Proceedings of Machine Learning Research vol 75:1–22, 2018 31st Annual Conference on Learning Theory

A Faster Approximation Algorithm for the Gibbs Partition Function

Vladimir Kolmogorov VNK@IST.AC.AT
IST Austria, Am Campus 1, Klosterneuburg 3400, Austria

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of estimating the partition function $Z(\beta) = \sum_x \exp(-\beta H(x))$ of a Gibbs distribution with a Hamilton $H(\cdot)$, or more precisely the logarithm of the ratio $q = \ln Z(0)/Z(\beta)$. It has been recently shown how to approximate q with high probability assuming the existence of an oracle that produces samples from the Gibbs distribution for a given parameter value in $[0, \beta]$. The current best known approach due to Huber (2015) uses $O(q \ln n \cdot [\ln q + \ln \ln n + \varepsilon^{-2}])$ oracle calls on average where ε is the desired accuracy of approximation and $H(\cdot)$ is assumed to lie in $\{0\} \cup [1, n]$. We improve the complexity to $O(q \ln n \cdot \varepsilon^{-2})$ oracle calls. We also show that the same complexity can be achieved if exact oracles are replaced with approximate sampling oracles that are within $O(\frac{\varepsilon^2}{q \ln n})$ variation distance from exact oracles. Finally, we prove a lower bound of $\Omega(q \cdot \varepsilon^{-2})$ oracle calls under a natural model of computation.

Keywords: Gibbs distribution, partition function, approximation

1. Introduction

It is known that for large classes of problems, e.g. *self-reducible problems* (Jerrum et al., 1986), there is an intimate connection between approximate counting and sampling: the ability to solve one problem allows solving the other one. This paper explores this connection for Gibbs distributions.

Let Ω be some finite set and $H(\cdot)$ be some real-valued function on Ω called a *Hamiltonian*. The *Gibbs distribution* for such a system is a family of distributions $\{\mu_\beta\}$ on Ω parameterized by β, where

$$\mu_\beta(x) = \frac{1}{Z(\beta)} \exp(-\beta H(x)) \qquad \forall x \in \Omega \tag{1}$$

The normalizing constant $Z(\beta)$ is called the *partition function*:

$$Z(\beta) = \sum_{x \in \Omega} \exp(-\beta H(x)) \tag{2}$$

Estimating this function for a given value of β is a widely studied computational problem with applications in many areas. In particular, it is a key computational task in statistical physics. Evaluations of $Z(\cdot)$ yield estimates of important thermodynamical quantities, such as the free energy. Note, parameter β corresponds to the *inverse temperature*. A classical example of a Gibbs distribution in physics is the *Ising model*.

Example 1 *Given an undirected graph (V, E), let $\Omega = \{-1, +1\}^V$ and $H(x) = \sum_{\{i,j\} \in E}[x_i \neq x_j]$ where $[\cdot]$ is 1 if its argument is true, and 0 otherwise. Distribution (1) for such a Hamiltonian is called the* Ising model. *It is* ferromagnetic *if $\beta > 0$, and* antiferromagnetic *if $\beta < 0$ (although* *in the latter case the function $H'(x) = -H(x)$ is usually treated as the Hamiltonian). Computing $Z(\beta)$ exactly is a #P-complete problem, and is even hard to approximate in the antiferromagnetic case (Jerrum and Sinclair, 1993).*

The problem of counting various combinatorial objects such as proper k-coloring and matchings in graphs can also be naturally phrased as estimating the partition function.

Example 2 *Let $\Omega = \{1, \ldots, k\}^{|V|}$ be the set of all colorings in an undirected graph $G = (V, E)$. Define $H(x) = \sum_{\{i,j\} \in E} [x_i = x_j]$, then $Z(+\infty)$ gives the number of proper k-colorings.*

Example 3 *Let Ω be the set of matchings $M \subseteq E$ in an undirected graph $G = (V, E)$. Define $H(M) = |M|$, then $Z(0) = |\Omega|$.*

A related problem is that of *sampling* from the distribution μ_β for a given value of β. There is a vast literature on designing sampling algorithms from Gibbs distributions, see e.g. (Metropolis et al., 1953; Swendsen and Wang, 1986; Huber, 2004; Fill and Huber, 2010) or (Brooks et al., 2011) for an overview. Here we only mention some known results related to Examples 1-3.

(1) *For the ferromagnetic Ising model there exists an exact sampling algorithm that appears to run efficiently at or above the* critical temperature *(Propp and Wilson, 1996). There are also polynomial-time approximate samplers for related models, namely the* Subgraphs-World model *(Jerrum and Sinclair, 1993) and the* Random Cluster model *(Guo and Jerrum, 2018). These are probability distribution over subsets of edges $S \subseteq E$ defined in such as a way that the corresponding partition functions equal the partition function of the ferromagnetic Ising model up to easily computable constants. The result of Guo and Jerrum (2018) has two implications: (i) the Swendsen-Wang dynamics for the ferromagnetic Ising model is rapidly mixing; (ii) applying the Coupling-from-the-Past technique (Propp and Wilson, 1996) to the Random Cluster dynamics (which is monotonic) gives a polynomial-time exact sampler.*

(2) *Approximate sampling of k-colorings in low-degree graphs is addressed in (Jerrum, 1995; Vigoda, 1999) (for $\beta = +\infty$, though techniques are potentially extendable to other values of β).*

(3) *For matchings polynomial-time approximate sampling is presented in (Jerrum and Sinclair, 1989) (for $\beta = +\infty$) and in (Matthews, 2008, Section 2.3.5).*

It is known that the ability to sample can be used for designing a *randomized approximation scheme* for estimating the partition function. By definition, it is an algorithm that for a given $\varepsilon > 0$ produces an estimate $\hat{Q}$ of the desired quantity Q such that $\hat{Q} \in \left[\frac{Q}{1+\varepsilon}, Q(1+\varepsilon)\right]$ with probability at least $3/4$. (The value $3/4$ is arbitrary: by repeating the algorithm multiple times and taking the median of the outputs the probability can be boosted to any other constant in $(0, 1)$). This paper studies the following question: how many samples are needed to approximate $Z(\beta)$ with a given accuracy ε?

Formal description To state the complexity of different approaches, we need to introduce several quantities. First, we assume that $H(x) \in \{0\} \cup [1, n]$ for any $x \in \Omega$ where n is a known number. Non-negativity of the Hamiltonian implies that $Z(\cdot)$ is a decreasing function. Our goal will be to estimate the ratio $Q = Z(\beta_{\min})/Z(\beta_{\max})$ for given values $\beta_{\min} < \beta_{\max}$. Note that computing $Z(\beta)$ for some specific value of β is usually an easy task, so this will allow estimating $Z(\beta)$ for any other β. In particular, in Examples 1, 2 and 3 we have $Z(0) = 2^{|V|}$, $Z(0) = k^{|V|}$ and $Z(+\infty) = 1$ respectively.

Let us denote $q = \log Q$, and assume that there exists an oracle that can produce a sample $X \sim \mu_\beta$ for a given value $\beta \in [\beta_{\min}, \beta_{\max}]$. When stating asymptotic complexities, we will always assume that $q = \Omega(1)$, $n = 1 + \Omega(1)$ and $\varepsilon = O(1)$ to simplify the expressions. Bezáková et al. (2008) showed that Q can be estimated using $O(q^2(\ln n)^2)$ oracle calls in the worst case (for a fixed ε). This was improved to $O(q(\ln q + \ln n)^5\varepsilon^{-2})$ expected number of calls by Štefankovič et al. (2009) and then to $O(q \ln n \cdot [\ln q + \ln\ln n + \varepsilon^{-2}])$ by Huber (2015).[1]

The first contribution of this paper is to improve the complexity further to $O(q \ln n \cdot \varepsilon^{-2})$ oracle calls (on average). This is achieved by a better analysis of the algorithm in (Huber, 2015). The formal statement of our result is given in Section 3 as Theorems 6 and 8.

In many applications we only have an access to approximate sampling oracles. Using a standard coupling argument, in Section 3.1 we show that the same complexity can be achieved with approximate oracles assuming that they are within $O(\frac{\varepsilon^2}{q \ln n})$ variation distance from exact oracles.

As our final contribution, we prove a lower bound of $\Omega(q \cdot \varepsilon^{-2})$ oracle calls under a natural model of computation. The precise statement of the result is given as Theorem 10 in Section 3.2. To our knowledge, this is the first lower bound for this problem. We are only aware of a lower bound for a specific class of algorithms, namely those based on $O(1)$-*Chebyshev cooling schedules*. Such schedule can be used to estimate the desired quantity using $\Theta(\ell^2/\varepsilon^2)$ oracle calls, where ℓ is the length of the schedule. Štefankovič et al. (2009) proved that $\ell = \Omega(q \log n)$ for a *non-adaptive* schedule and $\ell = \Omega(\sqrt{n})$ for an *adaptive* schedule. Note that algorithms in (Huber, 2015) and in the current paper are not based on $O(1)$-Chebyshev schedules, and therefore this lower bound does not apply to them. In contrast, our model covers a very general class of algorithms; the only essential assumption is that the algorithm receives values $H(x)$ from the oracle, and not states x (see Section 3.2).

Remark 1 *The assumption that $H(\cdot)$ lies in $\{0\} \cup [1, n]$ can be relaxed using a standard trick. Suppose, for example, that $H(x) \in \{h_{\min}, h_{\min} + 1, \dots, h_{\max}\}$ where $h_{\min}$ and $h_{\max}$ are known integers. Let $n = h_{\max} - h_{\min}$. We claim that the problem can be solved using $O(q' \ln n \cdot \varepsilon^{-2})$ oracle calls (on average), where either (i) $q' = q - (\beta_{\max} - \beta_{\min}) \cdot h_{\min}$, or (ii) $q' = -q + (\beta_{\max} - \beta_{\min}) \cdot h_{\max}$.*

Indeed, to achieve the first complexity, define new Hamiltonian $H'(x) = H(x) - h_{\min}$. The partition function for the new problem is $Z'(\beta) = e^{\beta h_{\min}} \cdot Z(\beta)$, and so q' is as defined in (i). (We use "primes" to denote all quantities related to the new problem). We have $H'(x) \in \{0, 1, \dots, n\}$, so the algorithm claimed above can be applied to give an estimate of q' and thus of q. Note that distributions μ'_β and μ_β coincide, and so sampling from μ_β allows to sample from μ'_β.

To achieve the second complexity, define $H'(x) = -H(x) + h_{\max}$ and also change the bounds: $\beta'_{\min} = -\beta_{\max}$ and $\beta'_{\max} = -\beta_{\min}$. There holds $Z'(\beta) = e^{-\beta h_{\max}} \cdot Z(-\beta)$, and q' is as defined in (ii). We again have $H'(x) \in \{0, 1, \dots, n\}$, and distributions μ'_β and $\mu_{-\beta}$ coincide. We can now use the same argument as before.

1. The $O(q \ln q)$ algorithm (for fixed n, ε) was first announced in (Huber and Schott, 2011, Section 7) and appeared online in (Huber, 2012). Improving the complexity from $O(q \ln q)$ to $O(q)$ was posed as an open problem in (Huber and Schott, 2011, Section 7).

