Learning theory from first principles

Lecture 9: Special topics

Francis Bach

November 20, 2020

Class summary

- Implicit regularization of gradient descent
- Double descent
- Global convergence of gradient descent for two-layer neural networks

In this lecture, we will cover three recent topics within learning theory, all partially related to high-dimensional models (such as neural networks) in the "over-parameterized" regime, where the number of parameters is larger than the number of observations.

The number of parameters is not what characterizes in general the generalization capabilities of regularized learning methods.

1 Implicit bias of gradient descent

Given an optimization problem whose aim is to minimize some function $F(\theta)$ over some $\theta \in \mathbb{R}^d$, if there is a unique global minimizer θ_*, then the goal of optimization algorithms is to find this minimizer, that is, we want that the t-th iterate θ_t converges to that θ_*. When there are multiple minimizers (thus for a function which cannot be strongly convex), we showed only that $F(\theta_t) - \inf_{\theta \in \mathbb{R}^d} F(\theta)$ is converging to zero (and only if a minimizer exists, see Lecture 4).

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

This is interesting in machine learning because, when $F(\theta)$ is the empirical loss on n observations, and d is much larger than n, and no regularization is used, there are multiple minimizers, and an arbitrary

empirical risk minimizer is not expected to work well on unseen data. A classical solution is to use explicit regularization (e.g., ℓ_2-norms like in Lecture 2 and 6, or ℓ_1-norms like in Lecture 7). In this section, we show that optimization algorithms have a similar regularizing effect. In a nutshell, gradient descent usually leads to minimum ℓ_2-norm solutions. This shows that the chosen empirical risk minimizer is not arbitrary.

This will be explicitly shown for the quadratic loss, and partially only for the logistic loss. These results will be used in subsequent sections.

1.1 Least-squares

We consider $F(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2$, with $\Phi \in \mathbb{R}^{n\times d}$ such that $d > n$ and (for simplicity) $\Phi\Phi^\top \in \mathbb{R}^{n\times n}$ invertible (this is the kernel matrix). There are thus infinitely many (a whole affine subspace) solutions such that $y = \Phi\theta$, since the column space of Φ is the entire space $\mathbb{R}^n$ and θ has dimension $d > n$. We apply gradient descent with step-size $\gamma \leqslant \dfrac{1}{\lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)} = \dfrac{1}{\lambda_{\max}(\frac{1}{n}\Phi\Phi^\top)}$ starting from $\theta_0 = 0$. Thus, for any θ solution of $y = \Phi\theta$, we have, as shown in Lecture 4:

$$\theta_t - \theta = \big(I - \frac{\gamma}{n}\Phi^\top\Phi\big)^t(\theta_0 - \theta) = -\big(I - \frac{\gamma}{n}\Phi^\top\Phi)^t\theta,$$

leading to

$$\theta_t = \big[I - \big(I - \frac{\gamma}{n}\Phi^\top\Phi\big)^t\big]\theta.$$

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

If $\Phi = U\,\mathrm{Diag}(s)V^\top$ is the SVD decomposition of Φ, where $U \in \mathbb{R}^{n\times n}$ is orthonormal, and $V \in \mathbb{R}^{d\times n}$ has orthonormal columns and $s \in (\mathbb{R}_+^*)^n$, we can take $\theta = V\,\mathrm{Diag}(s)^{-1}U^\top y$ as one of the solutions (since then $\Phi\theta = U\,\mathrm{Diag}(s)V^\top V\,\mathrm{Diag}(s)^{-1}U^\top y = U\,\mathrm{Diag}(s)\,\mathrm{Diag}(s)^{-1}U^\top y = UU^\top y = y$) and get:

$$\theta_t = V\,\mathrm{Diag}((1 - (1 - \gamma s_i^2/n)^t)s_i^{-1})U^\top y.$$

Since each $s_i > 0$, and $\gamma_i \leqslant \frac{n}{\max_i s_i^2}$, we have

$$s_i^{-1} \leqslant (1 - (1 - \gamma s_i^2/n)^t)s_i^{-1} \leqslant s_i^{-1}(1 - (1 - \gamma \min_i s_i^2/n)^t),$$

and thus

$$\big\|\theta_t - V\,\mathrm{Diag}(s)^{-1}U^\top y\big\|_2 \leqslant (1 - \gamma \min_i s_i^2/n)^t\|V\,\mathrm{Diag}(s)^{-1}U^\top y\big\|_2.$$

We thus get linear convergence to $V\,\mathrm{Diag}(s)^{-1}U^\top y$, which happens to be the minimum ℓ_2-norm solution, because all solutions to $y = \Phi\theta$ can be written as $V\,\mathrm{Diag}(s)^{-1}U^\top y$ plus a vector which is orthogonal to the column space of V.

