A Convergence Theory for SVGD in the Population Limit under Talagrand's Inequality T1

Adil Salim Lukang Sun Peter Richtárik

Microsoft Research, USA. KAUST, Saudi Arabia

ICML 2022

Sampling framework

- F smooth and nonconvex

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

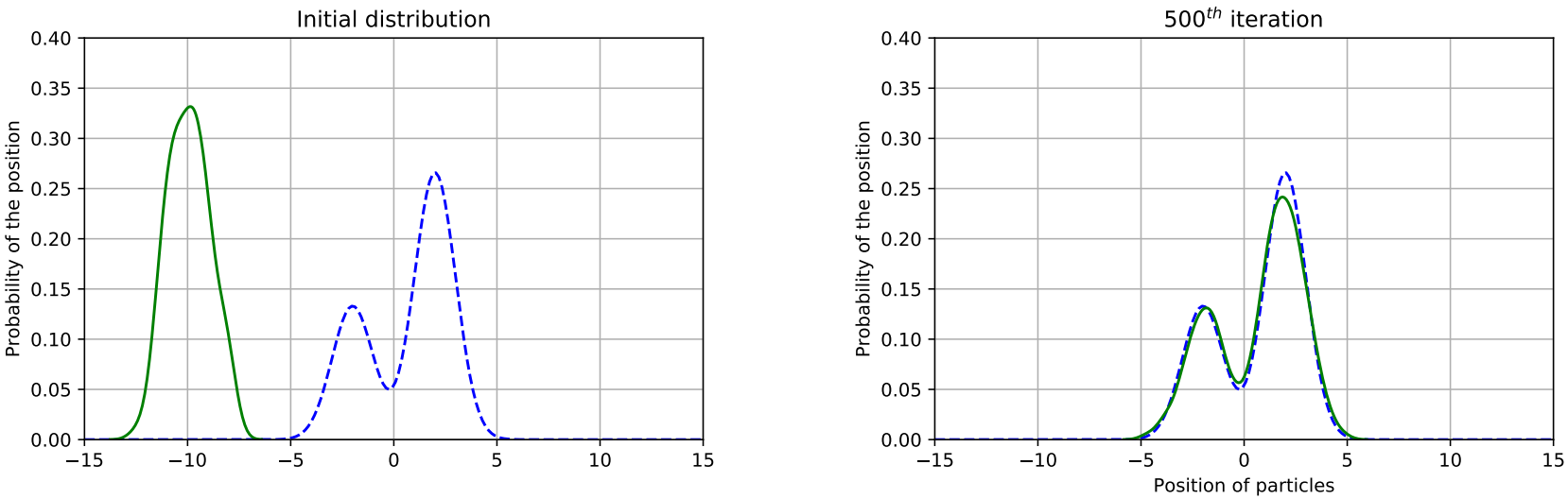


Figure: Simulation from [KSA+20] (Code from Q. Liu)

Sampling as optimization

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

where

$$\mathrm{KL}(\mu|\mu^\star) := \int \log\left(\frac{d\mu}{d\mu^\star}(x)\right) d\mu(x) \text{ if } \mu \ll \mu^\star, +\infty \text{ else,}$$

satisfies $\mathrm{KL}(\mu|\mu^\star) \geq 0$ and $\mathrm{KL}(\mu|\mu^\star) = 0$ iff $\mu = \mu^\star$.

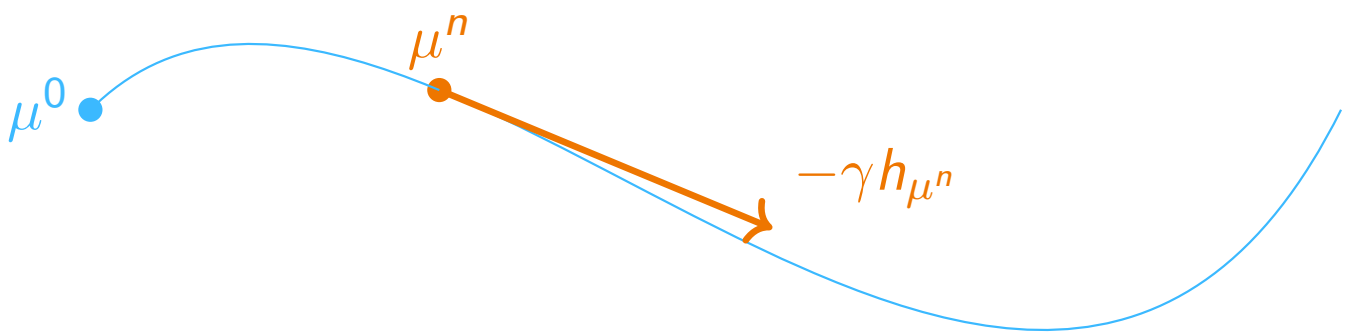
Gradient descent for $\mathrm{KL}(\,\cdot\,|\mu^\star)$?

Stein Variational Gradient Descent (SVGD)

$$\mu^{n+1} = (I - \gamma h_{\mu^n}) \# \mu^n, \tag{1}$$

where $h_\mu := \int \nabla F(x) k(\cdot, x) - \nabla_2 k(\cdot, x) d\mu(x)$ and $k(x, y)$ kernel of a RKHS H.

- When μ^0 is discrete then μ^n is discrete and (1) is called SVGD algorithm [Liu and Wang, 2016].
- When μ^0 has a density then μ^n has a density [SSR22] and (1) is called SVGD in the population limit.



Kernelized Stein Discrepancy

Remark: $\mathrm{KSD}(\mu|\mu^\star) := \|h_\mu\|_H$
[Liu et al., 2016, Chwialkowski et al., 2016, Oates et al., 2019, Gorham and Mackey, 2017].

If H rich enough, $\mathrm{KSD}(\mu|\mu^\star) = 0 \Longrightarrow \mu = \mu^\star$.

Analysis in the population limit

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., 2022, Gorham et al., 2020], [KSA+20, SSR22].

Theorem 1 (Complexity and convergence of SVGD in the population limit [KSA+20, SSR22])

Convergence: If μ^0 has a density and if $\int \exp(\beta\|x-a\|^2)d\mu^\star(x) < \infty$, then $W_1(\mu^n, \mu^\star) \to 0$.
Moreover, complexity $n = \mathcal{O}(d^{3/2}/\varepsilon)$ to output μ s.t. $\mathrm{KSD}^2(\mu|\mu^\star) < \varepsilon$.

Proof

$$\mathrm{KL}(\mu^{n+1}|\mu^\star) \leq \mathrm{KL}(\mu^n|\mu^\star) - \gamma(1-\gamma C) \underbrace{\|h_{\mu^n}\|_H^2}_{=\mathrm{KSD}^2(\mu^n|\mu^\star)} .$$

References I

[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 (ICML)*, pages 1292–1301.

[Gorham et al., 2020] Gorham, J., Raj, A., and Mackey, L. (2020).
Stochastic stein discrepancies.
Advances in Neural Information Processing Systems (NeurIPS), 33:17931–17942.

[Liu, 2017] Liu, Q. (2017).
Stein variational gradient descent as gradient flow.
In *Advances in Neural Information Processing Systems (NIPS)*, volume 30.

[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.

[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.

References II

[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., 2022] Shi, J., Liu, C., and Mackey, L. (2022).
Sampling with mirrored stein operators.
In *International Conference on Learning Representations (ICLR)*.

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.

[SSR22] Adil Salim, Lukang Sun, and Peter Richtárik.
A convergence theory for SVGD in the population limit under Talagrand's inequality T1.
In *International Conference on Machine Learning (ICML)*, 2022.