Class 5
Stability

Carlo Ciliberto
Department of Computer Science, UCL

November 3, 2017

Uniform Stability - Notation

Let $\mathcal{Z}$ be a set. For any set

$$S = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$$

and for any $z \in \mathcal{Z}$ and $i = 1, \ldots, n$, we denote

$$S^{i,z} = \{z_1, \ldots, z_{i-1}, z, z_{i+1}, \ldots, z_n\} \in \mathcal{Z}^n$$

the set obtained by substituting the i-th element in S with z.

Uniform Stability

We denote input-output pairs as $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and for any $f : \mathcal{X} \to \mathcal{Y}$ we denote $\ell(f, z) = \ell(f(x), y)$.

For an algorithm $\mathcal{A}$ and for any dataset $S = (z_i)_{i=1}^n$ we write $f_S = \mathcal{A}(S)$.

Uniform β-Stability. An algorithm $\mathcal{A}$ is $\beta(n)$-stable with $n \in \mathbb{N}$ and $\beta(n) > 0$, if for all $S \in \mathcal{Z}^n$, $z \in \mathcal{Z}$ and $i = 1, \dots, n$

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

Stability and Generalization Error

Theorem. Let $\mathcal{A}$ be a uniform $\beta(n)$-stable algorithm. For any dataset $S \in \mathcal{Z}^n$ denote $f_S = \mathcal{A}(S)$. Then

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

where $S \sim \rho^n$ denotes a random dataset with n points sampled independently from ρ.

The above result shows that uniform stability of an algorithm allows to directly control its *generalization error*.

Note that this result relies only on the properties of the learning algorithm but does not require any knowledge about the complexity of the hypotheses space (however it is indirectly related).

Stability and Generalization Error (Continued)

We begin by providing alternative formulation for:

1) The expectation of the empirical risk $\mathbb{E}_S[\ \mathcal{E}_S(f_S)\ ]$

$$
\begin{aligned}
\mathbb{E}_S[\ \mathcal{E}_S(f_S)\ ] &= \mathbb{E}_S\left[\ \frac{1}{n}\sum_{i=1}^{n}\ell(f_S, z_i)\ \right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[\ \ell(f_S, z_i)\ ] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S\mathbb{E}_{z_i'}[\ \ell(f_S, z_i)\ ] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S\mathbb{E}_{z_i'}[\ \ell(f_{S^{i,z_i'}}, z_i')\ ] = \mathbb{E}_S\mathbb{E}_{S'}\left[\ \frac{1}{n}\sum_{i=1}^{n}\ell(f_{S^{i,z_i'}}, z_i')\ \right]
\end{aligned}
$$

2) The expected risk $\mathcal{E}(f_S)$

$$
\mathcal{E}(f_S) = \mathbb{E}_{z'}\ell(f_S, z') = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z'}\ell(f_S, z') = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i'}\ell(f_S, z_i') = \mathbb{E}_{S'}\left[\ \frac{1}{n}\sum_{i=1}^{n}\ell(f_S, z_i')\ \right]
$$

Stability and Generalization Error (Continued)

Putting the two together

$$\begin{aligned} \left| \; \mathbb{E}_S \left[\mathcal{E}(f_S) - \mathcal{E}_n(f_S) \right] \; \right| &\leq \left| \mathbb{E}_S \mathbb{E}_{S'} \left[\; \frac{1}{n} \sum_{i=1}^{n} \ell(f_{S^{i,z_i'}}, z_i') - \ell(f_S, z_i') \; \right] \; \right| \\ &\leq \mathbb{E}_S \mathbb{E}_{S'} \; \frac{1}{n} \sum_{i=1}^{n} \left| \; \ell(f_{S^{i,z_i'}}, z_i') - \ell(f_S, z_i') \; \right| \leq \beta(n) \end{aligned}$$

Stability of Tikhonov Regularization

In the following we focus on the Tikhonov regularization algorithm $\mathcal{A} = \mathcal{A}_\lambda$ with $\lambda > 0$. In particular, for any $S \in \mathcal{Z}^n$

$$\mathcal{A}(S) = f_S = \operatorname*{argmin}_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(f, z_i) + \lambda \|f\|_{\mathcal{H}}^2$$

