Supplementary Material for Data-Dependent Stability of Stochastic Gradient Descent

Ilja Kuzborskij [1] **Christoph H. Lampert** [2]

1. Proofs

In this section we present proofs of all the statements.

Proof of Theorem 2. Indicate by $S = \{z_i\}_{i=1}^m$ and $S' = \{z'_i\}_{i=1}^m$ independent training sets sampled i.i.d. from $\mathcal{D}$, and let $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_m\}$, such that $z'_i \overset{\text{iid}}{\sim} \mathcal{D}$. We relate expected empirical risk and expected risk by

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{A}\left[\widehat{R}_S(A_S)\right] &= \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{A}\left[\frac{1}{m}\sum_{i=1}^m f(A_S, z_i)\right] \\
&= \mathop{\mathbb{E}}_{S,S'}\mathop{\mathbb{E}}_{A}\left[\frac{1}{m}\sum_{i=1}^m f(A_{S^{(i)}}, z'_i)\right] \\
&= \mathop{\mathbb{E}}_{S,S'}\mathop{\mathbb{E}}_{A}\left[\frac{1}{m}\sum_{i=1}^m f(A_S, z'_i)\right] - \delta \\
&= \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{A}\left[R(A_S)\right] - \delta \,,
\end{aligned}$$

where

$$\begin{aligned}
\delta &= \mathop{\mathbb{E}}_{S,S'}\mathop{\mathbb{E}}_{A}\left[\frac{1}{m}\sum_{i=1}^m \left(f(A_S, z'_i) - f(A_{S^{(i)}}, z'_i)\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{S,z'_i}\mathop{\mathbb{E}}_{A}\left[f(A_S, z'_i) - f(A_{S^{(i)}}, z'_i)\right] \,.
\end{aligned}$$

Renaming z'_i as z and taking $\sup$ over i we get that

$$\delta \leqslant \sup_{i\in[m]}\left\{\mathop{\mathbb{E}}_{S,z}\mathop{\mathbb{E}}_{A}\left[f(A_S, z) - f(A_{S^{(i)}}, z)\right]\right\} \,.$$

This completes the proof. □

1.1. Preliminaries

We say that the Stochastic Gradient Descent (SGD) gradient update rule is an operator $G_t : \mathcal{H} \mapsto \mathcal{H}$, such that

$$G_t(\boldsymbol{w}) := \boldsymbol{w} - \alpha_t \nabla f(\boldsymbol{w}, z_{i_t}) \,,$$

and it is also a function of the training set S and a random index set I. Then, $\boldsymbol{w}_{t+1} = G_t(\boldsymbol{w}_t)$, throughout $t = 1, \ldots, T$. Recall the use of notation $\boldsymbol{w}_{S,t}$ to indicate the output of SGD ran on a training set S, at step t, and define

$$\delta_t(S, z) := \|\boldsymbol{w}_{S,t} - \boldsymbol{w}_{S^{(i)},t}\| \,.$$

Next, we summarize a few instrumental facts about G_t and few statements about the loss functions used in our proofs.

Definition 1 (Expansiveness)**.** *A gradient update rule is η-expansive if for all $\boldsymbol{w}, \boldsymbol{v}$,*

$$\|G_t(\boldsymbol{w}) - G_t(\boldsymbol{v})\| \leqslant \eta\|\boldsymbol{w} - \boldsymbol{v}\| \,.$$

The following lemma characterizes expansiveness for the gradient update rule under different assumptions on f.

Lemma 1 (Lemma 3.6 in (Hardt et al., 2016))**.** *Assume that f is β-smooth. Then, we have that:*

1) *G_t is $(1 + \alpha_t\beta)$-expansive,*
2) *If f in addition is convex, then, for any $\alpha_t \leqslant \frac{2}{\beta}$, the gradient update rule G_t is 1-expansive.*

An important consequence of β-smoothness of f is self-boundedness (Shalev-Shwartz & Ben-David, 2014), which we will use on many occasions.

Lemma 2 (Self-boundedness)**.** *For β-smooth non-negative function f we have that*

$$\|\nabla f(\boldsymbol{w}, z)\| \leqslant \sqrt{2\beta f(\boldsymbol{w}, z)} \,.$$

Self-boundedness in turn implies the following boundedness of a gradient update rule.

Corollary 1. *Assume that f is β-smooth and non-negative. Then,*

$$\begin{aligned}
\|\boldsymbol{w} - G_t(\boldsymbol{w})\| &= \alpha_t\|\nabla f(\boldsymbol{w}, z_{j_t})\| \\
&\leqslant \alpha_t \min\left\{\sqrt{2\beta f(\boldsymbol{w}, z_{j_t})}, L\right\} \,.
\end{aligned}$$

Proof. By Lemma 2

$$\|\alpha_t \nabla f(\boldsymbol{w}, z_{j_t})\| \leqslant \alpha_t\sqrt{2\beta f(\boldsymbol{w}, z_{j_t})} \,,$$

[1]University of Milan, Italy [2]IST Austria. Correspondence to: Ilja Kuzborskij <ilja.kuzborskij@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

and also by Lipschitzness of f, $\|\alpha_t \nabla f(\boldsymbol{w}, z_{j_t})\| \leqslant \alpha_t L$. □

Next we introduce a bound that relates the risk of the output at step t to the risk of the initialization point $\boldsymbol{w}_1$ through the variance of the gradient. Given an appropriate choice of step size, this bound will be crucial at stating stability bounds that depend on the risk at $\boldsymbol{w}_1$. The proof idea is similar to the one of (Ghadimi & Lan, 2013). In particular, it does not require convexity of the loss function.

