Stochastic Convex Optimization **Fall 2021**

Lecture 10: Algorithmic Generalization Lower Bounds

Lecturer: Roi Livni *Scribe:*

In Lecture 9 we analyzed the generalization performance of Gradient Descent, and provided stability analysis under different structural assumptions. We observed several different phenomena. For example, in smooth optimization, our bounds required *fast training* – As the number of iterations grows to infinity, the stability analysis breaks. For strongly convex functions our bounds behave differently – let T grow, increases stability and improves accuracy up to accuracy $O(1/m)$, when the statistical error starts to dominate the optimization error).

These bounds are perhaps what makes SCO such an interesting model to study learnability. Once our bounds cannot rely on uniform convergence, we are required to study the algorithms, and the exact algorithmic behaviour the induce generalization. And in fact, all of the above phenomena are observed, to some extent in real-life cases. Where sometimes *early stopping* of the training process helps to achieve generalization, and in many cases interpolation of the data set (i.e. achieving zero empirical error) is recommended for learning. We next wish to analyze and see how many of these phenomena are real, and which are *artifacts* of our stability analysis. In this lecture we will see that all of these phenomena are real, to some extent (though, some of the lower bounds will not be tight – leaving room from further research).

As in Lecture 9 we divide our study to three cases: strongly convex, smooth and the general case. We start with smooth functions.

Benchmark– optimal statistical rate Before we begin the analysis, notice that we already provided in Lecture 8 two generalization lower bounds that applies to *all* algorithms. In particular, we showed that any algorithm must exhibit (on some distribution) a mistake bound of

$$F(w_S) - F(w^\star) \geq \Omega\left(\frac{RL}{\sqrt{m}}\right),$$

and in the strongly convex case:

$$F(w_S) - F(w^\star) \geq \left(\frac{L}{\alpha m}\right),$$

These bounds are the *optimal statistical rate*. They have nothing to do with the algorithm at choice, and they are the *information-theoretic bound*, and they follow from inherent uncertainty due to lack of data.

Since we cannot expect any algorithm to break it (at least not in the general case), our analysis will mostly be concerned on when and how GD can *match* the optimal statistical rate. Hence, throughout we assume that we observe m examples and we we let $\epsilon = O(RL/\sqrt{m})$ be our benchmark, the desired accuracy we wish to achieve ($O(L/\alpha m)$ in the strong convex case).

10.1 Smooth Functions

In Lecture 9 (section 9.2) we devised a stability bound for smooth optimization that deteriorates as $T \to \infty$. For the bound to be meaningful we need $T = o(m)$, and to achieve optimal rate we require $T = \Theta(\sqrt{m})$. We next provide a lower bound that shows that T indeed needs to be finite for achieving generalization. Again, compare with the case of strongly convex functions where the bounds only improve with T.

Our lower bound will not match the upper bound, and it is still an open question how many iterations are "okay" before GD starts to overfit over a smooth optimization problem. Nevertheless, the next theorem does show that it can potentially happen. The construction below is due to Shalev-Shwartz et al. [3], the analysis is taken from [2].

Theorem 10.1. *For every T, η and m, there exists a γ-smooth, $\gamma \leq 3$, 2-Lipschitz, convex function $f(w, z)$ defined over $\mathcal{W} = B(0, 1) \subseteq \mathbb{R}^n$, such that $n = \Omega(2^m)$, and: if we run GD with step size $\eta = 1/\gamma$ for T iterations then:*

$$\mathop{\mathbb{E}}_{S \sim D^m}[F(w_S)] - \min_{w \in \mathcal{W}} F(w) = \Omega\left(\left(1 - \frac{n}{T}\right)^2\right).$$

Open Question 1. *Is there, for every m, a choice $T = \Omega(\sqrt{m})$, and a 1-smooth, 1-Lipschitz, convex function $f(w, z)$ defined over the unit ball $\mathcal{W} = B(0, 1)$ in $\mathbb{R}^n$ where*

$$n = \text{poly}(m)$$

such that: if we run GD with step size $\eta = O(1)$, for T iterations then w.p. $1/3$ over the sample S:

$$F(w_S) - \min_{w \in \mathcal{W}} F(w) = \Omega(1).$$

Open Question 2. *Is there, for every m,a 1-smooth, 1-Lipschitz, convex function $f(w, z)$ defined over the unit ball $\mathcal{W} = B(0, 1)$ in $\mathbb{R}^n$ where n may depend on m (could be exponential as in theorem 10.1 such that for $\eta = O(1)$:*

if we run GD with step size $\eta = O(1)$, for T iterations then :

$$\mathop{\mathbb{E}}_{S \sim D^m}[F(w_S)] - \min_{w} F(w) = \Omega\left(\frac{T}{m}\right).$$

