Conclusion on least squares

- now we can look at the solutions:

Theorem (James, 1978): Let $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. If $AA^\dagger b = b$, the complete set of solutions of $Ax = b$ is given by

$$z = A^\dagger b + (\mathrm{I}_d - A^\dagger A)w\,,$$

for $w \in \mathbb{R}^d$.

- $A^\dagger A$ is an orthogonal projection, $I_d - A^\dagger A$ is the orthogonal projection on $\mathrm{Im}(A^\dagger A)^\perp$ and

$$\begin{aligned}\|A^\dagger b + (\mathrm{I}_d - A^\dagger A)w\|^2 &= \|(A^\dagger A)A^\dagger b + (\mathrm{I}_d - A^\dagger A)w\|^2 \\ &= \|A^\dagger b\|^2 + \|(\mathrm{I}_d - A^\dagger A)w\|^2\,.\end{aligned}$$

- taking the Moore-Penrose pseudo-inverse guarantees that **we take the solution with smallest Euclidean norm**.

Gradient descent

- yet another possibility: gradient descent
- **Idea:** minimize $\hat{\mathcal{R}}$ following the steepest descent line
- formally, build the sequence of iterates

$$\begin{cases} \theta^{(0)} & = \theta_0 \\ \theta^{(t+1)} & = \theta^{(t)} - \gamma \nabla \hat{\mathcal{R}}(\theta^{(t)}) \end{cases}$$

  with $\gamma > 0$ the *stepsize*
- if convergence, then $\nabla \hat{\mathcal{R}} = 0$: minimizer
- computational complexity for each step is reduced to $\mathcal{O}(d)$
- it T steps, with $T \ll d^2$, much faster

3.3. Fixed design analysis

Setting

- **Fixed design:** in this section, we assume that Φ is *deterministic*
- namely, fixed, deterministic $x_1, \ldots, x_n \in \mathcal{X}$
- **Assumption I:** there exists $\theta^\star \in \mathbb{R}^d$ such that

$$\forall i \in [n], \qquad Y_i = \varphi(x_i)^\top \theta^\star + \varepsilon_i \,,$$

  with ε_i noise variables
- in matrix notation, we still have:

$$Y = \Phi\theta^\star + \varepsilon \,.$$

- **Assumption II:** the ε_is are independent, have zero mean, and variance $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$
- **Remark (i):** we do not assume identically distributed
- **Remark (ii):** variance assumption is sometimes called *homoscedasticity*

Mahalanobis distance

- for any positive-definite matrix A, we set

$$\forall u \in \mathbb{R}^d, \qquad \|u\|_A^2 := u^\top A u \,.$$

- **Remark (i):** taking $A = \mathrm{I}$, we recover the Euclidean norm
- **Remark (ii):** intuition when A is diagonal: weighting the features
- the function

$$d_A(x, y) := \|x - y\|_A$$

  is often called *Mahalanobis distance*

Excess risk

- under our assumptions, we now turn to the computation of the Bayes risk and excess risk of ordinary least squares
- **Definition:** excess risk $=$ true risk $-$ Bayes risk
- **Notation:** we set $\hat{\Sigma} := \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ the (empirical) covariance matrix

Proposition (excess risk of OLS): under assumptions I and II, for any $\theta \in \mathbb{R}^d$, we have $\mathcal{R}^\star = \sigma^2$ and

$$\mathcal{R}(\theta) - \mathcal{R}^\star = \|\theta - \theta^\star\|_{\hat{\Sigma}}^2 \, .$$

- **Remark (i):** in the presence of noise ($\sigma^2 > 0$), the Bayes risk is positive
- **Remark (ii):** excess risk is the squared distance between our parameter and the true parameter in the geometry defined by $\hat{\Sigma}$

Excess risk, ctd.

Proof: we know that $Y = \Phi\theta^\star + \varepsilon$, thus

$$
\begin{aligned}
\mathcal{R}(\theta) &= \mathbb{E}\left[\frac{1}{n}\left\|Y - \Phi\theta\right\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\left\|\Phi\theta^\star + \varepsilon - \Phi\theta\right\|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left\|\Phi(\theta^\star - \theta)\right\|^2 + 2\varepsilon^\top\Phi(\theta^\star - \theta) + \left\|\varepsilon\right\|^2\right] \\
&= \sigma^2 + \frac{1}{n}(\theta - \theta^\star)^\top\Phi^\top\Phi(\theta - \theta^\star)\,. \qquad (\mathbb{E}\left[\varepsilon_i\right] = 0,\ \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2)
\end{aligned}
$$

Since $\hat{\Sigma}$ is invertible, $\theta^\star$ is the unique global minimizer and the minimum value is σ^2. □

Bias / variance decomposition

Proposition (bias-variance): Let $\hat{\theta} \in \mathbb{R}^d$. Then, under assumption I and II,

