PAC-Bayes Learning Bounds for Sample-Dependent Priors

Pranjal Awasthi
Google Research and
Rutgers University
pranjalawasthi@google.com

Satyen Kale
Google Research
satyenkale@google.com

Stefani Karp
Google Research and
Carnegie Mellon University
stefanik@google.com

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

Abstract

We present a series of new PAC-Bayes learning guarantees for randomized algorithms with sample-dependent priors. Our most general bounds make no assumption on the priors and are given in terms of certain covering numbers under the infinite-Rényi divergence and the ℓ_1 distance. We show how to use these general bounds to derive leaning bounds in the setting where the sample-dependent priors obey an infinite-Rényi divergence or ℓ_1-distance sensitivity condition. We also provide a flexible framework for computing PAC-Bayes bounds, under certain stability assumptions on the sample-dependent priors, and show how to use this framework to give more refined bounds when the priors satisfy an infinite-Rényi divergence sensitivity condition.

1 Introduction

The PAC-Bayesian framework provides generalization bounds for the performance of randomized learning algorithms [McAllester, 1999b,a, Shawe-Taylor and Williamson, 1997]. Rather than outputting a single hypothesis, such algorithms output a probability distribution Q (the posterior) over a hypothesis set $\mathcal{H}$. In the PAC-Bayes framework, the generalization guarantees associated with Q are typically expressed in terms of the relative entropy, $\mathsf{D}(Q \parallel P)$, where P is a fixed prior distribution over the hypothesis set. In the traditional framework, the prior P must be selected before receiving a training sample [Langford and Caruana, 2002, Langford and Seeger, 2001, Seeger, 2002].

In recent years, there have been efforts to establish more refined PAC-Bayes bounds in which the prior can depend on the distribution generating the data or a separate sample drawn from the same distribution [Catoni, 2007, Ambroladze et al., 2007, Parrado-Hernández et al., 2012, Lever et al., 2013]. However, in practice, information about the underlying data distribution is available only via the training sample and discarding a fraction of that data to compute a generalization bound can be wasteful, motivating the study of PAC-Bayes bounds for *sample-dependent* priors.[1]

In the context of overparameterized deep neural networks, where deriving non-vacuous generalization bounds is notoriously hard, it has been argued that sample-dependent priors can lead to finer generalization bounds [Nagarajan and Kolter, 2019, Dziugaite and Roy, 2017, Neyshabur et al., 2018].

[1] The same notion has also been called *data-dependent* priors in the literature [Dziugaite and Roy, 2018a, Negrea et al., 2019, Dziugaite et al., 2020, Haghifam et al., 2020].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Sample-dependent priors can also lead to new learning methods. For instance, when training deep neural networks via stochastic gradient descent, a standard choice for the prior P is a Gaussian centered around the parameters at random initialization. In this case, $\mathsf{D}(Q \parallel P)$ is related to the distance from initialization of the final iterate's parameters. This, however, can be large in most realistic settings and it is therefore more appealing to choose as prior a Gaussian centered around parameters obtained by running some amount of pre-training using the training data, and subsequently use that prior as a guide for fine-tuning the parameters with additional training. This combination of pre-training followed by fine-tuning is common practice and clearly such a choice would be sample-dependent. Sample-dependent priors are also relevant in emerging scenarios such as adversarial training. A common practice here is to smooth a given classifier by injecting Gaussian noise into the inputs. This results in a classifier with a more favorable Lipschitz property, thereby improving robustness [Lecuyer et al., 2019, Cohen et al., 2019]. While the choice of the noise magnitude depends on the input, typically, these methods choose a priori a uniform noise magnitude across all inputs. It is much more appealing instead to choose a posterior over the noise magnitudes and inform this choice by carefully selecting a prior P based on the sample, over the noise magnitudes, and using the prior P as a regularizer to guide the search for the posterior.

From a theoretical perspective, there has been little work on generalization bounds for sample-dependent priors. The recent work of [Dziugaite and Roy, 2018a,b] took an important step in this direction by showing that for sample-dependent priors chosen via a differentially private mechanism PAC-Bayesian generalization bounds can be derived. They also showed that weaker conditions where the sample-dependent prior need only be "close" to a differentially private prior suffice for the bounds. We also recently became aware of [Rivasplata et al., 2020], which will appear at NeurIPS 2020 as well; this work also discusses general sample-dependent priors, although it is not yet apparent how the results compare. The following are our main contributions:

1. **General bounds via covering numbers**. We give general PAC-Bayes bounds, with no assumption on the sample-dependent priors, in terms of certain covering numbers of the priors. We provide two such bounds using covering numbers computed with the infinite-Rényi divergence and the ℓ_1 distance.

2. **Bounds for stable priors**. We say that sample-dependent priors satisfy *prior stability*, if for any two samples S and S' that differ in exactly one input, the corresponding sample-dependent priors P_S and $P_{S'}$ are close. *Closeness* here is measured either in terms of the infinite-Rényi divergence or the ℓ_1 distance. For both cases, we show that our general covering number based bounds already give non-trivial generalization bounds.

3. **Framework for PAC-Bayes bounds under prior stability**. Building on the work of Foster et al. [2019] on *hypothesis set stability*, we provide a general method for deriving PAC-Bayes bounds assuming prior stability. We show how this method leads to refinements of the PAC-Bayes bound mentioned above for infinite-Rényi divergence prior stability.

Related Work. Our work builds on a strong line of work using algorithmic stability to derive generalization bounds, in particular [Bousquet and Elisseeff, 2002, Feldman and Vondrak, 2018, 2019, Bousquet et al., 2019]. Most significantly, our work builds on the recent notion of *hypothesis set stability* introduced by Foster et al. [2019].

We note that our work is not the first to combine PAC-Bayesian bounds and stability-like notions. Rivasplata et al. [2018] derive PAC-Bayesian bounds by randomizing the learned hypothesis output by a stable learning algorithm. However, their priors are only distribution-dependent (vs. sample-dependent), and they do not invoke any stability of the *priors*. London [2017] combines PAC-Bayes bounds and algorithmic stability, but remains in the setting of fixed, sample-independent priors.

The work of Dziugaite and Roy [2018a] is perhaps the most closely related to ours. Specifically, to our knowledge, Dziugaite and Roy [2018a] presents the first example of actually using the full m-item sample S in order to generate a prior P_S. In particular, they assume that the priors are generated from samples via a *randomized* differentially-private mechanism. They then use results from the differential privacy literature (specifically [Dwork et al., 2015]) to show that, with high probability over both the choice of the sample *and* the sample-dependent prior, their PAC-Bayesian bounds hold with respect to this sample-dependent prior. In contrast, in this work we assume that priors are generated from samples in a *deterministic* manner, and furthermore, the mapping from samples to priors is *stable*, either in infinite-Rényi divergence or ℓ_1 distance. In the case of infinite-Rényi divergence stable priors, the priors themselves define a differentially-private mechanism for generating *hypotheses*. Thus, our setting is fundamentally different from that of Dziugaite and Roy [2018a].

2 Preliminaries

We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the input and output spaces, respectively. For convenience, we define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and denote by $\mathcal{D}$ a distribution over $\mathcal{Z}$ from which samples are drawn. We let $\mathcal{H}$ denote a hypothesis set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}'$, and use $\Delta(\mathcal{H})$ throughout to denote the set of distributions on $\mathcal{H}$.

We consider a loss function $\ell\colon \mathcal{Y}' \times \mathcal{Y} \to [0,1]$ and use $L(h,z)$ as shorthand to denote the composition $\ell(h(x),y)$. The expected loss of a randomized classifier parameterized by a distribution $Q \in \Delta(\mathcal{H})$ is the following expectation: $\mathbb{E}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)]$. For simplicity, we use L_z to denote the vector $(L(h,z))_{h\in\mathcal{H}}$, allowing us to rewrite $\mathbb{E}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)]$ as $\mathbb{E}_{z\sim\mathcal{D}}\left[\langle Q, L_z\rangle\right]$.

Since the priors in this paper are sample-dependent, we denote by $P_S \in \Delta(\mathcal{H})$ a prior obtained after seeing the sample S. For two distributions $\mathcal{P}, \mathcal{Q}$ defined on the same discrete domain[2] Ω, we use $\mathsf{D}(\mathcal{P} \parallel \mathcal{Q}) = \mathbb{E}_{\omega\sim\mathcal{P}}\left[\log\left(\frac{\mathcal{P}(\omega)}{\mathcal{Q}(\omega)}\right)\right]$ to denote the relative entropy (or KL divergence) of $\mathcal{P}$ from $\mathcal{Q}$, and we use $\mathsf{D}_\infty(\mathcal{P} \parallel \mathcal{Q})$ to denote the infinite-Rényi divergence (or max-divergence, as often seen in the differential privacy literature) of $\mathcal{P}$ from $\mathcal{Q}$, defined as follows: $\mathsf{D}_\infty(\mathcal{P} \parallel \mathcal{Q}) = \sup_{\omega\in\Omega}\log\left(\frac{\mathcal{P}(\omega)}{\mathcal{Q}(\omega)}\right)$. We will also need the notion of γ-approximate infinite-Rényi divergence, denoted $\mathsf{D}_\infty^\gamma(\mathcal{P} \parallel \mathcal{Q})$ for any two distributions $\mathcal{P}, \mathcal{Q}$: $\mathsf{D}_\infty^\gamma(\mathcal{P} \parallel \mathcal{Q}) := \sup_{A\subseteq\Omega\colon \mathcal{P}(A)\geq\gamma}\log\left(\frac{\mathcal{P}(A)-\gamma}{\mathcal{Q}(A)}\right)$. Finally, we use $\|\mathcal{P}-\mathcal{Q}\|_{\mathrm{TV}} = \frac{1}{2}\|\mathcal{P}-\mathcal{Q}'\|_1$ to denote the total variation distance between $\mathcal{P}$ and $\mathcal{Q}$.

Our bounds are stated in terms of Rademacher complexity, defined as follows. Let $S = (z_1, z_2, \ldots, z_m) \in \mathcal{Z}^m$ be a sample set sampled from $\mathcal{D}^m$, and let $\boldsymbol{\sigma} \in \{-1,1\}^m$ be a vector of independent Rademacher variables. The notions of Rademacher complexity we need are:[3]

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i L(h,z_i)\right] \quad\text{and}\quad \mathfrak{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S}[\widehat{\mathfrak{R}}_S(\mathcal{H})].$$

3 General sample-dependent priors

In this section, we present general PAC-Bayes bounds for sample-dependent priors. Our bounds are in terms of certain covering numbers for sample-dependent priors, defined as follows.

Definition 1. *Let $\rho : \Delta(\mathcal{H}) \times \Delta(\mathcal{H}) \to \mathbb{R}$ be a divergence function taking values in non-negative reals. Let m, n be positive integers. For a given sample U of size $m+n$, and a scale $\alpha \geq 0$, $C \subseteq \Delta(\mathcal{H})$ is called a cover for U at scale α under ρ if for all subsamples $S \subseteq U$ with $|S| = m$, there exists a distribution $P \in C$ such that $\rho(P_S, P) \leq \alpha$. Define the covering number $\mathcal{N}(\alpha, m, U, \rho)$ to be the size of the smallest such cover. Define $\mathcal{N}(\alpha, m, n, \rho) = \max_{U\in\mathcal{Z}^{m+n}}[\mathcal{N}(\alpha, m, U, \rho)]$. When $m = n$, we use the notation $\mathcal{N}(\alpha, m, \rho)$ to mean $\mathcal{N}(\alpha, m, m, \rho)$.*

We now provide our general PAC-Bayes bounds with sample-dependent priors. These bounds are based on D_∞ and ℓ_1 covering numbers, and the most general forms depend on two sample size parameters, m and n. To keep the presentation clean, here we present the learning bounds using the $O(\cdot)$ notation for the special case $m = n$. The detailed bound without this assumption and proof can be found in Appendix A.1.

Theorem 1. *Let $P_S \in \Delta(\mathcal{H})$ be a prior over $\mathcal{H}$ determined by the choice of $S \in \mathcal{Z}^m$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$ and all $\alpha \geq 0$: if $D = \max\{\mathsf{D}(Q \parallel P_S), 2\}$,*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + O\left(\sqrt{\left(D + \alpha + \log\mathcal{N}(\alpha, m, \mathsf{D}_\infty) + \log(\tfrac{D}{\delta})\right)\left(\tfrac{1}{m}\right)}\right). \qquad (1)$$

[2] Extension to continuous domains is straightforward using standard measure-theoretic formulations.

[3] Technically, this is the Rademacher complexity of the class $\mathcal{G} = \{z \mapsto L(h,z)\colon h \in \mathcal{H}\}$, however we define it in this way by absorbing the loss function for clarity of notation.

Similarly, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$ and all $\alpha \geq 0$: if $D = \max\{\mathsf{D}(Q \parallel P_S), 2\}$,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + O\left((\sqrt{D}+\alpha)\mathfrak{R}_{2m}(\mathcal{H}) + \sqrt{\left(\log\mathcal{N}(\alpha, m, \ell_1) + \log\left(\tfrac{D}{\delta}\right)\right)\left(\tfrac{1}{m}\right)}\right). \quad (2)$$

In order to establish the above theorem, we build upon the recently proposed framework of Foster et al. [2019] that provides generalization bounds for sample-dependent hypothesis sets. In particular, [Foster et al., 2019] consider a family of hypothesis sets $\mathscr{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ and show that generalization bounds for this family can be obtained via a notion of transductive Rademacher complexity. Formally, for a sample set U of size $(m+n)$, define $\overline{\mathcal{H}}_{U,m} = \bigcup_{\substack{S\subset U,\\ |S|=m}} \mathcal{H}_S$. Then the transductive Rademacher complexity $\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{H})$ is defined for any $U = (z_1, \ldots, z_{m+n}) \in \mathcal{Z}^{m+n}$ as follows: if $\boldsymbol{\sigma}$ is a vector of $(m+n)$ independent random variables taking value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$ and value $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$, then

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{H}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\overline{\mathcal{H}}_{U,m}} \frac{1}{m+n}\sum_{i=1}^{m+n}\sigma_i L(h, z_i)\right]. \quad (3)$$

Foster et al. [2019] gave the following generalization bound for $\mathscr{H}$ in terms of the maximum transductive Rademacher complexity, over all sample sets U of size $m+n$.

