Stochastic Convex Optimization **Fall 2021**

Lecture 9: Stability analysis for Gradient Descent

Lecturer: Roi Livni *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

So far we concluded that tight generalization bounds in SCO are *algorithmic specific*. Namely, it is not enough to collect sufficent data and obtain small empirical risk. Different algorithms may behave differently. In the next two lectures we will focus on analyzing Gradient Descent. In this lecture, we will focus on upper bounds and in the next lecture we will discuss lower bounds. Notice that, distinctively from our previous discussion on GD, where we analyzed its convergence to the minimizer of the empirical risk, our focus now is on the *true risk*.

Like before, we will divide our analysis to three different cases: strongly convex, smooth convex and the general case.

We recall the algorithm GD which we analyzed in Lecture 4.

Algorithm 1 Projected subgradient descent

Given a fixed sequence $\eta_1, \ldots, \eta_T$

Define $x_1 = 0$

for $t = 1, \ldots, T$ **do**

$y_{t+1} = x_t - \eta_t g_t$, $g_t \in \partial f(x_t)$

$x_{t+1} = \Pi_{\mathcal{W}}(y_{t+1})$

end for

Throughout, we assume we observe $z_1, \ldots, z_m$ data points from some unknown distribution and we consider the empirical risk:

$$\hat{F}_m(w) = \frac{1}{m} \sum_{i=1}^{m} f(w, z_i).$$

We consider here a setting where we run the iterates of GD over $\hat{F}_m(w)$, receive the iterates $w_1, \ldots, w_T$ and

observe

$$w_S = \frac{1}{T}\sum_{t=1}^{T} w_t.$$

The learning rate η_t will depend on the exact setting, but eventually we would like to analyze the performance over the true risk

$$F(w) = \mathop{\mathbb{E}}_{z\sim D}[f(w,z)].$$

Next, we recall corollary 7.3 from Lecture 7 that states that an $\beta(m)$-stable algorithm A that optimizes over the empirical risk with $\epsilon(m)$-bounded error enjoys a generalization bound of

$$\mathop{\mathbb{E}}_{S\sim D}[F(w_S)] \leq \min_{w\in\in\mathcal{W}} F(w) + \epsilon(m) + \beta(m). \tag{9.1}$$

In light of eq. (9.1), throughout this section we will provide stabiilty guarantees for GD, and utilize the optimization error for the empirical risk we proved in Lecture 4 to obtain optimization error over the true risk.

9.1 Strongly Convex functions

We start with the setting which we, in fact, already analyzed in Lecture 7. The key difference is that in Lecture 7 we discussed how we can induce stability by *adding* regularization which made our objective function strongly convex. Here we will assume that we get strong convexity "for free". In turn we also don't pay additional cost for alternating the objective function.

The analysis of Lecture 7 provides, then, the following result:

Theorem 9.1. *Given $z_1,\ldots,z_m$ i.i.d samples from some unknown distribution D, assume f is L-Lipschitz. If f are α-strongly convex and we run GD with stepsize $\eta_t = \frac{1}{\alpha t}$, then*

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq \min_{w\in W} F(w) + \tilde{O}\left(\frac{L^2}{\alpha\sqrt{T}} + \frac{L^2}{\alpha m}\right) \tag{9.2}$$

If we want to optimize the above bound we would like to take $T = n^2$ iterations, this will yield an error:

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq \min_{w\in W} F(w) + \left(\frac{L^2}{\alpha m}\right)$$

Notice, that the above rate is distinctively slower than the optimization error of the empirical risk. In Lecture 4, we obtained a rate of $T = 1/\epsilon$ for the empirical risk, and not $T = 1/\epsilon^2$.

For the proof we first provide the following stability bound.

