Sampling with Mirrored Stein Operators

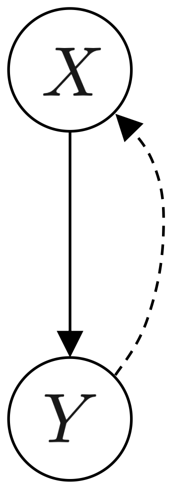
Jiaxin Shi

Microsoft Research New England

Aug 30, 2021

Joint work with Chang Liu, Lester Mackey

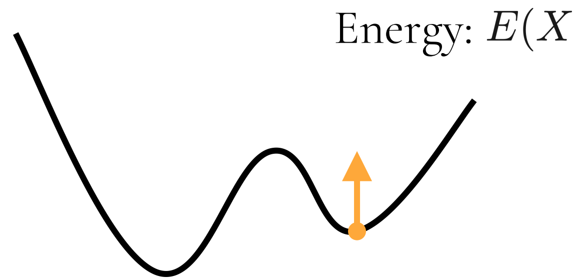
Sampling from an Unnormalized Distribution



$$p(X|Y) \propto p(Y|X)p(X)$$

Bayesian inference

$$p_\theta(X) \propto e^{-E_\theta(X)}$$

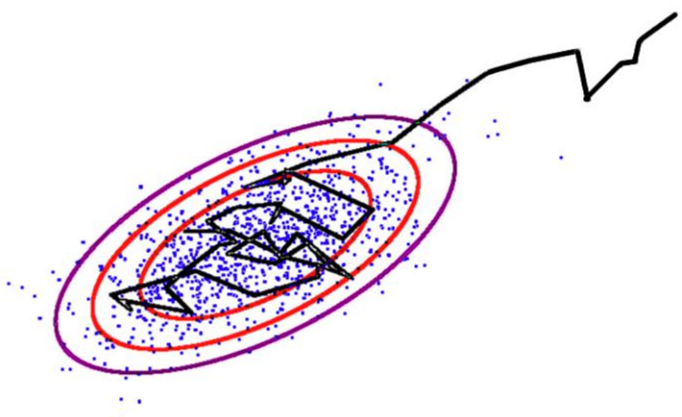


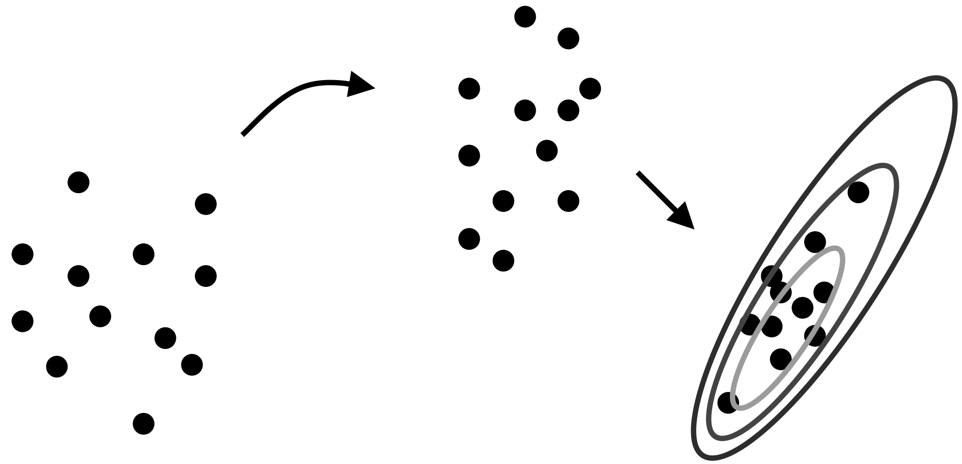
$$\nabla_\theta \log p_\theta(X) = \mathbb{E}_{X' \sim p}[E(X')] - E(X)$$

Learning unnormalized models

Solutions

Fig. from Murray (2009)



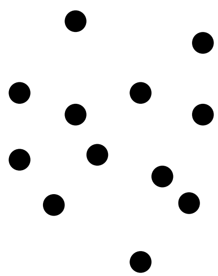


MCMC

Particle evolution methods

Stein Variational Gradient Descent (SVGD)

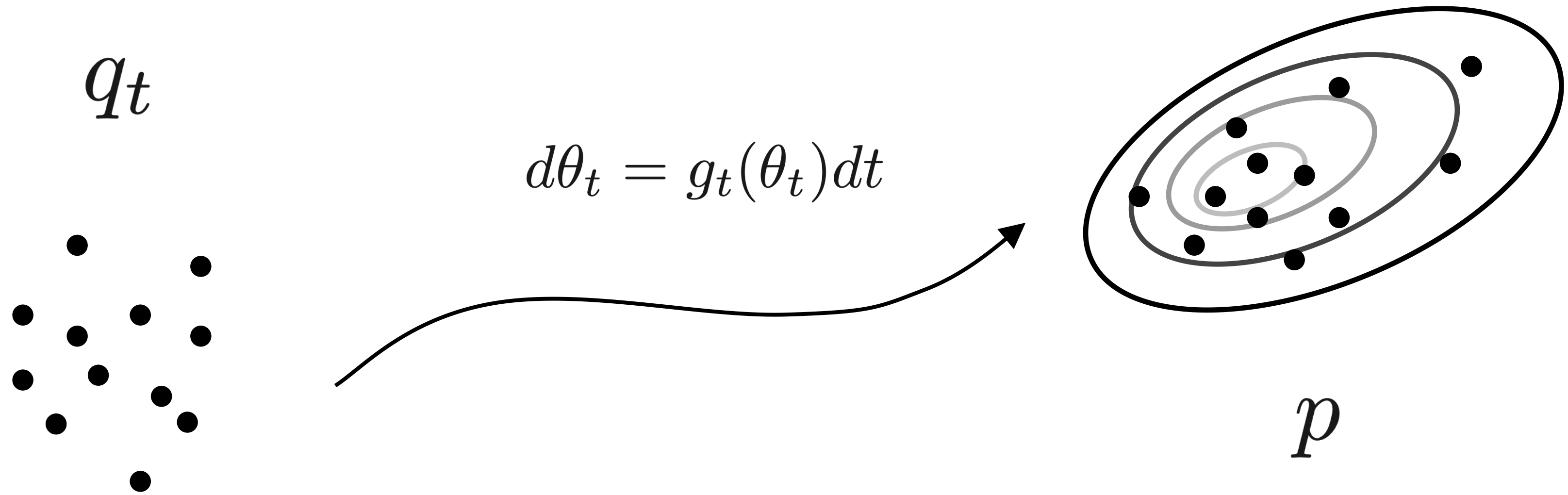
$$T(\theta) = \theta + \epsilon g(\theta)$$



$$\epsilon \to 0$$

$$\theta^1, \theta^2, \ldots, \theta^n$$

Stein Variational Gradient Descent (SVGD)



