Neural networks, mean field limit and optimization

François-Xavier Vialard

LIGM, Université Gustave Eiffel

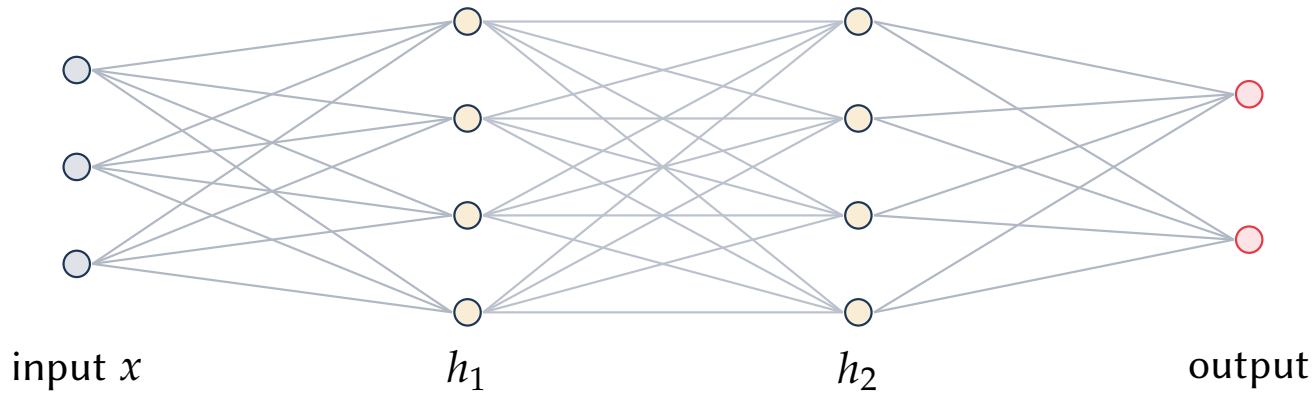
Important architectures

1. **Important architectures**
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Architecture I: multilayer perceptron (MLP)

An MLP alternates affine maps and coordinatewise nonlinearities:

$$h_0 = x, \qquad h_{\ell+1} = \sigma(W_\ell h_\ell + b_\ell), \qquad f_\theta(x) = W_L h_L + b_L.$$



Design principle: dense composition

Every unit in one layer communicates with every unit in the next. Depth builds hierarchical features, but information and gradients must cross every intermediate transformation.

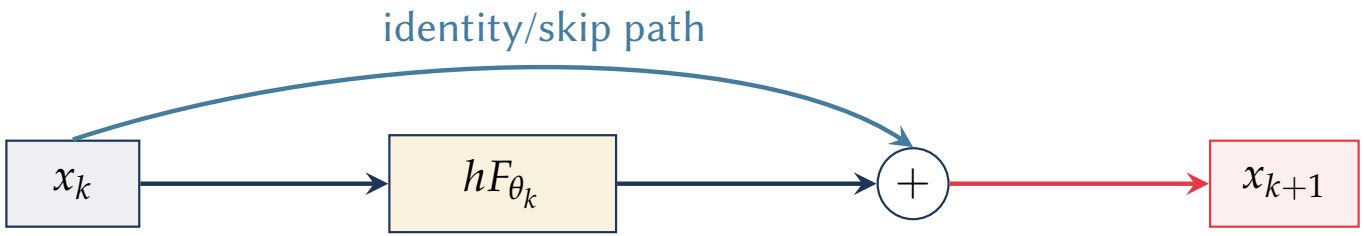
Typical use

Generic vector-valued data, tabular prediction, and learned heads on top of structured feature extractors.

Architecture II: residual network (ResNet)

A residual block learns an increment rather than a complete replacement:

$$x_{k+1} = x_k + hF_{\theta_k}(x_k).$$



Design principle: perturb the identity

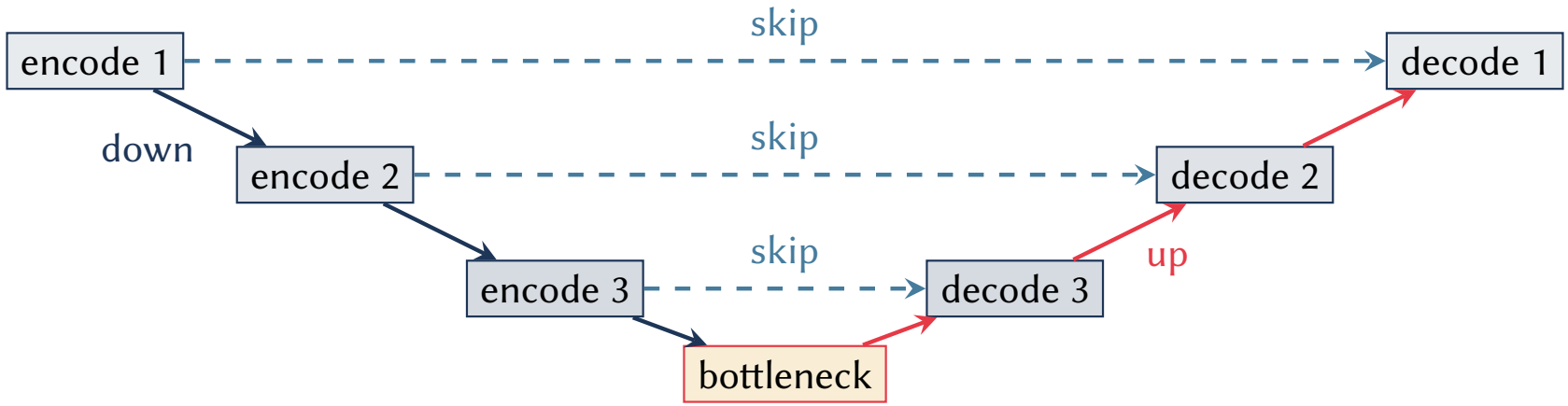
Setting $F_{\theta_k} = 0$ preserves the current representation. Deepening the network adds corrections without forcing new layers to reconstruct an identity map.

Optimization consequence

Forward and backward Jacobians contain an identity term $I + hDF_{\theta_k}$, improving signal propagation. As $h \to 0$, depth becomes the time discretization of a controlled differential equation.

Architecture III: U-Net

U-Net combines a contracting encoder with an expanding decoder and connects features at equal spatial resolutions.



Design principle: global context plus precise localization

Downsampling enlarges the receptive field and extracts semantic context; upsampling restores resolution. Skip connections concatenate fine encoder features with coarse decoder features.

Typical use

Dense prediction: biomedical and semantic segmentation, image restoration, and diffusion-model backbones.

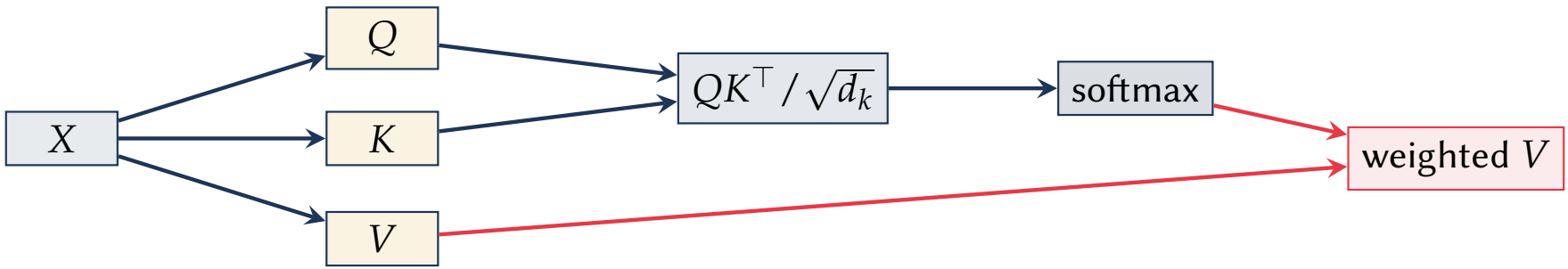
Ronneberger–Fischer–Brox, MICCAI (2015).

Architecture IV: attention

Given token representations $X \in \mathbb{R}^{n \times d}$, self-attention forms queries, keys, and values:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V,$$

$$\boxed{\mathrm{Attn}(X) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.}$$



Design principle: content-dependent communication

Each token computes a data-dependent weighted average of all value tokens. Multi-head attention performs several such interactions in parallel.

Architectural consequence

Without positional encodings, self-attention is permutation equivariant: the position of each token is lost. Position information must therefore be supplied separately whenever token order matters.

Vaswani et al., "Attention Is All You Need," NeurIPS (2017).

