Stochastic Convex Optimization **Fall 2021**

Lecture 7: Stability

Lecturer: Roi Livni *Scribe:*

We so far discussed uniform-convergence generalization bounds, and provided dimension-dependent bounds for general convex functions and improved rates for general linear models. We now move towards a new technique that, surprisingly does not go through the uniform convergence argument.

7.1 Uniform Stability

The overall intuition behind stability arguments is to measure the sensitivity of the algorithm to a perturbation of a single point. In other words, we want to measure the difference between the algorithm's prediction rule given that it *has observed* a certain point, to it's decision *had it not* observed that point.

Today there are many stability arguments which largely differ on the process in which one excludes a point out of the sample (e.g. exclusion, replacement, randomly, worst-case etc...). We will work here with a slight variant of the following, original notion, of *uniform stability* introduced in [1]. Given a sample $S = \{z_1, \ldots, z_m\}$ and a sample S', let us write $|S - S'| \leq 1$ that is S and S' differ by at most one sample point, and given an algorithm A, we denote by w_S^A the output of algorithm A on sample S:

Definition 7.1 (Uniform Stability)**.** *An algorithm A is said to be $\beta(m)$-uniform stable if given a sample S and S' of size m, such that*

$$|S - S'| \leq 1,$$

then

$$\sup_{z \in \mathcal{Z}} |f(w_S^A, z) - f(w_{S'}^A, z)| \leq \beta(m).$$

Theorem 7.2. *Suppose algorithm A is $\beta(m)$ stable. Let w_S^A be the output of algorithm A on an i.i.d sample*

S drawn from distribution D. Then:

$$\mathop{\mathbb{E}}_{S\sim D}\left[\left|F(w_S^A)-\hat{F}_m(w_S^A)\right|\right]=\mathop{\mathbb{E}}_{S\sim D}\left[\left|\frac{1}{m}\sum_{i=1}^m f(w_S^A,z_i)-\mathbb{E}[f(w_S^A,z)]\right|\right]\leq\beta(m).$$

Corollary 7.3. *Suppose algorithm A is $\beta(m)$ stable, and given sample S returns w_S such that*

$$\hat{F}_m(w_S)\leq\min_{w\in\mathcal{W}}\hat{F}_m(w)+\epsilon(m).$$

Then

$$\mathop{\mathbb{E}}_{S\sim D}[F(w_S)]\leq\min_{w\in\mathcal{W}}F(w)+\epsilon(m)+\beta(m).$$

Proof of corollary 7.3. Let $w^\star$ be the minimizer of $F(w)$ on the set $\mathcal{W}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim D}[F(w_S)-F(w^\star)] &\leq \mathop{\mathbb{E}}_{S\sim D}[F(w_S)-\hat{F}_m(w_S)+\hat{F}_m(w_S)-F(w^\star)] \\
&\leq \beta(m)+\mathop{\mathbb{E}}_{S\sim D}[\hat{F}_m(w_S)-F(w^\star)] && \text{theorem 7.2} \\
&\leq \beta(m)+\epsilon(m)+\mathop{\mathbb{E}}_{S\sim D}[\hat{F}_m(w^\star)-F(w^\star)] && \hat{F}_m(w_S)\leq\hat{F}_m(w^\star)+\epsilon(m) \\
&= \beta(m)+\epsilon(m)+\mathop{\mathbb{E}}_{S\sim D}\left[\frac{1}{m}\sum_{i=1}^m f(w^\star,z_i)\right]-F(w^\star) \\
&= \beta(m)+\epsilon(m)+\frac{1}{m}\sum_{i=1}^m\mathop{\mathbb{E}}_{S\sim D}[f(w^\star,z_i)]-F(w^\star) \\
&= \beta(m)+\epsilon(m) && \mathop{\mathbb{E}}_{S\sim D}[f(w^\star,z_i)]=F(w^\star),\ \forall i=1,\dots,m
\end{aligned}$$

□

7.1.1 Proof of theorem 7.2

Given a sample $S=\{z_1,\dots,z_m\}$ and a point $z'\in\mathcal{Z}$, define for any $i\in\{1,\dots,m\}$

$$\hat{F}^{(i)}(w)=\frac{1}{m}\left(\sum_{j\neq i}f(w,z_j)+f(w,z')\right),$$

and also denote by ${w_S}^{(i)}$ the output of A given sample $S^{(i)}$ which is the that same as the original sample, except for example z_i that is replaced by z'.

