Sharpening high probability generalization bounds for almost everywhere stable learning algorithms

David Wu and Eshaan Nichani

May 19, 2021

1 Introduction

In the analysis of generalization bounds for machine learning algorithms, a central theme is the notion of algorithmic stability — a measurement of how sensitive a learning algorithm is to perturbations to its input. The notion that generalization is closely tied to stability is rather intuitive; indeed, under most formulations *average-case stability* is equivalent to expected generalization error, but is often unwieldy to work with. Introduced in Bousquet and Elisseeff [2], the *uniform stability* of an algorithm quantifies how much – in the worst case – an algorithm depends on the training data, and thus can be used to bound the expected generalization error of a learning algorithm. They also presented a high-probability generalization bound over the randomness of the data; however the presented bound is only meaningful when the stability parameter γ is $O(1/n)$, where n is the number of training samples.

This high-probability generalization bound was improved in Feldman and Vondrak [4], which allows for stability up to $O(1/\sqrt{n})$. Such an improvement was notable as it allowed for generalization bounds in a number of algorithms with $O(1/\sqrt{n})$-stability such as Empirical Risk Minimization [8] and Stochastic Gradient Descent [5]. A streamlined proof which removes another $\log n$ factor was recently shown in Bousquet et al. [3].

In this paper we attempt to sharpen with high probability generalization bounds for learning algorithms that do not quite possess the worst-case notion of uniform stability. To that end, we consider the notion of *almost-everywhere stability*, first studied by Kutin and Niyogi [6]. Informally speaking, almost everywhere stability captures the scenario where a learning algorithm is stable with high probability over the input, but occasionally we get unlucky and potentially have a much larger stability coefficient. Depending on how stability is measured and over what events the probability is taken over, one obtains a whole zoo of weakened notion of stability. Intuitively, the effectiveness of the generalization bound depends on the relative sizes of the failure probability and the sizes of the two stability coefficients. In a certain regime, Kutin and Niyogi [6] are able to show with high probability generalization when the stability coefficient is $O(1/n)$ and the failure probability is $O(2^{-\Omega(n)})$.

The paper is organized as follows. In Section 2, we introduce the notation for almost-everywhere algorithmic stability and some key technical lemmas useful for the proof. In Section 3, we prove analogues of the main results in Bousquet et al. [3] and arrive at our improved generalization bounds for almost-everywhere stable learning algorithms. In Section 4, we construct a few toy examples where our results are applicable and highlight some of the difficulties with constructing applications. We conclude in Section 5 with the takeaways of our analysis and potential future directions.

2 Preliminaries

First, we introduce the notation to facilitate the discussion. We find the notation in the literature to be rather varied and confusing, so we attempt to unify the discussion with judicious choice of notation.

A learning algorithm is a function that maps training sets $S \in \mathcal{Z}^n$ to a learned hypothesis $A_S \in \mathcal{H}$. We measure the performance of the hypothesis via a uniformly bounded loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, L]$. The true risk is defined as $R(A_S) = \mathbb{E}_z[\ell(A_S(x), y)]$, and the empirical risk is defined as $R_{\text{emp}}(A_S) = \frac{1}{n}\sum_{i=1}^{n} \ell(A_S(x_i), y_i)$.

Central to the semantics of stability is precisely what a learning algorithm is stable against, and in what sense stability is measured. To that end, we introduce the notation $S^{i \leftarrow z}$ to denote the same training set with sample i replaced with $z = (x, y)$. We say that a learning algorithm is *uniformly stable* with stability parameter γ if

$$\sup_{i,z,z',S} |\ell(A_S(x), y) - \ell(A_{S^{i \leftarrow z'}}(x), y)| \leq \gamma.$$

Note that uniform stability is asking quite a bit out of our learning algorithm. It needs to be robust not only to training sets S, but also to adversarial corruptions z' and adversarial test points z. It was recently shown Feldman and Vondrak [4], Bousquet et al. [3] that this notion of stability yields nearly optimal high probability bounds on the generalization gap.

Theorem 2.1 (Bousquet et al. [3])**.** *If A_S is uniformly stable with parameter γ, then with probability $1 - \delta$ over the choice of S,*

$$|R(A_S) - R_{\text{emp}}(A_s)| \leq c_1 \gamma \log n \log\left(\frac{1}{\delta}\right) + c_2 L \sqrt{\frac{\log(1/\delta)}{n}},$$

where c_1 and c_2 are absolute constants (i.e. independent of n, δ, L).

To relax the notion of uniform stability, we introduce the following two notions of almost-everywhere hypothesis stability, à la Kutin and Niyogi [6].

Definition 2.2 (Strong hypothesis stability)**.** *We say that A_S is (γ, δ) strong hypothesis stable if with probability at least $1 - \delta$ over the choice of S, we have*

$$\sup_{i,z,z'} |\ell(A_S(x), y) - \ell(A_{S^{i \leftarrow z'}}(x), y)| \leq \gamma.$$

By further relaxing the requirement that the algorithm must be robust to adversarial data corruption, we obtain weak hypothesis stability.

Definition 2.3 (Weak hypothesis stability)**.** *We say that A_S is (γ, δ) weak hypothesis stable if with probability at least $1 - \delta$ over the choice of S and z', we have*

$$\sup_{i,z} |\ell(A_S(x), y) - \ell(A_{S^{i \leftarrow z'}}(x), y)| \leq \gamma.$$

Uniform stability plays well with bounded differences, since if we view the inputs as deterministic the loss function satisfies a bounded differences property. Since we have relaxed uniform stability, we also need an analogous relaxation of bounded differences. To that end, we have the following notion.

Definition 2.4 (Strong bounded differences). *Let $f(S) = f(Z_1, \ldots, Z_n)$ be a measurable function of random variables. We say that f is strongly difference-bounded by (M, γ, δ) if with probability at least $1 - \delta$ over the choice of S, any corruption $S^{i \leftarrow z'}$ satisfies*

$$\left| f(S) - f(S^{i \leftarrow z'}) \right| \leq \gamma.$$

Furthermore, we have that for any $S, S^{i \leftarrow z'}$ that

$$\left| f(S) - f(S^{i \leftarrow z'}) \right| \leq M,$$

that is, M is a uniform bound on the change one stability of f.

The weak difference bounded property can be defined similarly, where we also have randomness in the data corruption z'. With these definitions in mind, we can now state a corresponding result for the bounded differences inequality (also known in the literature as McDiarmid/Azuma-Hoeffding inequality).

