Algorithmic Foundations of Learning **Lecture 12**

High-Dimensional Statistics. Gaussian Complexity

Lecturer: Patrick Rebeschini *Version: November 21st 2019*

12.1 Introduction

So far in this course we have investigated the *prediction* problem in supervised learning, where the observed examples correspond to pairs of points $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, and the goal is to learn a predictor $a : \mathcal{X} \to \mathcal{Y}$ (as a function of the training data $\{(X_i, Y_i)\}_{i\in[n]}$, among predictors in a certain class $\mathcal{A}$ that we can choose) that minimizes the expected loss with respect to a *new* data point, i.e., that minimizes the expected risk $r(a) = \mathbf{E}\,\ell(a, (X, Y)) = \mathbf{E}\,\phi(a(X), Y)$, for a *prediction* loss function $\phi : \mathcal{Y} \times \mathcal{Y} \to \infty$ and a new (independent of the training data) data point (X, Y) coming from the same *unknown* distribution. The results that we derived, either in expectation or with high probability, hold with respect to the randomness in the training data (as the randomness in the test data (X, Y) is already taken into account in the definition of the risk function r) and the randomness in the algorithm we use (for instance, the uniform random variables $I_2, I_3, \ldots$ in the application of stochastic gradient descent to the empirical risk minimization problem, in the so-called multiple passes setup).

In this lecture we start to investigate the *parameter estimation* problem in supervised learning, where the goal is to use the training data to learn an estimator for the parameters of interest. In what we are going to describe, the two main differences compared to the prediction setting are:

- there is *no* test data, only training data.
- the underlying distribution is *not* completely unknown, and a parametric model is assumed.

Given a parametric model for the underlying distribution of the training data, the goal is to use the training data to learn an estimator for the parameters of interest. As for the prediction case where the predictor is a random variable, function of the training data, also the parameter estimator is a random variable, function of the training data. Hence, also in this case the results that we derive, either in expectation or with high probability, hold with respect to the randomness in the training data and, possibly, also the randomness in the algorithms we use.

In particular, we consider the high-dimensional setting in statistical estimation, where the number of parameters to be inferred is larger than the number of data samples available.

Remark 12.1 (Estimation Error and Prediction Error) *As we saw in* **Problem 1.4**, *while estimation and prediction errors are related in some settings, in general these are two very different quantities. In the case of the square loss with linear predictors, where the underlying data generating distribution is given by $Y = \langle X, w^\star\rangle + \sigma\xi$, the prediction error is given by*

$$\mathbf{E}[(\langle X, W\rangle - Y)^2|W] - \mathbf{E}[(\langle X, w^\star\rangle - Y)^2] = (W - w^\star)^\top \mathbf{E}[XX^\top](W - w^\star),$$

while the estimation error is given by

$$\|W - w^\star\|_2^2 = (W - w^\star)^\top (W - w^\star).$$

The population covariance matrix *of the feature vector $\mathbf{E}[XX^\top]$ is what differentiates the two notions of error. Unless X is isotropic (i.e. $\mathbf{E}[XX^\top] = I$), we see that there is a difference between prediction and*

estimation. In particular, note that the estimation deviation $W - w^\star$ could be very large in a direction aligned with a very small eigenvalue of the population covariance matrix $\mathbf{E}[XX^\top]$. In this case, the estimation error $\|W - w^\star\|_2^2$ can be very large while the prediction error can be very small!

12.2 High-Dimensional Statistics: Motivating Examples

To motivate what follows, we consider the classical problem of high-dimensional linear 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 $w^\star \in \mathbb{R}^d$ (unknown to us) such that the observations Y_i's satisfy

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

where $\xi_i \sim \mathcal{N}(0,1)$ 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. Throughout, we use the notation $\langle a, b \rangle = a^\top b$. 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 $\mathbf{x}$ and Y) that minimizes $\|W - w^\star\|_2$, in expectation and with high probability.

