Algorithmic Stability and Hypothesis Complexity

Tongliang Liu [1] **Gábor Lugosi** [2 3 4] **Gergely Neu** [5] **Dacheng Tao** [1]

Abstract

We introduce a notion of algorithmic stability of learning algorithms—that we term *argument stability*—that captures stability of the hypothesis output by the learning algorithm in the normed space of functions from which hypotheses are selected. The main result of the paper bounds the generalization error of any learning algorithm in terms of its argument stability. The bounds are based on martingale inequalities in the Banach space to which the hypotheses belong. We apply the general bounds to bound the performance of some learning algorithms based on empirical risk minimization and stochastic gradient descent.

1. Introduction

Many efforts have been made to analyze various notions of algorithmic stability and prove that a broad spectrum of learning algorithms are stable in some sense. Intuitively, a learning algorithm is said to be stable if slight perturbations in the training data result in small changes in the output of the algorithm, and these changes vanish as the data set grows bigger and bigger (Bonnans & Shapiro, 2013). For example, Devroye & Wagner (1979), Lugosi & Pawlak (1994), and Zhang (2003) showed that several non-parametric learning algorithms are stable; Bousquet & Elisseeff (2002) proved that ℓ_2 regularized learning algorithms are uniformly stable; Wibisono et al. (2009) generalized Bousquet and Elisseeff's results and proved that regularized learning algorithms with strongly convex penalty functions on bounded domains, e.g., ℓ_p regularized learning algorithms for $1 < p \leq 2$, are also uniformly stable; Hardt et al. (2015) showed that parametric models trained by stochastic gradient descent algorithms are uniformly stable; and Liu et al. (2017) proved that tasks in multi-task learning can act as regularizers and that multi-task learning in a very general setting will therefore be uniformly stable under mild assumptions.

The notion of algorithmic stability has been an important tool in deriving theoretical guarantees of the generalization abilities of learning algorithms. Various notions of stability have been introduced and have been exploited to derive generalization bounds. For some examples, Mukherjee et al. (2006) proved that a statistical form of leave-one-out stability is a sufficient and necessary condition for the generalization and learnability of empirical risk minimization learning algorithms; Shalev-Shwartz et al. (2010) defined a weaker notion, the so-called "on-average-replace-one-example stability", and showed that this condition is both sufficient and necessary for the generalization and learnability of a general learning setting.

In this paper we study learning algorithms that select a hypothesis (i.e., a function used for prediction) from a certain fixed class of functions belonging to a separable Banach space. We introduce a notion of *argument stability* which measures the impact of changing a single training example on the hypothesis selected by the learning algorithm. This notion of stability is stronger than uniform algorithmic stability of Bousquet & Elisseeff (2002) that is only concerned about the change in the loss but not the hypothesis itself. However, as we will show, the new notion is still quite natural and holds for a variety of learning algorithms. On the other hand, it allows one to exploit martingale inequalities (Boucheron et al., 2013) in the Banach space of the hypotheses. Indeed, the performance bounds we derive for stable algorithms depend on characteristics related to the *martingale type* of the Banach space.

Generalization bounds typically depend on the complexity of a class of hypotheses that can be chosen by the learning algorithm. Exploiting the local estimates of the complexity of the predefined hypothesis class is a promising way to obtain sharp bounds. Building on martingale inequalities in the Banach space of the hypotheses, we define a subset of the predefined hypothesis class, whose elements will (or will have a high probability to) be output by a

[1] UBTech Sydney AI Institute, School of IT, FEIT, The University of Sydney, Australia [2] Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain [3] ICREA, Pg. Llus Companys 23, 08010 Barcelona, Spain [4] Barcelona Graduate School of Economics [5] AI group, DTIC, Universitat Pompeu Fabra, Barcelona, Spain. Correspondence to: Tongliang Liu <tliang.liu@gmail.com>, Gábor Lugosi <gabor.lugosi@upf.edu>, Gergely Neu <gergely.neu@gmail.com>, Dacheng Tao <dacheng.tao@sydney.edu.au>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP. Copyright 2017 by the author(s).

learning algorithm, as the *algorithmic hypothesis class*, and study the complexity of the algorithmic hypothesis class of argument-stable learning algorithms. We show that, if the hypotheses belong to a Hilbert space, the upper bound of the Rademacher complexity of the algorithmic hypothesis class will converge at a fast rate of order $O(1/n)$, where n is the sample size.

The rest of the paper is organized as follows. Section 2 introduces the mathematical framework and the proposed notion of algorithmic stability. Section 3 presents the main results of this study, namely the generalization bounds in terms of argument stability. Section 4 specializes the results to some learning algorithms, including empirical risk minimization and stochastic gradient descent. Section 5 concludes the paper.

