Class 6
Early Stopping

Carlo Ciliberto
Department of Computer Science, UCL

November 16, 2018

Computational Regularization

So far we have mostly focused on studying and characterizing the generalization (i.e. statistical) properties of a learning algorithm.

We have observed that good learning rates can be achieved by: $i)$ *limiting* the expressiveness of the hypotheses class from which the estimator is obtained in order to avoid *overfitting* and then, $ii)$ increasing such expressiveness only as we see more training points.

Computational regularization is a paradigm that aims to implement regularization (i.e. control the expressiveness of an hypotheses class) by limiting the computational resources available to the algorithm. The idea is that in this way we can obtain an algorithm that is both statistically efficient *and* much faster to train.

Computational Regularization: Early Stopping

In this class will consider an instance of computational regularization known as *early stopping*.

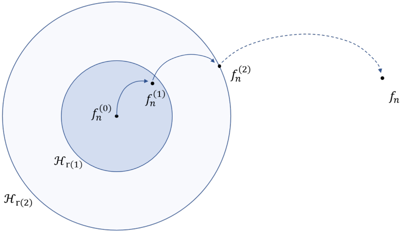
Intuition: Consider an iterative method such as gradient descent applied to *unregularized* ERM with n training points. Let f_n be the ERM solution and $(f_n^{(t)})_{t\in\mathbb{N}}$ the sequence of iterates obtained by gradient descent (for instance starting from $f_n^{(0)} = 0$).

Clearly, the first few iterates (i.e. for small values of t) will not fit well the training data, while for t that grows to infinity, the predictors $f_n^{(t)}$ will get closer to the ERM solution f_n and will start to overfit.

Early stopping aims to find the sweet spot between performing "too many" and "too few" iterations.

Early Stopping: Further Intuition

Every step of gradient descent allows to move from the previous iterate only by a certain amount (i.e. $f_n^{(t)} \in \mathcal{H}_{r(t)}$ for some radius $r(t)$).



Therefore, intuitively, in early stopping the number t of iterations takes the role of regularization parameter. Indeed, similarly to what we observed for Ivanov/Tikhonov regularization, t *controls the expressiveness of the class of estimators* $f_n^{(t)}$.

Early Stopping: Further Intuition

Let's make the above intuition more rigorous:

Lemma. Let $F : \mathcal{H} \to \mathbb{R}$ be L-Lipschitz, convex and differentiable. Then,

$$\|\nabla F(f)\| \leq L \qquad \forall f \in \mathcal{H}.$$

Therefore, at step t of gradient descent on F, with step-size $\gamma > 0$

$$\|f_t\|_{\mathcal{H}} = \|f_{t-1} - \gamma \nabla F(f_{t-1})\|_{\mathcal{H}} \leq \|f_{t-1}\|_{\mathcal{H}} + \gamma L \leq t\gamma L.$$

Namely, after t steps, we are at least in a ball $\mathcal{H}_{r(t)}$ of radius $r(t) = t\gamma L$.

Early Stopping and Generalization Error

Let us assume $f_* \in \mathcal{H}$. Consider the decomposition of the expected risk $\mathcal{E}(f_S^{(T)}) - \mathcal{E}(f_*)$ as

$$\underbrace{\mathcal{E}(f_S^{(T)}) - \mathcal{E}_S(f_S^{(T)})}_{\text{Generalization error}} + \underbrace{\mathcal{E}_S(f_S^{(T)}) - \mathcal{E}_S(f_*)}_{\text{Optimization error}} + \mathcal{E}_S(f_*) - \mathcal{E}(f_*)$$

Note that:

- Optimization error: $\mathcal{E}_S(f_S^{(T)}) - \mathcal{E}_S(f_*) \leq O(1/T)$, and
- $\mathbb{E}_S \ \mathcal{E}_S(f_*) - \mathcal{E}(f_*) = 0$

We are left with studying the generalization error of the estimator $f_S^{(T)}$. In this class we will address this question (in expectation only) in terms of the *stability* of gradient descent.

