Notions in Optimal Transport for Sigmoid Neural Networks

A beginners' analysis of: *"On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport"* - Chizat, Bach

Simone Maria Giancola[1]

[1]Bocconi University, Milan, Italy

Real Analysis II, Bocconi University, January 2023

Lecture Contents

Lecture Path

Content

- Mostly an exploration of the results of [CB18]
- Also a video presentation of the publication [Ins19] and two blog posts made by the authors [Bac20a; Chi20]

Content

- The focus is on two layer sigmoid neural networks, and all the theoretical results needed to understand them.
- Ideally, a sufficient explanation for a beginner
- The doc at this link has the proofs, a wide Appendix section and lots of references (80 pages)

Boxes I

This is a definition

Here I define something

This is a theorem

Something is gnihtemoS backwards

This is an assumption

assumptions are purple boxes

A remark an observation or an example

for example, I observe or remark that this is an observation

Partial Notation

- in $\mathbb{R}^d$ scalar products $\cdot$, norms $|\cdot|$
- in a Hilbert space $\mathcal{F}$ scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$
- norms of nonlinear operators $\|\cdot\|$
- differential of f at x as df_x
- $\mathcal{M}(\mathbb{R}^d)$ the set of finite signed Borel measures on $\mathbb{R}^d$
- δ_x a dirac mass at x
- $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures endowed with Wasserstein distance:

Symbols and colors instead of proofs

Some parts are advanced, and even the 80 pages document avoids the discussion. For the sake of the presentation, tecnnical aspects are left aside, instead we use:

- ☺ means good for what we want to do
- ☹ means bad for what we want to do

Symbols and colors instead of proofs

Some parts are advanced, and even the 80 pages document avoids the discussion. For the sake of the presentation, tecnnical aspects are left aside, instead we use:

- ☠ means difficult, overlooked, taken as granted

Symbols and colors instead of proofs

Some parts are advanced, and even the 80 pages document avoids the discussion. For the sake of the presentation, tecnnical aspects are left aside, instead we use:

- orange to highlight things that are connected in the exposition

A motivating example

Consider a dataset of images where $\mathcal{Y} = \{-1, 1\} \rightsquigarrow \{dogs, cats\}$. The sizes usually exceed $n, d > 10^6$. A **neural network** (NN) is implemented. It could be described as a **nonlinear** predictor with general form:

$$h(x, \theta) = \theta_l^T \sigma(\theta_{l-1}^T \sigma(\ldots \sigma(\theta_2^T \sigma(\theta_1^T x)))$$

Where l denotes the number of layers before the output and σ is a nonlinearity (e.g. a sigmoid). Observe that the nonlinearity is in the parameters in this case.

Cats VS Dogs NN visualized

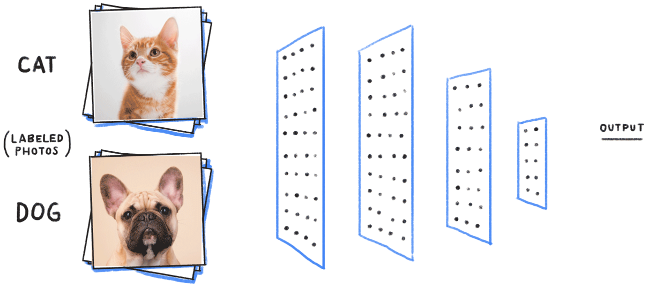


Figure: Idealized Animation of a simple Neural Network. Source Github

Solving Cats VS Dogs

Assume our data sample is a collection of pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}$ and $y_i \in \mathcal{Y} \subset \mathbb{R}$. The two signals come from an unknown distribution $\rho(x, y)$. We aim to build a prediction function $h : \mathbb{R}^{d-2} \times \mathbb{R}^{d-1} \to \mathbb{R}$ parametrized by $\theta \in \mathbb{R}^{d-1}$. Such function $h(\cdot, \theta)$ is fitted against:

Regularized Empirical Risk Minimization

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i, \theta)) + \lambda \Xi(\theta) \quad (1)$$

Where:

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function
- $\Xi : \mathbb{R}^{d-1} \to \mathbb{R}_+$ is an (optional) regularization function
- λ (optional) is a *Lagrange coefficient*

Mimicking the "world" of Cats VS Dogs

Since we observe a sample $\mathcal{D}$ of the underlying distribution $\rho(x, y)$ what we actually wish to mimic is a minimization of the test error wrt θ.

Expected Risk

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\Big[\ell(y, h(x, \theta))\Big] \tag{2}$$

which is **in most reasonable cases** convex by the convexity of ℓ. Here, $\mathcal{F}$ is a Hilbert space[a]

[a] Complete wrt to the distance induced by an inner product

This problem is **convex** in the function but **non convex** in the parameters!

Mimicking the "world" of Cats VS Dogs

Since we observe a sample $\mathcal{D}$ of the underlying distribution $\rho(x, y)$ what we actually wish to mimic is a minimization of the test error wrt θ.

Expected Risk

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\Big[\ell(y, h(x, \theta))\Big] \tag{3}$$

which is **in most reasonable cases** convex by the convexity of ℓ. Here, $\mathcal{F}$ is a Hilbert space[a]

[a] Complete wrt to the distance induced by an inner product

This problem is **convex** in the function but **non convex** in the parameters!

Mimicking the "world" of Cats VS Dogs

Since we observe a sample $\mathcal{D}$ of the underlying distribution $\rho(x, y)$ what we actually wish to mimic is a minimization of the test error wrt θ.

Expected Risk

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\Big[\ell(y, h(x, \theta))\Big] \tag{4}$$

which is **in most reasonable cases** convex by the convexity of ℓ. Here, $\mathcal{F}$ is a Hilbert space[a]

[a] Complete wrt to the distance induced by an inner product

This problem is **convex** in the function but **non convex** in the parameters!

Linear VS nonlinear

A plethora of research questions have been solved when considering linear models of the form $h(x, \theta) = \theta^T \Phi(x)$

- Theory and practice meld together beautifully
- Gradient Descent and faster techniques lead to satisfactory results

Linear VS nonlinear

A plethora of research questions have been solved when considering linear models of the form $h(x, \theta) = \theta^T \Phi(x)$

- Theory and practice meld together beautifully
- Gradient Descent and faster techniques lead to satisfactory results

This is not happening in nonlinear parametric optimization, where the optimization is non convex. Gradient descent suffers from many issues, including but not limited to:

- stationary points
- local minima
- plateaux
- bad initialization

Results in the nonlinear setting

There are local guarantees [Jin+18; Lee+], but global efficient convergence is **impossible to prove a priori**. Some results up to **very strong** assumptions are:

- Most local minima are equivalent [Cho+15]
- no spurrious local minima [SJL22]
- other results up to different assumptions [JK17]

Why and What in one slide

- Neural Networks proved to be instrumental for hard tasks where linear models do not perform well, and open the door to higher flexibility in terms of model design.

Why and What in one slide

- A theoretical work on one of the simplest models will be analyzed. We will see how **two layer sigmoid neural networks** of the form

$$\phi(\theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \theta_{d-1}\right)$$

fall under the umbrella of a much broader class of optimization problems which has global optimization guarantees up to conditions to be specified.

Why and What in one slide

- Such results are achieved thanks to techniques involving Wasserstein Gradient Flows, a *byproduct of Optimal Transport* [CB18].

Recap

The problem of (1)

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i, \theta)) + \lambda \Xi(\theta)$$

seen as the empirical version for a sample $\mathcal{D}$ from a distribution ρ as (2):

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\Big[\ell(y, h(x, \theta))\Big] \tag{5}$$

is *difficult* but *interesting* for nonlinear parametric functions such as Sigmoid NNs $\phi(\theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \theta_{d-1}\right)$ but:

Recap

The problem of (1)

$$\theta^* = \underset{\theta \in \mathbb{R}^{d-1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i, \theta)) + \lambda \Xi(\theta)$$

seen as the empirical version for a sample $\mathcal{D}$ from a distribution ρ as (2):

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\left[\ell(y, h(x, \theta))\right] \tag{5}$$

is *difficult* but *interesting* for nonlinear parametric functions such as Sigmoid NNs $\phi(\theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \theta_{d-1}\right)$ but:

- we need to understand how [CB18] describes them and under which principles
- we do not now why this holds

Lecture Path

Functional Optimization Perspective

We save our discussion on Neural Networks for the last section and focus on a functional optimization problem. Informally:

- Instead of minimizing in terms of parameters, we minimize in terms of functions arising from parameters using $R : \mathcal{F} \to \mathbb{R}_+$
- A solution will be a combination of elements from the parametric space $\{\phi(\theta)\}_{\theta \in \Theta} \subset \mathcal{F}$.

Later we will show why this is **reasonable**.

Functional Optimization Perspective

We save our discussion on Neural Networks for the last section and focus on a functional optimization problem. Informally:

- Instead of minimizing in terms of parameters, we minimize in terms of functions arising from parameters using $R : \mathcal{F} \to \mathbb{R}_+$
- A solution will be a combination of elements from the parametric space $\{\phi(\theta)\}_{\theta \in \Theta} \subset \mathcal{F}$.

Later we will show why this is **reasonable**.

On the form of ϕ

Assume that ϕ parametrized by $\theta \in \Theta$ lives in the Hilbert space $\mathcal{F}$ and is differentiable.

Optimizing by means of Choosing

Think about finding the optimal choice of θ in the $\mathbb{R}^d$ space as to minimize the functional loss. Endowing $\Theta = \mathbb{R}^{d-1}$ with a measure $\mu \in \mathcal{M}(\Theta)$ it is possible to restate the task.

Measure Optimization Problem

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu \right) + G(\mu) \tag{6}$$

Where:

- $G(\mu) : \mathcal{M}(\Theta) \to \mathbb{R}$ is the regularizer of the functional J, just like $\lambda \Xi(\theta)$. Usually, the total variation norm for sparse solutions.
- $|\Theta| = d - 1$, features + bias

Optimizing by means of Choosing

Think about finding the optimal choice of θ in the $\mathbb{R}^d$ space as to minimize the functional loss. Endowing $\Theta = \mathbb{R}^{d-1}$ with a measure $\mu \in \mathcal{M}(\Theta)$ it is possible to restate the task.

Measure Optimization Problem

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu \right) + G(\mu) \tag{7}$$

Where:

- $G(\mu) : \mathcal{M}(\Theta) \to \mathbb{R}$ is the regularizer of the functional J, just like $\lambda \Xi(\theta)$. Usually, the total variation norm for sparse solutions.
- $|\Theta| = d - 1$, features + bias

Interpretation

We look among all possible allocations of choices of the parameters for the best combination to obtain a function that attains minimal risk/maximum fit with the dataset $\mathcal{D}$.