2. Algorithmic Stability and Hypothesis Class

We consider the classical statistical learning problem, where the value of a real random variable Y is to be predicted based on the observation of an another random variable X. Let S be a training sample of n i.i.d. pairs of random variables $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ drawn from a fixed distribution P on a set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the so-called *feature space*. A learning algorithm $\mathcal{A} : S \in \mathcal{Z}^n \mapsto h_S \in H$ is a mapping from $\mathcal{Z}^n$ to a hypothesis class H that we assume to be a subset of a separable Banach space $(\mathfrak{B}, \|\cdot\|)$. We focus on *linear* prediction problems, that is, when $h(x)$ is a linear functional of x. We write $h(x) = \langle h, x\rangle$. In other words, we assume that the feature space $\mathcal{X}$ is the algebraic dual of the Banach space $\mathfrak{B}$. We denote the norm in $\mathcal{X}$ by $\|\cdot\|_*$. The output h_S of the learning algorithm is a hypothesis used for predicting the value for Y.

An important special case is when $\mathfrak{B}$ is a Hilbert space. In that case we may assume that $\mathcal{X} = \mathfrak{B}$ and that $\langle h, x\rangle$ is the inner product in $\mathfrak{B}$.

The quality of the predictions made by any hypothesis will be measured by a loss function $\ell : \mathfrak{B} \times \mathcal{Z} \to \mathbb{R}_+$ (where $\mathbb{R}_+$ denotes the set of positive reals). Specifically, $\ell(h, Z)$ measures the loss of predicting an example Z using a hypothesis h.

The *risk* of $h \in H$ is defined by

$$R(h) = \mathbb{E}\ell(h, Z) ;$$

while the *empirical risk* is

$$R_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i) .$$

For the output h_S of a learning algorithm $\mathcal{A}$, the generalization error is defined as

$$R(h_S) - R_S(h_S) . \tag{1}$$

The notion of *algorithmic stability* was proposed to measure the changes of outputs of a learning algorithm when the input is changed. Various ways have been introduced to measure algorithmic stability. Here we recall the notion of *uniform stability* defined by Bousquet & Elisseeff (2002) for comparison purposes. This notion of stability relies on the altered sample $S^i = \{Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n\}$, the sample S with the i-th example being replaced by an independent copy of Z_i.

Definition 1 (Uniform Stability)**.** *A learning algorithm $\mathcal{A}$ is $\beta(n)$-uniformly stable with respect to the loss function ℓ if for all $i \in \{1, \ldots, n\}$,*

$$|\ell(h_S, Z) - \ell(h_{S^i}, Z)| \le \beta(n) ,$$

with probability one, where $\beta(n) \in \mathbb{R}_+$.

We propose the following, similar, notion that “acts” on the hypotheses directly, as opposed to the losses.

Definition 2 (Uniform Argument Stability)**.** *A learning algorithm $\mathcal{A}$ is $\alpha(n)$-uniformly argument stable if for all $i \in \{1, \ldots, n\}$,*

$$\|h_S - h_{S^i}\| \le \alpha(n) .$$

with probability one, where $\alpha(n) \in \mathbb{R}_+$.

The two notions of stability are closely related: Intuitively, if the loss $\ell(h, z)$ is a sufficiently smooth function of h, then uniform argument stability should imply uniform stability. To make this intuition precise, we define the notion of *Lipschitz-continuous* loss functions below.

Definition 3 (L-Lipschitz Loss Function)**.** *The loss function $\ell : \mathfrak{B} \times \mathcal{Z} \to \mathbb{R}_+$ is L-Lipschitz for an $L > 0$ if*

$$|\ell(h, z) - \ell(h', z)| \le L\,|\langle h, x\rangle - \langle h', x\rangle|$$

holds for all $z \in \mathcal{Z}$ and $h, h' \in H$.

Additionally assuming that $\|X\|_*$ is bounded by some $B > 0$ with probability one, it is easy to see that an $\alpha(n)$-uniformly argument stable learning algorithm is uniformly stable with $\beta(n) = LB\alpha(n)$, since

$$\|h_S - h_{S^i}\| = \sup_{x\in\mathcal{X}:\|x\|_*\le 1} \left(\langle h_S, x\rangle - \langle h_{S^i}, x\rangle\right) .$$

However, the reverse implication need not necessarily hold and hence uniform argument stability is a stronger notion.

In the rest of the paper, we will focus on L-Lipschitz loss functions and assume that $\|X\|_* \le B$ holds almost surely.

These assumptions are arguably stronger than those made by Bousquet & Elisseeff (2002) who only require that the loss function be bounded. In contrast, our results will require that the loss $\ell(h, z)$ be Lipschitz in the linear form $\langle h, x\rangle$, which is only slightly more general than assuming generalized linear loss functions. Nevertheless, these stronger assumptions will enable us to prove stronger generalization bounds.

