Stability beyond bounded differences: sharp generalization bounds under finite L_p moments

Qianqian Lei[1]   Soham Bonnerjee[1]   Yuefeng Han[2]

Wei Biao Wu[1]

[1]Department of Statistics, University of Chicago

[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

January 29, 2026

Abstract

While algorithmic stability is a central tool for understanding generalization of learning algorithms, existing high-probability guarantees typically rely on uniform boundedness or sub-Gaussian/sub-Weibull tail assumptions, which can be overly restrictive for modern settings with heavy-tailed or unbounded losses. We develop a stability-based framework that requires only a finite L_p moment condition. Our first contribution is sharp concentration inequalities for functions of independent random variables under L_p constraints, extending McDiarmid's bounded-differences techniques beyond the classical regime. Leveraging these results, we derive sharp high-probability generalization bounds across a range of learning paradigms, including empirical risk minimization, transductive regression, and meta-learning. These guarantees show that L_p stability suffices for robust generalization even when boundedness fails, substantially weakening the standard assumptions in the stability literature.

1 Introduction

Algorithmic stability has emerged as a fundamental tool for the theoretical analysis of machine learning algorithms, providing a principled framework for establishing generalization bounds. Seminal work by [9] established that uniform stability is sufficient to bound the discrepancy between empirical and true risk for symmetric learning algorithms. Unlike complexity-based measures such as VC-dimension [58, 8, 55, 57, 48, 47] and Rademacher complexity [4, 3, 29], stability analysis quantifies how perturbations in training data affect the output hypothesis, thereby providing insight into why machine learning algorithms trained on finite samples generalize to unseen data. Subsequent work has extended these ideas to broader settings, including stochastic optimization [23, 33, 35], federated learning [52, 39], and reinforcement learning [53, 51]. A significant portion of this literature derives high-probability control over generalization error via uniform stability [9, 41, 61, 10, 66] or its relaxations through distribution-dependent surrogates such as sub-Gaussian or sub-exponential conditions [30, 42, 36, 37, 19].

Though powerful, these approaches fail to capture the complexities of modern data science and deep learning. Recent empirical and theoretical studies [12, 44, 50, 40] highlight the inadequacy of sub-Gaussian assumptions for data with heavy tails. Instead, these studies reveal a prevalence of power law behavior–that is, polynomial tail in the large deviation regimes–reminiscent of the

t-distribution. To accommodate such behavior, in this work, we introduce a further relaxation to (L_p, β)-*Lipschitz stability*, which requires only a finite p-th moment rather than uniform constant bounds or tail specifications such as sub-Gaussian or sub-Weibull conditions. Drawing on the classical Nagaev-type inequalities [46] as a powerful alternative when exponential concentration is unattainable, we derive novel sharp concentration bounds for generalization error and apply them to a range of settings, including empirical risk minimization, transductive learning, and meta-learning.

1.1 Main Contributions

Our contributions can be summarized as follows.

- In Section 2, we establish new *sharp* concentration inequalities for general functions of independent random variables under L_p moment. The resulting bounds exhibit a clear two-regime structure: a sub-Gaussian tail governing moderate deviations and a polynomial correction capturing rare large fluctuations. Informally:

  Theorem 1.1 (Theorem 2.2, informal). *Let $x_1, ..., x_n$ be independent random variables, and let x_i' be an independent copy of x_i. Suppose there exist functions f and $\{H_i\}_{i=1}^n$ such that, almost surely, $|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq H_i(x_i, x_i')$. If $H_i(x_i, x_i')$ has a finite p-th moment for $p \geq 2$, then for $y > 0$,*

  $$\mathbb{P}(|f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)]| > y)$$
  $$\lesssim \frac{\sum_{i=1}^n \mathbb{E}[|H_i|^p]}{y^p} + \exp\Big(- \frac{y^2}{\sum_{i=1}^n \mathbb{E}[|H_i|^2]}\Big).$$

  We also develop the result for the complementary heavy-tailed regime $p \in (1, 2)$ in Theorem 2.5. This two-regime behavior arises directly from working under weaker L_p moment assumptions rather than bounded or sub-Weibull conditions, yielding non-asymptotic guarantees that remain valid without exponential moments.

