A Non Asymptotic Analysis for Stein Variational Gradient Descent

Anna Korba[1], Adil Salim[2], Michael Arbel[1], Giulia Luise[3], Arthur Gretton[1]

[1]Gatsby Unit, University College London. [2]Visual Computing Center, KAUST. [3] Department of Computer Science, University College London.

Goal

Stein Variational Gradient Descent (SVGD) [2, 3] is a sampling algorithm that builds a sequence of probability measures $(\mu_n)_n$ targeting a distribution $\pi(x) \propto \exp(-V(x))$, where $V : \mathbb{R}^d \to \mathbb{R}$, in the Kullback Leibler (KL) sense.

Goal : Get convergence rates for SVGD.

Idea : Use optimization ideas on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

Background

Wasserstein distance.
Let $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite second moments on $\mathbb{R}^d$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W^2(\nu,\mu) := \inf_{s \in S(\mu,\nu)} \int \|x-y\|^2 ds(x,y).$$

$S(\mu,\nu)$ is the set of couplings between μ and ν.

KL divergence.
Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\mu \ll \pi$, then

$$\mathrm{KL}(\mu|\pi) := \int \log(\frac{d\mu}{d\pi}(x))d\mu(x)$$

and $\mathrm{KL}(\mu|\pi) := +\infty$ else.

Kernel integral operator.
Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a p.s.d. kernel and $\mathcal{H}_0$ its corresponding RKHS of real-valued on $\mathbb{R}^d$. Denote by $\mathcal{H} = \mathcal{H}_0^{\otimes d}$ the product RKHS equipped with standard inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int \|f\|^2 d\mu < \infty\}$.

$S_\mu : L^2(\mu) \to \mathcal{H}$ is defined by

$$\boldsymbol{S_\mu f = \int k(.,x) f(x) d\mu(x)}, \quad \forall f \in L^2(\mu).$$

Assume $\int k(x,x)d\mu(x) < \infty$. Then $\mathcal{H} \subset L^2(\mu)$. Denote the inclusion $\iota : \mathcal{H} \to L^2(\mu)$ with $\iota^* = S_\mu$ its adjoint, and define $P_\mu : L^2(\mu) \to L^2(\mu)$ the operator:

$$\boldsymbol{P_\mu := \iota S_\mu}$$

SVGD as KL minimization

Sampling from π is equivalent to sampling from the minimizer of $\mu \mapsto \mathrm{KL}(\mu|\pi)$. In the infinite number of particles regime, SVGD [2] can be seen as a gradient-descent like algorithm in the space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ where **at each iteration** $n \geq 0$:

$$\mu_{n+1} = \left(I - \gamma \boldsymbol{P_{\mu_n}} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right)_{\#} \mu_n, \tag{1}$$

where $\gamma > 0$, I identity map, μ_n, π also denote densities and $\#$ is the pushforward operation, i.e. in $\mathbb{R}^d$:

$$X_0 \sim \mu_0 \quad \Longrightarrow \quad X_{n+1} = X_n - \gamma \boldsymbol{P_{\mu_n}} \nabla \log\left(\frac{\mu_n}{\pi}\right)(X_n) \sim \mu_{n+1}. \tag{2}$$

Non Asymptotic Analysis of SVGD

Definition. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The *Stein Fisher Information* of μ relative to π is defined by :

$$I_{Stein}(\mu|\pi) = \|S_\mu \nabla \log\left(\frac{\mu}{\pi}\right)\|^2_{\mathcal{H}}. \tag{3}$$

Also referred to as the squared Kernel Stein Discrepancy (KSD) in the literature, separates the measures under mild assumptions [1].

We assume the following.

- $(\mathbf{A}_1)$ Assume that $\exists B > 0$ s.t. for all $x \in$, $\|k(x,.)\|_{\mathcal{H}_0} \leq B$ and $\|\nabla_x k(x,.)\|_{\mathcal{H}} = (\sum_{i=1}^d \|\partial_{x_i} k(x_i,.)\|^2_{\mathcal{H}_0})^{\frac{1}{2}} \leq B$.
- $(\mathbf{A}_2)$ The Hessian H_V of $V = -\log \pi$ is well-defined and $\exists M > 0$ s.t. $\|H_V\|_{op} \leq M$.
- $(\mathbf{A}_3)$ Assume that $\exists$ is $C > 0$ s.t. $I_{Stein}(\mu_n|\pi) < C$ for all n.

