Hypothesis Set Stability and Generalization

Dylan J. Foster
Massachusetts Institute of Technology
dylanf@mit.edu

Spencer Greenberg
Spark Wave
admin@sparkwave.tech

Satyen Kale
Google Research
satyen@satyenkale.com

Haipeng Luo
University of Southern California
haipengl@usc.edu

Mehryar Mohri
Google Research and Courant Institute
mohri@google.com

Karthik Sridharan
Cornell University
sridharan@cs.cornell.edu

Abstract

We present a study of generalization for data-dependent hypothesis sets. We give a general learning guarantee for data-dependent hypothesis sets based on a notion of transductive Rademacher complexity. Our main result is a generalization bound for data-dependent hypothesis sets expressed in terms of a notion of *hypothesis set stability* and a notion of Rademacher complexity for data-dependent hypothesis sets that we introduce. This bound admits as special cases both standard Rademacher complexity bounds and algorithm-dependent uniform stability bounds. We also illustrate the use of these learning bounds in the analysis of several scenarios.

1 Introduction

Most generalization bounds in learning theory hold for a fixed hypothesis set, selected before receiving a sample. This includes learning bounds based on covering numbers, VC-dimension, pseudo-dimension, Rademacher complexity, local Rademacher complexity, and other complexity measures [Pollard, 1984, Zhang, 2002, Vapnik, 1998, Koltchinskii and Panchenko, 2002, Bartlett et al., 2002]. Some alternative guarantees have also been derived for specific algorithms. Among them, the most general family is that of uniform stability bounds given by Bousquet and Elisseeff [2002]. These bounds were recently significantly improved by Feldman and Vondrak [2019], who proved guarantees that are informative, even when the stability parameter β is only in $o(1)$, as opposed to $o(1/\sqrt{m})$. The $\log^2 m$ factor in these bounds was later reduced to $\log m$ by Bousquet et al. [2019]. Bounds for a restricted class of algorithms were also recently presented by Maurer [2017], under a number of assumptions on the smoothness of the loss function. Appendix A gives more background on stability.

In practice, machine learning engineers commonly resort to hypothesis sets depending on the *same sample* as the one used for training. This includes instances where a regularization, a feature transformation, or a data normalization is selected using the training sample, or other instances where the family of predictors is restricted to a smaller class based on the sample received. In other instances, as is common in deep learning, the data representation and the predictor are learned using the same sample. In ensemble learning, the sample used to train models sometimes coincides with the one used to determine their aggregation weights. However, standard generalization bounds are not directly applicable for these scenarios since they assume a fixed hypothesis set.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

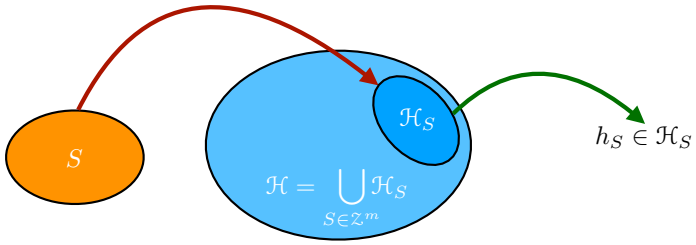

Figure 1: Decomposition of the learning algorithm's hypothesis selection into two stages. In the first stage, the algorithm determines a hypothesis $\mathcal{H}_S$ associated to the training sample S which may be a small subset of the set of all hypotheses that could be considered, say $\mathcal{H} = \bigcup_{S \in \mathcal{Z}^m} \mathcal{H}_S$. The second stage then consists of selecting a hypothesis h_S out of $\mathcal{H}_S$.

1.1 Contributions of this paper.

1. Foundational definitions of data-dependent hypothesis sets. We present foundational definitions of learning algorithms that rely on *data-dependent* hypothesis sets. Here, the algorithm decomposes into two stages: a first stage where the algorithm, on receiving the sample S, chooses a hypothesis set $\mathcal{H}_S$ dependent on S, and a second stage where a hypothesis h_S is selected from $\mathcal{H}_S$. Standard generalization bounds correspond to the case where $\mathcal{H}_S$ is equal to some fixed $\mathcal{H}$ independent of S. Algorithm-dependent analyses, such as uniform stability bounds, coincide with the case where $\mathcal{H}_S$ is chosen to be a singleton $\mathcal{H}_S = \{h_S\}$. Thus, the scenario we study covers both existing settings and other intermediate scenarios. Figure 1 illustrates our general scenario.

2. Learning bounds via transductive Rademacher complexity. We present general learning bounds for data-dependent hypothesis sets using a notion of transductive Rademacher complexity (Section 3). These bounds hold for arbitrary bounded losses and improve upon previous guarantees given by Gat [2001] and Cannon et al. [2002] for the binary loss, which were expressed in terms of a notion of shattering coefficient adapted to the data-dependent case, and are more explicit than the guarantees presented by Philips [2005][corollary 4.6 or theorem 4.7]. Nevertheless, such bounds may often not be sufficiently informative, since they ignore the relationship between hypothesis sets based on similar samples.

3. Learning bounds via hypothesis set stability. We provide finer generalization bounds based on the key notion of *hypothesis set stability* that we introduce in this paper. This notion admits algorithmic stability as a special case, when the hypotheses sets are reduced to singletons. We also introduce a new notion of *Rademacher complexity for data-dependent hypothesis sets*. Our main results are generalization bounds (Section 4) for stable data-dependent hypothesis sets expressed in terms of the hypothesis set stability parameter, our notion of Rademacher complexity, and a notion of *cross-validation stability* that, in turn, can be upper-bounded by the diameter of the family of hypothesis sets. Our learning bounds admit as special cases both standard Rademacher complexity bounds and algorithm-dependent uniform stability bounds.

4. New generalization bounds for specific learning applications. In section 5 (see also Appendix G), we illustrate the generality and the benefits of our hypothesis set stability learning bounds by using them to derive new generalization bounds for several learning applications. To our knowledge, there is no straightforward analysis based on previously existing tools that yield these generalization bounds. These applications include: (a) *bagging* algorithms that may employ non-uniform, data-dependent, averaging of the base predictors, (b) *stochastic strongly-convex optimization* algorithms based on an average of other stochastic optimization algorithms, (c) *stable representation* learning algorithms, which first learn a data representation using the sample and then learn a predictor on top of the learned representation, and (d) *distillation* algorithms, which first compute a complex predictor using the sample and then use it to learn a simpler predictor that is close to it.

1.2 Related work on data-dependent hypothesis sets.

Shawe-Taylor et al. [1998] presented an analysis of structural risk minimization over data-dependent hierarchies based on a concept of *luckiness*, which generalizes the notion of margin of linear classifiers. Their analysis can be viewed as an alternative and general study of data-dependent hypothesis sets,

using luckiness functions and ω-smallness (or ω-smoothness) conditions. A luckiness function helps decompose a hypothesis set into *lucky sets*, that is sets of functions *luckier* than a given function. The luckiness framework is attractive and the notion of luckiness, for example margin, can in fact be combined with our results. However, finding pairs of truly data-dependent luckiness and ω-smallness functions, other than those based on the margin and the empirical VC-dimension, is quite difficult, in particular because of the very technical ω-smallness condition [see Philips, 2005, p. 70]. In contrast, hypothesis set stability is simpler and often easier to bound. The notions of luckiness and ω-smallness have also been used by Herbrich and Williamson [2002] to derive algorithm-specific guarantees. In fact, the authors show a connection with algorithmic stability (not hypothesis set stability), at the price of a guarantee requiring the strong condition that the stability parameter be in $o(1/m)$, where m is the sample size [see Herbrich and Williamson, 2002, pp. 189-190].

Data-dependent hypothesis classes are conceptually related to the notion of data-dependent priors in PAC-Bayesian generalization bounds. Catoni [2007] developed localized PAC-Bayes analysis by using a prior defined in terms of the data generating distribution. This work was extended by Lever et al. [2013] who proved sharp risk bounds for stochastic exponential weights algorithms. Parrado-Hernández et al. [2012] investigated the possibility of learning the prior from a separate data set, as well as priors obtained via computing a data-dependent bound on the KL term. More closely related to this paper is the work of Dziugaite and Roy [2018a,b], who develop PAC-Bayes bounds by choosing the prior via a data-dependent differentially private mechanism, and also showed that weaker notions than differential privacy also suffice to yield valid bounds. In Appendix H, we give a more detailed discussion of PAC-Bayes bounds, in particular to show how finer PAC-Bayes bounds than standard ones can be derived from Rademacher complexity bounds, here with an alternative analysis and constants than [Kakade et al., 2008] and how data-dependent PAC-Bayes bounds can be derived from our data-dependent Rademacher complexity bounds. More discussion on data-dependent priors can be found in Appendix F.3.

2 Definitions and Properties

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space, and define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ We denote by $\mathcal{D}$ the unknown distribution over $\mathcal{X} \times \mathcal{Y}$ according to which samples are drawn.

The hypotheses h we consider map $\mathcal{X}$ to a set $\mathcal{Y}'$ sometimes different from $\mathcal{Y}$. For example, in binary classification, we may have $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{Y}' = \mathbb{R}$. Thus, we denote by $\ell \colon \mathcal{Y}' \times \mathcal{Y} \to [0, 1]$ a loss function defined on $\mathcal{Y}' \times \mathcal{Y}$ and taking non-negative real values bounded by one. We denote the loss of a hypothesis $h \colon \mathcal{X} \to \mathcal{Y}'$ at point $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ by $L(h, z) = \ell(h(x), y)$. We denote by $R(h)$ the generalization error or expected loss of a hypothesis $h \in \mathcal{H}$ and by $\widehat{R}_S(h)$ its empirical loss over a sample $S = (z_1, \ldots, z_m)$:

$$R(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h, z)] \qquad \widehat{R}_S(h) = \mathop{\mathbb{E}}_{z \sim S}[L(h, z)] = \frac{1}{m} \sum_{i=1}^{m} L(h, z_i).$$

In the general framework we consider, a hypothesis set depends on the sample received. We will denote by $\mathcal{H}_S$ the hypothesis set depending on the labeled sample $S \in \mathcal{Z}^m$ of size $m \geq 1$. We assume that $\mathcal{H}_S$ is invariant to the ordering of the points in S.

Definition 1 (Hypothesis set uniform stability)**.** *Fix $m \geq 1$. We will say that a family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ is* β-uniformly stable *(or simply β-stable) for some $\beta \geq 0$, if for any two samples S and S' of size m differing only by one point, the following holds:*

$$\forall h \in \mathcal{H}_S, \exists h' \in \mathcal{H}_{S'} \colon \forall z \in \mathcal{Z}, |L(h, z) - L(h', z)| \leq \beta. \tag{1}$$

Thus, two hypothesis sets derived from samples differing by one element are close in the sense that any hypothesis in one admits a counterpart in the other set with β-similar losses. A closely related notion is the *sensitivity* of a function $f \colon \mathcal{Z}^m \to \mathbb{R}$. Such a function f is called β-sensitive if for any two samples S and S' of size m differing only by one point, we have $|f(S) - f(S')| \leq \beta$.

We also introduce a new notion of Rademacher complexity for data-dependent hypothesis sets. To introduce its definition, for any two samples $S, T \in \mathcal{Z}^m$ and a vector of Rademacher variables $\boldsymbol{\sigma}$, denote by $S_{T,\boldsymbol{\sigma}}$ the sample derived from S by replacing its ith element with the ith element of T, for all $i \in [m] = \{1, 2, \ldots, m\}$ with $\sigma_i = -1$. We will use $\mathcal{H}^{\sigma}_{S,T}$ to denote the hypothesis set $\mathcal{H}_{S_{T,\boldsymbol{\sigma}}}$.

Definition 2 (Rademacher complexity of data-dependent hypothesis sets)**.** *Fix $m \geq 1$. The* empirical Rademacher complexity $\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H})$ and the Rademacher complexity $\mathfrak{R}^{\diamond}_{m}(\mathcal{H})$ of a family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ *for two samples* $S = (z_1^S, \ldots, z_m^S)$ *and* $T = (z_1^T, \ldots, z_m^T)$ *in* $\mathcal{Z}^m$ *are defined by*

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right] \qquad \mathfrak{R}^{\diamond}_{m}(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{S,T \sim \mathcal{D}^m \atop \boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right]. \tag{2}$$

When the family of data-dependent hypothesis sets $\mathcal{H}$ is β-stable with $\beta = O(1/m)$, the empirical Rademacher complexity $\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{G})$ is sharply concentrated around its expectation $\mathfrak{R}^{\diamond}_{m}(\mathcal{G})$, as with the standard empirical Rademacher complexity (see Lemma 4).

Let $\mathcal{H}_{S,T}$ denote the union of all hypothesis sets based on $\mathcal{U} = \{U : U \subseteq (S \cup T), U \in \mathcal{Z}^m\}$, the subsamples of $S \cup T$ of size m: $\mathcal{H}_{S,T} = \bigcup_{U \in \mathcal{U}} \mathcal{H}_U$. Since for any $\boldsymbol{\sigma}$, we have $\mathcal{H}^{\boldsymbol{\sigma}}_{S,T} \subseteq \mathcal{H}_{S,T}$, the following simpler upper bound in terms of the standard empirical Rademacher complexity of $\mathcal{H}_{S,T}$ can be used for our notion of empirical Rademacher complexity:

$$\mathfrak{R}^{\diamond}_{m}(\mathcal{H}) \leq \frac{1}{m} \mathop{\mathbb{E}}_{S,T \sim \mathcal{D}^m \atop \boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}_{S,T}} \sum_{i=1}^{m} \sigma_i h(z_i^T) \right] = \mathop{\mathbb{E}}_{S,T \sim \mathcal{D}^m} \left[\widehat{\mathfrak{R}}_T(\mathcal{H}_{S,T}) \right],$$

where $\widehat{\mathfrak{R}}_T(\mathcal{H}_{S,T})$ is the standard empirical Rademacher[1] complexity of $\mathcal{H}_{S,T}$ for the sample T. Some properties of our notion of Rademacher complexity are given in Appendix B.

Let $\mathcal{G}_S$ denote the family of loss functions associated to $\mathcal{H}_S$:

$$\mathcal{G}_S = \{z \mapsto L(h, z) : h \in \mathcal{H}_S\}, \tag{3}$$

and let $\mathcal{G} = (\mathcal{G}_S)_{S \in \mathcal{Z}^m}$ denote the family of hypothesis sets $\mathcal{G}_S$. Our main results will be expressed in terms of $\mathfrak{R}^{\diamond}_{m}(\mathcal{G})$. When the loss function ℓ is μ-Lipschitz, by Talagrand's contraction lemma [Ledoux and Talagrand, 1991], in all our results, $\mathfrak{R}^{\diamond}_{m}(\mathcal{G})$ can be replaced by $\mu \mathbb{E}_{S,T \sim \mathcal{D}^m}[\widehat{\mathfrak{R}}_T(\mathcal{H}_{S,T})]$.

