A DATA PROCESSING INEQUALITY FRAMEWORK FOR GENERALIZATION BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a change-of-measure framework based on the data-processing inequality (DPI) for deriving novel high-probability generalization bounds for general learning algorithms. Compared to classical bounds based on Rademacher complexity or covering numbers, which often require sophisticated task-specific bounding techniques, our bounds are easy to evaluate and effective. Furthermore, our framework allows one to produce a family of generalization bounds by adopting different DPI-based change of measure inequalities and different assumptions on the loss function. We evaluate our bounds for examples of Gaussian mean estimation, linear regression, and multiclass classification with deep neural networks. Compared to state-of-the-art bounds from prior literature, our bounds provide significantly tighter guarantees at practically relevant sizes of the training dataset.

1 INTRODUCTION

Generalization bounds quantify how well a model's performance on training data translates to unseen data. A model's empirical risk is its average loss on the training sample, whereas its population risk is its expected loss on new observations from the same but unknown underlying distribution. Population risk cannot be calculated directly because of the unknown data distribution, and is commonly estimated using a held-out test set. Generalization bounds provide an alternative way to assess the population risk: they control the generalization gap, namely the difference between the population risk and the empirical risk, and thereby upper-bound population risk by the observed training loss plus a term bounding this gap. Under suitable assumptions, these guarantees hold with high probability over the training sample without requiring a separate test set, allowing all available observations to contribute to training. This is particularly valuable when labeled data are limited. Additionally, by linking empirical risk to population risk, generalization bounds provide statistical guarantees that can inform model selection and the analysis of learning algorithms (Bartlett & Mendelson, 2002).

We consider hypotheses (i.e., models) $w \in \mathcal{W} \subseteq \mathbb{R}^d$ and a loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z}$ is the sample space of data points. For any $w \in \mathcal{W}$, we define the population and empirical risks as $L(w) := \mathbb{E}_{Z\sim P_Z}[\ell(w, Z)]$ and $\hat{L}(s, w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$, respectively. The focus of this work is worst-case high-probability generalization bounds, which typically take the following form: for any $\delta \in (0, 1]$,

$$P_S\left(\sup_{w\in\mathcal{W}}\left(L(w) - \hat{L}(S, w)\right) \geq f\left(\mathcal{W}, \delta, n\right)\right) \leq \delta,$$

where $S = (Z_1, \ldots, Z_n) \sim P_Z^{\otimes n}$ is a training sample of n i.i.d. observations, $P_S := P_Z^{\otimes n}$ and $f\left(\mathcal{W}, \delta, n\right)$ is a function depending on the hypothesis class, the confidence parameter and also the data size. Their practical usefulness, however, is not entirely determined by asymptotic convergence rates (namely the decay rate of $f\left(\mathcal{W}, \delta, n\right)$ with respect to n) alone. At practically relevant sample sizes, constants and problem-dependent terms that quantify the size or complexity of the hypothesis space may dominate and render a bound vacuous; moreover, a bound provides limited practical insight if its right-hand side depends on quantities that cannot be easily computed or tightly estimated. We therefore focus on two complementary goals: numerical tightness at practically relevant sample sizes and direct evaluability. An informative bound should be tight at the available sample size and also be expressible in terms of directly computable quantities, without requiring a separate problem-specific relaxation of an intractable complexity measure.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are summarized as follows:

- To the best of our knowledge, we introduce the first framework based on the data processing inequality (DPI) for deriving worst-case, high-probability generalization bounds for general learning algorithms with continuous hypothesis spaces. Our framework allows one to produce a family of generalization bounds, by adopting different DPI-based change of measure inequalities, and different assumptions on the loss function. A more detailed discussion of this flexibility is provided in Section 3.
- Unlike Rademacher-complexity- or covering-number-based bounds, whose direct evaluation is often difficult because these complexity measures are intractable to compute exactly and require additional task-specific upper bounds, our bounds can be evaluated directly without introducing extra complexity measures.
- We empirically evaluate our bounds on Gaussian mean estimation, linear regression, and multiclass classification with deep neural networks (DNNs). Across the finite-sample regimes considered, our bounds are numerically tighter than the corresponding existing bounds from previous literature.

1.1 Related Works

One of the classic results is the Rademacher complexity bound proposed by Bartlett & Mendelson (2002). They introduced Rademacher complexity $\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}_{S\sim P_S,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]$ to quantify the complexity of the hypothesis class, where the σ_i independently take values in $\{-1,+1\}$ with equal probability. Although the bound applies in very general settings, a practical difficulty is the evaluation of $\mathfrak{R}_n(\mathcal{F})$. For rich hypothesis classes, neither the supremum nor the subsequent expectation over the Rademacher variables is generally available in a simple closed form. Consequently, obtaining an explicit guarantee typically requires a problem-specific upper bound on $\mathfrak{R}_n(\mathcal{F})$, derived from additional structural assumptions on the hypothesis class, the data distribution, or both. For example, Corollary 5 of Kakade et al. (2008) derives generalization bounds for norm-constrained linear predictors by upper-bounding $\mathfrak{R}_n(\mathcal{F})$. In their L_p/L_q formulation, the inputs are uniformly bounded in the p-norm and the predictor weights are bounded in the dual q-norm, where $1/p + 1/q = 1$. Similarly, Theorem 1 of Awasthi et al. (2020) provides a data-dependent upper bound on the empirical Rademacher complexity of linear predictors with p-norm-bounded weights. Although this result does not itself require uniformly bounded inputs, it still presupposes a norm-bounded hypothesis class. Finally, Corollary 10 of Maurer & Pontil (2021) derives a Rademacher-complexity-based generalization bound using concentration inequalities. Their linear regression specialization assumes finite sub-exponential Orlicz norms for both the inputs and outputs, together with a norm constraint on the predictor.

PAC-Bayes bounds (McAllester, 1998) are another related class of high-probability bounds. Such bounds control the population risk of hypotheses generated using stochastic learning algorithms, and are generally not applicable to deterministic learning algorithms. PAC-Bayesian bounds hold simultaneously for every posterior distribution ρ over $\mathcal{W}$, relative to a data-independent prior π, and control the population risk of a stochastic hypothesis through a bound of the form $\mathbb{E}_{W\sim\rho}[L(W)] \leq \mathbb{E}_{W\sim\rho}[\hat{L}(S,W)] + f(\mathrm{KL}(\rho\|\pi), n, \delta)$ (McAllester, 1998), where the expectation is taken over the entire data-dependent distribution ρ output by the learning algorithm. In particular, for a deterministic learning algorithm, the PAC-Bayes posterior reduces to a Dirac measure at the learned hypothesis, which has infinite KL divergence with respect to a continuous prior on a continuous hypothesis space (Foong et al., 2021, Appendix A), rendering the bound vacuous. Recent work attempts to extend PAC-Bayes bounds to handle deterministic algorithms (Leblanc & Germain, 2025), but its applicability is restricted to classification problems.

The DPI has emerged as a useful tool for generalization error analysis. Esposito et al. (2021) were, to the best of our knowledge, the first to leverage the DPI for generalization error analysis, deriving generalization bounds characterized by maximal leakage. Building on this perspective, Guan et al. (2025) developed a family of change-of-measure inequalities based on the DPI and applied them to derive generalization bounds for countable hypothesis spaces. In contrast, our bounds can cover uncountable hypothesis spaces arising in modern machine learning, such as DNNs with real-valued parameters. More recently, Liu et al. (2026) extended the use of the DPI to broader generalization settings, including PAC-Bayesian and information-theoretic bounds. Our work instead uses the DPI

to derive high-probability generalization bounds for general learning algorithms. To the best of our knowledge, our framework is the first DPI-based approach to provide high-probability generalization bounds for any general (deterministic or stochastic) learning algorithm with a continuous hypothesis space, which is arguably the most important use case for generalization bounds.

2 MAIN RESULTS

We state a few definitions and assumptions before presenting our main results. For $p \in [1, \infty]$, let $\|\cdot\|_p$ denote the p-norm on $\mathbb{R}^d$, where $\|w\|_p := \left(\sum_{j=1}^d |w_j|^p\right)^{1/p}$ for $1 \le p < \infty$, and $\|w\|_\infty := \max_{1\le j\le d} |w_j|$ for $p = \infty$.

Assumption 1. *($C_{\ell,p}$-Lipschitz Continuous Loss). The loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is $C_{\ell,p}$-Lipschitz continuous in its first argument with respect to the p-norm, for some $C_{\ell,p} > 0$. That is, for all $w, w' \in \mathcal{W}$ and all $z \in \mathcal{Z}$,*

$$|\ell(w, z) - \ell(w', z)| \le C_{\ell,p}\|w - w'\|_p.$$

Assumption 2. *(σ_ℓ-Sub-Gaussian Loss). For every fixed $w \in \mathcal{W}$, the loss $\ell(w, Z)$ satisfies*

$$\mathbb{E}_Z\left[\exp\left(t\left(\ell(w, Z) - \mathbb{E}_Z[\ell(w, Z)]\right)\right)\right] \le \exp\left(\frac{t^2\sigma_\ell^2}{2}\right), \quad \forall t \in \mathbb{R}.$$

With the above assumptions, we can show the following novel generalization error bound.

Theorem 1. *(DPI Generalization Bound for Sub-Gaussian Loss). Suppose that Assumption 1 and Assumption 2 hold. For $w \in \mathbb{R}^d$, define $\overline{W}(w) := \max\{\lceil \|w\|_p \rceil, 1\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over S, the following inequality holds simultaneously for all $w \in \mathbb{R}^d$ and all $\gamma \ge 1$:*

$$\left|L(w) - \hat{L}(S, w)\right| \le \sqrt{\frac{2\sigma_\ell^2\left[d\log\left(\overline{W}(w)\sqrt{\gamma n} + 1\right) + \log\left(\frac{4\overline{W}(w)^2}{\delta}\right)\right]}{n}} + \frac{2C_{\ell,p}}{\sqrt{\gamma n}}. \tag{1}$$

Remark 1. *Since Theorem 1 holds simultaneously for all $\gamma \ge 1$, the best choice of γ is obviously the one that minimizes the RHS of the inequality. However, one interesting choice of γ (albeit suboptimal in general) is $\gamma = C_{\ell,p}^2$. This choice results in the following bound: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over S, the following inequality holds simultaneously for all $w \in \mathbb{R}^d$:*

$$\left|L(w) - \hat{L}(S, w)\right| \le \sqrt{\frac{2\sigma_\ell^2\left[d\log\left(\overline{W}(w)C_{\ell,p}\sqrt{n} + 1\right) + \log\left(\frac{4\overline{W}(w)^2}{\delta}\right)\right]}{n}} + \frac{2}{\sqrt{n}}.$$

One interesting feature of this bound is its logarithmic dependence on the Lipschitz constant $C_{\ell,p}$. This is particularly advantageous when we apply the bound to DNNs, where the Lipschitz constant of the entire DNN depends on products of layer-wise Lipschitz constants, such as spectral norms of the weight operators, and therefore grows rapidly with network depth. We will demonstrate this through our numerical examples in Section 4.3.

We will show the proof sketch of Theorem 1 in Section 3, and its entire proof appears in Appendix A.2. Beyond the sub-Gaussian setting, we also derive the result for sub-exponential losses, extending the applicability of our bound to this regime. Section 4.2 provides a concrete linear-regression example in which the loss is sub-exponential but not sub-Gaussian. This extension further shows the flexibility of our framework in handling loss functions satisfying different distributional assumptions, and we discuss the source of this flexibility in detail in Section 3.

Assumption 3. *($(\nu_\ell^2, \alpha_\ell)$-Sub-Exponential Loss). For every fixed $w \in \mathcal{W}$, the loss $\ell(w, Z)$ satisfies*

$$\mathbb{E}_Z\left[\exp\left(t\big(\ell(w, Z) - \mathbb{E}_Z[\ell(w, Z)]\big)\right)\right] \le \exp\left(\frac{t^2\nu_\ell^2}{2}\right), \quad |t| < \frac{1}{\alpha_\ell}.$$

Theorem 2. *(DPI Generalization Bound for Sub-Exponential Loss). Suppose that Assumptions 1 and 3 hold. For $w \in \mathbb{R}^d$, define $\overline{W}(w) := \max\{\lceil \|w\|_p \rceil, 1\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, the following inequality holds simultaneously for all $w \in \mathbb{R}^d$ and all $\gamma \geq 1$:*

$$\left|L(w) - \hat{L}(S,w)\right| \leq \begin{cases} \sqrt{\frac{2\nu_\ell^2\left(\log\frac{1}{\delta}+\psi(w)\right)}{n}} + \frac{2C_{\ell,p}}{\sqrt{\gamma n}}, & \text{if } n \geq \frac{2\alpha_\ell^2}{\nu_\ell^2}\left(\psi(w)+\log\left(\frac{1}{\delta}\right)\right), \\ \frac{2\alpha_\ell\left(\log\frac{1}{\delta}+\psi(w)\right)}{n} + \frac{2C_{\ell,p}}{\sqrt{\gamma n}}, & \text{if } 0 < n < \frac{2\alpha_\ell^2}{\nu_\ell^2}\left(\psi(w)+\log\left(\frac{1}{\delta}\right)\right), \end{cases} \tag{2}$$

where $\psi(w) = d\log\left(\overline{W}(w)\sqrt{\gamma n}+1\right) + \log\left(4\overline{W}(w)^2\right)$.

The proof of Theorem 2 appears in Appendix A.3.

2.1 Discussion

We next examine how these bounds balance direct evaluability with finite-sample tightness.

We view simplicity as one advantage of our bound. The RHS of the bounds in Theorems 1 and 2 are straightforward to evaluate (the optimization over γ is a simple one-dimensional problem which can be solved numerically very easily). Additionally, unlike some other bounds (e.g., (Kakade et al., 2008; Awasthi et al., 2020; Maurer & Pontil, 2021)) our bound does not require a bounded hypothesis space. Instead, it depends on the norm of the selected hypothesis.

By choosing $\gamma = 1$, our bounds achieve the asymptotic rate $O\left(\sqrt{\frac{\log n}{n}}\right)$ with respect to the sample size n. Although Rademacher-complexity-based bounds can attain the slightly faster rate $O\left(\frac{1}{\sqrt{n}}\right)$ in some settings, asymptotic convergence rates alone do not determine the tightness of generalization guarantees at practically relevant sample sizes. In our experiments on Gaussian mean estimation (Section 4.1) and linear regression (Section 4.2), the proposed bounds are numerically tighter over the considered finite-sample regimes than the corresponding bounds obtained from Kakade et al. (2008) and Awasthi et al. (2020). These results show that our bounds can be tighter at practically relevant sample sizes despite their slower asymptotic rate, and our bounds also remain directly evaluable and require weaker assumptions on the data and hypothesis space.

3 The DPI framework and comparison with other bounds

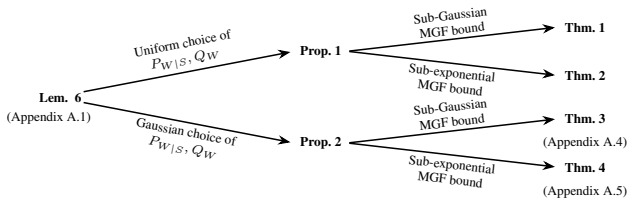


Figure 1: Proof roadmap and logical dependencies among the main results. We demonstrate the main technical ingredients used to derive the propositions from the lemma and the final theorems from the propositions.

Although our results take a similar form to the bounds obtained via Rademacher complexity, our proof technique, however, explores a very different path, which could be of interest by itself. In this section, we discuss the key ingredients of our proof and compare it with other proof techniques.

3.1 Proof sketch

We present an overview of the techniques used to prove Theorems 1 and 2. We also highlight the flexibility of the proposed approach and its potential to derive other related generalization bounds.

Our results bound the generalization gap simultaneously for all elements of the hypothesis space and therefore do not rely on any specific properties of the learning algorithm. In this sense they are similar to bounds found in the literature that are based on Rademacher complexity (e.g., Bartlett & Mendelson (2002),Kakade et al. (2008) and Awasthi et al. (2020)). However, the proof technique we use employs a change-of-measure argument that is more closely related to the PAC-Bayes setting (e.g., McAllester (1998), Ohnishi & Honorio (2021)).

As auxiliary devices for the proof, we introduce a conditional distribution $P_{W|S}$ and a "prior" Q_W on the hypothesis space. Contrary to the PAC-Bayes setting, where $P_{W|S}$ represents the learning algorithm and only Q_W can be chosen freely, in our approach neither the prior Q_W nor the conditional $P_{W|S}$ is a property of the analyzed learning scenario. Instead, both can be chosen freely, with each choice yielding a slightly different bound.

Core to our proof is a change-of-measure argument based on the DPI for Rényi divergence (Lemma 1), which we reproduce here: for all probability measures P, Q, all events E, and all $\alpha \in (1, \infty)$:

$$P(E) \leq Q(E)^{\frac{\alpha-1}{\alpha}} e^{\frac{\alpha-1}{\alpha} D_\alpha(P||Q)}. \tag{3}$$

We now briefly describe how we define an appropriate event E, auxiliary distributions $P_{W|S}$ and Q_W so that the results in Theorems 1 and 2 can be obtained. The detailed derivations are in Appendices A.2 and A.3.

Recall that we are interested in upper bounding the probability of the following event under the data distribution P_S:

$$A' := \left\{ \sup_{w \in \mathcal{W}} \left| L(w) - \hat{L}(S, w) \right| \geq f(\mathcal{W}, \delta, n) \right\}$$

where $f(\mathcal{W}, \delta, n)$ depends on the properties of the hypothesis space (or the hypothesis itself), the loss function, the confidence parameter, and the sample size.