Under Assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$, a sufficient condition for Assumption $(\mathbf{A}_3)$ is $\sup_n \int \|x\| \mu_n(x) dx < \infty$.

Descent lemma for SVGD. Let μ_n defined by (2). Assume that Assumptions $(\mathbf{A}_1)$ to $(\mathbf{A}_3)$ hold. Let $\alpha > 1$ and choose $\gamma \leq \frac{\alpha-1}{\alpha B C^{\frac{1}{2}}}$. Denote $\beta = 1 - \gamma \frac{(\alpha^2+M)B^2}{2}$. Then:

$$\mathrm{KL}(\mu_{n+1}|\pi) - \mathrm{KL}(\mu_n|\pi) \leq -\gamma \beta I_{stein}(\mu_n|\pi). \tag{4}$$

Consequence of (4). Let $\alpha > 1$ and $\gamma < \min\left(\frac{\alpha-1}{\alpha B C^{\frac{1}{2}}}, \frac{2}{(\alpha^2+M)B^2}\right)$. Then,

$$\min_{k=1,\dots,n} I_{Stein}(\mu_n|\pi) \leq \frac{1}{n}\sum_{k=1}^n I_{Stein}(\mu_k|\pi) \leq \frac{\mathrm{KL}(\mu_0|\pi)}{\gamma \beta n}. \tag{5}$$

$\Longrightarrow$ **Does not rely on the convexity of V!**

Proof of (4): In optimization, descent lemmas are usually obtained under a **smoothness** assumption on the objective. Here, the objective $\mu \mapsto \mathrm{KL}(\mu|\pi)$ is **nonsmooth**, since its (Wasserstein) Hessian at μ:

$$\langle v, H_{\mathrm{KL}(.|\pi)}(\mu) v\rangle_{L^2(\mu)} = \underbrace{\mathbb{E}_{X\sim\mu}[\langle v(X), H_V(X)v(X)\rangle]}_{(*)} + \underbrace{\mathbb{E}_{X\sim\mu}\left[\|Jv(X)\|^2_{HS}\right]}_{(**)}$$

is not bounded over the whole tangent space to $\mathcal{P}_2(\mathbb{R}^d)$ at μ (included in $L^2(\mu)$). However, we can control $(**)$ when restricted to $\mathcal{H}$ under $(\mathbf{A}_1)$ and $(\mathbf{A}_3)$, while $(*)$ is controlled by $(\mathbf{A}_2)$.

Finite number of particles regime

In the finite number of particles regime, SVGD [3] algorithm updates a set of N particles $(X_n^i)_{i=1,\dots,N}$, particles as:

$$X_{n+1}^i = X_n^i - \gamma P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \tag{6}$$

where

$$P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(\cdot) = \frac{1}{N}\left[\sum_{j=1}^N k(X_n^j, \cdot)\nabla_{X_n^j} \log \pi(X_n^j) + \nabla_{X_n^j} k(X_n^j, \cdot)\right],$$

and $\hat{\mu}_n = \frac{1}{N}\sum_{j=1}^N \delta_{X_n^j}$.

We assume the following.

- $(\mathbf{B}_1)$ Assume that $\exists C_V$ s.t. $\forall x \in$, $\|V(x)\| \leq C_V$.
- $(\mathbf{B}_2)$ Assume that $\exists D > 0$ s.t. :
  $|k(x,x') - k(y,y')| \leq D(\|x-y\| + \|x'-y'\|)$,
  $\|\nabla k(x,x') - \nabla k(y,y')\| \leq D(\|x-y\| + \|x'-y'\|)$
  for all $x, x', y, y' \in \mathbb{R}^d$.

Propagation of chaos result.
Let $n \geq 0$ and $T > 0$. Let μ_n and $\hat{\mu}_n$ be defined by (2) and (6) respectively. Under Assumption $(\mathbf{A}_1)$,$(\mathbf{A}_2)$,$(\mathbf{B}_1)$,$(\mathbf{B}_2)$; for any $0 \leq n \leq \frac{T}{\gamma}$ we have :

$$\mathbb{E}[W_2^2(\mu_n, \hat{\mu}_n)] \leq \frac{1}{2}\left(\frac{1}{\sqrt{N}}\sqrt{var(\mu_0)}e^{LT}\right)(e^{2LT} - 1)$$

where L is a constant depending on k and π.

References

[1] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

[2] Qiang Liu. Stein variational gradient descent as gradient flow. In *NIPS*, 2017.

[3] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, 2016.