Computational Gradient Flows and Optimal Transport: An Introduction

Unofficial Student Transcription (Draft)

Jia-Jie Zhu
KTH Royal Institute of Technology, Stockholm, Sweden

Summer School on Applied Mathematics
Beijing International Center for Mathematical Research (BICMR)
Peking University, China

July 20 – 31, 2026

Prepared by Lunji Zhu (UCLA), with assistance from Claude AI, from lectures and course materials by Jia-Jie Zhu. This unofficial version has not been reviewed or approved by the lecturer and may contain transcription or AI-generated errors or omissions.

Contents

Week 1

Gradient Flows, Dissipation Geometry, and Gradient Systems

1 Gradient Flows in Euclidean Space

1.1 The gradient-flow ODE

Let $x \in \mathbb{R}^d$. The (Euclidean) *gradient flow* of an energy $F : \mathbb{R}^d \to \mathbb{R}$ is the ODE

$$\dot{x} \;=\; -\nabla F(x). \tag{1}$$

It drives the state in the direction of steepest descent of F.

Remark. On the board ∇F appears as a row vector (the Jacobian DF), so the flow is written $\dot{x} = -\nabla F(x)^{\top}$. We adopt the convention that ∇F is a *column* vector, giving the clean form (1). The two agree.

Example 1.1 (Quadratic energy)**.** Let $F(x) = \frac{1}{2}\|x - m\|^2$. Then $\nabla F(x) = x - m$ and the flow is

$$\dot{x} = -(x - m), \qquad \text{with solution} \qquad x(t) = \left(1 - e^{-t}\right) m + e^{-t} x(0), \quad t \geq 0.$$

The energy contracts *exactly* exponentially:[1]

$$\boxed{\tfrac{1}{2}\|x(t) - m\|^2 \;=\; e^{-2t} \cdot \tfrac{1}{2}\|x(0) - m\|^2} \tag{2}$$

Exercise 1.2. Verify that the quadratic $F(x) = \frac{1}{2}\|x - m\|^2$ is 1-strongly convex.

Example 1.3 (Double well)**.** Let $F(x) = \frac{1}{4}(x^2 - 1)^2$. Then

$$\nabla F(x) = x^3 - x = x(x - 1)(x + 1),$$

so the flow $\dot{x}(t) = -\nabla F(x(t))$ has three equilibria $x \in \{-1, 0, 1\}$: the two wells ± 1 are stable, the hump at 0 is unstable.

1.2 Semiconvexity (λ-convexity)

Definition 1.4 (Semiconvexity in $(\mathbb{R}^d, \|\cdot\|)$)**.** A function $F : \mathbb{R}^d \to \mathbb{R}$ is *λ-convex* (with $\lambda \in \mathbb{R}$) if for all x, y and all $t \in [0, 1]$

$$F\big((1-t)x + ty\big) \;\leq\; (1-t)F(x) + tF(y) \;-\; \frac{\lambda}{2}\, t(1-t)\, \big\|x - y\big\|^2. \tag{3}$$

The parameter λ interpolates between familiar notions:

$\lambda = 0$	ordinary convexity;
$\lambda = 0$ with strict inequality	strict convexity;
$\lambda > 0$	strong convexity;
$\lambda < 0$	semiconvexity (allows some concavity).

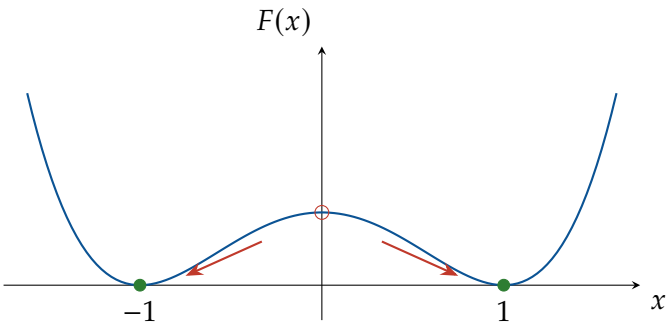


Figure 1. The double-well potential $F(x) = \frac{1}{4}(x^2-1)^2$. The gradient flow $\dot{x} = -(x^3 - x)$ has stable equilibria at the wells $x = \pm 1$ (filled) and an unstable equilibrium at the hump $x = 0$ (open); arrows show the downhill motion.

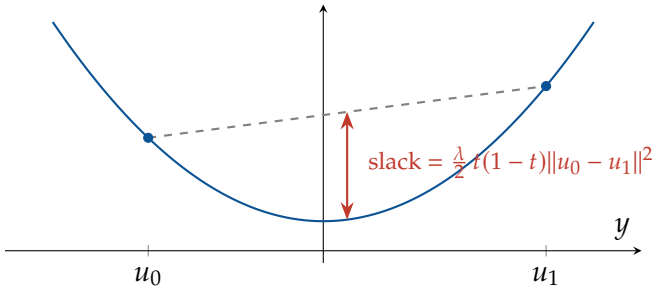


Figure 2. Semiconvexity: the graph lies below the chord between u_0 and u_1 by at least the quadratic slack $\frac{\lambda}{2}t(1-t)\,\|u_0 - u_1\|^2$ (here $\lambda > 0$). For $\lambda = 0$ this is ordinary convexity; $\lambda < 0$ permits the chord to dip below the curve.

The following differential characterizations are standard (proofs left as an exercise).

Proposition 1.5 (Differential characterizations)**.** *Let F be smooth. Then:*

(i) *If F is differentiable, F is λ-convex iff for all $x, y \in \mathbb{R}^d$*

$$F(y) \;\geq\; F(x) + \big\langle \nabla F(x),\, y - x \big\rangle + \frac{\lambda}{2}\,\big\|y - x\big\|^2 . \tag{4}$$

(ii) *If F is twice differentiable, F is λ-convex $\iff$ $\nabla^2 F(x) \succeq \lambda I$ for all $x \in \mathbb{R}^d$, where the board's Loewner-order convention is*

$$A \succeq B \;:\Longleftrightarrow\; v^{\top}\big(A - B\big)v \;\geq\; 0 \quad \forall\, v \in \mathbb{R}^d,$$

so that the criterion reads $v^{\top}\big(\nabla^2 F(x) - \lambda I\big)v \geq 0$ for all v.

1.3 Gradient dominance (the Polyak–Łojasiewicz inequality)

Theorem 1.6 (Gradient dominance / PL)**.** *If F is λ-convex with $\lambda > 0$, then F satisfies the* Polyak–Łojasiewicz (PL) *inequality*

$$\boxed{\tfrac{1}{2}\,\|\nabla F(x)\|^2 \;\geq\; \lambda\big(F(x) - \inf F\big), \qquad \forall x \in \mathbb{R}^d.} \tag{5}$$

[1]The whiteboard writes this identity with e^{-t}; the exact rate is 2. Indeed $x(t) - m = e^{-t}(x(0) - m)$ gives $\|x(t) - m\|^2 = e^{-2t}\,\|x(0) - m\|^2$, and equivalently $F = \frac{1}{2}\,\|x - m\|^2$ obeys $\frac{\mathrm{d}}{\mathrm{d}t}F = -\,\|\nabla F\|^2 = -2F$.

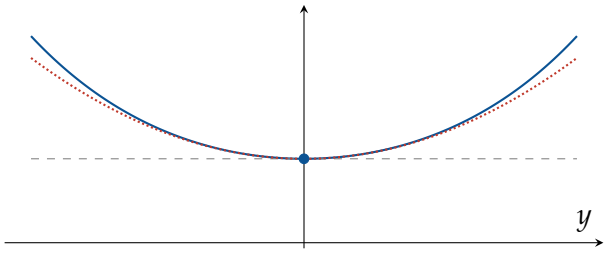


Figure 3. First-order characterization (4): a λ-convex F lies above the shifted parabola $F(x) + \langle \nabla F(x),\, y - x \rangle + \frac{\lambda}{2} \|y - x\|^2$ that touches its graph (and tangent line) at x.

Proof. By the first-order characterization (4), for all y,

$$F(y) \;\geq\; F(x) + \langle \nabla F(x),\, y - x \rangle + \frac{\lambda}{2} \|y - x\|^2 .$$

Complete the square in y (adding and subtracting $\frac{1}{2\lambda} \|\nabla F(x)\|^2$):

$$F(y) \;\geq\; F(x) + \frac{\lambda}{2} \left\| y - x + \tfrac{1}{\lambda} \nabla F(x) \right\|^2 - \frac{1}{2\lambda} \|\nabla F(x)\|^2 \;\geq\; F(x) - \frac{1}{2\lambda} \|\nabla F(x)\|^2 .$$

Taking the infimum over y on the left gives $\inf F \geq F(x) - \frac{1}{2\lambda} \|\nabla F(x)\|^2$, i.e. $\frac{1}{2\lambda} \|\nabla F(x)\|^2 \geq F(x) - \inf F$, which is (5). The second inequality is moreover an *equality* at exactly one y, as the board notes in the margin — namely where the completed square vanishes,

$$\left\| \lambda (y - x) + \nabla F(x) \right\|^2 = 0 \quad \Longleftrightarrow \quad y = x - \tfrac{1}{\lambda} \nabla F(x),$$

which is exactly one Newton-like step of size $1/\lambda$ from x. □

Fact 1.7 (Exponential decay of the energy)**.** *Gradient dominance of F implies that along any solution $x(t)$ of $\dot{x} = -\nabla F(x)$,*

$$F(x(t)) \;\leq\; e^{-2\lambda t} F(x(0)) + \left(1 - e^{-2\lambda t}\right) \inf F. \tag{6}$$

Proof. Along the flow,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(x(t)) = \langle \nabla F(x),\, \dot{x} \rangle = - \|\nabla F(x)\|^2 \overset{(5)}{\leq} -2\lambda \left(F(x(t)) - \inf F \right).$$

Writing $\varphi(t) := F(x(t)) - \inf F \geq 0$ we have $\dot{\varphi} \leq -2\lambda \varphi$, so by Gronwall's inequality $\varphi(t) \leq e^{-2\lambda t} \varphi(0)$, which rearranges to (6). □

Example 1.8 (A one-dimensional ODE)**.** Consider

$$\dot{x} = -x - x^3 = -\nabla F(x), \qquad \text{e.g. } F(x) = \tfrac{1}{2} x^2 + \tfrac{1}{4} x^4 + \text{const.}$$

Here $\nabla^2 F(x) = 1 + 3x^2 \geq 1$, so F is at least 1-strongly convex ($\lambda = 1$). By Fact 1.7, $F(x)$ decays exponentially with rate $2\lambda = 2$ along the ODE solution.

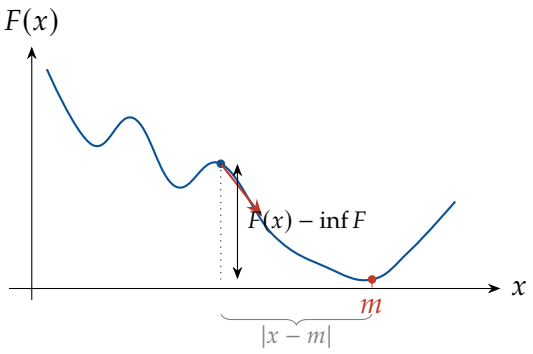


Figure 4. Gradient-flow descent on a (possibly non-convex) landscape toward a global minimizer m. Under gradient dominance / λ-convexity the height $F(x) - \inf F$ and the distance $|x - m|$ both decay, so $\frac{1}{2}\|\cdot - m\|^2$ acts as an exponentially decaying Lyapunov function.

1.4 Contraction to a minimizer and a Lyapunov view

Suppose m is a reference point and consider the squared distance $\frac{1}{2}\|x - m\|^2$. Along the flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\tfrac{1}{2}\|x - m\|^2\Big) = \langle x - m,\, \dot{x}\rangle = -\langle x - m,\, \nabla F(x)\rangle \overset{(*)}{\leq} \underbrace{F(m) - F(x)}_{\leq 0 \text{ if } m = \arg\min F} - \frac{\lambda}{2}\|x - m\|^2,$$

where $(*)$ is exactly the first-order λ-convexity inequality (4) rearranged.

Proposition 1.9 (Exponential contraction to the minimizer)**.** *If $m \in \arg\min_{z\in\mathbb{R}^d} F(z)$ and F is λ-convex with $\lambda > 0$, then*

$$\tfrac{1}{2}\|x(t) - m\|^2 \;\leq\; e^{-\lambda t}\Big(\tfrac{1}{2}\|x(0) - m\|^2\Big). \tag{7}$$

In particular $\frac{1}{2}\|\cdot - m\|^2$ is an exponentially decaying Lyapunov function.[2]

Proof. With m a minimizer, $F(m) - F(x) \leq 0$, so the display above gives $\frac{\mathrm{d}}{\mathrm{d}t}\big(\frac{1}{2}\|x - m\|^2\big) \leq -\lambda\big(\frac{1}{2}\|x - m\|^2\big)$, and Gronwall yields (7). □

For an *arbitrary* reference point $m \in \mathbb{R}^d$ and arbitrary $\lambda \in \mathbb{R}$, keeping the full term $F(m) - F(x)$ and integrating with Gronwall gives the sharper estimate

$$\begin{aligned}\tfrac{1}{2}\|x(t) - m\|^2 \;&\leq\; \frac{e^{-\lambda t}}{2}\|x_0 - m\|^2 + \int_0^t e^{-\lambda(t-s)}\big(F(m) - F(x(s))\big)\,\mathrm{d}s \\ &\leq\; \frac{e^{-\lambda t}}{2}\|x_0 - m\|^2 + M_\lambda(t)\big(F(m) - F(x(t))\big),\end{aligned} \tag{8}$$

where the second bound uses that $s \mapsto F(x(s))$ is nonincreasing, and

$$M_\lambda(t) := \int_0^t e^{-\lambda(t-s)}\,\mathrm{d}s = \begin{cases} \dfrac{1 - e^{-\lambda t}}{\lambda}, & \lambda \neq 0, \\ t, & \lambda = 0. \end{cases} \tag{9}$$

Corollary 1.10 (λ-evolutionary variational inequality, EVI_λ)**.** *Choosing $m = x_0$ in (8) kills the first term and yields*

$$\tfrac{1}{2}\|x(t) - x(0)\|^2 \;\leq\; M_\lambda(t)\big(F(x_0) - F(x(t))\big). \tag{10}$$

[2]The sign hypothesis is needed for the *name*, not for the bound: the displayed inequality is proved below for every $\lambda \in \mathbb{R}$, but for $\lambda \leq 0$ it is a growth estimate and the distance to a minimizer genuinely can increase. Take the double well of Example 1.3, $F(x) = \frac{1}{4}(x^2 - 1)^2$, which is (-1)-convex since $F'' = 3x^2 - 1 \geq -1$; with $m = +1 \in \arg\min F$ and $x(0) = -\frac{1}{2}$ the flow $\dot{x} = x - x^3$ runs to -1, so $\frac{1}{2}|x(t) - m|^2$ rises from $\frac{9}{8}$ to 2.

1.5 Dissipation estimates and the Energy Dissipation Balance

Along the flow $\dot{x}_t = -\nabla F(x_t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(x_t) = \langle \nabla F(x_t),\, \dot{x}_t \rangle = -\langle \nabla F(x_t),\, \nabla F(x_t) \rangle = -\|\nabla F(x_t)\|^2 \le 0.$$

The elementary Young inequality $-\langle a,\, b\rangle \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ (with equality iff $a = -b$) lets us rewrite this in the symmetric *dissipation* form. The key algebraic identity is the Fenchel–Young relation.

Proposition 1.11 (Fenchel–Young relation, quadratic case)**.** *For ξ, u in a Hilbert space the following are equivalent:*

1. $\xi = u$;
2. $\frac{1}{2}\|\xi\|^2 + \frac{1}{2}\|u\|^2 = \langle \xi,\, u\rangle$;
3. $u \in \arg\max_{u} \langle \xi,\, u\rangle - \frac{1}{2}\|u\|^2$.

Before saturating it, note what the Young inequality gives for an *arbitrary* absolutely continuous curve x_t — the board writes both directions:

$$\underbrace{-\langle \nabla F(x_t),\, \dot{x}_t\rangle}_{(*)} \;\le\; \tfrac{1}{2}\|\nabla F(x_t)\|^2 + \tfrac{1}{2}\|\dot{x}_t\|^2, \qquad \text{hence} \qquad \frac{\mathrm{d}}{\mathrm{d}t} F(x_t) \;\ge\; -\tfrac{1}{2}\|\nabla F(x_t)\|^2 - \tfrac{1}{2}\|\dot{x}_t\|^2.$$

This is the Young/chain-rule dissipation *estimate*, valid along any curve. It becomes an *identity* precisely when $\dot{x}_t = -\nabla F(x_t)$, i.e. precisely when x_t solves the gradient flow — which is the content of the boxed Energy Dissipation Balance below and the reason "EDB $\iff$ gradient flow".

Using it with $\xi = -\nabla F(x_t)$ and the flow $\dot{x}_t = -\nabla F(x_t)$ (so equality holds), we may split the dissipation symmetrically:

$$\frac{\mathrm{d}}{\mathrm{d}t} F(x_t) = -\langle -\nabla F(x_t),\, \dot{x}_t\rangle = -\tfrac{1}{2}\|\nabla F(x_t)\|^2 - \tfrac{1}{2}\|\dot{x}_t\|^2 \;\le\; 0,$$

and "completing the square" expresses the flow equation as

$$\tfrac{1}{2}\big\|\nabla F(x_t) + \dot{x}_t\big\|^2 = 0 \iff \dot{x}_t = -\nabla F(x_t).$$

Integrating in time from 0 to t gives the central identity.

Energy Dissipation Balance (EDB / EDI).

$$F(x_t) - F(x_0) \;=\; -\int_0^t \left(\tfrac{1}{2}\|\nabla F(x_s)\|^2 + \tfrac{1}{2}\|\dot{x}_s\|^2\right) \mathrm{d}s.$$

Moreover, a curve solves the gradient flow **if and only if** it satisfies the EDB:

$$\text{Solution to GF} \quad\Longleftrightarrow\quad \text{EDB}.$$

Remark. In the Euclidean ODE one has $\|\nabla F(x_t)\| = \|\dot{x}_t\|$, so the two dissipation terms coincide.

1.6 Convergence to an ε-stationary point

From the EDB (in the form $F(x_t) - F(x_0) = -\int_0^t \|\nabla F(x_s)\|^2 \,\mathrm{d}s$),

$$\frac{1}{t}\int_0^t \|\nabla F(x_s)\|^2 \,\mathrm{d}s = \frac{1}{t}\big(F(x_0) - F(x_t)\big) \le \frac{1}{t}\big(F(x_0) - \inf F\big), \qquad \forall t > 0.$$

Consequently, by Cauchy–Schwarz,

$$\min_{0\le s\le t} \|\nabla F(x_s)\| \;\le\; \frac{1}{\sqrt{t}}\sqrt{\int_0^t \|\nabla F(x_s)\|^2 \,\mathrm{d}s} \;\le\; \frac{1}{\sqrt{t}}\sqrt{F(x_0) - \inf F}\,.$$

Corollary 1.12 (Rate to stationarity)**.** *To reach an ε-stationary point $\|\nabla F(x)\| \le \varepsilon$ with the gradient flow, it suffices that*

$$\frac{1}{\sqrt{t}} \lesssim \varepsilon \iff t \gtrsim \frac{1}{\varepsilon^2}\,.$$

1.7 Two tools: Gronwall's lemma and Lyapunov functions

Lemma 1.13 (Gronwall)**.** *Let $u : [0,T] \to [0,+\infty)$ be absolutely continuous and a, b integrable (of either sign), with*

$$\dot u(t) \le a(t)\,u(t) + b(t) \quad \text{for a.e. } t \in [0,T],$$

Then for all $t \in [0,T]$,

$$u(t) \;\le\; u(0)\,e^{\int_0^t a(s)\,\mathrm{d}s} + \int_0^t e^{\int_s^t a(r)\,\mathrm{d}r}\, b(s)\,\mathrm{d}s.$$

Definition 1.14 (Lyapunov function)**.** Let $V : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. If $\frac{\mathrm{d}}{\mathrm{d}t}V(x_t) \le 0$ along the flow, then V is a *Lyapunov function* for the system.

Example 1.15. For $\dot x = -\nabla F(x)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}V(x_t) = \langle \nabla V(x_t),\, \dot x\rangle = -\langle \nabla V(x),\, \nabla F(x)\rangle\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}F(x_t) = \langle \nabla F(x_t),\, \dot x_t\rangle = -\|\nabla F(x_t)\|^2 \le 0.$$

Recall that if F is λ-convex ($\lambda > 0$) then $V(x) = \|x - m\|$, with $m = \arg\min F$, is a natural candidate.

Two refinements:

- If, in addition to $\frac{\mathrm{d}}{\mathrm{d}t}V(x) < 0$ strictly for all $x \notin \{0\}$, V is *positive definite* ($V(0) = 0$ and $V(x) > 0$ for $x \neq 0$), then V certifies asymptotic stability of the stationary point 0 (where $\nabla F(0) = 0$). Positive definiteness cannot be dropped: for $F(x) = -\frac{1}{2}x^2$ the flow is $\dot x = x$ and $V(x) = -x^2$ satisfies $\frac{\mathrm{d}}{\mathrm{d}t}V = -2x^2 < 0$ for $x \neq 0$, yet 0 is unstable.
- If $\frac{\mathrm{d}}{\mathrm{d}t}V(x) \le -\lambda\,V(x)$ with $\lambda > 0$ — recall $\frac{\mathrm{d}}{\mathrm{d}t}V = -\langle \nabla V(x), \nabla F(x)\rangle$ — then $V(x(t)) \le e^{-\lambda t}V(x(0))$;[3] i.e. V is an *exponentially decaying* Lyapunov function.

2 Gradient Flows in Banach and Hilbert Spaces

Let X be a reflexive Banach space with squared norm $\|\cdot\|_X^2$, and denote its dual by X^*.

[3] The whiteboard writes the hypothesis as $\frac{\mathrm{d}}{\mathrm{d}t}V \le \lambda V$; for $e^{-\lambda t}$ decay with $\lambda > 0$ the correct sign is $\frac{\mathrm{d}}{\mathrm{d}t}V \le -\lambda V$ (Gronwall with $a \equiv -\lambda$, $b \equiv 0$).

