Topics in Math of ML, Lecture 9
Algorithmic regularization of GD for linear models

Lénaïc Chizat*

November 16, 2025

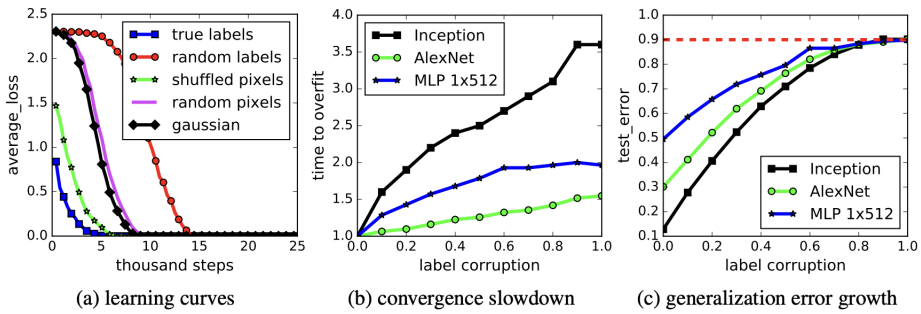
1 Introduction

In previous lectures we used either an explicit regularization term or a constraint to control the size of the "hypothesis space" and thus the estimation error. This approach was the paradigm in machine learning practice and theory before the deep learning era. However, large scale models are nowadays often trained without such regularization and still achieve state-of-the-art test performance in certain high-dimensional tasks. This is all the more counter-intuitive that these models are often *over-parameterized*, and optimized until empirical risk is minimized (see Figure 1). In this context, there are many functions in the hypothesis class that minimize the ERM, some of them will have a good test performance, some won't.

In today's lecture, we will explore how the optimization algorithm itself induces an implicit form of regularization. This regularization can be of several forms:

- **Early stopping regularization**: This refers to the situation where the optimization is stopped before convergence (this prevents the iterates from exploring the whole hypothesis space);
- **Implicit bias:** This refers to the situation where the empirical risk has multiple minimizers (some may generalize better than other) and the optimization algorithm, when run until convergence, selects a specific one;

*EPFL lenaic.chizat@epfl.ch



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10 (an image classification task). (a) shows the training loss of various experimental settings decaying with the training steps (for the "inception" model). (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions. From Zhang et al. [2021].

	Traditional ML	Modern Deep Learning
Capacity control	explicit: regularization or few params.	it depends(*)
Estimator $\hat{\theta}$	minimizer of ERM	output of training algorithm
Criterion for algo.	Computational efficiency for ERM	Good algorithmic bias & efficiency

Table 1: Some key differences between traditional ML and Deep Learning.

We summarize on Table 1 the main differences between traditional machine learning (ML) and modern deep learning[1] (DL).

One of the difficulty with DL theory is that the description of the learnt predictor is only implicit ("the output of the learning algorithm"). One technique used to overcome this difficulty is to express $\hat{\theta}$ (or directly the predictor $f_{\hat{\theta}}$) as the minimizer of a surrogate optimization problem that depends on the optimization algorithm but that is not explicitly formulated by the practitionner.

In this lecture, we begin our investigation of *implicit bias* with the study of linearly-parameterized models in two settings: regression (square-loss) and classification (exponentially tailed loss).

2 Ordinary Least Squares: dynamics of GD

Let us consider the behavior of gradient descent (GD) on the ordinary least squares (OLS) objective

$$F(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2$$

with $\Phi \in \mathbb{R}^{n\times d}$ the design matrix and $y \in \mathbb{R}^n$ the labels/covariates. For simplicity assume $d \leq n$ and Φ injective. Using results from Lecture 2, the excess empirical risk is, using $\Sigma = \frac{1}{n}\Phi^\top\Phi$ and θ^* the minimizer:

$$F(\theta) = \frac{1}{2}(\theta - \theta^*)\Sigma(\theta - \theta^*).$$

Trajectory of iterates. The gradient descent iterates with fixed step-size $\eta_t = \eta$ are:

$$\theta_{t+1} = \theta_t - \eta\nabla F(\theta_t) = \theta_t - \eta\Sigma(\theta_t - \theta^*)$$

This is a symmetric linear dynamical system and can be solved explicitly. We diagonalize $\Sigma = PDP^\top$ with $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$ are the eigenvalues of Σ and we define $v_t = P^\top(\theta_t - \theta^*)$. This vector evolves as

$$v_{t+1} = (\mathrm{Id} - \eta D)v_t \quad \Rightarrow v_t[j] = (1 - \eta\lambda_j)^t v_0[j].$$