Interpretation

We look among all possible allocations of choices of the parameters for the best combination to obtain a function that attains minimal risk/maximum fit with the dataset $\mathcal{D}$.
The problem is:

- ☺ linear in terms of μ
- ☺ convex
- ☹ infinite dimensional

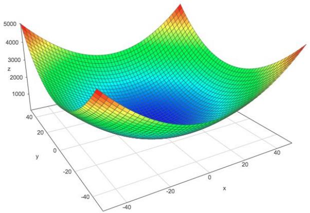


Figure: A convex landscape. Source[link]

Some Methods mentioned in [CB18]

Frank-Wolfe Algorithm: greedy approach of adding neurons at every iteration.

- ☺ connections with Conditional Gradient and Boosting [BSR15; Wan+15]
- ☹ decision problem of finding the optimal particle in general NP-Hard [BP13; Jag13; Bac16]
- ☹ **not practical**

Semidefinite hierarchy: based on expressing the measure in terms of its moments.

- ☺ belongs to larger class of *generalized moment problems* [Las09]
- ☹ asymptotic global convergence (nonquantitative)
- ☹ Only specific instances are covered [CDP17]
- ☹ increasing the dimension growth is exponential.
- ☹ **not practical**

Particle Gradient Descent (GD)

What is **actually used in practice** is Gradient Descent, allowed by the differentiability of ϕ. The measure μ is **discretized** to a finite set of *particles* against which backpropagation is performed.

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \underbrace{w_i}_{\text{weight}} \underbrace{\delta_{\theta_i}}_{\text{position}}$$

Particle Gradient Descent (GD)

What is **actually used in practice** is Gradient Descent, allowed by the differentiability of ϕ. The measure μ is **discretized** to a finite set of *particles* against which backpropagation is performed.

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \underbrace{w_i}_{\text{weight}} \underbrace{\delta_{\theta_i}}_{\text{position}}$$

- positions affect **choices** in the space of parameters
- weights represent **degree of importance** in determining the function to feed into R and G.

Particle GD objective function

The problem is then discretized as:

Discretized Measure Optimization Problem

$$\mu^* = arg \min_{\mathbf{w}\in\mathbb{R}^m \ \boldsymbol{\theta}\in\boldsymbol{\theta}^m} J_m(\mathbf{w}, \boldsymbol{\theta}) \qquad J_m(\mathbf{w}, \boldsymbol{\theta}) := J\left(\frac{1}{m}\sum_{i=1}^{m} w_i\delta_{\theta_i}\right) \tag{8}$$

There are m particles (later, hidden neurons) for which we have:

- weights w_i
- positions $\theta_i \in \mathbb{R}^{d-1}$.

Particle GD objective function

The problem is then discretized as:

Discretized Measure Optimization Problem

$$\mu^* = arg \min_{\mathbf{w}\in\mathbb{R}^m \ \boldsymbol{\theta}\in\boldsymbol{\theta}^m} J_m(\mathbf{w}, \boldsymbol{\theta}) \qquad J_m(\mathbf{w}, \boldsymbol{\theta}) := J\left(\frac{1}{m}\sum_{i=1}^{m} w_i\delta_{\theta_i}\right) \tag{8}$$

There are m particles (later, hidden neurons) for which we have:

- weights w_i
- positions $\theta_i \in \mathbb{R}^{d-1}$.

Discrete measures weakly approximate any measure, where by **weakly** we mean when measuring an integral with respect to a measure of continuous and bounded functions.

Pros, Cons

- ☺ Easy to implement
- ☹ **no a priori guarantees** that J_m is convex
- ☹ convergence is, in most cases, at a local minima.

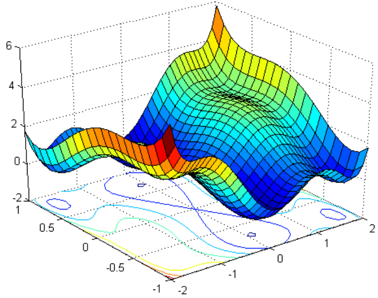


Figure: A nonconvex landscape. Source[StackOverflow]

Overview of Results

The results shown are mostly centered around two questions:

- evaluating the algorithmic limit as $m \to \infty$, known to be equivalent to a **Wasserstein Gradient Flow** [NS17]
- assessing Global Convergence to the optimal measure μ^*, subject to a *generic ideal dynamics that one can only hope to approximate* [CB18]

Overview of Results

The results shown are mostly centered around two questions:

- evaluating the algorithmic limit as $m \to \infty$, known to be equivalent to a **Wasserstein Gradient Flow** [NS17]
- assessing Global Convergence to the optimal measure μ^*, subject to a *generic ideal dynamics that one can only hope to approximate* [CB18]

We obtain:

- a link discretization-original convex problem at the divergent limit of the number of particles
- a non quantitative asymptotic convergence result subject to a criterion

Overview of Results

The results shown are mostly centered around two questions:

- evaluating the algorithmic limit as $m \to \infty$, known to be equivalent to a **Wasserstein Gradient Flow** [NS17]
- assessing Global Convergence to the optimal measure μ^*, subject to a *generic ideal dynamics that one can only hope to approximate* [CB18]

We obtain:

- a link discretization-original convex problem at the divergent limit of the number of particles
- a non quantitative asymptotic convergence result subject to a criterion

Remark

Namely, if criterion holds, then discrete measures converge to the optimal one from some m^* onwards. Unfortunately, no knowledge of a ϵ-bound on the loss in terms of m.

Idealized but also principled and practical

- SGD finds a global minimizer under very restrictive assumptions [LY17; SH17; VBB20; SJL22].
- discretization as a *child* also present in [NS17] but not explored in search of global optimality conditions.
- connection gradient flows and Gradient Descent is also extended to SGD [KY03](Thm. 2.1) and Accelerated gradient descent [Sci+17].

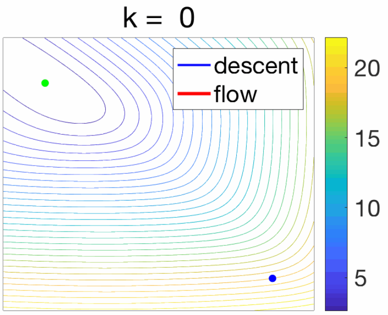


Figure: Animated GD vs gradient flow. Source [Bac20b]

A more general problem

consider the problem over **non negative finite measures** on $\Omega \subset \mathbb{R}^d$ of finding:

Lifted Problem

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\left(\int \Phi d\mu \right) + \int V d\mu \tag{9}$$

A more general problem

consider the problem over **non negative finite measures** on $\Omega \subset \mathbb{R}^d$ of finding:

Lifted Problem

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\left(\int \Phi d\mu \right) + \int V d\mu \tag{9}$$

What changed?

Recall $|\Theta| = d - 1 < d = |\Omega|$. Imagine we changed $\phi \rightsquigarrow \Phi$ and $\widetilde{V} \rightsquigarrow V$ both with one additional dimension.

Main Assumptions (MAs)

We do not stress too much on their formulation but the MAs are important throughout the presentation.

Main Assumptions (MAs)

Require the Hilbert space $\mathcal{F}$ to be separable and $\Omega \subset \mathbb{R}^d$ to be the closure of a convex open set. On top of this, establish that:

1. (*smooth loss*) $R : \mathcal{F} \to \mathbb{R}_+$ is differentiable and its differential dR is Lipschitz on bounded sets and bounded on sublevel sets
2. (basic regularity) the function $\Phi : \Omega \to \mathcal{F}$ is Fréchet differentiable, $V : \Omega \to \mathbb{R}_+$ is semiconvex

Main Assumptions (MAs)

continuation

3. (sublinear growth and locally Lipschitz derivatives) there exists a sequence $(Q_r)_{r\geq 0}$ of nested non empty closed convex subsets of Ω such that:
   a. a kind of matryoshka property
   $$\{u \in \Omega\ ;\ dist(u, Q_r) \leq r'\} \subset Q_{r+r'} \quad \forall r, r' > 0$$
   b. Φ and V are bounded and $d\Phi$ is Lipschitz on each Q_r
   c. denoting as $\|\partial V(u)\|$ the maximal norm of an element in $\partial V(u)$, the growth of the problem is sublinearly bounded as:
$$\exists C_1, C_2 > 0 \quad : \quad \sup_{u\in Q_r} \left\{ \|d\Phi_u\| + \|\partial V(u)\| \right\} \leq C_1 + C_2 r \quad \forall r > 0$$

Main Assumptions, forcing

Add that:

- (forcing in matryoshka) by convention, we set $F(\mu) = \infty$ if μ is not concentrated on Ω.

Main Assumptions, forcing

Add that:

- (forcing in matryoshka) by convention, we set $F(\mu) = \infty$ if μ is not concentrated on Ω.
- (forcing in Hilbert Space) the integral involving Φ is assumed to be a **Bochner integral**. In simple words, it maps to $\mathcal{F}$ whenever:
  - Φ is measurable
  - $\int \|\phi\| \, d|\mu| < \infty$

  Else $F(\mu) = \infty$

Why?

- avoid results in which part of the parameters are assigned outside of the region of optimization
- proper domain of R

Technical vs Reasonable points

Infinite matryoshkas

Q_r can be unbounded so 3-(c) is not only for local Lipschitzness and sublinear growth, but also as a **technical requirement** for the gradient flow analysis to be stable. Instrumental in proofs derived from [AGS05]

Technical vs Reasonable points

Infinite matryoshkas

Q_r can be unbounded so 3-(c) is not only for local Lipschitzness and sublinear growth, but also as a **technical requirement** for the gradient flow analysis to be stable. Instrumental in proofs derived from [AGS05]

Technical but not unreasonable

All the remaining are in line with common models such as:

- Sigmoid NNs (here)
- `ReLu` NNs
- Sparse Spikes Deconvolution
- Low Rank Tensor Decomposition

See original paper [CB18] for the others.

Homogeneous Lifting & Tools

Partially 1-homogeneous functions

For continuous functions:

$$\phi : \Theta \to \mathcal{F} \qquad \widetilde{V} : \Theta \to \mathbb{R}_+$$

assign $\Omega := \mathbb{R} \times \Theta \subset \mathbb{R}^d$, $\Phi(w, \theta) = w \cdot \phi(\theta)$ and $V(w, \theta) = |w|\widetilde{V}(\theta)$.
Notice that Φ and V are 1-homogeneous in the first entry i.e.
$f(\lambda w, \theta) = \lambda f(w, \theta) \forall w > 0$.

Homogeneous Lifting & Tools

Partially 1-homogeneous functions

For continuous functions:

$$\phi : \Theta \to \mathcal{F} \qquad \widetilde{V} : \Theta \to \mathbb{R}_+$$

assign $\Omega := \mathbb{R} \times \Theta \subset \mathbb{R}^d$, $\Phi(w, \theta) = w \cdot \phi(\theta)$ and $V(w, \theta) = |w|\widetilde{V}(\theta)$.
Notice that Φ and V are 1-homogeneous in the first entry i.e. $f(\lambda w, \theta) = \lambda f(w, \theta) \forall w > 0$.

