Algorithmic Foundations of Learning **Lecture 14**

Least Squares Regression. Implicit Bias and Implicit Regularization

Lecturer: Patrick Rebeschini *Version: December 8, 2021*

14.1 Introduction

In the previous lectures we investigated the problem of estimating an unknown vector from noisy linear measurements in high-dimension under sparsity constrains. The way we approached the problem was by means of *explicit* regularization. Recall the definition of the empirical risk:

$$R(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, w\rangle - Y_i)^2 = \frac{1}{n}\|\mathbf{x}w - Y\|_2^2.$$

The explicit regularization approach that we used last time is the following:

1. Consider the Lasso estimator $W^{p1} = \operatorname{argmin}_{w\in\mathbb{R}^d} R(w) + 2\lambda\|w\|_1$ (note that this definition deviates by a factor 2 compared to the one used in the last lecture; the minimizer remains unchanged).
2. Tune the regularization parameter, e.g. $\lambda = \|\nabla R(w^\star)\|_\infty = \sigma\frac{\|\mathbf{x}^\top\xi\|_\infty}{n}$.
3. Run a gradient descent method (e.g. ISTA) $(W_t)_{t\geq 0}$ to approximate W^{p1}.

Under restricted strong convexity one can control both terms in the following error decomposition:

$$\boxed{\|W_t - w^\star\|_2 \leq \underbrace{\|W_t - W^{p1}\|_2}_{\text{optimization error}} + \underbrace{\|W^{p1} - w^\star\|_2}_{\text{statistics error}}}.$$

Last time we only showed how to control the statistics error, while for the optimization error we referred to literature that establishes exponential convergence of the ISTA algorithm up the statistical error.

In this lecture we consider again the problem of estimating an unknown vector from noisy linear measurements (not necessarily in high-dimension; the argument we are going to presents is general), when we replace the sparsity assumption by the assumption that the unknown parameter $w^\star$ lies in the span of the data (as we are going to see below, this assumption makes the estimation problem well-posed). This time, however, we take a different approach based on *implicit* regularization:

1. Run the gradient descent algorithm $(W_t)_{t\geq 0}$ designed to find a minimizer of R.
2. Tune the parameters of gradient descent (i.e. choice of initial condition $W_0^\star$, learning rate $\eta^\star$, and stopping time $t^\star$) to directly solve the statistical problem and minimize the estimation error at the stopping time: $\|W_{t^\star} - w^\star\|_2$.

At first glance, this approach sounds magical. Note that we are running an algorithm $(W_t)_{t\geq 0}$ that is designed to converge to a minimizer of the empirical risk R, namely $\lim_{t\to\infty} W_t \in \operatorname{argmin}_{w\in\mathbb{R}^d} R(w)$. However, for our statistical purposes, we do not want to find a minimizer of R! The fact that we can tune the parameters of the algorithm and, in particular, the stopping time $t^\star$ so that $\|W_{t^\star} - w^\star\|_2$ is "small" (in fact, minimax optimal, using a notion of statistical optimality that we will define later on in this course) is indeed surprising. In other words, we are saying that on the optimization path $(W_t)_{t\geq 0}$ towards a minimizer of R, gradient descent visits a point $W_{t^\star}$ that is close to the unknown statistical parameter $w^\star$ that has generated the data! This is an instance of implicit/algorithmic regularization. In the process of establishing this result, we will also see that gradient descent has a certain *implicit bias*: the minimizer of R the algorithm converges towards, i.e. $\lim_{t\to\infty} W_t$, has a specific structure: it is the minimizer with the smallest ℓ_2 norm.

From a computational view point, implicit regularization is more advantageous than explicit regularization as it leads to a cheaper way to perform *model selection.* In fact, to perform explicit regularization we need to appropriately tune the penalty/constraint parameters (e.g. λ in the Lasso or Ridge Regession case). While the theory tells us how to tune these parameters, in practice exact tuning is not possible as we do not know all the parameters required to do it (e.g., the noise term ξ, which we do not observe). For this reasons, in practice one has to perform model selection, solving multiple optimization problems (one per each different choice of the regularization parameters) and choosing the parameters (i.e. the model) that performs better on a validation set. On the other hand, implicit regularization allows to treat time as a regularization parameter, so that the sequence of estimators $(W_t)_{t\geq 0}$ given by the iterates of gradient descent refer to different models, and one can perform model selection with respect to time. This means that in this case getting different models is very cheap, as each iteration of gradient descent yields a new model.

14.2 Least Square Regression

We assume that the data pairs $(x_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ have $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, that the feature vectors x_i's are deterministic, and that there exists a parameter $w^\star \in \mathbb{R}^d$ (unknown to us) such that the observations Y_i's are generated according to a linear model perturbed by noise:

$$Y_i = \langle x_i, w^\star \rangle + \sigma \xi_i.$$

Here, $\xi_i \sim \mathcal{N}(0,1)$ is the (unobserved) noise, a standard Gaussian random variable, independent of everything else in the model, and $\sigma > 0$ is the standard deviation of the noise. In matrix form, the above reads

$$Y = \mathbf{x} w^\star + \sigma \xi,$$

where $Y \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^{n\times d}$ is the matrix whose i-th row corresponds to the vector x_i.

The question of interest is the following: given the data pairs encoded in $\mathbf{x}$ and Y, we want to design an estimator W (function of the observable data $\mathbf{x}$ and Y) that minimizes the estimation error with respect to the ℓ_2 norm, i.e. $\|W - w^\star\|_2$, in expectation and with high probability.

In general, the data matrix $\mathbf{x}$ can have a non-zero null space. This is the case whenever we are dealing with the high-dimensional setting $n < d$, as we discussed in Lecture 12. In general, this can happen even in the low-dimensional case $n \geq d$. When $\mathbf{x}$ has a non-zero null space, there are infinitely many vectors w such that $\mathbf{x} w^\star = \mathbf{x} w$. In fact, note that for any vector $w^{\texttt{Span}}$ in the span of the data, i.e. $w^{\texttt{Span}} \in \texttt{Span}(x_1, \dots, x_n) := \{x \in \mathbb{R}^d : x = \mathbf{x}^\top \omega = \sum_{i=1}^n \omega_i x_i \text{ for some } \omega \in \mathbb{R}^n\}$ and for any vector $w^{\texttt{Span}^\perp}$ in its orthogonal complement we have

$$\mathbf{x}(w^{\texttt{Span}} + w^{\texttt{Span}^\perp}) = \mathbf{x} w^{\texttt{Span}}.$$