2.1 Fréchet derivative and subdifferential

Definition 2.1 (Fréchet derivative)**.** F is *Fréchet differentiable* at u if there exists $\mathrm{D}F : X \to X^*$ such that

$$F(u+h) = F(u) + \langle \mathrm{D}F(u),\, h\rangle_{X^*,X} + o\big(\|h\|_X\big) \qquad (h \to 0).$$

Definition 2.2 (Fréchet subdifferential)**.** The *Fréchet subdifferential* is the set

$$\partial^{\mathrm{F}} F(u) := \left\{\, \xi \in X^* \;\middle|\; F(u+h) \geq F(u) + \langle \xi,\, h\rangle_{X^*,X} + o(\|h\|_X),\ h \to 0 \,\right\} \subseteq X^*.$$

The *convex subdifferential* drops the $o(\|h\|_X)$ remainder and asks the inequality globally:

$$\partial F(u) := \left\{\, \xi \in X^* \;\middle|\; F(u+h) \geq F(u) + \langle \xi,\, h\rangle_{X^*,X} \,\right\}.$$

Geometrically, an element of $\partial^{\mathrm{F}} F(u)$ is a "lower subgradient": a slope ξ whose affine function $h \mapsto F(u) + \langle \xi,\, h\rangle$ lies below the graph up to an $o(\|h\|)$ error. The slack is what distinguishes it from a genuine (global) affine minorant, and it is why $\partial^{\mathrm{F}} F$ may be nonempty where no convex subgradient exists.

Example 2.3 ($\partial^{\mathrm{F}} F$ can be nonempty at a maximum)**.** Let $F(x) = -x^2$. Testing ξ at 0,

$$\frac{F(x) - F(0) - \langle \xi,\, x\rangle}{|x|} = \frac{-x^2}{|x|} - \left\langle \xi,\, \tfrac{x}{|x|} \right\rangle \xrightarrow[|x|\to 0]{} -\xi \cdot \operatorname{sign}(x).$$

Hence $\xi = 0$ satisfies the definition of the Fréchet subdifferential, so

$$\partial^{\mathrm{F}} F(0) = \{0\} \qquad \text{(but there is no convex subdifferential here).}$$

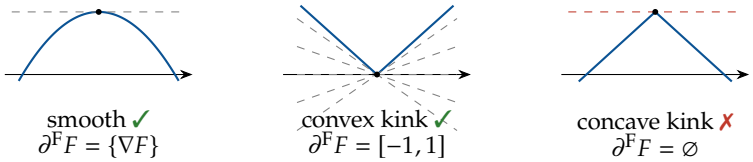


Figure 5. The Fréchet ("lower") subdifferential $\partial^{\mathrm{F}} F$ collects the slopes ξ whose affine function lies below the graph up to an $o(\|h\|)$ error. *Left*: at a smooth point it is the single tangent $\{\nabla F\}$ — as for $-x^2$, where the graph falls away only at second order. *Middle*: at a convex kink a whole fan of supporting lines qualifies. *Right*: at a concave kink the graph drops away at *first* order, which no $o(\|h\|)$ slack can absorb, so $\partial^{\mathrm{F}} F = \varnothing$.

Definition 2.4 (Variational derivative)**.** Suppose F is Fréchet differentiable at u and there is a function g such that

$$\langle \mathrm{D}F(u),\, h\rangle = \int_\Omega g(x)\, h(x)\, \mathrm{d}x \qquad \text{for every } h.$$

We call g the *variational derivative*, sometimes denoted $\dfrac{\delta F}{\delta u}$.

2.2 Examples of Hilbert spaces

- $(\mathbb{R}^d, \|\cdot\|)$: $\langle x,\, y\rangle = \sum_{i=1}^d x_i y_i$ $(= x^\top y)$, $\langle x,\, x\rangle = \|x\|^2$.
- $\mathbb{R}^d$ with a metric $G \succ 0$: $\langle x,\, y\rangle_G = x^\top G\, y$, $\langle x,\, x\rangle_G = \|x\|_G^2$.
- ℓ^2: $\langle x,\, y\rangle_{\ell^2} = \sum_i x_i y_i$, $\langle x,\, x\rangle_{\ell^2} = \sum_i x_i^2 < \infty$.
- $L^2(\Omega)$: $\langle f,\, g\rangle_{L^2(\Omega)} = \int_\Omega f(x) g(x)\, \mathrm{d}x$, $\langle f,\, f\rangle_{L^2} = \int_\Omega |f(x)|^2\, \mathrm{d}x < \infty$.

2.3 Reproducing kernel Hilbert spaces (RKHS)

Kernel ridge regression minimizes the functional

$$E(f) := \frac{1}{2n}\sum_{i=1}^{n}\big(f(x_i)-y_i\big)^2 + \frac{\lambda}{2}\,\|f\|_{\mathcal{H}}^2\,, \qquad \lambda>0,$$

over $f\in\mathcal{H}$, where $(x_i,y_i)\in(X,\mathbb{R})$, $i=1,\dots,n$, are the given data.

Definition 2.5 (RKHS). A Hilbert space $\mathcal{H}$ is a *reproducing kernel Hilbert space* if there exists a kernel $K: X\times X\to\mathbb{R}$ such that for all $x\in X$,

$$K(x,\cdot)\in\mathcal{H}, \qquad \text{and} \qquad f(x)=\big\langle f,\,K(x,\cdot)\big\rangle_{\mathcal{H}} \quad \text{for any } f\in\mathcal{H} \quad \text{(reproducing property)}.$$

A finite-dimensional subspace $\mathcal{H}_n\subsetneq\mathcal{H}$ consists of functions

$$\boxed{\widehat{f}(\cdot)=\sum_{i=1}^{n}\alpha_i\,K(x_i,\cdot)}$$

for given nodes $\{x_1,\dots,x_n\}$. If $\widehat{f},\widehat{g}\in\mathcal{H}_n\subseteq\mathcal{H}$ with $\widehat{g}=\sum_{i=1}^{n}\beta_i K(x_i,\cdot)$, then

$$\big\langle\widehat{f},\widehat{g}\big\rangle_{\mathcal{H}}=\big\langle\widehat{f},\widehat{g}\big\rangle_{\mathcal{H}_n}=\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\beta_j\,K(x_i,x_j)=\alpha^{\top}\mathbf{K}\,\beta, \qquad \mathbf{K}=\big[K(x_i,x_j)\big]_{i,j}.$$

Here $\mathbf{K}$ is the Gram matrix. Taking $\widehat{g}=\widehat{f}$ gives $\alpha^{\top}\mathbf{K}\alpha=\big\|\widehat{f}\big\|_{\mathcal{H}}^2\geq 0$ for every α and every choice of nodes: a *reproducing* kernel is necessarily positive semidefinite. Common kernel choices — the last of which is only *conditionally* positive definite, hence not reproducing for any $\mathcal{H}$ — are

$$K(x,y)=e^{-\|x-y\|^2/2b^2}, \qquad \underbrace{e^{-\|x-y\|^p/b}}_{\text{p.d. iff } 0<p\leq 2}, \qquad \underbrace{-\,|x-y|}_{\text{conditionally p.d.}}, \qquad \dots$$

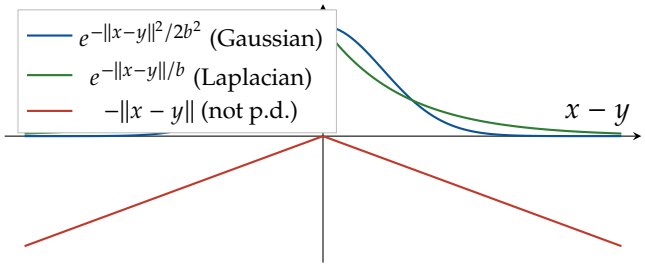


Figure 6. Several kernels as functions of the displacement $x-y$: the positive-definite Gaussian and Laplacian bumps, and the (non-positive-definite) negative-distance kernel.

3 The Abstract Gradient Structure

Definition 3.1 (Gradient system). A triple (X,F,R) is called a *gradient system*, specified by:

1. a *state space* X (e.g. $\mathbb{R}^d$, or functions on $\mathbb{R}^d$);
2. an *energy functional* $F: X\to\mathbb{R}\cup\{+\infty\}$;

3. a *dissipation geometry* R (e.g. a Banach/Hilbert norm or inner product; a Riemannian manifold with a metric $\mathsf{G}(u) : T_uM \to T_u^*M$).

The gradient system *generates* a gradient-flow equation.

Example 3.2 (Euclidean gradient system)**.** $(\mathbb{R}^d,\ F(x),\ \frac{1}{2}\|\cdot\|^2)$ generates the gradient-flow equation

$$\dot{x} = -\nabla F(x).$$

Example 3.3 (Hilbert gradient system)**.** $(\mathcal{H},\ F(x),\ \frac{1}{2}\|\cdot\|_{\mathcal{H}}^2)$ is the Hilbert gradient system. For a state $u(t) \in \mathcal{H}$ the flow is

$$\boxed{\dot{u}(t) = -\,\mathsf{G}^{-1}\mathrm{D}F\big(u(t)\big)} \tag{11}$$

where $\mathsf{G} : \mathcal{H} \to \mathcal{H}^*$ is the *Riesz map*, characterized by $\langle \mathsf{G}v,\ h\rangle_{\mathcal{H}^*,\mathcal{H}} = \langle v,\ h\rangle_{\mathcal{H}}$ for all h.

Remark. Sometimes one writes $\mathbb{K} := \mathsf{G}^{-1}$, so the flow reads $\dot{u} = -\mathbb{K}\,\mathrm{D}F(u)$.

Example 3.4 (G is the transpose)**.** Take $\mathsf{G} = (\cdot)^\top$, the map sending a column vector to the row vector (covector) that pairs with it. The gradient-flow equation then reads

$$\dot{x} = -\big(\nabla F(x)\big)^\top,$$

which is the row-vector convention of the Remark in §1.1 — now seen not as a notational accident but as a choice of Riesz map.

Example 3.5 (Metric $\mathbb{R}^d$)**.** Take $(\mathbb{R}^d,\ F,\ \langle\cdot,\ A\,\cdot\rangle)$ with $A \succ 0$, $A = A^\top$, so $\|\cdot\|_{\mathcal{H}}^2$ is the A-weighted energy. Following the board's chain, with the dimensions of each factor tracked,

$$\big\langle x,\ y\big\rangle_{\mathcal{H}} = \big\langle \underset{d\times 1}{x},\ \underbrace{Ay}_{\mathsf{G}\,y,\ d\times 1}\big\rangle_E = \big\langle A^\top x,\ y\big\rangle_E = \underbrace{x^\top A}_{=\,\mathrm{D}F\,\in\,\mathcal{H}^*}\ y = \big\langle x^\top A,\ y\big\rangle_{\mathcal{H}^*,\mathcal{H}}.$$

Two things are visible here: the Riesz map *is* A, and $\mathrm{D}F$ is the *covector* $x^\top A$, which is why $\mathsf{G}^{-1} = A^{-1}$ has to be applied to pull it back into $\mathcal{H}$. Hence the gradient flow becomes

$$\dot{x} = -A^{-1}\nabla F(x) \qquad \text{(vs. the Euclidean } \dot{x} = -\nabla F(x)\text{)}.$$

Notation. We write the metric gradient flow interchangeably as

$$\dot{u}(t) = -\operatorname{grad}_{\mathcal{H}} F(u(t)), \qquad \text{or} \qquad \dot{u}(t) = -\nabla_{\mathcal{H}} F(u(t)).$$

Definition 3.6 (Semiconvexity, general)**.** A function $F : X \to \mathbb{R} \cup \{+\infty\}$ is λ-*convex* ($\lambda \in \mathbb{R}$) if for all $u_0, u_1 \in X$ and $t \in [0,1]$,

$$F\big((1-t)u_0 + tu_1\big) \le (1-t)F(u_0) + tF(u_1) - \frac{\lambda}{2}\,t(1-t)\,\|u_0 - u_1\|_X^2\,.$$

Theorem 3.7 (Subdifferential of a semiconvex function)**.** *If F is λ-convex, then its Fréchet subdifferential satisfies*

$$\partial^{\mathrm{F}}F(u) = \left\{\ \xi \in X^* \,\middle|\, \forall v \in X:\ F(v) \ge F(u) + \langle \xi,\ v-u\rangle + \frac{\lambda}{2}\,\|v-u\|_X^2 \right\}.$$

3.1 Energy Dissipation Balance in a Hilbert space

Along the Hilbert gradient flow, identifying $\mathrm{D}F(u) \in \mathcal{H}^*$ with $\mathsf{G}^{-1}\mathrm{D}F(u) \in \mathcal{H}$ (the variational derivative δF),

$$\frac{\mathrm{d}}{\mathrm{d}t}F(u(t)) = \langle \mathrm{D}F(u),\, \dot{u}(t)\rangle_{\mathcal{H}} \overset{(Y)}{\geq} -\tfrac{1}{2}\,\|\mathrm{D}F(u)\|_{\mathcal{H}}^2 - \tfrac{1}{2}\,\|\dot{u}(t)\|_{\mathcal{H}}^2\,,$$

using the Young inequality $\langle a,\, b\rangle_{\mathcal{H}} \geq -\tfrac{1}{2}\,\|a\|_{\mathcal{H}}^2 - \tfrac{1}{2}\,\|b\|_{\mathcal{H}}^2$.

Proposition 3.8 (Fenchel–Young relation)**.** *Let ψ be proper, lower semicontinuous, and convex on a reflexive Banach space X, with convex conjugate*

$$\psi^*(\eta) := \sup_{u\in X}\left\{\,\langle \eta,\, u\rangle - \psi(u)\,\right\} \qquad (e.g.\ \psi = \tfrac{1}{2}\,\|\cdot\|^2\,,\ \psi^* = \tfrac{1}{2}\,\|\cdot\|^2).$$

For $(v_0, \eta_0) \in X \times X^*$ the following are equivalent:

1. *v_0 minimizes $\psi(v) - \langle \eta_0,\, v\rangle$ over X;*
2. *$\eta_0 \in \partial\psi(v_0)$;*
3. *$\psi(v_0) + \psi^*(\eta_0) = \langle \eta_0,\, v_0\rangle$;*
4. *$v_0 \in \partial\psi^*(\eta_0)$;*
5. *η_0 maximizes $\langle \eta,\, v_0\rangle - \psi^*(\eta)$ over X^*.*

For the quadratic $\psi = \tfrac{1}{2}\,\|\cdot\|^2$ on a Hilbert *space, equality forces $\eta_0 = v_0$ (after the Riesz identification $\mathcal{H}^* \cong \mathcal{H}$); on a general reflexive X it forces $\eta_0 \in J(v_0)$, the duality map.*

To *saturate* the Young inequality (Y) we require the pair $\big(\dot{u}(t),\, -\mathrm{D}F(u(t))\big)$ to satisfy the Fenchel–Young equivalence. By Proposition 3.8 equality then forces $-\mathrm{D}F(u(t)) = \dot{u}(t)$, i.e. the gradient-flow equation, and (Y) becomes an identity.[4]

Energy Dissipation Balance (Hilbert). Differential form:

$$\frac{\mathrm{d}}{\mathrm{d}t}F(u(t)) = -\tfrac{1}{2}\,\|\mathrm{D}F(u)\|_{\mathcal{H}}^2 - \tfrac{1}{2}\,\|\dot{u}(t)\|_{\mathcal{H}}^2 \qquad (\dot{u} = -\mathrm{D}F(u)).$$

Integrated form:

$$F(u(t)) - F(u(s)) = -\int_s^t \left(\tfrac{1}{2}\,\|\mathrm{D}F(u(r))\|_{\mathcal{H}}^2 + \tfrac{1}{2}\,\|\dot{u}(r)\|_{\mathcal{H}}^2\right)\mathrm{d}r.$$

By the Fenchel–Young equivalence this is equivalent to the gradient-flow equation $\dot{u} = -\mathrm{D}F(u)$.

For *nonsmooth* F the balance holds not for an arbitrary selection but for the one realizing the flow, $\xi(t) = -\dot{u}(t) \in \partial^{\mathrm{F}}F(u(t))$ (equivalently the minimal-norm element, the *local slope* $|\partial F|\,(u(t))$ — the right-hand side below depends on $\|\xi\|$, which a multivalued $\partial^{\mathrm{F}}F$ does not determine). With that selection,

$$F(u(t)) - F(u(s)) = -\int_s^t \left(\tfrac{1}{2}\,\|\xi(r)\|_{\mathcal{H}}^2 + \tfrac{1}{2}\,\|\dot{u}(r)\|_{\mathcal{H}}^2\right)\mathrm{d}r, \qquad \dot{u} \in -\partial^{\mathrm{F}}F(u).$$

The guiding principle:

Energy Dissipation Balance $\iff$ gradient-flow equation.

[4] The minus sign matters: the board's prose line annotates the pair as $(\dot{u}, \mathrm{D}F(u))$ and then corrects it in blue to $-\mathrm{D}F(u)$; read literally, the uncorrected form would give the ascent flow $\dot{u} = +\mathrm{D}F(u)$.

Proposition 3.9 (Helmholtz–Rayleigh principle)**.** *The velocity may be selected variationally:*

$$v \in \arg\min_{w} \left\{ \langle \mathrm{D}F(u),\, w\rangle + \tfrac{1}{2}\,\|w\|_X^2 \right\},$$

whose optimality condition $0 = \underline{\mathrm{D}F(u)} + \underline{w}$ *(the board underlines exactly these two terms — they are the objects to be matched) recovers* $w = -\mathrm{D}F(u)$.

Example 3.10 (Heat equation / Allen–Cahn as an L^2 gradient flow)**.** The board fills in the (X, F, R) template slot by slot, and asks for the flow as the *answer*:

State X	$L^2(\Omega)$ (a Hilbert space $\mathcal{H}$)
Geometry R	the same $L^2(\Omega)$ inner product
Energy F	$F_{\mathrm{Dir}}(u) = \frac{1}{2}\int_\Omega \lvert\nabla u\rvert^2 \,\mathrm{d}x$ (Dirichlet)
Question	what is the gradient-flow equation?

The variational derivative of F_{Dir} is $-\Delta u$, so the L^2-gradient flow is

$$\dot u = -\operatorname{grad}_{L^2(\Omega)} F(u) = \Delta u, \qquad \text{i.e.} \qquad \boxed{\partial_t u = \Delta u} \quad \text{(the heat equation).}$$

3.2 EVI_λ in a Hilbert space

Theorem 3.11 (EVI_λ)**.** *Suppose F is λ-convex in $\mathcal{H}$, and let $-\xi(t) = \mathsf{G}\,\dot u(t) \in -\partial^{\mathrm{F}} F(u(t))$ (via the Riesz map). Then for every $w \in \mathcal{H}$,*

$$\tfrac{1}{2}\,\|u(t) - w\|_{\mathcal{H}}^2 \;\le\; e^{-\lambda t}\,\tfrac{1}{2}\,\|u(0) - w\|_{\mathcal{H}}^2 + M_\lambda(t)\,\big(F(w) - F(u(t))\big),$$

with $M_\lambda(t) = \int_0^t e^{-\lambda s}\,\mathrm{d}s = \begin{cases} (1 - e^{-\lambda t})/\lambda, & \lambda \neq 0, \\ t, & \lambda = 0. \end{cases}$

Proof. For every $w \in \mathcal{H}$,

$$\tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|u(t) - w\|^2 = \langle u(t) - w,\, \dot u(t)\rangle_{\mathcal{H}} = \langle u(t) - w,\, -\xi(t)\rangle_{\mathcal{H}} \overset{\lambda\text{-cvx}}{\le} F(w) - F(u(t)) - \frac{\lambda}{2}\,\|u(t) - w\|_{\mathcal{H}}^2.$$

Hence, writing $y(t) := \frac{1}{2}\,\|u(t) - w\|^2$, we have $\dot y \le -\lambda y + \big(F(w) - F(u(t))\big)$, and Gronwall (Lemma 1.13, applied with $a \equiv -\lambda$ and the sign-indefinite $b(t) = F(w) - F(u(t))$) gives

$$y(t) \;\le\; e^{-\lambda t} y(0) + \int_0^t e^{-\lambda(t-s)}\big(F(w) - F(u(s))\big)\,\mathrm{d}s.$$

Finally $s \mapsto F(u(s))$ is nonincreasing along the flow — by the Energy Dissipation Balance of §3.1, $\frac{\mathrm{d}}{\mathrm{d}t}F(u(t)) = -\|\dot u(t)\|_{\mathcal{H}}^2 \le 0$ — so $F(w) - F(u(s)) \le F(w) - F(u(t))$ for $s \le t$, and the integral is at most $M_\lambda(t)\big(F(w) - F(u(t))\big)$. □

4 Gradient Flows on Riemannian Manifolds

Why leave the linear setting at all? In a Hilbert space the interpolation between two states is a straight segment that never leaves the space, and a velocity is simply a vector. On a curved state space neither is true: motion has to follow curves *inside* the space, and velocities must be read in the tangent space at the current point. That is the whole content of the next definition.

Definition 4.1 (Riemannian manifold)**.** (M, G) is a Riemannian manifold:

- M is a smooth (finite-dimensional) manifold;

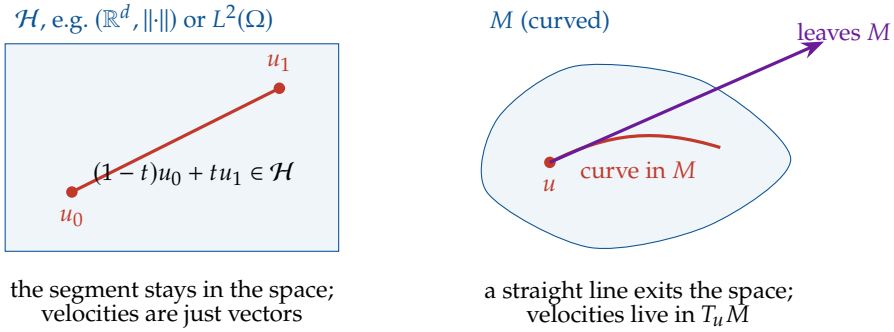


Figure 7. The step from §2 to this section. *Left*: in a Hilbert space, states can be combined linearly and the straight segment between two of them is again a state, so "direction" and "difference of two points" are the same thing. *Right*: on a curved state space they are not — the straight line through u leaves M — so the metric $\mathsf{G}(u)$ must be specified pointwise on the tangent space T_uM, and motion follows curves inside M. Figure 24 shows the same distinction concretely, for the space of probability measures.

