ICTP, Fall 2018
Winter School on Learning and AI

Class 02: Implicit Regularization

Lorenzo Rosasco

Statistical learning

- $X \times Y$ probability space, with measure P.

Problem: Solve

$$\min_{f:X \to Y} \mathbb{E}_{(x,y)\sim P}[(y - f(x))^2],$$

given only

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

sampled i.i.d. with P fixed, but unknown.

Learning algorithm design so far

- ERM, penalized/constrained

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

- Direct solver

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

Non linear extensions via features/kernels.

Beyond ERM

- Are there other algorithm design principles?

Today we will see how *optimization regularizes implicitly*.

Least squares (recap)

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

$$\underbrace{\min_{w\in\mathbb{R}^d} \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2}_{n>d} \qquad \underbrace{\min_{w\in\mathbb{R}^d} \left\| w \right\|^2, \quad \text{subj. to} \quad \widehat{X} w = \widehat{Y}}_{n<d}$$

$$\Rightarrow \quad \widehat{w}^\dagger = \widehat{X}^\dagger \widehat{Y}.$$

Iterative solvers for least squares

Let

$$\widehat{L}(w) = \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2 .$$

The gradient descent iteration is

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma \frac{2}{n} \widehat{X}^\top (\widehat{X} \widehat{w}_t - \widehat{Y}).$$

For suitable γ

$$\widehat{L}(\widehat{w}_t) \to \min \widehat{L}(w)$$

Implicit bias/regularization

It is easy to see that gradient descent

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma \frac{2}{n} \widehat{X}^\top (\widehat{X} \widehat{w}_t - \widehat{Y}),$$

converges to the minimal norm solution for suitable w_0.

Reminder: the minimal norm solution $\widehat{w}^\dagger$ satisfies

$$\widehat{w}^\dagger = \widehat{X}^\top c, \quad c \in \mathbb{R}^n \qquad \text{that is} \qquad \widehat{w}^\dagger \perp \mathrm{Null}(\widehat{X}).$$

Implicit bias/regularization

Then,

$$\widehat{w}_t \mapsto \widehat{w}^\dagger.$$

Gradient descent explores solutions with a *bias* towards small norms.

Regularization is not achieved via explicit constraint/penalties.

In this sense it is *implicit*.

Terminology: regularization and pseudosolutions?

- In signal processing minimal norm solutions are called regularization.
- In classical regularization theory, they are called pseudosolutions.
- Regularization refers to a family of solutions converging to pseudosolutions, e.g. Tikhonov's. See later.

Terminology: implicit or iterative regularization?

- In machine learning, implicit regularization has recently become fashionable.
- It refers to regularization achieved without imposing constraints or adding penalties.
- In classical regularization theory, it is called *iterative* regularization and it is a classic idea.
- We will see the idea of early stopping is also very much related.

Back for more regularization

According to classical regularization theory: among different regularized solutions, one ensuring stability should be selected.

- For example, in Tikhonov regularization

  $$\widehat{w}^{\lambda} \to \widehat{w}^{\dagger}$$

  as $\lambda \to 0$.

- But in practice $\lambda \neq 0$ is chosen, when data are noisy/sampled.

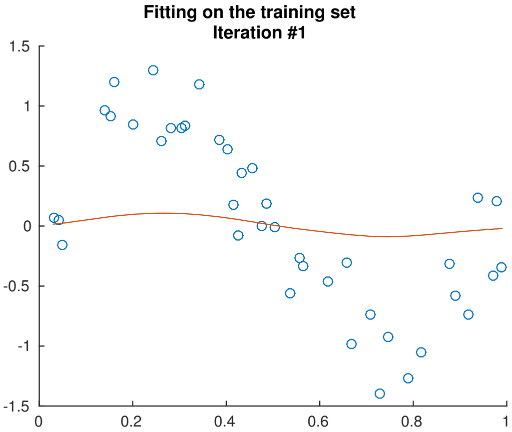
Regularization by gradient descent?

Gradient descent converges to the minimal norm solution, but:

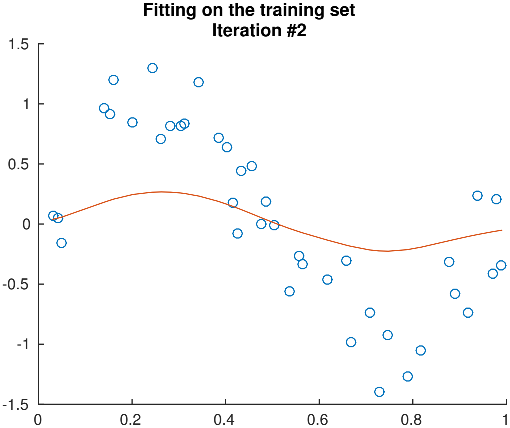
- does it define meaningful regularized solutions?