Use the projection operator for $B \subset \Theta$ measurable:

$$h^1 : \mathcal{M}_+(\Omega) \to \mathcal{M}(\Theta) \quad h^1(\mu)(B) = \int_{\mathbb{R}} w\mu(dw, B) \;\; \forall \mu \in \mathcal{P}(\Omega)$$

On the pushforward lifted measure:

$$\nu = \underbrace{f}_{\in L^1(\sigma)} \underbrace{\sigma}_{\in P(\Theta)} \qquad \mu := (f \times \mathtt{id})_{\#}\sigma = \sigma \circ (f \times \mathtt{id})^{-1} \in \mathcal{P}(\Omega)$$

Notation

Alert slide

To avoid potential confusion, we use the following notation:

	smaller space Θ	bigger space Ω
dimension	$d-1$	d
measures	ν	μ
functions	$\phi, \widetilde{V}$	Φ, V
risk functional	J	F

Both have R and G as cost and regularizer.
Takes time to digest as there are many objects at the same time.

Results

Lifted problem is equivalent

1. (normalization) $\exists \mu_{norm} \in \mathcal{P}(\Omega) \quad : \quad F(\mu_{norm}) = F(\mu) \quad \forall \mu \in \mathcal{M}_+(\Omega)$
   i.e. we can use probability measures

Proof Strategy. Construction. ◇

Results

Lifted problem is equivalent

2. (surjectivity of h^1) $h^1(\mathcal{P}(\Omega)) \supset \mathcal{M}(\Theta)$ i.e. we cover all ν

Proof Strategy. Construction. ◇

Results

Lifted problem is equivalent

3. (equality condition) for appropriate Θ-regularizers $G(\nu)$, $\nu \in \mathcal{M}(\Theta)$ minimizing J:

$$\exists \mu \in \mathcal{P}(\Omega) : \mu = \operatorname*{arg\,min}_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \tag{10}$$

Proof Strategy. Construction. ◇

Results

Lifted problem is equivalent

4. (Total Variation is included) $V(w,\theta) = |w|, \mu \in \mathcal{P}(\Omega)$ pushlifted as before $\implies |h^1(\mu)| = \int V d\mu$ is appropriate as per #3

Proof Strategy. Construction. ◇

Addenda & OT view

To avoid confusion, we recap below the symbols:

$$\mathcal{P}(\Omega) \ni \mu \overset{h^1(\cdot)}{\rightsquigarrow} \nu \in \mathcal{M}(\Theta)$$

$$\int \phi d\nu \overset{\cdot w}{\rightsquigarrow} \int \Phi d\mu \qquad G(\nu) = \int \widetilde{V}(\theta) d\nu \overset{\cdot |w|}{\rightsquigarrow} \int V(w, \theta) d\mu = G(\mu)$$

Addenda & OT view

To avoid confusion, we recap below the symbols:

$$\mathcal{P}(\Omega) \ni \mu \overset{h^1(\cdot)}{\rightsquigarrow} \nu \in \mathcal{M}(\Theta)$$

$$\int \phi d\nu \overset{\cdot w}{\rightsquigarrow} \int \Phi d\mu \qquad G(\nu) = \int \widetilde{V}(\theta) d\nu \overset{\cdot |w|}{\rightsquigarrow} \int V(w,\theta) d\mu = G(\mu)$$

We also need this side result:

F continuity

Under (MAs) F is continuous for the Wasserstein Metric below:

$$W_2(\mu_1, \mu_2) = \sqrt{\inf_{\gamma \in \Pi(\mu_1,\mu_2)} \int_{\Omega\times\Omega} |y-x|^2 d\gamma(x,y)}$$

Proof Strategy. (MAs) and F form. ◇

Recap

- we can see the problem from a measure choice perspective as in (6):

$$\nu^* = \underset{\nu \in \mathcal{M}(\Theta)}{\arg\min} J(\nu) \quad J(\nu) := R\left(\int \phi d\nu \right) + G(\nu)$$

Recap

- we can see the problem from a measure choice perspective as in (6):

$$\nu^* = \arg\min_{\nu \in \mathcal{M}(\Theta)} J(\nu) \quad J(\nu) := R\left(\int \phi d\nu \right) + G(\nu)$$

- this is lifted to the **equivalent** version (9)

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\left(\int \Phi d\mu \right) + \int V d\mu$$

for a **reasonable** choice of regularizer

Recap

- we can see the problem from a measure choice perspective as in (6):

$$\nu^* = \arg\min_{\nu \in \mathcal{M}(\Theta)} J(\nu) \quad J(\nu) := R\Big(\int \phi d\nu \Big) + G(\nu)$$

- this is lifted to the **equivalent** version (9)

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\Big(\int \Phi d\mu \Big) + \int V d\mu$$

  for a **reasonable** choice of regularizer

The discretized version (8) becomes:

$$F_m(\mathbf{u}) := F\left(\frac{1}{m} \sum_{i=1}^{m} \delta_{\mathbf{u}_i} \right) = R\left(\frac{1}{m} \sum_{i=1}^{m} \Phi(\mathbf{u}_i) \right) + \frac{1}{m} \sum_{i=1}^{m} V(\mathbf{u}_i). \tag{11}$$

Recap

- we can see the problem from a measure choice perspective as in (6):

$$\nu^* = \underset{\nu \in \mathcal{M}(\Theta)}{\arg\min} J(\nu) \quad J(\nu) := R\left(\int \phi d\nu\right) + G(\nu)$$

- this is lifted to the **equivalent** version (9)

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\left(\int \Phi d\mu\right) + \int V d\mu$$

  for a **reasonable** choice of regularizer

The discretized version (8) becomes:

$$F_m(\mathbf{u}) := F\left(\frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{u}_i}\right) = R\left(\frac{1}{m}\sum_{i=1}^{m} \Phi(\mathbf{u}_i)\right) + \frac{1}{m}\sum_{i=1}^{m} V(\mathbf{u}_i). \qquad (11)$$

where μ encapsulates weights w_i and positions $\boldsymbol{\theta}_i$ in the same dirac of $\mathbf{u} = (w_i, \boldsymbol{\theta}_i)$.

Alert slide

Differentiated notation stop

From now onwards, ν and μ will not be restricted to the notation we used in the lifting. This may be confusing.

What now?

We have that:

- ☺ the problem is **feasible** in practice (GD)
- ☺ $\mu \in \mathcal{P}(\Omega)$ is a probability measure $\implies$ we will see that we can use Wasserstein Gradient Flows (wide results)
- ☺ Gradient Flow and GD have analogies
- ☺ weights and positions are not decoupled, both under δ_u
- ☺ F is continuous under the (MAs)

What now?

We have that:

- ☹ still non convex

What now?

We have that:

- ☺ the problem is **feasible** in practice (GD)
- ☺ $\mu \in \mathcal{P}(\Omega)$ is a probability measure $\implies$ we will see that we can use Wasserstein Gradient Flows (wide results)
- ☺ Gradient Flow and GD have analogies
- ☺ weights and positions are not decoupled, both under δ_u
- ☺ F is continuous under the (MAs)
- ☹ still non convex

☹ is not drastically bad

At this point, we obtained a well posed problem. Now, we use the Theory of Wasserstein Gradient Flows to tackle the issue.

Lecture Path

Overview

- main theoretical results presented from an intuitive point of View.

Overview

- first subsection: dynamics on the parameters can be seen in terms of a probability measure over the parameters that moves according to a Wasserstein Gradient Flow (Wgf)
- second subsection: Wgfs are instrumental to design a **criterion** on the starting measure **to escape local minima**

Intuition [Bac20b]

1. GD as discrete update of parameters of a differentiable function
   $\mathbf{u}_{n+1} = \mathbf{u}_n - \epsilon \nabla F_m(\mathbf{u}_n) \quad \epsilon > 0$

Intuition [Bac20b]

1. GD as discrete update of parameters of a differentiable function
   $\mathbf{u}_{n+1} = \mathbf{u}_n - \epsilon \nabla F_m(\mathbf{u}_n) \quad \epsilon > 0$
2. See u_n as
   $X : \mathbb{R}_+ \to \Omega \quad \mathbf{u}_n = X(n\epsilon).$

Intuition [Bac20b]

1. GD as discrete update of parameters of a differentiable function
$\mathbf{u}_{n+1} = \mathbf{u}_n - \epsilon \nabla F_m(\mathbf{u}_n) \quad \epsilon > 0$
2. See u_n as
$X : \mathbb{R}_+ \to \Omega \quad \mathbf{u}_n = X(n\epsilon)$.
3. GF as ODE for $t\epsilon = n, \epsilon \to 0$:

$$X'(t) = -\nabla F_m(X(t))$$

Up to **verified** regularity assumptions.

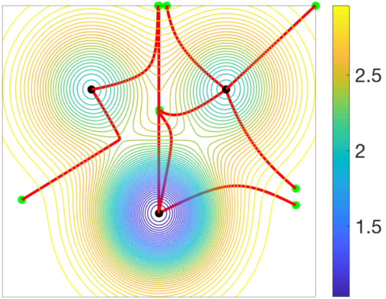


Figure: Gradient flows, Source [Bac20b]

Flow properties and specifications

- Function decreases along the trajectory (chain rule):

$$\frac{d}{dt}F_m(X(t)) = -\|\nabla F_m(X(t))\|_2^2$$

- if convergence, it is necessarily at a stationary point s.t. $\nabla F_m(X(t)) = 0$.

Remark

- convergence specifics later
- construction for Wgfs more elaborate, doc has refs.

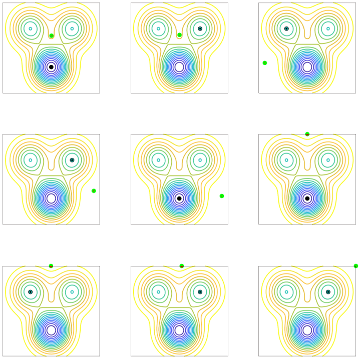


Figure: Animation of previous image. Source [Bac20b]

On parameters

Particle Gradient flow

A dynamics for F_m:

$$\mathbf{u} : \mathbb{R}_+ \to \Omega^m \quad t \to \mathbf{u}(t) \in \Omega^m$$

is a particle gradient flow if:

1. absolute continuity
2. rescaled gradient flow equation

$$\mathbf{u}'(t) = -m\partial F_m(\mathbf{u}(t)) \; a.e. \; t \geq 0$$

On parameters

Particle Gradient flow

A dynamics for F_m:
$\mathbf{u} : \mathbb{R}_+ \to \Omega^m \quad t \to \mathbf{u}(t) \in \Omega^m$ is a particle gradient flow if:

1. absolute continuity
2. rescaled gradient flow equation $\mathbf{u}'(t) = -m\partial F_m(\mathbf{u}(t))$ *a.e.* $t \geq 0$

Remarks

Notice that in #2 we have:

- a.e. conditions by the absolute continuity requirement #1
- subdifferentials by potential non differentiability of V (only semiconvex)
- rescaling by m for convenience at limit, each atom has $\frac{1}{m}$ mass

On parameters

Particle Gradient flow

A dynamics for F_m:
$\mathbf{u} : \mathbb{R}_+ \to \Omega^m \quad t \to \mathbf{u}(t) \in \Omega^m$ is a particle gradient flow if:

1. absolute continuity
2. rescaled gradient flow equation $\mathbf{u}'(t) = -m\partial F_m(\mathbf{u}(t))$ $a.e.\ t \geq 0$

Remarks

Notice that in #2 we have:

- a.e. conditions by the absolute continuity requirement #1
- subdifferentials by potential non differentiability of V (only semiconvex)
- rescaling by m for convenience at limit, each atom has $\frac{1}{m}$ mass

On parameters

Particle Gradient flow

A dynamics for F_m:
$\mathbf{u} : \mathbb{R}_+ \to \Omega^m \quad t \to \mathbf{u}(t) \in \Omega^m$ is a particle gradient flow if:

1. absolute continuity
2. rescaled gradient flow equation $\mathbf{u}'(t) = -m\partial F_m(\mathbf{u}(t))$ $a.e.\ t \geq 0$

Remarks

Notice that in #2 we have:

- a.e. conditions by the absolute continuity requirement #1
- subdifferentials by potential non differentiability of V (only semiconvex)
- rescaling by m for convenience at limit, each atom has $\frac{1}{m}$ mass

On parameters

Particle flow in F_m properties

1. existance and uniqueness for any initialization
2. for a.e. $t > 0$
   $\frac{d}{ds} F_m(\mathbf{u}(s))|_{s=t} = -\frac{1}{m} |\mathbf{u}'(t)|^2$
3. particle velocity $v_t(u)$ is

$$\widetilde{v}_t(u) - \mathtt{proj}_{\partial V(u)}(\widetilde{v}_t(u)) \quad (12)$$

for a general particle u

Remarks

recognize that:

- expressions below, basically chain rule

$$[\widetilde{v}_t(\mathbf{u}_i)]_{i=1}^m = -\nabla R \left(\frac{1}{m} \sum_{i=1}^m \Phi(\mathbf{u}_i) \right) \qquad \widetilde{v}_t(u) = [\langle R'(\textstyle\int \Phi d\mu_{m,t}),\ \partial_j \Phi(u) \rangle]_{j=1}^d$$

On parameters

Particle flow in F_m properties

1. existance and uniqueness for any initialization
2. for a.e. $t > 0$
$\frac{d}{ds} F_m(\mathbf{u}(s))|_{s=t} = -\frac{1}{m}|\mathbf{u}'(t)|^2$
3. particle velocity $v_t(u)$ is

$$\widetilde{v}_t(u) - \mathtt{proj}_{\partial V(u)}(\widetilde{v}_t(u)) \quad (12)$$

for a general particle u

Remarks

recognize that:

- $R'(f)$ denotes the gradient of R at $f \in \mathcal{F}$
- $\partial_j \Phi \in \mathcal{F}$ differential $d\Phi(u)$ applied to the j^{th} vector of the canonical basis of $\mathbb{R}^d$.

$$[\widetilde{v}_t(\mathbf{u}_i)]_{i=1}^m = -\nabla R\left(\frac{1}{m}\sum_{i=1}^m \Phi(\mathbf{u}_i)\right) \qquad \widetilde{v}_t(u) = [\langle R'(\textstyle\int \Phi d\mu_{m,t}),\ \partial_j \Phi(u)\rangle]_{j=1}^d$$

On parameters

Particle flow in F_m properties

1. existance and uniqueness for any initialization
2. for a.e. $t > 0$
   $\frac{d}{ds} F_m(\mathbf{u}(s))|_{s=t} = -\frac{1}{m}|\mathbf{u}'(t)|^2$
3. particle velocity $v_t(u)$ is

$$\widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)}(\widetilde{v}_t(u))$$

   for a general particle u

Remarks

recognize that:

- expressions below, basically chain rule
- $R'(f)$ denotes the gradient of R at $f \in \mathcal{F}$
- $\partial_j \Phi \in \mathcal{F}$ differential $d\Phi(u)$ applied to the j^{th} vector of the canonical basis of $\mathbb{R}^d$.
- $\texttt{proj}$ by regularization

$$[\widetilde{v}_t(\mathbf{u}_i)]_{i=1}^m = -\nabla R\left(\frac{1}{m}\sum_{i=1}^m \Phi(\mathbf{u}_i)\right) \qquad \widetilde{v}_t(u) = [\langle R'(\textstyle\int \Phi d\mu_{m,t}),\ \partial_j \Phi(u)\rangle]_{j=1}^d$$

On measures

Wasserstein Gradient Flow

For the functional F and an interval $[0, T)$ a Wasserstein gradient flow is a path $t \to \mu_t$ on $[0, T)$ such that:

1. it is absolutely continuous
2. $(\mu_t)_{t\in[0,T)} \in \mathcal{P}_2(\Omega)$
3. for $[0, T) \times \Omega^d$ satisfies the continuity equation:

$$\partial_t \mu_t = -\mathtt{div}(v_t \mu_t) \quad v_t \in \partial F'(\mu_t) \tag{13}$$

On measures

Wasserstein Gradient Flow

For the functional F and an interval $[0, T)$ a Wasserstein gradient flow is a path $t \to \mu_t$ on $[0, T)$ such that:

1. it is absolutely continuous
2. $(\mu_t)_{t \in [0,T)} \in \mathcal{P}_2(\Omega)$
3. for $[0, T) \times \Omega^d$ satisfies the continuity equation:

$$\partial_t \mu_t = -\mathtt{div}(v_t \mu_t) \quad v_t \in \partial F'(\mu_t)$$

Remark

In a distributional sense since densities are not necessarily smooth. A broader presentation is given in the Appendix.

Particles flow as discrete measures

Link gradient flow and atomic Wasserstein gradient flow

For a gradient flow $\mathbf{u} : \mathbb{R}_+ \to \Omega^m$ of F_m the map:

$$t \to \mu_{m,t} := \frac{1}{m} \sum_{i=1}^{m} \delta_{\mathbf{u}_i(t)}$$

is a Wasserstein gradient flow for the non particle version of F_m, denoted as F.

Proof Strategy. show continuity equation satisfied distributionally $\diamond$

Remarks

- dynamics are in t at m fixed

Particles flow as discrete measures

Link gradient flow and atomic Wasserstein gradient flow

For a gradient flow $\mathbf{u} : \mathbb{R}_+ \to \Omega^m$ of F_m the map:

$$t \to \mu_{m,t} := \frac{1}{m} \sum_{i=1}^{m} \delta_{\mathbf{u}_i(t)}$$

is a Wasserstein gradient flow for the non particle version of F_m, denoted as F.

Proof Strategy. show continuity equation satisfied distributionally $\diamond$

Remarks

- dynamics are in t at m fixed
- if F does not admit an m-atomic minimizer, $\mu_{m,t}$ converges to a measure that **does not** minimize F.

Particles flow as discrete measures

Link gradient flow and atomic Wasserstein gradient flow

For a gradient flow $\mathbf{u} : \mathbb{R}_+ \to \Omega^m$ of F_m the map:

$$t \to \mu_{m,t} := \frac{1}{m} \sum_{i=1}^{m} \delta_{\mathbf{u}_i(t)}$$

is a Wasserstein gradient flow for the non particle version of F_m, denoted as F.

Proof Strategy. show continuity equation satisfied distributionally $\diamond$

Remarks

- dynamics are in t at m fixed
- if F does not admit an m-atomic minimizer, $\mu_{m,t}$ converges to a measure that **does not** minimize F.
- still not covering *diffuse* measures theory

On measures, general properties

Existance and uniqueness of Wgf for F

Under (MAs), if $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on $Q_{r_0} \subset \Omega$:

$$\exists!(\mu_t)_{t\geq 0} \quad Wgf: \quad \text{velocities as (12)}$$

Proof Strategy. Detour on matryoshka concentrated $F^{(r)}$ from [AGS05] with many subproofs. Details in publication [CB18] and doc. ◇

On measures, general properties

Existance and uniqueness of Wgf for F

Under (MAs), if $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on $Q_{r_0} \subset \Omega$:

$$\exists!(\mu_t)_{t\geq 0} \quad Wgf : \quad \text{velocities as (12)}$$

Proof Strategy. Detour on matryoshka concentrated $F^{(r)}$ from [AGS05] with many subproofs. Details in publication [CB18] and doc. ◇

Interpretation

For any starting point concentrated on a matryoshka we always identify unambiguously the Wgf.

Particles flowing to measures

Many-particle limit

Under (MAs), consider a sequence in m of gradient flows for F_m $(t \to \mathbf{u}_m(t))_{m\in\mathbb{N}}$ initialized at $\mu_{m,0}$ concentrated in $Q_{r_0} \subset \Omega$. If

$$\lim_{m\to\infty} \|\mu_{m,0} - \mu_0\|_{W_2} = 0$$

with $\mu_0 \in \mathcal{P}_2(\Omega)$ Then :

$$(\mu_{m,t})_{t\geq 0} \underset{W_2}{\overset{m\to\infty}{\rightrightarrows}} (\mu_t)_{t\geq 0}$$

Remarks

Where:

- $(\mu_t)_{t\geq 0}$ is the unique (and existent) Wgf of F which starts at μ_0

Proof Strategy. find limit curve, show it is Wgf by subsequences ◇

Particles flowing to measures

Many-particle limit

Under (MAs), consider a sequence in m of gradient flows for F_m $(t \to \mathbf{u}_m(t))_{m\in\mathbb{N}}$ initialized at $\mu_{m,0}$ concentrated in $Q_{r_0} \subset \Omega$. If

$$\lim_{m\to\infty} \|\mu_{m,0} - \mu_0\|_{W_2} = 0$$

with $\mu_0 \in \mathcal{P}_2(\Omega)$ Then :

$$(\mu_{m,t})_{t\geq 0} \underset{W_2}{\overset{m\to\infty}{\rightrightarrows}} (\mu_t)_{t\geq 0}$$

Proof Strategy. find limit curve, show it is Wgf by subsequences ◇

Remarks

Where:

- $(\mu_t)_{t\geq 0}$ is the unique (and existent) Wgf of F which starts at μ_0
- Namely, if our discrete starting point converges to $\mu_0 \in \mathcal{P}_2(\Omega)$ then the whole discrete sequence converges to the continuous version of the same problem

Practical Example

Empirical Measure

As an example, consider a measure $\mu_0 \in \mathcal{P}_2(Q_{r_0})$. If we want to build a sequence converging in W_2 to it we can simply choose a flow in the parameters governed by the size m:

$$\mathbf{u}_m(0) = (u_1, \dots, u_m) \quad u_i \overset{iid}{\sim} \mu_0 \quad \forall i = 1, \dots, m$$

Namely, parameters picked at random from the diffuse measure μ_0. Then by the CLT the sequence:

$$\mu_{m,0} = \frac{1}{m} \sum_{i=1}^{m} \delta_{u_i} \quad \mu_{m,0} \xrightarrow[W_2]{a.s.} \mu_0$$

Recap

We outlined:

- main properties of particle gradient flows over parameters
- main properties of Wasserstein gradient flows over probability measures

Recap

We outlined:

- main properties of particle gradient flows over parameters
- main properties of Wasserstein gradient flows over probability measures

We link the two whenever:

- (MAs) hold
- the discrete measure at the start W_2-converges to a measure

Overview

Need:

- Suited Assumptions (SAs), more technical, where (SAs) $\Longrightarrow$ (MAs), so all previous results are inherited.

