Raphaël Barboni
raphael.barboni@ens.fr

September 2024

Training infinitely deep and wide ResNets

with Conditional Optimal Transport

with G. Peyré and F-X. Vialard

Outline

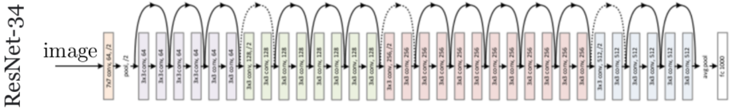


Figure: The ResNet-34 architecture (He et al., ’16)

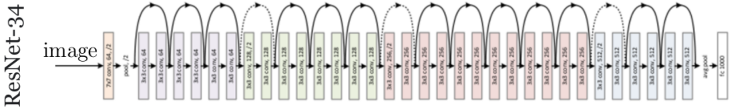


Figure: The ResNet-34 architecture (He et al., '16)

$\mathrm{F} : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ is a Neural Network (the *residual*).

Definition (Residual Neural Network (ResNet))

For parameterization $\theta = (\theta(1), ..., \theta(S)) \in \Theta^S$ and input $x \in \mathbb{R}^d$:

$$\mathrm{ResNet}_\theta(x) := x(S) \quad \text{with} \quad \begin{cases} x(0) = x \\ x(s+1) = \underbrace{x(s)}_{\text{skip connection}} + \underbrace{\frac{1}{S}\mathrm{F}_{\theta(s+1)}(x(s))}_{\text{residual}} \end{cases}$$

We consider the **infinite depth** limit $S \to +\infty$:

Definition (Neural ODE (Chen et al.'18))

For parameterization $\theta \in \Theta^{[0,1]}$ and input $x \in \mathbb{R}^d$:

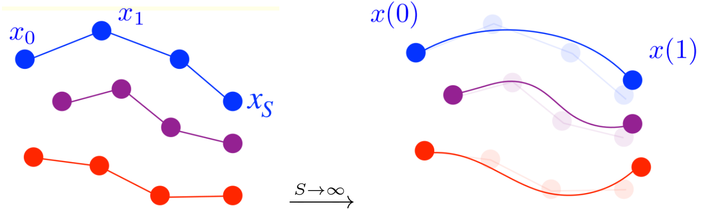
$$\mathrm{NODE}_\theta(x) := x(1) \quad \text{with} \quad \begin{cases} x(0) &= x \\ \frac{\mathrm{d}}{\mathrm{d}s} x(s) &= \mathrm{F}_{\theta(s)}(x(s)) \end{cases}$$

We consider the **infinite depth** limit $S \to +\infty$:

Definition (Neural ODE (Chen et al.'18))

For parameterization $\theta \in \Theta^{[0,1]}$ and input $x \in \mathbb{R}^d$:

$$\mathrm{NODE}_\theta(x) := x(1) \quad \text{with} \quad \begin{cases} x(0) & = & x \\ \frac{\mathrm{d}}{\mathrm{d}s} x(s) & = & \mathrm{F}_{\theta(s)}(x(s)) \end{cases}$$

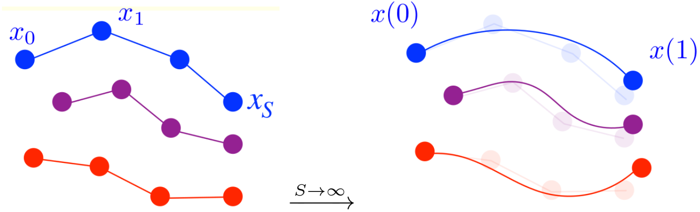


We consider the **infinite depth** limit $S \to +\infty$:

Definition (Neural ODE (Chen et al.'18))

For parameterization $\theta \in \Theta^{[0,1]}$ and input $x \in \mathbb{R}^d$:

$$\mathrm{NODE}_\theta(x) := x(1) \quad \text{with} \quad \begin{cases} x(0) &= x \\ \frac{\mathrm{d}}{\mathrm{d}s} x(s) &= \mathrm{F}_{\theta(s)}(x(s)) \end{cases}$$