2. Background and preliminaries

We will assume for simplicity that $H(\cdot) \neq const$. Let us denote $z(\beta) = \ln Z(\beta)$. It can be easily checked that

$$z'(\beta) = \mathbb{E}_{X \sim \mu_\beta}[-H(X)]$$

Since $H(\cdot)$ is non-negative and non-constant, we have $z'(\beta) < 0$ for any β and thus $z(\cdot)$ and $Z(\cdot)$ are strictly decreasing functions. It is also known (Wainwright and Jordan, 2008, Proposition 3.1) that function $z(\cdot)$ is convex for any $H(\cdot)$, and in fact strictly convex if $H(\cdot) \neq const$.

Next, we discuss previous approaches for estimating $Z(\beta_{\min})/Z(\beta_{\max})$, closely following Huber (2015).

It is well-known that for given values β_1, β_2 an unbiased estimator W of $Z(\beta_2)/Z(\beta_1)$ can be obtained as follows: first sample $X \sim \mu_{\beta_1}$ and then set $W = \exp((\beta_1 - \beta_2)H(X))$. Indeed,

$$\mathbb{E}[W] = \sum_{x \in \Omega} \frac{\exp(-\beta_1 H(x))}{Z(\beta_1)} \cdot \exp((\beta_1 - \beta_2)H(x)) = \sum_{x \in \Omega} \frac{\exp(-\beta_2 H(x))}{Z(\beta_1)} = \frac{Z(\beta_2)}{Z(\beta_1)}$$

Applying this estimator directly to $(\beta_1, \beta_2) = (\beta_{\min}, \beta_{\max})$ or to $(\beta_1, \beta_2) = (\beta_{\max}, \beta_{\min})$ is problematic since it usually has a huge relative variance. A standard approach to reduce the relative variance is via the *multistage sampling* method of Valleau and Card (1972). First, a sequence $\beta_{\min} = \beta_0 \leq \beta_1 \leq \ldots \leq \beta_\ell = \beta_{\max}$ is selected; it is called a *cooling schedule*. We then have

$$\frac{Z(\beta_{\min})}{Z(\beta_{\max})} = \frac{Z(\beta_0)}{Z(\beta_1)} \cdot \frac{Z(\beta_1)}{Z(\beta_2)} \cdot \ldots \cdot \frac{Z(\beta_{\ell-1})}{Z(\beta_\ell)}$$

Throughout the paper we refer to $[\beta_i, \beta_{i+1}]$ as "interval i", where $i \in \{0, 1, \ldots, \ell - 1\}$. The ratio $Z(\beta_i)/Z(\beta_{i+1})$ for each such interval can be estimated independently as described above, and then multiplied to give the final estimate. Fishman calls an estimate of this form a *product estimator* (Fishman, 1994). Its mean and variance are given by the lemma below. In this lemma we use the following notation: if X is a random variable then $\mathbb{S}[X] \overset{\text{def}}{=} \frac{\mathbb{E}[X^2]}{(\mathbb{E}[X])^2} = \frac{\texttt{Var}(X)}{(\mathbb{E}[X])^2} + 1$ (the relative variance of X plus 1).

Lemma 2 ((Dyer and Frieze, 1991, page 136)) *For $P = \prod_i P_i$ where the P_i are independent,*

$$\mathbb{E}[P] = \prod_i \mathbb{E}[P_i], \qquad \mathbb{S}[P] = \prod_i \mathbb{S}[P_i]$$

Using a fixed cooling schedule, Bezáková et al. (2008) obtained an approximation algorithm that needs $O(q^2(\ln n)^2)$ samples in the worst case (for a fixed ε). Štefankovič et al. (2009) asymptotically improved this to $10^8 q(\ln q + \ln n)^5 \varepsilon^{-2}$ samples on average. They used an *adaptive* cooling schedule where the values β_i depend on the outputs of sampling oracles. A further improvement to $O(q \ln n \cdot [\ln q + \ln \ln n + \varepsilon^{-2}])$ was given by Huber (2015). One of the key ideas in (Huber, 2015) was to replace the product estimator with the *paired product* estimator, which is described next.

2.1. Paired product estimator

One run of this estimator can be described as follows:

- sample $X_i \sim \mu_{\beta_i}$ for each $i \in [0, \ell]$

- for each interval $i \in [0, \ell - 1]$ compute

$$W_i = \exp(-\tfrac{\beta_{i+1}-\beta_i}{2} H(X_i)), \qquad V_i = \exp(\tfrac{\beta_{i+1}-\beta_i}{2} H(X_{i+1}))$$

- compute $W = \prod_i W_i$ and $V = \prod_i V_i$.

An easy calculation (see (Huber, 2015)) shows that

$$\mathbb{E}[W_i] = \frac{Z(\bar{\beta}_{i,i+1})}{Z(\beta_i)}, \qquad \mathbb{E}[V_i] = \frac{Z(\bar{\beta}_{i,i+1})}{Z(\beta_{i+1})}, \qquad \mathbb{E}[V_i]/\mathbb{E}[W_i] = \frac{Z(\beta_i)}{Z(\beta_{i+1})}, \qquad \mathbb{E}[V]/\mathbb{E}[W] = \frac{Z(\beta_{\min})}{Z(\beta_{\max})}$$

where we denoted $\bar{\beta}_{i,i+1} = \frac{\beta_i+\beta_{i+1}}{2}$. Also,

$$\mathbb{S}[W_i] = \mathbb{S}[V_i] = \frac{Z(\beta_i)Z(\beta_{i+1})}{Z(\bar{\beta}_{i,i+1})^2}, \qquad \mathbb{S}[W] = \mathbb{S}[V] = \prod_i \frac{Z(\beta_i)Z(\beta_{i+1})}{Z(\bar{\beta}_{i,i+1})^2} \tag{3}$$

Although $\mathbb{E}[V]/\mathbb{E}[W] = \frac{Z(\beta_{\min})}{Z(\beta_{\max})} = Q$, using V/W as the estimator of Q would be a poor choice since it is biased in general. Instead, Huber (2015) uses the following procedure.

Algorithm 1 Paired product estimator. **Input**: schedule $(\beta_0, \dots, \beta_\ell)$, integer $r \geq 1$.

1: compute r independent samples of (W, V) as described above
2: take their sample averages $\bar{W}$ and $\bar{V}$ and output $\hat{Q} = \bar{V}/\bar{W}$ as the estimator of Q

The argument from (Huber, 2015) gives the following result.

Lemma 3 *Suppose that*

$$\mathbb{S}[W] = \mathbb{S}[V] \quad \leq \quad 1 + \tfrac{1}{2}\gamma r \tilde{\varepsilon}^2 \tag{4}$$

where $\tilde{\varepsilon} = 1 - (1+\varepsilon)^{-1/2} = \frac{1}{2}\varepsilon + O(\varepsilon^2)$ *and* $\gamma > 0$*. Then* $\mathbb{P}(\hat{Q}/Q \in (\frac{1}{1+\varepsilon}, 1+\varepsilon)) \geq 1 - \gamma$.

Proof We have $\mathbb{E}[\bar{W}] = \mathbb{E}[W]$ and $\mathtt{Var}(\bar{W}) = \frac{1}{r}\mathtt{Var}(W)$, and so $\mathbb{S}[\bar{W}] = \frac{1}{r}(\mathbb{S}[W] - 1) + 1$. By Chebyshev's inequality, $\mathbb{P}(|\bar{W}/\mathbb{E}[\bar{W}] - 1| \geq \tilde{\varepsilon}) \leq (\mathbb{S}[\bar{W}] - 1)/\tilde{\varepsilon}^2 = \frac{1}{r}(\mathbb{S}[W] - 1)/\tilde{\varepsilon}^2 \leq \gamma/2$. Similarly, $\mathbb{P}(|\bar{V}/\mathbb{E}[\bar{V}] - 1| \geq \tilde{\varepsilon}) \leq \gamma/2$.

Denote $S = \bar{W}/\mathbb{E}[\bar{W}]$ and $T = \bar{V}/\mathbb{E}[\bar{V}]$. The union bound gives $\mathbb{P}(\max\{|S-1|, |T-1|\} \geq \tilde{\varepsilon}) \leq \gamma$. Observe that condition $\max\{|S-1|, |T-1|\} < \tilde{\varepsilon}$ implies $\{S, T\} \subset (1-\tilde{\varepsilon}, 1+\tilde{\varepsilon}) \subseteq (\frac{1}{(1+\varepsilon)^{1/2}}, (1+\varepsilon)^{1/2})$ and thus $\frac{\hat{Q}}{Q} = \frac{T}{S} \in (\frac{1}{1+\varepsilon}, 1+\varepsilon)$. The claim follows. ■

Recall that $\mathbb{S}[W] = \mathbb{S}[V]$ is a deterministic function of the schedule $(\beta_0, \dots, \beta_\ell)$ (see eq. (3)). We say that the schedule is *good* (with respect to fixed constants r and γ) if the resulting quantity $\mathbb{S}[W] = \mathbb{S}[V]$ satisfies (4). Huber (2015) presented a randomized algorithm that produces a good schedule with probability at least 0.95 (with respect to $r = \Theta(\varepsilon^{-2})$ and $\gamma = 0.2$). By Lemma 3, the output $\hat{Q}$ of the resulting algorithm lies in $(\frac{Q}{1+\varepsilon}, Q(1+\varepsilon))$ with probability at least $0.95 \cdot (1-\gamma) > 0.75$, as desired.

Huber's algorithm for producing schedule $(\beta_0, \dots, \beta_\ell)$ is reviewed in the next section. It makes $O(q \ln n \cdot [\ln q + \ln \ln n])$ calls to the sampling oracle (on average). Then in Section 3 we will describe how to reduce the number of oracle calls to $O(q \ln n)$ while maintaining the desired guarantees.

2.2. TPA method

The algorithm of Huber (2015) for producing a schedule is based on the *TPA method* of Huber and Schott (2010, 2014). (The abbreviation stands for the "Tootsie Pop Algorithm"). Let us review the application of the method to the Gibbs distribution with a non-negative Hamiltonian $H(\cdot)$.

Its key subroutine is procedure $\texttt{TPAstep}(\beta)$ that for a given constant β produces a random variable in $[\beta, +\infty]$ as follows:

- *sample $X \sim \mu_\beta$, draw $U \in [0,1]$ uniformly at random, return $\beta - \ln U / H(X)$ (or $+\infty$ if $H(X)=0$).*

The motivation for this sampling rule comes from the following fact (which we prove here for completeness).