- $\mathsf{G}(u) : T_uM \to T_u^*M$ is symmetric and positive definite.

Thus $\mathsf{G}(u)$ plays the role $X \to X^*$, mapping a tangent vector $v \in T_uM$ to a covector, and

$$g_u(v, w) := \langle \mathsf{G}(u)v,\, w\rangle_{T_u^*M,\, T_uM}$$

is a symmetric 2-tensor – the *metric*.

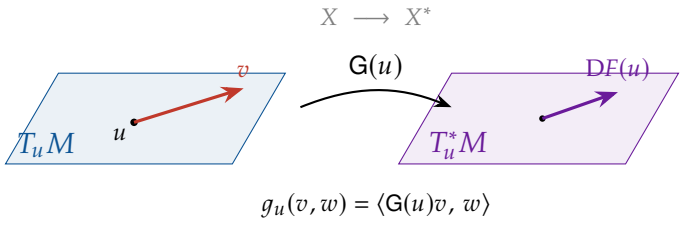


Figure 8. The metric $\mathsf{G}(u)\colon T_uM \to T_u^*M$ raises/lowers indices between the tangent space (holding velocities v) and the cotangent space (holding differentials $\mathrm{D}F(u)$); $\mathbb{K}(u) = \mathsf{G}(u)^{-1}$ runs the other way, and $\operatorname{grad}_{\mathsf{G}} F(u) = \mathbb{K}(u)\mathrm{D}F(u)$.

Definition 4.2 (Riemannian gradient and flow)**.** The *Riemannian gradient* of F is

$$\operatorname{grad}_{\mathsf{G}} F(u) := \mathsf{G}(u)^{-1}\mathrm{D}F(u), \qquad \mathbb{K}(u) := \mathsf{G}(u)^{-1} : T_u^*M \to T_uM.$$

The gradient-flow equation is

$$T_uM \ni \dot{u} = -\operatorname{grad}_{\mathsf{G}} F(u) = -\mathbb{K}(u)\,\mathrm{D}F(u).$$

The dissipation is monotone:

$$\frac{\mathrm{d}}{\mathrm{d}t}F(u(t)) = \langle \mathrm{D}F(u),\, \dot{u}\rangle = -\langle \mathrm{D}F(u),\, \mathbb{K}(u)\mathrm{D}F(u)\rangle = -\left\langle \mathsf{G}(u)\operatorname{grad}_{\mathsf{G}} F(u),\, \operatorname{grad}_{\mathsf{G}} F(u)\right\rangle \le 0.$$

The same PDE can arise from different gradient structures. For instance, $\dot{u} = \Delta u$ is:

- the $L^2(\Omega)$ gradient flow of the Dirichlet energy F_{Dir};
- the H^{-1} gradient flow of $\frac{1}{2}\|u\|_{L^2}^2$;
- the Otto–Wasserstein gradient flow (Riemannian) of the Boltzmann entropy (when u is a probability measure).

4.1 The Otto–Wasserstein gradient flow

Consider the gradient system $(\mathcal{P}(\Omega),\, H,\, \mathsf{G}_{\mathrm{OT}})$, where:

- $\mathcal{P}(\Omega)$ is the set of probability measures on Ω;
- $H(\rho)$ is the *Boltzmann entropy*

$$H(\rho) := \begin{cases} \displaystyle\int_\Omega \psi_B(\rho) = \int_\Omega \big(\rho(x)\log\rho(x) - \rho(x) + 1\big)\,\mathrm{d}x, & \rho \ll \text{Lebesgue}, \\ +\infty, & \text{otherwise}, \end{cases}$$

  with $\psi_B'(\rho) = \log\rho$; here Ω is taken of finite Lebesgue measure, so that the constant $+1$ is integrable. On unbounded Ω replace the reference density 1 by a probability density π, i.e. $\psi_B(\rho) \rightsquigarrow \rho\log\frac{\rho}{\pi} - \rho + \pi$; nothing below changes, since only ψ_B' enters the flow;
- the (inverse-metric) *Onsager operator* is

$$\boxed{\mathbb{K}_{\mathrm{OT}}(\rho)\,\xi := -\nabla\cdot\big(\rho\,\nabla\xi\big), \qquad \mathbb{K}_{\mathrm{OT}}(\rho) = \mathsf{G}_{\mathrm{OT}}(\rho)^{-1}.}$$

The gradient-flow equation is the continuity equation

$$\dot\rho = -\operatorname{grad}_{\mathrm{OT}} F(\rho) = -\mathbb{K}_{\mathrm{OT}}(\rho)\,\mathrm{D}F(\rho) = \nabla\cdot\big(\rho\,\nabla\mathrm{D}F(\rho)\big).$$

For the Boltzmann entropy $F = H$, we have $\mathrm{D}H(\rho) = \psi_B'(\rho) = \log\rho$, hence

$$\dot\rho = \nabla\cdot(\rho\,\nabla\log\rho) = \nabla\cdot\Big(\rho\cdot\tfrac{1}{\rho}\nabla\rho\Big) \quad\Longrightarrow\quad \boxed{\dot\rho = \Delta\rho}\ \text{(the heat equation)}.$$

Example 4.3 (Adding a potential: Fokker–Planck)**.** Take the energy

$$E(\rho) = \int V(x)\,\mathrm{d}\rho(x) + H(\rho), \qquad \mathrm{D}E(\rho) = V(x) + \log\rho(x).$$

The board highlights the arguments x in red, and the point is worth keeping: $\mathrm{D}E(\rho)$ is a *function on* Ω, not a number. That is exactly what makes $\nabla\,\mathrm{D}E(\rho)$ meaningful when it is fed to $\mathbb{K}_{\mathrm{OT}}(\rho)\xi = -\nabla\cdot(\rho\nabla\xi)$. Then $(\mathcal{P}(\Omega), E, \mathbb{K}_{\mathrm{OT}})$ generates the continuity equation

$$\dot\rho = -\operatorname{grad}_{\mathrm{OT}} E(\rho) = \nabla\cdot\big(\rho(\nabla V + \nabla\log\rho)\big) = \nabla\cdot(\rho\,\nabla V) + \Delta\rho \qquad \text{(Fokker–Planck, FPE)}.$$

4.2 The Hellinger gradient flow

Now let $\mathcal{M}^+(\Omega)$ be the positive measures on $\Omega \subseteq \mathbb{R}^d$ (mass need not equal 1, in contrast to $\rho \in \mathcal{P}(\Omega)$ where $\int\rho = 1$). With the Boltzmann entropy H and the Hellinger geometry

$$\mathbb{K}_{\mathrm{He}}(\rho)\,\xi := \rho\,\xi,$$

the gradient flow is

$$\dot\rho = -\operatorname{grad}_{\mathrm{He}} H(\rho) = -\rho\,\psi_B'(\rho) = -\rho\log\rho.$$

Since $\frac{\mathrm{d}}{\mathrm{d}t}(\log\rho) = -\log\rho$, we get $\log\rho(t) = e^{-t}\log\rho(0)$, i.e.

$$\boxed{\rho(t) = \rho(0)^{\,e^{-t}}.}$$

Remark. In the Otto–Wasserstein system $(\mathcal{P}(\Omega), H, \mathbb{K}_{\mathrm{OT}})$, a Dirac initial datum $\rho(0) = \delta_{x_0}$ spreads out (mass transport), consistent with the heat-equation dynamics.

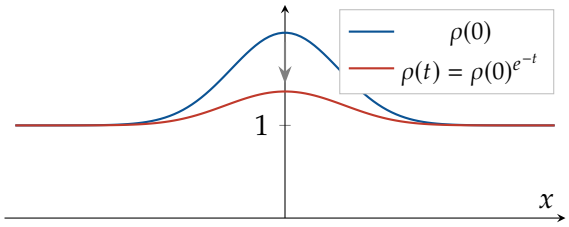


Figure 9. Hellinger flow flattens the density *pointwise* toward the value 1 (as $t \to \infty$, $\rho(t) = \rho(0)^{e^{-t}} \to 1$), with no mass transport – in contrast to the Otto–Wasserstein flow below.

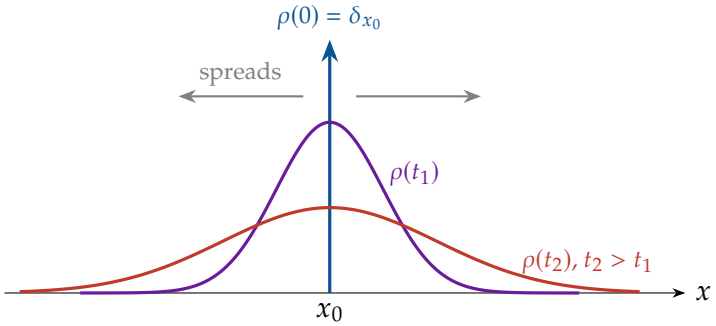


Figure 10. Under the Otto–Wasserstein (heat) flow the point mass δ_{x_0} instantaneously becomes the smooth density $\mathcal{N}(x_0, 2t)$, which widens *symmetrically about* x_0: the barycentre never moves and the total mass stays 1 (both bumps are drawn to the same scale, so they enclose equal area). Contrast Figure 9, where the Hellinger flow rescales the density pointwise and transports nothing; and Figure 16, where an added drift b is what makes the bump translate.

4.3 Relative entropy, and the spherical Hellinger (Fisher–Rao) flow

The computation above used the plain Boltzmann entropy. Run the same geometry against the *relative* entropy and the flow acquires a target. For a general energy F on $\mathcal{M}^+(\Omega)$ the Hellinger gradient system $(\mathcal{M}^+(\Omega), F, \mathbb{K}_{\mathrm{He}})$ generates

$$\dot\rho = -\mathbb{K}_{\mathrm{He}}(\rho)\,\frac{\delta F}{\delta \rho}(\rho) = -\,\rho\,\frac{\delta F}{\delta \rho}(\rho).$$

Example 4.4 (Hellinger flow of the relative entropy)**.** Let $F : \rho \mapsto H(\rho \,\|\, \pi)$. Because (14) carries the mass-correction terms $-\mu + \nu$, its first variation is exactly the log-ratio,

$$\frac{\delta F}{\delta \rho}(\rho)(x) = \log \rho(x) - \log \pi(x),$$

so the flow is $\dot\rho = -\rho\big(\log\rho - \log\pi\big)$. Setting $u := \log\rho - \log\pi$ turns this into $\dot u = \dot\rho/\rho = -u$, whence $u(t) = e^{-t}u(0)$ and

$$\boxed{\rho_t \;=\; \rho_0^{\,e^{-t}}\,\pi^{\,1-e^{-t}}.} \tag{12}$$

Taking $\pi \equiv 1$ recovers the boxed formula of §4.2. The flow is a pointwise geometric interpolation from ρ_0 to π: mass is destroyed where $\rho_0 > \pi$ and created where $\rho_0 < \pi$, and *nothing moves*. Contrast the Otto–Wasserstein flow of the *same* energy, which is Fokker–Planck: same state space, same energy, different dissipation geometry R, completely different equation.

The Hellinger flow does not preserve mass — which is why its state space was $\mathcal{M}^+(\Omega)$. Projecting it back onto the probability simplex gives the last geometry of the course.

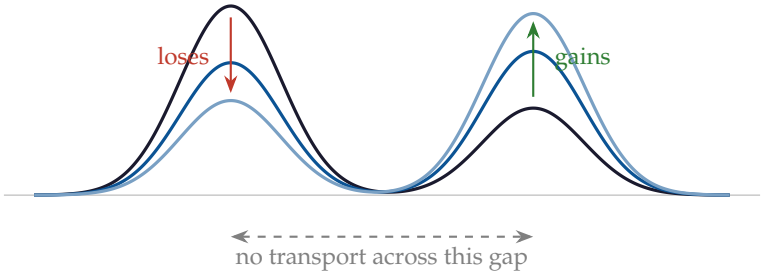


Figure 11. The Hellinger geometry *teleports* mass rather than moving it: the modes stay put and only their heights change (here toward a target π weighted to the right). Compare Figure 9 (pointwise flattening toward 1, the case $\pi \equiv 1$) and Figure 10 (Otto–Wasserstein, where mass genuinely travels). This is why the Hellinger flow crosses a low-density barrier instantly while Wasserstein flows must push mass through it.

Definition 4.5 (Spherical Hellinger / Fisher–Rao / Bhattacharyya)**.** On $\mathcal{P}(\Omega)$ put

$$\boxed{\mathbb{K}_{\mathrm{B}}(\rho)\,\xi \;:=\; \rho\Big(\xi - \int_\Omega \rho\,\xi\Big).}$$

The gradient system $\big(\mathcal{P}(\Omega), F, \mathbb{K}_{\mathrm{B}}\big)$ generates

$$\dot\rho = -\mathbb{K}_{\mathrm{B}}(\rho)\frac{\delta F}{\delta \rho}(\rho) = -\,\rho\Big(\frac{\delta F}{\delta \rho}(\rho) - \int_\Omega \rho\,\frac{\delta F}{\delta \rho}(\rho)\Big).$$

The subtracted mean is precisely what keeps the flow on the simplex: writing $\xi = \frac{\delta F}{\delta \rho}(\rho)$ and using $\int_\Omega \rho = 1$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\int_\Omega \rho\Big) = \int_\Omega \dot\rho = -\int_\Omega \rho\,\xi + \Big(\int_\Omega \rho\Big)\int_\Omega \rho\,\xi = 0, \qquad \Longrightarrow \qquad \text{mass} = \text{const.}$$

This is the replicator equation of evolutionary game theory, and it is the measure-level form of the Fisher–Rao natural gradient met in the Bernoulli example of §4.5. We will meet it once more in §10, restricted to a Gaussian family.

4.4 The continuity equation

Two equivalent descriptions of transport:

$$\partial_t u = -\nabla\cdot(u\,v) \qquad (\mathcal{E}), \qquad\qquad \dot x(t) = v(x(t)), \quad x(0) = x_0 \qquad (\ell).$$

Theorem 4.6 (Divergence theorem, Gauß–Green)**.** *Let $E \subset \mathbb{R}^d$ be a bounded domain with C^1 boundary $S = \partial E$ and outward normal n. For sufficiently smooth w,*

$$\int_S w\cdot n\;\mathrm{d}S = \int_E \nabla\cdot w\;\mathrm{d}x.$$

Derivation of the continuity equation. Track the mass $M(t) = \int_E \rho_t\,\mathrm{d}x$. The flux through the boundary gives

$$\frac{\mathrm{d}}{\mathrm{d}t}M(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_E \rho_t\;\mathrm{d}x = -\int_{\partial E}\rho_t\,\vec v\cdot\vec n\;\mathrm{d}S \overset{\text{Div. Thm}}{=} -\int_E \nabla\cdot(\rho v)\;\mathrm{d}x.$$

Therefore, for arbitrary E,

$$\int_E \Big(\;\underbrace{\dot\rho_t + \nabla\cdot(\rho_t v)}_{\text{pointwise continuity eq.}}\;\Big)\,\mathrm{d}x = 0 \quad\Longrightarrow\quad \dot\rho_t + \nabla\cdot(\rho_t v) = 0.$$

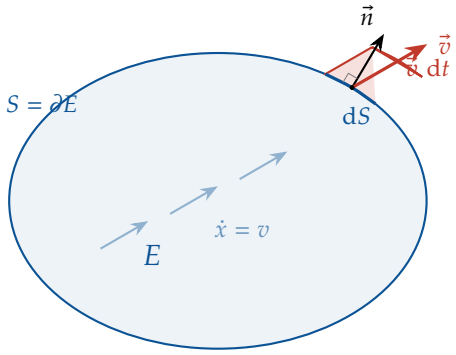


Figure 12. Flux picture behind the continuity equation. In time $\mathrm{d}t$ the mass crossing a boundary element $\mathrm{d}S$ fills the thin sliver swept by the flow, of signed height $\vec{v}\cdot\vec{n}\;\mathrm{d}t$ along the outward normal $\vec{n}$. Summing $\rho\,\vec{v}\cdot\vec{n}$ over $S=\partial E$ and applying the divergence theorem gives $\frac{\mathrm{d}}{\mathrm{d}t}\int_E \rho = -\int_S \rho\,\vec{v}\cdot\vec{n}\;\mathrm{d}S = -\int_E \nabla\cdot(\rho v)\;\mathrm{d}x$.

Recall $\mathbb{K}_{\mathrm{OT}}(\rho)\xi = -\nabla\cdot(\rho\,\nabla\xi)$, matching the velocity $v=\nabla\xi$. In weak form, for test functions $\varphi\in C_c^\infty$,

$$\int_0^\infty\int_{\mathbb{R}^d}\big(\partial_t\varphi+\nabla\varphi\cdot v\big)\,\rho(x,t)\;\mathrm{d}x\;\mathrm{d}t=0.$$

4.5 An information-geometry example: Bernoulli MLE

Consider the Bernoulli model $P_p(X=x)=p^{\,x}(1-p)^{1-x}$, $x\in\{0,1\}$. The log-likelihood of one sample is

$$\ell(p,x)=x\log p+(1-x)\log(1-p).$$

Maximum likelihood estimation minimizes the (average) negative log-likelihood

$$p^*\in\arg\min_p\underbrace{\left\{-\tfrac{1}{n}\log\prod_{i=1}^n P_p(X=x_i)\right\}}_{=:F(p)}.$$

The *score* and *Fisher information* are

$$\nabla_p\log P_p(x)=\frac{x-p}{p(1-p)}=:S_p(x),\qquad I(p)=\mathbb{E}_x\big[S_p(x)^2\big]=\frac{1}{p(1-p)}=:\mathsf{G}(p).$$

Endow the parameter space with the Fisher–Rao metric $g_p(u,v)=u\,I(p)\,v$, so that the Onsager operator is the inverse Fisher information,

$$\mathbb{K}(p)=\mathsf{G}(p)^{-1}=I(p)^{-1}=p(1-p).$$

Writing $\bar{x}=\frac{1}{n}\sum_i x_i$ for the sample mean, one computes $\nabla F(p)=\dfrac{p-\bar{x}}{p(1-p)}$,[5] so the Riemannian (natural-gradient) flow is

$$\dot{p}=-\operatorname{grad}F(p)=-\mathbb{K}(p)\nabla F(p)=-\,p(1-p)\cdot\frac{p-\bar{x}}{p(1-p)}=-(p-\bar{x}).$$

[5]Using the $\frac{1}{n}$-averaged log-likelihood keeps F consistent with the per-observation score S_p and information $I(p)$ above, and makes the natural-gradient flow the clean $\dot{p}=-(p-\bar{x})$. The whiteboard omits the $\frac{1}{n}$.

The Fisher–Rao natural gradient turns the curved MLE problem into the *Euclidean* gradient flow of $\frac{1}{2}\left\|p - \bar{x}\right\|^2$:

$$\dot{p} = -(p - \bar{x}) \qquad \Longrightarrow \qquad p(t) = e^{-t}\, p(0) + \left(1 - e^{-t}\right) \bar{x}.$$

Week 2

Divergences, Sampling, and Optimal Transport

5 Comparing Measures: f-Divergences

Week 1 built gradient systems (X, F, R) and identified the energy F as the object being dissipated. When the state space is a space of measures, the energies that matter in statistics and machine learning are almost always *discrepancies* between measures. This section collects the standard ones.

5.1 The Csiszár f-divergence

The question is simply: given $\mu, \nu \in \mathcal{P}(\Omega)$ (or, more generally, in $\mathcal{M}^+(\Omega)$), how do we compare them?

Definition 5.1 (Divergence generator)**.** A *divergence generator* is a proper, lower semicontinuous, convex function

$$f : [0, +\infty) \longrightarrow \mathbb{R} \cup \{+\infty\}$$

normalized by

$$f(1) = 0, \qquad 0 \in \partial f(1),$$

and strictly convex at 1. When f is C^2 near 1 — true for every generator below except the total-variation one, whose graph has a corner at $s = 1$ — this is exactly the form written on the board,

$$f(1) = f'(1) = 0, \qquad f''(1) > 0.$$

The normalization $f(1) = 0$ makes the divergence vanish when $\mu = \nu$; $0 \in \partial f(1)$ (i.e. $f'(1) = 0$ in the smooth case) is a harmless calibration — replacing $f(s)$ by $f(s) - p\,(s-1)$ for a subgradient p changes nothing when μ, ν have the same mass; and strict convexity at 1 makes the divergence nondegenerate. That last curvature is exactly the Fisher information of §4.5: for a smooth model $\theta \mapsto \mu_\theta$, writing $r = \mathrm{d}\mu_{\theta+\mathrm{d}\theta}/\mathrm{d}\mu_\theta = 1 + \delta$ with $\delta \approx \langle \nabla_\theta \log \mu_\theta, \mathrm{d}\theta \rangle$ and Taylor-expanding $f(1+\delta) \approx \frac{1}{2} f''(1)\delta^2$ gives

$$D_f\big(\mu_{\theta+\mathrm{d}\theta} \,\|\, \mu_\theta\big) \;\approx\; \tfrac{1}{2}\, f''(1)\; \mathrm{d}\theta^{\mathsf{T}} I(\theta)\, \mathrm{d}\theta, \qquad I(\theta) = \mathbb{E}\big[\nabla_\theta \log \mu_\theta\, \nabla_\theta \log \mu_\theta^{\mathsf{T}}\big]. \tag{13}$$

So every f-divergence whose generator is C^2 near 1 induces the same Fisher–Rao metric infinitesimally, up to the scalar $f''(1)$; they differ only in the large. (Total variation, $f(s) = \frac{1}{2}|s-1|$, is excluded: it has a corner at $s = 1$, no $f''(1)$, and it is already first order in $\mathrm{d}\theta$.)

Definition 5.2 (f-divergence, Csiszár)**.** For μ, ν with Lebesgue decomposition $\mu = r\nu + \mu^\perp$, $r = \dfrac{\mathrm{d}\mu}{\mathrm{d}\nu}$ the Radon–Nikodym derivative,

$$D_f(\mu \,\|\, \nu) := \begin{cases} \displaystyle\int_\Omega f\!\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu, & \mu \ll \nu, \\ +\infty, & \mu \not\ll \nu. \end{cases}$$

Remark. The singular part can be kept rather than discarded: writing $f'_\infty := \lim_{s\to\infty} f(s)/s \in (0, +\infty]$ for the recession slope, the extended definition is $D_f(\mu\|\nu) = \int_\Omega f(r)\,\mathrm{d}\nu + f'_\infty\,\mu^\perp(\Omega)$, which reduces to Definition 5.2 exactly when $f'_\infty = +\infty$.

