Chapter 6

Interpolation

Contents

This chapter heavily relies on the article [BMR21] and the lecture notes of our spiritual father Francis Bach.

The performance of deep learning is remarkable and surprising, especially since it seems to contradict the statistical theory that has guarded against overfitting for decades: while being complex models, NN seem to still provide excellent predictive accuracy.

The training of NN is usually performed via stochastic gradient strategies (SGD). The conjecture is therefore that overparameterization allows gradient methods to find “good" interpolating solutions: the overfitting would be then “benign" as not harmful for the optimization of NN nor for the generalization abilities of the found solution.

6.1 Implicit bias of SGD in interpolation regimes

Statistical wisdom suggests that a method that takes advantage of too many degrees of freedom by perfectly interpolating noisy training data will be poor at predicting new outcomes. In deep learning, training algorithms appear to induce a bias that breaks the equivalence among all the models that interpolate the observed data.

6.1.1 Preliminary on optimization

The goal of an optimization problem is generally to minimize a function F over some parameter space Θ. If the global minimizer $\theta^\star$ is unique, even if the initial goal is to minimize F, one should expect that that the t-th iterate θ_t given by some optimization algorithm converges to that $\theta^\star$. When there are multiple minimizers (preventing the function to minimize to be strongly convex), one can only expect that $F(\theta_t) - \inf_{\theta\in\Theta} F(\theta)$ is converging to zero (and only if a minimizer exists). Note that when F is a convex function, the set of minimizers is a convex set.

With some extra assumptions, one can show that the algorithm is converging to one of the multiple minimizers of F. But which one? This is what is referred to as the implicit regularization properties of optimization algorithms, and here gradient descent and its variants.

Imagine now that, for a learning purpose, F stands for an empirical loss associated to n observations with $\Theta \subset \mathbb{R}^d$ and d much larger than n. No regularization being used, there are multiple minimizers achieving a zero training error (usually referred as the overfitting regime). Therefore, an arbitrary empirical risk minimizer is not expected to work well on unseen data. To reduce the complexity of the model (embodied by d here), a classical way to prevent overfitting is to use explicit regularization (e.g. ℓ^2-norms - Ridge/Tikhonov penalties- or ℓ^1-norms -Lasso penalties).

In this section, we show that optimization algorithms have a similar regularizing effect, without appealing to explicit penalties. In a nutshell, gradient descent usually leads to particular solutions of minimum ℓ^2-norm, meaning that the chosen empirical risk minimizer is not arbitrary.

6.1.2 Quadratic loss and linear models

Setting. To better understand this phenomenon, we restrict ourselves in a first time to the case of linear models. Choosing a quadratic loss for the empirical risk minimization boils down in building a least-square estimator:

$$F(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i^\top\theta)^2 = \frac{1}{2n}\left\|y - \mathbb{X}\theta\right\|_2^2 \tag{6.1}$$

where $\mathbb{X} = (X_1|X_2|\ldots|X_n)^\top \in \mathbb{R}^{n\times p}$ with $n \ll p$. The (kernel) matrix $\mathbb{X}\mathbb{X}^\top$ is assumed to be invertible. Therefore there is an infinity of minimizers of F corresponding to the solutions of the system $y = \mathbb{X}\theta$: the set of minimizers is actually an affine space (given a solution θ_0 to $y = \mathbb{X}\theta$, $\theta_0 + \text{null}(\mathbb{X})$ is also solution).

Running GD. Let's do the thought experiment that you are not able to write a particular solution of $y = \mathbb{X}\theta$, one can use a gradient descent strategy instead to minimize F. F being convex, the GD is going to converge to a global minimizer (and we know there exists since there is an infinity of solutions).

The gradient algorithm (GD) used to minimize F can be written as follows: for some initial parameter θ_0, the iterates of GD are

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

When the function F is assumed to be L-smooth (meaning that its gradient is assumed to be L-Lipschitz), the gradient algorithm is a descent method provided that the step η is chosen such that $\eta \le 1/L$.

Therefore applying GD with $\theta_0 = 0$ (zero initialization) and $\eta \le 1/\lambda_{\max}(\mathbb{X}^\top\mathbb{X}/n)$ and considering a solution θ of $y = \mathbb{X}\theta$ leads to

$$\theta_t - \theta = \left(I - \frac{\eta}{n}\mathbb{X}^\top\mathbb{X}\right)^t(\theta_0 - \theta) = -\left(I - \frac{\eta}{n}\mathbb{X}^\top\mathbb{X}\right)^t\theta$$

and

$$\theta_t = \left[I - \left(I - \frac{\eta}{n} \mathbb{X}^\top \mathbb{X} \right)^t \right] \theta \tag{6.2}$$

Note that it is not entirely obvious that the formula above is independent of the choice of θ (but it is).

Proposition 6.1

The solution of $y = \mathbb{X}\theta$ with the minimal ℓ^2-norm is

$$\mathbb{X}^\dagger y = V \mathrm{diag}(s^{-1}) U^T y$$

where

- $\mathbb{X}^\dagger$ is the pseudo-inverse of X, and
- $\mathbb{X} = U \mathrm{diag}(s) V^T$ is the SVD decomposition of X, so that
  - $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthonormal ($U^T U = I_n$ and $V^T V = I_d$),
  - $\mathrm{diag}(s) \in \mathbb{R}^{n \times p}$ ⚠ with the singular values $(s_i)_{1 \le i \le n}$ of $\mathbb{X}$.
    NB: $\mathrm{diag}(s^{-1}) \in \mathbb{R}^{p \times n}$

Proof. Left in exercise. □

Proposition 6.2

Considering the gradient descent initialized at 0 (6.2), one gets

$$\left\| \theta_t - V \mathrm{diag}(s^{-1}) U^T y \right\|_2 \le \left(1 - \frac{\min_i s_i^2}{\max_i s_i^2} \right)^t \left\| V \mathrm{diag}(s^{-1}) U^T y \right\|_2 . \tag{6.3}$$

Proof. Choosing $\theta = V \mathrm{diag}(s^{-1}) U^T y$ in (6.2) leads to

$$\theta_t = V \mathrm{diag}\left(\left(1 - \left(1 - \frac{\eta s_i^2}{n} \right)^t \right) s_i^{-1} \right) U^\top y. \tag{6.4}$$

Since each $s_i > 0$ for $i = 1, \dots, n$ (X is assumed of full rank) and $\eta \le 1/\lambda_{\max}(\mathbb{X}^\top \mathbb{X}/n) = n/\max_i s_i^2$, one gets

$$0 \le \left(1 - \left(1 - \frac{\eta s_i^2}{n} \right)^t \right) s_i^{-1} \le s_i^{-1} \left(1 - \left(1 - \frac{\eta \min_i s_i^2}{n} \right)^t \right).$$

This shows that

$$\left\| \theta_t - V \mathrm{diag}(s^{-1}) U^T y \right\|_2 \le \left(1 - \frac{\eta \min_i s_i^2}{n} \right)^t \left\| V \mathrm{diag}(s^{-1}) U^T y \right\|_2 . \tag{6.5}$$

Fixing η to be the largest step size allowed, i.e. $\eta = n/\max_i s_i^2$ gives the result. □

Note that $\frac{\min_i s_i^2}{\max_i s_i^2}$ can be seen as the inverse of the conditioning number of $\mathbb{X}$.

In the case of overparameterized linear regression, the gradient descent (with constant step size, initialized at 0) linearly converges towards the solution of $y = \mathbb{X}\theta$ of minimal ℓ^2-norm.

Question: how important is the initialization at zero?

Exercise: http://fa.bianp.net/blog/2022/implicit-bias-regression/

Consider the optimization problem where the objective function is a generalized linear model with a data matrix $\mathbb{X} = (X_1|\dots|X_n)^\top \in \mathbb{R}^{n\times p}$ and a target vector $y \in \mathbb{R}^n$:

$$\min_{\theta\in\mathbb{R}^p} f(\theta) = \sum_{i=1}^{n} \varphi(X_i^\top \theta, y_i) \tag{6.6}$$

where $\varphi(z, y)$ is a differentiable real-valued function verifying the "unique finite root condition", which is that it has a unique minimizer at $z = y$. These losses are usually used for regression and includes the quadratic loss or Huber functions. Assume that $p > n$ and that $\mathbb{X}$ is of full-rank.