Lemma 4 *Consider random variable $Y = Z(\texttt{TPAstep}(\beta))$. If $H(\cdot)$ is strictly positive (implying that $Z(+\infty) = 0$) then Y has the uniform distribution on $[0, Z(\beta)]$. If $H(x) = 0$ for some $x \in \Omega$ (implying that $Z(+\infty) > 0$)) then Y has the same distribution as the following random variable Y': sample $Y' \in [0, Z(\beta)]$ uniformly at random and set $Y' \leftarrow \max\{Y', Z(+\infty)\}$.*

Proof It suffices to prove $\mathbb{P}(\texttt{TPAstep}(\beta) \geq \alpha) = Z(\alpha)/Z(\beta)$ for any $\alpha \in [\beta, +\infty)$. We have

$$[\texttt{TPAstep}(\beta) \geq \alpha] = [\ln U / H(X) \leq \beta - \alpha] = [\ln U \leq (\beta - \alpha) H(X)] = [U \leq \exp((\beta - \alpha) H(X))]$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\texttt{TPAstep}(\beta) \geq \alpha) &= \sum_{x \in \Omega} \mathbb{P}(\texttt{TPAstep}(\beta) \geq \alpha | X = x) \mathbb{P}(X = x) \\
&= \sum_{x \in \Omega} \mathbb{P}(U \leq \exp((\beta - \alpha) H(x))) \cdot \frac{\exp(-\beta H(x))}{Z(\beta)} \\
&= \sum_{x \in \Omega} \exp((\beta - \alpha) H(x)) \cdot \frac{\exp(-\beta H(x))}{Z(\beta)} \\
&= \sum_{x \in \Omega} \frac{\exp(-\alpha H(x))}{Z(\beta)} \quad = \quad \frac{Z(\alpha)}{Z(\beta)}
\end{aligned}$$

■

Roughly speaking, the TPA method counts how many steps are needed to get from $\beta_{\min}$ to $\beta_{\max}$.

Algorithm 2 One run of TPA. **Output:** a multiset $\mathcal{B}$ of values in the interval $[\beta_{\min}, \beta_{\max}]$.

set $\beta_0 = \beta_{\min}$, let $\mathcal{B}$ be the empty multiset
for $i = 1$ **to** $+\infty$ **do**
  sample $\beta_i = \texttt{TPAstep}(\beta_{i-1})$
  if $\beta_i \in [\beta_{\min}, \beta_{\max}]$ then add β_i to $\mathcal{B}$, otherwise output $\mathcal{B}$ and terminate
end

The output of Algorithm 2 will be denoted as $\texttt{TPA}(1)$, and the union of k independent runs of $\texttt{TPA}(1)$ as $\texttt{TPA}(k)$. For a multiset $\mathcal{B}$ we define multiset $z(\mathcal{B}) \stackrel{\text{def}}{=} \{z(\beta) \mid \beta \in \mathcal{B}\}$ in a natural way.

(Recall that $z(\cdot)$ is a continuous strictly decreasing function). It is known (Huber and Schott, 2010, 2014) that $z(\mathtt{TPA}(k))$ is a Poisson Point Process (PPP) on $[z(\beta_{\max}), z(\beta_{\min})]$ of rate k, starting from $z(\beta_{\min})$ and going downwards. In other words, the random variable $\boldsymbol{z} = z(\mathtt{TPA}(k))$ is generated by the following process.

Algorithm 3 Equivalent process for generating $z(\mathtt{TPA}(k))$.

1: set $z_0 = z(\beta_{\min})$, let $\boldsymbol{z}$ be the empty multiset
2: **for** $i = 1$ **to** $+\infty$ **do**
3: draw η from the exponential distribution of rate k (and with the mean $\frac{1}{k}$), set $z_i = z_{i-1} - \eta$
4: if $z_i \in [z(\beta_{\max}), z(\beta_{\min})]$ then add z_i to $\boldsymbol{z}$, otherwise output $\boldsymbol{z}$ and terminate
5: **end**

One way to use the TPA method is to simply count the number of points in $\mathtt{TPA}(k)$. Indeed, $|\mathtt{TPA}(k)|$ is distributed according to the Poisson distribution with rate $k\cdot(z(\beta_{\min})-z(\beta_{\max})) = k\cdot q$, so $\frac{1}{k}|\mathtt{TPA}(k)|$ is an unbiased estimator of q. Unfortunately, obtaining a good estimate of q with this approach requires a fairly large number of samples, namely $O(q^2)$ for a given accuracy and the probability of failure (Huber and Schott, 2010, 2014). A better application of TPA was proposed in (Huber, 2015), where the method was used for generating a schedule $(\beta_0, \ldots, \beta_\ell)$ as follows.

Algorithm 4 Generating a schedule $(\beta_0, \ldots, \beta_\ell)$. **Input:** integers $k, d \geq 1$.

1: sample $\mathcal{B} \sim \mathtt{TPA}(k)$
2: sort the values in $\mathcal{B}$ and then keep every dth successive value
3: add values $\beta_{\min}$ and $\beta_{\max}$ and output the resulting sequence $(\beta_0, \ldots, \beta_\ell) = (\beta_{\min}, \ldots, \beta_{\max})$

Note that the resulting sequence $(z_1, \ldots, z_{\ell-1})$=$(z(\beta_1), \ldots, z(\beta_{\ell-1}))$ can be described by a process in Algorithm 3 where η is drawn as the sum of d exponential distributions each of rate k; this is the gamma (Erlang) distribution with shape parameter d and rate parameter k.

Huber (2015) showed that if $d = \Theta(\ln q + \ln\ln n)$ and $k = \Theta(d \ln n)$ (with appropriate constants) then Algorithm 4 produces a good schedule with high probability. Since q is unknown in practice, Huber (2015) uses a two-stage procedure: first an estimate $\hat{q} = 2 \cdot \frac{|\mathtt{TPA}(5)|}{5} + 1$ is computed, which is shown to be an upper bound on q with probability at least 0.99. This estimate is then used for setting d and k.

In the next section we prove that the algorithm has desired guarantees for smaller parameter values, namely $d = \Theta(1)$ and $k = \Theta(\ln n)$. This allows to reduce the complexity of Algorithm 4 by a factor of $\Theta(\ln q + \ln\ln n)$, and also eliminates the need for a two-stage procedure.

3. Our results

For technical reasons we will need to make the following assumption for line 2 of Algorithm 4: if $\beta_1, \beta_2, \ldots$ is the sorted sequence of points in $\mathcal{B}$ then the index of the first point to be taken is sampled uniformly from $\{1, \ldots, d\}$ (and after that the index is always incremented by d).

Denote $m = \frac{k}{d}$ and $z_i = z(\beta_i)$ for $i \in [0, \ell]$. We treat m and d as being fixed, and $k = md$ as their function. Also let $\delta = \ln \mathbb{S}[W] = \ln \mathbb{S}[V]$. From (3) we get

$$\delta = \sum_i \delta_i\ , \qquad \delta_i = z(\beta_i) - 2z\left(\tfrac{\beta_i+\beta_{i+1}}{2}\right) + z(\beta_{i+1}). \tag{5}$$

Since $z(\cdot)$ is convex, we have $\delta_i \geq 0$ for all i.

Case I: $H(x) \in [1,n]$ **for all** $x \in \Omega$ First, let us assume that $H(\cdot)$ does not take value 0. In this case the proofs become somewhat simpler, and we will get slightly smaller constants.

Huber showed that for $d = \Theta(\ln(q \ln n))$ the schedule is *well-balanced* with probability $\Theta(1)$, meaning that **all** intervals i satisfy $z_i - z_{i+1} \leq \tau \cdot \frac{1}{m}$ for a constant $\tau = \frac{4}{3}$. (Note that $\mathbb{E}[z_i - z_{i+1}] \approx \frac{1}{m}$, ignoring boundary effects). It was then proved[2] that a well-balanced schedule satisfies $\delta \leq \frac{\tau}{2} \cdot \frac{\ln n}{m}$, leading to condition (4). We improve on this result as follows.

Choose a constant $\tau > 0$ (to be specified later), and say that interval i is *large* if $z_i - z_{i+1} > \tau \cdot \frac{1}{m}$, and *small* otherwise. Let δ^+ be the sum of δ_i over large intervals and δ^- be the sum of δ_i over small intervals (so that $\delta = \delta^+ + \delta^-$). In Section A.1 we prove the following fact. (Recall that δ^+, δ^- are deterministic functions of the schedule $(\beta_0, \ldots, \beta_\ell)$).

Lemma 5 *There holds $\delta^- \leq \frac{\tau}{2} \cdot \frac{\ln n}{m}$ and $\mathbb{E}[\delta^+] \leq \frac{\Gamma(d+2,\tau d)}{2d \cdot d!} \cdot \frac{\ln n}{m}$ for the schedule $(\beta_0, \ldots, \beta_\ell)$ produced by Algorithm 4 with parameters $k = md$ and d, where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function:*

$$\Gamma(a,b) = \int_b^{+\infty} t^{a-1} e^{-t} dt \qquad (\textit{with } \Gamma(a,0) = \Gamma(a) = (a-1)!)$$

Using Markov's inequality, we can now conclude that for any $\tau^+ > 0$ we have

$$\mathbb{P}(\delta^+ \geq \tfrac{\tau^+}{2} \cdot \tfrac{\ln n}{m}) \quad \leq \quad \frac{1}{\frac{\tau^+}{2} \cdot \frac{\ln n}{m}} \cdot \mathbb{E}[\delta^+] \quad \leq \quad \frac{\Gamma(d+2,\tau d)}{\tau^+ \cdot d \cdot d!}$$

Thus, with probability at least $1 - \frac{\Gamma(d+2,\tau d)}{\tau^+ \cdot d \cdot d!}$ Algorithm 4 produces a schedule satisfying $\delta \leq \frac{\tau + \tau^+}{2} \cdot \frac{\ln n}{m}$.

Recall that we want Algorithm 4 to succeed with probability at least $\rho = \frac{0.75}{1-\gamma}$ to make the overall probability of success at least 0.75. (Here γ is the constant from Lemma 3). Let us define function $\tau_\rho(d)$ as follows:

$$\tau_\rho(d) = \min_{\tau \geq 0, \tau^+ > 0} \left\{ \tau + \tau^+ \mid \tfrac{\Gamma(d+2,\tau d)}{\tau^+ \cdot d \cdot d!} \leq 1 - \rho \right\} = \min_{\tau \geq 0} \left[\tau + \tfrac{\Gamma(d+2,\tau d)}{(1-\rho) \cdot d \cdot d!} \right]$$

The table below shows some values of this function for $\gamma = 0.24$ and $\rho = \frac{0.75}{1-\gamma} = \frac{75}{76}$ (computed with the code of Bhattacharjee (1970)).

d	1	2	4	8	16	32	64	128	256	512
upper bound on $\tau_\rho(d)$	9.903	6.052	4.000	2.860	2.197	1.794	1.539	1.372	1.260	1.184
achieved with $\tau =$	8.645	5.384	3.634	2.653	2.075	1.720	1.492	1.342	1.241	1.170

We can now formulate our main result for case I.

