MATHÉMATIQUES
VISION
APPRENTISSAGE

Stein Variational Gradient Descent: Main ideas

Gaëtan Serré and Perceval Beja-Battais

École Normale Supérieure Paris-Saclay
Master Mathématiques, Vision, Apprentissage
{gaetan.serre, perceval.beja-battais}@ens-paris-saclay.fr

1 Introduction

In this paper we will contextualize and describe the article Korba et al. [2020]. The goal is to be able to sample from an unknown distribution π thanks to an iterative procedure that is called Stein Variational Gradient Descent (SVGD). It was first introduced by Liu and Wang [2016]. Starting from an initial distribution μ_0, this algorithm can be seen as a gradient descent in the Wasserstein space of distributions $\mathcal{P}_2(\mathcal{X})$, with $\mathcal{X} \in \mathbb{R}^d$. **TODO METTRE NOS CONTRIBS**

2 SVGD Context (Liu and Wang [2016])

In all what follows, $\mathcal{X} = \mathbb{R}^d$.
We fix here π an objective distributions from which we want to sample, and an initial distribution from which we know how to sample μ_0.

2.1 Notations

We will denote as earlier by $\mathcal{P}_2(\mathcal{X})$ the Wasserstein space of distributions i.e. the set of distributions such that $\int ||x||^2 d\mu(x) < \infty$. We assume the objective distribution π lives in $\mathcal{P}_2(\mathcal{X})$, and define the Kullback-Liebler divergence between π_1 and π_2 by

$$\mathrm{KL}(\pi_1||\pi_2) := \mathbb{E}_{\pi_1}[\log \pi_1(x)] - \mathbb{E}_{\pi_1}[\log \pi_2(x)]$$

Let $\mathcal{A}_\pi$ the Stein Operator defined by $\forall \phi \in \mathcal{H}, \forall x \in \mathcal{X}, \mathcal{A}_\pi \phi(x) = \nabla \log \pi(x) \phi(x)^\top + \nabla \cdot \phi(x)$, for some $\mathcal{H}$ we will precise later on.
We define the Stein class of π the subset of functions ϕ such that $\lim_{||x|| \longrightarrow \infty} \phi(x)\pi(x) = 0$. Note that for every function in the Stein class of π, we have

$$\mathbb{E}_{x \sim \pi}[\mathcal{A}_\pi \phi(x)] = 0 \quad (1)$$

(See A.1 for the proof)
Also, we define the pushforward measure of μ by $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ by $\int \phi(T(x)) d\mu(x) = \int \phi(x) dT_\# \mu(x)$ for any bounded and measurable function ϕ.
The gradient of $\mathrm{KL}(\cdot||\pi)$ at μ in $\mathcal{P}_2(\mathcal{X})$ is given by $\nabla_{W_2} \mathrm{KL}(\mu||\pi) = \nabla \log(\frac{\mu}{\pi})$. The idea behind SVGD is to iteratively follow the descent direction given by $\nabla_{W_2} \mathrm{KL}(\cdot||\pi)$.
Finally, for $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, we denote by $||\phi||_{op}$ the operator norm.

2.2 Context

Let $\mu \in \mathcal{P}_2(\mathcal{X})$. Given a smooth function $\phi = [\phi_1, ..., \phi_d]^\top$, a small perturbation of μ in the direction of ϕ is given by

$$\mu_{[T]} = \mu_\# (I + \gamma \phi) \quad (2)$$

for a small $\gamma > 0$.
Recall that $\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] = \int_{\mathcal{X}} \big(\nabla \log \pi(x)^\top \phi(x) + \nabla \cdot \phi(x) \big) \mu(x) \,\mathrm{d}x$. As soon as μ is not in the Stein class of π, one can show that $\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] > 0$, increasing w.r.t. the difference between μ and π (Proof here: A.3)..
Therefore, the problem we want to solve is to find

$$\mu^* = \arg\min_{\mu} \mathbb{S}(\mu, \pi) = \arg\min_{\mu} \max_{\phi \in \mathcal{H}} \{ \mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] \}, \quad (3)$$

for a certain class $\mathcal{H}$ of functionals. A question now raises: how to choose $\mathcal{H}$?