Lemma 9.2. *Suppose f is L-Lipschitz and α-strongly convex. Any algorithm that receives a sample S and returns output w_S such that*

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

is $\beta(m)$-stable where:

$$\beta(m) = 2L\sqrt{\frac{2\epsilon(m)}{\alpha}} + \frac{4L^2}{\alpha m}. \tag{9.3}$$

We can now use corollary 7.3. In particular, replace $\beta(m)$ in eq. (9.1) with the RHS in eq. (9.3) and obtain the following immediate corollary:

Corollary 9.3. *Suppose $f(w, z)$ is L-Lipschitz and α-strongly convex. Any algorithm that receives a sample S and returns output w_S such that*

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

guarantees:

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

Theorem 9.1 now follows from the optimization error we derived for GD with respect to a strongly convex objective. Recall theorem 4.5 that states the for gradient descent over an α-strongly convex and L-Lipschitz function, if $\eta_t = 1/\alpha t$ we obtain that:

$$f\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)-f(x^\star)\leq\frac{L^2(1+\log T)}{\alpha\cdot T}$$

Hence, plugging

$$\epsilon(m)=\frac{L^2(1+\log T)}{\alpha T}$$

into corollary 9.3, we obtain theorem 9.1. We are left then with proving lemma 9.2

9.1.1 Proof of lemma 9.2

Like in Lecture 7, our analysis follows the footsteps of Bousquet and Elisseeff [1] and we begin by showing that the minimizer of the empirical risk is *stable*. This is done very similarly to theorem 7.4, only this time we assume that the objective is strongly convex and we don't add any regularization term.

Suppose, then, that S and S' are two samples that are distinct on example i and denote $\hat{F}_S$ and $\hat{F}_{S'}$ the empirical risks, and $w_S^\star$, $w_{S'}^\star$ their respective minimizers:

$$\begin{aligned}
\hat{F}_S\big(w_{S'}^\star\big)-\hat{F}_S\big(w_S^\star\big) &= \frac{f\left(w_{S'}^\star,z_i\right)-f\left(w_S^\star,z_i\right)}{m}+\sum_{j\neq i}\frac{f\left(w_{S'}^\star,z_j\right)-f\left(w_S^\star,z_j\right)}{m}\\
&= \frac{f\left(w_{S'}^\star,z_i\right)-\ell\left(w_S^\star,z_i\right)}{m}+\frac{f\left(w_S^\star,z_i'\right)-f\left(w_{S'}^\star,z_i'\right)}{m}+\sum_{j=1}^{m}\frac{f(w_{S'}^\star,z_j')-f(w_S^\star,z_j')}{m}\\
&= \frac{f\left(w_{S'}^\star,z_i\right)-\ell\left(w_S^\star,z_i\right)}{m}+\frac{f\left(w_S^\star,z_i'\right)-f\left(w_{S'}^\star,z_i'\right)}{m}+\underbrace{\hat{F}_{S'}(w_{S'}^\star)-\hat{F}_{S'}(w_S^\star)}_{\leq 0}\\
&\leq \frac{|f(w_{S'}^\star,z_i)-f(w_S^\star,z_i)|}{m}+\frac{|f(w_{S'}^\star,z_i')-f(w_S^\star,z_i')|}{m} && \text{triangle inequality}\\
&\leq \frac{2L}{m}\|w_{S'}^\star-w_S^\star\|
\end{aligned}$$

By the α-strongly convex property of the empirical loss:

$$\begin{aligned}
F_S\big(w_{S'}^\star\big)-F_S(w_S^\star) &\geq \nabla F_S(w_S^\star)^T\left(w_S^\star-w_{S'}^\star\right)+\frac{\alpha}{2}\|w_{S'}^\star-w_S^\star\|^2 && (9.4)\\
&\geq \frac{\alpha}{2}\|w_{S'}^\star-w_S^\star\|^2 && \text{(first order optimality conditions (theorem 2.12))}
\end{aligned}$$

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

$$\frac{\alpha}{2}\|w^\star_{S'} - w^\star_S\|^2 \leq \frac{2L}{m}\|w^\star_{S'} - w^\star_S\| \Rightarrow \|w^\star_{S'} - w^\star_S\| \leq \frac{4L}{\alpha m}.$$

