Example 1.80: Continuous flow for density estimation (Tabak and Vanden-Eijnden 2010)

Let p be a known density, e.g., standard normal, and q be a density we wish to learn. We know there exist mappings T so that

$$q = \mathsf{T}_{\#}^{-1} p, \quad i.e., \quad \mathsf{T}^{-1}(\mathsf{X}) \sim q \text{ if } \mathsf{X} \sim p.$$

Let us build T^{-1} through a continuous family T_t^{-1}, leading to

$$p_t := (\mathsf{T}_t^{-1})_{\#} p, \quad p_t(\mathbf{x}) = |\det \mathsf{T}_t' \mathbf{x}| \cdot p(\mathsf{T}_t \mathbf{x}).$$

Conversely, we also have

$$p = \mathsf{T}_{\#} q, \quad q_t := (\mathsf{T}_t)_{\#} q, \quad q_t(\mathsf{T}_t \mathbf{x}) \cdot |\det \mathsf{T}_t' \mathbf{x}| = q(\mathbf{x}).$$

We use the KL divergence as our objective of learning:

$$\mathsf{KL}(q \| p_t) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p_t(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \mathsf{KL}(q_t \| p) = \int q(\mathbf{x}) \log \frac{q_t(\mathsf{T}_t \mathbf{x})}{p(\mathsf{T}_t \mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

We take the (functional) derivative w.r.t. T_t:

$$\frac{\delta \mathsf{KL}}{\delta \mathsf{T}_t}(\mathbf{x}) = [\mathbf{s}_{q_t}(\mathsf{T}_t \mathbf{x}) - \mathbf{s}_p(\mathsf{T}_t \mathbf{x})] q(\mathbf{x}) = [\mathbf{s}_{q_t}(\mathsf{T}_t \mathbf{x}) - \mathbf{s}_p(\mathsf{T}_t \mathbf{x})] \cdot q_t(\mathsf{T}_t \mathbf{x}) \cdot |\det \mathsf{T}_t' \mathbf{x}| \tag{1.24}$$

and evolve T_t according to the ODE (that guarantees decrease of our KL objective):

$$\mathrm{d}\mathsf{T}_t = -\mathbf{b}(\mathsf{T}_t), \quad \text{where} \quad \mathbf{b}(\mathbf{z}) = [\mathbf{s}_{q_t}(\mathbf{z}) - \mathbf{s}_p(\mathbf{z})] \cdot q_t(\mathbf{z}). \tag{1.25}$$

We have dropped the Jacobian $|\det \mathsf{T}_t' \mathbf{x}|$ in (1.24) for better interpretation. Essentially, we seek an infinitesimal improvement T over our current estimate T_t, so we compute $\delta \mathsf{KL}(\mathsf{T} \circ \mathsf{T}_t)/\delta \mathsf{T} \restriction_{\mathsf{T}=\mathrm{Id}}$, which leads exactly to (1.25). This Lagrangian view allows us to "forget" the past and focus "myopically" on deforming the *current* q_t to the target p. In fact, we have the following continuity equation corresponding to (1.25):

$$\partial_t q_t = \nabla \cdot (q_t \mathbf{b}) = \nabla \cdot [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)].$$

Alternatively, we may use the reverse KL divergence as our objective:

$$\mathsf{LK}(q \| p_t) = \mathsf{LK}(q_t \| p) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

Interestingly, the dynamics (1.25) also decreases the reverse KL divergence:

$$\begin{aligned} \frac{\mathrm{d}\mathsf{LK}}{\mathrm{d}t} &= -\int \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \partial_t q_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} = -\int \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \nabla \cdot [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)] \, \mathrm{d}\mathbf{x} = \int [q_t^2 (\mathbf{s}_{q_t} - \mathbf{s}_p)] \cdot \nabla \frac{p(\mathbf{x})}{q_t(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\ &= -\int p q_t \|\mathbf{s}_{q_t} - \mathbf{s}_p\|_2^2 \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Tabak, E. G. and E. Vanden-Eijnden (2010). "Density estimation by dual ascent of the log-likelihood". *Communications in Mathematical Sciences*, vol. 8, no. 1, pp. 217–233.