We are interested in the high-dimensional setting where the number of parameters to be inferred d is greater than the number of samples at our disposal n, namely, $d > n$. In this case, the matrix $\mathbf{x}$ has a non-empty null space, so the problem is ill-posed as $w^\star$ is not uniquely defined: there are infinitely many vectors w that can give rise to the same observation vector Y. In particular, this means that an "adversary" can choose and hide the "signal" $w^\star$ in the null space of $\mathbf{x}$, taking it as big as they want, and we, as statistician, are never going to find it.

To overcome the ill-posedness of the problem, we need to introduce extra assumptions on the model. Two common types of assumptions involve:

- **Sparsity:** the number of non-zero components of $w^\star$ is bounded: $\|w^\star\|_0 := \sum_{i=1}^d \mathbf{1}_{|w_i^\star| > 0} \le k$. Note that the ℓ_0 "norm" is really a pseudo-norm as the property $\|cx\|_0 = |c| \|x\|_0$ holds only if $|c| = 1$.
- **Low-rank:** in applications where the parameter $w^\star \in \mathbb{R}^d$ can be naturally rearranged as a matrix in $\mathbb{R}^{d_1 \times d_2}$, with $d = d_1 \times d_2$, the rank of $w^\star$ is bounded: $\texttt{Rank}(w^\star) \le k$.

In the lecture notes we will only cover the sparsity case. For the low-rank case we refer to **Problem 4.6** in the Problem Sheets.

12.3 Non-Convex Estimator. Restricted Eigenvalues Condition

Henceforth, we consider the setting of sparsity where $\|w^\star\|_0 \le k$ for a *known* k, and we want to construct an estimator $W \in \mathbb{R}^d$ with $\|W\|_0 \le k$ so that the quantity $\|W - w^\star\|_2$ can be controlled and driven to zero as the number of data points n increases. To this end, we first consider the following estimator:

$$\boxed{W^0 := \operatorname*{argmin}_{w : \|w\|_0 \le k} \frac{1}{2n} \|\mathbf{x} w - Y\|_2^2} \tag{12.1}$$

where the superscript ^{0} refers to the fact that we are considering the ℓ_0 pseudo-norm.

We consider the following assumption on the so-called *design matrix* $\mathbf{x}$ (the name "design matrix" comes from the fact that in some applications, such as in compress sensing, we can design the measurement device $\mathbf{x}$).

Assumption 12.2 (Restricted Eigenvalues condition) *There exists $\alpha > 0$ such that*

$$\boxed{\alpha\|w\|_2^2 \leq \frac{1}{2n}\|\mathbf{x}w\|_2^2 \qquad \text{for any } w \in \mathbb{R}^d \text{ with } \|w\|_0 \leq 2k}$$

Remark 12.3 *If we define the* empirical (or sample) covariance matrix *as*

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

the restricted eigenvalue condition can be rewritten as

$$w^\top \mathbf{c} w \geq 2\alpha w^\top w \qquad \text{for any } w \in \mathbb{R}^d \text{ with } \|w\|_0 \leq 2k,$$

that is,

$$\frac{w^\top \mathbf{c} w}{w^\top w} \geq 2\alpha \qquad \text{for any } w \in \mathbb{R}^d \setminus \{0\} \text{ with } \|w\|_0 \leq 2k.$$

As the matrix $\mathbf{c}$ is symmetric and real-valued, all its eigenvalues are real. The condition $\frac{w^\top \mathbf{c} w}{2w^\top w} \geq \alpha$ for any $w \in \mathbb{R}^d \setminus \{0\}$ corresponds to the condition that the minimum eigenvalue of $\mathbf{c}/2$ is greater than α. The additional requirement $\|w\|_0 \leq 2k$ yields the "restriction" on this eigenvalue condition.

