Stein Variational Gradient Descent, an optimization algorithm for sampling

Adil Salim

Microsoft Research

April 2022

Sampling

- Bayesian inference in machine learning
- GANs, etc.

Figure: `thispersondoesnotexist.com`

Vision

We can leverage the extensive toolbox of optimization to design and analyze sampling algorithms.

Outline

Outline

Outline

A quote by Andrew Wiles

"Perhaps I can best describe my experience of doing mathematics in terms of a journey through a dark unexplored mansion."

Gradient Descent (GD)

Problem:

$$\min_{x \in \mathbb{R}^d} F(x),$$

where F differentiable.

Gradient descent algorithm: $x^0 = a$ and

$$x^{n+1} = x^n - \gamma^n \nabla F(x^n), \quad \gamma^n > 0.$$

Derivation

Taylor expansion:

$$F(x^n + h) = F(x^n) + \langle \nabla F(x^n), h \rangle + o(h),$$

Then,

$$h \propto \arg\min_{h:\|h\|=1} F(x^n) + \langle \nabla F(x^n), h \rangle \propto -\nabla F(x^n),$$

And $x^{n+1} = x^n + h$.

Outline

Wasserstein distance

The Reference: [Ambrosio et al., 2008].

- $\mathcal{P}(\mathbb{R}^d)$ of probability measures μ over $\mathbb{R}^d$ with finite second moment.
- The 2-Wasserstein distance W.

$$W^2(\mu,\nu) := \inf_{(X,Y)} \mathbb{E}(\|X-Y\|^2), \text{ s.t. } \quad X \sim \mu, Y \sim \nu.$$

- Brenier's theorem: If μ has density, inf achieved by $(X, T_\mu^\nu(X))$, where T_μ^ν Brenier's map ($T_\mu^\nu \# \mu = \nu$).
- $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d \ s.t. \ \int \|f\|^2 d\mu < \infty\}$.

The Wasserstein space: $(\mathcal{P}(\mathbb{R}^d), W)$.

Examples of functionals

Let $\mu^\star \propto \exp(-F)$.

- **Potential**. $\mathcal{E}_F(\mu) := \int F d\mu$.
- **Negative entropy**. $\mathcal{H}(\mu) := \int \mu(x)\log(\mu(x))dx$.
- **Kullback-Leibler (KL)**. $\mathcal{V} := \mathcal{H} + \mathcal{E}_F$.

$$0 \leq \mathrm{KL}(\mu|\mu^\star) := \int \log\left(\frac{d\mu}{d\mu^\star}(x)\right) d\mu(x) = \mathcal{V}(\mu) - \mathcal{V}(\mu^\star).$$

Sampling as optimization

Goal: Sample from $\mu^\star \propto \exp(-F)$.

$$\mu^\star = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\mu^\star) = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{V}(\mu)$$

Gradient descent for $\mathcal{V}$?

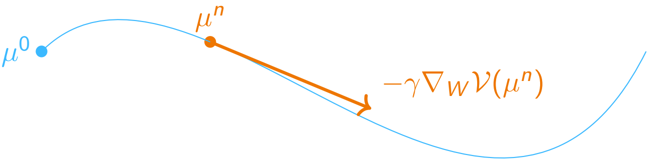
Examples of Wasserstein gradients

Wasserstein gradients belong to tangent space:
$\nabla_W \mathcal{V}(\mu) \in L^2(\mu)$.

- **Potential**. $\nabla_W \mathcal{E}_F(\mu) = \nabla F$.
- **Negative entropy**. $\nabla_W \mathcal{H}(\mu) = \nabla \log(\mu)$.
- **Kullback-Leibler (KL)**.
  $\nabla_W \mathcal{V}(\mu) = \nabla_W \mathcal{H}(\mu) + \nabla_W \mathcal{E}_F(\mu) = \nabla \log \frac{d\mu}{d\mu^\star}$.

