

MIT 9.520/6.860,
Statistical Learning Theory and Applications

Class 03: Regularized Least Squares

Lorenzo Rosasco

Learning problem and algorithms

Solve

$$\min_{f \in \mathcal{F}} L(f), \quad L(f) = \mathbb{E}_{(x,y) \sim P}[\ell(y, f(x))],$$

given only

$$S_n = (x_1, y_1), \dots, (x_n, y_n) \sim P^n.$$

Learning algorithm

$$S_n \rightarrow \widehat{f} = \widehat{f}_{S_n},$$

\widehat{f} estimates f_P given the observed examples S_n .

Learning problem and algorithms

Solve

$$\min_{f \in \mathcal{F}} L(f), \quad L(f) = \mathbb{E}_{(x,y) \sim P}[\ell(y, f(x))],$$

given only

$$S_n = (x_1, y_1), \dots, (x_n, y_n) \sim P^n.$$

Learning algorithm

$$S_n \rightarrow \widehat{f} = \widehat{f}_{S_n},$$

\widehat{f} estimates f_P given the observed examples S_n .

How can we design a learning algorithm?

Algorithm design: complexity and regularization

The design of most algorithms proceed as follows:

- ▶ Pick a (possibly large) class of function \mathcal{H} , ideally

$$\min_{f \in \mathcal{H}} L(f) = \min_{f \in \mathcal{F}} L(f).$$

- ▶ Define a procedure $A_\gamma(S_n) = \hat{f}_\gamma \in \mathcal{H}$ to *explore* the space \mathcal{H} .

Empirical risk minimization

A classical example (called M-estimation in statistics).

Consider $(\mathcal{H}_\gamma)_\gamma$ such that

$$\mathcal{H}_1 \subset \mathcal{H}_2, \dots, \mathcal{H}_\gamma \subset \dots \mathcal{H}.$$

Then, let

$$\hat{f}_\gamma = \min_{f \in \mathcal{H}_\gamma} \widehat{L}(f), \quad \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)).$$

This is the idea we discuss next.

Linear functions

Let \mathcal{H} be the space of linear functions

$$f(x) = w^\top x.$$

Then,

- ▶ $f \leftrightarrow w$ is one to one,
- ▶ inner product $\langle f, \bar{f} \rangle_{\mathcal{H}} := w^\top \bar{w}$,
- ▶ norm/metric $\|f - \bar{f}\|_{\mathcal{H}} := \|w - \bar{w}\|$.

Linear functions are the conceptual building block of most functions.

Linear least squares

ERM with least squares also called ordinary least squares (OLS)

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2}_{\widehat{L}(w)}.$$

- ▶ Statistics later...
- ▶ ...now **computations**.

Matrices and linear systems

Let $\widehat{X} \in \mathbb{R}^{nd}$ and $\widehat{Y} \in \mathbb{R}^n$. Then

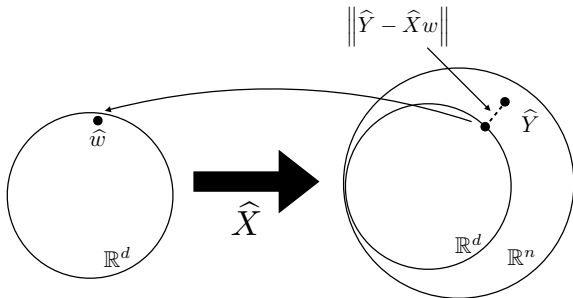
$$\frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2.$$

This is the least squares problem associated to the linear system

$$\widehat{X}w = \widehat{Y}.$$

Overdetermined lin. syst.

$$n > d$$



$$\nexists \hat{w} \text{ s.t. } \hat{X}\hat{w} = \hat{Y}$$

Least squares solutions

From the optimality conditions

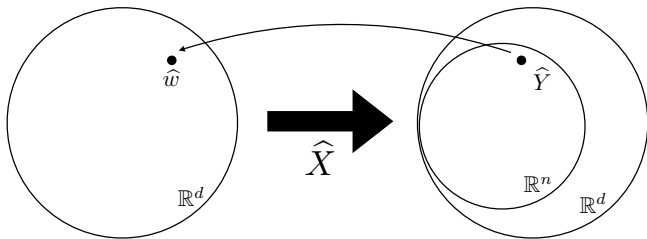
$$\nabla_w \frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2 = 0$$

we can derive the **normal equation**

$$\widehat{X}^T \widehat{X}w = \widehat{X}^T \widehat{Y} \quad \Leftrightarrow \quad \widehat{w} = (\widehat{X}^T \widehat{X})^{-1} \widehat{X}^T \widehat{Y}.$$

Underdetermined lin. syst.

$$n < d$$



$$\exists \hat{w} \text{ s.t. } \hat{X}\hat{w} = \hat{Y}$$

possibly not unique...

Minimal norm solution

There can be many solutions

$$\widehat{X}\widehat{w} = \widehat{Y}, \quad \text{and} \quad \widehat{X}w_0 = 0 \quad \Rightarrow \quad \widehat{X}(\widehat{w} + w_0) = \widehat{Y}.$$

Consider

$$\min_{w \in \mathbb{R}^d} \|w\|^2, \quad \text{subj. to} \quad \widehat{X}w = \widehat{Y}.$$

Using the method of Lagrange multipliers, the solution is

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top)^{-1} \widehat{Y}.$$

Pseudoinverse

$$\widehat{w} = \widehat{X}^\dagger \widehat{Y}$$

For $n > d$, (independent columns)

$$\widehat{X}^\dagger = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top.$$

For $n < d$, (independent rows)

$$\widehat{X}^\dagger = \widehat{X}^\top (\widehat{X} \widehat{X}^\top)^{-1}.$$

Spectral view

Consider the SVD of \widehat{X}

$$\widehat{X} = USV^T \quad \Leftrightarrow \quad \widehat{X}w = \sum_{j=1}^r s_j (v_j^T w) u_j,$$

here $r \leq n \wedge d$ is the rank of \widehat{X} .