→ several results about time discretization (Marion, Wu et al. '23)

(unregularized) Empirical Risk:

Let $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$ be training data samples:

$$\forall \theta \in \Theta^{[0,1]}, \quad \mathcal{L}(\theta) \coloneqq \frac{1}{2N} \sum_{i=1}^{N} \left| \mathrm{NODE}_\theta(x^i) - y^i \right|^2 .$$

(unregularized) Empirical Risk:

Let $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$ be training data samples:

$$\forall \theta \in \Theta^{[0,1]}, \quad \mathcal{L}(\theta) \coloneqq \frac{1}{2N} \sum_{i=1}^{N} \left| \mathrm{NODE}_\theta(x^i) - y^i \right|^2 .$$

Gradient Flow (GF)

For initialization $\theta_0 \in \Theta$: $\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = -\nabla \mathcal{L}(\theta_t)$

(unregularized) Empirical Risk:

Let $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$ be training data samples:

$$\forall \theta \in \Theta^{[0,1]}, \quad \mathcal{L}(\theta) := \frac{1}{2N} \sum_{i=1}^{N} \left| \mathrm{NODE}_\theta(x^i) - y^i \right|^2 .$$

Gradient Flow (GF)

For initialization $\theta_0 \in \Theta$: $\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = -\nabla \mathcal{L}(\theta_t)$

Question

Does GF find $\theta^* \in \arg\min_{\theta \in \Theta^{[0,1]}} \mathcal{L}(\theta)$?

→ minimization of a **non-convex** and **non-coercive** loss in **high dimension**.

- $\Omega = \{(u, w, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\}$ is the space of weights,
- $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation

- $\Omega = \{(u, w, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\}$ is the space of weights,
- $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation

A **Single Hidden Layer (SHL) Perceptron** of width M is defined as:

$$\mathrm{F}_{\{(u_i,w_i,b_i)\}_{i=1}^M}(x) = \frac{1}{M}\sum_{i=1}^{M} u_i \sigma(w_i^\top x + b_i)$$

- $\Omega = \{(u, w, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\}$ is the space of weights,
- $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation

A **Single Hidden Layer (SHL) Perceptron** of width M is defined as:

$$\begin{aligned} \mathrm{F}_{\{(u_i,w_i,b_i)\}_{i=1}^M}(x) &= \frac{1}{M}\sum_{i=1}^{M} u_i \sigma(w_i^\top x + b_i) \\ &= \int_\Omega u\sigma(w^\top x + b)\mathrm{d}\hat{\nu}(u,w,b), \quad \hat{\nu} = \frac{1}{M}\sum \delta_{(u_i,w_i,b_i)}. \end{aligned}$$

- $\Omega = \{(u, w, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\}$ is the space of weights,
- $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation

A **Single Hidden Layer (SHL) Perceptron** of width M is defined as:

$$\begin{aligned} \mathrm{F}_{\{(u_i,w_i,b_i)\}_{i=1}^M}(x) &= \frac{1}{M}\sum_{i=1}^{M} u_i\sigma(w_i^\top x + b_i) \\ &= \int_\Omega u\sigma(w^\top x + b)\mathrm{d}\hat{\nu}(u,w,b), \quad \hat{\nu} = \frac{1}{M}\sum \delta_{(u_i,w_i,b_i)}. \end{aligned}$$

We consider **arbitrarily wide** NNs:

Definition (Mean-Field Neural Network (Chizat'18, Mei'19))

For every ν in the space $\mathcal{P}(\Omega)$ of probability measures over Ω:

$$\mathrm{F}_\nu(x) := \int_\Omega u\sigma(w^\top x + b)\mathrm{d}\nu(u,w,b).$$

Parameters are measures over $[0,1] \times \Omega$ with uniform marginal on $[0,1]$:

$$\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega) \coloneqq \left\{ \mu \in \mathcal{P}_2([0,1] \times \Omega),\ s.t.\ (\pi_s)_{\#}\mu = \mathrm{Leb}([0,1]) \right\}.$$

then for $\mathrm{d}s$-a.e. $s \in [0,1]$, $\mu(.|s) \in \mathcal{P}_2(\Omega)$.