Lemma 3. *Suppose SGD is ran with step sizes $\alpha_1, \ldots, \alpha_{t-1} \leqslant \frac{1}{\beta}$ w.r.t. the β-smooth loss f. Then we have that*

$$\begin{aligned}
&\sum_{k=1}^{t-1} \left(\alpha_k - \frac{\alpha_k^2 \beta}{2}\right) \mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_{S,k})\|^2\right] \\
&\leqslant R(\boldsymbol{w}_1) - R(\boldsymbol{w}_t) \\
&+ \frac{\beta}{2} \sum_{k=1}^{t-1} \alpha_k^2 \mathop{\mathbb{E}}_{S} \left[\|\nabla f(\boldsymbol{w}_{S,k}, z_{j_k}) - \nabla R(\boldsymbol{w}_{S,k})\|^2\right] .
\end{aligned}$$

Proof. For brevity denote $f_k(\boldsymbol{w}) \equiv f(\boldsymbol{w}, z_{j_k})$. By β-smoothness of R and racalling that the SGD update rule $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f_k(\boldsymbol{w}_k)$, we have

$$\begin{aligned}
&R(\boldsymbol{w}_{k+1}) - R(\boldsymbol{w}_k) \\
&\leqslant \nabla R(\boldsymbol{w}_k)^\top (\boldsymbol{w}_{k+1} - \boldsymbol{w}_k) + \frac{\beta}{2} \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|^2 \\
&= -\alpha_k \nabla R(\boldsymbol{w}_k)^\top \nabla f_k(\boldsymbol{w}_k) + \frac{\beta \alpha_k^2}{2} \|\nabla f_k(\boldsymbol{w}_k)\|^2 \\
&= -\alpha_k \nabla R(\boldsymbol{w}_k)^\top \nabla f_k(\boldsymbol{w}_k) \\
&+ \frac{\beta \alpha_k^2}{2} \|\nabla f_k(\boldsymbol{w}_k) - \nabla R(\boldsymbol{w}_k) + \nabla R(\boldsymbol{w}_k)\|^2 \\
&= -\left(\alpha_k + \alpha_k^2 \beta\right) \nabla R(\boldsymbol{w}_k)^\top \nabla f_k(\boldsymbol{w}_k) \\
&+ \frac{3\alpha^2 \beta}{2} \|\nabla R(\boldsymbol{w}_k)\|^2 + \frac{\beta \alpha_k^2}{2} \|\nabla f_k(\boldsymbol{w}_k) - \nabla R(\boldsymbol{w}_k)\|^2 .
\end{aligned}$$

Taking expectation w.r.t. S on both sides, recalling that $\mathbb{E}_{z_k}[\nabla f_k(\boldsymbol{w}_k)] = \nabla R(\boldsymbol{w}_k)$ and rearranging terms we get

$$\begin{aligned}
\left(\alpha_k - \frac{\alpha^2 \beta}{2}\right) \mathbb{E} \left[\|\nabla R(\boldsymbol{w}_k)\|^2\right] &\leqslant R(\boldsymbol{w}_k) - R(\boldsymbol{w}_{k+1}) \\
&+ \frac{\beta \alpha_k^2}{2} \mathbb{E} \left[\|\nabla f_k(\boldsymbol{w}_k) - \nabla R(\boldsymbol{w}_k)\|^2\right] ,
\end{aligned}$$

and summing above over $k = 1, \ldots, t-1$ we get the statement. □

Lemma 4. *Suppose SGD is ran with step sizes $\alpha_1, \ldots, \alpha_{t-1} \leqslant \frac{1}{\beta}$ on the β-smooth loss f. Assume that the variance of stochastic gradients obeys*

$$\mathop{\mathbb{E}}_{S,z} \left[\|\nabla f(\boldsymbol{w}_{S,k}, z) - \nabla R(\boldsymbol{w}_{S,k})\|^2\right] \leqslant \sigma^2 \quad \forall k \in [T] .$$

Then we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S} \left[\sum_{k=1}^{t-1} \alpha_k \|\nabla f(\boldsymbol{w}_{S,k}, z_k)\|\right] \\
&\leqslant 2 \sqrt{\left(\sum_{k=1}^{t-1} \alpha_k\right) \left(R(\boldsymbol{w}_1) - R^\star + \frac{\beta \sigma^2}{2} \sum_{k=1}^{t-1} \alpha_k^2\right)} \\
&+ \sigma \sum_{k=1}^{t-1} \alpha_k .
\end{aligned}$$

Proof. First we perform the decomposition,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S} \left[\sum_{k=1}^{t-1} \alpha_k \|\nabla f(\boldsymbol{w}_{S,k}, z_k)\|\right] \\
&= \sum_{k=1}^{t-1} \alpha_k \mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_{S,k})\|\right] \\
&+ \sum_{k=1}^{t-1} \alpha_k \mathop{\mathbb{E}}_{S} \left[\|\nabla f(\boldsymbol{w}_{S,k}, z_k) - \nabla R(\boldsymbol{w}_{S,k})\|\right] \\
&\leqslant \sum_{k=1}^{t-1} \alpha_k \mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_{S,k})\|\right] + \sigma \sum_{k=1}^{t-1} \alpha_k . \qquad (1)
\end{aligned}$$

Introduce

$$Q_t := \sum_{k=1}^{t-1} \left(\alpha_k - \frac{\alpha_k^2 \beta}{2}\right) .$$

Now we invoke the stationary-point argument to bound the first term above as

$$\begin{aligned}
&\sum_{k=1}^{t-1} \alpha_k \mathop{\mathbb{E}}_{S} \left[\sqrt{\|\nabla R(\boldsymbol{w}_k)\|^2}\right] \qquad (2) \\
&\leqslant \sum_{k=1}^{t-1} \frac{\left(1 - \frac{\alpha_k \beta}{2}\right)}{\left(1 - \frac{\alpha_k \beta}{2}\right)} \cdot \alpha_k \sqrt{\mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_k)\|^2\right]} \\
&\qquad \text{(By Jensen's inequality)} \\
&\leqslant 2 \sum_{k=1}^{t-1} \left(\alpha_k - \frac{\alpha_k^2 \beta}{2}\right) \sqrt{\mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_k)\|^2\right]} \\
&\qquad \text{(Assuming that } \alpha_k \leqslant \tfrac{1}{\beta}\text{)} \\
&= \frac{2Q_t}{Q_t} \sum_{k=1}^{t-1} \left(\alpha_k - \frac{\alpha_k^2 \beta}{2}\right) \sqrt{\mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_k)\|^2\right]} \qquad (3) \\
&\leqslant 2\sqrt{Q_t} \sqrt{\sum_{k=1}^{t-1} \left(\alpha_k - \frac{\alpha_k^2 \beta}{2}\right) \mathop{\mathbb{E}}_{S} \left[\|\nabla R(\boldsymbol{w}_k)\|^2\right]} \\
&\qquad \text{(By Jensen's inequality)} \\
&\leqslant 2\sqrt{Q_t} \sqrt{R(\boldsymbol{w}_1) - R(\boldsymbol{w}_t) + \frac{\beta \sigma^2}{2} \sum_{k=1}^{t-1} \alpha_k^2} . \\
&\qquad \text{(By Lemma 3)}
\end{aligned}$$

Combining this with (1) gives the statement and completes the proof. □

The following lemma is similar to Lemma 3.11 of (Hardt et al., 2016), and is instrumental in bounding the stability of SGD. However, we make an adjustment and state it in expectation over the data. Note that it does not require convexity of the loss function.