Proposition 5.3 (Nonnegativity)**.** *$D_f(\mu\|\nu) \geq 0$ for all $\mu, \nu \in \mathcal{M}^+(\Omega)$. If moreover f is strictly convex at 1, then $D_f(\mu\|\nu) = 0$ iff $\mu = \nu$.*

Proof. Convexity gives $f(s) \geq f(1) + p\,(s-1)$ for every subgradient $p \in \partial f(1)$, and the normalization supplies $f(1) = 0$ with $p = 0$ admissible, whence the pointwise bound

$$f(s) \;\geq\; 0 \qquad \text{for all } s \geq 0,$$

so $D_f(\mu\|\nu) = \int_\Omega f(r)\,\mathrm{d}\nu \geq 0$ with $r = \mathrm{d}\mu/\mathrm{d}\nu$. If f is strictly convex at 1 then $f(s) > 0$ for every $s \neq 1$, so $\int f(r)\,\mathrm{d}\nu = 0$ forces $r = 1$ ν-a.e., i.e. $\mu = \nu$. □

Remark (The Jensen route). When *both* $\mu, \nu \in \mathcal{P}(\Omega)$ the same bound follows from Jensen's inequality applied to the convex f and the probability measure ν,

$$\int_\Omega f\!\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu \;\geq\; f\!\left(\int_\Omega \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu\right) = f\big(\mu(\Omega)\big) = f(1) = 0,$$

which is the argument on the board. Note that it uses $\mu(\Omega) = 1$ in the last step, so on $\mathcal{M}^+(\Omega)$ — where the mass-comparison terms of Example 5.4 are exactly the point — one must argue pointwise as above.

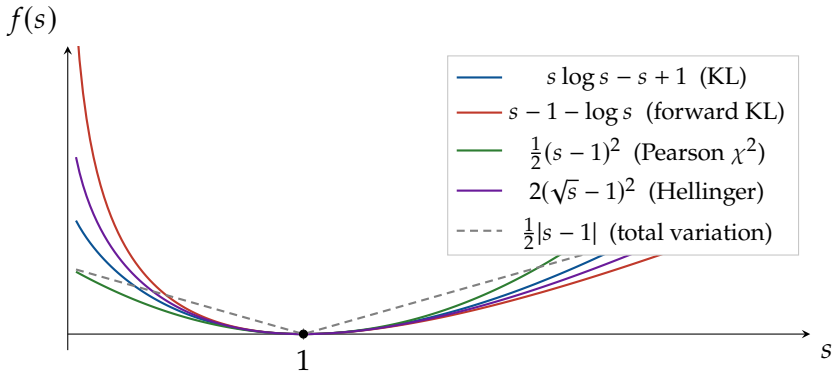


Figure 13. Divergence generators. Every curve is convex and vanishes at $s = 1$ with 0 in its subdifferential there — tangentially for the four smooth ones (so $f(1) = f'(1) = 0$), with a corner for the dashed TV generator. For the smooth ones the local curvature $f''(1)$ is the Fisher information carried by the divergence, cf. (13). Behaviour on mutually singular measures is decided by the *pair* $(f(0), f'_\infty)$, since there $D_f(\mu\|\nu) = f(0)\nu(\Omega) + f'_\infty\mu(\Omega)$: only TV $(\frac{1}{2}, \frac{1}{2})$ and Hellinger $(2, 2)$ stay finite. KL and χ^2 have $f'_\infty = +\infty$; forward KL grows only linearly ($f'_\infty = 1$) yet still blows up because $f(0) = +\infty$ — visible as the curve escaping the frame at the left edge.

5.2 Kullback–Leibler divergence (relative entropy)

Example 5.4 (KL)**.** Take $f(s) = s\log s - s + 1$ (indeed $f(1) = 0$, $f'(s) = \log s$ so $f'(1) = 0$, and $f''(s) = 1/s$ gives $f''(1) = 1 > 0$). Then

$$H(\mu\,\|\,\nu) \;=\; \int_\Omega \Big(\mu\log\frac{\mu}{\nu} - \mu + \nu\Big). \tag{14}$$

The terms $-\mu + \nu$ *compare the masses*; they cancel when $\mu, \nu \in \mathcal{P}(\Omega)$ but are what makes (14) the right object on $\mathcal{M}^+(\Omega)$.

Notation. This quantity travels under many names and symbols: D_{KL}, KL, H, D_1 (the $\alpha = 1$ member of the Rényi/Tsallis family), H_B (relative Boltzmann entropy). We write $H(\mu\|\nu)$ throughout.

Example 5.5 (Forward KL)**.** Reversing the arguments is again an f-divergence: with

$$f_{\mathrm{FKL}}(s) = s - 1 - \log s$$

one gets $D_f(\mu\|\nu) = H(\nu\,\|\,\mu)$. Indeed $\int(\frac{\mu}{\nu} - 1 - \log\frac{\mu}{\nu})\,\mathrm{d}\nu = \int(\mu - \nu + \nu\log\frac{\nu}{\mu})$.

Remark (Mode-seeking vs. mass-covering). $H(\mu\|\nu)$ — with μ in the first slot, the "reverse" KL of the variational-inference literature — charges $+\infty$ where $\mu > 0$ but $\nu = 0$: it forces μ to hide inside the support of ν, and is *mode-seeking*. The "forward" KL $H(\nu\|\mu)$, the MLE objective of §5.4, forces μ to cover the support of ν, and is *mass-covering*.

Example 5.6 (Multivariate Gaussians)**.** Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ on $\mathbb{R}^d$. Then

$$H(\mu\,\|\,\nu) = \frac{1}{2}\Big[(m_0 - m_1)^{\top}\Sigma_1^{-1}(m_0 - m_1) + \operatorname{tr}\big(\Sigma_1^{-1}\Sigma_0\big) - d + \log\frac{\det\Sigma_1}{\det\Sigma_0}\Big]. \tag{15}$$

If both covariances are diagonal, $\Sigma_0 = \operatorname{diag}(b_0^i)$, $\Sigma_1 = \operatorname{diag}(b_1^i)$, and $m_0 = m_1$, this collapses to the scalar sum[6]

$$H(\mu\,\|\,\nu) = \frac{1}{2}\left[\sum_{i=1}^{d}\frac{b_0^i}{b_1^i} - d + \sum_{i=1}^{d}\log\frac{b_1^i}{b_0^i}\right].$$

Each summand is $\frac{1}{2}(t - 1 - \log t)$ with $t = b_0^i/b_1^i$: nonnegative, vanishing iff $t = 1$.

5.3 A catalogue of divergences

Example 5.7 (Pearson χ^2)**.** $f(s) = \frac{1}{2}(s-1)^2$ gives

$$\chi^2(\mu\,\|\,\nu) = \tfrac{1}{2}\int_{\Omega}\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big)^2\,\mathrm{d}\nu.$$

Here $\mathrm{d}\mu/\mathrm{d}\nu$ is the Radon–Nikodym density, i.e. $\mu(\mathrm{d}x) = r(x)\,\nu(\mathrm{d}x)$.

Example 5.8 (Squared Hellinger distance)**.** $f(s) = 2(\sqrt{s} - 1)^2$ gives

$$D_f(\mu\,\|\,\nu) = 2\int_{\Omega}\Big(\sqrt{\tfrac{\mathrm{d}\mu}{\mathrm{d}\nu}} - 1\Big)^2\,\mathrm{d}\nu = 2\int_{\Omega}\big(\sqrt{p} - \sqrt{q}\big)^2\,\mathrm{d}x \;=\; 2\,\mathrm{He}^2(\mu,\nu),$$

writing p, q for densities of μ, ν and $\mathrm{He}^2(\mu,\nu) := \int(\sqrt{p} - \sqrt{q})^2$.[7]

Example 5.9 (Total variation)**.** $f(s) = \frac{1}{2}|s - 1|$ gives

$$\big\|\mu - \nu\big\|_{\mathrm{TV}} = \tfrac{1}{2}\int_{\Omega}\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big|\,\mathrm{d}\nu = \tfrac{1}{2}\int_{\Omega}\big|p - q\big|\,\mathrm{d}x = \sup_{A\subseteq\Omega}\big|\mu(A) - \nu(A)\big|.$$

This is the only generator in Figure 13 that is not differentiable at $s = 1$. It also has *both* $f(0) = \frac{1}{2} < \infty$ and $f'_\infty = \frac{1}{2} < \infty$, and that pair of conditions — not either one alone — is what keeps a divergence finite on mutually singular measures: for $\mu \perp \nu$ the extended formula gives $D_f(\mu\|\nu) = f(0)\,\nu(\Omega) + f'_\infty\,\mu(\Omega)$, hence $\big\|\mu - \nu\big\|_{\mathrm{TV}} = 1$. Among the five generators of Figure 13 only TV and Hellinger ($f(0) = f'_\infty = 2$, giving $2\mathrm{He}^2 = 4$) qualify. Forward KL has $f'_\infty = 1 < \infty$ but $f(0) = +\infty$, so it still blows up; KL and χ^2 have $f'_\infty = +\infty$.

[6] The whiteboard writes the diagonal case without the overall factor $\frac{1}{2}$; it is inherited from (15).

[7] Constants in front of the Hellinger distance differ between authors (the board writes $\frac{1}{2}\mathrm{He}^2$ for the same quantity, i.e. its He^2 is four times ours); only the fixed choice matters, not the constant. Note that He itself (the square root) is a genuine *metric* on $\mathcal{P}(\Omega)$, as is the total-variation distance below; the KL and χ^2 entries are not.

5.4 Maximum likelihood is forward-KL minimization

Let $\{\mu_\theta\}_{\theta\in\Theta}$ be a parametric model, e.g. $\theta = (m, \Sigma)$ with $\mu_\theta \in \mathcal{P}(\mathbb{R}^d)$, and let the data $x_1, \dots, x_n$ be i.i.d. samples from a data-generating distribution ν.

Proposition 5.10 (MLE = forward KL)**.** *Maximum likelihood estimation minimizes the average negative log-likelihood,*

$$\min_\theta \underbrace{-\frac{1}{n}\sum_{i=1}^n \log \mu_\theta(x_i)}_{\text{empirical}} \xrightarrow[n\to\infty]{} -\int_\Omega \log \mu_\theta(x)\, \mathrm{d}\nu(x),$$

and since $\int \nu \log \nu$ *does not depend on* θ*,*

$$\int_\Omega \nu \log \nu - \int_\Omega \nu \log \mu_\theta = H(\nu \,\|\, \mu_\theta) \qquad \Longrightarrow \qquad \boxed{\theta^* \in \arg\min_\theta \; H(\nu \,\|\, \mu_\theta).}$$

(For probability measures the mass-comparison term $\int(-\mu_\theta + \nu)$ *vanishes.)*

So MLE is a *forward*-KL projection of the (unknown) data distribution onto the model family: the estimator is pushed to cover all of the data, and pays $+\infty$ for assigning zero likelihood to an observed point (Figure 14).

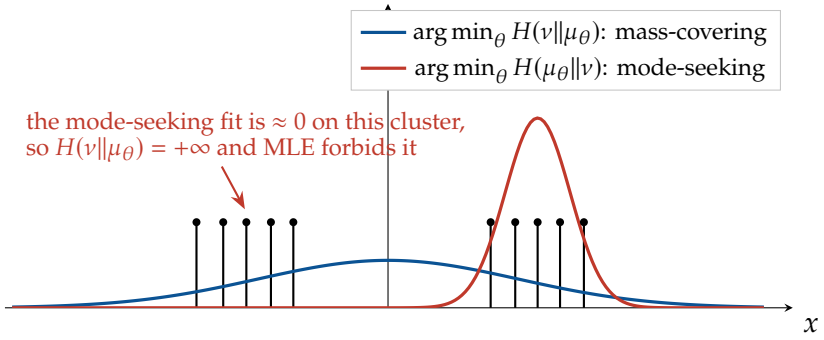


Figure 14. Why MLE is *mass-covering*. The data (spikes) come from a bimodal ν; the model class here is single Gaussians. Minimizing the forward KL $H(\nu\|\mu_\theta)$ — which is what maximum likelihood does — forces μ_θ to put mass wherever ν does, because a single observed sample in a region where μ_θ vanishes costs $+\infty$; the fit therefore stretches across both clusters and fits neither well. The reverse KL $H(\mu_\theta\|\nu)$ has the opposite failure mode: it may sit happily on one mode and ignore the other.

5.5 Bayesian inference: the posterior as a Gibbs measure

Example 5.11 (Bayesian logistic regression)**.** Given data $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$, and parameter $\beta \in \mathbb{R}^d$, the model is

$$y_i \mid \beta \;\sim\; \mathrm{Bern}\big(b(x_i^\top \beta)\big), \qquad b(t) = \frac{1}{1+e^{-t}}.$$

With a standard Gaussian prior the (unnormalized) posterior is

$$p(\beta) \;\propto\; e^{-\frac{\|\beta\|^2}{2}} \prod_{i=1}^n \Big(\frac{1}{1+e^{-x_i^\top\beta}}\Big)^{y_i} \Big(\frac{e^{-x_i^\top\beta}}{1+e^{-x_i^\top\beta}}\Big)^{1-y_i}.$$

Define the *potential* $V(\beta) := -\log p(\beta) + \text{const}$, so that $p \propto e^{-V}$. Explicitly[8]

$$V(\beta) = \tfrac{1}{2}\|\beta\|^2 + \sum_{i=1}^{n}\Big[y_i \log\big(1 + e^{-x_i^\top \beta}\big) + (1 - y_i)\log\big(1 + e^{x_i^\top \beta}\big)\Big], \tag{16}$$

with gradient $\nabla V(\beta) = \beta + \sum_{i=1}^{n}\big(b(x_i^\top \beta) - y_i\big)x_i$.

Now *lift* the finite-dimensional problem to a problem over measures. Take the energy

$$F(\mu) := H(\mu \,\|\, p) = -\int \mu \log p + \int \mu \log \mu = \underbrace{\int_\Omega V(\beta)\,\mathrm{d}\mu(\beta)}_{\text{potential}} + \underbrace{\int_\Omega \mu \log \mu}_{\text{entropy}} + \text{const}.$$

Gibbs measure as the minimizer. The variational derivative is

$$\frac{\delta F}{\delta \mu}(x) = V(x) + \log \mu(x) + \text{const},$$

and setting it equal to a constant (the Lagrange multiplier for $\int \mu = 1$) gives

$$\mu^* = \frac{1}{Z} e^{-V}, \qquad Z = \int_\Omega e^{-V},$$

the *Gibbs measure*. Thus $\mu^* \in \arg\min_\mu F(\mu)$ and $\mu^* \propto p$: sampling the posterior *is* minimizing F.

Remark (The lift). Dropping the entropy leaves the *linear* functional $\mu \mapsto \int V\,\mathrm{d}\mu$, whose minimizers over $\mathcal{P}(\Omega)$ are exactly the measures supported on $\arg\min V$ — in particular δ_{β^*} with $\beta^* = \arg\min V$. So $\min_\beta V(\beta)$ lifts to a *linear program over measures*, and the entropy term is a regularizer that smears the Dirac into the Gibbs measure. This is the same lift that turns Monge's problem into Kantorovich's in §7.

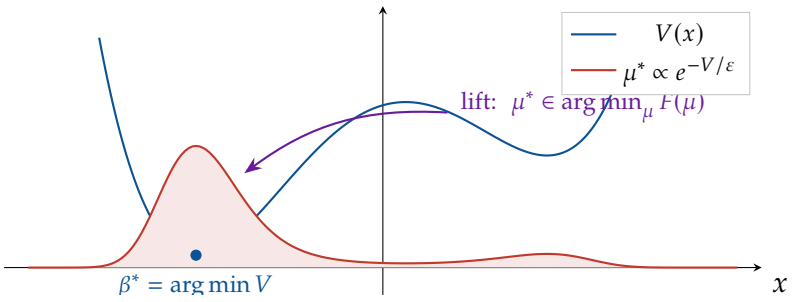


Figure 15. Lifting optimization to the space of measures. The minimizer β^* of the potential V becomes, after adding the entropy, the Gibbs measure $\mu^* \propto e^{-V/\varepsilon}$ minimizing $F(\mu) = \int V\,\mathrm{d}\mu + \varepsilon \int \mu \log \mu$. As $\varepsilon \downarrow 0$ the Gibbs measure concentrates on $\arg\min V$; the secondary well is visited with exponentially small probability.

5.6 Variational representation and f-GANs

Convex duality turns an f-divergence into a supremum over test functions — which is what makes it estimable from samples, and hence trainable.

[8] The whiteboard keeps the log-likelihood terms with the signs they carry inside $\log p$; taking $V = -\log p$ flips them, which is the form written here. Both $\log(1 + e^{-t})$ and $\log(1 + e^{t})$ are convex, so V is 1-strongly convex thanks to the prior term — a fact we will use in §8.5.

Let $f^*(t) = \sup_s\{st - f(s)\}$ be the convex conjugate, so that $f(r) = \sup_{t\in\mathbb{R}}\{rt - f^*(t)\}$ by biconjugation (recall f is proper, lsc, convex). For the quadratic $f(s) = \frac{1}{2}\|s\|_2^2$ one has $f = f^*$.

Proposition 5.12 (f-GAN variational bound)**.** *Let ν_θ be a generated distribution, μ the data distribution, $r_\theta = \mathrm{d}\mu/\mathrm{d}\nu_\theta$, and assume $\mu \ll \nu_\theta$ (otherwise read D_f in the extended sense of the remark after Definition 5.2, which is what the right-hand side computes). Then*

$$D_f(\mu\,\|\,\nu_\theta) = \int_\Omega f(r_\theta)\,\mathrm{d}\nu_\theta = \sup_T \int_\Omega \big(r_\theta T - f^*(T)\big)\,\mathrm{d}\nu_\theta \;=\; \sup_T \Big\{\int_\Omega T\,\mathrm{d}\mu - \int_\Omega f^*(T)\,\mathrm{d}\nu_\theta\Big\},$$

the supremum being over measurable $T : \Omega \to \operatorname{dom} f^*$. Restricting T to a parametric class $\{T_\omega\}$ (a neural network, the "discriminator") turns "=" into "≥" and gives the f-GAN objective

$$\min_\theta \max_\omega \; \mathbb{E}_{x\sim\mu}\big[T_\omega(x)\big] - \mathbb{E}_{x\sim\nu_\theta}\big[f^*(T_\omega(x))\big].$$

Both integrals are now expectations, so both are estimable from samples — μ is never needed in closed form. This is the mechanism by which every divergence in Figure 13 becomes an implementable training loss.

6 Sampling as a Gradient Flow

We now put the energies of §5.5 into the Otto–Wasserstein gradient structure of Week 1 and read off algorithms.

6.1 Fokker–Planck, revisited

Recall from Week 1 the Otto (inverse-metric) Onsager operator

$$\mathbb{K}_{\mathrm{OT}}(\rho):\ \xi \longmapsto -\nabla\cdot\big(\rho\,\nabla\xi\big).$$

The Fokker–Planck equation is steepest descent of the relative entropy $H\big(\rho\,\|\,\frac{1}{Z}e^{-V}\big)$ in the dissipation geometry $\mathbb{K}_{\mathrm{OT}}$, i.e. in 2-Wasserstein space:

$$\partial_t\rho = \nabla\cdot\big(\rho\,\nabla V\big) + \Delta\rho, \qquad \text{equivalently} \qquad \dot\rho = -\mathbb{K}_{\mathrm{OT}}(\rho)\,\mathrm{D}F(\rho).$$

Theorem 6.1 (Fundamental solution with constant drift)**.** *Let $b \in \mathbb{R}^d$ be constant. The Cauchy problem*

$$\begin{cases}\partial_t\rho = \nabla\cdot(\rho\,b) + \Delta\rho,\\ \rho(0) = \delta_{x_0},\end{cases}$$

has, for every $t > 0$, the (weak, fundamental) solution

$$\boxed{\rho(t) = \mathcal{N}\big(x_0 - bt,\ 2t\,I\big).}$$

Sketch. The associated SDE is $\mathrm{d}X_t = -b\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$, whose solution $X_t = x_0 - bt + \sqrt{2}\,W_t$ is Gaussian with mean $x_0 - bt$ and covariance $2t\,I$; the law of X_t solves the Fokker–Planck equation above. □

6.2 The unadjusted Langevin algorithm (ULA)

Discretizing the Fokker–Planck flow in time gives an algorithm — "noisy gradient descent".

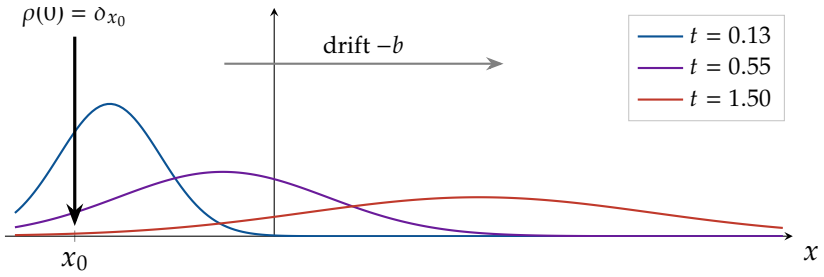


Figure 16. The fundamental solution of Theorem 6.1: the Dirac mass at x_0 instantaneously becomes a Gaussian which *translates* at velocity $-b$ (the drift) while its variance *grows* linearly, $\Sigma(t) = 2tI$ (the diffusion). Drift and diffusion are exactly the two terms $\nabla\cdot(\rho b)$ and $\Delta\rho$. The three curves are the exact densities $\mathcal{N}(x_0 - bt, 2t)$ at $t = 0.13, 0.55, 1.5$ for $x_0 = -2$ and drift speed $-b = 2.7$, so both the displacement $\propto t$ and the spread $\propto \sqrt{t}$ can be read off directly.

