Appendix of "Stein Variational Gradient Descent as Gradient Flow"

Qiang Liu
Department of Computer Science
Dartmouth College
Hanover, NH 03755
qiang.liu@dartmouth.edu

A Density Evolution of SVGD Dynamics

A.1 Proof of Lemma 3.1

Proof. Recall that $\boldsymbol{g}(x, x') = \mathcal{S}_p^{x'} \otimes k(x', x)$, and $\boldsymbol{\phi}^*_{\mu,p}(x) = \mathbb{E}_{x'\sim\mu}[\boldsymbol{g}(x, x')]$, we have $\boldsymbol{T}_{\mu,p}(x) = x + \epsilon\mathbb{E}_{x'\sim\mu}[\boldsymbol{g}(x, x')]$. Therefore,

$$\begin{aligned}||\boldsymbol{T}_{\mu,p}||_{\text{Lip}} &= \max_{x\neq y} \frac{||\boldsymbol{T}_{\mu,p}(x) - \boldsymbol{T}_{\mu,p}(y)||_2}{||x-y||_2} \\ &= \max_{x\neq y} \frac{||x - y + \epsilon\mathbb{E}_{x'\sim\mu}[\boldsymbol{g}(x, x') - \boldsymbol{g}(y, x')]||_2}{||x-y||_2} \\ &\leq 1 + \epsilon||\boldsymbol{g}||_{\text{Lip}}, \end{aligned} \tag{A.1}$$

and for $\forall\, x$,

$$||\boldsymbol{T}_{\mu,p}(x) - \boldsymbol{T}_{\nu,p}(x)||_2 = \epsilon||\mathbb{E}_{x'\sim\mu}\boldsymbol{g}(x, x') - \mathbb{E}_{x'\sim\nu}\boldsymbol{g}(x, x')||_2 \leq \epsilon||\boldsymbol{g}||_{\text{BL}}\ \text{BL}(\mu,\ \nu). \tag{A.2}$$

For any h with $||h||_{\text{BL}} = \max(||h||_\infty,\ ||h||_{\text{Lip}}) \leq 1$, we have

$$\begin{aligned}&\left|\mathbb{E}_{\Phi_p(\mu)}[h] - \mathbb{E}_{\Phi_p(\nu)}[h]\right| \\ &= \left|\mathbb{E}_{\mu}[h\circ\boldsymbol{T}_{\mu,p}] - \mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\nu,p}]\right| \\ &\leq \left|\mathbb{E}_{\mu}[h\circ\boldsymbol{T}_{\mu,p}] - \mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\mu,p}]\right| \ + \ \left|\mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\mu,p}] - \mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\nu,p}]\right|.\end{aligned}$$

We just need to bound these two terms. For the first term,

$$\begin{aligned}\left|\mathbb{E}_{\mu}[h\circ\boldsymbol{T}_{\mu,p}] - \mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\mu,p}]\right| &\leq ||h\circ\boldsymbol{T}_{\mu,p}||_{\text{BL}}\ \text{BL}(\mu,\ \nu) \\ &\leq \max\left(||h||_\infty,\ ||h||_{\text{Lip}}||\boldsymbol{T}_{\mu,p}||_{\text{Lip}}\right)\ \text{BL}(\mu,\ \nu) \\ &\leq (1+\epsilon||\boldsymbol{g}||_{\text{Lip}})\text{BL}(\mu,\ \nu), \qquad \text{//by Equation A.1.}\end{aligned}$$

For the second term,

$$\begin{aligned}\left|\mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\mu,p}] - \mathbb{E}_{\nu}[h\circ\boldsymbol{T}_{\nu,p}]\right| &\leq \max_x \left|h\circ\boldsymbol{T}_{\mu,p}(x) - h\circ\boldsymbol{T}_{\nu,p}(x)\right| \\ &\leq ||h||_{\text{Lip}} \max_x ||\boldsymbol{T}_{\mu,p}(x) - \boldsymbol{T}_{\nu,p}(x)||_2 \\ &\leq \epsilon||\boldsymbol{g}||_{\text{BL}}\ \text{BL}(\mu,\ \nu), \qquad \text{//by Equation A.2.}\end{aligned}$$