And we have that the minimizer of the empirical objective is $\frac{4L}{\alpha m}$ stable. Next we generalize eq. (9.4) as follows

Lemma 9.4. *Let F be an α-strongly convex function and $w^\star$ its minimizer over a set $\mathcal{W}$. Every $w \in \mathcal{W}$ satisfies:*

$$\|w - w^\star\| \leq \sqrt{\frac{2(F(w) - F(w^\star))}{\alpha}}$$

Proof.

$$\begin{aligned} F\big(w\big) - F(w^\star) &\geq \nabla F(w^\star)^T (w - w^\star) + \frac{\alpha}{2}\|w - w^\star\|^2 \\ &= \frac{\alpha}{2}\|w - w^\star\|^2 \qquad \text{first order optimality conditions (theorem 2.12)} \end{aligned}$$

□

Rearranging terms we obtain the desired result.

In particular, we have that if w_S is the output of the algorithm on sample S, $w_{S'}$ is the output of the algorithm on sample S' and $w^\star_S$ and $w^\star_{S'}$ are the minimizers of the respective empirical risks:

$$\begin{aligned} \|w_S - w_{S'}\| &\leq \|w_S - w^\star_S\| + \|w^\star_S - w^\star_{S'}\| + \|w^\star_{S'} - w_{S'}\| \\ &\leq \sqrt{\frac{2(F(w) - F(w^\star))}{\alpha}} + \frac{4L}{\alpha m} + \sqrt{\frac{2(F(w) - F(w^\star))}{\alpha}} \\ &\leq 2\sqrt{\frac{2\epsilon(m)}{\alpha}} + \frac{4L}{\alpha m}. \end{aligned}$$

and by Lipschitness we have that

$$\sup_{z \in \mathcal{Z}} |f(w_S, z) - f(w_{S'}, z)| \leq L\|w_S - w_{S'}\| \leq 2L\sqrt{\frac{2\epsilon(m)}{\alpha}} + \frac{4L^2}{\alpha m}.$$

Exercise 9.1. *Show that for an α-strongly convex function F, if $w^\star$ is the minimizer of F in $\mathbb{R}^n$ then:*

$$F(w) - F(w^\star) \leq \frac{2}{\alpha}\|\nabla F(w)\|^2.$$

Exercise 9.2. *Assume $f(w,z)$ is L-Lipschitz, convex and 1-strongly convex. Construct a procedure that given m i.i.d samples $z_1, \ldots, z_m$, and a parameter w, if $m = O(\frac{1}{\epsilon})$ the following happens with probability $2/3$*

1. *If w is the minimizer of $F(w) = \mathbb{E}_{z\sim D}[f(w,z)]$ the algorithm outputs ACCEPT.*
2. *If $F(w) - F(w^\star) > \epsilon$, the algorithm outputs REJECT.*

How many examples are needed if f is not strongly convex?

9.2 Smooth Functions

We next provide a generalization analysis for the *smooth* case. I.e. we assume that the function f are γ–smooth. Again for simplicity we assume the unconstrained case and $\mathcal{W} = \mathbb{R}^n$.

Theorem 9.5. *Suppose $\mathcal{W} \subseteq B(0,R)$. Given $z_1, \ldots, z_m$ i.i.d samples from some unknown distribution D. If $f(w,z)$ is γ-smooth convex and we run GD with stepsize $\eta = 1/\gamma$ and output $w_S = w_T$, then for $w^\star$ the minimizer of F:*

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq F(w^\star) + O\left(\frac{\gamma R^2 + LR}{T} + \frac{T}{\gamma m}L^2\right) \tag{9.5}$$

