Algorithmic Foundations of Learning

Lecture 12
High-Dimensional Statistics
Sparsity and the Lasso Algorithm

Patrick Rebeschini

Department of Statistics
University of Oxford
November 21st 2019
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

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

with $\ell$ is an prediction loss function and $Z$ is a new test data point

**Goal:** Minimize the estimation error defined by the following decomposition

$$ r(A) - \inf_{a \in \mathcal{B}} r(a) = r(A) - \inf_{a \in \mathcal{A}} r(a) + \inf_{a \in \mathcal{A}} r(a) - \inf_{a \in \mathcal{B}} r(a) $$

as a function of $n$ and notions of “complexity” of the set $\mathcal{A}$ of the function $\ell$

**Note:** Estimation/Approximation trade-off, a.k.a. complexity/bias
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 \( a^* \in A \)
2. Choose a parameter \( A \in A \)
3. Suffer a loss \( \ell(A, a^*) \) where \( \ell \) is an estimation loss function

**Goal:** Minimize the estimation loss \( \ell(A, a^*) \) as a function of \( n \) and notions of “complexity” of the set \( A \) of the function \( \ell \)

**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 $Z_1 = (x_1, Y_1), \ldots, Z_n = (x_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from distr. parametrized by $w^* \in \mathbb{R}^d$:

   $$Y_i = \langle x_i, w^* \rangle + \sigma \xi_i \quad i \in [n]$$
   $$Y = xw^* + \sigma \xi$$

   (data in matrix form: $Y \in \mathbb{R}^n$ and $x \in \mathbb{R}^{n \times d}$)

2. Choose a parameter $W \in \mathcal{W}$

3. **Goal:** Minimize loss $\ell(W, w^*) = \|W - w^*\|_2$

**High-dimensional setting:** $n < d$ (dimension greater than no. of data)

Assumptions (otherwise problem is ill-posed):

- **Sparsity:** $\|w^*\|_0 := \sum_{i=1}^d 1_{|w_i^*| > 0} \leq k$
- **Low-rank:** $\text{Rank}(w^*) \leq k$, when $w^*$ can be thought of as a matrix
Non-Convex Estimator. Restricted Eigenvalue Condition

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

Algorithm:

$$W^0 := \underset{w : \|w\|_0 \leq k}{\text{argmin}} \frac{1}{2n} \|xw - Y\|_2^2$$

Restricted eigenvalues (Assumption 12.1)

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

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

Statistical Guarantees $\ell_0$ Recovery (Theorem 12.4)

If the restricted eigenvalue assumption holds, then

$$\|W^0 - w^*\|_2 \leq \sqrt{2} \frac{\sigma \sqrt{k} \|x^\top \xi\|_\infty}{\alpha n}$$
Bounds in Expectation. Gaussian Complexity

Recall: \[ \|W^0 - w^*\|_2 \leq \sqrt{2} \frac{\sigma \sqrt{k}}{\alpha} \frac{\|x^\top \xi\|_{\infty}}{n} \]

Gaussian complexity (Definition 12.5)

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

\[
\text{Gauss}(\mathcal{T}) := \mathbf{E} \sup_{t \in \mathcal{T}} 1/n \sum_{i=1}^{n} \xi_i t_i
\]

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

\[ A_1 := \{ x \in \mathbb{R}^d \rightarrow \langle u, x \rangle \in \mathbb{R} : u \in \mathbb{R}^d, \|u\|_1 \leq 1 \} \]

Bounds in Expectation (Corollary 12.6)

\[
\mathbf{E} \frac{\|x^\top \xi\|_{\infty}}{n} = \text{Gauss}(A_1 \circ \{x_1, \ldots, x_n\})
\]
Recall: \[ \| W^0 - w^* \|_2 \leq \sqrt{2} \frac{\sigma \sqrt{k}}{\alpha} \frac{\| x^\top \xi \|_\infty}{n} \]

Column normalization (Assumption 12.7)

\[ c_{jj} = \left( \frac{x^\top x}{n} \right)_{jj} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \leq 1 \]

Bounds in Probability (Corollary 12.8)

If the column normalization assumption holds, then

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