MIT 9.520/6.860
Statistical Learning Theory and Applications

Class 07: Implicit Regularization

Lorenzo Rosasco

Learning algorithm design so far

- ERM, penalized/constrained,

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

- Optimization by GD/SGD,

$$w_{t+1} = w_t - \gamma \nabla \widehat{L}^\lambda(w_t).$$

Non linear extensions via features/kernels.

Beyond ERM

- Are there other algorithm design principles?

- So far statistics/regularization separate from computations.

Today we will see how *optimization regularizes implicitly*.

Least squares recap: minimal norm

We recall that for least squares

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

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

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

Least squares recap: pernalized ERM

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

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

$$\lim_{\lambda \to 0} (\widehat{X}^\top \widehat{X} + \lambda n I)^{-1} \widehat{X}^\top = \widehat{X}^\dagger \quad \Rightarrow \quad \lim_{\lambda \to 0} \widehat{w}_\lambda = \widehat{w}^\dagger$$

Least squares learning with gradient descent

We want to understand the learning properties of

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

the gradient descent iteration for the empirical risk,

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

No penalties or constraints (!): what can we hope for?

Implicit bias of gradient descent

We know that, for suitable γ

$$\lim_{t \to \infty} \widehat{w}_t = \arg\min \widehat{L}(w).$$

In fact, we show that

$$\lim_{\lambda \to 0} \widehat{w}_t = \widehat{w}^{\dagger}.$$

Understanding the implicit bias of GD for LS

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

If $w_0 \in \text{Range}(\widehat{X}^\top)$ (e.g. $w_0 = 0$), then $\widehat{w}_t \in \text{Range}(\widehat{X}^\top)$ for all t.
$\Rightarrow$ GD converges to a solution in $\text{Range}(\widehat{X}^\top)$.

The minimal norm solution $\widehat{w}^\dagger$ is the unique solution in[1] $\text{Range}(\widehat{X}^\top)$.

Then,

$$\lim_{t \to \infty} \widehat{w}_t = \widehat{w}^\dagger.$$

[1] Recall that $\text{Range}(\widehat{X}^\top) = \text{Null}(\widehat{X})^\perp$

Implicit bias/regularization

Compare

$$\lim_{t\to\infty} \widehat{w}_t = \widehat{w}^\dagger,$$

to

$$\lim_{\lambda\to 0} \widehat{w}^\lambda = \widehat{w}^\dagger.$$

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

- The bias is *implicit* in the sense that there are no explicit constraint/penalties.

Implicit bias for other loss functions

Regression

Consider $\ell(x^\top w - y)$ and

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} x_i \ell'(x_i^\top w - y_i),$$

Then $w_t \in \text{Range}(\widehat{X}^\top)$, and if $\ell(0) = 0$ and $n > d$ then $\widehat{w}_t \to \widehat{w}^\dagger$ (why?).

Classification

Consider $\ell(x^\top wy)$ and

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} x_i \ell'(x_i^\top w y_i).$$

Then $w_t \in \text{Range}(\widehat{X}^\top)$. Convergence to max margin solutions can be ensured in some cases.

Terminology: regularization and pseudosolutions?

Recall that:

- 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.

Back for more regularization

- There are problems in which $\widehat{w}^{\dagger}$ is unstable.

- Regularization: define a family of solutions + select one ensuring stability.

Tikhonov regularization

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

as $\lambda \to 0$. Choose $\lambda \neq 0$ if needed.