Rademacher complexity is one way to measure the capacity of the family of data-dependent hypothesis sets. We also derive learning bounds in situations where a notion of *diameter* of the hypothesis sets is small. We now define a notion of *cross-validation stability* and diameter for data-dependent hypothesis sets. In the following, for a sample S, $S^{z \leftrightarrow z'}$ denotes the sample obtained from S by replacing $z \in S$ by $z' \in \mathcal{Z}$.

Definition 3 (Hypothesis set Cross-Validation (CV) stability, diameter)**.** *Fix $m \geq 1$. For some $\bar{\chi}, \chi, \bar{\Delta}, \Delta, \Delta_{\max} \geq 0$, we say that a family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ has* average CV-stability $\bar{\chi}$, CV-stability χ, average diameter $\bar{\Delta}$, diameter Δ *and* max-diameter $\Delta_{\max}$ *if the following hold:*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{z' \sim \mathcal{D}, z \sim S} \left[\sup_{h \in \mathcal{H}_S, h' \in \mathcal{H}_{S^{z \leftrightarrow z'}}} L(h', z) - L(h, z) \right] \leq \bar{\chi} \tag{4}$$

$$\sup_{S \in \mathcal{Z}^m} \mathop{\mathbb{E}}_{z' \sim \mathcal{D}, z \sim S} \left[\sup_{h \in \mathcal{H}_S, h' \in \mathcal{H}_{S^{z \leftrightarrow z'}}} L(h', z) - L(h, z) \right] \leq \chi \tag{5}$$

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{z \sim S} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \bar{\Delta} \tag{6}$$

$$\sup_{S \in \mathcal{Z}^m} \mathop{\mathbb{E}}_{z \sim S} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \Delta \tag{7}$$

$$\sup_{S \in \mathcal{Z}^m} \max_{z \in S} \left[\sup_{h, h' \in \mathcal{H}_S} L(h', z) - L(h, z) \right] \leq \Delta_{\max}. \tag{8}$$

CV-stability of hypothesis sets can be bounded in terms of their stability and diameter (see straightforward proof in Appendix C).

[1] Note that the standard definition of Rademacher complexity assumes that hypothesis sets are not data-dependent, however the definition remains valid for data-dependent hypothesis sets.

Lemma 1. *Suppose a family of data-dependent hypothesis sets $\mathcal{H}$ is β-uniformly stable. Then if it has diameter Δ, then it is $(\Delta+\beta)$-CV-stable, and if it has average diameter $\bar{\Delta}$ then it is $(\bar{\Delta}+\beta)$-average CV-stable.*

3 General learning bound for data-dependent hypothesis sets

In this section, we present general learning bounds for data-dependent hypothesis sets that do not make use of the notion of hypothesis set stability. One straightforward idea to derive such guarantees for data-dependent hypothesis sets is to replace the hypothesis set $\mathcal{H}_S$ depending on the observed sample S by the union of all such hypothesis sets over all samples of size m, $\overline{\mathcal{H}}_m = \bigcup_{S\in\mathcal{Z}^m} \mathcal{H}_S$. However, in general, $\overline{\mathcal{H}}_m$ can be very rich, which can lead to uninformative learning bounds. A somewhat better alternative consists of considering the union of all such hypothesis sets for samples of size m included in some supersample U of size $m+n$, with $n \geq 1$, $\overline{\mathcal{H}}_{U,m} = \bigcup_{\substack{S\in\mathcal{Z}^m \\ S\subseteq U}} \mathcal{H}_S$. We will derive learning guarantees based on the maximum *transductive Rademacher complexity* of $\overline{\mathcal{H}}_{U,m}$. There is a trade-off in the choice of n: smaller values lead to less complex sets $\overline{\mathcal{H}}_{U,m}$, but they also lead to weaker dependencies on sample sizes. Our bounds are more refined guarantees than the shattering-coefficient bounds originally given for this problem by Gat [2001] in the case $n = m$, and later by Cannon et al. [2002] for any $n \geq 1$. They also apply to arbitrary bounded loss functions and not just the binary loss. They are expressed in terms of the following notion of *transductive Rademacher complexity for data-dependent hypothesis sets*:

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \frac{1}{m+n}\sum_{i=1}^{m+n} \sigma_i L(h, z_i^U)\right],$$

where $U = (z_1^U, \ldots, z_{m+n}^U) \in \mathcal{Z}^{m+n}$ and where $\boldsymbol{\sigma}$ is a vector of $(m+n)$ independent random variables taking value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$, and $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$. Our notion of transductive Rademacher complexity is simpler than that of El-Yaniv and Pechyony [2007] (in the data-independent case) and leads to simpler proofs and guarantees. A by-product of our analysis is learning guarantees for standard transductive learning in terms of this notion of transductive Rademacher complexity, which can be of independent interest.

Theorem 1. *Let $\mathcal{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ be a family of data-dependent hypothesis sets. Let $\mathcal{G}$ be defined as in (3). Then, for any $\delta > 0$, with probability at least $1-\delta$ over the choice of the draw of the sample $S \sim \mathcal{Z}^m$, the following inequality holds for all $h \in \mathcal{H}_S$:*

$$R(h) \leq \widehat{R}_S(h) + \max_{U\in\mathcal{Z}^{m+n}} 2\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) + 3\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)\log\left(\tfrac{2}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn}.$$

Proof. (Sketch; full proof in Appendix D.) We use the following symmetrization result, which holds for any $\epsilon > 0$ with $m\epsilon^2 \geq 2$ for data-dependent hypothesis sets (Lemma 5, Appendix D):

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon\right] \leq 2 \mathop{\mathbb{P}}_{\substack{S\sim\mathcal{D}^m \\ T\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}_S} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right].$$

To bound the right-hand side, we use an extension of McDiarmid's inequality to sampling without replacement [Cortes et al., 2008] applied to $\Phi(S) = \sup_{h\in\overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h)$. Lemma 6 (Appendix D) is then used to bound $\mathbb{E}[\Phi(S)]$ in terms of our notion of transductive Rademacher complexity. □

4 Learning bound for stable data-dependent hypothesis sets

In this section, we present our main generalization bounds for data-dependent hypothesis sets.

Theorem 2. *Let $\mathcal{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ be a β-stable family of data-dependent hypothesis sets, with $\bar{\chi}$ average CV-stability, χ CV-stability and $\Delta_{\max}$ max-diameter. Let $\mathcal{G}$ be defined as in (3). Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample $S \sim \mathcal{Z}^m$, the following inequality*

holds for all $h \in \mathcal{H}_S$*:*

$$\forall h \in \mathcal{H}_S, R(h) \leq \widehat{R}_S(h) + \min\Bigg\{ \min\{2\mathfrak{R}^\circ_m(\mathcal{G}), \bar{\chi}\} + (1+2\beta m)\sqrt{\tfrac{1}{2m}\log(\tfrac{1}{\delta})}, \tag{9}$$

$$\sqrt{e}\chi + 4\sqrt{(\tfrac{1}{m}+2\beta)\log(\tfrac{6}{\delta})}, \tag{10}$$

$$48(3\beta + \Delta_{\max})\log(m)\log(\tfrac{5m^3}{\delta}) + \sqrt{\tfrac{4}{m}\log(\tfrac{4}{\delta})}\Bigg\}. \tag{11}$$

The proof of the theorem is given in Appendix E. The main idea is to control the sensitivity of the function $\Psi(S, S')$ defined for any two samples S, S', as follows:

$$\Psi(S, S') = \sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_{S'}(h).$$

To prove bound (9), we apply McDiarmid's inequality to $\Psi(S, S)$, using the $(\frac{1}{m} + 2\beta)$-sensitivity of $\Psi(S, S)$, and then upper bound the expectation $\mathbb{E}_{S\sim\mathcal{D}^m}[\Psi(S, S)]$ in terms of our notion of Rademacher complexity. The bound (10) is obtained via a differential-privacy-based technique, as in Feldman and Vondrak [2018], and (11) is a direct consequence of the bound of Feldman and Vondrak [2019] using the observation that an algorithm that chooses an arbitrary $h \in \mathcal{H}_S$ is $O(\beta + \Delta_{\max})$-uniformly stable in the classical [Bousquet and Elisseeff, 2002] sense.

Bound (9) admits as a special case the standard Rademacher complexity bound for fixed hypothesis sets [Koltchinskii and Panchenko, 2002, Bartlett and Mendelson, 2002]: in that case, we have $\mathcal{H}_S = \mathcal{H}$ for some $\mathcal{H}$, thus $\mathfrak{R}^\circ_m(\mathcal{G})$ coincides with the standard Rademacher complexity $\mathfrak{R}_m(\mathcal{G})$; furthermore, the family of hypothesis sets is 0-stable, thus the bound holds with $\beta = 0$.

Bounds (10) and (11) specialize to the bounds of Feldman and Vondrak [2018] and Feldman and Vondrak [2019] respectively for the special case of standard uniform stability of algorithms, since in that case, $\mathcal{H}_S$ is reduced to a singleton, $\mathcal{H}_S = \{h_S\}$, and so $\Delta = 0$, which implies that $\chi \leq \Delta + \beta = \beta$.

The Rademacher complexity-based bound (9) typically gives the tightest control on generalization error compared to the bounds (10) and (11), which rely on the cruder diameter notion. However the diameter may be easier to bound for some applications than the Rademacher complexity. To compare the diameter-based bounds, in applications where $\Delta_{\max} = O(\Delta)$, bound (11) may be tighter than (10). But, in several applications, we have $\beta = O(\frac{1}{m})$, and then bound (10) is tighter.

5 Applications

We now discuss several applications of our learning guarantees, with some others in Appendix G.

5.1 Bagging

Bagging [Breiman, 1996] is a prominent ensemble method used to improve the stability of learning algorithms. It consists of generating k new samples $B_1, B_2, \ldots, B_k$, each of size p, by sampling uniformly with replacement from the original sample S of size m. An algorithm $\mathcal{A}$ is then trained on each of these samples to generate k predictors $\mathcal{A}(B_i)$, $i \in [k]$. In regression, the predictors are combined by taking a convex combination $\sum_{i=1}^k w_i \mathcal{A}(B_i)$. Here, we analyze a common instance of bagging to illustrate the application of our learning guarantees: we will assume a regression setting and a uniform sampling from S *without replacement*.[2] We will also assume that the loss function is μ-Lipschitz in its first argument, that the predictions are in the range $[0, 1]$, and that all the mixing weights w_i are bounded by $\frac{C}{k}$ for some constant $C \geq 1$, in order to ensure that no subsample B_i is overly influential in the final regressor (in practice, a uniform mixture is typically used in bagging).

To analyze bagging, we cast it in our framework. First, to deal with the randomness in choosing the subsamples, we can equivalently imagine the process as choosing *indices* in $[m]$ to form the subsamples rather than samples in S, and then once S is drawn, the subsamples are generated by

[2] Sampling without replacement is only adopted to make the analysis more concise; its extension to sampling with replacement is straightforward.

filling in the samples at the corresponding indexes. For any index $i \in [m]$, the chance that it is picked in any subsample is $\frac{p}{m}$. Thus, by Chernoff's bound, with probability at least $1-\delta$, no index in $[m]$ appears in more than $t := \frac{kp}{m} + \sqrt{\frac{2kp\log(\frac{m}{\delta})}{m}}$ subsamples. In the following, we condition on the random seed of the bagging algorithm so that this is indeed the case, and later use a union bound to control the chance that the chosen random seed does not satisfy this property, as elucidated in section F.2.

Define the data-dependent family of hypothesis sets $\mathcal{H}$ as $\mathcal{H}_S := \left\{ \sum_{i=1}^k w_i \mathcal{A}(B_i) \colon w \in \Delta_k^{C/k} \right\}$, where $\Delta_k^{C/k}$ denotes the simplex of distributions over k items with all weights $w_i \leq \frac{C}{k}$. Next, we give upper bounds on the hypothesis set stability and the Rademacher complexity of $\mathcal{H}$. Assume that algorithm $\mathcal{A}$ admits uniform stability β_A [Bousquet and Elisseeff, 2002], i.e. for any two samples B and B' of size p that differ in exactly one data point and for all $x \in \mathcal{X}$, we have $|\mathcal{A}(B)(x) - \mathcal{A}(B')(x)| \leq \beta_{\mathcal{A}}$. Now, let S and S' be two samples of size m differing by one point at the same index, $z \in S$ and $z' \in S'$. Then, consider the subsets B_i' of S' which are obtained from the B_is by copying over all the elements except z, and replacing all instances of z by z'. For any B_i, if $z \notin B_i$, then $\mathcal{A}(B_i) = \mathcal{A}(B_i')$ and, if $z \in B_i$, then $|\mathcal{A}(B_i)(x) - \mathcal{A}(B_i')(x)| \leq \beta_{\mathcal{A}}$ for any $x \in \mathcal{X}$. We can now bound the hypothesis set uniform stability as follows: since L is μ-Lipschitz in the prediction, for any $z'' \in \mathcal{Z}$, and any $w \in \Delta_k^{C/k}$, we have

$$\left| L(\textstyle\sum_{i=1}^k w_i \mathcal{A}(B_i), z'') - L(\sum_{i=1}^k w_i \mathcal{A}(B_i'), z'') \right| \leq \left[\frac{p}{m} + \sqrt{\frac{2p\log(\frac{1}{\delta})}{km}} \right] \cdot C\mu\beta_{\mathcal{A}}.$$

It is easy to check the CV-stability and diameter of the hypothesis sets is $\Omega(1)$ in the worst case. Thus, the CV-stability-based bound (10) and standard uniform-stability bound (11) are not informative here, and we use the Rademacher complexity based bound (9) instead. Bounding the Rademacher complexity $\widehat{\mathfrak{R}}_S(\mathcal{H}_{S,T})$ for $S, T \in \mathcal{Z}^m$ is non-trivial. Instead, we can derive a reasonable upper bound by analyzing the Rademacher complexity of a larger function class. Specifically, for any $z \in \mathcal{Z}$, define the $d := \binom{2m}{p}$-dimensional vector $u_z = \langle \mathcal{A}(B)(z) \rangle_{B \subseteq S \cup T, |B|=p}$. Then, the class of functions is $\mathcal{F}_{S,T} := \{ z \mapsto w^\top u_z : \ w \in \mathbb{R}^d, \ \|w\|_1 = 1 \}$. Clearly $\mathcal{H}_{S,T} \subseteq \mathcal{F}_{S,T}$. Since $\|u_z\|_\infty \leq 1$, a standard Rademacher complexity bound (see Theorem 11.15 in [Mohri et al., 2018]) implies $\widehat{\mathfrak{R}}_S(\mathcal{F}_{S,T}) \leq \sqrt{\frac{2\log\left(2\binom{2m}{p}\right)}{m}} \leq \sqrt{\frac{2p\log(4m)}{m}}$. Thus, by Talagrand's inequality, we conclude that $\widehat{\mathfrak{R}}_S(\mathcal{G}_{S,T}) \leq \mu\sqrt{\frac{2p\log(4m)}{m}}$. In view of that, by Theorem 2, for any $\delta > 0$, with probability at least $1-2\delta$ over the draws of a sample $S \sim \mathcal{D}^m$ and the randomness in the bagging algorithm, the following inequality holds for any $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + 2\mu\sqrt{\frac{2p\log(4m)}{m}} + \left[1 + 2\left[p + \sqrt{\frac{2pm\log(\frac{1}{\delta})}{k}} \right] \cdot C\mu\beta_{\mathcal{A}} \right] \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

For $p = o(\sqrt{m})$ and $k = \omega(p)$, the generalization gap goes to 0 as $m \to \infty$, *regardless* of the stability of $\mathcal{A}$. This gives a new generalization guarantee for bagging, similar (but incomparable) to the one derived by Elisseeff et al. [2005]. One major point of difference is that unlike their bound, our bound allows for non-uniform averaging schemes.