Remark 12.4 (Connection with the Restricted Isometry Property) *The matrix $\mathbf{x} \in \mathbb{R}^{n\times d}$ is said to satisfy the* restricted isometry property with parameter $2k$ *if there exists $\delta \in (0,1)$ such that*

$$(1-\delta)\|w\|_2^2 \leq \frac{1}{2n}\|\mathbf{x}w\|_2^2 \leq (1+\delta)\|w\|_2^2 \qquad \text{for any } w \in \mathbb{R}^d \text{ with } \|w\|_0 \leq 2k.$$

This assumption is typically used in the compressed sensing literature. The condition $1-\delta \leq \frac{w^\top \mathbf{c} w}{2w^\top w} \leq 1+\delta$ for any $w \in \mathbb{R}^d \setminus \{0\}$ corresponds to the condition that all *the eigenvalues of $\mathbf{c}/2$ are in the interval $(1-\delta, 1+\delta)$. The additional requirement $\|w\|_0 \leq 2k$ yields the "restriction" on this eigenvalue condition. For our purposes, we will only need Assumption 12.2, which just deal with the lower bound on the eigenvalues of the matrix for sparse vectors.*

Assumption 12.2 provides a way to control the error $\Delta := W^0 - w^\star$, which is at most $2k$-sparse, and to yield the following bound for the ℓ_2 norm of the random vector $W^0 - w^\star$ in terms of the ℓ_∞ norm of the random vector $\frac{\mathbf{x}^\top \xi}{n}$, where $\mathbf{x}^\top$ denotes the transpose of the matrix $\mathbf{x}$.

Theorem 12.5 *Let W^0 be defined as in* (12.1)*. If Assumption 12.2 holds, then*

$$\boxed{\|W^0 - w^\star\|_2 \leq \sqrt{2}\frac{\sigma\sqrt{k}}{\alpha}\frac{\|\mathbf{x}^\top \xi\|_\infty}{n}}$$

Proof: Let $\Delta = W^0 - w^\star$. By the definition of W^0, we have

$$\|\mathbf{x}\Delta - \sigma\xi\|_2^2 = \|\mathbf{x}W^0 - Y\|_2^2 \leq \|\mathbf{x}w^\star - Y\|_2^2 = \|\sigma\xi\|_2^2,$$

so that, expanding the square, we find

$$\boxed{\|\mathbf{x}\Delta\|_2^2 \leq 2\sigma\langle \mathbf{x}\Delta, \xi\rangle}$$

This inequality is typically referred to as the *basic inequality* in the compressed sensing literature. Combining this inequality with the restricted eigenvalue assumption, Assumption 12.2, noticing that $\|\Delta\|_0 \leq 2k$, we find

$$\alpha\|\Delta\|_2^2 \leq \frac{1}{2n}\|\mathbf{x}\Delta\|_2^2 \leq \frac{\sigma}{n}\langle \mathbf{x}\Delta, \xi\rangle = \frac{\sigma}{n}\langle \Delta, \mathbf{x}^\top\xi\rangle \leq \frac{\sigma}{n}\|\Delta\|_1\|\mathbf{x}^\top\xi\|_\infty,$$

where the last inequality follows from Hölder's inequality. The proof follows by applying the Cauchy-Swartz's inequality:

$$\|\Delta\|_1 = \langle \operatorname{sign}(\Delta), \Delta\rangle \leq \|\operatorname{sign}(\Delta)\|_2\|\Delta\|_2 \leq \sqrt{2k}\|\Delta\|_2.$$

■

Being an inequality between random variables, the bound in Theorem 12.5 holds for *any* realization of the randomness. Recall that the only source of randomness in the setting that we are exploring is due to the random (noise) vector ξ, as W^0 defined in (12.1) is a function of ξ via the definition of the labels Y.

12.4 Bounds in Expectation. Gaussian Complexity

The derivation of bounds in expectation for the high-dimensional setting that we are considering, where the noise is assumed to be Gaussian, is related to a notion of complexity similar to the notion of Rademacher complexity that we saw earlier in the previous lectures. This notion is called *Gaussian complexity* (a.k.a. *Gaussian width*), and is obtained by using Gaussian random variables in place of Rademacher random variables.