We will show that when $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS), Tikhonov regularization is $\beta(n)$-stable with

$$\beta(n) = O\left(\frac{1}{n\lambda}\right)$$

Error Decomposition for Tikhonov Regularization

Define $f_\lambda = \operatorname{argmin}_{f \in \mathcal{H}}$ We can decompose the excess risk as

$$\mathcal{E}(f_S) - \mathcal{E}(f_*) = \mathcal{E}(f_S) \pm \mathcal{E}_S(f_S) \pm \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_*) \pm \lambda \|f_\lambda\|_{\mathcal{H}}^2$$

Moreover, since

- $\mathcal{E}(f_S) - \mathcal{E}(f_*) \leq \mathcal{E}(f_S) - \mathcal{E}(f_*) + \lambda \|f_S\|_{\mathcal{H}}^2$,
- f_S is the minimizer of the regularized empirical risk $\mathcal{E}_n(f_S) + \lambda \|f_S\|_{\mathcal{H}}^2 - \mathcal{E}_n(f_\lambda) - \lambda \|f_\lambda\|_{\mathcal{H}}^2 \leq 0$,
- $\mathbb{E}_S \ \mathcal{E}_S(f_\lambda) = \mathcal{E}(f_\lambda)$

We can conclude

$$\mathbb{E}_S \ \mathcal{E}(f_S) - \mathcal{E}(f_*) \leq \mathbb{E}_S \ [\mathcal{E}(f_S) - \mathcal{E}_n(f_S)] + \mathcal{E}(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2$$

Error Decomposition for Tikhonov Regularization

$$\mathbb{E}_S\ \mathcal{E}(f_S) - \mathcal{E}(f_*) \leq \underbrace{\mathbb{E}_S\ [\mathcal{E}(f_S) - \mathcal{E}_n(f_S)]}_{\text{Generalization Error}} + \underbrace{\mathcal{E}(f_\lambda) - \mathcal{E}(f_*) + \lambda\|f_\lambda\|^2_{\mathcal{H}}}_{\substack{\text{(related to) Interpolation Error} \\ \text{and Approximation Error}}}$$

Stability of Tikhonov regularization $O(1/(n\lambda))$ + assuming the interpolation/approximation error to be bounded by λ^s with $s > 0$ lead to

$$\mathbb{E}_S\ \mathcal{E}(f_S) - \mathcal{E}(f_*) \leq O(1/(n\lambda)) + \lambda^s$$

We can choose the optimal $\lambda(n)$ and (expected) error rates $\epsilon(n)$ as

$$\lambda(n) = O(n^{-\frac{1}{s+1}}) \qquad \mathbb{E}_S\ \mathcal{E}(f_S) - \mathcal{E}(f_*) \leq O(n^{-\frac{s}{s+1}})$$

Note. If $f_* \in \mathcal{H}$ it is easy to show that $s = 1$ and therefore that the expected excess risk goes to zero at least as $O(n^{-1/2})$.

Stability of Tikhonov Regularization

Let $\mathcal{H}$ be a RKHS with associated kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We want to show that for any $S \in \mathcal{Z}^n$, $z' \in \mathcal{Z}$ and $i = 1, \dots, n$

$$\sup_{z \in \mathcal{Z}} \ |\ell(f_S, z) - \ell(f_{S^{i,z'}}, z)| \leq \frac{2L^2 k^2}{n\lambda}$$

where $L > 0$ is the Lipschitz constant of $\ell(\cdot, y)$ (uniformly w.r.t. $y \in \mathcal{Y}$) and $k^2 = \sup_{x \in \mathcal{X}} k(x, x)$.

Reproducing Property

Recall the reproducing property of RKHS $\mathcal{H}$: $\forall f \in \mathcal{H}$, $\forall x \in \mathcal{X}$

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$$

In particular, $|f(x)| \leq \sqrt{k(x,x)} \|f\|_{\mathcal{H}}$.

Therefore,

$$\begin{aligned} \sup_{z \in \mathcal{Z}} |\ell(f_S, z) - \ell(f_{S^{i,z'}}, z)| \leq & \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |\ell(f_S(x), y) - \ell(f_{S^{i,z'}}(x), y)| \\ & \leq L \sup_{x \in \mathcal{X}} |f_S(x) - f_{S^{i,z'}}(x)| \leq Lk \|f_S - f_{S^{i,z'}}\|_{\mathcal{H}} \end{aligned}$$

We need to control $\|f_S - f_{S^{i,z'}}\|_{\mathcal{H}}$. We will exploit the *strong convexity* of Tikhonov regularization.