The main technical argument goes through the following chain of inequalities

$$P_S(A') = \underbrace{P_{S,W}(A, D_S) \leq P_{S,W}(E) \leq Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}} e^{\frac{\alpha-1}{\alpha} D_\alpha(P_{S,W}||Q_{S,W})}}_{\text{Lemma 6}}.$$

where the second inequality is just an application of equation 3 (Lemma 1), resulting in Lemma 6 (Appendix A.1), which is the main technical device underlying all of our results. Here we choose $P_{S,W} := P_{W|S} P_S$ and $Q_{S,W} := Q_W P_S$ where the auxiliary distributions $P_{W|S}$ and Q_W are defined in Proposition 1 (Appendix A.1). Under the distribution $P_{S,W}$, the event A (defined in equation 4) is essentially the same as the event A', the event D_s (defined in equation 5) is an auxiliary event which allows us to pass from the distribution P_S to the distribution $P_{S,W}$, and the measurable event $E \subseteq \mathcal{S} \times \mathcal{W}$ is a superset of the joint event $\{A, D_s\}$. To obtain the final result, we need to calculate $D_\alpha(P_{S,W}||Q_{S,W})$ and upper bound $Q_{S,W}(E)$. Specifically, we choose $\alpha = \infty$ for our final bound as this is the best choice. Figure 1 (upper branch) sketches the roadmap going from Lemma 6 to Theorems 1 and 2.

Now we highlight a few aspects of the proof technique that can be explored further in future work:

- (Change-of-measure argument) In place of equation 3, it is in principle also possible to instead use similar change-of-measure lemmas based on other divergences such as the Hellinger p-divergence inequality (Guan et al., 2025, Lemma 2) or the Pearson chi-squared divergence inequality (Guan et al., 2025, Lemma 3). We consider an exploration of the resulting family of bounds as a promising avenue of future research that is opened up by this paper.
- (Choice of auxiliary distributions) Our framework does not prescribe particular forms for $P_{W|S}$ and Q_W. Any probability measures that have controllable divergence are applicable. The main results Theorems 1 and 2 are obtained by choosing $P_{W|S}$ and Q_W to be uniform distributions over a ball. In principle, we can choose other distributions with a suitable parametrization. To showcase the modularity of our proof technique, we illustrate an extension (where $P_{W|S}$ and Q_W are chosen to be Gaussian distributions) in Proposition 2 and Theorems 3 and 4 in Appendices A.4 and A.5, as illustrated in the lower branch in

Figure 1. Remark 2 in Appendix A.6 shows that Theorems 1 and 2 are tighter bounds than Theorems 3 and 4.

- (Assumption on loss functions) Finally, as is common with these types of bounds, we require an assumption that the loss is well-behaved. However, our approach is flexible in the sense that any assumption that implies a bound on the moment generating function of the generalization gap can be used. In this work, we illustrate the flexibility of our method by giving bounds relying on an assumption of sub-Gaussianity (Assumption 2 and Theorems 1 and 3) and on an assumption of sub-exponentiality (i.e., Assumption 3 and Theorems 2 and 4), also reflected in Figure 1. The same proof strategy can be adapted to other concentration regimes, such as Bernstein-type or sub-gamma conditions.

3.2 Comparison with Rademacher complexity-based proof

Comparing the failure events reveals how our construction avoids an explicit Rademacher complexity term. Our proof uses the set A defined as:

$$A := \left\{ s \in \mathcal{S} : \sup_{w \in \mathcal{W}} \left| L(w) - \hat{L}(s,w) \right| \geq \underbrace{2C_{\ell,p}\beta + \sqrt{\frac{\log(1/\delta)}{2n}}}_{\text{threshold}} \right\},$$

which is a failure event where the worst-case generalization gap exceeds a threshold determined by the Lipschitz constant $C_{\ell,p}$ and the tunable parameter β which is later specialized to $\frac{1}{\sqrt{n\gamma}}$. If we inspect Rademacher-complexity based proofs, we can identify a similar failure event, defined as

$$\left\{ s \in \mathcal{S} : \sup_{w \in \mathcal{W}} \left(L(w) - \hat{L}(s,w) \right) \geq \underbrace{\mathbb{E}_S \left[\sup_{w \in \mathcal{W}} \left(L(w) - \hat{L}(S,w) \right) \right] + \sqrt{\frac{\log(1/\delta)}{2n}}}_{\text{threshold}} \right\},$$

where the corresponding threshold is governed by the expected worst-case generalization gap and the confidence term. Then Rademacher complexity is introduced as the upper bound of the expected worst-case generalization gap (i.e., $\mathbb{E}_S \left[\sup_{w \in \mathcal{W}} \left(L(w) - \hat{L}(S,w) \right) \right] \leq 2\mathcal{R}_n(\mathcal{F})$), which leads to the classic Rademacher generalization error bound. However, for many modern hypothesis classes, $\mathcal{R}_n(\mathcal{F})$ is difficult to compute, often limiting the practical evaluability of the resulting bound. In comparison, we introduce $C_{\ell,p}\beta$ so that we can control the threshold of our failure event by tuning β without imposing additional assumptions. This design is particularly useful in modern learning settings where Rademacher complexity is difficult to estimate sharply. β can be chosen freely within $(0, 1]$ to balance the terms in the final bound and thereby determine its convergence rate.

4 Numerical experiment

Having presented our bounds and the proof strategy underlying them, we now evaluate the practical applicability and numerical tightness of our bounds across three learning settings: Gaussian mean estimation (Theorem 1), linear regression (Theorem 2), and multiclass classification with DNNs (Corollary 1)[1]. The three examples demonstrate that our bounds can be easily evaluated across different problem settings, while providing competitive theoretical guarantees compared with existing generalization bounds. We only include numerical results using Theorems 1 and 2 (based on uniform distribution), as Theorems 3 and 4 give looser bounds. However, as mentioned in Section 3.1, exploring other auxiliary distributions in the proof remains an interesting research direction.

4.1 Example 1: Gaussian mean estimation

We consider a Gaussian mean estimation problem with Euclidean loss, for which the sub-Gaussian condition admits an explicit characterization. This example is designed to compare our bounds with

[1]This corollary is presented in Appendix D.1. It applies our bound to a ramp-loss upper bound on the 0–1 loss, yielding a margin-based generalization guarantee for DNNs.

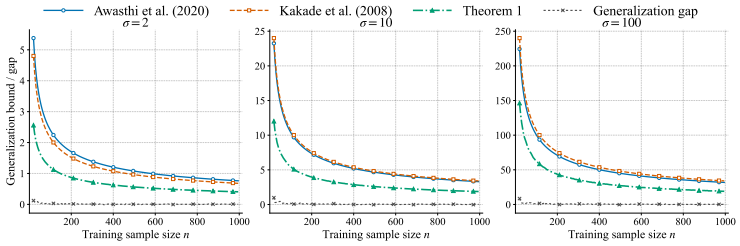


Figure 2: Comparison of generalization bounds for Gaussian mean estimation. We draw $Z^{(i)} \sim \mathcal{N}((1.5, 1.5)^\top, \sigma^2 I_2)$ and set $\delta = 0.05$. Generalization gap $:= L(\hat{w}_S) - \hat{L}(S, \hat{w}_S)$. For each n, we choose the γ that minimizes the bound. All results are averaged over 5 independent runs.

σ^2	$n_{\mathrm{A}}^{\times}$	$n_{\mathrm{K}}^{\times}$
2^2	3.95×10^{22}	5.92×10^{16}
10^2	5.91×10^{18}	1.97×10^{20}
100^2	8.97×10^{17}	1.33×10^{21}

Table 1: Crossover sample sizes for the Gaussian mean estimation example. We set $\delta = 0.05$, $\gamma = 1$ and $w = (1.5, 1.5)^\top$. The quantities $n_{\mathrm{A}}^{\times}$ and $n_{\mathrm{K}}^{\times}$ denote the real-valued solutions obtained by equating our bound with the bounds of Awasthi et al. (2020) and Kakade et al. (2008), respectively.

existing Rademacher complexity results and to highlight two practical advantages: (1) substantially tighter finite-sample guarantees; and (2) applicability to unbounded hypothesis and data spaces.

We consider a two-dimensional Gaussian mean-estimation problem. Let $Z \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2)^\top$ and $\Sigma = \mathrm{diag}(\sigma^2, \sigma^2)$. The mean vector μ is unknown to the learner, while Σ is known. The loss function is $\ell(w, z) := \|w - z\|_2$. We choose the projected sample-mean estimator $\hat{w}_S = \Pi_{\mathcal{W}}\big(\frac{1}{n}\sum_{i=1}^n Z^{(i)}\big)$ as the learning algorithm, where $\Pi_{\mathcal{W}}$ denotes Euclidean projection onto $\mathcal{W}$.

For comparison with the Rademacher complexity bounds of Awasthi et al. (2020) and Kakade et al. (2008), which require a prior norm constraint on the hypothesis class, we set $W_1 = 3\sigma$ and restrict their hypothesis space to $\mathcal{W} = \big\{w \in \mathbb{R}^2 : \|w\|_1 \le 3\sigma\big\}$. This restriction is imposed only to instantiate the existing bounds from previous literature; our bound does not require a bounded hypothesis space and continues to hold over $\mathbb{R}^2$. In this setting, the Lipschitz constant and sub-Gaussian parameter can be easily calculated as $C_{\ell,1} = 1$ and $\sigma_\ell = \sigma$, respectively. We also provide the specialized form of our bound (Theorem 1) and the two specialized existing bounds from previous literature in Appendix B. As shown in Figure 2, our bound is consistently tighter than the prior bounds from the literature across $\sigma^2 \in \{2^2, 10^2, 100^2\}$, including the high-variance regime $\sigma^2 = 100^2$ for sample sizes up to 8.97×10^{17}. As discussed in Section 2.1, the scaling behavior of our bound w.r.t n is of the form $\sqrt{\frac{\log n}{n}}$, which is asymptotically worse than the other two bounds in Kakade et al. (2008) and Awasthi et al. (2020), which scale as $\sqrt{\frac{1}{n}}$ for this particular example. However, as we show in Table 1, for data of practically relevant size, our bound is tighter than both bounds. In Table 1, we calculate the value of n (called the crossover point), above which the bounds in Kakade et al. (2008) and Awasthi et al. (2020) become tighter than Theorem 1. It can be seen that the other two bounds become tighter only at extremely large sample sizes. This shows that our bound has a substantial finite-sample advantage despite the slower asymptotic convergence rate.

4.2 Example 2: Linear Regression

To illustrate the usefulness of our sub-exponential generalization bound in Theorem 2 beyond the sub-Gaussian regime, we consider a linear regression example with absolute loss and Laplace noise, for which the centered loss is sub-exponential but not sub-Gaussian. This example also allows us to examine whether the finite-sample advantage of our bounds extends beyond sub-Gaussian losses.

Here we consider a two-dimensional linear regression problem. Let $X = (X_1, X_2)^\top$, where X_1 and X_2 are drawn from a uniform distribution and a truncated Gaussian distribution, respectively.

$$X_1 \sim \mathrm{Unif}(\mu - 2s, \mu + 2s) \text{ and } X_2 \sim \mathcal{N}(\mu, \sigma^2) \text{ conditioned on } |X_2 - \mu| \le 2s,$$

with $(s, \sigma^2) \in \{(1, 1), (\sqrt{10}, 10), (10, 98)\}$. The response is generated according to $Y = (w^\star)^\top X + \varepsilon$, where $w^\star = (0.5, -0.4)^\top$ and $\varepsilon \sim \mathrm{Laplace}(0, \frac{1}{\sqrt{2}})$ independently of X. We consider the hypothesis space $\mathcal{W} = \{w \in \mathbb{R}^2 : \|w\|_1 \le 1\}$, and use the absolute-error loss $\ell(w, Z) = |(w - w^\star)^\top X - \varepsilon|$. We use the empirical risk minimizer $\hat{w} = \arg\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n |w^\top X_i - Y_i|$ as the learning algorithm. In the linear regression setting, the parameters appearing in our bound are explicitly calculated as $C_{\ell,1} = \mu + 2s$, $\nu_\ell^2 = 16s^2 + 4$, and $\alpha_\ell = 1$. The specialized forms of our bound and the two existing bounds from prior work are provided in Appendix C.

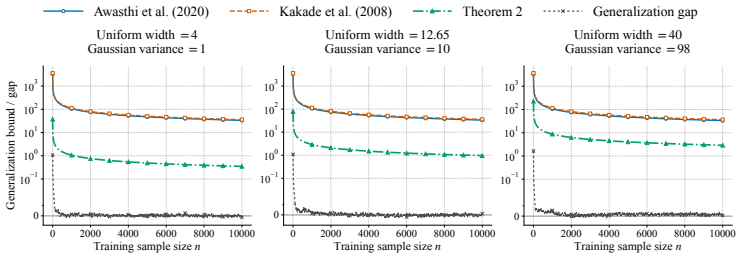


Figure 3: Comparison of generalization bounds for linear regression. We set $\mu = 10^3$ and $\delta = 0.05$. For each n, we choose the γ that minimizes the bound. All results are averaged over 5 independent runs.

Figure 3 shows that our bound yields tighter guarantees than the two existing bounds in the linear regression setting. More importantly, the example illustrates the flexibility of our framework in accommodating different distributional assumptions on the loss.

4.3 Example 3: Multiclass classification using DNNs

We examine the numerical tightness of our bound and its sensitivity to network size using DNNs on MNIST with $60,000$ training samples.[2] We compare it with existing bounds developed specifically for DNNs, including the third bound in Long & Sedghi (2020, Theorem 3.1), the first bound in Graf et al. (2022, Theorem 3.5), Galanti et al. (2023, Theorem 3.2), and Ledent et al. (2026, Theorem G.10). Note that all the bounds above are derived by directly upper-bounding either Rademacher complexities or covering numbers. However, our bound relies on neither, offering a distinct perspective for deriving generalization guarantees for DNNs. The PAC-Bayes bounds are not directly applicable to the scenarios where DNNs are trained directly using SGD-like algorithms without introducing the surrogate distributions over weights for calculating its KL divergence term.

Network architecture. We use convolutional networks with L convolutional layers and H channels per layer, denoted by CNN-L-H. We omit all bias parameters to match the bias-free formulation of our Lipschitz analysis and express $C_{\ell,2}$ through layer-wise operator norms. Under the imposed spectral-norm constraint, we have the following explicit expression: $C_{\ell,2} = 2B_x \times \lambda^L \sqrt{1 + \sum_{k=1}^L H_k^{\mathrm{out}} W_k^{\mathrm{out}}}$, where B_x denotes an upper bound on the 2-norm of the feature vector, L denotes the number of convolutional layers, and H_k^{out} and W_k^{out} denote the spatial height and width of the output feature map of k-th convolutional layer. The quantity λ is the common upper bound imposed on the spectral norm of each layer. The total number of trainable parameters is $4(L-1)H^2 + 494H$.

Optimization process. The training objective combines cross-entropy with a spectral-norm penalty computed using the power-iteration surrogate described in Appendix D.2. During training, we constrain the spectral norm of the weight operator associated with each trainable layer to be

[2]Detailed calculation of the Lipschitz constant $C_{\ell,2}$, the optimization process, and also additional experimental results on the CIFAR-10 dataset are provided in Appendix D.

at most λ. At the end of each epoch, when a layer exceeds the prescribed spectral norm bound, we rescale its weight tensor so that the resulting spectral norm is constrained to be at most λ. We set $\lambda = 10$ in all experiments on MNIST. The spectral norm constraint enables an explicit evaluation of the Lipschitz constant ($C_{\ell,2}$) appearing in our bound, and it facilitates a fair comparison with existing generalization bounds on DNNs, for which constraining or penalizing the spectral norms of the model during training is a standard experimental practice; see, for example, the experimental setups of Long & Sedghi (2020), Graf et al. (2022), and Ledent et al. (2026).

Experiment. For the bound in Ledent et al. (2026), we set $p_\ell = 1$ for all layers, while for Long & Sedghi (2020, Theorem 3.1) we set the unspecified universal constant to $C = 1$, as done by Galanti et al. (2023). We evaluate our bound (i.e., Corollary 1) by choosing the optimal γ that minimizes it. As shown in Figure 4, our bound shows the tightest guarantee under highly overparameterized CNNs with large widths and depths. In particular, the results shown in Figure 4 indicate that increasing the depth of CNNs leads to rapid growth in all evaluated bounds, but our bound exhibits a substantially slower growth rate than the existing bounds. We also investigate the behavior of our bounds by varying the strength of the spectral norm constraint, which is shown in Figure 5 in Appendix D.2.

Although all the existing bounds, including our bound, are vacuous in this regime, our bound is significantly tighter than others and scales more favorably with increasing network complexity. Additionally, our bound requires fewer training samples than existing bounds to become non-vacuous. For example, for the CNN-3-100 configuration, holding the parameters and empirical data statistics fixed, our bound requires 1.5×10^6 training samples to become non-vacuous, whereas the evaluated bounds of Ledent et al. (2026), Galanti et al. (2023), Graf et al. (2022), and Long & Sedghi (2020) require at least 3.7×10^{19}, 7.8×10^{14}, 1.4×10^{24}, and 3.2×10^{20} samples, respectively.

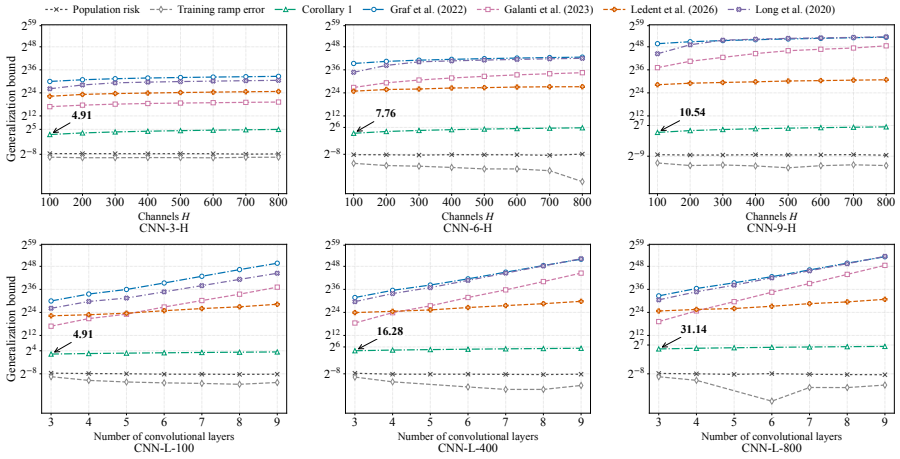


Figure 4: Varying depth L and channels H for CNN-L-H architectures trained on MNIST. Each annotation marks the minimum of our bound in that panel.