Moreover, with $\gamma_i = \frac{n}{\max_i s_i^2}$ (largest allowed step-size), we get a rate of $\Big(1 - \gamma\frac{\min_i s_i^2}{\max_i s_i^2}\Big)^t$.

Lojasiewicz's inequality (♦). It turns out that linear convergence here can be shown directly for any L-smooth function, for which we have the so-called Lojasiewicz's inequality:

$$\forall\theta \in \mathbb{R}^d,\ F(\theta) - F(\theta_*) \leqslant \frac{1}{2\mu}\|F'(\theta)\|_2^2, \tag{1}$$

for some $\mu > 0$.

We have seen in Lecture 4 that this is a consequence of μ-strong-convexity, but this can be satisfied without strong convexity. For example, for any least-squares example, we have, for any minimizer θ_*:

$$\|F'(\theta)\|_2^2 = \|\frac{1}{n}\Phi^\top\Phi(\theta-\theta_*)\|_2^2 = \frac{1}{n^2}(\theta-\theta_*)^\top\Phi^\top\Phi\Phi^\top\Phi(\theta-\theta_*) \geqslant \frac{\lambda_{\min}^+(\Phi\Phi^\top)}{n^2}(\theta-\theta_*)^\top\Phi^\top\Phi(\theta-\theta_*),$$

where $\lambda_{\min}^+(\Phi\Phi^\top) = \lambda_{\min}^+(\Phi^\top\Phi)$ is the smallest non-zero eigenvalue of $\Phi\Phi^\top$ (which is also the one of $\Phi^\top\Phi$). Thus, we have

$$\|F'(\theta)\|_2^2 \geqslant \frac{\lambda_{\min}^+(K)}{n^2}\|\Phi(\theta-\theta_*)\|_2^2 = \frac{2\lambda_{\min}^+(K)}{n}[F(\theta)-F(\theta_*)].$$

Thus, Eq. (1) is satisfied with $\mu = \frac{1}{n}\lambda_{\min}^+(K)$, where $K = \Phi\Phi^\top \in \mathbb{R}^{n\times n}$ is the kernel matrix. Note that this includes also the strongly-convex case since $\lambda_{\min}^+(\Phi^\top\Phi) \geqslant \lambda_{\min}(\Phi^\top\Phi)$.

When Eq. (1) is satisfied, we have for the t-th iterate of gradient descent with step-size $\gamma = 1/L$, following the analysis of Lecture 4:

$$F(\theta_t) - F(\theta_*) \leqslant F(\theta_{t-1}) - F(\theta_*) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2 \leqslant \big(1-\frac{\mu}{L}\big)\big[F(\theta_{t-1})-F(\theta_*)\big].$$

Moreover, we can then show that the iterates x_t are also converging to a minimizer of F (see, [1, 2] for more details).

Alternative proof. If started at $\theta_0 = 0$, gradient descent techniques (stochastic or not) will always have iterates θ_t which are linear combinations of row of Φ, that is, of the form $\theta_t = \Phi^\top\alpha_t$ for some $\alpha_t \in \mathbb{R}^n$. This is an alternative algorithmic version of the representer theorem from Lecture 6.

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

It turns out that this is exactly the minimum ℓ_2-norm solution as, by standard Lagrangian duality:

$$\begin{aligned}
\inf_{\theta\in\mathbb{R}^d}\frac{1}{2}\|\theta\|_2^2 \text{ such that } y=\Phi\theta &= \inf_{\theta\in\mathbb{R}^d}\sup_{\alpha\in\mathbb{R}^n}\frac{1}{2}\|\theta\|_2^2 + \alpha^\top(y-\Phi\theta) \\
&= \sup_{\alpha\in\mathbb{R}^n}\alpha^\top y - \frac{1}{2}\|\Phi^\top\alpha\|_2^2 \text{ with } \theta = \Phi^\top\alpha \text{ at optimum,} \\
&= \sup_{\alpha\in\mathbb{R}^n}\alpha^\top y - \frac{1}{2}\alpha^\top K\alpha.
\end{aligned}$$

The last problem is exactly solved for $\alpha = K^{-1}y$. Note that in Lecture 6, we used this formula for function interpolation to compare different RKHSs.

1.2 Separable classification

We now consider logistic regression, that is,

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

with $\Phi \in \mathbb{R}^{n\times d}$ the design matrix such that $d > n$ and $\Phi\Phi^\top$ invertible.

