Accelerated Stein Variational Gradient Flow

Viktor Stein, Wuchen Li

TECHNISCHE UNIVERSITÄT BERLIN

Goal: Task: sample density $\pi \propto e^{-V}$ with *potential* $V\colon \mathbb{R}^d \to \mathbb{R}$, e.g., for generative modeling and Bayesian inference.
Contribution. Novel interacting particle algorithm converging faster than SVGD with deterministic behavior, without need for score estimation, e.g. by kernel density estimation (KDE).

Stein Variational Gradient Descent (SVGD)

Liu & Wang, NeurIPS'16

$K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **symmetric, positive definite**, **differentiable** "kernel".
SVGD update in the n-th iteration with step size $\tau > 0$:

$$x_i^{n+1} \leftarrow x_i^n - \frac{\tau}{N}\Bigg(\underbrace{\sum_{j=1}^N K(x_i^n, x_j^n)\nabla V(x_j^n)}_{\text{attraction}} - \underbrace{\sum_{j=1}^N \nabla_2 K(x_i^n, x_j^n)}_{\text{repulsion}}\Bigg), \quad i \in \{1,\dots,N\}.$$

$\rightsquigarrow$ SVGD is gradient descent of $\mathrm{KL}(\cdot \mid \pi)$ with respect to a "Stein geometry" (induced by K) when replacing ρ by a sample average, where $\mathrm{KL}(\rho \mid \pi) \coloneqq \int_{\mathbb{R}^d} \rho(x) \log\left(\frac{\rho(x)}{\pi(x)}\right) \mathrm{d}x$.
We focus on: generalized bilinear kernel: $K(x,y) = x^\mathsf{T} A y + 1$, $A \in \mathrm{Sym}_+(d)$, Gauss kernel $K(x,y) \coloneqq \exp\left(-\frac{1}{2\sigma^2}\|x-y\|_2^2\right)$, $\sigma > 0$.

Nesterov's accelerated gradient descent

Nesterov, 1983

to find minimizer x^* of convex, differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ with L-Lipschitz gradient use gradient descent $x^{n+1} \leftarrow x^n - \tau\nabla f(x^n)$ with step size $\tau \in \left(0, \frac{2}{L}\right)$
converges linearly: $f(x^n) - f(x^*) \in O(n^{-1})$.
Nesterov's accelerated gradient descent:

$$\begin{cases} x^{n+1} = y^n - \tau\nabla f(y^n) & \text{gradient step} \\ y^{n+1} = x^{n+1} + \alpha_{n+1}(x^{n+1} - x^n) & \text{momentum step} \end{cases}$$

converges **quadratically**: $f(x^n) - f(x^*) \in O(n^{-2})$.
Damping: $\alpha_n = \frac{n-1}{n+2}$ or $\alpha_n = \frac{\sqrt{L}-\sqrt{\beta}}{\sqrt{L}+\sqrt{\beta}}$ if f is β-strongly convex.
Then, $f(x^n) - f(x^*) \in O\left(e^{-n\sqrt{\frac{L}{\beta}}}\right)$ vs. $O\left(\left(1 - \frac{\beta}{L}\right)^n\right)$ for gradient descent.

The density manifold

Lafferty, 1988; Otto, 2001

The set of **smooth positive probability densities**

$$\widetilde{\mathcal{P}}(\Omega) \coloneqq \left\{\rho \in \mathcal{C}^\infty(\Omega) : \rho(x) > 0 \ \forall x \in \Omega, \int_\Omega \rho(x)\,\mathrm{d}x = 1\right\}.$$

forms an **infinite-dimensional C^∞-Fréchet manifold** (under regularity assumptions on Ω) with $T_\rho\widetilde{\mathcal{P}}(\Omega) \coloneqq \left\{\sigma \in \mathcal{C}^\infty(\Omega) : \int_\Omega \sigma(x)\,\mathrm{d}x = 0\right\}$ and $T_\rho^*\widetilde{\mathcal{P}}(\Omega) \coloneqq \mathcal{C}^\infty(\Omega)/\mathbb{R}$.
Metric tensor field *on* $\widetilde{\mathcal{P}}(\Omega)$ = smooth map $G\colon \widetilde{\mathcal{P}}(\Omega) \ni \rho \mapsto G_\rho\colon T_\rho\widetilde{\mathcal{P}}(\Omega) \leftrightarrow T_\rho^*\widetilde{\mathcal{P}}(\Omega)$.