Interpretation. The columns of P associated to λ_i large – the principal components of Σ – can be interpreted as "low complexity" components while λ_i small are "high complexity" components. If we suppose that the step-size is small enough so that $\eta\lambda_1 < 1$ then:

- Targets θ^* which concentrate on low complexity components are learnt faster;
- For any kind of target θ^*, GD will first learn the low complexity components;

[1] Regarding capacity control in deep learning, the situation remains unclear, subtle, and task dependent: LLMs are often under-parameterized, regularized and trained with online SGD; image classifiers are often over-parameterized and solve ERM; denoising diffusion models are over-parameterized and early-stopped

- Observe that $v_t[j] = f(t, \lambda_j) v_0[j]$, for some f monotonous in t satisfying $f(0, \lambda_j) = 1$ and $f(\infty, \lambda_j) = 0$. An expression of the same form holds for "ridge regression" (Lecture 2) with $1/\lambda$ playing the role of t. This shows that before convergence, GD induces an *algorithmic regularization* which might be statistically beneficial (see Ali et al. [2020] for more on this link).

Decrease of objective. In terms of excess risk, we have

$$F(\theta_t) = \frac{1}{2} v_t^\top D v_t = \frac{1}{2} \sum_{j=1}^{d} \lambda_j |1 - \eta \lambda_j|^{2t} v_0[j]^2 \leq \Big(\max_j |1 - \eta \lambda_j| \Big)^{2t} F(\theta_0).$$

Choice of step-size. If we want the fastest asymptotic rate, we need to choose η that minimizes the contraction ratio. Writing $\alpha = \min\{\lambda_j\}$ and $\beta = \max\{\lambda_j\}$ and the *condition number* $\kappa = \beta/\alpha$, we obtain

$$\min_\eta \max_j |1 - \eta \lambda_j| = \min_\eta \max\{\eta\beta - 1, 1 - \eta\alpha\} = \frac{\beta - \alpha}{\beta + \alpha} = \frac{\kappa - 1}{\kappa + 1}$$

with the minimizer $\eta = 2/(\beta + \alpha)$. In practice, we do not know α and β, but we have exponential convergence as long as $0 < \eta < 2/\beta$ so it is sufficient to know an upper bound on β.

3 Over-parameterized Least-squares: implicit bias

Let us continue the study of gradient descent (GD) on the linear least-squares objective but we now assume that $d > n$ (over-parameterized setting). We also assume that Φ is full rank, i.e. the kernel matrix $\Phi\Phi^\top \in \mathbb{R}^{n \times n}$ is invertible, for simplicity. The objective is

$$F(\theta) = \frac{1}{2n} \|y - \Phi\theta\|_2^2$$

and there are infinitely many minimizers such that $y = \Phi\theta$ since the column space of Φ is the entire space $\mathbb{R}^n$ and θ has dimension $d > n$. Consider again the iterates of GD with $\theta_0 \in \mathbb{R}^d$ and

$$\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t).$$

Proposition 3.1. *Suppose that θ_t converges towards θ_∞ and that θ_∞ satisfies $y = \Phi\theta_\infty$. Then θ_∞ is the (unique) solution to*

$$\theta_\infty = \arg\min \Big\{ \|\theta - \theta_0\|_2^2 \ ; \ \theta \in \mathbb{R}^d \ s.t. \ y = \Phi\theta \Big\}. \tag{1}$$

which is the ℓ_2-projection of the initialization on the set of solutions. When $\theta_0 = 0$, this is the minimum ℓ_2-norm minimizer.

Note: In fact we could deduce this result from the explicit form of the dynamics above, but we give here an alternative proof that applies to other dynamics and uses important techniques.

Proof. One can formulate the problem (1) in Lagrangian form as

$$\inf_{\theta \in \mathbb{R}^p} \sup_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|\theta - \theta_0\|_2^2 + \alpha^\top (y - \Phi\theta) \tag{2}$$

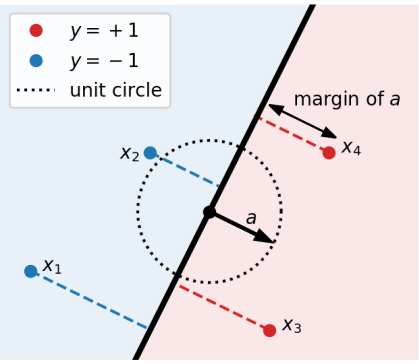


Figure 2: $a \in \mathbb{R}^2$ is the max-margin classifier of the training set $(x_i, y_i)_{i=1}^4$

and the unique minimizer θ^* is characterized by the KKT conditions

$$\begin{cases} y = \Phi\theta \\ \theta - \theta_0 = \Phi^\top \alpha \text{ for some } \alpha \in \mathbb{R}^n. \end{cases} \tag{3}$$