Theorem 2 (Theorem 1 in [Foster et al., 2019])**.** *Let $\mathscr{H} = (\mathcal{H}_S)_{S\in\mathcal{Z}^m}$ be a family of data-dependent hypothesis sets. 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 $h \in \mathcal{H}_S$:*

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}[L(h,z)] \leq \mathop{\mathbb{E}}_{z\sim S}[L(h,z)] + 2\max_{U\in\mathcal{Z}^{m+n}}\left[\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{H})\right] + 3\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)\log\left(\tfrac{2}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn},$$

To apply the above result in our setting, recall that we interpret any distribution $Q \in \Delta(\mathcal{H})$ as a randomized hypothesis whose loss on any given point $z \in \mathcal{Z}$ is $\langle Q, L_z\rangle$. For a given $\mu > 0$, we then apply Theorem 2 to the following family of sample-dependent (randomized) hypothesis sets $\mathscr{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$ as

$$\mathcal{Q}_{S,\mu} = \left\{Q \in \Delta(\mathcal{H}) \colon \mathsf{D}(Q \parallel P_S) \leq \mu\right\}. \quad (4)$$

To apply Theorem 2, we need to bound the transductive Rademacher complexity of this family, $\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{Q\in\overline{\mathcal{Q}}_{U,m,\mu}} \frac{1}{m+n}\sum_{i=1}^{m+n}\sigma_i\langle Q, L_{z_i}\rangle\right]$, where $\boldsymbol{\sigma}$ is a vector of random variables as defined just before (3). We establish such upper bounds (Lemmas 1 and 2 below) via the covering numbers from Definition 1, which leads to a bound similar to that of Theorem 1 in terms of μ. The following lemma, proved in Appendix A.2, bounds the transductive Rademacher complexity using D_∞-covering numbers:

Lemma 1. *For any $\alpha \geq 0$, we have*

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) \leq \sqrt{\left(\frac{\mu + \alpha + \log\mathcal{N}(\alpha, m, U, \mathsf{D}_\infty)}{2}\right)\left(\frac{1}{m}+\frac{1}{n}\right)^3 mn}.$$

We now give a bound (proved in Appendix A.3) in terms of ℓ_1-covering numbers using a bit of notation. Let m, n be two positive integers, and let $U = (z_1, z_2, \ldots, z_{m+n}) \in \mathcal{Z}^{m+n}$ be a sample set. Then we define a notion of Rademacher complexity $\widetilde{\mathfrak{R}}_{U,m}(\mathcal{H})$ as follows: if $\boldsymbol{\sigma}$ is a vector of $(m+n)$ independent random variables taking value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$ and value $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$, then

$$\widetilde{\mathfrak{R}}_{U,m}(\mathcal{H}) := \frac{1}{m+n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{m+n}\sigma_i L(h, z_i)\right|\right]. \quad (5)$$

Lemma 2. *For any $\alpha \geq 0$, we have*

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) \leq (\sqrt{2\mu} + \alpha)\widetilde{\mathfrak{R}}_{U,m}(\mathcal{H}) + \sqrt{\frac{\log\mathcal{N}(\alpha, m, U, \ell_1)}{2}\left(\frac{1}{m}+\frac{1}{n}\right)^3 mn}.$$

We obtain a uniform bound (proved in Appendix A.4) over all values of μ by using a standard doubling argument:

Lemma 3. *Suppose the following bound holds with probability at least $1-\delta$ over the choice of S: for all $Q \in \mathcal{Q}_{S,\mu}$,*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(\mu) + g(\delta),$$

where f is an increasing function of μ and g is a decreasing function of δ. Then, the following holds with probability at least $1-\delta$ for all $Q \in \Delta(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(2\max\{D(Q \parallel P_S), 2\}) + g\left(\tfrac{\delta}{\max\{D(Q\parallel P_S),2\}}\right).$$

4 Stable sample-dependent priors

We now provide PAC-Bayes bounds for the setting where the sample-dependent prior P_S satisfies a sensitivity assumption; i.e., for two samples S and S' of size m that differ in only a single data point, the priors P_S and $P_{S'}$ are close in some divergence defined on the pair of distributions over hypotheses. The precise definition of sensitivity follows.

Definition 2. *Let $\rho : \Delta(\mathcal{H}) \times \Delta(\mathcal{H}) \to \mathbb{R}$ be a divergence function taking values in non-negative reals. The family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ is said to have sensitivity ϵ w.r.t. ρ if for all samples $S, S' \in \mathcal{Z}^m$ differing in a single data point, $\rho(P_S, P_{S'}) \leq \epsilon$.*

The specific divergences we will consider are the infinite Rényi divergence D_∞ and the ℓ_1 distance. The bounds of Theorem 1 imply the following learning bounds under assumptions of D_∞ and ℓ_1 sensitivity:

Corollary 1. *Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q \parallel P_S), 2\}$,*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + O\left(\sqrt{\frac{D}{m} + \epsilon + \log\left(\frac{D}{\delta}\right)\frac{1}{m}}\right). \tag{6}$$

Suppose instead that the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has ℓ_1 sensitivity ϵ. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q \parallel P_S), 2\}$,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + O\left((\sqrt{D} + \epsilon m)\mathfrak{R}_{2m}(\mathcal{H}) + \sqrt{\log\left(\frac{D}{\delta}\right)\frac{1}{m}}\right). \tag{7}$$

Proof. Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Let $U \in \mathcal{Z}^{2m}$, and let S be an arbitrary subset of U of size m. It is then easy to see that $\{P_S\}$ is a cover for U at scale ϵm under D_∞, and (6) follows by immediately by applying (1) from Theorem 1. The bound for the ℓ_1 case is exactly analogous. □

We can obtain more nuanced bounds than the ones in Corollary 1 by exploiting the sensitivity of the priors via the concept of *hypothesis set stability* from [Foster et al., 2019]. In order to obtain PAC-Bayesian learning bounds using this framework, we first define several quantities in terms of a *family* of sample-dependent sets of distributions $\mathcal{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}, \mathcal{Q}_S \subseteq \Delta(\mathcal{H})$. This construction is analogous to (4); the only difference is that we have temporarily dropped μ for now, to emphasize that the following definitions are applicable to a general sample-dependent family. Specifically, we will assume that the family $\mathcal{Q}_m$ satisfies a certain *stability* property defined below:

Definition 3. *We say that $\mathcal{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ is β-uniformly stable for some $\beta \geq 0$ if $\forall S, S' \in \mathcal{Z}^m$ differing by exactly one point, for every $Q \in \mathcal{Q}_S$, there exists a $Q' \in \mathcal{Q}_{S'}$ such that $\|Q - Q'\|_{\mathrm{TV}} \leq \beta$.*

To describe our learning bounds, we need a bit of notation from [Foster et al., 2019]. We denote by $\boldsymbol{\sigma} \in \{-1, 1\}^m$ a vector of independent Rademacher variables. For two samples $S, T \in \mathcal{Z}^m$, we denote by $S_T^{\boldsymbol{\sigma}}$ the sample obtained from S by replacing the i-th element of S by the corresponding element of T for all i such that $\boldsymbol{\sigma}_i = -1$. Finally, we define the following notion of Rademacher complexity for a family of sample-dependent sets of distributions $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$:

$$\mathfrak{R}^{\diamond}_m(\mathscr{Q}_m) = \frac{1}{m}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{Q\in\mathcal{Q}_{S_T^{\boldsymbol{\sigma}}}}\sum_{i=1}^{m}\sigma_i\langle Q, L_{z_i}\rangle\right]. \tag{8}$$

With these definitions, we have the following learning bound. This is analogous to a bound from [Foster et al., 2019] and proven using similar techniques, but it is tighter because of the Rademacher complexity term multiplying the stability term in the bound. The proof appears in Appendix B.1.

Theorem 3. *Suppose $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ is β-uniformly stable. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following holds for all $Q \in \mathcal{Q}_S$:*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}L(h,z_i)\right] + O\left(\mathfrak{R}^{\diamond}_m(\mathscr{Q}_m) + (1+\beta\mathfrak{R}_m(\mathcal{H})m)\sqrt{\tfrac{1}{m}\log(\tfrac{1}{\delta})} + \beta\log(\tfrac{m}{\delta})\right).$$

The proof of the theorem above is along the lines of the proof of Theorem 2 in [Foster et al., 2019]. Specifically, for two samples $S, S' \in \mathcal{Z}^m$, define the function $\Psi(S,S')$ as follows:

$$\Psi(S,S') = \sup_{Q\in\mathcal{Q}_S}\langle Q,\ell\rangle - \langle Q,\hat{\ell}_{S'}\rangle,$$

where $\ell, \hat{\ell}_{S'} \in \mathbb{R}^{\mathcal{H}}$ defined as $\ell(h) = \mathbb{E}_{z\sim\mathcal{D}}[L(h,z)]$ and $\hat{\ell}_{S'}(h) = \mathbb{E}_{z\sim S'}[L(h,z)]$, where $z \sim S'$ indicates uniform sampling from S'. The proof of the bound consists of applying McDiarmid's inequality to $\Psi(S,S)$. To do this, we need to analyze the sensitivity of this function, i.e., compute a bound on $|\Psi(S,S) - \Psi(S',S')|$ where S' is a sample differing from S in exactly one point. We provide a refined bound on the sensitivity using the fact that the map $Q \mapsto \langle Q,\ell\rangle$ is linear.

This bound leads to refined PAC-Bayesian bounds via the following template. (1) We define an appropriate sample-dependent family of distributions $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$. Typically $\mathcal{Q}_S$ will be set to $\mathcal{Q}_{S,\mu} := \{Q \in \Delta(\mathcal{H})\colon \mathsf{D}(Q \parallel P_S) \leq \mu\}$ for some parameter μ, as in (4). (2) Then, assuming that the priors P_S are chosen to have ϵ sensitivity, we show that $\mathscr{Q}_m$ is β-stable for some small β depending on ϵ. (3) We also derive bounds on the Rademacher complexity $\mathfrak{R}^{\diamond}_m(\mathscr{Q}_m)$ in terms of μ and ϵ. Using these bounds in Theorem 3 gives us a learning bound that depends on μ. (4) We obtain a uniform bound over all possible values of μ via a standard union bound argument (Lemma 3).

We now proceed to instantiate this template for D_∞-sensitive priors. This leads to a better bound than the one in Corollary 1, with an extra assumption. The following theorem (precise bound spelled out in Appendix B.2) shows how to apply Theorem 3 to obtain refined PAC-Bayes bounds.

Theorem 4. *Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Also assume that for some $\eta > 0$, we have $P_S(h) \geq \eta$ for all $h \in \mathcal{H}$, and all $S \in \mathcal{Z}^m$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q \parallel P_S), 2\}$,*

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}L(h,z_i)\right] + O\Bigg(&\sqrt{\tfrac{D}{m} + \epsilon^2 + \epsilon\sqrt{\tfrac{\log(m/\eta)}{m}}}\\ &+ (1+\epsilon\mathfrak{R}_m(\mathcal{H})m)\sqrt{\tfrac{1}{m}\log(\tfrac{D}{\delta})} + \epsilon\log(\tfrac{mD}{\delta})\Bigg).\end{aligned}$$

Proof. Define a sample-dependent family of distributions $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ where $\mathcal{Q}_S = \{Q\colon \mathsf{D}_\infty(Q \parallel P_S) \leq \mu\}$ for some parameter μ. We now apply the bound in Theorem 3, using the bound on the Rademacher complexity from Lemma 4, and the bound $\beta \leq 2\epsilon$ from Lemma 6. Finally, a uniform bound over all values of μ follows by an application of Lemma 3. □

At first glance, the statement of Theorem 4 may look qualitatively similar to Theorem 4.2 of [Dziugaite and Roy, 2018a]. Indeed, both bounds make use of the same tools developed in the differential privacy literature (specifically, [Dwork et al., 2015]), but the two analyses are completely different, since the two settings are fundamentally different, as stated in the Introduction. This also makes the two results incomparable.

The following is the key technical lemma needed in the proof of Theorem 4 to bound the Rademacher complexity term from Theorem 3 which employs the tools from [Dwork et al., 2015].

Lemma 4. *If* $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ *for all* $S, S' \in \mathcal{Z}^m$ *differing by exactly one point, and for some* $\eta > 0$*, we have* $P_S(h) \ge \eta$ *for all* $h \in \mathcal{H}$*, and all* $S \in \mathcal{Z}^m$*. Then*

$$\mathfrak{R}^\diamond_m(\mathcal{Q}_{m,\mu}) \le \sqrt{\frac{2\mu}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2m/\eta)}{m}}} + \sqrt{\frac{2}{m}} + \frac{\eta}{m}.$$

Proof. Fix the value of $\boldsymbol{\sigma}$. Consider the distribution $\mathcal{P}_{\boldsymbol{\sigma}}$ on (S, T, h) induced by sampling $S, T \sim \mathcal{D}^m$, and then conditioned on the values of S and T, sampling $h \sim P_{S^{\boldsymbol{\sigma}}_T}$. Consider the marginal distribution of h induced by $\mathcal{P}_{\boldsymbol{\sigma}}$. The probability assigned to h in this marginal distribution is

$$\mathop{\mathbb{E}}_{S,T\sim\mathcal{D}^m}[P_{S^{\boldsymbol{\sigma}}_T}(h)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[P_S(h)],$$

by symmetry, so this marginal distribution is independent of $\boldsymbol{\sigma}$. Call the marginal distribution $\mathcal{P}$.

Since sampling $h \sim P_{S^{\boldsymbol{\sigma}}_T}$ is ϵ-differentially private, by Theorem 20 in [Dwork et al., 2015], for any $\gamma > 0$, we have[4] $\mathsf{D}^\gamma_\infty(\mathcal{P}_{\boldsymbol{\sigma}} \| \mathcal{D}^{2m} \otimes \mathcal{P}) \le \kappa := \epsilon^2 m + \epsilon\sqrt{m \log(2/\gamma)}$. Thus, by Lemma 3.17, part 1, in [Dwork and Roth, 2014], there exists a distribution $\mathcal{P}'_{\boldsymbol{\sigma}}$ on (S, T, h) such that $\|\mathcal{P}_{\boldsymbol{\sigma}} - \mathcal{P}'_{\boldsymbol{\sigma}}\|_{\mathrm{TV}} \le \gamma$ and $\mathsf{D}_\infty(\mathcal{P}'_{\boldsymbol{\sigma}} \| \mathcal{D}^{2m} \otimes \mathcal{P}) \le \kappa$.