Notice that in the strongly convex case the stability relied on the proximity of the solution to the minimizer. In the smooth case we will see a reversed effect. Note the tradeoff between the two terms as T grows. So, in the strong convex case we want T to be as large as possible, in order to obtain small optimization error and large stability. We will later see that, in the smooth case, large T causes *instability*. And while for strongly convex functions we want *slow* training, here we want *fast* training. Overall the number of iterates that optimizes the upper bound is then given by

$$T = \frac{\gamma\|w^\star\|}{L}\sqrt{m},$$

which yields

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq F(w^\star) + O\left(L\sqrt{\frac{\|w^\star\|}{m}}\right).$$

The proof of theorem 9.5 again relies on theorem 7.2: and we provide both bounds to the optimization error and stability term. The optimization error can be bounded above by the following result which can be found in [1], Thm. 3.7 (theorem 9.6 is a slight strengthening of theorem 4.7, which we proved in Lecture 4, but extends to the constrained case):

Theorem 9.6. *Let $\mathcal{W} \subseteq B(0,R)$, and f be an L-Lipschitz convex and γ-smooth on $\mathbb{R}^n$. Then gradient descent (initialized at zero) with $\eta_t = \eta \leq \frac{1}{\gamma}$ satisfies for any $x^\star$:*

$$f(x_t) - f(x^\star) \leq \frac{(3\gamma R^2 + LR)}{t}$$

From eq. (9.1), theorem 9.5 follows from the following Lemma:

Lemma 9.7. *Suppose we run GD with $\eta \leq 2/\gamma$ and output $w_S = w_T$. Then if $f(w,z)$ is γ-smooth and L-Lipschitz for every z, then GD is $\beta(m)$-stable, where*

$$\beta(m) = \frac{2\eta T}{m} L^2.$$

We can now observe that the first term in the RHS in theorem 9.5 is the optimization error and the second term is the stability term. However, note that distinctively from before now the stability term deteriorates with the number of iterations.

9.2.1 Proof of lemma 9.7

The stability analysis we provide here follows the work of Hardt et al. [4] where the key ingredients in the analysis lies in the notion of *expansiveness* which we next introduce

Expansiveness We next consider update rule G which can be simply considered as functions from $\mathcal{W}$ to $\mathcal{W}$, and define the expansiveness of an update rule:

Definition 9.8. *We call any function from $\mathcal{W}$ to $\mathcal{W}$ an update rule. Let G be an update rule, we say that G is θ-expansive if $\|G(w) - G(u)\| \leq \theta\|w - u\|$, and given two update rules G, G' we denote:*

$$\|G - G'\| = \sup_{w \in W} |G(w) - G'(w)|.$$

We will be concerned, naturally, with the update rule that involves a single iterate of GD. Our analysis basically tracks the divergence in the trajectory of GD, given that we change one single function in the empirical risk. In other words, we will compare how w_t diverges from w'_t where w_t follows an update rule G induced by sample S, and w'_t follows an update rule G' induced by sample S':

Lemma 9.9. *Let G, G' be two θ-expansive update rules. Let $w_0 = 0$, and define recursively the sequences*

$$w_t = G'(w'_{t-1}), \qquad w'_t = G'(w_{t-1}),$$

and let

$$\delta_t = \|w_t - w'_t\|,$$

then:

$$\|w_t - w'_t\| \leq \theta\delta_{t-1} + \|G - G'\|.$$

In particular,

$$\|w_T - w'_T\| \leq \sum_{t=1}^{T} \theta^t \|G - G'\|.$$

Proof. The proof is by induction, and clearly $\delta_0 = 0$:

$$\begin{aligned}
\|w_t - w'_t\| &\leq \|G(w_{t-1}) - G(w'_{t-1}) + G(w'_{t-1}) - G'(w'_{t-1})\| \\
&\leq \|G(w_{t-1}) - G(w'_{t-1})\| + \|G(w'_{t-1}) - G'(w'_{t-1})\| \\
&\leq \theta\delta_{t-1} + \|G - G'\|.
\end{aligned}$$

