On the global convergence of gradient descent using optimal transport

L. Chizat, F. Bach, NeurIPS 2018

Matilde Dolfato

Introduction to Real Analysis II, Prof. G. Savaré

January 16, 2026

Overview

Introduction

Classical task in machine learning:

$$\min_{\mu \in \mathscr{P}(\Theta)} j(\mu) \qquad j(\mu) = \ell\left(\int \phi \, \mathrm{d}\mu \right) + g(\mu) \tag{1}$$

$\ell : \mathcal{F} \to \mathbb{R}_+$ smooth, convex *loss function*
data lives in a large parametrized set with parameters $p \in \Theta \subset \mathbb{R}^d$
$g : \mathscr{P}(\Theta) \to \mathbb{R}$ optional convex *regularizer*

we look for a linear combination ϕ mapping features to labels
$\phi(p) : x \mapsto \sigma(\sum_i p_i x_i + p_d)$, $p \in \Theta$

Introduction

Classical task in machine learning:

$$\min_{\mu \in \mathscr{P}(\Theta)} j(\mu) \qquad j(\mu) = \ell\left(\int \phi \, \mathrm{d}\mu\right) + g(\mu) \tag{1}$$

$\ell : \mathcal{F} \to \mathbb{R}_+$ smooth, convex *loss function*
data lives in a large parametrized set with parameters $p \in \Theta \subset \mathbb{R}^d$
$g : \mathscr{P}(\Theta) \to \mathbb{R}$ optional convex *regularizer*

we look for a linear combination ϕ mapping features to labels
$\phi(p) : x \mapsto \sigma(\sum_i p_i x_i + p_d)$, $p \in \Theta$

$\star$ Minimize over the space of measures $\mathscr{P}(\Theta)$ $\star$

Introduction

$$\min_{\mu \in \mathscr{P}(\Theta)} j(\mu) = \ell\left(\int \phi \, \mathrm{d}\mu \right) + g(\mu) \tag{1}$$

Convex problem, but intractable

Introduction

$$\min_{\mu \in \mathscr{P}(\Theta)} j(\mu) = \ell\left(\int \phi \, \mathrm{d}\mu \right) + g(\mu) \tag{1}$$

Convex problem, but intractable

$\Rightarrow$ discretize the measure into m particles parametrized by *weights* and *positions*

$$\mu_m := \frac{1}{m} \sum_{i=1}^{m} w_i \delta_{p_i}$$

$$\min_{\substack{w \in \mathbb{R} \\ p \in \Theta^m}} j_m(w,p) \qquad j_m(w,p) := j(\mu_m) \tag{2}$$

Introduction

$$\min_{\mu \in \mathscr{P}(\Theta)} j(\mu) = \ell\left(\int \phi \, \mathrm{d}\mu \right) + g(\mu) \tag{1}$$

Convex problem, but intractable
$\Rightarrow$ discretize the measure into m particles parametrized by *weights* and *positions*

$$\mu_m := \frac{1}{m} \sum_{i=1}^{m} w_i \delta_{p_i}$$

$$\min_{\substack{w \in \mathbb{R}^m \\ p \in \Theta^m}} j_m(w,p) \qquad j_m(w,p) := j(\mu_m) \tag{2}$$

Non-convex problem
$\star$ Idea: use tools from optimal transport to exploit convexity of (1) to study global convergence in (2) $\star$

Conceptual key points

1. Problem reformulation (lifting) $j \to f$
2. To be able to study gradient flow of f_m, $(\mu_{m,t})_t$
3. Look at it as a particular case of infinite-dimensional case

$$\text{as } m \to \infty,\ (\mu_{m,t})_t \to (\mu_t)_t$$

4. Exploit convex structure of inf-dim case

$$\lim_{m,t\to\infty} f(\mu_{m,t}) = \min_{\mu_t \in \mathscr{P}(\Omega)} f(\mu_t)$$

Overview

Lifting

$$\min_{\nu \in \mathscr{P}(\Theta)} j(\nu) = \ell\left(\int \phi \, \mathrm{d}\nu\right) + g(\nu) \tag{1}$$

$$\downarrow \; h_1^{-1}$$

$$\min_{\mu \in \mathscr{P}(\Omega)} f(\mu) = \ell\left(\int \varphi \, \mathrm{d}\mu\right) + \int v \, \mathrm{d}\mu \tag{3}$$

Recovering the lifting:

- $\Omega = \Theta \times \mathbb{R}$
- $\varphi(p, w) = w\phi(p)$
- projection map: $h_1 : \Theta \times \mathbb{R} \to \Theta$, $h_1(\mu)(p) = \int_{\mathbb{R}} w\mu(\mathrm{d}w, p)$
- $g(\nu) = \inf_{\mu \in h_1^{-1}(\nu)} \int v \, \mathrm{d}\mu$
- then, $\inf_\nu j = \inf_\mu f$

New problem

$$\min_{\mu \in \mathscr{P}(\Omega)} f(\mu) = \ell\left(\int \varphi \, \mathrm{d}\mu \right) + \int v \, \mathrm{d}\mu$$