Theorem 2.5 (Kutin and Niyogi [6]). *Suppose f is strongly difference-bounded by (M, γ, δ). Let $K \geq \gamma$. Then for any $\tau > 0$ we have*

$$\mathbb{P}[|f - \mathbb{E}[f]| \geq \tau] \leq 2 \exp\left(\frac{-\tau^2}{8n\gamma^2}\right) + \frac{2nM\delta}{\gamma}.$$

Given a uniformly stable learning algorithm, one can apply the (uniform) bounded differences inequality to obtain a generalization bound [2]. However, these bounds are only optimal in the regime where $\gamma = O(1/n)$, which is too restrictive in certain settings. For example, regularized convex SGD yields a stability parameter $O(\log n / n^{1/2})$ [8]. In the recent work of Feldman and Vondrak [4], this gap was closed, with a nearly optimal generalization bound (compared to the sample efficiency) up to $\log n$ factors. Following the breakthrough of Feldman and Vondrak [4], an alternative simplified proof was given by Bousquet et al. [3]. The key difference in proof strategy is to work instead with moment bounds, and then translate to tail bounds. As the following standard result shows, tail bounds are equivalent to moment bounds, thus allowing us to interchange between the two.

Lemma 2.6 (Equivalence of tails and moments [3, Lemma 1]). *If for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$ that $|Y| \leq a\sqrt{\log(e/\delta)} + b\log(e/\delta)$, then for any $p \geq 1$, we have*

$$\|Y\|_p \leq 3\sqrt{p}a + 9pb.$$

Conversely, if $\|Y\|_p \leq \sqrt{p}a + pb$ for $p \geq 1$, then with probability at least $1 - \delta$ we have

$$|Y| \leq e(a\sqrt{\log(e/\delta)} + b\log(e/\delta)).$$

Directly converting theorem 2.5 into a moment inequality is a bit troublesome because of the second term's independence of τ. However, for our purposes, not only will the change-one differences be uniformly bounded by M, but also we have $|f(S)| \leq M$ a.s.. In Bousquet et al. [3], they note that a moment form of the bounded differences inequality follows immediately from Theorem 15.4 of Boucheron et al. [1]. Unfortunately, the lack of an almost sure bound on the stability complicates the details, so that the proof is not quite as immediate. However, the proof strategy of Theorem 15.4 can be accommodated for our setting, at the cost of more tedious analysis of Theorem 15.4.

Lemma 2.7 (Moment form of high probability bounded differences inequality)**.** *Suppose f is strongly difference bounded by (M, γ, δ), and further that f is bounded a.s. by M. Then*

$$\|f(S) - \mathbb{E}[f(S)]\|_q \leq 2(\gamma + \delta^{\frac{1}{2(q-1)}} M)\sqrt{qn}.$$

Similarly, if f is weakly difference bounded by (M, γ, δ) and f is bounded a.s. by M, then

$$\|f(S) - \mathbb{E}[f(S)]\|_q \leq 2(\gamma + 2\delta^{\frac{1}{4(q-1)}} M)\sqrt{qn}.$$

Proof. We follow the notation and proof of Theorem 15.4 in Boucheron et al. [1]. First define

$$V^+ = \sum_{i=1}^{n} \mathbb{E}_{Z'}[(f(S) - f(S^{i \leftarrow z'})_+^2].$$

The key is to use Lemma 15.1 of Boucheron et al. [1], the generalization of the Efron-Stein inequality, to bound higher order moments. Define $m_q = \|(f(S) - \mathbb{E}[f(S)])_+\|_q$. Also, we will find it convenient to introduce the notation

$$c_k = \gamma^k(1 - \delta) + M^k\delta.$$

The Efron-Stein inequality precisely states that $m_2^2 \leq nc_2$.

Then Lemma 15.1 implies that

$$m_q^q \leq m_{q-1}^q + (q-1)\,\mathbb{E}[V^+(f(S) - \mathbb{E}[f(S)])_+^{q-2}].$$

Applying Hölder's inequality with $(q-1, \frac{q-1}{q-2})$ to the latter expectation, we obtain

$$\mathbb{E}[V^+(f(S) - \mathbb{E}[f(S)])_+^{q-2}] \leq \mathbb{E}[(V^+)^{q-1}]^{1/(q-1)} m_{q-1}^{q-2}.$$

Expanding out the expectation, noting that we are summing over all $(q-1)$ tuples and applying strong bounded differences yields an upper bound of

$$n(\sup_{i,z'}(f(S) - f(S^{i \leftarrow z'}))^{2(q-1)})^{1/(q-1)} m_{q-1}^{q-2} \leq nc_{2(q-1)}^{1/(q-1)} m_{q-1}^{q-2}.$$

Hence we obtain the recursive inequality

$$m_q^q \leq m_{q-1}^q + n(q-1)c_{2(q-1)}^{1/(q-1)} m_{q-1}^{q-2}.$$

Define the new constant $b_k = c_k^{1/k}$. It is easy to verify that b_k is increasing in k by the L^p norm inequalities. The recursive inequality thus becomes

$$m_q^q \leq m_{q-1}^q + n(q-1)b_{2(q-1)}^2 m_{q-1}^{q-2}.$$

Let's first compute a few base cases. For $q = 2$, the Efron-Stein inequality implies that $m_2^2 \leq nc_2$. Hence for $q = 3$, as Hölder implies that $m_1 \leq m_2 \leq \sqrt{nc_2} = b_2\sqrt{n}$, the inequality implies that

$$m_3^3 \leq n^{3/2}(b_2^3 + 2b_4^2 b_2) \leq 3n^{3/2}b_4^3.$$

We will show by induction that $m_q \le 2b_{2(q-1)}\sqrt{qn}$. Clearly the base cases of $q = 2, 3$ are satisfied. For the inductive step, we have

$$\begin{aligned}
m_q^q &\le m_{q-1}^q + n(q-1)b_{2(q-1)}^2 m_{q-1}^{q-2} \\
&\le n^{q/2}\left(2^q(q-1)^{q/2}b_{2(q-2)}^q + (q-1)2^{q-2}b_{2(q-1)}^2 b_{2(q-2)}^{q-2}(q-1)^{q/2-1}\right) \\
&\le n^{q/2}\left(2^q(q-1)^{q/2}b_{2(q-1)}^q + 2^{q-2}(q-1)^{q/2}b_{2(q-1)}^q\right) \\
&= n^{q/2}b_{2(q-1)}^q 2^q \cdot \frac{5}{4}(q-1)^{q/2} \\
&\le n^{q/2}b_{2(q-1)}^q 2^q q^{q/2}.
\end{aligned}$$

Hence $m_q \le 2b_{2(q-1)}\sqrt{qn}$. Using the inequality $(x^p + y^p)^{1/p} \le x + y$ yields $b_{2(q-1)} = (\gamma^{2(q-1)}(1-\delta) + M^{2(q-1)}\delta)^{1/(2(q-1))} \le \gamma + \delta^{1/(2(q-1))}M$, which gives the desired bound on the moment.