Example 1.81: Score matching as gradient flow of KL (e.g., Lyu 2009)

The Fisher divergence between two densities is the square L_2 distance between their score functions:

$$\mathsf{F}(p \| q) := \mathbb{E}_{\mathsf{X} \sim p} \|\mathbf{s}_p(\mathsf{X}) - \mathbf{s}_q(\mathsf{X})\|_2^2.$$

Recall from Example 1.79 that the Gaussian density satisfies the heat equation (1.23). Thus, upon convolu-

tion

$$p_t = p * \mathcal{N}(0, 2t) \quad \text{and} \quad q_t = q * \mathcal{N}(0, 2t)$$

also satisfy the heat equation. We are now ready to prove the following result due to Lyu (2009):

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(p_t \| q_t) = -\mathsf{F}(p_t \| q_t).} \tag{1.26}$$

Indeed, expanding the first term:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(p_t\|q_t) = \int \left[\partial_t p_t \cdot (\log \tfrac{p_t}{q_t}) + p_t \frac{\partial_t p_t}{p_t} - p_t \frac{\partial_t q_t}{q_t}\right] \mathrm{d}\mathbf{x} &= \left\langle \Delta p_t; \log \tfrac{p_t}{q_t} \right\rangle + \partial_t \int p_t \,\mathrm{d}\mathbf{x} - \left\langle \Delta q_t; \tfrac{p_t}{q_t} \right\rangle \\
&= -\left\langle \nabla p_t; \nabla \log \tfrac{p_t}{q_t} \right\rangle + 0 + \left\langle \nabla q_t; \nabla \tfrac{p_t}{q_t} \right\rangle \\
&= -\left\langle p_t \tfrac{\nabla p_t}{p_t}; \nabla \log \tfrac{p_t}{q_t} \right\rangle + \left\langle p_t \tfrac{\nabla q_t}{q_t}; \tfrac{\nabla \frac{p_t}{q_t}}{\frac{p_t}{q_t}} \right\rangle \\
&= -\left\langle p_t(\mathbf{s}_{p_t} - \mathbf{s}_{q_t}); \mathbf{s}_{p_t} - \mathbf{s}_{q_t} \right\rangle = -\mathsf{F}(p_t\|q_t),
\end{aligned}$$

where we recall the score $\mathbf{s}_p := \nabla \log p = \frac{\nabla p}{p}$.

The relation (1.26), after changing the base measure to $q_t \,\mathrm{d}\mathbf{x}$, is known as the de Brujin identity (Stam 1959). Extension to the f-divergence is immediate (Valero-Toranzo et al. 2018).

Lyu, S. (2009). "Interpretation and generalization of score matching". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 359–366.

Stam, A. J. (1959). "Some inequalities satisfied by the quantities of information of Fisher and Shannon". *Information and Control*, vol. 2, no. 2, pp. 101–112.

Valero-Toranzo, I., S. Zozor, and J.-M. Brossier (2018). "Generalization of the de Bruijn Identity to General ϕ-Entropies and ϕ-Fisher Informations". *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6743–6758.

Definition 1.82: Denoising auto-encoder (Vincent 2011; Vincent et al. 2010)

Consider the joint densities $p(\mathbf{x}, \mathbf{z})$ and $q(\mathbf{x}, \mathbf{z})$, with marginals $p(\mathbf{x})$ and $q(\mathbf{x})$, respectively. Their Fisher divergence can be decomposed as:

$$\begin{aligned}
\mathsf{F}\big(p(\mathbf{x},\mathbf{z})\|q(\mathbf{x},\mathbf{z})\big) &= \tfrac{1}{2}\mathbb{E}_{(X,Z)\sim p}[\|\nabla_{\mathbf{x}} \log q(X,Z) - \nabla_{\mathbf{x}} \log p(X,Z)\|_2^2 + \|\nabla_{\mathbf{z}} \log q(X,Z) - \nabla_{\mathbf{z}} \log p(X,Z)\|_2^2] \\
&= \tfrac{1}{2}\mathbb{E}_{(X,Z)\sim p}[\|\nabla_{\mathbf{x}} \log q(X|Z) - \nabla_{\mathbf{x}} \log p(X|Z)\|_2^2 + \|\nabla_{\mathbf{z}} \log q(Z|X) - \nabla_{\mathbf{z}} \log p(Z|X)\|_2^2] \\
&= \mathbb{E}\mathsf{F}\big(p(\mathbf{x}|\mathbf{z})\|q(\mathbf{x}|\mathbf{z})\big) + \mathbb{E}\mathsf{F}\big(p(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}|\mathbf{x})\big).
\end{aligned}$$