Theorem 6 *Let $\hat{Q}$ be the estimate given by Algorithm 1 (with parameter r) applied to the schedule produced by Algorithm 4 (with parameters $k = md$ and d). Suppose that*

$$m \geq \frac{\tau_\rho(d) \cdot \ln n}{2 \ln \left(1 + \frac{1}{2} \gamma r \tilde{\varepsilon}^2\right)} \qquad \textit{for some } \gamma \in (0, 0.25) \textit{ and } \rho = \tfrac{0.75}{1-\gamma} \tag{6}$$

2. More precisely, this is what the argument of Huber (2015) would give assuming that $H(\cdot)$ does not take value 0.

where $\tilde{\varepsilon} = 1 - (1+\varepsilon)^{-1/2} = \frac{1}{2}\varepsilon + O(\varepsilon^2)$*. Then* $\hat{Q} \in (\frac{Q}{1+\varepsilon}, Q(1+\varepsilon))$ *with probability at least* 0.75*. The expected number of oracle calls that this algorithm makes is* $mq(r+d)+2r+1$*.*

In particular, (6) *will be satisfied for* $d = 64$*,* $m \geq 3.6 \cdot \ln n$ *and* $r = \lceil 2\tilde{\varepsilon}^{-2} \rceil = 8(1+o(1))\varepsilon^{-2}$*.*

Proof As we just showed,

$$\mathbb{P}\left(\delta \leq \frac{\tau_\rho(d)}{2} \cdot \frac{\ln n}{m}\right) \quad \geq \quad \rho \tag{7}$$

Condition $\delta \leq \frac{\tau_\rho(d)}{2} \cdot \frac{\ln n}{m}$ implies condition $\delta \leq \ln\left(1 + \frac{1}{2}\gamma r \tilde{\varepsilon}^2\right)$ (by (6)), which is in turn equivalent to $\mathbb{S}[W] \leq 1 + \frac{1}{2}\gamma r \tilde{\varepsilon}^2$. Thus, from Lemma 3 we get

$$\mathbb{P}\left(\hat{Q} \in (\tfrac{Q}{1+\varepsilon}, Q(1+\varepsilon)) \mid \delta \leq \frac{\tau_\rho(d)}{2} \cdot \frac{\ln n}{m}\right) \geq 1 - \gamma \tag{8}$$

Multiplying (7) and (8) gives the first claim.

A PPP of rate k on an interval $[z(\beta_{\max}), z(\beta_{\min})]$ produces $k[z(\beta_{\min}) - z(\beta_{\max})] = mdq$ points on average. Thus, Algorithm 4 makes $mdq + 1$ oracle calls on average and produces a sequence $(\beta_0, \ldots, \beta_\ell)$ with $\mathbb{E}[\ell] = mq + 1$. Algorithm 1 then makes $(\ell+1)r$ oracle calls, i.e. $(mq+2)r$ calls on average. This gives the second claim. ∎

$\boxed{\textbf{Case II: } H(x) \in \{0\} \cup [1,n] \textbf{ for all } x \in \Omega}$ We now consider the general case. In Section A.2 we prove the following fact.

Lemma 7 *For any constant* $\lambda \in (0,1)$ *there exists a decomposition* $\delta = \delta^- + \delta^+$ *with* $\delta^-, \delta^+ \geq 0$ *such that*

$$\delta^- \leq \ln \frac{1}{1-\lambda} + \frac{\tau}{2} \cdot \frac{2 + \ln \frac{n}{\lambda}}{m} \qquad\qquad \mathbb{E}[\delta^+] \leq \frac{\Gamma(d+2, \tau d)}{2d \;\cdot\; d!} \cdot \frac{2 + \ln \frac{n}{\lambda}}{m}$$

As in the first case, we conclude from the Markov's inequality that with probability at least $1 - \frac{\Gamma(d+2,\tau d)}{\tau^+ \cdot d \cdot d!}$ Algorithm 4 produces a schedule satisfying $\delta \leq \ln \frac{1}{1-\lambda} + \frac{\tau + \tau^+}{2} \cdot \frac{2 + \ln \frac{n}{\lambda}}{m}$. This leads to

Theorem 8 *The conclusion of Theorem 6 holds if*

$$m \geq \frac{\tau_\rho(d) \cdot (2 + \ln \frac{n}{\lambda})}{2 \ln \left[\left(1 + \frac{1}{2}\gamma r \tilde{\varepsilon}^2\right)(1-\lambda)\right]} \qquad \textit{for some } \gamma \in (0, 0.25),\ \rho = \tfrac{0.75}{1-\gamma} \textit{ and } \lambda \in (0,1) \tag{9}$$

For example, (9) *will be satisfied for* $d = 64$*,* $m \geq 3.6 \cdot (9 + \ln n)$ *and* $r = \lceil 2\tilde{\varepsilon}^{-2} \rceil = 8(1+o(1))\varepsilon^{-2}$ *(where we used* $\gamma = 0.24$ *and* $\lambda = e^{-7}$*).*

3.1. Approximate sampling oracles

So far we assumed that exact sampling oracles μ_β are used. For many applications, however, we only have approximate sampling oracles $\tilde{\mu}_\beta$ that are sufficiently close to μ_β in terms of the variation distance $||\cdot||_{TV}$ defined via

$$||\tilde{\mu}_\beta - \mu_\beta||_{TV} = \max_{A \subseteq \Omega} |\tilde{\mu}_\beta(A) - \mu_\beta(A)| = \frac{1}{2}\sum_{x \in \Omega} |\tilde{\mu}_\beta(x) - \mu_\beta(x)|.$$

The analysis can be extended to approximate oracles using a standard trick (see e.g. (Štefankovič et al., 2009, Remark 5.9)).

Theorem 9 *Let $\hat{Q}$ be the output of the algorithm with parameters d, m, r satisfying the conditions of Theorem 6 or 8 (depending on whether $H(\cdot) \in [1, n]$ or $H(\cdot) \in \{0\} \cup [n]$), where exact sampling oracles μ_β are replaced with approximate sampling oracles $\tilde{\mu}_\beta$ satisfying $||\mu_\beta - \tilde{\mu}_\beta||_{TV} \le \frac{\kappa}{mq(r+d)+3r+1}$. Then $\hat{Q} \in (\frac{Q}{1+\varepsilon}, Q(1+\varepsilon))$ with probability at least $0.75 - \kappa$.*

As mentioned in the introduction, probability $0.75 - \kappa$ can be boosted to any other probability in $(0.5, 1)$ by repeating the algorithm a constant number of times and taking the median (assuming that κ is a constant in $(0, 0.25)$). Alternatively, one can tweak parameters in Theorems 6 and 8 to get the desired probability directly.

Proof It is known that there exists a coupling between μ_β and $\tilde{\mu}_\beta$ such that they produce identical samples with probability at least $1 - ||\tilde{\mu}_\beta - \mu_\beta||_{TV} \ge 1 - \delta$, where we denoted $\delta = \frac{\kappa}{mq(r+d)+3r+1}$. Let $\mathbb{A}$ and $\tilde{\mathbb{A}}$ be the algorithms that use respectively exact and approximate samples, where the k-th call to μ_β in $\mathbb{A}$ is coupled with the k-th call to $\tilde{\mu}_{\tilde{\beta}}$ in $\tilde{\mathbb{A}}$ when $\beta = \tilde{\beta}$. We say that the k-th call is *good* if the produced samples are identical. Note, $\mathbb{P}[k\text{-th call is good} \mid \text{all previous calls were good}] \ge 1 - \delta$, since the conditioning event implies $\beta = \tilde{\beta}$. Also, if all calls are good then $\mathbb{A}$ and $\tilde{\mathbb{A}}$ give identical results.

Let N be the number of points inside $[z(\beta_{\max}), z(\beta_{\min})]$ produced by the call $\mathtt{TPA}(md)$ in Algorithm 4. Then N follows the Poisson distribution of rate $\lambda = mdq$, i.e. $\mathbb{P}(N = n) = \frac{\lambda^n e^{-\lambda}}{n!}$. Algorithm 4 makes $N+1$ oracle calls, and produces a sequence $(\beta_0, \ldots, \beta_\ell)$ with $\ell \le \frac{N}{d} + 2$. Thus, the total number of oracle calls is $N + 1 + (\ell + 1)r \le Nc + 3r + 1$ where $c = 1 + \frac{r}{d}$. Denoting $\mu = \lambda(1 - \delta)^c$, we can write

$$
\begin{aligned}
\mathbb{P}[\text{all calls are good}] &\ge \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot (1-\delta)^{nc+3r+1} = \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot (1-\delta)^{nc+3r+1} \\
&= \sum_{n=0}^{\infty} \frac{\mu^n e^{-\mu}}{n!} \cdot e^{\mu-\lambda}(1-\delta)^{3r+1} = e^{\mu-\lambda}(1-\delta)^{3r+1} = e^{-\lambda(1-(1-\delta)^c)}(1-\delta)^{3r+1} \\
&\ge e^{-\lambda(1-(1-c\delta))}(1-\delta)^{3r+1} \ge (1-\lambda c\delta)(1-\delta)^{3r+1} \ge 1-\lambda c\delta-(3r+1)\delta \ge 1-\kappa
\end{aligned}
$$

where we used the facts that $(1 - x)^c \ge 1 - cx$ and $e^{-x} \ge 1 - x$ for $x \ge 0$ and $c \ge 1$. Using the union bound, we obtain the claim of the theorem. ■

3.2. Lower bound

In this section we establish a lower bound on the number of calls to the sampling oracles for estimating $q = \ln \frac{Z(\beta_{\min})}{Z(\beta_{\max})}$. First, we describe our model of computation and the set of instances that we allow.

We assume that the estimation algorithm only receives values $H(x)$ from the sampling oracle, and not individual states $x \in \Omega$. This means that an instance can be defined by counts $c_h = |\{x \in \Omega \mid H(x) = h\}|$ for values h in the range of H; these counts uniquely specify the partition function $Z(\beta) = \sum_h c_h e^{-\beta h}$ and the distribution of sampling oracle outputs for a given β. We will thus view an instance as a triplet $\Gamma = (c[\Gamma], \beta_{\min}[\Gamma], \beta_{\max}[\Gamma])$ where $c[\Gamma] : \mathbb{R} \to \mathbb{Z}_{\ge 0}$ is a function with a finite non-empty support. When the instance is clear from the context, we will omit the square brackets and write simply $\Gamma = (c, \beta_{\min}, \beta_{\max})$. For a value $h \in \mathtt{supp}(c)$ let $\psi(\beta, h \mid \Gamma)$ be the probability that the sampling oracle returns value h when queried at β in instance Γ:

$$\psi(\beta, h \mid \Gamma) = c_h e^{-\beta h} / Z(\beta)$$

For a finite subset $\mathcal{H} \subseteq \mathbb{R}$ let $\mathbb{I}(\mathcal{H})$ be the set of instances $\Gamma = (c, 0, \beta_{\max})$ satisfying $\mathtt{supp}(c) \subseteq \mathcal{H}$. Also for a subset $\mathcal{Q} \subseteq \mathbb{R}$ let $\mathbb{I}(\mathcal{H}, \mathcal{Q}) = \{\Gamma \in \mathbb{I}(\mathcal{H}) \mid q^*(\Gamma) \in \mathcal{Q}\}$, where we denoted $q^*(\Gamma) = \ln \frac{Z(0)}{Z(\beta_{\max})}$.

