What are ... generative probability flows for sampling?

Viktor Stein, Technical University of Berlin

"What is ... ?" Seminar

07.11.2025

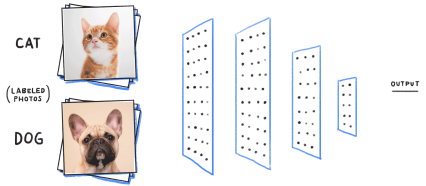
1. Two types of generative modelling

2. Gradient flows - from Euclidean to metric

3. Optimal transport and the Wasserstein metric

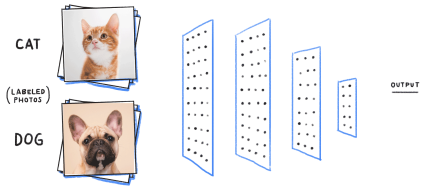
4. Wasserstein gradient flow

Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.



Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

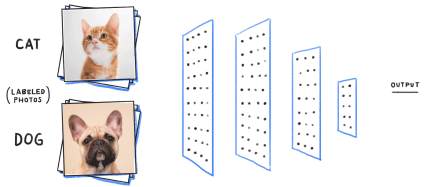
Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.



Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

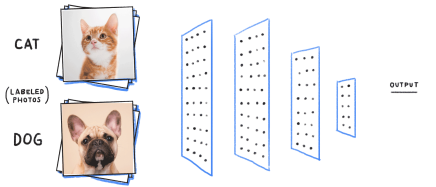
Goal: learn best *distribution* of parameter θ to fit D.



Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

Goal: learn best *distribution* of parameter θ to fit D.



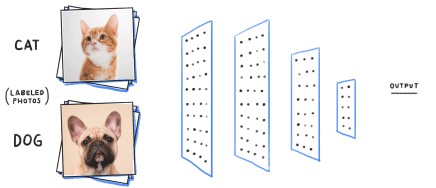
Solution: Since $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, the *likelihood* is

$$p(D \mid \theta) = \prod_{i=1}^m p(y_i \mid \theta, w_i) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2\sigma^2}\|y_i - g_\theta(w_i)\|_2^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m \|y_i - g_\theta(w_i)\|_2^2\right), \quad \theta \in \Theta.$$

Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

Goal: learn best *distribution* of parameter θ to fit D.



Solution: Since $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, the *likelihood* is

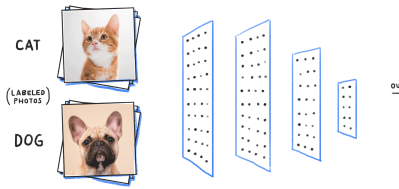
$$p(D \mid \theta) = \prod_{i=1}^m p(y_i \mid \theta, w_i) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2\sigma^2}\|y_i - g_\theta(w_i)\|_2^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m \|y_i - g_\theta(w_i)\|_2^2\right), \quad \theta \in \Theta.$$

After *choosing* prior p on θ,

Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

Goal: learn best *distribution* of parameter θ to fit D.



Solution: Since $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, the *likelihood* is

$$p(D \mid \theta) = \prod_{i=1}^m p(y_i \mid \theta, w_i) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2\sigma^2}\|y_i - g_\theta(w_i)\|_2^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m \|y_i - g_\theta(w_i)\|_2^2\right), \quad \theta \in \Theta.$$

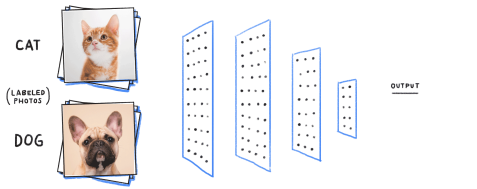
After *choosing* prior p on θ, **Bayes' rule** yields

$$\pi(\theta) \coloneqq p(\theta \mid D) \propto p(D \mid \theta)p(\theta) =: e^{-V(\theta)},$$

Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

Goal: learn best *distribution* of parameter θ to fit D.



