On the geometry of Stein variational gradient descent (SVGD)

Nikolas Nüsken

Universität Potsdam, CRC 1114

April 3, 2019

Joint work with **Andrew Duncan** (Imperial College London) and **Lukasz Szpruch** (University of Edinburgh).
This work was partially funded by the **Alan Turing Institute** in London.

Stein variational gradient descent (SVGD)

Consider the **interacting particle system**

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

where

- N number of particles,
- k positive definite *kernel*, e.g. $k(x,y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right)$,
- $V : \mathbb{R}^d \to \mathbb{R}$ is called the *potential*.

Fact (informal):

As $N \to \infty$ and $t \to \infty$, the distribution ρ of the particles approaches

$$\pi := \frac{1}{Z} e^{-V},$$

where Z is a normalisation constant.

Measure transport

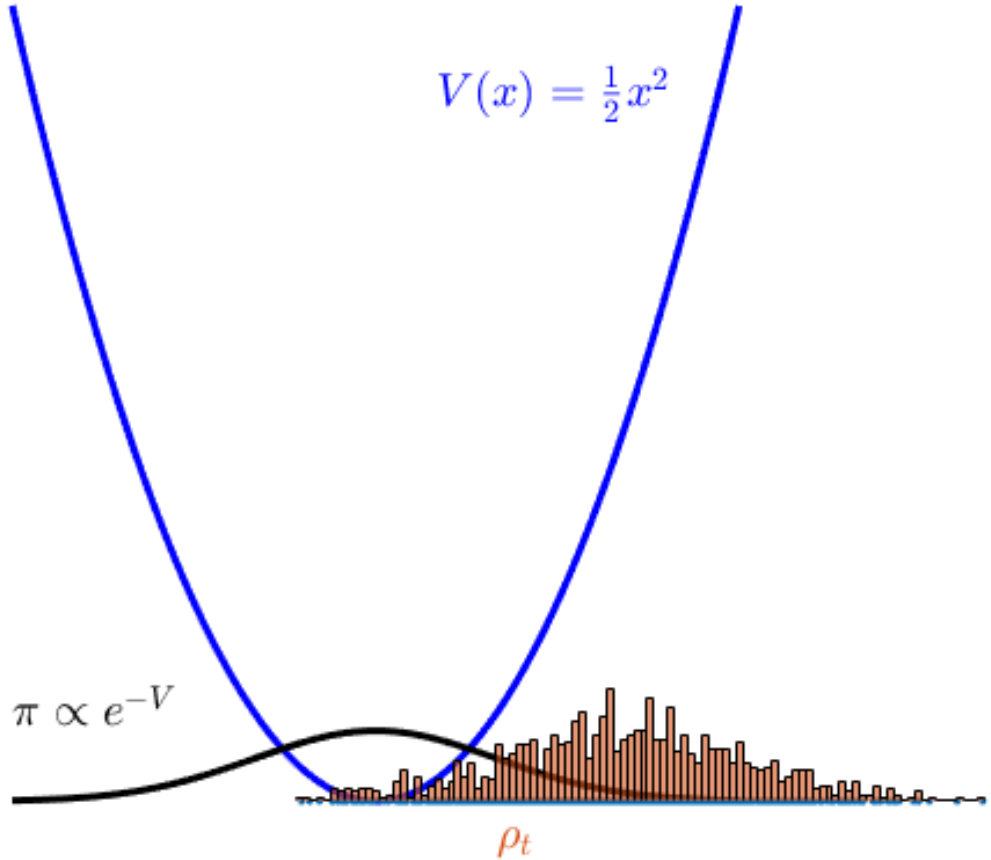
$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

convergence: $\rho_t \xrightarrow[t\to\infty]{N\to\infty} \pi := \frac{1}{Z}e^{-V}$

Why is this good? – Because...