Now, fix the values of S, T, and let $\mathcal{P}'_{\boldsymbol{\sigma}|S,T}$ be the distribution of h conditioned on these values of S and T. Assume $m \ge 2$, since the bound in the lemma statement holds trivially if $m = 1$. Set $\gamma = \frac{\eta}{m}$. Since $\|\mathcal{P}_{\boldsymbol{\sigma}} - \mathcal{P}'_{\boldsymbol{\sigma}}\|_{\mathrm{TV}} \le \gamma$, we have $\|\mathcal{P}'_{\boldsymbol{\sigma}|S,T} - P_{S^{\boldsymbol{\sigma}}_T}\|_{\mathrm{TV}} \le \gamma$, and so $\mathcal{P}'_{\boldsymbol{\sigma}|S,T}(h) \ge \eta - \gamma$ for all $h \in \mathcal{H}$. Using this fact in the analysis in the proof of Lemma 5, we conclude that for every Q such that $\mathsf{D}(Q \| P_{S^{\boldsymbol{\sigma}}_T}) \le \mu$, there exists Q' such that $\mathsf{D}(Q' \| \mathcal{P}'_{\boldsymbol{\sigma}|S,T}) \le \mu$ and $\|Q - Q'\|_{\mathrm{TV}} \le \sqrt{\frac{\gamma}{(\eta-\gamma)}}$. Thus,

$$\sup_{\mathsf{D}(Q\|P_{S^{\boldsymbol{\sigma}}_T})\le\mu} \langle Q, u_{\boldsymbol{\sigma}}\rangle - \sup_{\mathsf{D}(Q'\|\mathcal{P}'_{\boldsymbol{\sigma}|S,T})\le\mu} \langle Q, u_{\boldsymbol{\sigma}}\rangle \le \sqrt{\frac{\gamma}{(\eta-\gamma)}} \cdot \sup_h |u_{\boldsymbol{\sigma}}(h)| \le \sqrt{\frac{2\gamma}{\eta}} \cdot m,$$

since $\gamma \le \eta/2$. Thus, if $\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}$ is the marginal distribution of $\mathcal{P}'_{\boldsymbol{\sigma}}$ on (S, T), we have

$$\begin{aligned}
\mathfrak{R}^\diamond_m(\mathcal{Q}_{m,\mu}) = \frac{1}{m}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|P_{S^{\boldsymbol{\sigma}}_T})\le\mu} \langle Q, u_{\boldsymbol{\sigma}}\rangle\right] &\le \frac{1}{m}\mathop{\mathbb{E}}_{S,T,\boldsymbol{\sigma}}\left[\sup_{\mathsf{D}(Q\|\mathcal{P}'_{\boldsymbol{\sigma}|S,T})\le\mu} \langle Q, u_{\boldsymbol{\sigma}}\rangle\right] + \sqrt{\frac{2\gamma}{\eta}} \\
&\le \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}}\left[\sup_{\mathsf{D}(Q\|\mathcal{P}'_{\boldsymbol{\sigma}|S,T})\le\mu} \langle Q, u_{\boldsymbol{\sigma}}\rangle\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma,
\end{aligned}$$

where the last inequality follows because $\|\mathcal{D}^{2m} - \mathcal{P}'_{\boldsymbol{\sigma},(S,T)}\|_{\mathrm{TV}} \le \|\mathcal{P}_{\boldsymbol{\sigma}} - \mathcal{P}'_{\boldsymbol{\sigma}}\|_{\mathrm{TV}} \le \gamma$. Now define $\Psi_{\boldsymbol{\sigma}|S,T}(Q)$ by $\Psi_{\boldsymbol{\sigma}|S,T}(Q) = \mathsf{D}(Q \| \mathcal{P}'_{\boldsymbol{\sigma}|S,T})$ if $\mathsf{D}(Q \parallel \mathcal{P}'_{\boldsymbol{\sigma}|S,T}) \le \mu$ and $+\infty$ otherwise. The conjugate function $\Psi^*_{\boldsymbol{\sigma}|S,T}(u) = \log\left(\mathbb{E}_{h\in\mathcal{P}'_{\boldsymbol{\sigma}|S,T}}[e^{u(h)}]\right)$, for all $u \in \mathbb{R}^{\mathcal{H}}$ [Mohri et al., 2018, Lemma B.37]. Continuing the bound on $\mathfrak{R}^\diamond_m(\mathcal{Q}_{m,\mu})$ above, for any $t > 0$,

$$\begin{aligned}
\mathfrak{R}^\diamond_m(\mathcal{Q}_{m,\mu}) &\le \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}}\left[\sup_{\mathsf{D}(Q\|\mathcal{P}'_{\boldsymbol{\sigma}|S,T})\le\mu} \langle Q, tu_{\boldsymbol{\sigma}}\rangle\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \\
&\le \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}}\left[\sup_{\Psi_{\boldsymbol{\sigma}|S,T}(Q)\le\mu} \Psi_{\boldsymbol{\sigma}|S,T}(Q) + \Psi^*_{\boldsymbol{\sigma}|S,T}(tu_{\boldsymbol{\sigma}})\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \quad \text{(Fenchel inequality)} \\
&\le \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}}\left[\Psi^*_{\boldsymbol{\sigma}|S,T}(tu_{\boldsymbol{\sigma}})\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \\
&= \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}}\left[\log\left(\mathop{\mathbb{E}}_{h\sim\mathcal{P}'_{\boldsymbol{\sigma}|S,T}}\left[e^{tu_{\boldsymbol{\sigma}}(h)}\right]\right)\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \quad \text{(definition of } \Psi^*) \\
&\le \frac{\mu}{mt} + \frac{1}{mt}\left[\log\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{P}'_{\boldsymbol{\sigma},(S,T)}} \mathop{\mathbb{E}}_{h\sim\mathcal{P}'_{\boldsymbol{\sigma}|S,T}}\left[e^{tu_{\boldsymbol{\sigma}}(h)}\right]\right)\right] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \quad \text{(Jensen's inequality)}
\end{aligned}$$

[4] Theorem 20 in [Dwork et al., 2015] is stated in terms of approximate max-information; here we state the equivalent bound in terms of approximate infinite-Rényi divergence: see Definition 10 in [Dwork et al., 2015].

$$
\begin{aligned}
&= \frac{\mu}{mt} + \frac{1}{mt}\Big[\log\Big(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}\Big[e^{tu_{\boldsymbol{\sigma}}(h)}\Big]\Big)\Big] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \\
&\le \frac{\mu}{mt} + \frac{1}{mt}\Big[\log\Big(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathcal{P}}\Big[e^{\kappa}e^{tu_{\boldsymbol{\sigma}}(h)}\Big]\Big)\Big] + \sqrt{\frac{2\gamma}{\eta}} + \gamma \quad (\text{since } \mathsf{D}_\infty(\mathcal{P}'_{\boldsymbol{\sigma}} \| \mathcal{D}^{2m}\otimes\mathcal{P}) \;\le\; \kappa) \\
&= \frac{\mu}{mt} + \frac{1}{mt}\Big[\log\Big(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathcal{P}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[e^{tu_{\boldsymbol{\sigma}}(h)}\Big]\Big)\Big] + \frac{\kappa}{mt} + \sqrt{\frac{2\gamma}{\eta}} + \gamma \\
&\le \frac{\mu}{mt} + \frac{1}{mt}\Big[\log\Big(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathcal{P}} e^{\frac{t^2 m}{2}}\Big)\Big] + \frac{\kappa}{mt} + \sqrt{\frac{2\gamma}{\eta}} + \gamma \qquad (\text{Hoeffding's lemma}) \\
&= \frac{\mu}{mt} + \frac{t}{2} + \frac{\epsilon^2}{t} + \frac{\epsilon}{t}\sqrt{\frac{\log(2/\gamma)}{m}} + \sqrt{\frac{2\gamma}{\eta}} + \gamma.
\end{aligned}
$$

Now we choose $\gamma = \frac{\eta}{m}$, and $t = \sqrt{\frac{2\mu}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2m/\eta)}{m}}}$, to obtain the claimed bound. □

The requirement of the minimum probability $\eta > 0$ in Theorem 4 limits the applicability of the theorem to finite hypothesis sets $\mathcal{H}$. Via a similar proof technique, we can also derive the following PAC-Bayes bound in the case the priors have D_∞ stability without any requirement of a minimum probability. The precise bound and the proof appear in Appendix B.3.

Theorem 5. *Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q \,\|\, P_S), 2\}$,*

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right] + O\Bigg(&\sqrt{\tfrac{D}{m} + \epsilon^2 + \tfrac{\epsilon}{\sqrt{m}}} + \epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3} + \epsilon^{4/5} \\
&+ \Big(\epsilon\mathfrak{R}_m(\mathcal{H}) + \epsilon\sqrt{\tfrac{\log(m^{1.5}D/\delta)}{m}} + \tfrac{1}{m}\Big)\sqrt{m\log\left(\tfrac{D}{\delta}\right)}\Bigg).
\end{aligned}
$$

The above bound can also be extended to the case where the priors define an (ϵ,δ)-differentially private mechanism for some $\delta > 0$, instead of a pure ϵ-differentially private mechanism, as required by Theorem 5. The precise bound in the theorem below and proof can be found in Appendix B.4.

Theorem 6. *Assume that $\epsilon \ge 0$ and $\delta \in [0, \frac{Ce^{-16m}\epsilon}{m^2}]$ for some constant C. Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ satisfy the property that $\mathsf{D}^{\delta}_\infty(P_S \| P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing in exactly one point. Then, for any $\nu > 0$, with probability at least $1-\nu$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q \,\|\, P_S), 2\}$,*

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right] + O\Bigg(&\sqrt{\tfrac{D}{m} + \epsilon^2 + \tfrac{\epsilon}{\sqrt{m}}} + \epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3} + \epsilon^{4/5} + \tfrac{\sqrt{\delta}}{\epsilon^{3/2}} \\
&+ \Big(\epsilon\mathfrak{R}_m(\mathcal{H}) + \epsilon\sqrt{\tfrac{\log(m^{1.5}D/\nu)}{m}} + \tfrac{1}{m}\Big)\sqrt{m\log\left(\tfrac{D}{\nu}\right)}\Bigg).
\end{aligned}
$$

Similarly, one can consider studying other variants of the above theorems based on different sensitivity assumptions via the template above. A crucial component in implementing the template is effective control of the stability term β in various settings. Below, we give a few lemmas which provide such control, some of which are used in the proof of Theorem 4. Proofs of these lemmas can be found in Appendices B.5, B.6 and B.7 respectively.

Lemma 5. *Suppose $\|P_S - P_{S'}\|_1 \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \ge 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} := \{Q\colon \mathsf{D}(Q \,\|\, P_S) \le \mu\}$, and the corresponding family to be $\mathcal{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$. Then $\mathcal{Q}_{m,\mu}$ is β-stable for $\beta = \min\left\{\frac{\epsilon d_\infty}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon d_\infty}{2}}\right\}$, where $d_\infty \coloneqq \sup_{S,S',Q\in\mathcal{Q}_{S,\mu}} \left\|\frac{Q}{P_{S'}}\right\|_\infty$.*

Lemma 6. *Suppose $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \ge 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} \coloneqq \{Q\colon \mathsf{D}(Q \parallel P_S) \le \mu\}$, and the corresponding family to be $\mathcal{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$. Then $\mathcal{Q}_{m,\mu}$ is β-stable for $\beta = \min\left\{2\epsilon, \frac{\epsilon}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon}{2}}\right\}$.*

Lemma 7. *Suppose $\|P_S - P_{S'}\|_1 \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \ge 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} \coloneqq \{Q\colon \|Q - P_S\|_1 \le \mu\}$, and the corresponding family to be $\mathcal{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$. Then $\mathcal{Q}_{m,\mu}$ is β-stable for $\beta = \frac{\epsilon}{2}$.*

We conclude with a brief discussion of some important considerations when applying these bounds.

Choice of ϵ. Depending on the choice of the function $S \mapsto P_S$ (for $S \in \mathcal{Z}^m$) and divergence function ρ, the family of priors $(P_S)_{S\in\mathcal{Z}^m}$ can be made to have varying sensitivity ϵ. A larger value of ϵ provides greater freedom in selecting a sample-dependent prior, thus making it possible to cleverly choose a prior P_S *closer* to the posterior Q in order to decrease $\mathsf{D}(Q \parallel P_S)$. However, the price of such flexibility in our bounds is captured via additive terms that increase with ϵ. In Corollary 1, we see that to derive $O\left(\frac{1}{\sqrt{m}}\right)$ rates, one must choose $\epsilon = O\left(\frac{1}{m}\right)$. In our refined bounds based on hypothesis set stability, a weaker $\epsilon = O\left(\frac{1}{\sqrt{m}}\right)$ suffices to obtain such $O\left(\frac{1}{\sqrt{m}}\right)$ rates, assuming that the Rademacher complexity of the base hypothesis class $\mathcal{H}$, $\mathfrak{R}_m(\mathcal{H})$, scales as $O\left(\frac{1}{\sqrt{m}}\right)$.

Choice of μ. Our bounds are intentionally presented as standard PAC-Bayes bounds - i.e., for all $Q \in \Delta(\mathcal{H})$ - eliminating the need for an explicit choice of μ to control the size of the sample-dependent sets of priors. However, if one were to fix μ (possibly as a function of m) a priori and then only consider those posteriors contained in the sample-dependent set $\mathcal{Q}_{S,\mu}$, then the application of Lemma 3 would be unnecessary.

A general application recipe. A general strategy for obtaining an ϵ-sensitive family of priors is to leverage any existing algorithm known to generate "parameter-stable" hypotheses (i.e., those for which the final parameters are close in some metric upon swapping a single element of the training set, under some parameterization of the hypothesis class). A notable example is the application of gradient descent with a limited number of parameter updates, under certain conditions discussed in [Feldman and Vondrak, 2018, 2019, Hardt et al., 2016]. Let $w_S = \mathcal{A}(S)$ denote the parameters found by running such a parameter-stable algorithm $\mathcal{A}$ on sample S. One natural prior P_S is then a Gaussian distribution centered at w_S. Concretely, in a neural network setting, one could imagine running gradient descent on a sample S for a limited number of iterations to obtain a parameter-stable prior P_S and then continuing training on S to generate a posterior Q. The art is in choosing an appropriate family of priors sufficiently close to the posteriors for the application in question; in the above neural network setting, this involves choosing the number of iterations to use in generating P_S, which highlights the tradeoff between ϵ and $\mathsf{D}(Q \parallel P_S)$.

5 Conclusion

We presented a general framework for deriving PAC-Bayesian learning bounds with sample-dependent priors, by leveraging the recently introduced notion of hypothesis set stability [Foster et al., 2019]. Our bounds include covering number-based bounds, as well as bounds specifically tailored to priors satisfying a sensitivity condition, upon swapping an element of the sample.

This provides a broad framework for deriving PAC-Bayesian bounds that takes advantage of the full training sample when generating a prior. Much of our results can be further extended to the use of arbitrary Bregman divergences, instead of the specific (unnormalized) relative entropy. In particular, our Rademacher complexity analysis can be used similarly to derive upper bounds in that case.

A by-product of our analysis is a finer learning guarantee for sample-dependent hypothesis sets without any specific assumption about the closeness of these sample-dependent sets. This can provide a powerful tool for the analysis of broad collections of learning algorithms, or the design of new algorithms. While we leveraged these guarantees here, in particular by considering a closeness based on hypothesis set stability implied by the sensitivity of the priors, an important research direction is that of exploring alternative notions of closeness that can be significant for the theory of sample-dependent learning guarantees.

Broader Impact

Due to the theoretical nature of this paper, we currently cannot foresee any short-to-medium-term negative societal impact. In general, we believe that the *short*-term societal impact is extremely limited. However, we hope that, in the medium-to-long term, such bounds - or other theoretical work that follows from such bounds - will play a role in furthering our understanding of the differences in generalization performance among various algorithms. We believe that such understanding is very important when deploying models in real-world settings and when attempting to design new algorithms that generalize even better. PAC-Bayes bounds, in particular, have shown some promise in explaining the generalization performance of neural networks, which are widely used in practice. Thus, beyond a general claim about various machine learning algorithms, we think it is possible that our bounds or those inspired by them can contribute to the community's understanding of (and expectations for) neural networks. Due to the widespread use of neural networks, such improved understanding can have a significant positive impact.