In particular, we have

$$\begin{aligned}
\mathsf{F}\big(p(\mathbf{x})\|q(\mathbf{x})\big) &= \tfrac{1}{2}\mathbb{E}_{X\sim p}\|\nabla_{\mathbf{x}} \log q(X) - \nabla_{\mathbf{x}} \log p(X)\|_2^2 \\
&= \mathsf{F}\big(p(\mathbf{x},\mathbf{z})\|q(\mathbf{x})p(\mathbf{z}|\mathbf{x})\big) \\
&= \tfrac{1}{2}\mathbb{E}_{(X,Z)\sim p}\|\nabla_{\mathbf{x}} \log q(X|Z) - \nabla_{\mathbf{x}} \log p(X|Z)\|_2^2 \\
&= \tfrac{1}{2}\mathbb{E}_{(X,Z)\sim p}\|\nabla_{\mathbf{x}} \log q(X) - \nabla_{\mathbf{x}} \log p(X|Z)\|_2^2 + \\
&\quad + \tfrac{1}{2}\mathbb{E}_{(X,Z)\sim p}\|\nabla_{\mathbf{x}} \log p(X)\|_2^2 - \|\nabla_{\mathbf{x}} \log p(X|Z)\|_2^2,
\end{aligned}$$

where the last equality follows from the fact that

$$\begin{aligned}
\mathbb{E}_{(X,Z)\sim p}\left\langle \nabla_{\mathbf{x}} \log q(X), \nabla_{\mathbf{x}} \log p(X|Z) \right\rangle &= \mathbb{E}_{(X,Z)\sim p}\left\langle \nabla_{\mathbf{x}} \log q(X), \nabla_{\mathbf{x}} \log p(X,Z) \right\rangle \\
&= \int \left\langle \nabla_{\mathbf{x}} \log q(\mathbf{x}), \nabla_{\mathbf{x}} p(\mathbf{x},\mathbf{z}) \right\rangle \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z} \\
&= \int \left\langle \nabla_{\mathbf{x}} \log q(\mathbf{x}), \nabla_{\mathbf{x}} \int p(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z} \right\rangle \mathrm{d}\mathbf{x}
\end{aligned}$$

$$= \int \langle \nabla_{\mathbf{x}} \log q(\mathbf{x}), \nabla_{\mathbf{x}} p(\mathbf{x}) \rangle \, \mathrm{d}\mathbf{x}.$$

Vincent, P. (2011). "A Connection Between Score Matching and Denoising Autoencoders". *Neural Computation*, vol. 23, no. 7, pp. 1661–1674.
Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". *Journal of Machine Learning Research*, vol. 11, no. 110, pp. 3371–3408.

Example 1.83: Neural ODE (Chen et al. 2018)

The popular residual network (He et al. 2016)

$$\mathsf{X}_{t+1} = \mathsf{X}_t + \mathbf{f}(t, \mathsf{X}_t, \mathbf{u}_t)$$

can be treated as an Euler discretization of the ODE:

$$\mathrm{d}\mathsf{X}_t = \mathbf{f}(t, \mathsf{X}_t, \mathbf{u}_t), \tag{1.27}$$

where $\mathbf{u}_t$, the weights of the network, will play the role of a control variable. Consider the control minimization problem (Liberzon 2012)

$$\min_{\mathbf{u}} \; g(\mathbf{x}_1) + \int_0^1 \ell(t, \mathbf{x}_t, \mathbf{u}_t) \, \mathrm{d}t, \quad \text{s.t.} \quad (1.27),$$

where X_0 is our training data. To train the network, we need to compute the derivative w.r.t. $\mathbf{u}$, for which we introduce the Lagrangian multiplier p_t and consider

$$L(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) := g(\mathbf{x}_1) + \int_0^1 (\langle \mathbf{p}_t, \dot{\mathbf{x}}_t \rangle - H(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{p}_t)) \, \mathrm{d}t, \text{ where the Hamilton } H(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{p}_t) := \langle \mathbf{p}, \mathbf{f}(t, \mathbf{x}, \mathbf{u}) \rangle .$$