- Using these probabilistic tools, we derive high-probability generalization bounds for three learning paradigms: empirical risk minimization (Section 3.1), transductive regression (Section 3.2), and meta-learning (Section 3.3). For each setting, we formulate an appropriate (L_p, β)-Lipschitz stability notion and obtain corresponding two-regime bounds. The transductive setting requires concentration inequalities for sampling without replacement, developed in Theorem 3.9. Compared to bounded uniform stability, the weaker L_p framework relaxes tail assumptions but imposes stronger decay requirements on the relevant stability notion to achieve vanishing generalization bounds.

- Our numerical experiments confirm the sharpness of the theory and, in particular, highlight the necessity of the polynomial term in our generalization bounds. The results clearly exhibit the predicted transition between the sub-Gaussian and polynomial-tail regimes.

1.2 Related Work

Algorithmic Stability and Generalization. Algorithmic stability offers an algorithm-dependent route to generalization bounds, initiated by [9] who combined uniform stability and bounded losses via McDiarmid's inequality [43]; later works sharpened rates [20, 21, 10, 28, 66] and connected stability to optimization [23, 33, 35]. Since expectation bounds can mask non-negligible probabilities of large deviations for a single learned model, high-probability guarantees

are needed [41, 10, 63], but typically rely on bounded differences. For unbounded stability, existing approaches either tolerate rare large deviations under high-probability boundedness [31, 32, 61], or replace worst-case bounds by distributional surrogates such as sub-Gaussian/sub-exponential assumptions [30, 42, 36, 37, 19], often via Efron-Stein-type inequalities [1], or provide moment bound [64]. Another route avoids exponential moments by robustifying empirical risk (or gradients) using robust mean estimators: Catoni-type M-estimators for robust ERM [12, 11], median-of-means extensions [26, 34], and truncation/clipping methods such as robust gradient descent [24], which can yield exponential-type deviation bounds under low-order moments but modify the learning rule. Our L_p stability framework instead *relaxes tail assumptions directly*, accommodating heavy-tailed regimes where exponential moments may not exist, without changing the learning rule.

Transductive Regression. In transductive learning, the learner observes unlabeled test inputs in advance and predicts only on this fixed set [56]. [15] provided systematic VC dimension bounds for transductive regression, and [16] derived algorithm-dependent generalization bounds via stability analysis. We extend this framework to the L_p setting, requiring new concentration inequalities for sampling without replacement.

Meta-Learning. Meta-learning studies how training on past tasks improves performance on new tasks [7, 54, 41, 25]. Theoretical work focuses on convergence [65, 27, 45] and generalization [5, 6, 17]; see [60] for a review. Stability-based analyses were initiated by [41] and recently advanced by [13, 18, 2, 22] under various meta-stability notions. [60] proposed uniform meta-stability coupling task-level and within-task perturbations with high-probability bounds. Our L_p framework generalizes these results under weaker moment assumptions.

2 Main Results

In this section, as a key technical contribution, we present sharp concentration inequalities that substantially broaden the scope of algorithmic stability. Classical high-probability stability bounds rely on uniform stability [9], which requires uniform boundedness of the loss, or sub-Gaussian/sub-Weibull tail assumptions. We replace these restrictions with a single, weaker requirement: a finite L_p moment for the one-sample perturbation. This relaxation is essential in modern applications, such as deep learning with heavy-tailed gradients or regression over unbounded domains, where the effect of replacing one training sample is not uniformly bounded but can be controlled in L_p norm. Our framework thus provides a powerful theoretical tool for analyzing generalization in learning algorithms.