Definition 12.6 *The Gaussian complexity of a set $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as*

$$\boxed{\texttt{Gauss}(\mathcal{T}) := \mathbf{E}\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n \xi_i t_i}$$

where $\xi_1, \ldots, \xi_n$ are independent standard Gaussian random variables.

Let us define the class of functions

$$\mathcal{A}_1 := \{x \in \mathbb{R}^d \to \langle u, x\rangle \in \mathbb{R} : u \in \mathbb{R}^d, \|u\|_1 \leq 1\}.$$

Recall that the notation $\mathcal{A}_1 \circ \{x_1, \ldots, x_n\}$ denotes the subset of $\mathbb{R}^n$ that is obtained by applying the functions in $\mathcal{A}_1$ to each element in $\{x_1, \ldots, x_n\}$, namely,

$$\mathcal{A}_1 \circ \{x_1, \ldots, x_n\} := \{(\langle u, x_1\rangle, \ldots, \langle u, x_n\rangle) \in \mathbb{R}^n : u \in \mathbb{R}^d, \|u\|_1 \leq 1\}.$$

Corollary 12.7 *We have*

$$\boxed{\mathbf{E}\frac{\|\mathbf{x}^\top\xi\|_\infty}{n} = \texttt{Gauss}(\mathcal{A}_1 \circ \{x_1, \ldots, x_n\})}$$

Let $\|w^\star\|_0 \leq k$ and let W^0 be defined as in (12.1). If Assumption 12.2 holds, then

$$\boxed{\mathbf{E}\|W^0 - w^\star\|_2 \leq \sqrt{2}\frac{\sigma\sqrt{k}}{\alpha}\texttt{Gauss}(\mathcal{A}_1 \circ \{x_1, \ldots, x_n\})}$$

Proof: First of all, recall that the ℓ_∞ norm is the dual of the ℓ_1 norm, namely,

$$\|\mathbf{x}^\top \xi\|_\infty = \sup_{u \in \mathbb{R}^d : \|u\|_1 \le 1} \langle \mathbf{x}u, \xi \rangle.$$

To see this, note that Hölder's inequality yields $\langle \mathbf{x}u, \xi \rangle = \langle u, \mathbf{x}^\top \xi \rangle \le \|u\|_1 \|\mathbf{x}^\top \xi\|_\infty$ for any vector u, so that

$$\|\mathbf{x}^\top \xi\|_\infty \ge \sup_{u \in \mathbb{R}^d : \|u\|_1 \le 1} \langle \mathbf{x}u, \xi \rangle.$$

On the other hand, note that the choice $u = e_j$, $j \in [d]$, satisfies $\|u\|_1 = 1$ and yields $\langle \mathbf{x}e_j, \xi \rangle = \langle e_j, \mathbf{x}^\top \xi \rangle = (\mathbf{x}^\top \xi)_j$, so that the inequality is achieved by at least one of the vectors e_j, $j \in [d]$. Recalling that the i-th row of the matrix $\mathbf{x}$ corresponds to the feature vector of the i-th data point in our sample set, we get

$$\langle \mathbf{x}u, \xi \rangle = \sum_{i=1}^n (\mathbf{x}u)_i \xi_i = \sum_{i=1}^n \langle u, x_i \rangle \xi_i,$$

so that

$$\frac{1}{n} \mathbf{E} \|\mathbf{x}^\top \xi\|_\infty = \mathbf{E} \sup_{u \in \mathbb{R}^d : \|u\|_1 \le 1} \frac{1}{n} \sum_{i=1}^n \xi_i \langle u, x_i \rangle = \texttt{Gauss}(\mathcal{A}_1 \circ \{x_1, \dots, x_n\}).$$

The proof follows by Theorem 12.5. ∎

12.5 Bounds in Probability. Gaussian Concentration

We now derive a bound that holds with high probability, under the additional assumption of column normalization.