Gap between statistics and optimization

1. Important architectures
2. **Gap between statistics and optimization**
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Cybenko: one hidden layer is universal

Cybenko's theorem: continuous sigmoidal σ

For every $f \in C([0,1]^d)$ and $\varepsilon > 0$, there exist $M \in \mathbb{N}$ and parameters (a_j, w_j, b_j) such that

$$\boxed{\left\| f - \sum_{j=1}^{M} a_j \sigma(w_j^\top x + b_j) \right\|_\infty < \varepsilon.}$$

Main proof mechanism

Let $\mathcal{S}$ be the span of the ridge functions $\sigma(w^\top x + b)$.

1. If $\overline{\mathcal{S}} \neq C([0,1]^d)$, Hahn–Banach gives a nonzero continuous linear functional L that vanishes on $\mathcal{S}$.
2. By the Riesz representation theorem, $L(g) = \int g \, \mathrm{d}\mu$ for a nonzero signed measure μ.
3. Thus $\int \sigma(w^\top x + b) \, \mathrm{d}\mu(x) = 0$ for every w, b. A sigmoid is *discriminatory*: after rescaling it approaches half-space indicators, forcing $\mu = 0$, a contradiction.

G. Cybenko, *Math. Control Signals Systems* 2 (1989), 303–314.

Statistical optimality does not solve training

Minimax benchmark for $Y_i = f_*(X_i) + \varepsilon_i$

For a β-smooth function of d variables, under standard assumptions,

$$\inf_{\widehat{f}} \sup_{f_* \in \mathcal{H}^\beta} \mathbb{E}\|\widehat{f} - f_*\|_{L^2}^2 \asymp n^{-\frac{2\beta}{2\beta+d}}.$$

Neural-network estimators can attain optimal rates

- Deep ReLU: near-minimax on compositional classes (up to logs).
- Norm-constrained shallow ReLU networks are minimax up to logarithmic factors for the specified Hölder and neural variation classes.

The statistics–optimization gap

Minimax theory controls a global or constrained empirical-risk minimizer. It does not show that SGD finds it in the nonconvex parameter landscape.

Does the design of neural network matters for optimization?

Stone, *Ann. Statist.* 10 (1982); Schmidt–Hieber, *Ann. Statist.* 48 (2020); Yang–Zhou, *JMLR* 25 (2024).

Training a neural network means minimizing a finite sum

A shallow neural network

For parameters $\theta = \{(\omega_j, u_j)\}_{j=1}^{M}$,

$$F_\theta(x) = \frac{1}{M} \sum_{j=1}^{M} u_j \, \phi(\omega_j, x), \qquad x \in \mathbb{R}^d,$$

$$\mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta), \qquad \ell_i(\theta) = \frac{1}{2} |F_\theta(x_i) - y_i|^2,$$

using the samples $\{(x_i, y_i)\}_{i=1}^{N}$.

Gradient descent: all samples

$$\theta^{k+1} = \theta^k - \eta_k \nabla \mathcal{L}_N(\theta^k)$$

$$\nabla \mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \nabla \ell_i(\theta).$$

SGD: one cheap, unbiased sample

Draw $I_k \sim \mathrm{Unif}\{1, \ldots, N\}$ and update

$$\theta^{k+1} = \theta^k - \eta_k \nabla \ell_{I_k}(\theta^k).$$

$$\mathbb{E}\left[\nabla \ell_{I_k}(\theta)\right] = \nabla \mathcal{L}_N(\theta).$$

Gradient schemes: explicit and implicit

1. Important architectures
2. Gap between statistics and optimization
3. **Gradient schemes: explicit and implicit**
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Gradient flow is continuous-time steepest descent

Euclidean gradient flow

For a differentiable objective $F : \mathbb{R}^d \to \mathbb{R}$,

$$\dot{x}(t) = -\nabla F(x(t)), \qquad x(0) = x_0.$$

Along the trajectory, the objective dissipates at the rate

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

Continuous flow versus discrete algorithm

Gradient flow evolves in continuous time t along an ODE.

Gradient descent evolves in discrete time k and depends on the step size.

$$x(t + \tau) = x(t) - \tau \nabla F(x(t)) + o(\tau).$$

Gradient descent is the explicit Euler discretization.

Explicit and implicit steps discretize the same flow

Explicit Euler

Evaluate the gradient at the current iterate:

$$x_{k+1} = x_k - \tau \nabla F(x_k).$$

- One gradient evaluation per step.
- For an L-smooth objective, stability requires a controlled step size.

Implicit Euler

Evaluate the gradient at the next iterate:

$$x_{k+1} = x_k - \tau \nabla F(x_{k+1}).$$

For convex F, equivalently,

$$x_{k+1} \in \operatorname*{argmin}_x \left\{ F(x) + \frac{\|x - x_k\|^2}{2\tau} \right\}.$$

The price and benefit of the implicit step

It requires solving a proximal subproblem, but every step satisfies

$$F(x_{k+1}) + \frac{\|x_{k+1} - x_k\|^2}{2\tau} \leq F(x_k).$$

Implicit Euler is explicit descent on the Moreau envelope

For a proper lower-semicontinuous convex F and $\tau > 0$, define

$$p_\tau(x) := \operatorname*{argmin}_y \left\{ F(y) + \frac{\|y - x\|^2}{2\tau} \right\}, \qquad F_\tau(x) := \min_y \left\{ F(y) + \frac{\|y - x\|^2}{2\tau} \right\}.$$

The Moreau envelope F_τ is differentiable and

$$\nabla F_\tau(x) = \frac{x - p_\tau(x)}{\tau}.$$

Explicit for F_τ is implicit for F

$$\boxed{x_{k+1} = x_k - \tau \nabla F_\tau(x_k) = p_\tau(x_k).}$$

Proximal optimality gives

$$0 \in \partial F(x_{k+1}) + \frac{x_{k+1} - x_k}{\tau}, \qquad \text{or} \qquad \frac{x_{k+1} - x_k}{\tau} = -\nabla F(x_{k+1})$$

when F is differentiable.

The implicit scheme survives without coordinates

Let $(\mathcal{X}, d)$ be a metric space and $F : \mathcal{X} \to (-\infty, +\infty]$.

Minimizing movement scheme

Starting from x_0^τ, define

$$x_{k+1}^\tau \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F(x) + \frac{d^2(x, x_k^\tau)}{2\tau} \right\}.$$

Only the energy F and the distance d are needed: there is no vector subtraction and no classical gradient.

Energy controls the motion

Comparing the minimizer with x_k^τ gives

$$F(x_{k+1}^\tau) + \frac{d^2(x_{k+1}^\tau, x_k^\tau)}{2\tau} \leq F(x_k^\tau).$$

Under standard lower-semicontinuity and compactness assumptions, the interpolated iterates converge as $\tau \downarrow 0$ to a metric gradient flow.

Discretization bias of gradient descent

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. **Discretization bias of gradient descent**
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Which flow does explicit gradient descent really follow?

Let $E : \mathbb{R}^m \to \mathbb{R}$ be smooth and write $f = -\nabla E$. Gradient descent is

$$\theta_{n+1} = \theta_n + hf(\theta_n).$$

The flow of the original ODE misses the discrete step

The exact solution of $\dot{\theta} = f(\theta)$ satisfies

$$\theta(h) = \theta + hf(\theta) + \frac{h^2}{2} f'(\theta) f(\theta) + \mathcal{O}(h^3).$$

Hence one Euler step and the time-h flow of $-\nabla E$ differ by $\mathcal{O}(h^2)$.

Backward error analysis

Instead of asking how the numerical scheme departs from the given ODE, seek a nearby vector field

$$\widetilde{f} = f + hf_1 + h^2 f_2 + \cdots$$

whose exact flow interpolates the discrete iterates to higher order.

Matching the modified flow determines the correction

For $\dot{\theta} = \widetilde{f}(\theta)$ with $\widetilde{f} = f + hf_1 + \mathcal{O}(h^2)$, Taylor expansion gives

$$\theta(h) = \theta + hf(\theta) + h^2 \left(f_1(\theta) + \frac{1}{2} f'(\theta) f(\theta) \right) + \mathcal{O}(h^3).$$

Match one gradient-descent step

Since the discrete update is exactly $\theta + hf(\theta)$, cancellation of the order-h^2 term forces

$$\boxed{f_1 = -\frac{1}{2} f' f.}$$

For $f = -\nabla E$, symmetry of the Hessian gives

$$f_1 = -\frac{1}{2} D^2 E \, \nabla E = -\frac{1}{4} \nabla \|\nabla E\|^2.$$

The modified vector field is again a gradient field.