Gradient flows in the density manifold

If it exists, the **first functional derivative of** $E\colon \widetilde{\mathcal{P}}(\Omega) \to \mathbb{R}$ is $\delta E\colon \widetilde{\mathcal{P}}(\Omega) \to \mathcal{C}^\infty(\Omega)/\mathbb{R}$ with

$$\langle \delta E(\rho), \phi\rangle_{L^2(\Omega)} = \partial_t\Big|_{t=0} E(\rho + t\phi), \quad \forall \phi \in \mathcal{C}^\infty(\Omega) : \rho + t\phi \in \widetilde{\mathcal{P}}(\Omega), |t| \text{ small}.$$

A smooth curve $\rho\colon [0,\infty) \to \widetilde{\mathcal{P}}(\Omega), t \mapsto \rho_t$ is a $(\widetilde{\mathcal{P}}(\Omega), G)$**-gradient flow of** E starting at $\rho(0)$ if

$$\partial_t\rho_t = -G_{\rho_t}^{-1}[\delta E(\rho_t)], \qquad \forall t > 0.$$

Stein metric defined via:

$$(G_\rho^{(K)})^{-1}(\Phi) \coloneqq \left(x \mapsto -\nabla_x \cdot \left(\rho(x)\int_\Omega K(x,y)\rho(y)\nabla\Phi(y)\,\mathrm{d}y\right)\right).$$

The $(\widetilde{\mathcal{P}}(\Omega), G^{(K)})$-gradient flow of $E = \mathrm{KL}(\cdot \mid Z^{-1}e^{-V})$ is

$$\partial_t\rho_t(x) = \nabla_x \cdot \left(\rho_t(x)\int_\Omega (K(x,y)\nabla V(y) - \nabla_2 K(x,y))\,\rho_t(y)\,\mathrm{d}y\right).$$

Accelerated Stein variational gradient flow on densities

The Hamiltonian on $\widetilde{\mathcal{P}}(\Omega)$ is

$$H\colon T\widetilde{\mathcal{P}}(\Omega) \to \mathbb{R}\cup\{\infty\}, \qquad (\rho,\Phi) \mapsto \frac{1}{2}\int_\Omega \Phi(x) G_\rho^{-1}[\Phi](x)\,\mathrm{d}x + E(\rho).$$

$\rho \triangleq$ position, $\Phi \triangleq$ momentum.
$\rightsquigarrow$ accelerated Stein variational gradient flow is

$$\begin{cases} \partial_t\rho_t + \nabla\cdot\left(\rho_t\int_\Omega K(\cdot,y)\rho_t(y)\nabla\Phi_t(y)\,\mathrm{d}y\right) = 0, \\ \partial_t\Phi_t + \alpha_t\Phi_t + \int_\Omega K(y,\cdot)\langle\nabla\Phi_t(y), \nabla\Phi_t(\cdot)\rangle\rho_t(y)\,\mathrm{d}y + \delta E(\rho_t) = 0. \end{cases}$$

Discretization of ASVGD

Use particle momentum $Y\colon (0,\infty) \to \mathbb{R}^d, t \mapsto \dot{X}_t$,

$$\dot{X}_t = Y_t = \int_\Omega K(X_t,y)\nabla\Phi_t(y)\rho_t(y)\,\mathrm{d}y.$$

Key idea of SVGD, using **integration by parts** to shift derivative from density ρ to the kernel K, can also be applied in the accelerated scheme:

Lemma (ASVG flow with particles' momenta)

The associated deterministic interacting particle system is

$$\begin{cases} \dot{X}_t = Y_t, \\ \dot{Y}_t = -\alpha_t Y_t + \int_\Omega (K(X_t,y)\nabla V(y) - \nabla_2 K(X_t,y))\,\rho_t(y)\,\mathrm{d}y + \int_{\Omega^2}\langle\nabla\Phi_t(z), \nabla\Phi_t(y)\rangle \\ \qquad\qquad \cdot\Big[K(y,z)(\nabla_2 K)(X_t,y) + K(X_t,z)(\nabla_1 K)(X_t,y) - K(X_t,y)(\nabla_2 K)(z,y)\Big]\rho_t(y)\,\mathrm{d}y\,\rho_t(z)\,\mathrm{d}z. \end{cases}$$

Then, replace expectation w.r.t. ρ_t by sample averages.