Algorithm (ULA). Draw $X_0 = x_0 \sim \mu_0$. Fix a step size $\tau = t_{k+1} - t_k$. At time t_k,

$$X_{t_{k+1}} \longleftarrow X_{t_k} - \tau\,\nabla V(X_{t_k}) + \sqrt{2\tau}\,Z_k, \qquad Z_k \sim \mathcal{N}(0, I) \text{ i.i.d.}$$

At the level of laws this is one pushforward followed by one Gaussian smoothing:

$$\rho^{t_{k+1}} = \big(\mathrm{Id} - \tau\nabla V\big)_{\#}\rho^{t_k} \;*\; \mathcal{N}(0, 2\tau I).$$

Remark. The two operations are precisely the two halves of the energy: the pushforward by $\mathrm{Id} - \tau\nabla V$ is a step of the Wasserstein gradient flow of the *potential* $\int V\,\mathrm{d}\rho$; the Gaussian convolution is a step of the heat flow, i.e. the Wasserstein gradient flow of the *entropy*. ULA is an operator splitting of $F = \int V\,\mathrm{d}\rho + \int \rho\log\rho$.

Definition 6.2 (Langevin SDE)**.** Let W_t denote Brownian motion (the Wiener process), so that increments satisfy $X_t - X_s = -b\,(t-s) + \sqrt{2}\,(W_t - W_s)$ in the constant-drift case. In differential form

$$\mathrm{d}X_t = -b\;\mathrm{d}t + \sqrt{2}\;\mathrm{d}W_t,$$

and the special choice $b = \nabla V(X_t)$ gives the *Langevin SDE*

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

whose law solves the Fokker–Planck equation and converges to $\pi \propto e^{-V}$.

Example 6.3 (ULA for Bayesian logistic regression)**.** Plug the potential (16) of Example 5.11 and its gradient $\nabla V(\beta) = \beta + \sum_i \big(b(x_i^\top\beta) - y_i\big)x_i$ into ULA. Each iteration costs one pass over the data plus one Gaussian draw, and the iterates approximate posterior samples.

Remark ("Unadjusted"). The time discretization introduces an $O(\tau)$ bias: the chain converges not to π but to a nearby measure. Adding a Metropolis–Hastings accept/reject step removes the bias (MALA); leaving it out is what the word *unadjusted* records.

6.3 Stein geometry and SVGD

Changing the *geometry* $\mathbb{K}$ while keeping the *energy* F produces a different algorithm for the same target. This is the whole point of the gradient-system formalism.

Let $\mathcal{P}(\Omega)$ be the probability measures on Ω and let

$$H(\rho\,\|\,\pi) = \begin{cases} \displaystyle\int_\Omega \rho\log\frac{\rho}{\pi} - \rho + \pi, & \rho \ll \pi, \\ +\infty, & \text{otherwise,} \end{cases}$$

be the relative entropy against the target $\pi \propto e^{-V}$.

Definition 6.4 (Stein/Onsager operator)**.** Precompose the Otto operator with a kernel smoothing:

$$\mathbb{K}_{\mathrm{OT}}(\rho)\xi := -\nabla\cdot\big(\rho\,\nabla\xi\big) \qquad \rightsquigarrow \qquad \boxed{\mathbb{K}_{\mathrm{Stein}}(\rho)\,\xi := -\nabla\cdot\big(\rho\,K_\rho\nabla\xi\big),}$$

where K is a symmetric positive-definite kernel (§2.3) and K_ρ is the kernel integral operator

$$K_\rho:\ \eta \longmapsto \int_\Omega K(y,\cdot)\,\eta(y)\,\mathrm{d}\rho(y).$$

The resulting gradient-flow equation is

$$\partial_t\rho = \nabla\cdot\big(\rho\,K_\rho\nabla\,\mathrm{D}H(\rho)\big), \qquad \mathrm{D}H(\rho) = \log\rho - \log\pi.$$

Unwinding the velocity field and using $\rho\nabla\log\rho = \nabla\rho$ together with an integration by parts:

$$\begin{aligned} K_\rho\nabla\,\mathrm{D}H(\rho)(\cdot) &= \int K(y,\cdot)\big[\nabla\log\rho(y) - \underbrace{\nabla\log\pi(y)}_{=-\nabla V(y)}\big]\rho(y)\,\mathrm{d}y \\ &= -\int K(y,\cdot)\nabla\log\pi(y)\,\rho(y)\,\mathrm{d}y + \int K(y,\cdot)\nabla\rho(y)\,\mathrm{d}y \\ &= K_\rho\nabla V - \int \nabla_y K(y,\cdot)\,\rho(y)\,\mathrm{d}y. \end{aligned}$$

Replacing ρ by the empirical measure of n particles $x_t^1,\dots,x_t^n$ gives the *Stein velocity field*[9]

$$\boxed{v_{\mathrm{Stein}}(\cdot) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\Big[-K(x_t^i,\cdot)\,\nabla V(x_t^i) + \nabla_1 K(x_t^i,\cdot)\Big].} \tag{17}$$

For the Gaussian kernel

$$K(x,y) = e^{-\frac{|x-y|^2}{2h^2}}, \qquad \nabla_1 K(x,y) = -\frac{1}{h^2}(x-y)\,e^{-\frac{|x-y|^2}{2h^2}}.$$

Two particle schemes for the same target $\pi \propto e^{-V}$:

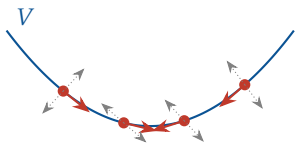
$$X_{k+1}^i \longleftarrow X_k^i + \tau_k\, v_{\mathrm{Stein}}(X_k^i) \qquad \text{(Stein / SVGD — deterministic, interacting)}$$

$$X_{t_{k+1}} \longleftarrow X_{t_k} - \tau\,\nabla V(X_{t_k}) + \sqrt{2\tau}\,Z_k \qquad \text{(Langevin / ULA — stochastic, independent)}$$

Same energy $H(\cdot\|\pi)$, different dissipation geometry $\mathbb{K}$.

Remark (Where the noise went). In (17) the first term is a kernel-smoothed drift towards low V (attraction), and the second term $\nabla_1 K$ pushes particles apart (repulsion): it is the deterministic surrogate for the Brownian noise of Langevin. Without it all particles would collapse onto $\arg\min V$.

[9] The board boxes the ascent direction $\frac{1}{n}\sum_i K(x_t^i,\cdot)\nabla V(x_t^i) - \frac{1}{n}\sum_i \nabla_1 K(x_t^i,\cdot)$ and then writes the update with a "+". The gradient *flow* moves along the negative of it, which is the sign fixed in (17) — and it is the sign that makes the second term *repulsive*.



Langevin / ULA
drift $-\nabla V$ + i.i.d. noise

Stein / SVGD
drift $-K_\rho \nabla V$ + kernel repulsion

Figure 17. Two discretizations of the same energy $H(\cdot\|\pi)$ under two dissipation geometries. Langevin particles are independent and explore by *random* kicks (dotted); Stein particles are coupled and explore by a *deterministic* kernel repulsion $\nabla_1 K$ (green). Both drift downhill in V.

6.4 Otto–Wasserstein flow of a kernel discrepancy (MMD flow)

Now keep the geometry $\mathbb{K}_{\mathrm{OT}}$ and change the *energy*.

Definition 6.5 (Kernel/maximum mean discrepancy)**.** For a symmetric *positive-definite* kernel K (e.g. a Gaussian),

$$D_K(\rho,\pi) := \big\langle \rho-\pi,\ K(\rho-\pi)\big\rangle_{L^2} = \iint_{\Omega\times\Omega} K(x,y)\,\mathrm{d}(\rho-\pi)(x)\,\mathrm{d}(\rho-\pi)(y).$$

This is the squared *kernel mean discrepancy* (KMD), also called maximum mean discrepancy (MMD) or the RKHS distance — it is $\|m_\rho - m_\pi\|^2_{\mathcal{H}}$ for the kernel mean embeddings $m_\rho = \int K(x,\cdot)\,\mathrm{d}\rho$.[10]

The gradient flow is $\dot\rho = -\mathbb{K}_{\mathrm{OT}}(\rho)\,\mathrm{D}D_K(\rho,\pi)$. Since D_K is *quadratic* in ρ, its first variation is[11]

$$\mathrm{D}D_K(\rho,\pi)(\cdot) = 2\int_\Omega K(x,\cdot)\big(\rho(x)-\pi(x)\big)\,\mathrm{d}x \ \approx\ 2\Big(\frac{1}{N}\sum_{i=1}^N K(x_i,\cdot) - \frac{1}{M}\sum_{j=1}^M K(y_j,\cdot)\Big) =: 2\,\widehat{v}_K(\cdot),$$

for samples $x_i\sim\rho$, $y_j\sim\pi$. Hence

$$\dot\rho = 2\,\nabla\cdot\big(\rho\,\nabla\widehat{v}_K\big), \qquad \nabla\widehat{v}_K(z) = \frac{1}{N}\sum_{i=1}^N \nabla_2 K(x_i,z) - \frac{1}{M}\sum_{j=1}^M \nabla_2 K(y_j,z),$$

and the corresponding particle system is $\dot x = -2\,\nabla\widehat{v}_K(x)$ — a purely sample-based scheme in which *both* ρ and π enter only through samples. (Contrast Langevin and SVGD, which need $\nabla\log\pi$ in closed form.)

6.5 The JKO / minimizing-movement scheme

Time discretization can also be done *implicitly*, in a way that never mentions the velocity field.

[10] Positive definiteness is what makes $D_K \ge 0$ and gives the mean-embedding reading; the board also writes the quadratic form $K(x,y) = (x-y)^\top A(x-y)$ next to K (with $A = A^\top \ge 0$), and that one must be read with a minus sign. Write $\bar m_\rho := \int x\,\mathrm{d}\rho(x) \in \mathbb{R}^d$ for the *first moment* — not the embedding $m_\rho \in \mathcal{H}$ above. Since $\sigma := \rho-\pi$ has total mass 0, the two pure terms of $x^\top Ax - 2x^\top Ay + y^\top Ay$ integrate away and $\iint (x-y)^\top A(x-y)\,\mathrm{d}\sigma\,\mathrm{d}\sigma = -2\,(\bar m_\rho - \bar m_\pi)^\top A(\bar m_\rho - \bar m_\pi) \le 0$ — so with the + sign the "discrepancy" is nonpositive and the flow below would *ascend*. With $K = -(x-y)^\top A(x-y)$ one gets $+2(\bar m_\rho - \bar m_\pi)^\top A(\bar m_\rho - \bar m_\pi) \ge 0$, a valid but degenerate discrepancy that sees only the means. Cf. the non-positive-definite kernel $-\|x-y\|$ in Figure 6.

[11] The board omits the factor 2 here *and* in the flow below, which is self-consistent; we keep it in both places. It only rescales time, so it can equally be absorbed into the step size τ.

JKO scheme (Jordan–Kinderlehrer–Otto).

$$\mu^{k+1} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\Omega)} \left\{ F(\mu) + \frac{1}{2\tau_k} W_2^2(\mu, \mu^k) \right\}.$$

This is the exact analogue of the Euclidean proximal step $x^{k+1} = \arg\min_x \{F(x) + \frac{1}{2\tau}\|x - x^k\|^2\}$, with $\|\cdot\|$ replaced by the Wasserstein distance W_2. As $\tau_k \to 0$ the interpolated iterates converge to the gradient flow of F in $(\mathcal{P}_2, W_2)$; for $F = H(\cdot\|\pi)$ one recovers Fokker–Planck. The scheme requires no differential structure at all — only a metric — which is what makes W_2 the central object of the next two sections.

7 Optimal Transport: Monge, Kantorovich, Brenier

Everything so far used W_2 as a black box. We now open it.

The motivating question is the one from Figure 18: given two distributions μ and ν — say, the distribution of images of cats () and the distribution of images of dogs () — what is $\mathrm{Dist}(\mu, \nu)$, and is there a *map* T carrying one to the other? The f-divergences of §5 answer the first question badly: they *saturate* as soon as the supports are disjoint — at $+\infty$ for KL and χ^2 (whose generators have $f'_\infty = +\infty$), at their finite maximum for total variation and Hellinger — so they cannot tell a near miss from a total one, and they never see the *geometry* of the underlying space. Optimal transport fixes both defects.

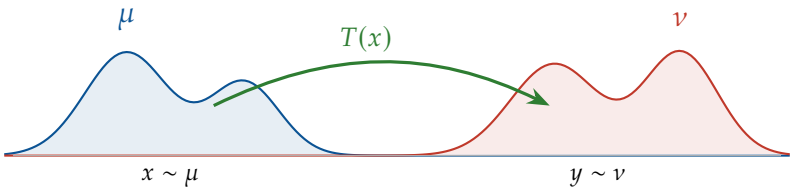


Figure 18. The transport question. Two multimodal distributions with (nearly) disjoint support: every f-divergence of §5 saturates, but a transport map T with $T_\#\mu = \nu$ and a transport *cost* still measure how far apart they are. In the lecture: μ = cats , ν = dogs , $\mathrm{Dist}(\mu, \nu)$ =?

7.1 Monge's problem (1781, "200+ years old")

Definition 7.1 (Pushforward and transport map)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ — e.g. $\mathcal{X} = \mathcal{Y} = \Omega \subseteq \mathbb{R}^d$. A map $T : \mathcal{X} \to \mathcal{Y}$ transports μ to ν if $T_\#\mu = \nu$, where the *pushforward* is

$$(T_\#\mu)(B) := \mu\big(T^{-1}(B)\big) \qquad \text{for Borel } B \subseteq \mathcal{Y}.$$

Equivalently $\int \varphi \, \mathrm{d}(T_\#\mu) = \int \varphi \circ T \, \mathrm{d}\mu$ for all test functions φ.

Definition 7.2 (Transport cost)**.** $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$ is a *transport cost*; the canonical choice is $c(x, y) = \frac{1}{2}\|x - y\|^2$. We work on

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

which guarantees finiteness of the quadratic cost.

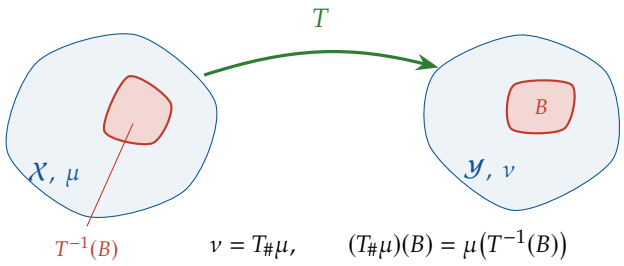


Figure 19. The pushforward is defined through a *preimage*, and that is the point students most often invert. To weigh a set B in the target, you do not push B forward — you collect the whole source region $T^{-1}(B)$ that lands in it and weigh *that* with μ. This is what makes $T_\#$ well defined even when T is neither injective nor surjective, and it is why $T_\#$ can merge mass (two source regions landing on one target set) but never split it — the obstruction of Example 7.3.

Monge's problem.

$$\inf_{T:\, T_\#\mu=\nu} \int_{\mathcal{X}} c\big(x, T(x)\big)\, \mathrm{d}\mu(x).$$

"Move the pile of earth μ into the excavation ν at least cost."

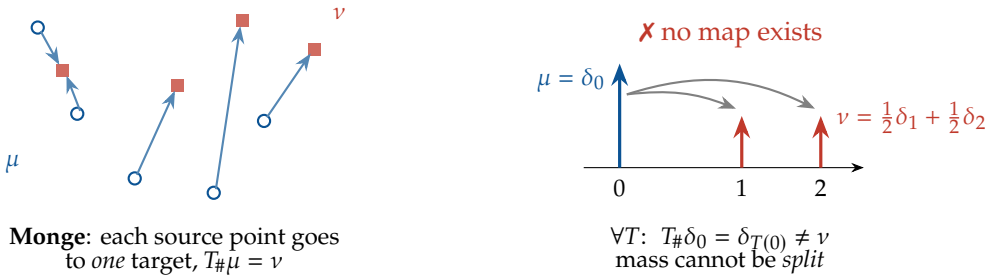


Figure 20. *Left*: a Monge map assigns each source (circles) to a single target (squares). *Right*: the obstruction. A Dirac cannot be split by any map, so Monge's problem may have *no feasible point at all* — the infimum is over the empty set. Kantorovich's relaxation repairs exactly this.

Example 7.3 (Monge's problem can be infeasible)**.** Let $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$. For *any* map T we have $T_\#\mu = \delta_{T(0)} \neq \nu$. The feasible set is empty.

7.2 Kantorovich's relaxation

The fix is the lift already met in §5.5: replace the nonlinear constraint "$T_\#\mu = \nu$" by a linear one on a *joint distribution*, which is allowed to split mass.

Definition 7.4 (Transport plan)**.** A *transport plan* is a probability measure (a joint distribution) $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ subject to the marginal constraints

$$\pi(A \times \mathcal{Y}) = \mu(A), \qquad \pi(\mathcal{X} \times B) = \nu(B), \qquad \forall A, B,$$

or, in the density form the board also writes and which is what the duality below dualizes,

$$\int_{\mathcal{Y}} \pi(x, y)\, \mathrm{d}y = \mu(x), \qquad \int_{\mathcal{X}} \pi(x, y)\, \mathrm{d}x = \nu(y), \qquad \text{for a.a. } x \in \mathcal{X},\ y \in \mathcal{Y}.$$

The feasible set is

$$\Pi(\mu, \nu) := \big\{\, \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) :\ \pi \text{ has marginals } \mu, \nu \,\big\}.$$

Kantorovich's problem.

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y).$$

Proposition 7.5 (The relaxation is genuine)**.**

(i) $\Pi(\mu,\nu)$ *is never empty: the product coupling* $\mu\otimes\nu\in\Pi(\mu,\nu)$.

(ii) *Every Monge map is a special plan: given* T *with* $T_\#\mu=\nu$, *set*

$$\bar\pi := (\mathrm{Id},T)_\#\mu, \qquad \textit{i.e.} \qquad \pi(x,y)=\mu(x)\,\delta_{T(x)}(y),$$

which lies in $\Pi(\mu,\nu)$ *and has the same cost. Hence* $\inf_{Kantorovich}\le\inf_{Monge}$.

(iii) *The problem is a* linear *program: linear objective, linear (affine) constraints, over a convex, weakly-* compact set. A minimizer exists whenever* c *is lower semicontinuous.*

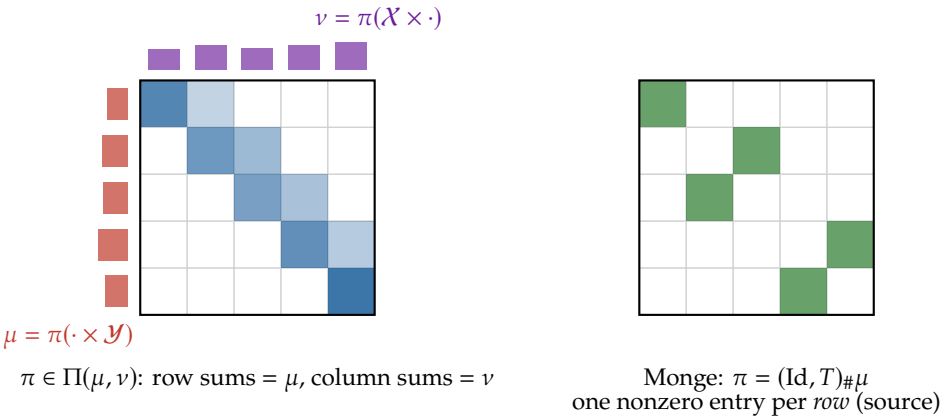


Figure 21. Kantorovich's plan as a coupling matrix. *Left*: a general plan spreads the mass of one source over several targets, subject to prescribed row and column marginals. *Right*: a Monge map is the degenerate case with exactly one nonzero entry per source — a vertex of the transport polytope. Because the objective is linear, an optimum is always attained at *some* vertex; that this vertex is a Monge map is a separate and much deeper fact, and it needs a hypothesis on the cost: for $c=\frac12|x-y|^2$ and μ absolutely continuous it is Brenier's theorem (Theorem 7.9). For a general cost no Monge map need exist at all.

7.3 Kantorovich duality

Write the constraints with dual functions $f\in C_b(\mathcal{X})$, $g\in C_b(\mathcal{Y})$ and use the pairing $\langle f,\,\mu\rangle=\int f\,\mathrm{d}\mu$. The Lagrangian is

$$\mathcal{L}(\pi;f,g):=\int c\,\mathrm{d}\pi+\Big\langle f,\;\mu-\int\pi(\cdot,y)\,\mathrm{d}y\Big\rangle+\Big\langle g,\;\nu-\int\pi(x,\cdot)\,\mathrm{d}x\Big\rangle.$$

Taking $\sup_{f,g}$ first reproduces the primal (an infeasible π costs $+\infty$). Exchanging sup and inf and regrouping,

$$\max_{\substack{f,g\\ \in C_b(\mathcal{X})\times C_b(\mathcal{Y})}}\ \min_{\pi\ge0}\ \underbrace{\int\big[c(x,y)-f(x)-g(y)\big]\,\mathrm{d}\pi(x,y)}_{=\,0\text{ or }-\infty\text{ (Lemma 7.6)}}+\big\langle f,\,\mu\big\rangle+\big\langle g,\,\nu\big\rangle.$$

Lemma 7.6 (The inner minimum forces the constraint). *If $c(x_0, y_0) - f(x_0) - g(y_0) < 0$ for some (x_0, y_0), then picking $\pi = t\,\delta_{(x_0,y_0)}$ and letting $t \to +\infty$ makes the first term $\to -\infty$. Hence at the optimum*

$$c - f \oplus g \;\geq\; 0, \qquad \text{where } (f \oplus g)(x, y) := f(x) + g(y),$$

and then the inner minimum equals 0 (attained at $\pi = 0$).

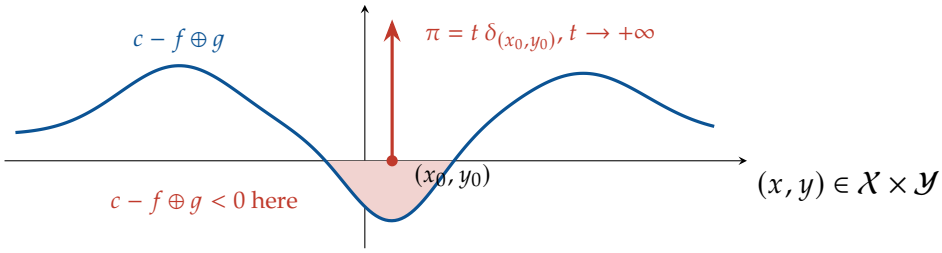


Figure 22. The whole duality argument in one picture. The dual variables are only constrained through the inner minimization, and that minimization is over the *cone* of nonnegative measures. So if $c - f \oplus g$ dips below zero anywhere (shaded), putting a Dirac of mass t at that point and letting $t \to +\infty$ gives $\int (c - f \oplus g)\,\mathrm{d}\pi = t\,\big[c - f \oplus g\big](x_0, y_0) \to -\infty$. A pair (f, g) therefore survives the outer maximization only if $f \oplus g \leq c$ *everywhere* — and then the inner minimum is 0, attained at $\pi = 0$.