Wasserstein Gradient Descent [SKL20]

$$\mu^{n+1} = (I - \gamma \nabla_W \mathcal{V}(\mu^n))\#\mu^n$$



Outline

Sampling framework

$$\mu^\star(x) \propto \exp(-F(x))$$

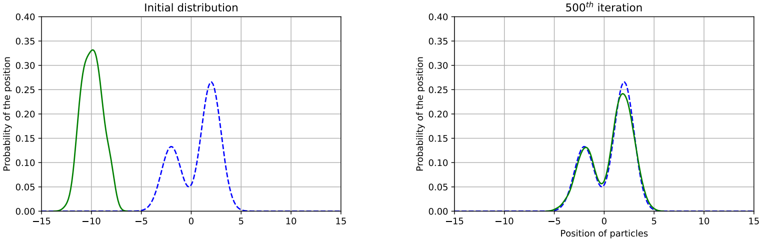


Figure: Simulation from [KSA+20] (Code from Q. Liu)

Outline

Stein Variational Gradient Descent

SVGD [Liu and Wang, 2016] to sample from $\mu^\star \propto \exp(-F)$.
SVGD maintains a set of N particles $x_1, \ldots, x_N$.

$$x_i^{n+1} = x_i^n - \frac{\gamma}{N} \sum_{j=1}^{N} \nabla F(x_j^n) k(x_i^n, x_j^n) - \nabla_2 k(x_i^n, x_j^n),$$

where $k(x, y)$ is a kernel associated to a **Reproducing Kernel Hilbert Space** H.

Reproducing kernel Hilbert Space

- Hilbert space of functions H (here, $H \subset L^2(\mu)$ for every μ)
- For every x, $k(x, \cdot) \in H$
- Reproducing property: for every $f \in H$, $\langle f, k(x, \cdot) \rangle_H = f(x)$.

Applications of SVGD

Stein variational policy gradient
Y Liu, P Ramachandran, Q Liu, J Peng - arXiv preprint arXiv:1704.02399, 2017 - arxiv.org
… We then propose a novel **Stein variational** pol- icy gradient method (SVPG) which combines ex- isting policy … Recent advances in policy gradient methods and deep **learning** have demonstrated their applicability for com- plex **reinforcement learning** problems …
☆ 99 Cited by 76 Related articles All 6 versions

VAE learning via **Stein variational** gradient descent
Y Pu, Z Gan, R Henao, C Li, S Han, L **Carin** - arXiv preprint arXiv …, 2017 - arxiv.org
A new method for learning variational autoencoders (VAEs) is developed, based on Stein variational gradient descent. A key advantage of this approach is that one need not make parametric assumptions about the form of the encoder distribution. Performance is further …
☆ 99 Cited by 43 Related articles All 7 versions

Scalable thompson sampling via optimal transport
R **Zhang**, Z Wen, C Chen, L Carin - arXiv preprint arXiv:1902.07239, 2019 - arxiv.org
Thompson sampling (TS) is a class of algorithms for sequential decision-making, which requires maintaining a posterior distribution over a model. However, calculating exact posterior distributions is intractable for all but the simplest models. Consequently, efficient …
☆ 99 Cited by 9 Related articles All 7 versions

[PDF] Accelerated first-order methods on the wasserstein space for bayesian inference
C Liu, J Zhuo, P Cheng, R Zhang, J Zhu, L **Carin** - stat, 2018 - researchgate.net
We consider doing Bayesian inference by minimizing the KL divergence on the 2-Wasserstein space P2. By exploring the Riemannian structure of P2, we develop two inference methods by simulating the gradient flow on P2 via updating particles, and an …
☆ 99 Cited by 10 Related articles All 3 versions

Understanding and accelerating particle-based **variational** inference
C Liu, J Zhuo, P Cheng, R Zhang… - … Conference on Machine …, 2019 - proceedings.mlr.press
… Correspondence to: Jun Zhu <dcszj@tsinghua.edu.cn>, **Lawrence Carin** <lcarin@duke.edu> … 2.3. Particle-Based **Variational** Inference Methods **Stein Variational** Gradient Descent (SVGD) (Liu & Wang, 2016) uses a vector field v to update particles: x (i) k+1 …
☆ 99 Cited by 20 Related articles All 14 versions

Policy optimization as wasserstein gradient flows
R Zhang, C Chen, C Li, L **Carin** - … Conference on Machine …, 2018 - proceedings.mlr.press
Policy optimization is a core component of reinforcement learning (RL), and most existing RL methods directly optimize parameters of a policy based on maximizing the expected total reward, or its surrogate. Though often achieving encouraging empirical success, its …
☆ 99 Cited by 32 Related articles All 9 versions

Variational annealing of GANs: A Langevin perspective
C Tao, S Dai, L Chen, K Bai, J Chen… - International …, 2019 - proceedings.mlr.press
… Correspondence to: Chenyang Tao <chenyang.tao@duke.edu>, **Lawrence Carin** <lcarin@duke.edu> … Recent attempts have been made to fill this gap: **Stein variational** gradient descent (SVGD) used the kernel trick to analytically derive the functional gradient of an implicitly …
☆ 99 Cited by 5 Related articles All 7 versions