5.2 Stochastic strongly-convex optimization

Here, we consider data-dependent hypothesis sets based on stochastic strongly-convex optimization algorithms. As shown by Shalev-Shwartz et al. [2010], uniform convergence bounds do not hold for the stochastic convex optimization problem in general.

Consider K stochastic strongly-convex optimization algorithms $\mathcal{A}_j$, each returning vector $\widehat{w}_j^S$, after receiving sample $S \in \mathcal{Z}^m$, $j \in [K]$. As shown by Shalev-Shwartz et al. [2010], such algorithms are $\beta = O(\frac{1}{m})$ sensitive in their output vector, i.e. for all $j \in [K]$, we have $\|\widehat{w}_j^S - \widehat{w}_j^{S'}\| \leq \beta$ if S and S' differ by one point.

Assume that the loss $L(w, z)$ is μ-Lipschitz with respect to its first argument w. Let the data-dependent hypothesis set be defined as follows: $\mathcal{H}_S = \left\{ \sum_{j=1}^K \alpha_j \widehat{w}_j^S \colon \alpha \in \Delta_K \cap \mathsf{B}_1(\alpha_0, r) \right\}$, where

α_0 is in the simplex of distributions Δ_K and $\mathsf{B}_1(\alpha_0, r)$ is the L_1 ball of radius $r > 0$ around α_0. We choose $r = \frac{1}{2\mu D\sqrt{m}}$. A natural choice for α_0 would be the uniform mixture.

Since the loss function is μ-Lipschitz, the family of hypotheses $\mathcal{H}_S$ is $\mu\beta$-stable. In this setting, bounding the Rademacher complexity is difficult, so we resort to the diameter based bound (10) instead. Note that for any $\alpha, \alpha' \in \Delta_K \cap \mathsf{B}_1(\alpha_0, r)$ and any $z \in \mathcal{Z}$, we have

$$L\Big(\sum_{j=1}^K \alpha_j \widehat{w}_j^S, z\Big) - L\Big(\sum_{j=1}^K \alpha'_j \widehat{w}_j^S, z\Big) \leq \mu \Big\| \sum_{j=1}^K (\alpha_i - \alpha'_j)\widehat{w}_j^S \Big\|_2 \leq \mu \big\|[w_1^S \cdots w_K^S]\big\|_{1,2} \|\alpha - \alpha'\|_1 \leq 2\mu r D,$$

where $\big\|[w_1^S \cdots w_K^S]\big\|_{1,2} := \max_{x \neq 0} \frac{\|\sum_{j=1}^k x_j w_j^S\|_2}{\|x\|_1} = \max_{i \in [K]} \|w_i^S\|_2 \leq D$. Thus, the average diameter admits the following upper bound: $\widehat{\Delta} \leq 2\mu r D = \frac{1}{\sqrt{m}}$. In view of that, by Theorem 2, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\alpha \in \Delta_K \cap \mathsf{B}_1(\alpha_0, r)$:

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}}\left[L\Big(\sum_{j=1}^K \alpha_j \widehat{w}_j^S, z\Big)\right] \leq \frac{1}{m}\sum_{i=1}^m L\Big(\sum_{j=1}^K \alpha_i \widehat{w}_j^S, z_i^S\Big) + \sqrt{\frac{e}{m}} + \sqrt{e}\mu\beta + 4\sqrt{\Big(\frac{1}{m} + 2\mu\beta\Big)\log\Big(\frac{6}{\delta}\Big)}.$$

The second stage of an algorithm in this context consists of choosing α, potentially using a non-stable algorithm. This application both illustrates the use of our learning bounds using the diameter and its application even in the absence of uniform convergence bounds.

As an aside, we note that the analysis of section 5.1 can be carried over to this setting, by setting $\mathcal{A}$ to be a stochastic strongly-convex optimization algorithm which outputs a weight vector $\hat{w}$. This yields generalization bounds for aggregating over a larger set of mixing weights, albeit with the restriction that each algorithm uses only a small part of S.

5.3 Δ-sensitive feature mappings

Consider the scenario where the training sample $S \in \mathcal{Z}^m$ is used to learn a non-linear feature mapping $\Phi_S \colon \mathcal{X} \to \mathbb{R}^N$ that is Δ-sensitive for some $\Delta = O(\frac{1}{m})$. Φ_S may be the feature mapping corresponding to some positive definite symmetric kernel or a mapping defined by the top layer of an artificial neural network trained on S, with a stability property.

To define the second state, let $\mathcal{L}$ be a set of γ-Lipschitz functions $f \colon \mathbb{R}^N \to \mathbb{R}$. Then we define $\mathcal{H}_S = \big\{x \mapsto f(\Phi_S(x)) \colon f \in \mathcal{L}\big\}$. Assume that the loss function ℓ is μ-Lipschitz with respect to its first argument. Then, for any hypothesis $h \colon x \mapsto f(\Phi_S(x)) \in \mathcal{H}_S$ and any sample S' differing from S by one element, the hypothesis $h' \colon x \mapsto f(\Phi_{S'}(x)) \in \mathcal{H}_{S'}$ admits losses that are β-close to those of h, with $\beta = \mu\gamma\Delta$, since, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, by the Cauchy-Schwarz inequality, the following inequality holds:

$$\ell(f(\Phi_S(x)), y) - \ell(f(\Phi_{S'}(x)), y) \leq \mu |f((\Phi_S(x)) - f(\Phi_{S'}(x))| \leq \mu\gamma \|\Phi_S(x) - \Phi_{S'}(x)\| \leq \mu\gamma\Delta.$$

Thus, the family of hypothesis set $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ is uniformly β-stable with $\beta = \mu\gamma\Delta = O(\frac{1}{m})$. In view of that, by Theorem 2, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S \sim \mathcal{D}^m$, the following inequality holds for any $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}_m^\circ(\mathcal{G}) + (1 + 2\mu\gamma\Delta m)\sqrt{\tfrac{1}{2m}\log(\tfrac{1}{\delta})}. \tag{12}$$

Notice that this bound applies even when the second stage of an algorithm, which consists of selecting a hypothesis h_S in $\mathcal{H}_S$, is not stable. A standard uniform stability guarantee cannot be used in that case. The setting described here can be straightforwardly extended to the case of other norms for the definition of sensitivity and that of the norm used in the definition of $\mathcal{H}_S$.

5.4 Distillation

Here, we consider *distillation algorithms* which, in the first stage, train a very complex model on the labeled sample. Let $f_S^* \colon \mathcal{X} \to \mathbb{R}$ denote the resulting predictor for a training sample S of size m. We will assume that the training algorithm is β-sensitive, that is $\|f_S^* - f_{S'}^*\| \leq \beta = O(\frac{1}{m})$ for S and S' differing by one point.

In the second stage, the algorithm selects a hypothesis that is γ-close to f^*_S from a less complex family of predictors $\mathcal{H}$. This defines the following sample-dependent hypothesis set: $\mathcal{H}_S = \{h \in \mathcal{H} \colon \|(h - f^*_S)\|_\infty \leq \gamma\}$.

Assume that the loss ℓ is μ-Lipschitz with respect to its first argument and that $\mathcal{H}$ is a subset of a vector space. Let S and S' be two samples differing by one point. Note, f^*_S may not be in $\mathcal{H}$, but we will assume that $f^*_{S'} - f^*_S$ is in $\mathcal{H}$. Let h be in $\mathcal{H}_S$, then the hypothesis $h' = h + f^*_{S'} - f^*_S$ is in $\mathcal{H}_{S'}$ since $\|h' - f^*_{S'}\|_\infty = \|h - f^*_S\|_\infty \leq \gamma$. Figure 2 illustrates the hypothesis sets. By the μ-Lipschitzness of the loss, for any $z = (x, y) \in \mathcal{Z}$, $|\ell(h'(x), y) - \ell(h(x), y)| \leq \mu \|h'(x) - h(x)\|_\infty = \mu \|f^*_{S'} - f^*_S\| \leq \mu\beta$. Thus, the family of hypothesis sets $\mathcal{H}_S$ is $\mu\beta$-stable.

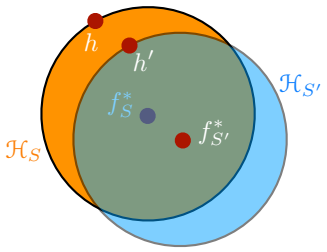

Figure 2: Illustration of the distillation hypothesis sets. Notice that the diameter of a hypothesis set $\mathcal{H}_S$ may be large here.

In view of that, by Theorem 2, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S \sim \mathcal{D}^m$, the following inequality holds for any $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}^\diamond_m(\mathcal{G}) + (1 + 2\mu\beta m)\sqrt{\tfrac{1}{2m}\log(\tfrac{1}{\delta})}.$$

Notice that a standard uniform-stability argument would not necessarily apply here since $\mathcal{H}_S$ could be relatively complex and the second stage not necessarily stable.

6 Conclusion

We presented a broad study of generalization with data-dependent hypothesis sets, including general learning bounds using a notion of transductive Rademacher complexity and, more importantly, learning bounds for stable data-dependent hypothesis sets. We illustrated the applications of these guarantees to the analysis of several problems. Our framework is general and covers learning scenarios commonly arising in applications for which standard generalization bounds are not applicable. Our results can be further augmented and refined to include model selection bounds and local Rademacher complexity bounds for stable data-dependent hypothesis sets (to be presented in a more extended version of this manuscript), and further extensions described in Appendix F. Our analysis can also be extended to the non-i.i.d. setting and other learning scenarios such as that of transduction. Several by-products of our analysis, including our proof techniques, new guarantees for transductive learning, and our PAC-Bayesian bounds for randomized algorithms, both in the sample-independent and sample-dependent cases, can be of independent interest. While we highlighted several applications of our learning bounds, a tighter analysis might be needed to derive guarantees for a wider range of data-dependent hypothesis classes or scenarios.

Acknowledgements. HL is supported by NSF IIS-1755781. The work of SG and MM was partly supported by NSF CCF-1535987, NSF IIS-1618662, and a Google Research Award. KS would like to acknowledge NSF CAREER Award 1750575 and Sloan Research Fellowship.

References

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning*, 3, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized Rademacher complexities. In *Proceedings of COLT*, volume 2375, pages 79–97. Springer-Verlag, 2002.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of STOC*, pages 1046–1059, 2016.

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

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *CoRR*, abs/1910.07833, 2019. URL http://arxiv.org/abs/1910.07833.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Adam Cannon, J. Mark Ettinger, Don R. Hush, and Clint Scovel. Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, 2:335–358, 2002.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics, 2007.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. In *Proceedings of NIPS*, pages 359–366, 2001.

Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In *Proceedings of ICML*, pages 176–183, 2008.

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

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Proceedings of NeurIPS*, pages 8440–8450, 2018a.

Gintare Karolina Dziugaite and Daniel M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-sgd and data-dependent priors. In *Proceedings of ICML*, pages 1376–1385, 2018b.

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher complexity and its applications. In *Proceedings of COLT*, pages 157–171, 2007.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Proceedings of NeurIPS*, pages 9770–9780, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of COLT*, 2019.

Yoram Gat. A learning generalization bound with an application to sparse-representation classifiers. *Machine Learning*, 42(3):233–239, 2001.

Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2002.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, pages 793–800, 2008.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Proceedings of Innovations in Computer Science*, pages 487–495, 2011.

Michael J. Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of COLT*, pages 152–162. ACM, 1997.

Vladmir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of UAI*, pages 275–282, 2002.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York, 1991.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

Andreas Maurer. A second-order look at stability and generalization. In *Proceedings of COLT*, pages 1461–1475, 2017.

David A. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of COLT*, pages 203–215, 2003.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of FOCS*, pages 94–103, 2007.

Kohei Miyaguchi. PAC-Bayesian transportation bound. *CoRR*, abs/1905.13435, 2019.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary phi-mixing and beta-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of ICLR*, 2018.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(Dec):3507–3531, 2012.

Petra Philips. *Data-Dependent Analysis of Learning Algorithms*. PhD thesis, Australian National University, 2005.

David Pollard. *Convergence of Stochastic Processess*. Springer, 1984.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 05 1978.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21, 2007.

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

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Information Theory*, 44(5):1926–1940, 1998.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *NIPS*, 2010.

Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *CoRR*, abs/1701.03493, 2017. URL http://arxiv.org/abs/1701.03493.