Overview

Need:

- Φ and V need to have a **homogeneity direction**

Overview

Need:

- $\mathtt{spt}\,\mu_0$ for the initial measure of the Wgf has to satisfy a **separation property**, which is **preserved** along the path.

Overview

Need:

- Suited Assumptions (SAs), more technical, where (SAs) $\Longrightarrow$ (MAs), so all previous results are inherited.
- Φ and V need to have a **homogeneity direction**
- $\mathrm{spt}\,\mu_0$ for the initial measure of the Wgf has to satisfy a **separation property**, which is **preserved** along the path.

Show:

- difference stationary - optimal measures
- criteria to escape stationary points
- convergence implies null dynamics
- condition for the starting measure to be always capable of escaping across dynamics

Assuming convergence, we **craft** a **discrete** measure that, after some m^*, escapes all local minimas!

Minimizers (general property)

Minimizers with convexity characterization

Assume R is convex, μ is a minimizer if and only if:

1. $F'(\mu) \geq 0$
2. $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$

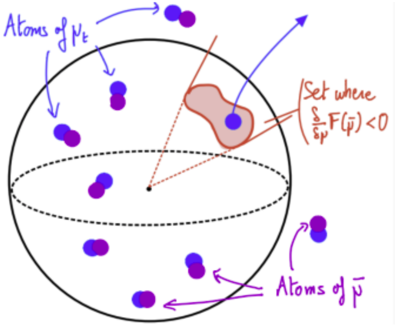


Figure: $\overline{\mu}$ is not a minimizer if it does not sat #2. Source [Chi21]

Minimizers (general property)

Minimizers with convexity characterization

Assume R is convex, μ is a minimizer if and only if:

1. $F'(\mu) \geq 0$
2. $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$

Remarks

- We solve the PDE
- intuition: no abstract direction of improvement
- stronger than stationarity, particle as backpropagation [Bac20a]

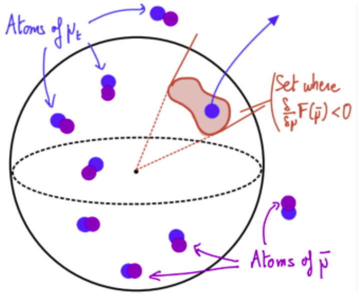


Figure: $\overline{\mu}$ is not a minimizer if it does not sat #2. Source [Chi21]

Minimizers (general property)

Minimizers with convexity characterization

Assume R is convex, μ is a minimizer if and only if:

1. $F'(\mu) \geq 0$
2. $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$

Remarks

- We solve the PDE
- intuition: no abstract direction of improvement
- stronger than stationarity, particle as backpropagation [Bac20a]

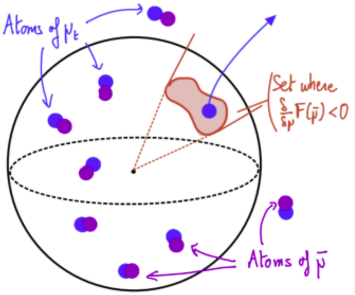


Figure: $\overline{\mu}$ is not a minimizer if it does not sat #2. Source [Chi21]

Minimizers (general property)

Minimizers with convexity characterization

Assume R is convex, μ is a minimizer if and only if:

1. $F'(\mu) \geq 0$
2. $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$

Remarks

- We solve the PDE
- intuition: no abstract direction of improvement
- stronger than stationarity, particle as backpropagation [Bac20a]

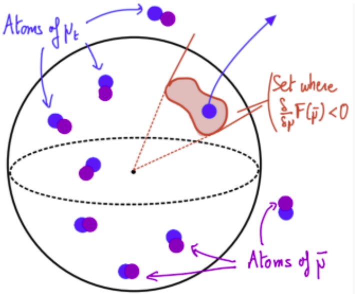


Figure: $\overline{\mu}$ is not a minimizer if it does not sat #2. Source [Chi21]

Flows over Homogeneous functions

- imaginary Level sets of $F'(\mu)$

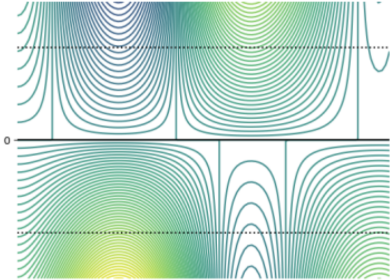


Figure: Source [CB18]

Flows over Homogeneous functions

- imaginary Level sets of $F'(\mu)$
- $\Omega = \mathbb{R}^2$ and weights on vertical axis

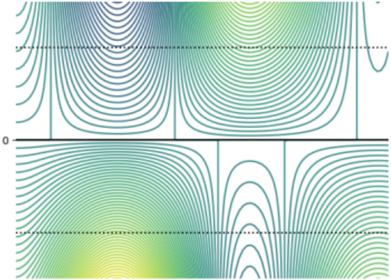


Figure: Source [CB18]

Flows over Homogeneous functions

- imaginary Level sets of $F'(\mu)$
- $\Omega = \mathbb{R}^2$ and weights on vertical axis
- a Wgf flows over them but the landscape dependends on μ

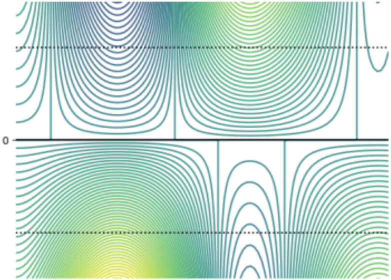


Figure: Source [CB18]

Flows over Homogeneous functions

- imaginary Level sets of $F'(\mu)$
- $\Omega = \mathbb{R}^2$ and weights on vertical axis
- a Wgf flows over them but the landscape dependends on μ
- minimizers are nonnegative and null on the support

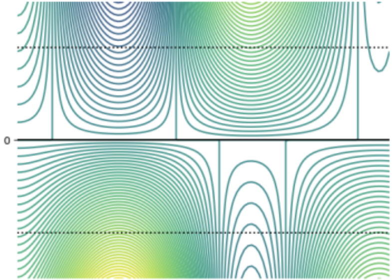


Figure: Source [CB18]

Flows over Homogeneous functions

- imaginary Level sets of $F'(\mu)$
- $\Omega = \mathbb{R}^2$ and weights on vertical axis
- a Wgf flows over them but the landscape dependends on μ
- minimizers are nonnegative and null on the support
- by homogeneity, only the dotted lines are studied

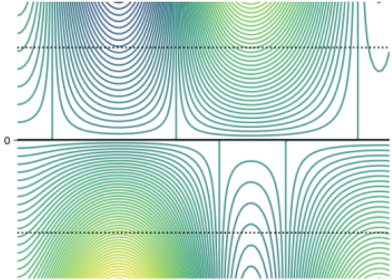


Figure: Source [CB18]

Escaping condition

Criteria to escape local minima

Under (SAs) a Wgf which gets ϵ-$\|\cdot\|_{BL}$ close in h^1-projection to a local minima escapes at a later time if $\mu_t(A) > 0$ for
$A = (\mathbb{R}_+ \times K^+) \cup (\mathbb{R}_- \times K^-)$
Where:

- K^+ is the $-\eta$ sublevel set of $\theta \to F'(\mu)(1, \theta)$
- K^- is the $-\eta$ sublevel set of $\theta \to F'(\mu)(-1, \theta)$

With $\eta > 0$ arbitrarily small.

Remark

The objective is finding a condition at the start that preserves the escaping criteria across dynamics.

Stability

Separation property

A a closed set $K \subset [-r, r] \times \Theta$ that separates (continuous paths across it) $\{-r\} \times \Theta$ and $\{r\} \times \Theta$ for some $r > 0$.

Stability of the separation property

Under (SAs), let $(\mu_t)_t$ be a Wgf for F. If $\operatorname{spt} \mu_0$ satisfies the separation property, then $\operatorname{spt} \mu_t$ does $\forall t > 0$.

Stability

Separation property

A a closed set $K \subset [-r, r] \times \Theta$ that separates (continuous paths across it) $\{-r\} \times \Theta$ and $\{r\} \times \Theta$ for some $r > 0$.

Remark

Reached with a detour on topological degree theory. [CB18]

Stability of the separation property

Under (SAs), let $(\mu_t)_t$ be a Wgf for F. If $\operatorname{spt} \mu_0$ satisfies the separation property, then $\operatorname{spt} \mu_t$ does $\forall t > 0$.

Remark

We have a condition on the support satisfied for all t in a Wgf, we will use it later

A Projection result

Nullity at convergence

Under (SAs), consider a Wgf $(\mu_t)_t$ for F. Then:

$$h^1(\mu_t) \xrightarrow{w} \nu \implies F'(\nu) = 0 \quad \nu\text{-a.e.}$$

Where $\nu \in \mathcal{M}_+(\Theta)$

Remark

The flow imposes that we always improve fit, if we converge, it must be at a measure at which we cannot decrease F.

Main Results: Convergence

General case

Under (SAs), for some $r_0 > 0$ let:

- (concentration) $\operatorname{spt} \mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ be a Wgf of F such that $\operatorname{spt} \mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then:

$$h^1(\mu_t) \xrightarrow{w} \nu \implies F(\mu_t) \xrightarrow{t\to\infty} F^* = \min_{\mathcal{M}_+(\Omega)} F$$

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

Main Results: Convergence

General case

Under (SAs), for some $r_0 > 0$ let:

- (concentration) $\operatorname{spt}\mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ be a Wgf of F such that $\operatorname{spt}\mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then if $h^1(\mu_t) \stackrel{w}{\to} \nu$:

$$F(\mu_t) \stackrel{t\to\infty}{\to} F^* = \min_{\mathcal{M}_+(\Omega)} F$$

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

Proof Strategy. The separation is satisfied throughout (§Stability). ◇

Main Results: Convergence

General case

Under (SAs), for some $r_0 > 0$ let:

- (concentration) $\operatorname{spt}\mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ be a Wgf of F such that $\operatorname{spt}\mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then if $h^1(\mu_t) \xrightarrow{w} \nu$:

$$F(\mu_t) \overset{t\to\infty}{\rightarrow} F^* = \min_{\mathcal{M}_+(\Omega)} F$$

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

Proof Strategy. The separation is satisfied throughout (§Stability), convergence ensures that we reach a point where we have $F'(\nu) = 0$ (§Projection result). Assume we reach a local minima by contradiction. ◇

Main Results: Convergence

General case

Under (SAs), for some $r_0 > 0$ let:

- (concentration) $\operatorname{spt}\mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ be a Wgf of F such that $\operatorname{spt}\mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then if $h^1(\mu_t) \xrightarrow{w} \nu$:

$$F(\mu_t) \overset{t\to\infty}{\rightarrow} F^* = \min_{\mathcal{M}_+(\Omega)} F$$

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

Proof Strategy. The separation is satisfied throughout (§Stability), convergence ensures that we reach a point where we have $F'(\nu) = 0$ (§Projection result). Assume we reach a local minima by contradiction. With additional notions from [CB18], it is possible to show that the flow satisfies the escaping criteria throughout (§Escaping condition), so given convergence, it **must be at a global minima**. $\diamond$

Main Results: Order

Limit order is not important

Under (MAs), if:

- $(\mu_t)_t : \mu_0$ is concentrated on Q_{r_0} and $F(\mu_t) \overset{t\to\infty}{\to} F^*$
- $(\mu_{0,m})_m$ concentrated on $Q_{r_0} : \mu_m \underset{m\to\infty}{\overset{W_2}{\to}} \mu_0$

Then, limits can be exchanged:

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

Limit switch is fundamental

The divergent indexes m, t do not influence each other in the convergence to F^*.