Variance reduction in stochastic particle-optimization sampling
J **Zhang**, Y Zhao, C Chen - International Conference on …, 2020 - proceedings.mlr.press
Stochastic particle-optimization sampling (SPOS) is a recently-developed scalable Bayesian sampling framework unifying stochastic gradient MCMC (SG-MCMC) and Stein variational gradient descent (SVGD) algorithms based on Wasserstein gradient flows. With a rigorous …
☆ 99 Cited by 1 Related articles All 5 versions

Stochastic particle-optimization sampling and the non-asymptotic convergence theory
J Zhang, R Zhang, L **Carin**… - … Conference on Artificial …, 2020 - proceedings.mlr.press
Particle-optimization-based sampling (POS) is a recently developed effective sampling technique that interactively updates a set of particles. A representative algorithm is the Stein variational gradient descent (SVGD). We prove, under certain conditions, SVGD …
☆ 99 Cited by 18 Related articles All 5 versions

Stein variational gradient descent: A general purpose bayesian inference algorithm
Q Liu, D Wang - arXiv preprint arXiv:1608.04471, 2016 - arxiv.org
Page 1. **Stein Variational** Gradient Descent: A General Purpose Bayesian Inference Algorithm Qiang Liu Dilin Wang Department of Computer Science Dartmouth College Hanover, NH 03755 {qiang.liu, dilin.wang.gr}@dartmouth.edu … 3.2 **Stein Variational** Gradient Descent …
☆ 99 Cited by 371 Related articles All 8 versions

Outline

Taylor expansion of KL [Liu and Wang, 2016, Liu, 2017]

$$\mu^\star = \arg\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\mu^\star) = \arg\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{V}(\mu) \propto \exp(-F)$$

$$\mathrm{KL}((I+h)\#\mu^n|\mu^\star) = \mathrm{KL}(\mu^n|\mu^\star) - \mathbb{E}_\mu\left(S^\star h\right) + o(h),$$

where $S^\star$ is the Stein operator

$$S^\star h = -\langle \nabla F, h \rangle + \mathrm{div}(h).$$

Then,

$$h \propto \arg\min_{h:\|h\|=1} \mathrm{KL}(\mu^n|\mu^\star) - \mathbb{E}_\mu\left(S^\star h\right) \propto \quad ?$$

A key formula

$$
\begin{aligned}
\mathbb{E}_\mu S^\star h &= -\int \langle \nabla F, h \rangle d\mu + \int \mathrm{div}(h) d\mu \\
&= -\int \langle \nabla F, h \rangle d\mu + \int \mathrm{div}(h)(x)\mu(x) dx \\
&= -\int \langle \nabla F, h \rangle d\mu - \int \langle h(x), \nabla \mu(x) \rangle dx \\
&= -\int \langle \nabla F, h \rangle d\mu - \int \langle h(x), \nabla \log \mu(x) \rangle d\mu(x) \\
&= -\int \left\langle \nabla \log \frac{d\mu}{d\mu^\star}, h \right\rangle d\mu \\
&= -\langle \nabla_W \mathcal{V}(\mu), h \rangle_{L^2(\mu)}
\end{aligned}
$$

Consequences

- Stein's identity: if $\mu = \mu^\star$, $\mathbb{E}_\mu S^\star h = 0$
- $h \propto \arg\min_{h:\|h\|_{L^2(\mu)}=1} \mathrm{KL}(\mu|\mu^\star) - \mathbb{E}_\mu (S^\star h) \propto \nabla_W \mathcal{V}(\mu)$
- $h \propto \arg\min_{h:\|h\|_{H}=1} \mathrm{KL}(\mu|\mu^\star) - \mathbb{E}_\mu (S^\star h) \propto$?

A key formula (cont.)

Assume $h \in H$.

$$
\begin{aligned}
\mathbb{E}_\mu S^\star h &= \int \left\langle \nabla \log \frac{d\mu}{d\mu^\star}(x), h(x) \right\rangle d\mu(x) \\
&= \int \left\langle \nabla \log \frac{d\mu}{d\mu^\star}(x), \langle h, k(x,\cdot)\rangle_H \right\rangle d\mu(x) \\
&= \left\langle h, \int \nabla \log \frac{d\mu}{d\mu^\star}(x) k(x,\cdot) d\mu(x) \right\rangle_H \\
&= \langle h, P_\mu \nabla_W \mathcal{V}(\mu)\rangle_H ,
\end{aligned}
$$

where $P_\mu f := \int f(x)k(x,\cdot)d\mu(x)$.

