CCIMI Literature Review
On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport
(Lénaïc Chizat and Francis Bach 2018)

Parley Ruogu Yang & Harry Goulbourne

UNIVERSITY OF CAMBRIDGE
Faculty of Mathematics

16 Nov 2020

Cambridge, UK

Outline of the presentation and Remarks

1. Motivation and Overview (PY)
2. Gradient Flow (G.F.) (HG)
3. Convergence theorem (PY)
4. Examples (HG & PY)
5. Conclusion (HG)
6. Q & A (HG & PY)

Remark: we inherit all numbering of results from the paper. Many of the analytical results and precise assumptions, unless necessary, are omitted due to the time constraint. One could further engage with those by reading the particular parts of the paper.
Requests to clarify the notations are welcomed.

Motivation from neural network

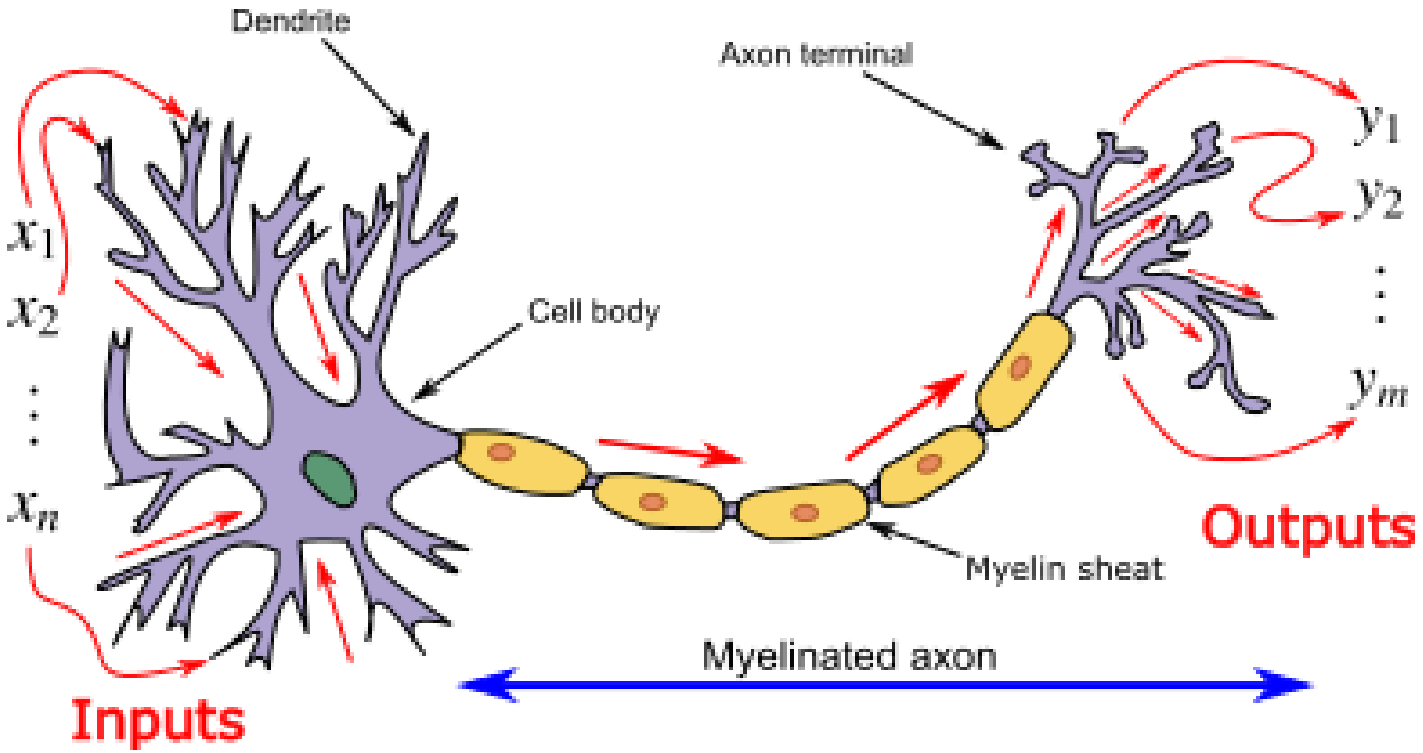


Figure: Neural Network. Courtesy: Wikipedia.

