Efficient and adaptive non-log-concave sampling in fixed dimension via reverse diffusion.

Adrien Vacher[1], Omar Chehab[1], Anna Korba[1]
[1] Department of Statistics, CREST, ENSAE, IP Paris

The Bayesian sampling problem

- Goal: given $\mu \propto e^{-V}$ with $\int_{\mathbb{R}^d} V(x) < +\infty$ and an oracle access to V (and/or to its higher order derivatives), generate a sample $X \sim p$ such that p is ϵ-close to μ w.r.t. some probability divergence while keeping the number of queries to V (and/or its derivatives) as small as possible.
- Popular approach: Langevin algorithm
$$X_{n+1} = X_n - h\nabla V(X_n) + \sqrt{2h}z\,,$$
with $z \sim \mathcal{N}(0, I_d)$.
- Guarantees: As in the euclidean case, if V is L-smooth and α-strongly convex, $\tilde{O}(L^2\alpha^{-2}d\epsilon^{-1})$ queries to ∇V are sufficient to achieve ϵ-precision in KL for a well-chosen h. More broadly, if μ verifies an α-log-Sobolev inequality, the same guarantees hold [8].

Main issues

1. Multi-modality: heterogeneous data is not strongly log-concave and may have very poor log-Sobolev constants $\implies$ Langevin is stuck in local modes in practice and the complexity guarantees degrade exponentially with the distance between modes.

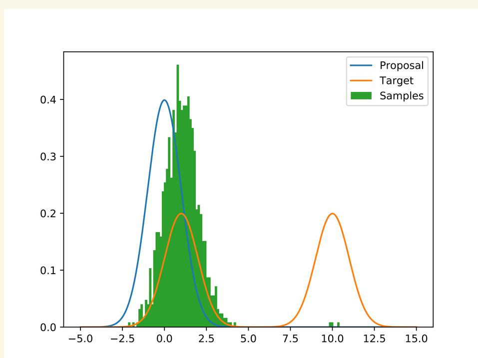

Figure: Metastable behavior of Langevin: particles get stuck in the first encountered mode.

E.g.: when uniform α-strong convexity is relaxed with α-strong convexity outside $B_R(0)$, the log-Sobolev constant degrades to $O(e^{-16RL^2}\alpha)$ and overall complexity degrades to $\tilde{O}(e^{-16RL^2}L^2\alpha^{-2}d\epsilon^{-1})$ [6].

2. Log-smoothness: popular multi-modal models, such as Gaussian Mixtures are *not* log-smooth.

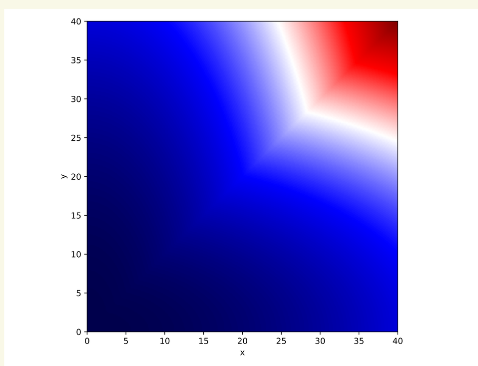

Figure: Laplacian of a non-smooth Gaussian Mixture.

E.g.: take $\mu = 0.5\mathcal{N}(0,\Sigma_1) + 0.5\mathcal{N}(0,\Sigma_2)$ with $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$. On the diagonal $-\nabla \log(\mu)(x,x) = 3(x/2, x/2)$ and right above, for $\eta > 0$ fixed, it holds asymptotically that $-\nabla \log(\mu)(x, x+\eta) \sim_{x\to+\infty} (2x, x)$.

3. Adaptivity: Langevin, and most alternatives, require *a priori* knowledge on the distribution (e.g. an upper-bound on the log-smoothness constant for Langevin, localization of the support for proposal-based methods) to achieve theoretical guarantees.

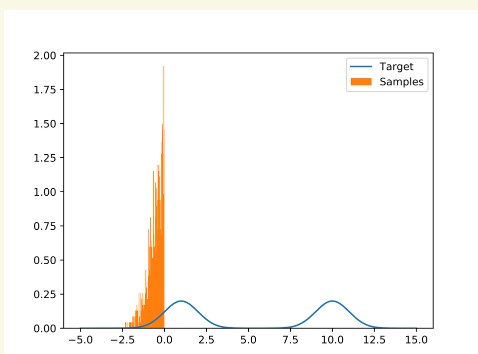

Figure: A rejection sampling algorithm with poorly chosen proposal.

Question

- Can we design a sampling algorithm that is 1) polynomial w.r.t. the constants of the problem 2) can handle Gaussian Mixtures 3) is adaptive?
- Unfortunately, existing lower bounds imply that multi-modal sampling has exponential complexity w.r.t. the dimension [5]. Still, can we address these three points when *the dimension is fixed*?

The reverse diffusion paradigm [7]

- Framework: Consider the *forward process*
$$\begin{cases} dX_t = -X_t + \sqrt{2}dB_t\,, \\ X_0 \sim \mu\,. \end{cases} \tag{1}$$
This process starts from μ, the density we wish to sample from, and targets a standard Gaussian $\implies$ quick convergence to equilibrium. Then, fix a horizon T, a N-discretization $0 = t_0 < t_1 < \cdots < t_N = T$ and implement the discretized reverse process as
$$\begin{cases} dY_t = Y_t dt + 2s_{t_k}(Y_k)dt + \sqrt{2}dB_t \quad t \in]T - t_{N-k}, T - t_{N-(k+1)}[\,, \\ Y_0 \sim \mathcal{N}(0, I_d)\,, \end{cases} \tag{2}$$
where s_{t_k} is a proxy of $-\nabla \log(p_{t_k})$ where p_{t_k} is the density of X_{t_k}, the forward process (1) at time t_k.
- Guarantees: under milder and milder assumptions [2, 1, 3] that notably allow for multi-modality, $Y_{t_N} \sim p$ is ensured to be close to μ whenever the proxies s_{t_k} provide a good approximation of the true intermediate scores $\implies$ the sampling problem is reduced to the problem of approximating the intermediate score functions.

