Imperial College London

Grassmann Stein Variational Gradient Descent

Xing Liu [1], Harrison Zhu [1], Jean-François Ton [2], George Wynne [1], Andrew Duncan [1]

[1]Imperial College London [2]University of Oxford

Introduction

Let P be a distribution over $\mathcal{X} \subset \mathbb{R}^d$ that admits a smooth density p. Assume p can be evaluated **up to a proportionality**. We want to **approximate** P **by particles** $\{x_i\}_i^n$.

- **Application:** Bayesian inference
- **Methods:** Markov chain Monte Carlo, Variational inference, **Stein Variational Gradient Descent**.

Summary:

1. **Stein Variational Gradient Descent** (SVGD) is a promising Bayesian inference method, but suffers from **under-estimation of variance** in high dimensions.
2. Recent advances address this issue via 1-dimensional projections (slices), which might be sub-optimal in terms of uncertainty estimation.
3. We propose **Grassmann Stein Variational Gradient Descent** (GSVGD), which tackles this sub-optimality by projecting onto arbitrary subspaces.

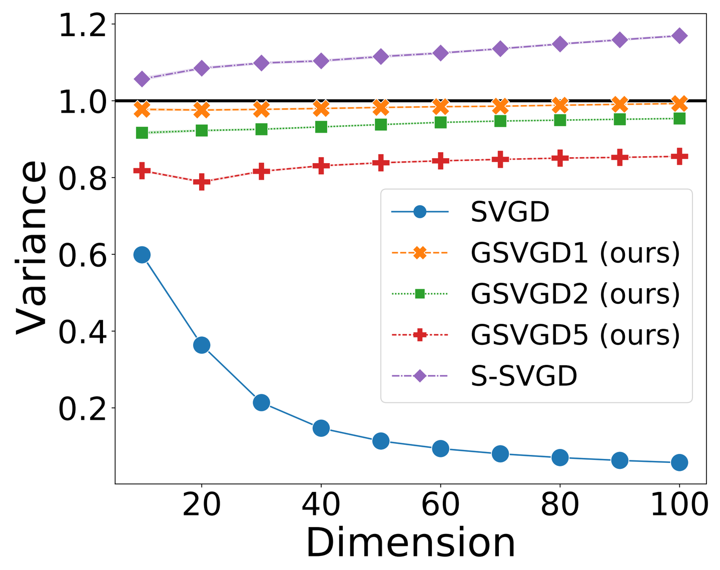

Figure 1. Estimating the dimension-averaged marginal variance of a multivariate Gaussian density $p(x) = \mathcal{N}(x; 0, I_d)$ with different dimensions d.

Stein Variational Gradient Descent (SVGD)

SVGD [1] starts with i.i.d. particles $X := (x_1, \dots, x_n)$ drawn from an initial distribution Q, and iteratively updates X by minimizing the KL divergence of its empirical distribution from P:

$$T_\phi(x) = x + \epsilon\phi^*(x), \qquad \phi^* = \arg\min_{\phi \in \mathcal{B}_k^d} \mathrm{KL}(T_{\phi,\#}Q \| P),$$

where $\epsilon > 0$ is a small perturbation size, $\mathcal{B}_k^d := \{\phi \in \mathcal{H}_k^d : \|\phi\|_{\mathcal{H}_k^d} \le 1\}$ is the unit ball of the d-times product of RKHS $\mathcal{H}_k \times \cdots \times \mathcal{H}_k$ of RKHS $\mathcal{H}_k$ with a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and $T_{\phi,\#}Q$ is the pushforward of Q with respect to T_ϕ. The optimal ϕ^* can be derived (and estimated) explicitly

$$\phi^*(\cdot) = \mathbb{E}_Q[\mathcal{A}_p k(x, \cdot)] \approx \frac{1}{n}\sum_{i=1}^n k(x_i, \cdot)s_p(x_i) + \nabla_{x_i}k(x_i, \cdot) := \hat{\phi}^*(X, \cdot), \tag{1}$$

where $\mathcal{A}_p\phi(x) := s_p(x)\cdot\phi(x) + \nabla\cdot\phi(x)$ is the **(Langevin) Stein operator** and $s_p(x) := \nabla \log p(x)$ is the **score function** of p. The maximum rate of decay of the KL divergence given by ϕ^* coincides with the **kernelized Stein discrepancy** (KSD)

$$\mathrm{KSD}(Q, P) = \sup_{\phi \in \mathcal{B}_k^d} \mathbb{E}_Q[\mathcal{A}_p\phi(x)] = \sup_{\phi \in \mathcal{B}_k^d}\left\{-\frac{d}{d\epsilon}\mathrm{KL}(T_{\phi,\#}Q\|P)|_{\epsilon=0}\right\}. \tag{2}$$

Algorithm (SVGD [1]):

1. Start with $\{x_i^0\}_{i=1}^n$ drawn from some distribution Q.
2. For $t = 0, 1, \dots$, do $x_i^{t+1} = x_i^t + \epsilon\hat{\phi}^*(X^t, x_i^t)$, where $\hat{\phi}^*$ is given by Eq. 1.

