Lecture 19

Sample Compression. Stability.

Sasha Rakhlin

Nov 14, 2018

Outline

Compression Bounds

Algorithmic Stability

Compression Set

Let us use the shortened notation for data: $\mathscr{S} = \{Z_1, \ldots, Z_n\}$, and Let us make the dependence of the algorithm $\widehat{f}_n$ on the training set explicit: $\widehat{f}_n = \widehat{f}_n[\mathscr{S}]$. As before, denote $\mathcal{G} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$, and let us write $\widehat{g}_n(\cdot) = \ell(\widehat{f}_n(\cdot), \cdot)$. Let us write $\widehat{g}_n[\mathscr{S}](\cdot)$ to emphasize the dependence.

Suppose there exists a “compression function” C_k which selects from any dataset $\mathscr{S}$ of size n a subset of k examples $C_k(\mathscr{S}) \subseteq \mathscr{S}$ such that

$$\widehat{f}_n[\mathscr{S}] = \widehat{f}_k[C_k(\mathscr{S})]$$

That is, the learning algorithm produces the same function when given $\mathscr{S}$ or its subset $C_k(\mathscr{S})$.

One can keep in mind the example of support vectors in SVMs.

Then,

$$\begin{aligned}
\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n) &= \mathbb{E}\widehat{g}_n - \frac{1}{n}\sum_{i=1}^{n}\widehat{g}_n(Z_i) \\
&= \mathbb{E}\widehat{g}_k[C_k(\mathscr{S})](Z) - \frac{1}{n}\sum_{i=1}^{n}\widehat{g}_k[C_k(\mathscr{S})](Z_i) \\
&\leq \max_{I\subseteq\{1,\ldots,n\},|I|\leq k}\left\{\mathbb{E}\widehat{g}_k[\mathscr{S}_I](Z) - \frac{1}{n}\sum_{i=1}^{n}\widehat{g}_k[\mathscr{S}_I](Z_i)\right\}
\end{aligned}$$

where $\mathscr{S}_I$ is the subset indexed by I.

Since $\widehat{g}_k[\mathscr{S}_I]$ only depends on k out of n points, the empirical average is "mostly out of sample". Adding and subtracting loss functions on for an additional set of i.i.d. random variables $W = \{Z'_1, \ldots, Z'_k\}$ results in an upper bound

$$\max_{I \subseteq \{1,\ldots,n\}, |I| \leq k} \left\{ \mathbb{E}\widehat{g}_k[\mathscr{S}_I](Z) - \frac{1}{n} \sum_{Z' \in \mathscr{S}'} \widehat{g}_k[\mathscr{S}_I](Z') \right\} + \frac{(b-a)k}{n}$$

where $[a, b]$ is the range of functions in $\mathcal{G}$ and $\mathscr{S}'$ is obtained from $\mathscr{S}$ by replacing $\mathscr{S}_I$ with the corresponding subset W_I.

For each fixed I, the random variable

$$\mathbb{E}\widehat{g}_k[\mathscr{S}_I](Z) - \frac{1}{n} \sum_{Z' \in \mathscr{S}'} \widehat{g}_k[\mathscr{S}_I](Z')$$

is zero mean with standard deviation $O((b-a)/\sqrt{n})$. Hence, the expected maximum over I with respect to $\mathscr{S}, W$ is at most

$$c\sqrt{\frac{(b-a)k\log(en/k)}{n}}$$

since $\log\binom{n}{k} \leq k\log(en/k)$.

Conclusion: compression-style argument limits the bias

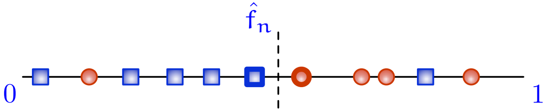
$$\mathbb{E}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right] \leq O\left(\sqrt{\frac{k \log n}{n}}\right),$$

which is non-vacuous if $k = o(n/\log n)$.

Recall that this term was the upper bound (up to log) on expected excess loss of ERM if class has VC dimension k. However, a possible equivalence between compression and VC dimension is still being investigated.

Example: Classification with Thresholds in 1D

- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$
- $\mathcal{F} = \{f_\theta : f_\theta(x) = \mathbf{I}\{x \geq \theta\}, \theta \in [0,1]\}$
- $\ell(f_\theta(x), y) = \mathbf{I}\{f_\theta(x) \neq y\}$



For any set of data $(x_1, y_1), \ldots, (x_n, y_n)$, the ERM solution $\widehat{f}_n$ has the property that the first occurrence x_l on the left of the threshold has label $y_l = 0$, while first occurrence x_r on the right – label $y_r = 1$.

Enough to take $k = 2$ and define $\widehat{f}_n[\mathscr{S}] = \widehat{f}_2[(x_l, 0), (x_r, 1)]$.

Further examples/observations:

- Compression of size d for hyperplanes (realizable case)
- Compression of size $1/\gamma^2$ for margin case
- Bernstein bound gives $1/n$ rate rather than $1/\sqrt{n}$ rate on realizable data (zero empirical error).