Let us perturb $\tilde{\mathbf{u}} = \mathbf{u} + \epsilon \mathbf{v}$ and obtain $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \mathbf{y} + o(\epsilon)$. Then,

$$\frac{\mathrm{d}L}{\mathrm{d}\epsilon} \restriction_{\epsilon=0} = \langle \nabla g(\mathbf{x}_1) + \mathbf{p}_1, \mathbf{y}_1 \rangle - \int_0^1 \langle \dot{\mathbf{p}} + H_{\mathbf{x}}(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{p}_t), \mathbf{y} \rangle + \langle H_{\mathbf{u}}(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{p}_t), \mathbf{v} \rangle \, \mathrm{d}t,$$

which motivates us to choose the Lagrangian multiplier (a.k.a. adjoint) $\mathbf{p}_t$ as follows:

$$\begin{aligned} \dot{\mathbf{x}}_t &= H_{\mathbf{p}}, \text{ with initial data } \mathbf{x}_0 \\ \dot{\mathbf{p}}_t &= -H_{\mathbf{x}}, \text{ with end momentum } \mathbf{p}_1 = -\nabla g(\mathbf{x}_1) \\ \nabla_{\mathbf{u}} L &= -H_{\mathbf{u}}. \end{aligned}$$

When the control $\mathbf{u}_t \equiv \mathbf{u}$, we obtain

$$\nabla_{\mathbf{u}} L = -\int_0^1 H_{\mathbf{u}} \, \mathrm{d}t = -\int_0^1 \nabla_{\mathbf{u}} \mathbf{f}(t, \mathbf{x}_t, \mathbf{u}) \cdot \mathbf{p} \, \mathrm{d}t.$$

After training $\mathbf{u}$, Chen et al. (2018) also showed the following:

$$\frac{\mathrm{d} \log p_t(\mathbf{x}_t)}{\mathrm{d}t} = \tfrac{1}{p_t}[\partial_t p_t(\mathbf{x}_t) + \langle \nabla_{\mathbf{x}} p_t(\mathbf{x}_t), \mathbf{f}_t(\mathbf{x}_t) \rangle] = \tfrac{1}{p_t}[-\nabla \cdot (p_t \mathbf{f}_t) + \langle \nabla_{\mathbf{x}} p_t(\mathbf{x}_t), \mathbf{f}_t(\mathbf{x}_t) \rangle] = -\nabla \cdot \mathbf{f}_t(\mathbf{x}_t),$$

which can be useful in computing the log-likelihood along the trajectory of $\mathbf{x}_t$.

Chen, T. Q., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural ordinary differential equations". In: *Advances in Neural Information Processing Systems*, pp. 6572–6583.
He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
Liberzon, D. (2012). "Calculus of Variations and Optimal Control Theory: A Concise Introduction". Princeton University Press.
Chen, S., S. Ding, Y. Karayiannidis, and M. Björkman (2023). "Learning Continuous Normalizing Flows For Faster Convergence To Target Distribution via Ascent Regularizations". In: *The Eleventh International Conference on Learning Representations*.

Example 1.84: Variational optimal transport (Benamou and Brenier 2000)

Let ρ_0 and ρ_1 be two given densities, with Wasserstein distance

$$\mathbb{W}_2^2(\rho_0,\rho_1)=\min_{\mathsf{T}:\mathsf{T}_{\#}\rho_0=\rho_1}\mathbb{E}_{\mathsf{X}\sim\rho_0}\|\mathsf{T}\mathsf{X}-\mathsf{X}\|_2^2,\quad\text{where}\quad\rho_0(\mathbf{x})=\rho_1(\mathsf{T}\mathbf{x})\cdot\det\mathsf{T}'\mathbf{x}.$$

Brenier (1991) proved the existence of such a (unique) $\mathsf{T}=\nabla\varphi$ for some convex potential function φ.

Let us consider the ODE:

$$\mathrm{d}\mathsf{X}_t(\mathbf{x})=\mathbf{b}_t(\mathsf{X}_t),\quad\text{where}\quad\mathsf{X}_0=\mathbf{x}.$$