Refresher on Stability

Notation. Let $S = (z_i)_{i=1}^n \in \mathcal{Z}^n$ with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and consider the empirical risk $\mathcal{E}_S(f) = \frac{1}{n} \sum_{i=1} \ell(f, z_i)$ for a function $f : \mathcal{X} \to \mathcal{Y}$. We denote $\ell(f, z_i) = \ell(f(x_i), y_i)$ where $z_i = (x_i, y_i) \in \mathcal{Z}$.

Recall that an algorithm $\mathcal{A} : S \mapsto f_S$ is said uniform $\beta(n)$-stable if, for any $S \in \mathcal{Z}^n$, $z \in \mathcal{Z}$ and $i = 1, \dots, n$

$$\sup_{\bar{z} \in \mathcal{Z}} \; |\ell(f_S, \bar{z}) - \ell(f_{S^{i,z}}, \bar{z})| \leq \beta(n)$$

where $S^{i,z}$ is the set obtained by substituting z_i with z in S.

In last class we observed that stability directly implies bounds on the generalization error, namely

$$|\mathbb{E}_{S \sim \rho^n} \; [\mathcal{E}(f_S) - \mathcal{E}_S(f_S)]| \leq \beta(n)$$

Stability of Gradient Descent

In this class we will prove the following characterization for the stability of gradient descent.

Theorem. Let $\ell(\cdot, y) : \mathcal{H} \to \mathbb{R}$ be convex, L-Lipschitz and M-smooth uniformly for $y \in \mathcal{Y}$. For any training set $S \in \mathcal{Z}^n$, let $f_S^{(T)}$ be obtained by applying gradient descent with steps-size $\gamma = 1/M$ on the empirical risk associated to S. The corresponding algorighm is $\beta(n, T)$-stable with

$$\beta(n, T) \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

Learning Rates for Early Stopping

We can therefore bound the excess risk as

$$\mathcal{E}(f_S^{(T)}) - \mathcal{E}(f_*) \leq O\left(\frac{T}{n} + \frac{1}{T}\right)$$

leading to the optimal choice for the number of iterations to be of the order $T(n) = O(\sqrt{n})$.

Note the similarity with the bound of the excess risk for Tikhonov regularization

$$\mathcal{E}(f_{S,\lambda}) - \mathcal{E}(f_*) \leq O\left(\frac{1}{n\lambda} + \lambda\right).$$

Again, we see how T playes the role of $1/\lambda$ and can indeed be interpreted as a regularization parmeter.

Benefits of Early Stopping

While achieving the same statistical performance as Tikhonov regularization

$$\mathcal{E}(f_S^{(T(n))}) - \mathcal{E}(f_*) \leq O(1/\sqrt{n}).$$

early stopping can often be advantageous from the computational perspective.

For instance, in Tikhonov/Ivanov regularization, one has to find the regularization parmeter λ by cross-validation over a set of candidate values $\lambda_1, \ldots, \lambda_m$. This requires solving m optimization problems (for instance by gradient descent).

On the contrary, the hyperparameter of early stopping *is* the number of iterations. This means that we need to run only one instance of gradient descent.

Stability of Gradient Descent (Auxiliary Results)

We now prove the stability of Gradient Descent.

Theorem. Let $\ell(\cdot, y) : \mathcal{H} \to \mathbb{R}$ be convex, L-Lipschitz and M-smooth uniformly for $y \in \mathcal{Y}$. For any training set $S \in \mathcal{Z}^n$, let $f_S^{(T)}$ be obtained by applying gradient descent with steps-size $\gamma = 1/M$ on the empirical risk associated to S. The corresponding algorighm is $\beta(n, T)$-stable with

$$\beta(n, T) \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

However, before doing that we need some auxiliary results.