By stability we have that $\forall z_1, \ldots, z_m, z', z \in \mathcal{Z}$, $|f(w_S, z) - f(w_S{}^{(i)}, z)| \leq \beta(m)$.

Consider a process where we randomly sample $z_1, \ldots, z_m$ and a point z' all i.i.d from the distribution D. Note that

$$\mathbb{E}\left[F(w_S)\right] = \mathbb{E}\left[F(w_S{}^{(i)})\right] = \mathbb{E}\left[f\left(w_S{}^{(i)}, z_i\right)\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f\left(w_S{}^{(i)}, z_i\right)\right].$$

In addition,

$$\mathbb{E}\left[\hat{F}_m(w_S)\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} f\left(w_S, z_i\right)\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f\left(w_S, z_i\right)\right]$$

Combining the two:

$$\mathbb{E}\left[F(w_S) - \hat{F}_m(w_S)\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f\left(w_S{}^{(i)}, z_i\right) - f\left(w_S, z_i\right)\right] \leq \beta(m)$$

7.2 Regularization

Next, we wish to apply the technique of stability to obtain improved generalization rate. In particular, we want dimension-independent bounds.

As we will later see, in general, an ERM algorithm or an Empirical Risk Minimzer need not be stable. So we will need to "stabilize" the algorithm. We will do it by adding a small *regularization* term. Given a sample S, let us denote the regularized empirical risk function by:

$$\hat{F}_\lambda(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} f(w, z_i).$$

Theorem 7.4. *Let $\mathcal{W} = B(0, R)$, and let f be an L lipschitz function bounded by 1. Given a sample S, let $w_S^\lambda \in \mathcal{W}$ be the minimizer of the objective function*

$$\hat{F}_\lambda(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} f(w, z_i). \tag{7.1}$$

Then

$$\mathop{\mathbb{E}}_{S\sim D}\left[F(w_S^\lambda)\right] \le \min_{w\in\mathcal{W}} F(w) + \frac{\lambda R^2}{2} + \frac{4L^2}{\lambda m}.$$

In particular, for a choice $\lambda = \frac{2L}{R\sqrt{m}}$*, we have:*

$$\mathop{\mathbb{E}}_{S\sim D}\left[F(w_S^\lambda)\right] \le \min_{w\in\mathcal{W}} F(w) + 3\frac{LR}{\sqrt{m}}.$$

Plugging the right choice of sample-size m yields the following corollary:

Corollary 7.5. *Any L-Lipschitz convex function is learnable (in expectation)*[1] *over the domain* $\mathcal{W} = B(0,R)$*, with sample complexity*

$$m(\epsilon) = O\left(\frac{L^2R^2}{\epsilon^2}\right).$$

7.2.1 Proof of theorem 7.4

The proof relies on bounding the sub-optimality of the solution with respect to the empirical risk, as well as the stability of the minimizer of the regularized empirical risk. Specifically, theorem 7.4 follows from corollary 7.3 as well as the next two claims (which we next prove):

Claim 7.6. *Let* S *be a sample and* $w_S^\lambda \in \mathcal{W}$ *be the minimizer of eq.* (7.1)*. Then*

$$\hat{F}_m(w_S^\lambda) \le \min_{w\in\mathcal{W}} \hat{F}_m(w) + \frac{\lambda R^2}{2}.$$

Claim 7.7. *The algorithm* A *that, given a sample* S*, returns* $w_S^\lambda \in \mathcal{W}$ *which is the minimizer of eq.* (7.1) *is* $\frac{4L^2}{\lambda m}$*-stable.*

Proof of claim 7.6. Again let $w^\star$ be the minimizer of $\hat{F}_m$ in $w \in \mathcal{W}$:

$$\begin{aligned}
\hat{F}_m(w_S^\lambda) &\le \frac{\lambda}{2}\|w_S^\lambda\|^2 + \hat{F}_m(w_S^\lambda) && \|w_S^\lambda\|^2 \ge 0\\
&\le \frac{\lambda}{2}\|w^\star\|^2 + \hat{F}_m(w^\star) && \text{minimality of } w_S^\lambda\\
&\le \frac{\lambda R^2}{2} + \hat{F}_m(w^\star) && \|w^\star\| \le R.
\end{aligned}$$

[1] Recall that a function is learnable if it is learnable in expectation (Exercise 5.2)

