Questions on global convergence of LDDMM with no regularization and ResNets
François-Xavier Vialard
Univ. Gustave Eiffel
LIGM
On global convergence of ResNets
Supervised learning

Setting: supervised learning.

Goal:

\[ G_\star = \min_\theta G(\theta) := \mathbb{E}[\|f_\theta(X) - Y\|^2], \]  

(1)

but only acess \( X, Y \) through samples: \((x_i, y_i)\).

\[ \implies \mathcal{L}(\theta) := \min_\theta \frac{1}{N} \sum_{i=1}^{N} \|f_\theta(x_i) - y_i\|^2. \]  

(2)

- Global convergence of gradient descent on \( \mathcal{L}(\theta) \), find \( \theta_\star \).
- Generalization, i.e. measure \( G(\theta_\star) - G_\star \).
Structure of $f_\theta$.

Define *Single Hidden Layer*

$$\text{SHL}_{\theta}(x) = \theta_1(\sigma(\theta_2(x))),$$ \hspace{1cm} (3)

with $\sigma(x)$ entrywise nonlinearity ($\max(0, x)$).

Deep networks

$$f_\theta(x) = \text{SHL}_{\theta_n} \circ \ldots \circ \text{SHL}_{\theta_1}(x).$$ \hspace{1cm} (4)

*ResNets*, encode residuals

$$f_\theta(x) = (\text{Id} + f_{\theta_n}) \circ \ldots \circ (\text{Id} + f_{\theta_1})(x).$$ \hspace{1cm} (5)
Structure of $f_\theta$.

Define *Single Hidden Layer*

$$\text{SHL}_{\theta}(x) = \theta_1(\sigma(\theta_2(x))),$$  \hspace{1cm} (3)

with $\sigma(x)$ entrywise nonlinearity ($\max(0, x)$).

Deep networks

$$f_\theta(x) = \text{SHL}_{\theta_n} \circ \ldots \circ \text{SHL}_{\theta_1}(x).$$  \hspace{1cm} (4)

*ResNets*, encode residuals

$$f_\theta(x) = (\text{Id} + f_{\theta_n}) \circ \ldots \circ (\text{Id} + f_{\theta_1})(x).$$  \hspace{1cm} (5)

- Very successful architecture (*Deep Residual Learning for Image Recognition*, [He et al.] $10^5$ citations)
- Resembles to an Euler integration scheme for ODE.
A glimpse at state of the art

Key info from deep learning

*Overparametrization is not harmful for generalization.*

Benefit of overparametrization,

$\mathcal{L}(\theta)$ can be made 0 on random data.

- Case of overparametrization SHL [Chizat, Bach], [Montanari et al.]
  - Represent $f_\mu(x) := \int_\theta f_\theta(x)d\mu(\theta)$
  - Show global convergence of this relaxation.

- Neural tangent kernel: [Jacot et al.] Linear regime of $f_\theta$.

- Similarly, global convergence with the last layer width, $m = \Omega(N^3)$.

$\implies$ So far, linear regime or shallow networks are treated.
Key tool for convergence
Identification of the key tool [Belkin et al.], Polyak-Lojasiewicz condition.

\[ \lambda(f(x) - f_*) \leq \frac{1}{2} \| \nabla f(x) \|^2. \]  

(6)

Example: \( \dot{x} = -\nabla f(x) \).

\[ \frac{d}{dt}(f(x) - f_*) = -\| \nabla f(x) \|^2 \leq 2\lambda(f(x) - f_*) . \]  

(7)

Therefore,

\[ f(x(t)) - f_* \leq (f(x(0)) - f_*)e^{-2\lambda t} , \]  

(8)

No need for convexity, nor Euclidean structure, applies to Riemannian manifolds.

Example: Log-Sobolev inequality and Wasserstein distance.
Key tool for convergence

Stability of PL:

Stability of PL

Let \( \varphi : \Omega \to \Omega \) be a \( C^1 \) diffeomorphism of the definition domain of \( f \), then \( \varphi^* f(y) \triangleq f \circ \varphi(y) \) satisfies \( PL(\lambda/M^2) \) if \( f \) satisfies \( PL(\lambda) \) for \( M = \sup_{x \in \Omega} \|d\varphi(x)^{-1}\| \)

PL says nothing on convergence of \( x(t) \).

Add regularity condition such as

\[
\| \nabla f(x) \|^2 \leq \beta (f(x) - f_*) ,
\]

\( \implies \) convergence towards \( x_* \in \arg \min f \).
Our set-up

Infinite depth and infinite width,

\[ \dot{q} = f_{\theta(t)}(q). \]  

(10)

with initial and final fixed layers \( A(x) = q \) and \( Bq = y \).

- Assume linearity wrt \( \theta \).
- Assume \( f_{\theta} \) lies in \( H \) RKHS.

Retains nonlinearity of deep networks.

Example: Finite dim vector space \( f_i(\cdot) \), Sobolev spaces.

Counter-example: SHL is not linear wrt hidden layer.\(^1\)

---

\(^1\)Chizat-Bach (Barron) relaxation is linear wrt parameter.
The continuous setting

**Group actions.** Let $G_V$ be a group acting on manifold $Q$.

\[ \Phi : G \times Q \to Q, \quad (g, q) \mapsto g \cdot q := \Phi_g(q). \quad (11) \]

$g_1 \cdot (g_2 \cdot q) = (g_1 g_2) \cdot q$ and $\text{Id} \cdot q = q$ for any $q \in Q$ and $g_1, g_2 \in G$.