Theorem ([3])

Assume that $\mu \propto e^{-V}$ has finite Fisher-information w.r.t. π the standard gaussian density in $\mathbb{R}^d$:
$$\mathcal{I}(\mu,\pi) = \int \|x - \nabla V(x)\|^2 d\mu(x) < +\infty\,.$$
Then, for the constant step-size discretization $t_k = kT/N$, denoting p the distribution of the sample Y_T output by (2), it holds that
$$KL(\mu,p) \lesssim (d + m_2)e^{-T} + \frac{1}{N}\sum_{k=1}^{N} \|\nabla \log(p_{t_k}) - s_{t_k}\|^2_{L^2(p_{t_k})} + \frac{T}{N}\mathcal{I}(\mu,\pi)\,,$$
where m_2 is the second order moment of μ and where $\lesssim$ hides a universal constant.

Estimator of the intermediate scores

Recall that the intermediate scores can be-rewritten as a ratio of Gaussian expectations
$$\nabla \log(p_t)(z) = \frac{-1}{1 - e^{-2t}} \frac{\mathbb{E}[Y_t e^{-V(e^t(z - Y_t))}]}{\mathbb{E}[e^{-V(e^t(z-Y_t))}]}\,,$$
where $Y_t \sim \mathcal{N}(0, (1 - e^{-2t})I_d) \implies$ cheap approximation as a ratio of empirical expectations yet, we must *correlate* the numerator and denominator
$$\hat{s}_{t,n}(z) = \frac{-1}{1 - e^{-2t}} \frac{\sum_{i=1}^n y_i e^{-V(e^t(z-y_i))}}{\sum_{i=1}^n e^{-V(e^t(z-y_i))}}\,. \tag{3}$$
Thanks to the correlation, this estimator is uniformly bounded with high probability: $\|\hat{s}_{t,n}(z)\| \leq \frac{\max_i \|y_i\|}{1-e^{-2t}} \sim \sqrt{\frac{d\log(n)}{1-e^{-2t}}}$.

Our assumptions

1. (Semi-log-convexity) The potential V is C^2 and verifies $\nabla^2 V \preceq \beta I_d$ for some $\beta \geq 0$.
2. (Dissipativity) There exists $a > 0, b \geq 0$ such that $\langle \nabla V(x), x\rangle \geq a\|x\|^2 - b$.

Note that Gaussian Mixtures verify both these assumptions.

Theorem

Under Assumptions 1-2, if we run algorithm (2) with $T = \log(1/\epsilon)$, $N = 1/\epsilon$, $t_k = kT/N$ and with the stochastic score estimators $\hat{s}_{n_k,t_k}$ defined in (3) with $n_k = d^2\epsilon^{-2(d+1)+1}$, then, denoting $\hat{p}$ the stochastic distribution of the output Y_{t_N}, it holds that
$$\mathbb{E}[KL(\mu,\hat{p})] \lesssim \epsilon\beta^{d+3}(b+d)/a^2\,,$$
where $\lesssim$ hides a universal constant as well as log factors with respect to $d, \epsilon^{-1}, a, b, \beta$. In particular, the error above can is achieved in $\sum_{k=1}^N n_k = d^2\epsilon^{-2(d+2)}$ queries to V.

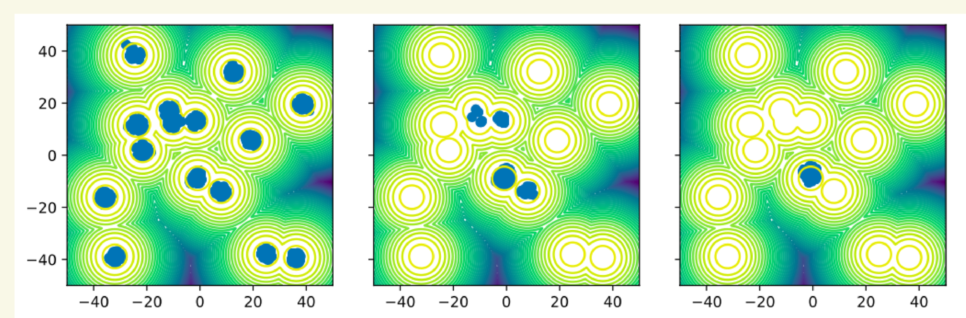

Figure: Our algorithm vs Langevin vs Reverse Diffusion Monte Carlo [4].

Bibliography

[1] Hongrui Chen, Holden Lee, and Jianfeng Lu. "Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions". In: *ICML*. 2023.
[2] Sitan Chen et al. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *ICLR*. 2023.
[3] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. "Score diffusion models without early stopping: finite Fisher information is all you need". In: *SIAM Journal on Mathematics of Data Science (SIMODS)* (2025).
[4] Xunpeng Huang et al. "Reverse diffusion Monte Carlo". In: *ICLR*. 2024.
[5] Holden Lee, Andrej Risteski, and Rong Ge. "Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo". In: *NeurIPS*. 2018.
[6] Yi-An Ma et al. "Sampling can be faster than optimization". In: *Proceedings of the National Academy of Sciences* (2019).
[7] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *ICLR*. 2021.
[8] Santosh Vempala and Andre Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices". In: *NeurIPS*. 2019.