Remarks:

- **SVGD update rule:** $\hat{\phi}^*$ leads to provable convergence to the target P under mild conditions, and each of the two terms in ϕ^* plays an intuitive role.
- **Curse of dimensionality:** suffers from **under-estimation of variance** for high dimensional problems. This attributes to the high dimensionality of both x and $s_p(x)$.

Sliced Stein Variational Gradient Descent (S-SVGD)

S-SVGD [2] is an extension of SVGD that tackles the curse of dimensionality by using **slices** (1-dim projections). In S-SVGD, the update rule is $\phi^* = (\phi_1^*, \phi_2^*, \dots, \phi_d^*)^\intercal$, where

$$\phi_j^*(\cdot) = \mathbb{E}_Q[r_j^\intercal s_p(x) k_{r_j g_j}(g_j^\intercal x, g_j^\intercal \cdot) + r_j^\intercal g_j \nabla_{g_j^\intercal x} k_{r_j g_j}(g_j^\intercal x, g_j^\intercal \cdot)],$$

where $O = (r_1, r_2, \dots, r_d)$ is a fixed orthonormal basis of $\mathbb{R}^d$, and $g_j \in \mathbb{S}^{d-1}$ are optimised by maximising a sliced discrepancy, called **max sliced KSD**.

Remarks:

- **Tackling curse of dimensionality:** S-SVGD sidesteps the under-estimation-of-variance issue of SVGD, as the particles are effectively transported along 1-dim subspaces at each step.
- **Fixed 1-dim slices:** However, the basis O is not optimised and both r_j and g_j are constrained to 1-dim, which may result in slower convergence and sub-optimal covariance estimation.

Grassmann Variational Gradient Step (GSVGD)

We propose **GSVGD**, which projects x and $s_p(x)$ onto subspaces of **an arbitrary dimension**, say m where $1 \le m \le d$.

- **Definition.** The **Grassmann kernelized Stein discrepancy**, $\mathrm{GKSD}(Q, P)$, between two distributions Q and P is

$$\mathrm{GKSD}(Q, P) = \sup_{[A] \in \mathrm{Gr}(d,m)} \mathrm{KSD}_A(Q, P), \quad \text{where} \tag{3}$$
$$\mathrm{KSD}_A(Q, P) = \sup_{\phi \in \mathcal{B}_{k_A}} \mathbb{E}_Q[\mathcal{A}_p\phi(x)] = \sup_{\phi \in \mathcal{B}_k^m} \mathbb{E}_Q[(A^\intercal s_p(x))\cdot\phi(A^\intercal x) + \nabla\cdot\phi(A^\intercal x)], \tag{4}$$

where $\mathcal{B}_{k_A}$ is a RKHS with kernel k_A, $\mathrm{Gr}(d, m) := \{\mathrm{Image}(A) \subset \mathbb{R}^d : AA^\intercal = I_m\}$ is the set of m-**dimensional subspaces** of $\mathbb{R}^d$ identified by **projector** A. $\mathrm{Gr}(d, m)$ is known as the **Grassmann manifold** (hence the name GSVGD).

The sup in Eq. 3 is taken over $\mathrm{Gr}(d, m)$ but not over all possible projectors A because we only care about **where** we project onto (subspace), but **not how** (projector A).

- The **GSVGD update rule** is

$$\phi_A^*(\cdot) = \mathbb{E}_Q[\mathcal{A}_p k_A(x, \cdot)] = \mathbb{E}_Q[AA^\intercal s_p(x)k(A^\intercal x, A^\intercal\cdot) + A\nabla_x k(A^\intercal x, A^\intercal \cdot)], \tag{5}$$

where the optimal A is sought using **Riemannian gradient descent + SDE**:

$$A \leftarrow \exp_{[A]}(\delta(I_m - AA^\intercal)\nabla\alpha([A]) + \sqrt{2T\delta}\xi), \tag{6}$$

where $\alpha([A]) := \mathrm{KSD}_A(Q, P)$ is the objective, $\delta > 0$ is the step size, ξ is $d \times m$ whose entries are i.i.d. $\mathcal{N}(0, 1)$ noise, $T > 0$ is the noise level, and $\exp_{[A]}(B)$ ensures A remains a projector.

