A Non Asymptotic Analysis of Stein Variational Gradient Descent

Anna Korba

CREST/ENSAE

Heriot-Watt University in Edinburgh online seminar
February 10, 2021

Joint work with Adil Salim (*KAUST*), Michael Arbel (*Gatsby Unit, UCL*), Giulia Luise (*CS Department, UCL*), Arthur Gretton (*Gatsby Unit, UCL*).

Outline

Problem : Sample from a target distribution π over $\mathcal{X} = \mathbb{R}^d$, whose density w.r.t. Lebesgue is written :

$$\pi(x) \propto \exp(-V(x))$$

where $V : \mathcal{X} \to \mathbb{R}$ is the potential function.

Problem : Sample from a target distribution π over $\mathcal{X} = \mathbb{R}^d$, whose density w.r.t. Lebesgue is written :

$$\pi(x) \propto \exp(-V(x))$$

where $V : \mathcal{X} \to \mathbb{R}$ is the potential function.

Motivation : Bayesian statistics.

- Let $\mathcal{D} = (x_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by θ (e.g. $p(y|x,\theta)$ gaussian)
  $\Longrightarrow$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^N p(y_i|\theta, x_i)$
- The parameter $\theta \sim p$ the prior distribution.

Bayes' rule : $p(\theta|\mathcal{D}) = \dfrac{p(\mathcal{D}|\theta)p(\theta)}{Z}$ where $Z = \displaystyle\int_{\mathbb{R}^d} p(\mathcal{D}|\theta)p(\theta)d\theta.$

How to sample from $\theta \mapsto p(\theta|\mathcal{D})$? (Z unknown).

Sampling as optimization of the KL

Assume $\pi \in \mathcal{P}_2(\mathcal{X}) = \{\mu, \int \|x\|^2 d\mu(x) < \infty\}$, hence $\pi \propto \exp(-V)$ is solution of :

$$\min_{\nu \in \mathcal{P}_2(\mathcal{X})} KL(\nu|\pi) \tag{1}$$

Sampling as optimization of the KL

Assume $\pi \in \mathcal{P}_2(\mathcal{X}) = \{\mu, \int \|x\|^2 d\mu(x) < \infty\}$, hence $\pi \propto \exp(-V)$ is solution of :

$$\min_{\nu \in \mathcal{P}_2(\mathcal{X})} KL(\nu|\pi) \tag{1}$$

1. **Langevin Monte Carlo (LMC)**

   [Dalalyan, 2017], [Durmus et al., 2017], [Durmus et al., 2019]

   - generates a Markov chain:

   $$x_{n+1} = x_n - \gamma \nabla V(x_n) + \sqrt{2\gamma} B_{n+1}, \quad \gamma > 0,\ B_n \sim N(0, I_d).$$

   - corresponds to a time-discretization of the gradient flow of the KL
   - asymptotic theory:
     if $x_n \sim \mu_n$ then $\mu_n \to \pi$ (weakly) as $n \to \infty, \gamma \to 0$.
   - non asymptotic theory (V smooth and strongly convex):
     it requires $\mathcal{O}(\frac{d}{\epsilon^2})$ iterations to get $W_2(\mu_{n+1}, \nu^*) \leq \epsilon$.
     $\implies$ converges at rate $\mathcal{O}(\sqrt{d/n})$, deteriorates quickly in high dimensions.

Sampling as optimization of the KL

Assume $\pi \in \mathcal{P}_2(\mathcal{X}) = \{\mu, \int \|x\|^2 d\mu(x) < \infty\}$, hence π is solution of :

$$\min_{\nu \in \mathcal{P}_2(\mathcal{X})} KL(\nu|\pi) \tag{2}$$

2. **Variational Inference (VI)**

   [Alquier and Ridgway, 2017], [Zhang et al., 2018]

   - restrict the search space in (2) to a parametric family
   - tractable in the large scale setting
   - only returns a parametric approximation of π

Sampling as optimization of the KL

Assume $\pi \in \mathcal{P}_2(\mathcal{X}) = \{\mu,\ \int \|x\|^2 d\mu(x) < \infty\}$, hence π is solution of :

$$\min_{\nu \in \mathcal{P}_2(\mathcal{X})} KL(\nu|\pi) \tag{3}$$

3. **Stein Variational Gradient Descent (SVGD)**

   [Liu and Wang, 2016],[Liu, 2017], [Duncan et al., 2019]

- "non parametric" VI, only depends on the choice of some kernel k
- corresponds to a time-discretization of the gradient flow of the KL under a metric depending on k
- uses a set of interacting particles to approximate π

`https://chi-feng.github.io/mcmc-demo/app.html?algorithm=HamiltonianMC&target=banana`

SVGD in the ML literature

- **Empirical performance** demonstrated in various tasks:
  - Bayesian inference [Liu and Wang, 2016, Feng et al., 2017, Liu and Zhu, 2018, Detommaso et al., 2018]
  - learning deep probabilistic models [Wang and Liu, 2016, Pu et al., 2017]
  - reinforcement learning [Liu et al., 2017]
- **Theoretical guarantees** :
  - asymptotic theory: (in continuous time, infinite number of particles) converges asymptotically to π [Lu et al., 2019] when V grows at most polynomially
  - non asymptotic theory: no rates of convergence.

This work : non asymptotic analysis of SVGD in the infinite particle regime but discrete time + finite sample approximation.