Strong convexity of $\|\cdot\|_{\mathcal{H}}^2$

Technical observation. For any $f, g \in \mathcal{H}$ and $\theta \in [0,1]$ we have

$$\begin{aligned}
\|\theta f + (1-\theta)g\|_{\mathcal{H}}^2 &= \theta^2 \|f\|_{\mathcal{H}}^2 = (1-\theta)^2 \|g\|_{\mathcal{H}}^2 + 2\theta(1-\theta)\langle f, g\rangle_{\mathcal{H}} \\
&= \theta(1-(1-\theta))\|f\|_{\mathcal{H}}^2 + (1-\theta)(1-\theta)\|g\|_{\mathcal{H}}^2 + 2\theta(1-\theta)\langle f, g\rangle_{\mathcal{H}} \\
&= \theta\|f\|_{\mathcal{H}}^2 + (1-\theta)\|g\|_{\mathcal{H}}^2 - \theta(1-\theta)(\|f\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{H}}^2 - 2\langle f, g\rangle_{\mathcal{H}}) \\
&= \theta\|f\|_{\mathcal{H}}^2 + (1-\theta)\|g\|_{\mathcal{H}}^2 - \theta(1-\theta)\|f-g\|_{\mathcal{H}}^2
\end{aligned}$$

In particular, for any $F' : \mathcal{H} \to \mathbb{R}$ convex, if we denote
$F(\cdot) = F'(\cdot) + \lambda\|\cdot\|^2$, we have

$$F(\theta f + (1-\theta)g) \leq \theta F(f) + (1-\theta)F(g) - 2\lambda\theta(1-\theta)\|f-g\|_{\mathcal{H}}^2$$

Strong convexity II

Let $\theta = 1/2$. Then we have

$$2F\left(\frac{f+g}{2}\right) \leq F(f) + F(g) - \frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2$$

By subtracting on both sides $2F(f)$ and adding $\lambda/2\ \|f-g\|_{\mathcal{H}}^2$ we have

$$\frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2 + 2F\left(\frac{f+g}{2}\right) - 2F(f) \leq F(f) - F(g)$$

Finally, note that if $f = \operatorname{argmin}_{f\in\mathcal{H}}\ F(f)$ we have $F\left(\frac{f+g}{2}\right) - F(f) \geq 0$ and therefore

$$\frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2 \leq F(f) - F(g)$$

Strong Convexity of Tikhonov Regularization

Let now define

- $F_1(\cdot) = \mathcal{E}_S(\cdot) + \lambda\|\cdot\|_{\mathcal{H}}^2$ and
- $F_1(\cdot) = \mathcal{E}_{S^{i,z'}}(\cdot) + \lambda\|\cdot\|_{\mathcal{H}}^2$

Furthermore, to simplify the notation denote $f_1 = f_S$ and $f_2 = f_{S^{i,z'}}$.

Recall that by construction

$$f_1 = \operatorname*{argmin}_{f\in\mathcal{H}} \ F_1(f) \qquad \text{and} \qquad f_2 = \operatorname*{argmin}_{f\in\mathcal{H}} \ F_2(f)$$

Strong Convexity of Tikhonov Regularization II

By our previous observation on strong convexity

$$\frac{\lambda}{2}\|f_1 - f_2\|_{\mathcal{H}}^2 \leq F_1(f_2) - F_1(f_1) \qquad \text{and} \qquad \frac{\lambda}{2}\|f - g\|_{\mathcal{H}}^2 \leq F_2(f_1) - F_2(f_2)$$

Summing the two inequalities (and rearranging the terms)

$$\begin{aligned}
\lambda\|f_1 - f_2\|_{\mathcal{H}}^2 &\leq F_1(f_2) - F_2(f_2) + F_2(f_1) - F_1(f_1) \\
&= \mathcal{E}_S(f_2) - \mathcal{E}_{S^{i,z'}}(f_2) + \mathcal{E}_{S^{i,z'}}(f_1) - \mathcal{E}_S(f_1) \\
&= \frac{1}{n}(\ell(f_2, z_i) - \ell(f_2, z') + \ell(f_1, z') - \ell(f_1, z_i)) \leq \frac{2}{n}\sup_z |\ell(f_1, z) - \ell(f_2, z)|
\end{aligned}$$

where we have used the definitions of F_1 and F_2.

