
Fast Doubly-Adaptive MCMC to Estimate the Gibbs Partition Function with Weak Mixing Time Bounds

(Supplementary material)

Shahrzad Haddadan *

Brown University
The Data Science Initiative
shahrzad.haddadan@gmail.com

Yue Zhuang 

Brown University
The Data Science Initiative
yue_zhuang1@brown.edu

Cyrus Cousins 


Brown University
Department of Computer Science
cyrus_cousins@brown.edu

Eli Upfal

Brown University
Department of Computer Science
eliezer_upfal@brown.edu

Abstract

We present a novel method for reducing the computational complexity of rigorously estimating the *partition functions* (normalizing constants) of Gibbs (Boltzmann) distributions, which arise ubiquitously in probabilistic graphical models. A major obstacle to practical applications of Gibbs distributions is the need to estimate their partition functions. The state of the art in addressing this problem is multi-stage algorithms, which consist of a cooling schedule, and a mean estimator in each step of the schedule. While the cooling schedule in these algorithms is adaptive, the mean estimation computations use MCMC as a black-box to draw approximate samples. We develop a *doubly adaptive* approach, combining the adaptive cooling schedule with an adaptive MCMC mean estimator, whose number of Markov chain steps adapts dynamically to the underlying chain. Through rigorous theoretical analysis, we prove that our method outperforms the state of the art algorithms in several factors: (1) The computational complexity of our method is smaller; (2) Our method is less sensitive to loose bounds on mixing times, an inherent component in these algorithms; and (3) The improvement obtained by our method is particularly significant in the most challenging regime of high-precision estimation. We demonstrate the advantage of our method in experiments run on classic factor graphs, such as voting models and Ising models.

*  indicates randomized ordering and equal contribution

Algorithm 1 RELMEANEST

```

1: procedure RELMEANEST
2:   Input: Markov chain  $\mathcal{M}$ , upper-bound on relaxation time  $T$ , real-valued function  $f$  with range  $[a, b]$ , letting  $R = b - a$ ,
   multiplicative precision  $\varepsilon$ , error probability  $\delta$ .
3:   Output: Multiplicative approximation  $\hat{\mu}$  of  $\mu = \mathbb{E}_\pi[f]$ .

4:    $T \leftarrow \left\lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \right\rceil$ ;  $\Lambda' \leftarrow \Lambda^T$  ▷ Choose  $T$  to be an upperbound on relaxation time
5:    $I \leftarrow 1 \vee \left\lceil \log_2 \left( \frac{bR}{2a^2} \cdot \frac{(1-\varepsilon)^2}{(1+\varepsilon)\varepsilon} \right) \right\rceil$ ;  $\alpha \leftarrow \frac{(1+\Lambda')R \ln \frac{3I}{2(1+\varepsilon)}}{(1-\Lambda')b\varepsilon}$ ;  $m_0 \leftarrow 0$  ▷ Initialize sampling schedule
6:    $T_{\text{unif}} \leftarrow \lceil T \cdot \ln(1/\pi_{\min}) \rceil$ ;  $(\vec{X}_{0,1}, \vec{X}_{0,2}) \leftarrow \mathcal{M}^{T_{\text{unif}}}(\perp)$  ▷ Warm-start two chains for  $T_{\text{unif}}$  steps from arbitrary  $\perp \in \Omega$ 
7:   for  $i \in 1, 2, \dots, I$  do
8:      $m_i \leftarrow \lceil \alpha r^i \rceil$  ▷ Total sample count at iteration  $i$ ;  $r$  is the geometric ratio (constant, usually 2) size
9:     for  $j \in (m_{i-1} + 1), \dots, m_i$  do
10:       $(\vec{X}_{j,1}, \vec{X}_{j,2}) \leftarrow (T \text{ steps of } \mathcal{M} \text{ starting at } \vec{X}_{j-1,1}, \vec{X}_{j-1,2})$  ▷ Run two independent copies of  $\mathcal{M}$  for  $T$  steps
11:       $\bar{f}(\vec{X}_{j,1}) \leftarrow \frac{1}{T} \sum_{t=1}^T f(\vec{X}_{j,1}(t))$ ;  $\bar{f}(\vec{X}_{j,2}) \leftarrow \frac{1}{T} \sum_{t=1}^T f(\vec{X}_{j,2}(t))$  ▷ Average  $f$  over  $T$ -traces
12:    end for
13:     $\hat{\mu}_i \leftarrow \frac{1}{2m_i} \sum_{i=1}^{m_i} (f(\vec{X}_{j,1}) + f(\vec{X}_{j,2}))$ ;  $\hat{v}_i \leftarrow \frac{1}{2m_i} \sum_{i=1}^{m_i} ((f(\vec{X}_{j,1}) - f(\vec{X}_{j,2})))^2$  ▷ Compute empirical mean; trace variance
14:     $u_i \leftarrow \hat{v}_i + \frac{(11 + \sqrt{21})(1 + \Lambda'/\sqrt{21})R^2 \ln \frac{3I}{\delta}}{(1 - \Lambda')m_i} + \sqrt{\frac{(1 + \Lambda')R^2 \hat{v}_i \ln \frac{3I}{\delta}}{(1 - \Lambda')m_i}}$  ▷ Variance upper bound
15:     $\hat{\varepsilon}_i^+ \leftarrow \frac{10R \ln \frac{3I}{\delta}}{(1 - \Lambda')m_i} + \sqrt{\frac{(1 + \Lambda')u_i \ln \frac{3I}{\delta}}{(1 - \Lambda')m_i}}$  ▷ Apply Bernstein bound
16:     $\hat{\mu}_i^\times \leftarrow \frac{(\hat{\mu}_i - \hat{\varepsilon}_i^+) \vee a + (\hat{\mu}_i + \hat{\varepsilon}_i^+) \wedge b}{2}$  ▷ Optimal mean estimate
17:     $\hat{\varepsilon}_i^\times \leftarrow \frac{((\hat{\mu}_i + \hat{\varepsilon}_i^+) \wedge b - (\hat{\mu}_i - \hat{\varepsilon}_i^+) \vee a)}{2\hat{\mu}_i^\times}$  ▷ Empirical relative error bound
18:    if  $(i = I) \vee (\hat{\varepsilon}_i^\times \leq \varepsilon)$  then ▷ Terminate if accuracy guarantee is met
19:      return  $\hat{\mu}_i^\times$ 
20:    end if
21:  end for
22: end procedure