G. W. Stewart. *Matrix Algorithms: Volume 1, Basic Decompositions*. Society for Industrial and Applied Mathematics, 1998.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

A Further background on stability

The study of stability dates back to early work on the analysis of k-neareast neighbor and other local discrimination rules [Rogers and Wagner, 1978, Devroye and Wagner, 1979]. Stability has been critically used in the analysis of stochastic optimization [Shalev-Shwartz et al., 2010] and online-to-batch conversion [Cesa-Bianchi et al., 2001]. Stability bounds have been generalized to the non-i.i.d. settings, including stationary [Mohri and Rostamizadeh, 2010] and non-stationary [Kuznetsov and Mohri, 2017] ϕ-mixing and β-mixing processes. They have also been used to derive learning bounds for transductive inference [Cortes et al., 2008]. Stability bounds were further extended to cover *almost stable* algorithms by Kutin and Niyogi [2002]. These authors also discussed a number of alternative definitions of stability, see also [Kearns and Ron, 1997]. An alternative notion of stability was also used by Kale et al. [2011] to analyze k-fold cross-validation for a number of stable algorithms.

B Properties of data-dependent Rademacher complexity

In this section, we highlight several key properties of our notion of data-dependent Rademacher complexity.

B.1 Upper-bound on Rademacher complexity of data-dependent hypothesis sets

Lemma 2. *For any sample $S = (x_1^S, \ldots, x_m^S) \in \mathbb{R}^N$, define the hypothesis set $\mathcal{H}_S$ as follows:*

$$\mathcal{H}_S = \left\{ x \mapsto w^S \cdot x \colon w^S = \sum_{i=1}^m \alpha_i x_i^S, \|\alpha\|_1 \leq \Lambda_1 \right\},$$

where $\Lambda_1 \geq 0$. Define r_T and $r_{S\cup T}$ as follows: $r_T = \sqrt{\frac{\sum_{i=1}^m \|x_i^T\|_2^2}{m}}$ and $r_{S\cup T} = \max_{x\in S\cup T} \|x\|_2$. Then, the empirical Rademacher complexity of the family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S\in\mathcal{X}^m}$ can be upper-bounded as follows:

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) \leq r_T\, r_{S\cup T} \Lambda_1 \sqrt{\frac{2\log(4m)}{m}} \leq r_{S\cup T}^2 \Lambda_1 \sqrt{\frac{2\log(4m)}{m}}.$$

Proof. The following inequalities hold:

$$\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) = \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}} \sum_{i=1}^m \sigma_i h(x_i^T)\right] &= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\alpha\|_1\leq\Lambda_1} \sum_{i=1}^m \sigma_i \sum_{j=1}^m \alpha_j x_j^{S_{T,\boldsymbol{\sigma}}} \cdot x_i^T\right] \\
&= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\alpha\|_1\leq\Lambda_1} \sum_{j=1}^m \alpha_j \left(x_j^{S_{T,\boldsymbol{\sigma}}} \sum_{i=1}^m \sigma_i \cdot x_i^T\right)\right] \\
&= \frac{\Lambda_1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[m]} \left| x_j^{S_{T,\boldsymbol{\sigma}}} \cdot \sum_{i=1}^m \sigma_i x_i^T \right|\right] \\
&\leq \frac{\Lambda_1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{\substack{x'\in S\cup T \\ \sigma'\in\{-1,+1\}}} \sum_{i=1}^m \sigma_i (\sigma' x' \cdot x_i^T)\right].
\end{aligned}$$

The norm of the vector $z' \in \mathbb{R}^m$ with coordinates $(\sigma' x' \cdot x_i^T)$ can be bounded as follows:

$$\sqrt{\sum_{i=1}^m (\sigma' x' \cdot x_i^T)^2} \leq \|x'\| \sqrt{\sum_{i=1}^m \|x_i^T\|^2} \leq r_{S\cup T}\sqrt{m}\, r_T.$$

Thus, by Massart's lemma, since $|S\cup T| \leq 2m$, the following inequality holds:

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) \leq r_T\, r_{S\cup T} \Lambda_1 \sqrt{\frac{2\log(4m)}{m}} \leq r_{S\cup T}^2 \Lambda_1 \sqrt{\frac{2\log(4m)}{m}},$$

which completes the proof. □

Notice that the bound on the Rademacher complexity in Lemma 2is non-trivial since it depends on the samples S and T, while a standard Rademacher complexity for non-data-dependent hypothesis set containing $\mathcal{H}_S$ would require taking a maximum over all samples S of size m.

Lemma 3. *Suppose $\mathcal{X} = \mathbb{R}^N$, and for every sample $S \in \mathcal{Z}^m$ we associate a matrix $A_S \in \mathbb{R}^{d\times N}$ for some $d > 0$, and let $\mathcal{W}_{S,\Lambda} = \{w \in \mathbb{R}^d : \|A_S^\intercal w\|_2 \leq \Lambda\}$ for some $\Lambda > 0$. Consider the hypothesis set $\mathcal{H}_S := \left\{ x \mapsto w^\top A_S x \colon w \in \mathcal{W}_{S,\Lambda} \right\}$. Then, the empirical Rademacher complexity of the family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ can be upper-bounded as follows:*

$$\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) \leq \frac{\Lambda\sqrt{\sum_{i=1}^m \|x_i^T\|_2^2}}{m} \leq \frac{\Lambda r}{\sqrt{m}},$$

where $r = \sup_{i\in[m]} \|x_i^T\|_2$.

Proof. Let $X_T = [x_1^T \cdots x_m^T]$. The following inequalities hold:

$$\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{H}) = \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\sum_{i=1}^{m}\sigma_i h(x_i^T)\right] &= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{w:\ \|A_S^\top w\|_2\le\Lambda} w^\top A_S X_T\boldsymbol{\sigma}\right] \\
&\le \frac{\Lambda}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\|X_T\boldsymbol{\sigma}\|_2\right] && \text{(Cauchy-Schwarz)} \\
&\le \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\|X_T\boldsymbol{\sigma}\|_2^2\right]} && \text{(Jensen's ineq.)} \\
&\le \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{m}\sigma_i\sigma_j(x_i^T\cdot x_j^T)\right]} \\
&= \frac{\Lambda\sqrt{\sum_{i=1}^{m}\|x_i^T\|_2^2}}{m},
\end{aligned}$$

which completes the proof. □

B.2 Concentration

Lemma 4. *Let $\mathcal{H}$ a family of β-stable data-dependent hypothesis sets. Then, for any $\delta > 0$, with probability at least $1-\delta$ (over the draw of two samples S and T with size m), the following inequality holds:*

$$\left|\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{G}) - \mathfrak{R}^{\diamond}_{m}(\mathcal{G})\right| \le \sqrt{\frac{[(m\beta+1)^2 + m^2\beta^2]\log\frac{2}{\delta}}{2m}}.$$

Proof. Let T' be a sample differing from T only by point. Fix $\eta > 0$. For any $\boldsymbol{\sigma}$, by definition of the supremum, there exists $h' \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T'}$ such that:

$$\sum_{i=1}^{m}\sigma_i L(h', z_i^T) \ge \sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T'}}\sum_{i=1}^{m}\sigma_i L(h, z_i^{T'}) - \eta.$$

By the β-stability of $\mathcal{H}$, there exists $h \in \mathcal{H}^{\boldsymbol{\sigma}}_{S,T}$ such that for any $z \in \mathcal{Z}$, $|L(h', z) - L(h, z)| \le \beta$. Thus, we have

$$\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T'}}\sum_{i=1}^{m}\sigma_i L(h, z_i^{T'}) \le \sum_{i=1}^{m}\sigma_i L(h', z_i^{T'}) + \eta \le \sum_{i=1}^{m}[\sigma_i(L(h, z_i^{T'}) + \beta)] + \eta.$$

Since the inequality holds for all $\eta > 0$, we have

$$\frac{1}{m}\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T'}}\sum_{i=1}^{m}\sigma_i L(h, z_i^{T'}) \le \frac{1}{m}\sum_{i=1}^{m}\sigma_i(L(h, z_i^{T'}) + \beta) \le \frac{1}{m}\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\sum_{i=1}^{m}\sigma_i L(h, z_i^T) + \beta + \frac{1}{m}.$$

Thus, replacing T by T' affects $\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{G})$ by at most $\beta + \frac{1}{m}$. By the same argument, changing sample S by one point modifies $\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{G})$ at most by β. Thus, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds:

$$\left|\widehat{\mathfrak{R}}^{\diamond}_{S,T}(\mathcal{G}) - \mathfrak{R}^{\diamond}_{m}(\mathcal{G})\right| \le \sqrt{\frac{[(m\beta+1)^2 + m^2\beta^2]\log\frac{2}{\delta}}{2m}}.$$

This completes the proof. □

C Proof of Lemma 1

Proof. Let $S \in \mathcal{Z}^m$, $z \in S$, and $z' \in \mathcal{Z}$. For any $h \in \mathcal{H}_S$ and $h' \in \mathcal{H}_{S^{z \leftrightarrow z'}}$, by the β-uniform stability of $\mathcal{H}$, there exists $h'' \in \mathcal{H}_S$ such that $L(h', z) - L(h'', z) \leq \beta$. Thus,

$$L(h', z) - L(h, z) = L(h', z) - L(h'', z) + L(h'', z) - L(h, z) \leq \beta + \sup_{h'', h \in \mathcal{H}_S} L(h'', z) - L(h, z).$$

This implies the inequality

$$\sup_{h \in \mathcal{H}_S, h' \in \mathcal{H}_{S^{z \leftrightarrow z'}}} L(h', z) - L(h, z) \leq \beta + \sup_{h'', h \in \mathcal{H}_S} L(h'', z) - L(h, z),$$

and the lemma follows. □

D Proof of Theorem 1

In this section, we present the proof of Theorem 1.

Proof. We will use the following symmetrization result, which holds for any $\epsilon > 0$ with $n\epsilon^2 \geq 2$ for data-dependent hypothesis sets (Lemma 5 below):

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon\right] \leq 2 \mathop{\mathbb{P}}_{\substack{S\sim\mathcal{D}^m \\ T\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}_S} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right]. \tag{13}$$

Thus, we will seek to bound the right-hand side as follows, where we write $(S,T) \sim U$ to indicate that the sample S of size m is drawn uniformly without replacement from U and that T is the remaining part of U, that is $(S,T) = U$:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\substack{S\sim\mathcal{D}^m \\ T\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}_S} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right] \\
&= \mathop{\mathbb{E}}_{U\sim\mathcal{D}^{m+n}}\left[\mathop{\mathbb{P}}_{\substack{(S,T)\sim U \\ |S|=m,|T|=n}}\left[\sup_{h\in\mathcal{H}_S} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right] \middle| U\right] \\
&\leq \mathop{\mathbb{E}}_{U\sim\mathcal{D}^{m+n}}\left[\mathop{\mathbb{P}}_{\substack{(S,T)\sim U \\ |S|=m,|T|=n}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right] \middle| U\right].
\end{aligned}$$

To upper bound the probability inside the expectation, we use an extension of McDiarmid's inequality to sampling without replacement [Cortes et al., 2008], which applies to symmetric functions. We can apply that extension to $\Phi(S) = \sup_{h\in\overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h)$, for a fixed U, since $\Phi(S)$ is a symmetric function of the sample points $z_1, \ldots, z_m$) in S. Changing one point in S affects $\Phi(S)$ at most by $\frac{1}{m} + \frac{1}{m} = \frac{m+u}{mu}$, thus, by the extension of McDiarmid's inequality to sampling without replacement, for a fixed $U \in \mathcal{Z}^{m+n}$, the following inequality holds:

$$\mathop{\mathbb{P}}_{\substack{(S,T)\sim U \\ |S|=m,|T|=n}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right] \leq \exp\left[-\frac{2}{\eta}\frac{mn}{m+n}\left(\frac{\epsilon}{2} - \mathbb{E}[\Phi(S)]\right)^2\right], \tag{14}$$

where $\eta = \frac{m+n}{m+n-\frac{1}{2}} \frac{1}{1-\frac{1}{2\max\{m,n\}}} \leq 3$. Plugging in the bound on $\mathbb{E}[\Phi(S)]$ of Lemma 6 below, and setting

$$\epsilon = \max_{U\in\mathcal{Z}^{m+n}} 2\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) + 3\sqrt{\left(\tfrac{1}{m} + \tfrac{1}{n}\right)\log\left(\tfrac{2}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m} + \tfrac{1}{n}\right)^3 mn},$$

which satisfies $n\epsilon^2 \geq 2$, it is easy to check that the RHS in (14) becomes smaller than $\frac{\delta}{2}$. This in turn implies, via (13), that the probability that $\sup_{h\in\mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon$ is at most δ, completing the proof. □

The following lemma shows that the standard symmetrization lemma holds for data-dependent hypothesis sets. This observation was already made by Gat [2001] (see also Lemma 2 in [Cannon et al., 2002]) for the symmetrization lemma of Vapnik [1998][p. 139], used by the author in the case $n = m$. However, that symmetrization lemma of Vapnik [1998] holds only for random variables taking values in $\{0, 1\}$ and its proof is not complete since the hypergeometric inequality is not proven.

Lemma 5. *Let $n \geq 1$ and fix $\epsilon > 0$ such that $n\epsilon^2 \geq 2$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon\right] \leq 2 \mathop{\mathbb{P}}_{\substack{S\sim\mathcal{D}^m \\ T\sim\mathcal{D}^n}}\left[\sup_{h\in\mathcal{H}_S} \widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}\right].$$

Proof. The proof is standard. Below, we are giving a concise version mainly for the purpose of verifying that the data-dependency of the hypothesis set does not affect its correctness.