Gradient descent implicitly regularizes the loss

Implicit gradient regularization

To first order in h, gradient descent follows the gradient flow of

$$\boxed{\widetilde{E}(\theta) = E(\theta) + \frac{h}{4}\|\nabla E(\theta)\|^2.}$$

One step agrees with the time-h flow of $-\nabla\widetilde{E}$ up to $\mathcal{O}(h^3)$, instead of $\mathcal{O}(h^2)$ for $-\nabla E$.

Interpretation

- Discretization creates a penalty on the local slope $\|\nabla E\|$, with strength proportional to h.
- Existing critical points remain critical, but the trajectory—and therefore the selected minimum—can change.

Scope: This is an asymptotic, moderate-step-size description; $E + \mu\|\nabla E\|^2$ decouples the penalty from h.

Barrett–Dherin, *Implicit Gradient Regularization*, ICLR 2021.

Convexity

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. **Convexity**
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Convexity makes descent globally meaningful

Convexity

For every $x, y \in \mathbb{R}^d$ and $s \in [0,1]$,

$$F((1-s)x + sy) \leq (1-s)F(x) + sF(y).$$

First-order characterization

If F is differentiable, convexity is equivalent to

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle.$$

Global consequence

If $\nabla F(x) = 0$, the first-order inequality gives $F(y) \geq F(x)$ for every y. Thus every stationary point is a global minimizer.

Smooth convex objectives give an $O(1/k)$ rate

L-smoothness

The gradient is L-Lipschitz:

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|.$$

With step size $1/L$, the descent lemma gives

$$x_{k+1} = x_k - \frac{1}{L}\nabla F(x_k), \qquad F(x_{k+1}) \leq F(x_k) - \frac{\|\nabla F(x_k)\|^2}{2L}.$$

Convergence under convexity

If F is also convex, $x_* \in \operatorname{argmin} F$, and $F_* = F(x_*)$, then

$$\boxed{F(x_k) - F_* \leq \frac{L\|x_0 - x_*\|^2}{2k}}, \qquad k \geq 1.$$

The objective values converge globally, but only at a sublinear rate.

Implicit descent converges without smoothness

Let $F : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lower-semicontinuous and convex, with $x_* \in \operatorname{argmin} F$, and fix $\tau > 0$.

Apply the smooth result to the Moreau envelope

The function F_τ is convex, $1/\tau$-smooth, and $\min F_\tau = \min F =: F_*$. Moreover,

$$x_{k+1} = p_\tau(x_k) = x_k - \tau \nabla F_\tau(x_k), \qquad F_\tau(x_k) - F_* \leq \frac{\|x_0 - x_*\|^2}{2\tau k}.$$

The envelope identity transfers this estimate to the original objective:

$$F_\tau(x_k) = F(x_{k+1}) + \frac{\|x_{k+1} - x_k\|^2}{2\tau} \geq F(x_{k+1}).$$

The proximal-point rate

$$\boxed{F(x_{k+1}) - F_* \leq \frac{\|x_0 - x_*\|^2}{2\tau k}, \qquad k \geq 1.}$$

No differentiability of F is required; in $\mathbb{R}^d$, x_k converges to a minimizer.

Not every optimization problem is convex

A basic example: shallow neural networks

For

$$f_\theta(x) = \frac{1}{M}\sum_{j=1}^{M} u_j \phi(\omega_j, x), \qquad \mathcal{L}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i\big(f_\theta(x_i)\big),$$

the empirical risk is generally nonconvex in $\theta = \{(\omega_j, u_j)\}_{j=1}^{M}$, because the hidden parameters ω_j enter nonlinearly.

A universal convex lift

For any $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$,

$$\boxed{\inf_{x \in \mathbb{R}^d} F(x) = \inf_{\rho \in \mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} F(x)\, \mathrm{d}\rho(x).}$$

The lifted objective is linear in ρ, and $\mathcal{P}(\mathbb{R}^d)$ is convex. The two problems have the same value because every Dirac mass δ_x is admissible.

Shallow networks become convex over measures

Encode the neurons by their empirical measure; the predictor is then linear:

$$\rho_M = \frac{1}{M}\sum_{j=1}^{M} \delta_{(\omega_j, u_j)}, \qquad f_{\rho_M}(x) = \int u\phi(\omega, x)\, \mathrm{d}\rho_M(\omega, u).$$

Convexity of the lifted risk

If every loss ℓ_i is convex in the prediction, then

$$\boxed{\mathcal{L}_N(\rho) := \frac{1}{N}\sum_{i=1}^{N} \ell_i\big(f_\rho(x_i)\big) \quad \text{is convex in } \rho.}$$

No free lunch

The convex variable ρ belongs to an infinite-dimensional space of measures. Restricting ρ to finitely many atoms recovers the original nonconvex particle parametrization.

PL condition

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. **PL condition**
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Circumventing convexity: the PL condition

Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and write $F_* = \inf_{x \in \mathbb{R}^d} F(x)$.

Polyak–Łojasiewicz inequality

For some $\mu > 0$,

$$\frac{1}{2}\|\nabla F(x)\|^2 \geq \mu\big(F(x) - F_*\big) \qquad \text{for every } x \in \mathbb{R}^d.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(F(x(t)) - F_*\big) = -\|\nabla F(x(t))\|^2 \leq -2\mu\big(F(x(t)) - F_*\big),$$

hence

$$\boxed{F(x(t)) - F_* \leq e^{-2\mu t}\big(F(x_0) - F_*\big).}$$

If F is also L-smooth, explicit gradient descent with step $1/L$ obeys

$$F(x_k) - F_* \leq (1 - \mu/L)^k\big(F(x_0) - F_*\big).$$

The PL condition extends to metric spaces

For $F : \mathcal{X} \to (-\infty, +\infty]$ on $(\mathcal{X}, d)$, the *descending slope* is

$$\boxed{|\partial F|(x) := \limsup_{\substack{y \to x \\ y \neq x}} \frac{\big(F(x) - F(y)\big)_+}{d(x,y)}}, \qquad (a)_+ := \max\{a, 0\}.$$

For differentiable F on $\mathbb{R}^d$, this gives $|\partial F|(x) = \|\nabla F(x)\|$.

Metric PL inequality

$$\frac{1}{2}|\partial F|^2(x) \geq \mu\big(F(x) - F_*\big).$$

If a metric gradient flow satisfies the energy-dissipation identity $-\frac{\mathrm{d}}{\mathrm{d}t}F(x(t)) = |\partial F|^2(x(t))$, then

$$\boxed{F(x(t)) - F_* \leq e^{-2\mu t}\big(F(x_0) - F_*\big).}$$

The same convergence mechanism therefore needs no linear structure.

Strong convexity implies PL, but PL is weaker

If F is differentiable and μ-strongly convex, then

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

Minimizing the right-hand side over y yields

$$F_* \geq F(x) - \frac{1}{2\mu}\|\nabla F(x)\|^2,$$

and therefore

$$\boxed{\mu\big(F(x) - F_*\big) \leq \frac{1}{2}\|\nabla F(x)\|^2.}$$

What is actually necessary?

- Convexity alone is not enough: $F(x) = x^4$ is convex, but its PL ratio $\|F'(x)\|^2/[2(F(x) - F_*)] = 8x^2$ vanishes near 0.
- Strong convexity is sufficient, not necessary; PL can also hold for nonconvex objectives.
- PL rules out stationary points above the minimum value.

When an optimization problem is PL (other than convex) ?

Example: overparameterized nonlinear least squares

Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$, with $p \geq n$, and define

$$r(\theta) = \Phi(\theta) - y, \qquad \mathcal{L}(\theta) = \frac{1}{2}\|r(\theta)\|^2, \qquad J(\theta) = D\Phi(\theta).$$

Then $\nabla \mathcal{L}(\theta) = J(\theta)^\top r(\theta)$.