$$\Phi(\theta) = \sigma(\theta \cdot x)$$

where $\theta, x \in \mathbb{R}^d$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$.
Then we have a parameterised set $\{\Phi(\theta)\}_{\theta \in \Theta} \subset \mathcal{F}$

Training example: Consider $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ to be squared or logistic loss, we may have the expected risk $\forall f \in \mathcal{F}$,

$$R(f) = \int l(f(x), y) d\rho(x, y)$$

Consider a minimisation problem over the set of non-negative finite measures on a domain $\Omega \subset \mathbb{R}^d$, denoted $M_+(\Omega)$.

$$F^* = \min_{\mu \in M_+(\Omega)} F(\mu)$$

where F is defined as

$$F(\mu) = R\left(\int \Phi d\mu\right) + \int V d\mu$$

The first term is the usual smooth and convex loss function and the second term is the convex regulariser.

⏩ Rigorous definition as per Normed Space

Particle gradient flow

$u \in \Omega^m$ of positions for m particles,

$$F_m(u) = R\left(m^{-1}\sum_{i\in[m]}\Phi(u_i)\right) + m^{-1}\sum_{i\in[m]}V(u_i)$$

With the sub-differential of $F_m(u)$ [1], we define the **particle gradient flow** as an absolutely continuous path $u : [0, +\infty) \to \Omega^m$ s.t.

$$u'(t) \in -m\partial F_m(u(t)) \quad \forall t \geq 0 \ a.s.$$

We then construct

$$\mu_{m(t)} := \mu_{m,t} := m^{-1}\sum_{i\in[m]}\delta_{u_i(t)}$$

where δ is a Dirac mass.

[1] $\partial f(u) := \{p | f(u) \geq f(u_0) + p \cdot (u - u_0) + \mathcal{O}(u - u_0)\}$

Overview

- Section 2:
  - Proposition 2.3: what is $u_i'(t)$ and particularly how it relates to $\Phi(u)$
  - Theorem 2.6: what is the behaviour of μ_m as $m \to \infty$?
- Section 3:
  - Theorem 3.3 / 3.5: is it the case that $F(\mu_{m,t}) \to F^*$ as $m, t \to \infty$?
- Section 4: example on Neural networks

Assumptions

- F is a separable Hilbert space, $\Omega \subset \mathbb{R}^d$ is the closure of a convex open set, and
- (smooth loss) $R : \mathcal{F} \to \mathbb{R}_+$ is differentiable, with a differential dR that is Lipschitz on bounded sets and bounded on sublevel sets,
- (basic regularity) $\Phi : \Omega \to \mathcal{F}$ is (Fréchet) differentiable, $V : \Omega \to \mathbb{R}_+$ is semiconvex, and
- (locally Lipschitz derivatives with sublinear growth) there exists a family $(Q_r)_{r>0}$ of nested nonempty closed convex subsets of Ω such that:
  - (a) $\{u \in \Omega; \operatorname{dist}(u, Q_r) \leq r'\} \subset Q_{r+r'}$ for all $r, r' > 0$
  - (b) Φ and V are bounded and $d\Phi$ is Lipschitz on each Q_r, and
  - (c) there exists $C_1, C_2 > 0$ such that $\sup_{u \in Q_r} (\|d\Phi_u\| + \|\partial V(u)\|) \leq C_1 + C_2 r$ for all $r > 0$ where $\|\partial V(u)\|$ stands for the maximal norm of an element in $\partial V(u)$

Existence and uniqueness of particle gradient flow

To each F_m and initial position vector $u(0)$ there exists a unique particle gradient flow. Further, one can describe the rate of improvement of loss and the velocity of each particle under the gradient flow.