Accelerated Stein variational gradient descent algorithm

Data: Number of particles $N \in \mathbb{N}$, number of steps $T \in \mathbb{N}$, step sizes $\tau > 0$, target score function $\nabla V\colon \mathbb{R}^d \to \mathbb{R}^d$. Either a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ for bilinear kernel or a bandwidth $\sigma^2 > 0$ for Gaussian kernel, regularization parameter $\varepsilon \geq 0$.
Result: Matrix $X^{T_{\max}}$, whose rows are particles that approximate the target distribution $\pi \sim \exp(-V)$.
Step 0. Initialize $Y^0 = 0 \in \mathbb{R}^{N\times d}$. .
for $k = 0,\dots,T_{\max}-1$ **do**
Step 1. Update particle positions using particle momenta.

$$X^{k+1} \leftarrow X^k + \sqrt{\tau}Y^k.$$

Step 2. Form kernel matrix and update momentum in density space.

$$K^{k+1} = \left(K(X_i^{k+1}, X_j^{k+1})\right)_{i,j=1}^N, \qquad M^{k+1} \leftarrow N(K^{k+1} + \varepsilon\,\mathrm{id}_N)^{-1}Y^k.$$

Step 3. Update damping parameter using speed restart and/or gradient restart for each particle.
Step 4. Update momenta.
For bilinear kernel:

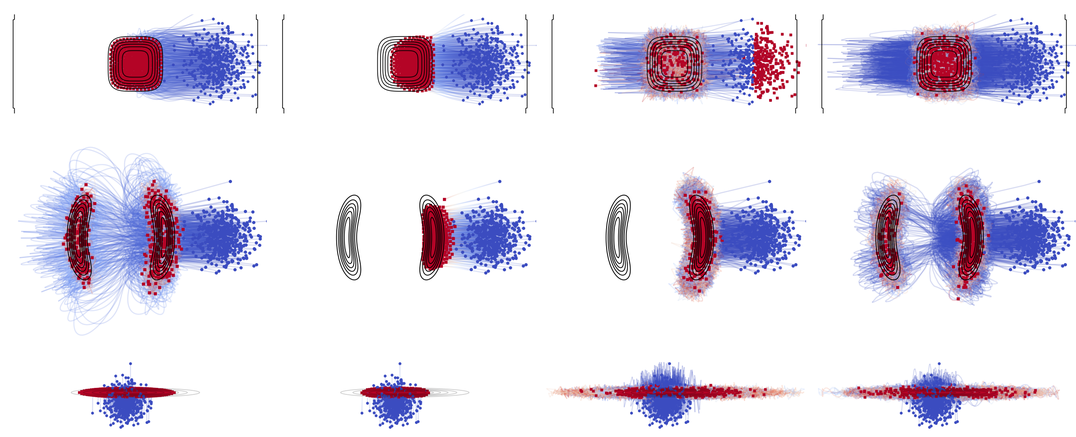
$$Y^{k+1} \leftarrow \alpha^k Y^k - \frac{\sqrt{\tau}}{N}K^{k+1}\nabla V(X^{k+1}) + \sqrt{\tau}\left(1 + N^{-2}\,\mathrm{tr}\left((M^{k+1})^\mathsf{T}K^{k+1}M^k\right)\right)X^{k+1}A.$$

For Gaussian kernel:

$$W^{k+1} \leftarrow NK^{k+1} + K^{k+1}(M^{k+1}(M^{k+1})^\mathsf{T}) \circ K^{k+1} - K^{k+1} \circ (K^{k+1}M^{k+1}(M^{k+1})^\mathsf{T}),$$
$$Y^{k+1} \leftarrow \alpha^k Y^k - \frac{\sqrt{\tau}}{N}K^{k+1}\nabla V(X^{k+1}) + \frac{\sqrt{\tau}}{2N^2\sigma^2}\left(\mathrm{diag}(W^{k+1}\mathbb{1}_N) - W^{k+1}\right)X^{k+1}.$$

end

Numerical Experiments



Particle trajectories. ASVGD, SVGD, with Gaussian kernel, MALA, and ULD (from left to right) for $V(x,y) = \frac{1}{4}(x^4 + y^4)$ (convex, non-Lipschitz) (top), the double bananas target (middle) and an anisotropic Gaussian target (bottom).
Double bananas target: constant high damping $\beta = 0.985$. Other targets: speed restart and gradient restart.