Outline

The Wasserstein space

The space $\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{X}),\ \int \|x\|^2 d\mu(x) < \infty\}$ is endowed with the Wassertein-2 distance from **Optimal transport** :

$$W_2^2(\nu, \mu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^2 \, ds(x, y) \qquad \forall \nu, \mu \in \mathcal{P}$$

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

The Wasserstein space

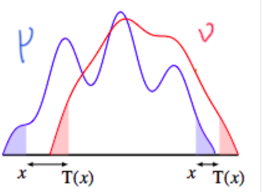
The space $\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{X}),\ \int \|x\|^2 d\mu(x) < \infty\}$ is endowed with the Wassertein-2 distance from **Optimal transport** :

$$W_2^2(\nu,\mu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^2\, ds(x,y) \qquad \forall \nu,\mu \in \mathcal{P}$$

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

Def (pushforward) : Let $\mu \in \mathcal{P}$, $T : \mathcal{X} \to \mathcal{X}$. The pushforward measure $T_{\#}\mu$ is characterized by:

- $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_{\#}\mu$



Continuity equations

For $\mu \in \mathcal{P}$, $L^2(\mu) = \{f : \mathcal{X} \to \mathcal{X},\ \int \|f(x)\|^2 d\mu(x) < \infty\}$.

It is a Hilbert space equipped with $\langle \cdot, \cdot \rangle_{L^2(\mu)}$ and $\| \cdot \|_{L^2(\mu)}$.

Continuity equations

For $\mu \in \mathcal{P}$, $L^2(\mu) = \{f : \mathcal{X} \to \mathcal{X},\ \int \|f(x)\|^2 d\mu(x) < \infty\}$.
It is a Hilbert space equipped with $\langle \cdot, \cdot \rangle_{L^2(\mu)}$ and $\| \cdot \|_{L^2(\mu)}$.

Consider a family $\mu : [0, \infty] \to \mathcal{P}, t \mapsto \mu_t$. It satisfies a continuity equation if there exists $(V_t)_{t \geq 0}$ such that $V_t \in L^2(\mu_t)$ and :

$$\boxed{\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0}$$

Density μ_t of particles $x_t \in \mathcal{X}$ driven by a vector field V_t:

$$\frac{dx_t}{dt} = V_t(x_t)$$

Riemannian interpretation [Otto, 2001]:
The tangent space of $\mathcal{P}$ at μ_t is $\mathcal{T}_{\mu_t}\mathcal{P} \subset L^2(\mu_t)$.

The KL defined over the Wasserstein space

For any $\mu, \pi \in \mathcal{P}$, the Kullback-Leibler divergence of μ w.r.t. π is defined by

$$KL(\mu|\pi) = \int_{\mathcal{X}} \log\left(\frac{d\mu}{d\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi$$

and is $+\infty$ otherwise.

We consider the functional $KL(\cdot|\pi) : \mathcal{P} \to [0, +\infty]$.

Recall that we want to solve:

$$\min_{\mu \in \mathcal{P}_2(\mathcal{X})} KL(\mu|\pi)$$

Wasserstein gradient flows [Ambrosio et al., 2008]

The **Wasserstein gradient flow of the functional** $KL(\cdot|\pi)$ is a curve $\mu : [0, \infty] \to \mathcal{P}, \; t \mapsto \mu_t$ that satisfies:

$$\frac{\partial \mu_t}{\partial t} = \text{“} - \nabla_{W_2} KL(\mu_t|\pi)\text{”}$$

Wasserstein gradient flows [Ambrosio et al., 2008]

The **Wasserstein gradient flow of the functional** $KL(\cdot|\pi)$ is a curve $\mu : [0, \infty] \to \mathcal{P},\ t \mapsto \mu_t$ that satisfies:

$$\frac{\partial \mu_t}{\partial t} = \text{“} - \nabla_{W_2} KL(\mu_t|\pi)\text{”}$$

Can be obtained as the limit when $\tau \to 0$ of the **JKO scheme** [Jordan et al., 1998] :

$$\mu(n+1) = \operatorname*{argmin}_{\mu \in \mathcal{P}} KL(\mu|\pi) + \frac{1}{2\tau} W_2^2(\mu, \mu(n))$$

Wassertein gradient flows

The Wassertein GF of $KL(\cdot|\pi)$ is written :

$$\boxed{\frac{\partial \mu_t}{\partial t} - div(\mu_t \nabla_{W_2} KL(\mu_t|\pi)) = 0}$$

where for μ_t regular enough,

$$\boxed{\nabla_{W_2} KL(\mu_t|\pi) = \nabla \frac{\partial KL(\mu_t|\pi)}{\partial \mu} = \nabla \log\left(\frac{d\mu_t}{d\pi}\right).}$$

Wassertein gradient flows

The Wassertein GF of $KL(\cdot|\pi)$ is written :

$$\boxed{\frac{\partial \mu_t}{\partial t} - div(\mu_t \nabla_{W_2} KL(\mu_t|\pi)) = 0}$$

where for μ_t regular enough,

$$\boxed{\nabla_{W_2} KL(\mu_t|\pi) = \nabla \frac{\partial KL(\mu_t|\pi)}{\partial \mu} = \nabla \log\left(\frac{d\mu_t}{d\pi}\right).}$$

$\frac{\partial KL(\mu|\pi)}{\partial \mu} : \mathcal{X} \to \mathbb{R}$: differential of $\mu \mapsto KL(\mu|\pi)$, evaluated at μ.