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \left\|\mathbb{E}[\hat{\theta}] - \theta^\star\right\|_{\hat{\Sigma}}^2 + \mathbb{E}\left[\left\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\right\|_{\hat{\Sigma}}^2\right]$$

$$\text{expected excess risk} = \text{bias} + \text{variance}$$

Proof: using the previous proposition:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\|\theta - \theta^\star\|_{\hat{\Sigma}}^2\right] \\ &= \mathbb{E}\left[\left\|\theta - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta^\star\right\|_{\hat{\Sigma}}^2\right],\end{aligned}$$

then develop. $\square$

Expectation and variance

- **Reminder:** the OLS solution is given by

$$\hat{\theta} = (\Phi^\top \Phi)^{-1}\Phi^\top Y = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top Y .$$

Proposition (mean and variance of OLS): Let $\hat{\theta}$ be the OLS solution. Assume I and II. Then $\hat{\theta}$ satisfies

$$\mathbb{E}[\hat{\theta}] = \theta^\star \qquad \text{and} \qquad \mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{n}\hat{\Sigma}^{-1} .$$

- **Remark (i):** in the language of statistics, we say that $\hat{\theta}$ is an *unbiased estimator* of $\theta^\star$
- **Remark (ii):** the matrix $\hat{\Sigma}^{-1}$ is sometimes called the *precision* matrix

Expectation and variance, proof

Proof: We know that $\mathbb{E}[Y] = \Phi\theta^\star$, thus

$$\mathbb{E}[\hat{\theta}] = (\Phi^\top\Phi)^{-1}\Phi^\top\Phi\theta^\star = \theta^\star .$$

We deduce that

$$\begin{aligned}\hat{\theta} - \theta^\star &= (\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\theta^\star + \varepsilon) - \theta^\star \\ &= (\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon ,\end{aligned}$$

from which we compute the variance

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathbb{E}\left[(\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi(\Phi^\top\Phi)^{-1}\right] \\ &= \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi)^{-1} && (\mathbb{E}[\varepsilon_i\varepsilon_j] = \sigma^2\mathbb{1}_{i=j}) \\ &= \sigma^2(\Phi^\top\Phi)^{-1} .\end{aligned}$$

□

Excess risk of OLS

Proposition (expected excess risk of OLS): Assume I and II. Then the (expected) excess risk of the ERM is equal to

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2 d}{n} \, .$$

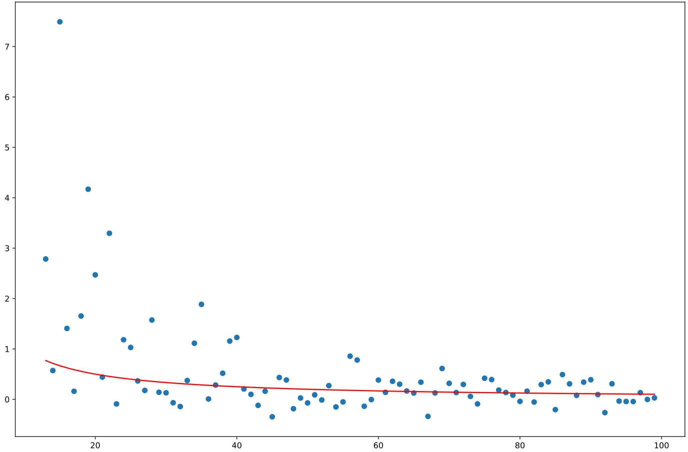
- **Remark (i):** decreasing when $n \to +\infty$ (consistency)
- **Remark (ii):** but, for fixed n, quite bad when $d \approx n$...
- **Remark (iii):** one can show that

$$\mathbb{E}\left[\hat{\mathcal{R}}(\hat{\theta})\right] = \frac{n-d}{n}\sigma^2 = \sigma^2 - \frac{d}{n}\sigma^2 \, ,$$

  thus training error *underestimates* test error, which is

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] = \sigma^2 + \frac{d}{n}\sigma^2 \, .$$

Excess risk of OLS, illustration



- **Figure:** excess risk as a function of n (one simulation per n). Gaussian noise, dimension 10, $\theta^\star = \mathbb{1}$. In red, the expected value $\sigma^2 d/n$.

Excess risk of OLS, proof

Proof: Using our previous computations:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left\|\hat{\theta} - \theta^\star\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\mathrm{trace}\left((\hat{\theta} - \theta^\star)^\top \hat{\Sigma}(\hat{\theta} - \theta^\star)\right)\right] && \text{(definition of } \|\cdot\|_{\hat{\Sigma}}) \\
&= \mathbb{E}\left[\mathrm{trace}\left((\hat{\theta} - \theta^\star)(\hat{\theta} - \theta^\star)^\top \hat{\Sigma}\right)\right] && \text{(cyclic property of the trace)} \\
&= \mathrm{trace}\left(\mathrm{Var}(\hat{\theta})\hat{\Sigma}\right) && \text{(linearity)} \\
&= \mathrm{trace}\left(\frac{\sigma^2}{n}\hat{\Sigma}^{-1}\hat{\Sigma}\right) && \text{(variance computation)} \\
&= \frac{\sigma^2}{n}\mathrm{trace}\left(\mathrm{I}\right)
\end{aligned}$$

$\square$