2.1 Preliminaries

We first introduce the mathematical framework used throughout the paper. Let $x_1, \ldots, x_n$ be independent random variables in a measurable space $\mathcal{X}$, and let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function. We study the concentration of $f(x_1, \ldots, x_n)$ around its expectation under an L_p-type moment condition. Fix $i \in [n]$ and define $g = f(x_1, \ldots, x_i, \ldots, x_n)$ and $g_i = f(x_1, \ldots, x_i', \ldots, x_n)$, where x_i' is an independent copy of x_i. Assume that for some nonnegative measurable functions $H_i : \mathcal{X}^2 \to \mathbb{R}_+$,

$$|g - g_i| \leq H_i(x_i, x_i'). \tag{1}$$

Each H_i is *deterministic*; randomness arises only through the pair (x_i, x_i'). We interpret H_i as measuring the sensitivity to resampling the i-th coordinate: the one-point perturbation bound

(1) is a Lipschitz-type condition with respect to this coordinate-wise distance, consistent with viewing stability assumptions as Lipschitz continuity in an appropriate metric structure [37].

When H_i is a metric (or is dominated by one), it induces the ℓ_1 product metric on $\mathcal{X}^n$, $\rho^{(n)}(x, x') := \sum_{i=1}^{n} H_i(x_i, x'_i)$, yielding a geometric interpretation: modifying a single coordinate incurs cost $H_i(x_i, x'_i)$, while multi-coordinate perturbations accumulate additively, as for Lipschitz functions on product spaces [30]. In unbounded settings, $(\mathcal{X}, H_i)$ may have infinite diameter, so we instead quantify the *distribution-dependent scale* via an independent pair (x_i, x'_i):

$$\|g - g_i\|_p < \|H_i(x_i, x'_i)\|_p < \infty.$$

This L_p-diameter is analogue of Orlicz/sub-Weibull diameters, except we require only a finite p-th moment rather than a ψ_α-norm bound [30, 37]; see also Remark 2.1. We refer to this as L_p-*Lipschitz stability*, which replaces bounded differences with an integrable, distribution-dependent surrogate.

Remark 2.1. The ψ_α Orlicz norm is defined as

$$\|X\|_{\psi_\alpha} = \inf\{\theta > 0 : \mathbb{E}[\exp(|X|/\theta)^\alpha] \leq 2\}. \tag{2}$$

This class includes sub-exponential ($\alpha = 1$) and sub-Gaussian ($\alpha = 2$) variables, but still requires an exponential moment, implying that all L_p moments exist and scale as $O(p^{1/\alpha})$. In contrast, we assume only a finite L_p moment for a fixed p, which is strictly weaker. For instance, heavy-tailed variables such as Pareto distributions with shape parameter p can have finite L_p moments yet fail every sub-Weibull condition due to the nonexistence of the moment-generating function. Therefore, our framework substantially weakens the tail requirements on the stability variable H_i, yielding high-probability guarantees in regimes where sub-Weibull-based stability is inapplicable.

2.2 Concentration Inequalities

With L_p-Lipschitz stability in place, we now state our main concentration bounds. We begin with the case $p \geq 2$, where the tail behavior is captured by a hybrid inequality combining a polynomial (moment) term with a sub-Gaussian term. This form is well suited to settings where the algorithm is typically stable but may exhibit occasional large deviations due to heavy-tailed data, contamination, or irregular loss landscapes.

Theorem 2.2. *Let $x_1, \ldots, x_n$ be independent random variables taking values in a measurable space $\mathcal{X}$, and let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function such that for some nonnegative measurable functions $H_i : \mathcal{X}^2 \to \mathbb{R}_+$, (1) holds. If $\|H_i(x_i, x'_i)\|_p < \infty$ for some $p \geq 2$ and all i, then for all $z > 0$,*

$$\mathbb{P}(|f(x_1, x_2, \ldots, x_n)| - \mathbb{E}[f(x_1, x_2, \ldots, x_n)]| > z)$$
$$\leq c_1 \frac{\sum_{i=1}^{n} \mathbb{E}|H_i(x_i, x'_i)|^p}{z^p} + 2 \exp\left\{-\frac{c_2 z^2}{\sum_{i=1}^{n} \mathbb{E}|H_i(x_i, x'_i)|^2}\right\}, \tag{3}$$

where $c_1 = 3(1 + 2/p)^p$ and $c_2 = 2((p + 2)^2 e^p)^{-1}$ depend only on p.