- Where is the regularization parameter?

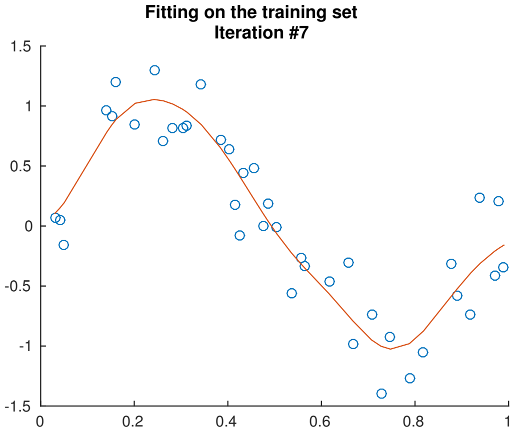
An intuition: early stopping



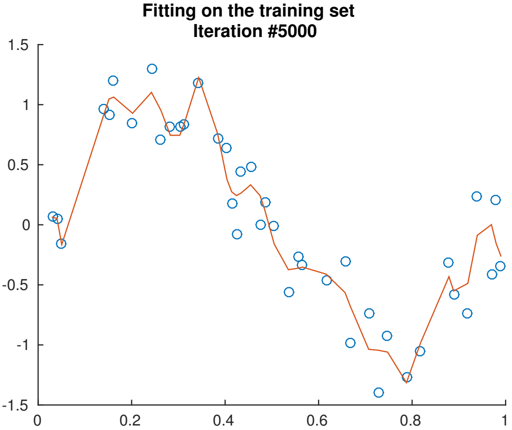
An intuition: early stopping



An intuition: early stopping



An intuition: early stopping



Is there a way to formalize this intuition?

Interlude: geometric series

Recall for $|a| < 1$

$$\sum_{j=0}^{\infty} a^j = (1-a)^{-1}, \qquad \sum_{j=0}^{t} a^j = (1-a^t)(1-a)^{-1}.$$

Equivalently for $|b| < 1$

$$\sum_{j=0}^{\infty} (1-b)^j = b^{-1}, \qquad \sum_{j=0}^{t} (1-b)^j = (1-(1-b)^t)b^{-1}.$$

Interlude II: Neumann series

Assume $I - A$ invertible matrix and $\|A\| < 1$

$$\sum_{j=0}^{\infty} A^j = (I - A)^{-1}, \qquad \sum_{j=0}^{t} A^j = (I - A^t)(I - A)^{-1}.$$

or equivalently B invertible[1] and $\|B\| < 1$

$$\sum_{j=0}^{\infty} (I - B)^j = B^{-1}, \qquad \sum_{j=0}^{t} (I - B)^j = (I - (I - B)^t) B^{-1}.$$

[1] Argument can be extended to pseudoinverses.

Rewriting GD

By induction

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma \frac{2}{n} \widehat{X}^\top (\widehat{X}\widehat{w} - \widehat{Y})$$

can be written as

$$\widehat{w}_{t+1} = \gamma \frac{2}{n} \sum_{j=0}^{t} (I - \gamma \widehat{X}^\top \widehat{X})^j \widehat{X}^\top \widehat{Y}.$$

Rewriting GD (cont.)

- Write

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma\frac{2}{n}\widehat{X}^\top(\widehat{X}\widehat{w} - \widehat{Y}) = (I - \gamma\frac{2}{n}\widehat{X}^\top\widehat{X})\widehat{w}_t + \gamma\frac{2}{n}\widehat{X}^\top\widehat{Y}.$$

- Assume

$$\widehat{w}_t = \gamma\frac{2}{n}\sum_{j=0}^{t-1}(I - \gamma\frac{2}{n}\widehat{X}^\top\widehat{X})^j\widehat{X}^\top\widehat{Y}.$$

- Then

$$\begin{aligned}\widehat{w}_{t+1} &= (I - \gamma\frac{2}{n}\widehat{X}^\top\widehat{X})\gamma\frac{2}{n}\sum_{j=0}^{t-1}(I - \gamma\frac{2}{n}\widehat{X}^\top\widehat{X})^j\widehat{X}^\top\widehat{Y} + \gamma\frac{2}{n}\widehat{X}^\top\widehat{Y} \\ &= \gamma\frac{2}{n}\sum_{j=0}^{t}(I - \gamma\frac{2}{n}\widehat{X}^\top\widehat{X})^j\widehat{X}^\top\widehat{Y}.\end{aligned}$$

Neumann series and GD

This is pretty cool

$$\widehat{w}_{t+1} = \gamma \frac{2}{n} \sum_{j=0}^{t} (I - \gamma \frac{2}{n} \widehat{X}^\top \widehat{X})^j \widehat{X}^\top \widehat{Y}.$$

GD is a truncated power series approximation of the pseudoinverse!

