4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–29

Sampling with Mirrored Stein Operators

Jiaxin Shi JIAXINSHI@MICROSOFT.COM
Microsoft Research, Cambridge MA
Chang Liu CHANG.LIU@MICROSOFT.COM
Microsoft Research, Beijing
Lester Mackey LMACKEY@MICROSOFT.COM
Microsoft Research, Cambridge MA

Abstract

Accurately approximating an unnormalized distribution with a discrete sample is a fundamental challenge in machine learning, probabilistic inference, and Bayesian inference. Particle evolution methods like Stein variational gradient descent have found great success in approximating unconstrained distributions but break down for constrained targets. We introduce a new family of particle evolution samplers suitable for constrained domains and non-Euclidean geometries. They minimize the Kullback-Leibler (KL) divergence to constrained target distributions by evolving particles in a dual space defined by a mirror map. We derive these samplers from a new class of mirrored Stein operators and adaptive kernels developed in this work. We establish the convergence of our new procedures under verifiable conditions on the target distribution. Finally, we demonstrate that these new samplers yield accurate approximations to distributions on the simplex and deliver valid confidence intervals in post-selection inference.

1. Background: Mirror Descent

Standard gradient descent can be viewed as optimizing a local quadratic approximation to the target function f: $\theta_{t+1} = \operatorname{argmin}_{\theta\in\Theta} \nabla f(\theta_t)^\top \theta + \frac{1}{2\epsilon_t}\|\theta - \theta_t\|_2^2$. When $\Theta \subseteq \mathbb{R}^d$ is constrained, it can be advantageous to replace $\|\cdot\|_2$ with a function Ψ that reflects the geometry of a problem (Nemirovskij and Yudin, 1983; Beck and Teboulle, 2003): $\theta_{t+1} = \operatorname{argmin}_{\theta\in\Theta} \nabla f(\theta_t)^\top \theta + \frac{1}{\epsilon_t}\Psi(\theta, \theta_t)$. The mirror descent (MD) algorithm chooses Ψ to be the Bregman divergence induced by a strongly convex, essentially smooth[1] function $\psi : \Theta \to \mathbb{R} \cup \{\infty\}$: $\Psi(\theta, \theta') = \psi(\theta) - \psi(\theta') - \nabla\psi(\theta')^\top(\theta - \theta')$. The solution of the problem is

$$\theta_{t+1} = \nabla\psi^*(\nabla\psi(\theta_t) - \epsilon_t \nabla f(\theta_t)), \tag{1}$$

where $\psi^*(\eta) \triangleq \sup_{\theta\in\Theta} \eta^\top\theta - \psi(\theta)$ is the convex conjugate of ψ and $\nabla\psi$ is a bijection from Θ to $\operatorname{dom}(\psi^*)$ with inverse map $(\nabla\psi)^{-1} = \nabla\psi^*$. We can view (1) as first mapping θ_t to η_t by $\nabla\psi$, applying the update $\eta_{t+1} = \eta_t - \epsilon_t \nabla f(\theta_t)$, and mapping back through $\theta_{t+1} = \nabla\psi^*(\eta_{t+1})$.

We can view MD as a discretization of the continuous-time dynamics $d\eta_t = -\nabla f(\theta_t)dt$, $\theta_t = \nabla\psi^*(\eta_t)$. It is equivalent to the Riemannian gradient flow (see App. B):

$$d\theta_t = -\nabla^2\psi(\theta_t)^{-1}\nabla f(\theta_t)dt, \quad \text{or equivalently,} \quad d\eta_t = -\nabla^2\psi^*(\eta_t)^{-1}\nabla_{\eta_t} f(\nabla\psi^*(\eta_t))dt, \tag{2}$$

where $\nabla^2\psi(\theta)$ and $\nabla^2\psi^*(\eta)$ are Riemannian metric tensors. In information geometry, the discretization of (2) is known as *natural gradient* descent (Amari, 1998). There is considerable theoretical and practical evidence (Martens, 2014) showing that natural gradient works efficiently in learning.

[1] ψ is continuously differentiable on the interior of Θ with $\|\nabla\psi(\theta_t)\| \to \infty$ whenever $\theta_t \to \theta \in \partial\Theta$.

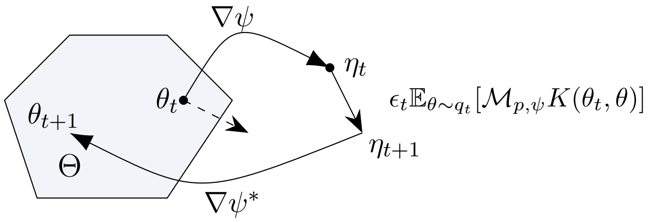

$\nabla\psi$
η_t
θ_t
$\epsilon_t \mathbb{E}_{\theta\sim q_t}[\mathcal{M}_{p,\psi} K(\theta_t,\theta)]$
θ_{t+1}
η_{t+1}
Θ
$\nabla\psi^*$

Figure 1: Updating particle approximations in constrained domains Θ. Standard updates like SVGD (dashed arrow) can push particles outside of the support. Our mirrored Stein updates in Alg. 1 (solid arrows) preserve the support.

2. Sampling with Mirrored Stein Operators

Particle evolution methods like Stein variational gradient descent (SVGD, Liu and Wang, 2016) apply deterministic updates to particles using operators based on Stein's method (Stein, 1972; Gorham and Mackey, 2015) and reproducing kernels (Berlinet and Thomas-Agnan, 2011) to sequentially minimize Kullback-Leibler (KL) divergence. Specifically, SVGD updates each particle in its approximation by applying $\theta_{t+1} = \theta_t + \epsilon_t g_t(\theta_t)$ for a chosen mapping $g_t : \mathbb{R}^d \to \mathbb{R}^d$. It chooses the mapping g_t^* that leads to the largest decrease in KL divergence to p in the limit as $\epsilon_t \to 0$. The following theorem summarizes their main findings.

Theorem 1 (Liu and Wang, 2016, Thm. 3.1) *Suppose $(\theta_t)_{t\geq 0}$ satisfies $d\theta_t = g_t(\theta_t)dt$ for bounded Lipschitz $g_t \in C^1 : \mathbb{R}^d \to \mathbb{R}^d$ and that θ_t has density q_t with $\mathbb{E}_{q_t}[\|\nabla \log q_t(\theta)\|_2] < \infty$. If $\mathrm{KL}(q_t\|p) \triangleq \mathbb{E}_{q_t}[\log(q_t(\theta)/p(\theta))]$ exists then,*

$$\tfrac{d}{dt}\mathrm{KL}(q_t\|p) = -\mathbb{E}_{q_t}[(\mathcal{S}_p g_t)(\theta)], \tag{3}$$

where $\mathcal{S}_p$ is the Langevin Stein operator *(Gorham and Mackey, 2015) given by*

$$(\mathcal{S}_p g)(\theta) = g(\theta)^\top \nabla \log p(\theta) + \nabla \cdot g(\theta). \tag{4}$$

For an *unconstrained* domain with $\mathbb{E}_p[\|\nabla \log p(\theta)\|_2] < \infty$, *Stein's identity* $\mathbb{E}_p[(\mathcal{S}_p g)(\theta)] = 0$ holds whenever $g \in C^1$ is bounded and Lipschitz by (Gorham et al., 2019, proof of Prop. 3), which ensures $q_t = p$ is a stationary point of the dynamics. To improve its current particle approximation, SVGD find the choice of g_t that most quickly decreases $\mathrm{KL}(q_t\|p)$ at time t, i.e., they minimize $\frac{d}{dt}\mathrm{KL}(q_t\|p)$ over a set $\mathcal{G}$ of candidate directions g_t. SVGD chooses g_t in a reproducing kernel Hilbert space (RKHS, Berlinet and Thomas-Agnan, 2011) norm ball $\mathcal{B}_{\mathcal{H}^d} = \{g : \|g\|_{\mathcal{H}^d} \leq 1\}$, where $\mathcal{H}^d$ is the product RKHS containing vector-valued functions with each component in the RKHS $\mathcal{H}$ of k. Then the optimal $g_t^* \in \mathcal{B}_{\mathcal{H}^d}$ that minimizes (3) is

$$g_t^* \propto g_{q_t,k}^* \triangleq \mathbb{E}_{q_t}[k(\theta,\cdot)\nabla \log p(\theta) + \nabla_\theta k(\theta,\cdot)] = \mathbb{E}_{q_t}[\mathcal{S}_p K_k(\cdot,\theta)],$$

where we let $K_k(\theta,\theta') = k(\theta,\theta')I$, and $\mathcal{S}_p K_k(\cdot,\theta)$ denotes applying $\mathcal{S}_p$ to each row of $K_k(\cdot,\theta)$.

2.1. Mirrored dynamics

SVGD encounters two difficulties when faced with a constrained support. First, the SVGD updates can push θ_t outside of its support Θ, rendering all future updates undefined. Second, Stein's identity $\mathbb{E}_p[(\mathcal{S}_p g)(\theta)] = 0$ often fails to hold for $\mathcal{B}_{\mathcal{H}^d}$ (cf. Ex. 1) if p is non-vanishing or explosive at the boundary. When this occurs, SVGD need not converge to p as p is not a stationary point of its dynamics. To fix this, we consider the following *mirrored* dynamics

$$\theta_t = \nabla\psi^*(\eta_t) \quad \text{for} \quad d\eta_t = g_t(\theta_t)dt, \quad \text{or, equivalently,} \quad d\theta_t = \nabla^2\psi(\theta_t)^{-1} g_t(\theta_t)dt, \tag{5}$$

where $g_t : \Theta \to \mathbb{R}^d$ now represents the update direction in η space. The inverse mirror map $\nabla\psi^*$ guarantees that θ_t belongs to the constrained domain Θ. Since ψ is strongly convex and $\nabla^2\psi^{-1}$ is bounded Lipschitz, from Thm. 1 it follows for any bounded Lipschitz g_t that

$$\tfrac{d}{dt}\mathrm{KL}(q_t \| p) = -\mathbb{E}_{q_t}[(\mathcal{M}_{p,\psi} g_t)(\theta)], \tag{6}$$

where we introduce the *mirrored Stein operator* $\mathcal{M}_{p,\psi}$[2]:

$$(\mathcal{M}_{p,\psi} g)(\theta) = g(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot (\nabla^2\psi(\theta)^{-1} g(\theta)), \tag{7}$$

Here ψ is as in Sec. 1 with $(\nabla^2\psi)^{-1}$ differentiable and Lipschitz on Θ. The following result, proved in App. L.1, shows that $\mathcal{M}_{p,\psi}$ generates mean-zero functions under p whenever $\nabla^2\psi^{-1}$ suitably cancels the growth of p at the boundary.

Proposition 2 *Suppose that $\nabla^2\psi^{-1}\nabla \log p$ and $\nabla\cdot\nabla^2\psi^{-1}$ are p-integrable. If $\lim_{r\to\infty} \int_{\partial\Theta_r} p(\theta)$ $\|\nabla^2\psi(\theta)^{-1} n_r(\theta)\|_2 d\theta = 0$ for $\Theta_r \triangleq \{\theta \in \Theta : \|\theta\|_\infty \le r\}$ and $n_r(\theta)$ the outward unit normal vectorto $\partial\Theta_r$ at θ, then $\mathbb{E}_p[(\mathcal{M}_{p,\psi} g)(\theta)] = 0$ if $g \in C^1$ is bounded Lipschitz.*

Example 1 (Dirichlet p, Negative entropy ψ) *When $\theta_{1:d+1} \sim \mathrm{Dir}(\alpha)$ for $\alpha \in \mathbb{R}_+^{d+1}$, even setting $g(\theta) = \mathbf{1}$ in (4) need not cause $\mathbb{E}_p[(\mathcal{S}_p g)(\theta)] = 0$ when $\exists j, \alpha_j \le 1$. However, we show in App. C that the conditions of Prop. 2 are met for any α if $\psi(\theta) = \sum_{j=1}^{d+1} \theta_j \log \theta_j$.*

We propose new deterministic sampling algorithms by seeking optimal directions g_t that minimizes (6) over different function classes. Thm. 3 forms the basis of our analysis.

Theorem 3 (Optimal mirror updates in RKHS, proved in App. L.2) *Let $(\theta_t)_{t\ge 0}$ follow the mirrored dynamics (5). Let $\mathcal{H}_K$ denote the RKHS of a matrix-valued kernel $K : \Theta \times \Theta \to \mathbb{S}^{d\times d}$ (Micchelli and Pontil, 2005). Then, the optimal direction of g_t that minimizes (6) in the norm ball $\mathcal{B}_{\mathcal{H}_K} \triangleq \{g : \|g\|_{\mathcal{H}_K} \le 1\}$ is*

$$g_t^* \propto g_{q_t,K}^* \triangleq \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi} K(\cdot, \theta)], \tag{8}$$

2.2. Mirrored Stein Variational Gradient Descent

We can simply choose the K to be $K_k(\theta, \theta') = k(\theta, \theta')I$. In this case, the update $g_{q_t,K_k}^*(\cdot) = \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi} K_k(\cdot, \theta)]$ is equivalent to running SVGD in the η space before mapping back to Θ.

Theorem 4 (Mirrored SVGD, proof is in App. L.3) *In the setting of Thm. 3, let $k_\psi(\eta, \eta') = k(\nabla\psi^*(\eta), \nabla\psi^*(\eta'))$, $p_H(\eta) = p(\nabla\psi^*(\eta)) \cdot |\det \nabla^2\psi^*(\eta)|$ denote the density of $\eta = \nabla\psi(\theta)$ when $\theta \sim p$, and $q_{t,H}$ denote the distribution of η_t under the mirrored dynamics (5). If $K_k = kI$,*

$$g_{q_t,K_k}^*(\theta') = \mathbb{E}_{\eta_t\sim q_{t,H}}[k_\psi(\eta_t, \eta')\nabla \log p_H(\eta_t) + \nabla_{\eta_t} k_\psi(\eta_t, \eta')] \quad \forall \theta' \in \Theta, \eta' = \nabla\psi(\theta'). \tag{9}$$

[2] We derive $\mathcal{M}_{p,\psi}$ from the (infinitesimal) generator of the mirror-Langevin diffusion. See App. D.

