Questions on global convergence of LDDMM with no regularization and ResNets

François-Xavier Vialard
Univ. Gustave Eiffel
LIGM

Contents

1 On global convergence of ResNets

Supervised learning

Setting: supervised learning.
Goal:

$$G_\star = \min_\theta \mathcal{G}(\theta) := \mathbb{E}[\|f_\theta(X) - Y\|^2], \tag{1}$$

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

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

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

Structure of f_θ.

Define *Single Hidden Layer*

$$\mathrm{SHL}_\theta(x) = \theta_1(\sigma(\theta_2(x)))\,, \tag{3}$$

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

$$f_\theta(x) = \mathrm{SHL}_{\theta_n} \circ \ldots \circ \mathrm{SHL}_{\theta_1}(x)\,. \tag{4}$$

ResNets, encode residuals

$$f_\theta(x) = (\mathrm{Id} + f_{\theta_n}) \circ \ldots \circ (\mathrm{Id} + f_{\theta_1})(x)\,. \tag{5}$$

Structure of f_θ.

Define *Single Hidden Layer*

$$\mathrm{SHL}_\theta(x) = \theta_1(\sigma(\theta_2(x)))\,, \tag{3}$$

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

$$f_\theta(x) = \mathrm{SHL}_{\theta_n} \circ \ldots \circ \mathrm{SHL}_{\theta_1}(x)\,. \tag{4}$$

ResNets, encode residuals

$$f_\theta(x) = (\mathrm{Id} + f_{\theta_n}) \circ \ldots \circ (\mathrm{Id} + f_{\theta_1})(x)\,. \tag{5}$$

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

$\Longrightarrow$ So far, linear regime or shallow networks are treated.

Key tool for convergence

Identification of the key tool [Belkin et al.], Polyak-Lojasiewicz condition.

$$\lambda(f(x) - f_*) \leq \frac{1}{2}\|\nabla f(x)\|^2 \,. \tag{6}$$

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

$$\frac{d}{dt}(f(x) - f_*) = -\|\nabla f(x)\|^2 \leq 2\lambda(f(x) - f_*) \,. \tag{7}$$

Therefore,

$$f(x(t)) - f_* \leq (f(x(0)) - f_*)e^{-2\lambda t} \,, \tag{8}$$

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

$$\|\nabla f(x)\|^2 \leq \beta(f(x) - f_*)\,, \tag{9}$$

$\implies$ convergence towards $x_\star \in \arg\min f$.

Our set-up

Infinite depth and infinite width,

$$\dot{q} = f_{\theta(t)}(q)\,. \tag{10}$$

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

- Assume linearity wrt θ.
- Assume f_θ 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.

$$\Phi : G \times Q \to Q\,, \quad (g, q) \mapsto g \cdot q := \Phi_g(q). \tag{11}$$

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

Infinitesimal generator

$$\xi_Q(q) := \frac{d}{dt}\bigg|_{t=0} \exp(t\xi) \cdot q\,. \tag{12}$$

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

Momentum map

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

$$J(p, q)(\xi) = \langle p, \xi \cdot q \rangle \tag{13}$$

Define $\mathrm{Ad}_h : V \mapsto V$ (and Ad_h^* by duality) by

$$\mathrm{Ad}_h(\xi) := h \cdot \xi \cdot h^{-1}\,. \tag{14}$$

Analytical setup

$$\partial_t \varphi(t,x) = \xi(t,\varphi(t,x)) \tag{15}$$

$$\varphi(0,x) = x \ \forall x \in D\,, \tag{16}$$

$\xi \in V \hookrightarrow W^{1,\infty}(D,\mathbb{R}^d)$.

$$\mathrm{Fl}_1(\xi) = \varphi(1) \text{ where } \varphi \text{ solves (15)}\,, \tag{17}$$

define

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

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

$\mathcal{G}_V$ is complete [Trouvé].

Examples of actions

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

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

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

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 $\mathcal{L}$

$$D\mathcal{L}(\xi)(\eta) = \int_0^1 \langle J(p,q), \eta \rangle \,\mathrm{d}t\,, \tag{21}$$

whith p, q satisfying

$$\begin{cases} \dot{p} = -\,\mathrm{d}\xi^\top(q)(p) \\ \dot{q} = \xi(q)\,, \end{cases} \tag{22}$$

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

---

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

$\Longrightarrow$ possible to integrate: $J(p(t), q(t)) = \mathrm{Ad}^*_{g(t)\cdot g(1)^{-1}}(J(p(1), q(1)))$.
But, $p(1) = B^*(B(q(1)) - y)$ and therefore, $\ell(q) = \frac{1}{2}\|p(1)\|^2_{[BB^*]^{-1}}$.

Local PL condition

Local PL

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

$$c\ell(\xi) \leq 2MRe^R\|\nabla\ell(\xi)\|^2 \tag{23}$$

$$\|\nabla\ell(\xi)\|^2 \leq 2MCRe^R\ell(\xi)\,. \tag{24}$$

- 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?