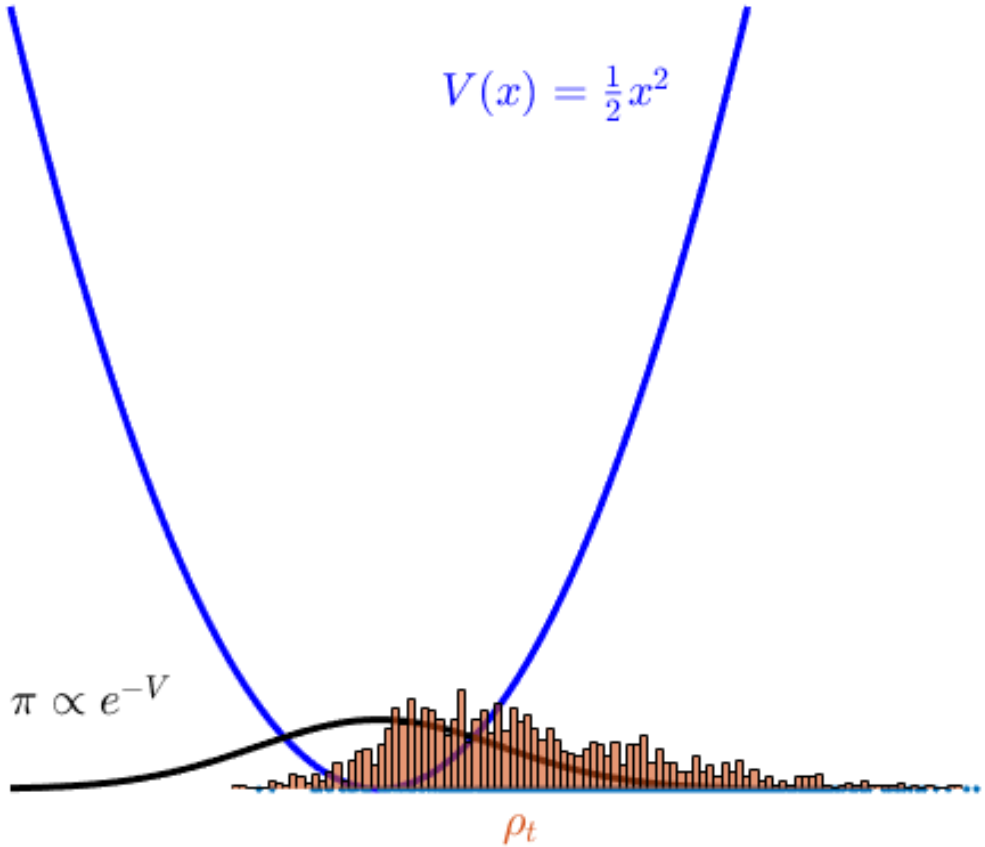
- ... by setting $V = -\log \pi$ we can approximate (expectations wrt.) π, having access only to $\nabla \log \pi$, without knowing Z.
- ... this is the typical setting in Bayesian inference (inverse problems, data assimilation, etc.).

$t = 0$



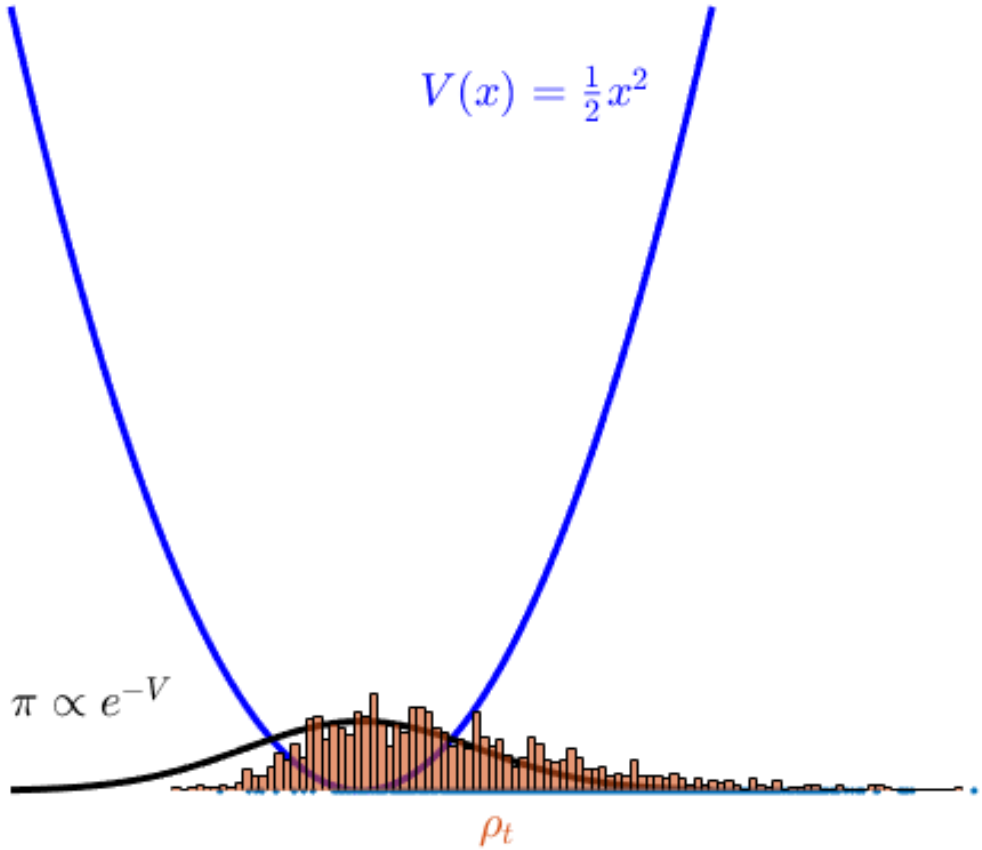
$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