It is the unique function s. t. for any $\mu, \mu' \in \mathcal{P}$, $\mu' - \mu \in \mathcal{P}$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(KL(\mu + \epsilon(\mu' - \mu)|\pi) - KL(\mu|\pi)) = \int_{\mathcal{X}} \frac{\partial KL(\mu|\pi)}{\partial \mu}(x)(d\mu' - d\mu)(x).$$

Wasserstein Gradient descent

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left(I - \gamma \nabla_{W_2} KL(\mu_n|\pi)\right)_{\#} \mu_n \tag{4}$$

where $\gamma > 0$ is a step-size.

- (Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla_{W_2} KL(\mu_t|\pi)(X_n), \quad X_n \sim \mu_n.$$

- (4) can be seen as RGD where $\phi \to (I + \phi)_{\#}\mu$ (defined on $L^2(\mu)$) is the exp. map at μ.

Wasserstein Gradient descent

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left(I - \gamma \nabla_{W_2} KL(\mu_n|\pi)\right)_{\#} \mu_n \tag{4}$$

where $\gamma > 0$ is a step-size.

- (Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla_{W_2} KL(\mu_t|\pi)(X_n), \quad X_n \sim \mu_n.$$

- (4) can be seen as RGD where $\phi \to (I + \phi)_{\#}\mu$ (defined on $L^2(\mu)$) is the exp. map at μ.

Problem: $\nabla_{W_2} KL(\mu_t|\pi) = \nabla \log\left(\frac{\mu_n}{\pi}\right)$.

While $\nabla \log \pi$ is known, $\nabla \log \mu_n$ has to be estimated from samples.

Outline

Background on kernels and RKHS [Steinwart and Christmann, 2008]

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive, semi-definite kernel

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \quad \phi : \mathcal{X} \to \mathcal{H}$$

- $\mathcal{H}$ its corresponding RKHS (Reproducing Kernel Hilbert Space):

$$\mathcal{H}_0 = \overline{\left\{ \sum_{i=1}^{m} \alpha_i k(\cdot, x_i);\; m \in \mathbb{N};\; \alpha_1, \ldots, \alpha_m \in \mathbb{R};\; x_1, \ldots, x_m \in \mathcal{X} \right\}}$$

- $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and norm $\|.\|_{\mathcal{H}}$.
  It satisfies the reproducing property:

$$\forall \quad f \in \mathcal{H},\; x \in \mathcal{X}, \quad f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}}$$

We assume $\int_{\mathcal{X}\times\mathcal{X}} k(x, x) d\mu(x) < \infty$ for any $\mu \in \mathcal{P}$.
$\implies \mathcal{H} \subset L^2(\mu)$.

The kernel integral operator

Then, the inclusion from $\iota : \mathcal{H} \to L^2(\mu)$ admits an adjoint $\iota^\star = S_\mu$, where $S_\mu : L^2(\mu) \to \mathcal{H}$ is defined by:

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

The kernel integral operator

Then, the inclusion from $\iota : \mathcal{H} \to L^2(\mu)$ admits an adjoint $\iota^\star = S_\mu$, where $S_\mu : L^2(\mu) \to \mathcal{H}$ is defined by:

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

We have for any $f, g \in L_2(\mu) \times \mathcal{H}$:

$$\langle f, \iota g \rangle_{L^2(\mu)} = \langle \iota^* f, g \rangle_{\mathcal{H}} = \langle S_\mu f, g \rangle_{\mathcal{H}}.$$

We will denote $P_\mu = \iota \circ S_\mu$.

SVGD algorithm

SVGD trick: applying this operator to the W_2 gradient of $KL(\cdot|\pi)$ leads to

$$P_\mu \nabla \log\left(\frac{\mu}{\pi}\right)(\cdot) = -\int [\nabla \log \pi(x) k(x,\cdot) + \nabla_x k(x,\cdot)] d\mu(x),$$

under appropriate boundary conditions on k and π, e.g. $\lim_{\|x\|\to\infty} k(x,\cdot)\pi(x) \to 0$.

SVGD algorithm

SVGD trick: applying this operator to the W_2 gradient of $KL(\cdot|\pi)$ leads to

$$P_\mu \nabla \log\left(\frac{\mu}{\pi}\right)(\cdot) = -\int [\nabla \log \pi(x) k(x,\cdot) + \nabla_x k(x,\cdot)] d\mu(x),$$

under appropriate boundary conditions on k and π, e.g. $\lim_{\|x\|\to\infty} k(x,\cdot)\pi(x) \to 0$.

Algorithm : Starting from N i.i.d. samples $(X_0^i)_{i=1,\dots,N} \sim \mu_0$, SVGD algorithm updates the N particles as follows :

$$X_{n+1}^i = X_n^i - \gamma \underbrace{\left[\frac{1}{N}\sum_{j=1}^{N} k(X_n^i, X_n^j)\nabla_{X_n^j} \log \pi(X_n^j) + \nabla_{X_n^j} k(X_n^j, X_n^i)\right]}_{P_{\hat{\mu}_n}\nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \quad \text{with } \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_n^j}}$$

SVGD algorithm

SVGD trick: applying this operator to the W_2 gradient of $KL(\cdot|\pi)$ leads to

$$P_\mu \nabla \log\left(\frac{\mu}{\pi}\right)(\cdot) = -\int [\nabla \log \pi(x) k(x,\cdot) + \nabla_x k(x,\cdot)] d\mu(x),$$

under appropriate boundary conditions on k and π, e.g. $\lim_{\|x\|\to\infty} k(x,\cdot)\pi(x) \to 0$.