Let's next walk through what happens for weakly bounded differences. For a training set S, let $\delta(S)$ be the fraction of bad z', i.e. z' such that $\sup_i |f(S) - f(S^{i\leftarrow z'})| > \gamma$. By definition, $\mathbb{E}_S[\delta(S)] = \delta$. We also have that

$$V^+ \le n((1-\delta(S))\gamma^2 + \delta(S)M^2)$$

Hence

$$\mathbb{E}_S[(V+)^{q-1}] \le n^{q-1}\,\mathbb{E}_S[((1-\delta(S))\gamma^2 + \delta(S)M^2)^{q-1}].$$

Fix some $\delta' > \delta$. $\mathbb{P}_S[\delta(S) > \delta'] \le \delta/\delta'$, and thus we can upper bound

$$\mathbb{E}_S[(V+)^{q-1}] \le n^{q-1}\left[\left(1 - \frac{\delta}{\delta'}\right)((1-\delta')\gamma^2 + \delta' M^2)^{q-1} + \frac{\delta}{\delta'}M^{2(q-1)}\right] := n^{q-1}(b'_{2(q-1)})^{2(q-1)}.$$

The rest of the recursion is identical, except with $b'_{2(q-1)}$ instead of $b_{2(q-1)}$, and so for weak bounded diffs we get $m_q \le 2b'_{2(q-1)}\sqrt{qn}$. To upper bound the b', set $\delta' = \sqrt{\delta}$; we then obtain

$$\begin{aligned}
b'_{2(q-1)} &= \left[\left(1 - \sqrt{\delta}\right)((1-\sqrt{\delta})\gamma^2 + \sqrt{\delta}M^2)^{q-1} + \sqrt{\delta}M^{2(q-1)}\right]^{\frac{1}{2(q-1)}} \\
&\le \sqrt{(1-\sqrt{\delta})\gamma^2 + \sqrt{\delta}M^2} + \delta^{\frac{1}{4(q-1)}}M \\
&\le \gamma + \delta^{\frac{1}{4}}M + \delta^{\frac{1}{4(q-1)}}M \\
&\le \gamma + 2\delta^{\frac{1}{4(q-1)}}M.
\end{aligned}$$

This yields the stated bound on the moment. □

Remark 2.8. The tension between γ, δ, M, and q is evident from the moment bound of lemma 2.7. For comparison, note the trivial upper bound of $2M\sqrt{qn}$ by using the uniform boundedness of M (this is just the moment form of bounded differences). Also, if $\delta = 0$ then we recover the lower bound corresponding to the uniform stability moment bound of $2\gamma\sqrt{qn}$. Now, δ must be small relative to q; otherwise the moment bound will be roughly $2(\gamma + M)\sqrt{qn}$, hence rendering it ineffective. If $\delta = \exp(-\Omega(n))$ and $q = O(\log n)$, then $\gamma + \delta^{O(1/(q-1))}M = \gamma + \exp(-\Omega(n/\log n))M$, which smoothly interpolates between the lower and upper limits of the moment bounds. Also, observe that we only lose $\delta^{O(1/(q-1))}$ factors when we assume weak difference bounded f rather than strongly difference bounded f. We will return to this point after proving the main result, theorem 3.5.

In the setting of uniform bounded differences ($\delta = 0$), note that we can use the inequality $V^+ \leq \gamma$. This is precisely how the proof of Theorem 15.4 in Boucheron et al. [1] proceeds. We point out the interpretation that leveraging this a.s. bound can be seen as an application of Hölder with the pair $(\infty, 1)$. For $\delta > 0$, as we seek a recursive inequality, using the pair $(q-1, \frac{q-1}{q-2})$ is the tightest application one can hope for.

3 Improved generalization bounds

Before we dive into the details of improving the generalization bound, we first show that the expected generalization gap can be bounded under strong or weak hypothesis stability.

Proposition 3.1. *Suppose A_S has strong (or weak) (γ, δ) hypothesis stability and the loss function ℓ is uniformly bounded by L. Then*

$$\mathbb{E}_S[R(A_S) - R_{\text{emp}}(A_S)] \leq (1-\delta)\gamma + \delta L.$$

Proof. First apply symmetry and the i.i.d. assumption to rewrite

$$\mathbb{E}_S[R(A_S)] = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(A_S(x_i'), y_i')\right].$$

Similarly, we can write

$$\mathbb{E}_S[R_{\text{emp}}(A_S)] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} \ell(A_S(x_i), y_i)\right] = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(A_{S^{i\leftarrow z_i'}}(x_i'), y_i')\right]$$

Hence we have

$$\mathbb{E}_S[R(A_S) - R_{\text{emp}}(A_S)] = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n} \left(\ell(A_S(x_i'), y_i') - \ell(A_{S^{i\leftarrow z_i'}}(x_i'), y_i')\right)\right].$$

Since $z_i' \sim D$, it follows that $S^{i\leftarrow z_i'}$ follows the same distribution as S. Hence by definition of strong hypothesis stability, $S^{i\leftarrow z_i'}$ is good with probability $1-\delta$, and the term in the parentheses is upper bounded by γ. Otherwise, with probability δ, the difference is at most L, since $\ell \leq L$. By symmetry, the result follows.

For weak hypothesis stability, the same proof goes through, since we are taking expectations over both S and S', which absorbs the probability over the choice (S, z') in definition 2.3. □

Of course, we seek high probability generalization bounds, which require a bit more effort. However, the bound in expectation is suggestive of the final form of the high probability result. With this motivation in mind, we set out to prove analogues of key statements in Bousquet et al. [3]. First, define

$$g_i = g_i(S) \triangleq \mathbb{E}_{z'}[\mathbb{E}_z[\ell(A_{S^{i\leftarrow z'}}(x), y)] - \ell(A_{S^{i\leftarrow z'}}(x_i), y_i)].$$

To break this down, note that $\ell(A_{S^{i\leftarrow z'}}(x), y)$ is the change-one loss estimate when we corrupt the data at position i with $z' = (x', y')$ and test on point x. Thus, g_i is looking at the average bias over data corruptions of the change-one test error estimate at data point $z_i = (x_i, y_i)$ compared to the expectation over all test points $z = (x, y)$.

The point is that g_i measures the loss sensitivity of the learning algorithm to corruptions at training point i. Perhaps unsurprisingly, then, it suffices to study the g_i in order to bound the generalization gap. We make this intuition precise by proving a couple key properties of g_i which are required in the main theorem.

Lemma 3.2 (cf. Lemma 7 in Bousquet et al. [3]). *Suppose A_S has strong or weak (γ, δ) hypothesis stability with parameter γ and a bounded loss function $\ell \leq L$. Then the following hold:*

(1) Under strong hypothesis stability, with probability at least $1 - \delta$ over the choice of S, we have

$$\left| |R(A_S) - R_{\text{emp}}(A_S)| - \left| \sum_{i=1}^{n} g_i \right| \right| \leq 2\gamma.$$

Under weak hypothesis stability, with probability at least $1 - \sqrt{\delta}$ over the choice of S, we have

$$\left| |R(A_S) - R_{\text{emp}}(A_S)| - \left| \sum_{i=1}^{n} g_i \right| \right| \leq 2(1 - \sqrt{\delta})\gamma + 2\sqrt{\delta}L.$$

In other words, to study the generalization error, it suffices to study the g_i.