This fact follows from the fundamental theorem of linear algebra, which states (among other things) that the null space of a matrix coincides with the orthogonal complement of its row space (and, in our case, the row space of the matrix $\mathbf{x}$ is the span of the data, as each data point is represented as a row vector in the matrix $\mathbf{x}$). To see this relationship, consider a generic matrix $\mathbf{m} \in \mathbb{R}^{n\times d}$ whose rows are made by the transpose of the d-dimensional column vectors $m_1, \dots, m_n$:

$$\mathbf{m} = \begin{bmatrix} m_1^\top \\ m_2^\top \\ \cdots \\ m_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$$

The product of the matrix $\mathbf{m}$ with any vector $w \in \mathbb{R}^d$ can be written in terms of the inner product of vectors:

$$\mathbf{m}w = \begin{bmatrix} m_1^\top w \\ m_2^\top w \\ \cdots \\ m_n^\top w \end{bmatrix}.$$

Hence, $\mathbf{m}w = 0$ if and only if w is orthogonal to each of the row vectors of $\mathbf{m}$. Equivalently, the null space of $\mathbf{m}$ coincides with the orthogonal complement of the row space of $\mathbf{m}$.

In the previous two lectures we saw how we can address the ill-posedness of the estimation problem by considering sparsity or low-rankedness assumptions. We showed how we can construct statistically sound and computationally feasible estimators by solving regularized forms of the empirical risk minimization problem. In this case, regularization was achieved *explicitly*, by either constraining the optimization problem on a particular subset $\mathcal{A}$ (i.e. $\mathcal{A} = \{w \in \mathbb{R}^d : \|w\|_0 \le k\}$ or $\mathcal{A} = \{w \in \mathbb{R}^d : \|w\|_1 \le k\}$ for convex relaxations, a case the latter that we only mentioned but not covered in detail) of by considering an unconstrained optimization problem with a penalty term (i.e. $R(w) + \mu\|w\|_1$).

Today we take a different approach to regularization, and investigate notions of implicit bias and implicit regularization via early stopping. We previously discussed early stopping in Lecture 11 in the context of the prediction error, via notion of stability. Today we discuss this approach in the context of the estimation error, using tools from linear algebra.

To make the estimation problem well-posed, henceforth we make the following assumption.

- **The unknown parameter lies in the span of the data:** there exists a vector $\omega = (\omega_1, \dots, \omega_n) \in \mathbb{R}^n$ such that the unknown parameter $w^\star \in \mathbb{R}^d$ can be written as:

$$\boxed{w^\star = \mathbf{x}^\top \omega = \sum_{i=1}^n \omega_i x_i}$$

14.3 Empirical Second Moment Matrix

A quantity that will play a fundamental role in the *analysis* we are going to present is the empirical (or sample) second moment matrix:

$$\boxed{\mathbf{c} := \frac{\mathbf{x}^\top \mathbf{x}}{n} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{d \times d}}$$

As the matrix $\mathbf{c}$ is symmetric positive semi-definite it admits the following orthonormal eigendecomposition:

$$\mathbf{c} = \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top,$$

where the columns of the matrix $\mathbf{u} = [u_1, \dots, u_d]$ constitute an orthonormal basis of eigenvectors of $\mathbf{c}$ (satisfying $\mathbf{u}^\top = \mathbf{u}^{-1}$ so that $\mathbf{u}^\top\mathbf{u} = \mathbf{u}\mathbf{u}^\top = I$) and the diagonal matrix $\boldsymbol{\mu}$ contains the corresponding real-valued eigenvalues of $\mathbf{c}$. We write

$$\boldsymbol{\mu} := \operatorname{diag}(\mu_1, \dots, \mu_r, \underbrace{0, \dots, 0}_{d-r}),$$

where $r \leq d$ is the rank of the matrix and the eigenvalues satisfy $0 < \mu_r \leq \cdots \leq \mu_1$. Given the possible presence of zero eigenvalues (if $r < d$), it is convenient to work with the following representation (check that this holds!):

$$\mathbf{c} = \mathbf{u}_{1:r} \boldsymbol{\mu}_{1:r} \mathbf{u}_{1:r}^\top,$$

where $\mathbf{u}_{j:k} := [u_j, \dots, u_k]$ and $\boldsymbol{\mu}_{j,k} := \operatorname{diag}(\mu_j, \dots, \mu_k)$ for any $j \leq k$. Note that by the orthonormality of the eigenvectors we have $\mathbf{u}_{j:k}^\top \mathbf{u}_{j:k} = I_{k-j+1}$, where I_i denotes the $i \times i$ identity matrix, but in general $\mathbf{u}_{j:k} \mathbf{u}_{j:k}^\top \neq I \equiv I_d$. Also note that

$$\boldsymbol{\pi} = \mathbf{u}_{1:r} \mathbf{u}_{1:r}^\top$$

is the orthogonal projection operator onto the range of $\mathbf{c}$ and

$$I - \boldsymbol{\pi} = \mathbf{u}_{r+1:d} \mathbf{u}_{r+1:d}^\top$$

is the orthogonal projection operator onto the null space of $\mathbf{c}$.

The pseudoinverse of the matrix $\mathbf{c}$ is the matrix $\mathbf{c}^+$ that in the present case is defined as:

$$\mathbf{c}^+ = \mathbf{u} \boldsymbol{\mu}^+ \mathbf{u}^\top,$$

where

$$\boldsymbol{\mu}^+ := \operatorname{diag}\left(\frac{1}{\mu_1}, \dots, \frac{1}{\mu_r}, \underbrace{0, \dots, 0}_{d-r} \right),$$

or, equivalently,

$$\mathbf{c}^+ = \mathbf{u}_{1:r} \boldsymbol{\mu}_{1:r}^{-1} \mathbf{u}_{1:r}^\top.$$

If $\mathbf{c}$ is invertible (which is equivalent to $r = d$), then the pseudoinverse coincides with the inverse: $\mathbf{c}^+ = \mathbf{c}^{-1}$.