Algorithm : Starting from N i.i.d. samples $(X_0^i)_{i=1,\dots,N} \sim \mu_0$, SVGD algorithm updates the N particles as follows :

$$X_{n+1}^i = X_n^i - \gamma \underbrace{\left[\frac{1}{N}\sum_{j=1}^N k(X_n^i, X_n^j)\nabla_{X_n^j} \log \pi(X_n^j) + \nabla_{X_n^j} k(X_n^j, X_n^i)\right]}_{P_{\hat{\mu}_n}\nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \quad \text{with } \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^N \delta_{X_n^j}}$$

This work : non asymptotic analysis of SVGD in the infinite particle regime + finite sample approximation.

Outline

Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017],[Lu et al., 2019]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017],[Lu et al., 2019]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \iota S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right), \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\|^2_{\mathcal{H}} \text{ since } \iota^* = S_{\mu_t} \\
&\leq 0.
\end{aligned}$$

Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017],[Lu et al., 2019]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \iota S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right), \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\|^2_{\mathcal{H}} \text{ since } \iota^* = S_{\mu_t} \\
&\leq 0.
\end{aligned}$$

On the r.h.s. we have the **Kernel Stein discrepancy** [Chwialkowski et al., 2016] or **Stein Fisher information** at μ_t.

Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017],[Lu et al., 2019]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \iota S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right), \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\|^2_{\mathcal{H}} \text{ since } \iota^* = S_{\mu_t} \\
&\leq 0.
\end{aligned}$$

On the r.h.s. we have the **Kernel Stein discrepancy** [Chwialkowski et al., 2016] or **Stein Fisher information** at μ_t.

Along the WGF of the KL (Langevin dynamics) we would have obtained the relative Fisher information $\|\nabla \log\left(\frac{\mu_t}{\pi}\right)\|^2_{L^2(\mu_t)}$.

Stein Fisher information

Stationary condition : $\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right) \right\|^2_{\mathcal{H}} = 0.$

Implies weak convergence of μ_t to π if [Gorham and Mackey, 2017]:

- π is distantly dissipative[1] (e.g. gaussian mixtures)
- k is translation invariant with a non-vanishing Fourier transform;
  or k is the IMQ kernel defined by $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for $c > 0$ and $\beta \in [-1, 0]$ (slow decay rate).

[1] $\liminf_{r \to \infty} \kappa(r) > 0$ for $\kappa(r) = \inf\{-2\langle \nabla \log \pi(x) - \nabla \log \pi(y), x - y \rangle / \|x - y\|_2^2; \|x - y\|_2^2 = r\}$

Stein Fisher information

Stationary condition : $\left\| S_{\mu_t} \nabla \log \left(\frac{\mu_t}{\pi}\right) \right\|^2_{\mathcal{H}} = 0.$

Implies weak convergence of μ_t to π if [Gorham and Mackey, 2017]:

- π is distantly dissipative[1] (e.g. gaussian mixtures)
- k is translation invariant with a non-vanishing Fourier transform;
  or k is the IMQ kernel defined by $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for $c > 0$ and $\beta \in [-1, 0]$ (slow decay rate).

We show that if k is bounded, $\pi \propto \exp(-V)$ with H_V bounded above and if $\exists C > 0$, $\int \|x\|^2 d\mu_t(x) < C$ for all $t > 0$, then $\left\| S_{\mu_t} \nabla \log \left(\frac{\mu_t}{\pi}\right) \right\|^2_{\mathcal{H}} \to 0$

[1] $\liminf_{r\to\infty} \kappa(r) > 0$ for
$\kappa(r) = \inf\{-2\langle \nabla \log \pi(x) - \nabla \log \pi(y), x - y\rangle / \|x - y\|_2^2; \|x - y\|_2^2 = r\}$

Convergence of continuous-time dynamics

The convergence of the Stein Fisher information to 0 can be slow. **When do we have fast convergence of SVGD dynamics?**

π satisfies the Stein log-Sobolev inequality [Duncan et al., 2019] with constant $\lambda > 0$ if for any μ:

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| S_\mu \nabla \log\left(\frac{\mu}{\pi}\right) \right\|_{\mathcal{H}}^2 .$$

Convergence of continuous-time dynamics

The convergence of the Stein Fisher information to 0 can be slow. **When do we have fast convergence of SVGD dynamics?**

π satisfies the Stein log-Sobolev inequality [Duncan et al., 2019] with constant $\lambda > 0$ if for any μ:

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| S_\mu \nabla \log\left(\frac{\mu}{\pi}\right) \right\|^2_{\mathcal{H}}.$$

If it holds,

$$\frac{dKL(\mu_t|\pi)}{dt} = -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right) \right\|^2_{\mathcal{H}} \leq -2\lambda KL(\mu_t|\pi)$$

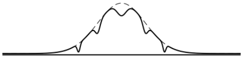
and by integrating :

$$KL(\mu_t|\pi) \leq e^{-2\lambda t} KL(\mu_0|\pi).$$

"Classic" log-Sobolev inequality upper bounds the KL by the Fisher divergence :

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| \nabla \log \left(\frac{\mu}{\pi} \right) \right\|^2_{L^2(\mu)} .$$

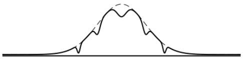
satisfied as soon as π is λ-log concave, but it's more general.



