Complexity Analysis of Normalizing Constant Estimation: from Jarzynski Equality to Annealed Importance Sampling and beyond

Wei Guo Molei Tao Yongxin Chen

Estimating Normalizing Constant (Partition Function, Free Energy)

Task: given an unnormalized probability density $\pi \propto \mathrm{e}^{-V}$, estimate its normalizing constant (a.k.a. partition function) $Z = \int_{\mathbb{R}^d} \mathrm{e}^{-V(x)}\mathrm{d}x$ or free energy $F = -\log Z$.

As a crucial problem in Bayesian statistics, statistical mechanics, and machine learning, it is challenging in high dimensions or when π is multimodal.

Importance sampling: with a prior $\mu = \frac{1}{Z_\mu}\mathrm{e}^{-U}$, we have the equality $\frac{Z_\pi}{Z_\mu} = \frac{1}{Z_\mu}\int \mathrm{e}^{-V}\mathrm{d}x = \mathbb{E}_\mu \frac{\mathrm{e}^{-V}}{\mathrm{e}^{-U}}$. Hence the ratio can be estimated by sampling from μ. However, this estimator suffers from high variance due to the mismatch between μ and π.

Annealing for Addressing Multimodality

Annealing: construct a sequence of intermediate distributions that bridge the target and the prior distributions. This idea motivates several popular methods:

- In statistics: path sampling, **annealed importance sampling**, sequential Monte Carlo, etc.
- In thermodynamics: thermodynamic integration, **Jarzynski equality**, etc.

Contributions: we aim to establish a rigorous non-asymptotic analysis of estimators based on JE and AIS, while introducing minimal assumptions on the target distribution. We also propose a new algorithm based on reverse diffusion samplers (RDS) to tackle a potential shortcoming of AIS.

Wasserstein Distance, Metric Derivative, and Action

For probability measures μ, ν on $\mathbb{R}^d$, the **Wasserstein-2 distance** is defined as $W_2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int \|x-y\|^2\gamma(\mathrm{d}x,\mathrm{d}y)\right)^{\frac{1}{2}}$, where $\Pi(\mu,\nu)$ is the set of all couplings of (μ,ν).

A vector field $v = (v_t : \mathbb{R}^d \to \mathbb{R}^d)_{t\in[a,b]}$ on $\mathbb{R}^d$ **generates** a curve of probability measures $\rho = (\rho_t)_{t\in[a,b]}$ if the **continuity equation** $\partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0$, $t \in [a,b]$ holds.

The **metric derivative** of ρ at $t \in [a,b]$ is defined as $|\dot\rho|_t := \lim_{\delta\to 0}\frac{W_2(\rho_{t+\delta},\rho_t)}{|\delta|}$, which can be interpreted as the "speed" of this curve. If $|\dot\rho|_t$ exists and is finite for a.e. $t\in[a,b]$, we say that ρ is **absolutely continuous (AC)**. Its **action** is defined as $\int_a^b |\dot\rho|_t^2\mathrm{d}t$, which is a key property characterizing the effectiveness of a curve in annealed sampling.

■ Lemma (Relationship between Metric Derivative and Continuity Equation [1])

For an AC curve of probability measures $(\rho_t)_{t\in[a,b]}$, any vector field $(v_t)_{t\in[a,b]}$ that generates $(\rho_t)_{t\in[a,b]}$ satisfies $|\dot\rho|_t \le \|v_t\|_{L^2(\rho_t)}$ for a.e. $t\in[a,b]$. Moreover, there exists a unique vector field $(v_t^*)_{t\in[a,b]}$ generating $(\rho_t)_{t\in[a,b]}$ that satisfies $|\dot\rho|_t = \|v_t^*\|_{L^2(\rho_t)}$ for a.e. $t\in[a,b]$.

Problem Setting

Criterion: given an accuracy threshold ε, study the oracle complexity required to obtain an estimator $\widehat Z$ of Z such that $\Pr\left(\left|\frac{\widehat Z}{Z}-1\right|\le\varepsilon\right)\ge\frac{3}{4}$. Note that the probability can be boosted to any $1-\zeta$ using the median trick.

Annealing curve: we define a curve of probability measures $\left(\pi_\theta = \frac{1}{Z_\theta}\mathrm{e}^{-V_\theta}\right)_{\theta\in[0,1]}$ from a prior distribution to the target distribution. $Z_1 = Z$ is what we need to estimate.

- **Assump. 1**: the potential $[0,1]\times\mathbb{R}^d \ni (\theta,x)\mapsto V_\theta(x)\in\mathbb{R}$ is jointly C^1, and the curve $(\pi_\theta)_{\theta\in[0,1]}$ is AC with finite action $\mathcal{A} := \int_0^1 |\dot\pi|_\theta^2\mathrm{d}\theta$.
- **Assump. 2**: V is β-smooth, and there exists x_*, with $\|x_*\| =: R \lesssim \frac{1}{\sqrt\beta}$ s.t. $\nabla V(x_*) = 0$. Let $m := \sqrt{\mathbb{E}_\pi\|\cdot\|^2} < +\infty$.