(2) For all i, we have $|g_i| \leq L$ and $\mathbb{E}[\mathbb{E}[g_i | z_{[n]\setminus i}]] = 0$.

(3) The functions g_i satisfy $(L, 2(1 - \sqrt{\delta})\gamma + 2\sqrt{\delta}L, \sqrt{\delta})$ strong or weak bounded differences.

Proof. Recalling the definitions of $R(A_S)$ and $R_{\text{emp}}(A_S)$, we have

$$|n(R(A_S) - R_{\text{emp}}(A_S))| = \left| \sum_{i=1}^{n} \mathbb{E}_z[\ell(A_S(x), y)] - \ell(A_S(x_i), y_i) \right|.$$

Now, for each i insert the terms

$$\mathbb{E}_{z'}[\ell(A_S(x_i), y_i) - \ell(A_{S^{i \leftarrow z'}}(x_i), y_i)] + \mathbb{E}_{z'}[\mathbb{E}_z[\ell(A_{S^{i \leftarrow z'}}(x), y)] - \ell(A_S(x), y)] \tag{1}$$

into the summation.

Note that these terms are precisely what is needed to transform the generalization gap on data point i to g_i. We can then apply the triangle inequality to isolate the absolute values of these terms.

$$\begin{aligned} |n(R(A_S) - R_{\text{emp}}(A_S))| \leq \left| \sum_{i=1}^{n} g_i \right| &+ \sum_{i=1}^{n} |\mathbb{E}_z[\mathbb{E}_{z'}[\ell(A_S(x), y) - \ell(A_{S^{i \leftarrow z'}}(x), y)]]| \\ &+ \sum_{i=1}^{n} |\mathbb{E}_{z'}[\ell(A_S(x_i), y_i) - \ell(A_{S^{i \leftarrow z'}}(x_i), y_i)]| \end{aligned}$$

Consider for example the terms

$$\sum_{i=1}^{n} |\mathbb{E}_{z'}[\ell(A_S(x_i), y_i) - \ell(A_{S^{i \leftarrow z'}}(x_i), y_i)]|. \tag{2}$$

Under strong hypothesis stability, with probability $1 - \delta$ over the choice of S, we have that the argument of the expectation for a particular i is uniformly bounded by γ. Hence the expectation is also bounded by γ with probability $1 - \delta$. Note also that once this holds for a particular i, it also holds for all i simultaneously by the definition of strong hypothesis stability. Hence this sum is at most γn with probability $1 - \delta$.

On the other hand, under weak hypothesis stability, we no longer have a uniform bound on the expectation over z'. We deal with this using a similar argument as in lemma 3.4. For a dataset S, define $\delta(S)$ to be the fraction of z' such that

$$\sup_{i,z} |\ell(A_S(x), y) - \ell(A_{S^{i\leftarrow z'}}(x), y)| > \gamma$$

By definition of weak hypothesis stability, we have that $\mathbb{E}_S[\delta(S)] = \delta$. For each S, we can thus bound (2) by $n((1-\delta(S))\gamma + \delta(S)L)$. With probability at least $1-\sqrt{\delta}$ over the choice of S, we therefore must have $\delta(S) \leq \sqrt{\delta}$ and hence we can upper bound (2) by $n((1-\sqrt{\delta})\gamma + \sqrt{\delta}L)$. Note that a similar argument applies to the second term of eq. (1). The expectation over z is irrelevant since in definitions 2.2 and 2.3 the bound holds uniformly over z.

Putting it all together, under strong hypothesis stability we obtain that

$$|n(R(A_S) - R_{\text{emp}}(A_S))| \leq 2\gamma n + \left|\sum_{i=1}^{n} g_i\right|,$$

with probability $1-\delta$. Under weak hypothesis stability,

$$|n(R(A_S) - R_{\text{emp}}(A_S))| \leq 2n\gamma(1-\sqrt{\delta}) + 2nL\sqrt{\delta} + \left|\sum_{i=1}^{n} g_i\right|,$$

with probability $1-\sqrt{\delta}$.

Next, since $\ell \in [0, L]$, point (2) immediately follows.

Finally, we show that g_i satisfies bounded differences wrt all variables $j \neq i$. We can write

$$\begin{aligned}
\left|g_i(S) - g_i(S^{j\leftarrow z'_j})\right| &= \Big|\mathbb{E}_{z'}[\mathbb{E}_z[\ell(A_{S^{i\leftarrow z'}}(x), y)] - \ell(A_{S^{i\leftarrow z'}}(x_i), y_i)] \\
&\quad - \mathbb{E}_{z'}[\mathbb{E}_z[\ell(A_{S^{i\leftarrow z', j\leftarrow z'_j}}(x), y)] - \ell(A_{S^{i\leftarrow z', j\leftarrow z'_j}}(x_i), y_i)]\Big| \\
&\leq \left|\mathbb{E}_z\,\mathbb{E}_{z'}[\ell(A_{S^{i\leftarrow z'}}(x), y) - \ell(A_{S^{i\leftarrow z', j\leftarrow z'_j}}(x), y)]\right| \\
&\quad + \left|\mathbb{E}_{z'}[\ell(A_{S^{i\leftarrow z'}}(x_i), y_i) - \ell(A_{S^{i\leftarrow z', j\leftarrow z'_j}}(x_i), y_i)]\right|
\end{aligned}$$

Under weak stability, we can apply a similar argument as we did earlier. For fixed S, define $\delta(S)$ to be the fraction of z' such that $\sup_{j,z,z'_j} \left|\ell(A_{S^{i\leftarrow z'}}(x), y) - \ell(A_{S^{i\leftarrow z', j\leftarrow z'_j}}(x), y)\right| > \gamma$. Since $S^{i\leftarrow z'}$ is distributed identically to S, by definition of strong hypothesis stability it follows that $\mathbb{E}_S[\delta(S)] = \delta$. So with probability at least $1-\sqrt{\delta}$ over z', we have that the difference is bounded by $\sqrt{\delta}$; otherwise, it's bounded in absolute value by L. Hence the first term can be bounded by $(1-\sqrt{\delta})\gamma + \sqrt{\delta}L$. An entirely analogous argument handles the second term of (2). In summary, the g_i have strong bounded difference $(L, 2(1-\sqrt{\delta})\gamma + 2\sqrt{\delta}L, \sqrt{\delta})$. □

3.1 Proving g_i^l has weak bounded differences

Before jumping into the moment bound for $\sum_{i=1}^n g_i$, we introduce a few new pieces of notation and address a few technical points that are needed for the proof to go through.