A well-conditioned tangent kernel gives PL

Assume interpolation is possible, so $\mathcal{L}_* = 0$, and suppose

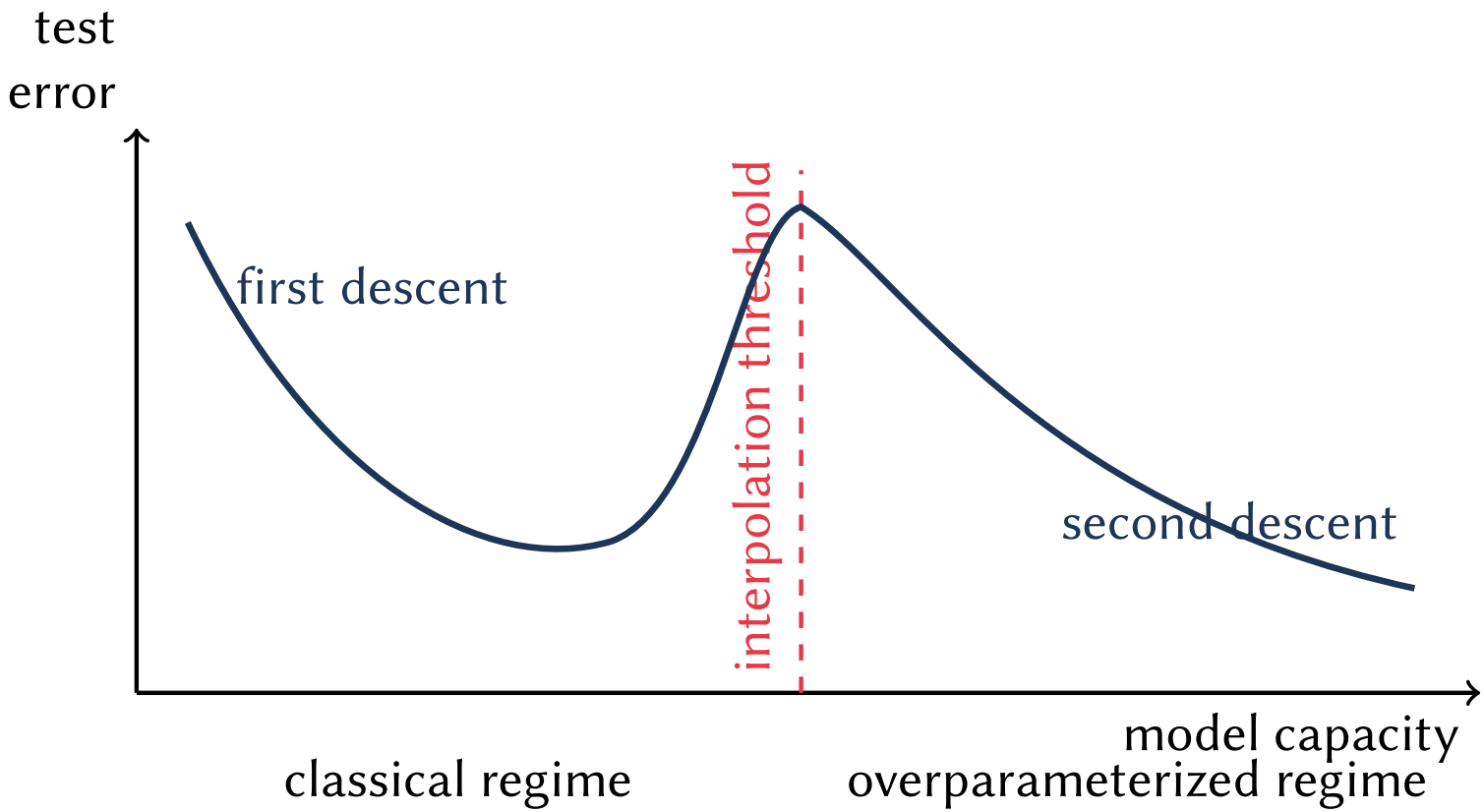
$$K(\theta) := J(\theta)J(\theta)^\top \succeq \mu I_n \qquad \text{along the optimization path.}$$

$$\boxed{\frac{1}{2}\|\nabla \mathcal{L}(\theta)\|^2 = \frac{1}{2} r(\theta)^\top K(\theta) r(\theta) \geq \mu \mathcal{L}(\theta).}$$

For $p \geq n$, full row rank is possible. In a wide neural network, $K(\theta)$ is the empirical tangent kernel, and width can help keep $\lambda_{\min}(K(\theta))$ away from zero.

Liu–Zhu–Belkin, arXiv:2003.00307 (2021).

Overparameterization need not hurt generalization



The double-descent observation

Test error may peak near the interpolation threshold and then decrease as the model becomes even larger. Extra parameters can therefore improve optimization without necessarily degrading generalization.

Belkin–Hsu–Ma–Mandal, *PNAS* 116 (2019), 15849–15854; Nakkiran et al., *ICLR* (2020).

Tangent controllability gives PL on every ball

Let $\mathcal{U}$ be a Hilbert space of controls and $\mathcal{E}(u) = x_u(T) \in \mathbb{R}^n$ the endpoint map. For a reachable target y, set $\mathcal{J}(u) = \frac{1}{2}\|\mathcal{E}(u) - y\|^2$, so $\mathcal{J}_* = 0$.

Quantitative tangent controllability on B_R

Suppose that for every $R > 0$ there is $\alpha_R > 0$ such that

$$D\mathcal{E}(u)D\mathcal{E}(u)^* \succeq \alpha_R I_n \qquad \text{whenever } \|u\|_{\mathcal{U}} \leq R.$$

Since $\nabla\mathcal{J}(u) = D\mathcal{E}(u)^*(\mathcal{E}(u) - y)$,

$$\boxed{\frac{1}{2}\|\nabla\mathcal{J}(u)\|^2 \geq \alpha_R \mathcal{J}(u) \qquad \text{on } B_R.}$$

PL on every ball is not a global PL inequality

The constants may degenerate, $\alpha_R \downarrow 0$ as $R \to \infty$. Without coercivity or bounded sublevel sets, a descent trajectory can leave every ball: **divergence at infinity can still occur.**

Why tangent controllability implies PL

Put $e(u) = \mathcal{E}(u) - y$. For $\mathcal{J}(u) = \frac{1}{2}\|e(u)\|^2$, the chain rule gives

$$\nabla \mathcal{J}(u) = D\mathcal{E}(u)^* e(u).$$

Proof on the control ball B_R

Quantitative tangent controllability means $D\mathcal{E}(u)D\mathcal{E}(u)^* \succeq \alpha_R I_n$. Therefore

$$\begin{aligned} \|\nabla \mathcal{J}(u)\|^2 &= \|D\mathcal{E}(u)^* e(u)\|^2 \\ &= \langle e(u), D\mathcal{E}(u)D\mathcal{E}(u)^* e(u)\rangle \\ &\geq \alpha_R \|e(u)\|^2 = 2\alpha_R \mathcal{J}(u). \end{aligned}$$

Hence

$$\boxed{\frac{1}{2}\|\nabla \mathcal{J}(u)\|^2 \geq \alpha_R\big(\mathcal{J}(u) - \mathcal{J}_*\big), \qquad \mathcal{J}_* = 0.}$$

Where the hypothesis enters

Surjectivity of $D\mathcal{E}(u)$ rules out a nonzero residual orthogonal to all reachable first-order variations. The lower bound by α_R makes this statement quantitative.

The optimization landscape of ResNets

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. **The optimization landscape of ResNets**
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. Mean field limit of transformers

Need an authority argument?

Residual connections is a subset of deep neural networks

Plain deep network

$$x_{k+1} = F_{\theta_k}(x_k),$$

$$f_\theta = F_{\theta_{L-1}} \circ \cdots \circ F_{\theta_0}.$$

Every layer must transmit the state through a new nonlinear map.

Residual network

$$x_{k+1} = x_k + h f_{\theta_k}(x_k),$$

$$k = 0, \ldots, L-1.$$

The skip connection carries x_k directly; the trainable block only learns a correction.

An architectural optimization prior

Setting $f_{\theta_k} = 0$ gives the identity map exactly. A deeper ResNet can therefore preserve a shallower solution and add useful transformations one increment at a time.

The architecture solves an optimization issue

He–Zhang–Ren–Sun, *CVPR* (2016).

Why linearize each layer around the identity?

Start with a deep *linear* network:

$$f_\theta(x) = \Theta_L \Theta_{L-1} \cdots \Theta_1 x.$$

Raw products amplify depth

$\|\Theta_k\| \le \rho$ gives $\|\Theta_L \cdots \Theta_1\| \le \rho^L$: contraction for $\rho < 1$, possible explosion for $\rho > 1$.
Backpropagation contains the same kind of product.

The order-one scaling

Keep each factor close to the identity:

$$\boxed{\Theta_k = I + \frac{1}{L} U_k}, \qquad \|U_k\| \le M, \qquad \left\| \prod_{k=1}^{L} \left(I + \frac{1}{L} U_k \right) \right\| \le \left(1 + \frac{M}{L} \right)^L \le e^M,$$

uniformly in depth. If $U_k \equiv U$, the product converges to e^U.

$I + L^{-1} U_k$ **is precisely a linear residual layer.**

The nonlinear replacement $x_{k+1} = x_k + L^{-1} f_{\theta_k}(x_k)$ preserves this stable scaling.