Now remark that $\nabla F(\theta) = \frac{1}{n}\Phi^\top(\Phi\theta - y)$ hence since $\theta_{t+1} = \theta_t - \eta\nabla F(\theta_t)$ we have that $\theta_t - \theta_0 \in \mathrm{Im}(\Phi^\top)$ for all $t \geq 0$. Since the later is closed, we have $\theta_\infty - \theta_0 \in \mathrm{Im}(\Phi^\top)$. As a consequence, θ_∞ satisfies the KKT conditions. □

- this result remains true for any algorithm that remains in the span of gradients such as SGD, accelerated GD...
- convergence is easy to prove in this context, but we skip it to focus on the implicit bias;
- the ℓ_2-norm appears here because we are using gradients for the ℓ_2 metric.

4 Classification with logistic loss in the separable case

In this section we are interested in the behavior of gradient descent for *unregularized* logistic regression in the *separable setting* [Soudry et al., 2018]. The objective to minimize is

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top \theta}).$$

Separable data and margin We assume that the training set $(x_i, y_i)_{i=1}^n$ is linearly separable, which means that there exists one linear classifier $\theta \in \mathbb{R}^d$ that makes no mistake on the training set: $\mathrm{sign}(x_i^\top \theta) = y_i$, $\forall i$. Equivalently, the ℓ_2-max-margin

$$\gamma := \max_{\|\theta\|_2 \leq 1} \min_i y_i x_i^\top \theta$$

satisfies $\gamma > 0$. For such a dataset, a natural predictor is the (unique) *max-margin predictor*, that achieves the maximum. See Figure 2 for an illustration.

To focus on the key aspect of the problem, let us make some simplifications:

- (i) we consider the gradient flow (GF), from some initialization[2] $\theta(0) = \theta_0 \in \mathbb{R}^d$ and

$$\theta'(t) = -\nabla F(\theta(t)).$$

[2]The initialization plays little role in this section, but that's specific to the fact that we are looking at the infinite time behavior in a classification setting.

- (ii) we replace the logistic loss by the exponential loss (which has the same tail as the logistic loss), that is the objective is

$$\frac{1}{n}\sum_{i=1}^{n} e^{-y_i x_i^\top \theta}.$$

Path and time-reparameterization. In this section, we only care about the optimization path $\{\theta(t)\ ;\ t \geq 0\} \subset \mathbb{R}^d$ (in particular about its limit), which is not changed by composing the objective with a differentiable function $h : \mathbb{R} \to \mathbb{R}$ with $h'(u) > 0$, $\forall u \in \operatorname{im} F$. Indeed letting $G = h \circ F$ and calling θ_h the corresponding GF, it holds

$$\theta_h'(t) = -\nabla G(\theta_h(t)) = -h'(F(\theta_h(t))) \cdot \nabla F(\theta_h(t)).$$

Letting $s(t) = \int_0^t h'(F(\theta_h(s))\mathrm{d}s$ we have $\frac{d}{dt}\theta(s(t)) = -\nabla F(\theta(s(t))) \cdot s'(t) = \frac{d}{dt}\theta_h(t)$, which shows that the paths $\{\theta_h(t)\}_{t\geq 0}$ and $\{\theta_h(s)\}_{s\geq 0}$ are the same. Keeping these remarks in mind, we consider the objective

$$F(\theta) = -\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{-y_i x_i^\top \theta}\right) \tag{4}$$

and the dynamics[3]

$$\theta'(t) = \nabla F(\theta(t))$$

which is just a time-reparameterization of the GF of ERM with exponential loss.

Theorem 4.1. *Assume $(x_i, y_i)_{i=1}^n$ is separable (i.e. $\gamma > 0$). For any initial point $\theta_0 \in \mathbb{R}^d$, then $\|\theta(t)\|_2 \to \infty$. Moreover, the renormalized predictor $\bar{\theta}(t) := \frac{\theta(t)}{\|\theta(t)\|_2}$ converges to the optimal margin at a $O(1/t)$ rate. Assuming $\theta_0 = 0$ for simplicity it holds for $t \geq t^* := \log(n)/\gamma^2$:*

$$\min_i y_i x_i^\top \bar{\theta}(t) \geq \gamma - \frac{\log(n)}{\gamma t}.$$

Let us start with some preliminary results.