"Classic" log-Sobolev inequality upper bounds the KL by the Fisher divergence :

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| \nabla \log \left(\frac{\mu}{\pi} \right) \right\|^2_{L^2(\mu)}.$$

satisfied as soon as π is λ-log concave, but it's more general.



When is Stein log-Sobolev satisfied? not as well known and understood [Duncan et al., 2019], but :

- it fails to hold if k is too regular with respect to π
- some working examples in dimension 1
- whether it holds in higher dimension is more challenging and subject to further research...

Outline

Proof of a descent lemma for GD of a smooth function

Gradient descent for $F : \mathbb{R}^d \to \mathbb{R}$ a $C^2(\mathbb{R}^d)$ s.t. $\|H_F(x)\| \leq M$ for any x.

$$x_{n+1} = x_n - \gamma \nabla F(x_n).$$

Proof of a descent lemma for GD of a smooth function

Gradient descent for $F : \mathbb{R}^d \to \mathbb{R}$ a $C^2(\mathbb{R}^d)$ s.t. $\|H_F(x)\| \leq M$ for any x.

$$x_{n+1} = x_n - \gamma \nabla F(x_n).$$

Denote $x(t) = x_n - t\nabla F(x_n)$ and $\varphi(t) = F(x(t))$. Using Taylor expansion :

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt.$$

Proof of a descent lemma for GD of a smooth function

Gradient descent for $F:\mathbb{R}^d \to \mathbb{R}$ a $C^2(\mathbb{R}^d)$ s.t. $\|H_F(x)\| \leq M$ for any x.

$$x_{n+1} = x_n - \gamma \nabla F(x_n).$$

Denote $x(t) = x_n - t\nabla F(x_n)$ and $\varphi(t) = F(x(t))$. Using Taylor expansion :

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt.$$

Since ($\ddot{x}(t) = 0$):

$$\varphi'(0) = \langle \nabla F(x(0)), \dot{x}(0)\rangle = \langle \nabla F(x(0)), -\nabla F(x_n)\rangle = -\|\nabla F(x_n)\|^2,$$

$$\varphi''(t) = \langle \dot{x}(t), H_F(x(t))\dot{x}(t)\rangle \leq M\|\dot{x}(t)\|^2 = M\|\nabla F(x_n)\|^2,$$

Proof of a descent lemma for GD of a smooth function

Gradient descent for $F : \mathbb{R}^d \to \mathbb{R}$ a $C^2(\mathbb{R}^d)$ s.t. $\|H_F(x)\| \leq M$ for any x.

$$x_{n+1} = x_n - \gamma \nabla F(x_n).$$

Denote $x(t) = x_n - t\nabla F(x_n)$ and $\varphi(t) = F(x(t))$. Using Taylor expansion :

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt.$$

Since ($\ddot{x}(t) = 0$):

$$\varphi'(0) = \langle \nabla F(x(0)), \dot{x}(0)\rangle = \langle \nabla F(x(0)), -\nabla F(x_n)\rangle = -\|\nabla F(x_n)\|^2,$$

$$\varphi''(t) = \langle \dot{x}(t), H_F(x(t))\dot{x}(t)\rangle \leq M\|\dot{x}(t)\|^2 = M\|\nabla F(x_n)\|^2,$$

we have

$$F(x_{n+1}) \leq F(x_n) - \gamma\|\nabla F(x_n)\|^2 + M\int_0^\gamma (\gamma - t)\|\nabla F(x_n)\|^2 dt$$

$$F(x_{n+1}) - F(x_n) \leq -\gamma\left(1 - \frac{M\gamma}{2}\right)\|\nabla F(x_n)\|^2.$$

A descent lemma for SVGD

Here, the Hessian operator of the KL at μ is an operator on $L^2(\mu)$:

$$\langle f, Hess_{KL(.|\pi)}(\mu) f\rangle_{L^2(\mu)} = \mathbb{E}_{X\sim\mu}\left[\langle f(X), H_V(X) f(X)\rangle + \|Jf(X)\|^2_{HS}\right]$$

and yet, this operator **is not bounded**.

A descent lemma for SVGD

Here, the Hessian operator of the KL at μ is an operator on $L^2(\mu)$:

$$\langle f, Hess_{KL(.|\pi)}(\mu) f\rangle_{L^2(\mu)} = \mathbb{E}_{X\sim\mu}\left[\langle f(X), H_V(X) f(X)\rangle + \|Jf(X)\|_{HS}^2\right]$$

and yet, this operator **is not bounded.**

In the case of SVGD one restricts the descent directions f to $\mathcal{H}$. Under several assumptions (boundedness of k and ∇k, of Hessian of V and moments on the trajectory) we could show for γ small enough:

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq -c_\gamma \underbrace{\left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right\|_{\mathcal{H}}^2}_{I_{Stein}(\mu_n|\pi)}.$$

Sketch of proof - 1

Fix $n \geq 0$. Denote $g = P_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)$, $\phi_t = I - tg$ for $t \in [0, \gamma]$ and $\rho_t = (\phi_t)_{\#}\mu_n$, which is ruled by the velocity field $w_t(x) = -g(\phi_t^{-1}(x))$.