An estimation algorithm $\mathcal{A}$ applied to instance $\Gamma = (c, 0, \beta_{\max}) \in \mathbb{I}(\mathcal{H})$ is assumed to have the following form. At step i (for $i = 1, 2, \ldots$) it does one of the following two actions:

1. Call the samping oracle for some value $\beta_i \in \mathbb{R}$. The oracle then returns a random variable $h_i \in \mathcal{H}$ with $\mathbb{P}(h_i = h) = \psi(\beta_i, h \mid \Gamma)$ for each $h \in \mathcal{H}$.

2. Output some estimate $\hat{q}$ and terminate.

The i-th action is a random variable that can depend only on the set $\mathtt{supp}(c)$, values $\beta_{\min}, \beta_{\max}$, and on the previously observed sequence $(\beta_1, h_1), \ldots, (\beta_{i-1}, h_{i-1})$. The output $\hat{q}$ of the algorithm will be denoted as $q^{\mathcal{A}}(\Gamma)$, and the expected number of calls to the sampling oracle as $T^{\mathcal{A}}(\Gamma)$.

We say that algorithm $\mathcal{A}$ is an (ε, δ)-*estimator for instance* Γ if $\mathbb{P}[|q^{\mathcal{A}}(\Gamma) - q^*(\Gamma)| > \varepsilon] < \delta$. We can now formulate our main theorem.

Theorem 10 *There exist positive numbers $q_{\min}, n_{\min}, c_1, c_2, c_3$ such that the following holds for all $q \geq q_{\min}$, $n \geq n_{\min}$ with $n \in \mathbb{Z}$, $\varepsilon \in (0, c_1 q)$, $\delta \in (0, \frac{1}{4})$.*

Denote $m = \left\lceil \frac{c_2 \sqrt{q}}{n} \right\rceil$ and $\mathcal{H}_n^m = \{h \in [1, n] \,:\, mh \in \mathbb{Z}\}$. Suppose that $\mathcal{A}$ is an (ε, δ)-estimator for all instances in $\mathbb{I}(\mathcal{H}_n^m, \left[\frac{2q}{3}, \frac{4q}{3}\right])$. Then there exists instance $\Gamma \in \mathbb{I}(\mathcal{H}_n^m, \left[\frac{2q}{3}, \frac{4q}{3}\right])$ such that $T^{\mathcal{A}}(\Gamma) \geq c_3 q \varepsilon^{-2} \ln \delta^{-1}$.

To prove this theorem, we construct three instances $\Gamma, \Gamma_+, \Gamma_-$ such that $|q^*(\Gamma_\pm) - q^*(\Gamma)| > 2\varepsilon$ and it is hard to differentiate between Γ and $\Gamma_\pm$ based on the outputs of sampling oracles. More precisely, we require that $\frac{\psi(\beta,h|\Gamma_-)\psi(\beta,h|\Gamma_+)}{\psi(\beta,h|\Gamma)^2} \geq \gamma$ for any β, h, where constant $\gamma < 1$ is sufficiently close to 1. Instances $\Gamma_\pm$ have functions $c_h^\pm = c_h \cdot e^{\pm h\nu}$ for some constant $\nu > 0$, where c_h is the function of Γ. The latter instance is defines so that its partition function has the form $Z(\beta) = e^{-\beta} \prod_{k=1}^{N} (a_k + e^{-\beta/m})$ where N is some integer in $[m(n-1)]$ and $a_1, \ldots, a_N$ are exponentially decreasing positive numbers. For further details we refer to Section A.3.

Acknowledgments

I thank Laszlo Erdös and Alexander Zimin for useful discussions. In particular, the link[3] provided by Alexander helped with the argument in Section A.1. The author is supported by the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 616160.

3. http://math.stackexchange.com/questions/74454/paradox-of-a-poisson-process-on-mathbb-r

References

I. Bezáková, D. Štefankovič, V. V. Vazirani, and E. Vigoda. Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM J. Comput.*, 37:1429–1454, 2008.

G. P. Bhattacharjee. Algorithm AS 32: The incomplete gamma integral. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 19(3):285–287, 1970.

Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC, 2011.

M. Dyer and A. Frieze. Computing the volume of convex bodies: A case where randomness provably helps. In *Proceedings of AMS Symposium on Probabilistic Combinatorics and Its Applications 44*, pages 123–170, 1991.

J. A. Fill and M. L. Huber. Perfect simulation of Vervaat perpetuities. *Electron. J. Probab.*, 15: 96–109, 2010.

G. S. Fishman. Choosing sample path length and number of sample paths when starting in the steady state. *Oper. Res. Lett.*, 16:209–219, 1994.

Heng Guo and Mark Jerrum. Random cluster dynamics for the Ising model is rapidly mixing. *The Annals of Applied Probability*, 28(2):1292–1313, 2018.

Mark Huber. Perfect sampling using bounding chains. *Annals of Applied Probability*, 14(2):734–753, 2004.

Mark Huber. Approximation algorithms for the normalizing constant of Gibbs distributions. *arXiv:1206.2689v1*, June 2012.

Mark Huber. Approximation algorithms for the normalizing constant of Gibbs distributions. *The Annals of Applied Probability*, 25(2):974–985, 2015.

Mark Huber and Sarah Schott. Using TPA for Bayesian inference. *Bayesian Statistics 9*, pages 257–282, 2010.

Mark Huber and Sarah Schott. Random construction of interpolating sets for high-dimensional integration. *arXiv:1112.3692*, December 2011.

Mark Huber and Sarah Schott. Random construction of interpolating sets for high-dimensional integration. *J. Appl. Prob.*, 51:92–105, 2014.

M. Jerrum. A very simple algorithm for estimating the number of k-colourings of a low-degree graph. *Random Structures and Algorithms*, 7:157–165, 1995.

M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22:1087–1116, 1993.

Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM J. COMPUT.*, 18(6):1149–1178, December 1989.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

J. F. C. Kingman. *Poisson Processes*. Clarendon Press, 1992.

James Matthews. *Markov Chains for Sampling Matchings*. PhD thesis, University of Edinburgh, School of Informatics, 2008.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

James G. Propp and David B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1-2):223–252, 1996.

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, 1986.

J. P. Valleau and D. N. Card. Monte Carlo estimation of the free energy by multistage sampling. *J. Chem. Phys.*, 57:5457–5462, 1972.

E. Vigoda. Improved bounds for sampling colorings. In *FOCS*, pages 51–59, 1999.

D. Štefankovič, S. Vempala, and E. Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *J. of the ACM*, 56(3):1–36, 2009.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, December 2008.

Appendix A. Proofs

A.1. Proof of Lemma 5

We will assume that the sequence $(\beta_0, \ldots, \beta_\ell)$ is strictly increasing (this holds with probability 1). Accordingly, the sequence $(z_0, \ldots, z_\ell)$ is strictly decreasing. The following has been shown in (Štefankovič et al., 2009; Huber, 2015).

Lemma 11 *For any $i \in [0, \ell - 1]$ there holds $\delta_i \le z_i - z_{i+1}$ and also*

$$\frac{-z'(\beta_i)}{-z'(\beta_{i+1})} \ge \exp(2\delta_i/(z_i - z_{i+1}))$$

Proof Denote $\bar{\beta} = (\beta_i + \beta_{i+1})/2$ and $\bar{z} = z(\bar{\beta})$, then $\delta_i = z_{i+1} - 2\bar{z} + z_i$. Since $z(\cdot)$ is a convex strictly decreasing function, we have

$$-z'(\beta_i) \ge \frac{z_i - \bar{z}}{\bar{\beta} - \beta_i} \qquad -z'(\beta_{i+1}) \le \frac{\bar{z} - z_{i+1}}{\beta_{i+1} - \bar{\beta}}$$

Since $\bar{\beta} - \beta_i = \beta_{i+1} - \bar{\beta}$, taking the ratio gives the second claim of the lemma:

$$\frac{-z'(\beta_i)}{-z'(\beta_{i+1})} \ge \frac{z_i - \bar{z}}{\bar{z} - z_{i+1}} = \frac{\frac{1}{2}(z_i - z_{i+1} + \delta_i)}{\frac{1}{2}(z_i - z_{i+1} - \delta_i)} = \frac{1+\lambda}{1-\lambda} \ge e^{2\lambda}$$

where we denoted $\lambda = \frac{\delta_i}{z_i - z_{i+1}} \ge 0$ and observed that $\lambda < 1$ since $1 - \lambda = 2\frac{\bar{z} - z_{i+1}}{z_i - z_{i+1}} > 0$. The fact that $\lambda < 1$ also gives the first claim of the lemma. ■

Let us define $s(\beta) = \ln[-z'(\beta)]$ and $s_i = \ln[-z'(\beta_i)]$ for $i \in [0, \ell]$, then function $s(\cdot)$ and the sequence $(s_0, \ldots, s_\ell)$ are strictly decreasing. Since $z(\beta)$ and $s(\beta)$ are continuous strictly decreasing functions of β, we can uniquely express z via s and define a continuous strictly increasing function $z(s)$ on the interval $S \overset{\text{def}}{=} [s_\ell, s_0]$ (Fig. 1(b)). Note, with some abuse of notation we use $z(\cdot)$ for two different functions: one of argument β, and another one of argument s. The exact meaning should always be clear from the context.

The inequality in the last lemma for an interval $i \in [0, \ell - 1]$ can be rewritten as follows:

$$2\delta_i \quad \le \quad (z_i - z_{i+1}) \cdot (s_i - s_{i+1}) \tag{10}$$

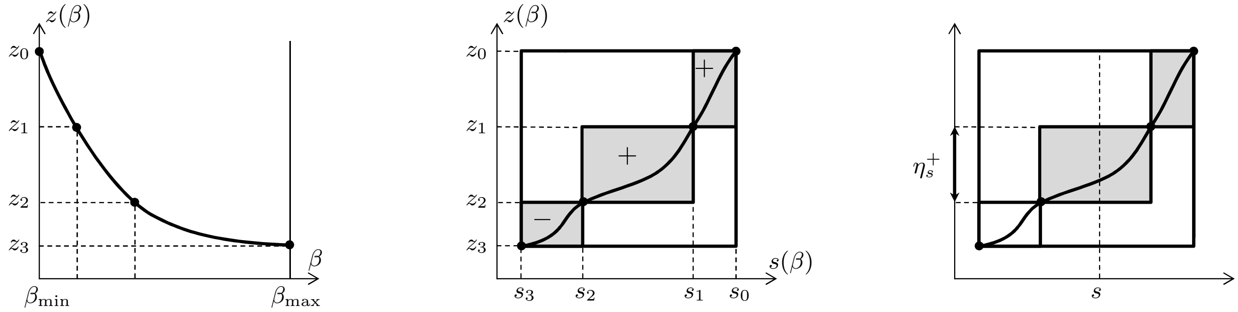
Equivalently, we have $2\delta_i \le Area(\Delta_i)$ where $\Delta_i \subseteq [s_\ell, s_0] \times [z_\ell, z_0]$ is the rectangle with the top right corner at (s_i, z_i) and the bottom left corner at (s_{i+1}, z_{i+1}) (Fig. 1(b)). Let Δ^+ be the union of rectangles Δ_i corresponding to large intervals i (with $|z_i - z_{i+1}| > \tau \cdot \frac{1}{m}$), and Δ^- be the union of Δ_i corresponding to small intervals i. Then $2\delta^+ \le Area(\Delta^+)$ and $2\delta^- \le Area(\Delta^-)$.