Exercise 10.1. *Show that if n is* not *sufficiently large then Open Questions 1 and 2 are false. Namely, show that for some dependence $n = f(m, \epsilon)$, for any choice of T and 1-smooth, 1-Lipschitz convex function f we can choose η so that:*

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

10.2 Proof of theorem 10.1

Our construction will rely on the idea in theorem 7.8 where we constructed a function that an ERM may fail. To make sure *GD* fails (and not some abstract ERM), the trick is to make sure that our function will also encourage GD to move towards such a bad minimizer. In more details, we consider the function

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

Our distribution is again uniformly distributed in the unit cube: namely $z \in \{0, 1\}^n$ is chosen uniformly. Note that

$$F(w) = \mathop{\mathbb{E}}_{z \sim D}\left[\sum_{j=1}^{n} z(j) w^2(j) + \frac{1}{n} \sum_{i=1}^{n} w(j)\right] = \frac{1}{2}\|w\|^2 + \frac{1}{n} \sum_{j=1}^{n} w(j),$$

Next, similar to theorem 7.8 we argue that if $n \geq 2^m \ln 3/2$ then with probability at least $1/3$, there exists $\hat{j}$ such that $z_i(\hat{j}) = 0$ for all $i = 1, \ldots, m$. We assume that this event happened and we argue that in this

case, GD will converge to the bad solution. Let w_T be the solution of GD. We know that $\|w_T\| \leq 1$. Now given w_T let us propose an improved solution:

and let

$$\hat{w}_T(j) = \begin{cases} w_T(j) - 1 + \|w_T\| & j = \hat{j} \\ w_T(j) & j \neq \hat{j} \end{cases},$$

By triangular inequality note that

$$\|\hat{w}_T(j)\| = \|w_T - (1 - \|w_T\|e_{\hat{j}})\| \leq \|w_T(j)\| + (1 - \|w_T\|) \leq 1.$$

And we have that:

$$\hat{F}_m(w_T) - \hat{F}_m(\hat{w}_T) \geq \frac{1}{n}(1 - \|w_T\|).$$

On the other hand, since GD enjoys the following optimization guarantee, theorem 9.6:

$$\hat{F}_m(w_T) - \hat{F}_m(\hat{w}_T) \leq \frac{3\gamma + 2L}{T},$$

bounding $\gamma, L \leq 3$ we have that

$$(1 - \|w_T\|) \leq \frac{15n}{T}.$$

And $\|W_T\| \geq 1 - \frac{15n}{T}$. Finally,

$$F(w_T) - F(w^\star) \geq F(w_T) - F(0) = F(w_T) \geq \frac{\|w_T\|^2}{2} \geq \frac{1}{2}\left(1 - \frac{15n}{T}\right)^2.$$

Then:

10.3 The general case

We next discuss the general case. In Lecture 9 we showed that for a $O(1)$-Lipschitz functions, with $\mathcal{W} = B(0,1)$, we can derive a generalization bound:

$$\mathop{\mathbb{E}}_{S\sim D^m} F(w_S) - \min_{w\in\mathcal{W}} F(w) \leq O\left(\eta\sqrt{T} + \frac{1}{\eta T} + \frac{\eta T}{m}\right),$$

To minimize the bound on the RHS, we would like to choose $\eta = m^{3/2}$, and $T = m^2$. It is then natural to ask if $T = m^2$ iterations are necessary for GD to converge to an optimal solution of the population risk. It turns out that this upper bound is indeed tight. In particular, Amir et al. [2] showed the following lower bound for GD:

Theorem 10.2. *Fix η, T and m and set $\mathcal{W} = B(0,1)$. For sufficiently large $n \geq \Omega(T \cdot 2^{m+5})$, there exists a $O(1)$-Lipschitz convex function $f(w;z)$, and a distribution D over $\mathcal{Z}$, such that if we run GD with step size η and for T iterations then:*

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right). \tag{10.1}$$

In particular, to achieve an optimal error rate of $O(1/\sqrt{m})$, $T = m^2$ iterations are necessary (for any choice of η). Compare eq. (10.1), to the optimization guarantee for GD we obtained at Lecture 4. With respect to the empirical risk we have the convergence rate (theorem 4.3):

$$\hat{F}_m(w_S) \leq \min_{w^\star \in \mathcal{W}} \hat{F}_m(w^\star) + O\left(\eta + \frac{1}{\eta T}\right).$$