Lemma 5. *Assume that the loss function $f(\cdot, z) \in [0,1]$ is L-Lipschitz for all z. Then, for every $t_0 \in \{0, 1, 2, \dots m\}$ we have that,*

$$\mathop{\mathbb{E}}_{S,z}\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right] \tag{4}$$

$$\leqslant L \mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{A}\left[\delta_T(S,z) \mid \delta_{t_0}(S,z) = 0\right]\right] + \mathop{\mathbb{E}}_{S,A}\left[R(A_S)\right]\frac{t_0}{m} . \tag{5}$$

Proof. We proceed with elementary decomposition, Lipschitzness of f, and using the fact that f is non-negative to have that

$$\begin{aligned} &f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z) \qquad (6)\\ &= \left(f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right)\mathbb{I}\left\{\delta_{t_0}(S,z) = 0\right\} \\ &+ \left(f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right)\mathbb{I}\left\{\delta_{t_0}(S,z) \neq 0\right\} \\ &\leqslant L\delta_T(S,z)\mathbb{I}\left\{\delta_{t_0}(S,z) = 0\right\} \\ &+ f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\delta_{t_0}(S,z) \neq 0\right\} . \qquad (7)\end{aligned}$$

Taking expectation w.r.t. algorithm randomization, we get that

$$\begin{aligned} &\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right] \\ &\leqslant L\mathop{\mathbb{E}}_{A}\left[\delta_T(S,z)\mathbb{I}\left\{\delta_{t_0}(S,z) = 0\right\}\right] \\ &+ \mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\delta_{t_0}(S,z) \neq 0\right\}\right] . \qquad (8)\end{aligned}$$

Recall that $i \in [m]$ is the index where S and $S^{(i)}$ differ, and introduce a random variable τ_A taking on the index of the first time step where SGD uses the example z_i or a replacement z. Note also that τ_A does not depend on the data. When $\tau_A > t_0$, then it must be that $\delta_{t_0}(S,z) = 0$, because updates on both S and $S^{(i)}$ are identical until t_0. A consequence of this is that $\mathbb{I}\left\{\delta_{t_0}(S,z) \neq 0\right\} \leqslant \mathbb{I}\left\{\tau_A \leqslant t_0\right\}$. Thus the rightmost term in (8) is bounded as

$$\begin{aligned} &\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\delta_{t_0}(S,z) \neq 0\right\}\right] \\ &\leqslant \mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right] .\end{aligned}$$

Now, focus on the r.h.s. above. Recall that we assume randomization by sampling from the uniform distribution over $[m]$ without replacement, and denote a realization by $\{j_i\}_{i=1}^m$. Then, we can always express our randomization as permutation function $\pi_A(S) = \{z_{j_i}\}_{i=1}^m$. In addition, introduce an algorithm $\mathrm{GD} : \mathcal{Z}^m \mapsto \mathcal{H}$, which is identical to A, except that it passes over the training set S sequentially without randomization. That said, we have that

$$\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right] = \mathop{\mathbb{E}}_{A}\left[f(\mathrm{GD}_{\pi_A(S)}, z)\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right] ,$$

and taking expectation over the data,

$$\begin{aligned} &\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z)\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right]\right] \\ &= \mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{S,z}\left[f(\mathrm{GD}_{\pi_A(S)}, z)\right]\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right] .\end{aligned}$$

Now observe that for any realization of A, $\mathbb{E}_{S,z}\left[f(\mathrm{GD}_{\pi_A(S)}, z)\right] = \mathbb{E}_A\mathbb{E}_{S,z}\left[f(A_S, z)\right]$ because expectation w.r.t. S and z does not change under our randomization [1]. Thus, we have that

$$\begin{aligned} &\mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{S,z}\left[f(\mathrm{GD}_{\pi_A(S)}, z)\right]\mathbb{I}\left\{\tau_A \leqslant t_0\right\}\right] \\ &= \mathop{\mathbb{E}}_{S,A}\left[R(A_S)\right]\mathbb{P}(\tau_A \leqslant t_0) .\end{aligned}$$

Now assuming that τ_A is uniformly distributed over $[m]$ we have that

$$\mathbb{P}\left(\tau_A \leqslant t_0\right) = \frac{t_0}{m} .$$

Putting this together with (6) and (7), we finally get that

$$\begin{aligned} &\mathop{\mathbb{E}}_{S,z}\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right] \\ &\leqslant L \mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{A}\left[\delta_T(S,z)\mathbb{I}\left\{\delta_{t_0}(S,z) = 0\right\}\right]\right] + \mathop{\mathbb{E}}_{S,A}\left[R(A_S)\right]\frac{t_0}{m} \\ &\leqslant L \mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{A}\left[\delta_T(S,z) \mid \delta_{t_0}(S,z) = 0\right]\right] + \mathop{\mathbb{E}}_{S,A}\left[R(A_S)\right]\frac{t_0}{m} .\end{aligned}$$

This completes the proof. □

We spend a moment to highlight the role of conditional expectation in (5). Observe that we could naively bound (4) by the Lipschitzness of f, but Lemma 5 follows a more careful argument. First note that t_0 is a free parameter. The expected distance in (5) between SGD outputs $\boldsymbol{w}_{S,t}$ and $\boldsymbol{w}_{S^{(i)},t}$ is conditioned on the fact that at step t_0 outputs of SGD are still the same. This means that the perturbed point is encountered after t_0. Then, the conditional expectation should be a decreasing function of t_0: the later the perturbation occurs, the smaller deviation between $\boldsymbol{w}_{S,t}$ and $\boldsymbol{w}_{S^{(i)},t}$ we should expect. Later we use this fact to minimize the bound (5) over t_0.

[1] Strictly speaking we could omit $\mathbb{E}_A[\cdot]$ and consider *any* randomization by reshuffling, but we keep expectation for the sake of clarity.

1.2. Convex Losses

In this section we prove on-average stability for loss functions that are non-negative, β-smooth, and convex.

