Class 4
Tikhonov Regularization

Carlo Ciliberto
Department of Computer Science, UCL

October 25, 2017

Last Class(es)

We have observed that key to to control the excess risk of an estimator is to limit the space of functions from which the estimator itself is chosen. In particular we have studied in detail how to control the *generalization error* of ERM on:

- Finite spaces (e.g. discretizations), or
- Compact spaces in $C(\mathcal{X})$ w.r.t. the $\|\cdot\|_\infty$ norm (under suitable assumptions on $\mathcal{X}$ and the loss).

Last Class(es)

We have considered the case where $\mathcal{H} = \bigcup_{\gamma \geq 0} \mathcal{H}_\gamma$ can be parametrized as the union of a family of spaces $\mathcal{H}_\gamma$ where the generalization error can be controlled uniformly (e.g. finite) and have decomposed the excess risk $\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_*)$ as

$$\underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)}_{\text{Sample error}} + \underbrace{\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)}_{\text{Approximation error}} + \underbrace{\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_*)}_{\text{Irreducible error}}$$

Where for any $\gamma \geq 0$

$$f_{\gamma,n} = \operatorname*{argmin}_{f \in \mathcal{H}_\gamma} \mathcal{E}_n(f) \qquad \text{and} \qquad f_\gamma = \operatorname*{argmin}_{f \in \mathcal{H}_\gamma} \mathcal{E}(f)$$

Last Class(es): Regularization

We controlled

- the sample error via bounds on the generalization error, e.g.

$$\mathcal{E}(f_{n,\gamma}) - \mathcal{E}(f_\gamma) \leq 2 \sup_{f \in \mathcal{H}_\gamma} |\mathcal{E}(f) - \mathcal{E}_n(f)| \leq \epsilon(n, \delta, \gamma)$$

  with probability $1 - \delta$ for any $\delta \in [0, 1)$.

- the approximation error by making assumption on the "regularity" of f_* (which depends on ρ)

$$\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq \mathcal{A}(\gamma, \rho)$$

We do not control the irreducible error since it depends by our initial choice of $\mathcal{H}$.

How to Choose $\mathcal{H}$ and the $\mathcal{H}_\gamma$?

We observed that performing ERM on a finite space of functions, albeit leading to good statistical performance, can become extremely expensive from the *computational* viewpoint (i.e. evaluating the empirical risk on *every* function in the space).

In the following of we will focus on (regularized) ERM over *linear* spaces of hypotheses $\mathcal{H}$. We consider the $\mathcal{H}_\gamma$ to be bounded convex subsets of $\mathcal{H}$, which are typically more amenable to computations while still allowing for good statistical performance.

Ivanov Regularization

Let $\mathcal{H}$ be a normed vector space of hypotheses (e.g. a reproducing kernel Hilbert space). For $\gamma \geq 0$ we consider

$$\mathcal{H}_\gamma = \{f \in \mathcal{H} \mid \|f\|_\mathcal{H} \leq \gamma\}$$

namely $\mathcal{H}_\gamma = B_\gamma(0) \subset \mathcal{H}$ are balls of radius γ in $\mathcal{H}$.

ERM on $\mathcal{H}_\gamma$ corresponds to

$$f_{\gamma,n} = \operatorname*{argmin}_{\|f\|_\mathcal{H} \leq \gamma} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

In particular, if $\ell(\cdot, y)$ is convex for any $y \in \mathcal{Y}$, ERM induces a *convex program* for which it is possible to find a minimizer in *polynomial time*.

Linear Functions

In this class we will focus on the case where $\mathcal{H}$ is a space of linear functions, namely: let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$ and

$$\mathcal{H} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \exists\, w \in \mathbb{R}^d,\ \text{s.t. } f(x) = x^\top w\ \forall x \in \mathbb{R}^d \right\}$$

with norm $\|f\|_{\mathcal{H}} = \|w\|_2$ and w the parameters corresponding to f.

What about Nonlinear Functions?

Studying the linear case is not limiting in that it can be naturally extended to richer spaces of functions by means of a collection (or *dictionary*) of nonlinear functions $\varphi_1, \ldots, \varphi_k : \mathbb{R}^d \to \mathbb{R}$ and

$$\mathcal{H} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \exists (w_i)_{i=1}^k \in \mathbb{R}, \text{ s.t. } f(x) = \sum_{i=1}^k \varphi_i(x) w_i \ \forall x \in \mathbb{R}^d \right\}.$$

We can still consider $\|f\|_{\mathcal{H}} = \|w\|_2$ with $w \in \mathbb{R}^k$ the vector with entries the w_is corresponding to f.