Problems of this form verify two key properties that make it easy to characterize the bias of gradient-based methods.

1. Show that iterates remain in the span of $\mathbb{X}$. The gradient of the i-th sample $\varphi(X_i^\top\theta, y_i)$ has the same direction as its data sample X_i:

$$\nabla_x \left[\varphi(X_i^\top\theta, y_i)\right] = \underbrace{X_i}_{\text{vector}} \underbrace{\varphi'(X_i^\top\theta, y_i)}_{\text{scalar}}$$

   This implies that any gradient-based method generates updates that stay in the span of the vectors $\{X_1, \dots, X_n\}$.

   It's no surprise then that the vector space generated by the samples $X_1, \dots, X_n$ plays a crucial role here. For convenience we'll denote this subspace by

$$\mathscr{X} := \text{span}\,(X_1, \dots, X_n)$$

   and its orthogonal complement $\mathscr{X}^\perp$.

2. What can you say about minimizers of f? Minimizers solve the linear system $\mathbb{X}\theta = y$. By the unique root condition of φ, the global minimizer is achieved when $X_i^\top\theta = y_i$ for all i. In other words, the global minimizers are the solutions to the linear system $\mathbb{X}\theta = y$, a set that is non-empty by the under-specification assumption.

3. Starting from θ_0, characterize the limit iterate of a gradient-based method. The main argument here is to show that the limit iterate belongs to the intersection of two affine spaces and then compute their intersection.

   By Property 2, the limit iterate must solve the linear system $\mathbb{X}\theta = y$. A classical linear algebra result states that all solutions of this problem are of the form $\theta + c$, with θ any solution of $\mathbb{X}\theta = y$ and $c \in \mathscr{X}^\top = \text{Ker}(\mathbb{X})$. Let's take $\theta = \mathbb{X}^\dagger y$ so that

$$\theta_\infty = \mathbb{X}^\dagger y + c, \text{ for some } c \in \mathscr{X}^\top$$

Let P denote the orthogonal projection onto $\mathscr{X}$. Then we can decompose the initialization as $\theta_0 = P\theta_0 + (I-P)\theta_0$. By the first property all updates are in $\mathscr{X}$, so the limit iterate can be written as

$$\theta_\infty = (I-P)\theta_0 + x \qquad \text{for some } x \in \mathscr{X}.$$

Combining the previous two equations, we have that $c = (I-P)\theta_0$ and $x = \mathbb{X}^\dagger y$. Hence we have arrived at the characterization

$$\theta_\infty = \mathbb{X}^\dagger y + (I-P)\theta_0. \tag{6.7}$$

4. Show that the limit iterate is actually the projection of θ_0 on the set of solutions of $\mathbb{X}\theta = y$. Let $\theta^\star$ denote the solution to

$$\theta_\infty = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\theta-\theta_0\|_2 \qquad \text{such that} \qquad \mathbb{X}\theta = y.$$

$\theta^\star$ is unique by strong convexity. We want to show that $\theta_\star = \theta_\infty$. For any solution θ of $\mathbb{X}\theta = y$, one has $\theta - \theta_\infty \in \mathscr{X}^\perp$ and

$$\begin{aligned}\|\theta-\theta_0\|_2 &= \|\theta-\theta_\infty+\theta_\infty-\theta_0\|_2 \\ &= \|\theta-\theta_\infty+\mathbb{X}^\dagger y - P\theta_0\|_2 \\ &= \sqrt{\|\theta-\theta_\infty\|_2^2 + \|\mathbb{X}^\dagger y - P\theta_0\|_2^2}\end{aligned}$$

where the last identity follows by orthogonality. Since $\theta^\star$ minimizes the distance $\|\theta-\theta_0\|_2$ on the set of solutions of $\mathbb{X}\theta = y$, we must have $\theta^\star - \theta_\infty$=0, and so $\theta^\star = \theta_\infty$. We have actually shown the following result.

Theorem 6.1. *Gradient-based methods started from θ_0 converge to the solution with smallest distance to θ_0. More precisely, assume that the iterates of a gradient-based method converge to a solution of (6.6), and let $\theta_\infty := \lim_{t\to+\infty}\theta_t$ denote this limit. Then θ_∞ solves*

$$\theta_\infty = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\theta-\theta_0\|_2 \qquad \textit{such that} \qquad \mathbb{X}\theta = y.$$

5. Conclude on the role of the initialization to converge towards the solution of minimal ℓ^2-norm. An immediate consequence of this Theorem is that when the initialization θ_0 is in $\mathscr{X}$, then its projection onto $\mathscr{X}^\perp$ is zero, and so from Eq. (6.7) we have $\theta_\infty = \mathbb{X}^\dagger y$ which corresponds to the minimal norm solution.

Corollary 6.2. *If $\theta_0 \in \mathscr{X}$, then the limit iterate θ_∞ solves the minimal norm problem*

$$\theta_\infty = \operatorname*{argmin}_{\theta} \|\theta\|_2 \qquad \textit{such that} \qquad \mathbb{X}\theta = y.$$

Remark The result holds for gradient-based methods, i.e. any method in which the updates are given by a linear combination of current and past gradients. This includes gradient descent, gradient descent with momentum, stochastic gradient descent (SGD), SGD with momentum, Nesterov's accelerated gradient method. It does not include however quasi-Newton methods or diagonally preconditioned methods such as Adagrad or Adam.

Alternative proof for convergence (in short). If started at $\theta_0 = 0$, gradient descent techniques (stochastic or not) will always have iterates θ_t which are linear combinations of rows of $\mathbb{X}$, that is, of the form $\theta_t = \mathbb{X}^\top \alpha_t$ for some $\alpha_t \in \mathbb{R}^n$. This is an alternative algorithmic version of the *representer theorem*.

If the method is converging, then we must have $\mathbb{X}\theta_t$ converging to y (because the standard squared Euclidean norm is strongly-convex, and $\mathbb{X}\theta$ is unique while θ may not be), and thus $\mathbb{X}\mathbb{X}^\top \alpha_t$ is converging to y. If $K = \mathbb{X}\mathbb{X}^\top$ is invertible, this means that α_t is converging to $K^{-1}y$, and thus $\theta_t = X^\top \alpha_t$ is converging to $\mathbb{X}^T K^{-1} y$.

For the story to be complete, one should check that $\mathbb{X}^T K^{-1} y$ is indeed the solution to $y = \mathbb{X}\theta$ of minimal ℓ^2-norm. By standard Lagrangian duality one gets

$$\begin{aligned}
\inf_{\theta\in\mathbb{R}^d} \frac{1}{2}\|\theta\|_2^2 \quad \text{such that} \quad y = \mathbb{X}\theta \qquad &= \qquad \inf_{\theta\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}^n} \underbrace{\frac{1}{2}\|\theta\|_2^2 + \alpha^\top (y - \mathbb{X}\theta)}_{\text{Lagrangian function } L(\theta,\alpha)} \\
&= \qquad \sup_{\alpha\in\mathbb{R}^n} \alpha^\top y - \frac{1}{2}\left\|\mathbb{X}^\top \alpha\right\|_2^2 \quad \text{(with } \theta = \mathbb{X}^\top \alpha \text{ at the optimum)} \\
&= \qquad \sup_{\alpha\in\mathbb{R}^n} \alpha^\top y - \frac{1}{2}\alpha^\top K \alpha.
\end{aligned}$$

The last problem is exactly solved for $\alpha = K^{-1}y$.

What about SGD? Note that in the overparameterized regime, SGD will also converge to the minimum-norm interpolator, even with a fixed learning rate. In contrast, under-parameterized SGD with a fixed learning rate does not converge at all (indeed the stochastic noise at the optimum is 0 only in the overparameterized setting).