Theorem 1. *Assume that f is convex, and that SGD's is ran with step sizes $\{\alpha_t\}_{t=1}^T$. Then, for every $t_0 \in \{0, 1, 2, \dots m\}$, SGD is $\epsilon(\mathcal{D}, \boldsymbol{w}_1)$-on-average stable with*

$$\epsilon(\mathcal{D}, \boldsymbol{w}_1) \leqslant \frac{2}{m} \sum_{t=t_0+1}^{T} \alpha_t \mathop{\mathbb{E}}_{S,z} \left[\|\nabla f(\boldsymbol{w}_t, z_{j_t})\|\right] + \mathop{\mathbb{E}}_{S,A} \left[R(A_S)\right] \frac{t_0}{m} .$$

Proof. For brevity denote $\Delta_t(S, z) := \mathbb{E}_A \left[\delta_t(S, z) \mid \delta_{t_0}(S, z) = 0\right]$. We start by applying Lemma 5:

$$\begin{aligned} &\mathop{\mathbb{E}}_{S,z} \mathop{\mathbb{E}}_{A} \left[f(\boldsymbol{w}_{S,T}, z) - f(\boldsymbol{w}_{S^{(i)},T}, z)\right] \\ &\leqslant L \mathop{\mathbb{E}}_{S,z} \left[\Delta_T(S, z)\right] + \mathop{\mathbb{E}}_{S,A} \left[R(A_S)\right] \frac{t_0}{m} . \end{aligned} \tag{9}$$

Our goal is to bound the first term on the r.h.s. as a decreasing function of t_0, so that eventually we can minimize the bound w.r.t. t_0. At this point we focus on the first term, and the proof partially follows the outline of the proof of Theorem 3.7 in (Hardt et al., 2016). The strategy will be to establish the bound on $\Delta_T(S, z)$ by using a recursive argument. In fact we will state the bound on $\Delta_{t+1}(S, z)$ in terms of $\Delta_t(S, z)$ and then unravel the recursion. Finally, we will take expectation w.r.t. the data after we obtain the bound by recursion.

To do so, we distinguish two cases: 1) SGD encounters a perturbed point at step t, that is $t = i$, and 2) the current point is the same in S and $S^{(i)}$, so $t \neq i$. For the first case, we will use data-dependent boundedness of the gradient update rule, Corollary 1, that is

$$\|G_t(\boldsymbol{w}_{S,t}) - G_t(\boldsymbol{w}_{S^{(i)},t})\| \leqslant \delta_t(S, z) + 2\alpha_t \|\nabla f(\boldsymbol{w}_{S,t}, z_{j_t})\| .$$

To handle the second case, we will use the expansiveness of the gradient update rule, Lemma 1, which states that for convex loss functions, the gradient update rule is 1-expansive, so $\delta_{t+1}(S, z) \leqslant \delta_t(S, z)$. Considering both cases of example selection, and noting that SGD encounters the perturbation w.p. $\frac{1}{m}$, we write $\mathbb{E}_A$ for a step t as

$$\begin{aligned} \Delta_{t+1}(S, z) &\leqslant \left(1 - \frac{1}{m}\right) \Delta_t(S, z) \\ &+ \frac{1}{m} \left(\Delta_t(S, z) + 2\alpha_t \|\nabla f(\boldsymbol{w}_{S,t}, z_{j_t})\|\right) \\ &= \Delta_t(S, z) + \frac{2\alpha_t \|\nabla f(\boldsymbol{w}_{S,t}, z_{j_t})\|}{m} . \end{aligned}$$

Unraveling the recursion from T to t_0 and plugging the above into (9) yields

$$\begin{aligned} &\mathop{\mathbb{E}}_{A} \mathop{\mathbb{E}}_{S,z} \left[\delta_T(S, z)\right] \\ &\leqslant \frac{2}{m} \sum_{t=t_0+1}^{T} \alpha_t \mathop{\mathbb{E}}_{S,z} \left[\|\nabla f(\boldsymbol{w}_t, z_{j_t})\|\right] + \mathop{\mathbb{E}}_{S,A} \left[R(A_S)\right] \frac{t_0}{m} . \end{aligned}$$

This completes the proof. □

Next statement is a simple consequence of Theorem 1 and Lemma 4.

Proof of Theorem 3. Consider Theorem 1 and set $t_0 = 0$.

$$\epsilon(\mathcal{D}, \boldsymbol{w}_1) \leqslant \frac{2}{m} \sum_{t=1}^{T} \alpha_t \mathop{\mathbb{E}}_{S,z} \left[\|\nabla f(\boldsymbol{w}_{S,t}, z_{j_t})\|\right] . \tag{10}$$

Bounding the sum using Lemma 4 recalling that $\alpha_t = c/\sqrt{t}$, we get

$$\begin{aligned} &\mathop{\mathbb{E}}_{S} \left[\sum_{t=1}^{T} \alpha_t \|\nabla f(\boldsymbol{w}_t, z_{j_t})\|\right] \\ &\leqslant 2\sqrt{\left(\sum_{t=1}^{T} \alpha_t\right) \left(R(\boldsymbol{w}_1) - R^\star + \frac{\beta\sigma^2}{2} \sum_{t=1}^{T} \alpha_t^2\right)} \\ &+ \sigma \sum_{t=1}^{T} \alpha_t \\ &\leqslant 2\sqrt{2c} \cdot \sqrt[4]{T} \cdot \sqrt{R(\boldsymbol{w}_1) - R^\star} \\ &+ 2c\sigma \left(\sqrt[4]{T} \sqrt{\frac{\beta}{2}} + \sqrt{T}\right) . \end{aligned}$$

Combining above with (10) completes the proof. □

1.3. Non-convex Losses

Our proof of a stability bound for non-convex loss functions, Theorem 4 (in the submission file), follows a general outline of (Hardt et al., 2016, Theorem 3.8). Namely, the outputs of SGD run on a training set S and its perturbed version $S^{(i)}$ will not differ too much, because by the time a perturbation is encountered, the step size has already decayed enough. So, on the one hand, stabilization is enforced by the diminishing the step size, and on the other hand, by how much updates expand the distance between the gradients after the perturbation. Since (Hardt et al., 2016) work with uniform stability, they capture the expansiveness of post-perturbation update by the Lipschitzness of the gradient. In combination with a recursive argument, their bound has exponential dependency on the Lipschitz constant of the gradient. We argue that the Lipschitz continuity of the gradient can be too pessimistic in general. Instead, we rely on a local

data-driven argument: considering that we initialize SGD at point $\boldsymbol{w}_1$, how much do updates expand the gradient under the distribution of interest? The following crucial lemma characterizes such behavior in terms of the curvature at $\boldsymbol{w}_1$.