Maximum margin classifier. Since $\Phi\Phi^\top$ is invertible, there exists $\eta \in \mathbb{R}^d$ of unit norm such that $\forall i \in \{1,\dots,n\}$, $y_i\varphi(x_i)^\top\eta > 0$. We denote by η_* the one such that

$$\min_{i\in\{1,\dots,n\}} y_i\varphi(x_i)^\top\eta$$

is maximal (and thus strictly positive). That is, η_* solves the following problem, which can be rewritten as, using Lagrange duality:

$$\begin{aligned}
\sup_{\|\eta\|_2\leqslant 1} \min_{i\in\{1,\dots,n\}} y_i\varphi(x_i)^\top\eta &= \sup_{\|\eta\|_2\leqslant 1, t\in\mathbb{R}} t \text{ such that } \forall i \in \{1,\dots,n\},\ y_i\varphi(x_i)^\top\eta \geqslant t \\
&= \inf_{\alpha\in\mathbb{R}^n_+} \sup_{\|\eta\|_2\leqslant 1, t\in\mathbb{R}} t + \sum_{i=1}^n \alpha_i(y_i\varphi(x_i)^\top\eta - t) \\
&= \inf_{\alpha\in\mathbb{R}^n_+} \Big\| \sum_{i=1}^n \alpha_i y_i\varphi(x_i)\Big\|_2 \text{ such that } \sum_{i=1}^n \alpha_i = 1,
\end{aligned}$$

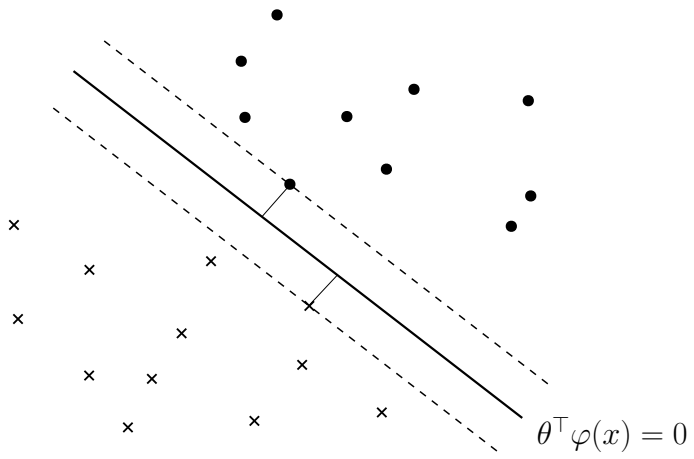
with $\eta \propto \sum_{i=1}^n \alpha_i y_i\varphi(x_i)$ at optimum. Moreover, by complementary slackness non-negative α_i is non zero only for i attaining the minimum in $\min_{i\in\{1,\dots,n\}} y_i\varphi(x_i)^\top\eta$.

Moreover, because of homogeneity, we want $\min_{i\in\{1,\dots,n\}} y_i\varphi(x_i)^\top\eta$ large and $\|\eta\|_2$ small, and we can decide to constrain the first and minimize the second one. In other words, we can see η_* as the direction of the solution θ_* of:

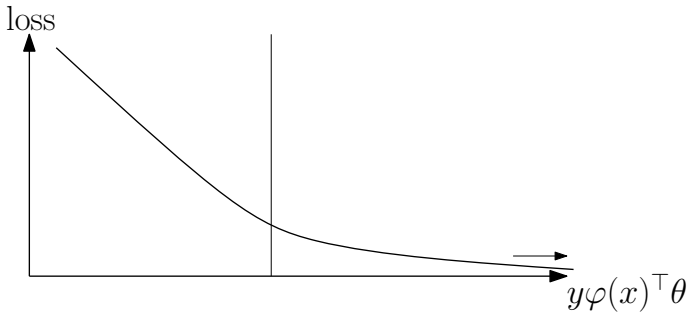
$$\begin{aligned}
\inf_{\theta\in\mathbb{R}^d} \frac{1}{2}\|\theta\|_2^2 \text{ such that } \operatorname{Diag}(y)\Phi\theta \geqslant 1_n &= \inf_{\theta\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}^n_+} \frac{1}{2}\|\theta\|_2^2 + \alpha^\top(1_n - \operatorname{Diag}(y)\Phi\theta) \\
&= \sup_{\alpha\in\mathbb{R}^n_+} \alpha^\top 1_n - \frac{1}{2}\|\Phi^\top \operatorname{Diag}(y)\alpha\|_2^2 \text{ with } \theta = \Phi^\top \operatorname{Diag}(y)\alpha \text{ at optimum.}
\end{aligned}$$

Note that above, $\operatorname{Diag}(y)\Phi\theta \geqslant 1_n$ is the compact formulation of $\forall i \in \{1,\dots,n\}$, $y_i\varphi(x_i)^\top\theta \geqslant 1$.