Solution: Since $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, the *likelihood* is

$$p(D \mid \theta) = \prod_{i=1}^m p(y_i \mid \theta, w_i) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2\sigma^2}\|y_i - g_\theta(w_i)\|_2^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m \|y_i - g_\theta(w_i)\|_2^2\right), \quad \theta \in \Theta.$$

After *choosing* prior p on θ, **Bayes' rule** yields

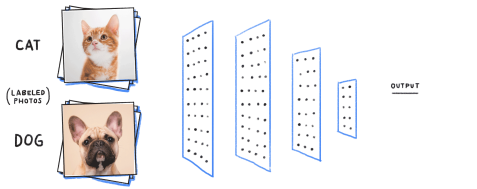
$$\pi(\theta) \coloneqq p(\theta \mid D) \propto p(D \mid \theta)p(\theta) =: e^{-V(\theta)},$$

where V **known, normalization unknown**. e^{-V} is called **Boltzmann density** (Chemseddine, Wald, et al. 2024).

Given: labelled i.i.d. data $D \coloneqq \big((w_i, y_i)\big)_{i=1}^m$.

Assume underlying model $y = g_\theta(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, and, e.g., g_θ is neural network with weights $\theta \in \Theta \subset \mathbb{R}^d$.

Goal: learn best *distribution* of parameter θ to fit D.



Solution: Since $\xi \sim \mathcal{N}(0, \sigma^2\,\mathrm{id})$, the *likelihood* is

$$p(D \mid \theta) = \prod_{i=1}^m p(y_i \mid \theta, w_i) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2\sigma^2}\|y_i - g_\theta(w_i)\|_2^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m \|y_i - g_\theta(w_i)\|_2^2\right), \quad \theta \in \Theta.$$

After *choosing* prior p on θ, **Bayes' rule** yields

$$\pi(\theta) \coloneqq p(\theta \mid D) \propto p(D \mid \theta)p(\theta) =: e^{-V(\theta)},$$

where V **known, normalization unknown**. e^{-V} is called **Boltzmann density** (Chemseddine, Wald, et al. 2024). $\rightsquigarrow$ Need π to predict new output $y_{new} = \int g_\theta(w_{new})\,\mathrm{d}\pi(\theta)$.

you have *some* images of cats $(\tilde{\theta}_j)_{j=1}^M$

Standard problem: you have *some* images of cats $(\tilde{\theta}_j)_{j=1}^M$, but that's not enough!

Standard problem: you have *some* images of cats $(\tilde{\theta}_j)_{j=1}^M$, but that's not enough!

How do we get more and how do we model this mathematically?

Standard problem: you have *some* images of cats $(\tilde{\theta}_j)_{j=1}^M$, but that's not enough!

How do we get more and how do we model this mathematically?

We assume that all images of cats (of fixed size, say with RGB values) follow some probability distribution π.

Standard problem: you have *some* images of cats $(\tilde{\theta}_j)_{j=1}^M$, but that's not enough!

How do we get more and how do we model this mathematically?

We assume that all images of cats (of fixed size, say with RGB values) follow some probability distribution π.

Another example. π represents hand-written digits (784-dimensional)

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

Solution: minimize $D(\cdot \mid \pi)\colon \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

Solution: minimize $D(\cdot \mid \pi)\colon \mathcal{P}(\mathbb{R}^d) \to [0,\infty]$, where

- $D(\cdot \mid \cdot)$ says how different both inputs are,

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

Solution: minimize $D(\cdot \mid \pi)\colon \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$, where

- $D(\cdot \mid \cdot)$ says how different both inputs are,
- $D(\mu \mid \pi) = 0 \iff \mu = \pi$ ("divergence" property).

To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

Solution: minimize $D(\cdot \mid \pi)\colon \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$, where

- $D(\cdot \mid \cdot)$ says how different both inputs are,
- $D(\mu \mid \pi) = 0 \iff \mu = \pi$ ("divergence" property).

with some gradient flow in the space of probability measures (Chen et al. 2018)

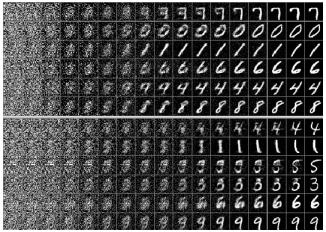
To generate new outputs (or quantify uncertainty), we want to "sample" from π, from which we only have samples, or we know V with $\pi \propto e^{-V}$.

$\rightsquigarrow$ Want $(\theta_j)_{j=1}^N \subset \mathbb{R}^d$ such that the *empirical measure* $\frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ **approximates** π **well** for $N \to \infty$.

Solution: minimize $D(\cdot \mid \pi)\colon \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$, where

- $D(\cdot \mid \cdot)$ says how different both inputs are,
- $D(\mu \mid \pi) = 0 \iff \mu = \pi$ ("divergence" property).

with some gradient flow in the space of probability measures (Chen et al. 2018) (image from (Hertrich, Wald, et al. 2024)).



Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

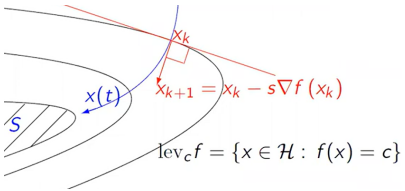
$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$

For generative modelling, the loss would be something like $\Phi = \| \cdot - \pi \|_p^q$.

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$

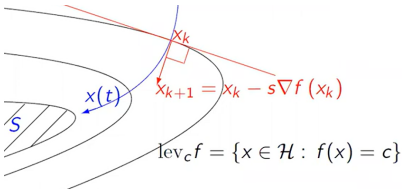


For generative modelling, the loss would be something like $\Phi = \|\cdot - \pi\|_p^q$.

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$



Energy dissipation equality: If u solves (1), then

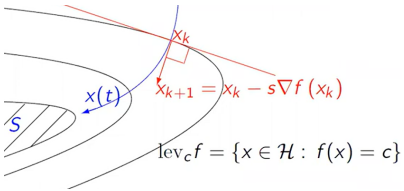
$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = \nabla\Phi(u(t)) \cdot u'(t)$$

For generative modelling, the loss would be something like $\Phi = \|\cdot - \pi\|_p^q$.

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$



Energy dissipation equality: If u solves (1), then

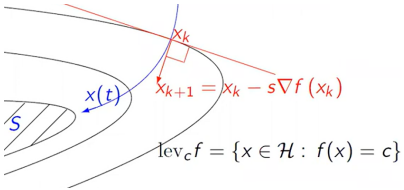
$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = \nabla\Phi(u(t)) \cdot u'(t) \overset{(1)}{=} -\|\nabla\Phi\big(u(t)\big)\|_2^2 \overset{(1)}{=} -\|u'(t)\|_2^2$$

For generative modelling, the loss would be something like $\Phi = \|\cdot - \pi\|_p^q$.

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$



Energy dissipation equality: If u solves (1), then

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = \nabla\Phi(u(t)) \cdot u'(t) \overset{(1)}{=} -\|\nabla\Phi\big(u(t)\big)\|_2^2 \overset{(1)}{=} -\|u'(t)\|_2^2 \leq 0,$$

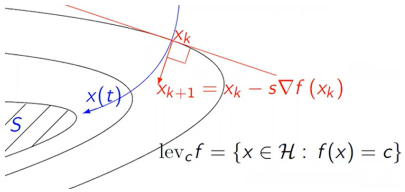
$\implies$ 1) Φ decreases along u.

For generative modelling, the loss would be something like $\Phi = \|\cdot - \pi\|_p^q$.

Definition (Gradient flow)

A *gradient flow* for a differentiable functional $\Phi \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ is a solution $u \in \mathrm{AC}_{\mathrm{loc}}((0,\infty); \mathbb{R}^d)$ to the ordinary initial value problem

$$\begin{cases} u'(t) = -\nabla\Phi\big(u(t)\big), & \text{for almost all } t > 0, \\ \lim_{t \searrow 0} u(t) = u_0. \end{cases} \tag{1}$$



Energy dissipation equality: If u solves (1), then

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = \nabla\Phi(u(t)) \cdot u'(t) \overset{(1)}{=} -\|\nabla\Phi\big(u(t)\big)\|_2^2 \overset{(1)}{=} -\|u'(t)\|_2^2 \leq 0,$$

$\implies$ 1) Φ decreases along u. 2) $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = 0 \iff u(t)$ critical point of Φ.

For generative modelling, the loss would be something like $\Phi = \|\cdot - \pi\|_p^q$.

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\rightsquigarrow$ can be *fully compensated by considering* $\|\nabla\Phi \circ u\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t))$.

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\rightsquigarrow$ can be *fully compensated by considering* $\|\nabla\Phi \circ u\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t))$.

Lemma (De Giorgi's Energy Dissipation principle)

For $\Phi \in \mathcal{C}^1(\mathbb{R}^d;\mathbb{R})$ *and a curve* $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty);\mathbb{R}^d\big)$*, the following are equivalent*

i) u *solves* (1) *for some* $u_0 \in \mathbb{R}^d$*.*

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\rightsquigarrow$ can be *fully compensated by considering* $\|\nabla\Phi \circ u\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t))$.