$t = 1$



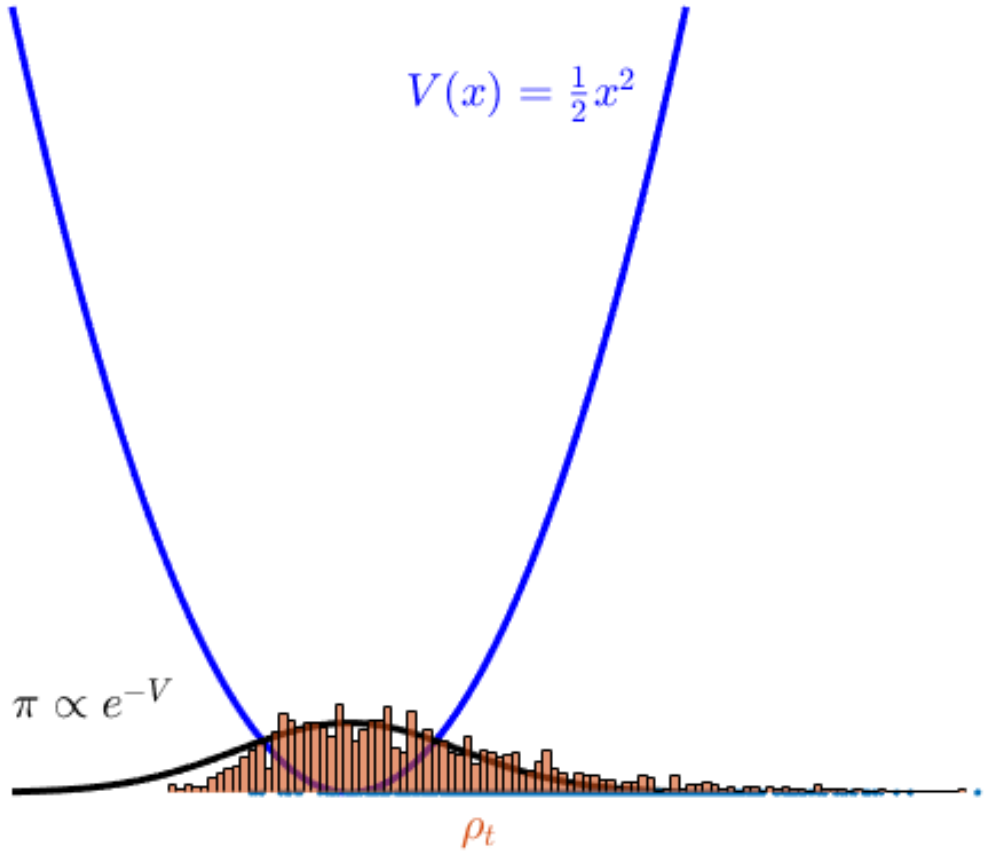
$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