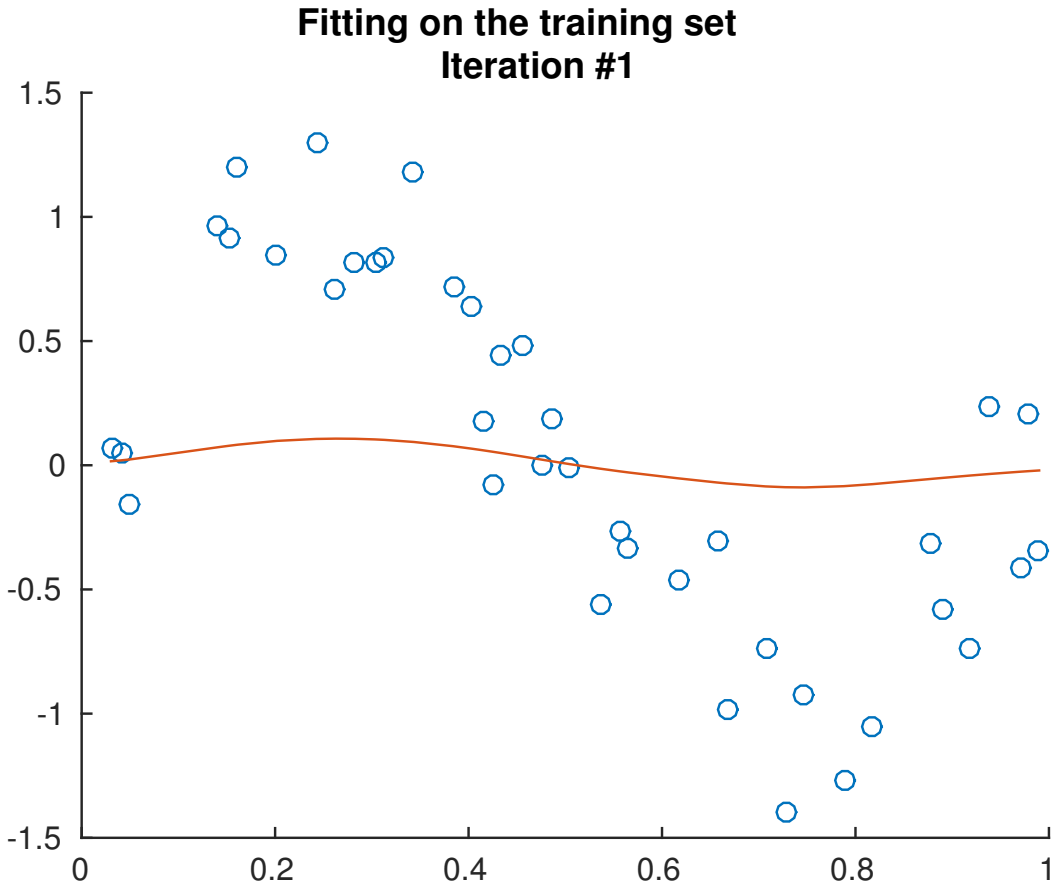
Regularization by gradient descent?

Gradient descent converges to the minimal norm solution, but:

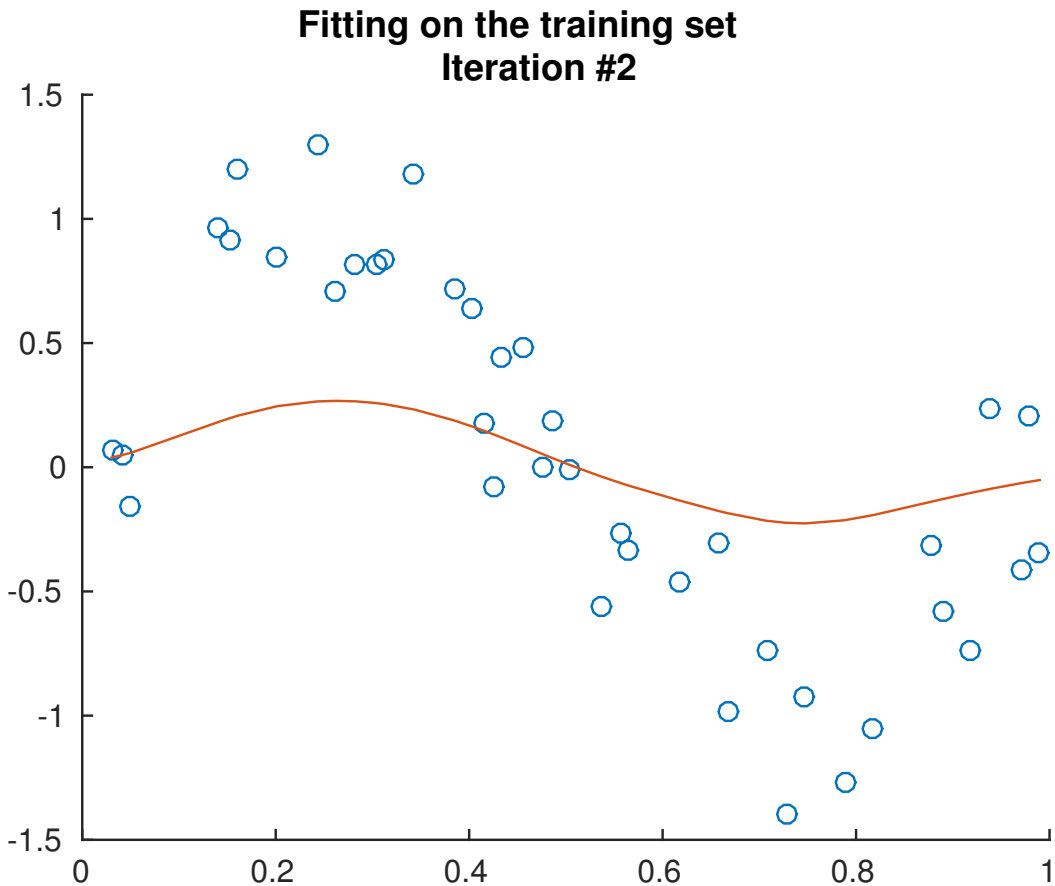
- does it define meaningful regularized solutions?