Therefore,

$$\text{BL}(\Phi_p(\mu),\ \Phi_p(\nu)) \leq (1 + \epsilon||\boldsymbol{g}||_{\text{Lip}} + \epsilon||\boldsymbol{g}||_{\text{BL}})\ \text{BL}(\mu,\ \nu) \leq (1 + 2\epsilon||\boldsymbol{g}||_{\text{BL}})\ \text{BL}(\mu,\ \nu).$$

□

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

A.2 Proof of Theorem 3.3

Proof. Denote by $\mu_\ell = \mu_\ell^\infty$ for notation convenience.

$$\begin{aligned}
&\mathrm{KL}(\mu_{\ell+1} \,||\, \nu_p) - \mathrm{KL}(\mu_\ell \,||\, \nu_p) \\
&= \mathrm{KL}(\boldsymbol{T}_{\mu_\ell,p}\mu_\ell \,||\, \nu_p) - \mathrm{KL}(\mu_\ell \,||\, \nu_p) \\
&= \mathrm{KL}(\mu_\ell \,||\, \boldsymbol{T}_{\mu_\ell,p}^{-1}\nu_p) - \mathrm{KL}(\mu_\ell \,||\, \nu_p) \qquad \text{//by Lemma A.2} \\
&= -\mathbb{E}_{x\sim\mu_\ell}[\log p(\boldsymbol{T}_{\mu_\ell,p}(x)) + \log\det(\nabla \boldsymbol{T}_{\mu_\ell,p}(x)) - \log p(x)]. \qquad \text{(A.3)}
\end{aligned}$$

Note that $\boldsymbol{T}_{\mu_\ell,p}(x) = x + \epsilon\boldsymbol{\phi}^*_{\mu_\ell,p}(x)$. We have the follow version of Taylor approximation:

$$\log p(x) - \log p(\boldsymbol{T}_{\mu_\ell,p}(x)) \le -\epsilon\nabla_x \log p(x)^\top \boldsymbol{\phi}^*_{\mu_\ell,p}(x) \;+\; \frac{\epsilon^2}{2}||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}||_2^2. \qquad \text{(A.4)}$$

This is because, defining $x_s = x + s\epsilon\boldsymbol{\phi}^*_{\mu_\ell,p}(x)$, $\forall s \in [0,1]$,

$$\begin{aligned}
&\log p(x) - \log p(\boldsymbol{T}_{\mu_\ell,p}(x)) \\
&= -\int_0^1 \nabla_s \log p(x_s)\mathrm{d}s \\
&= -\int_0^1 \nabla_x \log p(x_s)^\top (\epsilon\boldsymbol{\phi}^*_{\mu_\ell,p}(x))\,\mathrm{d}s \\
&= -\epsilon\nabla_x \log p(x)^\top \boldsymbol{\phi}^*_{\mu_\ell,p}(x) \;-\; \int_0^1 (\nabla_x \log p(x_s) - \nabla_x \log p(x))^\top (\epsilon\boldsymbol{\phi}^*_{\mu_\ell,p}(x))\mathrm{d}s \\
&\le -\epsilon\nabla_x \log p(x)^\top \boldsymbol{\phi}^*_{\mu_\ell,p}(x) \;+\; \epsilon^2||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_2^2 \int_0^1 s\mathrm{d}s \\
&= -\epsilon\nabla_x \log p(x)^\top \boldsymbol{\phi}^*_{\mu_\ell,p}(x) \;+\; \frac{\epsilon^2}{2}||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_2^2.
\end{aligned}$$

