A Stein Variational Newton (SVN) method

Gianluca Detommaso[1,2], Tiangang Cui[3], Alessio Spantini[4], Youssef Marzouk[4], Robert Scheichl[1,5]

[1]: University of Bath, [2]: The Alan Turing Institute, [3]: Monash University, [4]: MIT, [5]: Heidelberg University

A mathematical description

INGREDIENTS

- $\pi(\cdot)$: posterior distribution
- $p(\cdot)$: probability density of the particles
- $k(\cdot,\cdot)$: kernel of a Reproducing Kernel Hilbert Space (RKHS)
- $\mathcal{D}_{KL}(\cdot||\cdot)$: Kullback-Leibler (KL) divergence
- T_*p: pushforward map of density p

METHOD

- We want to find a transport map T that minimizes the KL divergence between T_*p, the pushforward map of p, and π
- Construct T as a composition of simple maps $T^{(\ell)}$ such that
$$T =: (T^{(1)} \circ T^{(2)} \circ \cdots \circ T^{(n)})_* p \xrightarrow{n\to\infty} \pi \quad \text{in distribution}$$
- Take $T^{(\ell)}$ to be a small perturbation of the identity map:
$$T_*^{(\ell)} = (I + \varepsilon Q^{(\ell)})_* \qquad \text{with } Q^{(\ell)} \in \text{RKHS}$$
- Define
$$J_\ell[Q] := \mathcal{D}_{KL}((I + \varepsilon Q)_* p_\ell \,||\, \pi)$$
- Take $Q^{(\ell)} \in$ RKHS such that $J_\ell[Q^{(\ell)}] < J[\mathbf{0}]$

Stein Variational Gradient Descent (SVGD) [Liu et al, NIPS 2016]

Choose $Q^{(\ell)}$ to be the gradient descent direction given by
$$Q^{(\ell)}(z) := -\nabla J_\ell[\mathbf{0}](z) = \mathbb{E}_{x\sim p_\ell}[\nabla_x \log \pi(x) k(x,z) + \nabla_x k(x,z)]$$

Stein Variational Newton (SVN)

Choose $Q^{(\ell)}$ to solve a Newton-like iteration with Hessian given by
$$H_\ell(y,z) := \mathbb{E}_{x\sim p_\ell}[\nabla_x^2 \log \pi(x) k(x,y) k(x,z) + \nabla_x k(x,y) \nabla_x k(x,z)^\top]$$

- Several possibilities: Newton, inexact Newton, Newton-CG, ...
- In practice, at every stage ℓ, update a set of particles $x_i^{(\ell)}$ in the directions $Q^{(\ell)}(x_i^{(\ell)})$ for $i = 1, \dots, N$
- Use Hessian information for automatically rescale the kernel at no extra cost:
$$k(x,z) = \exp(-(x-z)^\top M (x-z)), \qquad M \approx \mathbb{E}_{x\sim p}[\nabla^2 \log \pi(x)]$$

An intuitive description

- **WHAT IT DOES.** Sampling from a posterior density π
- **HOW IT DOES IT.** Transport a set of particles sequentially, from a reference density p to the posterior density π
- **WHY TO USE IT.** Fast convergence, deterministic, flexible and robust, simple to implement, embarransigly parallelizable

Main contributions and benefits

- Derivation of second-order information (Hessian)
- Access to Newton-like iterations, much faster convergence than gradient descent
- Automatic rescaling and reshaping of the kernel, more efficient particle spread
- Significantly improved scalability to high-dimensions

SCAN to watch SVN in action!

Watch video

Pseudo-algorithm: Stein Variational (block-diagonal) Newton

Input : Particles $\{x_i^{(\ell)}\}_{i=1}^N$ at stage ℓ; step size ε
Output: Particles $\{x_i^{\ell+1}\}_{i=1}^N$ at stage $\ell+1$

for $i = 1, 2, \dots, N$ **do**

Evaluate the gradient
$$-\nabla J_\ell[0](x_i^{(\ell)}) = \frac{1}{N}\sum_{j=1}^N \left[\nabla \log \pi(x_j^{(\ell)}) k(x_j^{(\ell)}, x_i^{(\ell)}) + \nabla_{x_j} k(x_i^{(\ell)}, x_j^{(\ell)})\right]$$

Evaluate the Hessian
$$H_\ell(x_i^{(\ell)}, x_i^{(\ell)}) = \frac{1}{N}\sum_{j=1}^N \left[H_\pi(x_j^{(\ell)}) k(x_j^{(\ell)}, x_i^{(\ell)})^2 + \nabla_{x_j} k(x_j^{(\ell)}, x_i^{(\ell)}) \nabla_{x_j} k(x_j^{(\ell)}, x_i^{(\ell)})^\top\right],$$
where H_π is Gauss-Newton approximation (positive-definite) of $\nabla^2 \log \pi$