5 Limitation and Future Work

As we discussed in Section 2.1, the convergence rate of our bound ($\sqrt{\frac{\log n}{n}}$) is not optimal in general, compared to some Rademacher-complexity based bounds, although for data sizes of practical relevance, our bound is still significantly tighter than these bounds (as shown in Section 4). It is still of theoretical interest to investigate whether an improved analysis within our framework could lead to bounds with optimal convergence rate. In the experiments with DNNs, although our bound is the tightest among all bounds, it seems that in order to obtain a non-vacuous and practically useful bound, a large training dataset would be required. For future work, we will further exploit the flexibility of the DPI framework to derive tighter generalization guarantees using alternative divergence measures and explore its applicability beyond traditional learning settings, such as error bounds for importance sampling.

6 AI USE STATEMENT

We used generative AI tools to polish the manuscript, brainstorm data distributions for the linear regression example (Section 4.2), assist with code for the DNN experiments (Section 4.3), and clarify parts of the mathematical exposition in Section 4.3. We did not use these tools to prepare the literature discussion or derive any of the main theorems. All AI-assisted work was reviewed by the authors. One author tested and verified the AI-generated code for correctness, and all authors reviewed the AI-assisted text. We take full responsibility for the final manuscript, including all text, claims, proofs, and accompanying artifacts.

7 REPRODUCIBILITY STATEMENT

We provide the source code for the DNN experiments (Section 4.3) in the supplementary material. All assumptions are explicitly stated in the main text and appendix, and complete proofs of all theorems and the corollary are provided in the appendix.

REFERENCES

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the rademacher complexity of linear hypothesis sets. 2020. URL https://arxiv.org/abs/2007.11045.

Peter Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8): 4986–5004, 2021. doi: 10.1109/TIT.2021.3085190.

Andrew Y. K. Foong, Wessel P. Bruinsma, David R. Burt, and Richard E. Turner. How tight can PAC-Bayes be in the small data regime? In *Advances in Neural Information Processing Systems*, volume 34, pp. 4093–4105, 2021.

Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for sparse neural networks. *Advances in Neural Information Processing Systems*, 36:42482–42501, 2023.

Manuel Gil. *On Rényi divergence measures for continuous alphabet sources*. PhD thesis, Queen's University at Kingston, 2011.

Jake Gipple. The volume of n-balls. *Rose-Hulman Undergraduate Mathematics Journal*, 15(1), 2014. URL https://scholar.rose-hulman.edu/rhumj/vol15/iss1/14. Article 14.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Florian Graf, Sebastian Zeng, Bastian Rieck, Marc Niethammer, and Roland Kwitt. On measuring excess capacity in neural networks. *Advances in Neural Information Processing Systems*, 35: 10164–10178, 2022.

Muhan Guan, Farhad Farokhi, and Jingge Zhu. A dpi-pac-bayesian framework for generalization bounds. In *2025 IEEE Information Theory Workshop (ITW)*, pp. 1–6, 2025. doi: 10.1109/ITW62417.2025.11240271.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Benjamin Leblanc and Pascal Germain. A framework for bounding deterministic risk with pac-bayes: Applications to majority votes. *arXiv preprint arXiv:2510.25569*, 2025.

Antoine Ledent, Rodrigo Alves, and Yunwen Lei. Generalization bounds for rank-sparse neural networks. *Advances in Neural Information Processing Systems*, 38:147927–147996, 2026.

Yanxiao Liu, Yijun Fan, and Deniz Gündüz. Tighter information-theoretic generalization bounds via a novel class of change of measure inequalities, 2026. URL https://arxiv.org/abs/2602.07999.

Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1e_FpNFDr.

Andreas Maurer and Massimiliano Pontil. Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34:7588–7597, 2021.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 230–234, 1998.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Yuki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. In *International conference on artificial intelligence and statistics*, pp. 1711–1719. PMLR, 2021.

Malempati Madhusudana Rao. *Measure Theory and Integration*. Pure and Applied Mathematics. Marcel Dekker, New York Basel, 2 edition, 2004.

Sahil Singla and Soheil Feizi. Fantastic four: Differentiable bounds on singular values of convolution layers. *arXiv preprint arXiv:1911.10258*, 2019.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

APPENDIX

A Proofs **12**

A.1 Preliminary Definitions and Auxiliary Results 12

A.2 Proof of Theorem 1 21

A.3 Proof of Theorem 2 21

A.4 Theorem 3 and its Proof 22

A.5 Theorem 4 and its Proof 23

A.6 Comparison of Uniform (Theorems 1 and 2) and Gaussian Choices (Theorems 3 and 4) 25

B Specialized bounds for Gaussian mean estimation **26**

C Specialized Bounds for Linear Regression **27**

D Additional Details for Multiclass Classification with DNNs **30**

D.1 Lipschitz Analysis for Convolutional Neural Networks 30

D.2 Experiment details on MNIST 35

D.3 Additional experimental results on CIFAR-10 35

A PROOFS

A.1 PRELIMINARY DEFINITIONS AND AUXILIARY RESULTS

Definition 1 introduces the notation used in the proof.

Definition 1. *For any $W_* \geq 1$, define the p-norm ball $B_{W_*,p} := \left\{w \in \mathbb{R}^d : \|w\|_p \leq W_*\right\}$. For a fixed sample s, we define a worst-case hypothesis over $B_{W_*,p}$ by $\mu_p(s) \in \arg\max_{w \in B_{W_*,p}} \left|L(w) - \hat{L}(s,w)\right|$, and denote the corresponding worst-case generalization gap by $\Delta(s) := \left|L(\mu_p(s)) - \hat{L}(s,\mu_p(s))\right|$.*

For any $0 < \beta \leq 1$, $0 < \delta < 1$, $n > 0$, $p \in [1,\infty]$ and Lipschitz constant $C_{\ell,p} > 0$, we define three sets below:

$$E := \left\{(s,w) \in \mathcal{S} \times \mathcal{W} : \left|L(w) - \hat{L}(s,w)\right| \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\},$$

$$A := \left\{s \in \mathcal{S} : \Delta(s) - 2C_{\ell,p}\beta \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\}, \tag{4}$$

and for any fixed s, we have

$$D_s := \left\{w \in \mathbb{R}^d : \|w - \mu_p(s)\|_p \leq \beta\right\}. \tag{5}$$

For any fixed w and $t \in \mathbb{R}$, define the moment-generating function (MGF) of the centered loss $\ell(w,Z) - L(w)$ by

$$M_w(t) := \mathbb{E}_{Z \sim P_Z}\left[\exp\left(t(\ell(w,Z) - L(w))\right)\right].$$

We write $P \ll Q$ if P is absolutely continuous with respect to Q, i.e., if every Q-null set is also a P-null set. Given probability measures P and Q, the Rényi divergence from P to Q is defined as $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int\left(\frac{dP}{dQ}\right)^\alpha dQ$, if $P \ll Q$, where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q. If $P \not\ll Q$, we use the convention $D_\alpha(P\|Q) = \infty$.

Lemma 1. *(Change of Measure with Rényi Divergence) (Guan et al., 2025). Let $(\Omega, \mathcal{F}, P)$, $(\Omega, \mathcal{F}, Q)$ be two probability spaces. For any $\alpha > 1$ and any event $E \in \mathcal{F}$, we have the bound*

$$P(E) \le Q(E)^{\frac{\alpha-1}{\alpha}} e^{\frac{\alpha-1}{\alpha} D_\alpha(P||Q)}.$$

Lemma 2. *Let Assumption 2 hold. Let $\delta \in (0,1)$, and let Q_W be a probability measure on $\mathcal{W}$. Then,*

$$\inf_{t\in\mathbb{R}} \left\{ \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n \right] \right\} \le 2\delta^{\frac{1}{4\sigma_\ell^2}}.$$

Proof. Applying the sub-Gaussian MGF bound in Assumption 2 and optimizing over $t \in \mathbb{R}$, we obtain

$$\begin{aligned} &\inf_{t\in\mathbb{R}} \left\{ \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[\big(M_W(t)\big)^n + \big(M_W(-t)\big)^n \right] \right\} \\ &\quad \le 2 \inf_{t\in\mathbb{R}} \left\{ \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} + \frac{nt^2\sigma_\ell^2}{2} \right) \right\} = 2\delta^{\frac{1}{4\sigma_\ell^2}}, \end{aligned}$$

where the last equality follows by choosing $t^* = \sigma_\ell^{-2}\sqrt{\log(1/\delta)/(2n)}$. □

Lemma 3. *Let Assumption 3 hold. Let $\delta \in (0,1)$, and let Q_W be a probability measure on $\mathcal{W}$. Then,*

$$\begin{aligned} &\inf_{|t|<\frac{1}{\alpha_\ell}} \left\{ \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n \right] \right\} \\ &\le \begin{cases} 2\exp\left(-\frac{\log(1/\delta)}{4\nu_\ell^2} \right), & n \ge \frac{\alpha_\ell^2 \log(1/\delta)}{2\nu_\ell^4}, \\ 2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell} \right), & 0 < n < \frac{\alpha_\ell^2 \log(1/\delta)}{2\nu_\ell^4}, \end{cases} \end{aligned}$$

Proof. By Assumption 3, we have

$$\begin{aligned} &\inf_{|t|<\frac{1}{\alpha_\ell}} \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n \right] \right) \\ &\le 2 \inf_{|t|<\frac{1}{\alpha_\ell}} \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[\exp\left(\frac{nt^2\nu_\ell^2}{2} \right) \right] \right) \\ &= 2 \inf_{|t|<\frac{1}{\alpha_\ell}} \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} + \frac{nt^2\nu_\ell^2}{2} \right), \end{aligned} \tag{6}$$

Since the exponential function is increasing, it suffices to evaluate

$$\inf_{|t|<\frac{1}{\alpha_\ell}} \phi(t) = \inf_{|t|<\frac{1}{\alpha_\ell}} \left(-t\sqrt{\frac{n\log(1/\delta)}{2}} + \frac{nt^2\nu_\ell^2}{2} \right).$$

The unconstrained minimizer is $t_0 := \frac{1}{\nu_\ell^2}\sqrt{\frac{\log(1/\delta)}{2n}}$. If $t_0 < 1/\alpha_\ell$, then the infimum is attained at $t^\star = t_0$. Otherwise, if $t_0 \ge 1/\alpha_\ell$, the infimum is not attained on the admissible interval. Instead, taking the limit as $t \uparrow 1/\alpha_\ell$ yields $\inf_{|t|<1/\alpha_\ell} \phi(t) = \lim_{t\to 1/\alpha_\ell} \phi(t) = -\frac{1}{\alpha_\ell}\sqrt{\frac{n\log(1/\delta)}{2}} + \frac{n\nu_\ell^2}{2\alpha_\ell^2}$. Combining this infimum evaluation with equation 6, we obtain

$$\begin{aligned} &\inf_{|t|<1/\alpha_\ell} \left\{ \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n \right] \right\} \\ &\le \begin{cases} 2\exp\left(-\frac{\log(1/\delta)}{4\nu_\ell^2} \right), & n \ge \frac{\alpha_\ell^2 \log(1/\delta)}{2\nu_\ell^4}, \\ 2\exp\left(\frac{n\nu_\ell^2}{2\alpha_\ell^2} - \frac{n}{\alpha_\ell}\sqrt{\frac{\log(1/\delta)}{2n}} \right) \overset{(a)}{\le} 2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell} \right), & 0 < n < \frac{\alpha_\ell^2 \log(1/\delta)}{2\nu_\ell^4}, \end{cases} \end{aligned}$$

where the step (a) follows from $n < \alpha_\ell^2 \log(1/\delta)/(2\nu_\ell^4)$, which implies

$$\frac{n\nu_\ell^2}{2\alpha_\ell^2} \leq \frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\,\alpha_\ell}.$$

□

Lemma 4. *Suppose Assumption 1 holds. Then, for any sample $S = (Z_1, \ldots, Z_n)$ and any $w, w' \in \mathcal{W}$, we have*

$$|L(w') - \hat{L}(S, w')| \leq |L(w) - \hat{L}(S, w)| + 2C_{\ell,p}\|w - w'\|_p.$$

Proof. We have

$$\begin{aligned} |L(w) - L(w')| &= |\mathbb{E}_Z(\ell(w, Z) - \ell(w', Z))| \\ &\overset{(a)}{\leq} \mathbb{E}_Z\,|(\ell(w, Z) - \ell(w', Z))| \\ &\overset{(b)}{\leq} C_{\ell,p}\|w - w'\|_p, \end{aligned} \tag{7}$$

where the step (a) inequality is by Jensen's inequality and the step (b) is by Assumption 1. Similarly, we have

$$\left|\hat{L}\,(S, w) - \hat{L}\,(S, w')\right| \leq C_{\ell,p}\|w - w'\|_p. \tag{8}$$

Then we have

$$\begin{aligned} \left|L(w) - \hat{L}\,(S, w) - L(w') + \hat{L}\,(S, w')\right| &= \left|(L(w) - L(w')) + \left(\hat{L}(S, w') - \hat{L}(S, w)\right)\right| \\ &\overset{(a)}{\leq} |(L(w) - L(w'))| + \left|\left(\hat{L}(S, w') - \hat{L}(S, w)\right)\right| \qquad (9) \\ &\overset{(b)}{\leq} 2C_{\ell,p}\|w - w'\|_p, \qquad (10) \end{aligned}$$

where the step (a) is by the triangle inequality, and the step (b) is by combining equation 7 with equation 8.

Then by the triangle inequality $|a| \leq |b| + |a - b|$, we get

$$\begin{aligned} |g_S(w')| &\overset{(a)}{\leq} |g_S(w)| + |g_S(w') - g_S(w)| \qquad (11) \\ &\overset{(b)}{\leq} |g_S(w)| + 2C_{\ell,p}\|w - w'\|_p, \qquad (12) \end{aligned}$$

where $g_S(\cdot) := L(\cdot) - \hat{L}(S, \cdot)$, the step (a) is by the triangle inequality, and the step (b) is by equation 10. □

Similar arguments have been used in deriving generalization bounds for DNNs (e.g., Bartlett et al. (2017); Galanti et al. (2023); Ledent et al. (2026)). For completeness, we state the corresponding result in our setting and provide a detailed proof below.

Lemma 5. *Under the notation of Definition 1, let $f : [0, \infty) \times (0, 1) \to [0, \infty)$ be a nonnegative function. Suppose that, for every $W_* \geq 1$ and every $\delta' \in (0, 1)$,*

$$P_S\left\{\forall w \in B_{W_*,p}, \left|L(w) - \hat{L}(S, w)\right| \leq f\,(W_*, \delta')\right\} \geq 1 - \delta',$$

then we have for all $\delta \in (0, 1)$,

$$P_S\left\{\forall w \in \mathbb{R}^d, \left|L(w) - \hat{L}(S, w)\right| \leq f\left(\max\{\lceil\|w\|_p\rceil, 1\}, \frac{\delta/2}{\max\{\lceil\|w\|_p\rceil, 1\}^2}\right)\right\} \geq 1 - \delta.$$

Proof. For each $W_* \in \mathbb{N}$, let $\delta_{W_*} \in (0, 1)$ be a confidence parameter, whose dependence on W_* will be specified later. Then, for every $W_* \in \mathbb{N}$, we define

$$\mathcal{B}_{W_*,p} = \left\{S : \forall w \in B_{W_*,p}, \left|L(w) - \hat{L}(S, w)\right| \leq f\,(W_*, \delta_{W_*})\right\},$$

and its complement set

$$\mathcal{B}^c_{W_*,p} = \left\{ S : \exists w \in B_{W_*,p}, \left| L(w) - \hat{L}(S,w) \right| > f\left(W_*, \delta_{W_*}\right) \right\}.$$

Then for every $W_* \in \mathbb{N}$, we have $P_S(\mathcal{B}_{W_*,p}) \geq 1 - \delta_{W_*}$, and also $P_S(\mathcal{B}^c_{W_*,p}) \leq \delta_{W_*}$.

Therefore, we have

$$P_S\left\{\left(\bigcap_{W_*\in\mathbb{N}} \mathcal{B}_{W_*,p}\right)^c\right\} \overset{(a)}{=} P_S\left\{\bigcup_{W_*\in\mathbb{N}} \mathcal{B}^c_{W_*,p}\right\} \overset{(b)}{\leq} \sum_{W_*\in\mathbb{N}} P_S\left\{\mathcal{B}^c_{W_*,p}\right\} \leq \sum_{W_*=1}^{\infty} \delta_{W_*}, \tag{13}$$

where the step (a) is due to De Morgan's Laws (i.e., $\bigcup_{W_*\in\mathbb{N}} \mathcal{B}^c_{W_*,p} = \left(\bigcap_{W_*\in\mathbb{N}} \mathcal{B}_{W_*,p}\right)^c$), and the step (b) is due to the union bound.