Fix $\eta > 0$. By definition of the supremum, there exists $h_S \in \mathcal{H}_S$ such that

$$\sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_S(h) - \eta \leq R(h_S) - \widehat{R}_S(h_S).$$

Since $\widehat{R}_T(h_S) - \widehat{R}_S(h_S) = \widehat{R}_T(h_S) - R(h_S) + R(h_S) - \widehat{R}_S(h_S)$, we can write

$$1_{\widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}} \geq 1_{\widehat{R}_T(h_S) - R(h_S) > -\frac{\epsilon}{2}} 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon} = 1_{R(h_S) - \widehat{R}_T(h_S) < \frac{\epsilon}{2}} 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon}.$$

Thus, for any $S \in \mathcal{Z}^m$, taking the expectation of both sides with respect to T yields

$$\begin{aligned}
\mathop{\mathbb{P}}_{T \sim \mathcal{D}^n}\left[\widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}\right] &\geq \mathop{\mathbb{P}}_{T \sim \mathcal{D}^n}\left[R(h_S) - \widehat{R}_T(h_S) < \frac{\epsilon}{2}\right] 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon} \\
&= \left[1 - \mathop{\mathbb{P}}_{T \sim \mathcal{D}^n}\left[R(h_S) - \widehat{R}_T(h_S) \geq \frac{\epsilon}{2}\right]\right] 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon} \\
&\leq \left[1 - \frac{4 \operatorname{Var}[L(h_S, z)]}{n\epsilon^2}\right] 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon} \qquad \text{(Chebyshev's ineq.)} \\
&\geq \left[1 - \frac{1}{n\epsilon^2}\right] 1_{R(h_S) - \widehat{R}_S(h_S) > \epsilon},
\end{aligned}$$

where the last inequality holds since $L(h_S, z)$ takes values in $[0, 1]$:

$$\begin{aligned}
\operatorname{Var}[L(h_S, z)] = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L^2(h_S, z)] - \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h_S, z)]^2 &\leq \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h_S, z)] - \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h_S, z)]^2 \\
&= \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h_S, z)](1 - \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L(h_S, z)]) \leq \frac{1}{4}.
\end{aligned}$$

Taking expectation with respect to S gives

$$\begin{aligned}
\mathop{\mathbb{P}}_{\substack{S \sim \mathcal{D}^m \\ T \sim \mathcal{D}^n}}\left[\widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}\right] &\geq \left[1 - \frac{1}{n\epsilon^2}\right] \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left[R(h_S) - \widehat{R}_S(h_S) > \epsilon\right] \\
&\geq \frac{1}{2} \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left[R(h_S) - \widehat{R}_S(h_S) > \epsilon\right] \qquad (n\epsilon^2 \geq 2) \\
&\geq \frac{1}{2} \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon + \eta\right].
\end{aligned}$$

Since the inequality holds for all $\eta > 0$, by the right-continuity of the cumulative distribution function, it implies

$$\mathop{\mathbb{P}}_{\substack{S \sim \mathcal{D}^m \\ T \sim \mathcal{D}^n}}\left[\widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}\right] \geq \frac{1}{2} \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_S(h) > \epsilon\right].$$

Since h_S is in $\mathcal{H}_S$, by definition of the supremum, we have

$$\mathop{\mathbb{P}}_{\substack{S \sim \mathcal{D}^m \\ T \sim \mathcal{D}^n}}\left[\sup_{h \in \mathcal{H}_S} \widehat{R}_T(h) - \widehat{R}_S(h) > \frac{\epsilon}{2}\right] \geq \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{D}^m \\ T \sim \mathcal{D}^n}}\left[\widehat{R}_T(h_S) - \widehat{R}_S(h_S) > \frac{\epsilon}{2}\right],$$

which completes the proof. □

The following lemma provides a bound on $\mathbb{E}[\Phi(S)]$:

Lemma 6. *Fix $U \in \mathcal{Z}^{m+n}$. Then, the following upper bound holds:*

$$\mathop{\mathbb{E}}_{\substack{(S,T) \sim U \\ |S|=m, |T|=n}}\left[\sup_{h \in \overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h)\right] \leq \widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) + \sqrt{\frac{\log(2e)(m+n)^3}{2(mn)^2}}.$$

For $m = n$, the inequality becomes:

$$\mathop{\mathbb{E}}_{\substack{(S,T) \sim U \\ |S|=m, |T|=n}}\left[\sup_{h \in \overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h)\right] \leq \widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G}) + 2\sqrt{\frac{\log(2e)}{m}}.$$

Proof. The proof is an extension of the analysis of *maximum discrepancy* in [Bartlett and Mendelson, 2002]. Let $|\boldsymbol{\sigma}|$ denote $\sum_{i=1}^{m+n}\sigma_i$ and let $I \subseteq \left[-\frac{(m+n)^2}{m}, \frac{(m+n)^2}{n}\right]$ denote the set of values $|\boldsymbol{\sigma}|$ can take. For any $q \in I$, define $s(q)$ as follows:

$$s(q) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \frac{1}{m+n}\sum_{i=1}^{m+n}\sigma_i L(h, z_i^U)\middle| |\boldsymbol{\sigma}| = q\right].$$

Let $|\boldsymbol{\sigma}|_+$ denote the number of positive σ_is, taking value $\frac{m+n}{n}$, then $|\boldsymbol{\sigma}|$ can be expressed as follows:

$$|\boldsymbol{\sigma}| = \sum_{i=1}^{m+n}\sigma_i = |\boldsymbol{\sigma}|_+\frac{m+n}{n} - (m+n-|\boldsymbol{\sigma}|_+)\frac{m+n}{m} = \frac{(m+n)^2}{mn}(|\boldsymbol{\sigma}|_+ - n). \tag{15}$$

Thus, we have $|\boldsymbol{\sigma}| = 0$ iff $|\boldsymbol{\sigma}|_+ = m$, and the condition $(|\boldsymbol{\sigma}| = 0)$ precisely corresponds to having the equality

$$\frac{1}{m+n}\sum_{i=1}^{m+n}\sigma_i L(h, z_i^U) = \widehat{R}_T(h) - \widehat{R}_S(h),$$

where S is the sample of size m defined by those z_is for which σ_i takes value $\frac{m+n}{n}$. In view of that, we have

$$\mathop{\mathbb{E}}_{\substack{(S,T)\sim U\\ |S|=m,|T|=n}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h)\right] = s(0).$$

Let $q_1, q_2 \in I$, with $q_1 = p_1\frac{m+n}{n} - (m+n-p_1)\frac{m+n}{m}$, $q_2 = p_2\frac{m+n}{n} - (m+n-p_2)\frac{m+n}{m}$ and $q_1 \leq q_2$. Then, we can write

$$s(q_1) = \mathbb{E}\left[\sup_{g\in G}\sum_{i=1}^{p_1}\frac{1}{n}L(h, z_i) - \sum_{i=p_1+1}^{m+n}\frac{1}{m}L(h, z_i)\right]$$

$$s(q_2) = \mathbb{E}\left[\sup_{g\in G}\sum_{i=1}^{p_1}\frac{1}{n}L(h, z_i) - \sum_{i=p_1+1}^{m+n}\frac{1}{m}L(h, z_i) + \sum_{i=p_1+1}^{p_2}\left[\frac{1}{n}+\frac{1}{m}\right]L(h, z_i)\right].$$

Thus, we have the following Lipschitz property:

$$\begin{aligned}
|s(q_2) - s(q_1)| \leq |p_2 - p_1|\left[\frac{1}{m}+\frac{1}{n}\right] &= |(p_2 - n) - (p_1 - n)|\left[\frac{1}{m}+\frac{1}{n}\right] \qquad \text{(using (15))}\\
&= |q_2 - q_1|\frac{mn}{(m+n)^2}\left[\frac{1}{m}+\frac{1}{n}\right]\\
&= \frac{|q_2 - q_1|}{m+n}.
\end{aligned}$$

By this Lipschitz property, we can write

$$\mathbb{P}\left[\left|s(|\boldsymbol{\sigma}|) - s(\mathbb{E}[|\boldsymbol{\sigma}|])\right| > \epsilon\right] \leq \mathbb{P}\left[\left||\boldsymbol{\sigma}| - \mathbb{E}[|\boldsymbol{\sigma}|]\right| > (m+n)\epsilon\right] \leq 2\exp\left[-2\frac{(mn)^2\epsilon^2}{(m+n)^3}\right],$$

since the range of each σ_i is $\frac{m+n}{n} + \frac{m+n}{m} = \frac{(m+n)^2}{mn}$. We now use this inequality to bound the second moment of $Z = s(|\boldsymbol{\sigma}|) - s(\mathbb{E}[|\boldsymbol{\sigma}|]) = s(|\boldsymbol{\sigma}|) - s(0)$, as follows, for any $u \geq 0$:

$$\begin{aligned}
\mathbb{E}[Z^2] &= \int_0^{+\infty}\mathbb{P}[Z^2 > t]\,dt\\
&= \int_0^u \mathbb{P}[Z^2 > t]\,dt + \int_u^{+\infty}\mathbb{P}[Z^2 > t]\,dt\\
&\leq u + 2\int_u^{+\infty}\exp\left[-2\frac{(mn)^2 t}{(m+n)^3}\right]dt\\
&\leq u + \left[\frac{(m+n)^3}{(mn)^2}\exp\left[-2\frac{(mn)^2 t}{(m+n)^3}\right]\right]_u^{+\infty}\\
&= u + \frac{(m+n)^3}{(mn)^2}\exp\left[-2\frac{(mn)^2 u}{(m+n)^3}\right].
\end{aligned}$$

Choosing $u = \frac{1}{2}\frac{\log(2)(m+n)^3}{(mn)^2}$ to minimize the right-hand side gives $\mathbb{E}[Z^2] \le \frac{\log(2e)(m+n)^3}{2(mn)^2}$. By Jensen's inequality, this implies $\mathbb{E}[|Z|] \le \sqrt{\frac{\log(2e)(m+n)^3}{2(mn)^2}}$ and therefore

$$\mathop{\mathbb{E}}_{\substack{(S,T)\sim U \\ |S|=m,|T|=n}} \left[\sup_{h \in \overline{\mathcal{H}}_{U,m}} \widehat{R}_T(h) - \widehat{R}_S(h) \right] = s(0) \le \mathbb{E}[s(|\boldsymbol{\sigma}|)] + \sqrt{\frac{\log(2e)(m+n)^3}{2(mn)^2}}.$$

Since we have $\mathbb{E}[s(|\boldsymbol{\sigma}|)] = \widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{G})$, this completes the proof. □

E Proof of Theorem 2

In this section, we present the full proof of Theorem 2. The proof of each of the three bounds (9), (10) and (11) are given in separate subsections.

E.1 Proof of bound (9)

Proof. For any two samples S, S', define the $\Psi(S, S')$ as follows:

$$\Psi(S, S') = \sup_{h \in \mathcal{H}_S} R(h) - \widehat{R}_{S'}(h).$$

The proof consists of applying McDiarmid's inequality to $\Psi(S, S)$. For any sample S' differing from S by one point, we can decompose $\Psi(S, S) - \Psi(S', S')$ as follows:

$$\Psi(S, S) - \Psi(S', S') = \big[\Psi(S, S) - \Psi(S, S')\big] + \big[\Psi(S, S') - \Psi(S', S')\big].$$

Now, by the sub-additivity of the $\sup$ operation, the first term can be upper-bounded as follows:

$$\begin{aligned}\Psi(S, S) - \Psi(S, S') &\leq \sup_{h \in \mathcal{H}_S} \big[R(h) - \widehat{R}_S(h)\big] - \big[R(h) - \widehat{R}_{S'}(h)\big] \\ &\leq \sup_{h \in \mathcal{H}_S} \frac{1}{m}\big[L(h, z) - L(h, z')\big] \leq \frac{1}{m},\end{aligned}$$

where we denoted by z and z' the labeled points differing in S and S' and used the 1-boundedness of the loss function.

We now analyze the second term:

$$\Psi(S, S') - \Psi(S', S') = \sup_{h \in \mathcal{H}_S} \big[R(h) - \widehat{R}_{S'}(h)\big] - \sup_{h \in \mathcal{H}_{S'}} \big[R(h) - \widehat{R}_{S'}(h)\big].$$

By definition of the supremum, for any $\epsilon > 0$, there exists $h \in \mathcal{H}_S$ such that

$$\sup_{h \in \mathcal{H}_S} \big[R(h) - \widehat{R}_{S'}(h)\big] - \epsilon \leq \big[R(h) - \widehat{R}_{S'}(h)\big]$$

By the β-stability of $(H_S)_{S \in \mathcal{Z}^m}$, there exists $h' \in \mathcal{H}_{S'}$ such that for all z, $|L(h, z) - L(h', z)| \leq \beta$. In view of these inequalities, we can write

$$\begin{aligned}\Psi(S, S') - \Psi(S', S') &\leq \big[R(h) - \widehat{R}_{S'}(h)\big] + \epsilon - \sup_{h \in \mathcal{H}_{S'}} \big[R(h) - \widehat{R}_{S'}(h)\big] \\ &\leq \big[R(h) - \widehat{R}_{S'}(h)\big] + \epsilon - \big[R(h') - \widehat{R}_{S'}(h')\big] \\ &\leq \big[R(h) - R(h')\big] + \epsilon + \big[\widehat{R}_{S'}(h') - \widehat{R}_{S'}(h)\big] \\ &\leq \epsilon + 2\beta.\end{aligned}$$

Since the inequality holds for any $\epsilon > 0$, it implies that $\Psi(S, S') - \Psi(S', S') \leq 2\beta$. Summing up the bounds on the two terms shows the following:

$$\Psi(S, S) - \Psi(S', S') \leq \frac{1}{m} + 2\beta.$$

Thus, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\Psi(S, S) \leq \mathbb{E}[\Psi(S, S)] + (1 + 2\beta m)\sqrt{\tfrac{1}{2m}\log(\tfrac{1}{\delta})}. \tag{16}$$

We now bound $\mathbb{E}[\Psi(S,S)]$ by $2\mathfrak{R}^{\diamond}_m(\mathcal{G})$ as follows:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\big[\Psi(S,S)\big] \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S}\big[R(h)-\widehat{R}_S(h)\big]\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S}\Big[\mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}\big[\widehat{R}_T(h)\big]-\widehat{R}_S(h)\Big]\right] && \text{(def. of } R(h)\text{)} \\
&\le \mathop{\mathbb{E}}_{S,T\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S}\widehat{R}_T(h)-\widehat{R}_S(h)\right] && \text{(sub-additivity of sup)} \\
&= \mathop{\mathbb{E}}_{S,T\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}_S}\frac{1}{m}\sum_{i=1}^m\Big[L(h,z_i^T)-L(h,z_i^S)\Big]\right] \\
&= \mathop{\mathbb{E}}_{S,T\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\frac{1}{m}\sum_{i=1}^m\sigma_i\Big[L(h,z_i^T)-L(h,z_i^S)\Big]\right]\right] && \text{(symmetry)} \\
&\le \mathop{\mathbb{E}}_{\substack{S,T\sim\mathcal{D}^m\\ \boldsymbol{\sigma}}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\frac{1}{m}\sum_{i=1}^m\sigma_iL(h,z_i^T)+\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\frac{1}{m}\sum_{i=1}^m-\sigma_iL(h,z_i^S)\right] && \text{(sub-additivity of sup)} \\
&= \mathop{\mathbb{E}}_{\substack{S,T\sim\mathcal{D}^m\\ \boldsymbol{\sigma}}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\frac{1}{m}\sum_{i=1}^m\sigma_iL(h,z_i^T)+\sup_{h\in\mathcal{H}^{-\boldsymbol{\sigma}}_{T,S}}\frac{1}{m}\sum_{i=1}^m-\sigma_iL(h,z_i^S)\right] && (\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}=\mathcal{H}^{-\boldsymbol{\sigma}}_{T,S}) \\
&= \mathop{\mathbb{E}}_{\substack{S,T\sim\mathcal{D}^m\\ \boldsymbol{\sigma}}}\left[\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{S,T}}\frac{1}{m}\sum_{i=1}^m\sigma_iL(h,z_i^T)+\sup_{h\in\mathcal{H}^{\boldsymbol{\sigma}}_{T,S}}\frac{1}{m}\sum_{i=1}^m\sigma_iL(h,z_i^S)\right] && \text{(symmetry)} \\
&= 2\mathfrak{R}^{\diamond}_m(\mathcal{G}). && \text{(linearity of expectation)}
\end{aligned}
$$