The θ_* above is the solution of the separable SVM with vanishing regularization parameter, that is, of $\frac{1}{2}\|\theta\|_2 + C\sum_{i=1}^n (1 - y_i\varphi(x_i)^\top\theta)_+$ for C large enough.



Divergence and convergence of directions. The function F has an infimum equal to zero, which is not attained. However, for any sequence θ_t such that all $y_i\varphi(x_i)^\top\theta_t$ tend to infinity, we have $F(\theta_t) \to \inf_{\theta\in\mathbb{R}^d} F(\theta) = 0$.



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, $\|\theta_t\|_2 \to +\infty$, with $\frac{1}{\|\theta_t\|_2}\theta_t \to \eta$ for some $\eta \in \mathbb{R}^d$ of unit ℓ_2-norm. See [3] for a proof. Here, we just show what the vector η is.

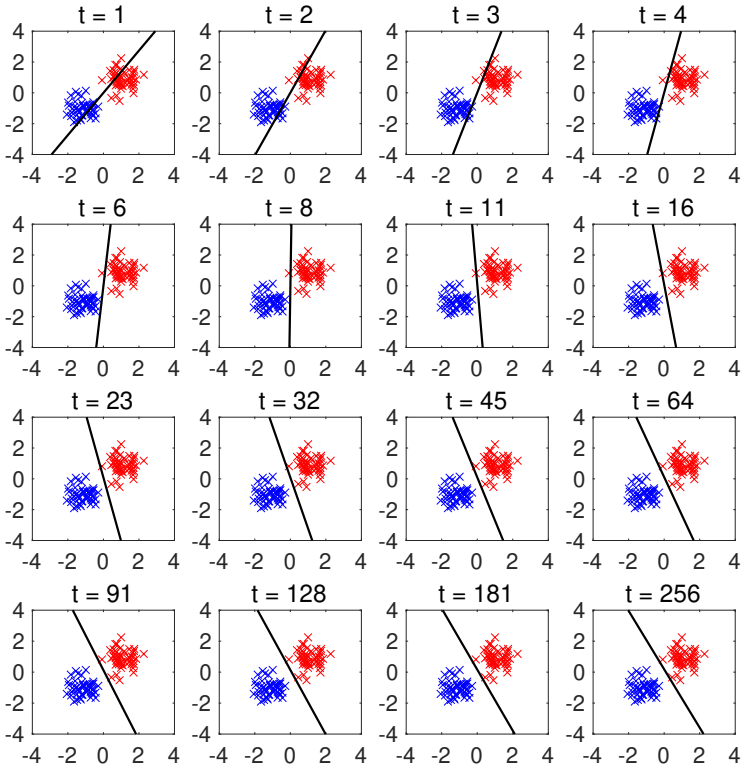
The gradient $F'(\theta)$ is equal to $F'(\theta) = -\dfrac{1}{n}\displaystyle\sum_{i=1}^n \frac{\exp(-y_i\varphi(x_i)^\top\theta)}{1+\exp(-y_i\varphi(x_i)^\top\theta)} y_i\varphi(x_i)$.

Asymptotically, θ_t will behave as $\|\theta_t\|\eta$, with $\|\theta_t\|_2$ tending to infinity. Thus, because we have a sum of exponentials with scale that goes to infinity, the dominant term in $F'(\theta_t)$ corresponds to the indices i for which $-y_i\varphi(x_i)^\top\eta$ is 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,

$$F'(\theta_t) \sim -\frac{1}{n}\sum_{i\in I} y_i \exp(-\|\theta_t\|_2 y_i\varphi(x_i)^\top\eta)\varphi(x_i).$$

Moreover, we must have $F'(\theta_t)$ along $-u$ to diverge in the direction u, thus u has to be proportional to a vector $\displaystyle\sum_{i\in I}\alpha_i y_i\varphi(x_i)$, where $\alpha \geqslant 0$, and $\alpha_i = 0$ as soon as i is not among the minimizers of $y_i\varphi(x_i)^\top\eta$. This is exactly the optimality condition for η_* above. Thus $\eta = \eta_*$.

Overall, we obtain a classifier corresponding to a minimum ℓ_2-norm. See examples in two dimensions below.



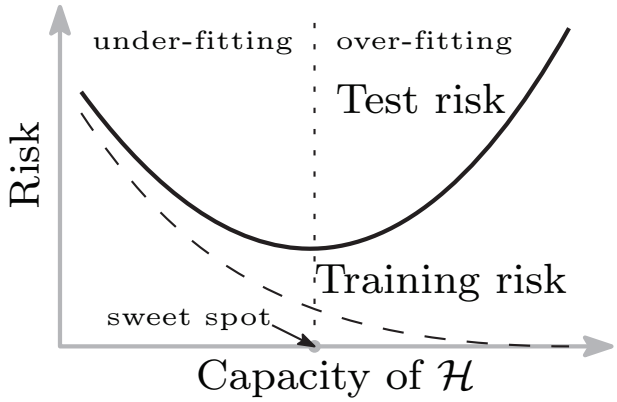
General result. The result above extends to more general situation beyond linear classification. See [4].