where we used the fundamental theorem of calculus, which holds for weakly differentiable functions [20, Theorem 3.60, page 77]. In addition, Take $B = \nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)$ in bound (A.9) of Lemma A.1, and take $\epsilon < 1/(2\rho(B + B^\top))$, we have

$$\begin{aligned}
\log|\det(\nabla \boldsymbol{T}_{\mu_\ell,p}(x))| &\ge \epsilon\,\mathrm{tr}(\nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)) - 2\epsilon^2||\nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_F^2 \\
&= \epsilon\,\nabla\cdot\boldsymbol{\phi}^*_{\mu_\ell,p}(x) - 2\epsilon^2||\nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_F^2. \qquad \text{(A.5)}
\end{aligned}$$

Combining (A.4) and (A.5) gives

$$\begin{aligned}
\mathrm{KL}(\mu_{\ell+1} \,||\, \nu_p) - \mathrm{KL}(\mu_\ell \,||\, \nu_p) &\le -\epsilon\mathbb{E}_{\mu_\ell}[\mathcal{S}_p\boldsymbol{\phi}^*_{\mu_\ell,p}] \;+\; \Delta \\
&= -\epsilon\mathbb{D}(\mu_\ell \,||\, \nu_p)^2 \;+\; \Delta,
\end{aligned}$$

where Δ is a residual term:

$$\Delta = \epsilon^2\mathbb{E}_{x\sim\mu_\ell}\left[\frac{1}{2}||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_2^2 + 2||\nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_F^2\right]$$

We need to bound $||\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_2$ and $||\nabla\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_F$. This can be done using the reproducing property: let $\boldsymbol{\phi}^*_{\mu_\ell,p} = [\phi_1, \cdots, \phi_d]^\top$; recall that $\phi_i \in \mathcal{H}_0$ and $\boldsymbol{\phi}^*_{\mu_\ell,p} \in \mathcal{H} = \mathcal{H}_0^d$, then

$$\phi_i(x) = \langle\phi_i(\cdot),\, k(x,\cdot)\rangle_{\mathcal{H}_0}, \quad \partial_{x_j}\phi_i(x) = \langle\phi_i(\cdot),\, \partial_{x_j}k(x,\cdot)\rangle_{\mathcal{H}_0}, \qquad \forall i,j = 1,\ldots,d, \;\; x \in X.$$

Also note that $||\boldsymbol{\phi}^*_{\mu_\ell,p}||_{\mathcal{H}}^2 = \sum_{i=1}^d ||\phi_i||_{\mathcal{H}_0}^2 = \mathbb{D}(\mu_\ell \,||\, \nu_p)^2$, we have by Cauchy-Swarchz inequality,

$$\begin{aligned}
||\boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_2^2 &= \sum_{i=1}^d \phi_i(x)^2 \\
&= \sum_{i=1}^d (\langle k(x,\cdot),\, \phi_i(\cdot)\rangle_{\mathcal{H}_0})^2 \\
&\le \sum_i ||k(x,\cdot)||_{\mathcal{H}_0}^2 \cdot ||\phi_i||_{\mathcal{H}_0}^2 \\
&= k(x,x) \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}||_{\mathcal{H}}^2 \\
&= k(x,x) \cdot \mathbb{D}(\mu_\ell \,||\, \nu_p)^2,
\end{aligned}$$