	RMSE		Log-likelihood		time (seconds)	
Dataset	ASVGD	SVGD	ASVGD	SVGD	ASVGD	SVGD
Concrete	$\mathbf{5.536}_{\pm\mathbf{0.060}}$	$7.349_{\pm 0.067}$	$\mathbf{-3.135}_{\pm\mathbf{0.016}}$	$-3.439_{\pm 0.010}$	$14.867_{\pm 0.040}$	$\mathbf{14.001}_{\pm\mathbf{0.044}}$
Energy	$\mathbf{0.899}_{\pm\mathbf{0.057}}$	$1.950_{\pm 0.028}$	$\mathbf{-1.268}_{\pm\mathbf{0.068}}$	$-2.088_{\pm 0.016}$	$14.870_{\pm 0.051}$	$\mathbf{13.942}_{\pm\mathbf{0.035}}$
Housing	$\mathbf{2.346}_{\pm\mathbf{0.077}}$	$2.386_{\pm 0.048}$	$\mathbf{-2.305}_{\pm\mathbf{0.020}}$	$-2.343_{\pm 0.014}$	$18.278_{\pm 0.045}$	$\mathbf{17.214}_{\pm\mathbf{0.077}}$
Kin8mn	$\mathbf{0.118}_{\pm\mathbf{0.001}}$	$0.165_{\pm 0.001}$	$\mathbf{0.71}_{\pm\mathbf{0.011}}$	$0.384_{\pm 0.004}$	$14.859_{\pm 0.029}$	$\mathbf{14.001}_{\pm\mathbf{0.033}}$
Naval	$\mathbf{0.005}_{\pm\mathbf{0.0}}$	$0.007_{\pm 0.0}$	$\mathbf{3.801}_{\pm\mathbf{0.01}}$	$3.504_{\pm 0.003}$	$20.912_{\pm 0.57}$	$\mathbf{19.704}_{\pm\mathbf{0.05}}$
power	$\mathbf{3.951}_{\pm\mathbf{0.005}}$	$4.035_{\pm 0.008}$	$\mathbf{-2.799}_{\pm\mathbf{0.001}}$	$-2.825_{\pm 0.002}$	$12.142_{\pm 0.053}$	$\mathbf{11.435}_{\pm\mathbf{0.054}}$
protein	$\mathbf{4.777}_{\pm\mathbf{0.007}}$	$4.987_{\pm 0.009}$	$\mathbf{-2.983}_{\pm\mathbf{0.001}}$	$-3.026_{\pm 0.002}$	$15.616_{\pm 0.04}$	$\mathbf{14.752}_{\pm\mathbf{0.047}}$
wine	$\mathbf{0.185}_{\pm\mathbf{0.013}}$	$0.191_{\pm 0.017}$	$\mathbf{0.201}_{\pm\mathbf{0.039}}$	$0.146_{\pm 0.046}$	$18.293_{\pm 0.097}$	$\mathbf{17.244}_{\pm\mathbf{0.077}}$

Table 1. Bayesian neural network experiment: test root mean square error (RMSE, lower is better) and log-likelihood (LL, higher is better) after 2000 iterations, without restarts and constant damping $\beta = 0.95$, with only 10 particles.

Further work

Find **best kernel parameters** A and σ^2 and best damping parameters (nearly done!). Conformal **symplectic** discretization instead of explicit Euler (retains structure of continuous dynamics), investigate **bias** of the algorithm, incorporate **annealing** strategy, (finite particle) convergence guarantees.

Ackknowledgements

V. Stein thanks his advisor, Gabi Steidl for her support and guidance throughout. W. Li's work is supported by the AFOSR YIP award No. FA9550-23-10087, NSF RTG: 2038080, and NSF DMS-2245097.