Lemma (De Giorgi's Energy Dissipation principle)

For $\Phi \in \mathcal{C}^1(\mathbb{R}^d;\mathbb{R})$ and a curve $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty);\mathbb{R}^d\big)$, the following are equivalent

i) u solves (1) for some $u_0 \in \mathbb{R}^d$.

ii) Both $\|u'(t)\|_2 = \|\nabla\Phi(u(t))\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = -\|\nabla\Phi(u(t))\|_2\|u'(t)\|_2$ hold for $t > 0$.

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\rightsquigarrow$ can be *fully compensated by considering* $\|\nabla\Phi \circ u\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t))$.

Lemma (De Giorgi's Energy Dissipation principle)

For $\Phi \in \mathcal{C}^1(\mathbb{R}^d;\mathbb{R})$ and a curve $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty);\mathbb{R}^d\big)$, the following are equivalent

i) u solves (1) for some $u_0 \in \mathbb{R}^d$.

ii) Both $\|u'(t)\|_2 = \|\nabla\Phi(u(t))\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = -\|\nabla\Phi(u(t))\|_2\|u'(t)\|_2$ hold for $t > 0$.

iii) $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) \le -\frac{1}{2}\big\|\nabla\Phi\big(u(t)\big)\big\|_2^2 - \frac{1}{2}\|u'(t)\|_2^2$ for $t > 0$.

$\|u'(t)\|_2 = \|\nabla\Phi\big(u(t)\big)\|_2$ carries less information than $u'(t) = -\nabla\Phi\big(u(t)\big)$.

$\rightsquigarrow$ can be *fully compensated by considering* $\|\nabla\Phi \circ u\|_2$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t))$.

Lemma (De Giorgi's Energy Dissipation principle)

For $\Phi \in \mathcal{C}^1(\mathbb{R}^d;\mathbb{R})$ *and a curve* $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty);\mathbb{R}^d\big)$*, the following are equivalent*

i) u *solves* (1) *for some* $u_0 \in \mathbb{R}^d$.

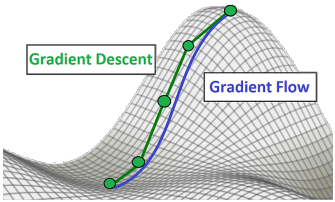
ii) Both $\|u'(t)\|_2 = \|\nabla\Phi(u(t))\|_2$ *and* $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) = -\|\nabla\Phi(u(t))\|_2\|u'(t)\|_2$ *hold for* $t > 0$.

iii) $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t)) \le -\frac{1}{2}\big\|\nabla\Phi\big(u(t)\big)\big\|_2^2 - \frac{1}{2}\|u'(t)\|_2^2$ *for* $t > 0$.

The **scalar** quantities $\|u'(t)\|_2,\ \|\nabla\Phi(u(t))\|_2$

can be given sense when replacing $\mathbb{R}^d$ by a metric space.

$\rightsquigarrow$ new scalar formulation of gradient flow



Let (X, d) be a complete metric space.

Let (X, d) be a complete metric space.

Definition (Absolute continuity)

A curve $u\colon (0,\infty) \to X$ is p-locally absolutely continuous for $p \in [1,\infty]$ and we write $u \in \mathrm{AC}^p_{\mathrm{loc}}\big((0,\infty); X\big)$ if there exists a $m \in L^p_{\mathrm{loc}}(0,\infty)$ with (if $p = 1$, we omit exponent)

$$d\big(u(t), u(s)\big) \leq \int_s^t m(r)\,\mathrm{d}r \qquad \forall 0 < s \leq t < \infty.$$

Let (X, d) be a complete metric space.

Definition (Absolute continuity)

A curve $u\colon (0,\infty) \to X$ is p-locally absolutely continuous for $p \in [1,\infty]$ and we write $u \in \mathrm{AC}^p_{\mathrm{loc}}\big((0,\infty); X\big)$ if there exists a $m \in L^p_{\mathrm{loc}}(0,\infty)$ with (if $p = 1$, we omit exponent)

$$d\big(u(t), u(s)\big) \leq \int_s^t m(r)\,\mathrm{d}r \qquad \forall 0 < s \leq t < \infty.$$

Definition (Metric derivative)

The *metric derivative* of a curve $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty); X\big)$ at $t \in (0,\infty)$ is

$$|u'|(t) \coloneqq \lim_{h \to 0} \frac{d(u(t+h), u(t))}{|h|}.$$

Let (X, d) be a complete metric space.