- Where is the regularization parameter?

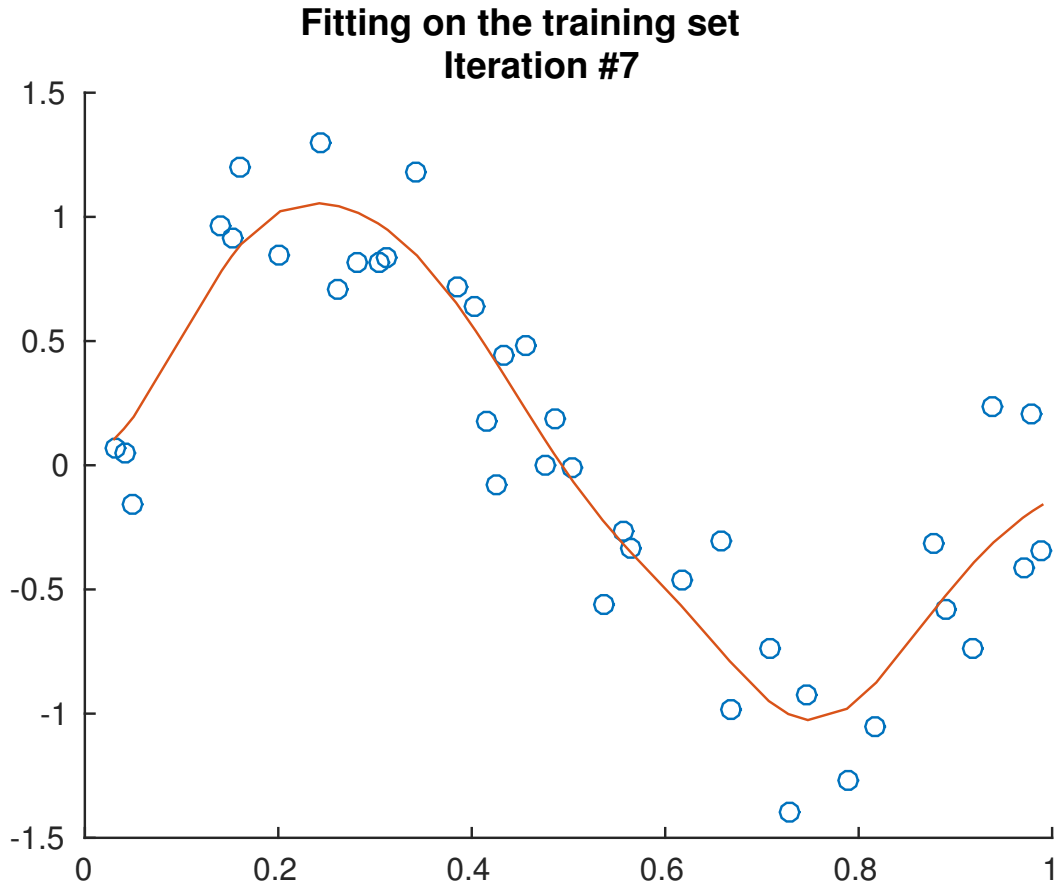
An intuition: early stopping



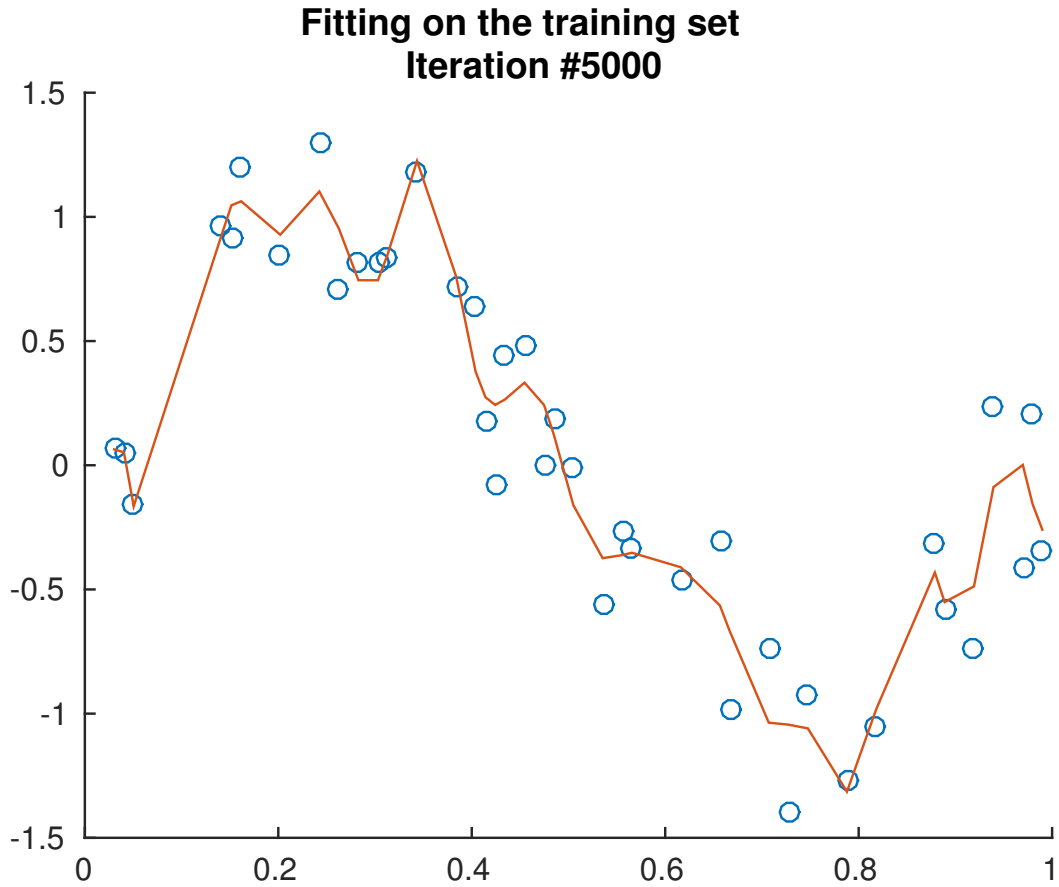
An intuition: early stopping



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 $\|A\| < 1$

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

Equivalently for B invertible[2] and $\|B\| < 1$

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

[2] Argument can be extended to pseudoinverses.

Rewriting GD

For $w_0 = 0$, by induction

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

can be written as

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

Rewriting GD (cont.)

- Write

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

- Assume

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

- Then

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

Neumann series and GD

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

GD is a truncated power series approximation of the pseudoinverse!

If γ is such that[3] $\left\| I - \frac{\gamma}{n} \widehat{X}^\top \widehat{X} \right\| < 1$, then for large t

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

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

[3] Compare to classic conditions.

Stability properties of GD

For any t

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

(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

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

For GD

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

Both methods can be seen as performing spectral filtering

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

for some suitable filter function F.

Implicit regularization and early stopping

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

