{f}(x) \coloneqq \log(2 \exp(f(x)) - 1)$ and $g(x) \coloneqq \log(2)$. We have

$$Z_{\tilde{f}} = \int_{\mathcal{X}} (2 \exp(f(x)) - 1) \, \mathrm{d}x = 2 Z_f - 1 = 1 ,$$
$$Z_g = \int_{\mathcal{X}} e^{\log(2)} \, \mathrm{d}x = 2 .$$

Let $\tilde{P}_f$ be the distribution of REJECTIONSAMPLING$(\tilde{f}, g, 1)$, which only uses one evaluation of $\tilde{f}$ and therefore only one evaluation of f. By Lemma 11, we have

$$\tilde{P}_f = \frac{Z_{\tilde{f}}}{Z_g} P_{\tilde{f}} + \left(1 - \frac{Z_{\tilde{f}}}{Z_g}\right) P_g = \frac{1}{2} P_{\tilde{f}} + \frac{1}{2} P_g = P_f .$$ □

D Proofs for Relations Between Different Problems

The proof of the following theorem adapts results from the literature, showing that they apply to our setting:

Theorem 15 (adapted from Li (2016) and Mirzaei (2015))**.** *Let $m, d \in \mathbb{N}_{\geq 1}$. Using the moving least squares method, it is possible to construct an approximation f_n of $f \in C^m(\mathcal{X})$ using n deterministic non-adaptive function evaluations such that*

(a) $\|f - f_n\|_{C^k} \leq O_{m,d}(\|f\|_{C^m} n^{-(m-k)/d})$ for $k \in \{0, 1, \dots, m\}$,
(b) the runtime of constructing f_n is zero (construction takes place on-the-fly during evaluation), and
(c) the runtime of evaluating f_n at a point $x \in \mathcal{X}$ is $O_{m,d}(1)$.

Proof. **Step 1: The method.** The idea of the moving least squares method (Lancaster and Salkauskas, 1981) is to obtain an approximation $g(x) = g_x(x)$ of $f(x)$ at an evaluation point x by determining g_x as the solution to a polynomial least-squares regression problem with data $(x_i, f(x_i))$, weighted with weights $\Phi(x, x_i)$ that (smoothly) vanish for large $\|x - x_i\|$. We will not state the exact method here but refer to the publications by Li (2016) and Mirzaei (2015), whose analysis we are using here. While Theorem 4.1 of Li (2016) essentially directly provides the result (a), it is unclear to us if the corresponding constants are independent of the evaluation points in the way that we need. Thus, in the following, we will try to verify the slightly stronger conditions of Theorem 3.12 of Mirzaei (2015) and explain how it can be adapted to our setting with minor modifications.

Step 2: Verifying the assumptions. Now, we list the major assumptions of Theorem 3.12 of Mirzaei (2015) and show that they are satisfied for a suitable choice of evaluation points and weighting function. The assumptions on smoothness are deferred until Step 3, where we will show how to adapt them to our setting. We define the number $N \coloneqq \lfloor n^{1/d} \rfloor$ of grid points along each axis. By dividing each axis into N equal intervals, we obtain a partition of $\mathcal{X}$ into N^d cubes. Let X be the set of midpoints of these cubes. Hence, $|X| = N^d = \lfloor n^{1/d} \rfloor^d \geq (n^{1/d}/2)^d = \Omega_{m,d}(n)$.

- The considered domain Ω is a bounded set with Lipschitz boundary. We want to consider $\Omega \coloneqq \mathcal{X}$, which is bounded and has a Lipschitz boundary.
- The maximum degree m of the polynomial basis satisfies $m \geq 1$. While m denotes the (known) smoothness of the target function f in our context, we will assume that the maximum degree of the polynomial basis is also m. While a maximum degree of $m - 1$ should be sufficient for our purposes (as it is in Li (2016)), using a maximum degree of m avoids notational confusion and simplifies the adaptation of the arguments of Mirzaei (2015).
- The fill distance $h_{X,\Omega} = \sup_{x \in \Omega} \min_{x' \in X} \|x - x'\|_2$ satisfies $h_{X,\Omega} \leq \min\{h_0, 1\}$ for some given constant $h_0 > 0$. In our case, the fill distance is $h_{X,\Omega} = \sqrt{d}/(2N) = \Theta_{m,d}(n^{-1/d})$, which satisfies the assumption for large enough values of n. The errors for smaller n do not affect the asymptotic rate.
- The weight function is defined through a radial function $\phi : [0, \infty) \to \mathbb{R}$, which is supported in $[0, 1]$ and its even extension belongs to $C^m(\mathbb{R})$. For this, we can just use the even and C^∞-smooth bump function b from Definition C.1 and set $\phi(x) \coloneqq b(x)$.
- The point set X is quasi-uniform with constant independent of f and n. This means that the separation distance
$$q_{X,\Omega} \coloneqq \frac{1}{2} \min_{x, x' \in X : x \neq x'} \|x - x'\|_2$$
satisfies $q_{X,\Omega} \leq h_{X,\Omega} \leq c_{\mathrm{qu}} q_{X,\Omega}$ for a constant c_{qu} independent of f and n. In our case, we have $q_{X,\Omega} = 1/(2N)$, and hence we can set $c_{\mathrm{qu}} \coloneqq \sqrt{d}$.

Step 3: Adapting the argument of Mirzaei (2015). Let f_n be the moving least squares approximation of f with evaluation points X. By Corollary 4.5 in Wendland (2004), f_n is in C^m since the weight function is also in C^m. Hence, the norms $\|f - f_n\|_{C^m}$ and $\|f - f_n\|_{W_\infty^m}$ are equivalent, where $W_p^m(\Omega)$ is the Sobolev space of smoothness m with the p-norm applied to the (weak) derivatives. Theorem 3.12 in Mirzaei (2015) shows that

$$\|f - f_n\|_{W_q^{|\alpha|}(\Omega)} \leq C h_{X,\Omega}^{m+s-|\alpha|-d\max\{0,1/p-1/q\}} \|f\|_{W_p^{m+s}(\Omega)}$$

for $p \in [1, \infty), q \in [1, \infty], s \in [0, 1)$ and a multi-index α satisfying $m > |\alpha| + d/p$. We would obtain (a) by setting $s = 0$, $p = q = \infty$, and $|\alpha| = k$. However, setting $p = \infty$ is not allowed by the assumptions of the theorem, and setting $|\alpha| = m$ for $m = k$ is also not allowed. Hence, we need to show that the theorem can be extended to $p = \infty$ and $|\alpha| = m$ in the special case $s = 0$ and $q = \infty$. The assumption $p < \infty$ is used for the Sobolev extension operator, but it is noted in the proof that $p = \infty$ is allowed for $s = 0$. The only other point where $p < \infty$ and $|\alpha| < m$ are required is in the invocation of Eq. (3.4) in Lemma 3.3 of Mirzaei (2015). However, for the special case $s = 0$, $p = q = \infty$ and $|\alpha| \leq m$, the statement of Lemma 3.3 also holds, as is shown by the Bramble-Hilbert lemma (cf. Lemma (4.3.8) in Brenner et al., 2008), which has also been employed by (Li, 2016) for the same purpose.

Step 4: Runtime bound. For (b) and (c), we note that due to the local support of the weight function, evaluating the moving least squares approximation at a point $x \in \mathcal{X}$ mainly requires the solution of a regression problem with $O_{m,d}(1)$ variables and evaluation points. This is shown, for example, above Lemma 3.6 in Mirzaei (2015). The moving least squares method does not require any pre-computations. □

D.1 Proofs for Relation between Sampling and Log-Partition Estimation

For analyzing thermodynamic integration, we are going to use Hoeffding's inequality in the form stated and proved in Theorem 6.10 in Steinwart and Christmann (2008).

Theorem D.1 (Hoeffding's inequality)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $a < b$ be two real numbers, $n \geq 1$ be an integer, and $\xi_1, \dots, \xi_n : \Omega \to [a, b]$ be independent random variables. Then, for all $\tau > 0$, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}_P\xi_i) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq e^{-\tau} .$$

Theorem 18 (Convergence of thermodynamic integration)**.** *Given $N \in \mathbb{N}_{\geq 1}$ and a sampling algorithm producing samples from approximate distributions $\tilde{P}_{\beta f}$, consider the following algorithm:*

- *Sample $\beta_1, \dots, \beta_N \sim \mathcal{U}([0,1])$ independently.*
- *Draw $X_i \sim \tilde{P}_{\beta_i f}$ independently.*
- *Output $\tilde{L}_f \coloneqq \frac{1}{N}\sum_{i=1}^{N} f(X_i)$.*

Then, for $\delta > 0$, we have

$$|\tilde{L}_f - L_f| \leq |L_f - \mathbb{E}\tilde{L}_f| + 2\|f\|_\infty\sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability $\geq 1 - \delta$, where

$$|L_f - \mathbb{E}\tilde{L}_f| \leq 2\|f\|_\infty \sup_{\beta\in[0,1]} D_{\mathrm{TV}}(P_{\beta f}, \tilde{P}_{\beta f}),$$
$$|L_f - \mathbb{E}\tilde{L}_f| \leq |f|_1 \sup_{\beta\in[0,1]} W_1(P_{\beta f}, \tilde{P}_{\beta f}).$$

Proof. Obviously, we have

$$|\tilde{L}_f - L_f| \leq |L_f - \mathbb{E}\tilde{L}_f| + |\tilde{L}_f - \mathbb{E}\tilde{L}_f| .$$

Step 1: Bounding the second term. For bounding the second term, we use Hoeffding's inequality (Theorem D.1) with $\xi_i \coloneqq f(X_i)$, $n = N$ and $\tau \coloneqq \log(2/\delta)$. Ignoring null sets, we can choose $b = \|f\|_\infty$ and $a = -\|f\|_\infty$. We then obtain

$$\tilde{L}_f - \mathbb{E}\tilde{L}_f \geq 2\|f\|_\infty\sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability $\leq \exp(-\log(2/\delta)) = \delta/2$. By applying the same argument to $\xi_i = -f(X_i)$ and applying the union bound, we obtain

$$|\tilde{L}_f - \mathbb{E}\tilde{L}_f| \leq 2\|f\|_\infty\sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability $\geq 1 - \delta$.