Remark 2.3. Theorem 2.2 admits an equivalent high-probability form. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|f(x_1, \ldots, x_n) - \mathbb{E}f(x_1, \ldots, x_n)\right|$$
$$\leq c_2^{-1/2} \sqrt{\log\frac{4}{\delta}} \left(\sum_{i=1}^{n} \|H_i(x_i, x'_i)\|_2^2\right)^{1/2} + (2c_1)^{1/p} \delta^{-1/p} \left(\sum_{i=1}^{n} \|H_i(x_i, x'_i)\|_p^p\right)^{1/p}. \tag{4}$$

Using $\left(\sum_{i=1}^n \|H_i\|_p^p\right)^{1/p} \le n^{1/p} \max_i \|H_i\|_p$, the second term in (4) simplifies to a max-type bound:

$$\left|f(x_1,\ldots,x_n) - \mathbb{E}f(x_1,\ldots,x_n)\right|$$
$$\le c_2^{-1/2}\sqrt{\log\frac{4}{\delta}}\left(\sum_{i=1}^n \|H_i(x_i,x_i')\|_2^2\right)^{1/2} + (2c_1)^{1/p}\, n^{1/p}\delta^{-1/p}\max_{1\le i \le n}\|H_i(x_i,x_i')\|_p.$$

The deviation thus decomposes into a sub-Gaussian term governed by $\sum_i \|H_i(x_i,x_i')\|_2^2$ and a heavy-tail correction controlled by the p-th moments.

Remark 2.4. We now compare Theorem 2.2 with several representative concentration results. Considering H_i as a metric function, define the sub-Weibull diameter [37] $\Delta_\alpha(\mathcal{X}_i) = \|H_i(x_i,x_i')\|_{\psi_\alpha}$, where $\|\cdot\|_{\psi_\alpha}$ is the Orlicz norm in (2). [30, Theorem 1] shows that if f is 1-Lipschitz function and sub-Gaussian diameter $\Delta_2(\mathcal{X}_i) < \infty$, then for any $z > 0$,

$$\mathbb{P}(|f - \mathbb{E}f| > z) \le 2\exp\left(-\frac{z^2}{2\sum_{i=1}^n \Delta_2^2(\mathcal{X}_i)}\right).$$

[42, Theorem 11] proves that if f is L-Lipschitz function and sub-exponential diameter $\Delta_1(\mathcal{X}_i) < \infty$, then a one sided bound holds for any $z > 0$,

$$\mathbb{P}(f - \mathbb{E}f > z) \le \exp\left(-\frac{z^2}{4eL^2\sum_{i=1}^n \Delta_1^2(\mathcal{X}_i) + 2e\max_i \Delta_1^2(\mathcal{X}_i)z}\right).$$

More generally, [37, Theorem 2.2] derives sub-Weibull bounds. If $0 < \alpha \le 1$, then

$$\mathbb{P}(|f - \mathbb{E}f| > z) \le \exp\left(-c_\alpha\frac{z^2}{\sum_{i=1}^n \|H_i(x_i,x_i')\|_{\psi_\alpha}^2}\right) + \exp\left(-\frac{z^\alpha}{\max_{1\le i\le n}\|H_i(x_i,x_i')\|_{\psi_\alpha}^\alpha}\right).$$

If $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$, then we have

$$\mathbb{P}(|f - \mathbb{E}f| > z) \le \exp\left(-c_\alpha\frac{z^2}{\sum_{i=1}^n \|H_i(x_i,x_i')\|_{\psi_\alpha}^2}\right) + \exp\left(-\frac{z^\alpha}{(\sum_{i=1}^n \|H_i(x_i,x_i')\|_{\psi_\alpha}^{\alpha^*})^{\alpha/\alpha^*}}\right).$$

In contrast, Theorem 2.2 yields a hybrid bound with a sub-Gaussian component and a polynomial term. For moderate deviations (small z), the second sub-Gaussian term in (3) dominates, matching McDiarmid-type tail behavior and agreeing with [30, 42, 37] up to constants. For large deviations (large z), the first polynomial term dominates, reflecting that our assumptions only impose finite L_p moments; correspondingly, the tail decays polynomially rather than sub-Weibull. Our simulations illustrate this transition between the sub-Gaussian and heavy-tail regimes in two application settings.