Note, then, that for a choice $\eta = O(1/\sqrt{m})$ and $T = m$, we obtain that GD converges to a solution with empirical error $O(1/\sqrt{m})$ but on the other hand, the generalization error is $\Omega(1)$. It remains an open problem whether the dependence on the dimension needs to be exponential

Open Question 3. *Is there, for every m, η, T a 1-Lipschitz, convex function $f(w,z)$ defined over the unit ball $\mathcal{W} = B(0,1)$ in $\mathbb{R}^n$ where*

$$n = \text{poly}(m).$$

such that: if we run GD with any step size η, for T iterations then:

$$\mathop{\mathbb{E}}_{S\sim D^m} F(w_S) - \min_{w\in\mathcal{W}} F(w) = \Omega\left(\eta\sqrt{T} + \frac{1}{\eta T}\right).$$

Open Question 4 (weaker version of Open Question 3)**.** *Is there a convex function $f(w,z)$ defined over the unit ball $\mathcal{W} = B(0,1)$ in $\mathbb{R}^n$ where*

$$n = \text{poly}(m).$$

such that, for $\eta = 1/\sqrt{T}$, if we run GD with any step size η, for T:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S) - \min_{w\in\mathcal{W}} F(w)\right] = \Omega\left(1\right).$$

We will not prove here theorem 10.2 in full generality and we refer the reader to [2] for a full proof. We will prove, instead, a slightly weaker result that builds on a similar, yet simpler construction, and shows that merely minimizing the empirical risk might lead to overfitting:

Theorem 10.3. *[Weaker version of theorem 10.2] Fix η, T and m. For sufficiently large $n \geq \Omega(T\cdot 2^{m+5})$, there exists a Lipschitz convex function $f(w;z)$, and a distribution D over $\mathcal{Z}$, such that if if we run GD with step size $\eta \leq 1/\sqrt{T}$, and for T iterations then:*

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S)\right] \geq \min_{w^\star\in\mathcal{W}} F(w^\star) + \Omega\left(\eta^2 T\right). \tag{10.2}$$

Importantly, choosing $\eta = O(1/\sqrt{m})$ and $T = m$ still leads to overfitting with this construction. So theorem 10.3 will prove that for wrong choice of parameters η and T, GD may overfit.

10.3.1 Proof of theorem 10.3

For the construction we define the following function defined for every $z \in \{0,1\}^n$:

$$f(w,z) = \sum_{i=1}^{n} z(j)w^2(j) + cw\cdot z + \max\{\max_{j\in[n]}\{w(j)\}, 0\},$$

where c is a sufficiently small scalar, to be determined later. In particular we assume $c < \frac{1}{\sqrt{n}}$, hence $\|c\cdot z\| \leq 1$.

As before, we assume a uniform distribution over $z \in \{0,1\}^n$. In particular, the population loss is given by:

$$F(w) = \frac{1}{2}\|w\|^2 + \mathbb{E}[z]\cdot w + \max\{\max_{j\in[n]}\{w(j)\}, 0\},$$

where $\mathbb{E}[z] = \frac{c}{2}\mathbf{1}$.

Note that the function $g(w) = \max\{\max_{j\in[n]}\{w(j)\}, 0\}$ is non-differentiable. So we need to define the exact first order oracle we will use for the proof.

First, observe that 0 is a subgradient at $w = 0$ (indeed $w = 0$ is an optimal solution and by first order optimality condition we have that $0 \in \partial g(0)$. Also observe that for any solution with $w \neq 0$, if

$$\hat{j} = \arg\max\{w(j)\}.$$

and if $w(\hat{j}) \geq 0$ then $e_{\hat{j}} \in \partial g(w)$. Indeed, we have that for for every u:

$$g(w) - g(u) = w(\hat{j}) - g(u) = w(\hat{j}) - \{\max \max_{j\in[n]}\{u(j)\}, 0\} \leq w(\hat{j}) - u(\hat{j}) \leq e_{\hat{j}}^{\top}(w - u).$$

So we can assume the oracle returns at 0, the subgradient 0 and at every $w \neq 0$, with some non-negative coordinate, the oracle returns the smallest $\hat{j}$ such that:

$$\hat{j} = \arg\max\{w(j)\}.$$

Exercise 10.2. *Let $f_1, \ldots, f_k$ be convex functions. Define*

$$f(w) = \max_k\{f_k(w)\}.$$

Show that for each w_0 $\nabla f_{k'}(w_0)$ where $f(w_0) = f_{k'}(w_0)$ is a subgradient at w_0.