$t = 2$



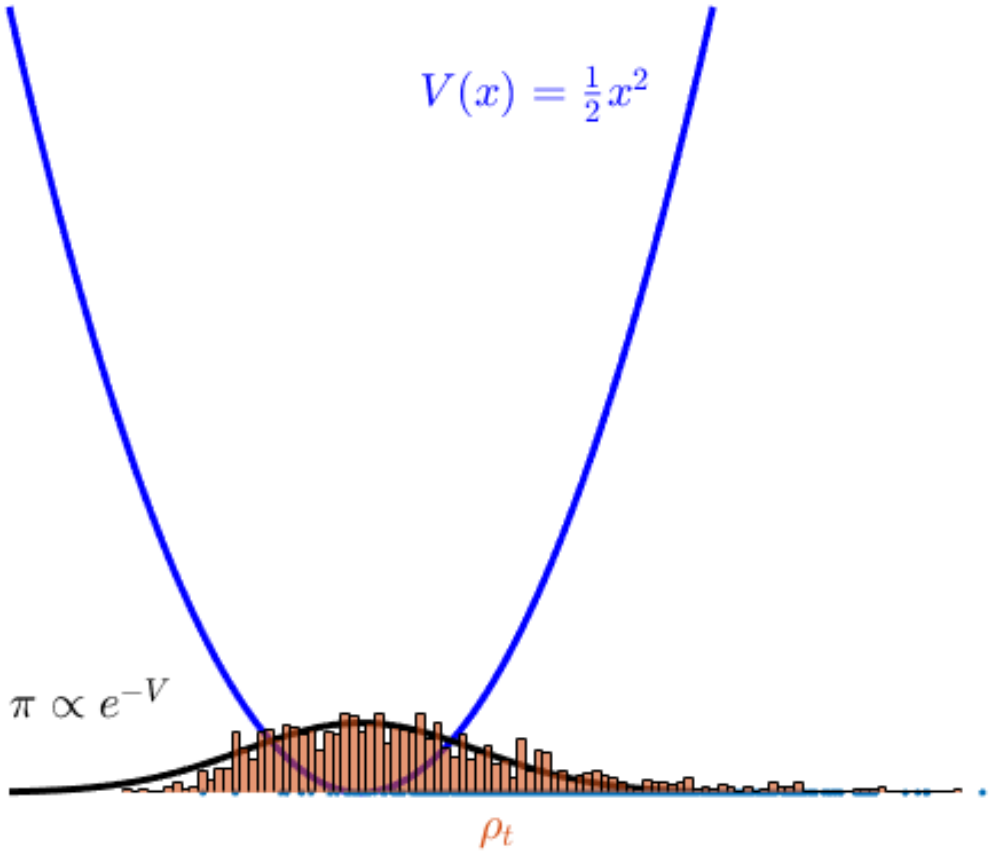
$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

$t = 3$



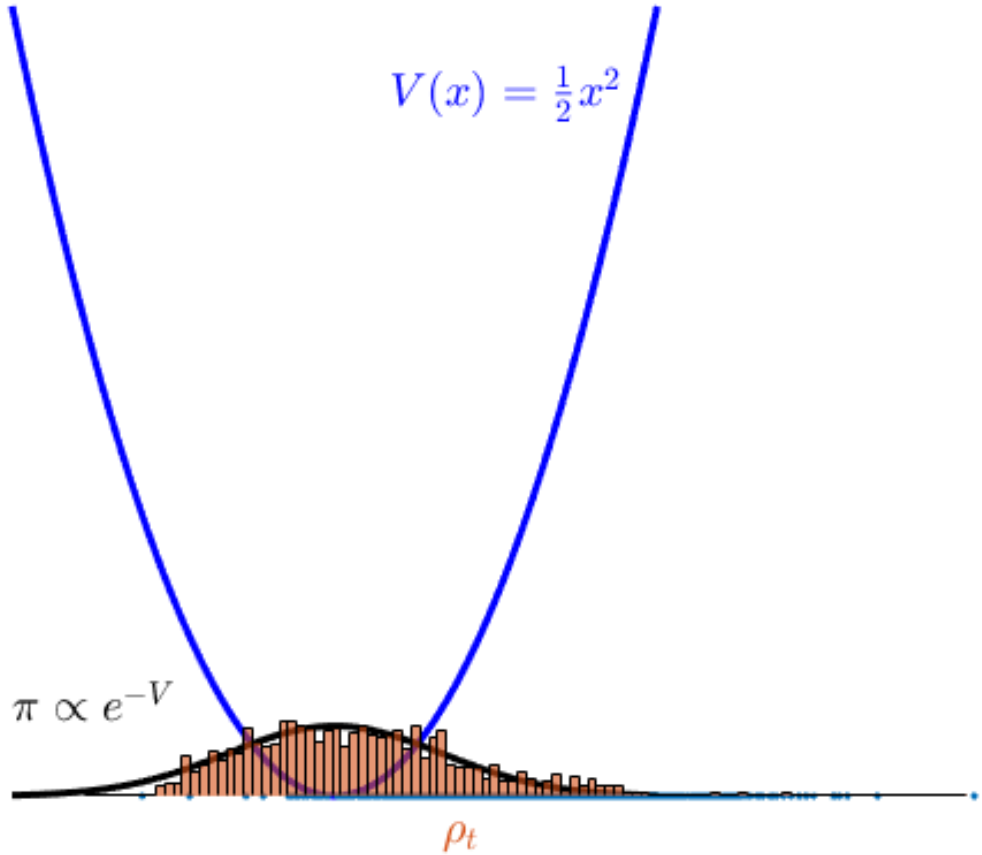
$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