Analysis of the Jarzynski Equality (JE)

We introduce a reparameterized curve $(\widetilde\pi_t = \pi_{\frac{t}{T}})_{t\in[0,T]}$ for some large T to be determined later. **Annealed Langevin diffusion (ALD)**:

$$\mathrm{d}X_t = \nabla\log\widetilde\pi_t(X_t)\mathrm{d}t + \sqrt2\mathrm{d}B_t,\ t\in[0,T];\ X_0\sim\widetilde\pi_0.$$

▶ Jarzynski Equality (JE) [5]

Let $\mathbb{P}^\rightarrow$ be the path measure of ALD. Then the following relation between the work functional W and free energy difference ΔF holds:

$$\mathbb{E}_{\mathbb{P}^\rightarrow}\mathrm{e}^{-W} = \mathrm{e}^{-\Delta F},\quad \textit{where } W(X) := \frac{1}{T}\int_0^T \partial_\theta V_\theta|_{\theta=\frac{t}{T}}(X_t)\mathrm{d}t,\ \textit{and}\ \Delta F := -\log\frac{Z_1}{Z_0}.$$

■ Theorem (Convergence Guarantee of JE)

$\widehat Z := Z_0\mathrm{e}^{-W(X)}$ with $X\sim\mathbb{P}^\rightarrow$ is an unbiased estimator of $Z = Z_0\mathrm{e}^{-\Delta F}$. Under Assump. 1, it suffices to choose $T = \frac{32\mathcal{A}}{\varepsilon^2}$ to obtain $\Pr\left(\left|\frac{\widehat Z}{Z}-1\right|\le\varepsilon\right)\ge\frac34$.

Analysis of the Annealed Importance Sampling (AIS)

▶ Annealed Importance Sampling (AIS) Equality [6]

Suppose we have probability distributions $\pi_\ell = \frac{1}{Z_\ell}f_\ell$, $\ell\in[\![0,M]\!]$ and transition kernels $F_\ell(x,\cdot)$, $\ell\in[\![1,M]\!]$, and assume that each π_ℓ is an invariant distribution of F_ℓ, $\ell\in[\![1,M]\!]$. Define the path measure $\mathbb{P}^\rightarrow(x_{0:M}) = \pi_0(x_0)\prod_{\ell=1}^M F_\ell(x_{\ell-1},x_\ell)$. Then we have

$$\mathbb{E}_{\mathbb{P}^\rightarrow}\mathrm{e}^{-W} = \mathrm{e}^{-\Delta F},\quad \textit{where } W(x_{0:M}) := \log\prod_{\ell=0}^{M-1}\frac{f_\ell(x_\ell)}{f_{\ell+1}(x_\ell)}\ \textit{and}\ \Delta F := -\log\frac{Z_M}{Z_0}.$$

For non-asymptotic analysis, we focus on the **geometric interpolation**:

$$\pi_\theta = \frac{1}{Z_\theta}f_\theta = \frac{1}{Z_\theta}\exp\left(-V - \frac{\lambda(\theta)}{2}\|\cdot\|^2\right),\ \theta\in[0,1]:\ \lambda(0) = 2\beta\searrow\lambda(1) = 0.$$

Introduce discrete time points $0 = \theta_0 < \theta_1 < \ldots < \theta_M = 1$, and define F_ℓ as running LD targeting π_{θ_ℓ} for time T_ℓ. In practice, we approximate this by running *one step* of **annealed Langevin Monte Carlo (ALMC)** using the exponential integrator discretization scheme with step size T_ℓ.

■ Theorem (Convergence Guarantee of AIS)

Under Assumps. 1 and 2, consider the annealing schedule $\lambda(\theta) = 2\beta(1-\theta)^r$ for some $1\le r\lesssim 1$. We use $\mathcal{A}_r$ to denote the action of $(\pi_\theta)_{\theta\in[0,1]}$. Then the oracle complexity for obtaining an estimate $\widehat Z$ that satisfies the criterion $\Pr\left(\left|\frac{\widehat Z}{Z}-1\right|\le\varepsilon\right)\ge\frac34$ is $\widetilde O\left(\frac{d^{\frac32}}{\varepsilon^2}\vee\frac{m\beta\mathcal{A}_r^{\frac12}}{\varepsilon^2}\vee\frac{d\beta^2\mathcal{A}_r^2}{\varepsilon^4}\right)$.