If γ is such that[2] $\left\| I - \gamma \frac{2}{n} \widehat{X}^\top \widehat{X} \right\| < 1$, then for large t

$$\gamma \frac{2}{n} \sum_{j=0}^{t} (I - \gamma \frac{2}{n} \widehat{X}^\top \widehat{X})^j \widehat{X}^\top \approx \widehat{X}^\dagger$$

and we recover $\widehat{w}_t \to \widehat{w}^\dagger$.

[2] Compare to classic conditions.

Stability properties of GD

For any t

$$\widehat{w}_t = (I - (I - \gamma \frac{2}{n} \widehat{X}^\top \widehat{X})^t)(\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{Y}$$

(assume invertibility for simplicity).

Then

$$\underbrace{\widehat{w}_t \approx (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{Y},}_{\text{large } t} \qquad \underbrace{\widehat{w}_t \approx \frac{\gamma}{n} \widehat{X}^\top \widehat{Y}}_{\text{small } t}.$$

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

$$\underbrace{\widehat{w}_\lambda \approx (\widehat{X}^\top \widehat{X})^{-1} \widehat{Y},}_{\text{small } \lambda} \qquad \underbrace{\widehat{w}_\lambda \approx \lambda n \widehat{X}^\top \widehat{Y}}_{\text{large } \lambda}.$$

Spectral view and filtering

Recall for Tikhonov

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

For GD

$$\widehat{w}^{\lambda} = \sum_{j=1}^{r} \frac{(1 - (1 - \gamma \frac{2}{n} \sigma_j^2)^t)}{\sigma_j} (u_j^\top \widehat{Y}) v_j.$$

Both methods can be seen as spectral filtering

$$\widehat{w}^{\lambda} = \sum_{j=1}^{r} F(\sigma_j)(u_j^\top \widehat{Y}) v_j,$$

for some suitable filter function F.

Implicit regularization and early stopping

The stability of GD decreases with t, i.e. higher condition number for

$$(I - (I - \gamma \frac{2}{n} \widehat{X}^\top \widehat{X})^t)(\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top.$$

Early-stopping the iteration as a (implicit) regularization effect.

Summary so far

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma \frac{2}{n} \widehat{X}^\top (\widehat{X}\widehat{w} - \widehat{Y}) = \gamma \frac{2}{n} \sum_{j=0}^{t} (I - \gamma \widehat{X}^\top \widehat{X})^j \widehat{X}^\top \widehat{Y}.$$

- Implicit bias: gradient descent converges to the minimal norm solution.

- Stability: the number of iteration is a regularization parameter.

Name game: gradient descent, Landweber iteration, L^2-Boosting.

A bit of history

These ideas are fashionable nowt but has also a long history.

- The idea that iterations converge to pseudosolutions is from the 50's.
- The observation that iterations control stability dates back at least to the 80's.

Classic name is iterative regularization (there are books about it).

Why is it back in fashion?

- Early stopping is used as a heuristic while training neural nets.

- Convergence to minimal norm solutions could help understanding generalization in deep learning?

- New perspective on algorithm design merging statistics and optimization.

Statistics meets optimization

GD offers a new a perspective on algorithm design.

- Training time= complexity?
- Iterations control statistical accuracy *and* numerical complexity.
- Recently, this kind of regularization is called computational or algorithmic.

Beyond least squares

- Other forms of optimization?
- Other loss functions?
- Other norms?
- Other class of functions?

Other forms of optimization

Largely unexplored there are results on:

- Accelerated methods and conjugate gradient.
- Stochastic/incremental gradient methods.

It is clear that other parameters control regularization/stability, e.g step-size, mini-batch-size, averaging etc.

Other loss functions

There are some results.

For ℓ convex, let

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i).$$

The gradient/subgradient descent iteration is

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \nabla \widehat{L}(\widehat{w}_t).$$

Other loss functions (cont.)

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \nabla \widehat{L}(\widehat{w}_t)$$

An intuition: note that, if $\sup_t \left\| \nabla \widehat{L}(\widehat{w}_t) \right\| \leq B$

$$\|\widehat{w}_t\| \leq \sum_t \gamma_t B,$$

the number of iterations/stepsize control the norm of the iterates.

Other norms

Largely unexplored.

- Gradient descent needs be replaced to bias iterations towards desired norms.
- Bregman iterations, mirror descent, proximal gradients can be used.

Other class of functions

Extensions using kernel/features are straight forward.

Considering neural nets is considerably harder.

In this context the following perspective has been considered:

- given a the function class (neural nets),
- given an algorithm (SGD),
- find which norm the iterates converge to.

Summary

A different way to design algorithms.

- Implicit/iterative regularization.
- Iterative regularization for least squares.
- Extensions.