Parameters are measures over $[0,1]\times\Omega$ with uniform marginal on $[0,1]$:

$$\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega) \coloneqq \{\mu \in \mathcal{P}_2([0,1]\times\Omega),\ s.t.\ (\pi_s)_\#\mu = \mathrm{Leb}([0,1])\}\,.$$

then for $\mathrm{d}s$-a.e. $s\in[0,1]$, $\mu(.|s)\in\mathcal{P}_2(\Omega)$.

Definition (Mean-field NODEs)

For every $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ and every input $x\in\mathbb{R}^d$:

$$\mathrm{NODE}_\mu(x) \coloneqq x_\mu(1), \quad \begin{cases} x_\mu(0) & = & x \\ \frac{\mathrm{d}}{\mathrm{d}s}x_\mu(s) & = & \mathrm{F}_{\mu(.|s)}(x_\mu(s)) \end{cases}$$

(unregularized) Empirical Risk:

Let $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$ be training data samples:

$$\forall \mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega), \quad \mathcal{L}(\mu) := \frac{1}{2N} \sum_{i=1}^{N} \left| \mathrm{NODE}_\mu(x^i) - y^i \right|^2 .$$

Question

Does GF find $\mu^* \in \arg\min \mathcal{L}$?

Outline

Conditional Optimal Transport distance, for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$:

$$\begin{aligned} \mathrm{D}^{\mathrm{COT}}(\mu, \mu')^2 &:= \int_0^1 \mathcal{W}_2(\mu(.|s), \mu'(.|s))^2 \mathrm{d}s \\ &= \min_{\substack{\gamma \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega^2) \\ \gamma(.|s) \in \Gamma(\mu(.|s), \mu'(.|s))}} \int_0^1 \int_{\Omega^2} \|\omega - \omega'\|^2 \mathrm{d}\gamma(s, \omega, \omega') \end{aligned}$$

Conditional Optimal Transport distance, for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$:

$$\mathrm{D}^{\mathrm{COT}}(\mu, \mu')^2 := \int_0^1 \mathcal{W}_2(\mu(.|s), \mu'(.|s))^2 \mathrm{d}s$$

$$= \min_{\substack{\gamma \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega^2) \\ \gamma(.|s) \in \Gamma(\mu(.|s), \mu'(.|s))}} \int_0^1 \int_{\Omega^2} \|\omega - \omega'\|^2 \mathrm{d}\gamma(s, \omega, \omega')$$

Proposition (Characterization of AC curves (analogous to the $\mathcal{W}_2$ case))

$(\mu_t)_{t>0}$ *is an absolutely continuous curve in* $(\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega), \mathrm{D}^{\mathrm{COT}})$ *iff:*

$$\partial_t \mu_t + \mathrm{div}_\omega(\mu_t v_t) = 0, \quad \textit{with} \quad v_t \in L^2(\mu_t)$$

$\rightarrow$ various applications: evolution PDEs with heterogeneities (Peszek&Poyato '22), bayesian flow matching (Chemseddine et al. '24), conditional generative modeling (Kerrigan et al. '24), ...

Numerically, backpropagation algorithm computes the **adjoint gradient** (Chen et al. '18) for every $s \in [0,1]$ and every $(u,w,b) \in \Omega$:

$$\nabla \mathcal{L}[\mu](s,(u,w,b)) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \sigma(w^\top x_\mu^i(s)+b)p_\mu^i(s) \\ \sigma'(w^\top x_\mu^i(s)+b)(u^\top p_\mu^i(s))x_\mu^i(s) \\ \sigma'(w^\top x_\mu^i(s)+b)(u^\top p_\mu^i(s)) \end{pmatrix} \quad \in \Omega$$

with the adjoint **adjoint states**:

$$p_\mu^i(s) = \nabla_{x_\mu^i(s)}\mathcal{L}$$

Numerically, backpropagation algorithm computes the **adjoint gradient** (Chen et al. '18) for every $s \in [0,1]$ and every $(u,w,b) \in \Omega$:

$$\nabla \mathcal{L}[\mu](s,(u,w,b)) := \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \sigma(w^\top x_\mu^i(s)+b)p_\mu^i(s) \\ \sigma'(w^\top x_\mu^i(s)+b)(u^\top p_\mu^i(s))x_\mu^i(s) \\ \sigma'(w^\top x_\mu^i(s)+b)(u^\top p_\mu^i(s)) \end{pmatrix} \in \Omega$$

with the adjoint **adjoint states**:

$$p_\mu^i(s) = \nabla_{x_\mu^i(s)} \mathcal{L}$$

Theorem (Conditional Wasserstein GF)