Definition (Absolute continuity)

A curve $u\colon (0,\infty) \to X$ is p-locally absolutely continuous for $p \in [1,\infty]$ and we write $u \in \mathrm{AC}^p_{\mathrm{loc}}\big((0,\infty); X\big)$ if there exists a $m \in L^p_{\mathrm{loc}}(0,\infty)$ with (if $p = 1$, we omit exponent)

$$d\big(u(t), u(s)\big) \leq \int_s^t m(r)\,\mathrm{d}r \qquad \forall 0 < s \leq t < \infty.$$

Definition (Metric derivative)

The *metric derivative* of a curve $u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty); X\big)$ at $t \in (0,\infty)$ is

$$|u'|(t) \coloneqq \lim_{h\to 0} \frac{d(u(t+h), u(t))}{|h|}.$$

The gradient norm $\|\nabla\Phi\|_2$ is replaced by an *upper gradient*, a function g such that

$$\big|\Phi(u(t)) - \Phi(u(s))\big| \leq \int_s^t g(u(s))|u'|(s)\,\mathrm{d}s \qquad \forall u \in \mathrm{AC}_{\mathrm{loc}}\big((0,\infty); X\big),\ 0 < s \leq t < \infty.$$

The **Wasserstein-2 space** is

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \right\}$$

The **Wasserstein-2 space** is

$$\mathcal{P}_2(\mathbb{R}^d) \coloneqq \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \right\}$$

and the **Wasserstein-2 metric** is given by

$$W_2(\mu, \nu)^2 = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, \mathrm{d}\pi(x, y), \;\; \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),$$

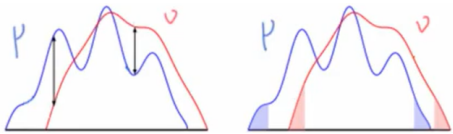
the transport polytope is $\Gamma(\mu, \nu) \coloneqq \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (P_1)_{\#}\pi = \mu, (P_2)_{\#}\pi = \nu\}$, and $f_{\#}\mu \coloneqq \mu \circ f^{-1}$.

The **Wasserstein-2 space** is

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \right\}$$

and the **Wasserstein-2 metric** is given by

$$W_2(\mu,\nu)^2 = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, \mathrm{d}\pi(x,y), \quad \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d),$$



Vertical (L^2) vs. horizontal (W_2) mass displacement.

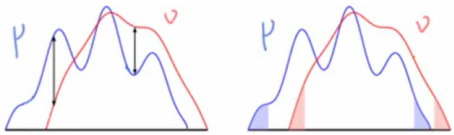
the transport polytope is $\Gamma(\mu,\nu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (P_1)_{\#}\pi = \mu, (P_2)_{\#}\pi = \nu\}$, and $f_{\#}\mu := \mu \circ f^{-1}$.

The **Wasserstein-2 space** is

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \right\}$$

and the **Wasserstein-2 metric** is given by

$$W_2(\mu,\nu)^2 = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, \mathrm{d}\pi(x,y), \quad \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d),$$



Vertical (L^2) vs. horizontal (W_2) mass displacement.

the transport polytope is $\Gamma(\mu,\nu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (P_1)_{\#}\pi = \mu, (P_2)_{\#}\pi = \nu\}$, and $f_{\#}\mu := \mu \circ f^{-1}$.

Theorem (Brenier (early 90s))

If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ has a density (and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is arbitrary), then there exists a transport map $T\colon \mathbb{R}^d \to \mathbb{R}^d$ with $T_{\#}\mu = \nu$ (and $T = \nabla\varphi$ for some convex function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$) solving the Monge problem

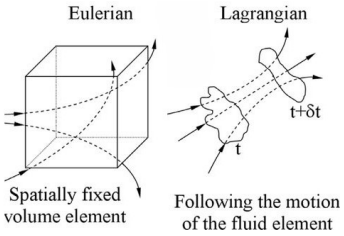
$$\inf_{\substack{T\colon \mathbb{R}^d \to \mathbb{R}^d, \\ T_{\#}\mu = \nu}} \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 \, \mathrm{d}\mu(x) = W_2(\mu,\nu)^2, \qquad \pi = (\mathrm{id}, T).$$

Density perspective: A curve $\mu\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is *absolutely continuous* if $\exists$ L^2-Borel velocity field $v\colon \mathbb{R}^d \times (0,\infty) \to \mathbb{R}^d$ s.t.