□

Proof of claim 7.7. Let $S = \{z_1, \ldots, z_m\}$ and $S' = \{z'_1, \ldots, z'_m\}$ be two samples that differ on example i. We set:

$$\hat{F}_{S,\lambda}(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{z\in S} f(w,z),$$

and for convenience of notation let us suppress the dependence of $w_S^\lambda, w_{S'}^\lambda$ and simply write $w_S, w_{S'}$:

$$\begin{aligned}
\hat{F}_{S,\lambda}\big(w_{S'}\big) - \hat{F}_{S,\lambda}\big(w_S\big) &= \frac{\lambda}{2}\left(\|w_{S'}\|^2 - \|w_S\|^2\right) + \frac{f\left(w_{S'}, z_i\right) - f\left(w_S, z_i\right)}{m} + \sum_{j\neq i}\frac{f\left(w_{S'}, z_j\right) - f\left(w_S, z_j\right)}{m} \\
&= \frac{f\left(w_{S'}, z_i\right) - f\left(w_S, z_i\right)}{m} + \frac{f\left(w_S, z'_i\right) - f\left(w_{S'}, z'_i\right)}{m} \\
&\quad + \frac{\lambda}{2}\left(\|w_{S'}\|^2 - \|w_S\|^2\right) + \sum_{j=1}^{m}\frac{f(w_{S'}, z'_j) - f(w_S, z'_j)}{m} \\
&= \frac{f\left(w_{S'}, z_i\right) - f\left(w_S, z_i\right)}{m} + \frac{f\left(w_S, z'_i\right) - f\left(w_{S'}, z'_i\right)}{m} \\
&\quad + \underbrace{\hat{F}_{S',\lambda}(w_{S'}) - \hat{F}_{S',\lambda}(w_S)}_{\leq 0} \\
&\leq \frac{|f(w_{S'}, z_i) - f(w_S, z_i)|}{m} + \frac{|f(w_{S'}, z'_i) - f(w_S, z'_i)|}{m} \\
&\leq \frac{2L}{m}\|w_{S'} - w_S\|
\end{aligned}$$

We next exploit the strong convexity of the regularized empirical risk. Recall (definition 4.4) that a function F is called λ-*strongly-convex* if

$$F(w) - F(w') \leq \nabla F(w)^\top (w - w') - \frac{\lambda}{2}\|w - w'\|^2,$$

and note that $F_{\lambda,S}$ is indeed λ -strongly convex (Exercise 4.2). By the λ-strongly convex property of the empirical loss:

$$\begin{aligned}
\hat{F}_{S,\lambda}\big(w_{S'}\big) - \hat{F}_{S,\lambda}(w_S) &\geq \nabla F_{\lambda,S}(w_S)^T\left(w_{S'} - w_S\right) + \frac{\lambda}{2}\|w_{S'} - w_S\|^2 \\
&= \frac{\lambda}{2}\|w_{S'} - w_S\|^2 \qquad \text{theorem 2.12}
\end{aligned}$$

Exploiting the upper as well as lower bound on the difference in the empirical loss we obtain that:

$$\frac{\lambda}{2}\|w_{S'} - w_S\|^2 \leq \frac{2L}{m}\|w_{S'} - w_S\| \Rightarrow \|w_{S'} - w_S\| \leq \frac{4L}{\lambda m}.$$

Finally, we exploit Lipschitzness of f:

$$\sup_{z \in Z} |f(w_S, z) - f(w'_S, z)| \leq L\|w_{S'} - w_S\| \leq \frac{4L^2}{\lambda m}.$$

□

Exercise 7.1. *Let $f(w, z)$ be an L-Lipschitz, λ-strongly convex function defined on $\mathcal{W} = B(0, 1)$ and bounded by 1. Assume D is an unknown distribution over z, and denote by $\hat{F}_m$, F the empirical and population risks respectively. Let $\hat{w}$ be such that:*

$$\hat{F}_m(\hat{w}) \leq \min_{w \in \mathcal{W}} \hat{F}_m(w) + \epsilon_0,$$

and let $\bar{w}$ be the minimizer of $\hat{F}_m$ in $\mathcal{W}$. Show that

1. $\|\hat{w} - \bar{w}\| \leq \sqrt{\frac{2\epsilon_0}{\lambda}}$.