For any initialization $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ the equation:

$$\partial_t \mu_t - \mathrm{div}_\omega(\mu_t \nabla \mathcal{L}[\mu_t]) = 0 \qquad \text{(Conditional WGF)}$$

is well-posed and is a (metric) gradient flow for $\mathcal{L}$.

Definition (Polyak-Łojasiewicz inequality)

$\mathcal{L}$ satisfies a (R,m)-P-Ł inequality around $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ if:

$$\forall \mu \in B(\mu_0, R), \quad \|\nabla \mathcal{L}[\mu]\|^2_{L^2(\mu)} \geq m\mathcal{L}(\mu),$$

Definition (Polyak-Łojasiewicz inequality)

$\mathcal{L}$ satisfies a (R, m)-P-Ł inequality around $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ if:

$$\forall \mu \in B(\mu_0, R), \quad \|\nabla \mathcal{L}[\mu]\|^2_{L^2(\mu)} \geq m\mathcal{L}(\mu),$$

Proposition (Hauer, Mazon '19, Dello Schiavo et al.'23)

Assume $\mathcal{L}$ satisfies a (R, m)-P-Ł inequality around μ_0 and $\mathcal{L}(\mu_0) < \frac{mR^2}{4}$. Then if $(\mu_t)_{t \geq 0}$ is the Conditional WGF for $\mathcal{L}$:

$$\forall t \geq 0, \quad \mathcal{L}(\mu_t) \leq \mathcal{L}(\mu_0)e^{-mt}.$$

and moreover $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$.

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{2\sqrt{\mathcal{L}(\mu_t)} + \sqrt{m}\int_0^t \|\nabla \mathcal{L}[\mu_{t'}]\|_{L^2(\mu_{t'})}\mathrm{d}t'\right\} \leq 0$$

□

For the empirical risk $\mathcal{L} : \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega) \to \mathbb{R}_+$:

$$\|\nabla \mathcal{L}[\mu]\|^2_{L^2(\mu)} = \int_0^1 \int_\Omega \left\| \frac{1}{N} \sum_{i=1}^N \begin{pmatrix} \sigma(w^\top x^i_\mu(s) + b) p^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) x^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) \end{pmatrix} \right\|^2 \mathrm{d}\mu(u, w, b | s) \mathrm{d}$$

For the empirical risk $\mathcal{L} : \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega) \to \mathbb{R}_+$:

$$\|\nabla \mathcal{L}[\mu]\|^2_{L^2(\mu)} = \int_0^1 \int_\Omega \left\| \frac{1}{N} \sum_{i=1}^N \begin{pmatrix} \sigma(w^\top x^i_\mu(s) + b) p^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) x^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) \end{pmatrix} \right\|^2 \mathrm{d}\mu(u, w, b|s)\mathrm{d}$$

$$\geq \frac{1}{N^2} \int_0^1 \sum_{1 \leq i,j \leq N} (p^i_\mu(s))^\top \underbrace{K_{\mu(.|s)}(x^i_\mu(s), x^j_\mu(s))}_{\text{kernel depending on } \mu(.|s)} p^j_\mu(s) \mathrm{d}s,$$