For any initialisation $u(0)$ there exists a unique particle gradient flow $u : \mathbb{R} \to \Omega^m$ for F_m. Moreover, for almost every $t \geq 0$ it holds that:

- $\frac{d}{ds} F_m(u(s))\big|_{s=t} = -|u'(t)|^2$
- $u_i'(t) = v_t(u_i(t))$

Where for $u \in \Omega$ and $\mu_{m,t} := \frac{1}{m} \sum_{i=1}^{n} \delta_{u_i(t)}$,
$v_t(u) = \tilde{v}_t(u) - proj_{\partial V(u)}(\tilde{v}_t(u))$, with
$\tilde{v}_t(u) = -\left[\left\langle R'\left(\int \Phi \mathrm{d}\mu_{m,t}\right), \partial_j \Phi(u)\right\rangle\right]_{j=1}^{d}$

When V is differentiable, we have $v_t(u) = \tilde{v}_t(u) - \nabla V$

Existence and uniqueness of particle gradient flow

As the sum of a continuously differentiable and a semiconvex function, F_m is locally semiconvex and the existence of a unique gradient flow on a maximal interval $[0, T]$ with the claimed properties is standard. Now, a general property of gradient flows is that for a.e $t \in \mathbb{R}_+, u \in \Omega$, the derivative is (minus) the subgradient of minimal norm. This leads to the explicit formula involving the velocity field with pointwise minimal norm:

$$\begin{aligned} v_t(u) &= \arg\min \left\{ |v|^2; \tilde{v}_t(u) - v \in \partial V(u) \right\} \\ &= \tilde{v}_t(u) - \arg\min \left\{ |\tilde{v}_t(u) - z|^2 ; z \in \partial V(u) \right\} \\ &= \left(\mathrm{id} - \mathrm{proj}_{\partial V(u)} \right) (\tilde{v}_t(u)) \, . \end{aligned}$$

Wasserstein metric

The 2-**Wasserstein distance** between two probability measures $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^d\right)$ is defined as

$$W_2(\mu,\nu) := \left(\inf \int |y-x|^2 d\gamma(x,y)\right)^{1/2}$$

Where the infimum is taken over all γ having marginals μ and ν.

Wasserstein gradient flow - motivation

Eventually we would like to take a "many particle limit" of a particle gradient flow. This would mean generalising the particle gradient flow to arbitrary measure-valued initialisations, not just atomic ones. To do this we introduce the Wasserstein gradient flow. What properties should it satisfy?

- The evolution of a time-dependent measure (μ_t) under instantaneous velocity fields (v_t) satisfies the *continuity equation* $\partial_t \mu_t = -\text{div}(v_t \mu_t)$.

Wasserstein gradient flow - motivation

- The differential of F evaluated at $\mu \in \mathcal{M}(\Omega)$ is represented by the function $F'(\mu) : \Omega \to \mathbb{R}$ defined as

$$F'(\mu)(u) := \left\langle R'\left(\int \Phi \mathrm{d}\mu\right), \Phi(u)\right\rangle + V(u)$$

Thus v_t (as defined before) is simply a field of (minus) subgradients of $F'(\mu_{m,t})$. We write this relation $v_t \in -\partial F'(\mu_{m,t})$. The set $\partial F'$ is called the Wasserstein subdifferential of F, as it can be interpreted as the subdifferential of F relative to the Wasserstein metric on $\mathcal{P}_2(\Omega)$.

Wasserstein gradient flow - definition

A **Wasserstein gradient flow** for the functional F on a time interval $[0,T]$ is an absolutely continuous path $(\mu_t)_{t\in[0,T]}$ in $\mathcal{P}_2(\Omega)$ that satisfies, distributionally on $\left[0,T\right]\times\Omega^d$

$$\partial_t\mu_t = -\operatorname{div}\left(v_t\mu_t\right) \qquad \text{where} \qquad v_t \in -\partial F'\left(\mu_t\right).$$

$F'(\mu)(u) := \left\langle R'\left(\int \Phi \mathrm{d}\mu\right), \Phi(u)\right\rangle + V(u)$ can be interpreted as the subdifferential of F relative to the Wasserstein metric on $\mathcal{P}_2(\Omega)$.

Wasserstein gradient flow - Existence and uniqueness

If $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on a set $Q_{r_0} \subset \Omega$, then there exists a unique Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ for F starting from μ_0.