By geometric considerations it should be clear that

$$Area(\Delta^-) \quad \le \quad \max\{|z_i - z_{i+1}| \ : \ i \text{ is small}\} \cdot |S| \quad \le \quad \tau \cdot \tfrac{1}{m} \cdot |S|$$

Observe that $-z'(\beta) = \mathbb{E}_{X \sim \mu_\beta}[H(X)] \in [1, n]$ for any β, and therefore $S = [s_\ell, s_0] \subseteq [0, \ln n]$ and so $|S| \le \ln n$. This establishes the first claim of Lemma 5. Next, we focus on proving the second claim.

For a point $s \in S$ let η_s be the length of the interval (z_{i+1}, z_i) into which $z(s)$ falls (or 0, if $z(s) \in \{z_\ell, \ldots, z_0\}$). Also let $\eta_s^+ = \psi[\eta_s]$ where $\psi[\cdot]$ is the following function: $\psi[a] = a$ if



(a) (b) (c)

Figure 1: (a) $z(\beta) = \ln Z(\beta)$ is a strictly convex decreasing function. Four dots show a possible output of Algorithm 4. Here $\ell = 3$ and $[\beta_0, \beta_\ell] = [\beta_{\min}, \beta_{\max}]$. (b) Definitions of the sets Δ_i (in gray). Intervals 0 and 1 are assumed to be large, while interval 2 is small. (c) Definition of the variable η_s^+ (in the case of a large interval).

$a > \tau \cdot \frac{1}{m}$, and $\psi[a] = 0$ otherwise. Thus, if $z(s) \in (z_{i+1}, z_i)$ for some large interval i then $\eta_s^+ = z_i - z_{i+1}$ (Fig. 1(c)), otherwise $\eta_s^+ = 0$. We have

$$Area(\Delta^+) = \int_S \eta_s^+ ds$$

The linearity of expectation gives

$$2\mathbb{E}[\delta^+] \quad \le \quad \mathbb{E}(Area(\Delta^+)) \quad = \quad \int_S \mathbb{E}[\eta_s^+] ds \quad \le \quad \max_{s \in S} \mathbb{E}[\eta_s^+] \cdot |S| \tag{11}$$

Now let $X_0, X_1, X_2, \ldots$ be a Poisson process on $[0, +\infty)$ and $X_{-1}, X_{-2}, \ldots$ be a Poisson process on $(-\infty, 0]$ (both with rate k). Thus, $X_i = \xi_0 + \ldots + \xi_i$ for $i \ge 0$ and $X_i = -\xi_{-1} - \xi_{-2} - \ldots - \xi_i$ for $i \le -1$, where ξ_j are i.i.d. variables from the exponential distribution of rate k. By the superposition theorem for Poisson processes (Kingman, 1992, page 16), bidirectional sequence $\mathbf{X} = \ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ is a Poisson process on $(-\infty, +\infty)$ (again with rate k), and in particular it is translation-invariant.

Let $\mathbf{Y} = \ldots, Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, \ldots$ be the following process: draw an integer $c \in \{0, \ldots, d-1\}$ uniformly at random and then set $Y_i = X_{di+c}$ for each i. It can be seen that $\mathbf{Y}$ models the output $(\beta_0, \ldots, \beta_\ell)$ of Algorithm 4 as follows: take the sequence $z(\beta_{\min}) - Y_0, z(\beta_{\min}) - Y_1, z(\beta_{\min}) - Y_2, \ldots$, restrict to $[z(\beta_{\max}), z(\beta_{\min})]$ and append $z(\beta_{\min})$ and $z(\beta_{\max})$. Then the resulting sequence has the same distribution as $(z(\beta_0), \ldots, z(\beta_\ell))$. We assume below that $(\beta_0, \ldots, \beta_\ell)$ is generated by this procedure.

For a point $a \in \mathbb{R}$ let θ_a be the length of the interval (Y_i, Y_{i+1}) into which a falls (or 0, if no such interval exists). Note, the distribution of random variable θ_a does not depend on a (since process $\mathbf{Y}$ is translation-invariant). We also denote $\theta_a^+ = \psi[\theta_a]$, and let θ and $\theta^+ = \psi[\theta]$ be random variables with the same distributions as θ_a and θ_a^+, respectively (for any fixed a). Clearly, for each $s \in [s_\ell, s_0]$ we have $\eta_s \le \theta_a$ and $\eta_s^+ \le \theta_a^+$ for a suitably chosen a, namely, $a = z(\beta_{\min}) - z(s)$. (Note, if $z(s) \in (z_{\ell-1}, z_1)$ then $\eta_s = \theta_a$ and $\eta_s^+ = \theta_a^+$, but at the boundaries the inequalities may be strict). We thus have

$$\mathbb{E}[\eta_s^+] \le \mathbb{E}[\theta^+] \tag{12}$$

Lemma 12 *Variable θ has the gamma (Erlang) distribution with shape parameter $d+1$ and rate k, whose probability density is $f(t) = k^{d+1}t^d e^{-kt}/d!$ for $t \ge 0$.*

Proof We prove this fact for variable θ_a with $a = 0$. We know that $Y_{-1} = X_{c-d} \le 0$ and $Y_0 = X_c \ge 0$, so $\theta_0 = Y_0 - Y_{-1}$ (with probability 1). By construction, $X_c - X_{c-d} = \xi_{c-d} + \xi_{c-d+1} + \ldots + \xi_c$, i.e. θ_0 is a sum of $d+1$ i.i.d. exponential random variables each of rate k. This implies the claim. ∎

Recall that $\theta^+ = \theta$ if $\theta > \tau/m$, and $\theta^+ = 0$ otherwise. Lemma 12 now gives

$$\begin{aligned}\mathbb{E}[\theta^+] &= \int_{\tau/m}^{+\infty} tf(t)dt = \int_{\tau/m}^{+\infty} \frac{k^{d+1}t^{d+1}e^{-kt}}{d!}dt = \int_{\tau d/k}^{+\infty} \frac{(kt)^{d+1}e^{-(kt)}}{k \cdot d!}d(kt) \\ &= \frac{1}{k \cdot d!}\int_{\tau d}^{+\infty} u^{d+1}\, e^{-u}du = \frac{\Gamma(d+2,\tau d)}{k \cdot d!} = \frac{\Gamma(d+2,\tau d)}{md \cdot d!}\end{aligned}$$

Combining this with (11) and (12) and observing again that $|S| \le \ln n$ finally gives the second claim of Lemma 5.

A.2. Proof of Lemma 7

We will use the same notation as in the previous section. Since $H(\cdot)$ can now take value 0, we have $-z'(\beta) = \mathbb{E}_{X\sim\mu_\beta}[H(X)] \in [0,n]$ and so $[s_\ell, s_0] \subseteq [-\infty, \ln n]$ (instead of $[s_\ell, s_0] \subseteq [0, \ln n]$, as in the previous section). We will deal with small values of $s(\beta)$ exactly as in (Huber, 2015).

Recall that $z'(\beta)$ is a strictly increasing function of β. Let $\hat\beta$ be the unique value with $z'(\hat\beta) = -\lambda$. (If it does not exist, then we take $\hat\beta \in \{-\infty, +\infty\}$ using the natural rule). Denote $\hat z = z(\hat\beta)$ and $\hat s = \ln[-z'(\hat\beta)]$. Now introduce the following terminology for an interval $i \in [0, \ell-1]$:

- interval i is *steep* if $\beta_{i+1} \le \hat\beta$, or equivalently $s_{i+1} \ge \hat s$;
- interval i is *flat* if $\beta_i \ge \hat\beta$, or equivalently $s_i \le \hat s$;
- interval i is *crossing* if $\hat\beta \in (\beta_i, \beta_{i+1})$, or equivalently $\hat s \in (s_{i+1}, s_i)$.

If steep intervals exist then $\hat\beta \ge \beta_{\min}$ and $z'(\hat\beta) \le -\lambda$. (The inequality may be strict if $\hat\beta = +\infty$). We thus have $[s_{i+1}, s_i] \subseteq [\hat s, s_0] \subseteq [\ln\lambda, \ln n]$ for all steep intervals i. The argument from the previous section gives that

$$\sum_{i\text{: } i \text{ is steep and small}} \delta_i \le \frac{\tau}{2} \cdot \frac{\ln\frac{n}{\lambda}}{m} \qquad\qquad \mathbb{E}\left[\sum_{i\text{: } i \text{ is steep and large}} \delta_i\right] \le \frac{\Gamma(d+2,\tau d)}{2d \cdot d!} \cdot \frac{\ln\frac{n}{\lambda}}{m}$$

(We just need to assume that $\beta_{\max}$ was replaced with $\min\{\beta_{\max}, \hat\beta\}$, then we would have $S = [s_\ell, s_0] \subseteq [\ln\lambda, \ln n]$ and $|S| \le \ln\frac{n}{\lambda}$ instead of $|S| \le \ln n$, the rest is the same as in the previous section).

Let us now consider flat intervals. The argument from Huber (2015) gives the following fact.

Lemma 13 *The sum of δ_i over flat intervals i is at most $\ln\frac{1}{1-\lambda}$.*

Proof Assume that flat intervals exist, then $\hat{\beta} \le \beta_{\max}$ and $z'(\hat{\beta}) \ge -\lambda$. (The inequality may be strict if $\hat{\beta} = -\infty$). Denote $\Omega_0 = \{x \in \Omega \mid H(x) = 0\}$ and $\Omega_+ = \{x \in \Omega \mid H(x) \ge 1\}$, then $\Omega = \Omega_0 \cup \Omega_+$ and

$$\mathbb{E}_{X\sim\mu_{\hat{\beta}}}[H(X)] = \frac{\sum_{x\in\Omega_+} \; H(x)e^{-\hat{\beta}H(x)}}{Z(\hat{\beta})} \ge \frac{\sum_{x\in\Omega_+} \; e^{-\hat{\beta}H(x)}}{Z(\hat{\beta})} = 1 - \frac{\sum_{x\in\Omega_0} \; e^{-\hat{\beta}H(x)}}{Z(\hat{\beta})} \ge 1 - \frac{Z(\beta_{\max})}{Z(\hat{\beta})}$$

On the other hand, $\mathbb{E}_{X\sim\mu_{\hat{\beta}}}[H(X)] = -z'(\hat{\beta}) \le \lambda$ and so $\frac{Z(\beta_{\max})}{Z(\hat{\beta})} \ge 1 - \lambda$ and $z(\hat{\beta}) - z(\beta_{\max}) \le \ln \frac{1}{1-\lambda}$. For all flat intervals i we have $[z_{i+1}, z_i] \subseteq [z(\beta_{\max}), z(\hat{\beta})]$ and also $\delta_i \le z_i - z_{i+1}$. This gives the claim of the lemma. ■

It remains to consider crossing intervals. Let us define values δ_c^- and δ_c^+ as follows. If there are no crossing intervals then $\delta_c^- = \delta_c^+ = 0$. Otherwise let i be the unique crossing interval; if i is small then set $(\delta_c^-, \delta_c^+) = (\delta_i, 0)$, and if i is large then set $(\delta_c^-, \delta_c^+) = (0, \delta_i)$. In all cases we have $\delta_c^- \le \frac{\tau}{m}$ (since $\delta_i \le z_i - z_{i+1}$). Also, $\mathbb{E}[\delta_c^+] \le \mathbb{E}[\psi(z_i - z_{i+1})] \le \mathbb{E}[\theta^+] \le \frac{\Gamma(d+2,\tau d)}{md \cdot d!}$ where function $\psi(\cdot)$ and random variable θ^+ were defined in the previous section.