Dual Kantorovich problem.

$$\max_{f,g \in C_b} \int_{\mathcal{X}} f\,\mathrm{d}\mu + \int_{\mathcal{Y}} g\,\mathrm{d}\nu \qquad \text{s.t.} \qquad f \oplus g \;\leq\; c.$$

At an optimum the constraint is active *where the mass sits* — so $\inf(c - f - g) = 0$, even though $f \oplus g < c$ typically holds off the support of π. Precisely:

Proposition 7.7 (Optimality / complementary slackness). *π and (f, g) are optimal iff $f \oplus g \leq c$ everywhere and*

$$c(x, y) = f(x) + g(y) \qquad \text{for } \pi\text{-a.e. } (x, y).$$

The set where equality holds is the contact set*; the optimal plan is concentrated on it.*

c-transforms

The constraint $f(x) \leq c(x, y) - g(y)$ must hold for all y, so the best admissible f given g is

$$g^c(x) := \inf_{y \in \mathcal{Y}} \left\{ c(x, y) - g(y) \right\},$$

the *c-transform* of g. Since $f \leq g^c$ pointwise implies $\int f\,\mathrm{d}\mu \leq \int g^c\,\mathrm{d}\mu$, replacing f by g^c never decreases the objective and preserves feasibility. Iterating on both sides, the dual collapses to a problem in *one* function.

Reduced dual.

$$\max_{f} \int_{\mathcal{X}} f\,\mathrm{d}\mu + \int_{\mathcal{Y}} f^c\,\mathrm{d}\nu, \qquad f^c(y) := \inf_{x} \left\{ c(x, y) - f(x) \right\},$$

the maximum being over c-concave f, i.e. those of the form $f = \gamma^c$ for some γ.[a]

[a] With the inf convention used here such functions are usually called *c-concave* (Santambrogio, Villani); the whiteboard writes "c-convex", which is the same class under the mirrored sup convention.

Feasibility of the reduced dual is automatic: by definition of the infimum,

$$f(x) + f^c(y) \;\le\; f(x) + \big[c(x,y) - f(x)\big] = c(x,y).$$

Remark (Positive vs. probability measures). The lemma above minimized over *all* nonnegative measures on $\mathcal{X} \times \mathcal{Y}$ — the marginal constraints having been dualized into f and g, so that only $\pi \ge 0$ survives. The board writes that set Π^+ and glosses it $\pi \in \mathcal{M}^+$; it is a *cone* ($\pi \in C \Rightarrow r\pi \in C$ for $r \in [0,\infty)$), and over a cone the inner minimum is 0 or $-\infty$. If instead the mass constraint is retained — minimizing over the *simplex* $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, so $\int \pi = 1$ — the inner minimum is $\lambda := \inf_{x,y}(c - f - g)$, a finite number, and the dual reads

$$\max_{f,g} \int f \,\mathrm{d}\mu + \int g \,\mathrm{d}\nu + \lambda \qquad \text{s.t.} \qquad c - f - g \ge \lambda.$$

Setting $\widetilde{f} := f + \lambda$ absorbs the multiplier and recovers the previous form, $\max_{\widetilde{f},g} \int \widetilde{f}\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu$ s.t. $c - \widetilde{f} - g \ge 0$. The two formulations agree — the mass constraint $\int \pi = 1$ only shifts the potentials by a constant. (On the board: adding the term $+\lambda(\int \pi - 1)$, $\lambda \ge 0$, to the Lagrangian.)

On the *primal* side the distinction collapses, and the board marks this with a large double bar: writing $\Pi^+(\mu,\nu) := \{\, \pi \in \mathcal{M}^+(\mathcal{X}\times\mathcal{Y}) : \pi \text{ has marginals } \mu,\nu \,\}$ for the marginal-constrained slice of the cone, one has, for $\mu,\nu \in \mathcal{P}$,

$$\min_{\pi \in \Pi^+(\mu,\nu)} \int c \,\mathrm{d}\pi \;=\; \min_{\pi \in \Pi(\mu,\nu)} \int c \,\mathrm{d}\pi,$$

because the marginal constraints already force $\int \pi = \int \mu = 1$: imposing $\mathbf{1}^\top \pi = 1$ on top of them is vacuous, so $\Pi^+(\mu,\nu)$ and $\Pi(\mu,\nu)$ are literally the same set. (Note that this slice is no longer a cone — scaling breaks the marginals; only the ambient $\mathcal{M}^+$ is.) The distinction matters only when μ,ν are unbalanced.

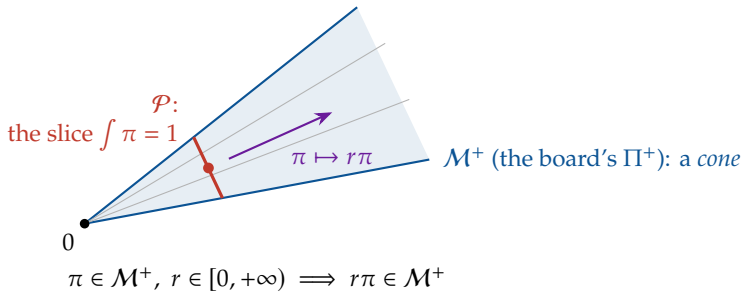


Figure 23. Cone versus simplex, *after the marginals have been dualized away*. What is left of the feasible set is the slice of total mass 1 (the warm segment: the probability measures $\mathcal{P}$); dropping the mass constraint too gives the ambient cone $\mathcal{M}^+$ — the board's Π^+ — that contains it. Minimizing a linear functional over the cone returns 0 or $-\infty$, which is what makes the constraint $f \oplus g \le c$ appear as a hard feasibility condition (Figure 22); over the slice one instead picks up the finite multiplier $\lambda = \inf(c - f - g)$, and absorbing λ into f recovers the same dual.

7.4 The quadratic cost and Brenier's theorem

Specialize to

$$c(x,y) = \tfrac{1}{2}\,|x-y|^2, \qquad \mu \ll \text{Lebesgue} \;\; (\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)).$$

Suppose we have an optimal dual pair (f, f^c). Define

$$\varphi(x) := \tfrac{1}{2}\,|x|^2 - f(x), \qquad \psi(y) := \tfrac{1}{2}\,|y|^2 - f^c(y).$$

Since $f \oplus f^c = c$ π-a.e., i.e. $f(x) + f^c(y) = \frac{1}{2}|x|^2 + \frac{1}{2}|y|^2 - \langle x, y\rangle$,

$$\varphi(x) + \psi(y) = \langle x, y\rangle \qquad \text{for } \pi\text{-a.e. } (x,y). \tag{18}$$

Lemma 7.8 (ψ is the Legendre conjugate of φ)**.** $\psi = \varphi^*$. *Indeed,*

$$\begin{aligned}\varphi^*(y) &= \max_x\big(\langle x, y\rangle - \varphi(x)\big) = \max_x\Big(\langle x, y\rangle - \tfrac{1}{2}|x|^2 + f(x)\Big)\\ &= \max_x\Big(-\tfrac{1}{2}|x-y|^2 + f(x)\Big) + \tfrac{1}{2}|y|^2 = -\min_x\Big(\underbrace{\tfrac{1}{2}|x-y|^2 - f(x)}_{\text{the } c\text{-transform}}\Big) + \tfrac{1}{2}|y|^2\\ &= \tfrac{1}{2}|y|^2 - f^c(y) \;=\; \psi(y).\end{aligned}$$

In particular φ is convex: the reduced dual may be taken over c-concave f, so $f = \gamma^c$ for some γ, whence $\varphi(x) = \frac{1}{2}|x|^2 - f(x) = \sup_y\big(\langle x, y\rangle - \frac{1}{2}|y|^2 + \gamma(y)\big)$ is a supremum of affine functions of x — hence convex and lower semicontinuous.

Now (18) says the Fenchel–Young inequality $\varphi(x) + \varphi^*(y) \geq \langle x, y\rangle$ holds with *equality* π-a.e. By Proposition 3.8 this is the same as saying the pair is in the subdifferential relation:

Fenchel–Young trichotomy . The following are equivalent for (x, y):

$$\varphi(x) + \varphi^*(y) = \langle x, y\rangle \iff y \in \partial\varphi(x)\ \ \big(\nabla\varphi(x) = y\big) \iff x \in \partial\varphi^*(y)\ \ \big(\nabla\varphi^*(y) = x\big).$$

The middle statement identifies the *transport map*: $y = \nabla\varphi(x)$.

Theorem 7.9 (Brenier, with McCann's refinement)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \mathcal{L}^d$ (McCann: it suffices that μ charges no $(d-1)$-rectifiable set), and $c(x,y) = \frac{1}{2}|x-y|^2$. Then there exists a convex, lower semicontinuous $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that*

$$\boxed{T = \nabla\varphi, \qquad T_\#\mu = \nu,}$$

and T is unique μ-a.e. Moreover $\pi = (\mathrm{Id}, T)_\#\mu$, i.e. $\pi(x,y) = \mu(x)\delta_{T(x)}(y)$, is the unique solution of the Kantorovich problem — so under this hypothesis on μ *the relaxation is* tight*: Monge = Kantorovich, and the optimal plan is induced by the map $\nabla\varphi$.*

Remark (Atomless is not enough). The regularity hypothesis on μ cannot be weakened to "atomless" in dimension $d \geq 2$. Take μ = the uniform measure on the segment $\{0\} \times [0,1] \subset \mathbb{R}^2$ (atomless, but singular, and carried by a 1-rectifiable set) and $\nu = \frac{1}{2}\,\mathrm{unif}(\{-1\}\times[0,1]) + \frac{1}{2}\,\mathrm{unif}(\{1\}\times[0,1])$. The unique optimal plan splits each point $(0,t)$ into $(\pm 1, t)$ with weight $\frac{1}{2}$ and is *not* induced by a map. For merely atomless μ only the equality of *values* inf(Monge) = min(Kantorovich) survives (Pratelli); the Monge infimum need not be attained.

Remark (Convention for W_2). We define $W_2^2(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2\,\mathrm{d}\pi$. With the cost $c = \frac{1}{2}|x-y|^2$ used above, the Kantorovich value is $\frac{1}{2}W_2^2$; the optimal plan and the Brenier map are the same either way.

7.5 Entropic regularization and the Schrödinger bridge

Example 7.10 (Entropic OT)**.** Add a relative entropy penalty against a reference coupling r (e.g. $r = \mu \otimes \nu$):

$$\min_{\pi\in\Pi} \int c\,\mathrm{d}\pi + \varepsilon\, H(\pi \,\|\, r), \qquad H(\pi\|r) = \int \pi\log\pi - \int \pi\log r - \int \pi + \int r,$$

subject to $\int \pi(x,y)\,\mathrm{d}x = \nu(y)$ and $\int \pi(x,y)\,\mathrm{d}y = \mu(x)$.

The regularized problem is strictly convex, hence has a unique solution of the factored form $\pi_\varepsilon = a(x)\, e^{-c(x,y)/\varepsilon}\, b(y)\, r$; alternating enforcement of the two marginals is Sinkhorn's algorithm. Letting $\varepsilon \downarrow 0$ recovers the unregularized optimal plan. In its dynamic formulation this is the *Schrödinger bridge* problem — the entropic analogue of the Benamou–Brenier dynamic formulation of W_2.

Remark. The lecture flagged the calculus of these objects (and the Schrödinger bridge) as the natural next step / homework. Note the closing identity of the section: W_2^2 *is* the value of a Kantorovich problem (Monge–Kantorovich). With the cost $c = \frac{1}{2}|x-y|^2$ used throughout §7 this reads $\int c\, \mathrm{d}\pi^* = \frac{1}{2} W_2^2(\mu,\nu)$, i.e. $W_2^2(\mu,\nu) = \int |x-y|^2\, \mathrm{d}\pi^*$ — which is the normalization appearing in the JKO scheme of §6.5.

8 The Wasserstein Space and Displacement Convexity

8.1 Geodesics in $(\mathcal{P}_2(\Omega), W_2)$

The metric space is

$$\Big(\mathcal{P}_2(\Omega),\ W_2\Big), \qquad \mathcal{P}_2(\Omega) = \left\{ \mu \in \mathcal{P}(\Omega) : \int |x|^2\, \mathrm{d}\mu < +\infty \right\},$$

which Week 1 already viewed as an (informal) Riemannian manifold with inverse metric $\mathbb{K}_{\mathrm{OT}}(\rho)$. Brenier's theorem lets us write its geodesics explicitly.

Definition 8.1 (McCann's displacement interpolation)**.** Let $\mu_0, \mu_1 \in \mathcal{P}_2(\Omega)$ with μ_0 absolutely continuous, and let $T = \nabla\varphi$ be the Brenier map from μ_0 to μ_1 (φ convex). The *displacement interpolant* is

$$\boxed{\mu_t := \big((1-t)\mathrm{Id} + tT\big)_{\#}\mu_0, \qquad 0 \le t \le 1.} \tag{19}$$

Writing $T_t(x) := (1-t)x + tT(x)$, each particle travels along the straight segment from x to $T(x)$ at constant speed.

Contrast this with the *linear* ("flat") interpolant $(1-t)\mu_0 + t\mu_1$, which does not move any mass at all: it merely fades one measure out while fading the other in.

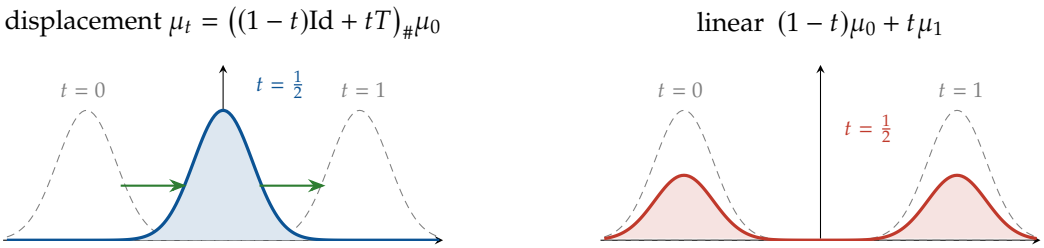


Figure 24. The two interpolations at $t = \frac{1}{2}$. *Left*: displacement interpolation *transports* the bump — this is the W_2-geodesic, and it is the reason entropy is convex along it. *Right*: linear interpolation only reweights, creating a bimodal mixture. Functionals behave completely differently along the two: $\mu \mapsto \int V\, \mathrm{d}\mu$ is affine on the right but inherits the convexity of V on the left.

Proposition 8.2 (Constant-speed geodesic)**.** *With μ_t as in* (19)*,*

$$W_2^2(\mu_t, \mu_0) = t^2\, W_2^2(\mu_1, \mu_0), \qquad \textit{and more generally}$$

$$W_2(\mu_t, \mu_s) = |t-s|\ W_2(\mu_0, \mu_1), \qquad s, t \in [0,1].$$

Hence $\{\mu_t\}_{t\in[0,1]}$ is a constant-speed geodesic *in $(\mathcal{P}_2(\Omega), W_2)$.*

Proof. The map $T_t = (1-t)\mathrm{Id} + tT = \nabla\left[(1-t)\frac{1}{2}|x|^2 + t\varphi\right]$ is the gradient of a convex function, so by Theorem 7.9 it is the *optimal* map from μ_0 to μ_t. Therefore

$$W_2^2(\mu_t,\mu_0) = \int_\Omega |T_t(x)-x|^2\,\mathrm{d}\mu_0(x) = \int_\Omega \left|(1-t)x + tT(x) - x\right|^2 \mathrm{d}\mu_0(x) = t^2\int_\Omega |T(x)-x|^2\,\mathrm{d}\mu_0(x),$$

which is $t^2 W_2^2(\mu_1,\mu_0)$. For the general statement, the coupling $(T_t, T_s)_\#\mu_0$ is a competitor for $W_2(\mu_t,\mu_s)$ and gives the upper bound $W_2(\mu_t,\mu_s) \le |t-s|\,W_2(\mu_0,\mu_1)$; the matching lower bound follows from the triangle inequality, since for $s < t$

$$W_2(\mu_0,\mu_1) \le W_2(\mu_0,\mu_s) + W_2(\mu_s,\mu_t) + W_2(\mu_t,\mu_1) = \big(s + (1-t)\big)W_2(\mu_0,\mu_1) + W_2(\mu_s,\mu_t),$$

i.e. $W_2(\mu_s,\mu_t) \ge (t-s)W_2(\mu_0,\mu_1)$. □

8.2 Displacement (λ-)convexity

Definition 8.3 (λ-displacement convexity)**.** A functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}\cup\{+\infty\}$ is *λ-displacement convex* if for every constant-speed geodesic (displacement interpolant) $\{\mu_t\}_{t\in[0,1]}$ from μ_0 to μ_1, and all $\mu_0, \mu_1 \in \mathcal{P}_2(\Omega)$,

$$\boxed{F(\mu_t) \;\le\; (1-t)F(\mu_0) + t\,F(\mu_1) - \frac{\lambda}{2}\,t(1-t)\,W_2^2(\mu_0,\mu_1).} \tag{20}$$

Synonyms: λ-convex along W_2-geodesics; geodesically λ-convex in the W_2 metric.

This is Definition 1.4 verbatim, with the straight line replaced by the geodesic and $\|u_0 - u_1\|$ by $W_2(\mu_0,\mu_1)$. Everything Week 1 proved from λ-convexity — EVI_λ, contraction, gradient dominance, exponential decay — transfers.

Remark (vs. flat convexity). *Flat* (or linear) convexity is the inequality $F\big((1-t)\mu_0 + t\mu_1\big) \le (1-t)F(\mu_0) + tF(\mu_1)$ along the linear interpolant. The two notions are genuinely different. For instance $\mathcal{V}(\mu) = \int V\,\mathrm{d}\mu$ is *affine* in the flat sense (so trivially 0-convex, never more), yet it is λ-displacement convex exactly when V is λ-convex (Proposition 8.6). Conversely the entropy is strictly convex in the flat sense but only 0-displacement convex.

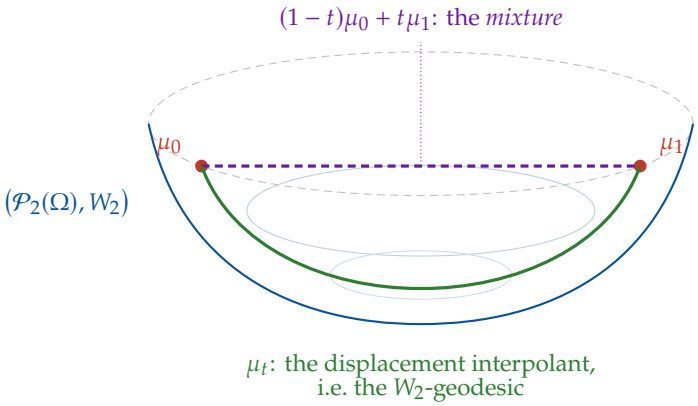


Figure 25. Why "convex along geodesics" is not "convex along mixtures". Both paths stay inside $\mathcal{P}_2(\Omega)$ — a mixture of probability measures is a probability measure — but only the green one is W_2-minimizing; the mixture is the straight chord of the *linear* structure, which the W_2 metric does not see as straight. The surface is a metaphor for that metric geometry, and it is the reason Definition 8.3 has real content: Figure 24 shows the two paths concretely as densities, and the entropy is convex along the green one but the potential energy is merely affine along the dashed one.

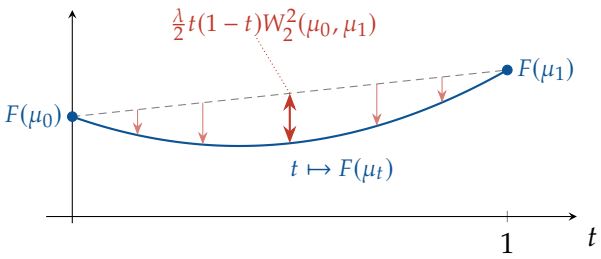


Figure 26. Displacement λ-convexity: along the W_2-*geodesic* $\{\mu_t\}$ the energy dips below the chord by at least the quadratic slack $\frac{\lambda}{2}t(1-t)W_2^2(\mu_0,\mu_1)$. The thin arrows trace that slack as a function of t: it vanishes at both endpoints and is maximal at $t=\frac{1}{2}$, and raising λ pushes the whole curve further below the chord. Compare Figure 2 — the picture is identical; only the notion of "straight line" changed.

Definition 8.4 (Generalized geodesics)**.** Fix a *base* μ_1 and let $T^{1\to 2}, T^{1\to 3}$ be the Brenier maps from μ_1 to μ_2, μ_3. The *generalized geodesic* from μ_2 to μ_3 with base μ_1 is

$$\mu_t^{2\to 3} := \big((1-t)T^{1\to 2} + t\,T^{1\to 3}\big)_{\#}\mu_1.$$

It coincides with the displacement interpolant when $\mu_1 = \mu_2$. Convexity along generalized geodesics (rather than just geodesics) is what makes the JKO functional $\mu \mapsto F(\mu) + \frac{1}{2\tau}W_2^2(\mu,\mu^k)$ convex, which is why $W_2^2(\cdot,\mu^k)$ enters the theory through this stronger notion.

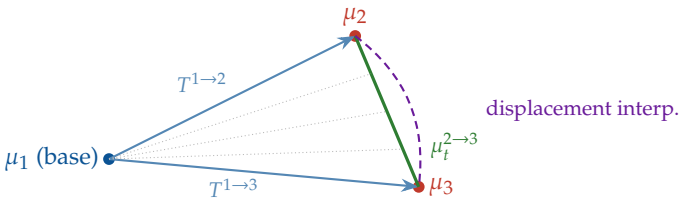


Figure 27. Generalized geodesics interpolate the *maps* $T^{1\to 2}, T^{1\to 3}$ emanating from a common base μ_1 (solid green), rather than following the intrinsic W_2-geodesic between μ_2 and μ_3 (dashed purple). They agree when the base is one of the endpoints.