Step 2: Bounding the first term. We use

$$\begin{aligned} |L_f - \mathbb{E}\tilde{L}_f| &= \left|\mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta f}}[f(x)] - \mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim\tilde{P}_{\beta f}}[f(x)]\right| \\ &\leq \sup_{\beta\in\mathcal{U}([0,1])}\left|\mathbb{E}_{x\sim P_{\beta f}}[f(x)] - \mathbb{E}_{x\sim\tilde{P}_{\beta f}}[f(x)]\right| . \end{aligned} \tag{14}$$

We can assume that $|f|_1 \neq 0$ since the bound is clear otherwise. Using the dual formulation of the 1-Wasserstein distance and that $f/|f|_1$ is 1-Lipschitz, we directly obtain

$$\left|\mathbb{E}_{x\sim P_{\beta f}}[f(x)] - \mathbb{E}_{x\sim\tilde{P}_{\beta f}}[f(x)]\right| = |f|_1\left|\mathbb{E}_{x\sim P_{\beta f}}[f(x)/|f|_1] - \mathbb{E}_{x\sim\tilde{P}_{\beta f}}[f(x)/|f|_1]\right| \leq |f|_1 W_1(\tilde{P}_{\beta f}, P_{\beta f}) .$$

Similarly, the bound on the TV distance follows from an alternative formulation of the TV distance (see e.g. Gibbs and Su, 2002) given by

$$D_{\mathrm{TV}}(P,Q) = \frac{1}{2} \sup_{g:\|g\|_\infty \le 1} \left| \int g \,\mathrm{d}P - \int g \,\mathrm{d}Q \right| . \qquad \square$$

Remark D.2. In Theorem 18, we can hope for a better bound in terms of the sup-log distance. For example, suppose that $\tilde{P}_{\beta f} = P_{\beta g}$, where g is an approximation of f that is independent of β. Since g is only determined up to a constant shift, we can assume that $L_g = L_f$. Then,

$$\begin{aligned}
|L_f - \mathbb{E}\tilde{L}_f| &= \left|\mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta f}}[f(x)] - \mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta g}}[f(x)]\right| \\
&\le \left|\mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta f}}[f(x)] - \mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta g}}[g(x)]\right| \\
&\quad + \left|\mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta g}}[g(x)] - \mathbb{E}_{\beta\sim\mathcal{U}([0,1])}\mathbb{E}_{x\sim P_{\beta g}}[f(x)]\right| \\
&\le |L_f - L_g| + \|g-f\|_\infty = 0 + D_{\text{sup-log}}(P_f, P_g) = D_{\text{sup-log}}(P_f, P_g) .
\end{aligned}$$

However, in the general case, the approach in Eq. (14) of taking the supremum over β cannot yield such a good bound. This can be seen by considering indicator functions $f = a\mathbb{1}_A$ and $g = (a+\delta)\mathbb{1}_A$. Instead, it appears that it would be necessary to obtain bounds depending on β and f and show that their integral over $\beta \in [0,1]$ is sufficiently small for all f. ◀

Theorem 19 (Convergence of bisection sampling)**.** *Let $m \ge 1, B \ge 0$ and $M \in \mathbb{N}_0$. Let $f \in \mathcal{F}_{d,m,B}$ and let $\tilde{L}$ be a log-partition estimator with worst-case error $E \ge 0$ on $\mathcal{F}_{d,m,B}$. Let $f \in C^m(\mathcal{X})$ and let $\tilde{P}_f$ be the distribution of samples produced by* BISECTIONSAMPLING$(f, \tilde{L}, M)$ *in Algorithm 2. Then,*

$$D_{\text{sup-log}}(\tilde{P}_f, P_f) \le 2MdE + 2^{-M} d \|f\|_{C^1} .$$

Proof. **Step 1: Log-density analysis.** We want to show that $\tilde{P}_f$ has a density $\tilde{p}_f$ and bound $\|\log \tilde{p}_f - \log p_f\|_\infty$. Partition $\mathcal{X}$ into 2^{Md} cubes of side length 2^{-M}. Since a density is only defined up to a null set, it suffices to consider an arbitrary x in the interior of one of these cubes, which we fix in the following. We denote the corresponding cube by $\mathcal{Z}^{(Md)}$. We can then find exactly one sequence $\mathcal{Z}^{(0)} = \mathcal{X}, \mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(Md)}$ of hyperrectangles which could have been visited during the execution of Algorithm 2 to obtain x. Since Algorithm 2 samples uniformly from $\mathcal{Z}^{(Md)}$ and the volume of $\mathcal{Z}^{(Md)}$ is 2^{-Md}, we have the density

$$\tilde{p}_f(x) = 2^{Md} \tilde{P}_f(\mathcal{Z}^{(Md)}) .$$

On the other hand, a simple integration argument shows that the target density satisfies

$$\inf_{x' \in \mathcal{Z}^{(Md)}} p_f(x') \le 2^{Md} P_f(\mathcal{Z}^{(Md)}) \le \sup_{x' \in \mathcal{Z}^{(Md)}} p_f(x') .$$

This yields

$$\begin{aligned}
|\log \tilde{p}_f(x) - \log p_f(x)| \le{}& \left|\log(2^{Md}\tilde{P}_f(\mathcal{Z}^{(Md)})) - \log(2^{Md}P_f(\mathcal{Z}^{(Md)}))\right| \\
&+ \left| \sup_{x' \in \mathcal{Z}^{(Md)}} \log p_f(x') - \inf_{x'' \in \mathcal{Z}^{(Md)}} \log p_f(x'') \right| .
\end{aligned}$$

Step 2: Bounding the second term. Since $\mathcal{Z}^{(Md)}$ is an axis-aligned cube with side length 2^{-M}, we have for $x', x'' \in \mathcal{Z}^{(Md)}$:

$$|f(x') - f(x'')| \le \sum_{i=1}^{d} \|\partial^i f\|_\infty 2^{-M} \le 2^{-M} d \|f\|_{C^1} .$$

Step 3: Bounding the first term. We can simplify

$$\left|\log(2^{Md}\tilde{P}_f(\mathcal{Z}^{(Md)})) - \log(2^{Md}P_f(\mathcal{Z}^{(Md)}))\right| = \left|\log(\tilde{P}_f(\mathcal{Z}^{(Md)})) - \log(P_f(\mathcal{Z}^{(Md)}))\right| .$$

We want to show by induction over $k \in \{0, \dots, Md\}$ that

$$\left|\log(\tilde{P}_f(\mathcal{Z}^{(k)})) - \log(P_f(\mathcal{Z}^{(k)}))\right| \le 2kE ,$$

which will then yield the desired error bound for $k = Md$. This is obviously true for $k = 0$. Now, suppose that it is true for some $k \in \{0, \dots, Md-1\}$. Consider a partition of $\mathcal{Z}^{(k)}$ into two equal-sized sub-hyperrectangles $\mathcal{Z}_1$ and $\mathcal{Z}_2$ as in Algorithm 2 such that $\mathcal{Z}^{(k+1)} = \mathcal{Z}_i$ for some $i \in \{1, 2\}$. Then,

$$\tilde{P}_f(\mathcal{Z}^{(k+1)}) = \sigma(\tilde{L}_{f_{\mathcal{Z}_i}} - \tilde{L}_{f_{\mathcal{Z}_{3-i}}})\tilde{P}_f(\mathcal{Z}^{(k)}) \,,$$

which also holds for $i = 2$ since the sigmoid function σ satisfies $\sigma(-u) = 1 - \sigma(u)$ for all $u \in \mathbb{R}$. Moreover, we have

$$P_f(\mathcal{Z}^{(k+1)}) = \frac{P_f(\mathcal{Z}_i)}{P_f(\mathcal{Z}_i) + P_f(\mathcal{Z}_{3-i})} P_f(\mathcal{Z}^{(k)}) = \frac{e^{L_{f_{\mathcal{Z}_i}}}}{e^{L_{f_{\mathcal{Z}_i}}} + e^{L_{f_{\mathcal{Z}_{3-i}}}}} P_f(\mathcal{Z}^{(k)}) = \sigma(L_{f_{\mathcal{Z}_i}} - L_{f_{\mathcal{Z}_{3-i}}}) P_f(\mathcal{Z}^{(k)}) \,.$$

By definition of the functions $f_{\mathcal{Z}_{i'}}$, $i' \in \{1, 2\}$ in Algorithm 2, since the side-lengths h_j of $\mathcal{Z}_{i'}$ satisfy $h_j \le 1$, we have $\|f_{\mathcal{Z}_{i'}}\|_{C^m} \le \|f\|_{C^m} \le B$, which means $f_{\mathcal{Z}_{i'}} \in \mathcal{F}_{d,m,B}$. Hence, by assumption, the log-partition error is

$$|\tilde{L}_{f_{\mathcal{Z}_{i'}}} - L_{f_{\mathcal{Z}_{i'}}}| \le \varepsilon \,.$$

Now, the log-sigmoid function $h(u) \coloneqq \log \sigma(u)$ satisfies $h'(u) = \frac{\sigma(u)(1-\sigma(u))}{\sigma(u)} = 1 - \sigma(u) \in (0, 1)$ and is therefore 1-Lipschitz. Hence,

$$\begin{aligned}
\left|\log\left(\tilde{P}_f(\mathcal{Z}^{(k+1)})\right) - \log\left(P_f(\mathcal{Z}^{(k+1)})\right)\right| &\le \left|\log\left(\tilde{P}_f(\mathcal{Z}^{(k)})\right) - \log\left(P_f(\mathcal{Z}^{(k)})\right)\right| \\
&\quad + \left|h(\tilde{L}_{f_{\mathcal{Z}_i}} - \tilde{L}_{f_{\mathcal{Z}_{3-i}}}) - h(L_{f_{\mathcal{Z}_i}} - L_{f_{\mathcal{Z}_{3-i}}})\right| \\
&\le 2kE + 2E = 2(k+1)E \,,
\end{aligned}$$

which completes the induction. □

D.2 Proofs for Relation to Optimization

Proposition 20 (Optimization by approximate sampling)**.** *Let Q be a probability distribution on $\mathcal{X}$. Then, for any $\delta \in (0, 1]$ and $\varepsilon > 0$,*

(a) $Q(\{x \in \mathcal{X} \mid f(x) \le \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta) - \varepsilon D_{\text{sup-log}}(P_{f/\varepsilon}, Q)\}) \le \delta$,
(b) $Q(\{x \in \mathcal{X} \mid f(x) \le \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)\}) \le \delta + D_{\text{TV}}(P_{f/\varepsilon}, Q)$,
(c) $Q(\{x \in \mathcal{X} \mid f(x) < \varepsilon L_{f/\varepsilon} - \varepsilon \log(2/\delta) - 2\delta^{-1}|f|_1 W_1(P_{f/\varepsilon}, Q)\}) \le \delta$.

