Topics in Math of ML, Lecture 10
Algorithmic regularization (II) : GD and non-linear parameterizations

Lénaïc Chizat*

November 24, 2025

1 Introduction

We consider the following setting:

- A linear space $\mathcal{F}$ of predictors $\mathbb{R}^d \to \mathbb{R}$. For simplicity, we take $\ell_2(\{x_1, \dots, x_n\}) \equiv \mathbb{R}^n$;
- A parameterized predictor $\theta \in \mathbb{R}^p \mapsto f_\theta \in \mathcal{F}$ (smooth)
- An objective function $R : \mathbb{R}^n \to \mathbb{R}$ (smooth, convex, lower bounded)

This leads to the following parameterized objective $F : \mathbb{R}^p \to \mathbb{R}$:

$$F(\theta) = R(f_\theta).$$

Simplest example Linear predictor $f_\theta(x) = x^\top \theta$ with square loss $R(f) = \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2$ where $(x_i, y_i)_{i=1}^n$ is the training set.

In Lecture 9, we motivated and studied the algorithmic regularization of gradient descent (GD). We have seen that for overparameterized linearly-parameterized models, GD converges:

- to the min ℓ_2-norm interpolator with the square loss;
- to the max ℓ_2-margin classifier with the logistic loss.

This week we turn our interest towards non-linear parameterizations.

2 A geometric remark (pushforward metric)

- Gradient Descent (GD) with step-size $\eta > 0$ induces the following dynamics in parameter space, starting from some $\theta(0) \in \mathbb{R}^p$:

$$\theta(t+1) = \theta(t) - \eta \nabla F(\theta(t)) \tag{1}$$
$$= \theta(t) - \eta Df(\theta(t))^\top \nabla R(f(\theta(t))) \tag{2}$$

  where $Df(\theta) \in \mathbb{R}^{n \times p}$ is the Jacobian of f at $\theta \in \mathbb{R}^p$.

*EPFL lenaic.chizat@epfl.ch

- The corresponding dynamics in predictor space, denoting $f_t = f_{\theta(t)}$ is

$$\begin{aligned} f_{t+1} &= f_t - \eta Df(\theta(t))Df(\theta(t))^\top \nabla R(f_t) + o(\eta) \\ &= f_t - \eta K(\theta(t))\nabla R(f_t) + o(\eta) \end{aligned}$$

  where $K(\theta) = Df(\theta)Df(\theta)^\top \in \mathbb{R}^{n\times n}$ is the so-called *tangent kernel* at θ.

Remark 2.1. *If $f : \mathbb{R}^p \to \mathrm{Im}(f)$ is a diffeomorphism, this dynamics is (up to a $o(\eta)$ term) Riemmannian GD:*

$$f_{t+1} = f_t - \eta G_{f_t}^{-1} \nabla R(f_t) + o(\eta)$$

where $G_{f_\theta} = K(\theta)^{-1}$ is called the pushforward metric *(of ℓ^2 on $\mathbb{R}^p$ via f). In general (e.g. for neural networks) $\theta \mapsto f_\theta$ is not a diffeomorphism because of a lack of smoothness and of injectivity, but the fact that a parameterization induces a certain "geometry" on function space is a useful point of view.*

In the rest of this lecture, we exploit this remark to understand the algorithmic regularization of GD in the case of non-linearly parameterized models.

3 Mirror parameterizations

Consider the following non-linearly parameterized linear model (this is a toy example)

$$f_\theta(x) = (\theta \odot \theta)^\top x = \sum_{j=1}^{d} \theta_j^2 x_j$$

for $\theta \in \mathbb{R}_+^d$. The corresponding class of predictor is just linear predictors with nonnegative weights (we could extend this to all linear predictors by including the features $(-x_j)_{j=1}^d$). This parameterization leads to the objective (with a slight change of notations):

$$F(\theta) \coloneqq R(\theta \odot \theta)$$

with $R(w) = \frac{1}{n}\sum_{i=1}^n \ell(w^\top x_i, y_i)$. Remember (from Lecture 9) that if the parameterization was linear GD would select the min-ℓ_2-norm solution in this context. What happens here?