2 Double descent

In this section, we consider a recent and interesting phenomenom described in several recent works [5, 6, 7, 8].

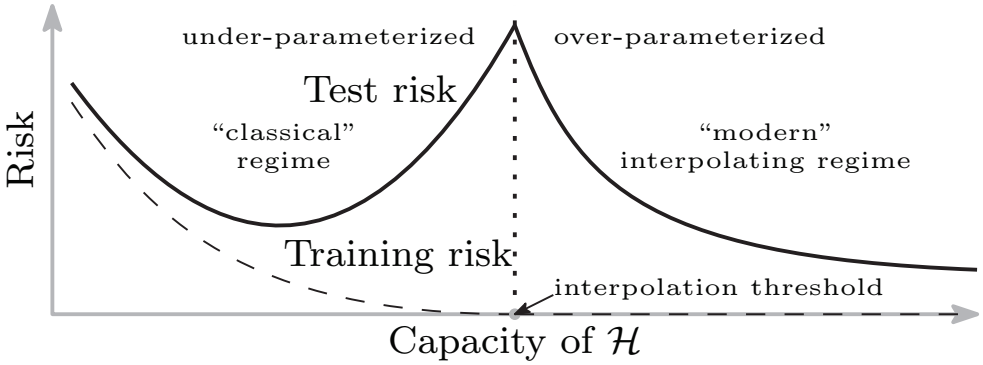
2.1 The double descent phenomenon

As seen in Lectures 1 and 3, typical learning curves look like the one below (figure taken from [5]):



Typically the "capacity" of the space of functions $\mathcal{H}$ is controlled either by the number of parameters, either 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 bad, and the bound that we have used in Lecture 3, such as Rademacher averages for $\mathcal{H}$ controlled by the ℓ_2-norm of some parameters (with a bound D), grows as $D/\sqrt{n}$, which can typically be quite large. These bounds were true for *all* empirical risk minimizers. In this section we will focus on a particular one, namely the one obtained by unconstrained gradient descent.

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 phenomenom occurs, where after the test error explodes as the capacity grows, it goes down again (figure also taken from [5]):



The goal of this section is to understand why. But before this let's present some empirical evidence, from toy examples and research papers.

There may be no double descent phenomenon if other empirical risk minimizers are used (instead of the one obtained by (stochastic) gradient descent).

2.2 Empirical evidence

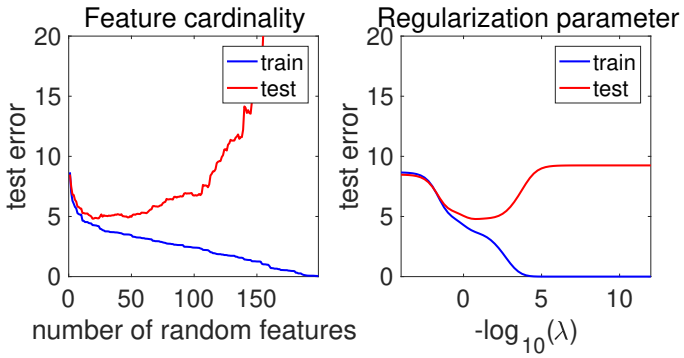
Toy example with random feature. We consider a random feature models like in Lectures 6 and 9, with the features $(v^\top x)_+$, for neurons v sampled uniformly on the unit spheres. We consider $n = 200$, $d = 5$

with input data distributed uniformly on the unit sphere, and we consider $y = \left(\frac{1}{4} + (v_*^\top x)^2\right)^{-1} + \mathcal{N}(0, \sigma^2)$, $\sigma = 2$, for some random v_*.

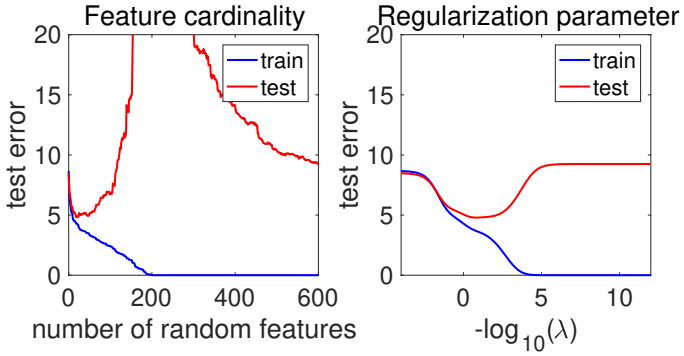
We sample m random features $v_1, \ldots, v_m$ uniformly on the sphere, and we learn parameters $\theta \in \mathbb{R}^m$ by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \theta_j (v_j^\top x_i)_+ \right)^2 + \lambda \|\theta\|_2^2. \tag{2}$$