$t = 4$



$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

$t = 5$



$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j),$$

Langevin

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$$

Fokker-Planck:

$$\partial_t \rho = \nabla \cdot (\rho \nabla V + \nabla \rho)$$

- noninteracting
- linear
- local
- stochastic[a]

[a] There are deterministic versions of Langevin.

Stein (SVGD)

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\frac{1}{N}\sum_{j=1}^{N} k(X_t^i, X_t^j)\nabla V(X_t^j) + \frac{1}{N}\sum_{j=1}^{N} \nabla_{X_t^j} k(X_t^i, X_t^j)$$

Stein pde:

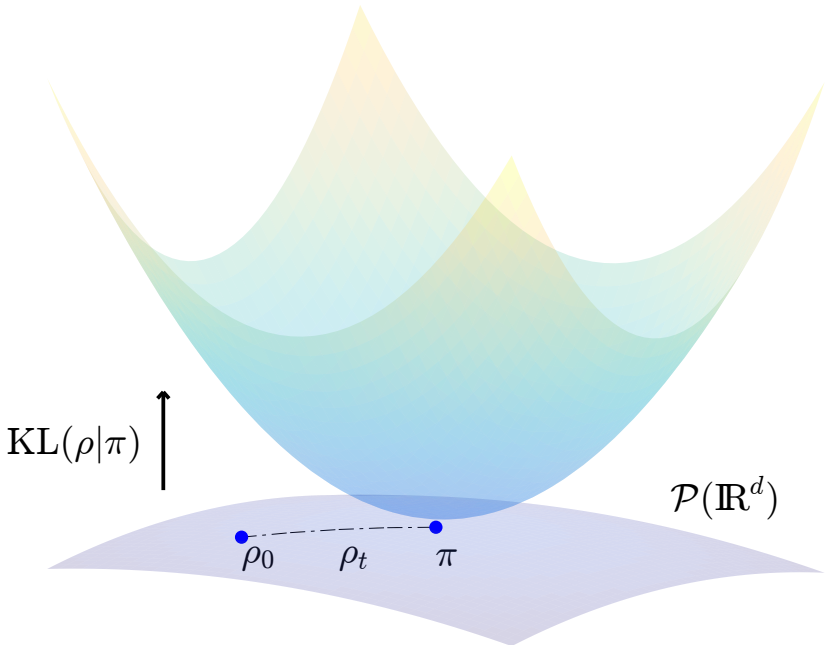
$$\partial_t \rho = \nabla \cdot (\rho\,(k * (\rho \nabla V + \nabla \rho)))$$

- interacting
- nonlinear
- nonlocal
- deterministic[a]

[a] There are stochastic versions of SVGD.

Gradient flows

$$\partial_t \rho_t = -\nabla_d \mathrm{KL}(\rho_t|\pi)$$



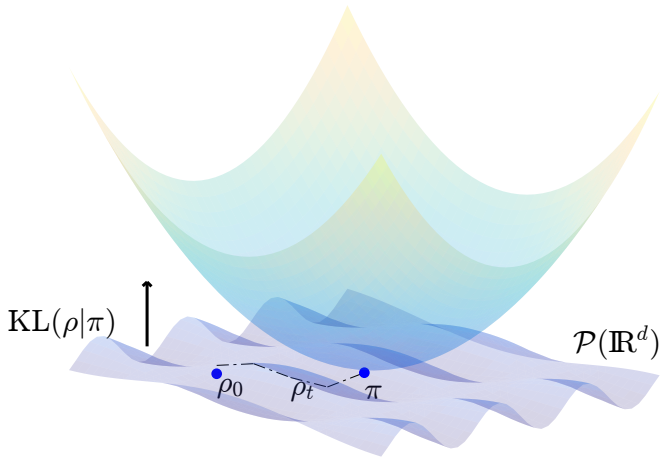
relative entropy/KL-divergence:

$$\begin{aligned}\mathrm{KL}(\rho|\pi) &= \int_{\mathbb{R}^d} \rho \log\left(\frac{\rho}{\pi}\right) \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \rho \log \rho \,\mathrm{d}x + \int_{\mathbb{R}^d} V \,\mathrm{d}\rho\end{aligned}$$