In this case $\mathcal{H}$ is a space of nonlinear functions on $\mathcal{X}$ but can also be interpreted as a space of linear functions on the image $\phi(\mathcal{X}) \subset \mathbb{R}^k$ of $\mathcal{X}$ via the *feature map* $\phi : \mathbb{R}^d \to \mathbb{R}^k$ with $\phi(x) = (\varphi_1(x), \ldots, \varphi_k(x))$.

ERM for Linear Functions

In the linear setting, ERM requires solving an optimization problem on $\mathbb{R}^d$

$$w_{n,\gamma} = \underset{\|w\|_2 \leq \gamma}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^\top w, y_i)$$

where the empirical risk minimizer $f_{\gamma,n} : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f_{\gamma,n}(x) = x^\top w_{\gamma,n} \qquad \forall x \in \mathbb{R}^d$$

Statistics and Computations...

We are faced with two questions:

- How to control the sample error of $\mathcal{H}_\gamma$?
- How to find $w_{\gamma,n}$ in practice?

Statistics and Computations...

We are faced with two questions:

- How to control the sample error of $\mathcal{H}_\gamma$?
- How to find $w_{\gamma,n}$ in practice?

Sample Error with Covering Numbers

Last class we observed that, for compact spaces $\mathcal{H}_\gamma \subset C(\mathcal{X})$ we could control the sample error of ERM on $\mathcal{H}_\gamma$ as

$$|\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)| \leq 2L\eta + 2\sqrt{\frac{2M^2 \log(2\mathcal{N}(\mathcal{H}_\gamma, \eta)/\delta)}{n}}$$

with probability $1-\delta$ for any $\delta \in [0,1)$. Where $L > 0$ is the Lipschitz constant of $\ell(\cdot, y)$ uniformly on $\mathcal{Y}$, and $M > 0$ is such that $\forall f \in \mathcal{H}_\gamma$, $|\ell(f(x), y)| \leq M$.

Note. $M = M(\gamma)$ depends on $\mathcal{H}_\gamma$ and could potentially increase as γ increases.

Covering Numbers of Balls in $\mathbb{R}^d$

The following provides estimates for the covering numbers of balls in $\mathbb{R}^d$

Theorem. For any $\gamma \geq 0$ and $B_\gamma(0) \subset \mathbb{R}^d$ the ball of radius γ centered in 0. Then

$$\mathcal{N}(B_\gamma(0), \eta) \leq \left(\frac{4\gamma}{\eta}\right)^d \qquad \forall \eta > 0$$

Since $\mathcal{H}$ is isometric to $\mathbb{R}^d$, we automatically have that

$$\mathcal{N}(\mathcal{H}_\gamma, \eta) \leq \left(\frac{4\gamma}{\eta}\right)^d \qquad \forall \eta > 0$$

as well.

Sample Error on $\mathcal{H}_\gamma$

Combining the covering number estimates with our previous bound on the sample error of ERM on $\mathcal{H}_\gamma$ and by choosing $\eta = \sqrt{\frac{M(\gamma)\log\gamma^d n/\delta}{n}}$, we can conclude that

$$|\mathcal{E}(w_{\gamma,n}) - \mathcal{E}(w_\gamma)| \leq O\left(\sqrt{\frac{M(\gamma)\log(\gamma^d n/\delta)}{n}}\right)$$

with probability not less than $1-\delta$.

Statistics and Computations...

We are faced with two questions:

- How to control the sample error of $\mathcal{H}_\gamma$?
- How to find $w_{\gamma,n}$ in practice?

Tikhonov Regularization

Instead of solving

$$w_{\gamma,n} = \operatorname*{argmin}_{\|w\|_{\mathcal{H}} \leq \gamma} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^\top w, y_i)$$

we will address the *Tikhonov regularization* problem

$$w_{\lambda,n} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^\top w, y_i) + \lambda \|w\|_{\mathcal{H}}^2$$

Indeed, it can be shown that for convex ℓ, the two problems are equivalent (for each γ there exists a $\lambda(\gamma)$ such that $w_{\gamma,n} = w_{\lambda(\gamma),n}$)

BUT Tikhonov regularization is an *unconstrained* optimization problem and therefore it is "easier" to design an optimization method to solve it.