The proof is in App. L.3. By discretizing the dynamics $d\eta_t = g^*_{q_t,K_k}(\theta_t)dt$ and initializing with any particle approximation $q_0 = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_0^i}$, we obtain *Mirrored SVGD (MSVGD)*, our first algorithm for sampling in constrained domains. The details are summarized in Alg. 1.

When a single particle is used and the kernel satisfies $\nabla k(\theta, \theta) = 0$, the MSVGD update (9) reduces to gradient descent on $-\log p_H(\eta)$. This however is not MD: although MD operates in η space, it uses the gradient of $-\log p(\theta)$ instead of $-\log p_H(\eta)$. Also note the modes of $p_H(\eta)$ need not match those of $p(\theta)$ (cf. examples in App. F). Since we are primarily interested in the θ-space density, it is natural to ask whether there exists a mirrored dynamics that reduces to finding the mode of $p(\theta)$ when $n = 1$. In the next section, we give an answer to this question by designing an adaptive kernel that yields a MD-like update.

2.3. Stein Variational Mirror Descent

Our second sampling algorithm for constrained problems is called *Stein Variational Mirror Descent (SVMD)*. We start by introducing a new adaptive matrix-valued kernel.

Definition 5 (Kernels for SVMD) *Given a scalar-valued kernel k, consider the Mercer representation*[3] $k(\theta, \theta') = \sum_{i\geq 1} \lambda_{t,i} u_{t,i}(\theta) u_{t,i}(\theta')$ *w.r.t. q_t, where $u_{t,i}$ is an eigenfunction:*

$$\mathbb{E}_{\theta_t \sim q_t}[k(\theta, \theta_t) u_{t,i}(\theta_t)] = \lambda_{t,i} u_{t,i}(\theta). \tag{10}$$

For $k_t^{1/2}(\theta, \theta') \triangleq \sum_{i\geq 1} \lambda_{t,i}^{1/2} u_{t,i}(\theta) u_{t,i}(\theta')$, we define the adaptive SVMD kernel at time t,

$$K_{\psi,t}(\theta, \theta') \triangleq \mathbb{E}_{\theta_t \sim q_t}[k_t^{1/2}(\theta, \theta_t) \nabla^2 \psi(\theta_t) k_t^{1/2}(\theta_t, \theta')]. \tag{11}$$

By Thm. 3, the optimal update direction for the SVMD kernel ball is $g^*_{q_t, K_{\psi,t}} = \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi} K_{\psi,t}(\cdot, \theta)]$. We obtain the SVMD algorithm (summarized in Alg. 1) by discretizing $d\eta_t = g^*_{q_t, K_{\psi,t}}(\theta_t)dt$ and initializing with $q_0 = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_0^i}$. Because of the discrete representation of q_t, $K_{\psi,t}$ takes the form $K_{\psi,t}(\theta, \theta') = \sum_{i=1}^n \sum_{j=1}^n \lambda_{t,i}^{1/2} \lambda_{t,j}^{1/2} u_{t,i}(\theta) u_{t,j}(\theta') \Gamma_{t,ij}$, for $\Gamma_{t,ij} = \frac{1}{n}\sum_{\ell=1}^n u_{t,i}(\theta_t^\ell) u_{t,j}(\theta_t^\ell) \nabla^2 \psi(\theta_t^\ell)$. $\lambda_{t,j}$, $u_{t,j}(\theta_t^i)$ and its gradients can be computed by solving a matrix eigenvalue problem involving the particle set $\{\theta_t^i\}_{i=1}^n$. We give the details in App. H.

Notably, SVMD differs from MSVGD only in its choice of kernel, but, whenever $\nabla k(\theta, \theta) = 0$, this change is sufficient to exactly recover MD when $n = 1$.

Proposition 6 (Single-particle SVMD is MD) *If $n = 1$, then one step of SVMD becomes $\eta_{t+1} = \eta_t + \epsilon_t(k(\theta_t, \theta_t)\nabla \log p(\theta_t) + \nabla k(\theta_t, \theta_t))$, $\theta_{t+1} = \nabla \psi^*(\eta_{t+1})$.*

In App. I, we establish convergence guarantees for our proposed algorithms. By leveraging the connection between MD and natural gradient descent, we further generalize SVMD to an algorithm for unscontrained domains that can exploit the geometry of the problems. We describe this algorithm in App. J.

3. Experiments

3.1. Approximation quality on the simplex

We first measure distributional approximation quality using two 20-dimensional simplex-constrained targets: the sparse Dirchlet posterior of Patterson and Teh (2013) and the

[3] See App. G for background on Mercer representations in non-compact domains.

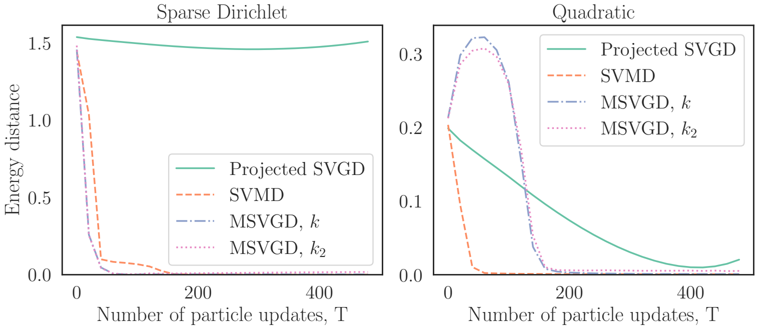


Figure 2: Quality of 50-particle approximations to 20-dimensional distributions on the simplex after T particle updates. (Left) Sparse Dirichlet posterior. (Right) Quadratic simplex target. Details of the target distributions are in App. K.1.

quadratic simplex target of Ahn and Chewi (2020). The Dirichlet mimics the multimodal conditionals that arise in latent Dirichlet allocation (Blei et al., 2003) but induces a log concave density in η space, while the quadratic is log-concave in θ space. To compare with standard SVGD and to prevent its particles from exiting the domain, we introduce a Euclidean projection onto Θ following each SVGD update. In Fig. 2, we compare MSVGD, SVMD, and projected SVGD with 50 particles and inverse multiquadric kernel k (Gorham and Mackey, 2017) by computing the energy distance (Székely and Rizzo, 2013) to a surrogate ground truth sample. We also compare to MSVGD with $k_2(\theta, \theta') = k(\nabla\psi(\theta), \nabla\psi(\theta'))$, a choice which corresponds to running SVGD in the dual space with kernel k by Thm. 4 and which ensures the convergence of MSVGD to p by Thms. 9, 11 and 12.

In the quadratic case, SVMD is favored over MSVGD as it is able to exploit the log-concavity of $p(\theta)$. In contrast, for the multimodal sparse Dirichlet with $p(\theta)$ unbounded near the boundary, MSVGD converges slightly more rapidly than SVMD by exploiting the log concave structure in η space. Projected SVGD fails to converge to the target in both cases and has particular difficulty in approximating the sparse Dirichlet target with unbounded density. MSVGD with k and k_2 perform similarly, but k yields better approximation quality upon convergence. Therefore, we employ k in the remaining MSVGD experiments.

3.2. Confidence intervals for post-selection inference

We next apply our methods to the constrained sampling problems that arise in post-selection inference (Taylor and Tibshirani, 2015). Specifically, we consider the task of forming valid confidence intervals (CIs) for regression parameters selected by the randomized Lasso (Tian and Taylor, 2018) with data $X \in \mathbb{R}^{\tilde{n}\times p}$ and $y \in \mathbb{R}^{\tilde{n}}$ and user-generated randomness $w \in \mathbb{R}^p$ from a log-concave distribution with density g. The Lasso returns $\hat{\beta} \in \mathbb{R}^p$ with non-zero coefficients denoted by $\hat{\beta}_E$ and their signs by s_E. It is common practice to report least squares CIs for β_E by running a linear regression on the selected features E. However, since E is chosen based on the same data, the resulting CIs are often invalid.

Post-selection inference solves this problem by conditioning the inference on the knowledge of E and s_E. To construct valid CIs, it suffices to approximate the *selective distribution*

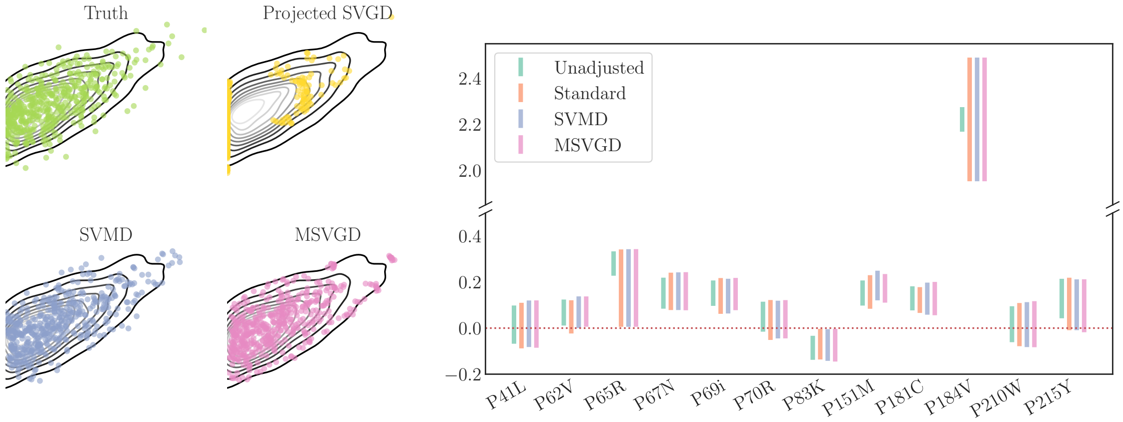


Figure 3: (*Left*) Sampling from a 2D selective density; (*Right*) Unadjusted and post-selection CIs for the mutations selected by the randomized Lasso as candidates for HIV-1 drug resistance (see Sec. 3.2).

with support $\{\hat\beta_E, u_{-E} :\ s_E \odot \hat\beta_E > 0,\ u_{-E} \in [-1,1]^{p-|E|}\}$ and density

$$\hat g(\hat\beta_E, u_{-E}) \propto g\big(X^\top y - \big(\begin{smallmatrix} X_E^\top X_E + \epsilon I_{|E|} \\ X_{-E}^\top X_E \end{smallmatrix}\big)\hat\beta_E + \lambda\big(\begin{smallmatrix} s_E \\ u_{-E} \end{smallmatrix}\big)\big).$$

In our experiments, we integrate out u_{-E} analytically, following Tian and Taylor (2018), and reparameterize $\hat\beta_E$ as $s_E \odot |\hat\beta_E|$ to obtain a log-concave density of $|\hat\beta_E|$ supported on the nonnegative orthant. We choose $\psi(\theta) = \sum_{j=1}^d (\theta_j \log\theta_j - \theta_j)$. In Fig. 3 (left) we show the example of a 2D selective distribution. We also plot the results by projected SVGD, SVMD, and MSVGD in this example. Projected SVGD fails to approximate the target, while the samples by MSVGD and SVMD closely resemble the truth.

We then compare our methods with the standard `norejection` MCMC approach of the `selectiveInference` R package (Tibshirani et al., 2019) using the example simulation setting described in Sepehri and Markovic (2017). To generate N total sample points we run MCMC for N iterations after burn-in or aggregate the particles from N/n independent runs of MSVGD or SVMD with $n = 50$ particles. As N ranges from 1000 to 3000 in Fig. 4(*a*), the MSVGD and SVMD CIs consistently yield higher coverage than the standard 90% CIs. This increased coverage is of particular value for smaller sample sizes, for which the standard CIs tend to undercover. For a much larger sample size of $N = 5000$ in Fig. 4(*b*), the SVMD and standard CIs closely track one another across confidence levels, while MSVGD overcovers.

We next apply our samplers to a post-selection inference task on the HIV-1 drug resistance dataset (Rhee et al., 2006), where we run randomized Lasso (Tian and Taylor, 2018) to find statistically significant mutations associated with drug resistance using susceptibility data on virus isolates. In Fig. 3 (right) we plot the CIs of selected mutations obtained with $N = 5000$ sample points. We see that the invalid unadjusted least squares CIs can lead to premature conclusions, e.g., declaring mutation 215Y significant when there is insufficient support after conditioning on the selection event. In contrast, mutation 184V, which has known association with drug resistance, is declared significant by all methods even after post-selection adjustment. The MSVGD and SVMD CIs mostly track those of the standard `selectiveInference` method, but their conclusions sometimes differ: e.g., 62Y is flagged as significant by MSVGD and SVMD but not by `selectiveInference`.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *arXiv preprint arXiv:2010.16212*, 2020.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

Andrew D Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, pages 175–184, 1988.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Rabi N Bhattacharya and Edward C Waymire. *Stochastic processes with applications*. SIAM, 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Peng Chen, Keyi Wu, Joshua Chen, Tom O'Leary-Roseberry, and Omar Ghattas. Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-Langevin diffusions. *arXiv preprint arXiv:2005.09669*, 2020.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689, 2017.

Gianluca Detommaso, Tiangang Cui, Youssef Marzouk, Alessio Spantini, and Robert Scheichl. A Stein variational Newton method. In *Advances in Neural Information Processing Systems*, volume 31, pages 9187–9197, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3):851–886, 2016.

JC Ferreira and VA Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301, 2017.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic stein discrepancies. *arXiv preprint arXiv:2007.02857*, 2020.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a: overview of mini-batch gradient descent. 2012.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript*, 2(1), 2009.

Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pages 31–35. IEEE, 2018.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1788–1794, 2016.