We can finally prove Lemma 7. Define δ^- as δ_c^- plus the sum of δ_i over small steep intervals i and flat intervals i. Define δ^+ as δ_c^+ plus the sum of δ_i over large steep intervals i. By collecting inequalities above we obtain the desired claim.

A.3. Proof of Theorem 10

The proof will be based on the following result. For brevity, we use notation $a \pm b$ to denote the closed interval $[a - b, a + b]$.

Lemma 14 *Suppose that $\mathcal{A}$ is an (ε, δ)-estimator for instances $\Gamma \in \mathbb{I}(\mathcal{H}, \{q\})$ and $\Gamma_1, \ldots, \Gamma_d \in \mathbb{I}(\mathcal{H}, \mathbb{R} \setminus (q \pm 2\varepsilon))$, where $\beta_{\max}[\Gamma_i] = \beta_{\max}[\Gamma]$ and $\mathtt{supp}(c[\Gamma_i]) = \mathtt{supp}(c[\Gamma])$ for $i \in [d]$. Suppose that*

$$\prod_{i\in[d]} \frac{\psi(\beta, h \mid \Gamma_i)}{\psi(\beta, h \mid \Gamma)} \ge \gamma \qquad \forall \beta \in \mathbb{R}, h \in \mathtt{supp}(c[\Gamma]) \tag{13}$$

for some constant $\gamma \in (0, 1)$. Then $T^{\mathcal{A}}(\Gamma) \ge \frac{(1-\delta'-\delta)d\ln(\delta'/\delta)}{\ln(1/\gamma)}$ for any constant $\delta' \in [\delta, 1 - \delta]$.

Proof The run of the algorithm can be described by a random variable $X = ((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q})$, where t is the number of oracle calls (possibly infinite, in which case $\hat{q}$ is undefined). Let $\mathcal{X}$ be the set all possible runs, and $\mathbb{P}^{\mathcal{A}}(\cdot \mid \tilde{\Gamma})$ be the probability measure of this random variable conditioned on $\tilde{\Gamma}$ being the input instance. The structure of the algorithm implies that this measure can be decomposed as follows:

$$d\mathbb{P}^{\mathcal{A}}(x \mid \tilde{\Gamma}) = \psi(x \mid \tilde{\Gamma})\, d\mu^{\mathcal{A}}(x) \qquad \forall x \in \{((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q}) \in \mathcal{X} \mid t \text{ is finite}\} \tag{14}$$

where $\mu^{\mathcal{A}}(\cdot)$ is some measure on $\mathcal{X}$ that depends only on the algorithm $\mathcal{A}$, and function $\psi(\cdot)$ is defined via

$$\psi((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q} \mid \tilde{\Gamma}) = \prod_{i\in[t]} \psi(\beta_i, h_i \mid \tilde{\Gamma})$$

Denote $\tau = \frac{d \ln(\delta'/\delta)}{\ln(1/\gamma)}$, and define the following subsets of $\mathcal{X}$:

$$
\begin{aligned}
\mathcal{X}^* &= \{((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q}) \in \mathcal{X} \mid t \le \tau \text{ and } \hat{q} \in q \pm \varepsilon\} \\
\mathcal{X}' &= \{((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q}) \in \mathcal{X} \mid t > \tau\} \\
\mathcal{X}'' &= \{((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q}) \in \mathcal{X} \mid t \le \tau \text{ and } \hat{q} \notin q \pm \varepsilon\}
\end{aligned}
$$

Suppose the claim of Lemma 14 is false, i.e. $T^{\mathcal{A}}(\Gamma) \le (1 - \delta' - \delta) \cdot \tau$. We have $T^{\mathcal{A}}(\Gamma) \ge \mathbb{P}^{\mathcal{A}}(\mathcal{X}' \mid \Gamma) \cdot \tau$, and therefore

$$
\mathbb{P}^{\mathcal{A}}(\mathcal{X}' \mid \Gamma) \quad \le \quad 1 - \delta - \delta' \tag{15}
$$

Since $\mathcal{A}$ is a (ε, δ)-estimator for instances $\Gamma, \Gamma_1, \ldots, \Gamma_d$, we have

$$
\mathbb{P}^{\mathcal{A}}(\mathcal{X}'' \mid \Gamma) \quad < \quad \delta \tag{16}
$$

$$
\mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma_i) \quad < \quad \delta \qquad\qquad \forall i \in [d] \tag{17}
$$

Set $\mathcal{X}$ is a disjoint union of $\mathcal{X}^*, \mathcal{X}', \mathcal{X}''$, therefore $\mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma) = 1 - \mathbb{P}^{\mathcal{A}}(\mathcal{X}' \mid \Gamma) - \mathbb{P}^{\mathcal{A}}(\mathcal{X}'' \mid \Gamma) > 1 - (1 - \delta - \delta') - \delta = \delta'$. Combining this with (17) gives

$$
\frac{1}{d} \sum_{i \in [d]} \mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma_i) \quad < \quad \frac{\delta}{\delta'} \, \mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma) \tag{18}
$$

Assumption (13) of the lemma gives that

$$
\prod_{i \in [d]} \frac{\psi(x \mid \Gamma_i)}{\psi(x \mid \Gamma)} \quad \ge \quad \gamma^t \quad \ge \quad \gamma^\tau \qquad\qquad \forall x = ((\beta_1, h_1), \ldots, (\beta_t, h_t), \hat{q}) \in \mathcal{X}^* \tag{19}
$$

We can now write

$$
\frac{1}{d} \sum_{i \in [d]} \psi(x \mid \Gamma_i) \;\ge\; \left(\prod_{i \in [d]} \psi(x \mid \Gamma_i) \right)^{1/d} \;\ge\; \gamma^{\tau/d} \psi(x \mid \Gamma) \;=\; \frac{\delta}{\delta'} \, \psi(x \mid \Gamma) \qquad \forall x \in \mathcal{X}^* \tag{20}
$$

where the first inequality is a relation between arithmetic and geometric means of non-negative numbers, and the second inequality follows from (19). We can write

$$
\frac{1}{d} \sum_{i \in [d]} \mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma_i) \;\overset{\text{(a)}}{=}\; \frac{1}{d} \sum_{i \in [d]} \int_{\mathcal{X}^*} \psi(x \mid \Gamma_i) \, d\mu^{\mathcal{A}}(x) \;\overset{\text{(b)}}{\ge}\; \frac{\delta}{\delta'} \int_{\mathcal{X}^*} \psi(x \mid \Gamma) \, d\mu^{\mathcal{A}}(x) \;\overset{\text{(c)}}{=}\; \frac{\delta}{\delta'} \, \mathbb{P}^{\mathcal{A}}(\mathcal{X}^* \mid \Gamma)
$$

where (a,c) follow from (14) and (b) follows from (20). We obtained a contradiction to (18).

■

Recall that by definition coefficients of instances should be non-negative integers. When using Lemma 14, we can relax this requirement to non-negative rationals (since multiplying coefficients by a constant does not affect quantities in Lemma 14) and further to non-negative reals (since they can be approximated by rationals with an arbitrary precision).

We will use Lemma 14 with $d = 2$ and three instances $\Gamma, \Gamma_+, \Gamma_-$. First, we will describe the construction of Γ_+ and Γ_- given an instance Γ. After stating some properties of this construction, we will define the instance Γ.

Instances Γ_+ and Γ_- Suppose that $\Gamma = (c, 0, \beta_{\max}) \in \mathbb{I}(\mathcal{H})$. We set $\Gamma_+ = (c^+, 0, \beta_{\max})$ and $\Gamma_- = (c^-, 0, \beta_{\max})$ where functions c^+, c^- are given by

$$c^+_h = c_h \cdot e^{h\nu}, \qquad c^-_h = c_h \cdot e^{-h\nu} \qquad\qquad \forall h \in \mathbb{R}$$

where $\nu > 0$ is some constant. Let $Z(\cdot)$, $Z_+(\cdot)$, $Z_-(\cdot)$ be the partition functions corresponding to Γ, Γ_+, Γ_-, respectively. One can check that

$$Z(\beta) = \sum_{h \in \mathtt{supp}(c)} c_h e^{-\beta h} \qquad Z_+(\beta) = Z(\beta - \nu) \qquad Z_-(\beta) = Z(\beta + \nu)$$

Denote $z(\beta) = \ln Z(\beta)$ and $z_{\mathtt{diff}}(\beta) = z(\beta) - z(\beta_{\max} + \beta)$. Then

$$q = q^*[\Gamma] = z_{\mathtt{diff}}(0) \qquad q^*[\Gamma_+] = z_{\mathtt{diff}}(-\nu) \qquad q^*[\Gamma_-] = z_{\mathtt{diff}}(\nu)$$

Condition $\Gamma^+, \Gamma^- \in \mathbb{I}(\mathcal{H}, \mathbb{R} \setminus (q \pm 2\varepsilon))$ can thus be written as follows:

$$|z_{\mathtt{diff}}(\pm\nu) - z_{\mathtt{diff}}(0)| > 2\varepsilon \tag{21}$$

Condition (13) after cancellations becomes

$$\frac{Z^2(\beta)}{Z(\beta - \nu) Z(\beta + \nu)} \geq \gamma \qquad\qquad \forall \beta \in \mathbb{R}$$

or equivalently

$$z(\beta - \nu) - 2z(\beta) + z(\beta + \nu) \leq \ln \frac{1}{\gamma} \qquad\qquad \forall \beta \in \mathbb{R} \tag{22}$$

Let us define the following quantities; note that they depend only on instance Γ:

$$\rho = |z'_{\mathtt{diff}}(0)| \qquad\qquad \kappa = \sup_{\beta \in \mathbb{R}} z''(\beta) \tag{23}$$

Lemma 15 *Let Γ be an instance with values $q = q^*(\Gamma), \rho, \kappa$ as described above. Fix $\varepsilon \in (0, \frac{\rho^2}{10\kappa})$. Suppose that algorithm $\mathcal{A}$ is an (ε, δ)-estimator for all instances in $\mathbb{I}(\mathcal{H}, q \pm 4\varepsilon)$. Then for any constant $\delta' \in [\delta, 1 - \delta]$ we have*

$$T^{\mathcal{A}}(\Gamma) \geq \frac{2(1 - \delta' - \delta)\rho^2 \ln(\delta'/\delta)}{9\kappa\varepsilon^2}$$

Proof Non-negativity of function c implies that function $z(\cdot)$ is convex, and so $z''(\beta) \in [0, \kappa]$ for all $\beta \in \mathbb{R}$. Define $\nu = 3\varepsilon/\rho$. For $\beta = \pm\nu$ we can write

$$|z_{\mathtt{diff}}(\beta) - z_{\mathtt{diff}}(0)| \overset{\text{(a)}}{=} \left| z'_{\mathtt{diff}}(0)\beta + z''_{\mathtt{diff}}(\tilde{\beta})\frac{\beta^2}{2} \right| \overset{\text{(b)}}{\in} [|\rho\beta| - \kappa\beta^2, |\rho\beta| + \kappa\beta^2]$$

where in (a) we used Taylor's theorem with the Lagrange form of the remainder (here $\tilde{\beta} \in \mathbb{R}$), and in (b) we used the fact that $|z''_{\mathtt{diff}}(\tilde{\beta})| = |z''(\tilde{\beta}) - z''(\beta_{\max} + \tilde{\beta})| \le 2\kappa$. Observing that $|\rho\beta| = 3\varepsilon$ and $\kappa\beta^2 = \varepsilon \cdot \frac{9\kappa\varepsilon}{\rho^2} < \varepsilon$, we get $|z_{\mathtt{diff}}(\beta) - z_{\mathtt{diff}}(0)| \in (2\varepsilon, 4\varepsilon)$. Thus, condition (21) holds, and $\Gamma_+, \Gamma_- \in \mathbb{I}(\mathcal{H}, q \pm 4\varepsilon)$.