Convex Optimization

Depending on the loss function ℓ we will be able to adopt different strategies to find the minimizer of the empirical risk.

- Closed form (e.g. Least squares loss).
- Iterative descent methods: Gradient Descent, Newton Method (smooth loss. E.g. logistic).
- Iterative methods: subgradient method (non-smooth loss. E.g. hinge).

Least Squares (a.k.a. Ridge Regression)

Let $\ell(f(x), y) = (y - f(x))^2$, then

$$w_{\lambda,n} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2 + \lambda \|w\|_2^2$$

or in vector notation

$$w_{\lambda,n} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \ \|y - Xw\|_2^2 + n\lambda \|w\|_2^2$$

with

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \qquad \text{and} \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Ridge Regression

For any *differentiable* convex function $F : \mathbb{R}^d \to \mathbb{R}$ we know that

$$w_* \in \mathbb{R}^d \text{ is a global minimizer for } F \qquad \Longleftrightarrow \qquad \nabla F(w_*) = 0$$

For ridge regression this characterization allows to recover the empirical risk estimator in closed form since

$$\nabla_w \|y - Xw\|_2^2 + n\lambda \|w\|_2^2 = 2X^\top Xw - 2X^\top y + 2n\lambda w = 0$$

if and only if

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

A Note on Computational Complexity

Note that in general for two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times p}$, the product $AB \in \mathbb{R}^{n \times p}$ requires $O(nmp)$ operations. Also, for a square invertible matrix $A \in \mathbb{R}^{n \times n}$ the cost of computing $A^{-1} \in \mathbb{R}^{n \times n}$ is $O(n^3)$.

Therefore, the cost of solving ridge regression is $O(nd^2 + d^3)$. In particular if $d > n$ this cost is $O(d^3)$.

Model selection. In practice we need to solve ridge regression for a number of candidate hyperparameters $\lambda_1, \ldots, \lambda_m$. So the total computational cost becomes $O(md^3)$.

Differentiable Loss Functions

In general, if $\ell(\cdot, y) : \mathbb{R} \to \mathbb{R}$ is differentiable for any $y \in \mathcal{Y}$, with $\ell'(t, y) = \frac{\partial}{\partial t}\ell(t, y)$, we have

$$\nabla \left(\mathcal{E}_n(w) + \lambda \|w\|_2^2 \right) = \frac{1}{n} \sum_{i=1}^{n} x_i \ell'(x_i^\top w, y_i) + 2\lambda w = 0$$

for which is not always possible to find a solution in close form.

In these settings we can resort to *iterative* descent optimization methods, which provide a sequence $(w^{(k)})_{k \in \mathbb{N}}$ that converge to the global minimizer.

Gradient Descent

Algorithm Let $F : \mathbb{R}^d \to \mathbb{R}$ differentiable and $w^{(0)} \in \mathbb{R}^d$. For any $k \in \mathbb{N}$ we define $w^{(k+1)} \in \mathbb{R}^d$ as

$$w^{(k+1)} = w^{(k)} - \sigma \nabla F(w^{(k)})$$

where $\sigma > 0$ represents the *step size* of the descent.

Assumption: Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function with Lipschitz gradient

$$\|\nabla F(w) - \nabla F(w')\|_2 \leq L\|w - w'\|_2 \qquad \forall w, w' \in \mathbb{R}^d$$

Theorem. Let $\sigma = 1/L$, then

$$F(w^{(k)}) - F(w_*) \leq \frac{L}{2k}\|w_*\|_2$$

Projected Gradient Descent

Problems such as Ivanov regularization,

$$\underset{\|w\|_2 \leq \gamma}{\text{minimize}} \; F(w)$$

which are constrained on a convex set can be solved by *projected gradient descent*. Namely, let $w^{(0)} \in \mathbb{R}^d$,

$$w^{(k+1)} = \Pi_{\mathcal{H}_\gamma}(w^{(k)} - \sigma \nabla F(w^{(k)}))$$

where $\Pi_{\mathcal{H}_\gamma} : \mathbb{R}^d \to \mathbb{R}^d$ denotes the euclidean projection onto $\mathcal{H}_\gamma$, namely

$$\Pi_{\mathcal{H}_\gamma}(w) = \underset{w' \in \mathcal{H}_\gamma}{\text{argmin}} \, \|w - w'\|_2^2 = \gamma \frac{w}{\|w\|_2}$$

Note. Projected gradient descent enjoys the same convergence rates of standard gradient descent.