Taking complements in equation 13, we obtain

$$P_S\left\{\bigcap_{W_*\in\mathbb{N}} \mathcal{B}_{W_*,p}\right\} \geq 1 - \sum_{W_*=1}^{\infty} \delta_{W_*},$$

which is equivalent to

$$P_S\left\{\forall W_* \in \mathbb{N}, \forall w \in B_{W_*,p}, \left| L(w) - \hat{L}(S,w) \right| \leq f\left(W_*, \delta_{W_*}\right)\right\} \geq 1 - \sum_{W_*=1}^{\infty} \delta_{W_*}. \tag{14}$$

By choosing each of the δ_{W_*} in equation 14 as $\delta_{W_*} = \frac{\delta/2}{{W_*}^2}$, then we have

$$\sum_{W_*=1}^{\infty} \delta_{W_*} = \sum_{W_*=1}^{\infty} \frac{\delta/2}{{W_*}^2} \leq \frac{\delta}{2}\left(1 + \int_1^{\infty} \frac{1}{x^2} dx\right) = \delta.$$

Therefore, we obtain

$$P_S\left\{\forall W_* \in \mathbb{N}, \forall w \in B_{W_*,p}, \left| L(w) - \hat{L}(S,w) \right| \leq f\left(W_*, \frac{\delta/2}{{W_*}^2}\right)\right\} \geq 1 - \delta. \tag{15}$$

Since equation 15 holds simultaneously for all $W_* \in \mathbb{N}$ and $w \in B_{W_*,p}$ with probability at least $1-\delta$, we then apply the bound with $W_* = \max\{\lceil \|w\|_p \rceil, 1\}$ for each w with $\|w\|_p$, we have

$$P_S\left\{\forall w \in \mathbb{R}^d, \left| L(w) - \hat{L}(S,w) \right| \leq f\left(\max\{\lceil \|w\|_p \rceil, 1\}, \frac{\delta/2}{\max\{\lceil \|w\|_p \rceil, 1\}^2}\right)\right\} \geq 1 - \delta,$$

which completes the proof. □

Lemma 6. *Follow definition 1 and suppose that Assumption 1 holds. Let $0 < \beta \leq 1$, $\alpha > 1$ and $t \in \mathbb{R}$. Define joint probability measures $P_{S,W} := P_S P_{W|S}$ and $Q_{S,W} := P_S Q_W$, where Q_W is a probability measure over $\mathcal{W}$ and $P_{W|S}$ is a conditional probability over $\mathcal{W}$ given S. Assume that $P_{W|S} \ll Q_W$ for P_S-almost every s, then we have*

$$\begin{aligned} &P_{S,W}\left(S \in A,\, W \in D_S\right) \\ &\leq \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right) \mathbb{E}_{Q_W}\left[M_W(t)^n + M_W(-t)^n\right]\right)^{\frac{\alpha-1}{\alpha}} \\ &\quad \times \left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)}\right]\right)^{\frac{1}{\alpha}}. \end{aligned} \tag{16}$$

Proof. We first derive the lower bound of $P_{S,W}(E)$.

$$\begin{aligned} P_{S,W}(E) &= P_{S,W}\left(\left\{\left| L(W) - \hat{L}(S,W) \right| \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\}\right) \\ &\overset{(a)}{\geq} P_{S,W}\left(\left\{\Delta(S) - 2C_{\ell,p}\left\|W - \mu_p(S)\right\|_p \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\}\right) \qquad (17) \\ &\overset{(b)}{\geq} P_{S,W}\left(\{W \in D_S,\, S \in A\}\right), \qquad (18) \end{aligned}$$

where (a) is due to Lemma 4, as $\mu_p(s) \in \mathcal{W}$ for any s, and (b) is due to set inclusion, as any (S, W) pair in the set of equation 18 is contained within the set of equation 17.

Therefore, by Lemma 1, we have

$$\begin{aligned} P_{S,W}\left(W \in D_S,\, S \in A\right) &\leq P_{S,W}(E) \\ &\leq Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}} e^{\frac{\alpha-1}{\alpha} D_\alpha(P_{S,W}||Q_{S,W})}. \end{aligned} \tag{19}$$

The right-hand side of equation 19 consists of an event-probability term $Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}}$ and a Rényi-divergence term. We bound these two factors separately, beginning with $Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}}$.

For every fixed w, we define

$$E_w := \left\{ s \in \mathcal{S} : \left| L(w) - \hat{L}(s,w) \right| \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right\},$$

and $X_i := \ell(w, Z_i) - \mathbb{E}_{Z\sim P_Z}[\ell(w,Z)]$; thus, $\frac{1}{n}\sum_{i=1}^n X_i = \hat{L}(S,w) - L(w)$.

Therefore, for any $\epsilon > 0$ and $t > 0$, we have

$$\begin{aligned} P_S\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) &= \mathbb{P}_{S\sim P_S}\left(e^{t\sum_{i=1}^n X_i} \geq e^{tn\epsilon}\right) \\ &\overset{(a)}{\leq} \frac{\mathbb{E}_{S\sim P_S}\left[e^{t\left(\sum_{i=1}^n X_i\right)}\right]}{e^{tn\epsilon}} \\ &= e^{-tn\epsilon}\prod_{i=1}^n \mathbb{E}_{Z\sim P_Z}\left[e^{tX_i}\right] \\ &= e^{-tn\epsilon}\left(M_w(t)\right)^n, \end{aligned}$$

where the step (a) follows from Markov's inequality.

Repeating the argument for $-\frac{1}{n}\sum_{i=1}^n X_i$ and letting $\epsilon = \sqrt{\frac{\log(1/\delta)}{2n}}$, we obtain the bound

$$\begin{aligned} P_S(E_w) &= P_S\left\{\left|L(w) - \hat{L}(S,w)\right| \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right\} \\ &\leq \exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\left(M_w(t)^n + M_w(-t)^n\right). \end{aligned} \tag{20}$$

Consequently, we have

$$\begin{aligned} Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}} &= \left(\mathbb{E}_{W\sim Q_W}\left[P_S(E_W)\right]\right)^{\frac{\alpha-1}{\alpha}} \\ &\overset{(a)}{\leq} \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n\right]\right)^{\frac{\alpha-1}{\alpha}}, \end{aligned} \tag{21}$$

where the step (a) is due to equation 20.

Having bounded $Q_{S,W}(E)^{\frac{\alpha-1}{\alpha}}$ in equation 21, we now rewrite the remaining Rényi-divergence factor in terms of the conditional divergences $D_\alpha(P_{W|S=s}\|Q_W)$. Specifically, we have

$$\begin{aligned} e^{\frac{\alpha-1}{\alpha}D_\alpha(P_{S,W}||Q_{S,W})} &= \left(\int_{\mathcal{S}\times\mathcal{W}}\left(\frac{dP_{S,W}}{dQ_{S,W}}(s,w)\right)^\alpha Q_{S,W}(ds,dw)\right)^{\frac{1}{\alpha}} \\ &\overset{(a)}{=} \left(\int_{\mathcal{S}}\int_{\mathcal{W}}\left(\frac{dP_S}{dQ_S}(s)\frac{dP_{W|S=s}}{dQ_W}(w)\right)^\alpha Q_W(dw)P_S(ds)\right)^{\frac{1}{\alpha}} \\ &\overset{(b)}{=} \left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)}\right]\right)^{\frac{1}{\alpha}}, \end{aligned} \tag{22}$$

where the step (a) follows from Proposition 1 in Section 6.3 of Rao (2004) and the step (b) follows from the definition of the Rényi divergence, together with $P_S = Q_S$ by construction:

$$\int_{\mathcal{W}} \left(\frac{dP_S}{dQ_S}(s) \frac{dP_{W|S=s}}{dQ_W}(w) \right)^{\alpha} Q_W(dw) = \exp\left((\alpha - 1) D_\alpha \left(P_{W|S=s} \| Q_W \right) \right),$$

Finally, substituting equation 21 and equation 22 into equation 19 yields equation 16. □

Proposition 1. *Follow definition 1 and suppose that Assumption 1 holds. For any $\gamma \geq 1$, choose $\beta = \frac{1}{\sqrt{\gamma n}}$, $P_{W|S} = \mathrm{Unif}\left(B_p\left(\mu_p(S), 1/\sqrt{\gamma n}\right)\right)$ and $Q_W = \mathrm{Unif}\left(B_p\left(0, W_* + 1/\sqrt{\gamma n}\right)\right)$ in Lemma 6. Then, we have*

$$\begin{aligned} &P_S\left(\sup_{w \in B_{W_*,p}} \left| L(w) - \hat{L}(S,w) \right| - \frac{2C_{\ell,p}}{\sqrt{\gamma n}} \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right) \\ &\leq \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}} \right) \mathbb{E}_{Q_W}\left[M_W(t)^n + M_W(-t)^n \right] \right) (1 + W_*\sqrt{\gamma n})^d, \end{aligned} \tag{23}$$

where for any center $c \in \mathbb{R}^d$ and radius $r > 0$, $B_p(c,r) := \left\{ w \in \mathbb{R}^d : \|w - c\|_p \leq r \right\}$.

Proof. Since $\|\mu_p(s)\|_p \leq W_*$ for every s, we have $B_p(\mu_p(s), \beta) \subseteq B_p(0, \beta + W_*)$ for every s. Moreover, $\mathrm{Vol}(B_p(c,r)) := \int_{\mathbb{R}^d} \mathbf{1}\{w \in B_p(c,r)\}\, dw$ denotes the d-dimensional volume of the p-norm ball $B_p(c,r)$. For $1 \leq p < \infty$, the d-dimensional volume of $B_p(c,r)$ is given by (Gipple, 2014)

$$\mathrm{Vol}(B_p(c,r)) = r^d \frac{\left(2\Gamma\left(1 + \frac{1}{p}\right) \right)^d}{\Gamma\left(1 + \frac{d}{p}\right)},$$

whereas, for $p = \infty$,

$$\mathrm{Vol}(B_\infty(c,r)) = (2r)^d.$$

Given Lemma 6, we analyze $P_{S,W}\left(\{S \in A,\, W \in D_S\}\right)$ and $\lim_{\alpha\to\infty} \left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)} \right] \right)^{\frac{1}{\alpha}}$.

For the joint-event probability, we obtain

$$\begin{aligned} P_{S,W}\left(\{S \in A,\, W \in D_S\}\right) &= \int_{\mathcal{S}} \mathbf{1}_{\{s\in A\}} P_{W|S=s}\left(D_s\right) P_S(ds) \\ &= \int_A \frac{\mathrm{Vol}\left(D_s \cap B_p\left(\mu_p(s), \beta\right)\right)}{\mathrm{Vol}\left(B_p\left(\mu_p(s), \beta\right)\right)} P_S(ds) \\ &\overset{(a)}{=} \int_A \frac{\mathrm{Vol}\left(B_p\left(\mu_p(s), \beta\right)\right)}{\mathrm{Vol}\left(B_p\left(\mu_p(s), \beta\right)\right)} P_S(ds) \\ &= P_S(A), \end{aligned} \tag{24}$$

where the step (a) is because of $D_s = B_p\left(\mu_p(s), \beta\right)$.

For the Rényi-divergence term, we obtain

$$\begin{aligned} &\lim_{\alpha\to\infty} \left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)} \right] \right)^{\frac{1}{\alpha}} \\ &\overset{(a)}{=} \lim_{\alpha\to\infty} \left(\mathbb{E}_{S\sim P_S}\left[\exp\left[(\alpha-1) d \log\left(\frac{\beta + W_*}{\beta} \right) \right] \right] \right)^{\frac{1}{\alpha}} \\ &= \lim_{\alpha\to\infty} \left[\exp\left((\alpha-1) d \log\left(\frac{\beta + W_*}{\beta} \right) \right) \right]^{\frac{1}{\alpha}} \\ &= \left(1 + \frac{W_*}{\beta} \right)^d, \end{aligned} \tag{25}$$

where the step (a) follows below

$$
\begin{aligned}
&D_\alpha\big(P_{W|S=s} \,\|\, Q_W\big) \\
&= \frac{1}{\alpha-1}\log \mathbb{E}_{Q_W}\left(\frac{p_{W|S}(W|s)}{q_W(W)}\right)^\alpha \\
&= \frac{1}{\alpha-1}\log \mathbb{E}_{Q_W}\left(\frac{1\{W\in B_p(\mu_p(s),\beta)\}}{\mathrm{Vol}\left(B_p(\mu_p(s),\beta)\right)^\alpha}\frac{1\{W\in B_p(0,\beta+W_*)\}}{\mathrm{Vol}\left(B_p(0,\beta+W_*)\right)^{-\alpha}}\right) \\
&\overset{(a)}{=} \frac{1}{\alpha-1}\log\left(\frac{\mathrm{Vol}(B_p(0,\beta+W_*))}{\mathrm{Vol}(B_p(\mu_p(s),\beta))}\right)^{\alpha-1} \\
&\overset{(b)}{=} d\log\left(1+\frac{W_*}{\beta}\right),
\end{aligned}
$$

where the step (a) follows from $B_p(\mu_p(s),\beta)\subseteq B_p(0,\beta+W_*)$, so both indicators equal 1 on the support of Q_W, while step (b) follows from the p-norm ball volume formula,

$$
\mathrm{Vol}(B_p(\mu_p(s),\beta)) = \beta^d\frac{\left(2\Gamma\left(1+\frac{1}{p}\right)\right)^d}{\Gamma\left(1+\frac{d}{p}\right)} \text{ and } \mathrm{Vol}(B_p(0,\beta+W_*)) = (\beta+W_*)^d\frac{\left(2\Gamma\left(1+\frac{1}{p}\right)\right)^d}{\Gamma\left(1+\frac{d}{p}\right)}.
$$

Moreover, when $\alpha\to\infty$, we also have

$$
\begin{aligned}
&\lim_{\alpha\to\infty}\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right)^{\frac{\alpha-1}{\alpha}} \\
&= \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right). \qquad (26)
\end{aligned}
$$

Finally, taking the limit $\alpha\to\infty$ in equation 16 and applying equation 24, equation 25, and equation 26 yields

$$
\begin{aligned}
&P_S\left(\left|L(\mu_p(S))-\hat{L}(S,\mu_p(S))\right|-2C_{\ell,p}\beta\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\leq\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right)\left(1+\frac{W_*}{\beta}\right)^d,
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
&P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w)-\hat{L}(S,w)\right|-2C_{\ell,p}\beta\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\leq\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right)\left(1+\frac{W_*}{\beta}\right)^d.
\end{aligned}
$$

By choosing $\beta=\frac{1}{\sqrt{\gamma n}}$ for any $\gamma\geq 1$, we obtain inequality 23 in Proposition 1. □

Proposition 2 is required for Theorems 3 and 4, but not Theorems 1 and 2 which are the main results of the paper.

Proposition 2. *Follow definition 1 and suppose that Assumption 1 holds. Let* $0<k<1$, $\sigma_q>0$, *and* $\gamma\geq 1$, *and choose* $\beta=\frac{1}{\sqrt{\gamma n}}$, $P_{W|S=s}=\mathcal{N}\big(\mu_p(s),(k\sigma_q)^2I_d\big)$, *and* $Q_W=\mathcal{N}\big(0,\sigma_q^2I_d\big)$ *in Lemma 6. Then, we have*

$$
\begin{aligned}
&P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w)-\hat{L}(S,w)\right|-\frac{2C_{\ell,p}}{\sqrt{\gamma n}}\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right)\leq \\
&\frac{\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{Q_W}\left[M_W\left(t\right)^n+M_W\left(-t\right)^n\right]\right)\left\{k^{-d}\exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2(1-k^2)}\right]\right\}}{\Psi_d\left(\frac{1/\sqrt{\gamma n}}{k\sigma_q}\right)}, \qquad (27)
\end{aligned}
$$

where

$$B_{W_*,p} := \{w \in \mathbb{R}^d : \|w\|_p \le W_*\},$$

and

$$\Psi_d\left(\frac{\beta}{k\sigma_q}\right) := \frac{1}{(2\pi)^{d/2}} \int_{\left\{u\in\mathbb{R}^d : \|u\|_p \le \frac{\beta}{k\sigma_q}\right\}} \exp\left(-\frac{\|u\|_2^2}{2}\right) du.$$

Proof. Given Lemma 6, we analyze $P_{S,W}\left(\{S \in A,\, W \in D_S\}\right)$ and $\lim_{\alpha\to\infty}\left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)}\right]\right)^{\frac{1}{\alpha}}$ under the choice of Gaussian distributions for Q_W and $P_{W|S}$

$$\begin{aligned}
P_{S,W}\left(S \in A,\, W \in D_S\right) &= \int_{\mathcal{S}} \mathbf{1}_{\{s\in A\}} P_{W|S=s}\left(D_s\right) P_S(ds) \\
&= \int_A P_{W|S=s}\left(D_s\right) P_S(ds) \\
&= \int_A \left[\int_{\|u\|_p \le \frac{\beta}{k\sigma_q}} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|u\|_2^2}{2}\right) du\right] P_S(ds) \qquad (28)\\
&= P_S(A)\Psi_d\left(\frac{\beta}{k\sigma_q}\right) \qquad (29)
\end{aligned}$$

In equation 28, we apply the change of variables $u := \left(w - \mu_p(s)\right)/\left(k\sigma_q\right)$. Since $W|S = s \sim \mathcal{N}(\mu_p(s), (k\sigma_q)^2 I_d)$, the transformed random vector $(W - \mu_p(s))/(k\sigma_q)$ follows $\mathcal{N}(0, I_d)$. Since the resulting Gaussian-ball probability does not depend on s, it can be factored out of the outer integral, yielding equation 29.