```

Algorithm 2 SUPERCHAINTRACEGIBBS and PARALLELTRACEGIBBS

```

1: procedure SUPERCHAINTRACEGIBBS(...)
2:    $(\beta_0, \beta_1, \dots, \beta_\ell) \leftarrow \text{TPA}(k, d)^a$ 
3:    $\varepsilon' \leftarrow \frac{\varepsilon}{2+\varepsilon}$ ;  $\delta' \leftarrow \frac{\delta}{2}$ 
4:   for  $i \in 1, 2, \dots, \ell$  do
5:      $f_i(x) \doteq \exp(-\frac{\beta_{i+1}-\beta_i}{2} H(x))$ 
6:      $g_i(x) \doteq \exp(\frac{\beta_i-\beta_{i-1}}{2} H(x))$ 
7:   end for
8:    $F \doteq \otimes_{i=1}^\ell f_i$ ;  $G \doteq \otimes_{i=1}^\ell g_i$ 
9:    $\mathcal{G}^\otimes \leftarrow \otimes_{i=1}^\ell \mathcal{G}_{H, \beta_i}$ , with  $\omega_i = \frac{1}{\ell}$ ,  $\forall i$ 
10:   $R_f \leftarrow \exp(-\frac{\beta-\beta_0}{2} H_{\min}) - \exp(-\frac{\beta-\beta_0}{2} H_{\max})$ 
11:   $R_g \leftarrow \exp(\frac{\beta-\beta_0}{2} H_{\max}) - \exp(\frac{\beta-\beta_0}{2} H_{\min})$ 
12:   $\hat{\mu} \leftarrow \text{RELMEANEST}(\mathcal{G}^\otimes, R_f, T, F, \varepsilon', \delta')$ 
13:   $\hat{v} \leftarrow \text{RELMEANEST}(\mathcal{G}^\otimes, R_g, T, G, \varepsilon', \delta')$ 
14:  return  $\hat{Z} \leftarrow \frac{\hat{v}}{\hat{\mu}}$ 
15: end procedure

16: procedure PARALLELTRACEGIBBS(...)
17:    $(\beta_0, \beta_1, \dots, \beta_\ell) = \text{TPA}(k, d)$ 
18:    $\varepsilon' \leftarrow \frac{\varepsilon \sqrt{1+\varepsilon}-1}{\sqrt{1+\varepsilon}+1}$ ;  $\delta' \leftarrow \frac{\delta}{2\ell}$ 
19:   for  $i \in 1, 2, \dots, \ell$  do
20:      $f_i(x) \doteq \exp(-\frac{\beta_{i+1}-\beta_i}{2} H(x))$ 
21:      $g_{i-1}(x) \doteq \exp(\frac{\beta_i-\beta_{i-1}}{2} H(x))$ 
22:      $R_f \leftarrow \exp(-\frac{\beta_{i+1}-\beta_i}{2} H_{\min}) - \exp(-\frac{\beta_{i+1}-\beta_i}{2} H_{\max})$ 
23:      $R_g \leftarrow \exp(\frac{\beta_{i+1}-\beta_i}{2} H_{\max}) - \exp(\frac{\beta_{i+1}-\beta_i}{2} H_{\min})$ 
24:      $\hat{\mu}_i \leftarrow \text{RELMEANEST}(\mathcal{G}_i, R_f, T_i, f_i, \varepsilon', \delta')$ 
25:      $\hat{v}_i \leftarrow \text{RELMEANEST}(\mathcal{G}_i, R_g, T_i, g_i, \varepsilon', \delta')$ 
26:   end for
27:   return  $\hat{Z} \leftarrow \prod_{i=1}^\ell \frac{\hat{v}_i}{\hat{\mu}_i}$ 
28: end procedure

```

^a $k = \Theta(\log H_{\max})$ and $d = 64$ as in [6]

A Appendix

A.1 Algorithms used in the literature

A.1.1 The TPA method [4, 6]

We refer to Huber and Schott's algorithm as the original TPA, and Kolmogorov's, which is used in our algorithms and referred to as TPA(k, d) in the main manuscript, as the TPA method.

Algorithm 3 THE ORIGINAL TPA-METHOD [4]

```
1: output a schedule  $(\beta_1, \dots, \beta_l)$  of values in the interval  $[\beta_{\min}, \beta_{\max}]$ .
2:  $\beta_0 \leftarrow \beta_{\min}$ 
3: for  $i = 0 : \infty$  do
4:   sample  $X \sim \pi_{\beta_i}$  draw  $U \in [0, 1]$  uniformly,  $\beta_{i+1} = \beta_i - \log U / H(X)$  (or  $+\infty$  if  $H(X) = 0$ ).
5:   if  $\beta_{i+1} \notin [\beta_{\min}, \beta_{\max}]$  then Terminate
6:   end if
7: end for
```

Algorithm 4 TPA-METHOD [6]

```
1: input integers  $k$  and  $d$ 
2: output a schedule  $(\beta_0, \beta_1, \dots, \beta_l)$  of values in the interval  $[\beta_{\min}, \beta_{\max}]$ .
3: for  $i = 1 : k$  do
4:    $\mathcal{B}_i \leftarrow$  THE ORIGINAL TPA-METHOD().
5:   let  $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_i$ 
6: end for
7: sort  $\mathcal{B}$ , keep one sample uniformly from the initial  $d$  elements, and keep every  $d$ th successive value in the remaining sequence.
8: add  $\beta_{\min}$  and  $\beta_{\max}$  to  $\mathcal{B}$  return  $\mathcal{B}$ 
```