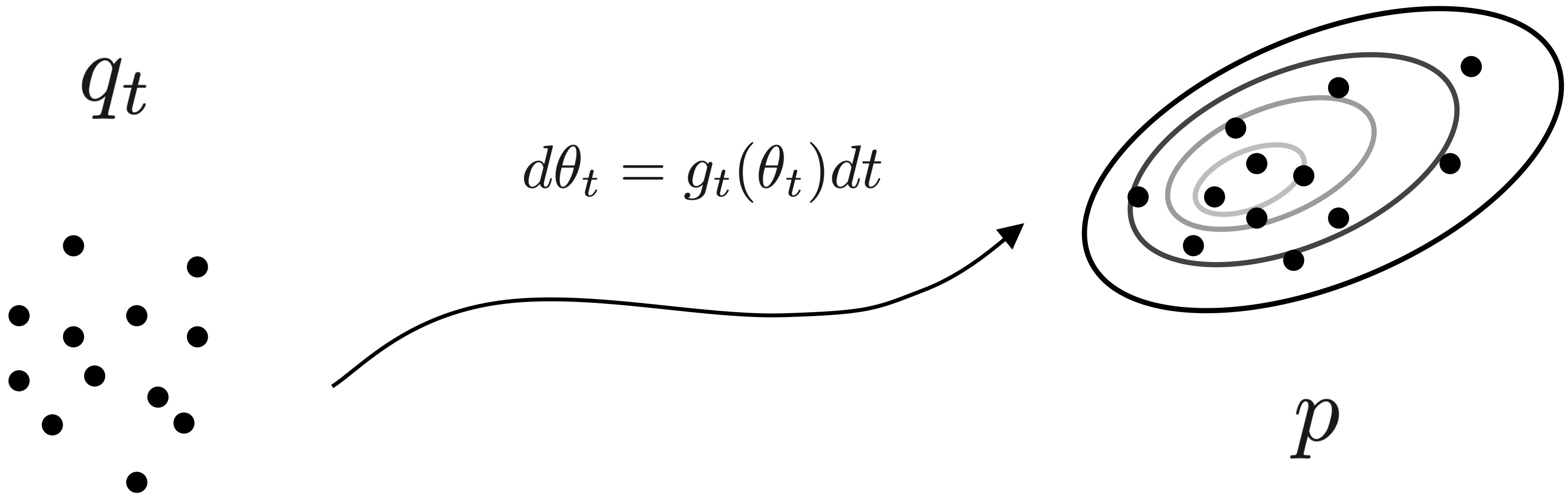
Langevin Stein Operator: $(\mathcal{S}_p g)(\theta) = g(\theta)^\top \nabla \log p(\theta) + \nabla \cdot g(\theta)$

(Liu & Wang, 2016)

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

Find the direction that **most quickly** decreases the KL divergence to p

Stein Variational Gradient Descent (SVGD)



Langevin Stein Operator: $(\mathcal{S}_p g)(\theta) = g(\theta)^\top \nabla \log p(\theta) + \nabla \cdot g(\theta)$

(Liu & Wang, 2016)

$$g_t^* = \underset{g_t \in \mathcal{H}, \|g_t\|_\mathcal{H} \leq 1}{\arg\min} \frac{d}{dt}\mathrm{KL}(q_t \| p) \propto \mathbb{E}_{q_t}[\mathcal{S}_p K(\cdot, \theta)]$$

Optimal direction in RKHS of K that **most quickly** decreases the KL divergence to p

Two Regimes of SVGD

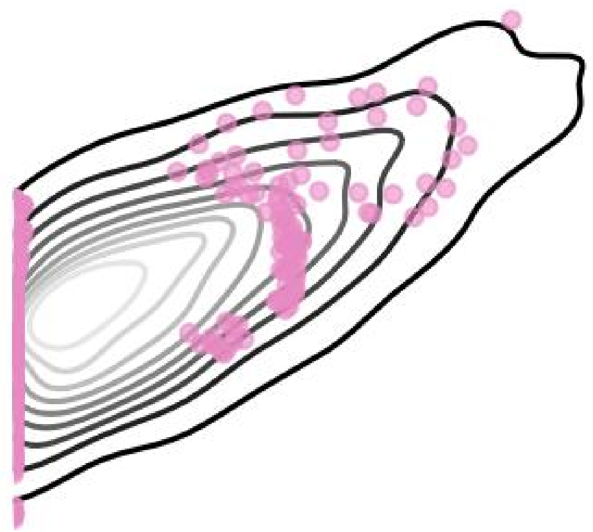
- $n = 1$: reduces to gradient descent on $-\log p(\theta)$ if $\nabla \cdot K(\theta, \theta) = 0$.
- $n \to \infty$: weak convergence to p under certain conditions.

(Gorham & Mackey, 2017; Liu 2017; Gorham et al., 2020)

(Liu & Wang, 2016)

$$\theta_{t+1}^i \leftarrow \theta_t^i + \epsilon_t \frac{1}{n} \sum_{j=1}^{n} \Big(K(\theta_t^i, \theta_t^j) \nabla \log p(\theta_t^j) + \nabla_{\theta_t^j} \cdot K(\theta_t^j, \theta_t^i) \Big)$$

SVGD Breaks Down for Constrained Targets



SVGD + Projection: Samples end up collecting on the boundary.

This Talk is About ..

Sampling

Optimization

Particle evolution samplers that work for <u>constrained targets</u> & exploit <u>non-Euclidean geometry</u>

Mirror descent

Natural gradient descent

Langevin Stein Operators

(Gorham & Mackey, 2015)

Under suitable boundary conditions, Langevin Stein Operator statisfies

$$\mathbb{E}_p[(\mathcal{S}_p g)(\theta)] = \mathbb{E}_p[g(\theta)^\top \nabla \log p(\theta) + \nabla \cdot g(\theta)]$$
$$= \int \nabla \cdot ((p(\theta)g(\theta))d\theta = 0$$

The last identity holds because of divergence theorem:

$$\int_\Theta \nabla \cdot ((p(\theta)g(\theta))d\theta = 0 \Leftrightarrow \int_{\partial\Theta} p(\theta)g(\theta)^\top n(\theta)d\theta = 0$$