Denote $\varphi(t) = KL(\rho_t|\pi)$. Using a Taylor expansion, $\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt$.

Step 1.

$$\varphi(0) = KL(\mu_n|\pi) \text{ and } \varphi(\gamma) = KL(\mu_{n+1}|\pi).$$

Step 2. Using the chain rule,

$$\varphi'(t) = \langle \nabla_{W_2} KL(\rho_t|\pi), w_t \rangle_{L^2(\rho_t)}.$$

Hence :

$$\varphi'(0) = -\langle \nabla \log\left(\frac{\mu_n}{\pi}\right), g \rangle_{L^2(\mu_n)} = -\left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \right\|_{\mathcal{H}}^2 .$$

Sketch of proof - 2

Step 3.

$$\varphi''(t) = \langle w_t, Hess_{KL(.|\pi)}(\rho_t) w_t \rangle_{L^2(\rho_t)} := \psi_1(t) + \psi_2(t),$$

$$\psi_1(t) = \mathbb{E}_{x\sim\rho_t}\left[\langle w_t(x), H_V(x) w_t(x)\rangle\right] \text{ and } \psi_2(t) = \mathbb{E}_{x\sim\rho_t}\left[\|J w_t(x)\|^2_{HS}\right]$$

where $\rho_t = (\phi_t)_{\#}\mu_n$, $w_t = -g \circ (\phi_t)^{-1}$.

Sketch of proof - 2

Step 3.

$$\varphi''(t) = \langle w_t, Hess_{KL(.|\pi)}(\rho_t) w_t \rangle_{L^2(\rho_t)} := \psi_1(t) + \psi_2(t),$$

$$\psi_1(t) = \mathbb{E}_{x\sim\rho_t}\left[\langle w_t(x), H_V(x) w_t(x)\rangle\right] \text{ and } \psi_2(t) = \mathbb{E}_{x\sim\rho_t}\left[\|Jw_t(x)\|^2_{HS}\right]$$

where $\rho_t = (\phi_t)_{\#}\mu_n$, $w_t = -g \circ (\phi_t)^{-1}$.

Step 3.a. Assuming $\|H_V\| \leq M$ and $k(.,.) \leq B$:

$$\psi_1(t) \leq M\|g\|^2_{L^2(\mu_n)} \leq MB^2 \left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right\|^2_{\mathcal{H}}.$$

Sketch of proof - 2

Step 3.

$$\varphi''(t) = \langle w_t, Hess_{KL(.|\pi)}(\rho_t) w_t \rangle_{L^2(\rho_t)} := \psi_1(t) + \psi_2(t),$$

$$\psi_1(t) = \mathbb{E}_{x\sim\rho_t}\left[\langle w_t(x), H_V(x) w_t(x)\rangle\right] \quad \text{and} \quad \psi_2(t) = \mathbb{E}_{x\sim\rho_t}\left[\|Jw_t(x)\|_{HS}^2\right]$$

where $\rho_t = (\phi_t)_{\#}\mu_n$, $w_t = -g \circ (\phi_t)^{-1}$.

Step 3.a. Assuming $\|H_V\| \leq M$ and $k(.,.) \leq B$:

$$\psi_1(t) \leq M\|g\|_{L^2(\mu_n)}^2 \leq MB^2 \left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right\|_{\mathcal{H}}^2.$$

Step 3.b. Since $\rho_t = (\phi_t)_{\#}\mu_n$, $w_t = -g \circ (\phi_t)^{-1}$,

$$\begin{aligned}\psi_2(t) = \mathbb{E}_{x\sim\mu_n}[\|Jw_t \circ \phi_t(x)\|_{HS}^2] &\leq \|Jg(x)\|_{HS}^2 \|(J\phi_t)^{-1}(x)\|_{op}^2 \\ &\leq B^2 \left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right\|_{\mathcal{H}}^2 \alpha^2,\end{aligned}$$

assuming $\|\nabla k(.,.)\| \leq B$ and choosing $\gamma \leq f(\alpha)$ with $\alpha > 1$.

From:

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt$$

we have:

$$\begin{aligned} KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq & -\gamma \| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \|_{\mathcal{H}}^2 \\ & + \frac{\gamma^2}{2}(\alpha^2 + M)B^2 \| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \|_{\mathcal{H}}^2 \end{aligned}$$

choosing γ small enough yields a descent lemma :

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq -c_\gamma \underbrace{\left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \right\|_{\mathcal{H}}^2}_{I_{Stein}(\mu_n|\pi)}.$$

Rates in terms of the Stein Fisher Information

Consequence of the descent lemma: for γ small enough,

$$\boxed{\min_{k=1,\ldots,n} I_{Stein}(\mu_n|\pi) \leq \frac{1}{n}\sum_{k=1}^{n} I_{Stein}(\mu_k|\pi) \leq \frac{KL(\mu_0|\pi)}{c_\gamma n}.}$$

This result does not rely on:

- Stein log Sobolev inequality
- nor on convexity of V
- only smoothness of V.

unlike most results on LMC which rely on Log Sobolev inequality or convexity of V.

Rates in terms of the KL objective?