- The GD iterates with step-size $\eta > 0$ read

$$\theta(t+1) = \theta(t) - 2\eta\theta(t) \odot \nabla R(\theta(t) \odot \theta(t)) \tag{3}$$

- For simplicity and to get rid of the annoying $o(\eta)$ terms, let us consider the continuous-time Gradient Flot (GF) dynamics instead:

$$\frac{d}{dt}\theta(t) = -2\theta(t) \odot \nabla R(\theta(t) \odot \theta(t))$$

- Now let us look the dynamics in predictor space. We pose $w(t) = \theta(t) \odot \theta(t)$ and we get

$$\begin{aligned} \frac{d}{dt}w(t) &= 2\theta(t) \odot \frac{d}{dt}\theta(t) \\ &= -4w(t) \odot \nabla R(w(t)) \\ &= -4[\nabla^2 \varphi(w(t))]^{-1} \nabla R(w(t)) \end{aligned}$$

where for $w \in \mathbb{R}^d_{++}$

$$\varphi(w) = \sum_{j=1}^{d} w_j \log(w_j) - w_j + 1 \tag{4}$$

$$\nabla\varphi(w) = [\log(w_j)]_{j=1}^d \tag{5}$$

$$\nabla^2\varphi(w) = \mathrm{diag}((w_j^{-1})_{j=1}^d) \tag{6}$$

Up to a simple change of time unit (accounting for the factor 4), we thus have a Riemmannian gradient flow with a Hessian metric $w \mapsto [\nabla^2\varphi(w)]$.

Link with mirror descent (MD) In exercises, we studied the MD algorithm. To minimize $R : \mathbb{R}^d \to \mathbb{R}$ over a domain $\mathcal{D}$, this algorithm uses a mirror map $\nabla\varphi : \mathcal{D} \to \mathbb{R}^d$ (assumed to be a homeomorphism, and the gradient of a convex function φ) and consists in the sequence

$$\nabla\varphi(w(t+1)) = \nabla\varphi(w(t)) - \eta\nabla R(w(t)). \tag{7}$$

Under basic assumptions on R (e.g. convex, smooth, lower bounded), this algorithm satisfies $R(w(t)) \to \min R$. As the step-size $\eta \to 0$, we obtain the so-called *mirror flow*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\varphi(w(t)) = -\nabla R(w(t)) \qquad \Leftrightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}t}w(t) = -[\nabla^2\varphi(w(t))]^{-1}\nabla R(w(t))$$

Hence, with the "square-parameterization", gradient flow is equivalent to a mirror flow with φ the negative entropy (this equivalence is only exact for the continous-time dynamics).

Let us now state the implicit bias of this continuous time dynamics[1].

Theorem 3.1. *Assume that R is the ERM with the square loss with training set $(x_i, y_i)_{i=1}^n$ and that $\nabla\varphi$ is continuous on its domain. If $w_t \mapsto w^*$ where $w^* \in \mathrm{dom}(\nabla\varphi) \subset \mathbb{R}^d$ satisfies $x_i^\top w^* = y_i$ for $i = 1, \dots, n$, then*

$$w^* = \arg\min D_\varphi(w|w_0) \qquad s.t. \qquad x_i^\top w^* = y_i, \quad i = 1, \dots, n. \tag{8}$$

where $D_\varphi(w|w_0)$ denotes the Bregman divergence defined as

$$D_\varphi(w|w_0) = \varphi(w) - \varphi(w_0) - \nabla\varphi(w_0)^\top(w - w_0).$$

Proof. The optimality conditions (KKT) for (8) are

$$\begin{cases} \nabla\varphi(w) - \nabla\varphi(w_0) + \sum_{i=1}^n \lambda_i x_i = 0 \text{ for some } \lambda \in \mathbb{R}^n \\ x_i^\top w = y_i \text{ for all } i \in \{1, \dots, n\} \end{cases}. \tag{9}$$

But by construction of the MD algorithm (in continuous or discrete time), we have $\forall t$:

$$\nabla\varphi(w_t) - \nabla\varphi(w_0) \in \mathrm{span}\{x_1, \dots, x_n\}$$

and this remains true in the limit by continuity of $\nabla\varphi$. Thus w^* satisfies the KKT conditions. □

[1] In fact the theorem also holds for the discrete-time dynamics of mirror descent (7), but not for (3) in general.