6.1.3 Interpolation in logistic regression

Context. Consider now the setting of binary classification (for the output Y living in $\{-1, 1\}$), based on the model of logistic regression, i.e. the prior on the distribution of $Y|X$ is

$$\mathbb{P}(Y = +1|X = x) = \sigma(\varphi(x)^\top \beta)$$

where σ is the sigmoid function, φ is an encoding of the input variables X with $\varphi(x) \in \mathbb{R}^p$, and the model parameters are $\beta \in \mathbb{R}^p$. Given a dataset $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. copies of (X, Y), the estimation of β is usually performed via MLE, resulting in solving the following problem:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-Y_i \varphi(X_i)^\top \beta\right)\right) =: F(\beta).$$

Define the design matrix as $\Phi := \left(\varphi(X_1)|\ldots|\varphi(X_n)\right)^\top \in \mathbb{R}^{n\times p}$ and consider the case where $d > n$, assuming in addition that $\Phi\Phi^\top$ is invertible.

Rewriting an SVM Since $\Phi\Phi^\top$ is invertible, there exists $\eta \in \mathbb{R}^p$ of unit-norm such that for all $i \in \{1, \ldots, n\}$, $Y_i \varphi(X_i)^\top \eta > 0$, meaning that the data is linearly separable. The distance of any point $\varphi(x) \in \mathbb{R}^p$ to a hyperplane defined by $\{x' : \eta^\top \varphi(x') + b = 0\}$ is given by

$$\frac{|\langle \eta, x\rangle + b|}{\|\eta\|}$$

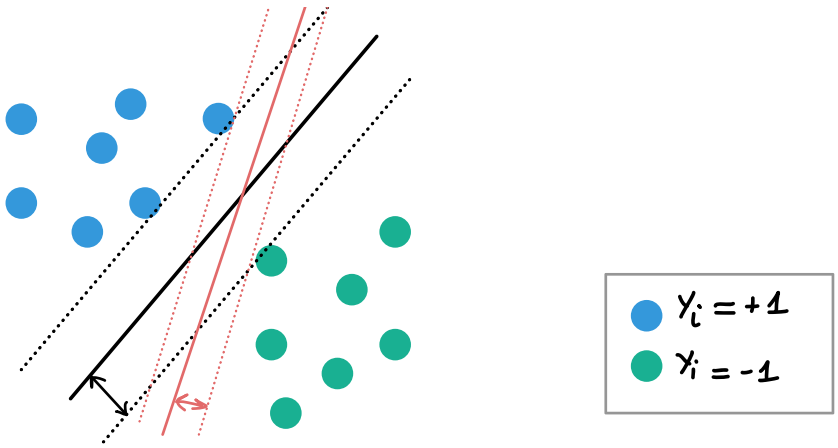


Figure 6.1: The maximum-margin classifier (in black) vs. a classifier based on an arbitrary separating hyperplane (in orange)

Therefore, the distance of a separating hyperplane to the closest points in the dataset, which is called the *margin* is given by

$$\min_{x\in\{X_1,\dots,X_n\}} \frac{|\eta^\top x + b|}{\|\eta\|} = \min_{1\le i\le n} Y_i\varphi(X_i)^\top \eta,$$

when no intercept is considered. One can thus search for a direction η of unit ℓ^2-norm that maximizes the margin:

$$\eta^\star \in \operatorname*{argmax}_{\|\eta\|_2\le 1} \min_{1\le i\le n} Y_i\varphi(X_i)^\top \eta.$$

$\eta^\star$ corresponds to the max-margin classifier (SVM). By Lagrange duality,

$$\begin{aligned}
\sup_{\|\eta\|_2\le 1} \inf_{i\in\{1,\dots,n\}} Y_i\varphi(X_i)^\top\eta &= \sup_{\|\eta\|_2\le 1} t \quad \text{such that} \quad \forall i\in\{1,\dots,n\}, Y_i\varphi(X_i)^\top\eta \ge t,\\
&= \inf_{\alpha\in\mathbb{R}^n_+} \sup_{\|\eta\|_2\le 1} t + \sum_{i=1}^n \alpha_i (Y_i\varphi(X_i)^\top\eta - t)\\
&= \inf_{\alpha\in\mathbb{R}^n_+} \left\| \sum_{i=1}^n \alpha_i Y_i \varphi(X_i) \right\|_2 \quad \text{such that} \quad \sum_{i=1}^n \alpha_i = 1.
\end{aligned}$$

where in the last step we used:

1. (Lagrangian for the constrained sup) $L(t,\eta,\mu) = t + \sum_i \alpha_i (Y_i\varphi(X_i)^t op\eta - t) + \mu(\|\eta\|_2^2 - 1)$
2. (KKT 1) $\nabla_t L = 0$, i.e. $\sum_i \alpha_i = 1$
3. (KKT 2) $\nabla_\eta L = 0$, i.e. $\sum_i \alpha_i Y_i \varphi(X_i) + 2\mu\eta = 0$
4. (KKT 3) $\mu = 0$ or $\|\eta\|_2^2 = 1$

so that $\eta \propto \sum_{i=1}^n \alpha_i Y_i \varphi(X_i)$ at the optimum. Besides, by complementary slackness, non-negative α_i is non-zero only for i such that at the optimum $t = Y_i\varphi(X_i)^\top\eta$, i.e. for i attaining the minimum $\min_{1\le i\le n} Y_i\varphi(X_i)^\top\eta$, corresponding to the so-called support vectors, see Figure 6.2.

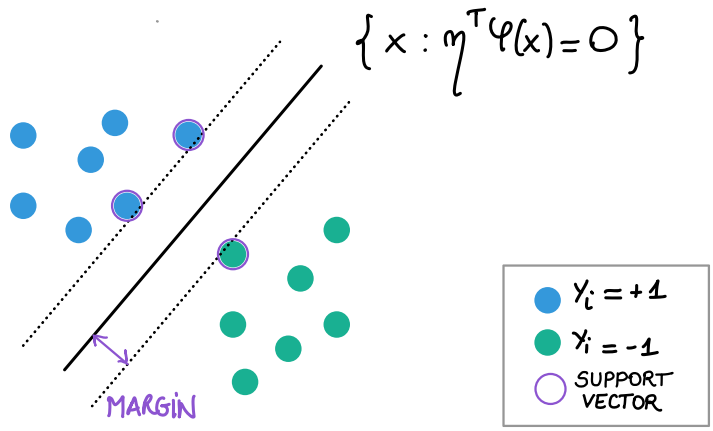


Figure 6.2: Support vectors.

Link with the traditional SVM Because of homogeneity, we want $\min_{1\le i\le n} Y_i\varphi(X_i)^\top\eta$ to be large and $\|\eta\|_2$ to be small. We can therefore constrain the former, and minimize the latter. In other words, we can see $\eta^\star$ as the direction of $\beta^\star$, solution of

$$\inf_{\beta\in\mathbb{R}^p}\frac{1}{2}\|\beta\|_2^2 \quad \text{such that} \quad \operatorname{diag}((Y_i)_i)\Phi\beta\ge\mathbb{1}_n \quad = \quad \inf_{\beta\in\mathbb{R}^p}\sup_{\alpha\in\mathbb{R}_+^n}\frac{1}{2}\|\beta\|_2^2+\alpha^\top\left(\mathbb{1}_n-\operatorname{diag}((Y_i)_i)\Phi\beta\right)$$

$$= \quad \sup_{\alpha\in\mathbb{R}_+^n}\alpha^\top\mathbb{1}_n-\frac{1}{2}\left\|\Phi^\top\operatorname{diag}((Y_i)_i)\alpha\right\|_2^2$$

with $\beta=\Phi^\top\operatorname{diag}((Y_i)_i)\alpha$ at the optimum.

Note that above, $\operatorname{diag}((Y_i)_i)\Phi\beta\ge\mathbb{1}_n$ is the compact formulation of: for all $i\in\{1,\dots,n\}$,

$$Y_i\varphi(X_i)^\top\beta\ge 1.$$

This amounts to fix the margin hyperplanes to be shifted by 1 and -1, see Figure 6.3.