Lemma 6. *Assume that the loss function $f(\cdot, z)$ is β-smooth and that its Hessian is ρ-Lipschitz. Then,*

$$\left\|G_t(\boldsymbol{w}_{S,t}) - G_t(\boldsymbol{w}_{S^{(i)},t})\right\| \leqslant (1 + \alpha_t \xi_t(S,z))\, \delta_t(S,z) \tag{11}$$

where

$$\begin{aligned}\xi_t(S,z) &:= \left\|\nabla^2 f(\boldsymbol{w}_1, z_t)\right\|_2 \\ &+ \frac{\rho}{2}\left\|\sum_{k=1}^{t-1} \alpha_k \nabla f(\boldsymbol{w}_{S,k}, z_k)\right\| \\ &+ \frac{\rho}{2}\left\|\sum_{k=1}^{t-1} \alpha_k \nabla f(\boldsymbol{w}_{S^{(i)},k}, z_{k'})\right\| .\end{aligned}$$

Furthermore, for any $t \in [T]$,

$$\begin{aligned}\mathop{\mathbb{E}}_{S,z}\left[\xi_t(S,z)\right] &\leqslant \mathop{\mathbb{E}}_{S,z}\left[\left\|\nabla^2 f(\boldsymbol{w}_1, z_t)\right\|_2\right] \\ &+2\rho\sqrt{(R(\boldsymbol{w}_1) - R^\star)\, c(1+\ln(T))} \\ &+\rho\sigma\left(\sqrt{2c\beta} + c(1+\ln(T))\right) .\end{aligned}$$

Proof. Recall that the randomness of the algorithm is realized through sampling without replacement from the uniform distribution over $[m]$. Apart from that we will not be concerned with the randomness of the algorithm, and given the set of random variables $\{j_i\}_{i=1}^m$, for brevity we will use indexing notation $z_1, z_2, \ldots, z_m$ to indicate $z_{j_1}, z_{j_2}, \ldots, z_{j_m}$. Next, let $S^{(i)} = \{z_i'\}_{i=1}^m$, and introduce a shorthand notation $f_k(\boldsymbol{w}) = f(\boldsymbol{w}, z_k)$ and $f_{k'}(\boldsymbol{w}) = f(\boldsymbol{w}, z_k')$. We start by applying triangle inequality to get

$$\begin{aligned}\left\|G_t(\boldsymbol{w}_{S,t}) - G_t(\boldsymbol{w}_{S^{(i)},t})\right\| &\leqslant \left\|\boldsymbol{w}_{S,t} - \boldsymbol{w}_{S^{(i)},t}\right\| \\ &+ \alpha_t \left\|\nabla f_t(\boldsymbol{w}_{S,t}) - \nabla f_t(\boldsymbol{w}_{S^{(i)},t})\right\| .\end{aligned}$$

In the following we will focus on the second term of r.h.s. above. Given SGD outputs $\boldsymbol{w}_{S,t}$ and $\boldsymbol{w}_{S^{(i)},t}$ with $t > i$, our goal here is to establish how much do gradients grow apart with every new update. This behavior can be characterized assuming that gradient is Lipschitz continuous, however, we conduct a local analysis. Specifically, we observe how much do updates expand gradients, given that we start at some point $\boldsymbol{w}_1$ under the data-generating distribution. So, instead of the Lipschitz constant, expansiveness rather depends on the curvature around $\boldsymbol{w}_1$. On the other hand, we are dealing with outputs at an arbitrary time step t, and therefore we first have to relate them to the initialization point $\boldsymbol{w}_1$. We do so by using the gradient update rule and telescopic sums, and conclude that this relationship is controlled by the sum of gradient norms along the update path. We further establish that this sum is controlled by the risk of $\boldsymbol{w}_1$ up to the noise of stochastic gradients, through stationary-point result of Lemma 4. Thus, the proof consists of two parts: 1) Decomposition into curvature and gradients along the update path, and 2) bounding those gradients.

1) Decomposition. Introduce $\boldsymbol{\delta}_t := \boldsymbol{w}_{S^{(i)},t} - \boldsymbol{w}_{S,t}$. By Taylor theorem we get that

$$\begin{aligned}&\nabla f_t(\boldsymbol{w}_{S,t}) - \nabla f_t(\boldsymbol{w}_{S^{(i)},t}) = \nabla^2 f_t(\boldsymbol{w}_1)\boldsymbol{\delta}_t \\ &+ \int_0^1 \left(\nabla^2 f_t(\boldsymbol{w}_{S,t} + \tau\boldsymbol{\delta}_t) - \nabla^2 f_t(\boldsymbol{w}_1)\right) \mathrm{d}\tau \boldsymbol{\delta}_t .\end{aligned}$$

Taking norm on both sides, applying triangle inequality, Cauchy-Schwartz inequality, and assuming that Hessians are ρ-Lipschitz we obtain

$$\begin{aligned}&\left\|\nabla f_t(\boldsymbol{w}_{S,t}) - \nabla f_t(\boldsymbol{w}_{S^{(i)},t})\right\| \\ &\leqslant \rho \int_0^1 \left\|\boldsymbol{w}_{S,t} - \boldsymbol{w}_1 + \tau\boldsymbol{\delta}_t\right\| \mathrm{d}\tau \left\|\boldsymbol{\delta}_t\right\| + \left\|\nabla^2 f_t(\boldsymbol{w}_1)\right\| \left\|\boldsymbol{\delta}_t\right\| .\end{aligned} \tag{12}$$

2) Bounding gradients. Using telescoping sums and SGD update rule we get that

$$\begin{aligned}&\boldsymbol{w}_{S,t} - \boldsymbol{w}_1 + \tau\boldsymbol{\delta}_t \\ &= \boldsymbol{w}_{S,t} - \boldsymbol{w}_1 + \tau\left(\boldsymbol{w}_{S^{(i)},t} - \boldsymbol{w}_1 + \boldsymbol{w}_1 - \boldsymbol{w}_{S,t}\right) \\ &= \sum_{k=1}^{t-1}\left(\boldsymbol{w}_{S,k+1} - \boldsymbol{w}_{S,k}\right) + \tau\sum_{k=1}^{t-1}\left(\boldsymbol{w}_{S^{(i)},k+1} - \boldsymbol{w}_{S^{(i)},k}\right) \\ &- \tau\sum_{k=1}^{t-1}\left(\boldsymbol{w}_{S,k+1} - \boldsymbol{w}_{S,k}\right) \\ &= (\tau - 1)\sum_{k=1}^{t-1}\alpha_k \nabla f_k(\boldsymbol{w}_{S,k}) - \tau\sum_{k=1}^{t-1}\alpha_k \nabla f_{k'}(\boldsymbol{w}_{S^{(i)},k}) .\end{aligned}$$

