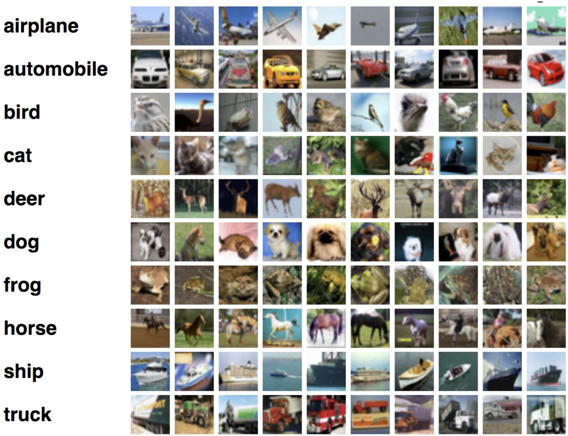
Algorithmic Foundations of Learning

Lecture 12
High-Dimensional Statistics
Sparsity and the Lasso Algorithm

Patrick Rebeschini

Department of Statistics
University of Oxford

Recall. Offline Statistical Learning: Prediction



Offline learning: prediction
Given a batch of observations (images & labels)
interested in predicting the label of a new image

Recall. Offline Statistical Learning: Prediction

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from unknown distribution
2. Choose action $A \in \mathcal{A} \subseteq \mathcal{B}$
3. Suffer an expected/population loss/risk $r(A)$, where

$$a \in \mathcal{B} \longrightarrow r(a) := \mathbf{E}\, \ell(a, Z)$$

with ℓ is an **prediction** loss function and Z is a new test data point

Goal: Minimize the estimation error defined by the following decomposition

$$\underbrace{r(A) - \inf_{a \in \mathcal{B}} r(a)}_{\text{excess risk}} = \underbrace{r(A) - \inf_{a \in \mathcal{A}} r(a)}_{\text{estimation error}} + \underbrace{\inf_{a \in \mathcal{A}} r(a) - \inf_{a \in \mathcal{B}} r(a)}_{\text{approximation error}}$$

as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Note: Estimation/Approximation trade-off, a.k.a. complexity/bias

Offline Statistical Learning: Estimation

User 1	★★★		★★★	
User 2	★★	★★★★		
User 3		★★★	★★	★★★★★

Offline learning: estimation

Given a batch of observations (users & ratings)

interested in estimating the missing ratings in a recommendation system

Offline Statistical Learning: Estimation

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from distr. parametrized by $a^\star \in \mathcal{A}$
2. Choose a parameter $A \in \mathcal{A}$
3. Suffer a loss $\ell(A, a^\star)$ where ℓ is an **estimation** loss function

Goal: Minimize the estimation loss $\ell(A, a^\star)$ as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Main differences:

- **No test data** (i.e., no population risk r).
  Only training data
- Underlying distribution is **not completely unknown**
  We consider a parametric model

Remark: We could also consider prediction losses with a new test data...

Supervised Learning. High-Dimensional Estimation

1. Observe training data $Z_1 = (x_1, Y_1), \dots, Z_n = (x_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from distr. parametrized by $w^\star \in \mathbb{R}^d$:

$$Y_i = \langle x_i, w^\star \rangle + \sigma \xi_i \qquad i \in [n]$$

$$Y = \mathbf{x} w^\star + \sigma \xi \qquad \text{(data in matrix form: } Y \in \mathbb{R}^n \text{ and } \mathbf{x} \in \mathbb{R}^{n \times d}\text{)}$$

2. Choose a parameter $W \in \mathcal{W}$
3. **Goal:** Minimize loss $\ell(W, w^\star) = \|W - w^\star\|_2$

High-dimensional setting: $\boxed{n < d}$ (dimension greater than no. of data)

Assumptions (otherwise problem is ill-posed):

- **Sparsity:** $\|w^\star\|_0 := \sum_{i=1}^d 1_{|w_i^\star|>0} \le k$
- **Low-rank:** $\texttt{Rank}(w^\star) \le k$, when $w^\star$ can be thought of as a matrix

Non-Convex Estimator. Restricted Eigenvalue Condition

Assume that we know k, the upper bound on the sparsity ($\|w^\star\|_0 \leq k$)

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

Restricted eigenvalues (Assumption 12.2)

There exists $\alpha > 0$ such that for any vector $w \in \mathbb{R}^d$ with $\|w\|_0 \leq 2k$ we have

$$\frac{1}{2n} \|\mathbf{x}w\|_2^2 \geq \alpha \|w\|_2^2$$

Statistical Guarantees ℓ_0 Recovery (Theorem 12.5)

If the restricted eigenvalue assumption holds, then

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

Proof of Theorem 12.5

- 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 the *basic inequality*:

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

- The restricted eigenvalue assumption yields, noticing that $\|\Delta\|_0 \leq 2k$:

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

  where the last inequality follows from Hölder's inequality.

- The proof follows by applying the Cauchy-Swartz's inequality:

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

Bounds in Expectation. Gaussian Complexity

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

Gaussian complexity (Definition 12.6)

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

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

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

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

Bounds in Expectation (Corollary 12.7)

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

Proof of Corollary 12.7

- The ℓ_∞ norm is the dual of the ℓ_1 norm: $\|\mathbf{x}^\top \xi\|_\infty = \sup_{u \in \mathbb{R}^d : \|u\|_1 \le 1} \langle \mathbf{x}u, \xi \rangle$

  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 u, so

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

  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, $j \in [d]$.

- We have

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

  so

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

Bounds in Probability. Gaussian Concentration

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

Column normalization (Assumption 12.8)

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

Bounds in Probability (Corollary 12.9)

If the column normalization assumption holds, then

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

Proof of Corollary 12.9 (Part I)

- Let $V = \frac{\mathbf{x}^\top \xi}{\sqrt{n}} \in \mathbb{R}^d$. 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 made of independent standard Gaussian components, so $\mathbf{E}[\xi\xi^\top] = I$

- 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

$$\mathbf{P}\left(\frac{\|\mathbf{x}^\top \xi\|_\infty}{\sqrt{n}} \geq \varepsilon\right) = \mathbf{P}(\|V\|_\infty \geq \varepsilon) = \mathbf{P}\left(\max_{i\in[n]} |V_i| \geq \varepsilon\right)$$

$$= \mathbf{P}\left(\bigcup_{i=1}^{d}\{|V_i| \geq \varepsilon\}\right) \leq \sum_{i=1}^{d} \mathbf{P}(|V_i| \geq \varepsilon) \leq d \max_{i\in[d]} \mathbf{P}(|V_i| \geq \varepsilon)$$

Proof of Corollary 12.9 (Part II)

- 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

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