□

Lemma 9.10. *Let f be any γ–smooth, and convex function, then the update rule:*

$$G(w_t) = \Pi_{\mathcal{W}}\left[w_t - \eta \nabla f(w_t)\right],$$

is 1 *expansive for every $\eta \leq 2/\gamma$.*

Proof. We recall from lemma 4.9 that smoothness implies

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2\gamma}\|\nabla f(x) - \nabla f(y)\|^2.$$

Applying the inequality twice (interchanging the roles of u and w with x and y, we have that

$$f(w) - f(u) \leq \nabla f(w)^\top (w - u) - \frac{1}{2\gamma}\|\nabla f(w) - \nabla f(u)\|,$$

$$f(u) - f(w) \leq \nabla f(u)^\top (u - w) - \frac{1}{2\gamma}\|\nabla f(u) - \nabla f(w)\|.$$

Summing up and rearranging terms we have that

$$(\nabla f(w) - \nabla f(u))^\top (w - u) \geq \frac{1}{\gamma}\|\nabla f(w) - \nabla f(u)\|^2$$

Thus we have that

$$\begin{aligned}
\|G(w_t) - G(w_t')\|^2 &\leq \|w_t - \eta \nabla f(w_t) - w_t' - \eta \nabla f(w_t')\|^2 && \text{lemma 4.1} \\
&= \|w_t - w_t'\|^2 - 2\eta (w_t - w_t')^\top (\nabla f(w_t) - \nabla f(w_t')) + \eta^2 \|\nabla f(w_t) - \nabla f(w_t')\|^2 \\
&\leq \|w_t - w_t'\|^2 - \frac{2\eta}{\gamma}\|\nabla f(w_t) - \nabla f(w_t')\|^2 + \eta^2 \|\nabla f(w_t) - \nabla f(w_t')\|^2 \\
&\leq \|w_t - w_t'\|^2 && \eta \leq 2/\gamma
\end{aligned}$$

□

Lemma 9.11. *Fix η, and given a sample S, denote G_S the update rule:*

$$G_S(w) = w - \frac{\eta}{|S|}\sum_{z \in S} \nabla f(w, z),$$

Suppose f is L-Lipschitz, then if S and S' are two samples of size m that differ on a single example, then $\|G_S - G_{S'}\| \leq \frac{2L}{m}$

Proof. Suppose S and S' differ on example i, then for every w:

$$\begin{aligned}\|G_S(w) - G_{S'}(w)\| &\leq \|w - \frac{\eta}{m}\sum_{j=1}^{m}\nabla f(w, z_j) - w - \frac{\eta}{m}\sum_{j=1}^{m}\nabla f(w, z'_j)\| \\ &= \|\frac{\eta}{m}\nabla f(w, z_i) + \frac{\eta}{m}\nabla f(w, z'_i)\| \\ &\leq \frac{\eta}{m}\|\nabla f(w, z_i)\| + \frac{\eta}{m}\|\nabla f(w, z'_i)\| \\ &\leq 2\frac{\eta L}{m}\end{aligned}$$

□

9.2.2 Proof of lemma 9.7

Denote G_S the update rule:

$$G_S(w) = w - \frac{\eta}{|S|}\sum_{z\in S}\nabla f(w, z),$$

Then by lemmas 9.9 to 9.11:

$$\|w_S - w_{S'}\| \leq \sum_{t=1}^{T}\|G_S - G_{S'}\| = T\|G_S - G_{S'}\| \leq \frac{2\eta T}{m}L.$$

Which yields the stability bound:

$$\sup_{z\in Z}|f(w_S, z) - f(w_{S'}, z)| \leq \frac{2\eta T L^2}{m}.$$

9.3 The General Case

We now move on to the general case, where we no longer assume f is strongly convex not smooth. In this case, we cannot rely on expansiveness or on the stability of the minimizer. Nevertheless we can still obtain a stability bound.