Now, we show that $\mathbb{E}_{S\sim\mathcal{D}^m}[\Psi(S,S)]\le\bar{\chi}$. To do so, first fix $\epsilon>0$. By definition of the supremum, for any $S\in\mathcal{Z}^m$, there exists h_S such that the following inequality holds:

$$\sup_{h\in\mathcal{H}_S}\big[R(h)-\widehat{R}_S(h)\big]-\epsilon\le R(h_S)-\widehat{R}_S(h_S).$$

Now, by definition of $R(h_S)$, we can write

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\big[R(h_S)\big]=\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\big(L(h_S,z)\right]=\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z\sim\mathcal{D}}}\big[L(h_S,z)\big].$$

Then, by the linearity of expectation, we can also write

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\big[\widehat{R}_S(h_S)\big]=\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z\sim S}}\big[L(h_S,z)\big]=\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}\\ z\sim S}}\big[L(h_{S^{z\leftrightarrow z'}},z')\big].$$

In view of these two equalities, we can now rewrite the upper bound as follows:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\big[\Psi(S,S)\big] &\le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\big[R(h_S)-\widehat{R}_S(h_S)\big]+\epsilon \\
&= \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}}}\big[L(h_S,z')\big]-\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}\\ z\sim S}}\big[L(h_{S^{z\leftrightarrow z'}},z')\big]+\epsilon \\
&= \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}\\ z\sim S}}\big[L(h_S,z')-L(h_{S^{z\leftrightarrow z'}},z')\big]+\epsilon \\
&= \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ z'\sim\mathcal{D}\\ z\sim S}}\big[L(h_{S^{z\leftrightarrow z'}},z)-L(h_S,z)\big]+\epsilon \\
&\le \bar{\chi}+\epsilon.
\end{aligned}
$$

Since the inequality holds for all $\epsilon>0$, it implies $\mathbb{E}_{S\sim\mathcal{D}^m}\big[\Psi(S,S)\big]\le\bar{\chi}$. Plugging in these upper bounds on the expectation in the inequality (16) completes the proof. □

E.2 Proof of bound (10)

The proof of bound (10) relies on recent techniques introduced in the differential privacy literature to derive improved generalization guarantees for stable data-dependent hypothesis sets [Steinke and Ullman, 2017, Bassily et al., 2016] (see also [McSherry and Talwar, 2007]). Our proof also benefits from the recent improved stability results of Feldman and Vondrak [2018]. We will make use of the following lemma due to Steinke and Ullman [2017, Lemma 1.2], which reduces the task of deriving a concentration inequality to that of upper bounding an expectation of a maximum.

Lemma 7. *Fix $p \geq 1$. Let X be a random variable with probability distribution $\mathcal{D}$ and $X_1, \ldots, X_p$ independent copies of X. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{X\sim\mathcal{D}}\left[X \geq 2 \mathop{\mathbb{E}}_{X_k\sim\mathcal{D}}\left[\max\left\{0, X_1, \ldots, X_p\right\}\right]\right] \leq \frac{\log 2}{p}.$$

We will also use the following result which, under a sensitivity assumption, further reduces the task of upper bounding the expectation of the maximum to that of bounding a more favorable expression.

Lemma 8 ([McSherry and Talwar, 2007, Bassily et al., 2016, Feldman and Vondrak, 2018])**.** *Let $f_1, \ldots, f_p : \mathcal{Z}^m \to \mathbb{R}$ be p functions with sensitivity Δ. Let $\mathcal{A}$ be the algorithm that, given a dataset $S \in \mathcal{Z}^m$ and a parameter $\epsilon > 0$, returns the index $k \in [p]$ with probability proportional to $e^{\frac{\epsilon f_k(S)}{2\Delta}}$. Then, $\mathcal{A}$ is ϵ-differentially private and, for any $S \in \mathcal{Z}^m$, the following inequality holds:*

$$\max_{k\in[p]}\left\{f_k(S)\right\} \leq \mathop{\mathbb{E}}_{k=\mathcal{A}(\mathsf{S})}\left[f_k(S)\right] + \frac{2\Delta}{\epsilon}\log p.$$

Notice that, if we define $f_{p+1} = 0$, then, by the same result, the algorithm $\mathcal{A}$ returning the index $k \in [p+1]$ with probability proportional to $e^{\frac{\epsilon f_k(S)1_{k\neq(p+1)}}{2\Delta}}$ is ϵ-differentially private and the following inequality holds for any $S \in \mathcal{Z}^m$:

$$\max\left\{0, \max_{k\in[p]}\left\{f_k(S)\right\}\right\} = \max_{k\in[p+1]}\left\{f_k(S)\right\} \leq \mathop{\mathbb{E}}_{k=\mathcal{A}(\mathsf{S})}\left[f_k(S)\right] + \frac{2\Delta}{\epsilon}\log(p+1). \tag{17}$$

Equipped with these lemmas, we can now turn to the proof of bound (10):

Proof. For any two samples S, S' of size m, define $\Psi(S, S')$ as follows:

$$\Psi(S, S') = \sup_{h\in\mathcal{H}_S} R(h) - \widehat{R}_{S'}(h).$$

The proof consists of deriving a high-probability bound for $\Psi(S, S)$. To do so, by Lemma 7 applied to the random variable $X = \Psi(S, S)$, it suffices to bound $\mathbb{E}_{\mathsf{S}\sim\mathcal{D}^{pm}}\left[\max\left\{0, \max_{k\in[p]}\left\{\Psi(S_k, S_k)\right\}\right\}\right]$, where $\mathsf{S} = (S_1, \ldots, S_p)$ with S_k, $k \in [p]$, independent samples of size m drawn from $\mathcal{D}^m$.

To bound that expectation, we can use Lemma 8 and instead bound $\mathbb{E}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[\Psi(S_k, S_k)\right]$, where $\mathcal{A}$ is an ϵ-differentially private algorithm.

Now, to apply Lemma 8, we first show that, for any $k \in [p]$, the function $f_k : \mathsf{S} \to \Psi(S_k, S_k)$ is Δ-sensitive with $\Delta = \frac{1}{m} + 2\beta$. Fix $k \in [p]$. Let $\mathsf{S}' = (S'_1, \ldots, S'_p)$ be in $\mathcal{Z}^{pm}$ and assume that S' differs from S by one point. If they differ by a point not in S_k (or S'_k), then $f_k(\mathsf{S}) = f_k(\mathsf{S}')$. Otherwise, they differ only by a point in S_k (or S'_k) and $f_k(\mathsf{S}) - f_k(\mathsf{S}') = \Psi(S_k, S_k) - \Psi(S'_k, S'_k)$. We can decompose this term as follows:

$$\Psi(S_k, S_k) - \Psi(S'_k, S'_k) = \left[\Psi(S_k, S_k) - \Psi(S_k, S'_k)\right] + \left[\Psi(S_k, S'_k) - \Psi(S'_k, S'_k)\right].$$

Now, by the sub-additivity of the sup operation, the first term can be upper-bounded as follows:

$$\begin{aligned}\Psi(S_k, S_k) - \Psi(S_k, S'_k) &\leq \sup_{h\in\mathcal{H}_{S_k}}\left[R(h) - \widehat{R}_{S_k}(h)\right] - \left[R(h) - \widehat{R}_{S'_k}(h)\right]\\ &\leq \sup_{h\in\mathcal{H}_{S_k}}\frac{1}{m}\left[L(h, z) - L(h, z')\right] \leq \frac{1}{m},\end{aligned}$$

where we denoted by z and z' the labeled points differing in S_k and S'_k and used the 1-boundedness of the loss function.

We now analyze the second term:

$$\Psi(S_k, S'_k) - \Psi(S'_k, S'_k) = \sup_{h \in \mathcal{H}_{S_k}} \left[R(h) - \widehat{R}_{S'_k}(h)\right] - \sup_{h \in \mathcal{H}_{S'_k}} \left[R(h) - \widehat{R}_{S'_k}(h)\right].$$

By definition of the supremum, for any $\eta > 0$, there exists $h \in \mathcal{H}_{S_k}$ such that

$$\sup_{h \in \mathcal{H}_{S_k}} \left[R(h) - \widehat{R}_{S'_k}(h)\right] - \eta \leq \left[R(h) - \widehat{R}_{S'_k}(h)\right]$$

By the β-stability of $(\mathcal{H}_S)_{S \in \mathcal{Z}^m}$, there exists $h' \in \mathcal{H}_{S'_k}$ such that for all z, $\left|L(h, z) - L(h', z)\right| \leq \beta$. In view of these inequalities, we can write

$$\begin{aligned}
\Psi(S_k, S'_k) - \Psi(S'_k, S'_k) &\leq \left[R(h) - \widehat{R}_{S'_k}(h)\right] + \eta - \sup_{h \in \mathcal{H}_{S'_k}} \left[R(h) - \widehat{R}_{S'_k}(h)\right] \\
&\leq \left[R(h) - \widehat{R}_{S'_k}(h)\right] + \eta - \left[R(h') - \widehat{R}_{S'_k}(h')\right] \\
&\leq \left[R(h) - R(h')\right] + \eta + \left[\widehat{R}_{S'_k}(h') - \widehat{R}_{S'_k}(h)\right] \\
&\leq \eta + 2\beta.
\end{aligned}$$

Since the inequality holds for any $\eta > 0$, it implies that $\Psi(S_k, S'_k) - \Psi(S'_k, S'_k) \leq 2\beta$. Summing up the bounds on the two terms shows the following:

$$\Psi(S_k, S_k) - \Psi(S'_k, S'_k) \leq \frac{1}{m} + 2\beta.$$

Having established the Δ-sensitivity of the functions f_k, $k \in [p]$, we can now apply Lemma 8. Fix $\epsilon > 0$. Then, by Lemma 8 and (17), the algorithm $\mathcal{A}$ returning $k \in [p+1]$ with probability proportional to $e^{\frac{\epsilon \Psi(S_k, S_k) 1_{k \neq (p+1)}}{2\Delta}}$ is ϵ-differentially private and, for any sample $\mathsf{S} \in \mathcal{Z}^{pm}$, the following inequality holds:

$$\max\left\{0, \max_{k \in [p]} \left\{\Psi(S_k, S_k)\right\}\right\} \leq \mathop{\mathbb{E}}_{k = \mathcal{A}(\mathsf{S})} \left[\Psi(S_k, S_k)\right] + \frac{2\Delta}{\epsilon} \log(p+1).$$

Taking the expectation of both sides yields

$$\mathop{\mathbb{E}}_{\mathsf{S} \sim \mathcal{D}^{pm}} \left[\max\left\{0, \max_{k \in [p]} \left\{\Psi(S_k, S_k)\right\}\right\}\right] \leq \mathop{\mathbb{E}}_{\substack{\mathsf{S} \sim \mathcal{D}^{pm} \\ k = \mathcal{A}(\mathsf{S})}} \left[\Psi(S_k, S_k)\right] + \frac{2\Delta}{\epsilon} \log(p+1). \tag{18}$$

We will show the following upper bound on the expectation: $\mathop{\mathbb{E}}_{\substack{\mathsf{S} \sim \mathcal{D}^{pm} \\ k = \mathcal{A}(\mathsf{S})}} \left[\Psi(S_k, S_k)\right] \leq (e^\epsilon - 1) + e^\epsilon \chi$. To do so, first fix $\eta > 0$. By definition of the supremum, for any $S \in \mathcal{Z}^m$, there exists $h_S \in \mathcal{H}_S$ such that the following inequality holds:

$$\sup_{h \in \mathcal{H}_S} \left[R(h) - \widehat{R}_S(h)\right] - \eta \leq R(h_S) - \widehat{R}_S(h_S).$$

In what follows, we denote by $\mathsf{S}^{k, z \leftrightarrow z'} \in \mathcal{Z}^{pm}$ the result of modifying $\mathsf{S} = (S_1, \ldots, S_p) \in \mathcal{Z}^{pm}$ by replacing $z \in S_k$ with z'.