Adapted from the HCERES talk, "Why is it a natural idea?"

Identity paths make deep optimization easier

The degradation problem in plain networks

A deeper model is more expressive, yet its *training* error can be worse: optimization requires correct initialization and non-vanishing/exploding gradient.

Without a skip connection

- More parameters/expressive than ResNets.
- Needs proper initialization.
- Vanishing/exploding gradients without proper techniques.

With a skip connection

- Less parameters/expressive.
- Simple initialization.
- Proper scale of gradients.

ResNets change the optimization geometry, not the training objective.

Skip connections stabilize backpropagation

For a plain network, the chain rule gives

$$D_{x_0} x_L = DF_{\theta_{L-1}}(x_{L-1}) \cdots DF_{\theta_0}(x_0).$$

If the factors are contractive, gradients vanish exponentially with depth; expansive factors can instead make them explode.

Residual Jacobians remain close to the identity

For $x_{k+1} = x_k + h f_{\theta_k}(x_k)$, write $A_k = Df_{\theta_k}(x_k)$. Then

$$D_{x_0} x_L = \prod_{k=0}^{L-1} (I + hA_k).$$

If $\|A_k\| \leq M$ and $hM < 1$, then

$$(1 - hM)^L \leq \sigma_{\min}(D_{x_0} x_L) \leq \|D_{x_0} x_L\| \leq (1 + hM)^L.$$

With $h = 1/L$, these bounds behave like e^{-M} and e^{M}, rather than deteriorating geometrically with the number of layers.

Infinite-depth ResNets are ODE flows

Scale each residual block by $h = 1/L$:

$$q_{k+1} = q_k + \frac{1}{L} v_{k/L}(q_k).$$

This is explicit Euler. As $L \to \infty$, depth becomes time and

$$\boxed{\partial_t q_t = v_t(q_t), \qquad q_0 = x.}$$

The parameters become a control

The time-dependent vector field $v = (v_t)_{t\in[0,1]}$ transports each input. Training chooses this control so that the endpoints $q_i(1)$ match the labels after the output layer.

Why the residual design matters

Depth refines a flow instead of repeatedly replacing the representation. Forward propagation is an ODE; backpropagation is its adjoint equation. Optimization can therefore be studied through stability and controllability.

What is a reproducing kernel Hilbert space?

An RKHS H is a Hilbert space of vector fields $v : \mathbb{R}^d \to \mathbb{R}^d$ in which point evaluation is continuous. It is described by a matrix-valued kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfying

$$\boxed{\langle v, K(\cdot, x)a\rangle_H = a^\top v(x)}, \qquad x, a \in \mathbb{R}^d.$$

Why the reproducing property matters

Cauchy–Schwarz controls every evaluation by the Hilbert norm:

$$\|v(x)\| \le \sqrt{\|K(x,x)\|_{\mathrm{op}}}\, \|v\|_H.$$

Finite kernel expansions $v(\cdot) = \sum_{\ell=1}^{p} K(\cdot, z_\ell)c_\ell$ are linear in their coefficients, smooth when K is smooth, and easy to evaluate.

A rich linear space of nonlinear vector fields

For example,

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2s^2}\right) I_d$$

gives a Gaussian RKHS. On compact sets, such kernels can approximate continuous vector fields while the control variable $v \in H$ still belongs to a linear Hilbert space.

A tractable model keeps the nonlinear flow

Let H be an RKHS of vector fields and optimize a time-dependent control

$$v \in L^2([0,1], H), \qquad \dot{q}_i(t) = v_t(q_i(t)), \qquad q_i(0) = A(x_i).$$

With a fixed output map B, train the terminal loss

$$\mathcal{L}(v) = \frac{1}{2N} \sum_{i=1}^{N} \|Bq_i(1) - y_i\|^2.$$

Linear at each instant

Each v_t belongs to the linear space H. Kernel expansions or an infinite-width outer layer provide concrete parametrizations.

Nonlinear end-to-end

The map $v \mapsto q_i(1)$ is nonlinear because the vector field is evaluated along the trajectory it generates.

Tangent controllability makes the ResNet loss locally PL

Compute the final positions $\mathcal{E}(v) = (Bq_1(1), \dots, Bq_N(1))$ and write $\mathcal{L}(v) = \frac{1}{2N}\|\mathcal{E}(v) - y\|^2$.

A quantitative tangent-surjectivity condition

If, on every ball $\|v\|_{L^2(H)} \leq R$,

$$D\mathcal{E}(v)D\mathcal{E}(v)^* \succeq \alpha_R I,$$

then the chain rule gives a local PL inequality

$$\boxed{\frac{1}{2}\|\nabla \mathcal{L}(v)\|^2 \geq \frac{\alpha_R}{N}\,\mathcal{L}(v).}$$

Optimization consequence

Every bounded critical point has zero loss. If the gradient-flow trajectory remains in a ball, the training loss converges exponentially to zero.

Barboni–Peyré–Vialard, *Global convergence of ResNets*, NeurIPS (2022).

Residual design helps, but is not magic

Why overparameterization helps

A rich space H supplies enough infinitesimal directions to move all training endpoints. Residual composition turns this expressivity into *tangent controllability*, which is precisely what yields local PL.

The remaining global obstruction

- The constant α_R may deteriorate as $R \to \infty$.
- Symmetries in data can favor a singular configuration: think to 1d alignment in $\mathbb{R}^2$.
- The loss may decrease while $\|v\|_{L^2(H)} \to \infty$; boundedness of the training trajectory is therefore a genuine assumption.

Favorable local geometry does not rule out escape to infinity.

The benefits of mean-field limits

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. **The benefits of mean-field limits**
9. Wasserstein gradient flows
10. Mean field limit of transformers

Mean-field representation of shallow neural networks

$$\rho_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{(a_j, w_j, b_j)}, \qquad f_m(x) = \int a\,\sigma(w^\top x + b)\,\mathrm{d}\rho_m(a, w, b).$$

Infinite width: optimize over measures

$$f_\rho(x) = \int a\,\sigma(w^\top x + b)\,\mathrm{d}\rho(a, w, b).$$

Width becomes the number of atoms approximating ρ. The predictor is linear in ρ, while optimization over finitely many particle locations remains nonconvex.

Barron space can break the curse of dimension

Barron controls a function through a weighted L^1 norm of its Fourier transform, measuring one derivative. Finite spectral Barron norm implies the existence of a width-m approximation satisfying

$$\boxed{\|f - f_m\|_{L^2(\mu)} \le \frac{C_X \|f\|_{\mathcal{B}}}{\sqrt{m}}.}$$

The exponent does not depend on d: this can avoid the usual curse of dimensionality. Caveat: $\|f\|_{\mathcal{B}}$ may itself grow with d.

A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Information Theory*, 1993.

Approximate Carathéodory

Classical Carathéodory. If $x \in \operatorname{conv}(T)$ for $T \subset \mathbb{R}^n$, then x is an exact convex combination of at most $n+1$ points of T. Exactness therefore costs a number of atoms that grows with the ambient dimension.

Approximate Carathéodory, or Maurey's lemma

Let T lie in a Hilbert space H, with $\operatorname{diam}(T) \le D$, and let $x = \int_T z \, \mathrm{d}\rho(z)$ for a probability measure ρ on T. For every $m \ge 1$, there exist $z_1, \dots, z_m \in T$, with repetitions allowed, such that

$$\boxed{\left\| x - \frac{1}{m} \sum_{j=1}^{m} z_j \right\|_H \le \frac{D}{\sqrt{m}}.}$$

Application to the mean-field representation

Take $H = L^2(\mu)$ and $T = \{\phi_\theta : \theta \in \Theta\}$, where $\phi_\theta(x) = a\sigma(w^\top x + b)$. Since $f_\rho = \int \phi_\theta \, \mathrm{d}\rho(\theta)$, there is an empirical measure $\rho_m = m^{-1} \sum_j \delta_{\theta_j}$ satisfying

$$\| f_\rho - f_{\rho_m} \|_{L^2(\mu)} \le D/\sqrt{m}.$$

The weights are equal and the exponent does not depend on dimension.