Furthermore, for the Gaussian distributions $P_{W|S} = \mathcal{N}\big(\mu_p(S), k^2\sigma_q^2 I_d\big)$ and $Q_W = \mathcal{N}\big(0, \sigma_q^2 I_d\big)$ on $\mathbb{R}^d$, applying the closed-form expression for the Rényi divergence in Gil (2011, Table 2.2) yields

$$D_\alpha(P_{W|S}\|Q_W) = \frac{\alpha\,\|\mu_p(S)\|_2^2}{2\sigma_q^2\left(\alpha(1-k^2)+k^2\right)} - \frac{d}{2(\alpha-1)}\log\left(\frac{\alpha(1-k^2)+k^2}{k^{2(1-\alpha)}}\right). \qquad (30)$$

For $\alpha > 1$, we require $\frac{\alpha}{k^2} + 1 - \alpha > 0$ (i.e., $0 < k < \sqrt{\alpha/(\alpha-1)}$) for $D_\alpha(P_{W|S}\|Q_W)$ to be finite; otherwise, $D_\alpha(P_{W|S}\|Q_W) = +\infty$. Consequently, when $\alpha \to \infty$, we have $k \in (0,1)$ and

$$\begin{aligned}
&\lim_{\alpha\to\infty}\left(\mathbb{E}_{S\sim P_S}\left[e^{(\alpha-1)D_\alpha\left(P_{W|S}\|Q_W\right)}\right]\right)^{\frac{1}{\alpha}} \\
&= \lim_{\alpha\to\infty}\left(\int_{\mathcal{S}} \exp\left[(\alpha-1)D_\alpha\left(P_{W|S}\,\|\,Q_W\right)\right] P_S(ds)\right)^{\frac{1}{\alpha}} \\
&\overset{(a)}{=} \lim_{\alpha\to\infty}\left\{\left(\frac{k^{2(1-\alpha)}}{\alpha(1-k^2)+k^2}\right)^{d/2} \int_{\mathcal{S}} \exp\left[\frac{\alpha(\alpha-1)\,\|\mu_p(s)\|_2^2}{2\sigma_q^2\left(\alpha(1-k^2)+k^2\right)}\right] P_S(ds)\right\}^{\frac{1}{\alpha}} \\
&\overset{(b)}{\le} \lim_{\alpha\to\infty}\left\{\left(\frac{k^{2(1-\alpha)}}{\alpha(1-k^2)+k^2}\right)^{\frac{d}{2\alpha}} \left\{\exp\left[\frac{\alpha(\alpha-1)\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2\left(\alpha(1-k^2)+k^2\right)}\right]\right\}^{\frac{1}{\alpha}} \left(\int_{\mathcal{S}} P_S(ds)\right)^{\frac{1}{\alpha}}\right\} \qquad (31)\\
&\overset{(c)}{=} k^{-d}\exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2(1-k^2)}\right], \qquad (32)
\end{aligned}$$

where step (a) follows by substituting equation 30, step (b) is due to the definition of $\mu_p(S)$ (i.e. $\|\mu_p(S)\|_2^2 \leq \sup_{w\in B_{W_*,p}} \|w\|_2^2$), step (c) follows from the limiting calculations below.

$$\begin{aligned}\lim_{\alpha\to\infty}\left(\frac{k^{2(1-\alpha)}}{\alpha(1-k^2)+k^2}\right)^{\frac{d}{2\alpha}} &= \lim_{\alpha\to\infty} k^{\frac{d(1-\alpha)}{\alpha}}\left(\alpha(1-k^2)+k^2\right)^{-\frac{d}{2\alpha}} \\ &= k^{-d}\cdot\lim_{\alpha\to\infty} k^{d/\alpha}\exp\left(-\frac{d}{2\alpha}\log\left(\alpha(1-k^2)+k^2\right)\right) \\ &= k^{-d}\cdot 1\cdot 1 \\ &= k^{-d},\end{aligned}$$

and

$$\begin{aligned}&\lim_{\alpha\to\infty}\left\{\exp\left[\frac{\alpha(\alpha-1)\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2\left(\alpha(1-k^2)+k^2\right)}\right]\right\}^{\frac{1}{\alpha}} \\ &= \lim_{\alpha\to\infty}\exp\left[\frac{(\alpha-1)\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2\left(\alpha(1-k^2)+k^2\right)}\right] \\ &= \exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2}\lim_{\alpha\to\infty}\frac{\alpha-1}{\alpha(1-k^2)+k^2}\right] \\ &= \exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2(1-k^2)}\right].\end{aligned}$$

Moreover, when $\alpha\to\infty$, we also have

$$\begin{aligned}&\lim_{\alpha\to\infty}\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right)^{\frac{\alpha-1}{\alpha}} \\ &= \left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right). \qquad (33)\end{aligned}$$

Taking the limit $\alpha\to\infty$ in equation 16 and applying equation 29, equation 32, and equation 33 yields

$$\begin{aligned}&P_S\left(\left|L(\mu_p(S))-\hat{L}(S,\mu_p(S))\right|-2C_{\ell,p}\beta\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right)\leq \\ &\frac{\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W(-t)^n\right]\right)\left\{k^{-d}\exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2(1-k^2)}\right]\right\}}{\Psi_d\left(\frac{\beta}{k\sigma_q}\right)},\end{aligned}$$

or equivalently,

$$\begin{aligned}&P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w)-\hat{L}(S,w)\right|-2C_{\ell,p}\beta\geq\sqrt{\frac{\log(1/\delta)}{2n}}\right)\leq \\ &\frac{\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n+M_W\left(-t\right)^n\right]\right)\left\{k^{-d}\exp\left[\frac{\sup_{w\in B_{W_*,p}}\|w\|_2^2}{2\sigma_q^2(1-k^2)}\right]\right\}}{\Psi_d\left(\frac{\beta}{k\sigma_q}\right)}.\end{aligned}$$

By choosing $\beta=\frac{1}{\sqrt{\gamma n}}$ for any $\gamma\geq 1$, we complete the proof. □

A.2 Proof of Theorem 1

Given Proposition 1, by applying Lemma 2 to upper bound $\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n\right]$ in Proposition 1, we obtain

$$P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w)-\hat{L}(S,w)\right| - \frac{2C_{\ell,p}}{\sqrt{\gamma n}} \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \left(1+W_*\sqrt{\gamma n}\right)^d 2\delta^{1/(4\sigma_\ell^2)}. \quad (34)$$

By defining $\mathcal{K} := 2\left(1+W_*\sqrt{\gamma n}\right)^d$ and also letting $\delta' = \left(1+W_*\sqrt{\gamma n}\right)^d 2\delta^{1/(4\sigma_\ell^2)}$, we have $\delta = \left(\frac{\delta'}{\mathcal{K}}\right)^{4\sigma_\ell^2}$.

Since $\mathcal{K}\geq 2$, for any $\delta'\in(0,1)$ and $\beta\in(0,1]$, the corresponding δ always belongs to $(0,1)$, which ensures that the probability is always smaller than 1.

Then, with the above reparameterization, we write equation 34 as

$$P_S\left\{\forall w\in B_{W_*,p}, \left|L(w)-\hat{L}(S,w)\right| \leq f\left(W_*,\delta'\right)\right\} \geq 1-\delta', \quad (35)$$

where

$$f\left(W_*,\delta'\right) = \inf_{\gamma\geq 1}\left\{\sqrt{\frac{2\sigma_\ell^2\left[d\log\left(W_*\sqrt{\gamma n}+1\right)+\log\left(\frac{2}{\delta'}\right)\right]}{n}} + \frac{2C_{\ell,p}}{\sqrt{\gamma n}}\right\},$$

which is deterministic with respect to S.

Applying Lemma 5 to inequality 35 yields inequality 1 in Theorem 1. □

A.3 Proof of Theorem 2

Proof. Given Proposition 1, by applying Lemma 3 to upper bound $\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{Q_W}\left[M_W(t)^n + M_W(-t)^n\right]$, we obtain

$$\begin{aligned} &P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w)-\hat{L}(S,w)\right| - \frac{2C_{\ell,p}}{\sqrt{\gamma n}} \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\ &\leq \begin{cases} \left(1+W_*\sqrt{\gamma n}\right)^d 2\exp\left(-\frac{\log(1/\delta)}{4\nu_\ell^2}\right), & n\geq \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \\ \left(1+W_*\sqrt{\gamma n}\right)^d 2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell}\right), & 0<n<\frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}. \end{cases} \end{aligned} \quad (36)$$

By defining $\mathcal{K} := 2\left(1+W_*\sqrt{\gamma n}\right)^d$ and letting

$$\delta' = \begin{cases} \left(\frac{W_*+\beta}{\beta}\right)^d 2\exp\left(-\frac{\log(1/\delta)}{4\nu_\ell^2}\right), & n\geq \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \\ \left(\frac{W_*+\beta}{\beta}\right)^d 2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell}\right), & 0<n<\frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \end{cases}$$

we have

$$\delta = \begin{cases} \left(\frac{\delta'}{\mathcal{K}}\right)^{4\nu_\ell^2}, & n\geq \frac{2\alpha_\ell^2\log(\mathcal{K}/\delta')}{\nu_\ell^2}, \\ \exp\left[-\frac{8\alpha_\ell^2}{n}\left(\log\frac{\mathcal{K}}{\delta'}\right)^2\right], & 0<n<\frac{2\alpha_\ell^2\log(\mathcal{K}/\delta')}{\nu_\ell^2}. \end{cases} \quad (37)$$

Since $\mathcal{K}\geq 2$, for the first case, for any $\delta'\in(0,1)$, it follows that $\delta = \left(\frac{\delta'}{\mathcal{K}}\right)^{4\nu_\ell^2} \in (0,1)$. For the second case, we have, for any $\delta'\in(0,1)$, $\log\frac{\mathcal{K}}{\delta'} \geq \log 2$, so we obtain $\exp\left[-\frac{8\alpha_\ell^2}{n}\left(\log\frac{\mathcal{K}}{\delta'}\right)^2\right] \in (0,1)$.

Consequently, for every $\delta' \in (0,1]$, the corresponding δ defined at equation 37 satisfies $\delta \in (0,1)$ in both regimes of n.

Then, with the above reparameterization, we write equation 36 as

$$P_S\left\{\forall w \in B_{W_*,p}, \left|L(w) - \hat{L}(S,w)\right| \le f\left(W_*, \delta'\right)\right\} \ge 1 - \delta', \tag{38}$$

where

$$f(W_*,\delta') := \inf_{\gamma \ge 1} \begin{cases} \sqrt{\dfrac{2\nu_\ell^2\left(\log\frac{1}{\delta'} + \log \mathcal{K}\right)}{n}} + \dfrac{2C_{\ell,p}}{\sqrt{\gamma n}}, & n \ge \dfrac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\dfrac{\mathcal{K}}{\delta'}\right), \\ \dfrac{2\alpha_\ell\left(\log\frac{1}{\delta'} + \log\mathcal{K}\right)}{n} + \dfrac{2C_{\ell,p}}{\sqrt{\gamma n}}, & 0 < n < \dfrac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\dfrac{\mathcal{K}}{\delta'}\right), \end{cases}$$

which is deterministic with respect to S. Applying Lemma 5 to inequality 38 yields inequality 2 in Theorem 2. □

A.4 Theorem 3 and its Proof

Theorem 3. *Suppose that Assumptions 1 and 2 hold. For $w \in \mathbb{R}^d$, define $\overline{W}(w) := \max\{\lceil \|w\|_\infty \rceil, 1\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, it holds that*

$\forall w \in \mathbb{R}^d$, $\forall \gamma \ge 1$,

$$\left|L(w) - \hat{L}(S,w)\right| \le \sqrt{\frac{2\sigma_\ell^2\left(d\log\left(\frac{1}{k^\star}\right) + \frac{\overline{W}(w)^2\gamma n d}{2}\frac{k^{\star 2}}{(1-k^{\star 2})} + \log\frac{(4\overline{W}(w)^2)/\delta}{[2\Phi(1)-1]^d}\right)}{n}} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}, \tag{39}$$

where $k^\star = \sqrt{\frac{2}{\left(2+\overline{W}(w)^2\gamma n + \sqrt{\overline{W}(w)^2\gamma n\left(\overline{W}(w)^2\gamma n + 4\right)}\right)}}$ and Φ denotes the standard normal cumulative distribution function, with $\Phi(1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{1} e^{-t^2/2}dt$.

Proof. The theorem follows by specializing the Gaussian parameters k and σ_q in Proposition 2 and also applying Lemma 2 to bound the MGF-dependent term

$$\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n\right].$$

With the choice $p = \infty$, we have $\sup_{w\in B_{W_*,p}} \|w\|_2^2 = dW_*^2$.

We have

$$\begin{aligned}\Psi_d(1) &= \frac{1}{(2\pi)^{d/2}}\int_{\{u\in\mathbb{R}^d:\|u\|_\infty\le 1\}} \exp\left(-\frac{\|u\|_2^2}{2}\right) du \\ &= \prod_{i=1}^{d}\left[\frac{1}{\sqrt{2\pi}}\int_{-1}^{1} e^{-u_i^2/2}du_i\right] \\ &= (2\Phi(1)-1)^d. \end{aligned} \tag{40}$$

We choose $\sigma_q = k^{-1}(\gamma n)^{-\frac{1}{2}}$ (i.e., $\frac{1/\sqrt{\gamma n}}{k\sigma_q} = 1$). Therefore, by substituting equation 40 into equation 27 and applying Lemma 2 to upper bound $\left(\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{W\sim Q_W}\left[M_W(t)^n + M_W(-t)^n\right]\right)$, we obtain

$$\begin{aligned} &P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w) - \hat{L}(S,w)\right| - \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}} \ge \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\ &\le \frac{2\,\delta^{\frac{1}{4\sigma_\ell^2}}\exp\left(\frac{dW_*^2k^2\gamma n}{2(1-k^2)}\right)}{\left[(2\Phi(1)-1)\,k\right]^d}, \end{aligned} \tag{41}$$

We next minimize the RHS by selecting an optimal $k \in (0,1)$. Ignoring terms independent of k, its k-dependent part is

$$f(k) = k^{-d} \exp\left(\frac{dAk^2}{2(1-k^2)}\right), \tag{42}$$

where $A := W_*^2 \gamma n \geq 0$. Since $\log(\cdot)$ is strictly increasing, minimizing $f(k)$ is equivalent to minimizing

$$h(k) := \frac{\log f(k)}{d} = -\log k + \frac{Ak^2}{2(1-k^2)}, \qquad k \in (0,1).$$

Its derivative is

$$h'(k) = -\frac{1}{k} + \frac{Ak}{(1-k^2)^2} = \frac{Ak^2 - (1-k^2)^2}{k(1-k^2)^2}.$$

Thus, by setting $h'(k) = 0$, we have

$$Ak^2 = (1-k^2)^2. \tag{43}$$

For $k \in (0,1)$, the LHS of equation 43 is strictly increasing, whereas the right-hand side is strictly decreasing. Hence, this equation has a unique solution in $(0,1)$, given by

$$\begin{aligned} k^\star &= \sqrt{\frac{2 + W_*^2\gamma n - \sqrt{W_*^2\gamma n\,(W_*^2\gamma n + 4)}}{2}} \\ &= \sqrt{\frac{2}{2 + W_*^2\gamma n + \sqrt{W_*^2\gamma n\,(W_*^2\gamma n + 4)}}}. \end{aligned} \tag{44}$$

Substituting $k^\star$ into equation 41, we define $\mathcal{K} := \frac{2\exp\left(\frac{dW_*^2 k^{\star 2}\gamma n}{2(1-k^{\star 2})}\right)}{[(2\Phi(1)-1)k^\star]^d}$ and let $\delta' = \frac{2\,\delta^{\frac{1}{4\sigma_\ell^2}}\exp\left(\frac{dW_*^2 k^{\star 2}\gamma n}{2(1-k^{\star 2})}\right)}{[(2\Phi(1)-1)k^\star]^d}$, then we have $\delta = \left(\frac{\delta'}{\mathcal{K}}\right)^{4\sigma_\ell^2}$.

Since $k^\star \in (0,1)$, we have $(k^\star)^d < 1$, while $2\Phi(1) - 1 < 1$ and the exponential term is at least 1; hence the denominator is strictly smaller than 1 and the numerator is at least 2, implying $\mathcal{K} > 2$. Therefore, for any $\delta' \in (0,1)$, the corresponding δ always belongs to $(0,1)$, which ensures that the probability is always smaller than 1.

Then, with the above reparameterization, equation 41 can be written as

$$P_S\left\{\forall w \in B_{W_*,p}, \left|L(w) - \hat{L}(S,w)\right| \leq g\left(W_*, \delta'\right)\right\} \geq 1 - \delta', \tag{45}$$

where

$$g(W_*, \delta') = \inf_{\gamma \geq 1}\left\{\sqrt{\frac{2\sigma_\ell^2\left(d\log\left(\frac{1}{k^\star}\right) + \frac{dW_*^2\gamma n}{2}\frac{k^{\star 2}}{(1-k^{\star 2})} + \log\frac{2/\delta'}{[2\Phi(1)-1]^d}\right)}{n}} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}\right\},$$

which is deterministic with respect to S. Applying Lemma 5 to inequality 45 yields inequality 39 in Theorem 3. □

A.5 Theorem 4 and its Proof

Theorem 4. *Suppose that Assumptions 1 and 3 hold. For $w \in \mathbb{R}^d$, define $\overline{W}(w) := \max\{\lceil \|w\|_\infty \rceil, 1\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, it holds that*

$$\forall w \in \mathbb{R}^d,\ \forall \gamma \geq 1,$$

$$\left|L(w) - \hat{L}(S,w)\right| \leq \begin{cases} \sqrt{\frac{2\nu_\ell^2\varphi(w)}{n}} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}, & n \geq \frac{2\alpha_\ell^2}{\nu_\ell^2}\varphi(w), \\ \frac{2\alpha_\ell\varphi(w)}{n} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}, & 0 < n < \frac{2\alpha_\ell^2}{\nu_\ell^2}\varphi(w), \end{cases} \tag{46}$$

where

$$\varphi(w) = d\log\left(\frac{1}{k^\star}\right) + \frac{d\overline{W}(w)^2\gamma n}{2}\frac{k^{\star 2}}{(1-k^{\star 2})} + \log\frac{4\overline{W}(w)^2}{[2\Phi(1)-1]^d} + \log\left(\frac{1}{\delta}\right),$$

$$k^\star = \sqrt{\frac{2}{\left(2+\overline{W}(w)^2\gamma n + \sqrt{\overline{W}(w)^2\gamma n\left(\overline{W}(w)^2\gamma n + 4\right)}\right)}} \text{ and}$$

$$\Phi(1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{1} e^{-t^2/2}dt.$$

Proof. Applying Lemma 3 to upper bound $\exp\left(-t\sqrt{\frac{n\log(1/\delta)}{2}}\right)\mathbb{E}_{Q_W}\left[M_W(t)^n + M_W(-t)^n\right]$ in Proposition 2 and choosing $\sigma_q = k^{-1}(\gamma n)^{-\frac{1}{2}}$ (i.e., $\frac{1/\sqrt{\gamma n}}{k\sigma_q} = 1$), we obtain

$$P_S\left(\sup_{w\in B_{W_*,p}}\left|L(w) - \hat{L}(S,w)\right| - \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}} \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right)$$
$$\leq \begin{cases} \frac{2\,\delta^{\frac{1}{4\nu_\ell^2}}\exp\left(\frac{dW_*^2k^2\gamma n}{2(1-k^2)}\right)}{[(2\Phi(1)-1)k]^d}, & n \geq \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \\ \frac{2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell}\right)\exp\left(\frac{dW_*^2k^2\gamma n}{2(1-k^2)}\right)}{[(2\Phi(1)-1)k]^d}, & 0 < n < \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}. \end{cases} \tag{47}$$

We next minimize the RHS by selecting an optimal $k \in (0,1)$. Ignoring terms independent of k, its k-dependent part is given by equation 42, and hence the unique optimizing k^* is given by equation 44.