14.4 Least Squares Regression: with and without Regularization

The gradient of the empirical risk is given by:

$$\nabla R(w) = \frac{2}{n} \mathbf{x}^\top (\mathbf{x} w - Y).$$

The first order optimality condition that characterizes the local minima $W^\star$'s (there may be infinitely many of them) is given by $\nabla R(W^\star) = 0$, namely,

$$\mathbf{c} W^\star = \frac{\mathbf{x}^\top Y}{n}.$$

If $\mathbf{c}$ is invertible the empirical risk minimization problem admits a unique solution given by

$$W^\star = \mathbf{c}^{-1} \frac{\mathbf{x}^\top Y}{n} = w^\star + \sigma \mathbf{c}^{-1} \frac{\mathbf{x}^\top \xi}{n},$$

where we used that $Y = \mathbf{x} w^\star + \sigma \xi$. If $\mathbf{c}$ is not invertible the empirical risk minimization problem admits infinitely many solutions (for the existence of the solution, c.f. Remark 14.1). In particular, the solution of smallest Euclidean norm is given by (this follows from a property of the pseudo-inverse that we will not prove):

$$\boxed{W^\star_{\text{l.s.}} = \mathbf{c}^+ \frac{\mathbf{x}^\top Y}{n} = \operatorname{argmin}\left\{\|w\|_2 : \mathbf{c}w = \frac{\mathbf{x}^\top Y}{n}\right\} = \operatorname{argmin}\left\{\|w\|_2 : w \in \operatorname*{argmin}_{w\in\mathbb{R}^d} R(w)\right\} = \boldsymbol{\pi} w^\star + \sigma \mathbf{c}^+ \frac{\mathbf{x}^\top \xi}{n}} \tag{14.1}$$

where we used that

$$\mathbf{c}^+ \frac{\mathbf{x}^\top \mathbf{x}}{n} = \mathbf{c}^+\mathbf{c} = \mathbf{u}\boldsymbol{\mu}^+\mathbf{u}^\top\mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top = \mathbf{u}\boldsymbol{\mu}^+\boldsymbol{\mu}\mathbf{u}^\top = \mathbf{u}\,\mathrm{diag}(1,\dots,1,\underbrace{0,\dots,0}_{d-r})\mathbf{u}^\top = \mathbf{u}_{1:r}\mathbf{u}_{1:r}^\top = \boldsymbol{\pi}. \tag{14.2}$$

Note that the assumption that $w^\star$ lies in the span of the data coincides with the statement that $\boldsymbol{\pi} w^\star = w^\star$. In what follows we derive results that hold for the general case $\boldsymbol{\pi} w^\star \neq w^\star$, making explicit the nature of the approximation term that relates the deviation of $w^\star$ from $\boldsymbol{\pi} w^\star$.

Remark 14.1 *In general, a matrix equation of the form* $\mathbf{m}w = b$ *may* not *admit a solution. Existence of a solution is equivalent to the fact that the vector b lies in the image (i.e. column span, or range) of the matrix* $\mathbf{m}$*. In our case, we are not interested in a solution of the equation* $\mathbf{c}w = b$ *for a generic matrix* $\mathbf{c}$ *and vector b, as we are given that* $\mathbf{c} = \frac{\mathbf{x}^\top\mathbf{x}}{n}$ *and* $b = \frac{\mathbf{x}^\top Y}{n}$*. Here, existence of a solution is guaranteed by the fact that the image of the matrix* $\mathbf{x}^\top\mathbf{x}$ *coincides with the image of the matrix* $\mathbf{x}^\top$*. To prove this, it is enough to show that for any vector v we can find a vector $\tilde v$ such that* $\mathbf{x}^\top v = \mathbf{x}^\top\mathbf{x}\tilde v$ *(the other direction is trivial). By the same argument used at the beginning of this lecture, we have that* $\mathbf{x}^\top v = \mathbf{x}^\top v^{\mathsf{Span}}$*, where* v^{Span} *is the projection of v on the span of the rows of the matrix* $\mathbf{x}^\top$*, which, by definition, can be written as* $v^{\mathsf{Span}} = \mathbf{x}\tilde v$*.*

Due to the presence of the noise term ξ, the solution of the *unregularized* empirical risk minimization problem does not coincide with the parameter that we want to infer, i.e. $w^\star$. This is the reason why we need to impose some form of regularization. The estimator considered by ridge regression is given by the minimization of the function

$$R(w) + \lambda\|w\|_2^2 = \frac{1}{n}\|\mathbf{x}w - Y\|_2^2 + \lambda\|w\|_2^2,$$

where $\lambda > 0$ controls the strength of the regularization. In this case the function to be minimized is strongly convex, and the equation $\nabla R(w) = 0$, which reads

$$(\mathbf{c} + \lambda I)w = \frac{\mathbf{x}^\top Y}{n}$$

admits the unique solution

$$W^\star_{ridge} = (\mathbf{c} + \lambda I)^{-1}\frac{\mathbf{x}^\top Y}{n}.$$

Classical statistical theory tells us how to tune λ to get optimal statistical rates. In what follows, we instead consider the algorithmic/implicit regularization approach.

14.5 Gradient Descent for Least Squares Regression

Storing the second moment matrix $\mathbf{c}$ into memory costs $O(nd^2)$ space, as there are d^2 entries and computing each entry involves taking the inner product of two n-dimensional vectors. In general, inverting the second moment matrix *exactly* costs $O(d^3)$ time, while computing an *approximate* inverse (up to precision ε) can be achieved (as the matrix is positive definite) by quasi-linear solvers with $\widetilde{O}(d^2 \log \frac{1}{\varepsilon})$ time, where the $\widetilde{O}(\,\cdot\,)$

notation hides poly-logarithmic terms. Note that the fast solvers are (quasi) linear in the degrees of freedom in the second moment matrix, and there are d^2 degrees of freedom [3].

We will show that when $\mu_r \geq c$ for a universal constant c, then gradient descent allows to solve the problem up to the optimal statistical rate (by which we mean achieving the fast rate $1/n$ for the square of the ℓ_2 loss) with the same computational complexity (time) required to read the data into memory (ignoring logarithmic terms), namely, $\widetilde{O}(nd)$. Recall that there are n data points, each involving a d-dimensional feature vector and a one-dimensional label.

The gradient descent update with step size $\eta/2$ (the factor $1/2$ is added for mathematical convenience so that the formulas below only depend on η and not 2η) applied to minimize the empirical risk R reads as follows:

$$\boxed{W_{t+1} = W_t - \frac{\eta}{2}\nabla R(W_t) = (I - \eta\mathbf{c})\, W_t + \eta\frac{\mathbf{x}^\top Y}{n}} \tag{14.3}$$

A single iteration of gradient descent requires $O(nd)$ space and $O(nd)$ time. In particular, the sample second moment matrix $\mathbf{c}$ does *not* need to be computed to run the algorithm, as the vector $\mathbf{c}W_t = \frac{\mathbf{x}^\top\mathbf{x}W_t}{n}$ can be computed by first computing $\widetilde{W_t} = \mathbf{x}W_t$ (which costs nd operations) and then computing $\mathbf{x}^\top\widetilde{W_t}$ (which costs another nd operations). The matrix $\mathbf{c}$ is only a tool that we will use for the theoretical analysis that we now develop.