Reverse Diffusion Sampler (RDS)

The choice of the curve $(\pi_\theta)_{\theta\in[0,1]}$ is crucial for the complexity of JE & AIS. The geometric interpolation is widely used due to the availability of the scores of π_θ. However, for general target distributions, the action of the curve can be large:

■ Lemma (Exponential Lower Bound on the Action of Geometric Annealing)

Consider $\pi = \frac12\mathcal{N}(0,1)+\frac12\mathcal{N}(m,1)$ on $\mathbb{R}$ for some large $m\gtrsim 1$, whose potential is $\frac{m^2}{2}$-smooth. Under the setting in AIS, define $\pi_\theta(x)\propto\pi(x)\mathrm{e}^{-\frac{\lambda(\theta)}{2}x^2}$, $\theta\in[0,1]$, where $\lambda(\theta) = m^2(1-\theta)^r$ for some $1\le r\lesssim 1$. Then, the action of the curve $(\pi_\theta)_{\theta\in[0,1]}$ is $\mathcal{A}_r\gtrsim m^4\mathrm{e}^{\frac{m^2}{40}}$.

Reverse diffusion samplers (RDS): a series of multimodal samplers inspired by diffusion models. The OU process $\mathrm{d}Y_t = -Y_t\mathrm{d}t+\sqrt2\mathrm{d}B_t$, $t\in[0,T]$; $Y_0\sim\pi$ transforms any target distribution π into $\phi := \mathcal{N}(0,I)$ as $T\to\infty$. Let $Y_t\sim\overline\pi_t$. The time-reversal $(Y_t^\leftarrow := Y_{T-t}\sim\overline\pi_{T-t})_{t\in[0,T]}$ satisfies the SDE $\mathrm{d}Y_t^\leftarrow = (Y_t^\leftarrow + 2\nabla\log\overline\pi_{T-t}(Y_t^\leftarrow))\mathrm{d}t+\sqrt2\mathrm{d}W_t$, $t\in[0,T]$; $Y_0^\leftarrow\sim\overline\pi_T(\approx\phi)$. We propose leveraging the curve along the OU process for normalizing constant estimation. The following proposition supports this idea:

■ Proposition (Polynomial Upper Bound of the Action of the OU curve)

Let $\overline\pi_t$ be the law of Y_t in the OU process initialized from $Y_0\sim\pi\propto\mathrm{e}^{-V}$, where V is β-smooth and let $m^2 := \mathbb{E}_\pi\|\cdot\|^2<\infty$. Then, $\int_0^\infty|\dot{\overline\pi}|_t^2\mathrm{d}t\le d\beta+m^2$.

■ Theorem (a Framework for Normalizing Constant Estimation via RDS)

Assume a total time duration T, an early stopping time $\delta\ge0$, and discrete time points $0 = t_0 < t_1 < \ldots < t_N = T-\delta\le T$. For $t\in[0,T-\delta)$, let t_- denote t_k if $t\in[t_k,t_{k+1})$. Let $s_\cdot\approx\nabla\log\overline\pi_\cdot$ be a score estimator, and $\phi = \mathcal{N}(0,I)$. Consider the following two SDEs on $[0,T-\delta]$ representing the sampling trajectory and the time-reversed OU process, respectively:

$$\mathbb{Q}^\dagger:\quad \mathrm{d}X_t = (X_t+2s_{T-t_-}(X_{t_-}))\mathrm{d}t+\sqrt2\mathrm{d}B_t,\qquad X_0\sim\phi;$$
$$\mathbb{Q}:\quad \mathrm{d}X_t = (X_t+2\nabla\log\overline\pi_{T-t}(X_t))\mathrm{d}t+\sqrt2\mathrm{d}B_t,\qquad X_0\sim\overline\pi_T.$$

Let $\widehat Z := \mathrm{e}^{-W(X)}$, $X\sim\mathbb{Q}^\dagger$ be the estimator of Z, where $X\mapsto W(X)$ is defined as

$$\log\phi(X_0)+V(X_{T-\delta})+(T-\delta)d+\int_0^{T-\delta}\left(\|s_{T-t_-}(X_{t_-})\|^2\mathrm{d}t+\sqrt2\left\langle s_{T-t_-}(X_{t_-}),\mathrm{d}B_t\right\rangle\right).$$

Then, to ensure $\widehat Z$ satisfies $\Pr\left(\left|\frac{\widehat Z}{Z}-1\right|\le\varepsilon\right)\ge\frac34$, it suffices that $\mathrm{KL}(\mathbb{Q}\|\mathbb{Q}^\dagger)\lesssim\varepsilon^2$, $\mathrm{TV}(\pi,\overline\pi_\delta)\lesssim\varepsilon$. We can use results in [3, 4, 2, 7] to derive the total complexity.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.*

[2] Y. He, K. Rojas, and M. Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion.

[3] X. Huang, H. Dong, Y. Hao, Y. Ma, and T. Zhang. Reverse diffusion Monte Carlo.

[4] X. Huang, D. Zou, H. Dong, Y.-A. Ma, and T. Zhang. Faster sampling without isoperimetry via diffusion-based Monte Carlo.

[5] C. Jarzynski. Nonequilibrium equality for free energy differences.

[6] R. M. Neal. Annealed importance sampling.

[7] A. Vacher, O. Chehab, and A. Korba. Polynomial time sampling from log-smooth distributions in fixed dimension under semi-log-concavity of the forward diffusion with application to strongly dissipative distributions.

Contact: `{wei.guo,mtao,yongchen}@gatech.edu`