In the proof, we reduce to the case where $n = 2^k$ by adding identically zero functions; this allows for a cleaner analysis. The first key idea of Bousquet et al. [3] is to create a filtration of set systems

$\mathcal{B}_0, \ldots, B_k$ with each $\mathcal{B}_i$ being a partition of $[2^k]$ into consecutive sets of size 2^i. In other words, we have

$$\mathcal{B}_0 = \{\{1\}, \ldots, \{2^k\}\}; \quad B_k = \{\{1, \ldots, 2^k\}\}.$$

Hence we have $|\mathcal{B}_i| = 2^{k-i}$, and for each set $B^i \in \mathcal{B}_i$, we have $\left|B^i\right| = 2^i$.

Now, for the g_i, we will find it useful to analysis g_i conditioned on various subsets of Z; in particular define for $l = 0, \ldots, k$ the random variables

$$g_i^l = \mathbb{E}[g_i | Z_i, Z_{[n]\setminus B^l(i)}].$$

Thus, as l increases, we condition on fewer and fewer of the Z_j. Put another way, the g_i^l form a reversed martingale.

In our general analysis, we assume that the g_i have strong or weak bounded differences (M, γ, δ). Under uniform bounded differences (i.e. $\delta = 0$), it is immediate that the g_i^l also have uniform bounded differences γ. However, when $\delta > 0$, which is the case we're interested in, it is not immediately obvious that the g_i^l also have a strong or weak bounded differences property, and if so, what those parameters would be. We now establish that in fact that the g_i^l possess weak bounded differences.

Lemma 3.3. *Assume that g_i has (M, γ, δ) weak bounded differences with respect to all but the ith variable. Then $g_i^l(Z_i, Z_{[n]\setminus B^l(i)})$ has $(M, (1-\sqrt{\delta})\gamma + \sqrt{\delta}M, \sqrt{\delta})$ weak bounded differences with respect to all but the ith variable.*

Proof. For fixed $T = (Z_i, Z_{[n]\setminus B^l(i)}, Z')$, let $\delta(T)$ be the fraction of $Z_{B^l(i)\setminus\{i\}}$ such that $\sup_{j\neq i} |g_i(S) - g_i(S^{j\leftarrow Z'})| > \gamma$. Then,

$$\begin{aligned}\left|g_i^l(Z_i, Z_{[n]\setminus B^l(i)}) - g_i^l(Z_i, Z_{[n]\setminus B^l(i)} \leftarrow Z')\right| &\leq \mathbb{E}_{Z_{B^l(i)\setminus\{i\}}}\left[\left|g_i(S) - g_i(S^{j\leftarrow Z'})\right| \mid Z_i, Z_{[n]\setminus B^l(i)}, Z'\right] \\ &\leq (1-\delta(T))\gamma + \delta(T)M\end{aligned}$$

Since $\mathbb{E}_T[\delta(T)] = \delta$, it follows that $\delta(T) \leq \sqrt{\delta}$ with probability at least $1 - \sqrt{\delta}$, and thus with probability $1 - \sqrt{\delta}$, the difference is bounded by $(1-\sqrt{\delta})\gamma + \sqrt{\delta}M$.

□

Finally, combining the same strategy as above and lemma 2.7 yields a moment bound on $g_i^l - g_i^{l+1}$.

Lemma 3.4. *Suppose $g_i^l(Z_i, Z_{[n]\setminus B^l(i)})$ has (M, γ, δ) weak bounded differences. Then*

$$\left\|g_i^l - g_i^{l+1}\right\|_p(Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq 2\sqrt{p2^l}(\gamma + 2\delta(Z_i, Z_{[n]\setminus B^{l+1}(i)})^{\frac{1}{4(p-1)}}M),$$

Furthermore,

$$\left\|g_i^l - g_i^{l+1}\right\|_p \leq 2\sqrt{p2^l}\left(\gamma + 3\delta^{\frac{1}{8(p-1)}}M\right)$$

Proof. For fixed $R = (Z_i, Z_{[n]\setminus B^{\ell+1}(i)})$, let $\delta(R)$ be the fraction of $(Z_{B^{\ell+1}(i)\setminus B^\ell(i)}, Z')$ such that g_i^l has weak bounded differences with parameter γ when replacing a variable in $B^{l+1}(i) \setminus B^l(i)$ with Z'. Then, g_i^l conditioned on R has $(M, \gamma, \delta(R))$ weak bounded diffs, so we can use 2.7 to bound the conditional moment as

$$\|g_i^l - g_i^{l+1}\|_p(Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq 2\sqrt{p2^l}(\gamma + 2\delta(R)^{\frac{1}{4(p-1)}}M).$$

We next integrate this bound to get a bound on $\left\|g_i^l - g_i^{l+1}\right\|_p$, keeping in mind that $\mathbb{E}_R[\delta(R)] = \delta$. Identically to the argument we've used before, we have that $\delta(R) > \sqrt{\delta}$ with probability at most $\sqrt{\delta}$. If $\delta(R) \leq \sqrt{\delta}$, then we have that

$$\|g_i^l - g_i^{l+1}\|_p(R) \leq 2\sqrt{p2^l}(\gamma + 2\delta^{\frac{1}{8(p-1)}}M).$$

Otherwise, we can naively bound

$$\|g_i^l - g_i^{l+1}\|_p(R) \leq 2M \leq 2\sqrt{p2^l}M.$$

Therefore

$$\left\|g_i^l - g_i^{l+1}\right\|_p^p = \mathbb{E}_R\left[\left\|g_i^l - g_i^{l+1}\right\|_p^p(R)\right] \leq (2\sqrt{p2^\ell})^p\Big[(1-\sqrt{\delta})(\gamma + 2\delta^{\frac{1}{8(p-1)}}M)^p + \sqrt{\delta}M^p\Big],$$

and hence

$$\left\|g_i^l - g_i^{l+1}\right\|_p \leq 2\sqrt{p2^\ell}\Big[\gamma + 2\delta^{\frac{1}{8(p-1)}}M + \delta^{\frac{1}{2p}}M\Big] \leq 2\sqrt{p2^\ell}\Big[\gamma + 3\delta^{\frac{1}{8(p-1)}}M\Big].$$

□

Theorem 3.5 (cf. Theorem 4 in Bousquet et al. [3]). *Let $Z = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ be independent random variables and $g_i : \mathcal{Z}^n \to \mathbb{R}$ for $i \in [n]$ be measurable functions such that*

(1) $|\mathbb{E}[g_i(Z)|Z_i]| \leq M$ *a.s.*

(2) $\mathbb{E}[g_i(Z)|Z_{[n]\setminus i}] = 0$ *a.s.*

(3) g_i *has weak bounded difference* (M, γ, δ) *with respect to all but the* i*th variable.*

Then for any $p \geq 2$, we have

$$\left\|\sum_{i=1}^n g_i(Z)\right\|_p \leq 12\sqrt{2}pn(\gamma + 3\delta^{\frac{1}{8(p-1)}}M)\log n + 4\sqrt{pn}M.$$

Proof. We summarize the steps of the proof and point out the key differences when g_i do not satisfy uniform bounded difference. We first note why each condition is useful. Conditions (1) and (2) allow us to apply a moment form of the bounded differences inequality, while (3) allows us to apply lemma 2.7.