Choice of $\mathcal{H}$ In all papers, $\mathcal{H}$ is chosen as the product RKHS of $\mathcal{X}$. We will quickly explain here why so.
Let $\mathcal{H}_0$ be a RKHS with a kernel $k(x, x')$ in the Stein class of π. Let $\mathcal{H} = (\mathcal{H}_0^{(1)}, \dots, \mathcal{H}_0^{(d)})$. The KSD maximizes ϕ in the unit ball of $\mathcal{H}$. The objective in (3) is then:

$$\mathbb{S}(\mu, \pi) = \max_{\phi \in \mathcal{H}} \{\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi(x)],\ s.t.\ \|\phi\|_{\mathcal{H}} \leq 1\}. \tag{4}$$

Within this framework, one can show that the optimal solution of (4) is given by:

$$\phi(x) = \frac{\phi^*(x)}{\|\phi^*\|_{\mathcal{H}}} \text{ , where } \phi^*(.) = \mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \otimes k(x,\cdot)] = \int_{\mathcal{X}} (k(x,\cdot)\nabla \log \pi(x) + \nabla k(x,\cdot))\, \mathrm{d}\mu(x), \tag{5}$$

where $\mathcal{A}_\pi \otimes f(x) = f(x)\nabla \log \pi(x) + \nabla f(x)$, is a variant of Stein operator. We also know that ϕ^* is in the Stein class of π as k is. Moreover, $\mathbb{S}(\mu, \pi) = \|\phi^*\|_{\mathcal{H}}$. (Complete proof in A.2)

Return to the problem For $\gamma < \frac{1}{||\phi||_{op}}$, $(I + \gamma\phi)$ is locally one-to-one. We then have that

$$\nabla_\gamma \mathrm{KL}(T_\# \mu || \pi)|_{\gamma=0} = -\mathbb{E}_{x\sim\pi}[\mathcal{A}_\pi \phi(x)]$$

Considering all descent directions on the ball $\{\phi \in \mathcal{H}, ||\phi||^2_{op} \leq \mathbb{S}(\mu, \pi)\}$, the one we will keep for our gradient descent is the one minimizing the gradient of KL, which writes ϕ^* as showed just earlier.
The Stein Variational Gradient Descent (SVGD) algorithm consists in an iterative procedure where one apply successive transformations to an initial density μ_0 following the trajectory ϕ^* that minimizes the gradient of the Kullback-Leibler divergence:

$$\mu_{n+1} = (I + \gamma\phi^*)_\# \mu_n \tag{6}$$

3 Non-asymptotic analysis of SVGD

In their paper (Korba et al. [2020]), under assumptions, the authors provide an exponential convergence rate for continuous time SVGD, and a convergence result between SVGD in the infinite particle setting and in the finite particle setting. This last result is very important as the latter setting is the one used in practice when implementing the SVGD algorithm and allows to make a link between the implementation and the theoretical results. They also reprove a descent lemma for discrete time SVGD, originally proved in 2017 (Liu [2017]), using the Wassertein gradient flow of the KL divergence.

3.1 Optimal transport reminders

Before going further, we will recall some notions of optimal transport that the authors used throughout their paper.

Definition 1 (Wasserstein distance). *Let μ and ν be two probability measures on $\mathcal{X}$ and*

$$\Gamma(\mu, \nu) = \left\{\gamma : \int_{\mathcal{X}} \gamma(x, y)\, \mathrm{d}y = \mu(x) \wedge \int_{\mathcal{X}} \gamma(x, y)\, \mathrm{d}x = \nu(y)\right\}.$$

The p-Wasserstein distance between μ and ν is defined by

$$\mathbb{W}_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}} \int_{\mathcal{X}} \|x - y\|^p \gamma(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

This define a distance on the space of probability measures as it is positive, symmetric, 0 if and only if $\mu = \nu$ and satisfies the triangle inequality.

Definition 2 (Continuity equation Villani [2003]). *Let $\mathcal{X}$ be $\mathbb{R}^d$ and $(T_t)_{0\leq t}$ a measurable map from $\mathcal{X}$ to $\mathcal{X}$ such that $T_t = I + \phi(t)$. Let v_t be the velocity field associated with the trajectories T_t. Let $\mu_0 \in \mathcal{P}_2(\mathcal{X})$ and $\mu_{t+1} = T_t \# \mu_t$. Then, μ_t is the unique solution of the following continuity equation:*

$$\begin{cases} \frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0 \\ v_t = \phi(t). \end{cases}$$

3.2 RKHS operators

In the entire paper, the authors let $\mathcal{X}$ be $\mathbb{R}^d$. They defined the RKHS $\mathcal{H}$ and $\mathcal{H}_0$ on real-valued function of $\mathcal{X}$ the same way as in Section 2.2
They start by defining operators on the RKHS.