Having defined f, let $z_1, \ldots, z_m$ be an i.i.d sample, then we consider the empirical risk:

$$\hat{F}_m(w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} z_i(j)w^2(j) + c\bar{z}\cdot w + \max\{\max_{j\in[n]}\{w(j)\}, 0\},$$

where $\bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$.

We, again, want to argue that for several coordinates we have that for all i, $z_i(j) = 0$. Let us call a coordinate *bad* if $z_i(j) = 0$ for all i. Now we want to estimate how many bad coordinates there are. If we let X_j be a random variable such that $X_j = 1$ if j is a bad coordinate, and $X_j = 0$ else, then by the same calculation as in theorems 10.1 and 7.8 we have that $\mathbb{E}[X_j] = \mathbb{P}(\text{j is bad}) = 2^{-m}$. Overall the expected number of bad

coordinates is then given by:

$$\mathbb{E}[\sum_{j=1}^{n} X_j] = \sum_{j=1}^{n} \mathbb{E}[X_j] = n2^{-m},$$

and if $n \geq 4T2^m$, we have by Markov's inequality that with probability at least $1/4$, the number of bad coordinates is at least T. We will assume that this event happened. and we denote the set of first T bad coordinates by $\hat{J} = \{j_1, \ldots, j_T\}$ First, observe that $\bar{z}(j) = 0$ for every bad coordinate and $\bar{z}(j) > 0$ for every *good* coordinate (i.e. not bad). Also note that

$$\nabla \hat{F}_m(0) = c \cdot \bar{z},$$

and in particular, we have that $w_1 = -\eta c\bar{z}$, hence $w_1(j) = 0$ for bad coordinates and $w_1(j) < 0$ for every good coordinate. We next claim by indcution the following:

1. For every good coordinate $-\eta ct < w_{t+1}(j) < 0$.
2. For $j_1, \ldots, j_t$: $w_{t+1}(j_k) = -\eta$.
3. For $j_t, \ldots, j_T$: $w_{t+1}(j_T) = 0$.

We showed the case $t = 0$, so we are left with proving the general case. Assume the correctness of the statement for t and prove it for $t + 1$: First note that for every good coordinate, since $w_{t+1}(j) < 0$, we have that:

$$w_t(j) - \eta \nabla \hat{F}_m(w_t)[j] = w_t(j) - 2\eta \cdot \bar{z}(j) w_t(j) - c\eta\bar{z}(j) = (1 - 2\eta\bar{z}(j)) w_{t+1}(j) - \eta ct < 0,$$

and by the same claculation $w_{t+1}(j) - c\eta(t + 1)$.

For bad coordinates such that $w_t(j_k) < 0$, similarly we have that

$$\nabla \hat{F}_m(w_t)[j_k] = 0.$$

Finally, by our choice of subgradient, since

$$\nabla \max\{\max_j(w_{t+1}(j)), 0\} = e_{j_t},$$

we have that for j_t:

$$w_t(j_t) - \eta \nabla \hat{F}_m(w_t)[j_t] = 0 - \eta,$$

and for all $j_{t+1}, \ldots, j_T$

$$w_t(j_t) - \eta \nabla \hat{F}_m(w_t)[j_t] = 0 - 0.$$

Note also, that since

$$\|w_t - \eta \nabla F_m(w_t)\| \leq \sum_{i=1}^{t+1} w^2(j_i) + \sum_{j \notin \hat{J}} w^2(j) \leq \eta^2 t + n \cdot \max_{j \notin \hat{J}} w^2(j) \leq \eta^2 t + n \cdot (\eta c T)^2,$$

we have again for sufficiently small c that the update rule stay in the unit ball and

$$w_{t+1} = w_t - \eta \nabla F_m(w_t).$$

To conclude, we have that for each bad coordinate j_k:

$$w_S(j_k) = \frac{1}{T} \sum_{t=1}^{T} w_t(j_k) = \frac{1}{T} \sum_{t=k}^{T} \eta = \frac{T-k}{T} \eta,$$

and

$$F(w_S) - F(w^\star) = F(w_S) - F(0) \geq F(w_S) \geq \frac{1}{2}\|w_S\|^2 - c\|w_S\| \geq \frac{1}{2} \sum_{t=1}^{T} \frac{(T-k)^2 \eta^2}{T^2} - c \geq \frac{1}{2} \sum_{t=1}^{T/2} \frac{T^2}{4T^2} \eta^2 - c \geq \frac{\eta^2 T}{8} - c,$$

and again we assume that c is sufficiently small (say smaller than $\frac{\eta^2}{16}$).