Assumption 12.8 (Column normalization) *The ℓ_2 norm of each column of $\mathbf{x}$ is less than $\sqrt{n}$, or, equivalently, each entry of the empirical covariance matrix $\mathbf{c} = \frac{\mathbf{x}^\top \mathbf{x}}{n}$ is less than 1:*

$$\boxed{\mathbf{c}_{jj} = \left(\frac{\mathbf{x}^\top \mathbf{x}}{n} \right)_{jj} = \frac{1}{n} \sum_{i=1}^n x_{ij}^2 \le 1}$$

The following result relies on concentration inequalities for sun-Gaussian random variables, as the only randomness in our model is due to the random vector ξ, which is Gaussian.

Corollary 12.9 *If Assumption 12.8 holds, then*

$$\boxed{\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_\infty}{\sqrt{n}} \ge \varepsilon \right) \le 2de^{-\frac{\varepsilon^2}{2}}}$$

Let $\|w^\star\|_0 \le k$ and let W^0 be defined as in (12.1)*. If Assumption 12.2 also holds, then for any $\tau > 2$ we have*

$$\boxed{\mathbf{P}\left(\|W^0 - w^\star\|_2 < \frac{\sigma}{\alpha} \sqrt{\frac{2\,k\,\tau \log d}{n}} \right) \ge 1 - \frac{2}{d^{\tau/2-1}}}$$

Proof: Let us define the random vector $V = \frac{\mathbf{x}^\top \xi}{\sqrt{n}} \in \mathbb{R}^d$. As each coordinate V_i is a linear combination of Gaussian random variables, V is a Gaussian random vector with mean

$$\mathbf{E}V = \frac{1}{\sqrt{n}}\mathbf{x}^\top \mathbf{E}\xi = 0,$$

and covariance matrix given by

$$\mathbf{E}[VV^\top] = \frac{1}{n}\mathbf{E}[\mathbf{x}^\top \xi\xi^\top \mathbf{x}] = \frac{1}{n}\mathbf{x}^\top \mathbf{E}[\xi\xi^\top]\mathbf{x} = \frac{\mathbf{x}^\top \mathbf{x}}{n} = \mathbf{c},$$

as ξ is a vector with independent standard Gaussian components, so that $\mathbf{E}[\xi\xi^\top] = I$. That is, $V \sim \mathcal{N}(0, \mathbf{c})$ and, in particular, the i-th component has distribution $V_i \sim \mathcal{N}(0, \mathbf{c}_{ii})$. By the union bound

$$\begin{aligned}\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_\infty}{\sqrt{n}} \geq \varepsilon\right) &= \mathbf{P}(\|V\|_\infty \geq \varepsilon) = \mathbf{P}\left(\max_{i\in[n]} |V_i| \geq \varepsilon\right) = \mathbf{P}\left(\bigcup_{i=1}^d \{|V_i| \geq \varepsilon\}\right) \leq \sum_{i=1}^d \mathbf{P}(|V_i| \geq \varepsilon) \\ &\leq d \max_{i\in[d]} \mathbf{P}(|V_i| \geq \varepsilon).\end{aligned}$$

By concentration for sub-Gaussian random variables (Proposition 6.6) and Assumption 12.8 we have

$$\mathbf{P}(|V_i| \geq \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2\mathbf{c}_{ii}}} \leq 2e^{-\frac{\varepsilon^2}{2}}.$$

Putting everything together we obtain

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_\infty}{\sqrt{n}} \geq \varepsilon\right) \leq 2de^{-\frac{\varepsilon^2}{2}}.$$

By setting $\varepsilon = \sqrt{\tau \log d}$ for $\tau > 2$, we have $2de^{-\frac{\varepsilon^2}{2}} = \frac{2}{d^{\tau/2-1}}$ so that

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_\infty}{n} < \sqrt{\frac{\tau \log d}{n}}\right) \geq 1 - \frac{2}{d^{\tau/2-1}}.$$

The proof follows by applying Theorem 12.5. ∎