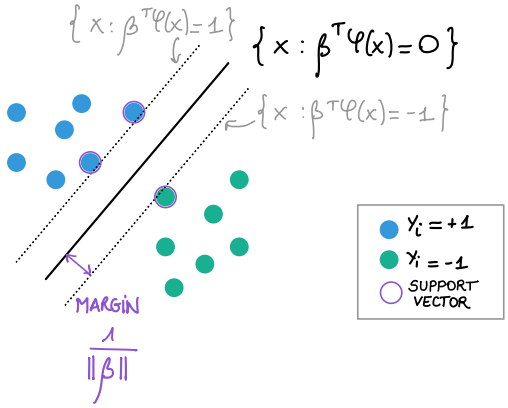


Figure 6.3: A traditional SVM with the margin hyperplanes shifted by 1 and -1.

Overall, the optimal $\beta^\star$ above is the solution of the separable SVM with vanishing regularization parameter, that is, of $\frac{1}{2}\|\beta\|_2 + C\sum_{i=1}^n (1 - Y_i\varphi(X_i)^\top\beta)_+$ for C large enough.

Divergence for the logistic regression with hands. The function F has an infimum equal to zero, which is not attained. However, for any sequence β_t such that all $Y_i\varphi(X_i)^\top\beta_t$ tend to infinity, we have $F(\beta_t) \to \inf_{\beta\in\mathbb{R}^d} F(\beta) = 0$.

In such a situation, gradient descent cannot converge to a point, and, to achieve small values of F, it has to diverge. It turns out that it diverges along a direction, that is, $\|\beta_t\|_2 \to +\infty$, with $\beta_t/\|\beta_t\|_2 \to \eta$ for some $\eta \in \mathbb{R}^d$ of unit ℓ_2-norm. See [SHN+18] for a proof. Here, we just show what the vector η is.

The gradient $\nabla F(\beta)$ is given by

$$\nabla F(\beta) = -\frac{1}{n}\sum_{i=1}^n \frac{\exp(-Y_i\varphi(X_i)^\top\beta)}{1+\exp(-Y_i\varphi(X_i)^\top\beta)} Y_i\varphi(X_i).$$

Asymptotically, β_t behaves as $\|\beta_t\|_2\eta$ with $\|\beta_t\|_2$ tending to infinity. By the structure of the sum of exponentials, the dominant term in $\nabla F(\beta_t)$ corresponds to the indices i for which $-Y_i\varphi(X_i)^\top\eta$ is the largest. Moreover, all of these values have to be negative (indeed we can only attain zero loss for well-classified training data). We denote by I this set. Thus,

$$\nabla F(\beta) \sim -\frac{1}{n}\sum_{i\in I} Y_i \exp\left(-\|\beta_t\|_2 Y_I\varphi(X_i)^\top\eta\right)\varphi(X_i).$$

Moreover, we must have $F(\beta_t)$ along $-u$ to diverge in the direction u, thus u has to be proportional to a vector $\sum_{i\in I}\alpha_i Y_i\varphi(X_i)$, where $\alpha \geq 0$ and $\alpha_i = 0$ as soon as i is not among the minimizers $Y_i\varphi(X_i)^\top\eta$. This is exactly the optimality condition for $\eta^\star$ above. Thus $\eta = \eta^\star$. Overall, we obtain a classifier corresponding to a minimum ℓ^2-norm.

See [LL19] for an extension beyond the linear classification case.

6.2 Interpolation is no longer synonym of bad generalization

The aim of this section is to present recent developments on the generalisation capabilities of neural networks, which in practice seem fantastic and which classical generalisation error bounds struggle to explain.

6.2.1 Preliminaries

A first lecture on machine/statistical learning traditionally warns the reader to the evils of overfitting, see Figure 6.4.

Typically the “capacity" of the space of learners $\mathscr{H}$ is controlled either by the number of parameters, or by some norms of its parameters. In particular, at the extreme right of the curve, when there is zero training error, the testing error may be arbitrarily large (bad), and the classical theoretical bound, such as Rademacher averages for $\mathscr{H}$ controlled by the ℓ^2-norm of some parameters (with a bound D), grows as $D/\sqrt{n}$, which can be typically quite large.

Proposition 6.3 (Estimation error)**.** *Assume a G-Lipschitz-continuous loss function ℓ, linear prediction functions with $\mathscr{F} = \{f_\theta(x) = \theta^\top\phi(x), \|\theta\|_2 \leq D\}$, where $\mathbb{E}\left[\|\phi(x)\|_2^2\right] \leq R^2$. Let $\hat{F} = f_{\hat{\theta}} \in \mathscr{F}$ be the minimizer of the empirical risk, then*

$$\mathbb{E}\left[\mathscr{R}(f_{\hat{\theta}})\right] \leq \inf_{\|\theta\|_2\leq D} \mathscr{R}(f_\theta) + \frac{2GRD}{\sqrt{n}}.$$

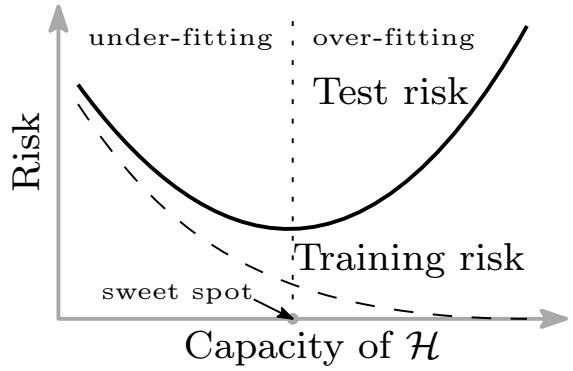


Figure 6.4: A typical learning curve about the bias-variance trade-off in the prediction when increasing the predictor complexity.

Model	Year	Nb of layers	Nb of param	Error
Shallow	<2012	-	-	> 25%
AlexNet	2012	8	61M	16.4%
VGG19	2014	19	144M	7.3%
GoogleNet	2014	22	7M	6.7%
ResNet-152	2015	152	60M	3.6%

Table 6.1: Performances of different architectures on the ImageNet dataset ($n = 500000$) in regard of the learning complexity captured here through the number of parameters or layers.

Here is a table summarizing the performances of different learners on the ImageNet dataset ($n = 500000$).

When the model is over-parameterized (in other words, the capacity gets very large), that is, when the number of parameters is large or the norm constraint allows for exact fitting, a new phenomenon occurs, where after the test error explodes as the capacity grows, it goes down again: this is the so-called *double descent* curve.

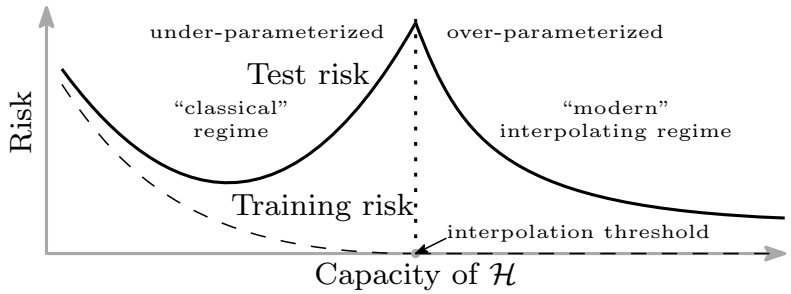


Figure 6.5: From [BHMM19] The learning story: to be continued.

6.2.2 Linear model

Montanari et al. has resolved this paradox in linear models [HMRT19], relying on non-asymptotic results for random matrices.

Consider a Gaussian random variable with mean 0 and covariance matrix identity, with n observations $X_1, \dots, X_n$, and responses $Y_i = X_i^\top \theta^\star + \varepsilon_i$, with ε_i normal with zero mean and variance $\sigma^2 I$.

We know the exact expression of the empirical risk minimizer (for which we know that gradient descent will converge to under proper initialization). Denote the design matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$, the non-centered covariance matrix $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$, and the kernel matrix $K = \mathbb{X}\mathbb{X}^\top$.