The relationship between argument stability and generalization performance hinges on a property of the Banach space $\mathfrak{B}$ that is closely related to the *martingale type* of the space—see Pisier (2011) for a comprehensive account. For concreteness we assume that the Banach space $\mathfrak{B}$ is $(2, D)$-smooth (or of martingale type 2) for some $D > 0$. This means that for all $h, h' \in \mathfrak{B}$,

$$\|h + h'\|^2 + \|h - h'\|^2 \le 2\|h\|^2 + 2D^2\|h'\|^2 .$$

Note that Hilbert spaces are $(2, 1)$-smooth. The property we need is described in the following result of (Pinelis, 1994):

Proposition 1. *Let $D_1, \dots, D_n$ be a martingale difference sequence taking values in a separable $(2, D)$-smooth Banach space $\mathfrak{B}$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{n\ge 1}\left\|\sum_{t=1}^{n} D_t\right\| \ge c\epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{2D^2}\right) ,$$

where c is a constant satisfying that $\sum_{t=1}^{\infty}\|D_t\|_\infty^2 \le c^2$ (and $\|D_t\|_\infty$ is the essential supremum of the random variable $\|D_t\|$).

Our arguments extend, in a straightforward manner, to more general Banach spaces whenever exponential tail inequalities for bounded martingale sequences similar to Proposition 1 are available. We stay with the assumption of $(2, D)$-smoothness for convenience and because it applies to the perhaps most important special case when $\mathfrak{B}$ is a Hilbert space. We refer to Rakhlin & Sridharan (2015) for more information of martingale inequalities of this kind.

A key property of stable algorithms, implied by the martingale inequality, is that the hypothesis h_S output by the algorithm is concentrated—in the Banach space $\mathfrak{B}$—around its expectation $\mathbb{E}h_S$. This is established in the next simple lemma.

Lemma 1. *Let the Banach space $\mathfrak{B}$ be $(2, D)$-smooth. If a learning algorithm $\mathcal{A}$ is $\alpha(n)$-uniformly argument stable, then, for any $\delta > 0$,*

$$\mathbb{P}\left(\|h_S - \mathbb{E}h_S\| \le D\alpha(n)\sqrt{2n\log(2/\delta)}\right) \ge 1 - \delta .$$

Proof. Introduce the martingale differences

$$D_t = \mathbb{E}(h_S|Z_1, \dots, Z_t) - \mathbb{E}(h_S|Z_1, \dots, Z_{t-1})$$

so that

$$h_S - \mathbb{E}h_S = \sum_{t=1}^{n} D_t .$$

We have

$$\begin{aligned}
&\sum_{t=1}^{\infty}\|D_t\|_\infty^2 \\
&\quad= \sum_{t=1}^{n}\|\mathbb{E}(h_S|Z_1, \dots, Z_t) - \mathbb{E}(h_S|Z_1, \dots, Z_{t-1})\|_\infty^2 \\
&\quad= \sum_{t=1}^{n}\|\mathbb{E}(h_S - h_{S^t}|Z_1, \dots, Z_t)\|_\infty^2 \\
&\quad\le \sum_{t=1}^{n}(\mathbb{E}(\|(h_S - h_{S^t}\|_\infty|Z_1, \dots, Z_t))^2 \\
&\quad\le n\alpha(n)^2 .
\end{aligned}$$

Thus, by Proposition 1, we have

$$\mathbb{P}\left(\|h_S - E_S h_S\| \ge \alpha(n)D\sqrt{2n\log(2/\delta)}\right) \le \delta$$

for $\delta = 2\exp\left(-\frac{\epsilon^2}{2D^2}\right)$. □

3. Algorithmic Rademacher Complexity and Generalization Bound

The concentration result of Lemma 1 justifies the following definition of the "algorithmic hypothesis class": since with high probability h_S concentrates around its expectation $\mathbb{E}h_S$, what matters in the generalization performance of the algorithm is the complexity of the ball centered at $\mathbb{E}h_S$ and *not that of the entire hypothesis class H*. This observation may lead to significantly improved performance guarantees.

Definition 4 (Algorithmic Hypothesis Class)**.** *For a sample size n and confidence parameter $\delta > 0$, let $r = r(n, \delta) = D\alpha(n)\sqrt{2n\log(2/\delta)}$ and define the algorithmic hypothesis class of a stable learning algorithm by*

$$B_r = \{h \in H|\, \|h - \mathbb{E}h_S\| \le r(n, \delta)\} .$$

Note that, by Lemma 1, $h_S \in B_r$ with probability at least $1 - \delta$.

We bound the generalization error (1) in terms of the Rademacher complexity (Bartlett & Mendelson, 2003) of the algorithmic hypothesis class. The Rademacher complexity of a hypothesis class H on the feature space $\mathcal{X}$ is defined as

$$\mathfrak{R}(H) = \mathbb{E}\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle h, X_i\rangle ,$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher variables that are uniformly distributed in $\{-1, +1\}$.

The next theorem shows how the Rademacher complexity of the algorithmic hypothesis class can be bounded. The bound depends on the *type* of the feature space $\mathcal{X}$. Recall that the Banach space $(\mathcal{X}, \| \cdot \|_*)$ is of type $p \geq 1$ if there exists a constant C_p such that for all $x_1, \ldots, x_n \in \mathcal{X}$,

$$\mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i \right\|_* \leq C_p \left(\sum_{i=1}^{n} \|x_i\|_*^p \right)^{1/p} .$$

In the important special case when $\mathcal{X}$ is a Hilbert space, the space is of type 2 with constant $C_2 = 1$.