8.3 The three canonical energies

Essentially every energy in this course is a sum of three terms:

$$F(\mu) = \underbrace{\int_\Omega V(x)\,\mathrm{d}\mu(x)}_{\mathcal{V}\text{: linear potential}} + \underbrace{\iint_{\Omega\times\Omega} K(x-y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)}_{\mathcal{W}\text{: interaction}} + \underbrace{\int_\Omega u\big(\mu(x)\big)\,\mathrm{d}x}_{\mathcal{U}\text{: internal}}, \tag{21}$$

with, typically, K convex and $u : s \mapsto s\log s$.

Example 8.5 (One canonical instance of each)**.** Each term in (21) has a canonical representative from the lecture:

- **Linear ($\mathcal{V}$) — inference.** Let μ be a distribution over a parameter $\theta \in \Theta$ and $\ell(\theta, x_i)$ a loss. Then
$$\min_\mu \sum_{i=1}^n \int_\Theta \ell(\theta,x_i)\,\mu(\theta)\,\mathrm{d}\theta$$
is linear in μ, i.e. of the form $\int V\,\mathrm{d}\mu$ with $V(\theta) = \sum_i \ell(\theta,x_i)$. Compare §5.5: adding entropy turns the Dirac minimizer into the Gibbs posterior.

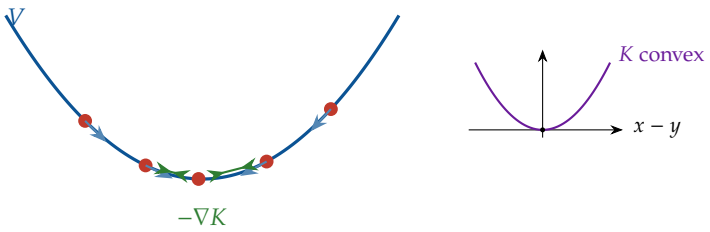


Figure 28. The energy (21) as a particle picture: $\mathcal{V}$ pulls each particle downhill in the external landscape V; $\mathcal{W}$ couples particles pairwise through $-\nabla K$ (the board draws the arrows as ∇K, pointing outward; the *force* is its negative) — *attractive* for the convex K drawn, since a convex symmetric K is minimized at 0 and therefore penalizes spread (for $K(z) = |z|^2$ one has $\mathcal{W}(\mu) = 2\,\mathrm{Var}\,\mu$), and repulsive for kernels decreasing in $|x - y|$; $\mathcal{U}$ is a pressure term penalizing concentration, invisible at the level of a single particle but decisive for the density.

- **Interaction ($\mathcal{W}$) — kernel discrepancy** (also KMD, MMD, or the RKHS distance). The object is an honest double integral,

$$\big\langle \mu - \nu,\ K_{\mathrm{id}}(\mu - \nu)\big\rangle_{L^2} = \iint_{\Omega\times\Omega} K(x-y)\,\mathrm{d}(\mu-\nu)(x)\,\mathrm{d}(\mu-\nu)(y),$$

  which, sampling $x_i \sim \mu$ ($i \le N$) and $y_j \sim \nu$ ($j \le M$), is estimated by

$$\approx \frac{1}{N^2}\sum_{i,j} K(x_i - x_j) + \frac{1}{M^2}\sum_{i,j} K(y_i - y_j) - \frac{2}{MN}\sum_{i,j} K(x_i - y_j),$$

  the V-statistic estimator of the MMD of §6.4: two self-interaction terms and one cross term.

- **Internal ($\mathcal{U}$) — Boltzmann entropy.** $u(r) = r\log r$ (the board's $r\log r - r + 1$ differs by an affine term, hence by a constant once $\int \rho = 1$ and $|\Omega| < \infty$; McCann's condition below is stated for the normalization $u(0) = 0$, which $r\log r$ satisfies and $r\log r - r + 1$ does not), or the power law $u(r) = \dfrac{r^m}{m-1}$ for $m \neq 1$. Relative entropy splits into a potential plus an internal term,

$$H(\rho\,\|\,\pi) = \int_\Omega \rho\log\frac{\rho}{\pi} = \int_\Omega \big(-\log\pi(x)\big)\rho(x)\,\mathrm{d}x + \int_\Omega \rho\log\rho,$$

  i.e. $\mathcal{V} + \mathcal{U}$ with $V = -\log\pi$ — which is exactly the decomposition used in §8.5.

8.4 When is a functional displacement λ-convex?

The potential energy

Proposition 8.6 (Potential energy)**.**

$$\mathcal{V}(\mu) = \int_\Omega V\,\mathrm{d}\mu \ \textit{ is displacement } \lambda\textit{-convex} \iff V : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\} \ \textit{ is } \lambda\textit{-convex}.$$

Proof. ($\Leftarrow$) Let $\mu_t = (T_t)_\#\mu_0$ with $T_t = (1-t)\mathrm{Id} + tT$. Pointwise λ-convexity of V gives, for every x,

$$V\big((1-t)x + tT(x)\big) \le (1-t)V(x) + t\,V(T(x)) - \frac{\lambda}{2}t(1-t)\,|x - T(x)|^2\,.$$

Integrate against μ_0 and use $T_\#\mu_0 = \mu_1$:

$$\mathcal{V}(\mu_t) = \int_\Omega V\big(T_t(x)\big)\,\mathrm{d}\mu_0 \le (1-t)\int V\,\mathrm{d}\mu_0 + t\int V\,\mathrm{d}\mu_1 - \frac{\lambda}{2}t(1-t)\underbrace{\int_\Omega |x - T(x)|^2\,\mathrm{d}\mu_0(x)}_{=\,W_2^2(\mu_0,\mu_1)}.$$

($\Rightarrow$) Take $\mu_0 = \delta_x$, $\mu_1 = \delta_y$. Here $\Pi(\delta_x, \delta_y) = \{\delta_{(x,y)}\}$ is a singleton, so the unique geodesic is $\mu_t = \delta_{(1-t)x+ty}$ (equivalently: transport by the constant map $T \equiv y$), and $W_2^2(\delta_x, \delta_y) = |x-y|^2$. Inequality (20) then becomes exactly (3) for V. □

The interaction energy

Proposition 8.7 (Interaction energy)**.** *If K is convex (and symmetric), then $\mathcal{W}(\mu) = \iint K(x-y)\,\mathrm{d}\mu\,\mathrm{d}\mu$ is 0-displacement convex.*

Proof. Along $\mu_t = (T_t)_\#\mu_0$ the displacement of a pair is $T_t(x) - T_t(y) = (1-t)(x-y) + t\big(T(x)-T(y)\big)$, a convex combination. Convexity of K and integration against $\mu_0 \otimes \mu_0$ give the claim. □

Remark. λ-convexity of K does *not* upgrade this to λ-displacement convexity: the slack produced is $\frac{\lambda}{2}t(1-t)\iint \big|(x-y)-(T(x)-T(y))\big|^2$, which is not $W_2^2(\mu_0,\mu_1)$. Hence the board's careful "(0−)convex".

The internal energy: McCann's condition

McCann's condition. Let $u(0) = 0$ and let d be the dimension. If

$$s \longmapsto s^d\, u\big(s^{-d}\big) \quad \text{is convex and nonincreasing on } (0,+\infty),$$

then $\mathcal{U}(\mu) = \int_\Omega u(\mu(x))\,\mathrm{d}x$ is W_2-geodesically (0-)convex.

Example 8.8 (Verification)**.**

- **Boltzmann**, $u(r) = r\log r$:

$$s^d u\big(s^{-d}\big) = s^d \cdot s^{-d}\log\big(s^{-d}\big) = -d\log s,$$

  which is convex and decreasing on $(0,\infty)$. ✓

- **Power (porous medium)**, $u(r) = \dfrac{r^m}{m-1}$, $m \neq 1$:

$$s^d u\big(s^{-d}\big) = \frac{s^d\, s^{-dm}}{m-1} = \frac{1}{m-1}\, s^{-d(m-1)}.$$

  For $m > 1$ this is a positive multiple of a decreasing convex power. ✓

Exercise 8.9. What is the exact relation between m and d for which the power law satisfies McCann's condition?[12]

8.5 Langevin as the W_2-gradient flow of the relative entropy

We can now close the loop opened in §6.

Example 8.10 (The Langevin equation is a gradient flow)**.** Let $\pi \propto e^{-V}$ and $\rho_t \in \mathcal{P}_2(\Omega)$. Then

$$H(\rho\,\|\,\pi) = \underbrace{\int_\Omega V\rho}_{\mathcal{V},\ \lambda\text{-convex iff } V \text{ is}} + \underbrace{\int_\Omega \rho\log\rho}_{\mathcal{U},\ 0\text{-convex by McCann}} + \text{const},$$

and the Langevin/Fokker–Planck equation is its W_2-gradient flow.

[12] Answer: $m \geq 1 - \frac{1}{d}$ (with $m \neq 1$). For $m < 1$ write $s^d u(s^{-d}) = -\frac{1}{1-m}s^\alpha$ with $\alpha = d(1-m) > 0$; this is always nonincreasing, and it is convex iff $s \mapsto s^\alpha$ is concave, i.e. iff $\alpha \leq 1$, i.e. iff $m \geq 1 - \frac{1}{d}$.

Convexity transfer. If V is λ-convex on $(\mathbb{R}^d, \|\cdot\|^2)$, then

$$H(\cdot \,\|\, \pi) \text{ is } \lambda\text{-convex along } W_2\text{-geodesics} \qquad (\lambda > 0).$$

Log-concavity of the target is exactly λ-displacement convexity of the energy — and the board records the threshold in a marginal note: already $\lambda \geq 0$ says that $\pi \propto e^{-V}$ is *log-concave*, while the strict $\lambda > 0$ printed in the box (strong log-concavity) is what the exponential rates below require.

For the Bayesian logistic posterior of Example 5.11 the potential (16) is 1-strongly convex (the Gaussian prior supplies $\frac{1}{2}\|\beta\|^2$, the log-likelihood terms are convex), so $\lambda = 1$ and everything below applies with rate 1.

8.6 First variations versus displacement perturbations

Before differentiating anything along a geodesic we must reconcile two notions of "perturbing a measure". The lecturer flagged this point on the board with an exclamation mark, and it is the commonest place to go wrong.

Definition 8.11 (First variation)**.** The variational derivative $\frac{\delta F}{\delta \mu}$ is defined through a *linear* perturbation: for admissible h (a signed measure with $\int_\Omega h = 0$, so that $\mu + \varepsilon h$ is again a probability measure),

$$\int_\Omega \frac{\delta F}{\delta \mu}(\mu)\, h \,\mathrm{d}x \;=\; \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} F\big(\mu + \varepsilon h\big), \qquad \mu_\varepsilon := \mu + \varepsilon h.$$

A displacement interpolant (19), by contrast, perturbs by a *pushforward*. Matching the two, the board writes

$$\mu_0 + t\, h \;=\; \mu_t \;=\; \big((1-t)\mathrm{Id} + tT\big)_\# \mu_0 \qquad \Longrightarrow \qquad \not{t}\, h = \not{t}\,(T - \mathrm{Id})_\# \mu_0,$$

cancels the common t, and records the moral in red: *a valid perturbation gives a displacement interpolant.*[13]

Remark (What the cancellation means, precisely). Read literally the last display cannot be an equality of measures: $(T - \mathrm{Id})_\#\mu_0$ is a probability measure (total mass 1), whereas an admissible h must have $\int_\Omega h = 0$; and $t \mapsto \mu_t$ is not affine in t. What the cancellation encodes is the identification of the *tangent vector*: the displacement field $v = T - \mathrm{Id}$ is the tangent vector to the geodesic at μ_0, and the corresponding perturbation of the measure is its image under the continuity equation,

$$\boxed{\; h \;=\; \dot\mu_0 \;=\; -\nabla\cdot\big(\mu_0\,(T - \mathrm{Id})\big), \;}$$

which does satisfy $\int_\Omega h = 0$. Along the gradient flow itself the tangent vector is $v = -\nabla \frac{\delta F}{\delta \mu}(\mu)$ and hence $h = \nabla\cdot\big(\mu\, \nabla \frac{\delta F}{\delta \mu}(\mu)\big)$ — the right-hand side of the gradient-flow equation, now read as a perturbation rather than as an evolution. It is exactly this substitution that produces the integration by parts in the proof of HWI below.

[13] The board writes the interpolant on this page as $\big((1-t)\mathrm{Id} - tT\big)_\#\mu_0$; the minus is a slip, inconsistent with the board's own next line and with (19).

8.7 HWI, log-Sobolev, and exponential convergence

Definition 8.12 (Relative Fisher information)**.**

$$\mathcal{I}(\mu\,\|\,\nu) := \int_\Omega \Big|\nabla \log \frac{\mu}{\nu}\Big|^2 \,\mathrm{d}\mu = \Big\|\nabla \log \tfrac{\mu}{\nu}\Big\|^2_{L^2_\mu}.$$

This is precisely the squared W_2-norm of the gradient of the entropy: $\mathcal{I}(\rho\|\pi) = \big\|\mathrm{grad}_{W_2} H(\rho\|\pi)\big\|^2_\rho$, since the metric velocity associated with the covector $\mathrm{D}H(\rho) = \log(\rho/\pi)$ is $\nabla\log(\rho/\pi)$.

Theorem 8.13 (HWI inequality, Otto–Villani)**.** *For $\mu,\nu \in \mathcal{P}_2(\Omega)$ and $\lambda \in \mathbb{R}$ (with $\nu \propto e^{-V}$, V λ-convex),*

$$\boxed{\;H(\mu\,\|\,\nu) \;\le\; W_2(\mu,\nu)\sqrt{\mathcal{I}(\mu\,\|\,\nu)} \;-\; \frac{\lambda}{2}\,W_2^2(\mu,\nu).\;}$$

The name records the three quantities involved: entropy H, Wasserstein distance W, and Fisher information I. It is a special case of a strictly more general statement, which is what the board actually derives: the measure-space version of the Euclidean first-order characterization (4).

The above-tangent inequality

Fix $\mu_0,\mu_1 \in \mathcal{P}_2(\Omega)$ with μ_0 absolutely continuous, let T be the Brenier map $\mu_0 \to \mu_1$, and let (μ_t) be the displacement interpolant (19). Each particle moves on a straight line,

$$X_t(x) = (1-t)\,x + t\,T(x), \qquad \dot X_t(x) = T(x) - x =: v(x), \qquad \dot\mu_t = -\nabla\cdot\big(\mu_t\, v\big).$$

Step 1 (divide by t and let $t \to 0$). Rearranging λ-displacement convexity (20),

$$\frac{F(\mu_t) - F(\mu_0)}{t} \;\le\; F(\mu_1) - F(\mu_0) - \frac{\lambda}{2}\,(1-t)\,W_2^2(\mu_0,\mu_1),$$

and letting $t \downarrow 0$ gives

$$F(\mu_1) \;\ge\; F(\mu_0) + \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t) + \frac{\lambda}{2}\,W_2^2(\mu_0,\mu_1). \tag{22}$$

Step 2 (chain rule and integration by parts). By Definition 8.11 and Remark 8.6,

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu_t) = \int_\Omega \frac{\delta F}{\delta \mu}(\mu_t)\,\dot\mu_t\,\mathrm{d}x = -\int_\Omega \frac{\delta F}{\delta \mu}(\mu_t)\,\nabla\cdot\big(\mu_t v\big)\,\mathrm{d}x = \int_\Omega \nabla\frac{\delta F}{\delta \mu}(\mu_t)\cdot\big(T(x)-x\big)\,\mathrm{d}\mu_t(x),$$

a pairing of the cotangent vector $\nabla\frac{\delta F}{\delta\mu} \in T^*_u M$ with the tangent vector $v = T - \mathrm{Id} \in T_u M$.

Step 3 (Cauchy–Schwarz in $L^2(\mu_0)$).

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\mu_t) \;\ge\; -\Big\|\nabla\tfrac{\delta F}{\delta\mu}(\mu_0)\Big\|_{L^2(\mu_0)} \underbrace{\Big(\int_\Omega |T(x)-x|^2\,\mathrm{d}\mu_0(x)\Big)^{1/2}}_{=\,W_2(\mu_0,\mu_1)}.$$

Above-tangent inequality in $(\mathcal{P}_2, W_2)$. For F λ-displacement convex and all $\mu_0,\mu_1 \in \mathcal{P}_2(\Omega)$,

$$\boxed{\;F(\mu_1) \;\ge\; F(\mu_0) - W_2(\mu_1,\mu_0)\Big\|\nabla\tfrac{\delta F}{\delta\mu}(\mu_0)\Big\|_{L^2(\mu_0)} + \frac{\lambda}{2}\,W_2^2(\mu_0,\mu_1).\;} \tag{23}$$

This is (4) with $\|\cdot\|$ replaced by W_2 and ∇F by $\nabla \frac{\delta F}{\delta \mu}$.

Proof of Theorem 8.13. Apply (23) to $F := H(\cdot \,\|\, \mu_1)$, which is λ-displacement convex whenever $\mu_1 \propto e^{-V}$ with V λ-convex. Then $F(\mu_1) = 0$, and since (14) carries its mass-correction terms, $\frac{\delta F}{\delta \mu}(\mu_0) = \log \frac{\mu_0}{\mu_1}$, so

$$\left\| \nabla \tfrac{\delta F}{\delta \mu}(\mu_0) \right\|^2_{L^2(\mu_0)} = \int_\Omega \left| \nabla \log \frac{\mu_0}{\mu_1} \right|^2 \mathrm{d}\mu_0 = \mathcal{I}(\mu_0 \,\|\, \mu_1).$$

Inequality (23) becomes $0 \geq H(\mu_0\|\mu_1) - W_2(\mu_0,\mu_1)\sqrt{\mathcal{I}(\mu_0\|\mu_1)} + \frac{\lambda}{2} W_2^2(\mu_0,\mu_1)$, which rearranges to HWI with $\mu = \mu_0$, $\nu = \mu_1$. □

Corollary 8.14 (Log-Sobolev inequality)**.** *If $\lambda > 0$, maximizing the right-hand side of Theorem 8.13 over $w = W_2(\mu,\nu) \geq 0$,*

$$\max_{w \geq 0} \left\{ w\sqrt{\mathcal{I}} - \frac{\lambda}{2} w^2 \right\} = \frac{\mathcal{I}}{2\lambda},$$

yields the logarithmic Sobolev inequality

$$\boxed{\tfrac{1}{2}\,\mathcal{I}(\mu \,\|\, \nu) \;\geq\; \lambda\, H(\mu \,\|\, \nu).}$$

Equivalently — and this is the way the board writes it — complete the square. Abbreviating $H = H(\mu\|\nu)$, $W = W_2(\mu,\nu)$, $I = \mathcal{I}(\mu\|\nu)$, add and subtract $\frac{1}{2\lambda} I$ in HWI:

$$H \;\leq\; \underbrace{W\sqrt{I} - \frac{\lambda}{2} W^2 - \frac{1}{2\lambda} I}_{= -\frac{\lambda}{2}\left(W - \frac{\sqrt{I}}{\lambda}\right)^2 \;\leq\; 0} + \frac{1}{2\lambda} I \;\leq\; \frac{1}{2\lambda}\, I \qquad \text{(LSI)}.$$

Definition 8.15 (Log-Sobolev inequality, stand-alone form)**.** π satisfies an *LSI* with constant $C > 0$ if[14]

$$C \underbrace{\int_\Omega \rho \log \frac{\rho}{\pi}}_{= H(\rho\|\pi) - \inf H \,\geq 0} \;\leq\; \underbrace{\int_\Omega \rho \left| \nabla \log \frac{\rho}{\pi} \right|^2}_{= \mathcal{I}(\rho\|\pi)} \qquad \forall\, \rho \in \mathcal{P}(\Omega),$$

i.e. $C\,H \leq I$. Corollary 8.14 says λ-log-concavity of π gives $C = 2\lambda$. Note that the right-hand side is the squared Wasserstein norm of the gradient of the entropy, $\mathcal{I}(\rho\|\pi) = \left\| \mathrm{grad}_{W_2} H(\rho\|\pi) \right\|^2_\rho = |\partial H(\cdot\|\pi)|^2(\rho)$, which is why LSI *is* gradient dominance.

Compare Theorem 1.6: this is *verbatim* the Polyak–Łojasiewicz / gradient dominance inequality $\frac{1}{2}\|\nabla F\|^2 \geq \lambda(F - \inf F)$, transported to $(\mathcal{P}_2, W_2)$ with $F = H(\cdot\|\nu)$, $\inf F = 0$, and $\|\nabla F\|^2 = \mathcal{I}$. LSI *is* W_2-gradient dominance.

Theorem 8.16 (Exponential convergence of Langevin)**.** *Let ρ_t solve the Fokker–Planck equation with target $\pi \propto e^{-V}$, V λ-convex, $\lambda > 0$. Then*

$$H(\rho_t \,\|\, \pi) \;\leq\; e^{-2\lambda t}\, H(\rho_0 \,\|\, \pi).$$

Proof. Along the gradient flow the energy-dissipation identity of Week 1 reads $\frac{\mathrm{d}}{\mathrm{d}t} H(\rho_t\|\pi) = -\left\| \mathrm{grad}_{W_2} H \right\|^2 = -\mathcal{I}(\rho_t\|\pi)$. By Corollary 8.14, $\mathcal{I} \geq 2\lambda H$, hence $\frac{\mathrm{d}}{\mathrm{d}t} H(\rho_t\|\pi) \leq -2\lambda\, H(\rho_t\|\pi)$, and Gronwall (Lemma 1.13) concludes. □

[14] The board writes $\exists\, C \gneq 0$, a $\geq$ with the equality bar struck out — the strictness matters, since $C = 0$ makes the inequality below vacuous. The same struck glyph reappears for the Poincaré constant of §8.8.

The plan, in one line.

$$\begin{aligned} H(\cdot\,\|\,\nu)\ \lambda\text{-displacement convex} &\implies \text{HWI} \\ &\implies \text{LSI } \big(= W_2\text{-gradient dominance}\big) \\ &\implies H(\mu_t\,\|\,\nu) \ \lesssim\ e^{-2\lambda t}. \end{aligned}$$

Which is Week 1, §1.3–1.4, read in the Wasserstein geometry: *λ-convexity $\Rightarrow$ PL $\Rightarrow$ exponential decay of the energy.* The only thing that changed between the two weeks is the dissipation geometry R in the gradient system (X, F, R).

8.8 The Poincaré inequality and χ^2 decay

One rung below LSI sits a weaker — and often easier to verify — functional inequality.