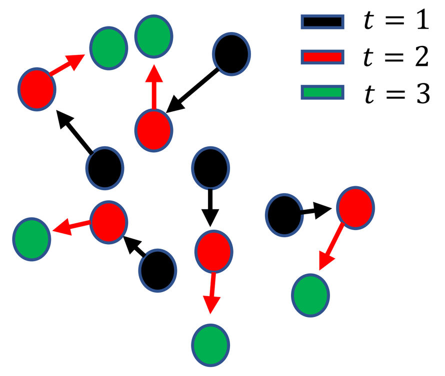

Figure 2. Two steps of particle descent (Eq. 5).

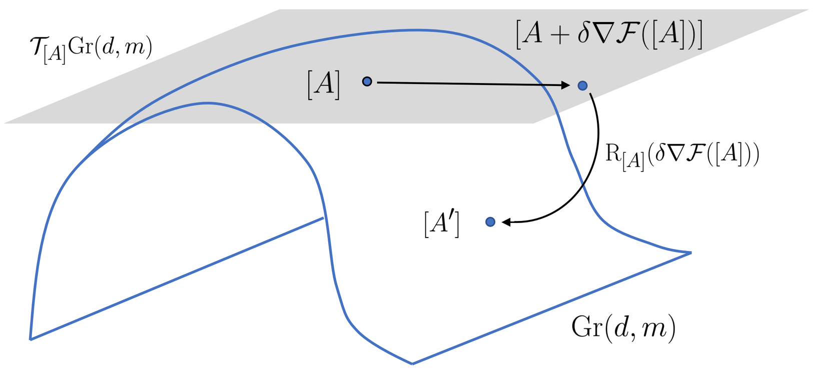

Figure 3. One step of Riemannian Gradient Descent + SDE (Eq. 6).

Algorithm (GSVGD; the proposed method)

1. Start with $\{x_i^0\}_{i=1}^n$ drawn from Q, and initialize M projectors $A_{t,1}, \dots, A_{t,M}$.
2. For $t = 0, 1, \dots$,
   i. Update each particle by $x_i^{t+1} = x_i^t + \epsilon\sum_{l=1}^M \hat{\phi}_{A_{t,l}}(x_i^t)$, where $\hat{\phi}_{A_{t,l}}$ is an estimate of Eq. 5.
   ii. Update each projector $A_{t,l}$ by Eq. 6.

Remarks:

- **Batched algorithm:** $M \ge 1$ projectors $A_1, \dots, A_M$ are used simultaneously to improve convergence.
- **Validity:** GKSD distinguishes distributions, meaning that $\mathrm{GKSD}(Q, P) = 0 \iff Q = P$.
- **Convergence:** can be established by viewing the update as a discretised ODE-SDE system.
- **Tackling curse of dimensionality:** solving the under-estimation-of-variance issue by transporting particles along lower dimensional subspaces.

Experiments

Experiment 1: Conditioned Diffusion Process

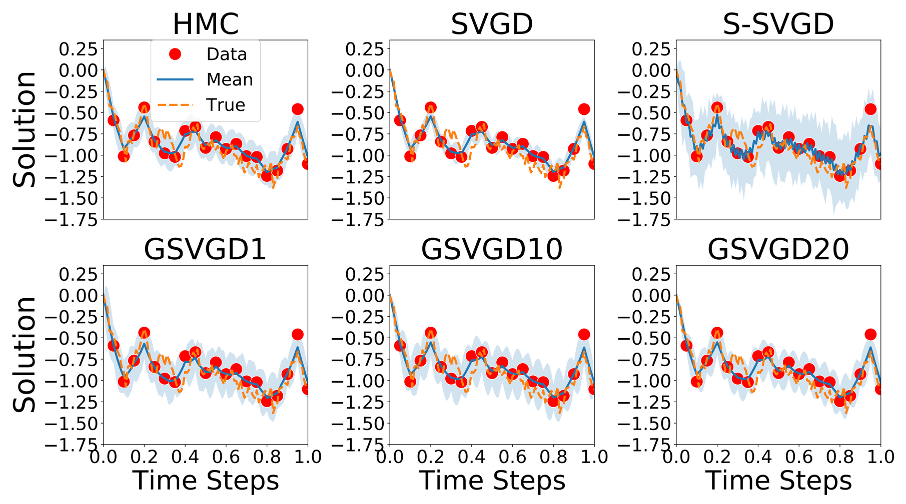

Figure 4. Estimating the posterior mean and variance of the conditioned diffusion SDE dynamic.