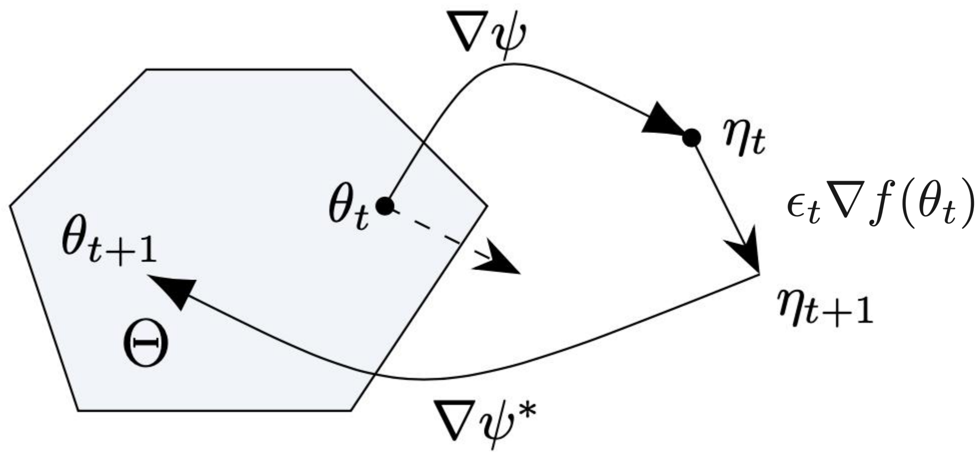
For unconstrained domain, since p vanishes at infinity, this holds under very mild conditions, such as bounded Lipschitz g.

Therefore, $q_t = p$ is a stationary point of the SVGD dynamics.

Two Problems of SVGD for Constrained Targets

- Standard SVGD updates can push the particles outside of its support
  - Result: Future updates undefined.
- The boundary conditions may fail to hold for g in the RKHS
  - This happens when p is non-vanishing or explosive on the boundary
  - Result: SVGD need not converge to p since p is not a stationary point.

Mirror Descent



Strictly convex $\psi : \Theta \to \mathbb{R} \cup \{\infty\}$

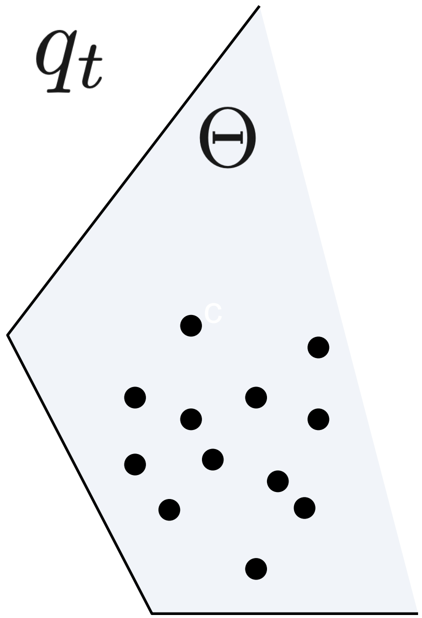
$(\nabla\psi)^{-1} = \nabla\psi^*$

Continuous time limit: mirror flow

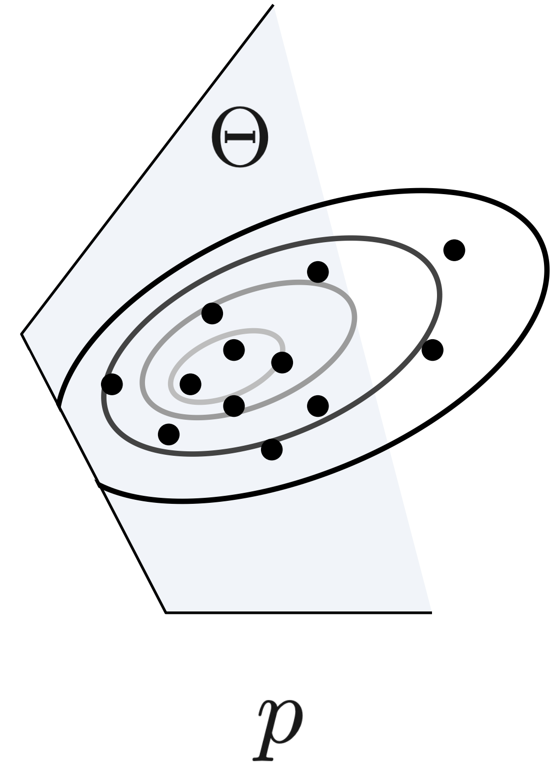
$$d\eta_t = -\nabla f(\theta_t)dt,\ \ \theta_t = \nabla\psi^*(\eta_t)$$

Equivalent Riemannian gradient flow: $\quad d\theta_t = -\nabla^2\psi(\theta_t)^{-1}\nabla f(\theta_t)dt$