References

Amiran Ambroladze, Emilio Parrado-hernández, and John S. Shawe-taylor. Tighter PAC-Bayes bounds. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 9–16, 2007.

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. *arXiv preprint arXiv:1910.07833*, 2019.

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

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

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.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. On the role of data in pac-bayes bounds. *CoRR*, abs/2006.10929, 2020.

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.

Dylan J. Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Proceedings of NeurIPS 2019*, pages 6726–6736, 2019.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *CoRR*, abs/2004.12983, 2020.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of ICML*, pages 1225–1234, 2016.

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.

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

John Langford and Rich Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, pages 809–816, 2002.

John Langford and Matthias Seeger. Bounds for averaging classifiers. 2001.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

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

Ben London. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2931–2940. 2017.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999a.

David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999b.

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

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11611–11622, 2019.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *NeurIPS*, pages 11013–11023, 2019.

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.

Pantelimon G Popescu, Sever S Dragomir, Emil I Sluşanschi, and Octavian N Stănăşilă. Bounds for Kullback-Leibler divergence. *Electronic Journal of Differential Equations*, 2016.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–417, 2005.

Omar Rivasplata, Emilio Parrado-Hernandez, John S Shawe-Taylor, Shiliang Sun, and Csaba Szepesvari. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9214–9224. 2018.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Proceedings of NeurIPS*, 2020.

Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *FOCS*, pages 487–494, 2016.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

John Shawe-Taylor and Robert C Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

A Proofs of results in Section 3

A.1 Proof of Theorem 1

Here, we present the full proof of Theorem 1, with the precise bound spelled out. To present the theorem, recall the definition of $\tilde{\mathfrak{R}}_{U,m}(\mathcal{H})$ in (5): let m, n be two positive integers, and let $U = (z_1, z_2, \ldots, z_{m+n}) \in \mathcal{Z}^{m+n}$ be a sample set. Then we define a notion of Rademacher complexity $\tilde{\mathfrak{R}}_{U,m}(\mathcal{H})$ as follows: if $\boldsymbol{\sigma}$ is a vector of $(m+n)$ independent random variables taking value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$ and value $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$, then

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

Furthermore, define $\tilde{\mathfrak{R}}_{m,n} = \mathbb{E}_U[\tilde{\mathfrak{R}}_{U,m}(\mathcal{H})]$.

The bound of Theorem 1 as stated in Section 3 is for the special case $m = n$, and is stated in terms of the standard Rademacher complexity $\mathfrak{R}_{2m}(\mathcal{H})$. This follows from the following bound:

Lemma 8. *If $m = n$, then $\tilde{R}_{U,m}(\mathcal{H}) \le 4\mathfrak{R}_U(\mathcal{H})$.*

Proof. Since $m = n$, $\boldsymbol{\sigma}$ is a vector of $2m$ variables taking values in $\{-2, 2\}$ uniformly at random.

$$\begin{aligned}
\tilde{R}_{U,m}(\mathcal{H}) &= \frac{1}{2m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{2m}\sigma_i L(h,z_i)\right|\right] \\
&= \frac{1}{2m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H}\\ s\in\{-1,+1\}}} s\sum_{i=1}^{2m}\sigma_i L(h,z_i)\right] \\
&\le \frac{1}{2m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{2m}\sigma_i L(h,z_i)\right] + \frac{1}{2m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{2m}-\sigma_i L(h,z_i)\right] \\
&= 4\mathfrak{R}_U(\mathcal{H}).
\end{aligned}$$

□

Theorem 1. *Let $P_S \in \Delta(\mathcal{H})$ be a prior over $\mathcal{H}$ determined by the choice of $S \in \mathcal{Z}^m$, and let n be a positive integer. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$, if $D := \max\{\mathsf{D}(Q\|P_S), 2\}$,*

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] &+ \inf_{\alpha\ge 0}\sqrt{2\left(2D+\alpha+\log\mathcal{N}(\alpha,m,n,\mathsf{D}_\infty)\right)\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn} \\
&+ 3\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)\log\left(\tfrac{4D}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn\log\left(\tfrac{8eD}{\delta}\right)}.
\end{aligned} \tag{9}$$

Similarly, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] &+ \inf_{\alpha\ge 0} 2(2\sqrt{D}+\alpha)\tilde{\mathfrak{R}}_{m,n}(\mathcal{H}) + \sqrt{2\log(\mathcal{N}(\alpha,m,n,\ell_1))\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn} \\
&+ 3\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)\log\left(\tfrac{4D}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn\log\left(\tfrac{8eD}{\delta}\right)}.
\end{aligned} \tag{10}$$

Proof. Fix $\mu > 0$ and define the sample-dependent hypothesis set as

$$\mathcal{Q}_{S,\mu} = \left\{Q \in \Delta(\mathcal{H}) \colon \mathsf{D}(Q\|P_S) \le \mu\right\},$$

where $\Delta(\mathcal{H})$ is the family of all distributions defined over $\mathcal{H}$. We define the loss of $Q \in \Delta(\mathcal{H})$ over the labeled sample $z = (x,y) \in \mathcal{Z}$ as $\ell(Q,z) = \langle Q, L_z\rangle$. Thus, the expected loss of Q is

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(Q,z)] = \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)].$$

We also define the sample-indexed family of sample-dependent hypothesis sets $\mathscr{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$ and the U-restricted union of sample-dependent hypothesis sets $\overline{\mathcal{Q}}_{U,m,\mu} = \bigcup_{\substack{S\in\mathcal{Z}^m\\ S\subseteq U}} \mathcal{Q}_{S,\mu}$.

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

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + 2\max_{U\in\mathcal{Z}^{m+n}} \widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) + 3\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)\log\left(\tfrac{2}{\delta}\right)} + 2\sqrt{\left(\tfrac{1}{m}+\tfrac{1}{n}\right)^3 mn},$$

where $\mathfrak{R}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu})$ is defined for any $U = (z_1, \ldots, z_{m+n}) \in \mathcal{Z}^{m+n}$ as follows: if $\boldsymbol{\sigma}$ is a vector of $(m+n)$ independent random variables taking value $\frac{m+n}{n}$ with probability $\frac{n}{m+n}$ and value $-\frac{m+n}{m}$ with probability $\frac{m}{m+n}$, then

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{Q\in\overline{\mathcal{Q}}_{U,m,\mu}} \frac{1}{m+n}\sum_{i=1}^{m+n} \sigma_i\langle Q, L_{z_i}\rangle\right].$$

Via covering number arguments for D_∞ (Lemma 1) and ℓ_1 (Lemma 2) we derive bounds on $\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu})$. The bounds in the theorem then follow by applying Lemma 3. □

A.2 Proof of Lemma 1

Lemma 1. *For any $\alpha \geq 0$, we have*

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) \leq \sqrt{\left(\frac{\mu + \alpha + \log\mathcal{N}(\alpha, U, \mathsf{D}_\infty)}{2}\right)\left(\frac{1}{m}+\frac{1}{n}\right)^3 mn}.$$

Proof. Let C be a covering for U under D_∞ at scale α of size $\mathcal{N}(\alpha, U, \mathsf{D}_\infty)$. Define $\mathcal{G}_{U,m,\mu+\alpha}$ as

$$\mathcal{G}_{U,m,\mu+\alpha} := \{Q \in \Delta(\mathcal{H}) : \exists P \in C \text{ s.t. } \mathsf{D}(Q\|P) \leq \mu + \alpha\}.$$

Now, let $Q \in \overline{\mathcal{H}}_{U,m,\mu}$. Then there exists a some subset S of U of size m, such that $\mathsf{D}(Q\|P_S) \leq \mu$. Since C is a covering for U under D_∞ at scale α, there exists a distribution $P' \in C$ such that $\mathsf{D}_\infty(P\|P') \leq \alpha$. We have $\mathsf{D}(Q\|P') \leq \mathsf{D}(Q\|P) + \mathsf{D}_\infty(P\|P') \leq \mu + \alpha$. Thus, $Q \in \mathcal{G}_{U,m,\mu+\alpha}$. This implies that $\overline{\mathcal{H}}_{U,m,\mu} \subseteq \mathcal{G}_{U,m,\mu+\alpha}$.

In the following derivation, we will use the shorthand $u_{\boldsymbol{\sigma}}(h) = \sum_{i=1}^{m+n}\sigma_i L(h,z_i)$, so that $\sum_{i=1}^{m+n}\sigma_i\langle Q, L_{z_i}\rangle = \langle Q, u_{\boldsymbol{\sigma}}\rangle$. For any $P \in C$ and $Q \in \Delta(\mathcal{H})$, define $\Psi_P(Q)$ by $\Psi_S(Q) = \mathsf{D}(Q\|P_S)$ if $\mathsf{D}(Q\|P_S) \leq \mu + \alpha$ and $+\infty$ otherwise. It is known that the conjugate function Ψ^*_P of Ψ_P is given by $\Psi^*_P(u) = \log\left(\mathbb{E}_{h\in P}[e^{u(h)}]\right)$, for all $u \in \mathbb{R}^{\mathcal{H}}$ (see for example [Mohri et al., 2018, Lemma B.37]). We now upper bound the transductive Rademacher complexity term as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathscr{Q}_{m,\mu}) &= \frac{1}{m+n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{Q\in\overline{\mathcal{H}}_{U,m,\mu}} \langle Q, u_{\boldsymbol{\sigma}}\rangle\right] && \text{(definition of } u_{\boldsymbol{\sigma}}\text{)}\\
&\leq \frac{1}{m+n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{Q\in\mathcal{G}_{U,m,\mu+\alpha}} \langle Q, u_{\boldsymbol{\sigma}}\rangle\right] && (\overline{\mathcal{H}}_{U,m,\mu} \subseteq \mathcal{G}_{U,m,\mu+\alpha})\\
&= \frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{Q\in\mathcal{G}_{U,m,\mu+\alpha}} \langle Q, tu_{\boldsymbol{\sigma}}\rangle\right] && (t > 0)\\
&= \frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\ \sup_{Q\colon \mathsf{D}(Q\|P)\leq\mu+\alpha} \langle Q, tu_{\boldsymbol{\sigma}}\rangle\right] && \text{(iterated sup)}\\
&\leq \frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\ \sup_{Q\colon \mathsf{D}(Q\|P)\leq\mu+\alpha} [\Psi_P(Q) + \Psi^*_P(tu_{\boldsymbol{\sigma}})]\right] && \text{(Fenchel inequality)}
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\left[\mu+\alpha+\Psi_S^*(tu_{\boldsymbol{\sigma}})\right]\right] && \text{(definition of } \Psi_P(Q)\text{)}\\
&= \frac{\mu+\alpha}{(m+n)t}+\frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\Psi_P^*(tu_{\boldsymbol{\sigma}})\right] && \text{(distribute)}\\
&= \frac{\mu+\alpha}{(m+n)t}+\frac{1}{(m+n)t}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\log\left(\mathop{\mathbb{E}}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right] && \text{(definition of } \Psi_P^*\text{)}
\end{aligned}
$$

We now upper bound $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\log\left(\mathbb{E}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right]$ as follows:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{P\in C}\log\left(\mathop{\mathbb{E}}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\log\left(\sup_{P\in C}\mathop{\mathbb{E}}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right] && \text{(log is mon. incr.)}\\
&\le \log\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\sup_{P\in C}\mathop{\mathbb{E}}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right] && \text{(Jensen's inequality)}\\
&\le \log\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\sum_{P\in C}\mathop{\mathbb{E}}_{h\sim P}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right)\right] && \text{(nonnegative terms)}\\
&= \log\left[\sum_{P\in C}\mathop{\mathbb{E}}_{h\sim P}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[e^{tu_{\boldsymbol{\sigma}}(h)}]\right] && \text{(lin. of expectation; } h,\boldsymbol{\sigma} \text{ indep.)}\\
&= \log\left[\sum_{P\in C}\mathop{\mathbb{E}}_{h\sim P}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[e^{t\sum_{i=1}^{m+n}\sigma_i L(h,z_i^U)}\right]\right] && \text{(def. of } u_{\boldsymbol{\sigma}}(h)\text{)}\\
&= \log\left[\sum_{P\in C}\mathop{\mathbb{E}}_{h\sim P}\left[\prod_{i=1}^{m+n}\mathop{\mathbb{E}}_{\sigma_i}e^{t\sigma_i L(h,z_i^U)}\right]\right] && \text{(indep. entries of } \boldsymbol{\sigma}\text{)}\\
&\le \log\left[\sum_{P\in C}\mathop{\mathbb{E}}_{h\sim P}\left[e^{\frac{t^2(m+n)^5}{8(mn)^2}}\right]\right] && \text{(Hoeffding's lemma)}\\
&= \log\left[\sum_{P\in C}e^{\frac{t^2(m+n)^5}{8(mn)^2}}\right] && \text{(no dep. on } h\text{)}\\
&= \log\left[|C|\cdot e^{\frac{t^2(m+n)^5}{8(mn)^2}}\right] && \text{(all terms equal)}\\
&= \log|C| + \frac{t^2(m+n)^5}{8(mn)^2}.
\end{aligned}
$$

Plugging this back in, we get:

$$
\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) &\le \frac{\mu+\alpha}{(m+n)t}+\frac{1}{(m+n)t}\left[\log|C|+\frac{t^2(m+n)^5}{8(mn)^2}\right]\\
&= \frac{\mu+\alpha+\log|C|}{(m+n)t}+\frac{t(m+n)^4}{8(mn)^2}.
\end{aligned}
$$

We find that $t=\sqrt{\frac{8(mn)^2(\mu+\alpha+\log|C|)}{(m+n)^5}}$ minimizes the bound.
Plugging this optimal t back in, we obtain:

$$
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) \le \sqrt{\frac{(\mu+\alpha+\log|C|)(m+n)^3}{2(mn)^2}} = \sqrt{\left(\frac{\mu+\alpha+\log|C|}{2}\right)\left(\frac{1}{m}+\frac{1}{n}\right)^3 mn}.
$$

□

A.3 Proof of Lemma 2

Lemma 2. *For any* $\alpha \ge 0$, *we have*

$$
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) \le (\sqrt{2\mu}+\alpha)\tilde{\mathfrak{R}}_{U,m}(\mathcal{H}) + \sqrt{\frac{\log\mathcal{N}(\alpha,U,\ell_1)}{2}\left(\frac{1}{m}+\frac{1}{n}\right)^3 mn}.
$$

Proof. Let C be a covering for U under ℓ_1 at scale α of size $\mathcal{N}(\alpha, U, \ell_1)$. Let $\mathcal{G}_{U,m,\sqrt{2\mu}+\alpha}$ be the union of all the ℓ_1 balls of radius $\sqrt{2\mu}+\alpha$ around distributions in C, i.e.

$$\mathcal{G}_{U,m,\sqrt{2\mu}} = \{Q \in \Delta(\mathcal{H}) : \exists P \in C \text{ s.t. } \|Q - P\|_1 \le \sqrt{2\mu} + \alpha\}.$$