Substituting $k^\star$ into equation 47, we define $\mathcal{K} = \frac{2\exp\left(\frac{dW_*^2k^{\star 2}\gamma n}{2(1-k^{\star 2})}\right)}{[(2\Phi(1)-1)k^\star]^d}$ and let

$$\delta' = \begin{cases} \frac{2\,\delta^{\frac{1}{4\nu_\ell^2}}\exp\left(\frac{dW_*^2k^{\star 2}\gamma n}{2(1-k^{\star 2})}\right)}{[(2\Phi(1)-1)k^\star]^d}, & n \geq \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \\ \frac{2\exp\left(-\frac{\sqrt{n\log(1/\delta)}}{2\sqrt{2}\alpha_\ell}\right)\exp\left(\frac{dW_*^2k^{\star 2}\gamma n}{2(1-k^{\star 2})}\right)}{[(2\Phi(1)-1)k^\star]^d}, & 0 < n < \frac{\alpha_\ell^2\log(1/\delta)}{2\nu_\ell^4}, \end{cases}$$

then we have

$$\delta = \begin{cases} \left(\frac{\delta'}{\mathcal{K}}\right)^{4\nu_\ell^2}, & n \geq \frac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\frac{\mathcal{K}}{\delta'}\right), \\ \exp\left[-\frac{8\alpha_\ell^2}{n}\left(\log\left(\frac{\mathcal{K}}{\delta'}\right)\right)^2\right], & 0 < n < \frac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\frac{\mathcal{K}}{\delta'}\right). \end{cases} \tag{48}$$

Since $k^\star \in (0,1)$, we have $[k^\star]^d < 1$, while $2\Phi(1) - 1 < 1$ and the exponential term is at least 1; hence the denominator is strictly smaller than 1 and the numerator is at least 2, implying $\mathcal{K} > 2$.

Therefore, for any $\delta' \in (0,1)$, we have

$$0 < \frac{\delta'}{\mathcal{K}} < \frac{1}{2} < 1, \qquad \log\frac{\mathcal{K}}{\delta'} > \log 2 > 0.$$

For the first case, for any $\delta' \in (0,1)$ and $\beta \in (0,1]$, it follows that

$$\delta = \left(\frac{\delta'}{\mathcal{K}}\right)^{4\nu_\ell^2} \in (0,1).$$

Similarly, for the second case, under the same conditions, we have

$$\delta = \exp\left[-\frac{8\alpha_\ell^2}{n}\left(\log\left(\frac{\mathcal{K}}{\delta'}\right)\right)^2\right] \in (0,1).$$

Consequently, the δ defined at equation 48 satisfies $\delta \in (0,1)$ for any $\delta' \in (0,1)$ in both regimes of n.

With the above reparameterization, we write equation 47 as

$$P_S\left\{\forall w \in B_{W_*,p}, \left|L(w) - \hat{L}(S,w)\right| \le f\left(W_*, \delta'\right)\right\} \ge 1 - \delta', \tag{49}$$

where

$$f(W_*, \delta') := \inf_{\gamma \ge 1} \begin{cases} \sqrt{\frac{2\nu_\ell^2\left(\log \mathcal{K} + \log\left(\frac{1}{\delta'}\right)\right)}{n}} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}, & n \ge \frac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\frac{\mathcal{K}}{\delta'}\right), \\ \frac{2\alpha_\ell\left(\log \mathcal{K} + \log\left(\frac{1}{\delta'}\right)\right)}{n} + \frac{2C_{\ell,\infty}}{\sqrt{\gamma n}}, & 0 < n < \frac{2\alpha_\ell^2}{\nu_\ell^2}\log\left(\frac{\mathcal{K}}{\delta'}\right), \end{cases}$$

which is deterministic with respect to S. Applying Lemma 5 to equation 49 yields inequality 46 in Theorem 4. □

A.6 Comparison of Uniform (Theorems 1 and 2) and Gaussian Choices (Theorems 3 and 4)

Remark 2. *[Comparison of uniform and Gaussian choices] For $p = \infty$ and the same loss parameters, the uniform-based bounds in Theorems 1 and 2 are tighter than their Gaussian-based counterparts in Theorems 3 and 4, respectively, for every fixed w, n, δ, and $\gamma \ge 1$.*

To establish this comparison, set $\overline{W}(w) := \max\left\{\lceil \|w\|_\infty \rceil, 1\right\}$ in both constructions. Their additive remainder $2C_{\ell,\infty}/\sqrt{\gamma n}$ is then identical. After removing the common term $\log\left(4\overline{W}(w)^2/\delta\right)$ from their complexity expressions and dividing by d, it suffices to compare

$$\underbrace{\log\left(\overline{W}(w)\sqrt{\gamma n} + 1\right)}_{\textit{uniform choice}}$$

with

$$\underbrace{\log\left(\frac{1}{k^\star}\right) + \frac{\overline{W}(w)^2\gamma n}{2}\frac{(k^\star)^2}{1-(k^\star)^2} - \log\left(2\Phi(1) - 1\right)}_{\textit{Gaussian choice}}.$$

Starting from the definition of $k^\star$, we obtain

$$\begin{aligned} k^\star &= \sqrt{\frac{2}{2 + \overline{W}(w)^2\gamma n + \sqrt{\overline{W}(w)^2\gamma n\left(\overline{W}(w)^2\gamma n + 4\right)}}} \\ &= \sqrt{\frac{2}{\frac{1}{2}\left(\sqrt{\overline{W}(w)^2\gamma n + 4} + \overline{W}(w)\sqrt{\gamma n}\right)^2}} \\ &= \sqrt{\frac{4}{\left(\sqrt{\overline{W}(w)^2\gamma n + 4} + \overline{W}(w)\sqrt{\gamma n}\right)^2}} \\ &= \frac{2}{\sqrt{\overline{W}(w)^2\gamma n + 4} + \overline{W}(w)\sqrt{\gamma n}} \\ &= \frac{\sqrt{\overline{W}(w)^2\gamma n + 4} - \overline{W}(w)\sqrt{\gamma n}}{2}. \end{aligned}$$

So we have

$$\begin{aligned} \frac{1}{k^\star} - k^\star &= \frac{\sqrt{\overline{W}(w)^2\gamma n + 4} + \overline{W}(w)\sqrt{\gamma n}}{2} - \frac{\sqrt{\overline{W}(w)^2\gamma n + 4} - \overline{W}(w)\sqrt{\gamma n}}{2} \\ &= \overline{W}(w)\sqrt{\gamma n}. \end{aligned} \tag{50}$$

Therefore, we obtain

$$\begin{aligned}
&\log\left(\overline{W}(w)\sqrt{\gamma n}+1\right)-\log\left(\frac{1}{k^\star}\right)\\
&\quad=\log\left(k^\star\left(\overline{W}(w)\sqrt{\gamma n}+1\right)\right)\\
&\quad\overset{(a)}{=}\log\left(1+k^\star-(k^\star)^2\right)\\
&\quad=\log\left(\frac{5}{4}-\left(k^\star-\frac{1}{2}\right)^2\right)\\
&\quad\le\log\left(\frac{5}{4}\right)\approx 0.2231,
\end{aligned}$$

where step (a) is due to equation 50.

On the other hand,

$$-\log\left(2\Phi(1)-1\right)\approx 0.3817>\log\left(\frac{5}{4}\right).$$

Moreover, since $k^\star\in(0,1)$, we have $\frac{\overline{W}(w)^2\gamma n}{2}\frac{(k^\star)^2}{1-(k^\star)^2}>0$.

Therefore,

$$\begin{aligned}
&\log\left(\overline{W}(w)\sqrt{\gamma n}+1\right)\\
&\quad<\log\left(\frac{1}{k^\star}\right)+\frac{\overline{W}(w)^2\gamma n}{2}\frac{(k^\star)^2}{1-(k^\star)^2}-\log\left(2\Phi(1)-1\right).
\end{aligned}$$

Since the sub-Gaussian bound is increasing in its complexity term, this proves that Theorem 1 gives a tighter bound than Theorem 3.

The same comparison applies for the sub-exponential bounds as they have the same remaining term after removing all common terms. Thus, Theorem 2 also gives a tighter bound than Theorem 4 under the matched settings above.

B Specialized bounds for Gaussian mean estimation

We first compute the constants $C_{\ell,p}$ and σ_ℓ for our bound in Theorem 1 under Gaussian mean estimation, and then specialize the two prior bounds to the same setting.

Under Gaussian mean estimation setting, we have $|\ell(w,z)-\ell(w',z)|\le 1\cdot\|w-w'\|_2\le 1\cdot\|w-w'\|_1$ (i.e.,$C_{\ell,2}=C_{\ell,1}=1$).

Writing $Z=\mu+\Sigma^{1/2}\widetilde{Z}$, where $\widetilde{Z}\sim\mathcal{N}(0,I_2)$. Then, we view $\ell(w,Z)=\left\|w-\mu-\Sigma^{1/2}\widetilde{Z}\right\|_2$ as a function of $\widetilde{Z}$, so its Euclidean Lipschitz constant is $\sup_{\|u\|_2=1}\left\|\Sigma^{1/2}u\right\|_2$, where $\{u\in\mathbb{R}^2,\|u\|_2=1\}$. And we have

$$\sup_{\|u\|_2=1}\|\Sigma^{1/2}u\|_2=\sup_{u_1^2+u_2^2=1}\sqrt{\sigma^2u_1^2+\sigma^2u_2^2}\le\sup_{u_1^2+u_2^2=1}\sqrt{\max\{\sigma^2,\sigma^2\}(u_1^2+u_2^2)}=\sigma.$$

Therefore, by Theorem 2.26 of Wainwright (2019), the centered loss (i.e., $\ell(w,Z)-\mathbb{E}_Z[\ell(w,Z)]$) is sub-Gaussian with parameter $\sigma_\ell=\sup_{\|u\|_2=1}\|\Sigma^{1/2}u\|_2\le\sigma$.

To instantiate the Rademacher-complexity baselines, we view the two coordinates as two independent intercept-only linear prediction tasks. For each coordinate $j\in\{1,2\}$, we set $X_{ij}=1$, $Y_{ij}=Z_{ij}$, and $f_{w_j}(X_{ij})=w_jX_{ij}=w_j$.

Therefore, specializing Theorem 1 to the Gaussian mean estimation setting, we obtain that, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of S, the following holds:

$$\forall w \in \mathbb{R}^2,\ \forall \gamma \geq 1,$$

$$L(w) - \hat{L}(S,w) \leq \sqrt{\frac{2\sigma^2\left[2\log\left(\overline{W}(w)\sqrt{\gamma n}+1\right)+\log\left(\frac{4\overline{W}(w)^2}{\delta}\right)\right]}{n}} + \frac{2}{\sqrt{\gamma n}},$$

where $\overline{W}(w) := \max\{\lceil \|w\|_1 \rceil, 1\}$.

Also, we get the bound of Kakade et al. (2008) (i.e., Corollary 5) with $d=2$, $L_\ell = 1$, $X=1$, and $W_1 = 3\sigma$:

$$P_S\left(\forall w : \|w\|_1 \leq 3\sigma,\ L(w) - \hat{L}(S,w) \leq 6\sigma\sqrt{\frac{2\log 2}{n}} + 3\sigma\sqrt{\frac{\log(1/\delta)}{2n}}\right) \geq 1-\delta.$$

Specializing Theorem 1 of Awasthi et al. (2020) with $p=1$, $p^*=\infty$, $d=2$, $W=3\sigma$, and $\|X^\top\|_{2,\infty} = \sqrt{n}$, we obtain

$$P_S\left(\forall w : \|w\|_1 \leq 3\sigma,\ L(w) - \hat{L}(S,w) \leq 6\sigma\sqrt{\frac{2\log 4}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}\right) \geq 1-\delta.$$

C Specialized Bounds for Linear Regression

We first verify that the linear regression setting in Section 4.2 satisfies the assumptions of our sub-exponential result and derive the corresponding constants in the bound of Theorem 2, i.e., $C_{\ell,p}$, α_ℓ and ν_ℓ^2, before specializing our bound and the alternative bounds from the literature. Due to the hypothesis space constraint, we have $\|w - w^\star\|_1 \leq \|w\|_1 + \|w^\star\|_1 \leq 2$. Since $X_1 \in [\mu - 2s, \mu + 2s]$ and $X_2 \in [\mu - 2s, \mu + 2s]$, we have $\|X\|_\infty \leq \mu + 2s$ and $\|x - x'\|_\infty \leq 4s$.

For the absolute-error loss $\ell(w,Z) = \left|(w-w^\star)^\top X - \varepsilon\right|$, we have

$$\begin{aligned}
|\ell(w,z) - \ell(w',z)| &= \left|\left|(w-w^\star)^\top X - \varepsilon\right| - \left|(w'-w^\star)^\top X - \varepsilon\right|\right| \\
&\leq \left|(w-w')^\top X\right| \\
&\leq \|X\|_\infty \|w - w'\|_1 \\
&\leq (\mu + 2s)\|w - w'\|_1
\end{aligned}$$

Hence, we have $C_{\ell,1} = \mu + 2s$.

We next verify the sub-exponential condition and derive the corresponding parameters ν_ℓ^2 and α_ℓ.

For any fixed ε, we have

$$\begin{aligned}
&\left||(w-w^\star)^\top x - \varepsilon| - |(w-w^\star)^\top x' - \varepsilon|\right| \\
&\quad\leq |(w-w^\star)^\top (x - x')| \\
&\quad\leq \|w - w^\star\|_1 \|x - x'\|_\infty \\
&\quad\leq 8s.
\end{aligned}$$

Therefore, conditionally on ε, the range length of $\ell(w,Z)$ as X varies is at most $8s$. By Hoeffding's lemma, for any fixed ε and any $t \in \mathbb{R}$, we have

$$\mathbb{E}_X\left[\exp\left(t\left[\ell(w,Z) - \mathbb{E}_X[\ell(w,Z)]\right]\right)\right] \leq \exp\left(\frac{t^2}{2}\,16s^2\right). \tag{51}$$

For any $\varepsilon, \varepsilon' \sim \text{Laplace}(0,b)$, we have

$$\begin{aligned}
&\left|\mathbb{E}_X\left[\left|(w-w^\star)^\top X - \varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X - \varepsilon'\right|\right]\right| \\
&= \left|\mathbb{E}_X\left[\left|(w-w^\star)^\top X - \varepsilon\right| - \left|(w-w^\star)^\top X - \varepsilon'\right|\right]\right| \\
&\leq \mathbb{E}_X\left[\left||(w-w^\star)^\top X - \varepsilon| - |(w-w^\star)^\top X - \varepsilon'|\right|\right] \\
&\leq |\varepsilon - \varepsilon'|.
\end{aligned} \tag{52}$$

For $\varepsilon \sim \text{Laplace}(0, b)$, its MGF satisfies

$$\mathbb{E}_\varepsilon[e^{t\varepsilon}] = \frac{1}{1-b^2t^2}, \quad |t| < \frac{1}{b}.$$

We denote (X', ε') as an independent copy of (X, ε). Since ε and ε' are independent, we have

$$\mathbb{E}_{\varepsilon,\varepsilon'}\left[e^{t(\varepsilon-\varepsilon')}\right] = \mathbb{E}_\varepsilon\left[e^{t\varepsilon}\right]\mathbb{E}_{\varepsilon'}\left[e^{-t\varepsilon'}\right] = \frac{1}{(1-b^2t^2)^2}. \tag{53}$$

Because ε and ε' are independent and identically distributed, swapping them does not change their joint distribution. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\varepsilon,\varepsilon'} \exp\left(t\left[\mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon\right| - \mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right) \\
&= \mathbb{E}_{\varepsilon,\varepsilon'} \exp\left(-t\left[\mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon\right| - \mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right).
\end{aligned}$$