Definition 3. *Let $S_\mu : L_2(\mu) \to \mathcal{H}$ be the operator defined by:*

$$S_\mu f = \int_{\mathcal{X}} k(x, \cdot) f(x) \, \mathrm{d}\mu(x).$$

They also make the assumption that $\int_{\mathcal{X}} k(x,x) \, \mathrm{d}\mu(x) < \infty$, $\forall \mu \in \mathcal{P}_2(\mathcal{X})$, which implies $\mathcal{H} \subset L^2(\mu)$ (proof in A.4).

They also defined the inclusion $\iota : L^2(\mu) \to \mathcal{H}$ and its adjoint $\iota^* : \mathcal{H} \to L^2(\mu) = S_\mu$. Finally, they defined $P_\mu = \iota S_\mu$. Thanks to these operators, we now have that:

$$\langle f, \iota g \rangle_{L^2(\mu)} = \langle \iota^* f, g \rangle_{\mathcal{H}} = \langle S_\mu f, g \rangle_{\mathcal{H}}, \ \forall f, g \in L^2(\mu) \times \mathcal{H}.$$

This allows to use proprieties of the scalar product of $\mathcal{H}$ for functions defined in $L^2(\mu)$ and to show that, if k is also in the Stein class of μ (see A.5):

$$P_\mu \nabla \log \frac{\mu}{\pi}(\cdot) = -\phi^*(\cdot). \tag{7}$$

3.3 Convergence of rates for continuous time SVGD

Definition 4 (Stein Fisher information). *Let $\mu \in \mathcal{P}_2(\mathcal{X})$. The Stein Fisher information of μ relative to π is defined as follows:*

$$I_{Stein}(\mu|\pi) = \left\| S_\mu \nabla \log \frac{\mu}{\pi} \right\|_{\mathcal{H}}^2.$$

Note that $I_{Stein}(\mu|\pi)$ is the square of the optimum value of the Kernelized Stein Discrepancy defined in (5).

The authors proved the following proposition:

Proposition 1. *The time-derivative (or dissipation) of the* KL *divergence between μ_t and π is*

$$\frac{\partial \mathrm{KL}(\mu_t || \pi)}{\partial t} = -I_{Stein}(\mu_t|\pi).$$

We provide a more complete proof in A.6.

Using this proposition, the authors proved the following convergence rate for the average of $I_{Stein}(\mu|\pi)$ over time:

$$\forall t, \min_{0 \leq s \leq t} I_{Stein}(\mu_s|\pi) \leq \frac{1}{t} \int_0^t I_{Stein}(\mu_s|\pi) \, \mathrm{d}s \leq \frac{\mathrm{KL}(\mu_0||\pi)}{t}. \tag{8}$$

(It can be easily shown by integrating A.6). However, for the convergence of $I_{Stein}(\mu_t|\pi)$ to be fast, π must satisfy the Stein log-Sobolev inequality:

Definition 5 (Stein log-Sobolev inequality). *Let $\lambda > 0$. We say π satisfies the Stein log-Sobolev inequality if:*

$$\mathrm{KL}(\mu||\pi) \leq \frac{1}{2\lambda} I_{Stein}(\mu|\pi).$$

This inequality holds if, for example, π has exponential tails and the derivative of k increases at most at a polynomial rate. E.g. π is a Mixture of Gaussians and k the RBF kernel.
Assuming this inequality holds for π, and by using Proposition 1 and the Gronwall's lemma, one can show that the KL divergence between μ_t and π exponentially converges to zero (complete proof in A.7):

$$\mathrm{KL}(\mu_t||\pi) \leq e^{-2\lambda t} \, \mathrm{KL}(\mu_0||\pi). \tag{9}$$

This last result is very interesting as it creates a direct link between the convergence of $\mathrm{KL}(\mu_t||\pi)$ and the convergence of $I_{Stein}(\mu_t|\pi)$, showing that the iterative process of SVGD minimizes the KL divergence between μ_t and π exponentially fast, assuming π satisfies the Stein log-Sobolev inequality.

3.4 SVGD in discrete time

The authors defined the following mild assumptions:

- **(A1)**: $\exists B > 0$ such that $\forall x \in \mathcal{X}$:

$$\|k(x,\cdot)\|_{\mathcal{H}_0} \leq B \text{ and } \|\nabla k(x,\cdot)\|_{\mathcal{H}} \leq B;$$

- **(A2)** the Hessian H_V of $V = \log \pi$ is well-defined and $\exists M > 0$ such that $\|H_V\|_{op} \leq M$:
- **(A3)**: $\exists C > 0$ such that $I_{Stein}(\mu_n||\pi) < C$ for all n.