Chang Liu and Jun Zhu. Riemannian Stein variational gradient descent for Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3627–3634, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, pages 3115–3123, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29:2378–2386, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Arkadij Semenovic Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2003.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L Brutlag, and Robert W Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

Amir Sepehri and Jelena Markovic. Non-reversible, tuning-and rejection-free Markov chain Monte Carlo via iterated random functions. *arXiv preprint arXiv:1711.07177*, 2017.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, pages 4644–4653, 2018.

Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-Newton Langevin Monte Carlo. In *International Conference on Machine Learning*, pages 642–651, 2016.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Ryan Tibshirani, Rob Tibshirani, Jonatha Taylor, Joshua Loftus, Stephen Reid, and Jelena Markovic. *selectiveInference: Tools for Post-Selection Inference*, 2019. URL https://CRAN.R-project.org/package=selectiveInference. R package version 1.2.5.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. In *Advances in Neural Information Processing Systems*, pages 7836–7846, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. *arXiv preprint arXiv:2002.04363*, 2020.

Algorithm 1: Mirrored SVGD & Stein Variational Mirror Descent

```
Input: density $p$ on $\Theta$, kernel $k$, mirror function $\psi$, particles $(\theta_0^i)_{i=1}^n \subset \Theta$, step sizes $(\epsilon_t)_{t=1}^T$;
Init: $\eta_0^i \leftarrow \nabla\psi(\theta_0^i)$ for $i \in [n]$;
for $t = 0 : T$ do
    if SVMD then $K_t \leftarrow K_{\psi,t}$ (11) else $K_t \leftarrow kI$ (MSVGD);
    for $i \in [n], \eta_{t+1}^i \leftarrow \eta_t^i + \epsilon_t \frac{1}{n}\sum_{j=1}^n \mathcal{M}_{p,\psi} K_t(\theta_t^i, \theta_t^j)$    (for $\mathcal{M}_{p,\psi} K_t(\cdot, \theta)$ defined in Thm. 3);
    for $i \in [n], \theta_{t+1}^i \leftarrow \nabla\psi^*(\eta_{t+1}^i)$;
end
return $\{\theta_{T+1}^i\}_{i=1}^n$.
```

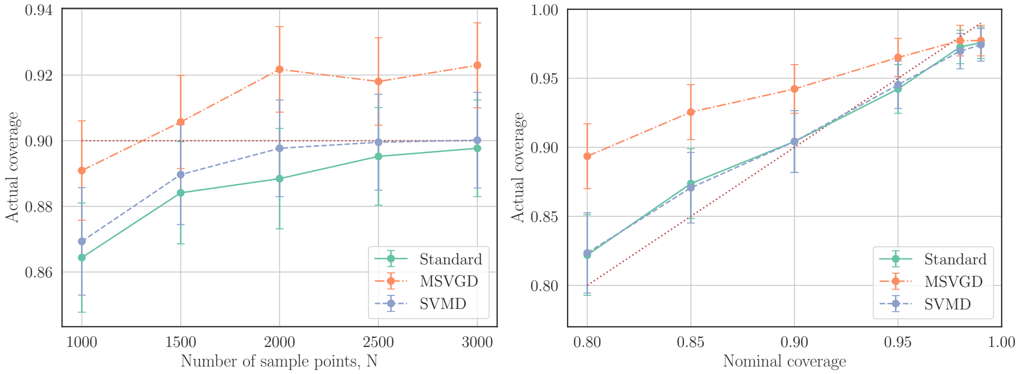
Algorithm 2: Stein Variational Natural Gradient (SVNG)

```
Input: density $p(\theta)$ on $\mathbb{R}^d$, kernel $k$, metric tensor $G(\theta)$, particles $(\theta_0^i)_{i=1}^n$, step sizes $(\epsilon_t)_{t=1}^T$;
for $t = 0 : T$ do
    for $i \in [n], \theta_{t+1}^i \leftarrow \theta_t^i + \epsilon_t G(\theta_t^i)^{-1} g_{G,t}^*(\theta_t^i)$, where
      $g_{G,t}^*(\theta) = \frac{1}{n}\sum_{j=1}^n [K_{G,t}(\theta, \theta_t^j) G(\theta_t^j)^{-1} \nabla \log p(\theta_t^j) + \nabla_{\theta_t^j} \cdot (K_{G,t}(\theta, \theta_t^j) G(\theta_t^j)^{-1})]$  (see (18));
end
return $\{\theta_{T+1}^i\}_{i=1}^n$.
```

Appendix A. Related Work

Our mirrored Stein operators (7) are instances of diffusion Stein operators in the sense of Gorham and Mackey (2017), but their specific properties have not been studied, nor have they been used to develop sampling algorithms. There is now a large body of work on transferring algorithmic ideas from optimization to sampling (see, e.g., Dalalyan, 2017; Welling and Teh, 2011; Durmus et al., 2018; Ma et al., 2019; Simsekli et al., 2016). The closest to our work in this space is the recent marriage of mirror descent and MCMC. For example, Hsieh et al. (2018) propose to run Langevin Monte Carlo (LMC, an Euler discretization of the Langevin diffusion) in a mirror space. Zhang et al. (2020) analyze the convergence properties of the mirror-Langevin diffusion, Chewi et al. (2020) demonstrate its advantages over the Langevin diffusion when using a Newton-type metric, and Ahn and Chewi (2020) study its discretization for MCMC sampling in constrained domains. Relatedly, Patterson and Teh (2013) proposed stochastic Riemannian LMC for sampling on the simplex.

Several modifications of SVGD have been proposed to incorporate geometric information. Riemannian SVGD (RSVGD, Liu and Zhu, 2018) generalizes SVGD to Riemannian manifolds, but, even with the same metric tensor, their updates are more complex than ours: notably they require higher-order kernel derivatives, do not operate in a mirror space, and do not reduce to natural gradient descent when a single particle is used. They also reportedly do not perform well when with scalable stochastic estimates of $\nabla \log p$. Stein Variational Newton (SVN, Detommaso et al., 2018; Chen et al., 2019) introduces second-order information into SVGD. Their algorithm requires an often expensive Hessian computation and need not lead to descent directions, so inexact approximations are employed in practice. Our SVNG can be seen as an instance of matrix SVGD (MatSVGD, Wang et al., 2019) with an adaptive time-dependent kernel discussed in App. J, a choice that is not explored in Wang et al. (2019) and which recovers natural gradient descent when $n = 1$ unlike the heuristic kernel



(a) Nominal coverage: 0.9 (b) $N = 5000$ sample points

Figure 4: Coverage of post-selection CIs across (a) 500 / (b) 200 replications of simulation of Sepehri and Markovic (2017).

constructions of Wang et al. (2019). None of the aforementioned works provide convergence guarantees, and neither SVN nor matrix SVGD deals with constrained domains.

Appendix B. Mirror Descent, Riemannian Gradient Flow, and Natural Gradient

The equivalence between the mirror flow $d\eta_t = -\nabla f(\theta_t)dt$, $\theta_t = \nabla\psi^*(\eta_t)dt$ and the Riemannian gradient flow in (2) is a direct result of the chain rule:

$$\frac{d\theta_t}{dt} = -\nabla_{\eta_t}\theta_t \frac{d\eta_t}{dt} = -(\nabla_{\theta_t}\eta_t)^{-1}\frac{d\eta_t}{dt} = -\nabla^2\psi(\theta_t)^{-1}\nabla f(\theta_t), \tag{12}$$

$$\frac{d\eta_t}{dt} = -\nabla f(\theta_t) = -\nabla_{\theta_t}\eta_t\nabla_{\eta_t} f(\nabla\psi^*(\eta_t)) = -\nabla^2\psi^*(\eta_t)^{-1}\nabla_{\eta_t} f(\nabla\psi^*(\eta_t)). \tag{13}$$

Depending on discretizing (12) or (13), there are two natural gradient descent (NGD) updates that can arise from the same gradient flow:

$$\text{NGD (a):} \quad \theta_{t+1} = \theta_t - \epsilon_t\nabla^2\psi(\theta_t)^{-1}\nabla f(\theta_t),$$
$$\text{NGD (b):} \quad \eta_{t+1} = \eta_t - \epsilon_t\nabla^2\psi^*(\eta_t)^{-1}\nabla_{\eta_t} f(\nabla\psi^*(\eta_t)).$$

With finite step sizes ϵ_t, their updates need not be the same and can lead to different optimization paths. Since $\nabla f(\theta_t) = \nabla^2\psi^*(\eta_t)^{-1}\nabla_{\eta_t} f(\nabla\psi^*(\eta_t))$, NGD (b) is equivalent to the dual-space update by mirror descent. This relationship was pointed out in Raskutti and Mukherjee (2015) and has been used for developing natural gradient variational inference algorithms (Khan and Nielsen, 2018). We emphasize, however, our SVNG algorithm developed in App. J corresponds to the discretization in the primal space as in NGD (a). Therefore, it does not require an explicit dual space, and allows replacing $\nabla^2\psi$ with more general information metric tensors.

Appendix C. Details of Example 1

For the entropic mirror map $\psi(\theta) = \sum_{j=1}^{d+1}\theta_j \log\theta_j$, we have $\nabla^2\psi(\theta)^{-1} = \mathrm{diag}(\theta) - \theta\theta^\top$. Note here θ denotes a d-dimensional vector and does not include $\theta_{d+1} = 1 - \sum_{j=1}^{d}\theta_d$. Since

Θ is a $(d+1)$-simplex, $\partial\Theta$ is composed of $d+1$ faces with θ in the j-th face satisfies $\theta_j = 0$. The outward unit normal vector $n(\theta)$ for the first d faces are $-e_j$ for $1 \le j \le d$, where e_j denotes the j-th standard basis of $\mathbb{R}^d$. The outward unit normal vector for the $(d+1)$-st face is a vector with $1/\sqrt{d}$ in all coordinates. Therefore, we have

$$\begin{aligned}\int_{\partial\Theta} p(\theta)g(\theta)^\top \nabla^2\psi(\theta)^{-1}n(\theta)d\theta =& \int_{\partial\Theta} p(\theta)g(\theta)^\top(\operatorname{diag}(\theta) - \theta\theta^\top)n(\theta)d\theta \\ =& \int_{\partial\Theta} p(\theta)(\theta \odot g(\theta) - \theta\theta^\top g(\theta))^\top n(\theta)d\theta \\ =& \sum_{j=1}^{d} \int_{\theta_j=0} p(\theta)(\theta^\top g(\theta) - g_j(\theta))\theta_j d\theta_{-j} \\ &+ \frac{1}{\sqrt{d}} \int_{\theta_{d+1}=0} p(\theta)\theta^\top g(\theta)\theta_{d+1} d\theta \\ =& 0,\end{aligned}$$

where in the second to last identity we used $\theta^\top \mathbf{1} = 1 - \theta_{d+1}$. Finally, we can verify the condition in Prop. 2 as

$$\lim_{r\to\infty} \int_{\partial\Theta_r} p(\theta)\|\nabla^2\psi(\theta)^{-1}n_r(\theta)\|_2 d\theta = \sup_{\|g\|_\infty \le 1} \int_{\partial\Theta} p(\theta)g(\theta)^\top\nabla^2\psi(\theta)^{-1}n(\theta)d\theta = 0.$$

Appendix D. Derivation of the Mirrored Stein Operator

We first review the (overdamped) Langevin diffusion – a Markov process that underlies many recent advances in Stein's method – along with its recent mirrored generalization. The Langevin diffusion with equilibrium density p on $\mathbb{R}^d$ is a Markov process $(\theta_t)_{t\ge0} \subset \mathbb{R}^d$ satisfying the stochastic differential equation (SDE)

$$d\theta_t = \nabla \log p(\theta_t)dt + \sqrt{2}dB_t \tag{14}$$

with $(B_t)_{t\ge0}$ a standard Brownian motion (Bhattacharya and Waymire, 2009, Sec. 4.5).

To identify Stein operators that generate mean-zero functions under p for broad classes of targets p, Gorham and Mackey (2015) proposed to build upon the generator method of Barbour (1988): First, identify a Markov process $(\theta_t)_{t\ge0}$ that has p as the equilibrium density; they chose the Langevin diffusion of (14). Next, build a Stein operator based on the (infinitesimal) generator A of the process (Øksendal, 2003, Def. 7.3.1):

$$(Af)(\theta) = \lim_{t\to0} \tfrac{1}{t}(\mathbb{E}f(\theta_t) - \mathbb{E}f(\theta_0)) \quad \text{for } f: \mathbb{R}^d \to \mathbb{R},$$

as the generator satisfies $\mathbb{E}_p[(Af)(\theta)] = 0$ under relatively mild conditions. We use the following theorem to derive the generator of the processes described by SDEs like (14):

Theorem 7 (Generator of Itô diffusion; Øksendal, 2003, Thm 7.3.3) *Let $(x_t)_{t\ge0}$ be the Itô diffusion in $\mathcal{X} \subseteq \mathbb{R}^d$ satisfying $dx_t = b(x_t)dt + \sigma(x_t)dB_t$. For any $f \in C_c^2(\mathcal{X})$, the (infinitesimal) generator A of $(x_t)_{t\ge0}$ is*

$$(Af)(x) = b(x)^\top\nabla f(x) + \tfrac{1}{2}\operatorname{Tr}(\sigma(x)\sigma(x)^\top\nabla^2 f(x)).$$

Substituting $\nabla \log p(\cdot)$ for $b(\cdot)$ and $\sqrt{2}I$ for $\sigma(\cdot)$ in Thm. 7, we obtain $Af = (\nabla \log p)^\top \nabla f + \nabla \cdot \nabla f$. Replacing ∇f with a vector-valued function g gives the Langevin Stein operator in (4).