Wasserstein gradient flow as a limit

This is a proper generalization of the gradient flow for an atomic measure since, whenever $(\mathbf{u}(t))_{t\geq 0}$ is a particle gradient flow for F_m, then $t \mapsto \mu_{m,t} := \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{u}_i(t)}$ is a Wasserstein gradient flow for F.

Many-particle limit

Consider $(t \mapsto \mathbf{u}_m(t))_{m\in\mathbb{N}}$ a sequence of classical gradient flows for F_m initialized in a set $Q_{r_0} \subset \Omega$. If $\mu_{m,0}$ converges to some $\mu_0 \in \mathcal{P}_2(\Omega)$ in the Wasserstein distance W_2, then $(\mu_{m,t})_t$ converges, as $m \to \infty$, to the unique Wasserstein gradient flow of F starting from μ_0.

Aim

Let the domain be $\Omega = \mathbb{R}^d$, $d \geq 2$ with smoothness, sphere-separability of the support of G.F. and regularity of F assumed.

Theorem 3.3

$$\mu_t \xrightarrow[t\to\infty]{W_2} \mu_\infty \implies F(\mu_{m,t}) \xrightarrow[m,t\to\infty]{|\cdot|_{\mathbb{R}}} F^*$$

A key Lemma

Point: to clarify what we want to achieve, i.e. the limit

$$\lim_{m,t\to\infty} F(\mu_{m,t}) = F^* \tag{1}$$

and to outline a step further.

Lemma C.15

Let (μ_t) be a G.F. which initialisation is on a set $Q_{r0} \subset \Omega$, convex and closed such that $F(\mu_t) \to F^*$. If $(\mu_{0,m})_m$ is a sequence of measures concentrated on a set Q_{r0} that converges to μ_0 in W_2, then

$$\lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = F^* = \lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t})$$

Back to the initial definition

Proof

Lemma A.1

F is continuous for W_2

Fun fact by construction

$t \mapsto F(\mu_{m,t})$ is decreasing

1. First equation follows from A.1 directly.
2. Second equation can be proved by a two-way sandwich. To demonstrate on the whiteboard.

White Board

Definitions and constructions

Back to the first Lemma

Definition (Norm and Weak convergence)

We work with $(M(\Omega), ||\cdot||_{BL})$ where[a]

$$||\mu||_{BL} := \sup_{\phi \in B_\infty(0,1)} \int \phi d\mu$$

$\mu_n \rightharpoonup \mu \in M(\Omega)$ if $\forall \phi : \mathbb{R}^d \to \mathbb{R}$ who's continuous and bounded,

$$\int \phi d\mu_n \to \int \phi d\mu$$

[a] $\phi : \mathbb{R}^d \to \mathbb{R}$ and some Lipschitz conditions are ignored here

Way towards the remaining proof

Construct an operator $h : M_+(\Omega) \to M_+(\mathbb{S}^{d-1})$ which, $\forall \phi : \mathbb{S}^{d-1} \to \mathbb{R}$, [2]

$$\int_{\mathbb{S}^{d-1}} \phi(\theta) dh(\mu)(\theta) = \int_{\mathbb{R}^d} |u|^2 \phi(\frac{u}{|u|}) d\mu(u)$$

Theorem C.16

If $h(\mu_t) \rightharpoonup \nu \in \mathbb{S}^{d-1}$, then ν is the global minimiser and

$$\lim_{t \to \infty} F(\mu_t) = F^*$$

◂ This is the final step

[2] Convention is in place for $\phi(\frac{0}{0}) = 0$

Some key results to use

Lemma C.3

If $h(\mu_t) \rightharpoonup \nu \in \mathbb{S}^{d-1}$, then $F'(\nu)$ vanishes $\nu - a.e.$

Proposition 3.1

$\mu \in M_+(\Omega)$ minimises F if and only if
$F'(\mu) \geq 0$ and $F'(\mu)(u) = 0 \quad \forall u \in \Omega \;\; \mu - a.e.$

Comment about what's being left in the proof.
Now we prove by contradiction that $F'(\nu) \geq 0$

Proof by contradiction

Proposition C.1

$\mu \in M_+(\Omega)$ s.t. $F'(\mu) < 0$. Then $\exists \varepsilon > 0, A \subset \Omega$ s.t.