With these condition satisfied, the authors were able to show the following descent lemma:

Lemma 1 (Descent lemma for SVGD in discrete time). *Let $\alpha > 1$ and $\gamma \leq \frac{\alpha-1}{\alpha BC^{\frac{1}{2}}}$. Then, for all $n \geq 0$:*

$$\mathrm{KL}(\mu_{n+1}||\pi) - \mathrm{KL}(\mu_n||\pi) \leq -\gamma\left(1 - \gamma\frac{(\alpha^2 + M)B^2}{2}\right)I_{Stein}(\mu_n||\pi).$$

A descent lemma has already been proved before (Liu [2017]), but the authors proved it using differential calculus in the Wasserstein space, showing a more direct link between the descent lemma and the Wasserstein gradient flow: $v_t = -P_{\mu_t}\nabla \log \frac{\mu_t}{\pi}$. This lemma also implies the convergence for the average of $I_{Stein}(\mu|\pi)$ defined in (8), but for discrete time (replacing the integral by a sum).

3.5 Finite particle setting

Finally, the authors proved a convergence bounds between μ_n and its equivalent in the finite particle setting. In practice, when implementing SVGD, one starts with N particles such that $X_i \sim \mu_0$ for $i \in \{1, \ldots, N\}$. Then, at each iteration n, SVGD computes the new particles X_i^{n+1} as follows:

$$X_i^{n+1} = X_i^n - \gamma P_{\hat{\mu}_n}\nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_i^n) \tag{10}$$

where $\hat{\mu}_n = \frac{1}{N}\sum_{i=1}^N \delta_{X_i^n}$ is the empirical distribution of the particles. Under some Lipschitz assumptions, the following proposition holds:

Proposition 2. *Let $n > 0$ and $T > 0$. Let μ_n and $\hat{\mu}_n$ defined in (6) and (10) respectively. Then, for any $0 \leq n \leq \frac{T}{\gamma}$:*

$$\mathbb{E}\big[\mathbb{W}_2^2(\mu_n, \hat{\mu}_n)\big] \leq \frac{1}{2}\left(\frac{1}{\sqrt{N}}\sqrt{var(\mu_0)}e^{LT}\right)\big(e^{2LT} - 1\big), \tag{11}$$

where T is a constant depending on π and k.

This last result connects the usual implementation of SVGD that is in the finite particle setting with the infinite setting. It allows to ensure that the theoretical results showed before asymptotically hold in practice. It means that the implementation of the SVGD algorithm minimizes the KL divergence between $\hat{\mu}_n$ and π, with a sufficient amount of time depending on the number of particles.

4 Experiences

To assess the performance of the SVGD algorithm and the theoretical results, the authors performed an experiment on a 1D Gaussian mixture model. They initialized 100 particles following a Gaussian distribution centered on -10 and with a variance of 1. Then, they used the SVGD algorithm to minimize the KL divergence between the empirical distribution of the particles and a Gaussian mixture model with two modes. They also empirically verified that the inequality (8) holds, which seems to be the case in their example. They used the code provided in the original paper (Liu and Wang [2016]).

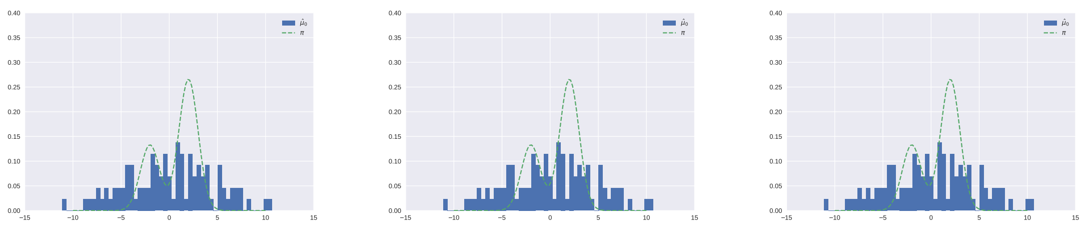
We decided to make a similar experiment but using our own code. Indeed, the original code is rather complex and implements many features that does not seem to be explicitly detailed in the paper. Therefore, we have implemented the SVGD algorithm only with the information provided in the studied paper (Korba et al. [2020]). We update the particles using (10) and we use PyTorch to compute the gradients of the kernel and $\log \pi$. We also wanted to verify that $\mathrm{KL}(\hat{\mu}_n || \pi)$ exponentially decreases, as shows (9). To do so, we made two experiments with different particles distributions. Both are Gaussian distributions, the first one is the same as the one used in the studied paper, and the second one is centered on 0 and with a variance of 5. π is a Gaussian mixture model with two modes centered on -5 and 5, a variance of 1, and weights of $2/3$ and $1/3$. The others hyperparameters are detailed in Table 1.