Proof.

(a) Suppose $D_{\text{sup-log}}(P_f, Q) < \infty$. Then, $Q = P_g$ for some g. For almost every $x \in \mathcal{X}$, we have the implications

$$\begin{aligned}
f(x) \le L_f - \log(1/\delta) - D_{\text{sup-log}}(P_f, Q) &\Leftrightarrow \bar{f}(x) \le L_{\bar{f}} - \log(1/\delta) - \|\bar{f} - \bar{g}\|_\infty \\
&\Rightarrow \bar{g}(x) \le L_{\bar{f}} - \log(1/\delta) \\
&\Leftrightarrow \bar{g}(x) \le L_{\bar{g}} - \log(1/\delta) \,.
\end{aligned}$$

Hence,

$$\begin{aligned}
&Q(\{x \in \mathcal{X} \mid f(x) \le L_f - \log(1/\delta) - D_{\text{sup-log}}(P_f, Q)\}) \\
&\le Q(\{x \in \mathcal{X} \mid \bar{g}(x) \le L_{\bar{g}} - \log(1/\delta)\}) \\
&= P_{\bar{g}}(\{x \in \mathcal{X} \mid \bar{g}(x) \le L_{\bar{g}} - \log(1/\delta)\}) \overset{\text{Lemma 3}}{\le} \delta \,.
\end{aligned}$$

By using f/ε instead of f and multiplying both sides of the inequality by ε, we obtain

$$Q(\{x \in \mathcal{X} \mid f(x) \le \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta) - \varepsilon D_{\text{sup-log}}(P_{f/\varepsilon}, Q)\}) \le \delta \,.$$

(b) The TV norm bound follows from Lemma 3 because for the considered event A,

$$Q(A) \le P_{f/\varepsilon}(A) + \sup_{A'} |Q(A) - P_{f/\varepsilon}(A')| = P_{f/\varepsilon}(A) + D_{\text{TV}}(P_{f/\varepsilon}, Q) \,.$$

(c) Let $\tilde{\varepsilon} > 0$. By definition of the Wasserstein distance, there exist random variables $X \sim P_{f/\varepsilon}$ and $Y \sim Q$ on a common probability space $(\Omega, \mathcal{F}, P_\Omega)$ such that $\mathbb{E}\|X - Y\|_2 \le W_1(P_{f/\varepsilon}, Q) + \tilde{\varepsilon}$. By the Markov inequality, we then have

$$\|X - Y\|_2 \le 2(W_1(P_{f/\varepsilon}, Q) + \tilde{\varepsilon})/\delta$$

with probability $\ge 1 - \delta/2$. Moreover, by Lemma 3, we have

$$f(X) > \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta)$$

with probability $\ge 1 - \delta/2$. By the union bound, we hence have

$$f(Y) > f(X) - |f|_1 \|X - Y\|_2 > \varepsilon L_{f/\varepsilon} - \varepsilon \log(1/\delta) - 2\delta^{-1}|f|_1(W_1(P_{f/\varepsilon}, Q) + \tilde{\varepsilon})$$

with probability $\ge 1 - \delta$. Since $\tilde{\varepsilon} > 0$ was arbitrary, the claim follows. □

E Proofs for Algorithms

E.1 Proofs for Approximation-based Algorithms

E.1.1 Proofs for Piecewise Constant Approximation

To study the error of piecewise constant approximation, we study the log-partition function of linear functions f. A first step is achieved using the following lemma:

Lemma E.1. *Let*

$$r : \mathbb{R} \to \mathbb{R}, t \mapsto \log\left(\int_{-1/2}^{1/2} \exp(tu)\,\mathrm{d}u\right) = \begin{cases} 0 & ,t = 0 \\ \log\left(t^{-1}(\exp(t/2) - \exp(-t/2))\right) & ,t \neq 0\,. \end{cases}$$

Then, r is even and $(1/2)$-Lipschitz with $r(t) \geq 0$ for all t and we have more generally for $a > 0$ and $t_1, \dots, t_d \in \mathbb{R}$:

$$\log\left(\int_{[-a/2,a/2]^d} \exp\left(\sum_{k=1}^d t_k u_k\right)\,\mathrm{d}u\right) = d\log(a) + \sum_{k=1}^d r(at_k)\,.$$

Proof. It follows from a simple symmetry argument that r is even. We have

$$\int_{-1/2}^{1/2} \exp(tu)\,\mathrm{d}u \geq \int_{-1/2}^{1/2} (1 + tu)\,\mathrm{d}u = 1\,,$$

which shows $r(t) \geq 0$. For $h > 0$ and $u \in [-1/2, 1/2]$, we have $\exp(hu) \in [\exp(-h/2), \exp(h/2)]$. Using the mean value theorem of integration, we obtain

$$r(t+h) - r(t) = \log\left(\int_{-1/2}^{1/2} \exp(tu)\exp(hu)\,\mathrm{d}u\right) - \log\left(\int_{-1/2}^{1/2} \exp(tu)\,\mathrm{d}u\right) \in [-h/2, h/2]\,,$$

which shows that r is $1/2$-Lipschitz.

For the more general integral, we use that the integrand is a product of one-dimensional functions to decompose

$$\begin{aligned} \log\left(\int_{[-a/2,a/2]^d} \exp\left(\sum_{k=1}^d t_k u_k\right)\,\mathrm{d}u\right) &= \log\left(\prod_{k=1}^d \int_{-a/2}^{a/2} \exp\left(t_k u_k\right)\,\mathrm{d}u_k\right) \\ &\overset{\text{Subst. } u_k = a v_k}{=} \log\left(\prod_{k=1}^d \int_{-1/2}^{1/2} \exp\left(t_k a v_k\right) a\,\mathrm{d}v_k\right) \\ &= d\log(a) + \sum_{k=1}^d r(at_k)\,. \end{aligned}$$

□

Another ingredient for the analysis of piecewise constant approximation is to analyze the global error through the errors on individual subcubes:

Lemma E.2. *Let $f, g : \mathcal{X} \to \mathbb{R}$ be bounded and measurable. Let $\mathcal{X}_i$ be a partition of $\mathcal{X}$. Let $L_f(\mathcal{X}_i) \coloneqq \log\left(\int_{\mathcal{X}_i} \exp(f(x))\,\mathrm{d}x\right)$. Then,*

$$\inf_i [L_f(\mathcal{X}_i) - L_g(\mathcal{X}_i)] \leq L_f - L_g \leq \sup_i [L_f(\mathcal{X}_i) - L_g(\mathcal{X}_i)]\,.$$

Proof. We prove the second inequality here, the first one follows analogously. Let $s \coloneqq \sup_i [L_f(\mathcal{X}_i) - L_g(\mathcal{X}_i)]$. Then,

$$L_f = \log\left(\sum_i \exp(L_f(\mathcal{X}_i))\right) \leq \log\left(\sum_i \exp(s)\exp(L_g(\mathcal{X}_i))\right) = L_g + s\,.$$

□

We now prove convergence rates for piecewise constant approximation, using a combination of different approaches:

Theorem 21 (Convergence rate of piecewise constant approximation)**.** *Let $m \geq 1$ and $n = N^d$ as above. If g_n is a piecewise constant interpolant as above, we have*

$$\sup_{f \in \mathcal{F}_{d,m,B}} |L_f - L_{g_n}| = \begin{cases} \Theta_{m,d}(Bn^{-1/d}) & \text{, if } m = 1 \text{ or } Bn^{-1/d} > 1 \\ \Theta_{m,d}(\max\{B, B^2\} n^{-2/d}) & \text{, otherwise.} \end{cases}$$
$$\sup_{f \in \mathcal{F}_{d,m,B}} D_{\text{sup-log}}(P_f, P_{g_n}) = \Theta_{m,d}(Bn^{-1/d}) \,.$$

Proof. Recall from Section 4.1.1 that g_n is piecewise constant on the cubes $\mathcal{X}_i$, interpolating f in the cube centers $x^{(i)}$.

Step 1: Lipschitz-type upper bounds. Let $f \in \mathcal{F}_{d,m,B}$. Since f is $Bd^{1/2}$-Lipschitz by Lemma B.1, it is easy to see that $\|f - g_n\|_\infty \leq O_{m,d}(B/N) = O_{m,d}(Bn^{-1/d})$. Then, it follows directly from Proposition 6 that $|L_f - L_{g_n}| \leq \|f - g_n\|_\infty \leq O_{m,d}(Bn^{-1/d})$ and $D_{\text{sup-log}}(P_f, P_{g_n}) \leq 2\|f - g_n\|_\infty \leq O_{m,d}(Bn^{-1/d})$.

Step 2: Lipschitz-type lower bound for log-partition with $m = 1$. If $m = 1$, it follows that

$$\sup_{f \in \mathcal{F}_{d,m,B}} |L_f - L_{g_n}| \geq e_n(\mathcal{F}_{d,m,B}, S_L, D_{\text{abs}}) \geq \Omega_{m,d}(Bn^{-1/d}) \,.$$

Step 3: Lipschitz-type lower bound for sampling. Take $f(x) = \beta(x_1 + \cdots + x_d)$, where $\beta = B/d$, such that $f \in \mathcal{F}_{d,m,B}$. Pick the cube $\mathcal{X}_1 = (0, 1/N)^d$. Then, we have

$$D_{\text{sup-log}}(P_f, P_{g_n}) = \|\bar{f} - \bar{g}_n\|_\infty \geq \frac{1}{2}\left(\sup_{x \in \mathcal{X}_1} f(x) - \inf_{x \in \mathcal{X}_1} f(x)\right) = \frac{1}{2} B/N = \frac{1}{2} Bn^{-1/d} \,.$$

Step 4: Lower bound for log-partition with $m \geq 2$. As in Step 3, take $f(x) = \beta(x_1 + \cdots + x_d)$, where $\beta = B/d$, such that $f \in \mathcal{F}_{d,m,B}$. To prove a lower bound on $L_f - L_{g_n}$, we follow Lemma E.2 and lower-bound the errors $L_f(\mathcal{X}_i) - L_{g_n}(\mathcal{X}_i)$ on individual subcubes $\mathcal{X}_i$.