We also provide an accompanying results when only lower moments exist.

Theorem 2.5. *Let $x_1,\ldots,x_n$ be independent random variables taking values in a measurable space $\mathcal{X}$, and let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function. For each $i \in [n]$, define $g = f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)$ and $g_i = f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)$, where x_i' is an independent copy of x_i. Assume there exist nonnegative measurable functions $H_i : \mathcal{X}^2 \to \mathbb{R}_+$, such that*

$$|g - g_i| \le H_i(x_i,x_i'), \quad i = 1,\ldots,n,$$

with $\|H_i(x_i,x_i')\|_p < \infty$ for some $1 < p < 2$ and all i. Then, for all $Q \ge 1$ and $z > 0$,

$$\mathbb{P}\left(|f(x_1,\ldots,x_n) - \mathbb{E}(x_1,\ldots,x_n| > z\right)$$
$$\le \sum_{i=1}^n \mathbb{P}(|H_i(x_i,x_i')| > \frac{z}{Q}) + 2\left(\frac{eQ^{p-1}\sum_{i=1}^n \mathbb{E}|H_i(x_i,x_i')|^p}{z^p}\right)^Q$$

Remark 2.6. Theorem 2.5 is essentially sharp for polynomially-tailed variables. Consider $n = 1$ with $H_1(x_1, x_1') \sim t_2$ (Student's t with $\nu = 2$ degrees of freedom). Then $\|H_1(x_1, x_1')\|_p < \infty$ for every $1 < p < 2$.

The t_2 distribution has the closed-form tail

$$\mathbb{P}(|H_1(x_1, x_1')| > \frac{z}{Q}) = 1 - \frac{z/Q}{\sqrt{z^2/Q^2 + 2}} \sim \frac{1}{z^2} \ (z \to \infty),$$

so the first term in the bound is of order z^{-2}. For any fixed $Q > 2/p$, the second term satisfies $O(z^{-pQ}) = o(z^{-2})$. Hence the bound yields an $O(z^{-2})$ tail rate, matching the true exponent of the t_2 distribution.

In particular, under only an L_p increment condition with $p < 2$, one cannot generally expect sub-Gaussian or sub-exponential decay. Compared with the naive Markov bound based solely on $\|H_1\|_p$, which gives order z^{-p}, Theorem 2.5 recovers the correct z^{-2} rate in this canonical heavy-tailed example.

3 Applications

3.1 Empirical Risk Minimization

In this subsection, we illustrate how our concentration results yield high-probability generalization bounds in the standard i.i.d. setting. Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output spaces, and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Consider a training sample

$$S = \{z_1 = (x_1, y_1), \ldots, z_m = (x_m, y_m)\} \sim \mathcal{D}^m,$$

drawn i.i.d. from an unknown distribution $\mathcal{D}$. A learning algorithm is a map $A : \mathcal{Z}^m \to \mathcal{F}$, producing a hypothesis $A_S \in \mathcal{F} \subseteq \mathcal{Y}$. For simplicity, we assume A is deterministic and permutation-invariant in S. We also adopt standard measurability assumptions.