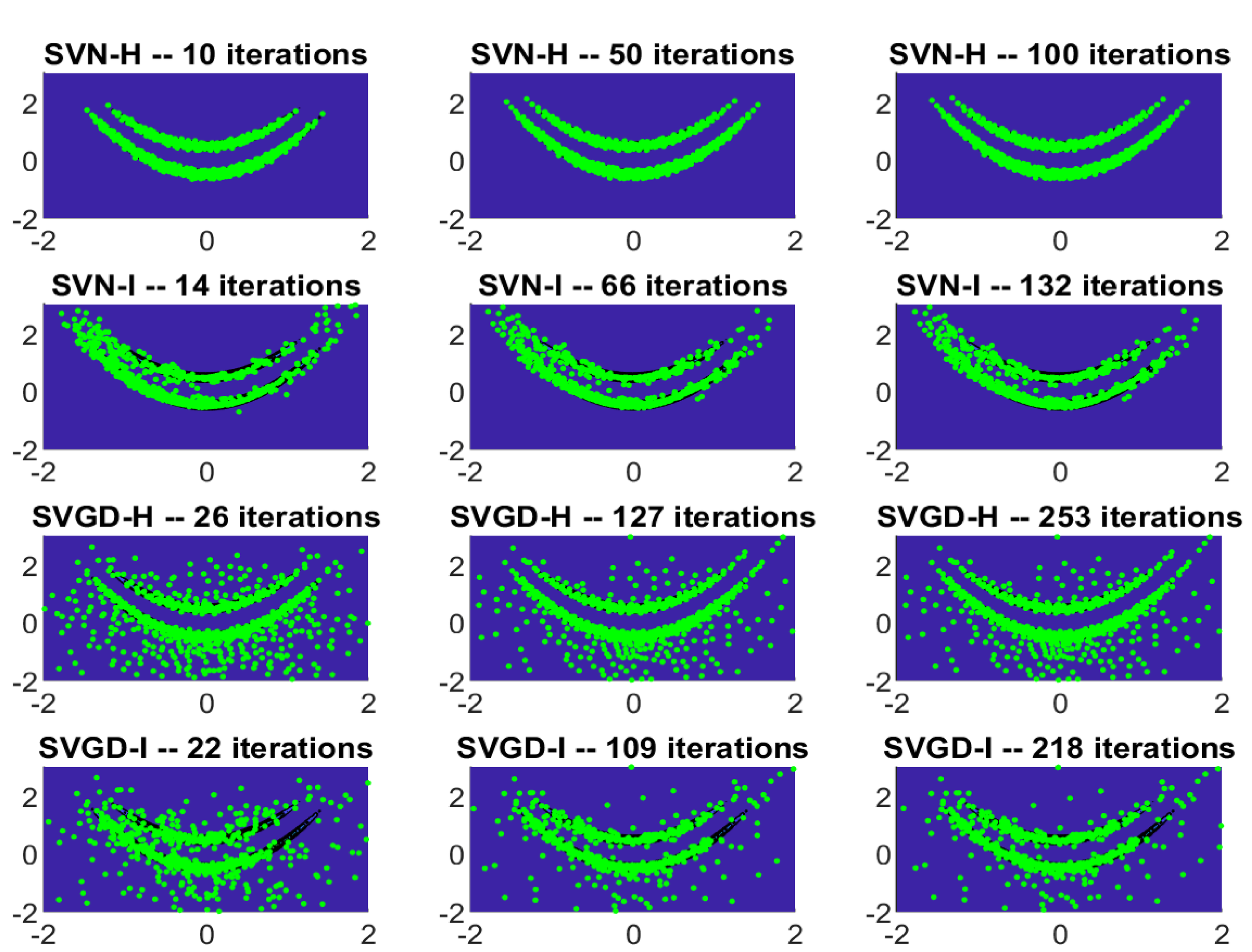
Solve the linear system
$$H_\ell(x_i^{(\ell)}, x_i^{(\ell)}) Q^{(\ell)}(x_i^{(\ell)}) = -\nabla J_\ell[0](x_i^{(\ell)})$$

Update the particle
$$x_i^{(\ell+1)} \leftarrow x_i^{(\ell)} + \varepsilon Q^{(\ell)}(x_i^{(\ell)})$$

end for

Test cases

2-dimensional double-banana



100-dimensional conditional diffusion