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \qquad (t,x) \in (0,\infty) \times \mathbb{R}^d, \text{ weakly.} \tag{Continuity Eq.}$$

Density perspective: A curve $\mu\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is *absolutely continuous* if $\exists$ L^2-Borel velocity field $v\colon \mathbb{R}^d \times (0,\infty) \to \mathbb{R}^d$ s.t.

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \qquad (t,x) \in (0,\infty) \times \mathbb{R}^d, \text{ weakly.} \qquad \text{(Continuity Eq.)}$$



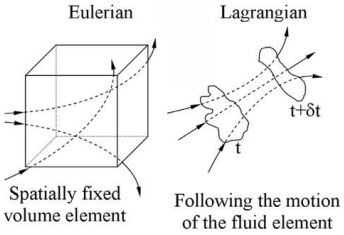
Density perspective: A curve $\mu\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is *absolutely continuous* if $\exists$ L^2-Borel velocity field $v\colon \mathbb{R}^d \times (0,\infty) \to \mathbb{R}^d$ s.t.

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \qquad (t,x) \in (0,\infty) \times \mathbb{R}^d, \text{ weakly.} \qquad \text{(Continuity Eq.)}$$

Particle perspective: Under regularity assumptions on $(v_t)_t$, there exists a solution ("flow map")

$$\partial_t \varphi(t,x) = v_t(\varphi(t,x)), \qquad \varphi(0,x) = x$$

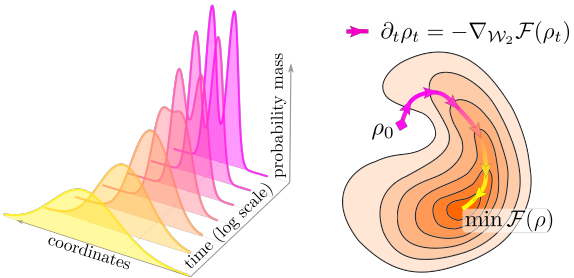
with $\mu_t = \varphi(t,\cdot)_{\#}\mu_0$.



Definition (Wasserstein gradient flow)

A locally absolutely continuous curve $\mu\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ with velocity field $v_t \in \mathrm{T}_{\mu_t}\,\mathcal{P}_2(\mathbb{R}^d)$ is a *Wasserstein gradient flow with respect to* $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ (assumed sufficiently regular) if

$$v_t \in -\partial_{W_2}\mathcal{F}(\mu_t), \quad \text{for a.e. } t > 0.$$



1) Start with “noise”: draw samples $(x_i^{(0)})_i$ from a Gaussian distribution (the easiest to sample from).

1) Start with "noise": draw samples $(x_i^{(0)})_i$ from a Gaussian distribution (the easiest to sample from).

2) Given vector field (topic of the next talk) v_t, evolve *particles* $x_i^{(0)}$ along the Lagrangian ODE.

1) Start with "noise": draw samples $(x_i^{(0)})_i$ from a Gaussian distribution (the easiest to sample from).

2) Given vector field (topic of the next talk) v_t, evolve *particles* $x_i^{(0)}$ along the Lagrangian ODE.

3) After finite time $x_i^{(0)}$ should look like a sample from the target distribution π.

1) Start with "noise": draw samples $(x_i^{(0)})_i$ from a Gaussian distribution (the easiest to sample from).

2) Given vector field (topic of the next talk) v_t, evolve *particles* $x_i^{(0)}$ along the Lagrangian ODE.

3) After finite time $x_i^{(0)}$ should look like a sample from the target distribution π.

(Altekrüger, Hertrich, et al. 2023)

Thank you for your attention!

I am happy to take any questions

Bibliography I

[1] J. Chemseddine, C. Wald, R. Duong, and G. Steidl, "Neural sampling from Boltzmann densities: Fisher-Rao curves in the Wasserstein geometry," in *The Thirteenth International Conference on Learning Representations*, 2024.

[2] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] J. Hertrich, C. Wald, F. Altekrüger, and P. Hagemann, "Generative sliced MMD flows with Riesz kernels," in *ICLR*, 2024.

[4] C. Villani, *Optimal transport: old and new* (Grundlehren der mathematischen Wissenschaften), 1st ed. Springer Berlin, Heidelberg, 2008, vol. 338, pp. XXII, 976. DOI: 10.1007/978-3-540-71050-9.