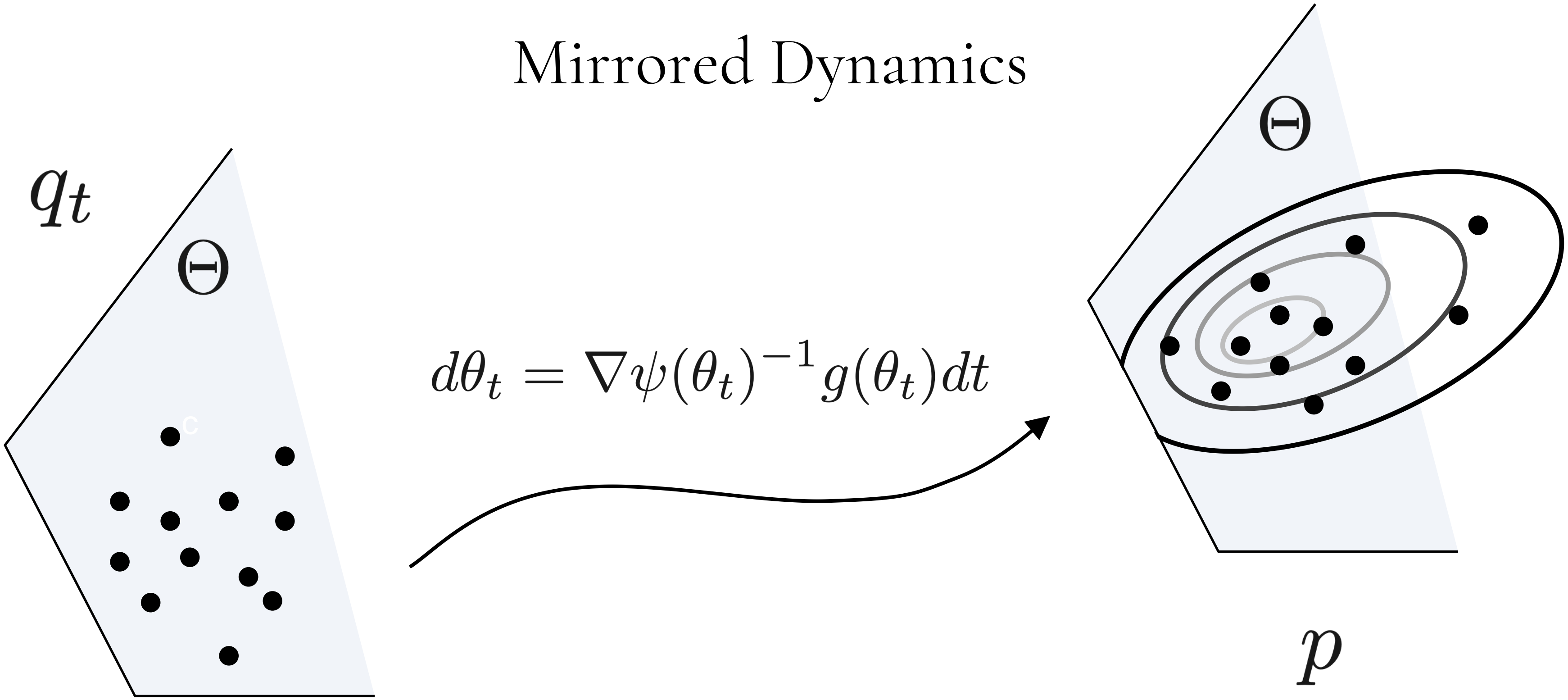
Mirrored Dynamics



$$d\eta_t = -g_t(\theta_t)dt,$$
$$\theta_t = \nabla\psi^*(\eta_t)$$



Mirrored Dynamics



$$d\theta_t = \nabla\psi(\theta_t)^{-1} g(\theta_t) dt$$

In analogy to Liu & Wang (2016)

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

Mirrored Stein Operator

A Stein Operator for Constrained Targets

Mirrored Stein Operator*

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

*Can be derived from the (infinitesimal) generator of Riemannian Langevin diffusion.

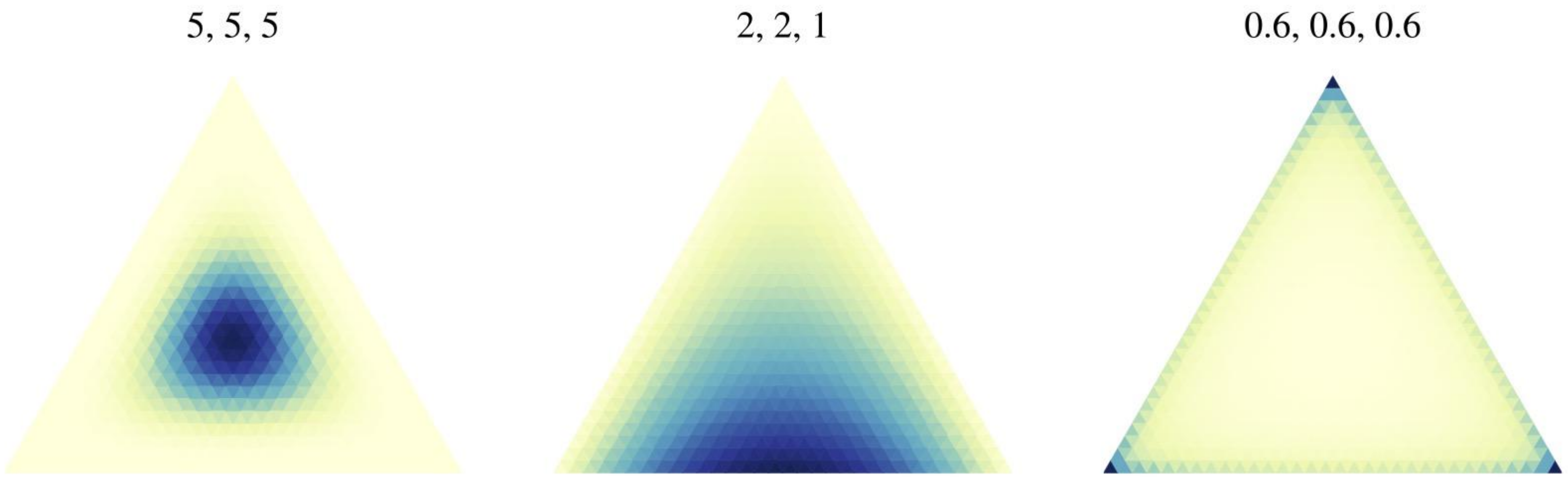
<u>**Proposition 1**</u> (informal) $\mathcal{M}_{p,\psi}$ generates mean-zero functions under p if

$$\int_{\partial\Theta} p(\theta) \|\nabla^2\psi(\theta)^{-1} n(\theta)\|_2 d\theta = 0$$

and $g \in C^1$ is bounded Lipschitz.

Intuitively, we expect $\nabla^2\psi(\theta)^{-1}$ to **cancel the growth** of p at the boundary.

Case Study: The Dirichlet Distribution



$$p(\theta) \propto \prod_{j=1}^{d+1} \theta_j^{\alpha_j - 1} \quad \begin{cases} \alpha_j < 1 : \theta_j \to 0, \theta_{-j} = \dfrac{1-\theta_j}{d} \Rightarrow p(\theta) \to \infty, \\ \alpha_j = 1 : \theta_j \to 0, \theta_{-j} = \dfrac{1-\theta_j}{d} \Rightarrow p(\theta) > 0. \end{cases}$$