and

$$
\begin{aligned}
||\nabla \boldsymbol{\phi}^*_{\mu_\ell,p}(x)||_F^2 &= \sum_{ij} \partial_{x_j}\phi_i(x)^2 \\
&= \sum_{ij} (\langle \partial_{x_j} k(x,\cdot),\ \phi_i(\cdot)\rangle_{\mathcal{H}_0})^2 \\
&\le \sum_{ij} ||\partial_{x_j} k(x,\cdot)||_{\mathcal{H}_0}^2 \cdot ||\phi_i||_{\mathcal{H}_0}^2 \\
&= \sum_{ij} \partial_{x_j,x_j'} k(x,x')|_{x=x'} \cdot ||\phi_i||_{\mathcal{H}_0}^2 \\
&= \nabla_{xx'} k(x,x) \cdot ||\boldsymbol{\phi}^*_{\mu_\ell,p}||_{\mathcal{H}}^2 \\
&= \nabla_{xx'} k(x,x) \cdot \mathbb{D}(\mu_\ell \,||\, \nu_p)^2. \qquad \text{(A.6)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Delta &\le \epsilon^2\, \mathbb{D}(\mu_\ell \,||\, \nu_p)^2 \left(\frac{1}{2} \mathbb{E}_{x\sim\mu_\ell}[||\nabla \log p||_{\mathrm{Lip}} k(x,x) + 2\nabla_{xx'} k(x,x)] \right) \\
&= \epsilon^2 R\, \mathbb{D}(\mu_\ell \,||\, \nu_p)^2.
\end{aligned}
$$

This gives

$$
\mathrm{KL}(\mu_{\ell+1} \,||\, \nu_p) - \mathrm{KL}(\mu_\ell \,||\, \nu_p) \le -\epsilon\,(1-\epsilon R)\, \mathbb{D}(\mu_\ell \,||\, \nu_p)^2.
$$

□

Lemma A.1. *Let B be a square matrix and $||B||_F = \sqrt{\sum_{ij} b_{ij}^2}$ its Frobenius norm. Let ϵ be a positive number that satisfies $0 \le \epsilon < \frac{1}{\rho(B+B^\top)}$, where $\rho(\cdot)$ denotes the spectrum radius. Then $I + \epsilon(B + B^\top)$ is positive definite, and*

$$
\log|\det(I+\epsilon B)| \ge \epsilon \mathrm{tr}(B) - \epsilon^2 \frac{||B||_F^2}{1-\epsilon\rho(B+B^\top)}. \qquad \text{(A.7)}
$$

Therefore, take an even smaller ϵ such that $0 \le \epsilon \le \frac{1}{2\rho(B+B^\top)}$, we get

$$
\log|\det(I+\epsilon B)| \ge \epsilon \mathrm{tr}(B) - 2\epsilon^2 ||B||_F^2. \qquad \text{(A.8)}
$$

Proof. When $\epsilon < \frac{1}{\rho(B+B^\top)}$, we have $\rho(I+\epsilon(B+B^\top)) \ge 1 - \epsilon\rho(B+B^\top) > 0$, so $I+\epsilon(B+B^\top)$ is positive definite.

By the property of matrix determinant, we have

$$
\begin{aligned}
\log|\det(I+\epsilon B)| &= \frac{1}{2}\log\det((I+\epsilon B)(I+\epsilon B)^\top) \\
&= \frac{1}{2}\log\det(I+\epsilon(B+B^\top)+\epsilon^2 BB^\top) \\
&\ge \frac{1}{2}\log\det(I+\epsilon(B+B^\top)), \qquad \text{(A.9)}
\end{aligned}
$$

where (A.10) holds because both $I+\epsilon(B+B^\top)$ and $\epsilon^2 BB^\top$ are positive semi-definite.

Let $A = B + B^\top$. We can establish

$$
\log\det(I+\epsilon A) \ge \epsilon \mathrm{tr}(A) - \frac{\epsilon^2}{2}\frac{||A||_F^2}{1-\epsilon\rho(A)}, \qquad \text{(A.10)}
$$

which holds for any symmetric matrix A and $0 \le \epsilon < 1/\rho(A)$. This is because, assuming $\{\lambda_i\}$ are the eigenvalues of A,