Plugging above into the integral of (12) we have

$$\begin{aligned}&\int_0^1 \left\|\sum_{k=1}^{t-1}\alpha_k\left((\tau-1)\nabla f_k(\boldsymbol{w}_{S,k}) - \tau\nabla f_{k'}(\boldsymbol{w}_{S^{(i)},k})\right)\right\| \mathrm{d}\tau \\ &\leqslant \frac{1}{2}\left\|\sum_{k=1}^{t-1}\alpha_k \nabla f_k(\boldsymbol{w}_{S,k})\right\| + \frac{1}{2}\left\|\sum_{k=1}^{t-1}\alpha_k \nabla f_{k'}(\boldsymbol{w}_{S^{(i)},k})\right\| \\ &\leqslant \frac{1}{2}\sum_{k=1}^{t-1}\alpha_k\left\|\nabla f_k(\boldsymbol{w}_{S,k})\right\| + \frac{1}{2}\sum_{k=1}^{t-1}\alpha_k\left\|\nabla f_{k'}(\boldsymbol{w}_{S^{(i)},k})\right\| .\end{aligned}$$

Plugging this result back into (12) completes the proof of the first statement. The second statement comes from Lemma 4 with $\alpha_t = c/t$. □

Next, we need the following statement to prove our stability bound.

Proposition 1 (Bernstein-type inequality)**.** *Let Z be a zero-mean real-valued r.v., such that $|Z| \leqslant b$ and $\mathbb{E}[Z^2] \leqslant \sigma^2$. Then for all $|c| \leqslant \frac{1}{2b}$, we have that $\mathbb{E}\left[e^{cZ}\right] \leqslant e^{c^2\sigma^2}$.*

Proof. Stated inequality is a consequence of a Bernstein-type inequality for moment generating functions, Theorem 2.10 in (Boucheron et al., 2013). Observe that zero-centered r.v. Z bounded by b satisfies Bernstein's condition, that is

$$|\,\mathbb{E}[(Z-\mathbb{E}[Z])^q]| \leqslant \frac{q!}{2}\sigma^2 b^{k-2} \qquad \text{for all integers } q \geqslant 3\,.$$

This in turn satisfies condition for Bernstein-type inequality stating that

$$\mathbb{E}\left[\exp\left(c(Z-\mathbb{E}[Z])\right)\right] \leqslant \exp\left(\frac{c^2\sigma^2/2}{1-b|c|}\right)\,.$$

Choosing $|c| \leqslant \frac{1}{2b}$ verifies the statement. □

Now we are ready to prove Theorem 4, which bounds the $\epsilon(\mathcal{D}, \boldsymbol{w}_1)$-on-average stability of SGD.

Proof of Theorem 4. For brevity denote

$$r := \mathop{\mathbb{E}}_{S,A}\left[R(A_S)\right]$$

and

$$\Delta_t(S,z) := \mathop{\mathbb{E}}_{A}\left[\delta_t(S,z) \mid \delta_{t_0}(S,z) = 0\right]\,.$$

By Lemma 5, for all $t_0 \in [m]$,

$$\mathop{\mathbb{E}}_{S,z}\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T},z) - f(\boldsymbol{w}_{S^{(i)},T},z)\right] \tag{13}$$

$$\leqslant L \mathop{\mathbb{E}}_{S,z}\left[\Delta_T(S,z)\right] + r\frac{t_0}{m}\,. \tag{14}$$

Most of the proof is dedicated to bounding the first term in (14). We deal with this similarly as in (Hardt et al., 2016). Specifically, we state the bound on $\Delta_T(S,z)$ by using a recursion. In our case, however, we also have an expectation w.r.t. the data, and to avoid complications with dependencies, we first unroll the recursion for the random quantities, and only then take the expectation. At this point the proof crucially relies on the product of exponentials arising from the recursion, and all relevant random quantities end up inside of them. We alleviate this by Proposition 1. Finally, we conclude by minimizing (14) w.r.t. t_0. Thus we have three steps: 1) recursion, 2) bounding $\mathbb{E}[\exp(\cdots)]$, and 3) tuning of t_0.

1) Recursion. We begin by stating the bound on $\Delta_T(S,z)$ by recursion. Thus we will first state the bound on $\Delta_{t+1}(S,z)$ in terms of $\Delta_t(S,z)$, and other relevant quantities and then unravel the recursion. As in the convex case, we distinguish two cases: 1) SGD encounters the perturbed point at step t, that is $t = i$, and 2) the current point is the same in S and $S^{(i)}$, so $t \neq i$. For the first case, we will use worst-case boundedness of G_t, Corollary 1, that is, $\|G_t(\boldsymbol{w}_{S,t}) - G_t(\boldsymbol{w}_{S^{(i)},t})\| \leqslant \delta_t(S,z) + 2\alpha_t L$. To handle the second case we will use Lemma 6, namely,

$$\left\|G_t(\boldsymbol{w}_{S,t}) - G_t(\boldsymbol{w}_{S^{(i)},t})\right\| \leqslant \left(1+\alpha_t\xi_t(S,z)\right)\delta_t(S,z)\,.$$

In addition, as a safety measure we will also take into account that the gradient update rule is at most $(1+\alpha_t\beta)$-expansive by Lemma 1. So we will work with the function $\psi_t(S,z) := \min\left\{\xi_t(S,z), \beta\right\}$ instead of $\xi_t(S,z)$. and decompose the expectation w.r.t. A for a step t. Noting that SGD encounters the perturbed example with probability $\frac{1}{m}$,

$$\begin{aligned}
\Delta_{t+1}(S,z) &\leqslant \left(1-\frac{1}{m}\right)\left(1+\alpha_t\psi_t(S,z)\right)\Delta_t(S,z) \\
&+ \frac{1}{m}\left(2\alpha_t L + \Delta_t(S,z)\right) \\
&= \left(1+\left(1-\frac{1}{m}\right)\alpha_t\psi_t(S,z)\right)\Delta_t(S,z) + \frac{2\alpha_t L}{m} \\
&\leqslant \exp\left(\alpha_t\psi_t(S,z)\right)\Delta_t(S,z) + \frac{2\alpha_t L}{m}\,,
\end{aligned} \tag{15}$$

where the last inequality follows from $1+x \leqslant \exp(x)$. This inequality is not overly loose for $x \in [0,1]$, and, in our case it becomes instrumental in handling the recursion.