Recall the g_i^l defined earlier:

$$g_i^l = \mathbb{E}[g_i|Z_i, Z_{[n]\setminus B^l(i)}].$$

Similar to martingale inequalities, we consider consecutive differences in the g_i^l. Since $g_i^k = \mathbb{E}[g_i|Z_i]$ and $g_i^0 = g_i$ (the original random variable), we obtain

$$g_i - \mathbb{E}[g_i|Z_i] = \sum_{l=0}^{k-1} g_i^l - g_i^{l+1}.$$

The point is after moving $\mathbb{E}[g_i|Z_i]$ to the RHS and applying an L^p norm, we can leverage the triangle inequality, along with various momen inequalities to upper bound the RHS. After summing over i and interchanging sums we obtain

$$\left\|\sum_{i=1}^n g_i\right\|_p \leq \left\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\right\|_p + \sum_{l=0}^{k-1}\left\|\sum_{i=1}^n g_i^l - g_i^{l+1}\right\|_p.$$

First, we apply bounded differences to $\mathbb{E}[g_i|Z_i]$ with conditions (1) and (2) to obtain that

$$\left\| \sum_i \mathbb{E}[g_i|Z_i] \right\|_p \leq 4\sqrt{pn}M.$$

For the other terms, under the strong or weak bounded differences property, it is tempting to condition on S being good. This has the benefit of ensuring that further conditioning preserve the same (uniform) bounded differences property. However, the pitfall to this approach is that conditioning on good S ruins independence of the Z_i. Since many of our moment inequalities heavily rely on independence, this is troublesome, although plausibly there could be workarounds if one could show that the Z_i are only weakly dependent after conditioning.

Hence, we instead continue without conditioning on the good set. One outstanding concern is that while the uniform bounded differences is immediately preserved under conditioning, the situation is rather delicate for strong or weak bounded differences. This was the point of Lemmas 3.3 and 3.4; to be explicit, the lemma implies that the g_i^l are each $(M, (1-\sqrt{\delta})\gamma + \sqrt{\delta}M, \sqrt{\delta})$ weakly difference bounded.

We now proceed with the analysis; the first trick is to further decompose the sum over i of $g_i^l - g_i^{l+1}$ into summing first over $i \in B^l$ and subsequently over all B^l sets. The same argument as in Bousquet et al. [3], by invoking the Marcinkiewicz-Zygmund inequality, yields

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq 3\sqrt{2p2^l} \left(\frac{1}{2^l} \sum_{i \in B^l} \left\| g_i^l - g_i^{l+1} \right\|_p^p \right)^{1/p}.$$

Now we are in the home stretch. By invoking lemma 3.4, we have that

$$\left\| g_i^l - g_i^{l+1} \right\|_p \leq 2\sqrt{p2^l} \left(\gamma + 3\delta^{\frac{1}{8(p-1)}} M \right),$$

and hence

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq 6\sqrt{2}p2^l(\gamma + 3\delta^{\frac{1}{8(p-1)}} M).$$

The rest of the proof goes through without issue, yielding a moment bound of the form

$$\sum_{l=0}^{k-1} \left\| \sum_{i=1}^{n} g_i^l - g_i^{l+1} \right\|_p \leq 12\sqrt{2}pn(\gamma + 3\delta^{\frac{1}{8(p-1)}} M) \log n.$$

□

Remark 3.6. We note that in the proof of theorem 3.5, the bound would be tighter for g_i strongly difference bounded, and indeed they are strongly difference bounded by lemma 3.2. However, it does not make any difference for asymptotic effectiveness. It is interesting to compare to the case of uniform stability, which immediately implies change-two stability and uniform bounded differences with the same stability coefficient. Clearly, the same cannot be said for hypothesis stability, unless one is willing to make the stronger assumption that the learning algorithm is change-two hypothesis stable.

We are finally ready to prove our tail bound by combining theorem 3.5 and lemmas 2.6 and 3.2. For a desired failure probability $\delta_2 \in (0,1)$, the proof of lemma 2.6 simply requires us to convert the moment bound for $p = \log(e/\delta_2)$. Notice also that the moment bound in theorem 3.5 deteriorates in quality as p grows large. This is to be expected, as the moments should be more dominated by the bad event where stability is as worse as M. Another minor technical detail is that theorem 3.5 only furnishes us with a moment bound for $p \geq 2$, but in the proof of lemma 2.6 as long as $p = \log(e/\delta_2) \geq 2$, the equivalence still holds.

Because of the sensitivity of theorem 3.5 to the size of δ_2, in order to get effective bounds we will typically require $\delta_2 = \Omega(\frac{1}{\text{poly}(n)})$ and $\delta = \exp(-\Omega(n))$.

Theorem 3.7 (Tighter generalization for almost everywhere stable algorithms)**.** *Suppose A_S is a (γ, δ_1) strongly (resp. weakly) hypothesis stable learning algorithm, and we measure the performance on a uniformly bounded loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, L]$. For any $\delta_2 \in (0,1)$, we have with probability at least $1 - \sqrt{\delta_1} - \delta_2$ that*

$$|R(A_S) - R_{\text{emp}}(A_s)| \leq c_1(\gamma + \delta_1^{\frac{1}{16\log(1/\delta_2)}} L)\log n \log\left(\frac{1}{\delta_2}\right) + c_2 L\sqrt{\frac{\log(1/\delta_2)}{n}} + 2\gamma + 2\sqrt{\delta_1}L.$$

Proof. From lemma 3.2, we have that the g_i satisfy $(L, 2(1-\sqrt{\delta_1})\gamma + L\sqrt{\delta_1}, \sqrt{\delta_1})$ weak bounded differences. Therefore by theorem 3.5, we can bound

$$\begin{aligned}\left\|\sum_{i=1}^n g_i(Z)\right\|_p &\leq 12\sqrt{2}pn(2\gamma + 2L\sqrt{\delta_1} + 3\delta_1^{\frac{1}{16(p-1)}} L)\log n + 4\sqrt{pn}L \\ &\leq 24\sqrt{2}pn(\gamma + 3\delta_1^{\frac{1}{16(p-1)}} L)\log n + 4\sqrt{pn}L\end{aligned}$$

By lemma 2.6, we have that with probability at least $1 - \delta_2$:

$$\left|\sum_{i=1}^n g_i(Z)\right| \leq c_1(\gamma + \delta_1^{\frac{1}{16\log(1/\delta_2)}} L)\log n \log\left(\frac{1}{\delta_2}\right) + c_2 L\sqrt{\frac{\log(1/\delta_2)}{n}}.$$