$$\nabla^2 \psi(\theta)^{-1} = \mathrm{diag}(\theta) - \theta\theta^\top$$

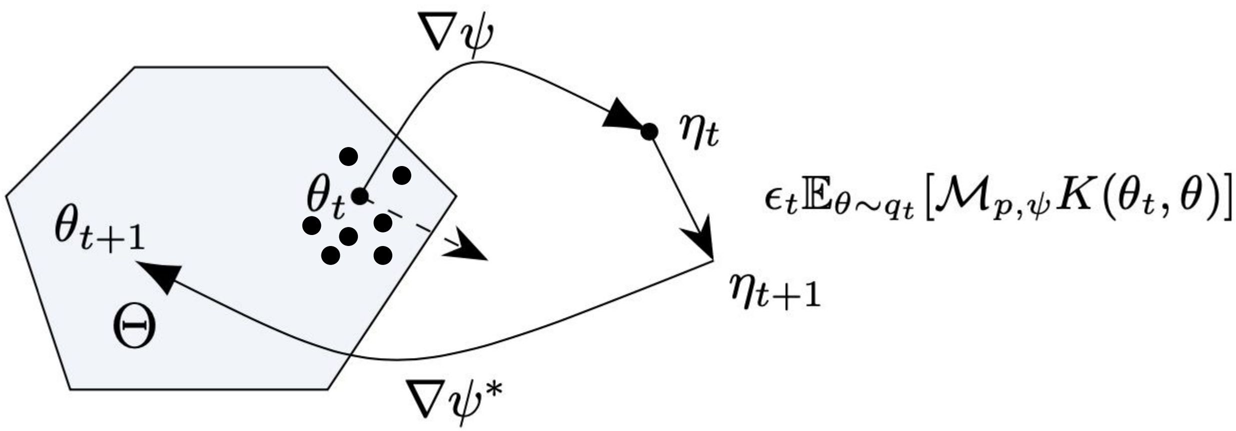
Negative entropy $\psi(\theta) = \sum_{j=1}^{d+1} \theta_j \log \theta_j$ meets the boundary condition

$$\int_{\partial\Theta} p(\theta) \|\nabla^2 \psi(\theta)^{-1} n(\theta)\|_2 d\theta = 0.$$

Sampling with Mirrored Stein Operators

Theorem 3 The optimal mirrored updates in the RKHS of K is

$$g_t^* = \underset{g_t \in \mathcal{H}, \|g_t\|_{\mathcal{H}} \le 1}{\arg\min} \frac{d}{dt}\mathrm{KL}(q_t \| p) \propto -\mathbb{E}_{q_t}[\mathcal{M}_{p,\psi} K(\cdot, \theta)]$$



Mirrored SVGD

<u>Theorem 4</u> If $K(\theta, \theta') = k(\theta, \theta')I$, then the optimal mirrored updates can alternatively be expressed as

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

where $k_\psi(\eta, \eta') = k(\nabla\psi^*(\eta), \nabla\psi^*(\eta'))$

transformed density of p
in dual space

- Mirrored SVGD is SVGD in η space with the **transformed kernel** k_ψ .
- When only a single particle is used $(n = 1)$, Mirrored SVGD reduces to gradient ascent on the log transformed density $\log p_H(\eta)$.

Single Particle MSVGD is Not Mirror Descent

Still want an algorithm that reduces to mirror descent when $n = 1$?

- θ space is the space we are primarily interested in.
- Mode in θ space need not match mode in η space
- Using $\log p(\theta)$ to guide the evolution could work better if $p(\theta)$ is better behaved than $p_H(\eta)$.

Stein Variational Mirror Descent (SVMD)

Key idea: Construct an **adaptive kernel** that

① incorporates the metric induced by ψ ② evolves with q_t

Definition (Kernels for SVMD)

Given a reference kernel k, we write it in Mercer's representation:

$$k(\theta, \theta') = \sum_{i \geq 1} \lambda_i u_i(\theta) u_i(\theta'),$$

where u_i is an eigenfunction satisfying:

$$\mathbb{E}_{q_t(\theta')}[k(\theta, \theta') u_i(\theta')] = \lambda_i u_i(\theta).$$

Kernels for SVMD:

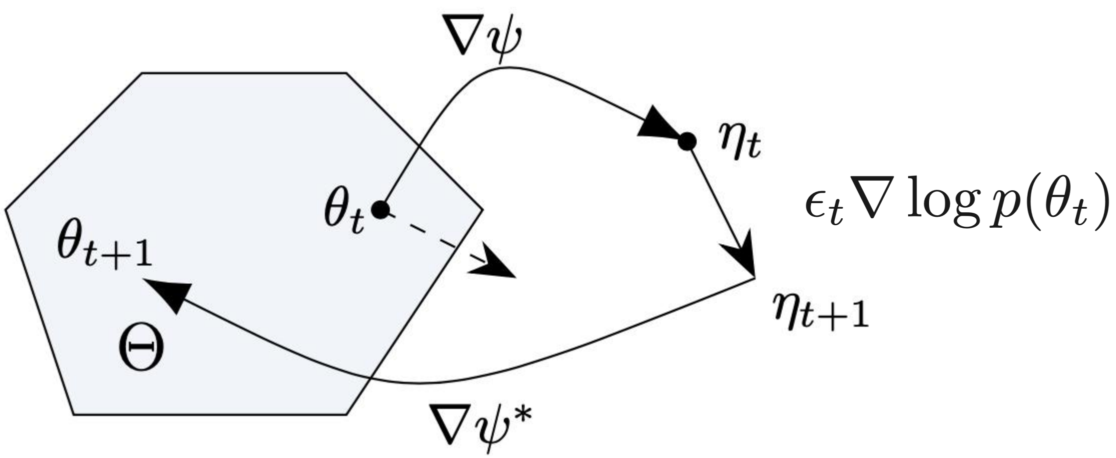
$$k^{1/2}(\theta, \theta') \triangleq \sum_{i \geq 1} \lambda_i^{1/2} u_i(\theta) u_i(\theta')$$