For the first experiment, the empirical distribution does not succeed to imitate π. Indeed, the algorithm get stuck in a local minimum, making $\hat{\mu}_t$ only follow the first modes of π. This behavior is illustrated in Figure 1, where one can see that, at first, the particles are updated in order to get closer to π, as expected. But after a few iterations, as soon as the particles are close enough to the first mode of π, the algorithm does not foresee the second mode and get stuck in a local minimum (as one can see with Figure 2). However, the KL divergence decreases exponentially, until it reaches the local minimum. This is illustrated in Figure 3, where we empirically set $\lambda = 0.01$.

On the other hand, the second experiment is more successful. Indeed, the particules are distributed following a Gaussian distribution with a large variance around 0, which allows the algorithm to take into account both mode of π. This is illustrated in Figure 4. The KL divergence also decreases exponentially, as shown in Figure 5, with $\lambda = 0.01$.

π	Mix Gaussian
N	100
γ	0.1
k	RBF kernel using median trick
# iter	500

Table 1: Hyperparameters used for the SVGD experiments.



(a) Iteration 1 (b) Iteration 100 (c) Iteration 500

Fig. 1: Several iterations of the SVGD algorithm of the first experiment. One can see that the algorithm get stuck in a local minimum, making $\hat{\mu}_t$ only follow the first modes of π.

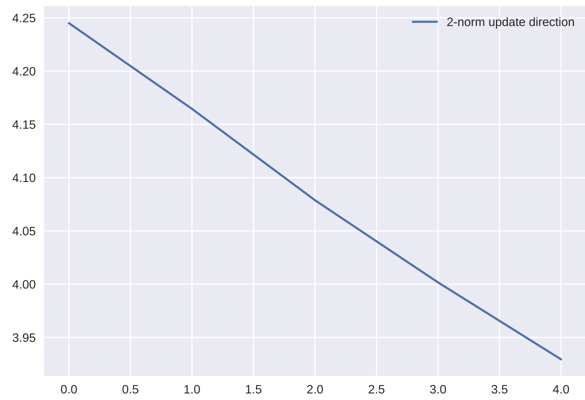


Fig. 2: 2-norm of the update direction of the particles for the first experiment. It corroborates the fact that the algorithm get stuck in a local minimum.

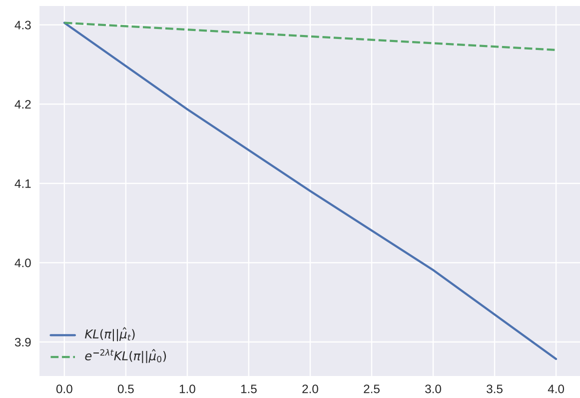
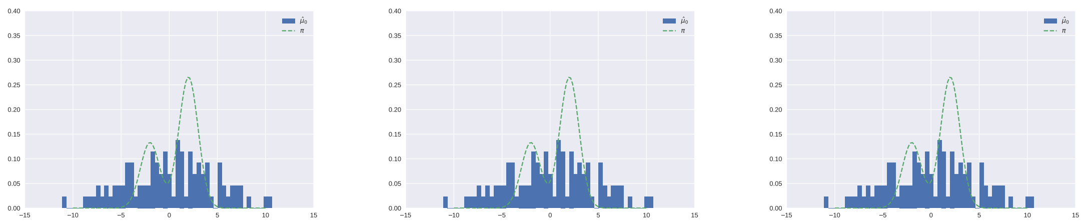


Fig. 3: Exponential decrease of the KL divergence over time for the first experiment. $\lambda = 0.01$.



(a) Iteration 1 (b) Iteration 100 (c) Iteration 500

Fig. 4: Several iterations of the SVGD algorithm of the second experiment. One can see that the algorithm is able to take into account both modes of π.

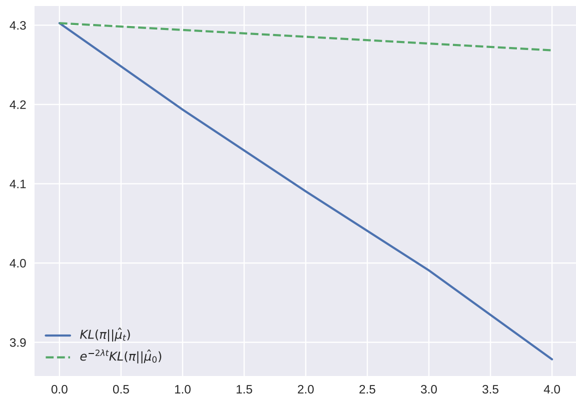


Fig. 5: Exponential decrease of the KL divergence over time for the second experiment. $\lambda = 0.001$.