Theorem 1. *Assume that $\mathfrak{B}$ is a $(2, D)$-smooth Banach space and that its dual $\mathcal{X}$ is of type p. Suppose that the marginal distribution of the X_i is such that $\|X_i\|_* \leq B$ with probability one, for some $B > 0$. If a learning algorithm is $\alpha(n)$-uniformly argument stable, then the Rademacher complexity of the algorithmic hypothesis class B_r on the feature space satisfies*

$$\mathfrak{R}(B_r) \leq DC_p B \sqrt{2 \log(2/\delta)} \alpha(n) n^{-1/2+1/p} .$$

In particular, when $\mathfrak{B}$ is a Hilbert space, the bound simplifies to

$$\mathfrak{R}(B_r) \leq B \sqrt{2 \log(2/\delta)} \alpha(n) .$$

Proof. We have

$$\begin{aligned}
&\mathfrak{R}(B_r) \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} (\sigma_i \langle h, X_i \rangle \\
&\quad -\sigma_i \mathbb{E} \langle h_S, X_i \rangle + \sigma_i \mathbb{E} \langle h_S, X_i \rangle) \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \mathbb{E} \langle h_S, X_i \rangle) \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left\langle h - \mathbb{E} h_S, X_i \right\rangle \\
&\leq \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \|h - \mathbb{E} h_S\| \left\| \sum_{i=1}^{n} \sigma_i X_i \right\|_* \\
&\leq \frac{r}{n} \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i X_i \right\|_* \\
&\leq \frac{1}{n} \alpha(n) D \sqrt{2n \log(2/\delta)} C_p \left(\sum_{i=1}^{n} \|X_i\|_*^p \right)^{1/p} \\
&\leq DC_p B \sqrt{2 \log(2/\delta)} \alpha(n) n^{-1/2+1/p} ,
\end{aligned}$$

concluding the proof. □

The theorem above may be easily used to bound the performance of an $\alpha(n)$-uniformly argument stable learning algorithm. For simplicity, we state the result for Hilbert spaces only. The extension to $(2, D)$-smooth Banach spaces with a type-p dual is straightforward.

Corollary 1. *Assume that $\mathfrak{B}$ is a separable Hilbert space. Suppose that the marginal distribution of the X_i is such that $\|X_i\|_* \leq B$ with probability one, for some $B > 0$ and that the loss function is bounded and Lipschitz, that is, $\ell(h, Z) \leq M$ with probability one for some $M > 0$ and $|\ell(h, z) - \ell(h', z)| \leq L |\langle h, x \rangle - \langle h', x \rangle|$ for all $z \in \mathcal{Z}$ and $h, h' \in H$. If a learning algorithm is $\alpha(n)$-uniformly argument stable, then its generalization error is bounded as follows. With probability at least $1 - 2\delta$,*

$$\begin{aligned}
&R(h_S) - R_S(h_S) \\
&\leq 2LB \sqrt{2 \log(2/\delta)} \alpha(n) + M \sqrt{\frac{\log(1/\delta)}{2n}} .
\end{aligned}$$

Proof. Note first that, by Lemma 1, with probability at least $1 - \delta$,

$$R(h_S) - R_S(h_S) \leq \sup_{h \in B_r} (R(h) - R_S(h)) .$$

On the other hand, by the boundedness of the loss function, and the bounded differences inequality, with probability at least $1 - \delta$,

$$\begin{aligned}
&\sup_{h \in B_r} (R(h) - R_S(h)) \\
&\leq \mathbb{E} \sup_{h \in B_r} (R(h) - R_S(h)) + M \sqrt{\frac{\log(1/\delta)}{2n}} \\
&\leq 2\mathfrak{R}(\ell \circ B_r) + M \sqrt{\frac{\log(1/\delta)}{2n}} ,
\end{aligned}$$

where $\ell \circ H$ denotes the set of compositions of functions ℓ and $h \in H$. By the Lipschitz property of the loss function and a standard contraction argument, i.e., Talagrand Contraction Lemma (Ledoux & Talagrand, 2013), we have,

$$\begin{aligned}
\mathfrak{R}(\ell \circ B_r) &\leq L \cdot \mathfrak{R}(B_r) \\
&\leq LB \sqrt{2 \log(2/\delta)} \alpha(n) .
\end{aligned}$$

□

Note that the order of magnitude of $\alpha(n)$ of many stable algorithms is of order $O(1/n)$. For the notion of uniform stability, such bounds appear in Lugosi & Pawlak (1994); Bousquet & Elisseeff (2002); Wibisono et al. (2009); Hardt et al. (2015); Liu et al. (2017). As we will show in the examples below, many of these learning algorithms even have uniform argument stability of order $O(1/n)$. In such cases the bound of Corollary 1 is essentially equivalent of the earlier results cited above. The bound is dominated by the term $M\sqrt{\frac{\log(1/\delta)}{2n}}$ present by using the bounded differences inequality. Fluctuations of the order of $O(n^{-1/2})$ are often inevitable, especially when $R(h_S)$ is not typically small. When small risk is reasonable to expect, one may use more advanced concentration inequalities with second-moment information, at the price of replacing the generalization error by the so-called "deformed" generalization error $R(h_S) - \frac{a}{a-1}R_S(h_S)$ where $a > 1$. The next theorem derives such a bound, relying on techniques developed by Bartlett et al. (2005). This result improves essentially on earlier stability-based bounds.