Now, by definition of the algorithm $\mathcal{A}$, we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[R(h_{S_k})\right] &= \mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[\mathop{\mathbb{E}}_{z'\sim\mathcal{D}}[L(h_{S_k},z')]\right] && \text{(def. of } R(h_{S_k})\text{)}\\
&= \mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z'\sim\mathcal{D}}}\left[\sum_{k=1}^{p}\mathbb{P}[\mathcal{A}(\mathsf{S})=k]\,L(h_{S_k},z')\right] && \text{(def. of } \mathop{\mathbb{E}}_{k=\mathcal{A}(\mathsf{S})}\text{)}\\
&= \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z'\sim\mathcal{D}}}\left[\mathbb{P}[\mathcal{A}(\mathsf{S})=k]\,L(h_{S_k},z')\right] && \text{(linearity of expect.)}\\
&\le \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z'\sim\mathcal{D},\ z\sim S_k}}\left[e^{\epsilon}\,\mathbb{P}[\mathcal{A}(\mathsf{S}^{k,z\leftrightarrow z'})=k]\,L(h_{S_k},z')\right] && (\epsilon\text{-diff. privacy of } \mathcal{A})\\
&= \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z'\sim\mathcal{D},\ z\sim S_k}}\left[e^{\epsilon}\,\mathbb{P}[\mathcal{A}(\mathsf{S})=k]\,L(h_{S_k^{z\leftrightarrow z'}},z)\right] && \text{(swapping } z' \text{ and } z\text{)}\\
&\le \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z'\sim\mathcal{D},\ z\sim S_k}}\left[e^{\epsilon}\,\mathbb{P}[\mathcal{A}(\mathsf{S})=k]\,L(h_{S_k},z)\right]+e^{\epsilon}\chi. && \text{(By Lemma 9 below)}
\end{aligned}$$

Now, observe that $\mathbb{E}_{z\sim S_k}[L(h_{S_k},z)]$ coincides with $\widehat{R}(h_{S_k})$, the empirical loss of h_{S_k}. Thus, we can write

$$\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[R(h_{S_k})\right] \le \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ z\sim S_k}}\left[e^{\epsilon}\,\mathbb{P}[\mathcal{A}(\mathsf{S})=k]\,\widehat{R}_{S_k}(h_{S_k})\right]+e^{\epsilon}\chi,$$

and therefore

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[\Psi(S_k,S_k)\right] &\le \sum_{k=1}^{p}\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}\left[(e^{\epsilon}-1)\widehat{R}_{S_k}(h_{S_k})\right]+e^{\epsilon}\chi+\eta\\
&\le (e^{\epsilon}-1)+e^{\epsilon}\chi+\eta.
\end{aligned}$$

Since the inequality holds for any $\eta>0$, we have

$$\mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm}\\ k=\mathcal{A}(\mathsf{S})}}[\Psi(S_k,S_k)] \le (e^{\epsilon}-1)+e^{\epsilon}\chi.$$

Thus, by (18), the following inequality holds:

$$\mathop{\mathbb{E}}_{\mathsf{S}\sim\mathcal{D}^{pm}}\left[\max\left\{0,\max_{k\in[p]}\left\{\Psi(S_k,S_k)\right\}\right\}\right] \le (e^{\epsilon}-1)+e^{\epsilon}\chi+\frac{2\Delta}{\epsilon}\log(p+1). \tag{19}$$

For any $\delta\in(0,1)$, choose $p=\frac{\log 2}{\delta}$, which implies $\log(p+1)=\log\left[\frac{2+\delta}{\delta}\right]\le\log\frac{3}{\delta}$. Then, by Lemma 7, with probability at least $1-\delta$ over the draw of a sample $S\sim\mathcal{D}^m$, the following inequality holds for all $h\in\mathcal{H}_S$:

$$R(h)\le\widehat{R}_S(h)+(e^{\epsilon}-1)+e^{\epsilon}\chi+\frac{2\Delta}{\epsilon}\log\left[\frac{3}{\delta}\right]. \tag{20}$$

For $\epsilon\le\frac{1}{2}$, the inequality $(e^{\epsilon}-1)\le 2\epsilon$ holds. Thus,

$$(e^{\epsilon}-1)+e^{\epsilon}\chi+\frac{2\Delta}{\epsilon}\log\left[\frac{3}{\delta}\right]\le 2\epsilon+\sqrt{e}\chi+\frac{2\Delta}{\epsilon}\log\left[\frac{3}{\delta}\right]$$

Choosing $\epsilon=\sqrt{\Delta\log\left[\frac{3}{\delta}\right]}$ gives

$$\begin{aligned}
R(h) &\le \widehat{R}_S(h)+\sqrt{e}\chi+4\sqrt{\Delta\log\left[\frac{3}{\delta}\right]}\\
&= \widehat{R}_S(h)+\sqrt{e}\chi+4\sqrt{\left[\tfrac{1}{m}+2\beta\right]\log\left[\tfrac{3}{\delta}\right]}.
\end{aligned}$$

Combining this inequality with the inequality of Theorem 2 related to the Rademacher complexity:

$$\forall h \in \mathcal{H}_S, R(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}^{\diamond}_m(\mathcal{G}) + [1 + 2\beta m]\sqrt{\frac{\log \frac{1}{\delta}}{2m}}, \tag{21}$$

and using the union bound complete the proof. □

The following is a helper lemma for the analysis in the above proof:

Lemma 9. *The following upper bound in terms of the CV-stability coefficient χ holds:*

$$\sum_{k=1}^{p} \mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm} \\ z'\sim\mathcal{D},\ z\sim S_k}} \left[e^{\epsilon}\, \mathbb{P}[\mathcal{A}(\mathsf{S}) = k]\, [L(h_{S_k^{z\leftrightarrow z'}}, z) - L(h_{S_k}, z)] \right] \leq e^{\epsilon}\chi.$$

Proof. Upper bounding the difference of losses by a supremum to make the CV-stability coefficient appear gives the following chain of inequalities:

$$\begin{aligned}
&\sum_{k=1}^{p} \mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm} \\ z'\sim\mathcal{D},\ z\sim S_k}} \left[e^{\epsilon}\, \mathbb{P}[\mathcal{A}(\mathsf{S}) = k]\, [L(h_{S_k^{z\leftrightarrow z'}}, z) - L(h_{S_k}, z)] \right] \\
&\leq \sum_{k=1}^{p} \mathop{\mathbb{E}}_{\substack{\mathsf{S}\sim\mathcal{D}^{pm} \\ z'\sim\mathcal{D},\ z\sim S_k}} \left[e^{\epsilon}\, \mathbb{P}[\mathcal{A}(\mathsf{S}) = k] \sup_{h\in\mathcal{H}_{S_k},\, h'\in\mathcal{H}_{S_k^{z\leftrightarrow z'}}} [L(h', z) - L(h, z)] \right] \\
&= \sum_{k=1}^{p} \mathop{\mathbb{E}}_{\mathsf{S}\sim\mathcal{D}^{pm}} \left[e^{\epsilon}\, \mathbb{P}[\mathcal{A}(\mathsf{S}) = k] \mathop{\mathbb{E}}_{z'\sim\mathcal{D},\ z\sim S_k} \left[\sup_{h\in\mathcal{H}_{S_k},\, h'\in\mathcal{H}_{S_k^{z\leftrightarrow z'}}} [L(h', z) - L(h, z)] \mid \mathsf{S} \right] \right] \\
&\leq \sum_{k=1}^{p} \mathop{\mathbb{E}}_{\mathsf{S}\sim\mathcal{D}^{pm}} \left[e^{\epsilon}\, \mathbb{P}[\mathcal{A}(\mathsf{S}) = k]\, \chi \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{S}\sim\mathcal{D}^{pm}} \left[\sum_{k=1}^{p} \mathbb{P}[\mathcal{A}(\mathsf{S}) = k] \right] \cdot e^{\epsilon}\chi \\
&= e^{\epsilon}\chi,
\end{aligned}$$

which completes the proof. □

E.3 Proof of bound (11)

Bound (11) is a simple consequence of the fact that the composition of the two stages of the learning algorithm is uniformly-stable in the classical sense. Specifically, consider a learning algorithm that consists of determining the hypothesis set $\mathcal{H}_S$ based on the sample S and then selecting an arbitrary (but fixed) hypothesis $h_S \in \mathcal{H}_S$. The following lemma shows that the uniform-stability coefficient of this learning algorithm can be bounded in terms of its hypothesis set stability and its max-diameter.

Lemma 10. *Suppose the family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ is β-uniformly stable and admits max-diameter $\Delta_{\max}$. Then, for any two samples $S, S' \in \mathcal{Z}^m$ differing in exactly one point, and for any $z \in \mathcal{Z}$, we have*

$$|L(h_S, z) - L(h_{S'}, z)| \leq 3\beta + \Delta_{\max}.$$

Proof. We first show that for any two hypotheses $h, h' \in \mathcal{H}_S$ and for any $z \in \mathcal{Z}$, we have $|L(h, z) - L(h', z)| \leq 2\beta + \Delta_{\max}$. Indeed, let S'' be a sample obtained by replacing an arbitrary point in S by z. Then, by β-uniform hypothesis set stability of $\mathcal{H}$, there exist hypotheses $g, g' \in \mathcal{H}_{S''}$ such that $|L(h, z) - L(g, z)| \leq \beta$ and $|L(h', z) - L(g', z)| \leq \beta$. Furthermore, since $z \in S''$, we have $|L(g, z) - L(g', z)| \leq \Delta_{\max}$. By combining these inequalities, we get that $|L(h, z) - L(h', z)| \leq 2\beta + \Delta_{\max}$, as required.

Now, let $h' \in \mathcal{H}_S$ be a hypothesis such that $|L(h', z) - L(h_{S'}, z)| \leq \beta$. Since $h', h_S \in \mathcal{H}_S$, by the analysis in the preceding paragraph, we have $|L(h_S, z) - L(h', z)| \leq 2\beta + \Delta_{\max}$. Combining these two inequalities, we have $|L(h_S, z) - L(h_{S'}, z)| \leq 3\beta + \Delta_{\max}$, completing the proof. □

Finally, bound (11) follows immediately from the following result of Feldman and Vondrak [2019], setting $\ell(S, z) := L(h_S, z)$ and $\gamma = 3\beta + \Delta_{\max}$, and the fact that any two hypotheses h and h' in $\mathcal{H}_S$ differ in loss on any point z by at most $\Delta_{\max}$ in order to get a bound which holds uniformly for all $h \in \mathcal{H}_S$.

Theorem 3 ([Feldman and Vondrak, 2019]). *Let $\ell : \mathcal{Z}^m \times \mathcal{Z} \to [0, 1]$ be a data-dependent function with uniform stability γ, i.e. for any $S, S' \in \mathcal{Z}^m$ differing in one point, and any $z \in \mathcal{Z}$, we have $|\ell(S, z) - \ell(S', z)| \leq \gamma$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the sample S, the following inequality holds:*

$$\left| \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(S, z)] - \mathop{\mathbb{E}}_{z \sim S}[\ell(S, z)] \right| \leq 47\gamma \log(m) \log(\tfrac{5m^3}{\delta}) + \sqrt{\tfrac{4}{m} \log(\tfrac{4}{\delta})}.$$

F Extensions

We briefly discuss here some extensions of the framework and results presented in the previous section.

F.1 Almost everywhere hypothesis set stability

As for standard algorithmic uniform stability, our generalization bounds for hypothesis set stability can be extended to the case where hypothesis set stability holds only with high probability [Kutin and Niyogi, 2002].

Definition 4. *Fix $m \geq 1$. We will say that a family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ is* weakly (β, δ)-stable *for some $\beta \geq 0$ and $\delta > 0$, if, with probability at least $1 - \delta$ over the draw of a sample $S \in \mathcal{Z}^m$, for any sample S' of size m differing from S only by one point, the following holds:*

$$\forall h \in \mathcal{H}_S, \exists h' \in \mathcal{H}_{S'} \colon \forall z \in \mathcal{Z}, |L(h, z) - L(h', z)| \leq \beta. \tag{22}$$

Notice that, in this definition, β and δ depend on the sample size m. In practice, we often have $\beta = O(\frac{1}{m})$ and $\delta = O(e^{-\Omega(m)})$. The learning bounds of Theorem 2 can be straightforwardly extended to guarantees for weakly (β, δ)-stable families of data-dependent hypothesis sets, by using a union bound and the confidence parameter δ.

F.2 Randomized algorithms

The generalization bounds given in this paper assume that the data-dependent hypothesis set $\mathcal{H}_S$ is *deterministic* conditioned on S. However, in some applications such as bagging, it is more natural to think of $\mathcal{H}_S$ as being constructed by a *randomized* algorithm with access to an independent source of randomness in the form of a random seed s. Our generalization bounds can be extended in a straightforward manner for this setting if the following can be shown to hold: there is a *good* set of seeds, G, such that (a) $\mathbb{P}[s \in G] \geq 1 - \delta$, where δ is the confidence parameter, and (b) conditioned on any $s \in G$, the family of data-dependent hypothesis sets $\mathcal{H} = (\mathcal{H}_S)_{S \in \mathcal{Z}^m}$ is β-uniformly stable. In that case, for any good set $s \in G$, Theorem 2 holds. Then taking a union bound, we conclude that with probability at least $1 - 2\delta$ over both the choice of the random seed s and the sample set S, the generalization bounds hold. This can be further combined with almost-everywhere hypothesis stability as in section F.1 via another union bound if necessary.

F.3 Data-dependent priors

An alternative scenario extending our study is one where, in the first stage, instead of selecting a hypothesis set $\mathcal{H}_S$, the learner decides on a probability distribution p_S on a fixed family of hypotheses $\mathcal{H}$. The second stage consists of using that *prior* p_S to choose a hypothesis $h_S \in \mathcal{H}$, either deterministically or via a randomized algorithm. Our notion of hypothesis set stability could then be extended to that of stability of priors and lead to new learning bounds depending on that stability parameter. This could lead to data-dependent prior bounds somewhat similar to the PAC-Bayesian bounds [Catoni, 2007, Parrado-Hernández et al., 2012, Lever et al., 2013, Dziugaite and Roy, 2018a,b], but with technically quite different guarantees.

G Other applications

G.1 Anti-distillation

A similar setup to distillation (section 5.4) is that of *anti-distillation* where the predictor f_S^* in the first stage is chosen from a simpler family, say that of linear hypotheses, and where the sample-dependent hypothesis set $\mathcal{H}_S$ is the subset of a very rich family $\mathcal{H}$. $\mathcal{H}_S$ is defined as the set of predictors that are close to f_S^*:

$$\mathcal{H}_S = \left\{h \in \mathcal{H} \colon (\|(h - f_S^*)\|_\infty \leq \gamma) \wedge (\|(h - f_S^*)1_S\|_\infty \leq \Delta)\right\},$$

with $\Delta = O(1/\sqrt{m})$. Thus, the restriction to S of a hypothesis $h \in \mathcal{H}_S$ is close to f_S^* in ℓ_∞-norm. As shown in section 5.4, the family of hypothesis sets $\mathcal{H}_S$ is $\mu\beta$-stable. However, here, the hypothesis sets $\mathcal{H}_S$ could be very complex and the Rademacher complexity $\mathfrak{R}_m^\diamond(\mathcal{H})$ not very favorable. Nevertheless, by Theorem 2, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample $S \sim \mathcal{D}^m$, the following inequality holds for any $h \in \mathcal{H}_S$:

$$R(h) \leq \widehat{R}_S(h) + \sqrt{e}\mu(\Delta + \beta) + 4\sqrt{(\tfrac{1}{m} + 2\mu\beta)\log(\tfrac{6}{\delta})}.$$

Notice that a standard uniform-stability does not apply here since the $(1/\sqrt{m})$-closeness of the hypotheses to f_S^* on S does not imply their global $(1/\sqrt{m})$-closeness.