The excess risk is

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}_X\left[(X^\top \hat{\theta} - X^\top \theta^\star)^2\right] = \mathbb{E}_X(\hat{\theta} - \theta^\star) X X^\top (\hat{\theta} - \theta^\star) = (\hat{\theta} - \theta^\star)\Sigma(\hat{\theta} - \theta^\star) \\ &= \|\hat{\theta} - \theta^\star\|_2^2. \end{aligned}$$

Underparameterized regime. In the underparameterized regime, then the minimum norm empirical risk minimizer is simply the ordinary least-squares estimator, which is unbiased, that is $\mathbb{E}\left[\hat{\theta}\right] = \theta^\star$, and we have an expected excess risk equal to

$$\mathbb{E}_{\mathscr{D}_n}\left[R(\hat{\theta})\right] = \frac{\sigma^2}{n}\mathbb{E}\left[tr\left(\Sigma\hat{\Sigma}^{-1}\right)\right]$$

Indeed,

$$\begin{aligned} \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[R(\hat{\theta})\right] &= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\|\hat{\theta} - \theta^\star\|_{2,\Sigma}^2\right] \\ &= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y - \theta^\star\|_{2,\Sigma}^2\right] = \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\left(\mathbb{X}\theta^\star + \varepsilon\right) - \theta^\star\|_{2,\Sigma}^2\right] \\ &= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\|(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\|_{2,\Sigma}^2\right] = \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\varepsilon^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\right] \\ &= \mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[\, tr\left(\varepsilon^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon\right)\right] \\ &= \sigma^2\mathbb{E}_{\mathscr{D}_n}\left[\, tr\left(\Sigma(\mathbb{X}^\top\mathbb{X})^{-1}\right)\right] = \frac{\sigma^2}{n}\mathbb{E}\left[tr\left(\Sigma\hat{\Sigma}^{-1}\right)\right]. \end{aligned}$$

In our case, $\Sigma = I$, so that the expected risk boils down to

$$\mathbb{E}_{\mathscr{D}_n,\varepsilon}\left[R(\hat{\theta})\right] = \sigma^2\mathbb{E}_{\mathscr{D}_n}\left[\, tr(\mathbb{X}^\top\mathbb{X})^{-1}\right]$$

The matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ is Gaussian, so that the matrix $\mathbb{X}^\top\mathbb{X} \in \mathbb{R}^{p \times p}$ has a Wishart distribution, with n degrees of freedom:

- it is almost surely invertible if $n > p$,
- $\mathbb{E}_{\mathscr{D}_n}\left[\, tr\left((\mathbb{X}^\top\mathbb{X})^{-1}\right)\right] = \frac{p}{n-p-1}$ if $n \geq p+2$. The expectation is infinite for $n = p$ or $n = p+1$.

Proof. Here is a simple way to derive the expectation of an inverse Wishart matrix W^{-1} where $W = \sum_{i=1}^n \Sigma^{1/2} g_i g_i^T \Sigma^{1/2}$ for a covariance $\Sigma \in R^{p \times p}$ and i.i.d. standard vectors $g_i \sim N(0, I_p)$. The covariance Σ is assumed invertible. The point follows [jh].

The first observation is that

$$E[W^{-1}] = \Sigma^{-1/2} E\left[\left(\sum_{i=1}^n g_i g_i^T\right)^{-1}\right]\Sigma^{-1/2}$$

so that it is enough to treat the case $\Sigma = I_p$. Assume $\Sigma = I_p$ here after.

Concerning the non-diagonal terms of $E[W^{-1}]$, note that with identity covariance, $\sum_i g_i g_i^T$ and $\sum_i \tilde{g}_i \tilde{g}_i^T$ with $\tilde{g}_i = D g_i$ have the same distribution where $D = \mathrm{diag}(1, ..., 1, -1, 1, ..1)$ (only one sign changes). This implies that

$$E[W^{-1}] = D^{-1} E[W^{-1}] D^{-1}$$

so that $E[W^{-1}]_{ij} = 0$ for $i \neq j$ (outside of the diagonal).

Concerning the diagonal terms, by symmetry,

$$E[W^{-1}]_{ii} = \frac{1}{d} E[\text{trace}[W^{-1}].$$

The trace is also the sum of the eigenvalues $\lambda_i(W^{-1})$ of W^{-1}, or the following Frobenius norm:

$$E\big[\,\text{trace}[W^{-1}]\big] = E\sum_{i=1}^{d} \lambda_i(W)^{-1} = E[\|G^\dagger\|_F^2]$$

where $G \in R^{n\times p}$ is the matrix with n rows $g_1, ..., g_n$, and $\dagger$ denotes the pseudo-inverse. At this point, if $c_1, ..., c_p$ are the columns of $G^\dagger$, the above display is $\sum_{j=1}^{p} \|c_j\|_2^2$. Furthermore by definition of the pseudo-inverse, with $z_1, ..., z_d$ the rows of G, we have $c_j^T z_j = 1$ and $c_j^T z_k 0$ for $j \neq k$. This implies that c_j belongs to the orthogonal complement of $\{z_k, k \in \{1, ..., p\} \setminus j\}$. Since c_j belongs to the span of $z_1, ..., z_p$, it must be that $c_j = \theta_j Q_j z_j$ with $Q_j \in R^{n\times n}$ the orthogonal projection onto $\{z_k, k \in \{1, ..., p\} \setminus j\}^\perp$ and θ_j a scalar. The condition $c_j^T z_j = 1$ then reveals $\theta_j = \|Q_j z_j\|_2^{-2}$. Finally, $\|Q_j z_j\|_2^2$ has χ^2_{n-p+1} distribution as Q_j and z_j are independent thanks to G having i.i.d. $N(0,1)$ entries, hence

$$E[\text{trace}[W^{-1}]] = E\sum_{j=1}^{d} \|c_j\|_2^2 = E\sum_{j=1}^{d} \|Q_j z_j\|_2^{-2} = \frac{p}{(n-p+1)-2} = \frac{p}{n-p-1}$$

provided that we already know that the expectation of an inverse χ^2_ν distribution has expectation $1/(\nu-2)$ for $\nu > 2$.

□

In conclusion, in the underparameterized regime when $n \geq p+2$, the excess risk is equal to

$$\mathbb{E}\big[R(\hat{\theta})\big] = \sigma^2 \frac{p}{n-p-1}.$$

Overparameterized regime. In the overparameterized regime, when $n \leq p$, then the kernel matrix is almost surely invertible, and the minimum ℓ^2-norm interpolator $\hat{\theta}$ is equal to

$$\hat{\theta} = \mathbb{X}^\dagger Y = \mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1} Y = \mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1} \left(\mathbb{X}\theta^\star + \varepsilon\right)$$

The expected excess risk can be decomposed into a variance and a bias term:

(i) The variance term is equal to

$$\begin{aligned}\mathbb{E}\left[\varepsilon^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1} \mathbb{X}\Sigma\mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1} \varepsilon\right] &= \sigma^2 \mathbb{E}\left[tr\left(\mathbb{X}^\top\right)^{-1} \mathbb{X}\mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\right] \\ &= \sigma^2 \mathbb{E}\left[tr\left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\right] \\ &= \sigma^2 \frac{n}{p-n+1}\end{aligned}$$

when $p \geq n+2$ since we recognize a similar Wishart matrix as before (with the role of n and p reversed).