Finally, applying claim (1) of lemma 3.2, we have that with probability at least $1 - \delta_1^{\frac{1}{2}} - \delta_2$ that

$$|R(A_S) - R_{\text{emp}}(A_s)| \leq c_1(\gamma + \delta_1^{\frac{1}{16\log(1/\delta_2)}} L)\log n \log\left(\frac{1}{\delta_2}\right) + c_2 L\sqrt{\frac{\log(1/\delta_2)}{n}} + 2\gamma + 2\sqrt{\delta_1}L.$$

□

4 Applications

Having improved generalization bounds for almost everywhere stability, we discuss scenarios where such improved generalization bounds are relevant. For uniform stability, the improvements of Feldman and Vondrak [4], Bousquet et al. [3] yield high probability generalization in the regime where $\gamma = O(1/\sqrt{n})$. On the other hand, in the extensive overview of almost everywhere algorithmic stability [6], most of the examples discussed are $(0, \delta)$ where δ is typically $\frac{1}{\text{poly}(n)}$. For example, the maximum margin algorithm is weakly $(0, \delta)$ hypothesis stable where $\delta = \frac{2\,\mathbb{E}[\#\text{ support points}]}{n+1}$. Since our generalization bound is effective in the regime where $\gamma = O(1/\sqrt{n})$, $\delta = \exp(-\Omega(n))$, and $\delta_2 = 1/\text{poly}(n)$, it behooves us to find applications where we indeed have (strong or weak) hypothesis stability with these settings of parameters.

4.1 1-nearest neighbor interpolation

We describe a simple scenario where we obtain $(\tilde{O}(\frac{1}{\sqrt{n}}), \frac{1}{\text{poly}(n)})$ strong hypothesis stability. Consider the uniform measure μ on the unit square $\mathcal{S} = [0,1] \times [0,1] \subset \mathbb{R}^2$. Our goal is to learn an unknown 1-Lipschitz function $f : \mathcal{S} \to \mathbb{R}$, given a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \sim \mu$ i.i.d and $y_i = f(x_i)$. The algorithm A_S we use to learn f is *1-nearest neighbor interpolation*, where $A_S(x)$ is equal to the value of its closest neighbor; concretely, we define $A_S(x) = f(x_j)$, where $j = \arg\min_{i \in [n]} \|x - x_i\|_2$. The loss is defined to be $\ell(A_S(x), y) = |A_S(x) - y|$.

It is easy to verify that the nearest neighbor interpolator is not uniformly stable. For example, if all of the training points in S are $(0, f(0))$, then the loss incurred on $z = (1, f(1))$ is $\ell(A_S(1), f(1)) = |f(1) - f(0)| = \Omega(1)$; however, if the perturbation is $z' = (1, f(1))$, then $\ell(A_{S^{i \leftarrow z'}}(1), f(1)) = 0$. Hence the uniform stability parameter is $\Omega(1)$.

This construction shows that the stability parameter can be large when our training S doesn't cover the unit square well. However, we hope that with high probability that S does cover the unit square.

Consider discretizing $\mathcal{S}$ into a grid of k^2 squares of size $\frac{1}{k} \times \frac{1}{k}$. Assume that S satisfies the property that for each grid square, there is $(x_i, y_i) \in S$ such that x_i lies in that grid square. For a fixed new data point $z = (x, f(x))$ and corruption z', let $x_j = \arg\min_{x' \in S} \|x - x'\|_2$, and $x_k = \arg\min_{x' \in S^{i \leftarrow z'}} \|x - x'\|_2$. Then, the stability at point z is

$$\begin{aligned} |\ell(A_S(x), f(x)) - \ell(A_{S^{i \leftarrow z'}}(x), f(x))| &= ||f(x) - f(x_j)| - |f(x) - f(x_k)|| \\ &\le \max\{|f(x) - f(x_j)|, |f(x) - f(x_k)|\} \\ &\le \max\{\|x - x_j\|_2, \|x - x_k\|_2\} \end{aligned}$$

Since S contains a point in each grid square, $\|x - x_j\|_2$ is at most the max distance between 2 points in the same square, so $\|x - x_j\|_2 \le \sqrt{2}/k$. Similarly, $\|x - x_k\|_2$ is at most the distance from x to the second closest point in S, which must be within 2 grid squares of x and hence $\|x - x_k\|_2 \le 2\sqrt{2}/k$. Therefore the stability parameter is $O(\frac{1}{k})$.

We calculate the probability that each grid square contains a point in S. For a fixed grid square, the probability that no point in S lies in it is $(1 - 1/k^2)^n \le \exp(-n/k^2)$. Taking the union bound over all grid squares tells us that the probability that a square is missing a point is at most $k^2 \exp(-n/k^2)$. Taking $k^2 = O(\frac{n}{\log n})$ gives us a failure probability of $\frac{1}{\text{poly}(n)}$; this yields a stability parameter of $O(\frac{\log n}{n})$. Hence the 1-nearest neighbors interpolation algorithm is $(\tilde{O}(\frac{1}{\sqrt{n}}), 1, \frac{1}{\text{poly}(n)})$ strongly hypothesis stable.

4.2 Strongly convex ERM

Following Shalev-Shwartz et al. [8], we consider the problem of solving convex ERM. We let our hypothesis space $\mathcal{H}$ be a closed, convex body in a Hilbert space and D be a distribution over points $z \in \mathcal{Z}$. Further, let $\ell : \mathcal{H} \times \mathcal{Z} \to [0, L]$ be a ρ-Lipschitz convex loss function in $\mathcal{H}$. The goal is to find the minimizer of the population risk, i.e. to find $h^* = \arg\min_h \mathbb{E}_z[\ell(h, z)] = \arg\min_h R(h)$. Now consider a learning algorithm A_S which performs ERM. In other words, A_S estimates h^* by minimizing the empirical risk $R_{\text{emp}}(A_S) = \frac{1}{n} \sum_{i=1}^n \ell(h, z_i)$ over $h \in \mathcal{H}$.

In this setting, stability is not guaranteed unless we add a strongly convex regularizer such as one of the form $\frac{\lambda}{2}\|h\|^2$; this turns the objective function into one that is λ-strongly convex. For such λ-strongly convex objective functions, Shalev-Shwartz et al. [8] prove that the learning algorithm has uniform stability $\frac{4\rho^2}{\lambda n}$. The choice of λ which minimizes $R(h^*) - R_{\text{emp}}(A_S)$ is $\lambda = \tilde{O}(1/\sqrt{n})$, hence placing us in the regime that requires stronger generalization bounds.

We note a few difficulties in turning this into a setting where we have almost everywhere stability. First of all, once we have convexity and ρ-Lipschitzness of ℓ, regularization is all that is needed to turn the problem into a strongly convex one. Since the loss function is usually predetermined, it does not make much sense for the choice of S to affect convexity.