Step 4.1: First lower bound. Fix a subcube $\mathcal{X}_i$ and set $a = 1/N$ and $t \coloneqq \frac{\partial f}{\partial x_k}(x^{(i)}) = (\beta, \ldots, \beta)$. Denote the volume of $\mathcal{X}_i$ by $V_n = 1/n = a^d$. Since g_n is constant on $\mathcal{X}_i$, we have

$$\begin{aligned} L_f(\mathcal{X}_i) - L_{g_n}(\mathcal{X}_i) &= \log\left(\exp(f(x^{(i)})) \int_{\mathcal{X}_i} \exp(\langle t, x - x^{(i)} \rangle) \,\mathrm{d}x\right) - \log\left(V_n \exp(f(x^{(i)}))\right) \\ &\overset{\text{Lemma E.1}}{=} d\log(a) + \left(\sum_{k=1}^d r(at_k)\right) - \log(V_n) = \sum_{k=1}^d r(at_k) = dr(\beta/N) \\ &\overset{\text{Lemma E.4}}{\geq} cd \min\{|\beta/N|, |\beta/N|^2\} \geq \Omega_{m,d}(\min\{Bn^{-1/d}, B^2 n^{-2/d}\}) \,, \end{aligned}$$

This lower bound is independent of i, hence by Lemma E.2, we obtain

$$L_f - L_{g_n} \geq \Omega_{m,d}(\min\{Bn^{-1/d}, B^2 n^{-2/d}\}) \,.$$

Step 4.2: Second lower bound. The lower bound above implicitly uses the strong convexity of $\exp(f)$. However, all of the curvature in $\exp(f)$ comes from $\exp$ and none from f. This is not sufficient in the case $B \ll 1$, where the quadratic dependency on B in $B^2 n^{-2/d}$ can be improved. For the case $B < 1$, we put the curvature into f by setting $f(x) \coloneqq \beta \sum_{k=1}^d x_k^2$. We then have $\partial_k f(x) = 2\beta x_k$ and $\partial_k^2 f(x) = 2\beta$. Hence, we set $\beta \coloneqq B/(2d)$ to ensure that $f \in \mathcal{F}_{d,m,B}$. We now again consider a subcube $\mathcal{X}_i$, for which we compute

$$\begin{aligned} \int_{\mathcal{X}_i} \exp(g_n(x)) \,\mathrm{d}x &= V_n \exp(f(x^{(i)})) \,, \\ \int_{\mathcal{X}_i} \exp(f(x)) \,\mathrm{d}x &= \exp(f(x^{(i)})) \int_{\mathcal{X}_i} \exp(f(x) - f(x^{(i)})) \,\mathrm{d}x \\ &\geq \exp(f(x^{(i)})) \int_{\mathcal{X}_i} \left(1 + (f(x) - f(x^{(i)}))\right) \,\mathrm{d}x \\ &= \exp(f(x^{(i)})) \left(V_n + \int_{\mathcal{X}_i} \left(\sum_{k=1}^d \beta(x_k - x_k^{(i)})^2\right) \,\mathrm{d}x\right) \end{aligned}$$

$$= \exp(f(x^{(i)})) \left(V_n + da^{d-1}\beta \left[\frac{1}{3}u^3 \right]_{-a/2}^{a/2} \right)$$
$$= \exp(f(x^{(i)}))V_n \left(1 + d\frac{2}{3 \cdot 2^3}\beta a^2 \right) .$$

Hence, we have

$$L_f(\mathcal{X}_i) - L_g(\mathcal{X}_i) \geq \log\left(1 + \frac{d}{12}\beta n^{-2/d}\right) = \log\left(1 + \frac{1}{24}Bn^{-2/d}\right) \geq \frac{1}{48}Bn^{-2/d} ,$$

where we used $\log'(x) = 1/x \geq 1/2$ for $x \in [1, 2]$ in the last step.

Step 5: Better upper bound for log-partition for $m \geq 2$**.** Let $m \geq 2$, $f \in \mathcal{F}_{d,m,B}$ and $Bn^{-1/d} \leq 1$. We define the piece-wise first-order approximant h_n, where for x in the interior of $\mathcal{X}_i$, we set

$$h_n(x) = f(x^{(i)}) + \langle \nabla f(x^{(i)}), x - x^{(i)} \rangle .$$

Our goal is to use

$$|L_f - L_{g_n}| \leq |L_f - L_{h_n}| + |L_{h_n} - L_g| .$$

Step 5.1: Bounding the first term. To bound the first term, we will bound $\|f - h_n\|_\infty$. Let $\delta(t) = f(x^{(i)} + t(x - x^{(i)})) - h_n(x^{(i)} + t(x - x^{(i)}))$. Then, we use Taylor's theorem to bound

$$f(x) - h_n(x) = \delta(1) = \delta(0) + 1 \cdot \delta'(0) + \frac{1^2}{2} \cdot \delta''(\xi) ,$$

where $\xi \in (0, 1)$. Since h_n is constructed such that $\delta(0) = \delta'(0) = 0$, we have

$$|f(x) - h_n(x)| = \frac{1}{2}|\delta''(\xi)| = \frac{1}{2}(x - x^{(i)})^\top [\nabla^2 f(x^{(i)} + \xi(x - x^{(i)}))](x - x^{(i)})$$
$$\leq d^2 \|x - x^{(i)}\|_\infty^2 \|f\|_{C^2} \leq \frac{d^2 B}{(2N)^2} = O_{m,d}(BN^{-2}) = O_{m,d}(Bn^{-2/d}) .$$

This shows $\|f - h_n\|_\infty \leq O_{m,d}(Bn^{-2/d})$ and therefore $|L_f - L_{h_n}| \leq O_{m,d}(Bn^{-2/d})$.

Step 5.2: Bounding the second term. To bound $|L_{h_n} - L_{g_n}|$, we follow Lemma E.2 and bound the errors $|L_{h_n}(\mathcal{X}_i) - L_{g_n}(\mathcal{X}_i)|$ on individual subcubes $\mathcal{X}_i$. Fix a subcube $\mathcal{X}_i$ and set $a = 1/N$ and $t \coloneqq \frac{\partial f}{\partial x_k}(x^{(i)})$. Denote the volume of $\mathcal{X}_i$ by $V_n = 1/n = a^d$. Since g_n is constant on $\mathcal{X}_i$, we have

$$|L_{h_n}(\mathcal{X}_i) - L_{g_n}(\mathcal{X}_i)| = \left| \log\left(\exp(f(x^{(i)})) \int_{\mathcal{X}_i} \exp(\langle t, x - x^{(i)} \rangle) \, \mathrm{d}x \right) - \log\left(V_n \exp(f(x^{(i)})) \right) \right|$$
$$\overset{\text{Lemma E.1}}{=} \left| d\log(a) + \left(\sum_{k=1}^d r(at_k) \right) - \log(V_n) \right| = \left| \sum_{k=1}^d r(at_k) \right|$$
$$\overset{\text{Lemma E.4}}{\leq} \sum_{k=1}^d C \min\{|at_k|, |at_k|^2\} \leq Cd \min\{Bn^{-1/d}, B^2 n^{-2/d}\} ,$$

where we used $|t_k| \leq \|f\|_{C^m}$ and $a = 1/N = n^{-1/d}$ in the last step. Using Lemma E.2, we now obtain

$$|L_{h_n} - L_{g_n}| \leq Cd \min\{Bn^{-1/d}, B^2 n^{-2/d}\} ,$$

which concludes the upper bound. □

The following two lemmas provide some additional bounds that have been used in the previous proof:

Lemma E.3. *Let*

$$H : \mathbb{R} \to \mathbb{R}, t \mapsto \sum_{k=0}^\infty \frac{t^k}{(k+1)!} = \begin{cases} \frac{\exp(t)-1}{t} & , t \neq 0 \\ 1 & , t = 0 . \end{cases}$$
$$h : \mathbb{R} \to \mathbb{R}, t \mapsto \sum_{k=0}^\infty \frac{t^k}{(k+2)!} = \begin{cases} \frac{H(t)-1}{t} & , t \neq 0 \\ 1/2 & , t = 0 . \end{cases}$$

Then, H and h are C^∞. Moreover,

- *H and H' are increasing on $[0,\infty)$ with $H(0)=1$, $H'(0)=1/2$, and $H(t)>0$ for all $t\in(0,\infty)$.*
- *h and h' are increasing on $[0,\infty)$ with $h(0)=1/2$, $h'(0)=1/6$.*

Proof. Using the series representation, it follows that H and h are C^∞. We also directly obtain $H(0)=1$ and $H'(0)=1/2$ as well as $h(0)=1/2$ and $h'(0)=1/6$. Moreover, it follows that $H', H''\geq 0$ on $[0,\infty]$, which implies that H' and H are increasing on $[0,\infty)$. The inequality $H(t)>0$ follows from the non-series representation together with $H(0)=1$. The results for h can be derived analogously. □

Lemma E.4. *Consider the function r from Lemma E.1. Then, there exist constants $c, C>0$ such that*

$$c\min\{|t|,|t|^2\}\leq r(t)\leq C\min\{|t|,|t|^2\}$$

for all $t\in\mathbb{R}$.

Proof. Since r is an even function, it suffices to prove the inequalities for $t\geq 0$.

Step 1: Simplifying the derivative. First, we compute the derivative of r for $t\neq 0$:

$$\begin{aligned}
r'(t) &= \frac{t}{\exp(t/2)-\exp(-t/2)}\cdot\left(\frac{\exp(t/2)+\exp(-t/2)}{2t}-\frac{\exp(t/2)-\exp(-t/2)}{t^2}\right)\\
&= \frac{1}{2}\frac{\exp(t)+1}{\exp(t)-1}-\frac{1}{t}=\frac{1}{2}+\frac{1}{\exp(t)-1}-\frac{1}{t}\\
&= \frac{1}{2}+\frac{1}{t}\left(\frac{1}{\left(\frac{\exp(t)-1}{t}\right)}-1\right)=\frac{1}{2}+\frac{1}{t}\left(\frac{1}{H(t)}-1\right)=\frac{1}{2}+\frac{1}{t}\cdot\frac{1-H(t)}{H(t)}=\frac{1}{2}-\frac{h(t)}{H(t)}\,,
\end{aligned}\tag{15}$$

where we used the functions H and h from Lemma E.3. Since h and H are also continuous in $t=0$, the equation

$$r'(t)=\frac{1}{2}-\frac{h(t)}{H(t)}$$

holds for all $t\in\mathbb{R}$.