Outline

Recall that compression was a way to upper bound $\mathbb{E}\left[\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right]$. Algorithmic stability is another path to the same goal.

Compare:

- Compression: $\widehat{f}_n$ depends only on a subset of k datapoints.
- Stability: $\widehat{f}_n$ does not depend on any of the datapoints too strongly.

As before, let's write shorthand $g = \ell \circ f$ and $\widehat{g}_n = \ell \circ \widehat{f}_n$.

We now write

$$\mathbb{E}_{\mathscr{S}} \mathbf{L}(\widehat{f}_n) = \mathbb{E}_{Z_1,\ldots,Z_n,Z} \left\{ \ \widehat{g}_n[Z_1, \ldots, Z_n](Z) \ \right\}$$

Again, the meaning of $\widehat{g}_n[Z_1, \ldots, Z_n](Z)$: train on $Z_1, \ldots, Z_n$ and test on Z.

On the other hand,

$$\begin{aligned} \mathbb{E}_{\mathscr{S}} \widehat{\mathbf{L}}(\widehat{f}_n) &= \mathbb{E}_{Z_1,\ldots,Z_n} \left\{ \ \frac{1}{n} \sum_{i=1}^{n} \widehat{g}_n[Z_1, \ldots, Z_n](Z_i) \ \right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_1,\ldots,Z_n} \left\{ \ \widehat{g}_n[Z_1, \ldots, Z_n](Z_i) \ \right\} \\ &= \mathbb{E}_{Z_1,\ldots,Z_n} \left\{ \ \widehat{g}_n[Z_1, \ldots, Z_n](Z_1) \ \right\} \end{aligned}$$

where the last step holds for symmetric algorithms (wrt permutation of training data). Of course, instead of Z_1 we can take any Z_i.

Now comes the renaming trick. It takes a minute to get used to, if you haven't seen it.

Note that $Z_1, \dots, Z_n, Z$ are i.i.d. Hence,

$$
\begin{aligned}
\mathbb{E}_{\mathscr{S}} \mathbf{L}(\widehat{f}_n) &= \mathbb{E}_{Z_1,\dots,Z_n,Z} \left\{ \ \widehat{g}_n[Z_1,\dots,Z_n](Z) \ \right\} \\
&= \mathbb{E}_{Z_1,\dots,Z_n,Z} \left\{ \ \widehat{g}_n[Z, Z_2,\dots,Z_n](Z_1) \ \right\}
\end{aligned}
$$

Therefore,

$$
\mathbb{E}\left\{ \mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n) \right\} = \mathbb{E}_{Z_1,\dots,Z_n,Z} \left\{ \ \widehat{g}_n[Z, Z_2,\dots,Z_n](Z_1) - \widehat{g}_n[Z_1,\dots,Z_n](Z_1) \ \right\}
$$

Of course, we haven't really done much except re-writing expectation. But the difference

$$\widehat{g}_n[Z, Z_2, \ldots, Z_n](Z_1) - \widehat{g}_n[Z_1, \ldots, Z_n](Z_1)$$

has a "stability" interpretation. If it holds that the output of the algorithm "does not change much" when one datapoint is replaced with another, then the gap $\mathbb{E}\left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\}$ is small.

Moreover, since everything we've written is an equality, this stability is equivalent to having small gap $\mathbb{E}\left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\}$.

NB: our aim of ensuring small $\mathbb{E}\left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\}$ only makes sense if $\widehat{\mathbf{L}}(\widehat{f}_n)$ is small (e.g. on average). That is, the analysis only makes sense for those methods that explicitly or implicitly minimize empirical loss (or a regularized variant of it).

It's not enough to be stable. Consider a learning mechanism that ignores the data and outputs $\widehat{f}_n = f_0$, a constant function. Then $\mathbb{E}\left\{\mathbf{L}(\widehat{f}_n) - \widehat{\mathbf{L}}(\widehat{f}_n)\right\} = 0$ and the algorithm is very stable. However, it does not do anything interesting.

Uniform Stability

Rather than the average notion we just discussed, let's consider a much stronger notion:

We say that algorithm is β uniformly stable if

$$\forall i \in [n], z_1, \ldots, z_n, z', z \quad \left|\widehat{g}_n[\mathscr{S}](z) - \widehat{g}_n[\mathscr{S}^{i,z'}](z)\right| \leq \beta$$

where $\mathscr{S}^{i,z'} = \{z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n\}$.

Uniform Stability

Clearly, for any realization of $Z_1, \ldots, Z_n, Z$,

$$\widehat{g}_n[Z, Z_2, \ldots, Z_n](Z_1) - \widehat{g}_n[Z_1, \ldots, Z_n](Z_1) \leq \beta,$$

and so expected loss of a β-uniformly-stable ERM is β-close to its empirical error (in expectation).

