1.1 High-dimensional statistics

We are interested in problems where the number of available examples $n$ can be much smaller than the number of free parameters $p$. In classical asymptotic analysis we analyze estimators by fixing $p$ and allowing $n \to \infty$. However, when $p \gg n$, such an analysis may not be useful. Instead, we will aim to analyze methods in a fixed sample regime. That is we will provide performance guarantees that shed light on the behavior of our methods based on fixing $n$, $p$, and other quantities of interest.

1.2 Linear Models

Suppose we let

$$y = X\theta + \epsilon,$$

where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ are known; and $\theta \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$ (problem noise) are unknown. Based on $X$ and $y$, our goal is to find the $\theta$ that best fits the above model. If the rank of $X$ is $p$, then we can simply calculate

$$\hat{\theta}_{OLS} = (X^TX)^{-1}X^Ty,$$

which is the solution to the least squares problem. However, if $p > n$, then the model is unidentifiable since the nullspace of $X$ is non-trivial. That is if we take a $v$ such that $Xv = 0$ and $v \neq 0$ (which exists since $p > n$), then there is no difference between $X\theta$ and $X(\theta + v)$.

1.3 Random Matrix Theory

In this section we will look at more subtle issues that arise in the high-dimensional setting. Let $X_i \in \mathbb{R}^p$, $\mathbb{E}X_i = 0$, and $\text{cov}(X_i) = \Theta \in \mathbb{R}^{p \times p}$ be i.i.d. and assume that $X_i$ has fourth moments. A natural problem is to estimate the covariance matrix $\Theta$. To that end we consider the sample covariance

$$\hat{\Theta} = \frac{1}{n} \sum_{j=1}^{n} X_iX_i^T.$$
Clearly, $E\hat{\Theta} = \Theta$. Classically, we also have that $\hat{\Theta} \xrightarrow{n \to \infty} \Theta$ by the (strong or weak) law of large numbers. Furthermore, we have that $\lambda_{\text{max}}(\hat{\Theta}) \to \lambda_{\text{max}}(\Theta)$ where $\lambda_{\text{max}}(\cdot)$ denotes the largest eigenvalue of its argument. Thus, some objects of interest are well-behaved.

1.3.1 High-dimensional scaling

Instead of fixing $p$ we let $p = \alpha n$ so that as $n$ grows so does $p$ for some $\alpha \in (0, 1)$. Asymptotic results with this scaling go back to Kolmogorov, Wigner, Marcenko & Pastur. Suppose that we consider the simple problem where the true underlying covariance matrix $\Theta = I_{p \times p}$ (the $p \times p$ identity matrix). Then it can be shown that the sample covariance is consistent only if $p/n \to 0$. Indeed, one can show that the asymptotic distribution of the eigenvalues is in the set $[(1 - \sqrt{\alpha})^2, (1 + \sqrt{\alpha})^2]$. Furthermore, $\lambda_{\text{max}}(\Theta) \xrightarrow{a.s.} (1 + \sqrt{\alpha})^2$. Therefore, when $p$ is of the same order as $n$, we already have problems with estimating the covariance matrix.

1.4 Structured estimation

One way to overcome these problems, which we will discuss in the remainder of the course, is to assume additional structure on the underlying parameters ($\theta$ in the setting of linear regression and $\Theta$ for covariance matrix estimation). For example, in the linear regression setting, rather than assume that $\theta$ can be any arbitrary vector, we suppose that $\theta$ belongs to the set of $k$-sparse vectors

$$\Omega_{p,k} = \{\theta | \theta_i \neq 0 \text{ only if } i \in S \text{ and } |S| = k\}$$

where $S \subset \{1, 2, \ldots, n\}$ and $|S|$ denotes the cardinality of the set $S$. Now, even if $p > n$, it is probably the case that for a vector $v$ in the nullspace, $\theta + v$ remains $k$-sparse. Hence, our chances of recovering $\theta$ have improved. In this setting, a popular method to estimate $\theta$ is the LASSO:

$$\hat{\theta}_{\text{LASSO}} = \arg \min_{\theta} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1.$$ 

The above optimization problem uses the $\ell_1$ norm to help encourage the solution to be sparse. This term is called a regularization term. We will see regularizers arise in many settings throughout the course to help control the issues that arise in the high-dimensional regime.