Now, let $Q \in \overline{\mathcal{H}}_{U,m,\mu}$. By Pinsker's inequality, for some subset S of U of size m, we have $\|Q - P_S\|_1 \le \sqrt{2\mu}$. Since C is a covering for U under ℓ_1 at scale α, there exists a distribution $P \in C$ such that $\|P_S - P\|_1 \le \alpha$. This implies that $\|Q - P\|_1 \le \sqrt{2\mu} + \alpha$, so $Q \in \mathcal{G}_{U,m,\sqrt{2\mu}+\alpha}$. Hence $\overline{\mathcal{H}}_{U,m,\mu} \subseteq \mathcal{G}_{U,m,\sqrt{2\mu}+\alpha}$. In the following derivation, we will use the shorthand $u_{\boldsymbol{\sigma}}(h) = \sum_{i=1}^{m+n} \sigma_i L(h, z_i)$, so that $\sum_{i=1}^{m+n} \sigma_i \langle Q, L_{z_i} \rangle = \langle Q, u_{\boldsymbol{\sigma}} \rangle$. We can now proceed the bound the Rademacher complexity as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) &= \frac{1}{m+n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{Q \in \overline{\mathcal{H}}_{U,m,\mu}} \langle Q, u_{\boldsymbol{\sigma}} \rangle \right] \\
&\le \frac{1}{m+n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{Q \in \mathcal{G}_{U,m,\sqrt{2\mu}+\alpha}} \langle Q, u_{\boldsymbol{\sigma}} \rangle \right] \\
&\le \frac{1}{m+n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{P \in C} \langle P, u_{\boldsymbol{\sigma}} \rangle \right] + (\sqrt{2\mu} + \alpha) \tilde{\mathfrak{R}}_{U,m}(\mathcal{H}).
\end{aligned}$$

The last inequality follows since for any $Q \in \mathcal{G}_{U,m,\sqrt{2\mu}+\alpha}$ there exists a distribution $P \in C$ such that $\|Q - P\|_1 \le \sqrt{2\mu} + \alpha$, and so we have

$$\mathop{\mathbb{E}}_{\sigma}[|\langle Q - P, u_{\boldsymbol{\sigma}} \rangle|] \le \mathop{\mathbb{E}}_{\sigma}[\|Q - P\|_1 \|u_{\boldsymbol{\sigma}}\|_\infty] \le (\sqrt{2\mu} + \alpha) \mathop{\mathbb{E}}_{\sigma}[\|u_{\boldsymbol{\sigma}}\|_\infty] = (\sqrt{2\mu} + \alpha)(m+n)\tilde{\mathfrak{R}}_{U,m}(\mathcal{H}).$$

Now, define $v : \Delta(\mathcal{H}) \to [0,1]^{m+n}$ as $v(P)_i = \mathbb{E}_{h \sim P}[L(h, z_i)]$. Note that $\langle P, u_{\boldsymbol{\sigma}} \rangle = \langle \boldsymbol{\sigma}, v(P) \rangle$, and so

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{P \in C} \langle P, u_{\boldsymbol{\sigma}} \rangle \right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{P \in C} \langle \boldsymbol{\sigma}, v(P) \rangle \right].$$

We can now bound $\mathbb{E}_{\boldsymbol{\sigma}} [\sup_{P \in C} \langle \boldsymbol{\sigma}, v(P) \rangle]$ by a version of Massart's lemma which applies to non-Rademacher (but still zero mean) random variables $\boldsymbol{\sigma}$, as follows: let $t > 0$ to be chosen momentarily. We have

$$\begin{aligned}
\exp\left(t \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{P \in C} \langle \boldsymbol{\sigma}, v(P) \rangle \right] \right) &\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\exp\left(t \sup_{P \in C} \langle \boldsymbol{\sigma}, v(P) \rangle \right) \right] && \text{(Jensen's inequality)} \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_{P \in C} \exp\left(\langle \boldsymbol{\sigma}, t v(P) \rangle \right) \right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_{P \in C} \prod_{i=1}^{m} \exp(t v(P)_i \sigma_i) \right] \\
&= \sum_{P \in C} \prod_{i=1}^{m+n} \mathop{\mathbb{E}}_{\sigma_i} \left[\exp(t v(P)_i \sigma_i) \right] \\
&\le |C| \exp\left(\tfrac{t^2 (m+n)^5}{8(mn)^2} \right) && \text{(Hoeffding's lemma).}
\end{aligned}$$

Thus,

$$\begin{aligned}
\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) &\le \frac{1}{m+n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{P \in C} \langle \boldsymbol{\sigma}, v(P) \rangle \right] + (\sqrt{2\mu} + \alpha) \tilde{\mathfrak{R}}_{U,m}(\mathcal{H}) \\
&\le \frac{\log |C|}{t(m+n)} + \frac{t(m+n)^4}{8(mn)^2} + 2(\sqrt{2\mu} + \alpha) \tilde{R}_{U,m}(\mathcal{H}).
\end{aligned}$$

Setting $t = \sqrt{\frac{8(mn)^2(\log|C|)}{(m+n)^5}}$ to minimize the bound, we obtain:

$$\widehat{\mathfrak{R}}^{\diamond}_{U,m}(\mathcal{Q}_{m,\mu}) \le \sqrt{\frac{(m+n)^3 \log |C|}{2(mn)^2}} + (\sqrt{2\mu} + \alpha) \tilde{\mathfrak{R}}_{U,m}(\mathcal{H}).$$

□

A.4 Proof of Lemma 3

Lemma 3. *Suppose the following bound holds with probability at least* $1 - \delta$ *over the choice of* S*: for all* $Q \in \mathcal{Q}_{S,\mu}$,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(\mu) + g(\delta),$$

where f *is an increasing function of* μ *and* g *is a decreasing function of* δ*. Then, the following holds with probability at least* $1 - \delta$ *for all* $Q \in \Delta(\mathcal{H})$*:*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(2\max\{D(Q\|P_S),2\}) + g\left(\tfrac{\delta}{\max\{D(Q\|P_S),2\}}\right).$$

Proof. The proof follows [Kakade et al., 2008][Corollary 8]. First, define the sequences $(\mu_j)_{j=0}^{\infty}$ and $(\delta_j)_{j=0}^{\infty}$. Let $a = 4$, $\mu_j := a2^j$ and $\delta_j := 2^{-(j+1)}\delta$, so that $\sum_{j=0}^{\infty}\delta_j = \delta$.

By the union bound, we thus have that with probability at least $1 - \delta$ over the draw of a sample $S \sim \mathcal{D}^m$, for all $Q \in \Delta(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(\mu_j) + g(\delta_j) \tag{11}$$

where μ_j is the smallest element of $(\mu_j)_{j=0}^{\infty}$ such that $D(Q\|P_S) \leq \mu_j$ (i.e., since we have a sequence of bounds holding for increasing values of μ_j, we choose the tightest applicable bound for each Q).

We now plug in the values of μ_j, δ_j:

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim S}}[L(h,z)] + f(a2^j) + g(2^{-(j+1)}\delta) \tag{12}$$

and try to upper bound the RHS in terms of $D(Q\|P_S)$, eliminating any appearances of j (i.e., we want a single bound that captures the sequence of bounds).

Upper bound μ_j: By the assumption that μ_j is the smallest element of $(\mu_j)_{j=0}^{\infty}$ such that $D(Q\|P_S) \leq \mu_j$, we necessarily have $D(Q\|P_S) > \mu_{j-1}$ for $j \geq 1$. (For $j = 0$, this simply yields $D(Q\|P_S) \geq 0$, which will not help, so we need to handle $j = 0$ separately.)

For $j \geq 1$, we thus have $D(Q\|P_S) > \mu_{j-1} = a2^{j-1}$, so $2D(Q\|P_S) > a2^j$.

For $j = 0$, $a2^j = a$.

This yields:

$$a2^j \leq \max\{2D(Q\|P_S), a\} = 2\max\{D(Q\|P_S), 2\}.$$

Lower bound δ_j: Since $\delta_j = 2^{-(j+1)}\delta$, we use the same assumption as above to obtain $4D(Q\|P_S) > a2^{j+1}$ and then use the definition of δ_j to obtain the lower bound: $\delta_j > \frac{a\delta}{4D(Q\|P_S)}$ for $j \geq 1$. For $j = 0$, we simply have $\delta_j = \delta/2$ by definition. This yields:

$$\delta_j \geq \min\left\{\frac{a\delta}{4D(Q\|P_S)}, \delta/2\right\} = \frac{\delta}{\max\{D(Q\|P_S),2\}}.$$

The stated bound follows from the monotonicities of f and g. □

B Proofs of results in Section 4

B.1 Proof of Theorem 3

We prove Theorem 3, with the exact bound explicitly spelled out:

Theorem 3. *Suppose $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ is β-uniformly stable. Then, for any $\delta > 0$, with probably at least $1-\delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following holds for all $Q \in \mathcal{Q}_S$:*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right]$$

$$+ 2\mathfrak{R}^{\diamond}_m(\mathscr{Q}_m) + \left(2\beta\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\tfrac{\log(4m^{1.5}/\delta)}{2m}}\right) + \tfrac{1}{m}\right)\sqrt{8m\log\left(\tfrac{4}{\delta}\right)}.$$

Proof. The proof is along the lines of the proof of Theorem 2 in [Foster et al., 2019] with a tighter analysis coming from the special structure in our setting. Specifically, for two samples $S, S' \in \mathcal{Z}^m$, define the function $\Psi(S,S')$ as follows:

$$\Psi(S,S') = \sup_{Q\in\mathcal{Q}_S} \langle Q,\ell\rangle - \langle Q,\hat{\ell}_{S'}\rangle,$$

where $\ell, \hat{\ell}_{S'} \in \mathfrak{R}^{\mathcal{H}}$ defined as $\ell(h) = \mathbb{E}_{z\sim\mathcal{D}}[L(h,z)]$ and $\hat{\ell}_{S'}(h) = \mathbb{E}_{z\sim S'}[L(h,z)]$, where $z \sim S'$ indicates uniform sampling from S'. The proof of the bound consists of applying McDiarmid's inequality to $\Psi(S,S)$. To do this, we need to analyze the sensitivity of this function, i.e. compute a bound on $|\Psi(S,S) - \Psi(S',S')|$ where S' is a sample differing from S in exactly one point. As in [Foster et al., 2019], we first observe that $\Psi(S,S) - \Psi(S,S') \leq \frac{1}{m}$, so now we turn to

$$\Psi(S,S') - \Psi(S',S') = \sup_{Q\in\mathcal{Q}_S} \langle Q,\ell\rangle - \langle Q,\hat{\ell}_{S'}\rangle - \sup_{Q\in\mathcal{Q}_{S'}} \langle Q,\ell\rangle - \langle Q,\hat{\ell}_{S'}\rangle.$$

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

$$\sup_{Q\in\mathcal{Q}_S} \langle Q,\ell\rangle - \langle Q,\hat{\ell}_{S'}\rangle - \epsilon \leq \sup_{Q\in\mathcal{Q}_S} \langle Q_\epsilon,\ell\rangle - \langle Q_\epsilon,\hat{\ell}_{S'}\rangle.$$

Using the β-stability of $\mathscr{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$, there exists a $Q'_\epsilon \in \mathcal{Q}_{S'}$ such that $\|Q_\epsilon - Q'_\epsilon\|_1 \leq 2\beta$. Thus, we have

$$\begin{aligned}
\Psi(S,S') - \Psi(S',S') &\leq \langle Q_\epsilon,\ell\rangle - \langle Q_\epsilon,\hat{\ell}_{S'}\rangle + \epsilon - \langle Q'_\epsilon,\ell\rangle - \langle Q'_\epsilon,\hat{\ell}_{S'}\rangle + \epsilon\\
&= \langle Q_\epsilon - Q'_\epsilon, \ell - \hat{\ell}_{S'}\rangle + \epsilon\\
&\leq \|Q_\epsilon - Q'_\epsilon\|_1 \|\ell - \hat{\ell}_{S'}\|_\infty + \epsilon\\
&\leq 2\beta \sup_h |\ell(h) - \hat{\ell}_{S'}(h)| + \epsilon.
\end{aligned}$$

Since this bound holds for any $\epsilon > 0$, we conclude that $\Psi(S,S') - \Psi(S',S') \leq 2\beta \sup_h |\ell(h) - \hat{\ell}_{S'}(h)|$, which implies that

$$\Psi(S,S) - \Psi(S',S') \leq 2\beta \sup_h |\ell(h) - \hat{\ell}_{S'}(h)| + \frac{1}{m} \leq 2\beta + \frac{1}{m}.$$

Now, via standard Rademacher complexity bounds Mohri et al. [2018], with probability at least $1-\delta$ over the choice of S', we have

$$\sup_h |\ell(h) - \hat{\ell}_{S'}(h)| \leq 2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2m}}.$$

Thus, with probability at least $1-\delta'$ over the choice of S', we have

$$\Psi(S,S) - \Psi(S',S') \leq 2\beta\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta')}{2m}}\right) + \frac{1}{m}.$$

Define $B := 2\beta\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(2/\delta')}{2m}}\right) + \frac{1}{m}$ for notational convenience. Now we can apply a variant of McDiarmid's inequality that allow almost-everywhere stability [Kutin and Niyogi, 2002] (using the explicit form in Theorem 5.2 in [Rakhlin et al., 2005] with $M = 2\beta + \frac{1}{m}$, $\beta_n = B$, and $\delta_n = \delta'$) to conclude that for any $t > 0$,

$$\mathbb{P}[|\Psi(S,S) - \mathbb{E}\,\Psi(S,S)| \ge t] \le 2\exp\left(\frac{-t^2}{8nB^2}\right) + \frac{2(2\beta + \frac{1}{m})m\delta'}{B} \le 2\exp\left(\frac{-t^2}{8nB^2}\right) + 2m^{1.5}\delta'.$$

Now, set $\delta' = \frac{\delta}{2m^{1.5}}$ and $t = B\sqrt{8m\log(\frac{4}{\delta})}$ so that $\mathbb{P}[|\Psi(S,S) - \mathbb{E}\,\Psi(S,S)| \ge t] \le \delta$. Finally, exactly as in [Foster et al., 2019], we have $\mathbb{E}_{S\sim\mathcal{D}^m}[\Psi(S,S)] \le 2\mathfrak{R}^{\diamond}_m(\mathcal{Q}_m)$. □

B.2 Explicit bound of Theorem 4

Theorem 4. *Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Also assume that for some $\eta > 0$, we have $P_S(h) \ge \eta$ for all $h \in \mathcal{H}$, and all $S \in \mathcal{Z}^m$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q\|P_S), 2\}$,*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right] + 2\sqrt{\frac{4D}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2m/\eta)}{m}}} + \sqrt{\frac{8}{m}} + \frac{2\eta}{m}$$
$$+ \left(4\epsilon\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(4m^{1.5}D/\delta)}{2m}}\right) + \frac{1}{m}\right)\sqrt{8m\log\left(\frac{4D}{\delta}\right)}.$$

B.3 Proof Theorem 5

We prove Theorem 5, with the exact bound explicitly spelled out:

Theorem 5. *Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the sample $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q\|P_S), 2\}$,*

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \le \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right]$$
$$+ \max\left\{4\sqrt{\frac{4D + 4\log(2)}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2)}{m}}}, 8\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 8\epsilon^{4/5}\right\}$$
$$+ \frac{2}{\sqrt{m}} + \left(4\epsilon\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(4m^{1.5}D/\delta)}{2m}}\right) + \frac{1}{m}\right)\sqrt{8m\log\left(\frac{4D}{\delta}\right)}.$$

Proof. Define a sample-dependent family of distributions $\mathcal{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ where $\mathcal{Q}_S = \{Q\colon \mathsf{D}_\infty(Q\|P_S) \le \mu\}$ for some parameter μ. We now apply the bound in Theorem 3, using the bound on the Rademacher complexity from Lemma 9, and the bound $\beta \le 2\epsilon$ from Lemma 6. Finally, a uniform bound over all values of μ follows by an application of Lemma 3. □

Lemma 9. *If $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point, then*

$$\mathfrak{R}^{\diamond}_m(\mathcal{Q}_{m,\mu}) \le \max\left\{2\sqrt{\frac{2\mu + 4\log(2)}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2)}{m}}}, 4\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 4\epsilon^{4/5}\right\} + \frac{1}{\sqrt{m}}.$$

Proof. Assume $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point.