$$\begin{aligned}
\log\det(I+\epsilon A) - \epsilon \mathrm{tr}(A) &= \sum_i [\log(1+\epsilon\lambda_i) - \epsilon\lambda_i] \\
&= \sum_i [\int_0^1 \frac{\epsilon\lambda_i}{1+s\epsilon\lambda_i}\mathrm{d}s - \epsilon\lambda_i] \\
&= -\sum_i \int_0^1 \frac{s\epsilon^2\lambda_i^2}{1+s\epsilon\lambda_i}\mathrm{d}s \\
&\ge -\sum_i \frac{\epsilon^2\lambda_i^2}{1-\epsilon\max_i|\lambda_i|}\int_0^1 s\mathrm{d}s \\
&\ge -\sum_i \frac{\epsilon^2\lambda_i^2}{2(1-\epsilon\max_i|\lambda_i|)} \\
&= -\frac{\epsilon^2}{2}\frac{||A||_F^2}{1-\epsilon\rho(A)}.
\end{aligned}$$

Take $A = B + B^\top$ in (A.11) and combine it with (A.10), we get

$$\begin{aligned}
\log|\det(I+\epsilon B)| &\ge \frac{1}{2}\log\det(I+\epsilon(B+B^\top)) \\
&\ge \frac{\epsilon}{2}\mathrm{tr}(B+B^\top) - \frac{\epsilon^2}{4}\frac{||B+B^\top||_F^2}{1-\epsilon\rho(B+B^\top)} \\
&\ge \epsilon\mathrm{tr}(B) - \epsilon^2\frac{||B||_F^2}{1-\epsilon\rho(B+B^\top)},
\end{aligned}$$

where we used the fact that $\mathrm{tr}(B) = \mathrm{tr}(B^\top)$ and $||B+B^\top||_F \le ||B||_F + ||B^\top||_F = 2||B||_F$. □

Lemma A.2. *Let $\boldsymbol{T}$ be a one-to-one map, and μ and ν two probability measures. We have*

$$\mathrm{KL}(\boldsymbol{T}\mu \,||\, \nu) = \mathrm{KL}(\mu \,||\, \nu),$$

given that the KL divergence between μ and ν exists.

Proof. We prove this for f-divergence in general, which includes KL divergence as a special case. Given a convex function f such that $f(1) = 0$, the f-divergence is defined

$$\mathrm{D}_f(\mu \,||\, \nu) = \mathbb{E}_\nu[f(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})].$$

Assume f^* is the convex conjugate of f, we have a variational representation for f-divergence:

$$\mathrm{D}_f(\mu \,||\, \nu) = \sup_g \left\{\mathbb{E}_\mu[g(x)] - \mathbb{E}_\nu[f^*(g(x))]\right\},$$

where g is over the set of all measurable functions. Therefore, we have

$$\begin{aligned}
\mathrm{D}_f(\boldsymbol{T}\mu \,||\, \nu) &= \sup_g \left\{\mathbb{E}_\mu[g\circ T(x)] - \mathbb{E}_\nu[f^*(g(x))]\right\} \\
&= \sup_{\tilde{g}} \left\{\mathbb{E}_\mu[\tilde{g}(x)] - \mathbb{E}_\nu[f^*(\tilde{g}\circ\boldsymbol{T}^{-1}(x))]\right\} \qquad \text{//Define } \tilde{g} = g\circ\boldsymbol{T}. \\
&= \mathrm{D}_f(\mu \,||\, \boldsymbol{T}^{-1}\nu).
\end{aligned}$$

□

A.3 Proof of Fokker-Planck Equation (13)

Proof. Recall that $\boldsymbol{T}_{\mu,p}(x) = x + \epsilon\boldsymbol{\phi}^*_{\mu,p}(x)$ and we denote by q the density of measure μ. Assume ϵ is sufficiently small so that $\nabla\boldsymbol{T}_{\mu_p}(x) = I + \epsilon\nabla\boldsymbol{\phi}^*_{\mu,p}(x)$ is positive definite (See Lemma A.1). By the implicit function theorem, we have

$$\boldsymbol{T}^{-1}_{\mu,p}(x) = x - \epsilon\boldsymbol{\phi}^*_{\mu,p}(x) + o(\epsilon).$$