Of course, it is unclear at this point whether a β-uniformly-stable ERM (or near-ERM) exists.

Kernel Ridge Regression

Consider

$$\widehat{f}_n = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + \lambda \|f\|_K^2$$

in RKHS $\mathcal{H}$ corresponding to kernel K.

Assume $K(x, x) \leq \kappa^2$ for any x.

Lemma: Kernel Ridge Regression is β-uniformly stable with $\beta = O\left(\frac{1}{\lambda n}\right)$

Proof (stability of Kernel Ridge Regression)

To prove this, first recall the definition of a σ-strongly convex function ϕ on convex domain $\mathcal{W}$:

$$\forall u, v \in \mathcal{W}, \quad \phi(u) \geq \phi(v) + \langle \nabla\phi(v), u - v \rangle + \frac{\sigma}{2} \|u - v\|^2 .$$

Suppose ϕ, ϕ' are both σ-strongly convex. Suppose w, w' satisfy $\nabla\phi(w) = \nabla\phi'(w') = 0$. Then

$$\phi(w') \geq \phi(w) + \frac{\sigma}{2} \|w - w'\|^2$$

and

$$\phi'(w) \geq \phi'(w') + \frac{\sigma}{2} \|w - w'\|^2$$

As a trivial consequence,

$$\sigma \|w - w'\|^2 \leq \left[\phi(w') - \phi'(w')\right] + \left[\phi'(w) - \phi(w)\right]$$

Proof (stability of Kernel Ridge Regression)

Now take

$$\phi(f) = \frac{1}{n} \sum_{i \in \mathscr{S}} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2$$

and

$$\phi'(f) = \frac{1}{n} \sum_{i \in \mathscr{S}'} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2$$

where $\mathscr{S}$ and $\mathscr{S}'$ differ in one element: (x_i, y_i) is replaced with (x_i', y_i').

Let $\widehat{f}_n, \widehat{f'_n}$ be the minimizers of ϕ, ϕ', respectively. Then

$$\phi(\widehat{f'_n}) - \phi'(\widehat{f'_n}) \leq \frac{1}{n}\left((\widehat{f'_n}(x_i) - y_i)^2 - (\widehat{f'_n}(x_i') - y_i')^2\right)$$

and

$$\phi'(\widehat{f}_n) - \phi(\widehat{f}_n) \leq \frac{1}{n}\left((\widehat{f}_n(x_i') - y_i')^2 - (\widehat{f}_n(x_i) - y_i)^2\right)$$

NB: we have been operating with f as vectors. To be precise, one needs to define the notion of strong convexity over $\mathcal{H}$. Let us sweep it under the rug and say that ϕ, ϕ' are 2λ-strongly convex with respect to $\|\cdot\|_K$.

Proof (stability of Kernel Ridge Regression)

Then $\left\|\widehat{f}_n - \widehat{f}'_n\right\|_K^2$ is at most

$$\frac{1}{2\lambda n}\left((\widehat{f}'_n(x_i) - y_i)^2 - (\widehat{f}_n(x_i) - y_i)^2 + (\widehat{f}_n(x'_i) - y'_i)^2 - (\widehat{f}'_n(x'_i) - y'_i)^2\right)$$

which is at most

$$\frac{1}{2\lambda n} C \left\|\widehat{f}_n - \widehat{f}'_n\right\|_\infty$$

where $C = 4(1 + c)$ if $|Y_i| \leq 1$ and $|\widehat{f}_n(x_i)| \leq c$.

On the other hand, for any x

$$f(x) = \langle f, K_x \rangle \leq \|f\|_K \|K_x\| = \|f\|_K \sqrt{\langle K_x, K_x \rangle} = \|f\|_K \sqrt{K(x, x)} \leq \kappa \|f\|_K$$

and so

$$\|f\|_\infty \leq \kappa \|f\|_K .$$

Proof (stability of Kernel Ridge Regression)

Putting everything together,

$$\left\|\widehat{f}_n - \widehat{f}'_n\right\|_K^2 \leq \frac{1}{2\lambda n} C \left\|\widehat{f}_n - \widehat{f}'_n\right\|_\infty \leq \frac{\kappa C}{2\lambda n} \left\|\widehat{f}_n - \widehat{f}'_n\right\|_K$$

Hence,

$$\left\|\widehat{f}_n - \widehat{f}'_n\right\|_K \leq \frac{1}{2\lambda n} C \left\|\widehat{f}_n - \widehat{f}'_n\right\|_\infty \leq \frac{\kappa C}{2\lambda n}$$

To finish the claim,

$$(\widehat{f}_n(x_i) - y_i)^2 - (\widehat{f}'_n(x_i) - y_i)^2 \leq C \left\|\widehat{f}_n - \widehat{f}'_n\right\|_\infty \leq \kappa C \left\|\widehat{f}_n - \widehat{f}'_n\right\|_K \leq O\left(\frac{1}{\lambda n}\right)$$