For the empirical risk $\mathcal{L} : \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega) \to \mathbb{R}_+$:

$$\|\nabla \mathcal{L}[\mu]\|^2_{L^2(\mu)} = \int_0^1 \int_\Omega \left\| \frac{1}{N} \sum_{i=1}^N \begin{pmatrix} \sigma(w^\top x^i_\mu(s) + b) p^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) x^i_\mu(s) \\ \sigma'(w^\top x^i_\mu(s) + b)(u^\top p^i_\mu(s)) \end{pmatrix} \right\|^2 \mathrm{d}\mu(u, w, b|s)\mathrm{d}$$

$$\geq \frac{1}{N^2} \int_0^1 \sum_{1 \leq i,j \leq N} (p^i_\mu(s))^\top \underbrace{K_{\mu(.|s)}(x^i_\mu(s), x^j_\mu(s))}_{\text{kernel depending on } \mu(.|s)} p^j_\mu(s) \mathrm{d}s,$$

where for $\nu \in \mathcal{P}(\Omega)$:

$$K_\nu(x, x') \coloneqq \int_\Omega \sigma(w^\top x + b)\sigma(w^\top x' + b)\mathrm{d}\nu(u, w, b).$$

$\to$ quantitative bounds on the conditioning of K_ν for specific choices of σ and ν.

Theorem (Convergence of Conditional Wasserstein GF)

Assume $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *is s.t.:*

$$\lambda_0 := \int_0^1 \lambda_{\min} \left((K_{\mu_0(.|s)}(x^i_{\mu_0}(s), x^j_{\mu_0}(s)))_{1 \leq i,j \leq N} \right) \mathrm{d}s > 0,$$

then if $\mathcal{L}(\mu_0)$ *is "sufficiently small"* $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *and:*

$$\mathcal{L}(\mu_t) \leq e^{-C(\lambda_0/N)t} \mathcal{L}(\mu_0)$$

Theorem (Convergence of Conditional Wasserstein GF)

Assume $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *is s.t.:*

$$\lambda_0 := \int_0^1 \lambda_{\min}\left((K_{\mu_0(.|s)}(x^i_{\mu_0}(s), x^j_{\mu_0}(s)))_{1 \le i,j \le N} \right) \mathrm{d}s > 0,$$

then if $\mathcal{L}(\mu_0)$ *is "sufficiently small"* $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *and:*

$$\mathcal{L}(\mu_t) \le e^{-C(\lambda_0/N)t} \mathcal{L}(\mu_0)$$

Example

For $\sigma = \cos$, and initialization μ_0 s.t. at each $s \in [0,1]$:

$$u \sim \delta_0, \quad w \sim \mu^w, \quad b \sim \mathcal{U}([0,\pi])$$

Theorem (Convergence of Conditional Wasserstein GF)

Assume $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ *is s.t.:*

$$\lambda_0 := \int_0^1 \lambda_{\min}\left((K_{\mu_0(.|s)}(x^i_{\mu_0}(s), x^j_{\mu_0}(s)))_{1\leq i,j\leq N}\right) \mathrm{d}s > 0,$$

then if $\mathcal{L}(\mu_0)$ *is "sufficiently small"* $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ *and:*

$$\mathcal{L}(\mu_t) \leq e^{-C(\lambda_0/N)t}\mathcal{L}(\mu_0)$$

Example

For $\sigma = \cos$, and initialization μ_0 s.t. at each $s \in [0,1]$:

$$u \sim \delta_0, \quad w \sim \mu^w, \quad b \sim \mathcal{U}([0,\pi])$$

- Gaussian: $\mu^w(w) \propto \exp(-\rho\|w\|^2)$ and $\mathcal{L}(\mu_0) < Ce^{-N^{2/d}}$,
- Heavy-tail: $\mu^w(w) \propto (1+\|w\|^2)^{-(d/2+\beta)}$ and $\mathcal{L}(\mu_0) < CN^{-3-6\beta/d}$,
- Random features: $\hat{\mu}^w = \frac{1}{M}\sum \delta_{w_i}$ with $w_1, \ldots, w_M \sim \mu^w$

Theorem (Convergence of Conditional Wasserstein GF)

