Convergence of Stein Variational Gradient Descent under a Weaker Smoothness Condition

Lukang Sun Avetik Karagulyan Peter Richtárik

Computer, Electrical and Mathematical Sciences and Engineering Division, KAUST

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Abstract

Stein Variational Gradient Descent (SVGD) is an important alternative to the Langevin-type algorithms for sampling from probability distributions of the form $\pi(x) \propto \exp(-V(x))$. In the existing theory of Langevin-type algorithms and SVGD, the potential function V is often assumed to be L-smooth. However, this restrictive condition excludes a large class of potential functions such as polynomials of degree greater than 2. Our paper studies the convergence of the SVGD algorithm for distributions with (L_0, L_1)-smooth potentials. This relaxed smoothness assumption was introduced by Zhang et al. [2019] for the analysis of gradient clipping algorithms. With the help of trajectory-independent auxiliary conditions, we provide a descent lemma establishing that the algorithm decreases the KL divergence at each iteration and prove a complexity bound for SVGD in the population limit in terms of the Stein Fisher information.

Introduction

- Our goal is to sample from a given target distribution π defined on $\mathbb{R}^d$ with a large value of d. The latter can be formulated as the minimization of the functional $\mathcal{F}(\cdot) := \mathrm{KL}(\cdot \mid \pi)$, where

$$\mathrm{KL}(\mu \mid \nu) := \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu(x)}{\nu(x)}\right) \mu(\mathrm{d}x), \text{ if } \mu \ll \nu; \\ +\infty, \text{ otherwise.} \end{cases}$$

  Thus, we need to generate a distribution μ defined on $\mathcal{B}(\mathbb{R}^d)$ that satisfies

$$\mathcal{F}(\mu) = \mathrm{KL}(\mu \mid \pi) \le \varepsilon. \tag{1}$$

- Important particular case: π has a density (w.r.t. the Lebesgue measure) given by

$$\pi(\theta) \propto \exp(-V(\theta)), \tag{2}$$

  with a "potential" $V : \mathbb{R}^d \to \mathbb{R}$.

Contributions

The main contribution of the paper relies on its weaker set of assumptions, that allow to treat a larger class of probability distributions which includes densities with polynomials. We enlarge the class of probability distributions two-fold.

- The gradient smoothness assumption is very common in the sampling literature. Mathematically, it is formulated as $\|\nabla^2 V(x)\|_{op} \le L, \quad \forall x \in \mathbb{R}^d$, where $\nabla^2 V$ corresponds to the Hessian of V which is assumed to be well defined on $\mathbb{R}^d$. The issues is that this condition imposes at most linear growth of the potential function. This leaves out the polynomials. We propose the relaxed smoothness assumption (L_0, L_1) to overcome this issue.
- In [Korba et al., 2020], trajectory dependent conditions are required to guarantee the convergence of the algorithm. Later, Salim et al. [2021] replaced that condition with the T_1 inequality, but they used the smoothness. In this paper we propose a more general class of functional inequalities which will allow to treat the log-polynomial distributions.

Definition of the SVGD

Let the map $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel and let $\mathcal{H}_0$ be its corresponding RKHS. Let $\mathcal{H}$ be the space of the d-dimensional maps $\{(f_1, \dots, f_d)^\top \mid f_i \in \mathcal{H}_0, i = 1, \dots, d\}$. For two vector functions $f = (f_1, \dots, f_d)^\top$ and $g = (g_1, \dots, g_d)^\top$ from $\mathcal{H}$, we define the scalar product as

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{d} \langle f_i, g_i \rangle_{\mathcal{H}_0}.$$

Each iterate of the SVGD is defined as a pushforward measure from the previous one in a way that it minimizes the KL distance the most:

$$g_\mu := \arg\max_{\|\psi\|_{\mathcal{H}} \le 1} \left\{ -\frac{\mathrm{d}}{\mathrm{d}\gamma} \mathrm{KL}\big((I - \gamma\psi)\#\mu \mid \pi\big)\Big|_{\gamma=0} \right\}. \tag{3}$$

The operator ψ will serve us as the direction or the perturbation, while as γ is the step-size. Liu and Wang [2016] have shown that g_μ is given by

$$g_\mu(\cdot) = -\int_{\mathbb{R}^d} \left[\nabla \log \pi(x) k(x, \cdot) + \nabla_x k(x, \cdot)\right] \mu(\mathrm{d}x).$$