To derive a Stein operator that works well for constrained domains, we consider the Riemannian Langevin diffusion (Patterson and Teh, 2013; Xifara et al., 2014; Ma et al., 2015) that extends the Langevin to non-Euclidean geometries encoded in a positive definite *metric tensor* $G(\theta)$:

$$d\theta_t = (G(\theta_t)^{-1}\nabla \log p(\theta_t) + \nabla \cdot G(\theta_t)^{-1})dt + \sqrt{2}G(\theta_t)^{-1/2}dB_t.^{4}$$

We show in App. E that the choice $G = \nabla^2\psi$ yields the recent mirror-Langevin diffusion (Zhang et al., 2020; Chewi et al., 2020)

$$\theta_t = \nabla\psi^*(\eta_t), \quad d\eta_t = \nabla \log p(\theta_t)dt + \sqrt{2}\nabla^2\psi(\theta_t)^{1/2}dB_t. \tag{15}$$

According to Thm. 7, the generator of the mirror-Langevin diffusion described by (16) is

$$\begin{aligned}(A_{p,\psi}f)(\theta) &= (\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot \nabla^2\psi(\theta)^{-1})^\top \nabla f(\theta) + \mathrm{Tr}(\nabla^2\psi(\theta)^{-1}\nabla^2 f(\theta)) \\ &= \nabla f(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot (\nabla^2\psi(\theta)^{-1}\nabla f(\theta)).\end{aligned}$$

Now substituting $g(\theta)$ for $\nabla f(\theta)$, we obtain the associated mirrored Stein operator:

$$(\mathcal{M}_{p,\psi}g)(\theta) = g(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot (\nabla^2\psi(\theta)^{-1}g(\theta)).$$

Appendix E. Riemannian Langevin Diffusions and Mirror-Langevin Diffusions

Zhang et al. (2020) pointed out (15) is a particular case of the Riemannian LD. However, they did not give an explicit derivation. The Riemannian LD (Patterson and Teh, 2013; Xifara et al., 2014; Ma et al., 2015) with $\nabla^2\psi(\cdot)$ as the metric tensor is

$$d\theta_t = (\nabla^2\psi(\theta_t)^{-1}\nabla \log p(\theta_t) + \nabla \cdot \nabla^2\psi(\theta_t)^{-1})dt + \sqrt{2}\nabla^2\psi(\theta_t)^{-1/2}dB_t. \tag{16}$$

To see the connection with mirror-Langevin diffusion, we would like to obtain the SDE that describes the evolution of $\eta_t = \nabla\psi(\theta_t)$ under the diffusion. This requires the following theorem that provides the analog of the "chain rule" in SDEs.

Theorem 8 (Itô formula; Øksendal, 2003, Thm 4.2.1) *Let $(x_t)_{t\geq 0}$ be an Itô process in $\mathcal{X} \subset \mathbb{R}^d$ satisfying $dx_t = b(x_t)dt + \sigma(x_t)dB_t$. Let $f(x) \in C^2 : \mathbb{R}^d \to \mathbb{R}^{d'}$. Then $y_t = f(x_t)$ is again an Itô process, and its i-th dimension satisfies*

$$dy_{t,i} = (\nabla f_i(x_t)^\top b(x_t) + \frac{1}{2}\mathrm{Tr}(\nabla^2 f_i(x_t)\sigma(x_t)\sigma(x_t)^\top)dt + \nabla f_i(x_t)^\top \sigma(x_t)dB_t.$$

Substituting $\nabla\psi$ for f in Thm. 8, we have the SDE of $\eta_t = \nabla\psi(\theta_t)$ as

$$d\eta_t = (\nabla \log p(\theta_t) + \nabla^2\psi(\theta_t)\nabla \cdot \nabla^2\psi(\theta_t)^{-1} + h(\theta_t))dt + \sqrt{2}\nabla^2\psi(\theta_t)^{1/2}dB_t,$$

[4] A matrix divergence $\nabla \cdot G(\theta)$ is the vector obtained by computing the divergence of each row of $G(\theta)$.

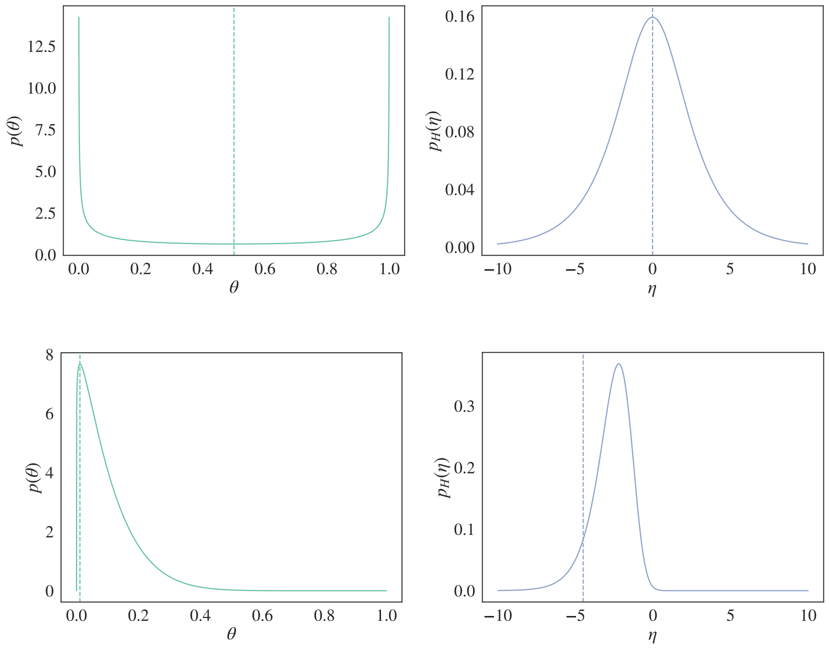


Figure 5: The density functions of the same distribution in θ (left) and η (right) space under the transformation $\eta = \nabla\psi(\theta)$. *Top*: $\theta \sim \mathrm{Beta}(0.5, 0.5)$. Dashed lines mark the mode of the transformed density $p_H(\eta)$ and the corresponding θ, which gives the lowest value of $p(\theta)$; *Bottom*: $\theta \sim \mathrm{Beta}(1.1, 10)$. Dashed lines mark the mode of the target density $p(\theta)$ and the corresponding η, which clearly does not match the mode of $p_H(\eta)$.

where $h(\theta_t)_i = \mathrm{Tr}(\nabla^2_{\theta_t}(\nabla_{\theta_{t,i}}\psi(\theta_t))\nabla^2\psi(\theta_t)^{-1})$. Moreover, we have

$$
\begin{aligned}
&[\nabla^2\psi(\theta_t)\nabla\cdot\nabla^2\psi(\theta_t)^{-1}]_i + \mathrm{Tr}(\nabla^2_{\theta_t}(\nabla_{\theta_{t,i}}\psi(\theta_t))\nabla^2\psi(\theta_t)^{-1}) \\
&= \sum_{\ell=1}^d \sum_{j=1}^d \nabla^2\psi(\theta_t)_{ij}\nabla_{\theta_{t,\ell}}[\nabla^2\psi(\theta_t)^{-1}]_{j\ell} + \sum_{\ell=1}^d \sum_{j=1}^d \nabla_{\theta_{t,\ell}}\nabla^2\psi(\theta_t)_{ij}[\nabla^2\psi(\theta_t)^{-1}]_{j\ell} \\
&= \sum_{\ell=1}^d \nabla_{\theta_{t,\ell}}\left(\sum_{j=1}^d \nabla^2\psi(\theta_t)_{ij}[\nabla^2\psi(\theta_t)^{-1}]_{j\ell}\right) = \sum_{\ell=1}^d \nabla_{\theta_{t,\ell}} I_{i\ell} = 0.
\end{aligned}
$$

Therefore, the η_t diffusion is described by the SDE:

$$
d\eta_t = \nabla\log p(\theta_t)dt + \sqrt{2}\nabla^2\psi(\theta_t)^{1/2}dB_t, \quad \theta_t = \nabla\psi^*(\eta_t).
$$

Appendix F. Mode Mismatch Under Transformations

In Fig. 5, we present two examples where the modes of density functions in θ and η space do not match. We assume θs follow Beta distributions on $[0, 1]$. And we choose the negative entropy $\psi(\theta) = \theta \log \theta + (1 - \theta) \log(1 - \theta)$. Then, the transformation is the logit function $\eta = \log(\theta/(1 - \theta))$ and its reverse is the sigmoid function $\theta = 1/(1 + e^{-\eta})$.

Appendix G. Background on Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{X}$ and taking their values in $\mathbb{R}$. We say k is a reproducing kernel (or kernel) of $\mathcal{H}$ if $\forall x \in \mathcal{X}, k(x, \cdot) \in \mathcal{H}$ and $\forall f \in \mathcal{H}, \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$. $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if it has a kernel. Kernels are positive definite (p.d.) functions, which means that matrices with the form $(k(x_i, x_j))_{ij}$ are positive semidefinite. For any p.d. function k, there is a unique RKHS with k as the reproducing kernel, which can be constructed by the completion of $\{\sum_{i=1}^n a_i k(x_i, \cdot), x_i \in \mathcal{X}, a_i \in \mathbb{R}, i \in \mathbb{N}\}$.

Now we assume $\mathcal{X}$ is a metric space, k is a bounded continuous kernel with the RKHS $\mathcal{H}$, and ν is a positive measure on $\mathcal{X}$. $L^2(\nu)$ denote the space of all square-integrable functions w.r.t. ν. Then the kernel integral operator $T_k : L^2(\nu) \to L^2(\nu)$ defined by

$$T_k g = \int_{\mathcal{X}} g(x) k(x, \cdot) d\nu$$

is compact and self-adjoint. Therefore, according to the spectral theorem, there exists an at most countable set of positive eigenvalues $\{\lambda_j\}_{j\in J} \subset \mathbb{R}$ with $\lambda_1 \geq \lambda_2 \geq \dots$ converging to zero and orthonormal eigenfunctions $\{u_j\}_{j\in J}$ such that

$$T_k u_j = \lambda_j u_j,$$

and k has the representation $k(x, x') = \sum_{j\in J} \lambda_j u_j(x) u_j(x')$ (Mercer's theorem on non-compact domains), where the convergence of the sum is absolute and uniform on compact subsets of $\mathcal{X} \times \mathcal{X}$ (Ferreira and Menegatto, 2009).

Appendix H. Computational Details of Stein Variational Mirror Descent

The matrix eigenvalue problem is defined as

$$B_t v_{t,j} = n\lambda_{t,j} v_{t,j},$$

where $B_t = (k(\theta_t^i, \theta_t^j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is the Gram matrix of pairwise kernel evaluations at particle locations, and the i-th element of $v_{t,j}$ is $u_{t,j}(\theta_t^i)$. To compute $\nabla_\theta K_{\psi,t}(\theta, \theta')$, we differentiate both sides of (10) to find that

$$\nabla u_{t,j}(\theta) = \frac{1}{\lambda_{t,j}} \sum_{i=1}^n v_{t,j,i} \nabla_\theta k(\theta, \theta_t^i).$$

This technique was used in Shi et al. (2018) to estimate gradients of eigenfunctions w.r.t. a continuous q. Following their recommendations, we truncate the sum at the J-th largest eigenvalues according to a threshold ($\tau \geq \sum_{j=1}^J \lambda_{t,j} / \sum_{j=1}^n \lambda_{t,j}$) to ensure numerical stability.

Appendix I. Convergence Guarantees

In this section, we turn our attention to the convergence properties of our proposed methods. For K_t and ϵ_t as in Alg. 1, let $(q_t^\infty, q_{t,H}^\infty)$ represent the distributions of the mirrored Stein updates (θ_t, η_t) when $\theta_0 \sim q_0^\infty$ and $\eta_{t+1} = \eta_t + \epsilon_t g^*_{q_t,K_t}(\theta_t)$ for $t \geq 0$. Our first result, proved in App. L.5, shows that if the Alg. 1 initialization $q_{0,H}^n = \frac{1}{n}\sum_{i=1}^n \delta_{\eta_0^i}$ converges in Wasserstein distance to a distribution $q_{0,H}^\infty$ as $n \to \infty$, then, on each round $t > 0$, the output of Alg. 1, $q_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_t^i}$, converges to q_t^∞.

Theorem 9 (Convergence of mirrored updates as $n \to \infty$) *Suppose Alg. 1 is initialized with $q_{0,H}^n = \frac{1}{n}\sum_{i=1}^n \delta_{\eta_0^i}$ satisfying $W_1(q_{0,H}^n, q_{0,H}^\infty) \to 0$ for W_1 the L^1 Wasserstein distance. Define the η-induced kernel $K_{\nabla\psi^*,t}(\eta,\eta') \triangleq K_t(\nabla\psi^*(\eta), \nabla\psi^*(\eta'))$. If, for some $c_1, c_2 > 0$,*

$$\|\nabla(K_{\nabla\psi^*,t}(\cdot,\eta)\nabla\log p_H(\eta) + \nabla\cdot K_{\nabla\psi^*,t}(\cdot,\eta))\|_{\mathrm{op}} \leq c_1(1+\|\eta\|_2),$$
$$\|\nabla(K_{\nabla\psi^*,t}(\eta',\cdot)\nabla\log p_H(\cdot) + \nabla\cdot K_{\nabla\psi^*,t}(\eta',\cdot))\|_{\mathrm{op}} \leq c_2(1+\|\eta'\|_2),$$

then, $W_1(q_{t,H}^n, q_{t,H}^\infty) \to 0$ and $q_t^n \Rightarrow q_t^\infty$ for each round t.