We will see that if the non-zero eigenvalues of the empirical second moment matrix $\mathbf{c}$ are bounded by universal constants and the signal-to-noise ratio $\frac{\|w^\star\|_2}{\sigma}$ is upper bounded by a universal constant, we only need to run gradient descent for a number of iterations that scales like $\log n$, hence implying the computational optimality of this method modulo logarithmic terms (and universal factors): the problem can be solved with the same time required to read the data.

By unraveling the iteration of gradient descent, assuming henceforth that $W_0 = 0$, we find

$$W_t = \left(\sum_{k=0}^{t-1}(I-\eta\mathbf{c})^k\right)\eta\frac{\mathbf{x}^\top Y}{n} = \mathrm{Inv}_t(\eta\mathbf{c})\eta\frac{\mathbf{x}^\top Y}{n}$$

where we have defined the quantity

$$\mathrm{Inv}_t(\eta\mathbf{c}) := \sum_{k=0}^{t-1}(I-\eta\mathbf{c})^k\,.$$

If all the non-zero eigenvalues of the matrix $\eta\mathbf{c}$ are strictly less than one, which is the case if the positive learning rate η is small enough, then, *on the image space of* $\mathbf{c}$, the matrix $\mathrm{Inv}_t(\eta\mathbf{c})$ approximates the pseudo-inverse of the matrix $\eta\mathbf{c}$ as $t \to \infty$ (recall that the pseudo-inverse coincides with the inverse if $\eta\mathbf{c}$ has no zero eigenvalues). This is the content of Proposition 14.2 below. Before stating this proposition, we use the data generating process to decompose the gradient descent iterate W_t into the sum of two terms: the mean $\mathbf{E}W_t$ and the deviation from the mean $W_t - \mathbf{E}W_t$.

Using that $Y = \mathbf{x}w^\star + \sigma\xi$ we obtain

$$\boxed{W_t = \underbrace{\mathrm{Inv}_t(\eta\mathbf{c})\eta\mathbf{c}w^\star}_{\mathbf{E}W_t} + \underbrace{\sigma\,\mathrm{Inv}_t(\eta\mathbf{c})\eta\frac{\mathbf{x}^\top\xi}{n}}_{W_t-\mathbf{E}W_t}} \tag{14.4}$$

where the identity $\mathbf{E}W_t = \mathrm{Inv}_t(\eta\mathbf{c})\eta\mathbf{c}w^\star$ holds as the noise has zero mean by assumption. This expression forms the basis for the two bias/variance decompositions that we will state below.

In the following, we define the *shrinkage matrix* as

$$\mathbf{s} := I - \eta\boldsymbol{\mu} = \operatorname{diag}(1-\eta\mu_1,\dots,1-\eta\mu_r,\underbrace{1,\dots,1}_{d-r})$$

Henceforth, let $\|\cdot\|$ denote the operator norm, defined for a generic matrix $\mathbf{m}$ (not necessarily a square matrix) as

$$\|\mathbf{m}\| := \sqrt{\mu_1(\mathbf{m}^\top\mathbf{m})},$$

where $\mu_d(\mathbf{m}^\top\mathbf{m}) \le \cdots \le \mu_1(\mathbf{m}^\top\mathbf{m})$ are the real eigenvalues of the symmetric matrix $\mathbf{m}^\top\mathbf{m}$. If a matrix $\mathbf{m} \in \mathbb{R}^{d\times d}$ is symmetric, then the operator norm coincides with the largest eigenvalue in magnitude:

$$\|\mathbf{m}\| = \max\{\mu_1(\mathbf{m}),\mu_d(\mathbf{m})\}.$$

For any matrix $\widetilde{\mathbf{m}}$ and vector v, the following properties hold (note that the first property implies the second one):

$$\begin{aligned}\|\mathbf{m}\widetilde{\mathbf{m}}\| &\le \|\mathbf{m}\|\|\widetilde{\mathbf{m}}\|,\\ \|\mathbf{m}v\|_2 &\le \|\mathbf{m}\|\|v\|_2.\end{aligned}$$

14.6 Implicit Bias of Gradient Descent

The following proposition expresses the gradient descent iterate at time t as a function of $\mathbf{s}^t$, the t-th power of the shrinkage matrix, and the pseudoinverse of the empirical second moment matrix $\mathbf{c}$. The fact that the gradient descent iterates are a function of the pseudoinverse $\mathbf{c}^+$ suggests that gradient descent is inherently connected to the least ℓ_2-norm solution of the empirical risk minimization problem, as we will see later on at convergence.

Proposition 14.2 *We have*

$$\begin{aligned}\mathrm{Inv}_t(\eta\mathbf{c}) &= (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)(\eta\mathbf{c})^+ + t(I-\boldsymbol{\pi}) = \sum_{i=1}^r \frac{1-(1-\eta\mu_i)^t}{\eta\mu_i}u_iu_i^\top + t(I-\boldsymbol{\pi}),\\ \mathrm{Inv}_t(\eta\mathbf{c})\eta\mathbf{c} &= (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top) = \sum_{i=1}^r (1-(1-\eta\mu_i)^t)u_iu_i^\top,\end{aligned}$$

and, from the identity (14.4) *we obtain*

$$\boxed{W_t = \underbrace{(I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)w^\star}_{\mathbf{E}W_t} + \underbrace{\sigma(I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)\mathbf{c}^+\frac{\mathbf{x}^\top\xi}{n}}_{W_t-\mathbf{E}W_t} = \underbrace{\sum_{i=1}^r(1-(1-\eta\mu_i)^t)u_iu_i^\top w^\star}_{\mathbf{E}W_t} + \underbrace{\sigma\sum_{i=1}^r\frac{1-(1-\eta\mu_i)^t}{\mu_i}u_iu_i^\top\frac{\mathbf{x}^\top\xi}{n}}_{W_t-\mathbf{E}W_t}}$$