Sample Error and Iterative Methods

An iterative optimization method provides vectors $w^{(k)}$ that get closer to $w_{\gamma,n}$ but we are not necessarily recovering it exactly.

Thus, if we decompose the sample error for the estimator after k iterations, we have

$$\begin{aligned}
\mathcal{E}(w^{(k)}) - \mathcal{E}(w_\gamma) &= \\
&= \mathcal{E}(w^{(k)}) - \mathcal{E}_n(w^{(k)}) + \mathcal{E}_n(w^{(k)}) - \mathcal{E}_n(w_{\gamma,n}) \\
&+ \underbrace{\mathcal{E}_n(w_{\gamma,n}) - \mathcal{E}_n(w_\gamma)}_{\leq 0} + \mathcal{E}_n(w_\gamma) - \mathcal{E}(w_\gamma) \\
&\leq \underbrace{\mathcal{E}(w^{(k)}) - \mathcal{E}_n(w^{(k)})}_{\text{sample error on } \mathcal{H}_\gamma} + \underbrace{\mathcal{E}_n(w^{(k)}) - \mathcal{E}_n(w_{\gamma,n})}_{\text{optimization error}} + \underbrace{\mathcal{E}_n(w_\gamma) - \mathcal{E}(w_\gamma)}_{\text{sample error on } \mathcal{H}_\gamma}
\end{aligned}$$

Differently from ERM, we are left with an *optimization error* term...

Sample Error and Optimization Error

$$\underbrace{\mathcal{E}(w^{(k)}) - \mathcal{E}_n(w^{(k)})}_{\text{generalization error on } \mathcal{H}_\gamma} + \underbrace{\mathcal{E}_n(w^{(k)}) - \mathcal{E}_n(w_{\gamma,n})}_{\text{optimization error}} + \underbrace{\mathcal{E}_n(w_\gamma) - \mathcal{E}(w_\gamma)}_{\text{generalization error on } \mathcal{H}_\gamma}$$

- We already know how the generalization error(s) can be controlled uniformly on $\mathcal{H}_\gamma$, leading to an error smaller than $\epsilon(n, \gamma, \delta)$ with probability no less than $1 - \delta$ (e.g. via Covering Numbers, next class: via "stability").

- We know that the optimization error decreases as $O(1/k)$.

Therefore it is sufficient to perform $k = O\left(\frac{1}{\epsilon(n,\gamma,\delta)}\right)$ iterations to attain the same accuracy of ERM.

Computational Complexity of Gradient Descent

Consider the computational complexity of gradient descent applied to ridge regression.

We have that for any $k \in \mathbb{N}$ a gradient step entails

$$w^{(k+1)} = w^{(k)} - \sigma(X^\top X + \lambda I)w^{(k)} + \sigma X^\top y$$

which requires $O(nd^2)$ to evaluate $X^\top X$ and $X^\top y$ (once) plus $O(d^2)$ for $(X^\top X + \lambda I)w^{(k)}$ (at each iteration).

This leads to a total of $O((k+n)d^2)$ operations for k gradient steps.

Benefits of Iterative Methods

Therefore, if $\epsilon(n, \gamma, \delta) \geq 1/n$, it is sufficient to perform $k = O(n)$ iterations to achieve the same excess risk as ERM leading to a *total cost* of $O(nd^2)$ operations.

We observed that the closed form solution of ridge regression requires $O(nd^2 + d^3)$ operations, which is dominated by $O(d^3)$ if $n < d$.

Therefore *stopping* the iterative method before convergence to the ERM solution can be potentially more appealing from the computational perspective without degrading statistical performances!

Appendix: Convergence Rates for Gradient Descent

We now prove the convergence rate for gradient descent when $F : \mathbb{R}^d \to \mathbb{R}$ is a differentiable convex function with Lipschitz gradient

$$\|\nabla F(w) - \nabla F(w')\|_2 \leq L\|w - w'\|_2 \qquad \forall w, w' \in \mathbb{R}^d$$

Step 1 (Gradient descent... descends!). We begin by showing that indeed for a suitable choice of step size the iterates $w^{(k)}$ of gradient descent are such that $F(w^{(k)}) > F(w^{(k+1)})$ for any $k \in \mathbb{N}$.