Maurey's empirical method

Draw independent random features

$$Z_1, \dots, Z_m \overset{\text{i.i.d.}}{\sim} \rho, \qquad \overline{Z}_m = \frac{1}{m} \sum_{j=1}^{m} Z_j, \qquad \mathbb{E}\, \overline{Z}_m = x.$$

Independence gives the dimension-free rate

Centering and independence make every cross term vanish:

$$\begin{aligned} \mathbb{E}\|\overline{Z}_m - x\|_H^2 &= \frac{1}{m^2} \sum_{j,k=1}^{m} \mathbb{E}\langle Z_j - x, Z_k - x\rangle_H \\ &= \frac{1}{m} \mathbb{E}\|Z_1 - x\|_H^2 \le \frac{D^2}{m}. \end{aligned}$$

The probabilistic method selects one finite network

Some realization must have error no larger than the expectation. Thus there are $z_1, \dots, z_m \in T$ with

$$\left\| x - \frac{1}{m} \sum_{j=1}^{m} z_j \right\|_H \le \frac{D}{\sqrt{m}}.$$

For neural networks, $Z_j = \phi_{\Theta_j}$ with $\Theta_j \overset{\text{i.i.d.}}{\sim} \rho$: the random objects are the neuron parameters, while μ only defines the norm over inputs.

Barron space is a broad approximation class

A large class described by one Fourier moment

Barron space consists of functions admitting an extension whose Fourier transform has an integrable first moment $\int \|\omega\| \, |\widehat{F}(\omega)| \, \mathrm{d}\omega$. It contains, in particular, restrictions of smooth compactly supported functions. On a compact domain, this already gives a dense class of continuous functions.

Proof architecture: Fourier representation plus Maurey

1. The Fourier representation expresses f as a signed integral of ridge features.
2. Absorb signs and amplitudes into the neuron parameters, then normalize the resulting measure into a probability law π_F.
3. Apply Maurey's empirical method to π_F: an average of m sampled neurons approximates f with error

$$\|f - f_m\|_{L^2(\mu)} \lesssim \frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}.$$

The Fourier representation defines the sampling law; Maurey's lemma provides the finite network.

The Barron relaxation convexifies the neural loss

Put one neuron at $\theta = (a, w, b)$ and write $\phi_\theta(x) = a\sigma(w^\top x + b)$. For $\rho \in \mathcal{P}(\Theta)$, define

$$f_\rho(x) = \int_\Theta \phi_\theta(x)\, \mathrm{d}\rho(\theta), \qquad \mathcal{L}(\rho) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\rho(x_i), y_i\big).$$

Convexity in the linear structure of measures

For $\rho_s = (1-s)\rho_0 + s\rho_1$,

$$f_{\rho_s}(x) = (1-s) f_{\rho_0}(x) + s f_{\rho_1}(x).$$

Hence, if $z \mapsto \ell(z, y)$ is convex for every y, then

$$\boxed{\mathcal{L}(\rho_s) \le (1-s)\mathcal{L}(\rho_0) + s\mathcal{L}(\rho_1).}$$

But how is the measure optimized in practice?

A finite network is the empirical measure $\rho_m = m^{-1} \sum_{j=1}^{m} \delta_{\theta_j}$. Gradient descent moves the atoms θ_j; it does not form affine mixtures of entire measures. We therefore need the geometry generated by *particle motion*.

Wasserstein gradient flows

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. **Wasserstein gradient flows**
10. Mean field limit of transformers

Chizat–Bach: can particle flow find the convex optimum?

Their starting point is a convex problem over measures,

$$J(\rho) = R\left(\int_\Theta \phi_\theta \, \mathrm{d}\rho(\theta)\right) + G(\rho),$$

covering shallow networks and sparse inverse problems. In practice, replace ρ by particles and run gradient descent on their weights and positions.

Their contribution

- Identify the many-particle, infinitesimal-step limit as a Wasserstein gradient flow of J.
- Use homogeneity to encode signed coefficients by positive measures on an enlarged parameter space.
- Under structural assumptions on the model and a sufficiently rich initialization, prove convergence of the limiting particle flow to global minimizers of the convex measure problem.

What this explains, and what it does not

Overparameterization lets the particle system approximate a measure flow that can escape nonglobal stationary configurations. The theorem concerns the many-particle continuous-time regime and requires nontrivial initialization/support assumptions; it is not a generic finite-width rate.

L. Chizat and F. Bach, “On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport,” NeurIPS 2018.

The geometry of particles is transport geometry

Let the particles move along $t \mapsto \theta_j(t)$ and set

$$\rho_t = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j(t)}.$$

If $v_t(\theta_j(t)) = \dot{\theta}_j(t)$, then ρ_t satisfies, weakly,

$$\boxed{\partial_t \rho_t + \nabla_\theta \cdot (\rho_t v_t) = 0.}$$

The kinetic norm of a tangent velocity

The Euclidean cost of moving all parameters becomes

$$\frac{1}{m} \sum_{j=1}^{m} \|\dot{\theta}_j(t)\|^2 = \int_\Theta \|v_t(\theta)\|^2 \, \mathrm{d}\rho_t(\theta).$$

Thus a tangent vector $\dot{\rho} = -\nabla_\theta \cdot (\rho v)$ is assigned the squared norm

$$\|\dot{\rho}\|_\rho^2 = \inf_{-\nabla \cdot (\rho v) = \dot{\rho}} \int_\Theta \|v\|^2 \, \mathrm{d}\rho.$$

Mass is transported through parameter space: this is the local W_2 geometry.

The Wasserstein metric measures particle displacement

For $\rho_0, \rho_1 \in \mathcal{P}_2(\Theta)$,

$$W_2^2(\rho_0, \rho_1) = \inf_{\gamma \in \Pi(\rho_0, \rho_1)} \int_{\Theta \times \Theta} \|\theta - \theta'\|^2 \, \mathrm{d}\gamma(\theta, \theta').$$

Benamou–Brenier dynamic formulation

$$\boxed{W_2^2(\rho_0, \rho_1) = \inf_{(\rho_t, v_t)} \int_0^1 \int_\Theta \|v_t(\theta)\|^2 \, \mathrm{d}\rho_t(\theta) \, \mathrm{d}t}$$

subject to

$$\partial_t \rho_t + \nabla_\theta \cdot (\rho_t v_t) = 0, \qquad \rho_{t=0} = \rho_0, \quad \rho_{t=1} = \rho_1.$$

For empirical measures

When particles can be matched without splitting, W_2^2 is the minimum average squared distance traveled by the parameters, up to permutation.

Monge transport and the Benamou–Brenier time-dependent formulation

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. A transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ is admissible when $T_\# \mu = \nu$, and Monge's problem is

$$\boxed{\inf_{T_\#\mu=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\mu(x).}$$

For quadratic cost and absolutely continuous μ, Brenier's theorem gives an optimal map $T = \nabla\varphi$ with φ convex.

From a velocity field to a transport map

If $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ and the flow solves $\dot{X}_t(x) = v_t(X_t(x))$, then $\rho_t = (X_t)_\# \mu$ and $T = X_1$ transports μ to ν. Moreover,

$$\int \|T(x) - x\|^2 \, \mathrm{d}\mu(x) \le \int_0^1 \int \|v_t\|^2 \, \mathrm{d}\rho_t \, \mathrm{d}t.$$

From an optimal map to the minimizing path

Set $X_t = (1-t)\,\mathrm{Id} + tT$ and $\rho_t = (X_t)_\# \mu$. Then $v_t(X_t(x)) = T(x) - x$, so the kinetic action equals the Monge cost. Hence

$$W_2^2(\mu, \nu) = \inf_{\substack{\partial_t \rho + \nabla\cdot(\rho v) = 0 \\ \rho_0 = \mu,\ \rho_1 = \nu}} \int_0^1 \int \|v_t\|^2 \, \mathrm{d}\rho_t \, \mathrm{d}t.$$

Wasserstein gradient flow is steepest transport descent

Let $\mathcal{F} : \mathcal{P}_2(\Theta) \to \mathbb{R}$ have first variation $\delta\mathcal{F}/\delta\rho$. Along $\partial_t\rho + \nabla_\theta \cdot (\rho v) = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) = \int_\Theta \left\langle \nabla_\theta \frac{\delta\mathcal{F}}{\delta\rho}(\theta), v_t(\theta) \right\rangle \mathrm{d}\rho_t(\theta).$$

Choose the steepest velocity

Minimize instantaneous dissipation plus kinetic cost:

$$\min_v \left\{ \frac{1}{2}\int \|v\|^2 \,\mathrm{d}\rho + \int \left\langle \nabla_\theta \frac{\delta\mathcal{F}}{\delta\rho}, v \right\rangle \mathrm{d}\rho \right\}.$$

The minimizer and induced continuity equation are

$$v = -\nabla_\theta \frac{\delta\mathcal{F}}{\delta\rho}, \qquad \boxed{\partial_t\rho = \nabla_\theta \cdot \left(\rho \nabla_\theta \frac{\delta\mathcal{F}}{\delta\rho}\right).}$$

Consequently, $\dfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) = -\displaystyle\int \left\| \nabla_\theta \frac{\delta\mathcal{F}}{\delta\rho} \right\|^2 \mathrm{d}\rho_t \le 0.$

Particle training is a Wasserstein gradient flow

For the relaxed neural loss, its first variation is

$$\frac{\delta \mathcal{L}}{\delta \rho}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \partial_1 \ell\big(f_\rho(x_i), y_i\big) \phi_\theta(x_i).$$

Particle dynamics

$$\dot{\theta}_j = -\nabla_\theta \frac{\delta \mathcal{L}}{\delta \rho}(\theta_j).$$

Then $\rho_m = m^{-1} \sum_j \delta_{\theta_j}$ follows the Wasserstein flow. Parameter GD gives the same ODE after rescaling time by m.