Theorem 2. *Assume that $\mathfrak{B}$ is a separable Hilbert space. Suppose that the marginal distribution of the X_i is such that $\|X_i\|_* \le B$ with probability one, for some $B > 0$ and that the loss function is bounded and Lipschitz, that is, $\ell(h, Z) \le M$ with probability one for some $M > 0$ and $|\ell(h, z) - \ell(h', z)| \le L\,|\langle h, x\rangle - \langle h', x\rangle|$ for all $z \in \mathcal{Z}$ and $h, h' \in H$. Let $a > 1$. If a learning algorithm is $\alpha(n)$-uniformly argument stable, then, with probability at least $1 - 2\delta$,*

$$
\begin{aligned}
&R(h_S) - \frac{a}{a-1}R_S(h_S)\\
&\quad\le 8LB\sqrt{2\log(2/\delta)}\alpha(n) + \frac{(6a+8)M\log(1/\delta)}{3n}\,.
\end{aligned}
$$

The proof of Theorem 2 relies on techniques developed by Bartlett et al. (2005). In particular, we make use of the following result.

Proposition 2. *(Bartlett et al., 2005, Theorem 2.1). Let F be a class of functions that map $\mathcal{X}$ into $[0, M]$. Assume that there is some $\rho > 0$ such that for every $f \in F$, $var(f(X)) \le \rho$. Then, with probability at least $1 - \delta$, we have*

$$
\begin{aligned}
&\sup_{f\in F}\left(\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)\\
&\le \left(4\mathfrak{R}(F) + \sqrt{\frac{2\rho\log(1/\delta)}{n}} + \frac{4M}{3}\frac{\log(1/\delta)}{n}\right).
\end{aligned}
$$

To prove the theorem, we also need to introduce the following auxiliary lemma.

Define

$$
\mathcal{G}_r(Z) = \left\{\frac{r}{\max\{r, \mathbb{E}\ell(h, Z)\}}\ell(h, Z) | h \in B_r\right\}.
$$

It is evident that $\mathcal{G}_r \subseteq \{\alpha\ell \circ h | h \in B_r, \alpha \in [0, 1]\}$. The following lemma is proven in (Bartlett et al., 2005).

Lemma 2. *Define*

$$
V_r = \sup_{g\in\mathcal{G}_r}\left(\mathbb{E}g(Z) - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right).
$$

For any $r > 0$ and $a > 1$, if $V_r \le r/a$ then every $h \in B_r$ satisfies

$$
\mathbb{E}\ell(h, Z) \le \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + V_r.
$$

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. First, we introduce an inequality to build the connection between algorithmic stability and hypothesis complexity. According to Lemma 1, for any $a > 1$ and $\delta > 0$, with probability at least $1 - \delta$, we have

$$
R(h_S) - \frac{a}{a-1}R_S(h_S) \le \sup_{h\in B_r}\left(R(h) - \frac{a}{a-1}R_S(h)\right). \tag{2}
$$

Second, we are going to upper bound the term $\sup_{h\in B_r}(R(h) - \frac{a}{a-1}R_S(h))$ with high probability. It is easy to check that for any $g \in \mathcal{G}_r$, $\mathbb{E}g(Z) \le r$ and $g(Z) \in [0, M]$. Then

$$
\text{var}(g(Z)) \le \mathbb{E}(g(Z))^2 \le M\mathbb{E}g(Z) \le Mr.
$$

Applying Proposition 2,

$$
V_r \le 4\mathfrak{R}(\mathcal{G}_r) + \sqrt{\frac{2Mr\log(1/\delta)}{n}} + \frac{4M}{3}\frac{\log(1/\delta)}{n}\,.
$$

Let

$$
4\mathfrak{R}(\mathcal{G}_r) + \sqrt{\frac{2Mr\log(1/\delta)}{n}} + \frac{4M}{3}\frac{\log(1/\delta)}{n} = \frac{r}{a}.
$$

We have

$$
r \le \frac{2Ma^2\log(1/\delta)}{n} + 8a\mathfrak{R}(\mathcal{G}_r) + \frac{4}{3}\frac{2aM\log(1/\delta)}{n},
$$

which means that there exists an $r^* \le \frac{2Ma^2\log(1/\delta)}{n} + 8a\mathfrak{R}(\mathcal{G}_r) + \frac{4}{3}\frac{2aM\log(1/\delta)}{n}$ such that $V_{r^*} \le r^*/a$ holds. According to Lemma 2, for any $h \in B_r$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
\mathbb{E}\ell(h, Z) &\le \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + V_{r^*}\\
&\le \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \frac{r^*}{a}\\
&\le \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \frac{2Ma\log(1/\delta)}{n}\\
&\quad + 8\mathfrak{R}(\mathcal{G}_r) + \frac{4}{3}\frac{2M\log(1/\delta)}{n}.
\end{aligned}
$$

It is easy to verify that $\mathcal{G}_r \subseteq \{\alpha \ell \circ h | h \in B_r, \alpha \in [0,1]\} \subseteq \text{conv}B_r$.