10.4 Strongly convex functions

Using a similar construction as in the general case, in Amir et al. [2] the authors showed an analogue result to one certain variant of GD over a strongly convex function[1]. The variant considered was introduced in [2] and is known to achieve optimal optimization rates and include the update rule and output w_S:

$$w_{t+1} = \Pi_{\mathcal{W}} \left[w_t - \frac{2}{\lambda(t+1)} \nabla F(w_t) \right], \quad w_S := \sum_{t=1}^{T} \frac{2t}{T(T+1)} w_t, \tag{10.3}$$

[1]The original paper, in fact showed a lower bound over the true loss, given a regularized objective but the same techniqe proves the following result

Theorem 10.4. *Fix m, $\lambda > 0$ and T, and assume $m \geq T \cdot 2^{m+5} \cdot m$. For the variant of GD depicted in eq.* (10.3) *there exist a distribution D over convex Lipschitz functions, $f(w; z)$ such that:*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S)] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\frac{1}{\lambda\sqrt{T}}, 1\right\}\right). \tag{10.4}$$

10.4.1 Revisiting the role of regularization

Recall that is Lecture 7 we showed that solving the regularized empirical risk problems induces stability and hence guarantees generalization. Combining the generalization lower bound in eq. (10.4) together with the optimization error inflicted by adding regularization we observe that, at least for the optimization protocol of [2] we obtain that the output of the regularized objective yields error of:

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S)] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\frac{1}{\lambda\sqrt{T}} + \lambda, 1\right\}\right).$$

In particular, to achieve an error rate of $O(1/\sqrt{m})$ we need to choose $\lambda = O(1/\sqrt{m})$ and $T = \Omega(m^2)$. Leading to the same performance as GD. Thus, at least apriori it is not clear whether regularization improves the performance once we consider a solution of a first-order algorithm.

It is natural to ask, though, whether by choosing slightly different learning rate, or by carefully devising the "right" regularization method, could we somehow perform as well as SGD, and perform better than $T = O(m^2)$?

In Amir et al. [1], the authors investigated *full-batch* methods in generality. A full-batch method is any method where the access of the learner to the data is only via the first order oracle over the full empirical-risk.

Formally, a *full-batch first-order oracle* is a procedure that has access to a fixed sample $z_1, \ldots, z_m$, and given input $w \in \mathcal{W}$, outputs

$$\mathcal{O}(w) := (\nabla \hat{F}_m(w); F(w)).$$

where $\nabla \hat{F}_m(w)$ is an empirical risk sub-gradient of the form

$$\nabla \hat{F}_m(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla f(w; z_i), \tag{10.5}$$

and each sub-gradient $\nabla f(w, z_i)$ is computed by the oracle as a function of w and z_i.

A full-batch algorithm is then any algorithm that optimizes the empirical risk only via access to a full-batch first order oracle. GD is an example of a full-batch method, but also GD over the regularized objective. SGD is not a full-batch method as it queries the gradient at single points.

Amir et al. [1] provided the following result that demonstrates that no full batch method can perform as well as SGD, and the $T = \Omega(m^2)$ bounds we observed is inherent in all full-batch methods.

Theorem 10.5. *For $m, T \in \mathbb{N}$; there exists $n = poly(2^m, T)$ such that the following holds. For any full-batch first-order algorithm with oracle complexity at most T, there exists a 1-Lipschitz convex function $f(w; z)$ in $\mathcal{W}$, the unit-ball in $\mathbb{R}^n$, and a distribution D over $\mathcal{Z}$ such that, for some universal constant $c > 0$:*

$$\mathbf{E}_{S \sim D^n}[F(w_S)] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \frac{1}{\sqrt{m}} + \Omega\left(\min\left\{1 - c\frac{\sqrt{T}}{m}, 0\right\}\right). \tag{10.6}$$

In words, every methods that optimizes the empirical risk by computing the exact gradient of the empirical risk at each iteration must perform at least $T = O(m^2)$ iterations in order to achieve the statistical rate of $O(1/\sqrt{m})$. Given that GD achieve this rate, we observe that adding regularization, smoothing, or any other regularization technique does not improve this rate.

References

[1] I. Amir, Y. Carmon, T. Koren, and R. Livni. Never go full batch (in stochastic convex optimization). *arXiv preprint arXiv:2107.00469*, 2021.

[2] I. Amir, T. Koren, and R. Livni. Sgd generalizes better than gd (and regularization doesn't help). In *COLT*, 2021.

[2] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

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