Step 2 (Rates) By studying the cumulative improvement of each such descent step, we derive the rate of convergence for the whole algorithm.

Appendix: Gradient Descent... Descends

Consider the function $h(t) = F(w + t(w - w'))$. We know that h is differentiable and that $\int_0^1 h'(t) = h(1) - h(0)$. In particular, since

$$h'(t) = \frac{\partial}{\partial t} F(w + t(w - w')) = (w - w')^\top \nabla F(w + t(w - w'))$$

we have

$$\begin{aligned}
F(w') &= F(w) + \int_0^1 (w - w')^\top \nabla F(w + t(w - w'))\, dt \\
&= F(w) + (w - w')^\top \nabla F(w) + \int_0^1 (w - w')^\top (F(w + t(w - w')) - F(w)) \\
&\leq F(w) + (w - w')^\top \nabla F(w) + L\|w - w'\|^2 \int_0^1 t dt \\
&= F(w) + (w - w')^\top \nabla F(w) + \frac{L}{2}\|w - w'\|^2
\end{aligned}$$

Appendix: Gradient Descent... Descends

In particular $F(w') \leq F(w) + (w - w')^\top \nabla F(w) + \frac{L}{2}\|w - w'\|^2$

If we plug the definition of gradient descent $w^{(k+1)} = w^{(k)} - \sigma \nabla F(w^{(k)})$ we have

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \sigma(1 - \frac{L}{2}\sigma)\|\nabla F(w^{(k)}\|^2$$

which implies that $F(w^{(k+1)}) < F(w^{(k)})$ whenever $0 < \sigma < \frac{2}{L}$. In particular, if $\sigma = \frac{1}{L}$ (which guarantees the maximum decrease)

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \frac{1}{2L}\|\nabla F(w^{(k)})\|^2$$

Appendix: Convergence Rates for Gradient Descent (Continued)

Leveraging the previous result:

$$
\begin{aligned}
F(w^{(k+1)}) &\leq F(w^{(k)}) - \frac{1}{2L}\|\nabla F(w^{(k)})\|^2 \\
&\leq F(w_*) + (w^{(k)} - w_*)\nabla F(w^{(k)}) - \frac{1}{2L}\|\nabla F(w^{(k)})\|^2 \qquad \text{(convexity)} \\
&= F(w_*) + \frac{L}{2}\left(\frac{2}{L}(w^{(k)} - w_*)\nabla F(w^{(k)}) - \frac{1}{L^2}\|\nabla F(w^{(k)})\|^2 \pm \|w^{(k)} - w_*\|^2\right) \\
&= F(w_*) + \frac{L}{2}\left(\|w^{(k)} - w_*\|^2 - \|w^{(k)} - \nabla F(w^{(k)}) - w_*\|^2\right) \\
&= F(w_*) + \frac{L}{2}\left(\|w^{(k)} - w_*\|^2 - \|w^{(k+1)} - w_*\|^2\right)
\end{aligned}
$$

In particular $F(w^{(k+1)}) - D(w_*) \leq \frac{L}{2}\left(\|w^{(k)} - w_*\|^2 - \|w^{(k+1)} - w_*\|^2\right)$

Appendix: Convergence Rates for Gradient Descent (Continued)

Therefore we have that for any $K \in \mathbb{N}$

$$\sum_{k=1}^{K} F(w_k) - F(w_*) \leq \frac{L}{2} \sum_{k=1}^{K} \left(\|w_{k-1} - w_*\|^2 - \|w^{(k)} - w_*\|^2 \right)$$
$$= \frac{L}{2} \left(\|w_0 - w_*\|^2 - \|w_K - w_*\|^2 \right) \leq \frac{L}{2} \|w_0 - w_*\|^2$$

In particular, since $F(w^{(k)})$ is a decreasing sequence we have

$$K(F(w_K) - F(w_*)) \leq \sum_{k=1}^{K} F(w_k) - F(w_*) \leq \frac{L}{2} \|w_0 - w_*\|^2$$

From which we conclude $F(w_K) - F(w_*) \leq \frac{L}{2K} \|w_0 - w_*\|^2$ as desired