Now, we fix the value of $\boldsymbol{\sigma} \in \{-1,1\}^m$ and introduce the following two distributions on $\mathcal{H}$:

(1) Let $\mathcal{P}_{\boldsymbol{\sigma}}$ be a joint distribution on (S,T,h) induced by sampling $S,T \sim \mathcal{D}^m$, and then, conditioned on the values of S and T, sampling $h \sim P_{S_T^{\boldsymbol{\sigma}}}$, using the notation $P_{S_T^{\boldsymbol{\sigma}}}$ introduced for Equation 8.

(2) Let $\mathcal{P}$ be the marginal distribution of h induced by $\mathcal{P}_{\boldsymbol{\sigma}}$. We have dropped $\boldsymbol{\sigma}$ from the notation because - since all elements of S and T are sampled i.i.d. - we have:

$$\mathop{\mathbb{E}}_{S,T\sim\mathcal{D}^m}[P_{S_T^{\boldsymbol{\sigma}}}(h)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[P_S(h)],$$

i.e., the marginal distribution of h is independent of $\boldsymbol{\sigma}$.

We first invoke several differential privacy results to show that, for the distributions $\mathcal{P}_{\boldsymbol{\sigma}}$ and $\mathcal{P}$ as defined above, and $\kappa := \epsilon^2 m + \epsilon\sqrt{m\log(2/\gamma)}$, we have:

$$\mathsf{D}_{\infty}^{\gamma}(\mathcal{P}_{\boldsymbol{\sigma}} \parallel \mathcal{D}^{2m} \otimes \mathcal{P}) \leq \kappa. \tag{13}$$

Specifically, consider $U = (S,T)$ and $U' = (S',T')$ for $S,T,S',T' \in \mathcal{Z}^m$ such that U and U' differ by only *one* of their $2m$ elements. Then $S_T^{\boldsymbol{\sigma}}$ and $S'^{\boldsymbol{\sigma}}_{T'}$ can only differ by at most one element, so by our main assumption: $\mathsf{D}_{\infty}(P_{S_T^{\boldsymbol{\sigma}}} \parallel P_{S'^{\boldsymbol{\sigma}}_{T'}}) \leq \epsilon$. Crucially, another way of saying this is: the algorithm $\mathcal{A}$ taking $U = (S,T)$ as input and outputting $h \sim P_{S_T^{\boldsymbol{\sigma}}}$ is an ϵ-differentially private algorithm, so we can apply Theorem 20 in [Dwork et al., 2015], with an input of size $2m$, and obtain (13).

We now use Lemma 3.17 (Part 1) in [Dwork and Roth, 2014] to convert (13) into a result concerning D_{∞} vs. $\mathsf{D}_{\infty}^{\gamma}$, so we can more easily use it below. Specifically, by Lemma 3.17 (Part 1), there exists a distribution $\mathcal{P}'_{\boldsymbol{\sigma}}$ on (S,T,h) such that $\|\mathcal{P}_{\boldsymbol{\sigma}} - \mathcal{P}'_{\boldsymbol{\sigma}}\|_{\mathrm{TV}} \leq \gamma$ and $\mathsf{D}_{\infty}(\mathcal{P}'_{\boldsymbol{\sigma}} \parallel \mathcal{D}^{2m} \otimes \mathcal{P}) \leq \kappa$.

Finally, we upper bound $\mathfrak{R}_m^{\diamond}(\mathcal{Q}_{m,\mu})$ as follows. For convenience, we use a variable $t > 0$ and the function $\Psi_P(Q)$, which is defined as $\mathsf{D}(Q \parallel P)$ if $\mathsf{D}(Q \parallel P) \leq \mu$ and $+\infty$ otherwise; thus, its conjugate function is $\Psi_P^*(u) = \log\left(\mathbb{E}_{h\in P}[e^{u(h)}]\right)$, for all $u \in \mathbb{R}^{\mathcal{H}}$ [Mohri et al., 2018, Lemma B.37]. We also use the shorthand $u_{\boldsymbol{\sigma}}(h) = \sum_{i=1}^m \sigma_i L(h,z_i)$, where z_i is element i of sample T, so that $\sum_{i=1}^m \sigma_i \langle Q, L_{z_i}\rangle = \langle Q, u_{\boldsymbol{\sigma}}\rangle$.

$$\begin{aligned}
\mathfrak{R}_m^{\diamond}(\mathcal{Q}_{m,\mu}) &= \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{(S,T)}\left[\sup_{\mathsf{D}(Q\parallel P_{S_T^{\boldsymbol{\sigma}}})\leq\mu}\langle Q, tu_{\boldsymbol{\sigma}}\rangle\right] \\
&\leq \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{(S,T)}\left[\sup_{\Psi_{P_{S_T^{\boldsymbol{\sigma}}}}(Q)\leq\mu}\Psi_{P_{S_T^{\boldsymbol{\sigma}}}}(Q) + \Psi^*_{P_{S_T^{\boldsymbol{\sigma}}}}(tu_{\boldsymbol{\sigma}})\right] \quad \text{(Fenchel inequality)} \\
&\leq \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{(S,T)}\left[\Psi^*_{P_{S_T^{\boldsymbol{\sigma}}}}(tu_{\boldsymbol{\sigma}})\right] \\
&= \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{(S,T)}\left[\log\left(\mathop{\mathbb{E}}_{h\sim P_{S_T^{\boldsymbol{\sigma}}}}\left[e^{tu_{\boldsymbol{\sigma}}(h)}\right]\right)\right] \qquad \text{(definition of } \Psi^*\text{)} \\
&\leq \frac{\mu}{mt} + \frac{1}{mt}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\log\left(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{P}_{\boldsymbol{\sigma}}}\left[e^{tu_{\boldsymbol{\sigma}}(h)}\right]\right) \qquad \text{(Jensen's inequality)}
\end{aligned} \tag{14}$$

In the following, to make the dependence of $u_{\boldsymbol{\sigma}}$ on the set T explicit, we now denote it as $u_{\boldsymbol{\sigma},T}$. For any sample T, define $\Psi(T)$ by $\Psi(T) = \frac{1}{m}\sup_{h\in\mathcal{H}}\left(u_{\boldsymbol{\sigma},T}(h) - \mathbb{E}_{T'\sim\mathcal{D}^m}[u_{\boldsymbol{\sigma},T'}(h)]\right)$. Changing one point in T affects $\Psi(T)$ by at most $1/m$, since the loss is bounded by one. Thus, by McDiarmid's inequality, for any fixed $\boldsymbol{\sigma}$ and for any $\delta > 0$, we have

$$\mathop{\mathbb{P}}_{T\sim\mathcal{D}^m}\left[\Psi(T) \leq \mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}[\Psi(T)] + \sqrt{\frac{2\log(\frac{1}{\delta})}{m}}\right] \geq 1 - \delta.$$

Now, $\mathbb{E}_{T\sim\mathcal{D}^m}[\Psi(T)]$ can be bounded in terms of the Rademacher complexity as in the standard analyses:

$$\begin{aligned}
\underset{T\sim\mathcal{D}^m}{\mathbb{E}}[\Psi(T)] &= \frac{1}{m}\underset{T\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\underset{T'\sim\mathcal{D}^m}{\mathbb{E}}[u_{\boldsymbol{\sigma},T}(h) - u_{\boldsymbol{\sigma},T'}(h)]\right] \\
&\leq \frac{1}{m}\underset{T,T'\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}} u_{\boldsymbol{\sigma},T}(h) - u_{\boldsymbol{\sigma},T'}(h)\right] && \text{(sub-additivity of sup)} \\
&\leq \frac{1}{m}\underset{T,T'\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\left(\sigma_i L(h, z_i^T) - \sigma_i L(h, z_i^{T'})\right)\right] \\
&\leq \frac{1}{m}\underset{T,T'\sim\mathcal{D}^m,\boldsymbol{\beta}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\beta_i\left(\sigma_i L(h, z_i^T) - \sigma_i L(h, z_i^{T'})\right)\right] && \text{(Rademacher variables } \beta_i\text{)} \\
&\leq \frac{2}{m}\underset{T\sim\mathcal{D}^m,\boldsymbol{\beta}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\beta_i\left(\sigma_i L(h, z_i^T)\right)\right] \\
&= \frac{2}{m}\underset{T\sim\mathcal{D}^m,\boldsymbol{\beta}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\beta_i L(h, z_i^T)\right] \\
&= 2\mathfrak{R}_m(\mathcal{H}).
\end{aligned}$$

Thus, for any fixed $\boldsymbol{\sigma}$ and for any $\delta > 0$, we have

$$\underset{T\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_h\left(u_{\boldsymbol{\sigma},T}(h) - \underset{T'\sim\mathcal{D}^m}{\mathbb{E}}[u_{\boldsymbol{\sigma},T'}(h)]\right) \leq 2m\mathfrak{R}_m(\mathcal{H}) + \sqrt{2m\log(1/\delta)}\right] \geq 1 - \delta. \tag{15}$$

Note that for any h, we have $\mathbb{E}_{T'\sim\mathcal{D}^m}[u_{\boldsymbol{\sigma},T'}(h)] = \sum_{i=1}^m \sigma_i \mathbb{E}_{z\sim D}[L(h,z)]$, and hence $|\mathbb{E}_{T'\sim\mathcal{D}^m}[u_{\boldsymbol{\sigma},T'}(h)]| \leq |\sum_{i=1}^m \sigma_i|$. Hence, we conclude that

$$\underset{T\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_h u_{\boldsymbol{\sigma},T}(h) \leq \left|\sum_{i=1}^{m}\sigma_i\right| + 2m\mathfrak{R}_m(\mathcal{H}) + \sqrt{2m\log(1/\delta)}\right] \geq 1 - \delta. \tag{16}$$

For notational convenience, define

$$B_{\boldsymbol{\sigma}} := \left|\sum_{i=1}^{m}\sigma_i\right| + 2m\mathfrak{R}_m(\mathcal{H}) + \sqrt{2m\log(1/\delta)}.$$

Now, let $\delta := e^{-tm}$, and let $G \subseteq \mathcal{Z}^m$ be the set of m-element samples T such that

$$G := \left\{T \in \mathcal{Z}^m \colon \sup_h u_{\boldsymbol{\sigma},T}(h) \leq B_{\boldsymbol{\sigma}}\right\}.$$

By (15), we have $\mathbb{P}_{T\sim\mathcal{D}^m}[G] \geq 1 - \delta$. Hence, we have

$$\begin{aligned}
\underset{(S,T,h)\sim\mathcal{P}_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma},T}(h)}\right] &\leq \underset{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma},T}(h)}\right] + \left(\sup_{T\in G}\sup_h e^{tu_{\boldsymbol{\sigma},T}(h)}\right)\cdot\left(\left|\underset{\mathcal{P}_{\boldsymbol{\sigma}}}{\mathbb{P}}[T\in G] - \underset{\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{P}}[T\in G]\right|\right) \\
&\quad + e^{tm}\cdot\underset{\mathcal{P}_{\boldsymbol{\sigma}}}{\mathbb{P}}[T\notin G] \\
&\leq \underset{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma},T}(h)}\right] + \gamma e^{tB_{\boldsymbol{\sigma}}} + e^{tm}\delta \\
&= \underset{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma},T}(h)}\right] + \gamma e^{tB_{\boldsymbol{\sigma}}} + 1 \\
&\leq \left(\underset{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma},T}(h)}\right] + 1\right)\cdot\left(\gamma e^{tB_{\boldsymbol{\sigma}}} + 1\right).
\end{aligned}$$

Using this bound in (14), we get

$$\mathfrak{R}_m^{\diamond}(\mathcal{Q}_{m,\mu}) \leq \frac{\mu}{mt} + \frac{1}{mt}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\log\left(\underset{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}}{\mathbb{E}}\left[e^{tu_{\boldsymbol{\sigma}}(h)}\right] + 1\right) + \log\left(\gamma e^{tB_{\boldsymbol{\sigma}}} + 1\right)\right]. \tag{17}$$

We bound the two terms involving the logarithm in (17) separately. First, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{P}'_{\boldsymbol{\sigma}}} \left[e^{tu_{\boldsymbol{\sigma}}(h)} \right] + 1 \right) \\
&\qquad \le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathscr{P}} \left[e^{\kappa} e^{tu_{\boldsymbol{\sigma}}(h)} \right] + 1 \right) \quad (\text{since } \mathsf{D}_\infty(\mathcal{P}'_{\boldsymbol{\sigma}} \| \mathcal{D}^{2m} \otimes \mathscr{P}) \ \le\ \kappa) \\
&\qquad \le \log\left(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathscr{P}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[e^{\kappa} e^{tu_{\boldsymbol{\sigma}}(h)} \right] + 1 \right) \qquad (\text{Jensen's inequality}) \\
&\qquad \le \log\left(\mathop{\mathbb{E}}_{(S,T,h)\sim\mathcal{D}^{2m}\otimes\mathscr{P}} e^{\kappa + mt^2/2} + 1 \right) \qquad (\text{Hoeffding's lemma}) \\
&\qquad \le \log\left(2e^{\kappa + mt^2/2} \right) \qquad (e^{k+mt^2/2} \ge 1) \\
&\qquad \le \kappa + \frac{mt^2}{2} + \log(2). \qquad\qquad (18)
\end{aligned}$$