A.1.2 Single site Gibbs sampler (Glauber dynamics chain)

Consider β and H defined as above. Let $X = (X_1, X_2, \dots, X_n)$ be the set of all variables in the Gibbs distribution with inverse temperature β and Hamiltonian H , thus, the domain of H is $\Omega = \Omega_1 \times \Omega_2 \times \dots \times \Omega_n$, and each Ω_i is the range of random variable X_i . At each time step t , assume the current state is $x^{(t)} = (x_1, x_2, \dots, x_n)$. Take $i \sim 1, \dots, n$ uniformly at random. Sample y from the following distribution:

$$\pi_{\beta}(y|x_{-i}^{(t)}) = \frac{\exp(-\beta H(x^{(t)}; x_i \leftarrow y))}{\sum_{\omega \in \Omega_i} \exp(-\beta H(x^{(t)}; x_i \leftarrow \omega))}, \quad (1)$$

where for an arbitrary $\omega \in \Omega_i$ we define $(x^{(t)}; x_i \leftarrow \omega)$ be the vector in which all the elements except the i th element are equal to x_i and the i th element is replaced with ω .

In other words, for any arbitrary vectors $x^{(t)}$ and $x^{(t+1)}$, the transition probability is:

$$\mathcal{G}_{H,\beta}(x^{(t)}, x^{(t+1)}) = \begin{cases} (1/n)\pi_{\beta}(y|x_{-i}^{(t)}), & \exists y, i \text{ such that } x_i \neq y \text{ and } x^{(t+1)} = (x^{(t)}; x_i \leftarrow y), \\ \sum_{i=1}^n (1/n)\pi_{\beta}(x_i|x_{-i}^{(t)}) & \text{if } x^{(t)} = x^{(t+1)}, \\ 0 & \text{otherwise.} \end{cases}$$

A.2 Missing proofs: TPA and relative trace variance properties

Lemma A.1. Let $z(\beta) \doteq \log(Z(\beta))$, d and k the parameters of the TPA method, and β_i and β_{i+1} two consecutive points generated by TPA(k, d), we have:

1. For any $\varepsilon \geq 0$, we have $\mathbb{P}(z(\beta_j) - z(\beta_{j+1}) \leq \varepsilon) \geq (1 - \exp(-\varepsilon k/d))^d \simeq 1 - d \exp(-\varepsilon k/d)$,
2. For any $\varepsilon \geq 0$, $\mathbb{P}(\Delta_i \geq \varepsilon/\mathbb{E}[H(x)]) \leq d \exp(-\varepsilon k/d)$, where the expectation of $H(x)$ is taken with respect to distribution $x \sim \pi_{\beta_{i+1}}$.

Proof of lemma A.1. Note that TPA(k, d) of [6] consists of k parallel runs of the original TPA of [4] and outputting a sub-sequence of elements which are d apart.

Let (b_i) be the sequence generated by k parallel copies of the original TPA, thus $\Delta_j = \beta_{j+1} - \beta_j = b_{j+d} - b_j$.

We first show item 1 by bounding $\mathbb{P}(z(b_j) - z(b_{j+d}) \geq \varepsilon)$, and using

$$\mathbb{P}(b_{j+d} - b_j < \varepsilon) \geq \prod_{i=1}^d \mathbb{P}(b_{j+i} - b_{j+i-1} < \varepsilon/d).$$

With the definition of the PPP, and using [3] we have $z(b_i) - z(b_{i+1})$ follows the exponential distribution with mean $1/k$, thus $\mathbb{P}(z(b_i) - z(b_{i+1}) \geq \varepsilon/d) = \exp(-\varepsilon k/d)$. Therefore,

$$\mathbb{P}(z(b_{j+d}) - z(b_j) < \varepsilon) \geq \prod_{i=1}^d \mathbb{P}(z(b_{j+i}) - z(b_{j+i-1}) < \varepsilon/d) = (1 - \exp(-\varepsilon k/d))^d.$$

To see item 2 of the Lemma let $z'(\beta)$ be the derivative of $z(\cdot)$ with respect to β , which is $z'(\beta) = \sum_{x \in \Omega} -H(x) \exp(-\beta H(x)) / Z(\beta)$, thus $z'(\beta) \leq 0$. Using the Cauchy-Schwarz inequality we have $z''(\beta) = (\sum_{x \in \Omega} H^2(x) \exp(-\beta H(x)) \sum_{x \in \Omega} \exp(-\beta H(x)) - (\sum_{x \in \Omega} -H(x) \exp(-\beta H(x)))^2) / Z^2(\beta) \geq 0$. Therefore,

$$z'(\beta_i) < \frac{z(\beta_{i+1}) - z(\beta_i)}{\beta_{i+1} - \beta_i} < z'(\beta_{i+1}),$$

Thus, $\beta_{i+1} - \beta_i < \frac{z(\beta_i) - z(\beta_{i+1})}{-z'(\beta_i)}$. Note that $-z'(\beta_i) = \mathbb{E}[H(x)]$, $x \sim \pi_{\beta_i}$. Therefore, we have:

$$\begin{aligned} \mathbb{P}\left(\Delta_i \leq \frac{\epsilon}{\mathbb{E}[H]}\right) &\geq \mathbb{P}\left(\frac{z(\beta_i) - z(\beta_{i+1})}{-z'(\beta_{i+1})} \leq \frac{\epsilon}{\mathbb{E}[H]}\right) \\ &= \mathbb{P}(z(\beta_i) - z(\beta_{i+1}) \leq \epsilon) \\ &\geq (1 - \exp(-\epsilon k/d))^d \end{aligned}$$