5 Discussions

6 Conclusion

Bibliography

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent, 2020. URL `https://arxiv.org/abs/2006.09797`.

Qiang Liu. Stein variational gradient descent as gradient flow, 2017. URL `https://arxiv.org/abs/1704.07520`.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2016. URL `https://arxiv.org/abs/1608.04471`.

C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003. ISBN 9781470418045. URL `https://books.google.fr/books?id=MyPjjgEACAAJ`.

A Proofs

A.1 Proof of 1

Proof.

$$
\begin{aligned}
\mathbb{E}_{x\sim\mu}[\mathcal{A}_\mu\phi(x)] &= \int_{\mathcal{X}} \big(\nabla \log\mu(\cdot)^\top \phi(\cdot) + \nabla\cdot\phi(\cdot)\big)\mu(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \nabla \log\mu(\cdot)^\top \phi(\cdot)\mu(x)\,\mathrm{d}x + \int_{\mathcal{X}} \nabla\cdot\phi(\cdot)\mu(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \nabla \log\mu(\cdot)^\top \phi(\cdot)\mu(x)\,\mathrm{d}x + \int_{\mathcal{X}} \sum_{k=1}^{d} \frac{\partial\phi_k}{\partial x_k}\mu(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \nabla \log\mu(\cdot)^\top \phi(\cdot)\mu(x)\,\mathrm{d}x + \sum_{k=1}^{d}\left(\int_{\partial X} (\pi(x)\phi_k(x))\cdot n\,\mathrm{d}n - \int_{\mathcal{X}} \frac{\partial\mu(x)}{\partial x_k}\phi_k(x)\,\mathrm{d}x\right) \\
&= \int_{\mathcal{X}} \nabla \log\mu(\cdot)^\top \phi(\cdot)\mu(x)\,\mathrm{d}x - \int_{\mathcal{X}} \sum_{k=1}^{d} \frac{\partial\mu(x)}{\partial x_k}\phi_k(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \mu(x)\sum_{k=1}^{d} \frac{\partial\log\mu(x)}{\partial x_k}\phi_k(x) - \mu(x)\sum_{k=1}^{d} \frac{\partial\log\mu(x)}{\partial x_k}\phi_k(x)\,\mathrm{d}x \quad \text{(log trick)} \\
&= 0
\end{aligned}
$$

∎

A.2 Proof of 5

Proof. We first need to prove that

$$\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] = \langle f, \phi^*\rangle_{\mathcal{H}},\ \forall f\in\mathcal{H}:$$

$$
\begin{aligned}
\langle f, \phi^*\rangle_{\mathcal{H}} &= \sum_{l=1}^{d}\left\langle f^{(l)}, \mathbb{E}_{x\sim\mu}\left[k(x,\cdot)\nabla\log\pi(x)^{(l)} + \nabla k(x,\cdot)^{(l)}\right]\right\rangle_{\mathcal{H}^0} \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{l=1}^{d}\left\langle f^{(l)}, k(x,\cdot)\nabla\log\pi(x)^{(l)} + \nabla k(x,\cdot)^{(l)}\right\rangle_{\mathcal{H}^0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{l=1}^{d}\nabla\log\pi(x)^{(l)}\left\langle f^{(l)}, k(x,\cdot)\right\rangle_{\mathcal{H}^0} + \left\langle f^{(l)}, \nabla k(x,\cdot)^{(l)}\right\rangle_{\mathcal{H}^0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{l=1}^{d}\nabla\log\pi(x)^{(l)} f^{(l)}(x) + \nabla_{x_l} f(x)^{(l)}\right] \text{ (see ?)} \\
&= \mathbb{E}_{x\sim\mu}\left[\nabla\log\pi(x)^\top f(x) + \nabla\cdot f(x)\right] \\
&= \mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)].
\end{aligned}
$$

Moreover, $\langle f, \phi * \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|\phi^*\|_{\mathcal{H}}$. Thus,

$$\mathbb{S}(\mu, \pi) = \max_{f \in \mathcal{H}} \{\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi f(x)] = \langle f, \phi^* \rangle_{\mathcal{H}},\ s.t.\ \|f\|_{\mathcal{H}} \leq 1\} \leq \|\phi^*\|_{\mathcal{H}}.$$