To obtain rates, one may combine a descent lemma (1) of the form

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq -c_\gamma \left\| S_{\mu_n} \nabla \log \left(\frac{\mu_n}{\pi} \right) \right\|_{\mathcal{H}}^2$$

and the Stein log-Sobolev inequality (2):

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \underbrace{\leq}_{(1)} -c_\gamma \left\| S_{\mu_n} \nabla \log \left(\frac{\mu_n}{\pi} \right) \right\|_{\mathcal{H}}^2 \underbrace{\leq}_{(2)} -c_\gamma 2\lambda KL(\mu_n|\pi).$$

Iterating this inequality yields $KL(\mu_n|\pi) \leq (1 - 2c_\gamma\lambda)^n KL(\mu_0|\pi)$.

"Classic" approach in optimization [Karimi et al., 2016] or in the analysis of LMC.

Not possible to combine both....

Given that both the kernel and its derivative are bounded, the equation

$$\int \sum_{i=1}^{d} [(\partial_i V(x))^2 k(x,x) \\ - \partial_i V(x)(\partial_i^1 k(x,x) + \partial_i^2 k(x,x)) + \partial_i^1 \partial_i^2 k(x,x)] d\pi(x) < \infty \quad (5)$$

reduces to a property on V which, as far as we can tell, always holds on $\mathcal{X} = \mathbb{R}^d$...

Not possible to combine both....

Given that both the kernel and its derivative are bounded, the equation

$$\int \sum_{i=1}^{d} [(\partial_i V(x))^2 k(x,x) \\ - \partial_i V(x)(\partial_i^1 k(x,x) + \partial_i^2 k(x,x)) + \partial_i^1 \partial_i^2 k(x,x)] d\pi(x) < \infty \quad (5)$$

reduces to a property on V which, as far as we can tell, always holds on $\mathcal{X} = \mathbb{R}^d$...

and this implies that Stein LSI does not hold [Duncan et al., 2019].

Remark : Equation (5) does not hold for :

- k polynomial of order ≥ 3, and
- π with exploding β moments with $\beta \geq 3$ (ex: a student distribution, which belongs to $\mathcal{P}$ the set of distributions with bounded second moment).

Experiments

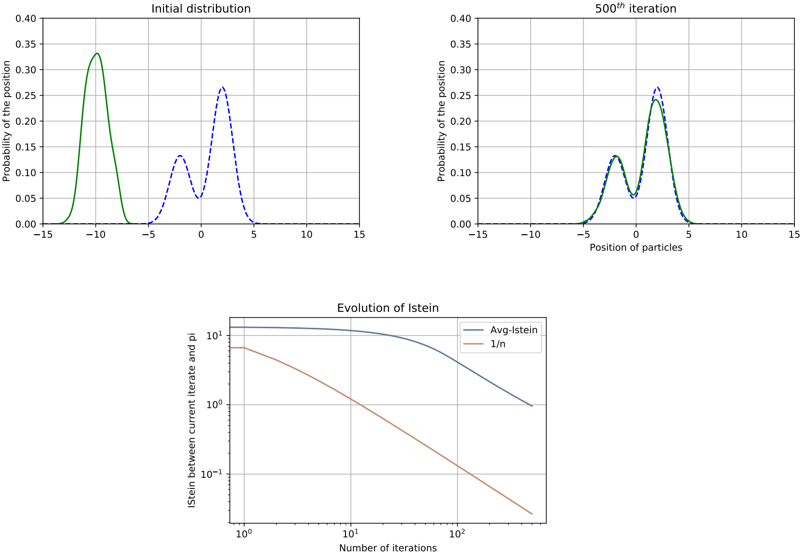


Figure: The particle implementation of the SVGD algorithm illustrates the convergence of $I_{Stein}(\mu_n|\pi)$ to 0.

Outline

We already have a bound on μ_n versus π. What about $\hat{\mu}_n$?

Recall that the practical SVGD implementation is :

$$X_{n+1}^i = X_n^i - \gamma P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \qquad \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_n^j}.$$

where $\hat{\mu}_n$ denotes the empirical distribution of the interacting particles.

We already have a bound on μ_n versus π. What about $\hat{\mu}_n$?

Recall that the practical SVGD implementation is :

$$X_{n+1}^i = X_n^i - \gamma P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \qquad \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_n^j}.$$

where $\hat{\mu}_n$ denotes the empirical distribution of the interacting particles.

Propagation of chaos result

Let $n \geq 0$ and $T > 0$. Under boundedness and Lipschitzness assumptions for all $k, \nabla k, V$; 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.

Contributions and openings

- First rates of convergence for SVGD, using techniques from optimal transport and optimization (discrete time - infinite number of particles)
- Propagation of chaos bound (finite number of particles regime)

Open questions

- Rates in KL?
- Propagation of chaos : weaker assumptions? uniform in time (UIT)?

Open questions

- Rates in KL?
- Propagation of chaos : weaker assumptions? uniform in time (UIT)?
- Is it possible to obtain a unified convergence bound (decreasing as $n, N \to \infty$)? (requires UIT)

$$D(\widehat{\mu}_n, \pi) \leq A_n + B_N$$

Open questions

- Rates in KL?
- Propagation of chaos : weaker assumptions? uniform in time (UIT)?
- Is it possible to obtain a unified convergence bound (decreasing as $n, N \to \infty$)? (requires UIT)

$$D(\widehat{\mu}_n, \pi) \leq A_n + B_N$$

- Other kernels?
  SVGD dynamics also appear in black-box variational inference and Gans [Chu et al., 2020], where the kernel is *the neural tangent kernel* and **depends on the current distribution** ($k \implies k_{\mu_n}$)

Some advertisement

Upcoming preprint : **Kernel Stein Discrepancy (KSD) Descent**

Joint work with Pierre-Cyril Aubin-Frankowski (*Les Mines ParisTech*), Szymon Majewski (*Ecole Polytechnique/ENSAE*), Pierre Ablin (*Ecole Normale Supérieure*).