Proof: Using that $\mathbf{u}\mathbf{u}^\top = \mathbf{u}^\top\mathbf{u} = I$ we have $\mathrm{Inv}_t(\eta\mathbf{c}) = \sum_{k=0}^{t-1}(\mathbf{u}(I-\eta\boldsymbol{\mu})\mathbf{u}^\top)^k = \mathbf{u}\sum_{k=0}^{t-1}(I-\eta\boldsymbol{\mu})^k\mathbf{u}^\top$. Using that $\sum_{k=0}^{t-1} x^k = \frac{1-x^t}{1-x}$ for any $x \in \mathbb{R} \setminus \{1\}$ and $\sum_{k=0}^{t-1} 1 = t$, we obtain

$$\begin{aligned}
\operatorname{Inv}_t(\eta \mathbf{c}) &= \mathbf{u} \operatorname{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1}, \dots, \frac{1-(1-\eta\mu_r)^t}{\eta\mu_r}, t, \dots, t\right)\mathbf{u}^\top \\
&= \mathbf{u} \operatorname{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1}, \dots, \frac{1-(1-\eta\mu_r)^t}{\eta\mu_r}, 0, \dots, 0\right)\mathbf{u}^\top + \mathbf{u}\operatorname{diag}(0,\dots,0,t,\dots,t)\mathbf{u}^\top \\
&= \mathbf{u}_{1:r} \operatorname{diag}\left(\frac{1-(1-\eta\mu_1)^t}{\eta\mu_1}, \dots, \frac{1-(1-\eta\mu_r)^t}{\eta\mu_r}\right)\mathbf{u}_{1:r}^\top + t\mathbf{u}_{r+1:d}\mathbf{u}_{r+1:d}^\top \\
&= \mathbf{u}_{1:r} \operatorname{diag}\left(1-(1-\eta\mu_1)^t, \dots, 1-(1-\eta\mu_r)^t\right)\mathbf{u}_{1:r}^\top \mathbf{u}_{1:r} \operatorname{diag}\left(\frac{1}{\eta\mu_1}, \dots, \frac{1}{\eta\mu_r}\right)\mathbf{u}_{1:r}^\top + t(I-\boldsymbol{\pi}) \\
&= \mathbf{u}(I-(I-\eta\boldsymbol{\mu})^t)\mathbf{u}^\top(\eta\mathbf{c})^+ + t(I-\boldsymbol{\pi}) \\
&= (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)(\eta\mathbf{c})^+ + t(I-\boldsymbol{\pi}).
\end{aligned}$$

By the properties of the pseudoinverse, it follows that

$$(I-\boldsymbol{\pi})\mathbf{x}^\top = 0.$$

If fact, for a generic matrix $\mathbf{m}$ it can be shown that the following properties hold:

$$\begin{aligned}
(\mathbf{m}^\top\mathbf{m})^+\mathbf{m}^\top &= \mathbf{m}^+, \\
\mathbf{m}^+\mathbf{m}\mathbf{m}^\top &= \mathbf{m}^\top.
\end{aligned}$$

As $\boldsymbol{\pi} = \boldsymbol{c}^+\boldsymbol{c}$ by (14.2) and $\boldsymbol{c} = \mathbf{x}^\top\mathbf{x}/n$ by definition, the two properties above yield

$$(I-\boldsymbol{\pi})\mathbf{x}^\top = (I-(\mathbf{x}^\top\mathbf{x})^+\mathbf{x}^\top\mathbf{x})\mathbf{x}^\top = (I-\mathbf{x}^+\mathbf{x})\mathbf{x}^\top = \mathbf{x}^\top - \mathbf{x}^+\mathbf{x}\mathbf{x}^\top = \mathbf{x}^\top - \mathbf{x}^\top = 0.$$

So, using that $\mathbf{c} = \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top$ we find

$$\operatorname{Inv}_t(\eta\mathbf{c})\eta\mathbf{c} = \mathbf{u}\operatorname{diag}\left(1-(1-\eta\mu_1)^t, \dots, 1-(1-\eta\mu_r)^t, 0, \dots, 0\right)\mathbf{u}^\top = (I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top),$$

and, by the identity (14.4),

$$W_t - \mathbf{E}W_t = \sigma \operatorname{Inv}_t(\eta\mathbf{c})\eta\frac{\mathbf{x}^\top\xi}{n} = \sigma(I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top)\mathbf{c}^+\frac{\mathbf{x}^\top\xi}{n}.$$

■

The following proposition shows that if the learning rate is small enough, gradient descent converges to the least ℓ_2-norm solution of the empirical minimization problem as given by (14.1). Namely, upon the infinitely many minimizer of the empirical risk available when $\mathbf{c}$ is not invertible (i.e., $r < d$), gradient descent chooses a particular solution, the one that minimizes the ℓ_2 norm. This is an instance of implicit/algorithmic bias.

Proposition 14.3 *If $\eta \le \frac{1}{\mu_1}$, then we have*

$$\lim_{t\to\infty}(I-\mathbf{u}\mathbf{s}^t\mathbf{u}^\top) = \mathbf{u}_{1:r}\mathbf{u}_{1:r}^\top = \boldsymbol{\pi},$$

and, from Proposition 14.2 we obtain

$$\boxed{\lim_{t\to\infty} W_t = \underbrace{\boldsymbol{\pi}w^\star}_{\lim_{t\to\infty}\mathbf{E}W_t} + \underbrace{\sigma\mathbf{c}^+\frac{\mathbf{x}^\top\xi}{n}}_{\lim_{t\to\infty}(W_t-\mathbf{E}W_t)} = W^\star_{l.s.}}$$

with

$$\boxed{\|W_t - W^\star_{l.s.}\|_2 \le (1-\eta\mu_r)^t \|w^\star\|_2 + \frac{\sigma}{\sqrt{n}} \frac{(1-\eta\mu_1)^t}{\mu_r} \left\| \frac{\mathbf{x}^\top \xi}{\sqrt{n}} \right\|_2}$$

Proof: If $\eta \le \frac{1}{\mu_1}$ we have

$$\lim_{t\to\infty} \mathbf{s}^t = \operatorname{diag}(0,\dots,0,\underbrace{1,\dots,1}_{d-r})$$

so that

$$\lim_{t\to\infty} (I - \mathbf{u}\mathbf{s}^t\mathbf{u}^\top) = \mathbf{u}\operatorname{diag}(1,\dots,1,\underbrace{0,\dots,0}_{d-r})\mathbf{u}^\top = \mathbf{u}_{1:r}\mathbf{u}_{1:r}^\top = \boldsymbol{\pi}$$

and, by Proposition 14.2,

$$\lim_{t\to\infty} W_t = \underbrace{\boldsymbol{\pi} w^\star}_{\lim_{t\to\infty} \mathbf{E} W_t} + \underbrace{\sigma \boldsymbol{\pi} \mathbf{c}^+ \frac{\mathbf{x}^\top \xi}{n}}_{\lim_{t\to\infty}(W_t - \mathbf{E}W_t)} .$$

The proof follows since $\boldsymbol{\pi}\mathbf{c}^+ = \mathbf{c}^+$, as the null space of the pseudoinverse of a matrix is equal the null space of the matrix transpose, and as $\mathbf{c}$ is symmetric, then the null space of $\mathbf{c}$ coincides with the null space of $\mathbf{c}^+$. Recall (14.1). ■

Proposition 14.3 also gives a rate of convergence towards $W^\star_{\text{l.s.}}$, establishing that the convergence of gradient descent with constant step size $\frac{1}{2\mu_1}$ (recall that by our current parametrization, the step size is given by the choice $\eta/2$) is exponential (a.k.a. *linear* in the optimization literature). This exponential rate of convergence is not a direct consequence of the general result we have previously seen for gradient descent on strongly convex and smooth function (Theorem 9.5). In fact, in general, the empirical risk is not strongly convex as its Hessian $\nabla R(w) = 2\mathbf{c}$ has zero eigenvalues whenever $r < d$. Also, note that $2\mu_1$ coincides with the smoothness parameter of the function R, so the learning rate is indeed tuned as the inverse of the smoothness parameter.