By elementary properties of the Rademacher complexity (see, e.g., Bartlett & Mendelson (2003)), $H' \subseteq H$ implies $\mathfrak{R}(H') \leq \mathfrak{R}(H)$. Then, with probability at least $1-\delta$, we have

$$\begin{aligned}&\sup_{h \in B_r} \left(\mathbb{E}\ell(h,X) - \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h,X_i) \right) \\ &\leq \frac{2Ma\log(1/\delta)}{n} + 8\mathfrak{R}(\ell \circ B_r) + \frac{4}{3}\frac{2M\log(1/\delta)}{n}.\end{aligned}$$

The proof of Theorem 2 is complete by combining the above inequality with inequality (2), the Talagrand Contraction Lemma, and Theorem 1. □

In the next section, we specialize the above results to some learning algorithms by proving their uniform argument stability.

4. Applications

Various learning algorithms have been proved to possess some kind of stability. We refer the reader to (Devroye & Wagner, 1979; Lugosi & Pawlak, 1994; Bousquet & Elisseeff, 2002; Zhang, 2003; Wibisono et al., 2009; Hardt et al., 2015; Liu et al., 2017) for such examples, including stochastic gradient descent methods, empirical risk minimization, and non-parametric learning algorithms such as k-nearest neighbor rules and kernel regression.

4.1. Empirical Risk Minimization

Regularized empirical risk minimization has been known to be uniformly stable (Bousquet & Elisseeff, 2002). Here we consider regularized empirical risk minimization (RERM) algorithms of the following form. The empirical risk (or the objective function) of RERM is formulated as

$$R_{S,\lambda}(h) = \frac{1}{n}\sum_{i=1}^{n}\ell(h,X_i) + \lambda N(h),$$

where $N : h \in H \mapsto N(h) \in \mathbb{R}^+$ is a convex function. Its corresponding expected counterpart is defined as

$$R_\lambda(h) = \mathbb{E}\ell(h,X) + \lambda N(h).$$

Bousquet & Elisseeff (2002) proved that ℓ_2-regularized learning algorithms are $\beta(n)$-uniformly stable. Wibisono et al. (2009) extended the result and studied a sufficient condition of the penalty term $N(h)$ to ensure uniform $\beta(n)$-stability. As we now show, both of their proof methods are applicable to the analysis of uniform argument stability. By exploiting their results, we show that stable RERM algorithms have strong generalization properties.

Theorem 3. *Assume that $\mathfrak{B}$ is a separable Hilbert space. Suppose that the marginal distribution of the X_i is such that $\|X_i\|_* \leq B$ with probability one, for some $B > 0$ and that the loss function is convex in h, bounded by M and L-Lipschitz. Suppose that for some constants C and $\xi > 1$, the penalty function $N(h)$ satisfies*

$$\begin{aligned}&N(h_S) + N(h_{S^i}) - 2N\left(\frac{h_S + h_{S^i}}{2}\right) \\ &\geq C\|h_S - h_{S^i}\|^\xi. \qquad (3)\end{aligned}$$

Then, for any $\delta > 0$, and $a > 1$, if h_S is the output of RERM, with probability at least $1 - 2\delta$, we have

$$\begin{aligned}&R(h_S) - \frac{a}{a-1}R_S(h_S) \\ &\leq \quad 8LB\left(\frac{LB}{C\lambda n}\right)^{\frac{1}{\xi-1}}\sqrt{2\log(2/\delta)} \\ &+ \quad \frac{(6a+8)M\log(1/\delta)}{3n}.\end{aligned}$$

Specifically, when $N(h) = \|h\|^2$, (3) holds with $\xi = 2$ and $C = \frac{1}{2}\left(\frac{M}{\lambda}\right)^{\frac{1}{2}}$.

Proof. The proof of Theorem 3 relies on the following result implied by Wibisono et al. (2009).

Proposition 3. *Assume the conditions of Theorem 3. Then the RERM learning algorithm is $\beta(n)$-uniformly stable with*

$$\beta(n) = \left(\frac{k^\xi L^\xi}{C\lambda n}\right)^{\frac{1}{\xi-1}},$$

and is $\alpha(n)$-uniformly argument stable with

$$\alpha(n) = \left(\frac{kL}{C\lambda n}\right)^{\frac{1}{\xi-1}}.$$

Specifically, when $N(h) = \|h\|_p^p$ and $1 < p \leq 2$, the condition 3 on the penalty function holds with $\xi = 2$ and $C = \frac{1}{4}p(p-1)\left(\frac{M}{\lambda}\right)^{\frac{p-1}{p}}$, where $\|h\|_p^p = \sum_r |h_r|^p$ and r is the index for the dimensionality.