Remark 10 *The pre-conditions hold, for example, whenever $\nabla\log p_H$ is Lipschitz, ψ is strongly convex, and $K_t = kI$ for k bounded with bounded derivatives.*

Given a mirrored Stein operator (7), an arbitrary Stein set $\mathcal{G}$, and an arbitrary matrix-valued kernel K we define the *mirrored Stein discrepancy* and *mirrored kernel Stein discrepancy*

$$\mathrm{MSD}(q,p,\mathcal{G}) \triangleq \sup_{g\in\mathcal{G}} \mathbb{E}_q[(\mathcal{M}_{p,\psi}g)(\theta)] \quad \text{and} \quad \mathrm{MKSD}_K(q,p) \triangleq \mathrm{MSD}(q,p,\mathcal{B}_{\mathcal{H}_K}). \tag{17}$$

The former is an example of a diffusion Stein discrepancy (Gorham et al., 2019) and the latter an example of a diffusion kernel Stein discrepancy (Barp et al., 2019). Since the MKSD optimization problem (17) matches that in Thm. 3, we have that $\mathrm{MKSD}_K(q,p) = \|g^*_{q,K}\|_{\mathcal{H}_K}$. Our next result, proved in App. L.6, shows that the infinite-particle mirrored Stein updates reduce the KL divergence to p whenever the step size is sufficiently small and drive MKSD to 0 if, for example, $\epsilon_t = \Omega(\mathrm{MKSD}_{K_t}(q_t^\infty, p)^\alpha)$ for any $\alpha > 0$.

Theorem 11 (Infinite-particle mirrored Stein updates decrease KL and MKSD) *Assume $\kappa_1 \triangleq \sup_\theta \|K_t(\theta,\theta)\|_{\mathrm{op}} < \infty$, and $\kappa_2 \triangleq \sum_{i=1}^d \sup_\theta \|\nabla^2_{i,d+i}K_t(\theta,\theta)\|_{\mathrm{op}} < \infty$, $\nabla\log p_H$ is L-Lipschitz, and ψ is α-strongly convex. If $\epsilon_t < 1/(\|\nabla_{\eta_t} g^*_{q_t^\infty,K_t}(\theta_t) + \nabla_{\eta_t} g^*_{q_t^\infty,K_t}(\theta_t)^\top\|_{\mathrm{op}})$, then*

$$\mathrm{KL}(q_{t+1}^\infty\|p) - \mathrm{KL}(q_t^\infty\|p) \leq -\big(\epsilon_t - \big(\tfrac{L\kappa_1}{2} + \tfrac{2\kappa_2}{\alpha^2}\big)\epsilon_t^2\big)\mathrm{MKSD}_{K_t}(q_t^\infty, p)^2.$$

Our last result, proved in App. L.7, shows that $q_t^\infty \Rightarrow p$ if $\mathrm{MKSD}_{K_k}(q_t^\infty, p) \to 0$. Hence, by Thms. 9 and 11, n-particle MSVGD converges weakly to p if ϵ_t decays at a suitable rate.

Theorem 12 (MKSD_{K_k} determines weak convergence) *Assume p_H is distantly dissipative (Eberle, 2016) with $\nabla\log p_H$ Lipschitz, ψ is strongly convex with continuous $\nabla\psi^*$, and $k(\theta,\theta') = \kappa(\nabla\psi(\theta), \nabla\psi(\theta'))$ for $\kappa(x,y) = (c^2 + \|x-y\|_2^2)^\beta$ with $\beta \in (-1,0)$. Then, $q_t^\infty \Rightarrow p$ if $\mathrm{MKSD}_{K_k}(q_t^\infty, p) \to 0$.*

Remark 13 *The pre-conditions hold, for example, for any Dirichlet target with negative entropy ψ.*

Appendix J. Stein Variational Natural Gradient

The fact that SVMD recovers mirror descent as a special case is not only of relevance in constrained problems. We next exploit the connection between MD and natural gradient descent discussed in Sec. 1 to design a new sampler – *Stein Variational Natural Gradient (SVNG)* – that more efficiently approximates unconstrained targets. The idea is to replace the Hessian $\nabla^2\psi(\cdot)$ in the SVMD dynamics $d\theta_t = \nabla^2\psi(\theta_t)^{-1} g^*_{q_t,K_{\psi,t}}(\theta_t)$ with a general metric tensor $G(\cdot)$. The result is the Riemannian gradient flow

$$d\theta_t = G(\theta_t)^{-1} g^*_{q_t,K_{G,t}}(\theta_t)dt \quad \text{with} \quad K_{G,t}(\theta,\theta') \triangleq \mathbb{E}_{\theta_t\sim q_t}[k^{1/2}(\theta,\theta_t)G(\theta_t)k^{1/2}(\theta_t,\theta')]. \tag{18}$$

Given any initial particle approximation $q_0 = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_0^i}$, we discretize these dynamics to obtain the unconstrained SVNG sampler of Alg. 2.

SVNG can be seen as an instance of MatSVGD (Wang et al., 2019) with a new adaptive time-dependent kernel $G^{-1}(\theta)K_{G,t}(\theta,\theta')G^{-1}(\theta')$. However, similar to Prop. 6 and unlike the heuristic kernels of Wang et al. (2019), SVNG reduces to natural gradient ascent for finding the mode of $p(\theta)$ when $n = 1$.

SVNG is well-suited to Bayesian inference problems where the target is a posterior distribution $p(\theta) \propto \pi(\theta)\pi(y|\theta)$. There, the metric tensor $G(\theta)$ can be set to the Fisher information matrix $\mathbb{E}_{\pi(y|\theta)}[\nabla \log\pi(y|\theta)\nabla\log\pi(y|\theta)^\top]$ of the data likelihood $\pi(y|\theta)$. Ample precedent from natural gradient variational inference (Hoffman et al., 2013; Khan and Nielsen, 2018) and Riemannian MCMC (Patterson and Teh, 2013) suggests that encoding problem geometry in this manner often leads to more rapid convergence. In App. K.3, we demonstrate the advantages of SVNG on unconstrained large-scale posterior inference with the Fisher information metric.

Appendix K. Supplementary Experimental Details and Additional Results

In this section, we report supplementary details and additional results from the experiments of Sec. 3. In Secs. 3.1 and 3.2, we use the inverse multiquadric input kernel $k(\theta,\theta') = (1 + \|\theta-\theta'\|_2^2/\ell^2)^{-1/2}$ due to its convergence control properties (Gorham and Mackey, 2017). In the unconstrained experiments of App. K.3, we use the Gaussian kernel $k(\theta,\theta') = \exp(-\|\theta-\theta'\|_2^2/\ell^2)$ for consistency with past work. The bandwidth ℓ is determined by the median heuristic (Garreau et al., 2017). We select τ from $\{0.98, 0.99\}$ for all SVMD experiments. For unconstrained targets, we report, for each method, results from the best fixed step size $\epsilon \in \{0.01, 0.05, 0.1, 0.5, 1\}$ selected on a separate validation set. For constrained targets, we select step sizes adaptively to accommodate rapid density growth near the boundary; specifically, we use RMSProp (Hinton et al., 2012), an extension of the AdaGrad algorithm (Duchi et al., 2011) used in Liu and Wang (2016), and report performance with the best learning rate. Results were recorded on an Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz and an NVIDIA Tesla P100 PCIe 16GB.

K.1. Approximation quality on the simplex

The sparse Dirichlet posterior of Patterson and Teh (2013) extended to 20 dimensions features a sparse, symmetric Dir(α) prior with $\alpha_k = 0.1$ for $k \in \{1,\ldots,20\}$ and sparse

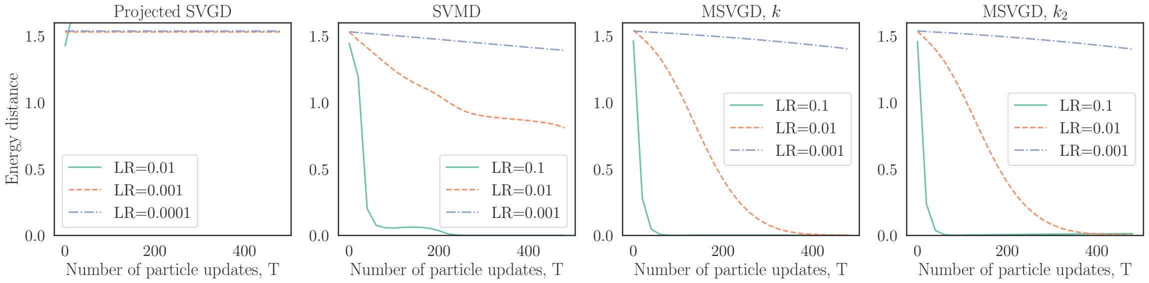


Figure 6: Sampling from a Dirichlet target on a 20-simplex. We plot the energy distance to a ground truth sample of size 1000.

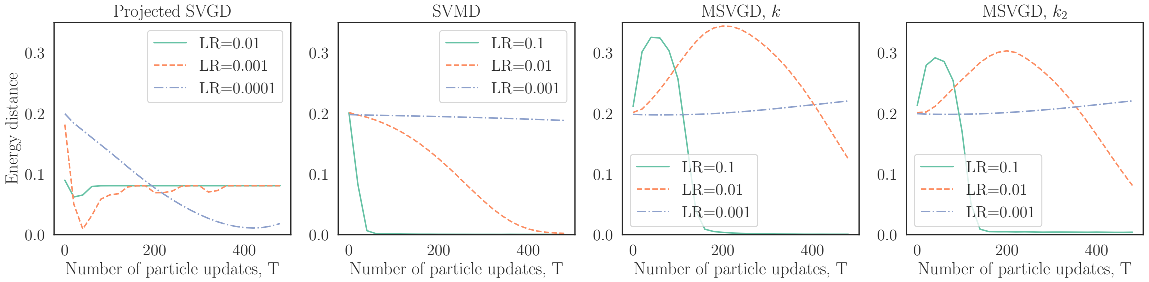


Figure 7: Sampling from a quadratic target on a 20-simplex. We plot the energy distance to a ground truth sample of size 1000 drawn by NUTS (Hoffman and Gelman, 2014).

count data $n_1 = 90,\ n_2 = n_3 = 5,\ n_j = 0\ (j > 3)$, modeled via a multinomial likelihood. The quadratic target satisfies $\log p(\theta) = -\frac{1}{2\sigma^2}\theta^\top A\theta + \text{const}$, where we slightly modify the target density of Ahn and Chewi (2020) to make it less flat by introducing a scale parameter $\sigma = 0.01$. $A \in \mathbb{R}^{19\times 19}$ is a positive definite matrix generated by normalizing products of random matrices with i.i.d. elements drawn from $\text{Unif}[-1, 1]$.

We initialize all methods with i.i.d samples from Dirichlet(5) to prevent any of the initial particles being too close to the boundary. For each method and each learning rate we apply 500 particle updates. For SVMD we set $\tau = 0.98$. We search the base learning rates of RMSProp in $\{0.1, 0.01, 0.001\}$ for SVMD and MSVGD. Since projected SVGD applies updates in the θ space, the appropriate learning rate range is smaller than those of SVMD and MSVGD. There we search the base learning rate of RMSProp in $\{0.01, 0.001, 0.0001\}$. For all methods the results under each base learning rate are plotted in Fig. 6.

K.2. Confidence intervals for post-selection inference

Given a dataset $X \in \mathbb{R}^{\tilde{n}\times p}$, $y \in \mathbb{R}^{\tilde{n}}$, the randomized Lasso (Tian and Taylor, 2018) solves the following problem:

$$\mathrm{argmin}_{\beta\in\mathbb{R}^p} \tfrac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 - w^\top\beta + \tfrac{\epsilon}{2}\|\beta\|_2^2, \quad w \sim \mathbb{G}.$$

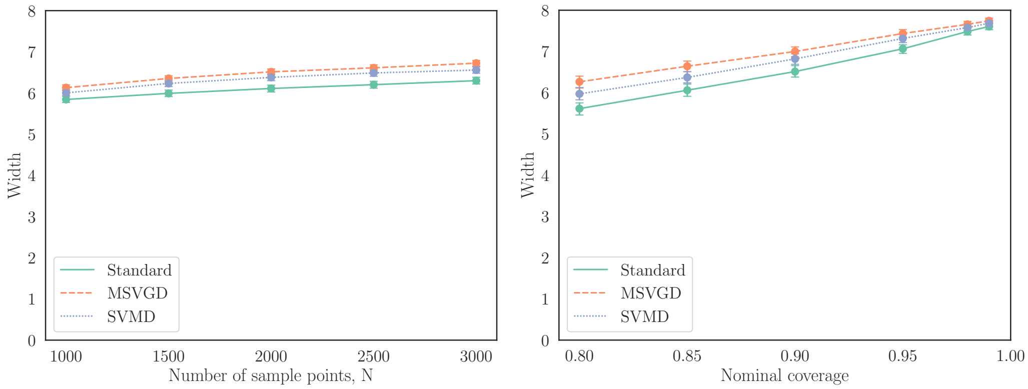
where $\mathbb{G}$ is a user-specified log-concave distribution with density g. We choose $\mathbb{G}$ to be zero-mean independent Gaussian distributions while leaving its scale and the ridge parameter ϵ to be automatically determined by the `randomizedLasso` function of the `selectiveInference` package. We initialize the particles of our SVMD and MSVGD in the following way: First, we map the solution $\hat{\beta}_E$ to the dual space by $\nabla\psi$. Next, we add i.i.d. standard Gaussian noise to n copies of the image in the dual space. Finally, we map the n particles back to the primal space by $\nabla\psi^*$ and use them as the initial locations. Below we discuss the remaining settings and additional results of the simulation and the HIV-1 drug resistance experiment separately.

Simulation In our simulation we mostly follow the settings of Sepehri and Markovic (2017) except using a different penalty level λ recommended in the `selectiveInference` R package. We set $\tilde{n} = 100$ and $p = 40$. The design matrix X is generated from an equi-correlated model, i.e., each datapoint $x_i \in \mathbb{R}^p$ is generated i.i.d. from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{ii} = 1, \Sigma_{ij} = 0.3$ $(i \neq j)$ and then normalized to have almost unit length. The normalization is done by first centering each dimension by subtracting the mean and dividing the standard deviation of that column of X, then additionally multiplying $1/\tilde{n}^{1/2}$. y is generated from a standard Gaussian which is independent of X, i.e., we assume the global null setting where the true value of β is zero. We set λ to be the value returned by `theoretical.lambda` of the `selectiveInference` R package multiplied a coefficient $0.7\tilde{n}$, where the 0.7 adjustment is introduced in the test examples of the R package to reduce the regularization effect so that we have a reasonably large set of selected features when $p = 40$. The base learning rates for SVMD and MSVGD are set to 0.01 and we run them for $T = 1000$ particle updates. τ is set to 0.98 for SVMD.