Remark 14.4 (Connection implicit bias and statistical bounds) *One might wonder what are the implication of gradient descent converging to the minimum ℓ_2 norm solution from the point of view of excess risk bounds. In the noiseless case ($\xi = 0$, so $Y = \mathbf{x}w^\star$), one such implication is easy to derive. Note that in this case the true parameter $w^\star$ is also a minimizer of the empirical risk R, as is $W^\star_{l.s.}$. By definition of $W^\star_{l.s.}$ we have $\|W^\star_{l.s.}\|_2 \le \|w^\star\|_2$, so we can derive excess risk bounds for $W^\star_{l.s.}$ using uniform convergence theory (i.e. Rademacher bounds) with respect to the class fo functions $\mathcal{A}_2 = \{x \in \mathbb{R}^d \to w^\top x : \|w\|_2 \le \|w^\star\|_2\}$. The excess risk bounds that we can establish in this way will depend explicitly (linearly) on $\|w^\star\|_2$ (this is not atypical in learning theory and optimization: recall, for instance, that the convergence rate of gradient descent in convex problems depends on the distance between the initialization and the solution of the optimisation problem.)*

14.7 Where does the Implicit Bias come from?

Why does gradient descent converge to the minimum ℓ_2-norm solution?

The connection between gradient descent and the ℓ_2-norm is not at all surprising, as we saw in the previous lectures. Recall that when applied to minimize a smooth function f, the gradient descent update with constant step size $\eta = 1/\beta$

$$x_{s+1} = x_s - \frac{1}{\beta}\nabla f(x_s)$$

corresponds to moving to the point that maximizes the guaranteed decrease given by the quadratic function with curvature β supported at x_s that uniformly upper-bounds f by smoothness, namely:

$$x_{s+1} = \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{f(x_s) + \nabla f(x_s)^\top (y - x_s) + \frac{\beta}{2}\|y - x_s\|_2^2\right\}.$$

The above interpretation of the gradient descent update holds more generally, even when the function f is not smooth. By the same argument as above, the gradient descent update with step size η_s:

$$x_{s+1} = x_s - \eta_s \nabla f(x_s)$$

is *defined* as the algorithm that at each time step minimizes the quadratic function (not necessarily a uniform upper bound!) with curvature $1/\eta_s$ supported at the current iterate:

$$\boxed{x_{s+1} = \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{f(x_s) + \nabla f(x_s)^\top (y - x_s) + \frac{1}{2\eta_s}\|y - x_s\|_2^2\right\}}$$

Hence, the ℓ_2 norm enters explicitly in the definition of gradient descent, which is what ultimately characterizes its implicit bias towards the smaller ℓ_2-norm solution. It is possible to connect implicit bias to the geometric properties of algorithms (not only gradient descent) more in general [1].

14.8 Implicit Regularization

The next proposition bounds the deviation of gradient descent from the true unknown parameter $w^\star$. It presents an error decomposition in terms of bias, concentration and approximation errors, where the approximation error corresponds to the implicit bias of the algorithm. This result shows that for a choice of learning rate η small enough, we can establish an upper bound for the bias error that decreases exponentially fast with time towards zero, and an upper bound for the variance term that increases exponentially fast towards the noise term $\frac{\sigma}{\sqrt{n}}\frac{1}{\mu_r}\frac{\|\mathbf{x}^\top\xi\|_2}{\sqrt{n}}$. In particular, using concentration results we can show that the quantity $\frac{1}{\mu_r}\frac{\|\mathbf{x}^\top\xi\|_2}{\sqrt{n}}$ is upper bounded with high probability by a function of the eigenvalues $\{\mu_1, \dots, \mu_r\}$, and we can use this to tune the stopping time $t^\star$ so that the bias term is of the same order as the correlation term.

Theorem 14.5 *We have*

$$\|W_t - w^\star\|_2 \le \underbrace{\|\mathbf{E}W_t - \boldsymbol{\pi}w^\star\|_2}_{\textit{bias error}} + \underbrace{\|W_t - \mathbf{E}W_t\|_2}_{\textit{concentration error}} + \underbrace{\|w^\star - \boldsymbol{\pi}w^\star\|_2}_{\textit{approximation error}}.$$

If $\eta \le \frac{1}{\mu_1}$, *then*

$$\boxed{\|W_t - w^\star\|_2 \le (1-\eta\mu_r)^t\|w^\star\|_2 + \frac{\sigma}{\sqrt{n}}\frac{1-(1-\eta\mu_1)^t}{\mu_r}\frac{\|\mathbf{x}^\top\xi\|_2}{\sqrt{n}} + \|w^\star - \boldsymbol{\pi}w^\star\|_2}$$

Furthermore, for any $c \in (0,1)$, *let* $t^\star$ *satisfy* $t^\star \ge \frac{1}{\log(1/(1-\eta\mu_r))}\log\left(\frac{\|w^\star\|_2}{\sigma}\frac{\sqrt{n}}{\tilde{c}}\right)$. *Then,*

$$\boxed{\mathbf{P}\left(\|W_{t^\star} - w^\star\|_2 \le 2\sigma\frac{\tilde{c}}{\sqrt{n}} + \|w^\star - \boldsymbol{\pi}w^\star\|_2\right) \ge 1-\delta}$$

with $\tilde{c} = \frac{1}{\mu_r}\sqrt{\sum_{i=1}^r \mu_i + c\sum_{i=1}^r \frac{\mu_i^2}{\mu_1}}$ *and* $\delta = e^{-\frac{c^2}{8}\sum_{i=1}^r (\mu_i/\mu_1)^2}$.