**Infinitesimal generator**

\[ \xi_Q(q) := \left. \frac{d}{dt} \right|_{t=0} \exp(t\xi) \cdot q. \quad (12) \]

Example: $G = \text{Diff}$ and $Q = \{ (x_1, \ldots, x_n) \mid x_i \neq x_j \in \mathbb{R}^d \}$.

**Momentum map**

The map $J : T^*Q \to V^*$ defined by

\[ J(p, q)(\xi) = \langle p, \xi \cdot q \rangle \quad (13) \]

Define $\text{Ad}_h : V \to V$ (and $\text{Ad}_h^*$ by duality) by

\[ \text{Ad}_h(\xi) := h \cdot \xi \cdot h^{-1}. \quad (14) \]
Analytical setup

\[ \partial_t \varphi(t, x) = \xi(t, \varphi(t, x)) \quad (15) \]
\[ \varphi(0, x) = x \; \forall x \in D, \quad (16) \]
\[ \xi \in V \hookrightarrow W^{1,\infty}(D, \mathbb{R}^d). \]
\[ \text{Fl}_1(\xi) = \varphi(1) \text{ where } \varphi \text{ solves (15)}, \quad (17) \]

define
\[ \mathcal{G}_V := \{ \varphi(1) : \exists \xi \in L^2([0, 1], V) \text{ s.t. Fl}_1(\xi) \}. \quad (18) \]

\[ \text{dist}(\psi_1, \psi_0)^2 = \inf \left\{ \int_0^1 \|\xi\|_V^2 \, dt : \xi \in L^2([0, 1], V) \text{ s.t. } \psi_1 = \text{Fl}_1(\xi) \circ \psi_0 \right\} \quad (19) \]

\( \mathcal{G}_V \) is complete [Trouvé].
Examples of actions

- $G_V \times [\mathbb{R}^d]^N \mapsto [\mathbb{R}^d]^N$ by composition $x_i \rightarrow \varphi(x_i)$.
- $G_V \times \text{Dens}(\mathbb{R}^d) \mapsto \text{Dens}(\mathbb{R}^d)$, $\varphi \cdot \mu = \varphi_{\#}(\mu)$.
- $G_V \times \text{Func}(\mathbb{R}^d) \mapsto \text{Func}(\mathbb{R}^d)$, $\varphi \cdot I = I \circ \varphi^{-1}$.

- $J(p, q) = \sum_{i=1}^{N} p_i \delta_{q_i}$.
- $J(p, \mu) = \mu \nabla p$ and $\langle J(p, q), \zeta \rangle = \int \langle \nabla p(x), \zeta(x) \rangle d\mu(x)$.
- $J(\mu, q) = (\nabla q)\mu$.

$$\|J(p, q)\|_{V^*}^2 = \sum_{i,j} p_i K(x_i, x_j) p_j.$$  \hspace{1cm} (20)

Equivalent norm with $\sum_i p_i^2$ if kernel matrix well-conditioned.
The goal

The loss is

\[ \ell(v) = \sum_{i=1}^{N} |\varphi(1)(x_i) - y_i|^2 \]

Is it possible to get a global minimum with gradient descent for almost every initialization?
Compute the gradient

Gradient of $\mathcal{L}$

$$D\mathcal{L}(\zeta)(\eta) = \int_0^1 \langle J(p, q), \eta \rangle \, dt,$$  \hspace{1cm} (21)

whith $p, q$ satisfying

$$\begin{cases} 
\dot{p} = -d\zeta^\top(q)(p) \\
\dot{q} = \zeta(q),
\end{cases}$$  \hspace{1cm} (22)

and initial conditions $p(1) = -\partial_q \ell(q(1))$.\(^a\)

\[^a\ell(q) = \sum \|B(q_i(1)) - y_i\|^2.\]

\[\implies \text{possible to integrate: } J(p(t), q(t)) = \text{Ad}_{g(t) \cdot g(1)^{-1}}(J(p(1), q(1))).\]

But, $p(1) = B^*(B(q(1)) - y)$ and therefore, $\ell(q) = \frac{1}{2} \|p(1)\|_{[BB^*]^{-1}}^2$.\]
Local PL condition

Local PL
Assuming $K$ the kernel of $V$ satisfies $\lambda(D, \delta) \text{Id} \preceq K(x_i, x_j) \preceq \Lambda(D, \delta) \text{Id}$. Then, a local PL is satisfied, on $B(R)$ in $L^2([0, 1], V)$, one has

\[
\begin{align}
c \ell(\xi) &\leq 2MRe^R \| \nabla \ell(\xi) \|^2 \\
\| \nabla \ell(\xi) \|^2 &\leq 2MCRe^R \ell(\xi).
\end{align}
\]

- All critical points are global.
- If loss is small enough, global convergence.
- If iterates are bounded, then global convergence.

Open question: global convergence.
Open questions

- Global convergence almost sure in initialization with no regularization
- What about convergence in the regularized case: "weight decay."

\[
\min \int_{0}^{1} \|\xi\|_{V}^{2} dt + \mathcal{L}(\varphi(x), y). \tag{25}
\]

- What about generalization?