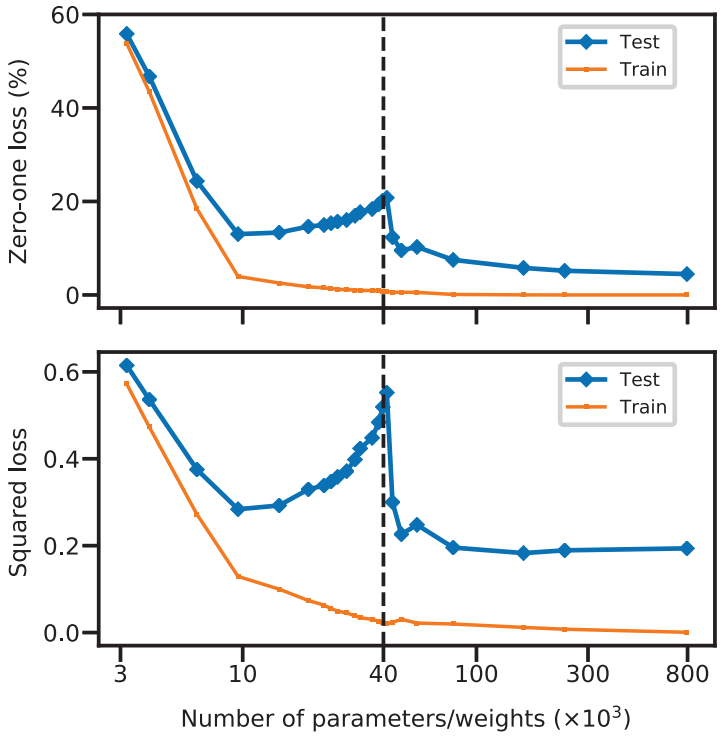
Below we report test errors after learning with gradient descent until convergence: (Left) varying m with $\lambda = 0$, (Right) varying λ with $m = +\infty$.



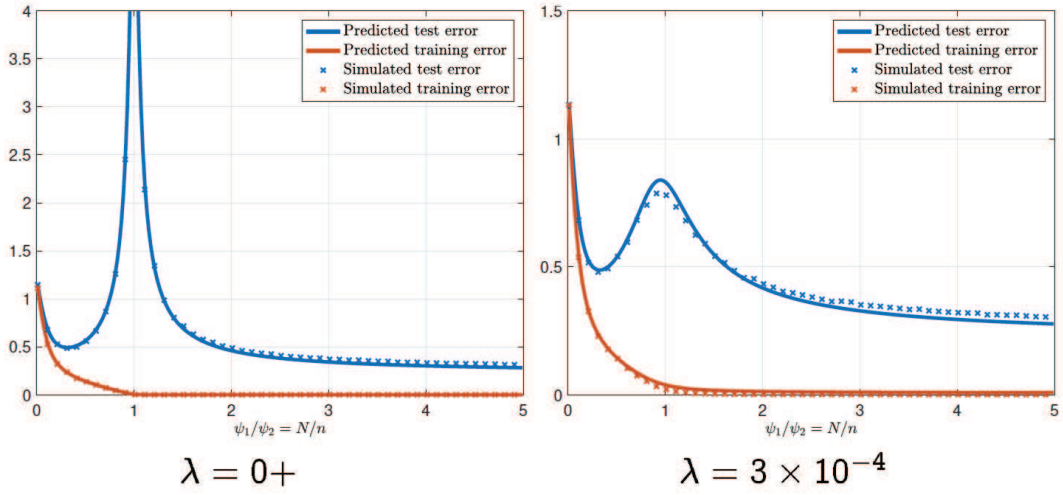
In the left curve above, the number of random features m is left less than n, as the test error diverges. But, when this number m is allowed to grow past n, we see the double descent phenomenon below (the right curve does not move). Similar experiments are shown in [5, 6].