Let $f = \frac{\phi^*}{\|\phi^*\|_{\mathcal{H}}}$, then $\|f\|_{\mathcal{H}} = 1$ and

$$\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] = \langle f, \phi^* \rangle_{\mathcal{H}} = \|\phi^*\|_{\mathcal{H}},$$

ending the proof. ■

A.3 Proof that $\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi(x)]$ measures the disperancy between μ and π

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi(x)] &= \int_{\mathcal{X}} \big(\nabla \log \pi(x)^\top \phi(x) + \nabla \cdot \phi(x)\big)\mu(x)\, \mathrm{d}x \\
&= \int_{\mathcal{X}} \nabla \log \pi(x)^\top \phi(x) \mu(x)\, \mathrm{d}x + \sum_{k=1}^{d} \left(\mathcal{R}_k - \int_{\mathcal{X}} \frac{\partial \mu(x)}{\partial x_k} \phi_k(x)\, \mathrm{d}x \right) \\
&= \sum_{k=1}^{d} \mathcal{R}_k + \int_{\mathcal{X}} \mu(x) \sum_{k=1}^{d} \frac{\partial \log \pi(x)}{\partial x_k} \phi_k(x) - \mu(x) \sum_{k=1}^{d} \frac{\partial \log \mu(x)}{\partial x_k} \phi_k(x)\, \mathrm{d}x \quad \text{(log trick)} \\
&= \sum_{k=1}^{d} \mathcal{R}_k + \sum_{k=1}^{d} [\mu(x)\phi_k(x)]_{\mathcal{X}} + \int_{\mathcal{X}} \mu(x) \left[\sum_{k=1}^{d} \phi_k(x) \left(\frac{\partial \log \pi(x)}{\partial x_k} - \frac{\partial \log \mu(x)}{\partial x_k} \right) \right] \mathrm{d}x \\
&= \sum_{k=1}^{d} \mathcal{R}_k + \sum_{k=1}^{d} [\mu(x)\phi_k(x)]_{\mathcal{X}} + \int_{\mathcal{X}} \mu(x) \left[\sum_{k=1}^{d} \phi_k(x) \left(\frac{\partial \log \frac{\pi(x)}{\mu(x)}}{\partial x_k} \right) \right] \mathrm{d}x,
\end{aligned}$$

■

A.4 Proof of $\mathcal{H} \subset L^2(\mu)$ (3)

Proof. We want to prove that, $\forall f \in \mathcal{H}$, $\forall \mu \in \mathcal{P}_2(\mathcal{X})$, $\int_{\mathcal{X}} f(x)^2\, \mathrm{d}\mu(x) < \infty$.

$$\begin{aligned}
\int_{\mathcal{X}} f(x)^2\, \mathrm{d}\mu(x) &= \int_{\mathcal{X}} \sum_{l=1}^{d} \left\langle f^{(l)}, k(x, \cdot) \right\rangle_{\mathcal{H}_0}^2 \mathrm{d}\mu(x) \\
&\leq \sum_{l=1}^{d} \int_{\mathcal{X}} \left\| f^{(l)} \right\|_{\mathcal{H}_0}^2 \|k(x, \cdot)\|_{\mathcal{H}_0}^2\, \mathrm{d}\mu(x) \text{ (by C.S)} \\
&= \sum_{l=1}^{d} \left\| f^{(l)} \right\|_{\mathcal{H}_0}^2 \int_{\mathcal{X}} \|k(x, \cdot)\|_{\mathcal{H}_0}^2\, \mathrm{d}\mu(x) \\
&= \sum_{l=1}^{d} \left\| f^{(l)} \right\|_{\mathcal{H}_0}^2 \int_{\mathcal{X}} \langle k(x, \cdot), k(x, \cdot) \rangle_{\mathcal{H}_0}\, \mathrm{d}\mu(x) \\
&= \sum_{l=1}^{d} \left\| f^{(l)} \right\|_{\mathcal{H}_0}^2 \int_{\mathcal{X}} k(x, x)\, \mathrm{d}\mu(x) \text{ (by reproducing propriety)} \\
&< \infty, \text{ as } \int_{\mathcal{X}} k(x, x)\, \mathrm{d}\mu(x) < \infty.
\end{aligned}$$