Therefore We have

$$\begin{aligned}\log q'(x) &= \log q(\boldsymbol{T}_{\mu,p}^{-1}(x)) + \log\det(\nabla_x \boldsymbol{T}_{\mu,p}^{-1}(x)) \\ &= \log q(x - \epsilon\cdot\boldsymbol{\phi}_{\mu,p}^*(x)) + \log\det(I - \epsilon\nabla_x\boldsymbol{\phi}_{\mu,p}^*(x)) + o(\epsilon) \\ &= \log q(x) - \epsilon\nabla_{x_i}\log q(x)^\top\boldsymbol{\phi}_{\mu,p}^*(x) - \epsilon q(x)\cdot\mathrm{tr}(\nabla_x\boldsymbol{\phi}_{\mu,p}^*(x)) + o(\epsilon) \\ &= \log q(x) - \epsilon\mathcal{S}_q\boldsymbol{\phi}_{\mu,p}^*(x) + o(\epsilon).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{q'(x)-q(x)}{\epsilon} &= \frac{q(\log q(x) - \log q(x))}{\epsilon} + o(\epsilon) \\ &= -q(x)\mathcal{S}_q\boldsymbol{\phi}_{q_\ell,p}^*(x) + o(\epsilon) \\ &= -\nabla\cdot(\boldsymbol{\phi}_{q_\ell,p}^*(x)q_\ell(x)) + o(\epsilon).\end{aligned}$$

Taking $\epsilon\to 0$ gives the result. □

A.4 Proof of Theorem 3.5

Proof. Since $q' = q + qf\mathrm{d}t$ is equivalent to transforming the variable by $\boldsymbol{T}(x) = x + \boldsymbol{\psi}_{q,f}\mathrm{d}t$, the corresponding change on KL divergence is

$$\begin{aligned}F(q + qf\mathrm{d}t) &= F(q) + \mathbb{E}_q[\mathcal{S}_p\boldsymbol{\psi}_{q,f}]\mathrm{d}t \\ &= F(q) + \langle\boldsymbol{\phi}_{q,p}^*,\ \boldsymbol{\psi}_{q,f}\rangle_{\mathcal{H}}\mathrm{d}t \\ &= F(q) + \langle\nabla\cdot(\boldsymbol{\phi}_{q,p}^*q),\ \nabla\cdot(\boldsymbol{\psi}_{q,f}q)\rangle_{q\mathcal{H}_q}\mathrm{d}t \\ &= F(q) + \langle\nabla\cdot(\boldsymbol{\phi}_{q,p}^*q),\ qf\rangle_{q\mathcal{H}_q}\mathrm{d}t\end{aligned}$$

This proves that $\nabla\cdot(\boldsymbol{\phi}_{q,p}^*q)$ is the covariant functional gradient. □

References

[1] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2016.

[2] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[3] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

[4] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-monte carlo way. *Acta Numerica*, 22:133–288, 2013.

[5] B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. In *The 19th International Conference on Artificial Intelligence and Statistics*, 2016.

[6] C. Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

[7] Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning (ICML)*, 2016.

[8] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness-of-fit. In *International Conference on Machine Learning (ICML)*, 2016.

[9] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 2017.

[10] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, 2017.

[11] C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein's identity. *arXiv preprint arXiv:1603.03220*, 2016.

[12] W. Braun and K. Hepp. The Vlasov dynamics and its fluctuations in the 1/n limit of interacting classical particles. *Communications in mathematical physics*, 56(2):101–113, 1977.

[13] A. A. Vlasov. On vibration properties of electron gas. *J. Exp. Theor. Phys*, 8(3):291, 1938.

[14] H. Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

[15] P. Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, 2013.

[16] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[17] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[18] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[19] J. Han and Q. Liu. Stein variational adaptive importance sampling. In *Uncertainty in Artificial Intelligence*, 2017.

[20] J. K. Hunter. Notes on partial differential equations. 2014. URL `https://www.math.ucdavis.edu/~hunter/m218a_09/pde_notes.pdf`.