Step 2: Upper bound. Since r is $1/2$-Lipschitz, we obtain $r(t)=r(t)-r(0)\leq t/2$ for $t\geq 0$. For $t\in[0,1]$, we can use $r'(0)=0$ to obtain

$$r(t)=r(0)+tr'(0)+\frac{t^2}{2}r''(\xi)\leq\frac{t^2}{2}\sup_{u\in[0,1]}r''(u)\leq Ct^2$$

for some $\xi\in[0,t]$ and the constant $C=\frac{1}{2}\sup_{u\in[0,1]}r''(u)>0$. This shows $r(t)\leq O(\min\{t,t^2\})$.

Step 3: Lower bound. We can now use Lemma E.3 to further simplify for $t\geq 0$

$$r'(t)=\frac{1}{2}-\frac{h(t)}{H(t)}\geq\frac{1}{2}-\frac{1/2}{H(t)}=\frac{1}{2}\left(1-\frac{1}{H(t)}\right)=:\tilde{r}(t)\,.$$

We find that

$$\tilde{r}'(t)=\frac{H'(t)}{2H(t)^2}$$

satisfies $\tilde{r}'(0)=1/4$ and $\tilde{r}'(t)>0$ for all $t\in[0,\infty)$.

Set $\tilde{c}\coloneqq\inf_{t\in[0,1]}\tilde{r}'(t)>0$. Since $H(0)=1$, we have $\tilde{r}(0)=0$. For $t\in[0,1]$, this yields

$$\tilde{r}(t)=r(0)+\int_0^t\tilde{r}(u)\,\mathrm{d}u\geq\tilde{c}t\,.$$

For $t>1$, since $\tilde{r}$ is increasing, we obtain $\tilde{r}(t)\geq\tilde{c}$. In total, this yields

$$\tilde{r}(t)\geq\tilde{c}\min\{1,t\}\qquad\text{for } t\in[0,\infty)\,.$$

Now, we obtain for $t\in[0,1]$

$$r(t)=r(0)+\int_0^t r'(u)\,\mathrm{d}u\geq\int_0^t\tilde{r}(u)\,\mathrm{d}u\geq\int_0^t\tilde{c}\min\{1,u\}\,\mathrm{d}u=\frac{\tilde{c}}{2}t^2$$

and for $t>1$

$$r(t)=r(1)+\int_1^t r'(u)\,\mathrm{d}u\geq r(1)+\int_1^t\tilde{r}(u)\,\mathrm{d}u\geq r(1)+\int_1^t\tilde{c}\min\{1,u\}\,\mathrm{d}u\geq\frac{\tilde{c}}{2}+\tilde{c}(t-1)\geq\frac{\tilde{c}}{2}t\,.$$ □

E.1.2 Proofs for Density-based Approximation

To prove bounds for density-based approximations, the following lemma will be helpful:

Lemma E.5. *The function*

$$f : [-1/2, 1/2] \to \mathbb{R}, x \mapsto 1 - \frac{1}{1+x}$$

satisfies

$$|f(x)| \leq 4|x| \,.$$

Proof. We have $f(0) = 0$ and $|f'(x)| = |(1+x)^{-2}| \leq (1/2)^{-2} = 4$. □

Now, we can analyze how log-partition and sampling errors can be bounded in terms of the underlying unnormalized densities:

Proposition 22 (Density approximation bounds)**.** *Let $p, q : \mathbb{R} \to [0, \infty)$ be bounded and measurable such that $I_p, I_q > 0$, where $I_p \coloneqq \int_{\mathcal{X}} p(x) \,\mathrm{d}x$. Define probability distributions P, Q with densities p/I_p and q/I_q, respectively. Then,*

$$|\log I_p - \log I_q| \leq \log\left(\frac{1}{1 - \|p-q\|_\infty / I_p}\right) \quad \text{if } \|p-q\|_\infty < I_p,$$
$$D_{\mathrm{TV}}(P, Q) \leq \frac{3\|p-q\|_\infty}{\max\{I_p, I_q\}} \leq \frac{3\|p-q\|_\infty}{I_p} \,.$$

Proof. **Step 1: Log-partition function.** We have

$$\log I_p - \log I_q \leq \log I_p - \log\left(\int_{\mathcal{X}} (p(x) - \|p-q\|_\infty) \,\mathrm{d}x\right) = \log\left(\frac{I_p}{I_p - \|p-q\|_\infty}\right) = \log\left(\frac{1}{1 - \|p-q\|_\infty / I_p}\right)$$
$$\log I_q - \log I_p \leq \log\left(\int_{\mathcal{X}} (p(x) + \|p-q\|_\infty) \,\mathrm{d}x\right) - \log I_p = \log\left(\frac{I_p + \|p-q\|_\infty}{I_p}\right) = \log(1 + \|p-q\|_\infty / I_p)$$
$$\leq \log\left(\frac{1}{1 - \|p-q\|_\infty / I_p}\right).$$

Step 2: Total variation distance. Without loss of generality, assume that $I_q \leq I_p = 1$. Define the normalized density function $\tilde{q} \coloneqq q / I_q$. We consider two cases:

- Suppose $\|p-q\|_\infty \geq 1/2$. Then,

$$D_{\mathrm{TV}}(P, Q) \leq 1 \leq 2\|p-q\|_\infty \,.$$

- Suppose $\|p-q\|_\infty \leq 1/2$. Then,

$$D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - \tilde{q}(x)| \,\mathrm{d}x \leq \frac{1}{2}\int_{\mathcal{X}} (|p(x) - q(x)| + |q(x)(1 - 1/I_q)|) \,\mathrm{d}x$$
$$\leq \frac{1}{2}\|p-q\|_\infty + \frac{1}{2}\left|1 - \frac{1}{1 + (I_q - 1)}\right| \,.$$

  To bound the second term, we first compute

$$|1 - I_q| = \left|\int_{\mathcal{X}} (p(x) - q(x)) \,\mathrm{d}x\right| \leq \int_{\mathcal{X}} |p(x) - q(x)| \,\mathrm{d}x \leq \|p-q\|_\infty \,.$$

  Since we are in the case where $\|p-q\|_\infty \leq 1/2$, we can apply Lemma E.5 to obtain

$$\left|1 - \frac{1}{1 + (I_q - 1)}\right| \leq 4|I_q - 1| \leq 4\|p-q\|_\infty \,,$$

  which completes the proof. □

Next, we turn to bounding $\|p_f\|$ in terms of $\|f\|$. Our first result will provide a bound for the sup-norm:

Lemma E.6. *For $f \in C^1(\mathcal{X})$, we have*

$$\|p_f\|_\infty \le O_{m,d}(\max\{1, \|f\|_{C^1}\}^d) \,.$$

Proof. By Lemma B.1, we have $|f|_1 \le d^{1/2}\|f\|_{C^1}$, and hence

$$\begin{aligned}\|p_f\|_\infty &= \frac{\exp(M_f)}{Z_f} = \exp(M_f - L_f) \\ &\overset{\text{Lemma 3}}{\le} \exp(d\log(1 + 3d^{-1/2}|f|_1)) = (1 + 3d^{-1/2}|f|_1)^d \le O_{m,d}(\max\{1, \|f\|_{C^1}\}^d) \,.\end{aligned}$$

□

The following lemma helps to bound the norms of products, which occur in the derivatives of $\exp(f)$:

Lemma E.7. *Let $f, g \in C^m(\mathcal{X})$ for $m \ge 0$. Then, $\|fg\|_{C^m} \le 2^m \|f\|_{C^m}\|g\|_{C^m}$.*

Proof. We use induction on m. This claim is obviously true for $m = 0$. Now suppose it is true for m and that $f, g \in C^{m+1}(\mathcal{X})$. Take any $\alpha \in \mathbb{N}_0^d$ with $|\alpha|_1 = m + 1$. Then, we can write $\partial_\alpha = \partial_\beta \partial_j$ for some $j \in \{1, \dots, d\}$ and $\beta \in \mathbb{N}_0^d$ with $|\beta| = m$. We then have

$$\begin{aligned}\|\partial_\alpha(fg)\|_\infty = \|\partial_\beta\partial_j(fg)\|_\infty &= \|\partial_\beta((\partial_j f)g + f(\partial_j g))\|_\infty \\ &\le \|(\partial_j f)g + f(\partial_j g)\|_{C^m} \le 2^m \left(\|\partial_j f\|_{C^m}\|g\|_{C^m} + \|f\|_{C^m}\|\partial_j g\|_{C^m}\right) \\ &\le 2^{m+1}\|f\|_{C^{m+1}}\|g\|_{C^{m+1}} \,.\end{aligned}$$

Moreover, for any $\alpha \in \mathbb{N}_0^d$ with $|\alpha|_1 \le m$, we have

$$\|\partial_\alpha(fg)\|_\infty \le \|fg\|_{C^m} \le 2^m\|f\|_{C^m}\|g\|_{C^m} \le 2^{m+1}\|f\|_{C^{m+1}}\|g\|_{C^{m+1}} \,.$$

This completes the proof of the induction step. □

Now, we can indeed bound higher-order norms of $\exp(f)$:

Lemma E.8. *Let $f \in C^m(\mathcal{X})$ for $m \ge 0$. Then,*

$$\|e^f\|_{C^m} \le 2^{m(m-1)/2}\max\{1, \|f\|_{C^m}\}^m\|e^f\|_\infty \,.$$

Proof. We prove this by induction on m. The claim is obviously true for $m = 0$. Now, suppose the claim is true for some $m \ge 0$ and let $f \in C^{m+1}(\mathcal{X})$. Take any $\alpha \in \mathbb{N}_0^d$ with $|\alpha|_1 = m + 1$. Then, we can write $\partial_\alpha = \partial_\beta\partial_j$ for some $j \in \{1, \dots, d\}$ and $\beta \in \mathbb{N}_0^d$ with $|\beta| = m$. Thus,

$$\begin{aligned}\|\partial_\alpha e^f\|_\infty = \|\partial_\beta\partial_j e^f\|_\infty &= \|\partial_\beta((\partial_j f)e^f\|_\infty \\ &\le \|(\partial_j f)e^f\|_{C^m} \overset{\text{Lemma E.7}}{\le} 2^m\|\partial_j f\|_{C^m}\|e^f\|_{C^m} \le 2^m\|f\|_{C^{m+1}}2^{m(m-1)/2}\max\{1, \|f\|_{C^m}\}^m\|e^f\|_\infty \\ &\le 2^{(m+1)m/2}\max\{1, \|f\|_{C^{m+1}}\}^{m+1}\|e^f\|_\infty \,.\end{aligned}$$