Stability of Gradient Descent (Auxiliary Result I)

Lemma. $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth with minimizer $w_* \in \mathcal{H}$. Then

$$F(w) - F(w_*) \geq \frac{1}{2M} \|\nabla F(w)\|_{\mathcal{H}}^2 \qquad \forall w \in \mathcal{H}$$

Proof. From a previous class (Lec 4) we know that for any $v, w \in \mathcal{H}$

$$F(v) \leq F(w) + \langle \nabla F(w), v - w \rangle_{\mathcal{H}} + \frac{M}{2} \|w - v\|_{\mathcal{H}}^2$$

By minimizing the left and right sides w.r.t. $v \in \mathcal{H}$, we have

$$F(w_*) \leq \inf_{v \in \mathcal{H}} F(w) + \langle \nabla F(w), v - w \rangle_{\mathcal{H}} + \frac{M}{2} \|w - v\|_{\mathcal{H}}^2 = F(w) - \frac{1}{2M} \|\nabla F(w)\|_{\mathcal{H}}^2$$

Which yields the desired result. (Note that the minimizer of the quadratic upper bound is indeed given by $v = w - \frac{1}{M} \nabla F(w)$).

Co-coercivity of the Gradient (Auxiliary Result II)

Proposition. $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth. Then $\forall v, w \in \mathcal{H}$

$$\langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \ \geq \ \frac{1}{M} \| \nabla F(w) - \nabla F(v) \|_{\mathcal{H}}^2$$

Proof. Define

$$F_w(z) = F(z) - \langle \nabla F(w), z \rangle_{\mathcal{H}} \qquad \text{and} \qquad F_v(z) = F(z) - \langle \nabla F(v), z \rangle_{\mathcal{H}}.$$

It is trivial to verify that F_w and F_v are M-smooth as well.

Moreover w and v are the minimizers of respectively F_w and F_v since

$$\nabla F_w(z) = \nabla F(z) - \nabla F(w) = 0 \iff z = w.$$

Therefore we can apply the previous Lemma.

Co-coercivity of the Gradient (Continued)

By applying the Lemma we have

- $\frac{1}{2M}\|\nabla F_w(v)\|_{\mathcal{H}}^2 \leq F_w(v) - F_w(w) = F(v) - F(w) - \langle \nabla F(w), v - w\rangle_{\mathcal{H}}$
- $\frac{1}{2M}\|\nabla F_v(w)\|_{\mathcal{H}}^2 \leq F_v(w) - F_v(v) = F(w) - F(v) - \langle \nabla F(v), w - v\rangle_{\mathcal{H}}$

Since $\|\nabla F_w(v)\|_{\mathcal{H}} = \|\nabla F_v(w)\|_{\mathcal{H}} = \|\nabla F(w) - \nabla F(w)\|_{\mathcal{H}}$, by summing the two inequalities we have

$$\frac{1}{M}\|\nabla F(w) - \nabla F(w)\|_{\mathcal{H}}^2 \leq \langle \nabla F(v) - \nabla F(w), v - w\rangle_{\mathcal{H}}$$

as desired.

Gradient Descent is Non-expansive (Aux. Res. III)

Theorem. $\ell : \mathcal{H} \to \mathbb{R}$ convex, differentiable and M-smooth. Let $0 \geq \gamma \geq 2/M$ and $G : \mathcal{H} \to \mathcal{H}$ be the gradient step operator $G(f) = f - \gamma \nabla \ell(f)$ for $f \in \mathcal{H}$. Then

$$\|G(f) - G(g)\|_{\mathcal{H}} \leq \|f - g\|_{\mathcal{H}}$$

Proof. By applying the co-coercivity of a convex M-smooth loss, we have

$$\begin{aligned}
\|G(f) - G(g)\|_{\mathcal{H}}^2 &= \|f - \gamma \nabla \ell(f) - g + \gamma \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&= \|f - g\|_{\mathcal{H}}^2 - 2\gamma \langle \nabla \ell(f) - \nabla \ell(g), f - g \rangle_{\mathcal{H}} + \gamma^2 \|\nabla \ell(f) - \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&\leq \|f - g\|_{\mathcal{H}}^2 - \gamma(\frac{2}{M} - \gamma)\|\nabla \ell(f) - \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&\leq \|f - g\|_{\mathcal{H}}^2
\end{aligned}$$

since $\gamma(\frac{2}{M} - \gamma) \leq 1$ for $\gamma \in [0, 2/M]$. This implies the desired result.