Now, observe that relation $x_{t+1} \leqslant a_t x_t + b_t$ with $x_{t_0} = 0$ unwinds from T to t_0 as $x_T \leqslant \sum_{t=t_0+1}^{T} b_t \prod_{k=t+1}^{T} a_k$. Consequently, having $\Delta_{t_0}(S,z) = 0$, we unwind (15) to get

$$\begin{aligned}
\Delta_T(S,z) &\leqslant \sum_{t=t_0+1}^{T}\left(\prod_{k=t+1}^{T}\exp\left(\frac{c\psi_k(S,z)}{k}\right)\right)\frac{2cL}{mt} \\
&= \sum_{t=t_0+1}^{T}\exp\left(c\sum_{k=t+1}^{T}\frac{\psi_k(S,z)}{k}\right)\frac{2cL}{mt}\,.
\end{aligned} \tag{16}$$

2) Bounding $\mathbb{E}[\exp(\cdots)]$. We take expectation w.r.t. S and z on both sides and focus on the expectation of the exponential in (16). First, introduce $\mu_k := \mathbb{E}_{S,z}[\psi_k(S,z)]$, and proceed as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,z}\left[\exp\left(c\sum_{k=t+1}^{T}\frac{\psi_k(S,z)}{k}\right)\right] \\
&= \mathop{\mathbb{E}}_{S,z}\left[\exp\left(c\sum_{k=t+1}^{T}\frac{\psi_k(S,z)-\mu_k}{k}\right)\right]\exp\left(c\sum_{k=t+1}^{T}\frac{\mu_k}{k}\right)\,.
\end{aligned} \tag{17}$$

Observe that zero-mean version of $\psi_k(S,z)$ is bounded as

$$\sum_{k=t+1}^{T} \frac{|\psi_k(S,z)-\mu_k|}{k} \leqslant 2\beta \ln(T) \,,$$

and assume the setting of c as $c \leqslant \frac{1}{2(2\beta\ln(T))^2}$. By Proposition 1, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(c\sum_{k=t+1}^{T}\frac{\psi_k(S,z)-\mu_k}{k}\right)\right] \\
&\leqslant \exp\left(c^2\,\mathbb{E}\left[\left(\sum_{k=t+1}^{T}\frac{\psi_k(S,z)-\mu_k}{k}\right)^2\right]\right) \\
&= \exp\left(\frac{c}{2}\,\mathbb{E}\left[\left(\frac{1}{2\beta\ln(T)}\sum_{k=t+1}^{T}\frac{\psi_k(S,z)-\mu_k}{k}\right)^2\right]\right) \\
&\leqslant \exp\left(\frac{c}{2}\,\mathbb{E}\left[\left|\sum_{k=t+1}^{T}\frac{\psi_k(S,z)-\mu_k}{k}\right|\right]\right) \\
&\leqslant \exp\left(\frac{c}{2}\sum_{k=t+1}^{T}\frac{\mathbb{E}\left[|\psi_k(S,z)-\mu_k|\right]}{k}\right) \\
&\leqslant \exp\left(c\sum_{k=t+1}^{T}\frac{\mu_k}{k}\right) \,.
\end{aligned}$$

Getting back to (17) we conclude that

$$\mathop{\mathbb{E}}_{S,z}\left[\exp\left(c\sum_{k=t+1}^{T}\frac{\psi_k(S,z)}{k}\right)\right] \leqslant \exp\left(c\sum_{k=t+1}^{T}\frac{2\mu_k}{k}\right) \,. \tag{18}$$

Next, we give an upper-bound on μ_k, that is $\mu_k \leqslant \min\{\beta, \mathbb{E}_{S,z}[\xi_k(S,z)]\}$. Finally, we bound $\mathbb{E}_{S,z}[\xi_k(S,z)]$ using the second result of Lemma 6, which holds for any $k \in [T]$, to get that $\mu_k \leqslant \gamma$, with γ defined in the statement of the theorem.

3) Tuning of t_0. Now we turn our attention back to (16). Considering that we took an expectation w.r.t. the data, we use (18) and the fact that $\mu_k \leqslant \gamma$ to get that

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,z}[\Delta_T(S,z)] &\leqslant \sum_{t=t_0+1}^{T}\exp\left(2c\gamma\sum_{k=t+1}^{T}\frac{1}{k}\right)\frac{2cL}{mt} \\
&\leqslant \sum_{t=t_0+1}^{T}\exp\left(2c\gamma\ln\left(\frac{T}{t}\right)\right)\frac{2cL}{mt} \\
&= \frac{2cL}{m}\left(T^{2c\gamma}\right)\sum_{t=t_0+1}^{T} t^{-2c\gamma-1} \\
&\leqslant \frac{1}{2c\gamma}\frac{2cL}{m}\left(\frac{T}{t_0}\right)^{2c\gamma} \,.
\end{aligned}$$

Plug the above into (14) to get

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,z}\mathop{\mathbb{E}}_{A}\left[f(\boldsymbol{w}_{S,T},z) - f(\boldsymbol{w}_{S^{(i)},T},z)\right] \\
&\leqslant \frac{L^2}{\gamma m}\left(\frac{T}{t_0}\right)^{2c\gamma} + r\frac{t_0}{m} \,.
\end{aligned} \tag{19}$$

Let $q = 2c\gamma$. Then, setting

$$t_0 = \left(\frac{2cL^2}{r}\right)^{\frac{1}{1+q}} T^{\frac{q}{1+q}}$$

minimizes (19). Plugging t_0 back we get that (19) equals to

$$\frac{1+\frac{1}{q}}{m}\left(2cL^2\right)^{\frac{1}{1+q}}(rT)^{\frac{q}{1+q}} \,.$$

This completes the proof. □

1.3.1. Optimistic Rates for Learning with Non-convex Loss Functions

Next we will prove an optimistic bound based on Theorem 4, in other words, the bound that demonstrates fast convergence rate subject to the vanishing empirical risk. First we will need the following technical statement.