Idea: compute gradient descent of the KSD :

$$KSD(\mu|\pi) = \|S_\mu \nabla \log\left(\frac{\mu}{\pi}\right)\|^2_{\mathcal{H}} = \iint k_\pi(x,y)d\mu(x)d\mu(y),$$

$$\begin{aligned} k_\pi(x,y) = \nabla V(x)^T \nabla V(y) k(x,y) + \nabla V(x)^T \nabla_2 k(x,y) \\ + \nabla_1 k(x,y)^T \nabla V(y) + \boldsymbol{\nabla} \cdot_1 \nabla_2 k(x,y). \end{aligned}$$

Pros:

- very simple update:

$$x_{n+1}^i = x_n^i - \frac{2\gamma}{N^2} \sum_{j=1}^{N} \nabla_2 k_\pi(x_n^j, x_n^i),$$

- closed-form cost function (KSD) enables to use L-BFGS [Liu and Nocedal, 1989] (fast, and does not require the choice of a step-size)
- works well on convex tasks (unimodal gaussian, bayesian logistic regression with gaussian priors)

Cons:

- KSD is not convex w.r.t. W_2, and no exponential decay near equilibrium holds
- does not work well on non-convex tasks (some mixture of gaussians, ICA)

As SVGD, a kernel-based sampling algorithm which is hard to analyze... (in particular with an unbounded kernel!)

Thank you for listening, questions?

A Non Asymptotic Analysis of Stein Variatonal Gradient Descent (2020).
A. Korba, A. Salim, M. Arbel, G. Luise, A. Gretton.
Advances in neural information processing systems.

Kernel Stein Discrepancy Descent (2021).
A. Korba, P-C. Aubin-Frankowski, S. Majewski, P. Ablin.
submitted, soon on Arxiv.

References I

Alquier, P. and Ridgway, J. (2017).
Concentration of tempered posteriors and of their variational approximations.
arXiv preprint arXiv:1706.09293.

Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media.

Chu, C., Minami, K., and Fukumizu, K. (2020).
The equivalence between stein variational gradient descent and black-box variational inference.
arXiv preprint arXiv:2004.01822.

References II

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International conference on machine learning*.

Dalalyan, A. S. (2017).
Theoretical guarantees for approximate sampling from smooth and log-concave densities.
Journal of the Royal Statistical Society.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. (2018).
A stein variational newton method.
In *Advances in Neural Information Processing Systems*, pages 9169–9179.

References III

Duncan, A., Nüsken, N., and Szpruch, L. (2019).
On the geometry of stein variational gradient descent.
arXiv preprint arXiv:1912.00894.

Durmus, A., Majewski, S., and Miasojedow, B. (2019).
Analysis of langevin monte carlo via convex optimization.
Journal of Machine Learning Research, 20(73):1–46.

Durmus, A., Moulines, E., et al. (2017).
Nonasymptotic convergence analysis for the unadjusted langevin algorithm.
Annals of Applied Probability.

Feng, Y., Wang, D., and Liu, Q. (2017).
Learning to draw samples with amortized stein variational gradient descent.
arXiv preprint arXiv:1707.06626.

References IV

Gorham, J. and Mackey, L. (2017).
Measuring sample quality with kernels.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

Karimi, H., Nutini, J., and Schmidt, M. (2016).
Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

References V

Liu, C. and Zhu, J. (2018).
Riemannian stein variational gradient descent for bayesian inference.
In *Thirty-second aaai conference on artificial intelligence*.

Liu, D. C. and Nocedal, J. (1989).
On the limited memory BFGS method for large scale optimization.
Mathematical programming, 45(1-3):503–528.

Liu, Q. (2017).
Stein variational gradient descent as gradient flow.
In *Advances in neural information processing systems*, pages 3115–3123.

References VI

Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances in neural information processing systems*, pages 2378–2386.

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017).
Stein variational policy gradient.
arXiv preprint arXiv:1704.02399.

Lu, J., Lu, Y., and Nolen, J. (2019).
Scaling limit of the stein variational gradient descent: The mean field regime.
SIAM Journal on Mathematical Analysis, 51(2):648–671.

References VII

Otto, F. (2001).
The Geometry of Dissipative Evolution Equations: The Porous Medium Equation.
Communications in Partial Differential Equations, 26(1-2):101–174.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. (2017).
Vae learning via stein variational gradient descent.
In *Advances in Neural Information Processing Systems*, pages 4236–4245.

Steinwart, I. and Christmann, A. (2008).
Support vector machines.
Springer Science & Business Media.

References VIII

Wang, D. and Liu, Q. (2016).
Learning to draw samples: With application to amortized mle for generative adversarial learning.
arXiv preprint arXiv:1611.01722.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018).
Advances in variational inference.
IEEE transactions on pattern analysis and machine intelligence.