There are a couple workarounds we considered:

- Consider the scenario where ℓ is *a priori* convex and ρ-Lipschitz. In the proof of uniform stability, λ-strong convexity is only used in the form

$$R_{\text{emp}}(A_S^{i\leftarrow z'}) \geq R_{\text{emp}}(A_S) + \frac{\lambda}{2}\left\|A_S^{i\leftarrow z'} - A_S\right\|^2.$$

  Instead of regularizing, suppose that the choice of S can also give ℓ (restricted) λ strong convexity. However, since ERM is invariant to scaling of the loss function, it seems difficult to produce a non-artificial situation where we have a strongly convex parameter that is $O(1/\sqrt{n})$. We attempt to address this concern in the following bullet points.

- Let's further investigate the notion of restricted strong convexity. In the setting of noisy matrix completion [7], we suppose that random indices (i, j) are sampled from a matrix $\boldsymbol{M}$ and we observe $y = M_{ij} + \epsilon_{ij}$, where ϵ_{ij} is a zero mean and finite variance noise term. We can rephrase the data observation model as a vector operator T_n (producing n i.i.d. noisy observations). It is known that in the setting of noisy matrix completion, that with high probability over the noise and observed matrix entries, the observation operator T_n is nice in search directions that are not too spiky or high rank, as measured by various matrix norms.

  The point is that under the quadratic loss $\|\boldsymbol{y} - T_n(\boldsymbol{M})\|_2^2$, strong convexity is governed by the smallest eigenvalue of the Hessian $T_n^\top T_n$. Hence the niceness of the observation operator implies strong convexity in the special search directions. Unfortunately, this niceness is governed by matrix norms which would unlikely lead to a $O(1/\sqrt{n})$ strong convexity parameter. One other difficulty with adapting this setting to a stability analysis is that matrix completion does not fit nicely into the supervised learning setting underlying algorithmic stability.

- Inspired by the Hessian lowest eigenvalue condition for quadratic loss functions, we recall that for a $N \times n$ random matrix with i.i.d. $\mathtt{N}(0, 1)$ entries, the lowest singular value is $\sqrt{N} - \sqrt{n}$ with high probability. With $n = N - d$, we have

$$\sqrt{N} - \sqrt{n} = \frac{d}{\sqrt{N} + \sqrt{N - d}},$$

  so taking $d = n^{1/4}$ the lowest singular value is $O(n^{-1/4})$ with high probability. If this was the observation operator, the Hessian would thus have lowest eigenvalue $O(1/\sqrt{n})$ with high probability.

- Switching gears entirely, let's take for granted that the final ERM will be $O(1)$-strongly convex. We first note that the crucial hypotheses that ℓ be Lipschitz and strongly convex in the hypothesis are somewhat conflicting. Indeed, if the range of the learning algorithm, a subset of $\mathcal{H}$, is unbounded, then it cannot be the case that the loss function is both λ strongly convex and ρ Lipschitz. Hence, another setting with almost everywhere hypothesis stability is if we know that for most (S, z') that hypotheses A_S and $A_{S^{i\leftarrow z'}}$ fall within a bounded subset of $\mathcal{H}$.

Take for example $\mathcal{Z} = \mathbb{R}$, D to be the standard normal distribution, and the loss function to be the squared loss $\frac{1}{2}\|y - A_S(x)\|_2^2$. By strong convexity we have

$$\ell(A_S(x), y) - \ell(A_{S^{i\leftarrow z'}}(x), y) \geq \|A_S - A_{S^{i\leftarrow z'}}\| \left(-\|\nabla_{A_S} \ell(A_S(x), y)\|_* + \frac{\lambda}{2}\|A_S - A_{S^{i\leftarrow z'}}\| \right).$$

Hence, to ensure this, with somewhat similar inspiration to the interpolation setting of section 4.2, suppose the learning algorithm A_S is $O(1)$-Lipschitz in the maximum distance between two points of S. Since a standard normal obeys the tail bound $\mathbb{P}[x \geq n^{1/4}] \leq \exp(-\sqrt{n}/2)$, by the union bound we obtain that the Lipschitz constant is $O(n^{1/4})$ with (very) high probability, so that overall with high probability the algorithm has stability $O(1/\sqrt{n})$.

5 Discussion

In this paper we strengthened generalization bounds for a relaxed notion of algorithmic stability that only requires stability with high probability over the training set. We primarily followed the high level strategy of Bousquet et al. [3], but the relaxation introduces several technicalities that must be addressed for the proof to go through. Our main insights are (1) a moment bound form of the bounded differences inequality when the bounded differences only hold with high probability (lemma 2.7) and (2) a unified way to pass from high probability bounded differences on a function $g(A, B)$ to high probability bounded differences (with slightly worse parameters) on $g(A, B)$ where A is fixed (i.e. we condition on A).

The former is interesting because previous tail bounds for almost-everywhere stable learning algorithms such as theorem 2.5 are not particularly user friendly as they are only effective in specific regimes of the parameters. The latter, obtained by repeatedly leveraging Markov's inequality, highlights the strength of uniform stability. For uniformly stable learning algorithms, point (2) comes for free, with the exact same parameter. While it is tempting to condition on the (high probability) good event for analysis, this introduces the thorny issue of dependence between the data points. However, taken together, (1) and (2) allow us to sidestep this thorny issue and obtain an unconditional moment bound.

Our main result, a high probability generalization bound for hypothesis stable learning algorithms, is stronger than theorem 2.5, and makes more explicit the relationship between the different parameters for high probability hypothesis stability and generalization. In particular, it applies when we want $1/\operatorname{poly}(n)$ guarantees on the generalization gap and the learning algorithm is $O(1/\sqrt{n})$ stable with failure probability $\exp(-\Omega(n))$.

Finding realistic applications for our results is no easy feat. The simple interpolation setting we described is indicative of how we searched for applications. By placing simple generative distributions on our data, we could identify a typical set of events. Within the typical set, the hope is that mild assumptions can be placed on the learning algorithm which set us up for $O(1/\sqrt{n})$ stability. To that end, we investigated Lipschitzness and strong convexity, both of which are crucial assumptions for ERM to be uniformly stable.

There are still several open directions that would be interesting to pursue. Finding a compelling and realistic application would be quite interesting; although Kutin and Niyogi [6] provide a few applications, the situation does not appear to be as universal as one would hope. Furthermore, despite uniform stability being a distribution-independent notion, distribution-dependence is clearly quite important for high probability stability. For example, one application we did not study was generalization bounds for SGD. Since SGD is a randomized algorithm (where the internal randomness is due to the choice of training points to perform gradient updates with respect to and the stability

coefficients associated with these updates), the generalization bound of Feldman and Vondrak [4] is only shown w.h.p. over the internal randomness but still holds in a distribution-free manner over the aggregate dataset S. It would be interesting to extend this to a distribution-dependent setting where we ostensibly need to involve the weakening of uniform stability studied in this paper.

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[2] Olivier Bousquet and Andre Elisseeff. Stability and generalization. *Journal of Machine Learning*, 2:499–526, 2002.

[3] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

[4] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, 2019.

[5] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

[6] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of UAI*, 2002.

[7] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

[8] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning*, 11:2635–2670, 2010.