Implicit/JKO scheme

$$\rho_{k+1} \in \operatorname*{argmin}_\rho \left\{ \mathcal{L}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho_k) \right\}.$$

This is proximal descent with Euclidean distance replaced by the cost of transporting neural parameters.

Key point

$\mathcal{L}$ is convex along affine mixtures $(1-s)\rho_0 + s\rho_1$, but particle optimization follows W_2 geodesics. This does not imply displacement convexity: the relaxed problem may remain hard. In favorable regimes, overparametrization can restore controllability.

Functionals on the space of probability measures

For $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, three basic classes are

$$\textbf{linear:} \quad \mathcal{V}(\rho) = \int_{\mathbb{R}^d} V(x)\,\mathrm{d}\rho(x),$$

$$\textbf{quadratic:} \quad \mathcal{W}(\rho) = \frac{1}{2}\iint_{\mathbb{R}^d\times\mathbb{R}^d} W(x,y)\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y),$$

$$\textbf{nonlinear/internal:} \quad \mathcal{U}(\rho) = \int_{\mathbb{R}^d} U(u(x))\,\mathrm{d}x, \qquad \rho = u\,\mathrm{d}x,$$

with $\mathcal{U}(\rho) = +\infty$ when ρ has no density.

λ-convexity in Wasserstein space

A functional $\mathcal{F}$ is λ-displacement convex if, along a constant-speed W_2 geodesic $(\rho_t)_{t\in[0,1]}$,

$$\boxed{\mathcal{F}(\rho_t) \le (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \frac{\lambda}{2}t(1-t)W_2^2(\rho_0,\rho_1).}$$

For an optimal map T from ρ_0 to ρ_1, one such geodesic is $\rho_t = ((1-t)\,\mathrm{Id} + tT)_{\#}\rho_0$.

Linear convexity $\neq$ displacement convexity

Wasserstein convexity tests transport geodesics, not affine mixtures $(1-t)\rho_0 + t\rho_1$. For example, $\mathcal{V}$ is displacement convex when V is convex; analogous conditions for $\mathcal{U}$ are subtler.

Convex inner potentials give displacement-convex energies

Let T be the optimal map from ρ_0 to ρ_1 and set $T_t = (1-t)\,\mathrm{Id} + tT$, so that $\rho_t = (T_t)_\# \rho_0$.

Linear potential energy

If $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex, then

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

Integrating in x gives

$$\boxed{\mathcal{V}(\rho) = \int V \,\mathrm{d}\rho \quad \text{is } \lambda\text{-displacement convex.}}$$

Conversely, test on Dirac masses: δ_x and δ_y.

Quadratic interaction energy

For $\mathcal{W}(\rho) = \frac{1}{2}\iint W(x-y)\,\mathrm{d}\rho(x)\mathrm{d}\rho(y)$,

$$\mathcal{W}(\rho_t) = \frac{1}{2}\iint W\big(T_t(x) - T_t(y)\big)\,\mathrm{d}\rho_0(x)\mathrm{d}\rho_0(y).$$

Because $T_t(x) - T_t(y) = (1-t)(x-y) + t(T(x)-T(y))$, convexity of W gives

$$\boxed{\mathcal{W}(\rho_t) \le (1-t)\mathcal{W}(\rho_0) + t\mathcal{W}(\rho_1).}$$

Entropy is convex along Wasserstein geodesics

McCann's criterion

If

$$s \longmapsto s^d U(s^{-d})$$

is convex and nonincreasing on $(0,\infty)$, then $\mathcal{U}(\rho) = \int U(u)\,\mathrm{d}x$, where $\rho = u\,\mathrm{d}x$, is displacement convex on $\mathcal{P}_2(\mathbb{R}^d)$.

For $U(r) = r\log r$, $s^d U(s^{-d}) = -d\log s$ is convex and decreasing. Therefore

$$\boxed{\mathrm{Ent}(\rho_t) \le (1-t)\,\mathrm{Ent}(\rho_0) + t\,\mathrm{Ent}(\rho_1),} \qquad \mathrm{Ent}(\rho) = \int u\log u\,\mathrm{d}x.$$

Why it is true: the Jacobian argument

Let $T = \nabla\varphi$ be the optimal map and $T_t = (1-t)\,\mathrm{Id} + tT$. The change-of-variables formula gives

$$\mathrm{Ent}(\rho_t) = \mathrm{Ent}(\rho_0) - \int \log\det\big((1-t)I + tDT(x)\big)\,\mathrm{d}\rho_0(x).$$

Since $DT \succeq 0$ and $A \mapsto \log\det A$ is concave, the last display lies below $(1-t)\,\mathrm{Ent}(\rho_0) + t\,\mathrm{Ent}(\rho_1)$.

R. J. McCann, *Advances in Mathematics* 128 (1997), 153–179.

Maximum mean discrepancy with a universal kernel

Let $k : \Omega \times \Omega \to \mathbb{R}$ be a continuous positive-definite kernel with RKHS $\mathcal{H}_k$. Embed a probability measure through its *kernel mean*

$$m_\mu := \int_\Omega k(x, \cdot)\, \mathrm{d}\mu(x) \in \mathcal{H}_k.$$

MMD is a dual distance between expectations

$$\begin{aligned} \mathrm{MMD}_k(\mu, \nu) &:= \|m_\mu - m_\nu\|_{\mathcal{H}_k} \\ &= \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu \right|, \end{aligned}$$

and, for independent $X, X' \sim \mu$ and $Y, Y' \sim \nu$,

$$\boxed{\mathrm{MMD}_k^2(\mu, \nu) = \mathbb{E}k(X, X') + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y).}$$

Why universality matters

If Ω is compact and $\mathcal{H}_k$ is dense in $C(\Omega)$, then

$$\mathrm{MMD}_k(\mu, \nu) = 0 \quad \Longrightarrow \quad \int f \,\mathrm{d}\mu = \int f \,\mathrm{d}\nu \quad \forall f \in C(\Omega),$$

hence $\mu = \nu$. Thus a universal kernel turns MMD into a genuine metric on probability measures.

The Coulomb kernel is the inverse Laplacian

On $\mathbb{R}^d$, $d \geq 3$, the Coulomb Green kernel is

$$\boxed{G_d(x-y) = \frac{1}{(d-2)|\mathbb{S}^{d-1}|}\,\frac{1}{\|x-y\|^{d-2}}, \qquad -\Delta G_d = \delta_0.}$$

On a closed manifold, take the zero-mean Green kernel of $-\Delta$. Since $\sigma := \mu - \nu$ has zero total mass, its Coulomb potential is

$$u_\sigma(x) = \int_M G(x,y)\,\mathrm{d}\sigma(y) = (-\Delta)^{-1}\sigma(x), \qquad -\Delta u_\sigma = \sigma.$$

The MMD quadratic form, now with a singular kernel

$$\boxed{\mathrm{MMD}_G^2(\mu,\nu) := \iint_{M\times M} G(x,y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) = \langle \sigma, (-\Delta)^{-1}\sigma\rangle,}$$

$$\boxed{2\mathcal{E}_\nu(\mu) = \mathrm{MMD}_G^2(\mu,\nu) = \int_M |\nabla u_\sigma|^2\,\mathrm{d}\,\mathrm{vol} = \|\sigma\|_{H^{-1}}^2.}$$

The energy space is $H^1(M)/\mathbb{R}$. Since G is singular, this is an extended MMD, finite only for measures of finite Coulomb energy.

Since $\delta\mathcal{E}_\nu/\delta\mu = u_\sigma$, its W_2 gradient flow is

$$\boxed{\partial_t \mu_t = \mathrm{div}(\mu_t \nabla u_{\mu_t - \nu}), \qquad -\Delta u_{\mu_t-\nu} = \mu_t - \nu.}$$

Beyond displacement convexity: PL for Coulomb flows

Main point

Global convergence can follow from an invariant PL region, without convexity of $\mathcal{E}_\nu$ along Wasserstein geodesics.