- IF G.F. (μ_t) satisfies $h(\mu_{t_1}) \in B_{BL}(h(\mu), \varepsilon)$ with $\mu_{t_1}(A) > 0$ for some $t_1 \geq 0$
- THEN $\exists t_2 > t_1$ s.t. $h(\mu_{t_2}) \notin B_{BL}(h(\mu), \varepsilon)$

Obtain ε, A as above. Then by $h(\mu_t) \rightharpoonup \nu$ we can find T s.t. $h(\mu_t) \in B_{BL}(\nu, \varepsilon) \forall t \geq T$. But by above we have [3] $t' > T$ s.t. $h(\mu_{t'}) \notin B_{BL}(\nu, \varepsilon)$. Therefore contradiction. □

[3] The requirement on the set A can be checked using C.10 and a complete statement of C.1.

Section 4.2: Neural Networks

Back to the first motivation

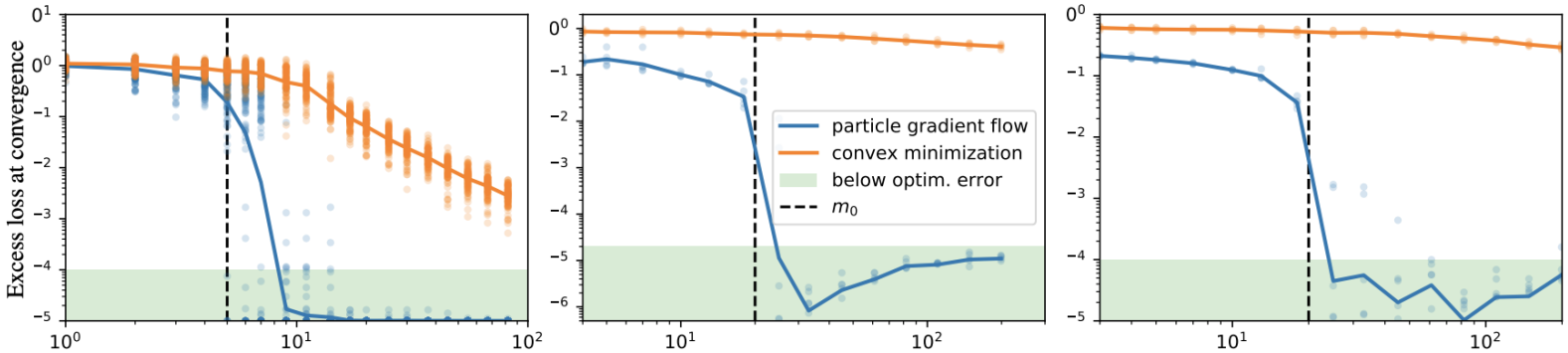
- Sigmoid activation $\sigma(s) = (1 + e^{-s})^{-1}$ and regularisation $V(w, \theta) = |w|$
- Relu activation $\sigma(s) = \max\{0, s\}$, and regularisation $V(w, \theta) = |w| \cdot |\theta|$
  - Further variations: $\Phi(\theta) = \sigma(s(\theta) \cdot x)$ where $s(\theta_i) = \theta_i |\theta_i|$, and $V(\theta) = |\theta|^2$

Proposition 4.2 & 4.3

In any of the three settings, if the Wasserstien G.F. of F converges in W_2 to μ_∞, then μ_∞ is a global minimiser of F.

Animation: `https://lchizat.github.io/PGF.html`

Empirical particle complexity



(a) Sparse deconvolution ($d = 1$) (b) ReLU activation ($d = 100$) (c) Sigmoid activation ($d = 100$)

Figure 3: Comparison of particle-complexity for particle gradient flow and convex minimization on a fixed grid: excess loss at convergence vs. number of particles. Simplest minimizer has m_0 particles.

Summary

1. Introduce the particle gradient flow and study the many-particle limit (a Wasserstein gradient flow).
2. Under suitable assumptions, if the Wasserstein gradient flow converges then it converges to a global minimiser of F.
3. Apply the results to training a neural network with a single hidden layer and ReLU activation function.
4. Numerical results show that the asymptotic behaviour of the particle gradient flow can be seen at fairly small m.