Our 2D example in Fig. 3 (left) is grabbed from one run of the simulation where there happen to be only 2 features selected by the randomized Lasso. The selective distribution in this case has log-density $\log p(\theta) = -8.07193((2.39859\theta_1 + 1.90816\theta_2 + 2.39751)^2 + (1.180999\theta_2 - 1.46104)^2) + \mathrm{const}, \theta_{1,2} \geq 0$.

The error bars for actual coverage levels in Fig. 4(a) and Fig. 4(b) are 95% Wilson intervals (Wilson, 1927), which is known to be more accurate than ±2 standard deviation intervals for binomial proportions like the coverage. In Fig. 8(a) and Fig. 8(b) we additionally plot the average length of the confidence intervals w.r.t. different sample size N and nominal coverage levels. For all three methods the CI widths are very close, although MSVGD consistently has wider intervals than SVMD and `selectiveInference`. This indicates that SVMD can be preferred over MSVGD when both methods produce coverage above the nominal level.

HIV-1 drug resistance We take the vitro measurement of log-fold change under the 3TC drug as response and include mutations that had appeared 11 times in the dataset as regressors. This results in $\tilde{n} = 663$ datapoints with $p = 91$ features. We choose λ to be the value returned by `theoretical.lambda` of the `selectiveInference` R package multiplied by $\tilde{n}$. The base learning rates for SVMD and MSVGD are set to 0.01 and we run them for $T = 2000$ particle updates. τ is set to 0.99 for SVMD.



(a) Nominal coverage: 0.9

(b) $N = 5000$ sample points

Figure 8: Width of post-selection CIs across (a) 500 / (b) 200 replications of simulation of Sepehri and Markovic (2017).

K.3. Large-scale posterior inference with non-Euclidean geometry

Finally, we demonstrate the advantages of exploiting non-Euclidean geometry by recreating the real-data large-scale Bayesian logistic regression experiment of Liu and Wang (2016) with 581,012 datapoints and $d = 54$ feature dimensions. Here, the target p is the posterior distribution over logistic regression parameters. We adopt the Fisher information metric tensor G, compare 20-particle SVNG to SVGD and its prior geometry-aware variants RSVGD (Liu and Zhu, 2018) and MatSVGD with average and mixture kernels (Wang et al., 2019), and for all methods use stochastic minibatches of size 256 to scalably approximate each log likelihood query. In Fig. 9, all geometry-aware methods substantially improve the log predictive probability of SVGD.[5] SVNG also strongly outperforms RSVGD and converges to its maximum test probability in half as many steps as MatSVGD (Avg) and more rapidly than MatSVGD (Mixture).

The Bayesian logistic regression model we consider is $\prod_{i=1}^{\tilde{n}} p(y_i|x_i, w)p(w)$, where $p(w) = \mathcal{N}(w|0, I)$, $p(y_i|x_i, w) = \text{Bernoulli}(\sigma(w^\top x_i))$. The bias parameter is absorbed into into w by adding an additional feature 1 to each x_i. The gradient of the log density of the posterior distribution of w is $\nabla_w \log p(w|\{y_i, x_i\}_{i=1}^N) = \sum_{i=1}^N x_i(y_i - \sigma(w^\top x_i)) - w$. We choose the

[5] Notably, on the same dataset, SVGD was shown to outperform preconditioned stochastic gradient Langevin dynamics (Li et al., 2016), a leading MCMC method imbued with geometric information (Wang et al., 2019).

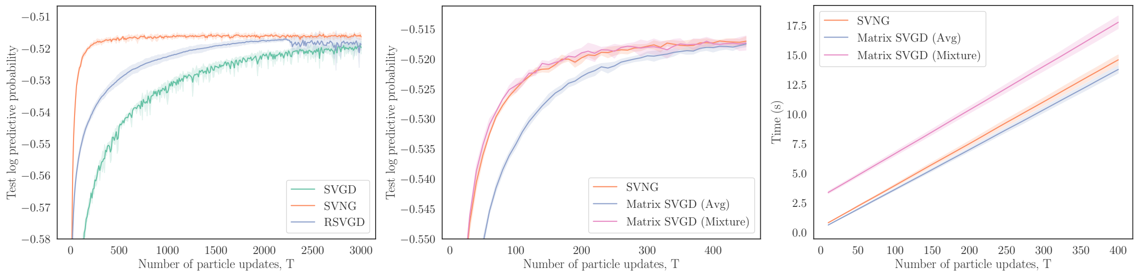


Figure 9: Value of non-Euclidean geometry in large-scale Bayesian logistic regression.

metric tensor $\nabla^2\psi(w)$ to be the Fisher information matrix (FIM) of the likelihood:

$$\begin{aligned} F &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathbb{E}_{p(y_i|w,x_i)}[\nabla_w \log p(y_i|x_i,w)\nabla_w \log p(y_i|x_i,w)^\top] \\ &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \sigma(w^\top x_i)(1-\sigma(w^\top x_i))x_i x_i^\top. \end{aligned}$$

Following Wang et al. (2019), for each iteration r ($r \geq 1$), we estimate the sum with a stochastic minibatch $\mathcal{B}_r$ of size 256: $\hat{F}_{\mathcal{B}_r} = \frac{\tilde{n}}{|\mathcal{B}_r|}\sum_{i\in\mathcal{B}_r}\sigma(w^\top x_i)(1-\sigma(w^\top x_i))x_i x_i^\top$ and approximate the FIM with a moving average across iterations:

$$\hat{F}_r = \rho_r \hat{F}_{r-1} + (1-\rho_r)\hat{F}_{\mathcal{B}_r}, \quad \text{where } \rho_r = \min(1-1/r, 0.95).$$

To ensure the positive definiteness of the FIM, a damping term $0.01I$ is added before taking the inverse. For RSVGD and SVNG, the gradient of the inverse of FIM is estimated with $\nabla_{w_j} F^{-1} \approx -\hat{F}_r^{-1}(\hat{\nabla}^r_{w_j}F)\hat{F}_r^{-1}$, where $\hat{\nabla}^r_{w_j}F = \rho_r\hat{\nabla}^{r-1}_{w_j}F + (1-\rho_r)\nabla_{w_j}\hat{F}_{\mathcal{B}_r}$.

We run each method for $T = 3000$ particle updates with learning rates in $\{0.01, 0.05, 0.1, 0.5, 1\}$ and average the results for 5 random trials. τ is set to 0.98 for SVNG. For each run, we randomly keep 20% of the dataset as test data, 20% of the remaining points as the validation set, and all the rest as the training set. The results of each method on validation sets with all choices of learning rates are plotted in Fig. 10. We see that the SVNG updates are very robust to the change in learning rates and is able to accommodate very large learning rates (up to 1) without a significant loss in performance. The results in Fig. 9 are reported with the learning rate that performs best on the validation set.

Appendix L. Proofs

L.1. Proof of Prop. 2

Proof Fix any $g \in \mathcal{G}_\psi$. Since g and ∇g are bounded and $\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta)$ and $\nabla \cdot \nabla^2\psi(\theta)^{-1}$ are p-integrable, the expectation $\mathbb{E}_p[(\mathcal{M}_{p,\psi}g)(\theta)]$ exists. Because Θ is convex, Θ_r

is bounded and convex with Lipschitz boundary. Since $p\nabla^2\psi^{-1}g \in C^1$, we have

$$\begin{aligned}
|\mathbb{E}_p[(\mathcal{M}_{p,\psi}g)(\theta)]| &= |\mathbb{E}_p[g(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot (\nabla^2\psi(\theta)^{-1}g(\theta))]| \\
&= \left| \int_\Theta \nabla p(\theta)^\top \nabla^2\psi(\theta)^{-1}g(\theta) + p(\theta)\nabla \cdot (\nabla^2\psi(\theta)^{-1}g(\theta))d\theta \right| \\
&= \left| \int_\Theta \nabla \cdot (p(\theta)\nabla^2\psi(\theta)^{-1}g(\theta))d\theta \right| \\
&= \left| \lim_{r\to\infty} \int_{\Theta_r} \nabla \cdot (p(\theta)\nabla^2\psi(\theta)^{-1}g(\theta))d\theta \right| \quad \text{(by dominated convergence)} \\
&= \left| \lim_{r\to\infty} \int_{\partial\Theta_r} (p(\theta)\nabla^2\psi(\theta)^{-1}g(\theta))^\top n_r(\theta)d\theta \right| \quad \text{(by the divergence theorem)} \\
&\le \lim_{r\to\infty} \int_{\partial\Theta_r} p(\theta)\|g(\theta)\|_2 \left\|\nabla^2\psi(\theta)^{-1}n_r(\theta)\right\|_2 d\theta \quad \text{(by Cauchy-Schwarz)} \\
&\le \|g\|_\infty \lim_{r\to\infty} \int_{\partial\Theta_r} p(\theta) \left\|\nabla^2\psi(\theta)^{-1}n_r(\theta)\right\|_2 d\theta = 0 \quad \text{(by assumption).}
\end{aligned}$$

■

L.2. Proof of Thm. 3: Optimal mirror updates in RKHS

Proof Let e_i denote the standard basis vector of $\mathbb{R}^d$ with the i-th element being 1 and others being zeros. Since $m \in \mathcal{H}_K$, we have

$$\begin{aligned}
m(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) &= \langle m, K(\cdot,\theta)\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta)\rangle_{\mathcal{H}_K} \\
\nabla \cdot (\nabla^2\psi(\theta)^{-1}m(\theta)) &= \sum_{i=1}^d \nabla_{\theta_i}(m(\theta)^\top \nabla^2\psi(\theta)^{-1}e_i) \\
&= \sum_{i=1}^d \langle m, \nabla_{\theta_i}(K(\cdot,\theta)\nabla^2\psi(\theta)^{-1}e_i)\rangle_{\mathcal{H}_K} \\
&= \langle m, \nabla_\theta \cdot (K(\cdot,\theta)\nabla^2\psi(\theta)^{-1})\rangle_{\mathcal{H}_K},
\end{aligned}$$

where we define the divergence of a matrix as a vector whose elements are the divergences of each row of the matrix. Then, we write (6) as

$$\begin{aligned}
&-\mathbb{E}_{q_t}[m(\theta)^\top \nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla \cdot (\nabla^2\psi(\theta)^{-1}m(\theta))] \\
&= -\mathbb{E}_{q_t}[\langle m, K(\cdot,\theta)\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla_\theta \cdot (K(\cdot,\theta)\nabla^2\psi(\theta)^{-1})\rangle_{\mathcal{H}_K}] \\
&= -\langle m, \mathbb{E}_{q_t}[K(\cdot,\theta)\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla_\theta \cdot (K(\cdot,\theta)\nabla^2\psi(\theta)^{-1})]\rangle_{\mathcal{H}_K} \\
&= -\langle m, \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi}K(\cdot,\theta)]\rangle_{\mathcal{H}_K}.
\end{aligned}$$

Therefore, the optimal direction in the $\mathcal{H}_K$ norm ball $\mathcal{B}_{\mathcal{H}_K} = \{g : \|g\|_{\mathcal{H}_K} \le 1\}$ that minimizes (6) is $g_t^* \propto g_{q_t,K}^* = \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi}K(\cdot,\theta)]$. ■

L.3. Proof of Thm. 4: Mirrored SVGD updates

Proof A p.d. kernel k composed with any map ϕ is still a p.d. kernel. To prove this, let $\{x_1, \dots, x_p\} = \{\phi(\eta_1), \dots, \phi(\eta_n)\}$, $p \le n$. Then

$$\sum_{i,j} \alpha_i \alpha_j k(\phi(\eta_i), \phi(\eta_j)) = \sum_{\ell,m} \beta_\ell \beta_m k(x_\ell, x_m) \ge 0,$$

where $\beta_\ell = \sum_{i \in S_\ell} \alpha_i$, $S_\ell = \{i : \phi(\eta_i) = x_\ell\}$. Therefore, $k_\psi(\eta, \eta') = k(\nabla\psi^*(\eta), \nabla\psi^*(\eta'))$ is a p.d. kernel. Plugging $K = kI$ into Lem. 14, for any $\theta' \in \Theta$ and $\eta' = \nabla\psi(\theta')$, we have

$$\begin{aligned}
g^*_{q_t,K_k}(\theta') &= \mathbb{E}_{\eta_t \sim q_{t,H}}[K_{\nabla\psi^*}(\nabla\psi(\theta'), \eta_t)\nabla \log p_H(\eta_t) + \nabla_{\eta_t} \cdot K_{\nabla\psi^*}(\nabla\psi(\theta'), \eta_t)] \\
&= \mathbb{E}_{\eta_t \sim q_{t,H}}[k(\nabla\psi^*(\eta'), \nabla\psi^*(\eta_t))\nabla \log p_H(\eta_t) + \sum_{j=1}^{d} \nabla_{\eta_{t,j}} k(\nabla\psi^*(\eta'), \nabla\psi^*(\eta_t)) e_j] \\
&= \mathbb{E}_{\eta_t \sim q_{t,H}}[k_\psi(\eta', \eta_t)\nabla \log p_H(\eta_t) + \nabla_{\eta_t} k_\psi(\eta', \eta_t)].
\end{aligned}$$

■

L.4. Proof of Prop. 6: Single-particle SVMD is mirror descent

Proof When $n = 1$, $\lambda_1 = k(\theta_t, \theta_t), u_1 = 1$, and thus $K_{\psi,t}(\theta_t, \theta_t) = k(\theta_t, \theta_t)\nabla^2\psi(\theta_t)$. ■