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

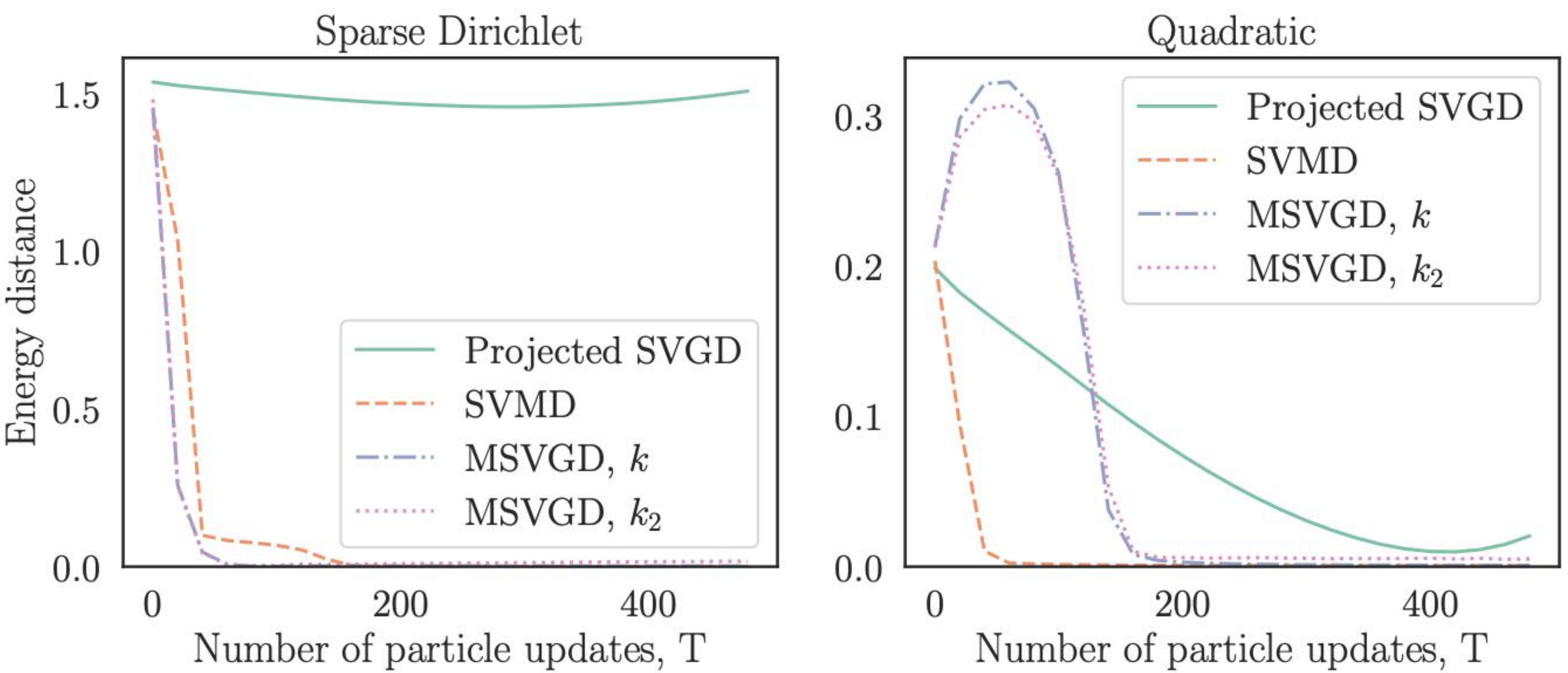
SVMD is a Multi-Particle Generalization of Mirror Descent

Proposition 5 If $n = 1$, then one-step of SVMD becomes

$$\eta_{t+1} = \eta_t + \epsilon_t \left(k(\theta_t, \theta_t)\nabla \log p(\theta_t) + \nabla k(\theta_t, \theta_t)\right),$$
$$\theta_{t+1} = \nabla \psi^*(\eta_{t+1}).$$



Approximation Quality on the Simplex

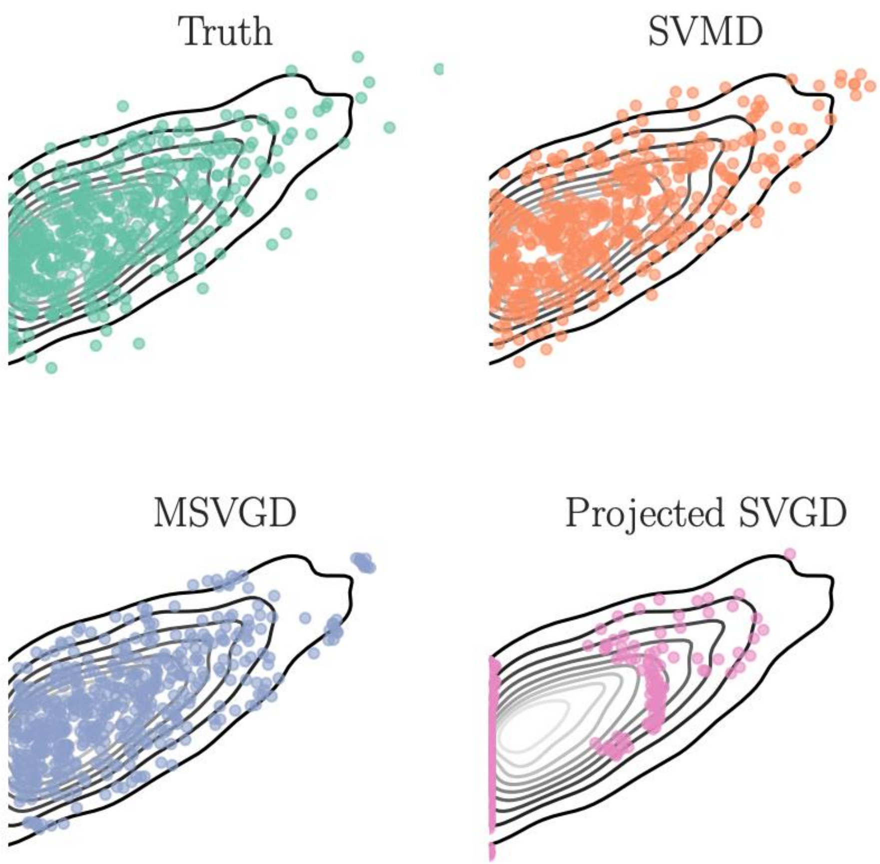


Quality of 50-particle approximations to 20-dimensional distributions on the simplex.

Application: Post-Selection Inference

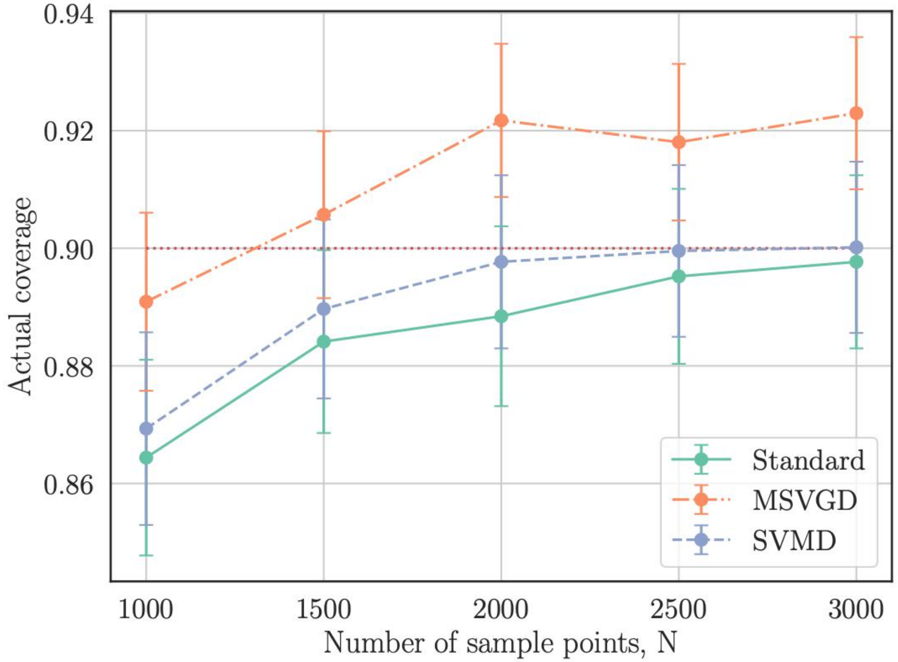
- Task: Generate valid confidence intervals (CIs) for regression parameters selected using the randomized LASSO
- Need to condition on the selection event
- Target distributions are log-concave and have constrained support