Thus $\mathbb{P}\left(\Delta_i \geq \frac{\epsilon}{\mathbb{E}[H]}\right) \geq 1 - (1 - \exp(-\epsilon k/d))^d \approx d \exp(-\epsilon k/d)$. □

Proof of Lemma 2.1. Note that by Thm 3.1. of [9] we have, $\mathbb{E}[(\bar{f}(\vec{X}_{1:\tau}) - \mathbb{E}(f))^2] \leq \frac{2\tau_{\text{rx}}}{\tau} \mathbb{V}[f]$. Dividing both sides by $(\mathbb{E}(f))^2$ we get the second part of the premise. The first part concludes from setting $\tau = \tau_{\text{rx}}$. □

A.3 RELMEANEST

RELMEANEST in summary To employ progressive sampling, we start by a small sample size and calculate the *empirical estimation* of the variance at each iteration. We estimate an upper bound on the trace variance based on its empirical estimation, and using that we check a termination condition.

Our variance estimator is what Cousins et al. introduced, and is based on running two independent chains. Each sample is obtained by taking a trace of length T (given upper-bound on relaxation time) and taking the average over all observed values on that trace. Thus, *half the square difference* of the averages on the two chains is an *unbiased estimate* of the *trace variance*.

Before showing the result, we state two key theorems from the literature, which describe how our tail bounds work.

Theorem A.2 (Hoeffding-Type Bounds for Mixing Processes, (see Thm. 2.1 of [2])). *For any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \sqrt{\frac{2(1+\lambda)(\frac{R^2}{4}) \ln(\frac{2}{\delta})}{(1-\lambda)m}}\right) \leq \delta. \quad (2)$$

This implies sample complexity

$$m_H(\lambda, R, \varepsilon, \delta) = \frac{1 + \lambda}{1 - \lambda} \ln\left(\frac{2}{\delta}\right) \frac{R^2}{2\varepsilon^2} \in \Theta\left(\tau_{\text{rx}} \ln\left(\frac{1}{\delta}\right) \frac{R^2}{\varepsilon^2}\right).$$

Theorem A.3 (Bernstein-Type Bound for Mixing Process [5, Thm. 1.2]). *For any $\delta \in (0, 1)$, we have*

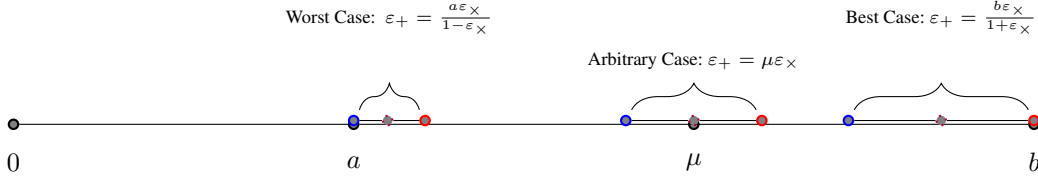
$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \frac{10R \ln\left(\frac{2}{\delta}\right)}{(1 - \lambda)m} + \sqrt{\frac{2(1 + \lambda)v_\pi \ln\left(\frac{2}{\delta}\right)}{(1 - \lambda)m}}\right) \leq \delta. \quad (3)$$

This implies sample complexity

$$m_B(\lambda, R, v, \varepsilon, \delta) = \frac{2}{1 - \lambda} \ln\left(\frac{2}{\delta}\right) \left(\frac{5R}{\varepsilon} + \frac{(1 + \lambda)v_\pi}{\varepsilon^2}\right) \in \Theta\left(\tau_{\text{rx}} \ln\left(\frac{1}{\delta}\right) \left(\frac{R}{\varepsilon} + \frac{v_\pi}{\varepsilon^2}\right)\right).$$

We now show the main result.

Proof of Theorem 2.2. Suppose confidence interval $[a, b]$. The interval endpoints, multiplicative error ε_\times , and additive error ε_+ are related as $2\varepsilon_+ = a \frac{1 + \varepsilon_\times}{1 - \varepsilon_\times} - a = a \frac{2\varepsilon_\times}{1 - \varepsilon_\times}$, depicted graphically below.



We derive a geometric progressive sampling schedule such that the algorithm draws sample sizes, ranging between optimistic and pessimistic (over unknown variance and mean) upper and lower bounds on the sufficient sample size.

Using the Markov chain Bennett inequality [5], the best-case complexity, assuming maximal expectation, and minimal variance, is

$$\begin{aligned} m^\downarrow &\geq m_B(\Lambda, R, 0, \varepsilon_+, \frac{2\delta}{3I}) \\ &\geq \frac{(1 + \Lambda)R \ln \frac{3I}{\delta}}{(1 - \Lambda)\varepsilon_+} = \frac{(1 + \Lambda)R \ln \frac{3I}{\delta} (1 + \varepsilon_\times)}{b(1 - \Lambda)\varepsilon_\times}. \end{aligned}$$

The worst-case complexity, then assuming minimal expectation, and maximal variance, is

$$\begin{aligned} m^\uparrow &\geq m_H(\Lambda, R, \varepsilon_+, \frac{2\delta}{3I}) \\ &\geq \frac{(1 + \Lambda)R^2 \ln \frac{3I}{\delta}}{2(1 - \Lambda)\varepsilon_+^2} = \frac{(1 + \Lambda)R^2 \ln \frac{3I}{\delta} (1 - \varepsilon_\times)^2}{2(1 - \Lambda)a^2\varepsilon_\times^2}, \end{aligned}$$

via the Markov chain Hoeffding's inequality [7].

Consequently, a doubling schedule requires $I = \left\lceil \log_2 \left(\frac{m^\uparrow}{m^\downarrow} \right) \right\rceil = \left\lceil \log_2 \left(\frac{bR}{2a^2} \cdot \frac{(1 - \varepsilon_\times)^2}{(1 + \varepsilon_\times)\varepsilon_\times} \right) \right\rceil$ steps.