Unfortunatelly, though, we will not be able to obtain stability bounds for our choice of parameter $\eta = O(1/\sqrt{T})$, and this will indeed lead to suboptimal rates. But we can obtain stability bounds for other choices of η, which as we will see stills lead to meaningful (albeit suboptimal) guarantees. We summarize the following result:

Theorem 9.12. *Given $z_1, \ldots, z_m$ i.i.d sample from some unknown distribution D, assume f is L-Lipschitz, and we run GD with stepsize $\eta_t = \eta$, then*

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq \min_{w\in W} F(w) + O\left(\frac{R^2}{T\eta} + \eta L^2\sqrt{T} + \frac{\eta L^2 T}{m}\right) \tag{9.6}$$

Again, if we want to optimize the above bound one can show that wee need to choose $\eta T = O(L/(R\sqrt{m}))$, and $T = m^2$, leading to a generalization error of

$$\mathop{\mathbb{E}}_{S\sim D^m}[F(w_S)] \leq \min_{w\in W} F(w) + O\left(\frac{RL}{\sqrt{m}}\right) \tag{9.7}$$

which, as we discussed in Lecture 8 is the optimal statistical rate. Note however the suboptimality when compared with GD. Namely, the number of iterations we requires is $T = m^2$ (compared with $T = m$). The proof, again, follows a stability analysis, and we follow Bassily et al. [1] that analyzed the stability of Gradient Descent in the non-smooth case. We begin by bounding the otpimization error. Notice that we want to obtain a bound for a general η, while in Lecture 4 we analyzed the case that $\eta = O(1/\sqrt{T})$. We note, though, that in our proof of theorem 4.3 we've observed that for any choice of η the following bound holds

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^\star) \leq \frac{R^2}{2T\eta} + \frac{\eta L^2}{2}. \tag{9.8}$$

The proof is verbatim as in the proof of theorem 4.3, only we stop before the last line where we plug in our choice of $\eta = R/(L\sqrt{T})$. We next provide the stability analysis for GD. As before, theorem 9.12 follows from eq. (9.1) and eq. (9.8).

Lemma 9.13. *Suppose we run GD with step size η and output $w_S = \frac{1}{T}\sum_{t=1}^{T} w_t$. Then if $f(w, z)$ L-Lipschitz*

for every z, then GD is $\beta(m)$-stable, where

$$\beta(m) = 2\eta L^2\sqrt{T} + 2\frac{\eta L^2 T}{m}.$$

9.3.1 Proof of lemma 9.13

Let S and S' we two samples that differ on a single examples, say at index i, and we denote by w_t and w'_t the iterates of GD over the empirical risks, $F_S, F_{S'}$, induced by S and S' respectively. Let us denote by

$$\delta_t = \|w_t - w'_t\|.$$

Then we develop the following recursion formula:

$$\begin{aligned}
\delta_{t+1}^2 &= \|w_{t+1} - w'_{t+1}\|^2 \\
&= \|\Pi_{\mathcal{W}}[w_t - \eta\nabla F_S(w_t)] - \Pi_{\mathcal{W}}[w'_t - \eta\nabla F_{S'}(w'_t)]\|^2 \\
&\leq \|w_t - \eta\nabla F_S(w_t) - w'_t - \eta\nabla F_{S'}(w'_t)\|^2 \\
&\leq \|w_t - w'_t\|^2 - 2\eta(\nabla F_S(w_t) - \nabla F_{S'}(w'_t))^\top(w_t - w'_t) + \eta^2\|\nabla F_S(w_t) - \nabla F_{S'}(w'_t)\|^2 \\
&\leq \|w_t - w'_t\|^2 + \eta^2\|\nabla F_S(w_t) - \nabla F_{S'}(w'_t)\|^2 - 2\eta(\nabla F_S(w_t) - \nabla F_{S'}(w_t))^\top(w_t - w'_t) - 2\eta(\nabla F_{S'}(w_t) - \nabla F_{S'}(w'_t))^\top(w_t - w'_t) \\
&\leq \|w_t - w'_t\|^2 + \eta^2\|\nabla F_S(w_t) - \nabla F_{S'}(w'_t)\|^2 + 2\eta\|\nabla F_S(w_t) - \nabla F_{S'}(w_t))\|\|w_t - w'_t\| - 2\eta(\nabla F_{S'}(w_t) - \nabla F_{S'}(w'_t))^\top(w_t - w'_t) \\
&= \delta_t^2 + \eta^2\|\nabla F_S(w_t) - \nabla F_{S'}(w'_t)\|^2 + 2\eta\|\frac{1}{n}\nabla f(w_t, z_i) - \frac{1}{n}\nabla f(w_t, z'_i)\|\|w_t - w'_t\| - 2\eta(\nabla F_{S'}(w_t) - \nabla F_{S'}(w'_t))^\top(w_t - w'_t) \\
&\leq \delta_t^2 + 4\eta^2 L^2 + 2\frac{\eta L}{m}\delta_t - 2\eta(\nabla F_{S'}(w_t) - \nabla F_{S'}(w'_t))^\top(w_t - w'_t).
\end{aligned}$$

We next exploit *monotinicity of the subgradient*. Namely, for any convex function f, (in particular for $F_{S'}$, with $x = w_t$ and $y = w'_t$):

$$(\nabla f(x) - \nabla f(y))^\top(x - y) \geq 0. \tag{9.9}$$

Equation (9.9) simply follows from summing up the following two equations that follow from convexity:

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) \qquad f(y) - f(x) \leq \nabla f(y)^\top (y - x).$$

Thus, overall we obtain then:

$$\begin{aligned}
\delta_{t+1}^2 &\leq \delta_t^2 + 4\eta^2 L^2 + 2\frac{\eta L}{m}\delta_t \\
&\leq \delta_t^2 + 4\eta^2 L^2 + 2\frac{\eta L}{m}\delta_t - 2\eta(\nabla F_{S'}(w_t) - \nabla F_{S'}(w_t'))^\top (w_t - w_t') \\
&\leq \delta_t^2 + 4\eta^2 L^2 + 2\frac{\eta L}{m}\delta_t && \textit{eq. (9.9)} \\
&\leq 4\eta^2 L^2 \cdot (t+1) + 2\frac{\eta L}{m}\sum_{s=1}^{t}\delta_s. && \text{(by recursion)} \qquad (9.10)
\end{aligned}$$

Next denote $\delta^\star = \max_{t\in[T]} \delta_t$, then eq. (9.10) yields:

$$\delta^{\star 2} - 4\eta^2 L^2 T - 2\frac{\eta L T}{m}\delta^* \leq 0.$$

Solving the quadratic equation we can obtain a bound on $\delta^\star$, such that:

$$\begin{aligned}
\delta^\star &\leq \frac{2\frac{\eta L T}{m} + \sqrt{4\frac{\eta^2 L^2 T^2}{m^2} + 16\eta^2 L^2 T}}{2} \\
&= \frac{\eta L T}{m} + \sqrt{\frac{\eta^2 L^2 T^2}{m^2} + 4\eta^2 L^2 T} \\
&\leq 2\frac{\eta L T}{m} + 2\eta L\sqrt{T} && \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}
\end{aligned}$$

Finally, we have that:

$$\|w_S - w_{S'}\| \leq \frac{1}{T}\sum_{t=1}^{T}\|w_t - w_t'\| \leq \frac{1}{T}\sum \delta_t \leq \delta^* \leq 2\eta L\sqrt{T} + 2\frac{\eta L T}{m}.$$

And by Lipschitness we have that:

$$\sup_{z\in\mathcal{Z}} |f(w_S, z) - f(w_{S'}, z)| \leq 2\eta L^2\sqrt{T} + 2\frac{\eta L^2 T}{m}.$$

References

[1] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[1] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[4] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.