Both Langevin and Stein are gradient flows of KL.

Both Langevin and Stein are gradient flows of KL...

... but with respect to different geometries on $\mathcal{P}(\mathbb{R}^d)$.



Langevin:

$$d_{OT}^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)} \int_0^1 \|v_t\|_{L^2(\mu_t)}^2 \, \mathrm{d}t = W_2^2(\mu_0, \mu_1),$$

Stein:

$$d_k^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)} \int_0^1 \|v_t\|_{\mathcal{H}_k}^2 \, \mathrm{d}t,$$

both subject to

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0 \quad \text{(weakly)}.$$

Take-home message (recipe for sampling algorithms):

1. Choose a cost functional (here: KL),
2. Choose a geometry on $\mathcal{P}(\mathbb{R}^d)$,
3. Find a suitable simulation scheme for the ensuing gradient flow pde.

Second order: geodesic convexity and contraction rates

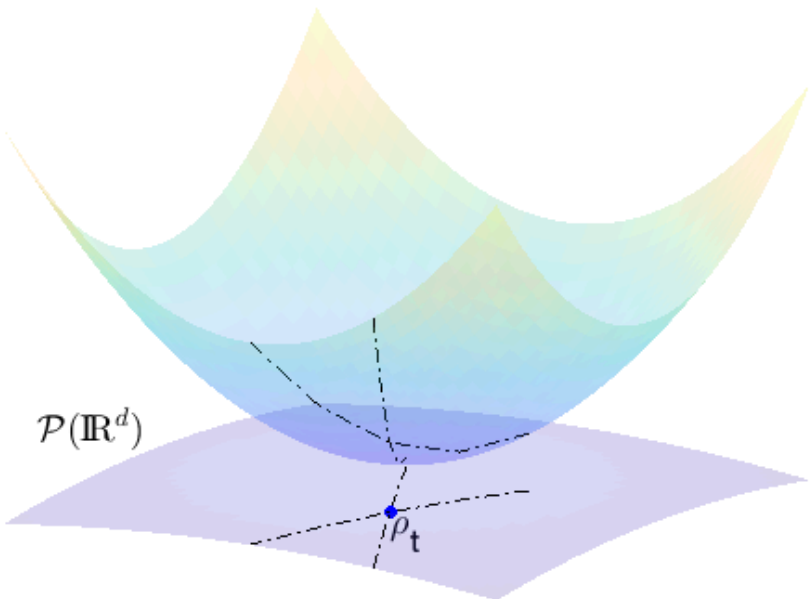
Theorem (Informal)

Assume that there exists $\lambda > 0$ *such that*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{KL}(\mu_t|\pi) > \lambda,$$

for all unit-speed geodesics $(\mu_t)_{t\in(-\varepsilon,\varepsilon)}$. *Then*

$$\mathrm{KL}(\rho_t|\pi) \leq e^{-2\lambda t}\mathrm{KL}(\rho_0|\pi).$$



Geodesic equations...

...for geodesics μ_t and their (generalised) velocity fields $\nabla\Psi_t$.