Lemma 7. *(Cucker & Zhou, 2007, Lemma 7.2) Let $c_1, c_2, \ldots, c_l > 0$ and $s > q_1 > q_2 > \ldots > q_{l-1} > 0$. Then the equation*

$$x^s - c_1 x^{q_1} - c_2 x^{q_2} - \cdots - c_{l-1}x^{q_{l-1}} - c_l = 0$$

has a unique positive solution $x^\star$. In addition,

$$x^\star \leqslant \max\left\{(lc_1)^{\frac{1}{s-q_1}}, (lc_2)^{\frac{1}{s-q_2}}, \cdots, (lc_{l-1})^{\frac{1}{s-q_{l-1}}}, (lc_l)^{\frac{1}{s}}\right\}.$$

Next we prove a useful technical lemma similarly as in (Orabona, 2014, Lemma 7).

Lemma 8. *Let $a, c > 0$ and $0 < \alpha < 1$. Then the inequality*

$$x - ax^\alpha - c \leqslant 0$$

implies

$$x \leqslant \max\left\{2^{\frac{\alpha}{1-\alpha}}a^{\frac{1}{1-\alpha}}, (2c)^\alpha a\right\} + c \,.$$

Proof. Consider a function $h(x) = x - ax^\alpha - c$. Applying Lemma 7 with $s=1$, $l=2$, $c_1 = a$, $c_2 = c$, and $q_1 = \alpha$ we get that $h(x) = 0$ has a unique positive solution $x^\star$ and

$$x^\star \leqslant \max\left\{(2a)^{\frac{1}{1-\alpha}}, 2c\right\} \,. \tag{20}$$

Moreover, the inequality $h(x) \leqslant 0$ is verified for $x = 0$, and $\lim_{x\to+\infty} h(x) = +\infty$, so we have that $h(x) \leqslant 0$ implies

$x \leqslant x^\star$. Now, using this fact and the fact that $h(x^\star) = 0$, we have that

$$x \leqslant x^\star = a\,(x^\star)^\alpha + c\,,$$

and upper-bounding $x^\star$ by (20) we finally have

$$x \leqslant a \max\left\{(2a)^{\frac{\alpha}{1-\alpha}}, (2c)^\alpha\right\} + c\,,$$

which completes the proof. □

Proof of Corollary 2. Consider Theorem 4 and observe that it verifies condition of Lemma 8 with $x = \mathbb{E}_{S,A}\left[R(A_S)\right]$, $c = \mathbb{E}_{S,A}\left[\widehat{R}_S(A_S)\right]$, $\alpha = \frac{c\gamma}{1+c\gamma}$, and

$$a = \frac{1+\frac{1}{c\gamma}}{m}\left(2cL^2\right)^{\frac{1}{1+c\gamma}} T^{\frac{c\gamma}{1+c\gamma}}\,.$$

Note that $\alpha/(1-\alpha) = c\gamma$ and $1/(1-\alpha) = 1 + c\gamma$. Then, we obtain that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,A}\left[R(A_S) - \widehat{R}_S(A_S)\right] \\
&\leqslant \max\left\{2^{c\gamma}\left(\frac{1+\frac{1}{c\gamma}}{m}\right)^{1+c\gamma}\left(2cL^2\right)T^{c\gamma},\right. \\
&\left.\left(2\mathop{\mathbb{E}}_{S,A}\left[\widehat{R}_S(A_S)\right]\right)^{\frac{c\gamma}{1+c\gamma}}\left(\frac{1+\frac{1}{c\gamma}}{m}\left(2cL^2\right)^{\frac{1}{1+c\gamma}} T^{\frac{c\gamma}{1+c\gamma}}\right)\right\} \\
&= \max\left\{\left(2+\frac{2}{c\gamma}\right)^{1+c\gamma}\left(cL^2\right)\left(\frac{T^{c\gamma}}{m^{1+c\gamma}}\right),\right. \\
&\left.\frac{1+\frac{1}{c\gamma}}{m}\left(2cL^2\right)^{\frac{1}{1+c\gamma}}\left(2\mathop{\mathbb{E}}_{S,A}\left[\widehat{R}_S(A_S)\right]\cdot T\right)^{\frac{c\gamma}{1+c\gamma}}\right\}.
\end{aligned}$$

This completes the proof. □

Proof of Proposition 1. Consider minimizing the bound given by Corollary 1 (in the submission file) over a discrete set of source hypotheses $\{\boldsymbol{w}_k^{\text{src}}\}_{k=1}^K$,

$$\begin{aligned}
&\min_{k\in[K]} \epsilon(\mathcal{D}, \boldsymbol{w}_k^{\text{src}}) \\
&\leqslant \min_{k\in[K]} \mathcal{O}\left(\frac{1+\frac{1}{c\gamma_k}}{m}\left(R(\boldsymbol{w}_k^{\text{src}})\cdot T\right)^{\frac{c\gamma_k}{1+c\gamma_k}}\right), \qquad (21)
\end{aligned}$$

and let

$$\begin{aligned}
\gamma_k &= \mathcal{O}\left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\|\nabla^2 f(\boldsymbol{w}_k^{\text{src}}, z)\|_2\right] + \sqrt{R(\boldsymbol{w}_k^{\text{src}})}\right), \\
\widehat{\gamma}_k &= \frac{1}{m}\sum_{i=1}^m \|\nabla^2 f(\boldsymbol{w}_k^{\text{src}}, z_i)\|_2 + \sqrt{\widehat{R}_S(\boldsymbol{w}_k^{\text{src}})}\,.
\end{aligned}$$

By Hoeffding inequality, with high probability, we have that $|\gamma_k - \widehat{\gamma}_k| \leqslant \mathcal{O}\left(\frac{1}{\sqrt[4]{m}}\right)$. Now we further upper bound (21) by upper bounding $R(\boldsymbol{w}_k^{\text{src}})$ and apply union bound to get

$$\begin{aligned}
&\min_{k\in[K]} \epsilon(\mathcal{D}, \boldsymbol{w}_k^{\text{src}}) \\
&\leqslant \min_{k\in[K]} \mathcal{O}\left(\left(1+\frac{1}{c\widehat{\gamma}_k^-}\right)\widehat{R}_S(\boldsymbol{w}_k^{\text{src}})^{\frac{c\widehat{\gamma}_k^+}{1+c\widehat{\gamma}_k^+}}\cdot\frac{\sqrt{\log(K)}}{m^{\frac{1}{1+c\widehat{\gamma}_k^+}}}\right),
\end{aligned}$$

where $\widehat{\gamma}_k^+ = \widehat{\gamma}_k \pm \frac{1}{\sqrt[4]{m}}$. This completes the proof. □

References

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Cucker, F. and Zhou, D. X. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learing (ICML)*, 2016.

Orabona, F. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Conference on Neural Information Processing Systems (NIPS)*, pp. 1116–1124, 2014.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.