Moreover, for any $\alpha \in \mathbb{N}_0^d$ with $|\alpha|_1 \le m$, we have by the induction hypothesis

$$\|\partial_\alpha e^f\|_\infty \le \|e^f\|_{C^m} \le 2^{m(m-1)/2}\max\{1, \|f\|_{C^m}\}^m\|e^f\|_\infty \le 2^{(m+1)m/2}\max\{1, \|f\|_{C^{m+1}}\}^{m+1}\|e^f\|_\infty \,.$$

This completes the proof of the induction step. □

It might be possible to improve the dependence on m in the previous lemma; we ignored this since we do not study the dependence on m. Combining the previous lemmas, we arrive at a higher-order norm bound for the density:

Theorem 23 (Density norm)**.** *For $m \ge 1$, we have*

$$\sup_{f \in \mathcal{F}_{d,m,B}} \|p_f\|_{C^m} = \Theta_{m,d}\left(\max\{1, B\}^{m+d}\right)$$

and this asymptotic rate is attained by $f_{d,m,B}(x) = Bd^{-1}(x_1 + \cdots + x_d)$.

Proof. **Step 1: Upper bound.** For $f \in \mathcal{F}_{d,m,B}$, we have

$$\|p_f\|_{C^m} = \|e^f/Z_f\|_{C^m} = \frac{1}{Z_f}\|e^f\|_{C^m}$$

$$\begin{aligned}
&\overset{\text{Lemma E.8}}{\leq} 2^{m(m-1)/2} \max\{1, \|f\|_{C^m}\}^m \|e^f / Z_f\|_\infty \\
&\overset{\text{Lemma E.6}}{\leq} 2^{m(m-1)/2} \max\{1, \|f\|_{C^m}\}^m O_d(\max\{1, \|f\|_{C^1}\}^d) \\
&\leq O_{m,d}(\max\{1, \|f\|_{C^m}\}^{d+m}) \,.
\end{aligned}$$

Step 2: Lower bound. Consider $f : \mathcal{X} \to \mathbb{R}, x \mapsto Bd^{-1}(x_1 + \cdots + x_d)$. A simple calculation yields $\|f\|_{C^m} = B$, hence $f \in \mathcal{F}_{d,m,B}$. Moreover, we have $M_f = f(1, \ldots, 1) = B$. We can also calculate

$$\begin{aligned}
Z_f &= \int_{\mathcal{X}} \left(\prod_{i=1}^d \exp(Bd^{-1} x_i) \right) \mathrm{d}x = \left(\int_0^1 \exp(Bd^{-1} x_1) \, \mathrm{d}x_1 \right)^d \\
&= (B^{-1} d[\exp(Bd^{-1}) - 1])^d \leq (B^{-1} d \exp(Bd^{-1}))^d = (B^{-1} d)^d \exp(B) \,.
\end{aligned}$$

Finally, we compute

$$\partial_{x_1}^m p_f(x)|_{x=(1,\ldots,1)} = \partial_{x_1}^m \frac{\exp(f(x))}{Z_f} \bigg|_{x=(1,\ldots,1)} = (Bd^{-1})^m \frac{f(1,\ldots,1)}{Z_f} \geq (Bd^{-1})^{d+m} \,.$$

We know that $\sup_{x \in \mathcal{X}} p_f(x) \geq 1$ because p_f is a density on a unit-volume domain. Hence, we conclude $\|p_f\|_{C^m} \geq \Omega_{m,d}(\max\{1, B\}^{d+m})$. □

E.2 Proofs for Simple Stochastic Algorithms

E.2.1 Proofs for Rejection Sampling

The following bound for rejection sampling with uniform proposal distribution is a consequence of the general rejection sampling bounds in Lemma 11.

Proposition 24 (Convergence of rejection sampling)**.** *Let $m \geq 1$ and let $f \in C^1(\mathcal{X})$. Then, the distribution $\tilde{P}_f$ produced by* RejectionSampling*(f, M_f, n) (see Algorithm 1) satisfies*

$$\begin{aligned}
D_{\text{sup-log}}(P_f, \tilde{P}_f) &\leq \min\left\{2\|f\|_\infty, \exp\left(2\|f\|_\infty - n/\|p_f\|_\infty\right)\right\} \\
D_{\text{TV}}(P_f, \tilde{P}_f) &\leq \min\{1, 2\|f\|_\infty\} \exp(-n/\|p_f\|_\infty) \leq O_{m,d}(\min\{1, \|f\|_\infty\} \max\{1, \|f\|_{C^1}\}^m n^{-m/d}) \,.
\end{aligned}$$

Proof. **Step 1: Rejection probability.** Set $g(x) \coloneqq M_f$. Since

$$\|p_f\|_\infty = \frac{\exp(M_f)}{Z_f} = \frac{Z_g}{Z_f} \,,$$

the overall rejection probability p_R from Lemma 11 satisfies

$$p_R \overset{\text{Lemma 11}}{\leq} \exp(-nZ_f/Z_g) = \exp(-n/\|p_f\|_\infty) \,.$$

Step 2: Sup-log distance. We have $D_{\text{sup-log}}(P_f, P_g) = \|\bar{f} - \bar{g}\|_\infty = \|\bar{f}\|_\infty \leq 2\|f\|_\infty$. We then obtain from Lemma 11 that

$$\begin{aligned}
D_{\text{sup-log}}(P_f, \tilde{P}_f) &\leq \min\{D_{\text{sup-log}}(P_f, P_g), p_R(\exp(D_{\text{sup-log}}(P_f, P_g)) - 1)\} \\
&\leq \min\left\{2\|f\|_\infty, \exp\left(2\|f\|_\infty - n/\|p_f\|_\infty\right)\right\} \,.
\end{aligned}$$

Step 3: TV distance. For the TV distance, we compute $D_{\text{TV}}(P_f, P_g) \leq D_{\text{sup-log}}(P_f, P_g) \leq 2\|f\|_\infty$ and also employ the trivial bound $D_{\text{TV}}(P_f, P_g) \leq 1$. From Lemma 11, we obtain

$$D_{\text{TV}}(P_f, \tilde{P}_f) = p_R D_{\text{TV}}(P_f, P_g) \leq \exp(-n/\|p_f\|_\infty) \min\{1, 2\|f\|_\infty\} \,.$$

Using that $\exp(-x) \leq O_{m,d}(x^{-m/d})$ for $x > 0$, we obtain

$$\exp(-n/\|p_f\|_\infty) \leq \|p_f\|_\infty^{m/d} n^{-m/d} \overset{\text{Theorem 23}}{\leq} O_{m,d}\left(\max\{1, \|f\|_{C^1}\}^m n^{-m/d}\right) \,.$$ □

E.2.2 Proofs for Monte Carlo Log-partition

For analyzing the Monte Carlo log-partition estimator, we are going to use Bernstein's inequality in the form stated and proved in Theorem 6.10 in Steinwart and Christmann (2008):

Theorem E.9 (Bernstein's inequality)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $B > 0, \sigma > 0, n \geq 1$. Moreover, let $\xi_1, \dots, \xi_n : \Omega \to \mathbb{R}$ be independent random variables with*

- $\mathbb{E}_P \xi_i = 0$
- $\|\xi_i\|_\infty \leq B$
- $\mathbb{E}_P \xi_i^2 \leq \sigma^2$

for all $i \in \{1, \dots, n\}$. Then,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}$$

for all $\tau > 0$.

We will use Bernstein's inequality since it yields better bounds than Hoeffding's inequality (Theorem D.1) when n is large and σ is significantly smaller than B, which will be the case in the following proof.

Theorem 25 (Upper bounds for MC log-partition)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be Lipschitz, let $X_1, \dots, X_n \sim \mathcal{U}(\mathcal{X})$ be independent and let*

$$\tilde{L}_n \coloneqq \log S_n, \qquad S_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \exp(f(X_i)).$$

Then, for any $\delta \in (0, 1]$, the following convergence rates hold:

(a) ***Optimization regime:*** *If $n \leq 4\log(2/\delta)(1 + 3d^{-1/2}|f|_1)^d$, we have*

$$|\tilde{L}_n - L_f| \leq d^{1/2}(\log(1/\delta))^{1/d}|f|_1 n^{-1/d} + \log(4\log(2/\delta)) + d\log(1 + 3d^{-1/2}|f|_1)$$

with probability $\geq 1 - \delta$.

(b) ***Quadrature regime:*** *If $n \geq 4\log(2/\delta)(1 + 3d^{-1/2}|f|_1)^d$, we have*

$$|\tilde{L}_n - L_f| \leq 4\log(2/\delta)^{1/2}(1 + 3d^{-1/2}|f|_1)^{d/2} n^{-1/2}$$

with probability $\geq 1 - \delta$.

Proof. Without loss of generality, we assume that f is shifted such that $M_f = 0$.

(a) **Step A.1: Simple one-sided bound.** We have

$$\tilde{L}_n - L_f \leq 0 - L_f \overset{\text{Lemma 3}}{\leq} d\log(1 + 3d^{-1/2}|f|_1)\,.$$

Step A.2: Bounding the other side. Define the empirical maximum

$$M_n \coloneqq \max_{i \in \{1,\dots,n\}} f(X_i)\,.$$

Since $\tilde{L}_n = \log(\sum_{i=1}^n \exp(f(X_i))) - \log(n) \geq M_n - \log(n)$, we obtain

$$L_f - \tilde{L}_n \leq 0 - \tilde{L}_n \leq \log(n) - M_n \leq \log(4\log(2/\delta)) + d\log(1 + 3d^{-1/2}|f|_1) - M_n\,.$$

It remains to provide a lower bound on M_n. Define $R \coloneqq (\log(1/\delta))^{1/d} n^{-1/d}$.