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles

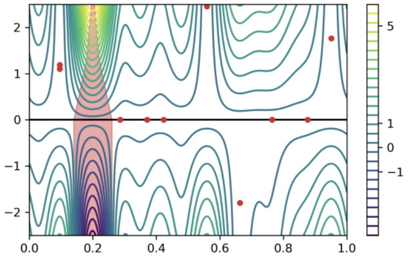


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles
- imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it

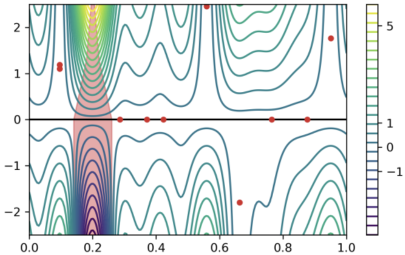


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles
- imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it
- to escape it should give positive weight to the red region

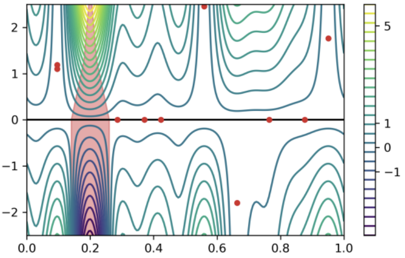


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles
- imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it
- to escape it should give positive weight to the red region
- part below, $F'(\nu)$ negative, use escaping criteria

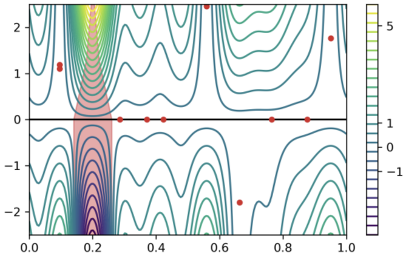


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles
- imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it
- to escape it should give positive weight to the red region
- part below, $F'(\nu)$ negative, use escaping criteria
- part above, $F'(\nu)$ positive, more technical [CB18](Lem. C.18).

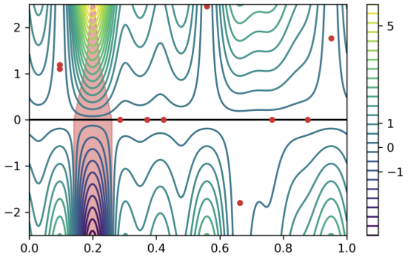


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Graphically escaping, 1-homogeneous case

- ν is non optimal, $F'(\nu) < 0$ at some particles
- imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it
- to escape it should give positive weight to the red region
- part below, $F'(\nu)$ negative, use escaping criteria
- part above, $F'(\nu)$ positive, more technical [CB18](Lem. C.18).
- Theorem uses both technical condition and 2-homogeneous notions

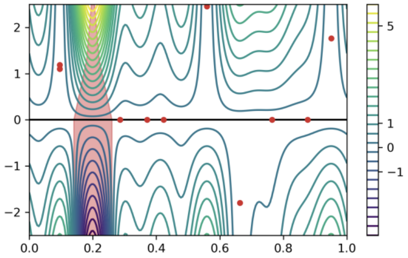


Figure: Level sets view of $F'(\mu)$, $\Omega = \mathbb{R}^2$. Vertical direction is w. Measure ν has support on the red dots. Source [CB18]

Use Case as a Corollary

Global Minimization Sufficient Conditions

Under (SAs) add that $(\mu_t)_t$ is a Wgf of F which for some $r_0 > 0$ satisfies

- (concentration) $\operatorname{spt}\mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ a Wgf of F such that $\operatorname{spt}\mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then:

1. $(\mu_t)_t \overset{W_2}{\to} \mu_\infty \implies F(\mu_t) \overset{t\to\infty}{\to} F^* = \arg\min_{\mathcal{M}_+(\Omega)} F$
2. for a given (parameter) classical Gradient flow $(\mathbf{u}_m(t))_{m\in\mathbb{N}, t\in\mathbb{R}_+}$ which is initialized at its Wgf in $[-r_0, r_0] \times \Theta$:

$$\mu_{m,0} \underset{m\to\infty}{\overset{W_2}{\to}} \mu_0 \implies \lim_{t,m\to\infty} F(\mu_{m,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$$

Use Case as a Corollary

Global Minimization Sufficient Conditions

Under (SAs) add that $(\mu_t)_t$ is a Wgf of F which for some $r_0 > 0$ satisfies

- (concentration) $\operatorname{spt}\mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ a Wgf of F such that $\operatorname{spt}\mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then:

1. $(\mu_t)_t \overset{W_2}{\to} \mu_\infty \implies F(\mu_t) \overset{t\to\infty}{\to} F^* = \arg\min_{\mathcal{M}_+(\Omega)} F$
2. for a given (parameter) classical Gradient flow $(\mathbf{u}_m(t))_{m\in\mathbb{N}, t\in\mathbb{R}_+}$ which is initialized at its Wgf in $[-r_0, r_0] \times \Theta$:

$$\mu_{m,0} \underset{m\to\infty}{\overset{W_2}{\to}} \mu_0 \implies \lim_{t,m\to\infty} F(\mu_{m,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$$

Use Case as a Corollary

Global Minimization Sufficient Conditions

Under (SAs) add that $(\mu_t)_t$ is a Wgf of F which for some $r_0 > 0$ satisfies

- (concentration) $\operatorname{spt} \mu_0 \subset [-r_0, r_0] \times \Theta$.
- (separation) $(\mu_t)_t$ a Wgf of F such that $\operatorname{spt} \mu_0$ separates $\{-r_0\} \times \Theta$ and $\{r_0\} \times \Theta$

Then:

1. $(\mu_t)_t \overset{W_2}{\to} \mu_\infty \implies F(\mu_t) \overset{t\to\infty}{\to} F^* = \arg\min_{\mathcal{M}_+(\Omega)} F$
2. for a given (parameter) classical Gradient flow $(\mathbf{u}_m(t))_{m\in\mathbb{N}, t\in\mathbb{R}_+}$ which is initialized at its Wgf in $[-r_0, r_0] \times \Theta$:

$$\mu_{m,0} \underset{m\to\infty}{\overset{W_2}{\to}} \mu_0 \implies \lim_{t,m\to\infty} F(\mu_{m,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$$

via #1 & (§Many-particle limit)

Recap

In the optimization setting we devised a condition on the starting measure:

- kept throughout dynamics
- always able to escape local minima

Recap

In the optimization setting we devised a condition on the starting measure:

- kept throughout dynamics
- always able to escape local minima

Using the results of the gradient flow - Wgf correspondence we can recover the behavior with the particle version, after some unquantified m^* large enough.

Weakenesses, comments

Convergence hypothesis

- General case: weak convergence of projection, difficult to check
- Use case: W_2 convergence, difficult to hold

Weakenesses, comments

Convergence hypothesis

- General case: weak convergence of projection, difficult to check
- Use case: W_2 convergence, difficult to hold

Nature of the Assumptions

- **instrumental**: homogeneity, separation
- **technical** ☹: Sard-type regularity (SAs), difficult to check
- **reasonable** ☺: convex smooth loss and classic regularity assumptions

Weakenesses, comments

Convergence hypothesis

- General case: weak convergence of projection, difficult to check
- Use case: W_2 convergence, difficult to hold

Nature of the Assumptions

- **instrumental**: homogeneity, separation
- **technical** ☹: Sard-type regularity (SAs), difficult to check
- **reasonable** ☺: convex smooth loss and classic regularity assumptions

Result

Non quantitative, only a limit, no ϵ-bound on F.

Lecture Path

Overview

Recall the discussion on NNs from the first Section. With the results in hand we show:

1. a quite general optimization task falls under the family of problems considered
2. two layer sigmoid NNs trained with GD satisfying can be embedded in it

Overview

Recall the discussion on NNs from the first Section. With the results in hand we show:

1. a quite general optimization task falls under the family of problems considered
2. two layer sigmoid NNs trained with GD satisfying can be embedded in it

Conclusion:

Sigmoid NNs with two hidden layers with a proper initialization, converge to the global minima of their loss if they meet a some conditions

Experiments

Promising results shown at the very end on synthetic data.

Loss level requirements

Loss structure

Choose as Hilbert space $\mathcal{F} = L^2(\rho)$ for $\rho : \mathcal{X} \to \mathbb{R}$ a probability measure with $X \subset \mathbb{R}^d$

$$R(f) = \int r(x, f(x)) d\rho(x) \quad r : \mathcal{X} \times \mathbb{R} \to$$

Sufficient Loss conditions

If:

1. r convex in the second variable
2. $\exists \partial_2 r$ Lipschitz uniformly in the first variable
3. $\partial_2 r \leq C_1 r + C_2 \; C_1, C_2 > 0$

Then R is convex, $\exists dR$ Lipschitz, bounded on sublevel sets

Remark

we meet (SAs)#1.

From Optimization to Optimization as Learning

We need a learning problem to embed NNs into the framework, for this we specify:

- $\rho(x, y) =$ labels y and features x, $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$ where $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$, via disintegration [AGS05](Thm. 5.3.1) $\rho(dx \otimes dy) = \rho(dy|x)\rho_x(dx)$ where $(\rho(\cdot|x))_{x \in \mathcal{X}} = \{p.m.\ on\ \mathcal{Y}\}$.