L.5. Proof of Thm. 9: Convergence of mirrored updates as $n \to \infty$

Proof The idea is to reinterpret our mirrored updates as one step of a matrix SVGD in η space based on Lem. 14 and then follow the path of Gorham et al. (2020, Thm. 7). Assume that $q^n_{t,H}$ and $q^\infty_{t,H}$ have integrable means. Let η^n, η^∞ be an optimal Wasserstein-1 coupling of $q^n_{t,H}$ and $q^\infty_{t,H}$. Let Φ_{q_t,K_t} denote the transform through one step of mirrored update: $\theta_t = \nabla\psi^\star(\eta_t)$, $\eta_{t+1} = \eta_t + \epsilon_t g^*_{q_t,K_t}(\theta_t)$. Then, with Lem. 14, we have

$$\begin{aligned}
&\|\Phi_{q_t,K_t}(\eta) - \Phi_{q_t,K_t}(\eta')\|_2 \\
&= \|\eta + \epsilon_t g^*_{q^n_t,K_t}(\theta) - \eta' - \epsilon_t g^*_{q^\infty_t,K_t}(\theta')\|_2 \\
&\le \|\eta - \eta'\|_2 + \epsilon_t \|g^*_{q^n_t,K_t}(\theta) - g^*_{q^\infty_t,K_t}(\theta')\|_2 \\
&\le \|\eta - \eta'\|_2 \\
&+ \epsilon_t \|\mathbb{E}_{\eta^n}[K_{\nabla\psi^*,t}(\eta, \eta^n)\nabla \log p_H(\eta^n) + \nabla_{\eta^n} \cdot K_{\nabla\psi^*,t}(\eta, \eta^n) \\
&\qquad - (K_{\nabla\psi^*,t}(\eta', \eta^n)\nabla \log p_H(\eta^n) + \nabla_{\eta^n} \cdot K_{\nabla\psi^*,t}(\eta', \eta^n))]\|_2 \\
&+ \epsilon_t \|\mathbb{E}_{\eta^n,\eta^\infty}[K_{\nabla\psi^*,t}(\eta', \eta^n)\nabla \log p_H(\eta^n) + \nabla_{\eta^n} \cdot K_{\nabla\psi^*,t}(\eta, \eta^n) \\
&\qquad - (K_{\nabla\psi^*,t}(\eta', \eta^\infty)\nabla \log p_H(\eta^\infty) + \nabla_{\eta^\infty} \cdot K_{\nabla\psi^*,t}(\eta', \eta^\infty))]\|_2 \\
&\le \|\eta - \eta'\|_2 + \epsilon_t c_1(1 + \mathbb{E}[\|\eta^n\|_2)]\|\eta - \eta'\|_2 + \epsilon_t c_2(1 + \left\|\eta'\right\|_2)\mathbb{E}_{\eta^n,\eta^\infty}[\|\eta^n - \eta^\infty\|_2] \\
&= \|\eta - \eta'\|_2 + \epsilon_t c_1(1 + \mathbb{E}_{q^n_{t,H}}[\|\cdot\|_2])\|\eta - \eta'\|_2 + \epsilon_t c_2(1 + \left\|\eta'\right\|_2)W_1(q^n_{t,H}, q^\infty_{t,H}).
\end{aligned}$$

Since $\Phi_{q_t,K_t}(\eta^n) \sim q^n_{t+1,H}$, $\Phi_{q_t,K}(\eta^\infty) \sim q^\infty_{t+1,H}$, we conclude

$$\begin{aligned}
&W_1(q^n_{t+1,H}, q^\infty_{t+1,H}) \\
&\leq \mathbb{E}[\|\Phi_{q_t,K}(\eta^n) - \Phi_{q_t,K}(\eta^\infty)\|_2] \\
&\leq (1 + \epsilon_t c_1(1 + \mathbb{E}_{q^n_{t,H}}[\|\cdot\|_2]))\mathbb{E}[\|\eta^n - \eta^\infty\|_2] + \epsilon_t c_2(1 + \left\|\eta'\right\|_2)W_1(q^n_{t,H}, q^\infty_{t,H})] \\
&\leq (1 + \epsilon_t c_1(1 + \mathbb{E}_{q^n_{t,H}}[\|\cdot\|_2]) + \epsilon_t c_2(1 + \mathbb{E}_{q^\infty_{t,H}}[\|\cdot\|_2]))W_1(q^n_{t,H}, q^\infty_{t,H}).
\end{aligned}$$

The final claim $q^n_t \Rightarrow q^\infty_t$ now follows by the continuous mapping theorem as $\nabla\psi^*$ is continuous. ■

L.6. Proof of Thm. 11: Infinite-particle mirrored Stein updates decrease KL and MKSD

Proof Let $T_{q^\infty_t,K_t}$ denote transform of the density function through one step of mirrored update: $\theta_t = \nabla\psi^\star(\eta_t)$, $\eta_{t+1} = \eta_t + \epsilon_t g^*_{q^\infty_t,K_t}(\theta_t)$. Then

$$\begin{aligned}
&\mathrm{KL}(q^\infty_{t+1}\|p) - \mathrm{KL}(q^\infty_t\|p) \\
&= \mathrm{KL}(q^\infty_t\|T^{-1}_{q^\infty_t,K_t}p) - \mathrm{KL}(q^\infty_t\|p) \\
&= \mathbb{E}_{\eta_t\sim q^\infty_{t,H}}[\log p_H(\eta_t) - \log p_H(\eta_t + \epsilon_t g^*_{q^\infty_t,K_t}(\theta_t)) - \log|\det(I + \epsilon_t\nabla_{\eta_t} g^*_{q^\infty_t,K_t}(\theta_t))|],
\end{aligned}$$

where we have used the invariance of KL divergence under reparameterization: $\mathrm{KL}(q_t\|p) = \mathrm{KL}(q_{t,H}\|p_H)$. Following Liu (2017), we bound the difference of the first two terms as

$$\begin{aligned}
&\log p_H(\eta_t) - \log p_H(\eta_t + \epsilon_t g^*_{q^\infty_t,K_t}(\theta_t)) \\
&= -\int_0^1 \nabla_s \log p_H(\eta_t(s))\, ds, \quad \text{where } \eta_t(s) \triangleq \eta_t + s\epsilon_t g^*_{q^\infty_t,K_t}(\theta_t) \\
&= -\int_0^1 \nabla \log p_H(\eta_t(s))^\top(\epsilon_t g^*_{q^\infty_t,K_t}(\theta_t))\, ds \\
&= -\epsilon_t\nabla\log p_H(\eta_t)^\top g^*_{q^\infty_t,K_t}(\theta_t) + \int_0^1 (\nabla\log p_H(\eta_t) - \nabla\log p_H(\eta_t(s)))^\top(\epsilon_t g^*_{q^\infty_t,K_t}(\theta_t))\, ds \\
&\leq -\epsilon_t\nabla\log p_H(\eta_t)^\top g^*_{q^\infty_t,K_t}(\theta_t) + \epsilon_t\int_0^1 \|\nabla\log p_H(\eta) - \nabla\log p_H(\eta_t(s))\|_2 \cdot \|g^*_{q^\infty_t,K_t}(\theta_t)\|_2\, ds \\
&\leq -\epsilon_t\nabla\log p_H(\eta_t)^\top g^*_{q^\infty_t,K_t}(\theta_t) + \frac{L\epsilon_t^2}{2}\|g^*_{q^\infty_t,K_t}(\theta_t)\|_2^2,
\end{aligned}$$

and bound the log determinant term using Lem. 16:

$$-\log|\det(I + \epsilon_t\nabla_{\eta_t} g^*_{q^\infty_t,K_t}(\theta_t)) \leq -\epsilon_t\,\mathrm{Tr}(\nabla_{\eta_t} g^*_{q^\infty_t,K_t}(\theta_t)) + 2\epsilon_t^2\|\nabla_{\eta_t} g^*_{q^\infty_t,K_t}(\theta_t)\|_F^2.$$

The next thing to notice is that $\mathbb{E}_{\eta_t\sim q^\infty_{t,H}}[\nabla\log p_H(\eta_t)^\top g^*_{q^\infty_t,K_t}(\theta_t) + \mathrm{Tr}(\nabla_{\eta_t} g^*_{q^\infty_t,K_t}(\theta_t))]$ is the square of the MKSD in (17). We can show this equivalence using the identity proved in

Lem. 15:

$$
\begin{aligned}
&\mathbb{E}_{\eta_t\sim q^\infty_{t,H}}[g^*_{q_t^\infty,K_t}(\theta_t)^\top\nabla\log p_H(\eta_t)+\mathrm{Tr}(\nabla_{\eta_t}g^*_{q_t^\infty,K_t}(\theta_t))]\\
&\quad=\mathbb{E}_{\theta_t\sim q_t^\infty}[g^*_{q_t^\infty,K_t}(\theta_t)^\top\nabla^2\psi(\theta_t)^{-1}\nabla_{\theta_t}(\log p(\theta_t)-\log\det\nabla^2\psi(\theta_t))\\
&\qquad+\mathrm{Tr}(\nabla^2\psi(\theta_t)^{-1}\nabla g^*_{q_t^\infty,K_t}(\theta_t))]\\
&\quad=\mathbb{E}_{\theta_t\sim q_t^\infty}[g^*_{q_t^\infty,K_t}(\theta_t)^\top\nabla^2\psi(\theta_t)^{-1}\nabla\log p(\theta_t)+\nabla\cdot(\nabla^2\psi(\theta_t)^{-1}g^*_{q_t^\infty,K_t}(\theta_t))]\quad\text{(Lem. 15)}\\
&\quad=\mathbb{E}_{\theta_t\sim q_t^\infty}[(\mathcal{M}_{p,\psi}g^*_{q_t^\infty,K_t})(\theta_t)]\\
&\quad=\mathrm{MKSD}_{K_t}(q_t^\infty,p)^2.
\end{aligned}
$$

Finally, we are going to bound $\|g^*_{q_t^\infty,K_t}(\theta_t)\|_2^2$ and $\|\nabla_{\eta_t}g^*_{q_t^\infty,K_t}(\theta_t)\|_F^2$. From the assumptions we have ψ is α-strongly convex and thus ψ^* is $\frac{1}{\alpha}$-strongly smooth (Kakade et al., 2009), therefore $\|\nabla^2\psi^*(\cdot)\|_2\le\frac{1}{\alpha}$. By Lem. 17, we know

$$
\begin{aligned}
\|g^*_{q_t^\infty,K_t}(\theta_t)\|_2^2&\le\|g^*_{q_t^\infty,K_t}\|^2_{\mathcal{H}_{K_t}}\|K(\theta_t,\theta_t)\|_{\mathrm{op}}=\mathrm{MKSD}_{K_t}(q_t^\infty,p)^2\|K_t(\theta_t,\theta_t)\|_{\mathrm{op}},\\
\|\nabla_{\eta_t}g^*_{q_t^\infty,K_t}(\theta_t)\|_F^2&=\|\nabla^2\psi^*(\eta_t)\nabla g^*_{q_t^\infty,K_t}(\theta_t)\|_F^2\\
&\le\|\nabla^2\psi^*(\eta_t)\|_2^2\|\nabla g^*_{q_t^\infty,K_t}(\theta_t)\|_F^2\\
&\le\frac{1}{\alpha^2}\|g^*_{q_t^\infty,K_t}\|^2_{\mathcal{H}_{K_t}}\sum_{i=1}^d\|\nabla^2_{i,d+i}K_t(\theta_t,\theta_t)\|_{\mathrm{op}}\\
&=\frac{1}{\alpha^2}\mathrm{MKSD}_{K_t}(q_t^\infty,p)^2\sum_{i=1}^d\|\nabla^2_{i,d+i}K_t(\theta_t,\theta_t)\|_{\mathrm{op}},
\end{aligned}
$$

where $\nabla^2_{i,d+i}K(\theta,\theta)$ denotes $\nabla^2_{\theta_i,\theta'_i}K(\theta,\theta')|_{\theta'=\theta}$. Combining all of the above, we have

$$
\begin{aligned}
&\mathrm{KL}(q_{t+1}^\infty\|p)-\mathrm{KL}(q_t^\infty\|p)\\
&\le-\left(\epsilon_t-\frac{L\epsilon_t^2}{2}\sup_\theta\|K_t(\theta,\theta)\|_{\mathrm{op}}-\frac{2\epsilon_t^2}{\alpha^2}\sum_{i=1}^d\sup_\theta\|\nabla^2_{i,d+i}K_t(\theta,\theta)\|_{\mathrm{op}}\right)\mathrm{MKSD}_{K_t}(q_t^\infty,p)^2.
\end{aligned}
$$

Plugging in the definition of κ_1 and κ_2 finishes the proof. ■

L.7. Proof of Thm. 12: MKSD_{K_k} determines weak convergence

Proof According to Thm. 4,

$$g^*_{q,K_k}=\mathbb{E}_{q_H}[k(\cdot,\nabla\psi^*(\eta))\nabla\log p_H(\eta)+\nabla_\eta k(\nabla\psi^*(\eta),\cdot)],$$

where $q_H(\eta)$ denotes the density of $\eta=\nabla\psi(\theta)$ under the distribution $\theta\sim q$. From the assumptions we have $k(\theta,\theta')=\kappa(\nabla\psi(\theta),\nabla\psi(\theta'))$. With this specific choice of k, the squared MKSD is

$$
\begin{aligned}
\mathrm{MKSD}_{K_k}(q,p)^2&=\|g^*_{q,K_k}\|^2_{\mathcal{H}_{K_k}}\\
&=\mathbb{E}_{\eta,\eta'\sim q_H}\left[\frac{1}{p_H(\eta)p_H(\eta')}\nabla_\eta\nabla_{\eta'}(p_H(\eta)k(\nabla\psi^*(\eta),\nabla\psi^*(\eta'))p_H(\eta'))\right]\\
&=\mathbb{E}_{\eta,\eta'\sim q_H}\left[\frac{1}{p_H(\eta)p_H(\eta')}\nabla_\eta\nabla_{\eta'}(p_H(\eta)\kappa(\eta,\eta')p_H(\eta'))\right].
\end{aligned}\tag{19}
$$