Consequence for square parameterization With φ the negative entropy as in (4), we have for $a, b \in \mathbb{R}^d_{++}$

$$D_\varphi(a|b) = \sum_j a_j \log(a_j) - a_j - \big(b_j \log(b_j) - b_j + \log(b_j)(a_j - b_j)\big) \tag{10}$$

$$= \sum_j a_j \log(a_j/b_j) - a_j + b_j \tag{11}$$

which is called the *relative entropy* between a and b (a.k.a. *Kullback-Leibler* divergence). We get, for an initialization $w_0 = \alpha \mathbf{1}$

$$D(w|w_0) = \sum_j w_i \Big(\log(w_i) - \log(\alpha) - 1 \Big) + \alpha d$$

and when $\alpha \to 0$ we have $D(w|w_0) \sim \log(1/\alpha)\|w\|_1$, so the implicit bias is towards the min ℓ_1-norm interpolator (instead of min ℓ_2-norm for linear parameterization). This is a general rule of thumb that "small" initialization for homogeneous models of degree ≥ 2 lead to "sparse" solutions, see e.g. Maennel et al. [2018].

Other examples There are other examples of reparameterizations which can be rewritten as mirror flows, such as the parameterization $(\theta_+, \theta_-) \mapsto w = \theta_+ \odot \theta_+ - \theta_- \odot \theta_-$ which leads to the hyperbolic entropy mirror map (see Exercises). However, these are very particular cases specifically designed to get an explicit form of the implicit bias. In general we do not have such an explicit result.

4 Lazy training

We now study a implicit bias which appears for "large" initializations called *lazy training*. The analysis that follows is very general and applies to modern ML models, but on the downside, state-of-the-art models are in general not trained in the regime where this implicit bias occurs. It can be thought as a "pitfall" of badly initialized neural networks.

Consider the following rescaled objective function, for $\alpha > 0$:

$$F_\alpha(\theta) = \frac{1}{\alpha^2} R(\alpha f_\theta)$$

The factor $1/\alpha^2$ does not change minimizers and is just the suitable normalization factor for large α. The gradient flow gives, in parameter space,

$$\frac{d}{dt}\theta(t) = -\nabla F_\alpha(\theta(t)) = -\frac{1}{\alpha} Df(\theta(t))^\top \nabla R(\alpha f_{\theta(t)})$$

and this gives, in predictor space (denoting $f_t = f_{\theta_t}$):

$$\frac{d}{dt}\alpha f_t = -Df(\theta(t))Df(\theta(t))^\top \nabla R(\alpha f_t)$$

This is just the computations of Section 2, except that now we track the factor α. At initialization, notice that if $\nabla R(\alpha f_0)$ is $\Theta(1)$ as $\alpha \to \infty$ (e.g. $f_0 = 0$ or $\|f_0\| = O(1/\alpha)$) then the change in parameter space is in $\Theta(1/\alpha)$ while it is $\Theta(1)$ in predictor space when the Jacobian is not degenerate.

What does α stand for? The factor α appears implicitly in various situations, such as:

- for models which are p-positively homogeneous for some $p > 1$, i.e. $f_{\beta\theta} = \beta^p f_\theta$ then scaling the initialization by a factor β is equivalent to setting $\alpha \sim \beta^p$
- for NNs with a certain standard choice of initialization, α is proportional to the square-root width (see next lectures).

Linearized dynamics Since the relative change of parameters is small, it makes sense to compare the GF with the GF of linearized model around the initialization θ_0:

$$\bar{F}_\alpha(\theta) = \frac{1}{\alpha^2} R(\alpha \bar{f}_\theta).$$

In this general analysis, let us assume for simplicity that $f_{\theta_0} = 0$, so that the linearized model is simply

$$\bar{f}_\theta = f_{\theta_0} + Df(\theta_0)(\theta - \theta_0) = Df(\theta_0)(\theta - \theta_0).$$

The GF in parameter space is

$$\frac{d}{dt}\bar{\theta}(t) = -\nabla \bar{F}_\alpha(\bar{\theta}(t)) = -\frac{1}{\alpha} Df(\theta_0)^\top \nabla R(\alpha \bar{f}_{\bar{\theta}_t})$$

and this gives, in predictor space (denoting $\bar{f}_t = \bar{f}_{\bar{\theta}_t}$):

$$\frac{d}{dt}\alpha \bar{f}_t = -Df(\theta_0) Df(\theta_0)^\top \nabla R(\alpha \bar{f}_t)$$