Discretized problem:

$$\min_{u \in \Omega} f_m(u) := f\Bigg(\underbrace{\frac{1}{m} \sum_{i=1}^{m} \delta_{u_i}}_{\mu_m} \Bigg) = \ell\left(\frac{1}{m} \sum_{i=1}^{m} \varphi(u_i) \right) + \frac{1}{m} \sum_{i=1}^{m} v(u_i)$$

$\star$ Weights w are another coordinate of a particle position $u \in \Omega$
$\Rightarrow$ study gradient flow of f_m $\star$

New problem

$$\min_{\mu \in \mathscr{P}(\Omega)} f(\mu) = \ell\left(\int \varphi \, \mathrm{d}\mu \right) + \int v \, \mathrm{d}\mu$$

Discretized problem:

$$\min_{u \in \Omega} f_m(u) := f\Bigg(\underbrace{\frac{1}{m} \sum_{i=1}^{m} \delta_{u_i}}_{\mu_m} \Bigg) = \ell\left(\frac{1}{m} \sum_{i=1}^{m} \varphi(u_i) \right) + \frac{1}{m} \sum_{i=1}^{m} v(u_i)$$

$\star$ Weights w are another coordinate of a particle position $u \in \Omega$
$\Rightarrow$ study gradient flow of f_m $\star$

Assumptions (high-level):

- ℓ is smooth, with Lispchitz and bounded differential $\mathrm{d}\ell$
- ϕ is differentiable and v is semiconvex
- there is a family of nested closed convex sets $(Q_r)_r$ on which the (sub)derivatives of ϕ, v are Lipschitz and grow sublinearly

Overview

Particle gradient flow I

Definition (Particle gradient flow)

A gradient flow for f_m is an absolutely continuous function $u : \mathbb{R}_+ \to \Omega^m$ such that

$$u'(t) \in -m\,\partial f_m(u(t))$$

for almost every $t \in \mathbb{R}_+$.

Properties:

(i) (existence and uniqueness) for any $u(0) \in \Omega^m$ starting point, there exists a unique gradient flow for f_m

(ii) (derivative of f_m) for a.e. $t \in \mathbb{R}_+$, it holds

$$\frac{\mathrm{d}}{\mathrm{d}s} f_m(u(s)) \Big|_{s=t} = -|u'(t)|^2$$

Particle gradient flow II

(iii) (form of the velocity) the velocity *of the i-th particle* $u_i(t)$ is a vector field $v_t : \Omega \to \mathbb{R}^d$ given by $u_i'(t) = v_t(u_i(t))$ where [2]

$$v_t(u_i) = \tilde{v}_t(u_i) - \mathrm{proj}_{\partial v(u_i)}(\tilde{v}_t(u_i))$$

$$\text{with} \quad \tilde{v}_t(u_i) = -\left[\left(\ell'\left(\int \varphi \, \mathrm{d}\mu_{m,t}\right), \partial_j \varphi(u_i)\right)\right]_{j=1}^{d}$$

from [2], gradient flow selects subgradients of minimal norm

Observations:

- velocity is the evaluation at each u_i of the same vector field v_t
- given an initialization $u(0) \in \Omega^m$, this defines an atomic measure $\mu_{m,0}$

$\longrightarrow$ makes sense to generalize to arbitrary measures μ_t

Generalize to Wasserstein gradient flow I

Since

1. Evolution of $(\mu_t)_t$ under $(v_t)_t$ satisfies:

$$\partial_t \mu_t = -\mathrm{div}(v_t \mu_t)$$

2. Link between v_t and f:

$$v_t \in -\partial f'(\mu_{m,t})$$

where $f'(\mu)(u) := \left(\ell\left(\int \varphi \, \mathrm{d}\mu\right), \varphi(u)\right) + v(u)$

$\Rightarrow$ we expect $(\mu_t)_t$ is a gradient flow on the space $\mathscr{P}_2(\Omega)$: *Wasserstein gradient flow*

Generalize to Wasserstein gradient flow II

Definition (Wasserstein gradient flow)

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 $\mathscr{P}_2(\Omega)$ that satisfies, distributionally on $[0,T[\times\Omega^d$,

$$\partial_t \mu_t = -\mathrm{div}(v_t \mu_t) \quad \textit{where } v_t \in -\partial f'(\mu_t)$$

Generalize to Wasserstein gradient flow II

Definition (Wasserstein gradient flow)

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 $\mathscr{P}_2(\Omega)$ that satisfies, distributionally on $[0, T[\times\Omega^d$,

$$\partial_t \mu_t = -\mathrm{div}(v_t \mu_t) \quad \textit{where } v_t \in -\partial f'(\mu_t)$$

$\rightarrow$ It is well-defined starting from μ_0 concentrated on a convex closed subset of Ω [1]

Generalize to Wasserstein gradient flow II

Definition (Wasserstein gradient flow)

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 $\mathscr{P}_2(\Omega)$ that satisfies, distributionally on $[0, T[\times\Omega^d$,

$$\partial_t \mu_t = -\mathrm{div}(v_t \mu_t) \quad \textit{where } v_t \in -\partial f'(\mu_t)$$

$\rightarrow$ It is well-defined starting from μ_0 concentrated on a convex closed subset of Ω [1]

W_2 gradient flow generalizes particle

Whenever $(u_t)_t$ is a gradient flow for f_m, $t \mapsto \mu_{m,t} := \sum_{i=1}^{m} \delta_{u_i(t)}$ is a Wasserstein gradient flow for f!