From Optimization to Optimization as Learning

We need a learning problem to embed NNs into the framework, for this we specify:

- $\rho(x, y) =$ labels y and features x, $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$ where $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$, via disintegration [AGS05](Thm. 5.3.1) $\rho(dx \otimes dy) = \rho(dy|x)\rho_x(dx)$ where $(\rho(\cdot|x))_{x \in \mathcal{X}} = \{p.m.\ on\ \mathcal{Y}\}$.
- as loss, we use the expected risk:

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(x), y) d\rho(x, y)$$

From Optimization to Optimization as Learning

We need a learning problem to embed NNs into the framework, for this we specify:

- $\rho(x, y) =$ labels y and features x, $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$ where $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$, via disintegration [AGS05](Thm. 5.3.1) $\rho(dx \otimes dy) = \rho(dy|x)\rho_x(dx)$ where $(\rho(\cdot|x))_{x \in \mathcal{X}} = \{p.m.\ on\ \mathcal{Y}\}$.
- as loss, we use the expected risk:

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(x), y) d\rho(x, y)$$

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ a convex loss function, either square or logistic loss

From Optimization to Optimization as Learning

We need a learning problem to embed NNs into the framework, for this we specify:

- $\rho(x, y) =$ labels y and features x, $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$ where $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$, via disintegration [AGS05](Thm. 5.3.1) $\rho(dx \otimes dy) = \rho(dy|x)\rho_x(dx)$ where $(\rho(\cdot|x))_{x \in \mathcal{X}} = \{p.m.\ on\ \mathcal{Y}\}$.
- as loss, we use the expected risk:

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(x), y) d\rho(x, y)$$

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ a convex loss function, either square or logistic loss
- as r function (slightly misleading order):

$$r(x, p) = \int_{\mathbb{R}} \ell(p, y)\rho(dy|x) \quad p : \mathcal{X} \to \mathbb{R}$$

Where p stands for "predictor" and we are **integrating out** $y \in \mathcal{Y}$.

Reconciliation with Original problem

ML functional Loss

In the framework of the previous slide, we *split* the integrals:

$$R : L^2(\rho_x) \to \mathbb{R} \quad R(f) = \int_{\mathcal{X}} \int_{\mathbb{R}} \ell(f(x), y)\rho(dy|x)\rho_x(dx)$$

Meeting SA#1

For ℓ as stated, the function r coupled with the optional $\widetilde{V} = 1$ satisfies the previous sufficient conditions.

One Layer Sigmoid Neural Networks, premise

- features are in $\mathbb{R}^{d-2}$ but we add a bias term so $z = (1, x) \sim \rho_x$, and the positions θ will be in $\mathbb{R}^{d-1} = \Theta$

One Layer Sigmoid Neural Networks, premise

- features are in $\mathbb{R}^{d-2}$ but we add a bias term so $z = (1, x) \sim \rho_x$, and the positions θ will be in $\mathbb{R}^{d-1} = \Theta$
- focus on Neural Networks with one hidden layer

One Layer Sigmoid Neural Networks, premise

- features are in $\mathbb{R}^{d-2}$ but we add a bias term so $z = (1, x) \sim \rho_x$, and the positions θ will be in $\mathbb{R}^{d-1} = \Theta$
- focus on Neural Networks with one hidden layer
- the functions we saw at the beginning is then

$$\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(z \cdot \theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \underbrace{\theta_{d-1}}_{bias}\right) \quad \widetilde{V} = 1$$

the hidden layer particles implement this function

One Layer Sigmoid Neural Networks, premise

- features are in $\mathbb{R}^{d-2}$ but we add a bias term so $z = (1, x) \sim \rho_x$, and the positions θ will be in $\mathbb{R}^{d-1} = \Theta$
- focus on Neural Networks with one hidden layer
- the functions we saw at the beginning is then

$$\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(z \cdot \theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \underbrace{\theta_{d-1}}_{bias}\right) \quad \widetilde{V} = 1$$

  the hidden layer particles implement this function
- σ is a **sigmoid**

One Layer Sigmoid Neural Networks in a nuthsell

Simplifying the dependence on $u = (w, \theta)$ which is implicitly present:

$$h(x) = \mathbf{w}^T \sigma(\boldsymbol{\theta}^T x)) = \sum_{i=1}^{m} w_i \cdot \sigma(\boldsymbol{\theta}(\cdot, i)^T x) \quad (14)$$

Where:

- m is the number of hidden neurons
- w_i is the outgoing weight of the i^{th} neuron
- $\boldsymbol{\theta}(\cdot, i)$ are the ingoing weights of the i^{th} neuron.

One Layer Sigmoid Neural Networks in a nuthsell

Simplifying the dependence on $u = (w, \theta)$ which is implicitly present:

$$h(x) = \mathbf{w}^T \sigma(\boldsymbol{\theta}^T x)) = \sum_{i=1}^{m} w_i \cdot \sigma(\boldsymbol{\theta}(\cdot, i)^T x) \quad (14)$$

Where:

- m is the number of hidden neurons
- w_i is the outgoing weight of the i^{th} neuron
- $\boldsymbol{\theta}(\cdot, i)$ are the ingoing weights of the i^{th} neuron.

Remark, on the one hidden layer structure

Formulation of Eqn. (14) interesting since:

- there is total independence of contributions, a linear combination of hidden neurons
- more layers do not have this peculiarity

One Sigmoid Layer Neural Networks graphically

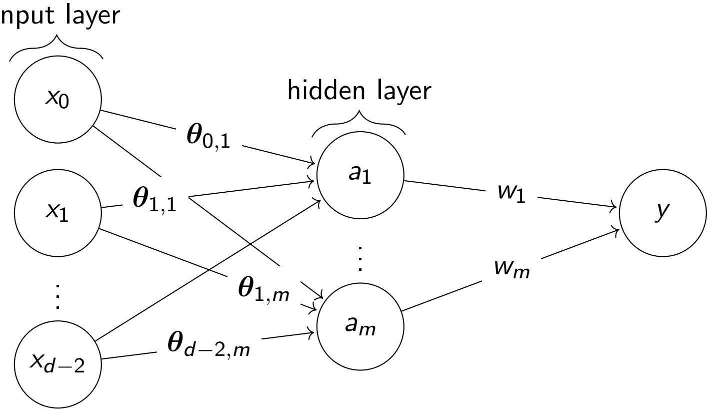


Figure: The diagram shows an intuitive representation of a two layer neural network. The inputs are $d-2$ dimensional, with an added bias. They are passed to activations a_i of the form $a_i(x) = \sigma(\boldsymbol{\theta}(\cdot, i)^T x)$. The final output is then determined by a weighted sum of activations.

Particle function level requirements

Aim

To embed Sigmoid NNs into (SAs),
ϕ and ρ_x need to have a structure.

Particle function level requirements

Aim

To embed Sigmoid NNs into (SAs), ϕ and ρ_x need to have a structure.

Remark

(SAs)#3-a boundary regularity assumed a priori as it is difficult to check

Particle function level requirements

Aim

To embed Sigmoid NNs into (SAs), ϕ and ρ_x need to have a structure.

Remark

(SAs)#3-a boundary regularity assumed a priori as it is difficult to check

Sufficient ϕ conditions

1. (SAs)#1 if ρ has finite 4^{th} moment then ϕ is differentiable with $d\phi_\theta$ Lipschitz (and known)
2. (SAs)#2 regularity condition if ρ has finite moments of order $2d-2$

One layer Sigmoid NNs Framework

Setting

Data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}, y_i \in \mathcal{Y} \subset \mathbb{R}$, unknown distribution $\rho(x, y)$.
Problem: in the form of (6)

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu \right) + G(\mu)$$

One layer Sigmoid NNs Framework

Setting

Data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}, y_i \in \mathcal{Y} \subset \mathbb{R}$, unknown distribution $\rho(x, y)$.
Problem: in the form of (6)

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu \right) + G(\mu)$$

Where:

- $\Theta = \mathbb{R}^{d-1}$
- $\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(\sum_{i=1}^{d-2} x_i \theta_i + \theta_{d-1})$
- R, risk of quadratic or logistic loss ℓ with functional loss sufficient assumptions
- G, total variation norm $G(\mu) = |\mu|(\Theta)$

One layer Sigmoid NNs Framework

Setting

Data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}, y_i \in \mathcal{Y} \subset \mathbb{R}$, unknown distribution $\rho(x, y)$.
Problem: in the form of (6)

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu\right) + G(\mu)$$

Where:

- $\Theta = \mathbb{R}^{d-1}$
- $\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(\sum_{i=1}^{d-2} x_i\theta_i + \theta_{d-1})$
- R, risk of quadratic or logistic loss ℓ with functional loss sufficient assumptions
- G, total variation norm $G(\mu) = |\mu|(\Theta)$

One layer Sigmoid NNs Framework

Setting

Data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}, y_i \in \mathcal{Y} \subset \mathbb{R}$, unknown distribution $\rho(x, y)$.
Problem: in the form of (6)

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) := R\left(\int \phi d\mu \right) + G(\mu)$$

Where:

- $\Theta = \mathbb{R}^{d-1}$
- $\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(\sum_{i=1}^{d-2} x_i \theta_i + \theta_{d-1})$
- R, risk of quadratic or logistic loss ℓ with functional loss sufficient assumptions
- G, total variation norm $G(\mu) = |\mu|(\Theta)$

One layer Sigmoid NNs convergence to Global Minimizers

Meta-Theorem, Wgf

Assume:

- (function (SAs)) $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$ has moments that are finite up to $\max\{4, 2d-2\}$
- (separation) $\operatorname{spt} \mu_0 = \{0\} \times \Theta$
- (boundary Sard) the condition of (SAs) #3-(a) is verified

Then a Wgf for the Problem $(\mu_t)_{t \in \mathbb{R}_+}$ is such that:

$$\mu_t \xrightarrow{W_2} \mu_\infty \implies \mu_\infty = \arg\min F$$

One layer Sigmoid NNs convergence to Global Minimizers

Meta-Theorem, particle gradient descent

Measure $\nu \in \mathcal{M}(\Theta)$ corresponding to $\mu \in \mathcal{P}(\Omega)$ finite particle dynamics:

$$\lim_{m,t\to\infty} J(\mu_{m,t}) = J^* \quad \mu_{m,t} = \frac{1}{m}\sum_{i=1}^{m} w_i^{(m)}(t)\delta_{\theta_i^{(m)}(t)}$$

are guaranteed to converge at some non-identified m^* to the global minima of J. The convergence is independent of the order of m, t, and we could simply increase the number of particles and let them flow in t until convergence

One layer Sigmoid NNs convergence to Global Minimizers

Meta-Theorem, particle gradient descent

Measure $\nu \in \mathcal{M}(\Theta)$ corresponding to $\mu \in \mathcal{P}(\Omega)$ finite particle dynamics:

$$\lim_{m,t\to\infty} J(\mu_{m,t}) = J^* \quad \mu_{m,t} = \frac{1}{m}\sum_{i=1}^{m} w_i^{(m)}(t)\delta_{\theta_i^{(m)}(t)}$$

are guaranteed to converge at some non-identified m^* to the global minima of J. The convergence is independent of the order of m, t, and we could simply increase the number of particles and let them flow in t until convergence

Theorem in words

For a sigmoid NN learning task, gradient descent, feasible in practice & widely used, converges to the global minima

Fixed number of particles dynamics

- $d = 2$
- dotted lines are global minimizer
- m fixed
- $\theta(0)$ Gaussian satisfies separation asymptotically [CB18] and is the *de facto choice in practice* [Bac20a]

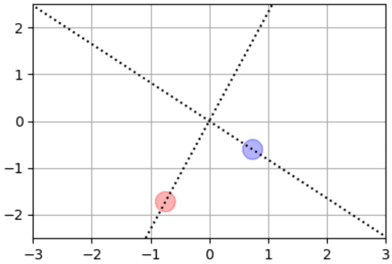
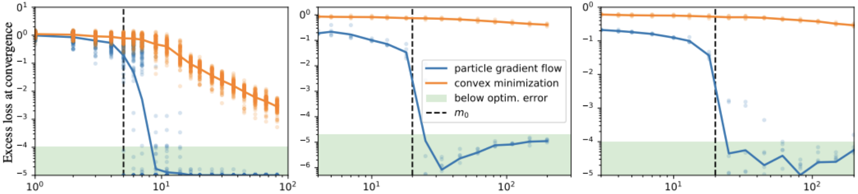


Figure: Sigmoid Dynamics. For more context, see the original source [CB18]

Performance



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

Figure: Particle-complexity, excess loss. For more context, see the original publication source [CB18].