Denote $f(\beta) = z(\beta) - z(\beta - \nu)$. Using twice the mean value theorem, we get

$$z(\beta-\nu) - 2z(\beta) + z(\beta+\nu) = f(\beta+\nu) - f(\beta) = f'(\tilde{\beta})\nu = [z'(\tilde{\beta}) - z'(\tilde{\beta}-\nu)]\nu = z''(\tilde{\tilde{\beta}})\nu^2 \le \kappa\nu^2$$

where $\tilde{\beta}, \tilde{\tilde{\beta}} \in \mathbb{R}$. Thus, condition (22) will be satisfied if we set $\gamma \in (0,1)$ so that $\ln\frac{1}{\gamma} = \kappa\nu^2 = \frac{9\kappa\varepsilon^2}{\rho^2}$. Lemma 15 now follows from Lemma 14. ■

Instance Γ We now need to construct instance Γ such that $q = q^*[\Gamma]$ is close to a given value $\bar{q}$, and the ratio $\frac{\rho^2}{\kappa}$ is large. We will use an instance with the following partition function:

$$Z(\beta) \quad = \quad e^{-\beta} \prod_{k=1}^{N} (a_k + e^{-\beta/m}) \tag{24}$$

where N is some integer in $[m(n-1)]$ and $a_1, \ldots, a_N$ are non-negative numbers. Expanding terms yields $Z(\beta) = \sum_{h \in \mathcal{H}_n^m} c_h e^{-\beta h}$ for some coefficients $c_h \ge 0$, so this is indeed a valid definition of an instance $\Gamma \in \mathbb{I}(\mathcal{H}_n^m)$. In Section A.4 we prove the following fact.

Lemma 16 *There exist values $a_1, \ldots, a_N, \beta_{\max} > 0$ such that $q = \frac{\ln 2}{2}N^2 \pm O(mN)$ and $\frac{\rho^2}{\kappa} > (\frac{N}{4} - 1)^2$.*

This will imply Theorem 10. Indeed, let $\bar{q}$ be the value chosen in Theorem 10. Set $\hat{N} = \sqrt{\frac{2}{\ln 2}\bar{q}}$ and $N = \lceil \hat{N} \rceil$. Note that

$$\frac{\hat{N}}{m(n-1)} \le \frac{\sqrt{\frac{2}{\ln 2}\bar{q}}}{\frac{c_2\sqrt{\bar{q}}}{n}(n-1)} = const \cdot \frac{n}{n-1} \qquad \text{with} \qquad const = \sqrt{\frac{2}{\ln 2}} \Big/ c_2$$

Thus, setting $c_2 > \sqrt{\frac{2}{\ln 2}}$ will ensure that $N \in [m(n-1)]$ for sufficiently large n.

We have $q = \frac{\ln 2}{2}N^2 \pm O(mN) = \frac{\ln 2}{2}\hat{N}^2 \pm O(m\hat{N}) = \bar{q} \pm O(m\sqrt{\bar{q}}) = \bar{q}\left(1 \pm O\left(\frac{m}{\sqrt{\bar{q}}}\right)\right)$. Recalling that $m = \left\lceil \frac{c_2\sqrt{\bar{q}}}{n} \right\rceil$, we conclude that $q \in \left[\frac{3\bar{q}}{4}, \frac{5\bar{q}}{4}\right]$ if $\bar{q}, n$ are sufficiently large. Furthermore, we have $\frac{\rho^2}{\kappa} > (\frac{N}{4} - 1)^2 > \frac{1}{6}\bar{q}$ if $\bar{q}$ is sufficiently large (note that $\frac{1}{6} < \frac{1}{8\ln 2}$).

We set $c_1 = \frac{1}{60}$, so that $\varepsilon \in (0, \frac{1}{60}\bar{q})$. Now suppose that the preconditions of Theorem 10 hold. It can be checked that $\varepsilon \in (0, \frac{\rho^2}{10\kappa})$ and $q \pm 4\varepsilon \subseteq \left[\frac{2\bar{q}}{3}, \frac{4\bar{q}}{3}\right]$, so the preconditions of Lemma 15 hold as well. Setting $\delta' = \frac{1}{2}$ and recalling that $\delta \in (0, \frac{1}{4})$, we obtain the desired result:

$$T^{\mathcal{A}}(\Gamma) \ge \frac{2(1 - \frac{1}{2} - \delta)\ln(\frac{1}{2}/\delta)}{9\varepsilon^2} \cdot \frac{\rho^2}{\kappa} \;\ge\; \frac{\ln\delta^{-1} - \ln 2}{18\varepsilon^2} \cdot \frac{\bar{q}}{6} \;\ge\; \frac{1 - \frac{\ln 2}{\ln 4}}{18 \cdot 6} \cdot \frac{\bar{q}\ln\delta^{-1}}{\varepsilon^2}$$

A.4. Proof of Lemma 16

Denote $u = u(\beta) = e^{-\beta/m}$ and $\eta = u(\beta_{\max}) = e^{-\beta_{\max}/m}$. (The choice of $\eta \in (0,1)$ will be specified later). We can write

$$z(\beta) = -\beta + \sum_{k=1}^{N} \ln(a_k + u) \qquad z'(\beta) = -\sum_{k=N}^{N} \frac{u}{m(a_k+u)} \qquad z''(\beta) = \sum_{k=1}^{N} \frac{a_k u}{m^2(a_k+u)^2} \tag{25}$$

$$q \;=\; z(0) - z(\beta_{\max}) \;=\; m \ln \frac{1}{\eta} + \sum_{k=1}^{N} \ln \frac{a_k+1}{a_k+\eta} \tag{26}$$

$$\rho \;=\; |z'(0) - z'(\beta_{\max})| \;=\; \frac{1}{m} \sum_{k=1}^{N} \left[\frac{1}{a_k+1} - \frac{\eta}{a_k+\eta} \right] \tag{27}$$

As for $\kappa = \max_{\beta \in \mathbb{R}} z''(\beta)$, we will use the following bound.

Lemma 17 *Suppose that $a_1 \geq a_2 \geq \ldots \geq a_N > 0$. Then $\kappa \leq \max_{r \in [N-1]} \kappa_r$ where we denoted*

$$\kappa_r \;=\; \frac{1}{m^2} \left[\sum_{k=1}^{r} \frac{a_r}{a_k} \;+\; \sum_{k=r+1}^{N} \frac{a_k}{a_{r+1}} \right]$$

Proof We need to show that

$$\sum_{k=1}^{N} \frac{a_k u}{m^2(a_k+u)^2} \;\leq\; \max_{r \in [N-1]} \kappa_r \qquad\qquad \forall u \in (0, +\infty)$$

By taking the derivative one can check that function $g_k(u) = \frac{a_k u}{(a_k+u)^2}$ is increasing on $[0, a_k]$ and decreasing on $[a_k, +\infty)$ (with the maximum at $u = a_k$). Therefore, function $g(u) = \sum_{k=1}^{N} g_k(u)$ attains a maximum at $[a_N, a_1]$. We can thus assume w.l.o.g. that $u \in [a_N, a_1]$.

Let $r \in [N-1]$ be an index such that $u \in [a_{r+1}, a_r]$. For $k \in [1, r]$ we have $g_k(u) \leq g_k(a_r) = \frac{a_r/a_k}{(1+a_r/a_k)^2} \leq \frac{a_r}{a_k}$, and for $k \in [r+1, N]$ we have $g_k(u) \leq g_k(a_{r+1}) = \frac{a_k/a_{r+1}}{(1+a_k/a_{r+1})^2} \leq \frac{a_k}{a_{r+1}}$. By summing these inequalities we get that $g(u) \leq m^2 \kappa_r$. ∎

We can now prove Lemma 16. Define $a_k = 2^{1-k}$ and $\eta = 2^{1-N}$. For each $k \in [N]$ we have $\ln \frac{a_k+1}{a_k+\eta} \geq \ln \frac{1}{2a_k} = (k-2)\ln 2$ and $\ln \frac{a_k+1}{a_k+\eta} < \ln \frac{a_k+1}{a_k} = \ln \frac{1}{a_k} + \ln(1+a_k) \leq \ln \frac{1}{a_k} + a_k = (k-1)\ln 2 + 2^{1-k}$, therefore

$$q \;>\; m(N-1)\ln 2 + \sum_{k=1}^{N} (k-2) \ln 2 \;=\; \left(m + \frac{N}{2}\right)(N-1)\ln 2 - N \ln 2$$

$$q \;<\; m(N-1)\ln 2 + \sum_{k=1}^{N} \left[(k-1)\ln 2 + 2^{1-k}\right] \;<\; \left(m + \frac{N}{2}\right)(N-1)\ln 2 + 2$$

The following inequalities imply the last two claims of Lemma 16:

$$\begin{aligned}
\rho &> \frac{1}{m}\sum_{k=1}^{N}\left[\frac{1}{1+1}-\frac{\eta}{a_k}\right] = \frac{1}{m}\left[\frac{N}{2}-\frac{2^N-1}{2^{N-1}}\right] > \frac{1}{m}\left[\frac{N}{2}-2\right] \\
\kappa_r &< \frac{1}{m^2}\left[\sum_{k=1}^{r}\frac{a_r}{a_k}+\sum_{k=r+1}^{+\infty}\frac{a_k}{a_{r+1}}\right] = \frac{1}{m^2}\left[\frac{2^r-1}{2^{r-1}}+2\right] < \frac{4}{m^2} \qquad \forall r\in[N-1]
\end{aligned}$$