All tail bounds on variances and means are hold simultaneously with probability at least $1 - \delta$ (by union bound), and the doubling schedule never overshoots the sufficient sample size by more than a constant factor, which yields the stated guarantees.

The proof consists of two parts, in both we make derive our new bounds by writing an ε_\times -multiplicative approximation in terms of an ε_+ -additive approximation.

In the worst-case, we *underestimate* the true mean μ by a factor $(1 - \varepsilon_\times)$, and thus require a radius $\varepsilon_+ = \varepsilon_\times(1 - \varepsilon_\times)\mu$ additive confidence interval.

We first show the *correctness guarantee*.

Observe that the sampling schedule is selected such that the final iteration I of the algorithm will draw a sufficiently large sample (size m^\uparrow) such that the Hoeffding inequality will yield such a confidence interval, even for worst-case (minimal) μ . Now observe that over the course of the algorithm, in each iteration, 3 tail bounds are applied; one to upper-bound the variance, and then two to upper and lower bound the mean in terms of the variance bound) as in [1]. By union bound, all $3I$ tail-bounds hold simultaneously with probability at least $1 - \delta$, thus when the algorithm terminates, it produces a correct answer with at least said probability.

We now show the *efficiency guarantee*. Suppose we get $\hat{\mu}$ from RELMEANEST, by guarantee of correctness of the algorithm, we have a lower bound on $\hat{\mu}$, $\hat{\mu} \geq \mu(1 - \varepsilon_x)$ with probability at least $1 - \delta$.

Furthermore, we have $\varepsilon_+ = \mu\varepsilon_x$ and $\text{trv}^{(\tau_{\text{rx}})} = (\text{Reltrv}^{\tau_{\text{rx}}} - 1) \times \hat{\mu}^2 \geq (\text{Reltrv}^{\tau_{\text{rx}}} - 1)\mu^2(1 - \varepsilon_x)^2$. For this ε_+ , we have via the Bernstein inequality that

$$m^* \in \mathcal{O} \left(\log \left(\frac{\log(R/(\mu\varepsilon_x))}{\delta} \right) \left(\frac{R/\mu}{(1 - \Lambda)\varepsilon_x} + \frac{\tau_{\text{rx}}(\text{Reltrv}^{\tau_{\text{rx}}} - 1)}{\varepsilon_x^2} \right) \right)$$

would be a sufficient sample size if (1) the algorithm were to draw a sample of this size, and (2) we were to use the *true trace variance* instead of the *estimated upper-bound on trace variance*.

Fortunately, correcting for (1) adds a constant factor to the sample complexity, as the first sample size α is selected to be twice the minimal sufficient sample size m^\downarrow (i.e., the sample size such that no smaller sample size would be sufficient), and at each iteration the sample size selected is double the previous (line 8). In other words, this geometric grid will never overshoot any sample size by more than a factor 2.

Resolving (2) is a bit more subtle, but we now show that there is no asymptotic change in replacing the variance with the estimated variance upper bound (w.h.p.). First, note that the Bernstein bound is *bidirectional*, so it can just as well be used to upper-bound empirical variance with true variance as to upper-bound true variance with empirical variance. We bound true variance in terms of empirical variance on line 14, and note that here we have

$$v \leq u \in \hat{v} + \mathcal{O} \left(\frac{R^2 \ln \frac{I}{\delta}}{m} + \sqrt{\frac{R^2 \hat{v} \ln \frac{I}{\delta}}{m}} \right).$$

Fortunately, the latter terms are negligible, as in line 15, we bound

$$\begin{aligned} \varepsilon_+ &\in \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{u \ln \frac{I}{\delta}}{m}} \right) \\ &= \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{\left(\hat{v} + \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{\hat{v} \ln \frac{I}{\delta}}{m}} \right) \right) \ln \frac{I}{\delta}}{m}} \right) \\ &= \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{\left(v + \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{v \ln \frac{I}{\delta}}{m}} \right) + \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{\hat{v} \ln \frac{I}{\delta}}{m}} \right) \right) \ln \frac{I}{\delta}}{m}} \right) \quad (\text{w.h.p.}) \\ &= \mathcal{O} \left(\frac{R \ln \frac{I}{\delta}}{m} + \sqrt{\frac{v \ln \frac{I}{\delta}}{m}} \right) \quad (\text{w.h.p.}) \end{aligned}$$

Putting these together, we thus have that, w.h.p., sample consumption is bounded as

$$\hat{m} \in 2\mathcal{O}(m^*) = \mathcal{O} \left(\log \left(\frac{\log(R/(\mu\varepsilon_x))}{\delta} \right) \left(\frac{R/\mu}{(1-\Lambda)\varepsilon_x} + \frac{\tau_{\text{rx}}(\text{Reltrv}^{\tau_{\text{rx}}}-1)}{\varepsilon_x^2} \right) \right).$$

To conclude, we need only relate $T(\text{Reltrv}^T - 1)$ and $\tau_{\text{rx}}(\text{Reltrv}^{\tau_{\text{rx}}}-1)$. Letting T as in line ??, note that since $T \geq \tau_{\text{rx}}$, it holds that $T(\text{Reltrv}^T - 1) \geq \tau_{\text{rx}}(\text{Reltrv}^{\tau_{\text{rx}}}-1)$, by the trace variance inequalities, which yields the result. \square

A.4 Missing proofs from analysis of SUPERCHAINTRACEGIBBS

Proof of thm 2.4. Follows immediately from thm. 2.2 and plugging in the values for paired product estimators and the product chain. \square

Full Proof of Lemma 2.8. Let $\bar{\beta}_{i,i+1} \doteq \frac{\beta_i + \beta_{i+1}}{2}$, we have $\mu_i = \frac{Z(\bar{\beta}_{i,i+1})}{Z(\beta_i)}$ and $\nu_i = \frac{Z(\bar{\beta}_{i,i+1})}{Z(\beta_{i+1})}$. Thus we have $\nu = \frac{\prod_{i=1}^{\ell-1} Z(\bar{\beta}_{i,i+1})}{\prod_{i=1}^{\ell-1} Z(\beta_{i+1})} > 1$, $\mu = \frac{\prod_{i=1}^{\ell-1} Z(\bar{\beta}_{i,i+1})}{\prod_{i=1}^{\ell-1} Z(\beta_i)} < 1$.