(ii) The bias term is equal to

$$\begin{aligned}\mathbb{E}\left[(\theta^\star)^\top \left(I - \mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1} \mathbb{X}\right)\theta^\star\right] &= \mathbb{E}\left[\left\|\text{Proj}_{\text{span}(X_1,...,X_n)^\perp}\left(\theta^\star\right)\right\|_2^2\right] = \mathbb{E}\left[\left\|\text{Proj}_{\text{Im}(\mathbb{X}^\top)^\perp}\left(\theta^\star\right)\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)\right\|_2^2\right]\end{aligned}$$

Indeed, the matrix $\mathbb{X}^\top \left(\mathbb{X}\mathbb{X}^\top\right)^{-1}\mathbb{X} \in \mathbb{R}^{p\times p}$ is the projection matrix on the rowspace of $\mathbb{X}$, which is a random subspace of dimension n corresponding to the linear span of the p-dimensional vectors $\{X_1,\dots,X_n\}$. By rotational invariance of the Gaussian distribution, this random subspace is uniformly distributed among all subspaces, and therefore, by rotational invariance, we can replace $\theta^\star$ by $\|\theta^\star\|_2 e_j$, for any of the canonical basis vector e_j in dimension p, that is

$$\mathbb{E}\left[(\theta^\star)^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\theta^\star\right] = \|\theta^\star\|_2\mathbb{E}\left[e_j^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}e_j\right]$$

and thus

$$\begin{aligned}\mathbb{E}\left[(\theta^\star)^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\theta^\star\right] &= \frac{\|\theta^\star\|_2}{p}\sum_{j=1}^{p}\mathbb{E}\left[e_j^\top\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}e_j\right]\\ &= \frac{\|\theta^\star\|_2}{p}\mathbb{E}\left[tr\left(\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\right)\right]\\ &= \|\theta^\star\|_2\frac{n}{p}.\end{aligned}$$

Thus the bias term leads to

$$\mathbb{E}\left[(\theta^\star)^\top\left(I-\mathbb{X}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}\right)\theta^\star\right] = \|\theta^\star\|_2\frac{p-n}{p}.$$

Therefore, the overall expected risk is

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2\frac{n}{p-n+1} + \|\theta^\star\|_2\frac{p-n}{p}.$$

Wrapping up One gets

$$\begin{cases}\text{if } p\le n-2, & \mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2\frac{p}{n-p-1},\\ \text{if } p\ge n+2 & \mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2\frac{n}{p-n+1} + \|\theta^\star\|_2\frac{p-n}{p},\end{cases}$$

as illustrated on Figure 6.7. The interpretation of these bounds are taken from [HMRT19]:

- The **bias** increases with p/n in the overparameterized regime, which is intuitive. When $p > n$, the min-norm least squares estimate is constrained to lie the row space of $\mathbb{X}$, the training feature matrix. This is a subspace of dimension n lying in a feature space of dimension p. Thus as p increases, so does the bias, since this row space accounts for less and less of the ambient p-dimensional feature space. Another way to see it is to note that the bias is nothing else than

$$\mathbb{E}\left[\left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)\right\|_2^2\right]$$

with $\dim(\text{Ker}(\mathbb{X})) = p-n$. Therefore

$$\min_{\theta\in\text{Ker}(\mathbb{X})}\|\theta-\theta^\star\|_2 = \left\|\text{Proj}_{\text{Ker}(\mathbb{X})}\left(\theta^\star\right)-\theta^\star\right\|_2^2$$

which $\searrow$ when $p\nearrow$ (the minimum is least on a larger subspace), so that $\text{Proj}_{\text{Ker}(\mathbb{X})}$ increases with p.

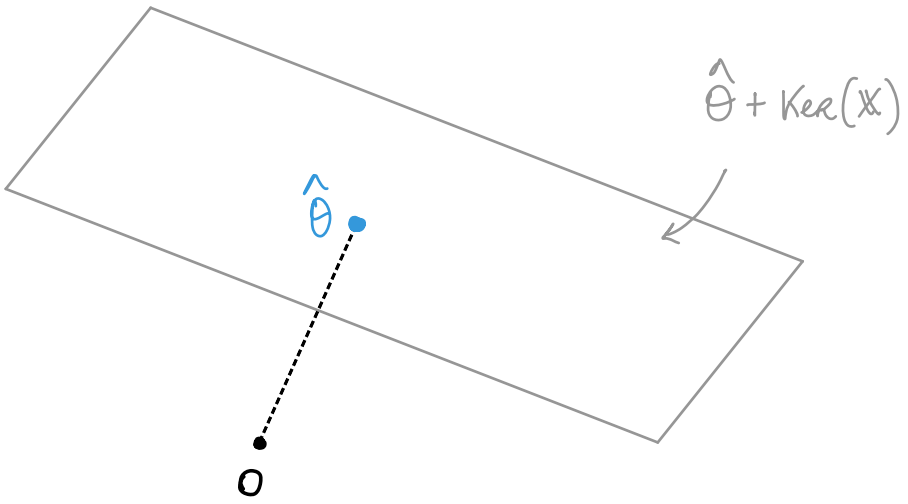


Figure 6.6: Intuition for the double descent phenomenon in linear models. When p increases, Ker($\mathbb{X}$) becomes larger, so there are more solutions to the system $Y = \mathbb{X}\theta$, so that $\|\hat{\theta}\|_2$ decreases, the bias increases with p/n and the variance decreases with p/n.

- In the overparameterized regime, the **variance** decreases with p/n. This may seem counter-intuitive at first, because it says, in a sense, that the min-norm least squares estimator becomes more regularized as p grows. However, this too can be explained intuitively: as p grows, the minimum ℓ^2-norm least squares solution—i.e., the minimum ℓ^2-norm solution to the linear system $\mathbb{X}\theta = Y$, for a training feature matrix $\mathbb{X}$ and response vector Y—will generally have decreasing ℓ^2-norm. Why? Compare two such linear systems: in each, we are asking for the min-norm solution to a linear system with the same Y, but in one instance we are given more columns in $\mathbb{X}$, so we can generally decrease the components of θ (by distributing them over more columns), and achieve a smaller ℓ^2-norm.
- Set the SNR= $\|\theta^\star\|_2^2/\sigma^2$. Note that the risk of the null estimator (i.e. $\hat{\theta} = 0$) is $\|\theta^\star\|_2^2$, which can be called the null risk. In the overparameterized regime, with an infinite sample size, and with $p/n \to \gamma$,
  - when SNR ≤ 1, the min-norm least squares risk is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above).
  - When SNR > 1, the min-norm least squares risk beats the null risk if and only if $\gamma >$ SNR$/$(SNR$-$ 1). It has a local minimum at $\gamma = \sqrt{\text{SNR}}/(\sqrt{SNR}-1)$, and approaches the null risk from below when $\gamma \to +\infty$.

Strikingly, interpolating predictors such as those studied here have been historically overlooked, at least for noisy data. Indeed, a classical prescription is to regularize the predictor by e.g., adding a ridge penalty "$\lambda\|\cdot\|_2^2$" (which adds λI to $\mathbb{X}\mathbb{X}^\top$), and leads to non-interpolating predictors.

In conclusion, this simple setting misses the approximation/variance trade-off. But the results are the best for $\gamma = 0$. Therefore, one can wonder if there exists linear settings for which there is a true benefit to go in the overparameterization regime? A partial answer is given in the misspecified case.

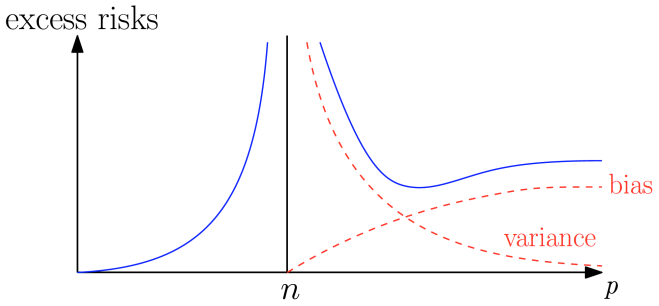


Figure 6.7: The double descent phenomenon in the linear case.

Conclusion on linear/kernel models in overparameterization regimes

Based on the previous sections, one can notice there is alignment of optimization and generalization in overparameterized linear (and kernel) models, since (S)GD converge to minimum norm interpolators having good generalization properties.