Furthermore, we can get

$$\begin{aligned}
&\mathbb{E}_{\varepsilon,\varepsilon'}\left[\exp\left(t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right)\right] \\
&= \frac{1}{2}\Bigg(\mathbb{E}_{\varepsilon,\varepsilon'} \exp\left(t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right) \\
&\qquad + \mathbb{E}_{\varepsilon,\varepsilon'} \exp\left(-t\left[\mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon\right| - \mathbb{E}_X\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right)\Bigg) \\
&= \mathbb{E}_{\varepsilon,\varepsilon'}\Bigg[\frac{1}{2}\Bigg\{\exp\left(t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right) \\
&\qquad + \exp\left(-t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right)\Bigg\}\Bigg] \\
&= \mathbb{E}_{\varepsilon,\varepsilon'}\left[\cosh\left[t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right]\right].
\end{aligned} \tag{54}$$

Then, for $|t| < \frac{1}{b}$, we have

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[e^{t(\mathbb{E}_X[\ell(w,Z)]-L(w))}\right] \\
&= \mathbb{E}_\varepsilon\left[\exp\left(t\left[\mathbb{E}_X\left[\ell\left(w,Z\right)\right] - \mathbb{E}_\varepsilon\left[\mathbb{E}_X\left[\ell\left(w,Z\right)\right]\right]\right]\right)\right] \\
&= \mathbb{E}_\varepsilon\left[\exp\left(\mathbb{E}_{\varepsilon'}\left[t\left(\mathbb{E}_X\left[\ell\left(w,Z\right)\right] - \mathbb{E}_{X'}\left[\ell\left(w,Z'\right)\right]\right)\right]\right)\right] \\
&\overset{(a)}{\leq} \mathbb{E}_\varepsilon\left[\mathbb{E}_{\varepsilon'}\left[\exp\left(t\left[\mathbb{E}_X\left[\ell\left(w,Z\right)\right] - \mathbb{E}_{X'}\left[\ell\left(w,Z'\right)\right]\right]\right)\right]\right] \\
&= \mathbb{E}_{\varepsilon,\varepsilon'}\left[\exp\left(t\left[\mathbb{E}_X\left[\ell\left(w,Z\right)\right] - \mathbb{E}_{X'}\left[\ell\left(w,Z'\right)\right]\right]\right)\right] \\
&\overset{(b)}{=} \mathbb{E}_{\varepsilon,\varepsilon'}\left[\cosh\left(t\left[\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right]\right)\right] \\
&\overset{(c)}{=} \mathbb{E}_{\varepsilon,\varepsilon'}\left[\cosh\left(|t|\left|\mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon\right|\right] - \mathbb{E}_X\left[\left|(w-w^\star)^\top X-\varepsilon'\right|\right]\right|\right)\right] \\
&\overset{(d)}{\leq} \mathbb{E}_{\varepsilon,\varepsilon'}\left[\cosh\left(t\left[\varepsilon-\varepsilon'\right]\right)\right] \\
&= \mathbb{E}_{\varepsilon,\varepsilon'}\left[e^{t\left(\varepsilon-\varepsilon'\right)}\right] \\
&\overset{(e)}{=} \frac{1}{\left(1-b^2t^2\right)^2}.
\end{aligned} \tag{55}$$

where step (a) is due to Jensen's inequality, step (b) is due to equation 54, step (c) is due to $\cosh(t) = \cosh(-t) = \cosh(|t|)$ for any $t \in \mathbb{R}$, step (d) is due to equation 52, and step (e) is due to equation 53.

Therefore, for $|t| < \frac{1}{\sqrt{2}b}$, we have

$$\begin{aligned}
&\mathbb{E}_Z\left[e^{t(\ell(w,Z)-\mathbb{E}_Z[\ell(w,Z)])}\right] \\
&= \mathbb{E}_{X,\varepsilon}\left[\exp\left(t[\ell(w,Z)-\mathbb{E}_X[\ell(w,Z)]]+t[\mathbb{E}_X[\ell(w,Z)]-L(w)]\right)\right] \\
&= \mathbb{E}_\varepsilon\left[e^{t(\mathbb{E}_X[\ell(w,Z)]-L(w))}\mathbb{E}_X\left[e^{t(\ell(w,Z)-\mathbb{E}_X[\ell(w,Z)])}\right]\right] \\
&\overset{(a)}{\leq} \exp\left(\frac{t^2}{2}16s^2\right)\mathbb{E}_\varepsilon\left[e^{t(\mathbb{E}_X[\ell(w,Z)]-L(w))}\right] \\
&\overset{(b)}{\leq} \exp\left(\frac{t^2}{2}16s^2\right)\frac{1}{(1-b^2t^2)^2} \\
&\overset{(c)}{\leq} \exp\left(\frac{t^2}{2}16s^2\right)\exp\left(\frac{t^2}{2}8b^2\right),
\end{aligned} \tag{56}$$

where step (a) is due to equation 51, step (b) is due to equation 55 and step (c) follows by applying $-2\log(1-b^2t^2) \leq 4b^2t^2$, valid for $0 \leq b^2t^2 < \frac{1}{2}$, which requires $|t| < 1/(\sqrt{2}\,b)$.

Comparing equation 56 with

$$M_w(t) \leq \exp\left(\frac{t^2\nu_\ell^2}{2}\right), \qquad |t| < \frac{1}{\alpha_\ell},$$

we have

$$\nu_\ell^2 = 16s^2 + 8b^2, \qquad \alpha_\ell = \sqrt{2}b.$$

In the current experiment, we have $\nu_\ell^2 = 16s^2 + 4$ and $\alpha_\ell = 1$.

Then we specialize Theorem 2 to the linear regression setting. Substituting $d = 2$, $p = 1$, $C_{\ell,1} = \mu + 2s$, $\nu_\ell^2 = 16s^2 + 4$, and $\alpha_\ell = 1$ into Theorem 2 , we obtain that, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, the following inequality holds simultaneously for all $w \in \mathbb{R}^d$ and all $\gamma \geq 1$:

$$\forall w \in \mathcal{W},\ \forall \gamma \geq 1,$$

$$L(w) - \hat{L}(S,w) \leq \begin{cases} \sqrt{\dfrac{2(16s^2+4)\psi_{\mathrm{LR}}(w)}{n}} + \dfrac{2(\mu+2s)}{\sqrt{\gamma n}}, & n \geq \dfrac{2\psi_{\mathrm{LR}}(w)}{(16s^2+4)}, \\ \dfrac{2\psi_{\mathrm{LR}}(w)}{n} + \dfrac{2(\mu+2s)}{\sqrt{\gamma n}}, & 0 < n < \dfrac{2\psi_{\mathrm{LR}}(w)}{(16s^2+4)}, \end{cases}$$

where

$$\psi_{\mathrm{LR}}(w) = 2\log\left(\overline{W}(w)\sqrt{\gamma n}+1\right) + \log\left(4\overline{W}(w)^2\right) + \log\frac{1}{\delta},$$

and

$$\overline{W}(w) := \max\{\lceil \|w\|_1 \rceil, 1\}.$$

For comparison, we next specialize Theorem 1 of Awasthi et al. (2020). Here, $d = 2$, $p = 1$, $p^* = \infty$, and $W_1 = 1$. Since the truncation implies $\|X\|_\infty \leq \mu + 2s$, then we have

$$\|X^\top\|_{2,\infty} = \max_{j\in\{1,2\}}\left(\sum_{i=1}^n X_{ij}^2\right)^{1/2} \leq (\mu+2s)\sqrt{n}.$$

Hence,

$$\hat{\mathfrak{R}}_S(\mathcal{F}_1) \leq (\mu+2s)\sqrt{\frac{2\log 4}{n}},$$

which yields the corresponding bound

$$P_S\left(\forall w \in \mathcal{W}, L(w) - \hat{L}(S,w) \leq 2(\mu+2s)\sqrt{\frac{2\log 4}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}\right) \geq 1-\delta.$$

Notation	Definition / Dimension	Description
K_k	$K_k \in \mathbb{R}^{c_k^{\text{out}} \times c_k^{\text{in}} \times r_k \times s_k}$, $k = 1, \ldots, L$	Convolutional kernel tensor at layer k, where c_k^{in} and c_k^{out} are the numbers of input and output channels, respectively, and $r_k \times s_k$ is the spatial kernel size.
$H_k^{\text{out}}, W_k^{\text{out}}$	$H_k^{\text{out}}, W_k^{\text{out}} \in \mathbb{N}$	Spatial height and width of the feature map produced by convolutional layer k.
$\text{op}(K_k)$	Linear operator induced by K_k	Matrix representation of convolution with K_k, satisfying $\text{vec}(K_k * u) = \text{op}(K_k)\,\text{vec}(u)$.
w	$w := \text{vec}(K_1, \ldots, K_L, W_{L+1}) \in \mathbb{R}^d$	Vector containing all trainable parameters of the network.
σ_k	$\sigma_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$	Activation following layer k. We assume $\sigma_k(0) = 0$ and that σ_k is non-expansive; e.g., ReLU, LeakyReLU with slope in $[0, 1]$, or $\tanh$.
g_k	$g_k(u) := \sigma_k(K_k * u)$	Mapping implemented by convolutional layer k, including its activation.
Lip_k	$\|g_k(u) - g_k(v)\|_2 \le \text{Lip}_k \|u - v\|_2$, $k = 1, ..., L + 1$	Layer-wise Lipschitz constant.
$f_w(x)$	$f_w(x) = g_{L+1} \circ g_L \circ \cdots \circ g_1(x) \in \mathbb{R}^C$	Predictive logit vector produced by the network with parameters w.
$f_{w'}(x)$	$f_{w'}(x) = g'_{L+1} \circ g'_L \circ \cdots \circ g'_1(x)$	Predictive logit vector produced by the perturbed network with parameters w'.
$f^{(k)}(x)$	$f^{(k)}(x) := g_{L+1} \circ \cdots \circ g_{k+1} \circ g'_k \circ \cdots \circ g'_1(x)$	Hybrid network in which the first k layers use the parameters from w', while the remaining layers use those from w. It is introduced for the layer-wise telescoping argument.
B_x	$\|x\|_2 \le B_x$	Uniform upper bound on the Euclidean norm of the network input.
$\|\cdot\|_F$	Frobenius norm	Euclidean norm of all entries of a matrix or convolutional kernel tensor; it coincides with the Euclidean norm after vectorization.
$\|W_{L+1}\|_2$, $\|\text{op}(K_k)\|_2$	Spectral norm	Spectral norm of the weight matrix of the final linear layer and the convolutional operator matrices.

Table 2: Notation used in the CNN Lipschitz analysis. (Appendix D)

Finally, specializing Corollary 5 of Kakade et al. (2008) with $d = 2$, $L_\ell = 1$, $\|X\|_\infty \le \mu + 2s$, and $W_1 = 1$ gives

$$P_S\left(\forall w \in \mathcal{W}, L(w) - \hat{L}(S, w) \le 2(\mu + 2s)\sqrt{\frac{2\log 2}{n}} + (\mu + 2s)\sqrt{\frac{\log(1/\delta)}{2n}}\right) \ge 1 - \delta.$$

D Additional Details for Multiclass Classification with DNNs

D.1 Lipschitz Analysis for Convolutional Neural Networks

Consider the convolutional neural network with L convolutional layers followed by an average pooling layer and a fully connected linear layer. Table 2 summarizes all the notations used in this section.

For each convolutional layer $k = 1, \ldots, L$, define

$$g_k(u) := \sigma_k(K_k * u) = \sigma_k(\text{op}(K_k)\,\text{vec}(u))\,,$$

where $\text{op}(K_k)$ is the matrix representation of the convolution kernel K_k, satisfying $\text{vec}(K_k * u) = \text{op}(K_k)\,\text{vec}(u)$.

Layer-wise Lipschitz constraint Lip_k: we assume that for $k = 1, ..., L+1$, g_k is Lip_k-Lipschitz with respect to the Euclidean norm; namely, for any admissible layer inputs u and v,

$$\|g_k(u) - g_k(v)\|_2 \le \mathrm{Lip}_k \|u - v\|_2. \tag{57}$$

Leaky-ReLU is defined as

$$\sigma_k(u) = \max\{u, au\} = \begin{cases} u, & u \ge 0, \\ au, & u < 0, \end{cases}$$

thus we have

$$\|\sigma_k(u) - \sigma_k(v)\| \le \|u - v\|. \tag{58}$$

For a convolutional layer, we have

$$\begin{aligned} \|g_k(u) - g_k(v)\|_2 &= \|\sigma_k(K_k * u) - \sigma_k(K_k * v)\|_2 \\ &\overset{(a)}{\le} \|K_k * u - K_k * v\|_2 \\ &= \|\mathrm{op}(K_k)\,\mathrm{vec}(u - v)\|_2 \\ &\le \|\mathrm{op}(K_k)\|_2 \,\|\mathrm{vec}(u - v)\|_2 \\ &= \underbrace{\|\mathrm{op}(K_k)\|_2}_{\mathrm{Lip}_k} \|u - v\|_2, \end{aligned} \tag{59}$$

where the step (a) follows from the 1-Lipschitz property of the LeakyReLU activation given in equation 58.

Let Pool denote the spatial average-pooling operator that maps the output of the final convolutional layer to a feature map of fixed size 7×7. For each output location (a, b) and channel c, let $\mathcal{R}_{a,b}$ denote the set of input spatial locations averaged to produce that output, and let $q_{a,b} := |\mathcal{R}_{a,b}|$ be the number of such locations.

Average pooling is applied independently to each channel and is given by

$$[\mathrm{Pool}(u)]_{c,a,b} = \frac{1}{q_{a,b}} \sum_{(i,j)\in\mathcal{R}_{a,b}} u_{c,i,j},$$

where $u_{c,i,j}$ denotes the value of the input feature map at channel c and spatial location (i, j). Hence, for any two feature maps u and v, by the Cauchy–Schwarz inequality,

$$\begin{aligned} |[\mathrm{Pool}(u) - \mathrm{Pool}(v)]_{c,a,b}|^2 &= \left| \frac{1}{q_{a,b}} \sum_{(i,j)\in\mathcal{R}_{a,b}} 1 \cdot (u_{c,i,j} - v_{c,i,j}) \right|^2 \\ &\le \frac{1}{q_{a,b}^2} \left(\sum_{(i,j)\in\mathcal{R}_{a,b}} 1^2 \right) \left(\sum_{(i,j)\in\mathcal{R}_{a,b}} (u_{c,i,j} - v_{c,i,j})^2 \right) \\ &= \frac{1}{q_{a,b}} \sum_{(i,j)\in\mathcal{R}_{a,b}} (u_{c,i,j} - v_{c,i,j})^2 . \end{aligned} \tag{60}$$

Summing over all channels and output spatial locations gives

$$\begin{aligned} \|\,\mathrm{Pool}(u) - \mathrm{Pool}(v)\|_2^2 &= \sum_c \sum_{a,b} |[\mathrm{Pool}(u) - \mathrm{Pool}(v)]_{c,a,b}|^2 \\ &\le \sum_c \sum_{a,b} \frac{1}{q_{a,b}} \sum_{(i,j)\in\mathcal{R}_{a,b}} |u_{c,i,j} - v_{c,i,j}|^2 \\ &= \sum_c \sum_{i,j} \left(\sum_{(a,b):(i,j)\in\mathcal{R}_{a,b}} \frac{1}{q_{a,b}} \right) |u_{c,i,j} - v_{c,i,j}|^2 \\ &\overset{(a)}{\le} \sum_c \sum_{i,j} |u_{c,i,j} - v_{c,i,j}|^2 \\ &= \|u - v\|_2^2. \end{aligned} \tag{61}$$

The step (a) is due to $\max_{i,j} \sum_{(a,b):(i,j)\in\mathcal{R}_{a,b}} \frac{1}{q_{a,b}} \leq 1$, which is because we implement a non-overlapping average pooling operation, so every input spatial location (i, j) belongs to exactly one pooling region $\mathcal{R}_{a,b}$, thus we have $\max_{(i,j)} \sum_{(a,b):(i,j)\in\mathcal{R}_{a,b}} \frac{1}{q_{a,b}} = \max_{a,b} \frac{1}{q_{a,b}} \leq 1$. Thus, spatial average pooling is non-expansive with respect to the Euclidean norm. Moreover, since Pool contains no trainable parameters, it does not contribute to w or $\|w - w'\|_2$.

For the $L + 1$-th final linear layer, we have

$$\begin{aligned}\|g_{L+1}(u) - g_{L+1}(v)\|_2 &\leq \|W_{L+1}\|_2 \|\mathrm{Pool}(u) - \mathrm{Pool}(v)\|_2 \\ &\overset{(a)}{\leq} \underbrace{\|W_{L+1}\|_2}_{\mathrm{Lip}_{L+1}} \|u - v\|_2, \end{aligned} \tag{62}$$

where the step (a) is due to equation 61.

Therefore, restricting the spectral norm of the convolutional operators and linear (i.e., $\|\mathrm{op}(K_k)\|_2 \leq \lambda$ and $\|W_{L+1}\|_2 \leq \lambda$) weight matrices is equivalent to restricting the layer-wise Euclidean Lipschitz constants Lip_k.

We define $w := \mathrm{vec}(K_1, ..., K_L, W_{L+1})$, so we have

$$\|w - w'\|_2 = \left(\sum_{k=1}^{L} \|K_k - K'_k\|_F^2 + \|W_{L+1} - W'_{L+1}\|_F^2 \right)^{1/2}, \tag{63}$$

where $\|\cdot\|_F$ denotes Frobenius norm.

For multiclass classification, we denote the output of the network parameterized by w as

$$f_w(x) = \big(f_w(x)_1, \ldots, f_w(x)_C\big) \in \mathbb{R}^C,$$

where C is the number of classes. For each $k = 0, \ldots, L + 1$, let $f^{(k)}$ denote the network obtained from f_w by replacing the parameters of its first k layers with the corresponding parameters from w', while keeping the parameters of the remaining layers unchanged. Accordingly, for $k = 1, \ldots, L+1$,

$$f^{(k)}(x) := g_{L+1} \circ \cdots \circ g_{k+1} \circ g'_k \circ \cdots \circ g'_1(x) \in \mathbb{R}^C.$$

Thus, each $f^{(k)}(x)$ is a predictive logit vector,

$$f^{(k)}(x) = \big(f^{(k)}(x)_1, \ldots, f^{(k)}(x)_C\big).$$

In particular, we have

$$f^{(0)}(x) = g_{L+1} \circ \cdots \circ g_1(x) \text{ and } f^{(L+1)}(x) = g'_{L+1} \circ \cdots \circ g'_1(x),$$

which represent the original network with no parameters replaced (i.e., $f^{(0)}(x) = f_w(x)$) and the network in which all layer parameters are replaced by $w' := \mathrm{vec}(K'_1, ..., K'_L, W'_{L+1})$ (i.e., $f^{(L+1)}(x) = f_{w'}(x)$), respectively.