Theorem 3 follows by combining Theorem 2 and Proposition 3. □

4.2. Stochastic Gradient Descent

Stochastic gradient descent (SGD) is one of the most widely used optimization methods in machine learning. Hardt et al. (2015) showed that parametric models trained by SGD methods are uniformly stable. Their results apply to both convex and non-convex learning problems and provide insights for why SGD performs well in practice, in particular, for deep learning algorithms.

Their results are based on the assumptions that the loss function employed is both Lipschitz and smooth. In order to avoid technicalities of defining derivatives in general Hilbert spaces, in this section we assume that $\mathfrak{B} = \mathcal{X} = \mathbb{R}^d$, the d-dimensional Euclidean space.

Definition 5 (Smooth)**.** *A differentiable loss function $\ell(h, \cdot)$ is s-smooth if for all $h, h' \in H$, we have*

$$\|\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot)\| \leq s\|h - h'\|,$$

where $\nabla_x f(x)$ denotes the derivative of $f(x)$ with respect to x and $s > 0$.

Definition 6 (Strongly Convex)**.** *A differentiable loss function $\ell(h, \cdot)$ is γ-strongly convex with respect to $\| \cdot \|$ if for all $h, h' \in H$, we have*

$$(\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot))^T (h - h') \geq \gamma\|h - h'\|^2,$$

where $\gamma > 0$.

Theorem 2 is applicable to the results of SGD when the general loss function $\ell(h, x)$ is L-Lipschitz, s-smooth, and h is linear with respect to x. Note that our definition of L-Lipschitzness requires the loss function to be Lipschitz in the linear form $\langle h, x \rangle$.

Theorem 4. *Let the stochastic gradient update rule be given by $h_{t+1} = h_t - \alpha_t \nabla_h \ell(h_t, X_{i_t})$, where $\alpha_t > 0$ is the learning rate and i_t is the index for choosing one example for the t-th update. Let h_T and h_T^i denote the outputs of SGD run on sample S and S^i, respectively. Assume that $\|X\|_* \leq B$ with probability one. Suppose that the loss function is L-Lipschitz, s-smooth, and upper bounded by M. Let SGD is run with a monotonically non-increasing step size $\alpha_t \leq c/t$, where c is a universal constant, for T steps. Then, for any $\delta > 0$ and $a > 1$, with probability at least $1 - 2\delta$, we have*

$$\begin{aligned} & R(h_T) - \frac{a}{a-1} R_S(h_T) \\ \leq \; & 8BL \frac{1 + 1/sc}{n - 1} (2cBL)^{\frac{1}{sc+1}} T^{\frac{sc}{sc+1}} \sqrt{2\log(2/\delta)} \\ & + \frac{(6a + 8)M \log(1/\delta)}{3n}. \end{aligned}$$

When the loss function ℓ is convex, L-admissible, s-smooth, and upper bounded by M, suppose that SGD is run with step sizes $\alpha_t \leq 2/s$ for T steps. Then, for any $\delta > 0$ and $a > 1$, with probability at least $1 - 2\delta$,

$$\begin{aligned} & R(h_T) - \frac{a}{a-1} R_S(h_T) \\ \leq \; & \frac{16B^2L^2}{n} \sum_{t=1}^{T} \alpha_t \sqrt{2\log(2/\delta)} \\ & + \frac{(6a + 8)M \log(1/\delta)}{3n}. \end{aligned}$$

Moreover, when the loss function ℓ is γ-strongly convex, s-smooth, and upper bounded by M, let the stochastic gradient update be given by $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, X_{i_t}))$, where Ω is a compact, convex set over which we wish to optimize and $\Pi_\Omega(\cdot)$ is a projection such that $\Pi_\Omega(f) = \arg\min_{h \in H} \|h - f\|$. If the loss function is further L-Lipschitz over the set Ω and the projected SGD is run with a constant step size $\alpha \leq 1/s$ for T steps. Then, for any $\delta > 0$ and $a > 1$, with probability at least $1 - 2\delta$, the projected SGD satisfies that

$$\begin{aligned} & R(h_T) - \frac{a}{a-1} R_S(h_T) \\ \leq \; & \frac{16DB^2L^2}{\gamma n} \sqrt{2\log(2/\delta)} + \frac{(6a + 8)M \log(1/\delta)}{3n}. \end{aligned}$$

Note that any ℓ_2 regularized convex loss function is strongly convex. Bousquet & Elisseeff (2002) studied the stability of batch methods. When the loss function is strongly convex, the stability of SGD is consistent with the result in (Bousquet & Elisseeff, 2002).

While the above result only applies to L-Lipschitz loss functions as defined in Definition 3, it does explain some generalization properties of *layer-wise* training of neural networks by stochastic gradient descent. In this once-common training scheme (see, e.g., Bengio et al., 2007), one freezes the parameters of the network before/after a certain layer and performs SGD for this single layer. It is easy to see that, as long as the activation function and the loss function (connected with the network) are Lipschitz-continuous in their inputs, the overall loss can easily satisfy the continuous conditions of Theorem 4. This implies that the parameters in each layer may generalize well in a certain sense if SGD is employed with an early stop.