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

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

Summary so far

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

- 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 now 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

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

Beyond linear least squares

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

Beyond linear LS with features

Consider $x \mapsto \Phi(x) \in \mathbb{R}^p$ and

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \ell'(\Phi(x_i)^\top \widehat{w}_t, y_i).$$

Iteration cost is $O(np)$ instead of $O(nd)$.

Beyond linear LS with kernels: representer

$$\widehat{w}_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} x_i \ell'(x_i^\top w, y_i).$$

It is easy to show that by induction that

$$\widehat{w}_t = \sum_{i=1}^{n} x_i (c_t)_i, \Leftrightarrow \hat{f}_t(x) = x^\top \widehat{w}_t = \sum_{i=1}^{n} x^\top x_i (c_t)_i$$

$$c_{t+1} = \widehat{w}_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} e_i \ell'(x_i^\top \widehat{w}_t, y_i),$$

Beyond linear LS with kernels

Let k be a postive definite kernel, then consder

$$\hat{f}_t(x) = \sum_{i=1}^{n} k(x, x_i)(c_t)_i$$

with

$$c_{t+1} = c_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} e_i \ell'(\hat{f}_t(x_i), y_i),$$

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.

Other forms of optimization

Largely unexplored, there are some results on:

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

Mostly for least squares.

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.

Summing up

- A different way to design algorithms.
- Implicit/iterative regularization.
- Iterative regularization for least squares.
- Extensions.