Note that $\nu = \mu \frac{Z(\beta_0)}{Z(\beta_{\max})}$, thus we proceed by bounding μ .

$$\begin{aligned} \log \prod_{i=1}^{\ell-1} Z(\bar{\beta}_{i,i+1}) &= \sum_{i=1}^{\ell-1} z(\bar{\beta}_{i,i+1}) && \text{TAKING log} \\ &\geq \sum_{i=1}^{\ell-1} z(\beta_i) - \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)] && \text{TAYLOR EXPANSION \& THAT } \frac{\partial^2}{\partial \beta^2} z(\beta) > 0 \end{aligned}$$

Thus, by taking exponents we get:

$$\begin{aligned} \prod_{i=1}^{\ell-1} Z(\bar{\beta}_{i,i+1}) &\geq \exp \left(\sum_{i=1}^{\ell-1} z(\beta_i) - \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)] \right) \\ &\geq \left(\prod_{i=1}^{\ell-1} Z(\beta_i) \right) \exp \left(- \sum_{i=1}^{\ell-1} \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)] \right) \end{aligned}$$

Therefore, $\mu = \frac{\prod_{i=1}^{\ell-1} Z(\bar{\beta}_{i,i+1})}{\prod_{i=1}^{\ell-1} Z(\beta_i)} \geq \exp \left(- \sum_{i=1}^{\ell-1} \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)] \right)$. Using this form, we now employ the fundamental theorem of calculus to prove the premise:

Let $\Delta_{\max} \doteq \max_i \Delta_i$.

$$\begin{aligned}
\mu &\geq \exp\left(-\sum_{i=1}^{\ell-1} \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)]\right) \\
&= \exp\left(-\sum_{i=1}^{\ell-1} \frac{\Delta_i}{2} \mathbb{E}_{x \sim \pi_{\beta_i}} [H(x)]\right) \\
&\geq \exp\left(\frac{1}{2} \int_{\beta_{\min} - \Delta_{\max}}^{\beta_{\max} - \Delta_{\max}} - \mathbb{E}_{x \sim \pi_{\beta}} [H(x)] d\beta\right) && \text{INCREASING INTEGRAND} \\
&= \exp\left(\frac{1}{2} (z(\beta_{\max} - \Delta_{\max}) - z(\beta_{\min} - \Delta_{\max}))\right) && \text{FTOC AND THAT } z'(\beta) = \mathbb{E}_{x \sim \pi_{\beta}} H \\
&\geq \exp\left(\frac{1}{2} (z(\beta_{\max}) - z(\beta_{\min} - \Delta_{\max}))\right) && z \text{ IS DECREASING} \\
&= \exp\left(\frac{1}{2} (z(\beta_{\max}) - z(\beta_{\min}) + z(\beta_{\min}) - z(\beta_{\min} - \Delta_{\max}))\right) \\
&\geq Q^{-\frac{1}{2}} \sqrt{\frac{Z(\beta_{\min})}{Z(\beta_{\min} - \Delta_{\max})}}.
\end{aligned}$$

From the above we also conclude that $\nu \geq Q^{1/2} \sqrt{\frac{Z(\beta_{\min})}{Z(\beta_{\min} - \Delta_{\max})}}$. Note that $\text{Range}(f) = \exp(-\frac{\Delta}{2} H_{\min}) - \exp(-\frac{\Delta}{2} H_{\max}) \leq \sqrt{\exp(-\Delta H_{\min})}$ and $\text{Range}(g) = \exp(\frac{\Delta}{2} H_{\max}) - \exp(\frac{\Delta}{2} H_{\min}) \leq \sqrt{\exp(\Delta H_{\max})}$. Thus the lemma is concluded. \square

Proof of Corollary 2.6. The corollary follows from thm 2.2 plugging in R from lemma 2.5 and setting $\tau_{\text{prx}} = \ell \max_{i=1}^{\ell} \tau_i$ (see, e.g., [8]). \square

A.5 Analysis of PARALLELTRACEGIBBS

Let $(\beta_0, \beta_1, \dots, \beta_{\ell})$ be a cooling schedule generated by TPA(k, d), where k and d are chosen as in [6]. For each i let $f_{\beta_i, \beta_{i+1}}, g_{\beta_{i-1}, \beta_i}$ be the paired estimators corresponding to this schedule, and $\mu_i = \mathbb{E}[f_{\beta_i, \beta_{i+1}}], \nu_i = \mathbb{E}[g_{\beta_{i-1}, \beta_i}]$. PARALLELTRACEGIBBS estimates Q by running RELMEANESTON each \mathcal{G}_{H, β_i} , to estimate μ_i and ν_i s each with precision $\varepsilon' = (\sqrt[1+\varepsilon]{1-\varepsilon}) / (\sqrt[1+\varepsilon]{1+\varepsilon})$. Note that by this setting, Q will be approximated within multiplicative factor of $(1+\varepsilon'/1-\varepsilon')^{\ell}$. Assume τ_i is the true relaxation time of \mathcal{G}_{H, β_i} and suppose Λ_i is a known upper bound on the second eigenvalue of \mathcal{G}_{H, β_i} , thus $(\Lambda_i - 1)^{-1} \log(2) \geq \tau_i$. The following hold and thm 2.7 is immediately concluded from it:

Lemma A.4. Let $H_{\max} \doteq \max_{x \in \Omega} H(x)$. we have:

1. for all $1 \leq i \leq \ell$, $\text{Range}(f_{\beta_i, \beta_{i+1}}) / \mu_i \leq \ell^{1/\log(n)}$,
2. for all $1 \leq i \leq \ell$, $\text{Range}(g_{\beta_{i-1}, \beta_i}) / \nu_i \leq \ell^{\alpha_0(i)/\log n}$, where $\alpha_0(i) = (H_{\max}/2\mathbb{E}[H(x)]) - 1$, $x \sim \pi_{\beta_i}$.