Our previous remarks lead to the following result:

Theorem 4.1. *Assume that $\theta \mapsto Df(\theta)$ and $y \mapsto \nabla R(y)$ are Lipschitz continuous. Given a fixed time horizon T and R lower bounded, it holds:*

$$\begin{aligned}
\sup_{t\in[0,T]} \|\theta(t) - \theta_0\|_2 &= O(1/\alpha) \\
\sup_{t\in[0,T]} \|\theta(t) - \bar{\theta}(t)\|_2 &= O(1/\alpha^2) \\
\sup_{t\in[0,T]} \|\alpha f_t - \alpha \bar{f}_t\|_2 &= O(1/\alpha)
\end{aligned}$$

(the last claim is not trivial because in general we have $\sup_{t\in[0,T]} \|\alpha f_t - \alpha f_0\| = \Theta(1)$ since the loss decrease is $\Theta(1)$).

Proof. Recall that for GF, $\frac{d}{dt} F_\alpha(\theta(t)) = -\|\nabla F_\alpha(\theta(t))\|_2^2$, thus for $t \geq 0$ it holds

$$\begin{aligned}
\|\theta(t) - \theta(0)\|_2 &\leq \int_0^t \|\theta'(s)\|_2 \mathrm{d}s \\
&\leq \sqrt{t}\Big(\int_0^t \|\theta'(s)\|_2^2 \mathrm{d}s\Big)^{1/2} \\
&\leq \sqrt{t}\Big(F_\alpha(\theta_0) - F_\alpha(\theta_t)\Big)^{1/2} \\
&\leq \frac{\sqrt{t}(R(0) - \inf R)^{1/2}}{\alpha}.
\end{aligned}$$

which proves the first claim. It also follows that $\|\alpha f_t\|$ and $\|\nabla R(\alpha f_t)\|$ are bounded independently of α.

Let us now control the deviation from the linearized dynamics. First, in predictor space, let $\Delta(t) := \|\alpha f_t - \alpha \bar{f}_t\|_2$. It satisfies $\Delta(0) = 0$ and, denoting $K(\theta) = Df(\theta)Df(\theta)^\top$ the tangent kernel,

$$\begin{aligned}\Delta'(t) &\leq \|K(\theta(t))\nabla R(\alpha f_t) - K(\theta(0))\nabla R(\alpha \bar{f}_t)\| \\ &\leq \|(K(\theta(t)) - K(\theta(0)))\nabla R(\alpha f_t)\| + \|K(\theta(0))(\nabla R(\alpha f_t) - \nabla R(\alpha \bar{f}_t))\| \\ &\leq C_1/\alpha + C_2\Delta(t)\end{aligned}$$

where $C_i > 0$ denote quantities independent of α and because $\theta \mapsto K(\theta)$ and $y \mapsto \nabla R(y)$ are locally Lischitz continuous. By Grönwall's lemma, it follows[2] that $\Delta(t) \leq C_3/\alpha$.

Finally, in parameter space, let $\delta(t) := \|\theta(t) - \bar{\theta}(t)\|_2$. It holds $\delta(0) = 0$ and

$$\begin{aligned}\delta'(t) &\leq \alpha^{-1}\|Df(\theta(t))^\top \nabla R(\alpha f_t) - Df(\theta(0))^\top \nabla R(\alpha \bar{f}_t)\| \\ &\leq \alpha^{-1}\|(Df(\theta(t)) - Df(\theta(0)))^\top \nabla R(\alpha f_t)\| + \alpha^{-1}\|Df(\theta(0))^\top(\nabla R(\alpha f_t) - \nabla R(\alpha \bar{f}_t))\| \\ &\leq C_4\alpha^{-2}\end{aligned}$$

by the (local) Lipschitz continuity of $\theta \mapsto Df(\theta)$ and $y \mapsto \nabla R(y)$. □

In certain contexts "lazy training" isn't just a transient phase, but holds until convergence.