- $S^{\backslash i}$, the leave-one-out sample of size $m - 1$ obtained by removing the i-th observation:

$$S^{\backslash i} = S \setminus z_i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m\}.$$

- S^i, the replacement sample of size m where z_i is replaced by an independent draw $z_i' \sim \mathcal{D}$:

$$S^i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m\}.$$

All expectations and probabilities are taken with respect to the data distribution $\mathcal{D}$. We use subscripts to specify the variables of integration: $\mathbb{E}_S[\cdot]$ denotes the expectation over the training sample $S \sim \mathcal{D}^m$, while $\mathbb{E}_z[\cdot]$ denotes the expectation over a single test instance $z \sim \mathcal{D}$.

To measure the discrepancy between a prediction $f(x)$ and the ground truth y, we use a nonnegative cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. For any hypothesis f and sample $z = (x, y)$, define the loss $\ell(f, z) = c(f(x), y)$. Given a training sample S, the population risk of the learned hypothesis A_S is $R(A, S) = \mathbb{E}_z[\ell(A_S, z)]$, where $z \sim \mathcal{D}$ is an independent test point. Since $\mathcal{D}$ is unknown, $R(A, S)$ cannot be computed directly. We therefore compare it to empirical surrogates: the empirical risk and leave-one-out risk, defined respectively as

$$R_{emp}(A, S) = \frac{1}{m} \sum_{i=1}^m \ell(A_S, z_i), \quad R_{loo}(A, S) = \frac{1}{m} \sum_{i=1}^m \ell(A_{S^{\backslash i}}, z_i).$$

When the algorithm A and sample S are clear from context, we write simply R, R_{emp}, and R_{loo}.

The classical route to high-probability generalization bounds typically relies on deterministic *uniform stability*: the pointwise change in loss under a one-sample perturbation is assumed to be bounded, often together with an almost surely bounded loss. These conditions make McDiarmid-type concentration inequalities straightforward, but they are too restrictive for many modern problems, such as regression with unbounded responses or heavy-tailed noise. Leveraging Theorem 2.2, we instead adopt an (L_p, β)-Lipschitz stability framework.

Assumption 3.1 $((L_p, \beta)$-Lipschitz stability). An algorithm A trained on set $S = \{z_1, \ldots, z_n\}$ is β-Lipschitz stable if the function $f(z_1, \ldots, z_n, z) = \ell(A_S, z)$ with respect to the loss ℓ is β-Lipschitz, and with respect to a measurable function $H : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}_+$ such that for all $i \in \{1, \ldots, m\}$,

$$|\ell(A_S, z) - \ell(A_{S^i}, z)| \leq \beta H(z_i, z_i'),$$

where $\|H(z)\|_p < \infty$ for some $p > 2$, and the norm $\|\cdot\|_p$ is computed with respect to $z \sim \mathcal{D}$.

Assumption 3.1 significantly generalizes the Lipschitz stability notion of [30, 37], and consequently also generalizes the uniform stability of [9], which in turn implies hypothesis stability. It can also be construed as generalizing the notions of argument stability [38, 59], random uniform stability [49].

Assumption 3.2. There exists a measurable function $G : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}_+$, such that for all $S = \{z_1, \ldots, z_m\} \in \mathcal{D}^m$, $z, z' \sim \mathcal{D}$,

$$|\ell(A_S, z) - \ell(A_S, z')| \leq \beta' G(z, z'), \ \beta' > 0,$$

where $\|G(z, z'))\|_p < \infty$ for some $p \geq 2$, and the norm $\|\cdot\|_p$ is computed with respect to $z, z' \sim \mathcal{D}^2$.

Combining these assumptions with Theorem 2.2 yields the following high-probability generalization bounds.

Theorem 3.3. *Suppose learning algorithm A satisfies Assumption 3.1 and the loss function $\ell(A_S, z)$ satisfies Assumption 3.2. Then, for any set S with $|S| = m \geq 1$ and any $y > 0$, the following two bounds hold:*

$$\mathbb{P}(|R - R_{emp}| > y + \beta \mathbb{E}H(z, z'))$$
$$\leq c_1 \frac{m\beta^p \mathbb{E}|H(z, z')|^p + (\beta')^p \mathbb{E}[|G(z, z')|^p]/m^{p-1}}{y^p} + 2\exp\left(-\frac{c_2 y^2}{\beta^2 m \mathbb{E}|H(z, z')|^2 + \frac{(\beta')^2}{m}\mathbb{E}[|G(z, z')|^2}\right),$$