- **Case 1:** $R \geq 1$**.** In this case, for all $x \in \mathcal{X}$, we have

$$f(x) = f(x) - M_f \geq -d^{1/2}|f|_1 \geq -d^{1/2}|f|_1 R = -d^{1/2}(\log(1/\delta))^{1/d}|f|_1 n^{-1/d}\,,$$

which implies that

$$-M_n \leq d^{1/2}(\log(1/\delta))^{1/d}|f|_1 n^{-1/d}$$

with probability 1.

- **Case 2:** $R \leq 1$**.** Let x^* be a maximizer of f. In this case, there exists an axis-aligned subcube $\tilde{\mathcal{X}}$ of $\mathcal{X}$ with side length R containing x^*. For each $x \in \tilde{\mathcal{X}}$, we have $f(x) = f(x) - f(x^*) \geq -|f|_1 d^{1/2} R$. Moreover,

$$\begin{aligned} P(M_n \leq -|f|_1 d^{1/2} R) &= P(f(X_1) \leq -|f|_1 d^{1/2} R)^n \\ &\leq P(X_1 \notin \tilde{\mathcal{X}})^n = (1 - P(X_1 \in \tilde{\mathcal{X}}))^n = (1 - R^d)^n \\ &\leq e^{-R^d n} = \delta \,. \end{aligned}$$

(b) Define $\tau \coloneqq \log(2/\delta)$. To apply Bernstein's inequality to $\xi_i \coloneqq e^{f(X_i)} - e^{L_f}$, we need a sup-bound and a variance bound.
Step B.1: Sup-bound. Since $e^{f(\mathcal{X})} \subseteq [0,1]$ by assumption on f, we have $\|\xi_i\|_\infty \leq 1$.
Step B.2: Variance bound. Since $e^{f(\tilde{\mathcal{X}})} \subseteq [0,1]$, we have:

$$\operatorname{Var}(\xi_i) = \operatorname{Var} e^{f(X_i)} \leq \mathbb{E}[(e^{f(X_i)})^2] \leq \mathbb{E} e^{f(X_i)} = e^{L_f} \,.$$

Step B.3: Non-logarithmic concentration. By Bernstein's inequality, we obtain for $\tau \in (0,1)$:

$$P\left(|S_n - e^{L_f}| \geq \sqrt{\frac{2e^{L_f}\tau}{n}} + \frac{2\tau}{3n}\right) \leq 2e^{-\tau} = \delta \,.$$

Step B.4: Lower bound on expectation. From Lemma 3, we obtain

$$e^{L_f} = e^{L_f - M_f} \geq e^{-d\log(1+3d^{-1/2}|f|_1)} = (1 + 3d^{-1/2}|f|_1)^{-d} \,.$$

Step B.5: Logarithmic concentration. By combining the previous steps, we obtain with probability $\geq 1 - 2e^{-\tau}$:

$$\left|\frac{S_n - e^{L_f}}{e^{L_f}}\right| \leq \sqrt{\frac{2\tau}{n(1+3d^{-1/2}|f|_1)^{-d}}} + \frac{2\tau}{3n(1+3d^{-1/2}|f|_1)^{-d}} \,.$$

Since we assumed $n \geq 4\log(2/\delta)(1+3d^{-1/2}|f|_1)^d = 4\tau(1+3d^{-1/2}|f|_1)^d$, the right-hand-side is less than $1/2$. Since the logarithm is 2-Lipschitz on $[1/2, 3/2]$, we obtain

$$\begin{aligned} |\tilde{L}_n - L_f| &= \left|\log\left(\frac{S_n}{e^{L_f}}\right)\right| \\ &= \left|\log\left(1 + \frac{S_n - e^{L_f}}{e^{L_f}}\right) - \log(1)\right| \\ &\leq 2\left|\frac{S_n - e^{L_f}}{e^{L_f}}\right| \\ &\leq 2\left(\sqrt{\frac{2\tau}{n(1+3d^{-1/2}|f|_1)^{-d}}} + \frac{2\tau}{3n(1+3d^{-1/2}|f|_1)^{-d}}\right) \end{aligned}$$

By using the assumption $n \geq 4\log(2/\delta)(1+3d^{-1/2}|f|_1)^d = 4\tau(1+3d^{-1/2}|f|_1)^d$ from (b), we can further bound

$$\begin{aligned} |\tilde{L}_n - L_f| &\leq 2\left(\sqrt{\frac{2\tau}{n(1+3d^{-1/2}|f|_1)^{-d}}} + \frac{2\tau}{3n^{1/2}(4\tau(1+3d^{-1/2}|f|_1)^d)^{1/2}(1+3d^{-1/2}|f|_1)^{-d}}\right) \\ &= (\sqrt{8} + 2/3)\sqrt{\frac{\tau}{n(1+3d^{-1/2}|f|_1)^{-d}}} \\ &\leq 4(1+3d^{-1/2}|f|_1)^{d/2}\tau^{1/2}n^{-1/2} \,. \end{aligned}$$

□

E.2.3 Proofs for Monte Carlo Sampling

We now prove a simple lower bound for a simple Monte Carlo sampling algorithm. We use the TV distance, but the general approach could also be used to prove lower bounds for the sup-log and 1-Wasserstein distances.

Theorem 26 (Lower bound for MC sampling)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be bounded and measurable. Let $X_1, \ldots, X_n \sim \mathcal{U}(\mathcal{X})$ and let the random index $I \in \{1, \ldots, n\}$ be distributed as*

$$P(I = i) = \frac{\exp(f(X_i))}{\sum_{j=1}^n \exp(f(X_j))}.$$

Consider the distribution $\tilde{P}_f$ of the random sample X_I. Then, for all $B > 0$ and $n \geq 1$ with $Bn^{-1/d} \geq 4d\log(4d)$,

$$\sup_{f \in \mathcal{F}_{d,m,B}} D_{\mathrm{TV}}(P_f, \tilde{P}_f) \geq \frac{1}{2} .$$

Proof. Set $f(x) \coloneqq -\beta(x_1 + \cdots + x_d)$, where $\beta \coloneqq Bd^{-1}$, such that $f \in \mathcal{F}_{d,m,B}$. For $\delta \in (0,1]$, set $\mathcal{X}_\delta \coloneqq [0,\delta]^d$. Then, similar as in Lemma E.1, we can compute

$$Z(\delta) \coloneqq \int_{\mathcal{X}_\delta} \exp(f(x))\,\mathrm{d}x = \left(\frac{1-\exp(-\beta\delta)}{\beta}\right)^d .$$

We then obtain

$$P_f(\mathcal{X}_\delta) = \frac{Z(\delta)}{Z(1)} = \left(\frac{1-\exp(-\beta\delta)}{1-\exp(-\beta)}\right)^d \geq (1-\exp(-\beta\delta))^d \geq 1 - d\exp(-\beta\delta) ,$$

where we used Bernoulli's inequality in the last step. Setting $\delta \coloneqq \log(4d)/\beta$, we obtain

$$P_f(\mathcal{X}_\delta) \geq 1 - d\exp(-\beta\delta) \geq \frac{3}{4} .$$

On the other hand, we have

$$\tilde{P}_f(\mathcal{X}_\delta) = P(X_I \in \mathcal{X}_\delta) \leq \sum_{j=1}^n P(X_j \in \mathcal{X}_\delta) = n\delta^d = n\frac{(\log(4d))^d}{\beta^d} = n\frac{(d\log(4d))^d}{B^d} .$$

Hence, if $Bn^{-1/d} \geq 4d\log(4d)$, then $\tilde{P}_f(\mathcal{X}_\delta) \leq 4^{-d}$, which implies

$$D_{\mathrm{TV}}(P_f, \tilde{P}_f) \geq P_f(\mathcal{X}_\delta) - \tilde{P}_f(\mathcal{X}_\delta) \geq \frac{3}{4} - 4^{-d} \geq \frac{1}{2} .$$ □

E.3 Proofs for Variational Formulation

The following simple lemma is central to our lower bound for the variational formulation:

Lemma 27. *For a model of the form $g(x) = \varphi(x)^* H \varphi(x)$ as above, we have*

$$L_g^{\mathrm{OPT}}(Q) = \sup_{\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{Q}} = \Sigma_Q} L_g(\tilde{Q}) .$$

Proof. We have

$$\begin{aligned}
L_g^{\mathrm{OPT}}(Q) &= \sup_{P \in \mathcal{P}(\mathcal{X})} \mathrm{tr}[H\Sigma_P] - \inf_{\tilde{P},\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{P}} = \Sigma_P, \Sigma_{\tilde{Q}} = \Sigma_Q} D_{\mathrm{KL}}(\tilde{P} \parallel \tilde{Q}) \\
&= \sup_{\Sigma \in \mathcal{K}} \mathrm{tr}[H\Sigma] - \inf_{\tilde{P},\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{P}} = \Sigma, \Sigma_{\tilde{Q}} = \Sigma_Q} D_{\mathrm{KL}}(\tilde{P} \parallel \tilde{Q}) \\
&= \sup_{\Sigma \in \mathcal{K}} \sup_{\tilde{P},\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{P}} = \Sigma, \Sigma_{\tilde{Q}} = \Sigma_Q} \mathrm{tr}[H\Sigma] - D_{\mathrm{KL}}(\tilde{P} \parallel \tilde{Q}) \\
&= \sup_{\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{Q}} = \Sigma_Q} \sup_{\tilde{P} \in \mathcal{P}(\mathcal{X})} \mathrm{tr}[H\Sigma_P] - D_{\mathrm{KL}}(\tilde{P} \parallel \tilde{Q}) \\
&= \sup_{\tilde{Q} \in \mathcal{P}(\mathcal{X}):\Sigma_{\tilde{Q}} = \Sigma_Q} L_g(Q).
\end{aligned}$$ □

Before proving the lower bound, we prove a Taylor-based bound on the cosine function, which is then used in the subsequent lemma to bound an integral of the form $\int_{\mathcal{X}} \exp(\beta\cos(x-z))\,\mathrm{d}x$.