As for the second term, setting $\gamma = e^{-(2mt\mathfrak{R}_m(\mathcal{H}) + \sqrt{2}mt^{3/2})}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(\gamma e^{tB_{\boldsymbol{\sigma}}} + 1 \right) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(\gamma e^{t(|\sum_{i=1}^m \sigma_i| + 2m\mathfrak{R}_m(\mathcal{H}) + \sqrt{2m\log(1/\delta)})} + 1 \right) \quad (\text{definition of } B_{\boldsymbol{\sigma}}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(e^{t|\sum_{i=1}^m \sigma_i|} + 1 \right) \quad (\text{using } \gamma = e^{-(2mt\mathfrak{R}_m(\mathcal{H}) + \sqrt{2}mt^{3/2})}) \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \log\left(2e^{t|\sum_{i=1}^m \sigma_i|} \right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[t \left| \sum_{i=1}^m \sigma_i \right| \right] + \log(2) \\
&= t \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sqrt{(\textstyle\sum_{i=1}^m \sigma_i)^2} \right] + \log(2) \\
&\le t \sqrt{ \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[(\textstyle\sum_{i=1}^m \sigma_i)^2 \right] } + \log(2) \quad (\text{Jensen's inequality}) \\
&= \sqrt{m}t + \log(2) \qquad\qquad (19)
\end{aligned}$$

Using bounds (18), (19), and the bound on k in (17) we get

$$\begin{aligned}
\mathfrak{R}^{\diamond}_m(\mathcal{Q}_{m,\mu}) &\le \frac{1}{mt}\left(\mu + \kappa + \frac{mt^2}{2} + \sqrt{m}t + 2\log(2) \right) \\
&\le \frac{1}{mt}\left(\mu + \epsilon^2 m + \epsilon\sqrt{m(2mt\mathfrak{R}_m(\mathcal{H}) + \sqrt{2}mt^{3/2})) + m\log(2)} + \frac{mt^2}{2} + \sqrt{m}t + 2\log(2) \right) \\
&\le \max\left\{ 2\sqrt{ \frac{2\mu + 4\log(2)}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2)}{m}} }, 4\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 4\epsilon^{4/5} \right\} + \frac{1}{\sqrt{m}},
\end{aligned}$$

setting $t = \max\left\{ \sqrt{ \frac{2\mu+4\log(2)}{m} + 2\epsilon^2 + 2\epsilon\sqrt{\frac{\log(2)}{m}} }, 2\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 2\epsilon^{4/5} \right\}$. □

B.4 Proof of Theorem 6

The requirement in Theorem 5 that the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ has D_∞ sensitivity ϵ is equivalent to saying that the priors define an ϵ-differentially private mechanism. Here, we give an extension to Theorem 5 which makes the weaker assumption that the priors define an (ϵ, δ)-differentially private mechanism, for some $\delta > 0$. The extension relies on the following theorem of Rogers et al. [2016]. The statement given below is an adaptation of Theorem 3.1 in [Rogers et al., 2016] that is implicit in their proof. We need this more nuanced statement for our analysis.

Theorem 7 (Theorem 3.1 in [Rogers et al., 2016]). *Let $\mathcal{A} : \mathcal{X}^m \to \mathcal{Y}$ be an (ϵ, δ)-differentially private algorithm for $\epsilon \in (0, \frac{1}{2}]$ and $\delta \in (0, \epsilon)$. Let $\mathcal{D}$ be any distribution on $\mathcal{X}$ and let $S \in \mathcal{X}^m$ be a dataset with elements sampled i.i.d. from $\mathcal{D}$. Let $\mathcal{P}$ be the joint distribution of $(S, \mathcal{A}(S))$, and $\mathscr{P}$ be the marginal distribution of $\mathcal{A}(S)$. Then there is a constant $c > 0$ such that for any $\gamma \in (0, 1]$ we have*

$$\mathsf{D}_\infty^{\delta + c\sqrt{\frac{\delta}{\epsilon}}m}(\mathcal{P} \,\|\, \mathcal{D}^m \otimes \mathscr{P}) \le 72\epsilon^2 m + 6\epsilon\sqrt{2m\log(1/\gamma)} + c\sqrt{\frac{\delta}{\epsilon}}m.$$

With this theorem, we can now prove the following theorem which is analogous to Theorem 5 but assumes only the priors define an (ϵ,δ)-differentially private mechanism.

Theorem 6. *Assume that $\epsilon \geq 0$ and $\delta \in [0, \frac{e^{-16m}}{4c^2m^2}\epsilon]$, where c is the constant from Theorem 7. Suppose the family of sample-dependent priors $(P_S)_{S\in\mathcal{Z}^m}$ satisfy the property that $\mathsf{D}^{\delta}_{\infty}(P_S\|P_{S'}) \leq \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing in exactly one point. Then, for any $\nu > 0$, with probability at least $1-\nu$ over the draw of the sample $S\sim\mathcal{D}^m$, the following inequality holds for all $Q\in\Delta(\mathcal{H})$: if $D = \max\{\mathsf{D}(Q\|P_S), 2\}$,*

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h,z_i)\right]&\\
&+ \max\left\{4\sqrt{\frac{4D+6\log(2)}{m}+300\epsilon^2}, 30\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 30\epsilon^{4/5}\right\}\\
&+ \frac{2}{\sqrt{m}} + \frac{c\sqrt{\delta}}{4\epsilon^{3/2}} + \left(4\epsilon\left(2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(4m^{1.5}D/\nu)}{2m}}\right) + \frac{1}{m}\right)\sqrt{8m\log\left(\tfrac{4D}{\nu}\right)}.
\end{aligned}
$$

Proof. Define a sample-dependent family of distributions $\mathcal{Q}_m = (\mathcal{Q}_S)_{S\in\mathcal{Z}^m}$ where $\mathcal{Q}_S = \{Q\colon \mathsf{D}_\infty(Q\|P_S)\leq\mu\}$ for some parameter μ. We now apply the bound in Theorem 3, using the bound on the Rademacher complexity from Lemma 10, and the bound $\beta\leq 2\epsilon$ from Lemma 6. Finally, a uniform bound over all values of μ follows by an application of Lemma 3. □

Lemma 10. *Assume that $\epsilon\geq 0$ and $\delta\in[0,\frac{e^{-16m}}{4c^2m^2}\epsilon]$, where c is the constant from Theorem 7. Suppose that $\mathsf{D}^{\delta}_{\infty}(P_S\|P_{S'})\leq\epsilon$ for all $S, S'\in\mathcal{Z}^m$ differing in exactly one point. Then,*

$$
\mathfrak{R}^{\diamond}_m(\mathcal{Q}_{m,\mu}) \leq \max\left\{2\sqrt{\frac{2\mu+6\log(2)}{m}+300\epsilon^2}, 15\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 15\epsilon^{4/5}\right\} + \frac{1}{\sqrt{m}} + \frac{c\sqrt{\delta}}{8\epsilon^{3/2}}.
$$

Proof. The proof is exactly along the lines of the proof of Lemma 9. Instead of using Theorem 20 in [Dwork et al., 2015], we use Theorem 7 above. Using this theorem, the proof of Lemma 10 follows with

$$
\kappa = 144\epsilon^2 m + 12\epsilon\sqrt{m\log(1/\gamma)} + 2c\sqrt{\tfrac{\delta}{\epsilon}}m
$$

and γ replaced by $\gamma + 2c\sqrt{\frac{\delta}{\epsilon}}m$. The bound (19) changes as follows: setting $\gamma = e^{-(2mt\mathfrak{R}_m(\mathcal{H})+\sqrt{2}mt^{3/2})}$ exactly as in the proof of Lemma 9, and assuming that we choose $t\leq 2$ ($t>2$ leads to a trivial bound), we note that $\gamma + 2c\sqrt{\frac{\delta}{\epsilon}}m \leq 2\gamma$ since we assumed that $\delta\leq\frac{e^{-16m}}{4c^2m^2}\epsilon$, and hence

$$
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\log\left((\gamma + 2c\sqrt{\tfrac{\delta}{\epsilon}}m)e^{tB_{\boldsymbol{\sigma}}}+1\right) \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\log\left(2\gamma e^{tB_{\boldsymbol{\sigma}}}+1\right) \leq \sqrt{m}t + \log(4).
$$

Finally, we have

$$
\begin{aligned}
\mathfrak{R}^{\diamond}_m(\mathcal{Q}_{m,\mu}) &\leq \frac{1}{mt}\left(\mu + \kappa + \frac{mt^2}{2} + \sqrt{m}t + 3\log(2)\right)\\
&\leq \frac{1}{mt}\Big(\mu + 144\epsilon^2 m + 12\epsilon\sqrt{m(2mt\mathfrak{R}_m(\mathcal{H}) + \sqrt{2}mt^{3/2}))} + 2c\sqrt{\frac{\delta}{\epsilon}}m + \frac{mt^2}{2} + \sqrt{m}t\\
&\qquad + 3\log(2)\Big)\\
&\leq \max\left\{2\sqrt{\frac{2\mu+6\log(2)}{m}+300\epsilon^2}, 15\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 15\epsilon^{4/5}\right\} + \frac{1}{\sqrt{m}} + \frac{c\sqrt{\delta}}{8\epsilon^{3/2}},
\end{aligned}
$$

setting $t = \min\left\{\max\left\{\sqrt{\frac{2\mu+6\log(2)}{m}+300\epsilon^2}, 15\epsilon^{2/3}\mathfrak{R}_m(\mathcal{H})^{1/3}, 15\epsilon^{4/5}\right\}, 2\right\}$ and using the bound $\frac{2c}{t}\sqrt{\frac{\delta}{\epsilon}} \leq \frac{2c}{\sqrt{300\epsilon^2}}\sqrt{\frac{\delta}{\epsilon}} \leq \frac{c\sqrt{\delta}}{8\epsilon^{3/2}}$. □

Remark. The stipulation that $\delta \leq \frac{e^{-16m}}{4c^2m^2}\epsilon$ in the statement of Lemma 10 is made simply to yield a clean statement. It should be evident from the proof that other values of δ also yield analogous bounds on the Rademacher complexity. For example, we can allow δ to be as large as $\frac{e^{-(4mt\mathfrak{R}_m(\mathcal{H})+2\sqrt{2}mt^{3/2})}}{4c^2m^2}\epsilon$ for the value of t in the proof above and retain the exact same bound.

B.5 Proof of Lemma 5

Lemma 5. *Suppose $\|P_S - P_{S'}\|_1 \leq \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \geq 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} := \{Q\colon \mathsf{D}(Q\|P_S) \leq \mu\}$, and the corresponding family to be $\mathscr{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$. Then $\mathscr{Q}_{m,\mu}$ is β-stable for $\beta = \min\left\{\frac{\epsilon d_\infty}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon d_\infty}{2}}\right\}$, where $d_\infty := \sup_{S,S',Q\in\mathcal{Q}_{S,\mu}} \left\|\frac{Q}{P_{S'}}\right\|_\infty$.*

Proof. Consider an arbitrary $Q \in \mathcal{Q}_{S,\mu}$.

Case (1): $\mathsf{D}(Q \parallel P_{S'}) \leq \mu$.
In this case, $Q \in \mathcal{Q}_{S',\mu}$, so we choose $Q' = Q$, and thus $\|Q' - Q\|_{\mathrm{TV}} = 0$.

Case (2): $\mathsf{D}(Q \parallel P_{S'}) > \mu$.
We consider $Q' = \lambda Q + (1-\lambda)P_{S'}$, for $\lambda = \frac{\mathsf{D}(Q\|P_S)}{\mathsf{D}(Q\|P_{S'})} < 1$. We show that $Q' \in \mathcal{Q}_{S',\mu}$ as follows:

$$\mathsf{D}(Q' \parallel P_{S'}) = \mathsf{D}(\lambda Q + (1-\lambda)P_{S'} \parallel P_{S'}) \leq \lambda \mathsf{D}(Q \parallel P_{S'}) + (1-\lambda)\mathsf{D}(P_{S'} \parallel P_{S'}) = \mathsf{D}(Q \parallel P_S) \leq \mu,$$

where the inequality is by the convexity of relative entropy.

We can upper bound $\|Q' - Q\|_{\mathrm{TV}}$ in two different ways.
One way is to directly upper bound the TV distance as follows:

$$\begin{aligned}
\|Q' - Q\|_{\mathrm{TV}} &= \|\lambda Q + (1-\lambda)P_{S'} - Q\|_{\mathrm{TV}} \\
&= (1-\lambda)\|Q - P_{S'}\|_{\mathrm{TV}} \\
&= \left[1 - \frac{\mathsf{D}(Q \parallel P_S)}{\mathsf{D}(Q \parallel P_{S'})}\right]\|Q - P_{S'}\|_{\mathrm{TV}} \\
&= \left[\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)\right]\frac{\|Q - P_{S'}\|_{\mathrm{TV}}}{\mathsf{D}(Q \parallel P_{S'})} \\
&\leq \frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}} \qquad \text{(Pinsker's inequality).}
\end{aligned}$$

Alternatively, we can upper bound the TV distance by upper bounding the KL divergence as follows:

$$\begin{aligned}
\mathsf{D}(Q \parallel Q') &= \mathsf{D}(Q \parallel \lambda Q + (1-\lambda)P_{S'}) \\
&\leq (1-\lambda)\mathsf{D}(Q \parallel P_{S'}) \qquad \text{(convexity of relative entropy)} \\
&= \left[1 - \frac{\mathsf{D}(Q \parallel P_S)}{\mathsf{D}(Q \parallel P_{S'})}\right]\mathsf{D}(Q \parallel P_{S'}) \\
&= \mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) \\
\implies \|Q' - Q\|_{\mathrm{TV}} &\leq \sqrt{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{2}} \qquad \text{(Pinsker's inequality).}
\end{aligned}$$

We upper bound the common term $\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)$ as follows:

$$\begin{aligned}
\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) &= \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{Q(h)}{P_{S'}(h)}\right] - \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{Q(h)}{P_S(h)}\right] && \text{(def. of relative entropy)}\\
&= \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{P_S(h)}{P_{S'}(h)}\right]\\
&\le \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{P_S(h)}{P_{S'}(h)} - 1\right] && (\log x \le x - 1)\\
&= \sum_{h\in\mathcal{H}} Q(h)\left[\frac{P_S(h)}{P_{S'}(h)} - 1\right]\\
&= \sum_{h\in\mathcal{H}} \frac{Q(h)}{P_{S'}(h)}\left[P_S(h) - P_{S'}(h)\right]\\
&\le \left\|\frac{Q}{P_{S'}}\right\|_\infty \|P_S - P_{S'}\|_1 && \text{(Hölder's inequality)}\\
&\le \epsilon d_\infty\left(\frac{Q}{P_{S'}}\right),
\end{aligned}$$

where $d_\infty(f) \coloneqq \|f\|_\infty$.

Putting this together, we obtain:

$$\begin{aligned}
\|Q' - Q\|_{\mathrm{TV}} &\le \min\left\{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}}, \sqrt{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{2}}\right\}\\
&\le \min\left\{\frac{\epsilon}{\sqrt{2\mu}} d_\infty\left(\frac{Q}{P_{S'}}\right), \sqrt{\frac{\epsilon}{2} d_\infty\left(\frac{Q}{P_{S'}}\right)}\right\}.
\end{aligned}$$

For convenience, define $d_\infty \coloneqq \sup_{S,S',Q\in\mathcal{Q}_{S,\mu}} d_\infty\left(\frac{Q}{P_{S'}}\right)$.

Thus, if we define $\beta \coloneqq \min\left\{\frac{\epsilon}{\sqrt{2\mu}} d_\infty, \sqrt{\frac{\epsilon}{2} d_\infty}\right\}$, then the family $\mathcal{Q}_{m,\mu}$ is β-uniformly stable.