Many-particle limit

Theorem (Many particle limit)

Consider $(t \mapsto u_m(t))_{m\in\mathbb{N}}$ *a sequence of classical gradient flows for* f_m *initialized in a closed convex set. If* $\mu_{m,0}$ *converges to some* $\mu_0 \in \mathscr{P}_2(\Omega)$ *for 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*.*

In a nutshell:

- if $u(0) = (u_1(0), \dots, u_m(0))$ are distributed according to μ_0, then $\mu_{m,0}$ converges to μ_0 by the law of large numbers
- $\lim_{m\to\infty}(\mu_{m,t})_t = (\mu_t)_t$

Global convergence result

Structural assumptions:

- φ, v are **2-homogeneous** (ϕ 1-homogeneous)
- the support of the initialization of the Wasserstein gradient flow satisfies a **"separation" property**: $B(0, r_b) \subset \mathbb{S}^{d-1}$ that separates $r_a \mathbb{S}^{d-1}$ and $r_b \mathbb{S}^{d-1}$ for $r_a < r_b$

Theorem (Global convergence of particle gradient descent)

Let $(\mu_t)_{t \geq 0}$ be a Wasserstein gradient flow of f such that the support of μ_0 is $S_0 \subset [-r_0, r_0] \times \Theta$ satisfies a separation property. If $(\mu_t)_t$ converges to μ_∞ in W_2, then μ_∞ is a global minimizer of f over $\mathscr{P}(\Omega)$. In particular, if $(u_m(t))_{m \in \mathbb{N}, t \geq 0}$ is a sequence of classical gradient flows initialized in $[-r_0, r_0] \times \Theta$ such that $\mu_{m,0}$ converges to μ_0 in W_2 then

$$\lim_{t,m \to \infty} f(\mu_{m,t}) = \min_{\mu \in \mathscr{P}(\Omega)} f(\mu).$$

Application to ReLu neural networks

Setting:

- features live in $\mathbb{R}^{d-2}$, labels in $\mathbb{R}$
- ℓ is either the square or logistic loss
- 2-homogeneous case
- domain Θ is the disjoint union of 2 copies of $\mathbb{R}^d$
- $\varphi(p) : x \mapsto \sigma(\sum_{i=1}^{d-1} s(p_i)x_i + s(p_d)), \quad s(p_i) = p_i|p_i|$
- regularizer: $v(p) = |p|^2$ (as if $w = |p|$)

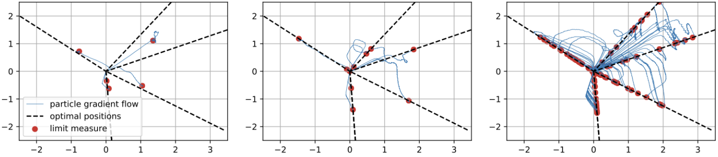

Figure: Training neural network with ReLU activation. Overfitting threshold $m = 4$. Failure for $m = 5$, success $m = 10, 100$.

Final remarks

- importance of initialization, as confirmed by extensive empirical literature
- particle gradient flow corresponds to continuous-time gradient descent: what can we say about discrete-step case? (double descent)

Thank you!

References I

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré.
Gradient flows: in metric spaces and in the space of probability measures.
Springer, 2005.

Filippo Santambrogio.
{Euclidean, metric, and Wasserstein} gradient flows: an overview.
Bulletin of Mathematical Sciences, 7(1):87–154, 2017.

Recovering the lifting of (3)

1. (equivalence of $\mathscr{M}$ and $\mathscr{P}$ under homogeneity)
2. surjectivity of h_1
3. define φ, g as above and prove equality of f, j

Prerequisits

Definition (Subgradient and subdifferential)

Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *subgradient of* f *at* x_0, $x_0 \in \mathbb{R}^d$, is $p \in \mathbb{R}^d$:

$$f(x) \geq f(x_0) + p \cdot (x - x_0) + o(x - x_0) \qquad \forall x \in \mathbb{R}^d$$

The set of all such ps is called the *subdifferential* of f at x_0 and we write $\partial f(x_0)$. The subdifferential is a closed and convex set.

Definition (Gradient flow)

A function $x : \mathbb{R}_+ \to \mathrm{Dom}(f)$ is a *gradient flow* of $f : \mathrm{Dom}(f) = \mathbb{R}^d \to \mathbb{R}$ if x is absolutely continuous over $\mathbb{R}_+$ and

$$x'(t) \in -\partial f(x(t)) \qquad \text{for a.e. } t \in \mathbb{R}_+$$