When $\mathsf{X}_0\sim p_0$, the intermediate densities $\mathsf{X}_t\sim p_t$ satisfy the continuity equation and boundary condition:

$$\partial_t p_t=-\nabla\cdot(p_t\mathbf{b}_t),\tag{1.28}$$

$$p_0=\rho_0,\ p_1=\rho_1.\tag{1.29}$$

In fact, Benamou and Brenier (2000) showed the following surprising result:

$$\begin{aligned}\int\int_0^1 p_t(\mathbf{x})\|\mathbf{b}_t(\mathbf{x})\|_2^2\,\mathrm{d}t\,\mathrm{d}\mathbf{x}&=\int\int_0^1\|\mathbf{b}_t(\mathsf{X}_t(\mathbf{x}))\|_2^2\,\mathrm{d}t\cdot p_0(\mathbf{x})\,\mathrm{d}\mathbf{x}=\int\int_0^1\|\,\mathrm{d}\mathsf{X}_t(\mathbf{x})\|_2^2\,\mathrm{d}t\cdot p_0(\mathbf{x})\,\mathrm{d}\mathbf{x}\\&\geq\int p_0(\mathbf{x})\|\mathsf{X}_1-\mathbf{x}\|_2^2\,\mathrm{d}\mathbf{x}\geq\mathbb{W}_2^2(\rho_0,\rho_1).\end{aligned}$$

The first inequality is attained if X_t is affine in t while the second is attained if $\mathsf{X}_1(\mathbf{x})=\nabla\varphi(\mathbf{x})$. Thus,

$$\begin{aligned}\mathbb{W}_2^2(\rho_0,\rho_1)&=\min_{p_t,\mathbf{b}_t}\int\int_0^1 p_t(\mathbf{x})\|\mathbf{b}_t(\mathbf{x})\|_2^2\,\mathrm{d}t\,\mathrm{d}\mathbf{x},\quad\text{s.t.}\quad(1.28)\text{ and }(1.29)\\&=\min_{p_t\ \text{s.t.}\ (1.29)}\int_0^1\min_{\mathbf{b}_t\ \text{s.t.}\ (1.28)}\int p_t(\mathbf{x})\|\mathbf{b}_t(\mathbf{x})\|_2^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\end{aligned}$$

Let us examine the minimizer $\mathbf{b}_t$ above. Consider the objective with slight perturbation:

$$f(\mathbf{b}_t+\epsilon\mathbf{d}_t):=\int p_t(\mathbf{x})\|\mathbf{b}_t(\mathbf{x})+\epsilon\mathbf{d}_t(\mathbf{x})\|_2^2\,\mathrm{d}\mathbf{x},\quad\text{s.t.}\quad\partial_t\mathbf{p}_t=-\nabla\cdot(p_t(\mathbf{b}_t+\epsilon\mathbf{d}_t))=-\nabla\cdot(p_t\mathbf{b}_t)-\epsilon\nabla\cdot(p_t\mathbf{d}_t).$$

To maintain the continuity equation we must have $\nabla\cdot(p_t\mathbf{d}_t)=0$. Take derivative w.r.t. ϵ at 0:

$$\frac{\mathrm{d}f}{\mathrm{d}\epsilon}\restriction_{\epsilon=0}=2\int p_t(\mathbf{x})\mathbf{d}_t(\mathbf{x})\cdot\mathbf{b}_t(\mathbf{x})\,\mathrm{d}\mathbf{x}=0.$$

Thus, $\mathbf{b}_t\perp(p_t\mathbf{d}_t)$ for any $p_t\mathbf{d}_t$ such that $\nabla\cdot(p_t\mathbf{d}_t)=0$. According to the Helmholtz decomposition, there exists a potential function φ such that the minimizer

$$\mathbf{b}_t=\nabla\varphi_t.$$

Therefore, we obtain the following nice representation of the Wasserstein distance:

$$\mathbb{W}_2^2(\rho_0,\rho_1)=\min_{p_t}\int_0^1\|\partial_t p_t\|_{p_t}^2\,\mathrm{d}t,\quad\text{s.t.}\quad\text{boundary condition (1.29) holds}\tag{1.30}$$

$$\|\partial_t p_t\|_{p_t}^2=\min_{\varphi_t}\int p_t(\mathbf{x})\|\nabla\varphi_t(\mathbf{x})\|_2^2\,\mathrm{d}\mathbf{x},\quad\text{s.t.}\quad\partial_t p_t=-\nabla\cdot(p_t\nabla\varphi_t),$$

where the formula (1.30) is strikingly similar to the geodesic distance on a Riemannian manifold!

Benamou, J.-D. and Y. Brenier (2000). "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". *Numerische Mathematik*, vol. 84, pp. 375–393.