Proof: The error decomposition immediately follows by the triangle inequality. From Proposition 14.2, using that $\boldsymbol{\pi} = \sum_{i=1}^r u_i u_i^\top$, we have the following bound for the bias term

$$\begin{aligned}\|\mathbf{E}W_t - \boldsymbol{\pi} w^\star\|_2 &= \left\| \sum_{i=1}^r (1-(1-\eta\mu_i)^t) u_i u_i^\top w^\star - \sum_{i=1}^r u_i u_i^\top w^\star \right\|_2 \\ &= \left\| - \sum_{i=1}^r (1-\eta\mu_i)^t u_i u_i^\top w^\star \right\|_2 \\ &\leq \left\| - \sum_{i=1}^r (1-\eta\mu_i)^t u_i u_i^\top \right\| \|w^\star\|_2 \\ &\leq (1-\eta\mu_r)^t \|w^\star\|_2,\end{aligned}$$

and the following bound for the concentration term

$$\begin{aligned}\|W_t - \mathbf{E}W_t\|_2 = \left\| \sigma \sum_{i=1}^r \frac{1-(1-\eta\mu_i)^t}{\mu_i} u_i u_i^\top \frac{\mathbf{x}^\top \xi}{n} \right\|_2 &\leq \sigma \left\| \sum_{i=1}^r \frac{1-(1-\eta\mu_i)^t}{\mu_i} u_i u_i^\top \right\| \frac{\|\mathbf{x}^\top \xi\|_2}{n} \\ &\leq \frac{\sigma}{\sqrt{n}} \frac{1-(1-\eta\mu_1)^t}{\mu_r} \frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}.\end{aligned}$$

The random vector $V := \frac{\mathbf{x}^\top \xi}{\sqrt{n}}$ is Gaussian with mean 0 and second moment matrix $\mathbf{c}$. We will now show that $\|V\|_2^2 = (\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}})^2$ has the same distribution as $\sum_{i=1}^r \mu_i Z_i^2$, where $Z_1, \ldots, Z_r$ are i.i.d. standard Gaussian random variables.

Let $\mathbf{c}^{1/2} = \mathbf{u}\boldsymbol{\mu}^{1/2}\mathbf{u}^\top$ be the square root of the matrix $\mathbf{c}$, with $\boldsymbol{\mu}^{1/2} = \operatorname{diag}(\sqrt{\mu_1}, \ldots, \sqrt{\mu_r}, 0, \ldots, 0)$. Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a Gaussian random vector with mean 0 and covariance I. Then, the random vector V has the same distribution as the random vector $T = \mathbf{c}^{1/2}\mathbf{u}Z$. In fact, T is Gaussian being a linear combination of a Gaussian vector and its variance is given by (using that $\mathbf{c}^{1/2}$ is symmetric)

$$\mathbf{E}TT^\top = \mathbf{E}[\mathbf{c}^{1/2}\mathbf{u}ZZ^\top\mathbf{u}^\top\mathbf{c}^{1/2}] = \mathbf{c}^{1/2}\mathbf{u}\mathbf{E}[ZZ^\top]\mathbf{u}^\top\mathbf{c}^{1/2} = \mathbf{c}^{1/2}\mathbf{u}\mathbf{u}^\top\mathbf{c}^{1/2} = \mathbf{c}.$$

Then, as $\mathbf{c} = \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top$, we find

$$\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2 = \|V\|_2^2 = V^\top V \sim T^\top T = Z^\top \mathbf{u}^\top \mathbf{c}\mathbf{u} Z = Z^\top \mathbf{u}^\top \mathbf{u}\boldsymbol{\mu}\mathbf{u}^\top \mathbf{u} Z = Z^\top \boldsymbol{\mu} Z = \sum_{i=1}^r \mu_i Z_i^2,$$

which is a weighted sum of independent chi-squared random variables with 1 degree of freedom. In particular,

$$\mathbf{E}\left[\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2\right] = \mathbf{E}[\|V\|_2^2] = \sum_{i=1}^r \mu_i \mathbf{E}[Z_i^2] = \sum_{i=1}^r \mu_i.$$

From **Problem 3.3** in the Problem Sheets, recall that each Z_i^2 is sub-exponential with parameters $\nu^2 = 4$ and $c = 4$, namely:

$$\mathbf{E}e^{t(Z_i^2-1)} \leq e^{\nu^2 t^2/2} \qquad \text{for any } t \in (-1/c, 1/c).$$

By Chernoff's bound we have, for any $\varepsilon, t > 0$,

$$\mathbf{P}(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2] \geq \varepsilon) \leq e^{-t\varepsilon}\mathbf{E}e^{t(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2])} = e^{-t\varepsilon}\mathbf{E}e^{t\sum_{i=1}^r \mu_i(Z_i^2-1)} = e^{-t\varepsilon}\prod_{i=1}^r \mathbf{E}e^{t\mu_i(Z_i^2-1)}.$$

If $t\mu_1 < 1/4$, then the previous result yields

$$\mathbf{P}(\|V\|_2^2 - \mathbf{E}[\|V\|_2^2] \geq \varepsilon) \leq e^{-t\varepsilon}\prod_{i=1}^r e^{2t^2\mu_i^2} = e^{-t\varepsilon + 2t^2\sum_{i=1}^r \mu_i^2}.$$

The smallest upper bound is obtained by choosing $t = \frac{\varepsilon}{4\sum_{i=1}^r \mu_i^2}$ (which satisfies the requirement $t\mu_1 < 1/4$ if and only if $\varepsilon < \sum_{i=1}^r \mu_i^2/\mu_1$) and yields the upper bound

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}} \geq \sqrt{\sum_{i=1}^r \mu_i + \varepsilon}\right) = \mathbf{P}\left(\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}}\right)^2 - \sum_{i=1}^r \mu_i \geq \varepsilon\right) \leq e^{-\varepsilon^2/(8\sum_{i=1}^r \mu_i^2)}.$$

Choosing $\varepsilon = c\sum_{i=1}^r \mu_i^2/\mu_1$, where c is any positive constant strictly less than 1, we find

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_2}{\sqrt{n}} < \sqrt{\sum_{i=1}^r \mu_i + c\sum_{i=1}^r \frac{\mu_i^2}{\mu_1}}\right) \geq 1 - e^{-\frac{c^2}{8}\sum_{i=1}^r (\mu_i/\mu_1)^2}.$$

Hence, so far we proved that for any $c \in (0,1)$ we have

$$\mathbf{P}\Big(\|W_t - w^\star\|_2 \leq (1-\eta\mu_r)^t\|w^\star\|_2 + \frac{\sigma}{\sqrt{n}}\tilde{c} + \|w^\star - \boldsymbol{\pi} w^\star\|_2\Big) \geq 1 - \delta,$$

with $\tilde{c} = \frac{1}{\mu_r}\sqrt{\sum_{i=1}^r \mu_i + c\sum_{i=1}^r \frac{\mu_i^2}{\mu_1}}$ and $\delta = e^{-\frac{c^2}{8}\sum_{i=1}^r (\mu_i/\mu_1)^2}$. Choosing $t^\star$ such that $(1-\eta\mu_r)^{t^\star}\|w^\star\|_2 = \frac{\sigma}{\sqrt{n}}\tilde{c}$ yields the final result.