The proof of Theorem 4 follows immediately from Theorem 2, combined with the following result implied by Hardt et al. (2015) (which is a collection of the results of Theorems 3.8, 3.9, and 3.12 therein).

Proposition 4. *Let the stochastic gradient update be given by $h_{t+1} = h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t})$, where $\alpha_t > 0$ is the learning rate and i_t is the index for choosing one example for the t-th update. Let h_T and h_T^i denote the outputs of SGD running on sample S and S^i respectively. When the loss function is L-Lipschitz and s-smooth, suppose that SGD is run with monotonically non-increasing step size $\alpha_t \leq c/t$, where c is a universal constant, for T steps. Then,*

$$\|h_T - h_T^i\| \leq \frac{1 + 1/sc}{n - 1} (2cBL)^{\frac{1}{sc+1}} T^{\frac{sc}{sc+1}}.$$

When the loss function ℓ is convex, L-Lipschitz, and s-smooth, suppose that SGD is run with step sizes $\alpha_t \leq 2/s$

for T steps. Then,

$$\|h_T - h_T^i\| \leq \frac{2BL}{n} \sum_{t=1}^{T} \alpha_t.$$

Moreover, when the loss function ℓ is γ-strongly convex and s-smooth, let the stochastic gradient update be given by $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$, where Ω is a compact, convex set over which we wish to optimize and $\Pi_\Omega(\cdot)$ is a projection such that $\Pi_\Omega(f) = \arg\min_{h \in H} \|h - f\|$. If the loss function is L-Lipschitz over the set Ω and the projected SGD is run with constant step size $\alpha \leq 1/s$ for T steps. Then, the projected SGD satisfies algorithmic argument stability with

$$\|h_T - h_T^i\| \leq \frac{2BL}{\gamma n}.$$

5. Conclusion

We introduced the concepts of uniform argument stability and algorithmic hypothesis class, defined as the class of hypotheses that are likely to be output by the learning algorithm. We proposed a general probabilistic framework to exploit local estimates for the complexity of hypothesis class to obtain fast convergence rates for stable learning algorithms. Specifically, we defined the algorithmic hypothesis class by observing that the output of stable learning algorithms concentrates around $\mathbb{E}h_S$. The Rademacher complexity defined on the algorithmic hypothesis class then converges at the same rate as that of the uniform argument stability in Hilbert space, which are of order $O(1/n)$ for various learning algorithms, such as empirical risk minimization and stochastic gradient descent. We derived fast convergence rates of order $O(1/n)$ for their deformed generalization errors. Unlike previously published guarantees of similar flavor, our bounds hold with high probability, rather than only in expectation.

Our study leaves some open problems and allows several possible extensions. First, the algorithmic hypothesis class defined in this study depends mainly on the property of learning algorithms but little on the data distribution. It would be interesting to investigate a way to define an algorithmic hypothesis class by considering both the algorithmic property and the data distribution. Second, it would be interesting to explore if there are some algorithmic properties other than stability that could result in a small algorithmic hypothesis class.

Acknowledgments

Liu and Tao were partially supported by Australian Research Council Projects FT-130101457, DP-140102164, LP-150100671. Lugosi was partially supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P, and FEDER, EU. Neu was partially supported by the UPFellows Fellowship (Marie Curie COFUND program 600387).

References

Bartlett, Peter L and Mendelson, Shahar. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

Bartlett, Peter L, Bousquet, Olivier, and Mendelson, Shahar. Local Rademacher complexities. *Annals of Statistics*, pp. 1497–1537, 2005.

Bengio, Yoshua, Lamblin, Pascal, Popovici, Dan, and Larochelle, Hugo. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

Bonnans, J Frédéric and Shapiro, Alexander. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

Bousquet, Olivier and Elisseeff, André. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526, 2002.

Devroye, Luc and Wagner, Terry J. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

Hardt, Moritz, Recht, Benjamin, and Singer, Yoram. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Ledoux, Michel and Talagrand, Michel. *Probability in Banach spaces: Isoperimetry and processes*. Springer Science & Business Media, 2013.

Liu, Tongliang, Tao, Dacheng, Song, Mingli, and Maybank, Stephen J. Algorithm-dependent generalization bounds for multi-task learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):227–241, February 2017.

Lugosi, Gábor and Pawlak, Miroslaw. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2):475–481, 1994.

Mukherjee, Sayan, Niyogi, Partha, Poggio, Tomaso, and Rifkin, Ryan. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

Pinelis, Iosif. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Pisier, Gilles. *Martingales in Banach spaces (in connection with type and cotype)*. IHP course notes, 2011.

Rakhlin, Alexander and Sridharan, Karthik. On equivalence of martingale tail bounds and deterministic regret inequalities. *arXiv preprint arXiv:1510.03925*, 2015.

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

Wibisono, Andre, Rosasco, Lorenzo, and Poggio, Tomaso. Sufficient conditions for uniform stability of regularization algorithms. *Techincal Report MIT-CSAIL-TR-2009-060*, 2009.

Zhang, Tong. Leave-one-out bounds for kernel methods. *Neural Computation*, 15(6):1397–1437, 2003.