Definition 8.17 (Poincaré inequality, PI)**.** ρ satisfies a *Poincaré inequality* with constant $C > 0$ if

$$\underbrace{C \int_\Omega \rho \Big| f - \int_\Omega \rho f \Big|^2}_{= \operatorname{Var}_\rho(f)} \ \le\ \int_\Omega \rho\, |\nabla f|^2 \qquad \text{for all smooth } f.$$

LSI controls the *entropy* by the Fisher information; PI controls the *variance* by the Dirichlet energy. Where LSI governs the decay of H, PI governs the decay of χ^2.

Theorem 8.18 (χ^2 decay along Langevin)**.** *If π satisfies PI with $C > 0$, then along the Fokker–Planck flow with target π,*

$$\boxed{\chi^2(\mu_t\,\|\,\pi) \ \le\ e^{-2Ct}\, \chi^2(\mu_0\,\|\,\pi).}$$

Proof. Write $h_t := \mathrm{d}\mu_t/\mathrm{d}\pi$, which solves $\partial_t h = \mathcal{L} h$ with $\mathcal{L} = \Delta - \nabla V \cdot \nabla$, self-adjoint in $L^2(\pi)$ with $\int_\Omega f(\mathcal{L} f)\,\mathrm{d}\pi = -\int_\Omega |\nabla f|^2\,\mathrm{d}\pi$. Since $\chi^2(\mu_t\|\pi) = \frac{1}{2}\operatorname{Var}_\pi(h_t)$ and $\int_\Omega h_t\,\mathrm{d}\pi = 1$ is conserved,

$$\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\mu_t\|\pi) = \int_\Omega (h_t - 1)\,\mathcal{L} h_t\,\mathrm{d}\pi = -\int_\Omega |\nabla h_t|^2\,\mathrm{d}\pi \ \le\ -C \operatorname{Var}_\pi(h_t) = -2C\,\chi^2(\mu_t\|\pi),$$

and Gronwall (Lemma 1.13) concludes. □

Proposition 8.19 (LSI $\Rightarrow$ PI)**.** *If π satisfies the LSI of Corollary 8.14 with modulus λ, it satisfies PI with $C = \lambda$.*

Proof (linearize, as on the board). Test the LSI on the perturbation the board writes,

$$\rho_\varepsilon := \big[1 + \varepsilon\big(f - \textstyle\int_\Omega \pi f\big)\big]\pi,$$

which is a probability density for small ε. Put $g := f - \int_\Omega \pi f$, so $\int_\Omega \pi g = 0$. Expanding to second order,

$$H(\rho_\varepsilon\|\pi) = \frac{\varepsilon^2}{2}\operatorname{Var}_\pi(g) + O(\varepsilon^3), \qquad \mathcal{I}(\rho_\varepsilon\|\pi) = \varepsilon^2 \int_\Omega \pi\,|\nabla g|^2 + O(\varepsilon^3),$$

and substituting into $\frac{1}{2}\mathcal{I} \ge \lambda H$ and dividing by $\varepsilon^2/2$ gives $\int_\Omega \pi\,|\nabla f|^2 \ge \lambda \operatorname{Var}_\pi(f)$. □

So the hierarchy of Week 2 extends one step further,

$$\lambda\text{-displacement convexity} \implies \text{HWI} \implies \text{LSI} \implies \text{PI},$$

mirroring the Euclidean chain *strong convexity $\Rightarrow$ PL $\Rightarrow$ quadratic growth.*

8.9 Perturbing the target: Holley–Stroock

Log-concavity is a strong hypothesis, and the targets of §6 are rarely log-concave. What survives if we tilt a good target by a bounded — but otherwise arbitrary, possibly wildly multimodal — factor?

It is convenient first to write the LSI in f-divergence generator form. With the KL generator of §5,

$$\psi_{\mathrm{KL}}(s) = s\log s - s + 1, \qquad \psi'_{\mathrm{KL}}(s) = \log s,$$

and $r := \rho/\pi$, "π satisfies an LSI with constant λ_{LSI}" reads

$$\lambda_{\mathrm{LSI}} \int_\Omega \pi\, \psi_{\mathrm{KL}}(r) \;\le\; \int_\Omega \pi\, r\, \left|\nabla \psi'_{\mathrm{KL}}(r)\right|^2 \qquad \forall\, \rho \in \mathcal{P}(\Omega), \tag{24}$$

because the left side is $H(\rho\|\pi)$ and the right side is $\int_\Omega \rho \left|\nabla \log \frac{\rho}{\pi}\right|^2 = \mathcal{I}(\rho\|\pi)$.[15]

Theorem 8.20 (Holley–Stroock perturbation lemma)**.** *Let π satisfy an LSI with constant λ_{LSI}, and let $\tilde\pi = e^{-\phi}\pi$ with ϕ bounded,*

$$m := \inf_\Omega \phi \;\le\; \sup_\Omega \phi =: M \qquad\Longrightarrow\qquad e^{-M} \le e^{-\phi} \le e^{-m}.$$

Then $\tilde\pi$ satisfies an LSI with constant

$$\boxed{\hat\lambda \;=\; e^{-(M-m)}\,\lambda_{\mathrm{LSI}}.}$$

Proof (the board's chain). Lower-estimate the Fisher-information side for $\tilde\pi$, writing $\psi := \psi_{\mathrm{KL}}$:

$$\int_\Omega \tilde\pi\, r\, \left|\nabla\psi'(r)\right|^2 = \int_\Omega e^{-\phi}\pi\, r\, \left|\nabla\psi'(r)\right|^2 \;\ge\; e^{-M} \underbrace{\int_\Omega \pi\, r\, \left|\nabla\psi'(r)\right|^2}_{(*)}.$$

Since π satisfies (24), and then since $e^{-\phi} \le e^{-m}$, i.e. $\pi \ge e^m\tilde\pi$,

$$(*) \;\ge\; \lambda_{\mathrm{LSI}} \int_\Omega \pi\, \psi(r) \qquad\Longrightarrow\qquad e^{-M}(*) \;\ge\; e^{-(M-m)}\lambda_{\mathrm{LSI}} \int_\Omega \tilde\pi\, \psi(r).$$

Chaining the two displays gives the claim. □

Remark. The board's chain is the clean skeleton. A fully rigorous statement also tracks the normalizing constant of $\tilde\pi$ and replaces $\psi_{\mathrm{KL}}(\rho/\pi)$ by $\psi_{\mathrm{KL}}(\rho/\tilde\pi)$ on the right-hand side, which costs at most another factor of the same shape. The point is untouched: *the LSI constant degrades exponentially in the oscillation $M - m$ of the perturbation, and not at all in its shape.* Combined with Theorem 8.18 and §8.5, this is what gives exponential convergence of Langevin for a large class of genuinely multimodal targets: take $\pi \propto e^{-V}$ with V λ-strongly convex as the reference, and tilt it by any bounded ϕ.

9 Metric Gradient Flows in $(\mathcal{P}_2(\Omega), W_2)$

Until now the Wasserstein gradient flow was defined through the Onsager operator $\mathbb{K}_{\mathrm{OT}}$ — that is, through a differential structure. We close by revisiting it *metrically*. Everything in this section uses only the distance W_2, which is what lets the theory survive on spaces carrying no linear or manifold structure at all.

9.1 Minimizing movements

[15] The board suppresses the ∇ on this page; it is restored here.

Minimizing movement scheme (MMS). On a metric space $(\mathcal{X}, d)$, with step $\tau > 0$,

$$u_\tau^{k+1} \in \operatorname*{arg\,min}_{u \in \mathcal{X}} \left\{ F(u) + \frac{1}{2\tau}\, d^2\big(u, u_\tau^k\big) \right\}.$$

Choosing $(\mathcal{X}, d) = \big(\mathcal{P}_2(\Omega), W_2\big)$ gives the **JKO** scheme of §6.5.

The sequence $\{u_\tau^k\}_{k=0}^T$ is itself regarded as a *discrete gradient flow*. In $\mathbb{R}^d$ with $d = \|\cdot\|$ it is implicit Euler for $\dot{x} = -\nabla F(x)$; no derivative of F and no tangent space is used anywhere in the definition.

9.2 The derivative of the Wasserstein distance

To pass from the scheme back to a PDE we must differentiate $\rho \mapsto W_2^2(\rho, \pi)$.

Proposition 9.1 (Derivative of W_2^2). *Let $T^{\rho\to\pi} = \nabla\varphi$ be the Brenier map from ρ to π, φ convex. Then*

$$\boxed{\tfrac{1}{2}\,\nabla_W W_2^2(\rho,\pi)(x) \;=\; x - T^{\rho\to\pi}(x),}$$

equivalently, in the flat L^2 sense, the first variation is Kantorovich's dual potential

$$\tfrac{1}{2}\,\mathrm{D}^{L^2} W_2^2(\rho,\pi)(x) \;=\; \tfrac{1}{2}\,|x|^2 - \varphi(x),$$

whose gradient is $x - \nabla\varphi(x) = x - T^{\rho\to\pi}(x)$; cf. (18).

Remark (On the orientation). The board writes both identities with the opposite sign, $\frac{1}{2}\nabla_W W_2^2 = T^{\rho\to\pi}(x) - x$ and $\frac{1}{2}\mathrm{D}^{L^2} W_2^2 = \varphi(x) - \frac{1}{2}|x|^2$. The orientation printed above is the one consistent with the board's own conclusion below, and it survives the simplest sanity check: for $\pi = \delta_0$ we have $W_2^2(\rho, \delta_0) = \int_\Omega |x|^2 \,\mathrm{d}\rho$, whose L^2 first variation is $|x|^2$, while $T \equiv 0$ gives $x - T(x) = x$. Mnemonic: the gradient of a squared distance points *away* from the target, so that $-$ grad moves ρ toward π.

9.3 From the JKO optimality condition to the PDE

Differentiate the JKO objective in u and set the result to zero. By Proposition 9.1 the optimality condition for $u = u_\tau^{k+1}$ is

$$0 \;=\; \nabla \frac{\delta F}{\delta \mu}(u)(x) \;+\; \frac{x - T^{\,u\to u_\tau^k}(x)}{\tau}. \tag{25}$$

The second term is the discrete velocity: mass sitting at $T^{\,u\to u_\tau^k}(x)$ at time t_k has arrived at x at time t_{k+1}, so $v(x) \approx \big(x - T^{\,u\to u_\tau^k}(x)\big)/\tau$. Thus (25) says exactly $v = -\nabla\frac{\delta F}{\delta\mu}(u)$, and feeding this into the continuity equation $\dot{u} = -\nabla\cdot(u\,v)$ gives, as $\tau \to 0$,

$$\dot{u} \;=\; \nabla\cdot\Big(u\,\nabla \frac{\delta F}{\delta \mu}(u)\Big) \qquad (W_2\text{-gradient-flow equation}).$$

Plugging in the free energy of §8.5, $F : \rho \mapsto \int_\Omega \rho V + \int_\Omega \rho\log\rho$, for which $\frac{\delta F}{\delta\mu} = V + \log\rho + 1$, returns

$$\dot{u} = \nabla\cdot\big(u\,\nabla V\big) + \Delta u \qquad (\text{FPE}),$$

so the implicit scheme and the explicit PDE describe the same flow: *JKO is Fokker–Planck, implicitly discretized.*

9.4 Curves of maximal slope and the metric EDB

The metric formulation of "being a gradient flow" mentions neither ∇F nor a velocity field.

Definition 9.2 (Metric derivative; Benamou–Brenier)**.** For a curve $t \mapsto u_t$ in $\mathcal{P}_2(\Omega)$,

$$|u'|_{W_2}^2 \;=\; \inf_{v} \left\{ \int_\Omega |v|^2 \,\mathrm{d}u \;:\; \partial_t u = -\nabla\cdot(u\,v) \right\}$$

— the "Wasserstein norm", an H^{-1}-type norm: the cheapest velocity field producing the observed change of u. The *metric slope* of F is $|\partial F|(u) = \left\|\nabla \frac{\delta F}{\delta u}(u)\right\|_{L^2(u)}$.

Curve of maximal slope. u is a curve of maximal slope for F if for all $s \le t$

$$F(u_t) - F(u_s) \;=\; -\tfrac12 \int_s^t |u'|^2(r)\,\mathrm{d}r \;-\; \tfrac12 \int_s^t |\partial F|^2(u_r)\,\mathrm{d}r.$$

This is verbatim the Energy Dissipation Balance of §3.1: the two halves are the kinetic term $\|\dot u\|_W^2$ and the potential term $\left\|\nabla \frac{\delta F}{\delta u}(u)\right\|_{L^2(u)}^2$, with the Hilbert norms replaced by their metric surrogates. Along the Wasserstein gradient flow Fenchel–Young is saturated, the two halves coincide, and the balance collapses to the single-term form

$$\text{(EDB)} \qquad F(u_t) - F(u_s) \;=\; -\int_s^t \left\|\nabla \frac{\delta F}{\delta u}(u_r)\right\|_{L^2(u_r)}^2 \mathrm{d}r.$$

The refrain of the whole course, one last time: *solution of the gradient flow* $\iff$ *EDB* — now in a space with no linear structure whatsoever.

9.5 EVI_λ in the Wasserstein space

Differential EVI_λ**.** For all $w \in \mathcal{P}_2(\Omega)$,

$$\tfrac12 \frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\rho_t, w) \;\le\; F(w) - F(\rho_t) - \frac{\lambda}{2} W_2^2(\rho_t, w).$$

Compare the Hilbert statement inside the proof in §3.2, $\frac12 \frac{\mathrm{d}}{\mathrm{d}t}\|u(t)-w\|_{\mathcal{H}}^2 \le F(w) - F(u(t)) - \frac{\lambda}{2}\|u(t)-w\|_{\mathcal{H}}^2$: only the metric has changed. Integrating by Gronwall (Lemma 1.13) reproduces the integrated EVI_λ with the same $M_\lambda(t)$ of (9). Applying the differential form in *each* argument for two solutions ρ_t, σ_t and adding — the two energy terms $F(\sigma_t) - F(\rho_t)$ and $F(\rho_t) - F(\sigma_t)$ cancel — gives $\frac12 \frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\rho_t, \sigma_t) \le -\lambda W_2^2(\rho_t, \sigma_t)$, i.e. the contraction $W_2(\rho_t, \sigma_t) \le e^{-\lambda t} W_2(\rho_0, \sigma_0)$ — exactly as in the Euclidean case of §1.4.

10 Gaussian Gradient Flows and Variational Inference

The flows above live on infinite-dimensional spaces. The practical move — the one that turns a PDE into an ODE one can actually run — is to *restrict* the geometry to a finite-dimensional submanifold of $\mathcal{P}(\Omega)$. Take the Gaussian family,

$$\big(\mathcal{P}_2(\Omega), W_2\big) \;\rightsquigarrow\; \big(\mathcal{N}^d(\Omega),\; W_2 \,/\, \mathrm{FR}\big), \qquad \Theta \ni \theta = \{m, \Sigma\},$$

with $\mathcal{N}^d(\Omega) \subset \mathcal{P}_2(\Omega)$ the nondegenerate Gaussians and $\Theta = \mathbb{R}^d \times \mathbb{S}^d_{++}$ the parameter space.

10.1 KL variational inference

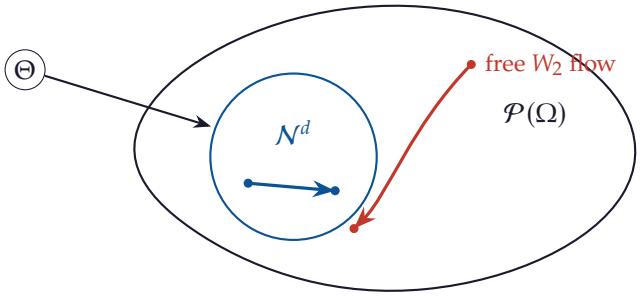


Figure 29. Restricting a gradient flow to a parametric family. The free flow (red) moves through all of $\mathcal{P}(\Omega)$; the restricted flow (blue) is confined to the Gaussian submanifold $\mathcal{N}^d$, parametrized by $\theta = (m, \Sigma) \in \Theta$. Restricting *different* geometries — Fisher–Rao or Bures–Wasserstein — yields different ODEs on Θ for one and the same energy.

$$\min_{(m,\Sigma)\in\Theta} \underbrace{H\big(\mathcal{N}(m,\Sigma)\,\big\|\,\pi\big)}_{=:\,F(m,\Sigma)}, \qquad \pi \propto e^{-V}.$$

This is the *reverse*, mode-seeking KL of §5.4; the alternative $H(\pi \,\|\, \mathcal{N}(m,\Sigma))$ is the forward, mass-covering choice (Figure 14). The target π is typically multimodal and far from Gaussian, which is the point: we are projecting it onto $\mathcal{N}^d$.

Proposition 10.1 (Gradients in (m,Σ))**.** *With $\rho = \mathcal{N}(m,\Sigma)$,*

$$\nabla_m F(m,\Sigma) = \mathbb{E}_\rho\big[\nabla V(x)\big], \qquad \nabla_\Sigma F(m,\Sigma) = \tfrac{1}{2}\Big(\underbrace{\mathbb{E}_\rho\big[\nabla^2 V(x)\big]}_{=:\,\Gamma^{-1}} - \Sigma^{-1}\Big).$$

Plain parameter-space gradient descent is then

$$m^{k+1} \leftarrow m^k - \tau\,\nabla_m F(m^k,\Sigma^k), \qquad \Sigma^{k+1} \leftarrow \Sigma^k - \tau\,\nabla_\Sigma F(m^k,\Sigma^k).$$

But Θ is curved, and — exactly as in the Bernoulli example of §4.5 — the right move is to precondition by an Onsager operator obtained by restricting a measure-space geometry to $\mathcal{N}^d$.

10.2 Reduced Onsager operators and their ODEs

Two restricted gradient structures on $\mathcal{N}^d$.

$$\mathbb{K}^{\mathrm{red}}_{\mathrm{FR}}(m,\Sigma) \simeq \begin{pmatrix} \Sigma\,\square & 0 \\ 0 & 2\,\Sigma\,\square\,\Sigma \end{pmatrix} \quad \text{i.e. } (a,A) \mapsto \big(\Sigma a,\ 2\Sigma A\Sigma\big),$$

$$\mathbb{K}^{\mathrm{red}}_{\mathrm{BW}}(m,\Sigma) \simeq \begin{pmatrix} \square & 0 \\ 0 & 2\big(\square\,\Sigma + \Sigma\,\square\big) \end{pmatrix} \quad \text{i.e. } (a,A) \mapsto \big(a,\ 2(A\Sigma + \Sigma A)\big),$$

obtained by restricting, respectively, the Fisher–Rao / spherical-Hellinger geometry of Definition 4.5 and the Otto–Wasserstein geometry to $\mathcal{N}^d(\Omega)$. The box $\square$ marks the slot into which the covector is inserted.

The gradient system $\big(\mathcal{N}^d, F, \mathbb{K}^{\mathrm{red}}_{\mathrm{FR}}\big)$ generates

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m \\ \Sigma \end{pmatrix} = -\,\mathbb{K}^{\mathrm{red}}_{\mathrm{FR}}(m,\Sigma)\,\nabla_{(m,\Sigma)} F(m,\Sigma) = -\begin{pmatrix} \Sigma\,\nabla_m F(m,\Sigma) \\ 2\,\Sigma\,\nabla_\Sigma F(m,\Sigma)\,\Sigma \end{pmatrix},$$

which, with Proposition 10.1, is the *KL–Fisher–Rao gradient-flow ODE*

$$\dot m = -\Sigma\,\mathbb{E}_\rho\big[\nabla V(x)\big], \qquad \dot\Sigma = \Sigma - \Sigma\,\Gamma^{-1}\Sigma, \qquad \Gamma^{-1} = \mathbb{E}_\rho\big[\nabla^2 V(x)\big].$$

Substituting $\mathbb{K}^{\mathrm{red}}_{\mathrm{BW}}$ instead, and using $-2\big(\nabla_\Sigma F\,\Sigma + \Sigma\,\nabla_\Sigma F\big) = -\big[(\Gamma^{-1}-\Sigma^{-1})\Sigma + \Sigma(\Gamma^{-1}-\Sigma^{-1})\big]$, gives the *Bures–Wasserstein* flow

$$\dot m = -\mathbb{E}_\rho\big[\nabla V(x)\big], \qquad \dot\Sigma = \big(\Sigma^{-1}-\Gamma^{-1}\big)\Sigma + \Sigma\big(\Sigma^{-1}-\Gamma^{-1}\big) \;=\; 2I - \Gamma^{-1}\Sigma - \Sigma\,\Gamma^{-1}.$$

Remark (Where they agree, and a note on the signs). Both flows have the same equilibrium: $\dot m = 0$ forces $\mathbb{E}_\rho[\nabla V] = 0$ and $\dot\Sigma = 0$ forces $\Sigma = \Gamma = \big(\mathbb{E}_\rho[\nabla^2 V]\big)^{-1}$ — the Gaussian whose mean sits at a critical point of the Σ-smoothed potential $V * \varphi_\Sigma$ (a critical point of V itself only when ∇V is affine) and whose covariance is the inverse averaged Hessian, i.e. a Laplace-type approximation of π. They differ in how they get there: Fisher–Rao preconditions the mean by Σ (large uncertainty, large steps), Bures–Wasserstein does not.

The board writes the two covariance equations as $\dot\Sigma = \Sigma\Gamma^{-1}\Sigma - \Sigma$ and $\dot\Sigma = 2(\Gamma^{-1}-\Sigma^{-1})\Sigma + \Sigma(\Gamma^{-1}-\Sigma^{-1})$, i.e. with the overall minus sign of $-\mathbb{K}\nabla F$ dropped — and, in the Bures–Wasserstein line, with a stray factor 2 on the first term only, which would make the right-hand side non-symmetric. With those signs the equilibrium $\Sigma = \Gamma$ is repelling rather than attracting: in one dimension $\dot\sigma = \sigma^2/\gamma - \sigma$ has $\partial_\sigma\dot\sigma = +1 > 0$ at $\sigma = \gamma$, whereas the descending form $\dot\sigma = \sigma - \sigma^2/\gamma$ gives $-1 < 0$. The orientation printed above is the descending one.

Both are ODEs on the finite-dimensional Θ, so the entire Week-1 apparatus — λ-convexity, EDB, EVI_λ, gradient dominance, Gronwall — applies to them verbatim. Only R changed.

End of notes.

Scribed by Lunji Zhu with the help of Claude AI · to be authenticated by Prof. Jia-Jie Zhu · not guaranteed correct · for internal circulation only