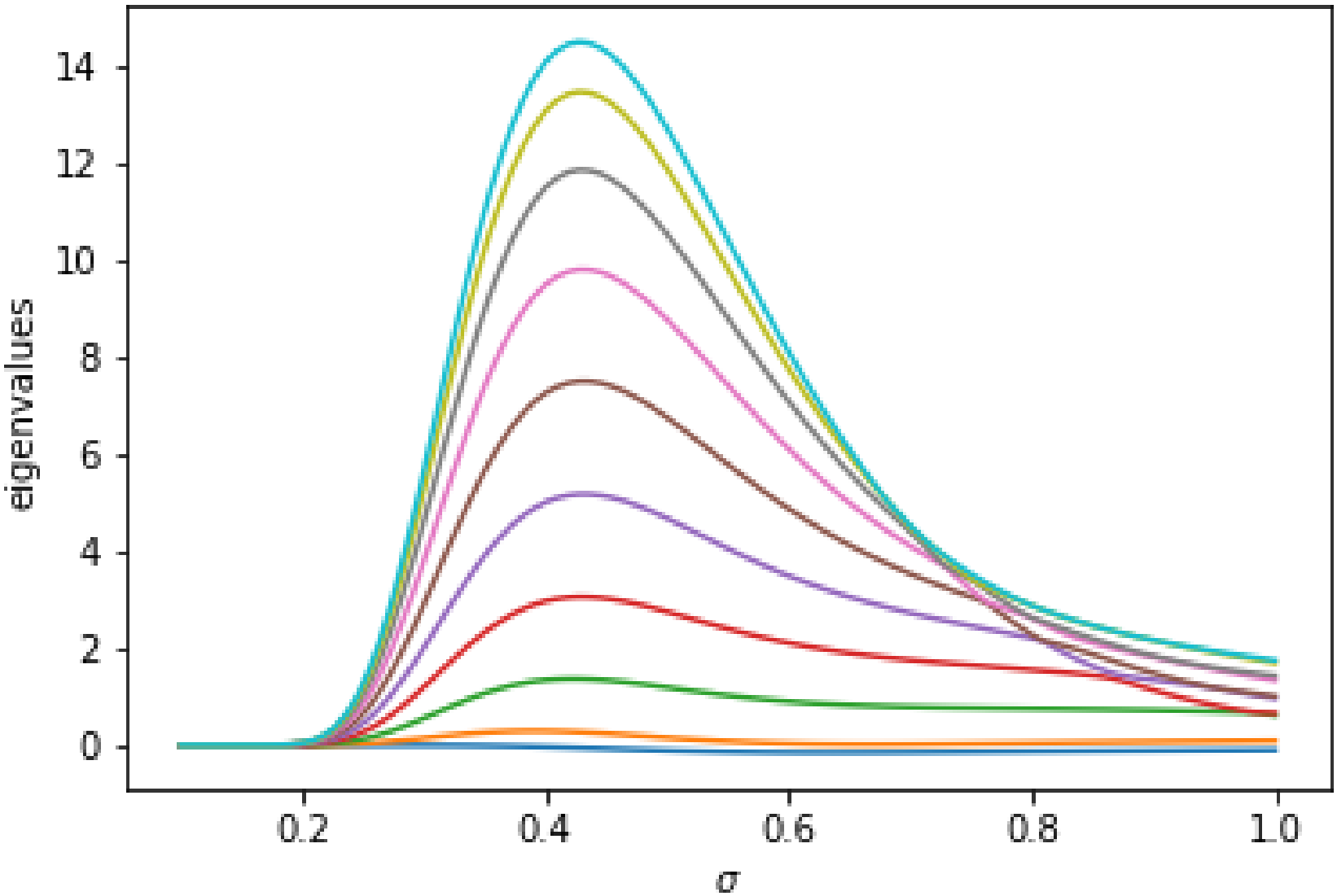
- Langevin (Wasserstein):

$$\partial_t \mu + \nabla \cdot (\mu \nabla \Psi) = 0,$$
$$\partial_t \Psi + \frac{1}{2}|\nabla \Psi|^2 = 0.$$

- Stein:

$$\partial_t \mu(x) + \nabla_x \cdot \left(\mu(x) \int_{\mathbb{R}^d} k(x,y) \nabla \Psi(y) \, \mathrm{d}\mu(y) \right) = 0,$$
$$\partial_t \Psi(x) + \nabla \Psi(x) \cdot \int_{\mathbb{R}^d} k(x,y) \, \nabla \Psi(y) \, \mathrm{d}\mu(y) = 0.$$

Curvature for a discrete measure, $V \equiv 0$



Conclusions:

- Probably there is no exponential decay for the Stein pde.
- The width of the kernel can (and should) be adjusted according to a 'mean curvature' criterion.

Future directions:

- Are there connections with the approximation theory in RKHS (bias-variance tradeoff, etc...)?
- Beyond gradient flows: Nesterov acceleration, Hamiltonian Monte Carlo, ...

Contact: nuesken@uni-potsdam.de