G.2 Principal Components Regression

Principal Components Regression is a very commonly used technique in data analysis. In this setting, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, with a loss function ℓ that is μ-Lipschitz in the prediction. Given a sample $S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \colon i \in [m]\}$, we learn a linear regressor on the data projected on the principal k-dimensional space of the data. Specifically, let $\Pi_S \in \mathbb{R}^{d \times d}$ be the projection matrix giving the projection of $\mathbb{R}^d$ onto the principal k-dimensional subspace of the data, i.e. the subspace spanned by the top k left singular vectors of the design matrix $X_S = [x_1, x_2, \cdots, x_m]$. The hypothesis space $\mathcal{H}_S$ is then defined as $\mathcal{H}_S = \{x \mapsto w^\top \Pi_S x \colon w \in \mathbb{R}^k, \ \|w\| \leq \gamma\}$, where γ is a predefined bound on the norm of the weight vector for the linear regressor. Thus, this can be seen as an instance of the setting in section 5.3, where the feature mapping Φ_S is defined as $\Phi_S(x) = \Pi_S x$.

To prove generalization bounds for this setup, we need to show that these feature mappings are stable. To do that, we make the following assumptions:

1. For all $x \in \mathcal{X}$, $\|x\| \leq r$ for some constant $r \geq 1$.
2. The data covariance matrix $\mathbb{E}_x[xx^\top]$ has a gap of $\lambda > 0$ between the k-th and $(k+1)$-th largest eigenvalues.

The matrix concentration bound of Rudelson and Vershynin [2007] implies that with probability at least $1 - \delta$ over the choice of S, we have $\|X_S X_S^\top - m\,\mathbb{E}_x[xx^\top]\| \leq cr^2\sqrt{m\log(m)\log(\frac{2}{\delta})}$ for some constant $c > 0$. Suppose m is large enough so that $cr^2\sqrt{m\log(m)\log(\frac{2}{\delta})} \leq \frac{\lambda}{2}m$. Then, the gap between the k-th and $(k+1)$-th largest eigenvalues of $X_S X_S^\top$ is at least $\frac{\lambda}{2}m$. Now, consider changing one sample point $(x, y) \in S$ to (x, y') to produce the sample set S'. Then, we have $X_{S'} X_{S'}^\top = X_S X_S^\top - xx^\top + x'x'^\top$. Since $\|-xx^\top + x'x'^\top\| \leq 2r^2$, by standard matrix perturbation theory bounds [Stewart, 1998], we have $\|\Pi_S - \Pi_{S'}\| \leq O(\frac{r^2}{\lambda m})$. Thus, $\|\Phi_S(x) - \Phi_{S'}(x)\| \leq \|\Pi_S - \Pi_{S'}\|\|x\| \leq O(\frac{r^3}{\lambda m})$.

Now, to apply the bound of (12), we need to compute a suitable bound on $\mathfrak{R}_m^\diamond(\mathcal{H})$. For this, we apply Lemma 3. For any $\|w\| \leq \gamma$, since $\|\Pi_S\| = 1$, we have $\|\Pi_S w\| \leq \gamma$. So the hypothesis set $\mathcal{H}'_S = \{x \mapsto w^\top \Pi_S x \colon w \in \mathbb{R}^k, \ \|\Pi_S w\| \leq \gamma\}$ contains $\mathcal{H}_S$. By Lemma 3, we have $\mathfrak{R}_m^\diamond(\mathcal{H}') \leq \frac{\gamma r}{\sqrt{m}}$. Thus, by plugging the bounds obtained above in (12), we conclude that with probability at least $1 - 2\delta$ over the choice of S, for any $h \in \mathcal{H}_S$, we have

$$R(h) \leq \widehat{R}_S(h) + O\left(\mu\gamma\frac{r^3}{\lambda}\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right).$$

H PAC-Bayesian Bounds

The PAC-Bayes framework assumes a prior distribution P over $\mathcal{H}$ and a posterior distribution Q selected after observing the training sample. The framework helps derive learning bounds for randomized algorithms with probability distribution Q, in terms of the relative entropy of Q and P.

In this section, we briefly discuss PAC-Bayesian learning bounds and present some key results. In Subsection H.1, we give PAC-Bayes learning bounds derived from Rademacher complexity bounds, which improve upon standard PAC-Bayes bounds [McAllester, 2003]. Similar bounds were already shown by Kakade et al. [2008] using elegant proofs based on strong convexity. Here, we give an alternative proof not invoking strong convexity. In Subsection H.2, we extend the PAC-Bayes framework to one where the prior distribution is selected after observing S and will denote by P_S that prior. Finally, in Subsection H.3, we briefly discuss derandomized PAC-Bayesian bounds, that is learning bounds derived for deterministic algorithms, using PAC-Bayes bounds.

H.1 PAC-Bayes bounds derived from Rademacher complexity bounds

We will denote by L_z the vector $(L(h,z))_{h\in\mathcal{H}}$. The expected loss of the randomized classifier Q can then be written as $\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\langle Q, L_z\rangle\right]$.

Define $\mathcal{G}_\mu$ via $\mathcal{G}_\mu = \{Q \in \Delta(\mathcal{H}) \colon \mathsf{D}(Q \,\|\, P) \leq \mu\}$, that is the family of distributions Q defined over $\mathcal{H}$ with μ-bounded relative entropy with respect to P. Then, by the standard Rademacher complexity bound [Koltchinskii and Panchenko, 2002, Mohri et al., 2018], for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample S of size m, the following holds for all $Q \in \mathcal{G}_\mu$:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\langle Q, L_z\rangle\right] \leq \mathop{\mathbb{E}}_{z\sim S}\left[\langle Q, L_z\rangle\right] + 2\mathfrak{R}_m(\mathcal{G}_\mu) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{23}$$

We now give an upper bound on $\mathfrak{R}_m(\mathcal{G}_\mu)$. For any Q, define $\Psi(Q)$ by $\Psi(Q) = \mathsf{D}(Q,P)$ if $Q \in \Delta(\mathcal{H})$ and $+\infty$ otherwise. It is known that the conjugate function Ψ^* of Ψ is given by $\Psi^*(U) = \log\left(\mathbb{E}_{h\in P}[e^{U(h)}]\right)$, for all $U \in \mathbb{R}^{\mathcal{H}}$ (see for example [Mohri et al., 2018, Lemma B.37]). Let $U_{\boldsymbol{\sigma}} = \sum_{i=1}^m \sigma_i L_{z_i}$. Then, for any $t > 0$, we can write:

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{G}_\mu) &= \frac{1}{m}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P)\leq\mu}\sum_{i=1}^m \sigma_i\langle Q, L_{z_i}\rangle\right]\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P)\leq\mu}\langle Q, U_{\boldsymbol{\sigma}}\rangle\right] && \text{(definition of } U_{\boldsymbol{\sigma}}\text{)}\\
&= \frac{1}{mt}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P)\leq\mu}\langle Q, tU_{\boldsymbol{\sigma}}\rangle\right] && (t>0)\\
&\leq \frac{1}{mt}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{\Psi(Q)\leq\mu}\Psi(Q) + \Psi^*(tU_{\boldsymbol{\sigma}})\right] && \text{(Fenchel inequality)}\\
&\leq \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\Psi^*(tU_{\boldsymbol{\sigma}})\right].
\end{aligned}$$

Now, we use the expression of Ψ^* to bound the second term as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\Psi^*(tU_{\boldsymbol{\sigma}})\right] &= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\log\left(\mathop{\mathbb{E}}_{h\sim P}\left[e^{t\sum_{i=1}^m \sigma_i L(h,z_i)}\right]\right)\right]\\
&\leq \mathop{\mathbb{E}}_{S}\left[\log\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma},h\sim P}\left[e^{t\sum_{i=1}^m \sigma_i L(h,z_i)}\right]\right)\right] && \text{(Jensen's inequality)}\\
&= \mathop{\mathbb{E}}_{S}\left[\log\left(\mathop{\mathbb{E}}_{h\sim P}\left[\prod_{i=1}^m \cosh(tL(h,z_i))\right]\right)\right]\\
&\leq \mathop{\mathbb{E}}_{S}\left[\log\left(\mathop{\mathbb{E}}_{h\sim P}\left[e^{m\frac{t^2}{2}}\right]\right)\right] = \frac{mt^2}{2}.
\end{aligned}$$

Choosing $t = \sqrt{\frac{2\mu}{m}}$ to minimize the bound on the Rademacher complexity gives $\mathfrak{R}_m(\mathcal{G}_\mu) \leq \sqrt{\frac{2\mu}{m}}$. In view of that, (23) implies:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\langle Q, L_z\rangle\right] \leq \mathop{\mathbb{E}}_{z\sim S}\left[\langle Q, L_z\rangle\right] + 2\sqrt{\frac{2\mu}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{24}$$

Proceeding as in [Kakade et al., 2008], by the union bound, the result can be extended to hold for any distribution Q, which is directly leading to the following result.

Theorem 4. *Let P be a fixed prior on $\mathcal{H}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample S of size m, the following holds for any posterior distribution Q over $\mathcal{H}$:*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^m L(h,z_i)\right] + \left(4 + \tfrac{1}{\sqrt{e}}\right)\sqrt{\frac{\max\{\mathsf{D}(Q\,\|\,P), 1\}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

This bound improves upon standard PAC-Bayes bounds (see for example [McAllester, 2003]) since it does not include an additive term in $\sqrt{(\log m)/m}$, as pointed by Kakade et al. [2008].

H.2 Data-dependent PAC-Bayes bounds

In this section, we extend the framework to one where the prior distribution is selected after observing S and will denote by P_S that prior. To analyze that scenario, we can both use the general data-dependent learning bounds of Section 3, or the hypothesis set stability bounds of Section 4. We will focus here on the latter.

Define the data-dependent hypothesis set $\mathcal{G}_{S,\mu} = \{Q \in \Delta(\mathcal{H}) \colon \mathsf{D}(Q\,\|\,P_S) \leq \mu\}$ and assume that the priors P_S are chosen so that $\mathcal{G}_\mu = (\mathcal{G}_{S,\mu})_S$ is β-stable. This may be by choosing P_S and $P_{S'}$ to be close in total variation or relative entropy for any two samples S and S' differing by one point. Then, by Theorem 2, for any $\delta > 0$, with probably at least $1-\delta$, the following holds for all $Q \in \mathcal{G}_{\mu,S}$:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^m L(h,z_i)\right] + \min\Bigg\{ &\min\left\{2\mathfrak{R}^\diamond_m(\mathcal{G}_\mu), \beta + \bar{\Delta}\right\} + (1+2\beta m)\sqrt{\tfrac{1}{2m}\log(\tfrac{1}{\delta})},\\ &\sqrt{e}(\beta+\Delta) + 4\sqrt{(\tfrac{1}{m}+2\beta)\log(\tfrac{6}{\delta})},\\ &48(3\beta+\Delta_{\max})\log(m)\log(\tfrac{5m^3}{\delta}) + \sqrt{\tfrac{4}{m}\log(\tfrac{4}{\delta})}\Bigg\}.\end{aligned}$$

The analysis of the Rademacher complexity $\mathfrak{R}^\diamond_m(\mathcal{G}_\mu)$ depends on the specific properties of the family of priors P_S. Here, we initiate its analysis and leave it to the reader to complete it for a choice of the priors.

Proceeding as in Subsection H.1, we define Ψ_S by $\Psi_S(Q) = \mathsf{D}(Q, P_S)$ for any $Q \in \Delta(\mathcal{H})$ and denote by Ψ^*_S its conjugate function. Let $U_{\boldsymbol{\sigma}} = \sum_{i=1}^m \sigma_i L_{z_i^T}$. Then, for any $t > 0$, we can write:

$$\begin{aligned}\mathfrak{R}^\diamond_m(\mathcal{G}_\mu) &= \frac{1}{m}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P_{S^{\boldsymbol{\sigma}}_T})\leq\mu}\sum_{i=1}^m \sigma_i\langle Q, L_{z_i}\rangle\right]\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P_{S^{\boldsymbol{\sigma}}_T})\leq\mu}\langle Q, U_{\boldsymbol{\sigma}}\rangle\right] && \text{(definition of } U_{\boldsymbol{\sigma}}\text{)}\\
&= \frac{1}{mt}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P_{S^{\boldsymbol{\sigma}}_T})\leq\mu}\langle Q, tU_{\boldsymbol{\sigma}}\rangle\right] && (t>0)\\
&\leq \frac{1}{mt}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\Psi_{S^{\boldsymbol{\sigma}}_T}(Q)\leq\mu}\Psi_{S^{\boldsymbol{\sigma}}_T}(Q) + \Psi^*_{S^{\boldsymbol{\sigma}}_T}(tU_{\boldsymbol{\sigma}})\right] && \text{(Fenchel inequality)}\\
&\leq \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\Psi^*_{S^{\boldsymbol{\sigma}}_T}(tU_{\boldsymbol{\sigma}})\right].\end{aligned}$$

Using the expression of the conjugate function $\Psi^*_{S^{\boldsymbol{\sigma}}_T}$, as in Subsection H.1, the second term can be bounded as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\Psi^*_{S^{\boldsymbol{\sigma}}_T}(tU_{\boldsymbol{\sigma}})\right] &= \mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\log\left(\mathop{\mathbb{E}}_{h\sim P_{S^{\boldsymbol{\sigma}}_T}}\left[e^{t\sum_{i=1}^m \sigma_i L(h,z_i)}\right]\right)\right] \\ &\leq \mathop{\mathbb{E}}_{S,T}\left[\log\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma},h\sim P_{S^{\boldsymbol{\sigma}}_T}}\left[e^{t\sum_{i=1}^m \sigma_i L(h,z_i)}\right]\right)\right] \qquad \text{(Jensen's inequality)}.\end{aligned}$$

This last term can be bounded using Hoeffding's inequality and the specific properties of the priors leading to an explicit bound on the Rademacher complexity as in Subsection H.1.

H.3 Derandomized PAC-Bayesian bounds

Derandomized versions of PAC-Bayesian bounds have been given in the past: margin bounds for linear predictors by McAllester [2003], more complex margin bounds by Neyshabur et al. [2018] where linear predictors are replaced with neural networks and where the norm-bound is replaced with a more complex norm condition, and chaining-based bounds by Miyaguchi [2019].

However, the benefit of these bounds is not clear since finer Rademacher complexity bounds can be derived for deterministic predictors. In fact, Rademacher complexity bounds can be used to derive finer PAC-Bayes bounds than existing ones. This was already pointed out by Kakade et al. [2008] and further shown here with an alternative proof and more favorable constants (Subsection H.1).

In fact, using the technique of obtaining KL-divergence between prior and posterior as upper bound on the Rademacher complexity, along with the optimistic rates in [Srebro et al., 2010], one can obtain just as in the previous section, an optimistic rate with data-dependent prior when one considers a non-negative smooth loss and, as predictor, the expected model under the posterior. As this is a straightforward application of the result of Srebro et al. [2010] combined with techniques presented here, we leave this for the reader to verify by themselves.