The m-ramp loss is defined as

$$\phi_m(m_w(x, y)) := \begin{cases} 1, & m_w(x, y) \leq 0, \\ 1 - \dfrac{m_w(x, y)}{m}, & 0 < m_w(x, y) < m, \\ 0, & m_w(x, y) \geq m, \end{cases}$$

where $m > 0$ and $m_w(x, y) := f_w(x)_y - \max_{j \neq y} f_w(x)_j$. Therefore, we have

$$\begin{aligned}
|\ell(w, z) - \ell(w', z)| &= |\phi_m(m_w(x, y)) - \phi_m(m_{w'}(x, y))| \\
&\leq \frac{1}{m} |m_w(x, y) - m_{w'}(x, y)| \\
&= \frac{1}{m} \left| f_w(x)_y - \max_{j \neq y} f_w(x)_j - f_{w'}(x)_y + \max_{j \neq y} f_{w'}(x)_j \right| \\
&\leq \frac{1}{m} \left(|f_w(x)_y - f_{w'}(x)_y| + \left| \max_{j \neq y} f_w(x)_j - \max_{j \neq y} f_{w'}(x)_j \right| \right) \\
&\leq \frac{1}{m} \left(\|f_w(x) - f_{w'}(x)\|_\infty + \|f_w(x) - f_{w'}(x)\|_\infty \right) \\
&= \frac{2}{m} \|f_w(x) - f_{w'}(x)\|_\infty \\
&\leq \frac{2}{m} \|f_w(x) - f_{w'}(x)\|_2. \qquad (64)
\end{aligned}$$

Additionally, we have $\ell_{01}(w, (x, y)) = \mathbf{1}\{m_w(x, y) \leq 0\} \leq \phi_m(m_w(x, y))$, thus

$$L_{01}(w) := \mathbb{E}_{X,Y}[\ell_{01}(w, (X, Y))] \leq \mathbb{E}_{X,Y}[\phi_m(m_w(X, Y))]. \qquad (65)$$

We consider a bias-free CNN and assume that the input has bounded Euclidean norm, i.e., $\|x\|_2 \leq B_x$. Moreover, each activation function is 1-Lipschitz and satisfies $\sigma_k(0) = 0$. Consequently, $g'_k(0) = 0$ for every layer k. Therefore, for any $k = 1, \ldots, L$, the representation entering the k-th layer satisfies

$$\begin{aligned}
\left\| g'_{k-1} \circ \cdots \circ g'_1(x) \right\|_2 &= \left\| g'_{k-1} \circ \cdots \circ g'_1(x) - g'_{k-1} \circ \cdots \circ g'_1(0) \right\|_2 \\
&\leq \lambda \left\| g'_{k-2} \circ \cdots \circ g'_1(x) - g'_{k-2} \circ \cdots \circ g'_1(0) \right\|_2 \\
&\leq \lambda^2 \left\| g'_{k-3} \circ \cdots \circ g'_1(x) - g'_{k-3} \circ \cdots \circ g'_1(0) \right\|_2 \\
&\quad \vdots \\
&\leq \lambda^{k-1} \|x - 0\|_2 \\
&\leq B_x \lambda^{k-1}. \qquad (66)
\end{aligned}$$

For any $k \in [1, L]$, we have

$$\begin{aligned}
\left\| f^{(k)}(x) - f^{(k-1)}(x) \right\|_2 &= \Big\| g_{L+1} \circ \cdots \circ g_{k+1} \circ g'_k \circ g'_{k-1} \circ \cdots \circ g'_1(x) \\
&\qquad - g_{L+1} \circ \cdots \circ g_{k+1} \circ g_k \circ g'_{k-1} \circ \cdots \circ g'_1(x) \Big\|_2 \\
&\overset{(a)}{\leq} \lambda^{L-k+1} \Big\| g'_k\big(g'_{k-1} \circ \cdots \circ g'_1(x)\big) - g_k\big(g'_{k-1} \circ \cdots \circ g'_1(x)\big) \Big\|_2 \\
&\leq \lambda^{L-k+1} \Big\| (K'_k - K_k) * \big(g'_{k-1} \circ \cdots \circ g'_1(x)\big) \Big\|_2 \\
&= \lambda^{L-k+1} \Big\| \big(\mathrm{op}(K'_k) - \mathrm{op}(K_k)\big) \mathrm{vec}\big(g'_{k-1} \circ \cdots \circ g'_1(x)\big) \Big\|_2 \\
&\leq \lambda^{L-k+1} \|\mathrm{op}(K'_k) - \mathrm{op}(K_k)\|_2 \left\| g'_{k-1} \circ \cdots \circ g'_1(x) \right\|_2 \\
&\leq \lambda^{L-k+1} \|\mathrm{op}(K'_k) - \mathrm{op}(K_k)\|_F \left\| g'_{k-1} \circ \cdots \circ g'_1(x) \right\|_2 \\
&\overset{(b)}{=} \lambda^{L-k+1} \sqrt{H_k^{\mathrm{out}} W_k^{\mathrm{out}}} \, \|K'_k - K_k\|_F \left\| g'_{k-1} \circ \cdots \circ g'_1(x) \right\|_2 \\
&\overset{(c)}{\leq} \lambda^{L-k+1} \sqrt{H_k^{\mathrm{out}} W_k^{\mathrm{out}}} \, \|(K'_k) - (K_k)\|_F \, B_x \lambda^{k-1} \\
&= \lambda^{L} \sqrt{H_k^{\mathrm{out}} W_k^{\mathrm{out}}} \, \|(K'_k) - (K_k)\|_F \, B_x \qquad (67)
\end{aligned}$$

where H_k^{out} and W_k^{out} denote the spatial height and width of the feature map produced by the k-th convolutional layer. The step (a) is due to the spectral norm constraint shown at 59 and 62, the step (b) follows from the weight-sharing structure of convolution, under which each kernel coefficient appears once at each of the $H_k^{\text{out}} W_k^{\text{out}}$ output spatial locations in $\text{op}(K_k' - K_k)$, and the step (c) is due to equation 66. Moreover, we derive the upper bound of $\|f_w(x) - f_{w'}(x)\|_2$.

$$
\begin{aligned}
\|f_w(x) - f_{w'}(x)\|_2 &= \left\| f^{(L+1)}(x) - f^{(0)}(x) \right\|_2 \\
&= \left\| f^{(L+1)}(x) - f^{(L)}(x) + ... + f^{(1)}(x) - f^{(0)}(x) \right\|_2 \\
&= \left\| \sum_{k=1}^{L+1} \left(f^{(k)}(x) - f^{(k-1)}(x) \right) \right\|_2 \\
&\leq \sum_{k=1}^{L+1} \left\| \left(f^{(k)}(x) - f^{(k-1)}(x) \right) \right\|_2 \\
&\overset{(a)}{\leq} B_x \sum_{k=1}^{L} \left[\left(\lambda^L\right) \sqrt{H_k^{\text{out}} W_k^{\text{out}}} \, \|(K_k') - (K_k)\|_F \right] \\
&\quad + B_x \left(\lambda^L\right) \left\| W_{L+1} - W_{L+1}' \right\|_F \\
&\overset{(b)}{\leq} B_x \left[\sum_{k=1}^{L} H_k^{\text{out}} W_k^{\text{out}} \left(\lambda^L\right)^2 + \left(\lambda^L\right)^2 \right]^{1/2} \\
&\quad \times \left[\sum_{k=1}^{L} \|K_k - K_k'\|_F^2 + \|W_{L+1} - W_{L+1}'\|_F^2 \right]^{1/2} \\
&\overset{(c)}{=} B_x \lambda^L \sqrt{1 + \sum_{k=1}^{L} H_k^{\text{out}} W_k^{\text{out}}} \, \|w - w'\|_2 \,,
\end{aligned} \tag{68}
$$

where the step (a) is by equation 67, the step (b) is by Cauchy–Schwarz inequality, the step (c) is due to the definition at equation 63.

Therefore, when the loss function is m-ramp loss, we have

$$
\begin{aligned}
|\phi_m(m_w(x,y)) - \phi_m(m_{w'}(x,y))| &\overset{(a)}{\leq} \frac{2}{m} \|f_w(x) - f_{w'}(x)\|_2 \\
&\overset{(b)}{\leq} \frac{2}{m} B_x \lambda^L \sqrt{1 + \sum_{k=1}^{L} H_k^{\text{out}} W_k^{\text{out}}} \, \|w - w'\|_2 \,,
\end{aligned}
$$

where the step (a) is due to equation 64, and the step (b) is due to equation 68.

So we obtain $C_{\ell,2} = \frac{2}{m} B_x \lambda^L \sqrt{1 + \sum_{k=1}^{L} H_k^{\text{out}} W_k^{\text{out}}}$. For the MNIST dataset, each input image satisfies $x \in [0,1]^{28 \times 28}$. Hence, $\|x\|_2 \leq \sqrt{28 \times 28} = 28$, and we may take $B_x = 28$. For CIFAR-10, we have $B_x = 32\sqrt{3}$. In particular, if we impose the same spectral norm constraint λ on all convolutional layers and the final linear layer, and we set the ramp-loss margin to $m = 1$ throughout all experiments, then

$$
C_{\ell,2} = 2B_x \times \lambda^L \sqrt{1 + \sum_{k=1}^{L} H_k^{\text{out}} W_k^{\text{out}}}.
$$

Because $0 \leq \phi_m(m_w(x,y)) \leq 1$, its sub-Gaussian parameter $\sigma_\ell = \frac{1}{2}$. Therefore, we can apply Theorem 1 to upper bound the generalization gap between expected ramp loss and empirical ramp

loss $|L_{\text{ramp}}(w) - \hat{L}_{\text{ramp}}(S, w)|$, where

$$L_{\text{ramp}}(w) := \mathbb{E}_{X,Y}\left[\phi_m(m_w(X, Y))\right] \text{ and } \hat{L}_{\text{ramp}}(S, w) := \sum_{i=1}^{n} \frac{\phi_m(m_w(x_i, y_i))}{n}.$$

By lower bounding $L_{\text{ramp}}(w)$ using $L_{01}(w)$ (see equation 65), we achieve the specialized bound under DNNs example:

Corollary 1. *(DPI bound for DNNs). Suppose that Assumption 1 and Assumption 2 hold. For $w \in \mathbb{R}^d$, define $\overline{W}(w) := \max\{\lceil \|w\|_2 \rceil, 1\}$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over S, the following inequality holds simultaneously for all $w \in \mathbb{R}^d$ and all $\gamma \geq 1$:*

$$\forall w \in \mathbb{R}^d,\ \forall \gamma \geq 1,$$

$$L_{01}(w) \leq \hat{L}_{ramp}(S, w) + \sqrt{\frac{\left[d \log\left(\overline{W}(w)\sqrt{\gamma n} + 1\right) + \log\left(\frac{4\overline{W}(w)^2}{\delta}\right)\right]}{2n}} + \frac{2C_{\ell,2}}{\sqrt{\gamma n}}. \tag{69}$$

D.2 Experiment details on MNIST

Each network consists of L convolutional layers with 2×2 kernels, unit stride, zero padding, and H output channels, followed by spatial average pooling to a fixed 7×7 resolution and a fully connected output layer. We use LeakyReLU activations with negative slope 0.1 after each convolutional layer.

Computing the spectral norm of a convolutional layer exactly can be costly because its associated linear operator is high-dimensional (e.g., the spectral norm of the second convolutional layer in an eight-layer MNIST CNN with 500 channels is the largest singular value of a $338{,}000 \times 364{,}500$ matrix). During bound evaluation, we therefore rescale the layer weights using the efficiently computable upper bound of Singla & Feizi (2019, Theorem 1). Constraining this upper bound to be at most λ ensures that the layer's spectral norm is also at most λ.

We optimize the network using AdamW with $(\beta_1, \beta_2) = (0.9, 0.99)$, zero weight decay, a batch size of 256 and 200 epochs. The peak learning rate is set to 5×10^{-3} for networks with at most four convolutional layers and 10^{-3} for deeper networks. We employ a step-wise learning-rate schedule with a linear warm-up over the first 10% of the training iterations, followed by cosine decay to 2% of the peak learning rate, i.e., 10^{-4} and 2×10^{-5}, respectively. The training objective is the cross-entropy loss combined with a spectral regularization term:

$$\mathcal{L}_{\text{train}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\big(f_{y_i}(x_i)\big)}{\sum_{c=1}^{10} \exp\big(f_c(x_i)\big)} + \theta \sum_{k} \widetilde{b}_k,$$

where we fix $\theta = 10^{-3}$ and $\widetilde{b}_k$ denotes a differentiable, warm-started power-iteration surrogate of the spectral quantity associated with the k-th trainable layer; one power-iteration step is used per mini-batch. Such training objectives have been widely adopted in studies of spectral norm control and regularization in deep neural networks. For details, see (Singla & Feizi, 2019), (Miyato et al., 2018), and (Gouk et al., 2021).

We additionally investigate the effect of varying the constraint strength in Figure 5. As λ increases, the spectral-norm constraint becomes less restrictive. The figure shows that our bound and the existing bounds increase concurrently as we reduce the constraints on spectral norm, but our bound exhibits more modest growth.

D.3 Additional experimental results on CIFAR-10

Here, we evaluate the tightness of our bound against existing bounds from the literature on CIFAR-10 by varying the network depth (Figure 6), width (Figure 7) and also the spectral-norm constraint (Figure 8). We observe a pattern similar to that obtained on MNIST across all evaluated bounds. Additionally, as the learning task becomes more challenging, all bounds become looser; nevertheless, our bound remains the tightest across the considered architectures.

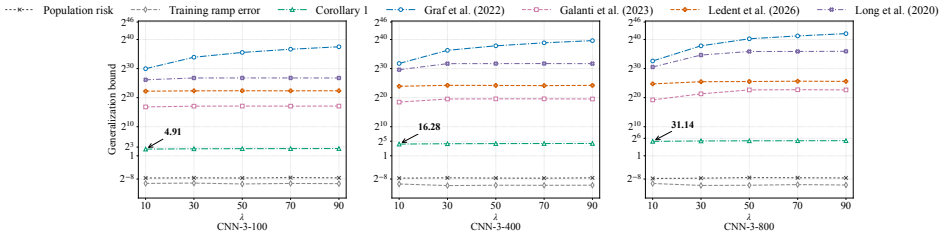


Figure 5: Varying the spectral-norm constraint λ for CNN-3-H architectures trained on MNIST, with different channel widths H.

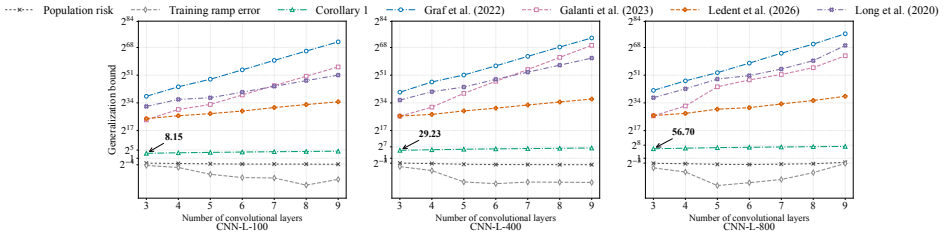


Figure 6: Varying depth L for CNN-L-H architectures trained on CIFAR-10, with different channel widths H. $\lambda = 50$ and $\theta = 0.001$.

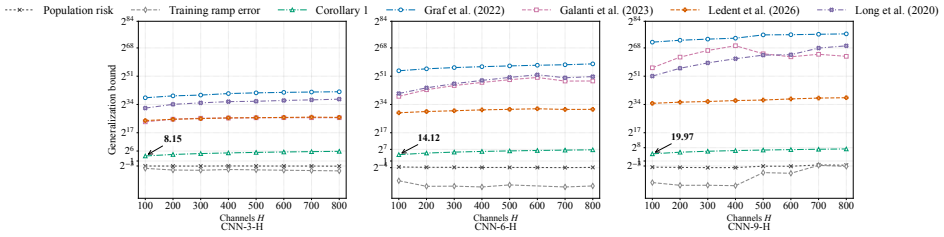


Figure 7: Varying channel width H for CNN-L-H architectures trained on the CIFAR-10 dataset, with different network depths L. $\lambda = 50$ and $\theta = 0.001$.

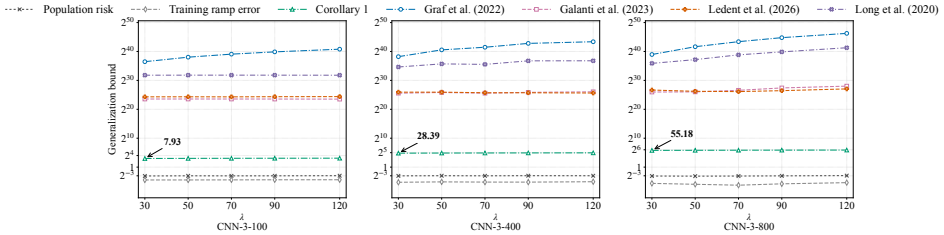


Figure 8: Varying the spectral-norm constraint λ used during bound evaluation for CNN-3-H architectures trained on CIFAR-10, with different channel widths H. $\theta = 0.001$