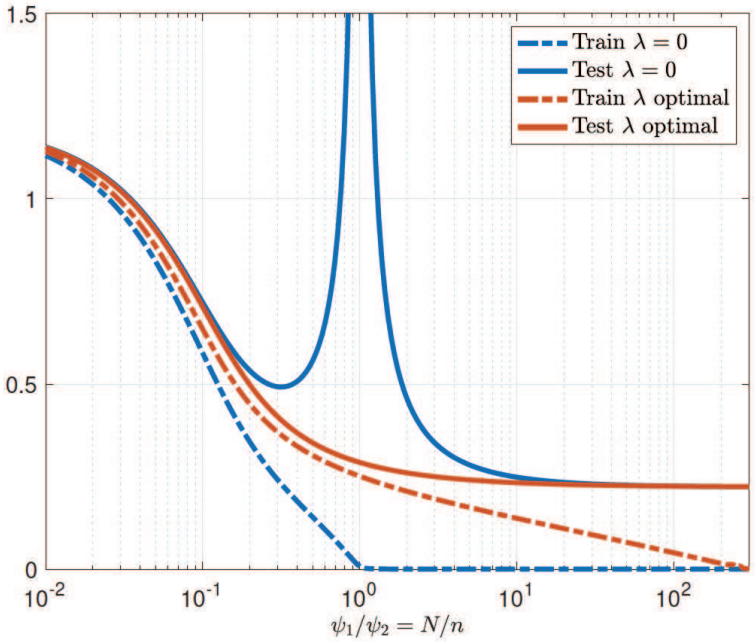
Neural networks. We consider here a single hidden-layer fully connected network, on the MNIST dataset of handwritten digits, trained by stochastic gradient descent. As shown below (figure taken from [5]), we see a similar spike in errors around $n = 40000$ which is the number of training data points.



No phenomenom when using regularization. When an extra regularizer is used, that is $\lambda \neq 0$ in Eq. (2), then the double descent phenomenom is reduced (see plots below from [6], in particular the right one), where "$\psi_1/\psi_2 = N/n$" is exactly m/n.



If the regularization parameter λ is adapted for each m, then the phenomenom totally disappears (plot below from [6]).



2.3 Simplest analysis

We consider a Gaussian random variable with mean 0 and covariance matrix identity, with n obervations $x_1, \dots, x_n$, and responses $y_i = x_i^\top \theta_* + \varepsilon_i$, with ε_i normal with mean zero and variance $\sigma^2 I$. We will compute an exact expectation of the risk of the minimum norm empirical risk minimizer (as detailed in Section 1.1), which is the one gradient descent converges to. We denote by $X \in \mathbb{R}^{n \times d}$ the design matrix, and $\hat{\Sigma} = \frac{1}{n} X^\top X$ the non-centered covariance matrix, and by $K = XX^\top \in \mathbb{R}^{n \times n}$ the kernel matrix.

The excess risk is $R(\hat{\theta}) = (\hat{\theta} - \theta_*)\Sigma(\hat{\theta} - \theta_*) = \|\hat{\theta} - \theta_*\|_2^2$.

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 $\mathbb{E}\big[\hat{\theta}\big] = \theta_*$, and we have an expected excess risk equal to (see the random design analysis from Lecture 2):

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

As seen in Lecture 2, the expected risk is equal to

$$\sigma^2 \mathbb{E}\big[\operatorname{tr}\big((X^\top X)^{-1}\big)\big],$$

where $X \in \mathbb{R}^{n \times d}$ is the associated design matrix. The matrix $X^\top X \in \mathbb{R}^{d \times d}$ has a Wishart distribution with n degrees of freedom. It is almost surely invertible if $n \geqslant d$, and is such that $\mathbb{E}\big[\operatorname{tr}\big((X^\top X)^{-1}\big)\big] = \frac{d}{n-d-1}$ if $n \geqslant d+2$. The expectation is infinite for $n = d$ and $n = d+1$.

Therefore, we have for $n \geqslant d+2$:

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

Overparameterized regime. In the overparameterized regime, when $n \leqslant d$, then the kernel matrix is almost surely invertible, and the minimum ℓ_2-norm interpolator $\hat{\theta}$ is equal to (using the formulas above) $\hat{\theta} = X^\top (XX^\top)^{-1} y = X^\top (XX^\top)^{-1} X\theta_* + X^\top (XX^\top)^{-1}\varepsilon$. The expected excess risk decomposes into a bias and a variance term.

The *variance* term is equal to, since $\Sigma = I$,

$$\mathbb{E}\big[\varepsilon^\top (XX^\top)^{-1} X\Sigma X^\top (XX^\top)^{-1}\varepsilon\big] = \sigma^2 \mathbb{E}\Big[\operatorname{tr}\big((XX^\top)^{-1} XX^\top (XX^\top)^{-1}\big)\Big] = \sigma^2 \mathbb{E}\Big[\operatorname{tr}\big((XX^\top)^{-1}\big)\Big],$$

which is now a Wishart related expectation with the order of n and d reversed, that is, $\sigma^2 \frac{n}{d-n-1}$ for $d \geqslant n+2$.