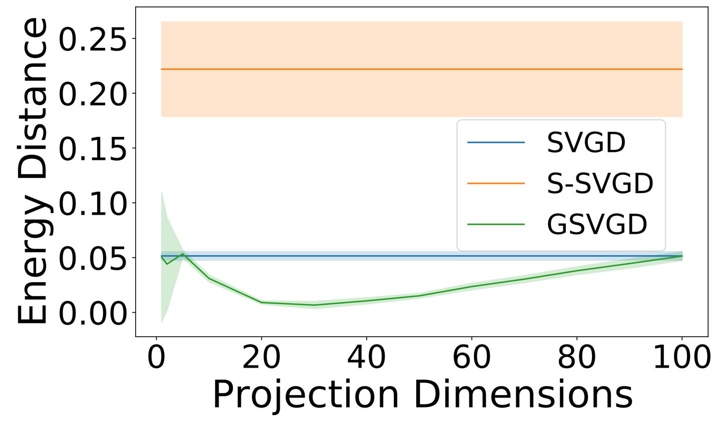

Figure 5. Estimation quality of GSVGD with various projection dimensions m compared with its competing methods.

Experiment 2: Bayesian Logistic Regression with the `covertype` Dataset

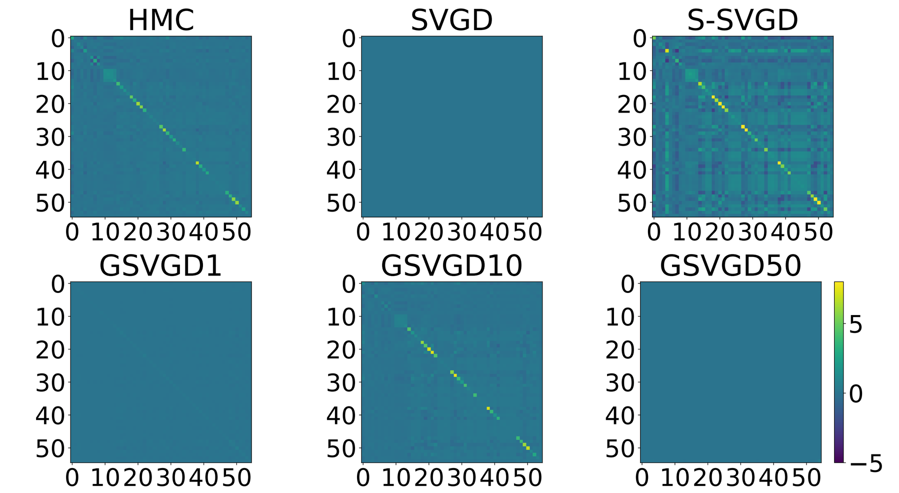

Figure 6. Estimating the posterior covariance matrix of the parameters of a Bayesian logistic regression model.

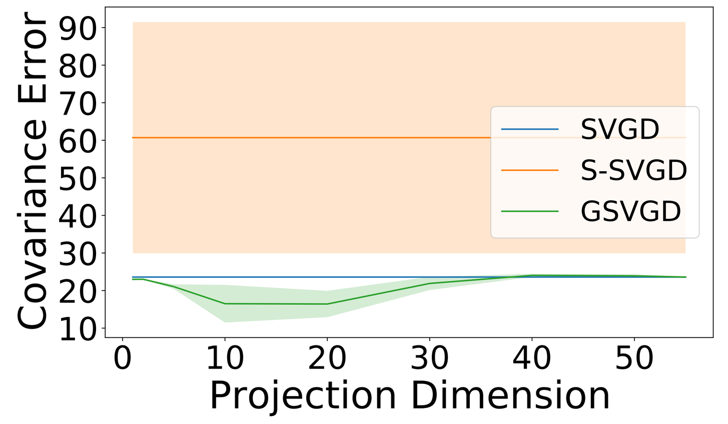

Figure 7. Covariance estimation error of GSVGD with various projection dimensions m compared with its competing methods.

References

[1] Q. Liu and D. Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, 2016.

[2] W. Gong, Y. Li, and J. M. Hernández-Lobato, "Sliced Kernelized Stein Discrepancy," in *International Conference on Learning Representations*, 2021.

 xing.liu16@imperial.ac.uk