1. Local PL. On densities satisfying $\mu \geq m$ vol,

$$\frac{1}{2}|\partial \mathcal{E}_\nu|^2(\mu) = \frac{1}{2}\int_M |\nabla u_{\mu-\nu}|^2 \, \mathrm{d}\mu \geq m\, \mathcal{E}_\nu(\mu).$$

2. Preserve this region. Along $\dot{X}_t = -\nabla u_{\mu_t-\nu}(X_t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t(X_t) = \mu_t(X_t)(\nu(X_t) - \mu_t(X_t)).$$

Thus $\mu_0 \geq m_0 > 0$ and $\nu \geq m_\nu > 0$ imply $\mu_t \geq m := \min\{m_0, m_\nu\}$ while the flow remains regular.
3. Prove existence and regularity. For positive $\mu_0, \nu \in C^{0,\alpha}(M)$, the Lagrangian formulation gives local well-posedness. Elliptic regularity and a priori Hölder bounds prevent finite-time breakdown, hence the solution remains regular globally.

The local PL estimate becomes global along the flow

Energy dissipation gives

$$\boxed{\mathcal{E}_\nu(\mu_t) \leq e^{-2mt}\mathcal{E}_\nu(\mu_0).}$$

S. Boufadène and F.-X. Vialard, *SIAM Journal on Mathematical Analysis* (2025).

Mean field limit of transformers

1. Important architectures
2. Gap between statistics and optimization
3. Gradient schemes: explicit and implicit
4. Discretization bias of gradient descent
5. Convexity
6. PL condition
7. The optimization landscape of ResNets
8. The benefits of mean-field limits
9. Wasserstein gradient flows
10. **Mean field limit of transformers**

From residual attention layers to a Transformer PDE

Let $x_i^\ell \in \mathbb{R}^d$ be token i at layer ℓ and let $\mu_\ell^n = n^{-1}\sum_{i=1}^n \delta_{x_i^\ell}$. A residual attention block with step h can be written

$$x_i^{\ell+1} = x_i^\ell + h\,\Gamma_{\mu_\ell^n}(\ell h, x_i^\ell).$$

For Softmax attention with prescribed layer-dependent matrices Q_t, K_t, V_t,

$$\boxed{\Gamma_\mu(t,x) = \frac{\int_{\mathbb{R}^d} e^{\langle Q_t x, K_t y\rangle} V_t y \,\mathrm{d}\mu(y)}{\int_{\mathbb{R}^d} e^{\langle Q_t x, K_t y\rangle}\,\mathrm{d}\mu(y)}.}$$

Many layers: $h \to 0$

Depth becomes time and the tokens solve an interacting ODE:

$$\dot{x}_i(t) = \Gamma_{\mu_t^n}(t, x_i(t)).$$

Many tokens: $n \to \infty$

The empirical distribution converges to a nonlinear transport PDE:

$$\boxed{\partial_t \mu_t + \mathrm{div}(\mu_t \Gamma_{\mu_t}) = 0.}$$

Here μ_t describes token representations at depth t, not a distribution of trainable parameters.

Castin–Ablin–Carrillo–Peyré, arXiv:2501.18322 (2025).

Sampled attention and its continuous limit

Draw tokens $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \nu$ and set $\nu_n = n^{-1} \sum_{j=1}^n \delta_{X_j}$. With $A = K^\top Q$, sampled Softmax attention at a query x is

$$\boxed{f_n(x) = f_{\nu_n}(x) = \frac{\frac{1}{n} \sum_{j=1}^n e^{\langle Ax, X_j \rangle} V X_j}{\frac{1}{n} \sum_{j=1}^n e^{\langle Ax, X_j \rangle}}.}$$

Its infinite-token counterpart replaces empirical averages by integrals:

$$\boxed{f(x) = f_\nu(x) = \frac{\int e^{\langle Ax, y \rangle} V y \, \mathrm{d}\nu(y)}{\int e^{\langle Ax, y \rangle} \, \mathrm{d}\nu(y)}.}$$

Two convergence questions

Map-level error. How fast does $f_n(x) \to f(x)$, uniformly over queries $x \in B_R$?

Output-statistic error. For $\widehat{X} \sim \nu_n$, how fast does $\mathbb{E}_n[h(f_n(\widehat{X}))]$ approach $\mathbb{E}[h(f(X))]$?

The parameters are fixed: n counts tokens in one input, not training examples.

L. Bohbot, C. Letrouit, G. Peyré and F.-X. Vialard, *Token Sample Complexity of Attention*, arXiv:2512.10656 (2026).

Bounded queries: the parametric rate survives

Assume $X \sim \nu$ is centered sub-Gaussian with parameter matrix $\Sigma \succ 0$. For every fixed radius R, with probability at least $1 - \delta$ and for n sufficiently large,

$$\sup_{x \in B_R} \|f_n(x) - f(x)\|_2 \leq \frac{q(\Sigma, A, V, R, d, \delta)}{\sqrt{n}} \, e^{8R^2 \|\Sigma^{1/2} A\|_2^2},$$

where q is polynomial in the displayed parameters.

What the theorem says

- On a fixed query ball: the usual Monte Carlo rate $n^{-1/2}$ holds uniformly.
- If $\operatorname{supp} \nu \subset B_R$, Lipschitz output statistics also converge as $C(R)n^{-1/2}$.
- But $C(R)$ grows exponentially: this is weak for unbounded token distributions and large attention scores.

Why a ratio is harder than an average

Write $f_n = N_n / D_n$ and $f = N/D$. Then

$$\left\| \frac{N_n}{D_n} - \frac{N}{D} \right\|_2 \leq \frac{\|N_n - N\|_2}{D_n} + \frac{\|N\|_2 |D_n - D|}{D D_n}.$$

Numerator and denominator are controlled uniformly by symmetrization, Rademacher complexity and Dudley's entropy bound; Jensen gives $D \geq 1$.

Bohbot–Letrouit–Peyré–Vialard, Theorem 4.2, arXiv:2512.10656 (2026).

The attention horizon controls the global rate

Token anisotropy and attention strength combine into the *attention horizon*

$$H := \|\Sigma^{1/2} A \Sigma^{1/2}\|_2 = \max_{\substack{\|x\|_{\Sigma^{-1}} \le 1 \\ \|y\|_{\Sigma^{-1}} \le 1}} \langle Qx, Ky \rangle.$$

If h is L_0-Lipschitz, then with high probability

$$\left\| \mathbb{E}_n[h(f_n(\widehat{X}))] - \mathbb{E}[h(f(X))] \right\|_2 \le \frac{P(\sqrt{\log n}, L_0)}{n^{\beta(H)}}, \qquad \beta(H) = \frac{1}{2(1+32H^2)}.$$

Where the slower exponent comes from

Truncate tokens at an effective radius R:

$$\text{inside error} \lesssim \frac{e^{cR^2}}{\sqrt{n}}, \qquad \text{tail error} \lesssim e^{-c'R^2}.$$

Balancing the two terms with $R \asymp \sqrt{\log n}$ produces the horizon-dependent power $n^{-\beta(H)}$.

From averaging to extreme selection

$H \simeq 0$ gives $\beta \simeq 1/2$. As H grows, Softmax emphasizes rare extreme tokens and convergence slows. In the one-dimensional Gaussian hardmax limit, for $\bar{f}_n = n^{-1} \sum_i f_n(X_i)$,

$$\mathbb{E}|\bar{f}_n| \sim \frac{\log 4}{2\sqrt{2}} \frac{\sigma}{\sqrt{\log n}}, \qquad \mathbb{E}[\bar{f}_n^2] \sim \frac{\pi^2}{24} \frac{\sigma^2}{\log n}.$$

Bohbot–Letrouit–Peyré–Vialard, Theorem 5.3 and Proposition 5.5, arXiv:2512.10656 (2026).

Token sample complexity: conclusions and limitations

- Theory assumes i.i.d. tokens.
- Real sequences are structured and dependent.
- Trained BigBird–RoBERTa confirms the slowdown.
- Empirical rates: error $\asymp n^{-\beta}$, with $\beta \in [.26, .49]$.
- More context gives diminishing returns.

Bohbot–Letrouit–Peyré–Vialard, Section 6 and Figure 6, arXiv:2512.10656 (2026).

Takeaway messages

- Greater expressivity does not necessarily imply better performance.
- Architectural design affects optimization.
- As expressivity increases, the PL condition becomes more realistic than convexity for neural networks.
- Continuous representations can help us understand limiting regimes.