Stability of Gradient Descent (Proof)

We can now prove the stability of gradient descent.

Proof. Let $S \in \mathcal{Z}^n$, $z \in \mathcal{Z}$ and $i \in \{1, \dots, n\}$. To simplify the notation, let us denote f_t (instead of $f_n^{(t)}$) the t-th iterate of gradient descent with step γ on S. Similarly, denote with $f'_t \in \mathcal{H}$ the t-th iterate of gradient descent with step γ on $S^{i,z}$.

Given a total number of iterations T, we want to control

$$\sup_{\bar{z} \in \mathcal{Z}} |\ell(f_T, \bar{z}) - \ell(f'_T, \bar{z})| \leq Lk \|f_T - f'_T\|_{\mathcal{H}}$$

Stability of Gradient Descent (Continued)

For any $t \in \mathbb{N}$, by construction $f_{t+1} = f_t - \gamma \nabla \mathcal{E}_S(f_t)$ and $f'_{t+1} = f_t - \gamma \nabla \mathcal{E}_{S^{i,z}}(f'_t)$. Therefore

$$\begin{aligned}
\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}} &= \left\| f_t - f'_t - \frac{\gamma}{n} \sum_{j \neq i} \left[\nabla \ell(f_t, z_j) - \nabla \ell(f'_t, z_j) \right] - \frac{\gamma}{n} \left[\nabla \ell(f_t, z_i) - \nabla \ell(f'_t, z) \right] \right\|_{\mathcal{H}} \\
&\leq \frac{1}{n} \sum_{j \neq i} \left\| f_t - \gamma \nabla \ell(f_t, z_j) - f'_t + \gamma \nabla \ell(f'_t, z_j) \right\|_{\mathcal{H}} \\
&\quad + \frac{1}{n} \|f_t - f'_t\|_{\mathcal{H}} + \frac{\gamma}{n} (\|\nabla \ell(f_t, z_i)\|_{\mathcal{H}} + \|\nabla \ell(f'_t, z)\|_{\mathcal{H}})
\end{aligned}$$

Recall that for $\gamma \in [0, 2/M]$, the gradient descent step $f - \gamma \nabla \ell(f, z)$ is non-expansive for any $f \in \mathcal{H}$ and $z \in \mathcal{Z}$. Therefore, for any $j \neq i$,

$$\left\| f_t - \gamma \nabla \ell(f_t, z_j) - f'_t + \gamma \nabla \ell(f'_t, z_j) \right\|_{\mathcal{H}} \leq \|f_t - f'_t\|_{\mathcal{H}}$$

Stability of Gradient Descent (Continued)

For the remaining terms, note that since ℓ is L-Lipschitz

$$\|\nabla \ell(f_t, z_j)\| \leq Lk \qquad \text{and} \qquad \|\nabla \ell(f_t', z)\| \leq Lk$$

Therefore, we have

$$\|f_{t+1} - f_{t+1}'\|_{\mathcal{H}} \leq \|f_t - f_t'\|_{\mathcal{H}} + \frac{2Lk}{n}\gamma = \frac{2Lk}{M}\frac{t+1}{n}$$

Finally, iterating on all $t = 1, \ldots, T$, we have

$$\sup_{\bar{z} \in \mathcal{Z}} |\ell(f_T, \bar{z}) - \ell(f_T', \bar{z})| \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

which concludes our proof.