■

A.5 Proof of 7

Proof. Let k in the Stein class of μ. Thus:

$$\begin{aligned}
P_\mu \nabla \log \frac{\mu}{\pi}(\cdot) &= \int_{\mathcal{X}} k(x,\cdot)\nabla \log \mu(x)\,\mathrm{d}\mu(x) - \int_{\mathcal{X}} k(x,\cdot)\nabla \log \pi(x)\,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} k(x,\cdot)\nabla \mu(x)\,\mathrm{d}x - \int_{\mathcal{X}} k(x,\cdot)\nabla \log \pi(x)\,\mathrm{d}\mu(x) \\
&= -\int_{\mathcal{X}} \nabla k(x,\cdot)\,\mathrm{d}\mu(x) - \int_{\mathcal{X}} k(x,\cdot)\nabla \log \pi(x)\,\mathrm{d}\mu(x) \\
&= -\int_{\mathcal{X}} k(x,\cdot)\nabla \log \pi(x) + \nabla k(x,\cdot)\,\mathrm{d}\mu(x) \\
&= -\phi^*(\cdot).
\end{aligned} \tag{12}$$

■

A.6 Proof of Proposition 1

Proof. The time derivative of the KL writes:

$$\begin{aligned}
\frac{\partial KL(\mu_t \| \pi)}{\partial t} &= \frac{\partial}{\partial t}\int_{\mathcal{X}} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}\mu_t(x) \\
&= \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{X}} \mu_t(x) \frac{\partial \log \frac{\mu_t(x)}{\pi(x)}}{\partial t}\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{X}} \mu_t(x) \frac{\partial \log \mu_t(x)}{\partial t}\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{X}} \mu_t(x) \frac{\frac{\partial \mu_t(x)}{\partial t}}{\mu_t(x)}\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t}\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial \mu_t(x)}{\partial t} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x, \left(\mu_t \text{ is a probability measure, so } \forall t, \int_{\mathcal{X}} \mathrm{d}\mu_t(x) = 1 \right)
\end{aligned}$$

Furthermore, as μ_t satisfies the continuity equation ((2)) where $v_t = -P_{\mu_t}\nabla \log \frac{\mu}{\pi}$, we have:

$$\begin{aligned}
\frac{\partial KL(\mu_t \| \pi)}{\partial t} &= -\int_{\mathcal{X}} \nabla \cdot (\mu_t(x) v_t(x)) \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x \\
&= -\sum_{l=1}^{d} \int_{\mathcal{X}} \frac{\partial \mu_t(x) v_t(x)}{\partial x_l} \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}x \\
&= -\int_{\partial X} \left(\mu_t(x) v_t(x) \log \frac{\mu_t(x)}{\pi(x)} \right) \cdot n\,\mathrm{d}n + \sum_{l=1}^{d} \int_{\mathcal{X}} \mu_t(x) v_t(x) \frac{\partial \log \frac{\mu_t(x)}{\pi(x)}}{\partial x_l}\,\mathrm{d}x
\end{aligned}$$

The first term cancels as probability densities tends to zero on the boundary.

$$\begin{aligned}
&= \int_{\mathcal{X}} v_t(x) \nabla \log \frac{\mu_t(x)}{\pi(x)}\,\mathrm{d}\mu_t(x) \\
&= \left\langle v_t, \nabla \log \frac{\mu_t}{\pi} \right\rangle_{L^2(\mu_t)} \\
&= \left\langle \iota^* v_t, \iota^* \nabla \log \frac{\mu_t}{\pi} \right\rangle_{\mathcal{H}} \\
&= \left\langle -\iota^* \iota S_{\mu_t} \nabla \log \frac{\mu_t}{\pi}, S_{\mu_t} \nabla \log \frac{\mu_t}{\pi} \right\rangle_{\mathcal{H}} \\
&= -\left\| S_{\mu_t} \nabla \log \frac{\mu_t}{\pi} \right\|_{\mathcal{H}}^2
\end{aligned}$$

■

A.7 Proof of 9

Proof. Assume that π satisfies the Stein log-Sobolev inequality. We have

$$\begin{aligned} \mathrm{KL}(\mu_t||\pi) &\leq \frac{1}{2\lambda} I_{Stein}(\mu_t||\pi) \\ -I_{Stein}(\mu_t||\pi) &\leq -2\lambda \mathrm{KL}(\mu_t||\pi). \end{aligned}$$

Now, using Proposition 1:

$$\begin{aligned} \frac{\partial \mathrm{KL}(\mu_t||\pi)}{\partial t} &\leq -2\lambda \mathrm{KL}(\mu_t||\pi) \\ \mathrm{KL}(\mu_t||\pi) &\leq \mathrm{KL}(\mu_0||\pi) \exp\left(\int_0^t -2\lambda \, \mathrm{d}s \right) \text{ (Gronwall's lemma)} \\ \mathrm{KL}(\mu_t||\pi) &\leq e^{-2\lambda t} \mathrm{KL}(\mu_0||\pi) \end{aligned}$$

■