The *bias term* is equal to

$$\mathbb{E}\Big[\|\Sigma^{1/2}\big(X^\top (XX^\top)^{-1} X\theta_* - \theta_*\big)\|_2^2\Big].$$

Since $\Sigma = I$, then we get a bias term equal to

$$\mathbb{E}\Big[\theta_*^\top \big(I - X^\top (XX^\top)^{-1} X\big)\theta_*\Big].$$

The matrix $X^\top (XX^\top)^{-1} X \in \mathbb{R}^{d\times d}$ is the projection matrix on a random subspace of size 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 θ_* by $\|\theta_*\|_2 e_j$, that is,

$$\mathbb{E}\Big[\theta_*^\top X^\top (XX^\top)^{-1} X\theta_*\Big] = \|\theta_*\|_2^2 \cdot \mathbb{E}\Big[e_j^\top X^\top (XX^\top)^{-1} e_j\Big]$$

for any of the d canonical basis vectors e_j, $j = 1, \dots, d$, and thus

$$\mathbb{E}\Big[\theta_*^\top X^\top (XX^\top)^{-1} X\theta_*\Big] = \frac{\|\theta_*\|_2^2}{d} \sum_{j=1}^d \mathbb{E}\Big[e_j^\top X^\top (XX^\top)^{-1} Xe_j\Big] = \frac{\|\theta_*\|_2^2}{d} \mathbb{E}\big[\operatorname{tr}\big[X^\top (XX^\top)^{-1} X\big]\big] = \frac{\|\theta_*\|_2^2 n}{d}.$$

Thus the bias term is equal to $\frac{d-n}{d}\|\theta_*\|_2^2$.

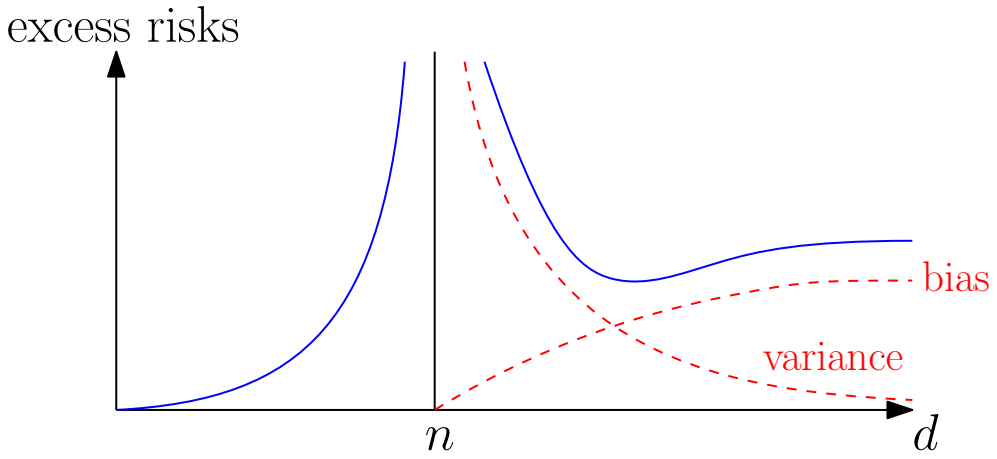
Therefore the overall expected risk is

$$\frac{\sigma^2 n}{d-n-1} + \|\theta_*\|_2^2 \frac{d-n}{d}.$$

Summary. We get

$$\begin{aligned} &\text{if } d \leqslant n-2, \qquad \mathbb{E}\big[R(\hat{\theta})\big] = \sigma^2 \frac{d}{n-d-1} \\ &\text{if } d \geqslant n+2, \qquad \mathbb{E}\big[R(\hat{\theta})\big] = \frac{\sigma^2 n}{d-n-1} + \|\theta_*\|_2^2 \frac{d-n}{d}. \end{aligned}$$

This leads to the following picture.



This extends to more general sampling models, see [8], and to random non-linear features [6].

3 Global convergence of gradient descent for two-layer neural networks

See the two blog posts based on [9, 10]:

- `https://francisbach.com/gradient-descent-neural-networks-global-convergence/`
- `https://francisbach.com/gradient-descent-for-wide-two-layer-neural-networks-implicit-bias/`

Acknowledgements

These class notes have been adapted from the notes of many colleagues I have the pleasure to work with, in particular Lénaïc Chizat, Pierre Gaillard, Alessandro Rudi and Simon Lacoste-Julien.

References

[1] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of lojasiewicz inequalities and applications. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[2] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[3] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *In International Conference on Machine Learning*, 2018.

[4] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

[5] 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.

[6] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[7] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019.

[8] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Technical Report 903.08560, arXiv, 2019.

[9] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

[10] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of COLT*, 2020.