Non quantitative results, but better performance VS the naïve convex optimization method.

Lecture Path

Recap

Results in [CB18] make use of:

- Wasserstein Gradient Flows
- analogy to Mean-field limit
- thougthful general problem construction

Recap

Results in [CB18] make use of:

- Wasserstein Gradient Flows
- analogy to Mean-field limit
- thougthful general problem construction

to show:

- that Sigmoid Neural Networks fall under the umbrella of problems that can be tuned to reach a global minimizer.
- good experimental results
- that the framework covers other cases (see [CB18]).

Recap

Results in [CB18] make use of:

- Wasserstein Gradient Flows
- analogy to Mean-field limit
- thougthful general problem construction

to show:

- that Sigmoid Neural Networks fall under the umbrella of problems that can be tuned to reach a global minimizer.
- good experimental results
- that the framework covers other cases (see [CB18]).

Pros

- ☺ gradient descent
- ☺ theoretical results
- ☺ mostly reasonable assumptions

Recap

Weaknesses

- ☹ non quantitative convergence
- ☹ Boundary Sard assumed
- ☹ Wgf convergence assumed

Additional/important refs:

- gradient flows on metric spaces book [AGS05]
- another NNs theory paper [MMN18]
- blog and paper (by authors) [Chi20; COB20].

Open Problems

- Promising results for Wgf convergence [BSR15; HM19]
- bigger networks adaptation
- quantitative result [MMM19]

Concluding

Any question/discussion, let me know!

Thank you!

simonegiancola09@gmail.com

personal webpage

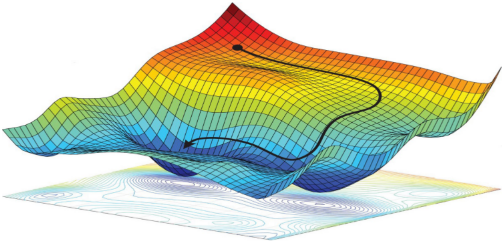


Figure: Source blog post

References I

[CB18] Lenaic Chizat and Francis Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport*. Oct. 29, 2018. DOI: 10.48550/arXiv.1805.09545. arXiv: 1805.09545 [cs, math, stat]. URL: http://arxiv.org/abs/1805.09545 (visited on 11/20/2022).

[Ins19] Institut Henri Poincaré, director. *On the Global Convergence of Gradient Descent for (...) - Bach - Workshop 3 - CEB T1 2019*. May 10, 2019. URL: https://www.youtube.com/watch?v=rQM5uh2EsHA (visited on 12/16/2022).

References II

[Bac20a] Francis Bach. *Gradient Descent for Wide Two-Layer Neural Networks – I : Global Convergence – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-descent-neural-networks-global-convergence/ (visited on 12/16/2022).

[Chi20] Lenaic Chizat. *Gradient Descent for Wide Two-Layer Neural Networks – II: Generalization and Implicit Bias – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-descent-for-wide-two-layer-neural-networks-implicit-bias/ (visited on 12/16/2022).

References III

[Jin+18] Chi Jin et al. "On the Local Minima of the Empirical Risk". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/da4902cb0bc38210839714ebdcf0efc3-Abstract.html (visited on 11/21/2022).

[Lee+] Jason D Lee et al. "Gradient Descent Only Converges to Minimizers". In: (), p. 12.

[Cho+15] Anna Choromanska et al. *The Loss Surfaces of Multilayer Networks*. Jan. 21, 2015. DOI: 10.48550/arXiv.1412.0233. arXiv: 1412.0233 [cs]. URL: http://arxiv.org/abs/1412.0233 (visited on 11/21/2022).

References IV

[SJL22] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. *Theoretical Insights into the Optimization Landscape of Over-Parameterized Shallow Neural Networks*. Aug. 23, 2022. DOI: 10.48550/arXiv.1707.04926. arXiv: 1707.04926 [cs, math, stat]. URL: http://arxiv.org/abs/1707.04926 (visited on 11/20/2022).

[JK17] Prateek Jain and Purushottam Kar. "Non-Convex Optimization for Machine Learning". In: *Foundations and Trends® in Machine Learning* 10.3-4 (2017), pp. 142–336. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000058. arXiv: 1712.07897 [cs, math, stat]. URL: http://arxiv.org/abs/1712.07897 (visited on 11/21/2022).

References V

[BSR15] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. *The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems*. July 6, 2015. arXiv: 1507.01562 [math]. URL: http://arxiv.org/abs/1507.01562 (visited on 11/19/2022).

[Wan+15] Chu Wang et al. *Functional Frank-Wolfe Boosting for General Loss Functions*. Oct. 8, 2015. DOI: 10.48550/arXiv.1510.02558. arXiv: 1510.02558 [cs, stat]. URL: http://arxiv.org/abs/1510.02558 (visited on 11/20/2022).

References VI

[BP13] Kristian Bredies and Hanna Katriina Pikkarainen. "Inverse Problems in Spaces of Measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), pp. 190–218. ISSN: 1262-3377. DOI: 10.1051/cocv/2011205. URL: http://www.numdam.org/item/COCV_2013__19_1_190_0/ (visited on 11/19/2022).

[Jag13] Martin Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Feb. 13, 2013, pp. 427–435. URL: https://proceedings.mlr.press/v28/jaggi13.html (visited on 11/19/2022).

References VII

[Bac16] Francis Bach. *Breaking the Curse of Dimensionality with Convex Neural Networks*. Oct. 31, 2016. DOI: 10.48550/arXiv.1412.8690. arXiv: 1412.8690 [cs, math, stat]. URL: http://arxiv.org/abs/1412.8690 (visited on 11/19/2022).

[Las09] Jean Bernard Lasserre. *Moments, Positive Polynomials and Their Applications*. Vol. 1. Series on Optimization and Its Applications. IMPERIAL COLLEGE PRESS, Oct. 2009. ISBN: 978-1-84816-445-1 978-1-84816-446-8. DOI: 10.1142/p665. URL: https://www.worldscientific.com/worldscibooks/10.1142/p665 (visited on 11/20/2022).

References VIII

[CDP17] Paul Catala, Vincent Duval, and Gabriel Peyré. "A Low-Rank Approach to Off-The-Grid Sparse Deconvolution". In: *Journal of Physics: Conference Series* 904 (Oct. 2017), p. 012015. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/904/1/012015. arXiv: 1712.08800 [cs, math]. URL: http://arxiv.org/abs/1712.08800 (visited on 11/19/2022).

[NS17] Atsushi Nitanda and Taiji Suzuki. *Stochastic Particle Gradient Descent for Infinite Ensembles*. Dec. 14, 2017. DOI: 10.48550/arXiv.1712.05438. arXiv: 1712.05438 [cs, math, stat]. URL: http://arxiv.org/abs/1712.05438 (visited on 11/20/2022).

References IX

[LY17] Yuanzhi Li and Yang Yuan. *Convergence Analysis of Two-layer Neural Networks with ReLU Activation*. Nov. 1, 2017. DOI: 10.48550/arXiv.1705.09886. arXiv: 1705.09886 [cs]. URL: http://arxiv.org/abs/1705.09886 (visited on 11/20/2022).

[SH17] Daniel Soudry and Elad Hoffer. *Exponentially Vanishing Sub-Optimal Local Minima in Multilayer Neural Networks*. Oct. 28, 2017. DOI: 10.48550/arXiv.1702.05777. arXiv: 1702.05777 [stat]. URL: http://arxiv.org/abs/1702.05777 (visited on 11/20/2022).

References X

[VBB20] Luca Venturi, Afonso S. Bandeira, and Joan Bruna. *Spurious Valleys in Two-layer Neural Network Optimization Landscapes*. June 16, 2020. DOI: 10.48550/arXiv.1802.06384. arXiv: 1802.06384 [cs, math, stat]. URL: http://arxiv.org/abs/1802.06384 (visited on 11/20/2022).

[KY03] Harold Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Vol. 35. Stochastic Modelling and Applied Probability. New York: Springer-Verlag, 2003. ISBN: 978-0-387-00894-3. DOI: 10.1007/b97441. URL: http://link.springer.com/10.1007/b97441 (visited on 11/20/2022).

References XI

[Sci+17] Damien Scieur et al. *Integration Methods and Accelerated Optimization Algorithms*. Feb. 22, 2017. DOI: 10.48550/arXiv.1702.06751. arXiv: 1702.06751 [math]. URL: http://arxiv.org/abs/1702.06751 (visited on 11/20/2022).

[Bac20b] Francis Bach. *Effortless Optimization through Gradient Flows – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-flows/ (visited on 12/26/2022).

References XII

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savarè. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Basel: Birkhäuser-Verlag, 2005. ISBN: 978-3-7643-2428-5. DOI: 10.1007/b137080. URL: http://link.springer.com/10.1007/b137080 (visited on 11/20/2022).

[Chi21] Lénaïc Chizat. "Analysis of Gradient Descent on Wide Two-Layer ReLU Neural Networks". In: (2021).

References XIII

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. "A Mean Field View of the Landscape of Two-Layer Neural Networks". In: *Proceedings of the National Academy of Sciences* 115.33 (Aug. 14, 2018), E7665–E7671. DOI: 10.1073/pnas.1806579115. URL: https://www.pnas.org/doi/10.1073/pnas.1806579115 (visited on 11/20/2022).

[COB20] Lenaic Chizat, Edouard Oyallon, and Francis Bach. *On Lazy Training in Differentiable Programming*. Jan. 7, 2020. DOI: 10.48550/arXiv.1812.07956. arXiv: 1812.07956 [cs, math]. URL: http://arxiv.org/abs/1812.07956 (visited on 11/21/2022).

References XIV

[HM19] Daniel Hauer and José Mazon. *Kurdyka-Lojasiewicz-Simon Inequality for Gradient Flows in Metric Spaces*. Jan. 24, 2019. arXiv: 1707.03129 [math]. URL: http://arxiv.org/abs/1707.03129 (visited on 11/19/2022).

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Mean-Field Theory of Two-Layers Neural Networks: Dimension-Free Bounds and Kernel Limit". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Conference on Learning Theory. PMLR, June 25, 2019, pp. 2388–2464. URL: https://proceedings.mlr.press/v99/mei19a.html (visited on 01/07/2023).