Brenier, Y. (1991). "Polar factorization and monotone rearrangement of vector-valued functions". *Communications on Pure and Applied Mathematics*, vol. 44, no. 4, pp. 375–417.

Definition 1.85: Wasserstein gradient (Otto 2001)

We can now define the following Wasserstein inner product on two functions h_1 and h_2 with $\int h_1 = \int h_2 = 0$:

$$\langle h_1, h_2 \rangle_p := \int \langle \nabla\varphi_1, \nabla\varphi_2 \rangle \cdot p \, \mathrm{d}\mathbf{x}, \quad \text{where} \quad \nabla \cdot (p\nabla\varphi_i) = -h_i.$$

(The condition $\int h_i = 0$ is needed in order for the continuity equation to have a solution φ_i. Recall also that $\int \partial_t p_t = \partial_t \int p_t = 0$.)

For a function $f : \mathcal{P}_2 \to \mathbb{R}$, we can define its Wasserstein gradient as the representation of the derivative w.r.t. the Wasserstein inner product:

$$\langle \nabla_{\mathbb{W}_2} f(p), \partial_t p_t \restriction_{t=0} \rangle_p = \frac{\mathrm{d} f(p_t)}{\mathrm{d}t} \restriction_{t=0},$$

where $p_t : (-\epsilon, \epsilon) \to \mathcal{P}_2$ is any smooth curve with $p_0 = p$. Suppose f admits a variational gradient (in $L_2(\mathrm{d}\mathbf{x})$) such that:

$$\frac{\mathrm{d} f(p_t)}{\mathrm{d}t} \restriction_{t=0} = \langle \nabla_{L_2} f(p), \partial_t p_t \restriction_{t=0} \rangle_{L_2(\mathrm{d}\mathbf{x})} = -\langle \nabla_{L_2} f(p), \nabla \cdot (p\nabla\varphi) \rangle_{L_2(\mathrm{d}\mathbf{x})} = \int \langle \nabla\nabla_{L_2} f(p), \nabla\varphi \rangle \cdot p \, \mathrm{d}\mathbf{x},$$

where $\nabla \cdot (p\nabla\varphi) = -\partial_t p_t \restriction_{t=0}$. Thus, comparing to the definition of the Wasserstein inner product, we know

$$\boxed{\nabla_{\mathbb{W}_2} f(p) = -\nabla \cdot (p \nabla \nabla_{L_2} f(p)).} \tag{1.31}$$

For a much more rigorous and thorough discussion, see Ambrosio et al. (2021, 2008).

Otto, F. (2001). "The geometry of dissipative evolution equations: the porous medium equation". *Communications in Partial Differential Equations*, vol. 26, no. 1-2, pp. 101–174.
Ambrosio, L., E. Brué, and D. Semola (2021). "Lectures on Optimal Transport". Springer.
Ambrosio, L., N. Gigli, and G. Savaré (2008). "Gradient Flows in Metric Spaces and in the Space of Probability Measures". 2nd. Springer.

Example 1.86: Typical Wasserstein gradients

We have reduced the computation of the Wasserstein gradient to its L_2 counterpart. Let us consider the following important example:

$$f(p) = \int u(p(\mathbf{x})) \, \mathrm{d}\mathbf{x}, \tag{1.32}$$

where $u : \mathbb{R}_+ \to \mathbb{R}$ is a smooth function. Assuming we may switch differentiation and integration,

$$\frac{\mathrm{d} f(p_t)}{\mathrm{d}t} \restriction_{t=0} = \int u'(p(\mathbf{x})) \partial_t p_t \restriction_{t=0} \, \mathrm{d}\mathbf{x} = \langle u'(p), \partial_t p_t \restriction_{t=0} \rangle_{L_2(\mathrm{d}\mathbf{x})} \implies \boxed{\nabla_{L_2} f(p) = u'(p)}$$

$$\boxed{\nabla_{\mathbb{W}_2} f(p) = -\nabla \cdot (p u''(p) \nabla p).}$$

When $u(p) = p \log p - p$ (negative entropy), we obtain $\nabla_{\mathbb{W}_2} f(p) = -\Delta p$. In other words, the heat equation in Example 1.79 can also be seen as the Wasserstein gradient flow:

$$\frac{\mathrm{d} p_t}{\mathrm{d}t} = -\nabla_{\mathbb{W}_2} f(p_t).$$

We leave the following computation as an exercise:

- $f(p) = \int u(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \implies \nabla_{L_2} f = u \implies \nabla_{\mathbb{W}_2} f = -\nabla \cdot (p \nabla u)$.
- $f(p) = \frac{1}{2} \int u(\mathbf{x} - \mathbf{y}) p(\mathbf{x}) p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}, u = u(-\cdot) \implies \nabla_{L_2} f = u * p \implies \nabla_{\mathbb{W}_2} f = -\nabla \cdot (p(p * \nabla u))$.