Assume $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ *is s.t.:*

$$\lambda_0 := \int_0^1 \lambda_{\min}\left((K_{\mu_0(.|s)}(x^i_{\mu_0}(s), x^j_{\mu_0}(s)))_{1\leq i,j\leq N}\right) \mathrm{d}s > 0,$$

then if $\mathcal{L}(\mu_0)$ *is "sufficiently small"* $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$ *and:*

$$\mathcal{L}(\mu_t) \leq e^{-C(\lambda_0/N)t}\mathcal{L}(\mu_0)$$

Example

For $\sigma = \cos$, and initialization μ_0 s.t. at each $s \in [0,1]$:

$$u \sim \delta_0, \quad w \sim \mu^w, \quad b \sim \mathcal{U}([0,\pi])$$

- Gaussian: $\mu^w(w) \propto \exp(-\rho\|w\|^2)$ and $\mathcal{L}(\mu_0) < Ce^{-N^{2/d}}$,
- Heavy-tail: $\mu^w(w) \propto (1+\|w\|^2)^{-(d/2+\beta)}$ and $\mathcal{L}(\mu_0) < CN^{-3-6\beta/d}$,
- Random features: $\hat{\mu}^w = \frac{1}{M}\sum \delta_{w_i}$ with $w_1, \ldots, w_M \sim \mu^w$

Theorem (Convergence of Conditional Wasserstein GF)

Assume $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *is s.t.:*

$$\lambda_0 := \int_0^1 \lambda_{\min} \left((K_{\mu_0(.|s)}(x^i_{\mu_0}(s), x^j_{\mu_0}(s)))_{1 \le i,j \le N} \right) \mathrm{d}s > 0,$$

then if $\mathcal{L}(\mu_0)$ *is "sufficiently small"* $\mu_t \to \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Omega)$ *and:*

$$\mathcal{L}(\mu_t) \le e^{-C(\lambda_0/N)t} \mathcal{L}(\mu_0)$$

Example

For $\sigma = \cos$, and initialization μ_0 s.t. at each $s \in [0,1]$:

$$u \sim \delta_0, \quad w \sim \mu^w, \quad b \sim \mathcal{U}([0,\pi])$$

- Gaussian: $\mu^w(w) \propto \exp(-\rho \|w\|^2)$ and $\mathcal{L}(\mu_0) < Ce^{-N^{2/d}}$,
- Heavy-tail: $\mu^w(w) \propto (1 + \|w\|^2)^{-(d/2+\beta)}$ and $\mathcal{L}(\mu_0) < CN^{-3-6\beta/d}$,
- Random features: $\hat{\mu}^w = \frac{1}{M} \sum \delta_{w_i}$ with $w_1, \ldots, w_M \sim \mu^w$

Contributions

- Proposed a **model of infinitely deep and wide ResNets** whose training is modeled by **GF for a Conditional OT metric**,

Contributions

- Proposed a **model of infinitely deep and wide ResNets** whose training is modeled by **GF for a Conditional OT metric**,
- We show this model satisfies a **(local) P-Ł property** and conclude to a **(local) convergence result**,

Contributions

- Proposed a **model of infinitely deep and wide ResNets** whose training is modeled by **GF for a Conditional OT metric**,
- We show this model satisfies a **(local) P-Ł property** and conclude to a **(local) convergence result**,

Open problems

- **Feature Learning:** no result about the feature representations learned during training,

Contributions

- Proposed a **model of infinitely deep and wide ResNets** whose training is modeled by **GF for a Conditional OT metric**,
- We show this model satisfies a **(local) P-Ł property** and conclude to a **(local) convergence result**,

Open problems

- **Feature Learning:** no result about the feature representations learned during training,
- **Global convergence:** we cannot prove that GF always succeeds in finding a global minimizer of the risk.
  $\rightarrow$ some trajectories are known to diverge but are never seen numerically...

Contributions

- Proposed a **model of infinitely deep and wide ResNets** whose training is modeled by **GF for a Conditional OT metric**,
- We show this model satisfies a **(local) P-Ł property** and conclude to a **(local) convergence result**,

Open problems

- **Feature Learning:** no result about the feature representations learned during training,
- **Global convergence:** we cannot prove that GF always succeeds in finding a global minimizer of the risk.
  $\rightarrow$ some trajectories are known to diverge but are never seen numerically...

Thanks for your attention!