□

B.6 Proof of Lemma 6

Lemma 6. *Suppose $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \ge 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} \coloneqq \{Q\colon \mathsf{D}(Q\|P_S) \le \mu\}$, and the corresponding family to be $\mathcal{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S\in\mathcal{Z}^m}$. Then $\mathcal{Q}_{m,\mu}$ is β-stable for $\beta = \min\left\{2\epsilon, \frac{\epsilon}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon}{2}}\right\}$.*

Proof. This follows from Lemmas 11 and 12. □

Lemma 11. *If $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point, then $\mathcal{Q}_{m,\mu}$ is β-uniformly stable with $\beta = \min\left\{\frac{\epsilon}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon}{2}}\right\}$.*

Proof. Consider an arbitrary $Q \in \mathcal{Q}_{S,\mu}$.

Case (1): $\mathsf{D}(Q \parallel P_{S'}) \le \mu$.
In this case, $Q \in \mathcal{Q}_{S',\mu}$, so we choose $Q' = Q$, and thus $\|Q' - Q\|_{\mathrm{TV}} = 0$.

Case (2): $\mathsf{D}(Q \parallel P_{S'}) > \mu$.
We consider $Q' = \lambda Q + (1-\lambda)P_{S'}$, for $\lambda = \frac{\mathsf{D}(Q\|P_S)}{\mathsf{D}(Q\|P_{S'})} < 1$. We show that $Q' \in \mathcal{Q}_{S',\mu}$ as follows:

$$\mathsf{D}(Q' \parallel P_{S'}) = \mathsf{D}(\lambda Q + (1-\lambda)P_{S'} \parallel P_{S'}) \le \lambda\mathsf{D}(Q \parallel P_{S'}) + (1-\lambda)\mathsf{D}(P_{S'} \parallel P_{S'}) = \mathsf{D}(Q \parallel P_S) \le \mu,$$

where the inequality is by the convexity of relative entropy.

We can upper bound $\|Q' - Q\|_{\mathrm{TV}}$ in two different ways.
One way is to directly upper bound the TV distance as follows:

$$\begin{aligned}
\|Q' - Q\|_{\mathrm{TV}} &= \|\lambda Q + (1-\lambda)P_{S'} - Q\|_{\mathrm{TV}} \\
&= (1-\lambda)\|Q - P_{S'}\|_{\mathrm{TV}} \\
&= \left[1 - \frac{\mathsf{D}(Q \parallel P_S)}{\mathsf{D}(Q \parallel P_{S'})}\right]\|Q - P_{S'}\|_{\mathrm{TV}} \\
&= [\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)]\frac{\|Q - P_{S'}\|_{\mathrm{TV}}}{\mathsf{D}(Q \parallel P_{S'})} \\
&\le \frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}} && \text{(Pinsker's inequality).}
\end{aligned}$$

Alternatively, we can upper bound the TV distance by upper bounding the KL divergence as follows:

$$\begin{aligned}
\mathsf{D}(Q \parallel Q') &= \mathsf{D}(Q \parallel \lambda Q + (1-\lambda)P_{S'}) \\
&\le (1-\lambda)\mathsf{D}(Q \parallel P_{S'}) && \text{(convexity of relative entropy)} \\
&= \left[1 - \frac{\mathsf{D}(Q \parallel P_S)}{\mathsf{D}(Q \parallel P_{S'})}\right]\mathsf{D}(Q \parallel P_{S'}) \\
&= \mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) \\
\implies \|Q' - Q\|_{\mathrm{TV}} &\le \sqrt{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{2}} && \text{(Pinsker's inequality).}
\end{aligned}$$

We upper bound the common term $\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)$ as follows:

$$\begin{aligned}
\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) &= \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{Q(h)}{P_{S'}(h)}\right] - \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{Q(h)}{P_S(h)}\right] && \text{(def. of relative entropy)} \\
&= \mathop{\mathbb{E}}_{h\sim Q}\left[\log\frac{P_S(h)}{P_{S'}(h)}\right] \\
&\le \mathsf{D}_\infty(P_S \parallel P_{S'}).
\end{aligned}$$

Putting this together, we obtain:

$$\|Q' - Q\|_{\mathrm{TV}} \le \frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}} < \frac{D_\infty(P_S \parallel P_{S'})}{\sqrt{2\mu}} \le \frac{\epsilon}{\sqrt{2\mu}}.$$

$$\begin{aligned}
\|Q' - Q\|_{\mathrm{TV}} &\le \min\left\{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}}, \sqrt{\frac{\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)}{2}}\right\} \\
&\le \min\left\{\frac{\mathsf{D}_\infty(P_S \parallel P_{S'})}{\sqrt{2\mu}}, \sqrt{\frac{D_\infty(P_S \parallel P_{S'})}{2}}\right\} \\
&\le \min\left\{\frac{\epsilon}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon}{2}}\right\}.
\end{aligned}$$

So if we define $\beta \coloneqq \min\left\{\frac{\epsilon}{\sqrt{2\mu}}, \sqrt{\frac{\epsilon}{2}}\right\}$, then the family $\mathcal{Q}_{m,\mu}$ is β-uniformly stable. □

Lemma 12. *If $\mathsf{D}_\infty(P_S \parallel P_{S'}) \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point, then $\mathcal{Q}_{m,\mu}$ is β-uniformly stable with $\beta = 2\epsilon$.*

Proof. For convenience, we measure stability using the total variation distance rather than ℓ_1, and then present the final bound in terms of ℓ_1 stability.

Consider an arbitrary $Q \in \mathcal{Q}_{S,\mu}$.

Case (1): $\mathsf{D}(Q \parallel P_{S'}) \leq \mathsf{D}(Q \parallel P_S)$.
In this case, $Q \in \mathcal{Q}_{S',\mu}$, so we choose $Q' = Q$, and thus $\|Q' - Q\|_{\mathrm{TV}} = 0$.

Case (2): $\mathsf{D}(Q \parallel P_{S'}) > \mathsf{D}(Q \parallel P_S)$.
We consider $Q' = \lambda Q + (1-\lambda) P_{S'}$, for $\lambda = \frac{\mathsf{D}(Q\|P_S)}{\mathsf{D}(Q\|P_{S'})} < 1$. We show that $Q' \in \mathcal{Q}_{S',\mu}$ as follows:

$$\mathsf{D}(Q' \parallel P_{S'}) = \mathsf{D}(\lambda Q + (1-\lambda)P_{S'} \parallel P_{S'}) \leq \lambda \mathsf{D}(Q \parallel P_{S'}) + (1-\lambda)\mathsf{D}(P_{S'} \parallel P_{S'}) = \mathsf{D}(Q \parallel P_S) \leq \mu,$$

where the inequality is by the convexity of relative entropy.

Next we will upper bound $\mathsf{D}(Q' \parallel P_S)$. For this we will use the fact that $\mathsf{D}(P_S \parallel P_{S'}) \leq 2\epsilon^2$. This fact is from [Popescu et al.] and we provide an alternate proof in Lemma 13 below. Given the lemma we have

$$\begin{aligned}
\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) &= \mathop{\mathbb{E}}_{h\sim Q}\Big[\log \frac{P_S(h)}{P_{S'}(h)}\Big] \\
&= \mathop{\mathbb{E}}_{h\sim P}\Big[\log \frac{P_S(h)}{P_{S'}(h)}\Big] + \big(\mathop{\mathbb{E}}_{h\sim Q} - \mathop{\mathbb{E}}_{h\sim P}\big)\Big[\log \frac{P_S(h)}{P_{S'}(h)}\Big] \\
&\leq \mathsf{D}(P_S, P_{S'}) + \epsilon\|Q - P\|_{\mathrm{TV}} \\
&\leq 2\epsilon^2 + \epsilon\|Q - P_S\|_{\mathrm{TV}} \\
&\leq 2\epsilon^2 + \epsilon\sqrt{\frac{\mathsf{D}(Q \parallel P_S)}{2}}. \text{ (Pinsker's inequality)} \qquad (20)
\end{aligned}$$

Next we show that Q and Q' are close in total variation distance. We consider two cases:

Case a: $\mathsf{D}(Q \parallel P_S) \leq 2\epsilon^2$. Using convexity of $\mathsf{D}(Q \parallel .)$ we have

$$\begin{aligned}
\mathsf{D}(Q \parallel Q') &\leq (1-\lambda)\mathsf{D}(Q \parallel P_{S'}) \\
&= \mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S) \\
&\leq 2\epsilon^2 + \epsilon\sqrt{\frac{\mathsf{D}(Q \parallel P_S)}{2}} \text{ [from (20)]} \\
&\leq 3\epsilon^2.
\end{aligned}$$

Using Pinsker's inequality we can conclude that $\|Q - Q'\|_{\mathrm{TV}} \leq 2\epsilon$.

Case b: $\mathsf{D}(Q \parallel P_S) > 2\epsilon^2$. We have

$$\begin{aligned}
\|Q - Q'\|_{\mathrm{TV}} &= (1-\lambda)\|Q - P_{S'}\|_{\mathrm{TV}} \\
&= \big(\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)\big)\frac{\|Q - P_{S'}\|_{\mathrm{TV}}}{\mathsf{D}(Q \parallel P_{S'})} \\
&\leq \big(\mathsf{D}(Q \parallel P_{S'}) - \mathsf{D}(Q \parallel P_S)\big)\frac{1}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}} \\
&\text{[from Pinsker's inequality and the fact that } \mathsf{D}(Q \parallel P_{S'}) > \mathsf{D}(Q \parallel P_S)] \\
&\leq \frac{2\epsilon^2}{\sqrt{2\mathsf{D}(Q \parallel P_{S'})}} + \frac{\epsilon}{2} \text{ [from (20)]} \\
&\leq 2\epsilon \text{ [since } \mathsf{D}(Q \parallel P_S) > 2\epsilon^2].
\end{aligned}$$

□

Lemma 13. *If* $\mathsf{D}_\infty(P_S, P_{S'}) \leq \epsilon$ *for all* $S, S' \in \mathcal{Z}^m$ *differing by exactly one point, then* $\mathsf{D}(P_S \parallel P_{S'}) \leq 2\epsilon^2$.

Proof. Suppose $\mathsf{D}_\infty(P_S, P_{S'}) \le \epsilon$ and $\mathsf{D}_\infty(P_{S'}, P_S) \le \epsilon$. Then,

$$\begin{aligned}
\mathsf{D}(P_S \parallel P_{S'}) + \mathsf{D}(P_{S'} \parallel P_S) &= \mathop{\mathbb{E}}_{x \sim P_S}\left[\log \frac{P_S(x)}{P_{S'}(x)}\right] + \mathop{\mathbb{E}}_{x \sim P_{S'}}\left[\log \frac{P_{S'}(x)}{P_S(x)}\right] \\
&= \mathop{\mathbb{E}}_{x \sim P_S}\left[\log \frac{P_S(x)}{P_{S'}(x)} + \log \frac{P_{S'}(x)}{P_S(x)}\right] + \mathop{\mathbb{E}}_{x \sim P_{S'} - P_S}\left[\log \frac{P_{S'}(x)}{P_S(x)}\right] \\
&= \epsilon \sum_x \left|P_{S'}(x) - P_S(x)\right| \qquad (\text{since } \mathsf{D}_\infty(P_S, P_{S'}), \mathsf{D}_\infty(P_{S'}, P_S) \le \epsilon) \\
&= \epsilon \sum_{P_S(x) > 0} P_S(x) \left|\frac{P_{S'}(x)}{P_S(x)} - 1\right|. \qquad (P_S(x) = 0 \text{ implies } P_{S'}(x) = 0)
\end{aligned}$$

Next, since both $\mathsf{D}_\infty(P_{S'}, P_S)$ and $\mathsf{D}_\infty(P_S, P_{S'})$ are bounded by ϵ, we have

$$\begin{aligned}
\left|\frac{P_{S'}(x)}{P_S(x)} - 1\right| &\le \max\left(e^\epsilon - 1, 1 - e^{-\epsilon}\right) \\
&\le e^\epsilon - 1.
\end{aligned}$$

Hence we can conclude that

$$\begin{aligned}
\mathsf{D}(P_S \parallel P_{S'}) + \mathsf{D}(P_{S'} \parallel P_S) &\le \epsilon(e^\epsilon - 1) \sum_{P_S(x) > 0} P_S(x) \\
&\le \epsilon(e^\epsilon - 1) \\
&\le 2\epsilon^2.
\end{aligned}$$

□

B.7 Proof of Lemma 7

Lemma 7. *Suppose $\|P_S - P_{S'}\|_1 \le \epsilon$ for all $S, S' \in \mathcal{Z}^m$ differing by exactly one point. For some $\mu \ge 0$, define the sample-dependent set of distributions as $\mathcal{Q}_{S,\mu} := \{Q \colon \|Q - P_S\|_1 \le \mu\}$, and the corresponding family to be $\mathscr{Q}_{m,\mu} = (\mathcal{Q}_{S,\mu})_{S \in \mathcal{Z}^m}$. Then $\mathscr{Q}_{m,\mu}$ is β-stable for $\beta = \frac{\epsilon}{2}$.*

Proof. For convenience, we do the computations using the total variation distance rather than ℓ_1.

Since $\|P_S - P_{S'}\|_{\mathrm{TV}} \le \frac{\epsilon}{2}$, there exists a coupling C_1 of P_S and $P_{S'}$ such that if $(X, X') \sim C_1$, we have $\mathbb{P}[X \ne X'] \le \frac{\epsilon}{2}$. Similarly, since $\|P_S - Q\|_{\mathrm{TV}} \le \frac{\mu}{2}$, there exists a coupling C_2 of P_S and Q such that if $(X, Y) \sim C_2$, we have $\mathbb{P}[X \ne Y] \le \frac{\mu}{2}$. Now construct a coupling C_3 as follows. First, sample $X \sim P_S$. Then, sample $X' \sim C_1$ conditioned on X, and independently, sample $Y \sim C_2$ conditioned on X. Set

$$Y' = \begin{cases} X' & \text{if } X = Y \\ Y & \text{otherwise.} \end{cases}$$

Let Q' be the distribution of Y'. Note that $\mathbb{P}[X = Y] \ge 1 - \frac{\mu}{2}$, so $\mathbb{P}[Y' = X'] \ge 1 - \frac{\mu}{2}$, which implies that $\|P_{S'} - Q'\|_{\mathrm{TV}} \le \frac{\mu}{2}$. Furthermore, by a union bound, we have

$$\mathbb{P}[Y' = Y] = \frac{\mu}{2} + \mathbb{P}[X' = X = Y] \ge \frac{\mu}{2} + 1 - (\mathbb{P}[X \ne Y] + \mathbb{P}[X \ne X']) \ge \frac{\mu}{2} + 1 - \left(\frac{\mu}{2} + \frac{\epsilon}{2}\right) = 1 - \frac{\epsilon}{2}.$$

So, $\|Q - Q'\|_{\mathrm{TV}} \le \frac{\epsilon}{2}$. □