Remark 6.4 (No phenomenon when using regularization)**.** *When an extra (ridge) regularizer is used, then the double descent phenomenon is reduced (see [MM19]). In particular, if the regularization parameter λ is adapted for each number of observations, then the phenomenon totally disappears (for more details, see [MM19]).*

6.2.3 Misspecified linear model

Suppose that the model is now

$$Y_i = X_i^\top \theta^\star + \underbrace{W_i^\top}_{\substack{\text{unobserved} \\ \text{i.i.d.}}} \zeta^\star + \varepsilon_i,$$

in which $(W_i)_i$'s are i.i.d. unobserved features that help to explain the outcome Y. In such a case, the risk is going to compare $X^\top\hat{\theta}$ to $\mathbb{E}[Y|X,W] = X^\top\theta^\star + W^\top\zeta^\star$.

$$\begin{aligned} R(\hat{\theta}) &:= \mathbb{E}\left[\left(X^\top\hat{\theta} - \mathbb{E}[Y|X,W]\right)^2 \middle| \mathbb{X}\right] \\ &= \mathbb{E}\left[\left(X^\top\hat{\theta} - \mathbb{E}[Y|X]\right)^2 \middle| \mathbb{X}\right] + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - \mathbb{E}[Y|X,W]\right)^2 \middle| \mathbb{X}\right]}_{=:M_{\zeta^\star} \text{ approximation bias}} \text{ (by Pythagoras)} \end{aligned}$$

where $M_{\zeta^\star}$ is complex in general.

When all entries of X and W are i.i.d. and isotropic,

$$\begin{aligned} M_{\zeta^\star} &= \mathbb{E}\left[\left(X^\top\theta^\star - \left(X^\top\theta^\star + W^\top\zeta^\star\right)\right)^2 \middle| \mathbb{X}\right] = \mathbb{E}\left[\left(W^\top\zeta^\star\right)^2 \middle| \mathbb{X}\right] = \|\zeta^\star\|^2 \\ &= r^2(1-\kappa) \end{aligned}$$

with

- $r^2 = \|\theta^\star\|_2^2 + \|\zeta^\star\|_2^2$ corresponds to the signal strength
- $\kappa = \|\theta^\star\|_2^2 / r^2$ is the fraction of the signal explained by the covariates X only.

Theorem 6.5. *Assume the misspecified linear model, and assume that* (X, W) *has i.i.d. entries with zero mean, unit variance, and a finite moment of order* $8+\eta$*, for some* $\eta > 0$*. Also assume that for all* n, p*,* $r^2 = \|\theta^\star\|_2^2 + \|\zeta^\star\|_2^2$ *and* $\kappa = \|\theta^\star\|_2^2 / r^2$*. Then for the min-norm least squares estimator* $\hat{\theta}$*, as* $n, p \to \infty$*, with* $p/n \to \gamma$ *, it holds almost surely that*

$$\mathbb{E}\big[R(\hat{\theta})\big] \to \begin{cases} r^2(1-\kappa) + \big(r^2(1-\kappa)+\sigma^2\big)\frac{\gamma}{1-\gamma} & \text{for} \quad \gamma < 1 \\ r^2(1-\kappa) + r^2\kappa(1-\frac{1}{\gamma}) + \big(r^2(1-\kappa)+\sigma^2\big)\frac{1}{\gamma-1} & \text{for} \quad \gamma > 1 \end{cases}$$

In the independence setting, the dimension of the unobserved feature space does not play any role: we may equally well take it equal to ∞ for all n, p (i.e., infinitely many unobserved features). Note that

1. The first term $r^2(1-\kappa)$ is the misspecification bias (irreducible).
2. The second term equal to 0 when $\gamma < 1$ or to $r^2\kappa(1-\frac{1}{\gamma})$ is the bias.
3. The third term is the misspecification variance.
4. The last term is the variance.

By considering a polynomial decay for the approximation bias, i.e.

$$1 - \kappa(\gamma) = (1+\gamma)^{-a}$$

for some $a > 0$, the global minimum of the risk is achieved in the overparameterized regime.

> In such a case, linear over-parameterized predictors are sometimes preferable to any "classical" under-parameterized model.

6.2.4 A first non-linear model with random features

Consider now a one-hidden-layer neural network, in which we optimize only the output weights, see Figure 6.8. The model is the following:

- assume the input vectors $X_i \in \mathbb{R}^p$ with i.i.d. centered Gaussian entries, $X_i \sim \mathcal{N}(0, I_p)$;
- assume that the weights $W \in \mathbb{R}^{q \times p}$ between the input layer and the hidden one are such that each entry of W is a random $\mathcal{N}(0, 1/d)$ variable;
- call φ the activation function used in the hidden layer, and assume it is purely non-linear, i.e.

$$\mathbb{E}\Big[\varphi(G)\Big] = \mathbb{E}\Big[G\varphi(G)\Big] = 0, \qquad \text{for} \qquad G \sim \mathcal{N}(0,1).$$

The hypothesis if purely non-linear is not common, it is satisfied for instance for $\varphi(t) = a(|t| - b)$, for $a = \sqrt{\pi}/\sqrt{\pi-2}$ and $b = \sqrt{2}/\sqrt{\pi}$.

We optimize the weights θ of the output layer in terms of quadratic risk minimization penalized by a ridge term:

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta_\lambda} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \varphi(WX_i)^\top\theta\big)^2 + \lambda\|\theta\|_2^2.$$

This amounts to penalized linear regression with transformed features (the $\varphi(WX_i)$'s instead of the X_i's).

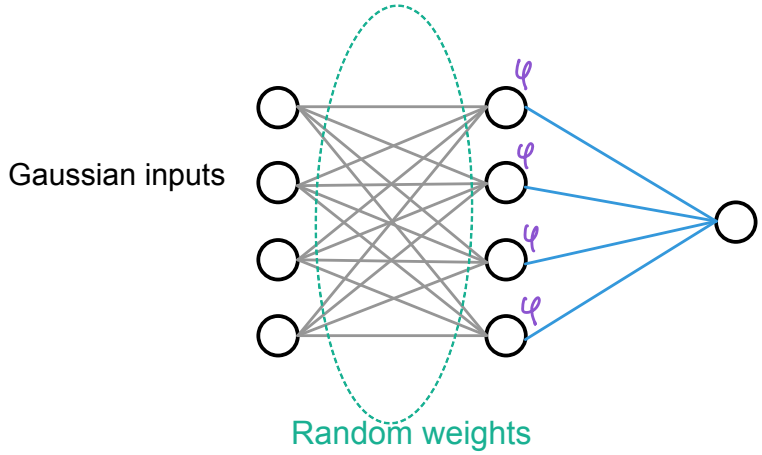


Figure 6.8: A first non-linear model, in which we optimize only the output weights (in blue). The input and the weights between the input layer and the hidden one are assumed to be Gaussian.

Theorem 6.6. *Assume that* $|\varphi(x)| \leq c_0(1+|x|)^{c_0}$ *for a constant* $c_0 > 0$. *Also, for* $G \sim N(0,1)$, *assume that the standardization conditions hold:* $\mathbb{E}[\varphi(G)] = 0$ *and* $\mathbb{E}[\varphi(G)^2] = 1$, $\mathbb{E}[G\varphi(G)] = 0$.

Then for $\gamma := p/n > 1$, *the variance satisfies, almost surely,*

$$\lim_{\lambda \to 0^+} \lim_{n,p,q \to \infty} V_X(\hat{\theta}_\lambda; \theta) = \frac{\sigma^2}{\gamma - 1},$$

which is precisely as in the case of linear isotropic features. Also, under a isotropic prior, namely $\mathbb{E}(\theta) = 0$, $Cov(\theta) = r^2 I_q/q$, *the Bayes bias* $B_X(\hat{\theta}_\lambda) := \mathbb{E}_\theta B_X(\hat{\theta}_\lambda; \theta)$ *satisfies, almost surely*

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} B_X(\hat{\theta}_\lambda) = \begin{cases} 0 & \text{for } \gamma < 1, \\ r^2(1 - 1/\gamma) & \text{for } \gamma > 1, \end{cases}$$

which is again as in the case of linear isotropic features

Note that this result is asymptotic, and heavily relies on the purely non-linear feature of the activation function that allows to retrieve standard asymptotics distribution got in the standard linear case.