Application: Post-Selection Inference

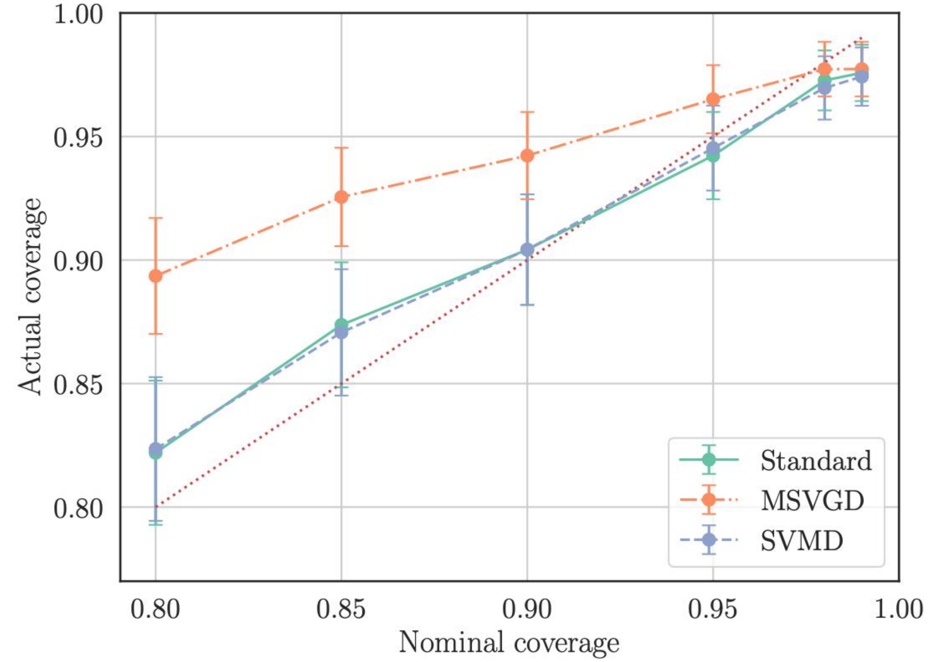


A 2D selective density example.

Application: Post-Selection Inference



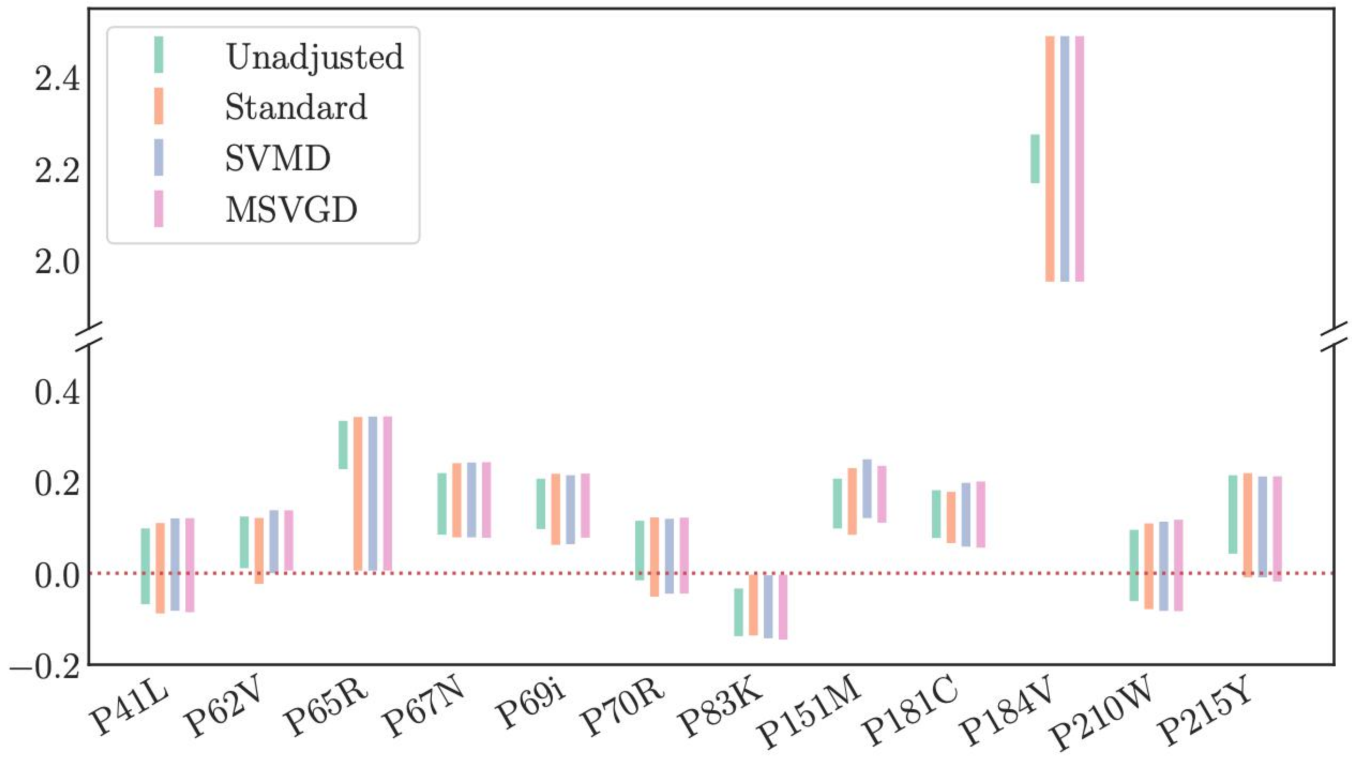
Nominal Coverage: 0.9



5000 sample points

Coverage of post-selection CIs.

Application: Post-Selection Inference



Unadjusted and post-selection CIs for the mutations selected by the randomized Lasso as candidates for HIV-1 drug resistance.