Proof. Let $\Delta_i = \beta_{i+1} - \beta_i$. Thus, $f_i(x) = \exp\left(\frac{-\Delta_i}{2} H(x)\right)$ and $g_i(x) = \exp\left(\frac{\Delta_i}{2} H(x)\right)$. So we have:

$$\text{Range}(f_i) = \exp\left(\frac{-\Delta_i}{2} \min_x H(x)\right) - \exp\left(\frac{-\Delta_i}{2} \max_x H(x)\right) \leq \exp\left(\frac{-\Delta_i}{2} \min_x H(x)\right)$$

and

$$\text{Range}(g_i) = \exp\left(\frac{\Delta_i}{2} \max_x H(x)\right) - \exp\left(\frac{\Delta_i}{2} \min_x H(x)\right) \leq \exp\left(\frac{\Delta_i}{2} \max_x H(x)\right)$$

$$\mu_i = Z(\beta_i + \Delta_i/2)/Z(\beta_i) \quad \& \quad \nu_i = Z(\beta_{i+1} - \Delta_i/2)/Z(\beta_{i+1})$$

$$\frac{\text{Range}(f_i)}{\mu_i} \leq \frac{\exp\left(-\frac{\Delta_i}{2} \min_x H(x)\right)}{\exp\left(z(\beta_i + \Delta_i/2) - z(\beta_i)\right)}, \quad \frac{\text{Range}(g_i)}{\nu_i} \leq \frac{\exp\left(\frac{\Delta_i}{2} \max_x H(x)\right)}{\exp\left(z(\beta_{i+1} - \Delta_i/2) - z(\beta_{i+1})\right)} \quad (4)$$

Writing $\Delta_i/2 = \frac{\Delta_i/2}{z(\beta_i + \Delta_i/2) - z(\beta_i)}(z(\beta_i + \Delta_i/2) - z(\beta_i))$, we get:

$$\begin{aligned} \frac{\text{Range}(f_i)}{\mu_i} &\leq \exp\left(-\frac{\Delta_i}{2} \min_x H(x) - (z(\beta_i + \Delta_i/2) - z(\beta_i))\right) \\ &\leq \exp\left((z(\beta_i + \Delta_i/2) - z(\beta_i)) \left(\frac{-\Delta_i \cdot \min_x H(x)}{2(z(\beta_i + \Delta_i/2) - z(\beta_i))} - 1\right)\right) \end{aligned}$$

and

$$\frac{\text{Range}(g_i)}{\nu_i} \leq \exp\left((z(\beta_{i+1} - \Delta_i/2) - z(\beta_{i+1})) \left(\frac{\Delta_i \cdot \max_x H(x)}{2(z(\beta_{i+1} - \Delta_i/2) - z(\beta_{i+1}))} - 1\right)\right)$$

Let z' and z'' be the first and second derivative of z with respect to β . Note that $z'(\beta) = \mathbb{E}_{x \sim \pi_\beta}[-H(x)]$. Since $z'' \geq 0$ we have:

$$z'(\beta_i) < \frac{z(\beta_i + \Delta_i/2) - z(\beta_i)}{\Delta_i/2} < z'(\beta_i + \Delta_i/2)$$

and

$$z'(\beta_{i+1} - \Delta_i/2) < \frac{z(\beta_{i+1}) - z(\beta_{i+1} - \Delta_i/2)}{\Delta_i/2} < z'(\beta_{i+1}).$$

Which are equivalent to $\frac{1}{z'(\beta_i + \Delta_i/2)} \leq \frac{\Delta_i/2}{z(\beta_i + \Delta_i/2) - z(\beta_i)} \leq \frac{1}{z'(\beta_i)}$ and $\frac{1}{z'(\beta_{i+1})} \leq \frac{\Delta_i/2}{z(\beta_{i+1}) - z(\beta_{i+1} - \Delta_i/2)} \leq \frac{1}{z'(\beta_{i+1} - \Delta_i/2)}$.

Therefore,

$$\frac{\text{Range}(f_i)}{\mu_i} \leq \exp\left((z(\beta_i + \Delta_i/2) - z(\beta_i)) \left(\frac{-\min_x H(x)}{2} \frac{1}{z'(\beta_i)} - 1\right)\right) \quad (5)$$

$$= \exp\left((z(\beta_i + \Delta_i/2) - z(\beta_i)) \left(\frac{\min_x H(x)}{2} \frac{1}{\mathbb{E}[H]} - 1\right)\right) \quad (6)$$

$$\leq \exp\left(z(\beta_i) - z(\beta_i + \Delta_i/2)\right) \quad (7)$$

Similarly for range of g_i s we have:

$$\frac{\text{Range}(g_i)}{\nu_i} \leq \exp\left((z(\beta_{i+1} - \Delta_i/2) - z(\beta_{i+1})) \left(\frac{-\max_x H(x)}{2} \frac{1}{z'(\beta_i)} - 1\right)\right) \quad (8)$$

$$\leq \exp\left((z(\beta_{i+1} - \Delta_i/2) - z(\beta_{i+1})) \left(\frac{\max_x H(x)}{2\mathbb{E}_{\pi_{\beta_i}}[H(X)]} - 1\right)\right) \quad (9)$$