SVGD

For a step-size $\gamma > 0$ and an initial distribution $\mu_0 \in \mathcal{P}_p(\mathbb{R}^d)$ the SVGD algorithm is defined as

$$\mu_{i+1} := (I - \gamma g_{\mu_i})\#\mu_i. \tag{SVGD}$$

We are going to the measure the error of the SVGD using the Stein-Fisher information. The Stein-Fisher information between μ and π is defined as

$$I_{\mathrm{Stein}}(\mu \mid \pi) := \max_{\|\psi\|_{\mathcal{H}} \le 1} \left\{ \int_{\mathbb{R}^d} \big(-V(x)\psi(x) + \operatorname{div} \psi(x) \big) \mu(\mathrm{d}x) \right\}^2.$$

Assumptions

(Smoothness) The Hessian $\nabla^2 V$ of $V = -\log \pi$ is well-defined and $\exists L_0, L_1 \ge 0$ s.t. for any $x \in \mathbb{R}^d$

$$\left\|\nabla^2 V(x)\right\|_{op} \le L_0 + L_1 \|\nabla V(x)\|. \tag{L_0, L_1}$$

(At most polynomial gradients) For some $p > 0$, there exists a polynomial with positive coefficients such that $\mathrm{ord}(\mathrm{Q}) = p$ and the following inequality is true:

$$\left\|\nabla V(x)\right\| \le \mathrm{Q}(\|x\|). \tag{poly, Q}$$

(Bounded kernel) There exists $B > 0$ such that $\|k(x, \cdot)\|_{\mathcal{H}_0} \le B$ and

$$\|\nabla_x k(x, \cdot)\|_{\mathcal{H}} = \left(\sum_{i=1}^{d} \|\partial_{x_i} k(x, \cdot)\|^2_{\mathcal{H}_0} \right)^{\frac{1}{2}} \le B, \tag{ker, B}$$

for all $x \in \mathbb{R}^d$.

Complexity result

Under these four assumptions we prove a descent lemma (Theorem 1 in the paper) for the functional $\mathcal{F}$. The latter result leads to the following complexity bound (Theorem 2 in the paper).

Theorem

Let assumptions (ker, B), (L_0, L_1), and (poly, Q) hold and let $\mu_0 = \mathcal{N}(0, \mathrm{I}_d)$. Then in order to have $\sum_{i=0}^{n} I_{\mathrm{Stein}}(\mu_i \mid \pi) \le \varepsilon$ it is sufficient to perform n iterations of the SVGD, where

- $n = \mathcal{O}\left(\varepsilon^{-1} Q(1)^3 \max(L_1, 1) \lambda_{BV}^p (pd)^{p+1}\right)$, if for every $\mu \in \mathcal{P}_p(\mathbb{R}^d)$

$$W_p(\mu, \pi) \le \lambda_{BV} (\mathrm{KL}(\mu \mid \pi)^{1/p} + (\mathrm{KL}(\mu \mid \pi)/2)^{1/2p}); \tag{i}$$

- $n = \mathcal{O}\left(\varepsilon^{-1} \mathrm{Q}(1)^{\frac{p+2}{2}} \max(L_1, 1) \lambda_{\mathrm{T}}^{-\frac{p}{2}} (pd)^{\frac{(p+1)(p+2)}{4}}\right)$, if for every $\mu \in \mathcal{P}_p(\mathbb{R}^d)$

$$W_p(\mu, \pi) \le \sqrt{2\,\mathrm{KL}(\mu \mid \pi)/\lambda_{\mathrm{T}}}. \tag{ii}$$

Remarks

- The condition (i) can be expressed as a tail condition (see Corollary 2.3 of [Bolley and Villani, 2005]) that can be easily verified for log-polynomial densities. However, simple calculations show that $\lambda_{BV} = \mathcal{O}(d^{1/p})$ for the density $\pi(x) \propto \exp(-\|x\|^p)$.
- The condition (ii) corresponds to the classical T_p inequality. The constant λ_T is known to be dimension independent.

Conclusion

We quantify the convergence of the SVGD algorithm in average Stein-Fisher information under certain assumptions, which generalize the previously known results. In particular, our analysis allows to treat the case of high order polynomial potential functions which remained out of the scope of the prior work.

Bibliography

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331--352, 2005.

A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4672--4682, 2020.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2378--2386, 2016.

A. Salim, L. Sun, and P. Richtárik. Complexity analysis of Stein variational gradient descent under Talagrand's inequality T1. *arXiv preprint arXiv:2106.03076*, 2021.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2019.