Stability of Tikhonov Regularization (Continued)

Since $\sup_z |\ell(f_1, z) - \ell(f_2, z)| \leq Lk\|f_1 - f_2\|_{\mathcal{H}}$, we have

$$\lambda\|f_1 - f_2\|_{\mathcal{H}}^2 \leq \frac{2kL}{n}\|f_1 - f_2\|_{\mathcal{H}}$$

which implies

$$\|f_1 - f_2\|_{\mathcal{H}} \leq \frac{2kL}{n\lambda}$$

and from which we can conclude that

$$\sup_{z \in \mathcal{Z}} |\ell(f_1, z) - \ell(f_2, z)| \leq \frac{2L^2k^2}{n\lambda}$$

proving the $\beta(n) = \frac{2L^2k^2}{n\lambda}$ uniform stability of Tikhonov regularization.

So far...

In previous classes we have studied the excess risk of an estimator (in particular its sample error) by controlling the complexity of the space of functions from which the estimator was sampled (e.g. by Covering numbers).

In this class we have investigated an alternative approach that focuses *exclusively on properties of the learning algorithm* (rather than of the whole space). In particular we have observed how the stability of an estimator allows to control its generalization error in expectation.

We have shown in particular that Tikhonov regularization is a stable algorithm. This allowed to immediately derive excess risk bounds.

Stability and Generalization (in Probability)

Ok but... what about controlling the generalization error in probability rather than in expectation?

We can exploit the following result

McDiarmid's Inequality. Let $F : \mathcal{Z}^n \times \mathcal{Z}^n \to \mathbb{R}$ such that for any $i = 1, \dots, n$ there exists a $c_i > 0$ for which $\sup_{S \in \mathcal{Z}^n, z \in \mathcal{Z}} \ |F(S) - F(S^{i,z})| \leq c_i$. Then,

$$\mathbb{P}_{S \sim \rho^n} \left(|F(S) - \mathbb{E}_{S' \sim \rho^n} F(S')| \geq \epsilon \right) \leq 2 \exp \left(- \frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right)$$

Stability and Generalization (Continued)

(Slides thanks to Lorenzo Rosasco and Tomaso Poggio)

For a $\beta(n)$ uniformly stable algorithm $\mathcal{A}$ we will apply McDiarmid's inequality to the excess risk of the estimator returned by the algorithm, namely

$$F(S) = \mathcal{E}(f_S) - \mathcal{E}_S(f_S).$$

Where for any $S \in \mathcal{Z}^n$ we have denoted $f_S = \mathcal{A}(S)$.

Recall that $\mid \mathbb{E}_S \; [F(S)] \mid = \mid \mathbb{E}_S \; \mathcal{E}(f_S) - \mathcal{E}_S(f_S) \mid \leq \beta(n)$.

Stability and Generalization (Continued)

By McDiarmid, for any $\delta \in (0, 1]$ we have

$$|F(S) - \mathbb{E}_{S'} F(S')| \leq \sqrt{\frac{\sum_{i=1}^{n} c_i \ \log(2/\delta)}{2}}$$

with probability no less than $1 - \delta$, where

$$\sup_{S \in \mathcal{Z}^n, z \in \mathcal{Z}} |F(S) - F(S^{i,z})| \leq c_i$$

for $i = 1, \ldots, n$.

Stability and Generalization (Continued)

In particular, since $|\mathbb{E}_{S'} F(S')| \leq \beta(n)$, and $F(S) = \mathcal{E}_S(f_S) - \mathcal{E}(f_S)$, we have

$$|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)| \leq \beta(n) + \sqrt{\frac{\sum_{i=1}^{n} c_i \;\; \log(2/\delta)}{2}}$$

with probability no less than $1 - \delta$.

We need to bound the c_i.