Theorem 4.2. *Let $R : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex, ν-smooth function with global minimizer y. Assume that $f_{\theta(0)} = 0$ and that $Df(\theta_0)Df(\theta_0)^\top \succ 0$. Then for α large enough, GF converges to a minimizer at an exponential rate and the conclusion of Theorem 4.1 holds with $T = +\infty$.*

This is the "over-parameterized setting as we require to be able to fit the training set with the linearized model. Note that this theorem can be made quantitative in all the quantities involved, see e.g. Bartlett et al. [2021]. This proof technique is behind most results in the literature that claim that optimization of deep NNs is "easy": however, these results apply in a regime where NNs behave like kernel methods, which has little practical relevance.

Lemma 4.3. *Let $R : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex, ν-smooth function with global minimizer y^* and $\Sigma(t) \succeq \sigma \mathrm{Id} \in \mathbb{R}^{n\times n}$ a time dependent matrix for $t \in [0, T]$. Then solutions on $[0, T]$ to the ODE*

$$\frac{d}{dt}y_t = -\Sigma(t)\nabla R(y_t)$$

satisfy for $0 \leq t \leq T$,

$$\|y_t - y^*\| \leq \sqrt{\frac{\nu}{\mu}}\|y_0 - y^*\|e^{-\mu\sigma t}.$$

Proof. By the Polyak-Łojasiewicz inequality (satisfied by strongly convex functions, see Lecture 7)

$$R(y_t) - R(y^*) \leq \frac{1}{2\mu}\|\nabla R(y_t)\|^2.$$

It follows

$$\frac{d}{dt}\big(R(y_t) - R(y)\big) = -\nabla R(y_t)^\top \Sigma(t)\nabla R(y_t) \leq -\sigma\|\nabla R(y_t)\|^2 \leq -2\mu\sigma\big(R(y_t) - R(y^*)\big).$$

It follows by integrating in time

$$\big(R(y_t) - R(y^*)\big) \leq e^{-2\mu\sigma t}\big(R(y_0) - R(y^*)\big).$$

[2] The ODE $u'(t) = C_1/\alpha + C_2 u(t)$ with $u(0) = 0$ has unique solution $u(t) = \frac{C_1}{\alpha C_2}(\exp(C_2 t) - 1)$ and upper bounds $\Delta(t)$ which is a subsolution.

Now we also have, since $\nabla R(y^*) = 0$ and by definition of strong convexity/smoothness, $\forall y' \in \mathbb{R}^n$,

$$\frac{\mu}{2}\|y' - y\|^2 \leq R(y') - R(y^*) \leq \frac{\nu}{2}\|y' - y\|^2$$

and this leads to

$$\|y_t - y\|^2 \leq \frac{\nu}{\mu} e^{-2\mu\sigma t}\|y_0 - y\|^2. \qquad \square$$

Proof of Thm. 4.2. Let $\sigma > 0$ be such that $\Sigma_0 \succ \sigma \mathrm{Id}$, $M > 0$ such that $\|Df(\theta(0))\|_{2\to 2} < M$ and let $T = \sup\{t \geq 0\ ;\ \Sigma_t \succeq \sigma \mathrm{Id} \text{ and } \|Df(\theta(t))\|_{2\to 2} \leq M\} > 0$. Our goal is to show that, for α large enough, $T = +\infty$, and exponential convergence of the predictor to the minimizer will follow by Lemma 4.3.

By Lemma 4.3, it holds for $t \in [0, T]$

$$\begin{aligned}
\|\theta_t - \theta_0\|_2 &\leq \int_0^t \|\theta'(s)\|_2 \mathrm{d}s \\
&\leq \frac{M}{\alpha}\int_0^t \|\nabla R(\alpha f_s)\|_2 \mathrm{d}s \\
&\leq \frac{M\nu}{\alpha}\int_0^t \|\alpha f_s - y\|_2 \mathrm{d}s \\
&\leq \frac{M\nu}{\alpha}\sqrt{\frac{\nu}{\mu}}\|y\|_2 \int_0^t e^{-\mu\sigma s}\mathrm{d}s \\
&\leq \frac{M}{\alpha\sigma}\Big(\frac{\nu}{\mu}\Big)^{3/2}\|y\|_2.
\end{aligned}$$

This quantity can be made arbitrarily small for α large, and thus for α large enough it holds $T = +\infty$. The fact that the conclusion of Theorem 4.1 holds with $T = +\infty$ can be found with similar techniques, see details in [Chizat et al., 2019, Thm. 2.4]. $\square$

References

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. 2018.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. *Advances in Neural Information Processing Systems*, 33: 253–264, 2020.