■

Consider the following two conditions:

1. The eigenvalues $\{\mu_1, \ldots, \mu_r\}$ are upper and lower bounded by universal constants, i.e.

$$a \leq \mu_r \leq \cdots \leq \mu_1 \leq b$$

for some $a, b > 0$ independent of the parameters in the model (i.e. independent of n, d, etc.).

2. The signal-to-noise ratio $\frac{\|w^\star\|_2}{\sigma}$ is upper bounded by a universal constant, i.e.

$$\frac{\|w^\star\|_2}{\sigma} \leq \tilde{b}.$$

for some $\tilde{b} > 0$ independent of the parameters in the model.

If these two conditions hold, then Theorem 14.5 shows that it is possible to run gradient descent for a number of iterations that scales like $\log n$, hence implying the computational optimality of this method modulo logarithmic terms (and universal factors): the problem can be solved up to the fast rate $O(1/n)$ (referring to the *square* of the ℓ_2 loss, i.e. $\|W_t - w^\star\|_2^2 \lesssim 1/n$) with the same memory and running time required to store and read the data, respectively.

It is possible to show that for many random ensembles condition 1. holds with high probability in the low-dimensional case $n > d$. For the high-dimensional case $n < d$, condition 1. does no longer apply and typically the rank of the matrix $\mathbf{c}$ equals the number of data points, i.e. $r = n$, and the smallest non-zero eigenvalue $\mu_r = \mu_n$ is a decreasing function of n (hence, depends on n). In this case, one can derive different bounds for the bias and concentration terms in Proposition 14.2, bounds that show a polynomial convergence rate instead than an exponential convergence rate. That is, instead of proceeding as in Theorem 14.5 by choosing the early stopping threshold as the time that matches the order of the two terms in the following type of upper bound:

$$\alpha e^{-\beta t} + \widetilde{\alpha},$$

one is left with choosing the early stopping time as the minimizer of the following type of upper bound:

$$\alpha \frac{1}{t^{\beta}} + \widetilde{\alpha} t^{\widetilde{\beta}}.$$

This type of argument yields a precise definition of the optimal stopping time $t^{\star}$, not just a lower bound as in Theorem 14.5.

When the unknown parameter $w^{\star}$ lies in the span of the data, a similar analysis to the one here presented can be performed with respect to the *empirical (or sample) kernel matrix* defined as

$$\boxed{\mathbf{k} := \frac{\mathbf{x}\mathbf{x}^{\top}}{n} = (x_i^{\top} x_j)_{i,j\in\{1,\dots,n\}} \in \mathbb{R}^{n\times n}}$$

This analysis is particularly convenient in the high-dimensional regime $n < d$, as in this case the $n \times n$ kernel matrix $\mathbf{k}$ is lower dimensional compared to the $d \times d$ second moment matrix $\mathbf{c}$. Of course, this analysis is *required* for kernel methods (particularly in the case $n < \infty, d \to \infty$), where the matrix $\mathbf{k}$ is not simply a theoretical tool for the analysis of gradient descent, but it is the very data object one has access to (i.e., instead of having access to $\mathbf{x} \in \mathbb{R}^{n\times d}$, one has access to $\mathbf{k} \in \mathbb{R}^{n\times n}$). We refer to the paper [2] for an analysis on these lines and connection to notions of localized Rademacher complexity.

14.9 Alternative decompositions for the square of the ℓ_2 norm

Other decompositions analogous to the one given in Theorem 14.5 can be derived for the square norm, as the next classical result attests (for simplicity, we only focus on the case where the approximation error is zero).

Proposition 14.6 (Bias-Variance Decompositions) *Assume $\boldsymbol{\pi} w^{\star} = 0$. Then,*

$$\mathbf{E}\|W_t - w^{\star}\|_2^2 \le \underbrace{\|\mathbf{E}W_t - w^{\star}\|_2^2}_{(bias\ term)^2} + \underbrace{\mathbf{E}\|W_t - \mathbf{E}W_t\|_2^2}_{variance\ error} = \underbrace{\|\mathbf{E}W_t - w^{\star}\|_2^2}_{(bias\ term)^2} + \underbrace{\sum_{i=1}^{d} \mathbf{Var} W_{t,i}}_{variance\ error},$$

$$\|W_t - w^{\star}\|_2^2 \le \underbrace{2\|\mathbf{E}W_t - w^{\star}\|_2^2}_{(bias\ error)^2} + \underbrace{2\|W_t - \mathbf{E}W_t\|_2^2}_{concentration\ error}.$$

Proof: The first inequality follows as a direct consequence of the classical bias/variance decomposition for the square loss (which holds for any estimator, not just W_t). This decomposition is proved by adding and subtracting $\mathbf{E}W_t$ and expanding the square norm so that

$$\|W_t - w^{\star}\|_2^2 = \|W_t - \mathbf{E}W_t + \mathbf{E}W_t - w^{\star}\|_2^2 = \|W_t - \mathbf{E}W_t\|_2^2 + \|\mathbf{E}W_t - w^{\star}\|_2^2 + (W_t - \mathbf{E}W_t)^{\top}(\mathbf{E}W_t - w^{\star})$$

which immediately yields

$$\mathbf{E}\|W_t - w^{\star}\|_2^2 = \underbrace{\|\mathbf{E}W_t - w^{\star}\|_2^2}_{\text{(Bias term)}^2} + \underbrace{\mathbf{E}\|W_t - \mathbf{E}W_t\|_2^2}_{\text{Variance term}}.$$

The second inequality follows from the basic inequality $\|a + b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ holding for any vector a and b, so that

$$\|W_t - w^{\star}\|_2^2 = \|W_t - \mathbf{E}W_t + \mathbf{E}W_t - w^{\star}\|_2^2 \le \underbrace{2\|\mathbf{E}W_t - w^{\star}\|_2^2}_{\text{(Bias term)}^2} + \underbrace{2\|W_t - \mathbf{E}W_t\|_2^2}_{\text{Variance term}}.$$

Plugging in (14.4) in the above two expressions yields the results. ■

References

[1] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 2018.

[2] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, January 2014.

[3] Nisheeth K Vishnoi et al. Lx= b. *Foundations and Trends® in Theoretical Computer Science*, 8(1–2):1–141, 2013.