The final expression in (19) is the squared kernel Stein discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017) between q_H and p_H with the kernel κ: $\mathrm{KSD}_\kappa(q_H, p_H)^2$. Recall that it is proved in Gorham and Mackey (2017, Theorem 8) that, for $\kappa(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ with $\beta \in (-1, 0)$ and distantly dissipative p_H with Lipschitz score functions, $q_H \Rightarrow p_H$ if $\mathrm{KSD}_\kappa(q_H, p_H) \to 0$. The advertised result ($q \Rightarrow p$ if $\mathrm{MKSD}_{K_k}(q, p) \to 0$) now follows by the continuous mapping theorem as $\nabla\psi^*$ is continuous. ■

Appendix M. Lemmas

Lemma 14 *Let $K_{\nabla\psi^*}(\eta, \eta') \triangleq K(\nabla\psi^*(\eta), \nabla\psi^*(\eta'))$. The mirrored updates $g^*_{q_t,K}$ in (8) can be equivalently expressed as*

$$g^*_{q_t,K} = \mathbb{E}_{q_{t,H}}[K_{\nabla\psi^*}(\nabla\psi(\cdot), \eta)\nabla \log p_H(\eta) + \nabla_\eta \cdot K_{\nabla\psi^*}(\nabla\psi(\cdot), \eta)].$$

Proof We will use the identity proved in Lem. 15.

$$\begin{aligned}
g^*_{q_t,K} &= \mathbb{E}_{q_t}[\mathcal{M}_{p,\psi}K(\cdot, \theta)] \\
&= \mathbb{E}_{q_t}[K(\cdot, \theta)\nabla^2\psi(\theta)^{-1}\nabla \log p(\theta) + \nabla_\theta \cdot (K(\cdot, \theta)\nabla^2\psi(\theta)^{-1})] \\
&= \mathbb{E}_{q_t}[K(\cdot, \theta)\nabla^2\psi(\theta)^{-1}\nabla_\theta(\log p_H(\nabla\psi(\theta)) + \log\det \nabla^2\psi(\theta)) + \nabla_\theta \cdot (K(\cdot, \theta)\nabla^2\psi(\theta)^{-1})] \\
&\qquad \text{(by change-of-variable formula)} \\
&= \mathbb{E}_{q_t}[K(\cdot, \theta)\nabla^2\psi(\theta)^{-1}\nabla_\theta \log p_H(\nabla\psi(\theta)) + \sum_{i,j=1}^d [\nabla^2\psi(\theta)^{-1}]_{ij}\nabla_{\theta_i}K(\cdot, \theta)_{:,j}] \\
&\qquad \text{(by applying Lem. 15 to each row of } K(\cdot, \theta)) \\
&= \mathbb{E}_{q_t}[K(\cdot, \theta)\nabla^2\psi(\theta)^{-1}\nabla_\theta \log p_H(\nabla\psi(\theta)) + \sum_{j=1}^d \nabla_{\eta_j}K(\cdot, \theta)_{:,j}] \\
&= \mathbb{E}_{q_{t,H}}[K(\cdot, \nabla\psi^*(\eta))\nabla \log p_H(\eta) + \sum_{j=1}^d \nabla_{\eta_j}K(\cdot, \nabla\psi^*(\eta))_{:,j}] \\
&= \mathbb{E}_{q_{t,H}}[K_{\nabla\psi^*}(\nabla\psi(\cdot), \eta)\nabla \log p_H(\eta) + \nabla_\eta \cdot K_{\nabla\psi^*}(\nabla\psi(\cdot), \eta)],
\end{aligned}$$

where $A_{:,j}$ denotes the j-th column of a matrix A. ■

Lemma 15 *For a strictly convex function $\psi \in C^2 : \mathbb{R}^d \to \mathbb{R}$ and any vector-valued $g \in C^1 : \mathbb{R}^d \to \mathbb{R}^d$, the following relation holds:*

$$\nabla \cdot (\nabla^2\psi(\theta)^{-1}g(\theta)) = \mathrm{Tr}(\nabla^2\psi(\theta)^{-1}\nabla g(\theta)) - g(\theta)^\top\nabla^2\psi(\theta)^{-1}\nabla_\theta \log\det \nabla^2\psi(\theta).$$

Proof By the product rule of differentiation:

$$\nabla \cdot (\nabla^2\psi(\theta)^{-1}g(\theta)) = \mathrm{Tr}(\nabla^2\psi(\theta)^{-1}\nabla g(\theta)) + g(\theta)^\top\nabla \cdot (\nabla^2\psi(\theta)^{-1}). \tag{20}$$

This already gives us the first term on the right side. Next, we have

$$\begin{aligned}
&[\nabla^2\psi(\theta)^{-1}\nabla \log\det \nabla^2\psi(\theta)]_i \\
&= \sum_{j=1}^d [\nabla^2\psi(\theta)^{-1}]_{ij}\operatorname{Tr}(\nabla^2\psi(\theta)^{-1}\nabla_{\theta_j}\nabla^2\psi(\theta)) \\
&= \sum_{j=1}^d [\nabla^2\psi(\theta)^{-1}]_{ij}\sum_{\ell,m=1}^d [\nabla^2\psi(\theta)^{-1}]_{\ell m}[\nabla_{\theta_j}\nabla^2\psi(\theta)]_{m\ell} \\
&= \sum_{j,\ell,m=1}^d [\nabla^2\psi(\theta)^{-1}]_{ij}[\nabla^2\psi(\theta)^{-1}]_{\ell m}\nabla_{\theta_j}\nabla^2\psi(\theta)_{m\ell} \\
&= \sum_{j,\ell,m=1}^d [\nabla^2\psi(\theta)^{-1}]_{ij}\nabla_{\theta_m}\nabla^2\psi(\theta)_{j\ell}[\nabla^2\psi(\theta)^{-1}]_{\ell m} \\
&= -\sum_{m=1}^d \nabla_{\theta_m}(\nabla^2\psi(\theta)^{-1})_{im} \\
&= -[\nabla\cdot\nabla^2\psi(\theta)^{-1}]_i.
\end{aligned}$$

Plugging the above relation into (20) proves the claimed result. ■

Lemma 16 (Liu, 2017, Lemma A.1) *Let A be a square matrix, and $0 < \epsilon < \frac{1}{2}\|A + A^\top\|_{\mathrm{op}}$. Then,*

$$\log|\det(I + \epsilon A)| \ge \epsilon \operatorname{Tr}(A) - 2\epsilon^2\|A\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Lemma 17 *Let K be a matrix-valued kernel and $\mathcal{H}_K$ be the corresponding RKHS. Then, for any $f \in \mathcal{H}_K$ (f is vector-valued), we have*

$$\|f(x)\|_2 \le \|f\|_{\mathcal{H}_K}\|K(x,x)\|_{\mathrm{op}}^{1/2}, \quad \|\nabla f(x)\|_F^2 \le \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^d \|\nabla^2_{x_i,x_i'}K(x,x')|_{x'=x}\|_{\mathrm{op}},$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a matrix induced by the vector 2-norm.

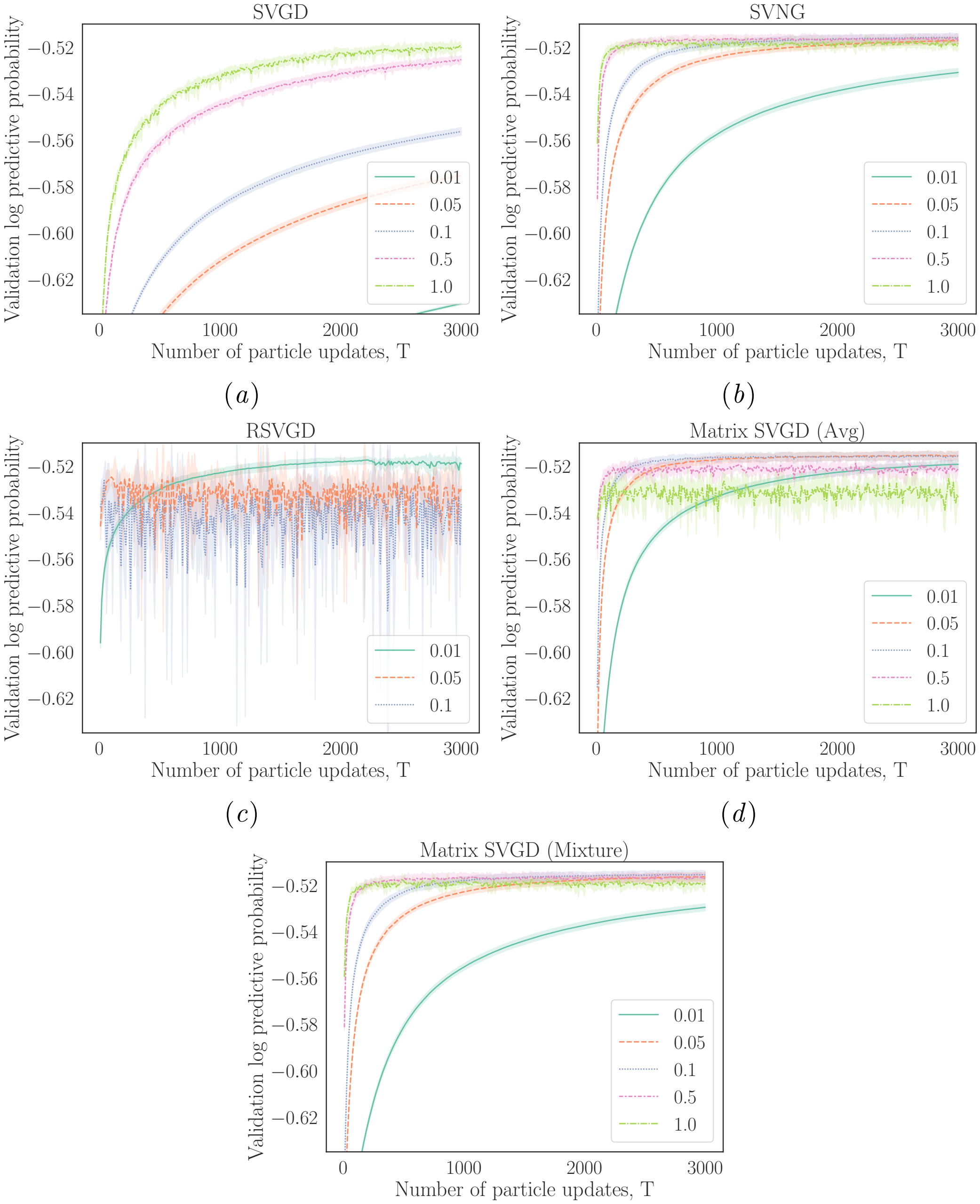
Proof We first bound the $\|f(x)\|_2$ as

$$\begin{aligned}
\|f(x)\|_2 &= \sup_{\|y\|_2=1} f(x)^\top y = \sup_{\|y\|_2=1}\langle f, K(\cdot,x)y\rangle_{\mathcal{H}_K} \le \|f\|_{\mathcal{H}_K}\sup_{\|y\|_2=1}\|K(\cdot,x)y\|_{\mathcal{H}_K} \\
&= \|f\|_{\mathcal{H}_K}\sup_{\|y\|_2=1}(y^\top K(x,x)y)^{1/2} \le \|f\|_{\mathcal{H}_K}\sup_{\|y\|_2=1}\sup_{\|u\|_2=1}(u^\top K(x,x)y)^{1/2} \\
&= \|f\|_{\mathcal{H}_K}\sup_{\|y\|_2=1}\|K(x,x)y\|_2^{1/2} = \|f\|_{\mathcal{H}_K}\|K(x,x)\|_{\mathrm{op}}^{1/2}.
\end{aligned}$$

The second result follows similarly,

$$
\begin{aligned}
\|\nabla f(x)\|_F^2 &= \sum_{i=1}^{d} \|\nabla_{x_i} f(x)\|_2^2 = \sum_{i=1}^{d} \sup_{\|y\|_2=1} (\nabla_{x_i} f(x)^\top y)^2 = \sum_{i=1}^{d} \sup_{\|y\|_2=1} (\nabla_{x_i} \langle f, K(\cdot, x) y \rangle_{\mathcal{H}_K})^2 \\
&= \sum_{i=1} \sup_{\|y\|_2=1} (\langle f, \nabla_{x_i} K(\cdot, x) y \rangle_{\mathcal{H}_K})^2 \leq \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^{d} \sup_{\|y\|_2=1} \|\nabla_{x_i} K(\cdot, x) y\|_{\mathcal{H}_K}^2 \\
&= \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^{d} \sup_{\|y\|_2=1} (y^\top \nabla^2_{x_i, x_i'} K(x, x')|_{x=x'} y) \\
&\leq \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^{d} \sup_{\|y\|_2=1} \sup_{\|u\|_2=1} (u^\top \nabla^2_{x_i, x_i'} K(x, x')|_{x=x'} y) \\
&= \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^{d} \sup_{\|y\|_2=1} \|\nabla^2_{x_i, x_i'} K(x, x')|_{x=x'} y\|_2 \\
&= \|f\|_{\mathcal{H}_K}^2 \sum_{i=1}^{d} \|\nabla^2_{x_i, x_i'} K(x, x')|_{x'=x}\|_{\mathrm{op}}.
\end{aligned}
$$

■



(a)

(b)

(c)

(d)

(e)

Figure 10: Logistic regression results on validation sets with learning rates in {0.01, 0.05, 0.1, 0.5, 1}. Running RSVGD with learning rates 0.5 and 1 produces numerical errors. Therefore, we did not include them in the plot.