SVGD in the population limit

$$h \propto \arg\min_{h:\|h\|_H=1} \mathrm{KL}(\mu^n|\mu^\star) - \mathbb{E}_\mu\left(S^\star h\right) \propto \quad P_\mu \nabla_W \mathcal{V}(\mu)$$

$$\mu^{n+1} = \left(I - \gamma P_{\mu^n} \nabla_W \mathcal{V}(\mu^n)\right) \# \mu^n.$$

Remark:
$\mathrm{KSD}(\mu|\mu^\star) := \max_{h:\|h\|_H=1} \mathbb{E}_\mu\left(S^\star h\right) = \|P_\mu \nabla_W \mathcal{V}(\mu)\|_H$ [Liu et al., 2016, Chwialkowski et al., 2016, Oates et al., 2019, Gorham and Mackey, 2017].

Why do we use H?

Iteration "tractable"!

$$P_\mu \nabla_W \mathcal{V}(\mu) = \int k(\cdot, x)\nabla F(x) - \nabla_x k(\cdot, x) d\mu(x),$$

using again integration by parts.

Algorithm:

$$x_i^{n+1} = x_i^n - \frac{\gamma}{N}\sum_{j=1}^{N} \nabla F(x_j^n) k(x_i^n, x_j^n) - \nabla_2 k(x_i^n, x_j^n).$$

Conclusion

$P_\mu \nabla_W \mathcal{V}(\mu)$ is the Wasserstein gradient of $\mathcal{V}$ at μ under the metric of H.

This change of metric makes the iterations tractable. $\nabla_W \mathcal{V}(\mu)$ is intractable but $P_\mu \nabla_W \mathcal{V}(\mu)$ is!

SVGD is a "Wasserstein gradient descent" in the metric of H [Liu, 2017, Duncan et al., 2019, Chewi et al., 2020, Nüsken and Renger, 2021, Shi et al., 2021], [KSA+20, SSR21].

Complexity and convergence of SVGD

In the population limit, we can analyze SVGD as a gradient descent:

Theorem 1 (Complexity and convergence of SVGD [KSA+20, SSR21])

Convergence: If $\int \exp(\beta\|x - a\|^2) d\mu^\star(x) < \infty$, then $W_1(\mu^n, \mu^\star) \to 0$.
Complexity: $n = \mathcal{O}(d^{3/2}/\varepsilon)$ in KSD^2.

References I

[Ambrosio et al., 2008] Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient Flows: In Metric Spaces and in the Space of Probability Measures.
Springer Science & Business Media.

[Chewi et al., 2020] Chewi, S., Gouic, T. L., Lu, C., Maunu, T., and Rigollet, P. (2020).
Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Chwialkowski et al., 2016] Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International Conference on Machine Learning (ICML)*, pages 2606–2615.

[Duncan et al., 2019] Duncan, A., Nuesken, N., and Szpruch, L. (2019).
On the geometry of stein variational gradient descent.
arXiv preprint arXiv:1912.00894.

[Gorham and Mackey, 2017] Gorham, J. and Mackey, L. (2017).
Measuring sample quality with kernels.
In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

[Liu, 2017] Liu, Q. (2017).
Stein variational gradient descent as gradient flow.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Liu et al., 2016] Liu, Q., Lee, J., and Jordan, M. (2016).
A kernelized Stein discrepancy for goodness-of-fit tests.
In *International Conference on Machine Learning (ICML)*, pages 276–284.

References II

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

[Nüsken and Renger, 2021] Nüsken, N. and Renger, D. (2021).
Stein variational gradient descent: many-particle and long-time asymptotics.
arXiv preprint arXiv:2102.12956.

[Oates et al., 2019] Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2019).
Convergence rates for a class of estimators based on stein's method.
Bernoulli, 25(2):1141–1159.

[Shi et al., 2021] Shi, J., Liu, C., and Mackey, L. (2021).
Sampling with mirrored stein operators.
arXiv preprint arXiv:2106.12506.

Selected publications I

[KSA+20] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton.
A non-asymptotic analysis for Stein variational gradient descent.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[SKL20] Adil Salim, Anna Korba, and Giulia Luise.
The Wasserstein proximal gradient algorithm.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[SSR21] Adil Salim, Lukang Sun, and Peter Richtárik.
Complexity analysis of stein variational gradient descent under talagrand's inequality t1.
arXiv preprint arXiv:2106.03076, 2021.