When considering standard linear features, the out-of-sample risk can be decomposed in a bias and a variance terms:

$$\begin{aligned} R_{\mathbb{X}}(\hat{\theta}) &= \mathbb{E}\left[(X^\top \theta - X^\top \theta^\star)^2 | \mathbb{X} \right] = \mathbb{E}\left[\|\hat{\theta} - \theta^\star\|_\Sigma^2 | \mathbb{X} \right] \\ &= \underbrace{\left\| \mathbb{E}\left[\hat{\theta}|\mathbb{X}\right] - \theta^\star \right\|_\Sigma^2}_{\text{Bias}} + \underbrace{\operatorname{tr}\left[\operatorname{Cov}\left[\hat{\theta}|\mathbb{X}\right] \right]}_{\text{Variance}} \end{aligned}$$

Ideas of proof for the variance. Focus on the variance term, for the regularized parameter:

$$V_{\mathbb{Z}}(\hat{\theta}_\lambda) = \frac{\sigma^2}{n}\mathrm{tr}\Big[\Sigma\frac{\mathbb{Z}^\top\mathbb{Z}}{n}\left(\lambda I + \frac{\mathbb{Z}^\top\mathbb{Z}}{n}\right)^{-2}\Big] \qquad \text{for } \mathbb{Z} = \begin{pmatrix}\varphi(WX_1)^\top\\ \vdots \\ \varphi(WX_n)^\top\end{pmatrix}$$

$$= \frac{\sigma^2}{n} p \sum_{i=1}^{p} \frac{1}{p}\frac{\mu_i}{(\mu_i+\lambda)^2} \qquad \text{for } (\mu_i)_i \text{ the singular values of } \mathbb{Z}$$

$$\xrightarrow[n,p\to\infty]{} \sigma^2\gamma\int \frac{t}{\lambda+t^2} dMP_\gamma(t)$$

where MP_γ denotes the Marchenko-Pastur law of parameter γ. This is true by [Péc19, Theorem 1.1], for purely non-linear activation functions (entailing $\theta_2(f) = 0$ in [Péc19]), and by using the convergence of the spectral measure of $\frac{\mathbb{X}^\top\mathbb{X}}{n}$ towards the Marchenko-Pastur law.

We actually know an explicit form for the Stieltjes transform of this distribution, i.e.

$$m(-\lambda) = \int \frac{1}{t-\lambda} dMP_\gamma(t)$$

so we can deduce

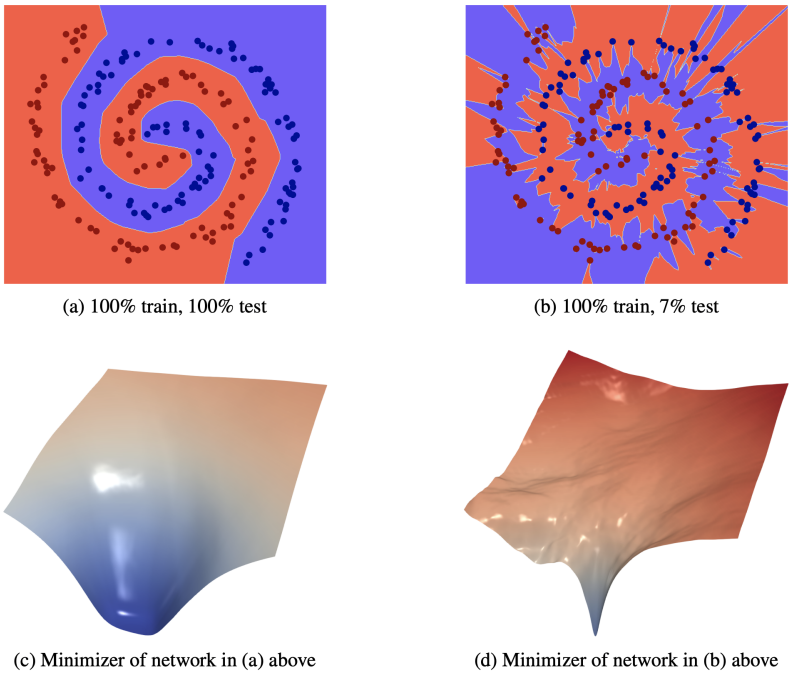
$$\begin{aligned}
M(\lambda) &= \int \frac{t}{\lambda+t^2} dMP_\gamma(t)\\
&= \frac{\partial}{\partial\lambda}\int \frac{-t}{\lambda+t} dMP_\gamma(t)\\
&= \frac{\partial}{\partial\lambda}\left(\underbrace{\int \frac{-t-\lambda}{\lambda+t} dMP_\gamma(t)}_{=-1} + \int \frac{\lambda}{\lambda+t} dMP_\gamma(t)\right)\\
&= \frac{\partial}{\partial\lambda}\left(\lambda m(-\lambda)\right)\\
&= \frac{\partial}{\partial\lambda}\left(\frac{-(1-\gamma+\lambda)+\sqrt{(1-\gamma+\lambda)^2+4\gamma\lambda}}{2\gamma}\right)\\
&= -\frac{1}{2\gamma} + \frac{1}{2\gamma}\left((1-\gamma+\lambda)^2+4\gamma\lambda\right)^{-1/2}(1+\gamma+\lambda)\\
&\xrightarrow[\lambda\to 0^+]{} \frac{1}{\gamma(\gamma-1)}.
\end{aligned}$$

Finally, when $\gamma > 1$,

$$V_{\mathbb{X}}(\hat{\theta}_\lambda) \xrightarrow[\lambda\to 0^+]{} \gamma\sigma^2\frac{1}{\gamma(\gamma-1)} = \frac{\sigma^2}{\gamma-1}.$$

6.2.5 Overparameterization/interpolation in neural network

Why overparameterizing in neural networks? It is often observed that for neural networks, depth efficiently helps to extract features in the dataset. Recent studies found that increasing both depth and width of a shallow model leads to very nice continuous limits, where PDE tools can be put in work. Besides, on the numerical side, one could argue that increasing the number of parameters could make harder the optimization/training of such complex architectures. However, networks with wide layers (larger than the sample size) can be shown to have no spurious minimizers [NMH18, Ngu19] (i.e. no local optima with bad generalization properties).



(a) 100% train, 100% test

(b) 100% train, 7% test

(c) Minimizer of network in (a) above

(d) Minimizer of network in (b) above

Figure 6.9: From [HEG+20]. Top: Decision boundaries of two networks with different parameters. Network (a) generalizes well. Network (b) generalizes poorly (perfect train accuracy, bad test accuracy). The flatness and large volume of (a) make it likely to be found by SGD, while the sharpness and tiny volume of (b) make this minimizer unlikely. Red and blue dots correspond to the training data. See Bottom: A slice through the loss landscapes around these minima reveals sharpness/flatness.

Bad consequences of overparameterization in neural networks? Beware, when training a NN with layers not wide enough, overparameterization usually entails existence of many local minimizers with potentially different statistical performances. Common practice advises to run stochastic gradient algorithm with random initialization and converges to parameters with very good practical prediction accuracy.Why is this simple approach actually often working? The goal of current research is to resolve these paradoxes.

Empirical observations: [HEG+20] on the importance of being flat. Flat minimizers (with a bad conditioned Hessian, and therefore with a flat attraction basin) are easier to reach and have better generalisation properties. Flatness seems to be nice for both generalization properties, and convergence of the used algorithms.

[BHMM19]:

6.2.6 What about non-parametric regression?

Other models in non-parametric regression have been addressed in [BRT19]. In this paper, the authors consider local-means (Nadaraya-Watson) estimator. They show that an interpolating kernel method using a singular kernel ($K(x) = ||x||^{-\alpha} \mathbb{1}_{||x||\leq 1}$) reaches minimax convergence rate for β-Hölder regular functions. By tuning the kernel bandwidth, the influence of the interpolation can be very limited and very localized around the training points. Anywhere else, the estimated function remains “smooth".

Bibliography

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[HEG+20] W Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. 2020.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[jh] jlewk (https://math.stackexchange.com/users/484640/jlewk). Simple(r) way to derive the expectation of an inverse wishart? Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3994137 (version: 2021-01-21).

[LL19] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[Ngu19] Quynh Nguyen. On connected sublevel sets in deep learning. In *International Conference on Machine Learning*, pages 4790–4799. PMLR, 2019.

[NMH18] Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. *arXiv preprint arXiv:1809.10749*, 2018.

[Péc19] S Péché. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

[SHN+18] 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.