We now use (5) together with lemma A.1. Setting $d = 1$ we have,

$$\begin{aligned} \mathbb{P}\left(z(\beta_i) - z(\beta_{i+1}) > \frac{\log(3l/4)}{\log n}\right) &= \exp\left(-\frac{\log(3l/4)}{\log n} \cdot k\right) = (3/4) \exp(-\log l / \log n(\log n)) \\ &= (3/4)(1/l). \end{aligned}$$

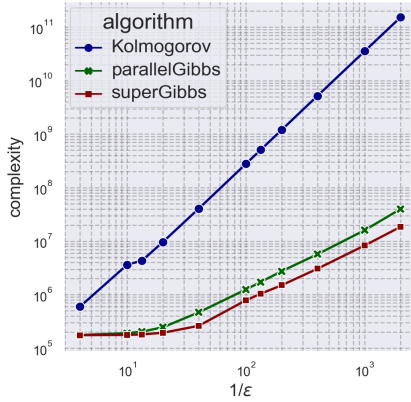
Using union bound over all $1 \leq i \leq \ell$ and that $z(\beta_i) - z(\beta_{i+1}) \geq z(\beta_i) - z(\beta_i + \Delta_i/2)$, we conclude that with probability at least $3/4$ we have that for all f_i , $\text{Range}(f_i)/\mu_i \leq \ell^{1/\log(n)}$.

Similarly using (8), the union bound, lemma A.1 and that $z(\beta_i) - z(\beta_{i+1}) \geq z(\beta_i - \Delta_i/2) - z(\beta_{i+1})$, we can show that with constant probability all g_i s generated by the TPA schedule obey: $\forall g_i; 1 \leq i \leq \ell$, $\text{Range}(g_i)/\nu_i \leq \exp((\log \ell / \log n) \cdot (\alpha)) = \ell^{\alpha_0 / \log n}$, where $\alpha_0 = \frac{\max_x H(x)}{2\mathbb{E}_{\pi_{\beta_i}}[H(X)]} - 1$. \square

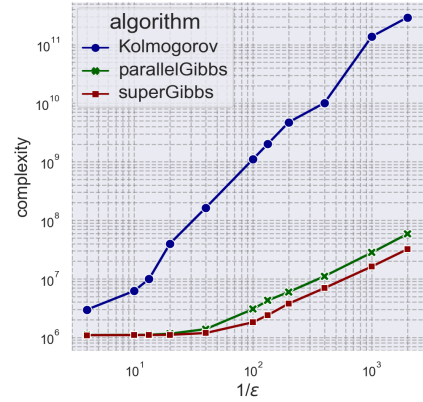
The following corollary is concluded from lemma A.4 and relative trace variance bounds:

Corollary A.5. *When $\varepsilon \leq \ell^{1/\log(n)}(1 + \ell^{\alpha_0(i)}) \cdot \frac{\ell\tau_{\beta_i}}{(1-\Lambda_i)^{-1}}$, RELMEANEST invoked on the i th iteration will stop using sample consumption of $\tilde{O}(\ell^2\tau_i\text{Reltrv}_i)$ note that this is improvement over classic bounds which are $\tilde{O}((1-\Lambda_i)^{-1}\mathbb{V}\text{rel}_i)$. In total the sample complexity of PARALLELTRACEGIBBS for $\varepsilon \leq \ell^{1/\log(n)}\min_i(1 + \ell^{\alpha_0(i)}) \cdot \frac{\ell\tau_{\beta_i}}{(1-\Lambda_i)^{-1}}$ is dominated by $\tilde{O}(\ell^2\sum_{i=1}^{\ell}\tau_i\text{Reltrv}_i)$.*

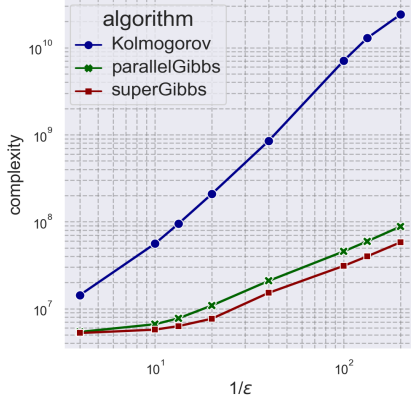
A.6 Further experimental results



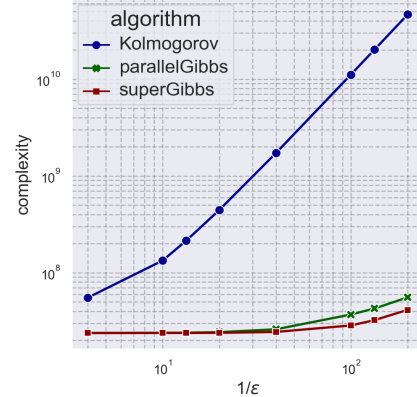
(a) $\beta = .05$, 2×2 lattice



(b) $\beta = .01$, 3×3 lattice



(c) $\beta = .02$, 4×4 lattice



(d) $\beta = .002$, 6×6 lattice

Figure 1: Comparison of sample complexity on Ising models.

References

- [1] C. Cousins, S. Haddadan, and E. Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynAMITE. *CoRR*, abs/2011.11129, 2020.

- [2] J. Fan, B. Jiang, and Q. Sun. Hoeffding's lemma for Markov chains and its applications to statistical learning. *arXiv:1802.00211*, 2018.
- [3] D. G. Harris and V. Kolmogorov. Parameter estimation for Gibbs distributions. *CoRR*, abs/2007.10824, 2020.
- [4] M. Huber, S. Schott, et al. Using TPA for Bayesian inference. *Bayesian Statistics*, 9:257–282, 2010.
- [5] B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general Markov chains. *arXiv:1805.10721*, 2018.
- [6] V. Kolmogorov. A faster approximation algorithm for the Gibbs partition function. In *Conference On Learning Theory*, pages 228–249. PMLR, 2018.
- [7] C. Leon and F. Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *The Annals of Applied Probability*, 14, 05 2004.
- [8] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- [9] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20, 2015.