Then,

$$\widehat{w}^+ = \widehat{X}^+ \widehat{Y} = \sum_{j=1}^r \frac{1}{s_j} (u_j^T \widehat{Y}) v_j.$$

Pseudoinverse and bias

$$\widehat{w}^{\dagger} = \widehat{X}^{\dagger} \widehat{Y} = \sum_{j=1}^r \frac{1}{s_j} (u_j^{\top} \widehat{Y}) v_j.$$

$(v_j)_j$ are principal components of \widehat{X} : OLS “likes” principal components.

Not all linear functions are the same for OLS!

The pseudoinverse introduces a bias towards certain solutions.

From OLS to ridge regression

Recall, it also holds,

$$\widehat{X}^\dagger = \lim_{\lambda \rightarrow 0_+} (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top = \lim_{\lambda \rightarrow 0_+} \widehat{X}^\top (\widehat{X} \widehat{X}^\top + \lambda I)^{-1}.$$

Consider for $\lambda > 0$,

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}.$$

This is called ridge regression.

Spectral view on ridge regression

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

Considering the SVD of \widehat{X} ,

$$\widehat{w}_\lambda = \sum_{j=1}^r \frac{s_j}{s_j^2 + \lambda} (u_j^\top \widehat{Y}) v_j.$$

Ridge regression as filtering

$$\widehat{w}_\lambda = \sum_{j=1}^r \frac{s_j}{s_j^2 + \lambda} (u_j^\top \widehat{Y}) v_j$$

The function

$$F(s) = \frac{s}{s^2 + \lambda},$$

acts as a low pass filter (low frequencies= principal components).

- ▶ For s small, $F(s) \approx 1/\lambda$.
- ▶ For s big, $F(s) \approx 1/s$.

Ridge regression as ERM

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

is the solution of

$$\min_{w \in \mathbb{R}^d} \underbrace{\|\widehat{Y} - \widehat{X}w\|^2 + \lambda \|w\|^2}_{\widehat{L}_\lambda(w)}.$$

It follows from,

$$\Delta \widehat{L}_\lambda(w) = -\frac{2}{n} \widehat{X}^\top (\widehat{Y} - \widehat{X}w) + 2\lambda w = 2\left(\frac{1}{n} \widehat{X}^\top \widehat{X} + \lambda I\right)w - \frac{2}{n} \widehat{X}^\top \widehat{Y}.$$

Ridge regression as ERM

ERM interpretation suggests the rescaling

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + n\lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

since

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2 + \lambda \|w\|^2}_{\widehat{L}_\lambda(w)}.$$

Related ideas

Tikhonov

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2 + \lambda \|w\|^2$$

Morozov

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{subj. to} \quad \frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2 \leq \delta$$

Ivanov

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{Y} - \widehat{X}w\|^2, \quad \text{subj. to} \quad \|w\|^2 \leq R$$

Ridge regression and SRM

The constraint

$$\|w\|^2 \leq R$$

- ▶ restricts the search of solution,
- ▶ shrinks the solution coefficients.

Different views on regularization

$$\widehat{w} = \widehat{X}^\dagger \widehat{Y}$$

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

$$\min_{w \in \mathbb{R}^d} \text{s.t. } \widehat{X}w = \widehat{Y} \quad \|w\|^2$$

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2 + \lambda \|w\|^2$$

- ▶ Introduces a *bias* towards certain solutions: small norm/principal components,
- ▶ controls the *stability* of the solution .

Complexity of ridge regression

Back to computations.

Solving

$$\widehat{w}^\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

requires essentially (using a direct solver)

- ▶ time $O(nd^2 + d^3)$,
- ▶ memory $O(nd \vee d^2)$.

What if $n \ll d$?

Representer theorem in disguise

A simple observation

Using SVD we can see that

$$(\widehat{X}^T \widehat{X} + \lambda I)^{-1} \widehat{X}^T = \widehat{X}^T (\widehat{X} \widehat{X}^T + \lambda I)^{-1}$$

More on complexity

Then

$$\widehat{w}_\lambda = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1} \widehat{Y}.$$

requires essentially (using a direct solver)

- ▶ time $O(n^2d + n^3)$,
- ▶ memory $O(nd \vee n^2)$.

Representer theorem

Note that

$$\widehat{\mathbf{w}}_\lambda = \widehat{\mathbf{X}}^\top \underbrace{(\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top + \lambda I)^{-1} \widehat{\mathbf{Y}}}_{\mathbf{c} \in \mathbb{R}^n} = \sum_{i=1}^n x_i c_i.$$

The coefficients vector is a linear combination of the input points.

Then

$$\widehat{f}_\lambda(x) = x^\top \widehat{\mathbf{w}}_\lambda = x^\top \widehat{\mathbf{X}}^\top \mathbf{c} = \sum_{i=1}^n x^\top x_i c_i$$

The function we obtain is a linear combination of *inner products*.

This will be the key to nonparametric learning.

Summing up

- ▶ From OLS to ridge regression
- ▶ Different views: (spectral) filtering and ERM
- ▶ Regularization and bias.

TBD

- ▶ Beyond linear models.
- ▶ Optimization.
- ▶ Model selection.