Figalli, A. and F. Glaudo (2021). "An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows". European Mathematical Society.

Definition 1.87: Displacement convexity (McCann 1997)

We say $f : \mathcal{P}_2 \to \mathbb{R}$ is λ-$\mathbb{W}_2$-convex (or displacement convex), if

$$[0,1] \ni t \mapsto f(p_t)$$

is λ-convex for any $\mathbb{W}_2$-geodesic $p : [0,1] \to \mathcal{P}_2$. For the integral function (1.32), exploiting the relation

$$p_t = [(1-t)\mathrm{Id} + t\nabla\varphi]_{\#}p_0 =: (\mathsf{T}_t)_{\#}p_0, \ \textit{i.e.}, \ p_t(\mathsf{T}_t\mathbf{x}) \det \mathsf{T}_t'(\mathbf{x}) = p_0(\mathbf{x})$$
$$f(p_t) = \int u(p_t(\mathbf{x}))\,\mathrm{d}\mathbf{x} = \int u(p_0(\mathbf{x})/\det \mathsf{T}_t'(\mathbf{x}))\det \mathsf{T}_t'(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad \text{where} \quad \mathsf{T}_t' = (1-t)\mathrm{Id} + t\nabla^2\varphi.$$

From Exercise 1.88 we know $t \mapsto f(p_t)$ is convex, i.e., f is $\mathbb{W}_2$-convex, if $u : \mathbb{R}_+ \to \mathbb{R}$ is convex.
Similarly, the following can be verified:

- $f(p) = \int u(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$ is λ-$\mathbb{W}_2$-convex if u is λ-convex.
- $f(p) = \int u(\mathbf{x}-\mathbf{y})p(\mathbf{x})p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$ is $\mathbb{W}_2$-convex if u is convex.

McCann, R. J. (1997). "A Convexity Principle for Interacting Gases". *Advances in Mathematics*, vol. 128, no. 1, pp. 153–179.

Exercise 1.88: Convexity

Prove the following:

- $(0,1) \ni t \mapsto \det^{1/d}[(1-t)\mathrm{Id} + tA]$ is concave for symmetric positive semidefinite $A \in \mathbb{R}^{d\times d}$.
- $\mathbb{R}_{++} \ni s \mapsto s^d u(1/s^d)$ is decreasing convex if u is convex.

Example 1.89: FPK as gradient flow (Jordan et al. 1998)

Consider the SDE

$$\mathrm{d}\mathsf{X}_t = -\nabla\varphi(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}\mathsf{B}_t,$$

and the FPK equation

$$\partial_t p = \nabla\cdot(p\nabla\varphi) + \beta^{-1}\Delta p = \nabla\cdot[p(\nabla\varphi + \beta^{-1}s)], \tag{1.33}$$

where $s(t,\mathbf{x}) = \nabla_{\mathbf{x}}\log p(t,\mathbf{x})$ is the score function. When the potential φ satisfies appropriate growth conditions, there is a unique stationary solution of FPK (the so-called Gibbs distribution):

$$s = -\beta\nabla\varphi \iff p \propto \exp(-\beta\varphi).$$

Comparing to the Wasserstein gradient formula (1.31), we are motivated to consider the Lyapunov function

$$f(p) = \int p\varphi + \beta^{-1}p\log p - \beta^{-1}p\,\mathrm{d}\mathbf{x} = \beta^{-1}\mathsf{KL}(p\|q) + \beta^{-1}c, \quad \text{where} \quad q \propto \exp(-\beta\varphi).$$

Then, the FPK equation (1.33) becomes the Wasserstein gradient flow:

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla_{\mathbb{W}_2} f(p_t).$$

Assuming φ is λ-convex, we have

$$\frac{\mathrm{d}f(p_t)}{\mathrm{d}t} = \langle \nabla_{\mathbb{W}_2} f(p_t), \partial_t p_t\rangle_{p_t} = -\langle \nabla_{\mathbb{W}_2} f(p_t), \nabla_{\mathbb{W}_2} f(p_t)\rangle_{p_t} \le -2\beta\lambda f(p_t) \implies \boxed{f(p_t) \le e^{-2\beta\lambda t} f(p_0)},$$

whence follows the following bound:

$$\mathbb{W}_2^2(p_t, q) \le \tfrac{2}{\lambda} f(p_t) \le \tfrac{2}{\lambda} e^{-2\beta\lambda t} f(p_0)$$
$$\tfrac{1}{2}\|p_t - q\|_1^2 \le \beta f(p_t) \le \beta e^{-2\beta\lambda t} f(p_0).$$

More generally, contractivity (Ambrosio et al. 2008) allows one to prove

$$\mathbb{W}_2^2(p_t, \rho_t) \le e^{-2\beta\lambda t} \mathbb{W}_2^2(p_0, \rho_0).$$

Jordan, R., D. Kinderlehrer, and F. Otto (1998). "The Variational Formulation of the Fokker–Planck Equation". *SIAM Journal on Mathematical Analysis*, vol. 29, no. 1, pp. 1–17.
Ambrosio, L., N. Gigli, and G. Savaré (2008). "Gradient Flows in Metric Spaces and in the Space of Probability Measures". 2nd. Springer.

Remark 1.90: Digesting λ-$\mathbb{W}_2$-convexity

From the definition we know $f : \mathcal{P}_2 \to \mathbb{R}$ is λ-$\mathbb{W}_2$-convexity iff

$$f(p_t) + \tfrac{\lambda}{2} t(1-t)\mathbb{W}_2^2(p_0, p_1) \le (1-t)f(p_0) + tf(p_1).$$

We can now apply the usual convex calculus. For instance, dividing t and then letting $t \downarrow 0$ we obtain

$$\langle \nabla_{\mathbb{W}_2} f(p_0), \partial_t p_t \restriction_{t=0} \rangle_{p_0} + \tfrac{\lambda}{2}\mathbb{W}_2^2(p_0, p_1) \le f(p_1) - f(p_0).$$

We are now ready to prove the logarithmic Sobolev inequality (Ledoux 2001). Consider the Gibbs density $q \propto \exp(-\beta\varphi)$ for some λ-convex φ. Then,

$$\beta f(p) := \mathsf{KL}(p\|q) \le \tfrac{1}{2\lambda} \langle \nabla_{\mathbb{W}_2} f(p), \nabla_{\mathbb{W}_2} f(p) \rangle_p = \tfrac{1}{2\lambda} \int \|\nabla\varphi + \beta^{-1}\mathbf{s}_p\|_2^2 p \,\mathrm{d}\mathbf{x} = \tfrac{1}{2\lambda\beta^2} \int \|\mathbf{s}_q - \mathbf{s}_p\|_2^2 p \,\mathrm{d}\mathbf{x}.$$

To put in a more succint and familiar form:

$$\boxed{\mathsf{KL}(p\|q) \le \tfrac{1}{2\lambda\beta^2}\mathsf{F}(p\|q)}, \quad \text{where} \quad q \propto \exp(-\beta\varphi) \text{ for some } \lambda\text{-convex } \varphi.$$

Ledoux, M. (2001). "The Concentration of Measure Phenomenon". American Mathematical Society.

Example 1.91: Langevin revisited

Let us apply the above theory to the Lagenvin equation:

$$\mathrm{d}\mathsf{X}_t = -b\mathsf{X}_t \,\mathrm{d}t + \sigma \,\mathrm{d}\mathsf{B}_t.$$

We have claimed before that $\mathsf{X}_t \to \mathcal{N}(0, \frac{\sigma^2}{2b})$ as $t \to \infty$.

We identity $\varphi(x) = \frac{b}{2}x^2$ and $\beta = \frac{2}{\sigma^2}$. Indeed, the (unique) stationary density $q = \mathcal{N}(0, \frac{\sigma^2}{2b})$. Moreover, we know now that the convergence is linear:

$$\mathbb{W}_2^2(p_t, q) \le \tfrac{2}{b} e^{-4bt/\sigma^2} f(p_0)$$
$$\tfrac{1}{2}\|p_t - q\|_1^2 \le \tfrac{2}{\sigma^2} e^{-4bt/\sigma^2} f(p_0).$$