[5] F. Santambrogio, "{Euclidean, metric, and Wasserstein} gradient flows: An overview," *Bulletin of Mathematical Sciences*, vol. 7, no. 1, pp. 87–154, 2017.

Bibliography II

[6] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, 2nd ed. Springer Science & Business Media, 2008.

[7] F. Altekrüger, J. Hertrich, and G. Steidl, "Neural Wasserstein gradient flows for discrepancies with Riesz kernels," in *International Conference on Machine Learning*, PMLR, 2023, pp. 664–690.

[8] F. Altekrüger, P. Hagemann, and G. Steidl, "Conditional generative models are provably robust: Pointwise guarantees for Bayesian inverse problems," *Trans. Mach. Learn. Res.*, 2023.

[9] J. Chemseddine, P. Hagemann, G. Steidl, and C. Wald, "Conditional Wasserstein distances with applications in Bayesian OT flow matching," *Journal of Machine Learning Research*, vol. 26, no. 141, pp. 1–47, 2025.

[10] C. Wald and G. Steidl, "Flow matching: Markov kernels, stochastic processes and transport plans," *Variational and Information Flows in Machine Learning and Optimal Transport*, pp. 185–254, 2025.

[11] P. Hagemann, J. Schütte, D. Sommer, M. Eigel, and G. Steidl, "Sampling from Boltzmann densities with physics informed low-rank formats," in *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 2025, pp. 374–386.

[12] P. Hagemann, J. Hertrich, F. Altekrüger, R. Beinert, J. Chemseddine, and G. Steidl, "Posterior sampling based on gradient flows of the MMD with negative distance kernel," in *The Twelfth International Conference on Learning Representations*, 2024.

[13] J. Hertrich, M. Gräf, R. Beinert, and G. Steidl, "Wasserstein steepest descent flows of discrepancies with Riesz kernels," *Journal of Mathematical Analysis and Applications*, vol. 531, no. 1, p. 127 829, 2024.

[14] R. Duong, J. Chemseddine, P. K. Friz, and G. Steidl, "Telegrapher's generative model via Kac flows," *arXiv preprint arXiv:2506.20641*, 2025.

[15] J. Chemseddine, G. Kornhardt, R. Duong, and G. Steidl, "Adapting noise to data: Generative flows from 1D processes," *arXiv preprint arXiv:2510.12636*, 2025.

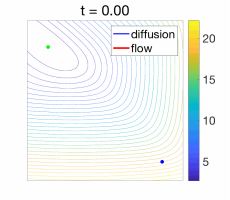
[16] P. L. Hagemann, J. Hertrich, and G. Steidl, *Generalized normalizing flows via Markov chains*. Cambridge University Press, 2023.

Definition (Fréchet subdifferential in Wasserstein space)

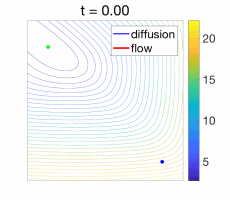
The (reduced) **Fréchet subdifferential** of $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ at $\mu \in \operatorname{dom}(\mathcal{F})$ is

$$\partial \mathcal{F}(\mu) \coloneqq \left\{ \xi \in L^2(\mathbb{R}^d; \mu) : \mathcal{F}(\nu) - \mathcal{F}(\mu) \geq \inf_{\pi \in \Gamma^{\mathrm{opt}}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \xi(x_1), x_2 - x_1 \rangle \, \mathrm{d}\pi(x, y) + o(W_2(\mu, \nu)) \right\}$$

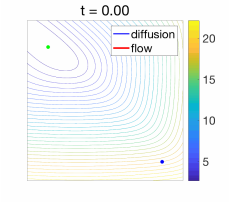
Gradient flows in $\mathbb{R}^d$



Gradient flows in $\mathbb{R}^d$



Gradient flows in $\mathbb{R}^d$



©Francis Bach

relevant AG papers (Altekrüger, P. Hagemann, et al. 2023; Chemseddine, P. Hagemann, et al. 2025; Chemseddine, Wald, et al. 2024; Wald and Steidl 2025; P. Hagemann, Schütte, et al. 2025; P. Hagemann, Hertrich, et al. 2024; Hertrich, Gräf, et al. 2024; Duong et al. 2025; Altekrüger, Hertrich, et al. 2023; Chemseddine, Kornhardt, et al. 2025; P. L. Hagemann et al. 2023)