Lemma E.10. *For all $x \in [-1/4, 1/4]$, we have*

$$\cos(2\pi x) \leq 1 - \frac{8}{5}\pi^2 x^2 .$$

Proof. For $f(x) \coloneqq \cos(2\pi x)$, we have

$$\begin{aligned} f'(x) &= -2\pi \sin(2\pi x) \\ f''(x) &= -4\pi^2 \cos(2\pi x) \\ f'''(x) &= 8\pi^3 \sin(2\pi x) \\ f^{(4)}(x) &= 16\pi^4 \cos(2\pi x) \end{aligned}$$

Applying Taylor's theorem with the Lagrange form of the remainder for $x \in [-1/4, 1/4]$ around $x_0 = 0$, we obtain for some $\xi \in [-1/4, 1/4]$:

$$f(x) = 1 - 2\pi^2 x^2 + \frac{2}{3}\pi^4 \cos(2\pi\xi) x^4 \, .$$

Therefore, we obtain for $x \in [-1/4, 1/4]$:

$$f(x) \leq 1 - 2\pi^2 x^2 + \left(\frac{2}{3}\pi^4 (1/4)^2\right) \cdot x^2 \leq 1 - 2\pi^2 (1 - 1/5) x^2 = 1 - \frac{8}{5}\pi^2 x^2 \, . \qquad \square$$

We can now bound the normalizing constant of a rescaled cosine function, which will be used in the lower bound for the variational formulation:

Lemma E.11. *For any $z \in \mathbb{R}^d$, define*

$$g_z : \mathcal{X} \to \mathbb{R}, x \mapsto \sum_{i=1}^{d} \cos(2\pi(x_i - z_i)) \, .$$

Then,

$$Z_{\beta g_z} \leq \beta^{-d/2} e^{\beta d} \, .$$

Proof. Since g_z is 1-periodic, $\exp(\beta g_z)$ is 1-periodic. Moreover, we have $\mathcal{X} = [0,1]^d$ and hence

$$\begin{aligned} Z_{\beta g_z} &= \int_{\mathcal{X}} \exp(\beta g_z(x)) \, \mathrm{d}x = \int_{\mathcal{X}} \exp(\beta g_0(x)) \, \mathrm{d}x \\ &= \int_{\mathcal{X}} \exp(\beta \cos(2\pi x_1)) \cdots \exp(\beta \cos(2\pi x_d)) \, \mathrm{d}x \\ &= \left(\int_0^1 \exp(\beta \cos(2\pi x_1)) \, \mathrm{d}x_1 \right)^d . \end{aligned} \tag{16}$$

From expanding the inequality $(1 - \sqrt{\beta})^2 \geq 0$, we obtain

$$e^\beta \geq 1 + \beta \geq 2\sqrt{\beta} \, . \tag{17}$$

This allows us to upper-bound the one-dimensional integral in Eq. (16) as

$$\begin{aligned} \int_0^1 e^{\beta \cos(2\pi x_1)} \, \mathrm{d}x_1 &= \int_{-1/4}^{3/4} e^{\beta \cos(2\pi x)} \, \mathrm{d}x \\ &\leq \int_{[-1/4,1/4]} e^{\beta \cos(2\pi x)} \, \mathrm{d}x + \int_{[1/4,3/4]} e^{\beta \cos(2\pi x)} \, \mathrm{d}x \\ &\overset{\text{Lemma E.10}}{\leq} \int_{[-1/4,1/4]} e^{\beta(1 - \frac{8}{5}\pi^2 x^2)} \, \mathrm{d}x + \int_{[1/4,3/4]} 1 \, \mathrm{d}x \\ &\leq \frac{1}{2} + e^\beta \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2 \cdot \frac{5}{16}\pi^{-2}\beta^{-1}} \right) \\ &= \frac{1}{2} + e^\beta \sqrt{2\pi \frac{5}{16}\pi^{-2}\beta^{-1}} \\ &= \frac{1}{2} + e^\beta \sqrt{\frac{5\beta^{-1}}{8\pi}} \\ &\overset{\text{Eq. (17)}}{\leq} e^\beta \left(\frac{\sqrt{\beta^{-1}}}{4} + \sqrt{\frac{5\beta^{-1}}{8\pi}} \right) \\ &\leq \beta^{-1/2} e^\beta \, . \end{aligned}$$

$\square$

Finally, we can use some elementary convex geometry to find a lower bound for the error of the variational formulation.

Theorem 28 (Lower bound for OPT relaxation)**.** *Let $\varphi : \mathcal{X} \to \mathbb{C}^N$ be continuous. Let*

$$n \coloneqq \dim_{\mathbb{C}} \mathcal{V}_{\mathrm{lin}}, \qquad \mathcal{V}_{\mathrm{lin}} \coloneqq \mathrm{Span}_{\mathbb{C}} \left\{ \varphi(x)\varphi(x)^* \mid x \in \mathcal{X} \right\} \subseteq \mathbb{C}^{N \times N} .$$

In other words, n is the number of effective degrees-of-freedom of the model $g(x) = \varphi(x)^* H \varphi(x)$, and hence corresponds to the maximum number of points where such a model can interpolate arbitrary function values. Then, there exists a point $z \in \mathcal{X}$ depending only on φ, such that the periodic and analytic function

$$f : \mathcal{X} \to \mathbb{R}, x \mapsto \sum_{i=1}^{d} \cos(2\pi(x_i - z_i))$$

satisfies

$$|L_g^{\mathrm{OPT}}(\mathcal{U}([0,1])) - L_{\beta f}(\mathcal{U}([0,1]))| \geq \log\left(\frac{\beta^{d/2}}{2n+1}\right) - \|g - \beta f\|_\infty \tag{7}$$

for any model $g(x) = \varphi(x)^* H \varphi(x)$ and any $\beta > 0$.

Proof. **Step 1: Representability by discrete distributions.** Let $Q \coloneqq \mathcal{U}([0,1])$. We want to find a discrete distribution $\tilde{Q} = \sum_{k=1}^{M} \lambda^{(k)} \delta_{x^{(k)}}$ with $\Sigma_{\tilde{Q}} = \Sigma_Q$. Using the feature map $\Phi : \mathcal{X} \to \mathbb{C}^{k \times k}, x \mapsto \varphi(x)\varphi(x)^*$, we can write

$$\Sigma_{\tilde{Q}} = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* \, \mathrm{d}\tilde{Q}(x) = \sum_{k=1}^{M} \lambda^{(k)} \Phi(x^{(k)}) ,$$

which shows that the matrices $\Sigma_{\tilde{Q}}$ are exactly those in the convex hull $\mathrm{conv}(\Phi(\mathcal{X}))$ of $\Phi(\mathcal{X})$. Since $\Sigma_Q \in \mathcal{K}$ by definition of $\mathcal{K}$, we first want to show that $\mathcal{K} = \mathrm{conv}(\Phi(\mathcal{X}))$. As we have just demonstrated, the inclusion $\mathcal{K} \supseteq \mathrm{conv}(\Phi(\mathcal{X}))$ is simple. Moreover, because the integral $\Sigma_Q = \int_{\mathcal{X}} \Phi(x) \, \mathrm{d}Q(x)$ is a limit of finite sums, we obtain $\mathcal{K} \subseteq \overline{\mathrm{conv}(\Phi(\mathcal{X}))}$. Since Φ is continuous and $\mathcal{X}$ is compact, $\Phi(\mathcal{X})$ is also compact. Hence, since we are in finite dimension, $\mathrm{conv}(\Phi(\mathcal{X}))$ is compact (see e.g. Proposition 2.3 in Gallier, 2008), which means that $\mathrm{conv}(\Phi(\mathcal{X})) \subseteq \mathcal{K} \subseteq \overline{\mathrm{conv}(\Phi(\mathcal{X}))} = \mathrm{conv}(\Phi(\mathcal{X}))$.

Step 2: Bounding the number of discrete points. By definition, $\mathcal{V}_{\mathrm{lin}}$ is the $\mathbb{C}$-linear span of $\Phi(\mathcal{X})$. Using Step 1, we conclude $\mathcal{K} \subseteq \mathcal{V}_{\mathrm{lin}}$. Hence, $\mathcal{K}$ is contained in the space $\mathcal{V}_{\mathrm{lin}}$ with $\dim_{\mathbb{R}} \mathcal{V}_{\mathrm{lin}} \leq 2n$. By Carathéodory's theorem (see e.g. Theorem 2.2 in Gallier, 2008), the matrix $\Sigma_Q \in \mathcal{K}$ is hence representable as a convex combination of $2n+1$ points:

$$\Sigma_Q = \sum_{k=1}^{2n+1} \lambda^{(k)} \Phi(x^{(k)}),$$

with $\lambda^{(k)} \geq 0, \sum_k \lambda^{(k)} = 1$. By setting $\tilde{Q} \coloneqq \sum_{k=1}^{2n+1} \lambda^{(k)} \delta_{x^{(k)}}$, we obtain $\Sigma_{\tilde{Q}} = \Sigma_Q$.

Step 3: Determining z. Choose an arbitrary index $k^* \in \{1, \ldots, 2n+1\}$ such that $\lambda^{(k^*)} \geq (2n+1)^{-1}$, which always exists. For such an index, we set $z \coloneqq x^{(k^*)}$.

Step 4: Lower-bounding the approximate log-partition function. By Lemma 27, we have

$$\begin{aligned} L_g^{\mathrm{OPT}}(\mathcal{U}([0,1])) &\geq L_g(\tilde{Q}) = \log\left(\sum_{k=1}^{2n+1} \lambda^{(k)} \exp(g(x^{(k)}))\right) \geq \log\left(\lambda^{(k^*)} \exp(g(x^{(k^*)}))\right) \\ &\geq \log\left((2n+1)^{-1} \exp(\beta f(x^{(k^*)}) - \|g - \beta f\|_\infty)\right) = \beta d - \log(2n+1) - \|g - \beta f\|_\infty . \end{aligned}$$

Step 5: Upper-bounding the true log-partition function. We have

$$L_{\beta f}(\mathcal{U}([0,1])) = \log Z_{\beta f} \overset{\text{Lemma E.11}}{\leq} \beta d - \log(\beta^{d/2}) .$$

Step 6: Putting it together. The previous two steps immediately yield the desired bound

$$|L_g^{\mathrm{OPT}}(\mathcal{U}([0,1])) - L_{\beta f}(\mathcal{U}([0,1]))| \geq L_g^{\mathrm{OPT}}(\mathcal{U}([0,1])) - L_{\beta f}(\mathcal{U}([0,1])) \geq \log\left(\frac{\beta^{d/2}}{2n+1}\right) - \|g - \beta f\|_\infty . \qquad \square$$