Bounding the Deviation of the Generalization Error

We have

$$
\begin{aligned}
|F(S) - F(S^{i,z})| &= |\mathcal{E}_S(f_S) - \mathcal{E}(f_S) + \mathcal{E}_{S^{i,z}}(f_{S^{i,z}}) - \mathcal{E}(f_{S^{i,z}})| \\
&\leq |\mathcal{E}_S(f_S) - \mathcal{E}(f_{S^{i,z}})| + |\mathcal{E}_{S^{i,z}}(f_{S^{i,z}}) - \mathcal{E}(f_S)| \\
&\leq \frac{1}{n}|\ell(f_S, z_i) - \ell(f_{S^{i,z}}, z)| + \frac{1}{n}\sum_{j \neq i} |\ell(f_S, z_j) - \ell(f_{S^{i,z}}, z_j)| + \beta(n) \\
&\leq 2\beta(n) + \frac{2}{n} \sup_{\substack{S \in \mathcal{Z}^n, \\ i=1,\ldots,n}} |\ell(f_S, z_i)|
\end{aligned}
$$

Depending on the algorithm $\mathcal{A}$ and loss function ℓ we can control the last term. Let us assume that there exists $M > 0$ such that

$$
\sup_{\substack{S \in \mathcal{Z}^n, \\ i=1,\ldots,n}} |\ell(f_S, z_i)| \leq M
$$

We will then provide an estimate of M for Tikhonov regularization.

Stability and Generalization (Continued)

We have shown that

$$\sum_{i=1}^{n} c_i^2 \leq 4 \sum_{i=1} (\beta(n) + M/n)^2 = 4n(\beta(n) + M/n)^2.$$

Plugging it into the previous bound, we have

$$|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)| \leq \beta(n) + (n\beta(n) + M)\sqrt{\frac{2\log(2/\delta)}{n}}$$

with probability no less than $1 - \delta$.

Stability of Tikhonov Regularization

The last term we need to control is

$$\sup_{\substack{S\in\mathcal{Z}^n,\\ i=1,\dots,n}} |\ell(f_S, z_i)|$$

We will show it for Tikhonov regularization.

Stability of Tikhonov Regularization (Continued)

Assume that for any $y \in \mathcal{Y}$, the loss $\ell(0, y) \leq C_0$ is uniformly bounded by a constant $C_0 \geq 0$. Since f_S is the minimizer of the Tikhonov regularized empirical risk, we have that for any $S \in \mathcal{Z}^n$

$$\mathcal{E}_S(f_S) + \lambda \|f_S\|^2 \leq \mathcal{E}_S(0) \leq C_0$$

In particular, if the loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is non-negative, this implies

$$\|f_S\| \leq \sqrt{\frac{C_0}{\lambda}}$$

Stability of Tikhonov Regularization (Continued)

Therefore, for any $S \in \mathcal{Z}^n$ and $z \in \mathcal{Z}$,

$$\begin{aligned}
|\ell(f_S, z)| &\leq |\ell(f_S, z) - \ell(0, z)| + |\ell(0, z)| \\
&\leq |\ell(f_S, z) - \ell(0, z)| + C_0 \\
&\leq kL\|f_S\|_{\mathcal{H}} + C_0 \\
&\leq kL\sqrt{\frac{C_0}{\lambda}} + C_0
\end{aligned}$$

Stability of Tikhonov Regularization (Continued)

By plugging our estimate of $M = kL\sqrt{\frac{C_0}{\lambda n}} + C_0$, and the $\beta(n) = \frac{2k^2L^2}{n\lambda}$ stability of Tikhonov regularization in the bound on the generalization error, we have

$$|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)| \leq \frac{2k^2L^2}{n\lambda} + (\frac{2k^2L^2}{\lambda} + kL\sqrt{\frac{C_0}{\lambda n}} + C_0)\sqrt{\frac{2\log(2/\delta)}{n}}$$

with probability no less than $1 - \delta$.

Stability of Tikhonov Regularization (Continued)

In particular, the generalization error of Tikhonov regularization will tighten as

$$|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)| \leq O\left(\frac{1}{\lambda\sqrt{n}}\right)$$

with high probability. As expected, the bound on the generalization error will decrease as we observe more point but will increase if we regularize less (e.g. make the algorithm less stable).

As already observed for the convergence in expectation, this can be combined with assumptions on the interpolation/approximation error in order to find the most suited estimate for λ

Wrapping up

This class:

- Stability & Generalization error
- Stability of Tikhonov Regularization

Next class: Stability of early stopping.