2. $|F(\hat{w}) - F(\bar{w})| \leq \sqrt{\frac{2L^2\epsilon_0}{\lambda}}$

3. $$\mathop{\mathbb{E}}_{S \sim D}[F(\bar{w})] \leq \min_{w \in \mathcal{W}} F(w) + \frac{4L^2}{\lambda m}.$$

Conclude that if $L = 1$, any algorithm that minimizes $\hat{F}_m$ to $\epsilon_0 = O(\lambda\epsilon^2)$ precision where $m = O(1/(\lambda\epsilon))$ then:

$$\mathop{\mathbb{E}}_{S \sim D}[F(\hat{w})] \leq \min_{w \in \mathcal{W}} F(w) + \epsilon. \tag{7.2}$$

How many first order oracle calls are needed if we want to use GD to find $\hat{w}$ that satisfies eq. (7.2)? Describe which variant of GD should be used (i.e. the step size).

7.3 No Uniform Convergence

The learnability proof we provided here avoids uniform convergence. Moreover, when we apply a uniform convergence argument we basically show that *any* minimizer of the empirical risk will generalize. Instead, here, we provided what is termed an *algorithmic*-dependent proof: we showed that there exists an algorithm (regularized empirical risk) that finds a solution that generalizes. A natural question is then, could we have used uniform convergence bounds to obtain similar rates. We next answer this question in the negative, due to a construction by [3]. We will show that *any* bound that is obtained via uniform convergence must be dimension dependent:

Theorem 7.8. *There exists a 1-Lipschitz convex function $f(w, z)$, defined over $\mathcal{W} = B(0, 1)$, and a distribution D over z such that if $m \leq O(\log n)$ then w.p. $2/3$ over a sample S drawn from the distribution D, there exists $\hat{w} \in \mathcal{W}$ such that*

$$\hat{F}_m(\hat{w}) = \min_{w \in \mathcal{W}} \hat{F}_m(w) = 0,$$

but:

$$F(\hat{w}) \geq \min_{w \in \mathcal{W}} F(w) + \frac{1}{2}.$$

In particular, $m_{uc}(1/4, 1/3) = \Omega(\log n)$.

Remark 1. *The dependence in the dimension can be improved to linear, which matches the bounds for uniform convergence that we obtained in the last lecture. We refer the reader to [2] for a slightly more involved construction.*

Proof. For our function we will let $z = \{0, 1\}^n$ and our distribution D over z is defined such that $z(i) = 0$ w.p $1/2$ and $z(i) = 1$ w.p. $1/2$ independent of other coordinates. Finally, the function f is defined to be

$$f(w, z) = \sum_{j=1}^{n} z(j) w^2(j).$$

Now, given a sample $z_1, \ldots, z_m$ let us denote by E_j the event such that:

$$\forall z_i, \ z_i(j) = 0.$$

Note that for any j we have that $P(E_j) = 2^{-m}$. However, we claim that if $n \geq 2^m \ln 3/2$, then w.p at least

1/3 for some $\hat{j}$, the event $E_{\hat{j}}$ occurs. Indeed, because the coordinates are sampled i.i.d we have that the events $E_1, \ldots, E_n$ are independent and:

$$\begin{aligned}
\mathbb{P}\left(\neg \cup_{j=1}^{n} E_j\right) &= \mathbb{P}\left(\cap_{j=1}^{n} \neg E_j\}\right) \\
&= \prod_{j=1}^{n} \mathbb{P}\left(\neg E_j\right) && E_j \text{ are independent} \\
&= \prod_{j=1}^{n} (1 - 2^{-m}) \\
&= (1 - 2^{-m})^n \\
&\leq (1 - \frac{1}{2^m})^{2^m \ln 3/2} \\
&\leq 2/3
\end{aligned}$$

Overall then, we obtain that w.p 1/3 for some $\hat{j}$ the event $E_{\hat{j}}$ occurs. For a sample S, we choose then $\hat{w} = e_{\hat{j}}$ the standard basis vector at the coordinate where $E_{\hat{j}}$ happens. We have that

$$\hat{F}_m(\hat{w}) = 0.$$

However,

$$\mathop{\mathbb{E}}_{z \sim D}[f(\hat{w}, z)] = \mathop{\mathbb{E}}_{z \sim D}[z(\hat{j})] = \frac{1}{2}.$$

□

References

[1] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar): 499–526, 2002.

[2] V. Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29:3576–3584, 2016.

[3] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.