From Constrained to Unconstrained Targets

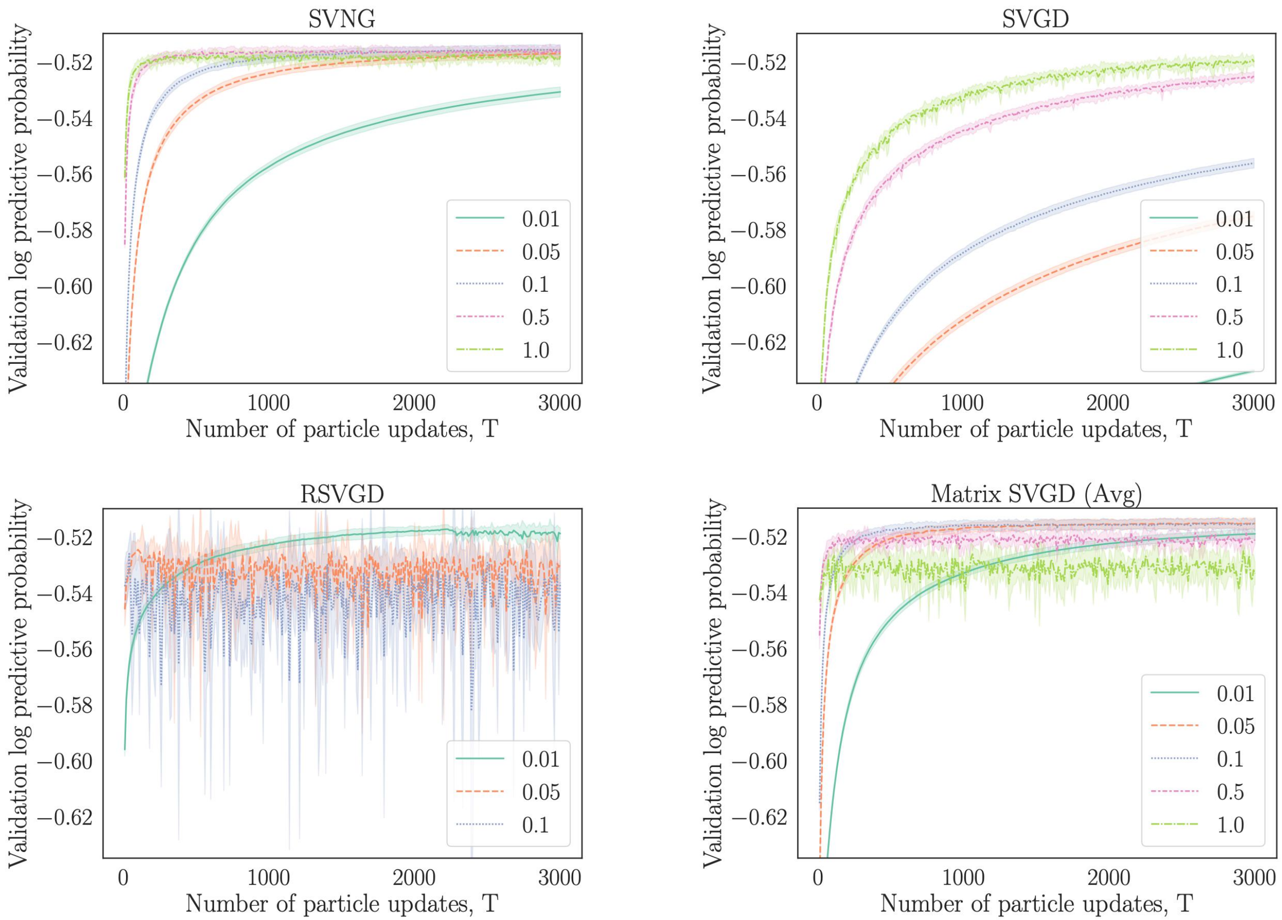
Continuous Time	Discretization
Mirror flow: $$d\eta_t = -\nabla f(\theta_t)dt,$$ $$\theta_t = \nabla\psi^*(\eta_t)$$	Mirror descent
Riemannian gradient flow with metric tensor $\nabla^2\psi$: $$d\theta_t = -\nabla^2\psi(\theta_t)^{-1}\nabla f(\theta_t)dt$$	Natural gradient descent with metric tensor $\nabla^2\psi$

Stein Variational Natural Gradient (SVNG)

- Replacing $\nabla^2\psi(\cdot)$ in SVMD with a general metric tensor $G(\cdot)$
- In Bayesian inference $p(\theta) \propto \pi(\theta)\pi(y|\theta)$, it is common to choose

$$\text{FIM: } G(\theta) = \mathbb{E}_{\pi(y|\theta)}[\nabla \log \pi(y|\theta)\nabla \log \pi(y|\theta)^\top]$$

Unconstrained Targets



Large-scale Bayesian Logistic Regression

$581,012$ datapoints, $d = 54$

Convergence Results

① Convergence of mirrored updates as $n \to \infty$.

② Infinite-particle mirrored Stein updates decrease KL with sufficiently small step size and drive Mirrored Kernel Stein Discrepancy (MKSD) to 0.

③ MKSD determines weak convergence under suitable conditions.

Conclusion

- We derive a new family of particle evolution samplers suitable for **constrained domains** and **non-Euclidean geometries**.
- SVMD is the first **multi-particle** generalization of mirror descent.
- SVNG is designed for unconstrained problems with informative metric tensors.

Future Work

- Reduce the $O(n^2)$ complexity of MSVGD via kernel approximation methods
- SVMD and SVNG are more costly than MSVGD due to the adaptive kernel construction
- Exponential convergence rates (Duncan et al., 2019, Korba et al., 2020, Chewi et al., 2020)

References

Liu, Q., & Wang, D. (2016). Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. Advances in Neural Information Processing Systems, 29, 2378-2386.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. Advances in Neural Information Processing Systems, 30, 3118-3126.

Gorham, J., & Mackey, L. (2015). Measuring Sample Quality with Stein's Method. Advances in Neural Information Processing Systems, 28, 226-234.

Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. In International Conference on Machine Learning (pp. 1292-1301).

Gorham, J., Raj, A., & Mackey, L. (2020). Stochastic Stein Discrepancies. Advances in Neural Information Processing Systems, 33, 17931-17942.

References

Murray, I. (2009). Markov chain Monte Carlo. Tutorial at Machine Learning Summer School, 2009

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

Korba, A., Salim, A., Arbel, M., Luise, G., & Gretton, A. (2020). A non-asymptotic analysis for Stein variational gradient descent. Advances in Neural Information Processing Systems, 33, 4672--4682.

Chewi, S., Gouic, T. L., Lu, C., Maunu, T., Rigollet, P., & Stromme, A. J. (2020). Exponential ergodicity of mirror-Langevin diffusions. arXiv preprint arXiv:2005.09669.