Lemma 4.2. *With F defined in Eq. (4) it holds:*

1. $\min_i y_i x_i^\top \theta \leq F(\theta) \leq \min_i y_i x_i^\top \theta + \log(n)$, $\forall \theta \in \mathbb{R}^d$
2. $\|\nabla F(\theta)\|_2 \geq \gamma$, $\forall \theta \in \mathbb{R}^d$

Proof. 1. Letting $m = \min_i y_i x_i^\top \theta$, we just apply $-\log$ to the following basic inequalities

$$e^{-m} \geq \frac{1}{n}\sum_{i=1}^{n} e^{-y_i x_i^\top \theta} \geq \frac{1}{n} e^{-m}.$$

2. Let $Z \in \mathbb{R}^{n\times d}$ be the matrix with rows $y_i x_i$ and let $\Delta_n = \{p \in \mathbb{R}_+^n\ ;\ \sum_{i=1}^n p_i = 1\}$ be the simplex. We have by the Minimax Theorem

$$\gamma = \max_{\|\theta\|_2 \leq 1} \min_{p\in\Delta_n} p^\top Z\theta = \min_{p\in\Delta_n} \max_{\|\theta\|_2\leq 1} p^\top Z\theta = \min_{p\in\Delta_n} \|Z^\top p\|_2.$$

On the other hand, we have

$$\nabla F(\theta) = \sum_{i=1}^{n} \frac{y_i x_i e^{-y_i x_i^\top \theta}}{\sum_{j=1}^d e^{-y_j x_j^\top \theta}}.$$

[3]We remove the minus sign in the definition of GF because $-\log$ has a negative derivative.

Hence $\nabla F(\theta) = Z^\top p$ with $p_i = \frac{e^{-y_i x_i^\top \theta}}{\sum_{j=1}^n e^{-y_j x_j^\top \theta}}$ for $i \in \{1, \dots, n\}$. Since $p \in \Delta_n$, it follows

$$\|\nabla F(\theta)\|_2 \geq \min_{p \in \Delta_n} \|Z^\top p\|_2 = \gamma. \qquad \square$$

Proof of the Theorem. • The equivalent of the descent lemma[4] in continuous time is

$$\frac{d}{dt} F(\theta(t)) = \nabla F(\theta(t))^\top \frac{d}{dt}\theta(t) = \|\nabla F(\theta(t))\|_2^2 \geq \gamma^2.$$

This means that F grows unbounded and thus $\|\theta(t)\|_2 \to \infty$ (e.g. by Lemma 3.1.(1)).

- It holds $F(\theta(t)) - F(\theta_0) = \int_0^t \|\nabla F(\theta(s))\|_2^2 \mathrm{d}s \geq \gamma \int_0^t \|\nabla F(\theta(s))\|_2 \mathrm{d}s$ and thus, since $F(\theta_0) = 0$,

$$\min_i y_i x_i^\top \theta(t) \geq F(\theta(t)) - \log(n) \geq \gamma \int_0^t \|\nabla F(\theta(s))\|_2 \mathrm{d}s - \log(n)$$

and this lower bound is larger than $\gamma^2 \cdot t - \log(n)$ which is nonnegative for $t \geq t^*$.

- Now notice that $\|\theta(t)\|_2 = \int_0^t \frac{d}{ds}\|\theta(s)\|_2 \mathrm{d}s \leq \int_0^t \|\frac{d}{ds}\theta(s)\|_2 \mathrm{d}s = \int_0^t \|\nabla F(\theta(s))\|_2 \mathrm{d}s$. We then divide the left-hand side by $\|\theta(t)\|_2$ and the right-hand side by the larger quantity $\int_0^t \|\nabla F(\theta(s))\|_2 \mathrm{d}s$ and we get (for $t \geq t^*$ so that all quantities are nonnegative):

$$\min_i y_i x_i^\top \bar{\theta}(t) \geq \gamma - \frac{\log(n)}{\int_0^t \|\nabla F(\theta(s))\|_2 \mathrm{d}s} \geq \gamma - \frac{\log(n)}{\gamma t}. \qquad \square$$

- For classification tasks, we only care about the sign of the predictor at test time; so the fact that $\|\theta(t)\| \to \infty$ is not an issue.
- Although we studied GF on the exponential loss, the "implicit bias" is the same for GD and the logistic loss.
- The max-margin classifier is unique, thus we have also convergence to the max-margin classifier.
- The initialization θ_0 does not play a role in the limit (our assumptions $\theta_0 = 0$ was only convenient to get simple non-asymptotic bounds).

References

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning*, pages 233–244. PMLR, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[4] here with the sign flipped, because we are following the gradient and not the negative gradient.