and,

$$\mathbb{P}(|R - R_{loo}| > y + R_m - R_{m-1})$$
$$\leq c_3 \frac{m\beta^p \mathbb{E}[|H(z, z')|^p] + (\beta')^p \mathbb{E}[|G(z, z')|^p]/m^{p-1}}{y^p} + 2\exp\left(-\frac{c_4 y^2}{m\beta^2 \mathbb{E}[|H(z, z')|^2] + \frac{(\beta')^2}{m}\mathbb{E}[|G(z, z')|^2]}\right),$$

where c_1, c_2, c_3, c_4 depend only on p, and $R_m := \mathbb{E}_{S \sim \mathcal{D}^m, z \sim \mathcal{D}}[\ell(A_S, z)] = \mathbb{E}_{S \sim \mathcal{D}^m}[R(A, S)]$.

Remark 3.4. We briefly recall classical generalization bounds for uniformly stable algorithms. A sharp high-probability result is Corollary 8 of [10], which assumes deterministic uniform stability with constant β (i.e., Assumption 3.1 with $H \equiv 1$) and almost sure bounded loss $0 \leq \ell \leq L$. It states that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$|R - R_{emp}| \lesssim \beta \log m \log(1/\delta) + L\sqrt{\frac{\log(1/\delta)}{m}}.$$

This bound is informative only when $\beta = o(1)$; moreover, it is tight (up to logarithmic factors) in the regime $\beta \lesssim m^{-1/2}$, which yields the canonical $m^{-1/2}$ rate. In contrast, Theorem 3.3 allows the stability increment to be random and data-dependent (through $H(z_i, z_i')$) and does not require bounded loss; it suffices that H and G have finite p-th moments for some $p \geq 2$. In particular, with probability at least $1 - \delta$,

$$|R - R_{emp}| \lesssim \left(\beta\|H\|_p m^{1/p} + \beta'\|G\|_p m^{-(1-1/p)}\right)\delta^{-1/p}$$
$$+ \left(\beta\|H\|_2\sqrt{m} + \beta'\frac{\|G\|_2}{\sqrt{m}}\right)\log\frac{1}{\delta} + \beta\|H\|_1.$$

For the bound to vanish, one needs $\|H\|_p = O(1)$, $\|G\|_p = O(1)$ and (up to logarithmic factors) $\beta \ll m^{-1/2}$, $\beta' \ll \sqrt{m}$; smaller δ further strengthens the required decay in β, β' through the factor $\delta^{-1/p}$. In this sense, moving from bounded uniform stability to weaker (L_p, β)-Lipschitz stability trades assumptions for stronger Lipschitz-decay requirements.

3.2 Transductive Regression Algorithms

We consider the transductive learning setting [16], where the learner is given a fixed finite population $\mathcal{X}$ of size $N = m + u$. A training set S of m labeled samples is drawn uniformly at random *without replacement* from $\mathcal{X}$; the remaining u points form the unlabeled test set $T := \mathcal{X} \setminus S$. We denote this random partition by $\mathcal{X} \vdash (S, T)$.

The goal is to predict the labels of the test points in T using only the labeled data in S. Unlike inductive learning, which aims to learn a function that generalizes to arbitrary future inputs, transductive learning only targets a fixed, known set of test inputs. Access to the unlabeled set T during training allows the algorithm to exploit the geometry or manifold structure of the test set to regularize learning and improve prediction.

Let $\ell(h, z) \geq 0$ be a nonnegative loss measuring the error of a hypothesis h on sample $z = (x, y)$; for regression, a canonical choice is the squared loss $\ell(h, z) = (h(x) - y)^2$. Define the training and test (transductive) risks by

$$\hat{R}(h) = \frac{1}{m}\sum_{z \in S}\ell(h, z), \quad R(h) = \frac{1}{u}\sum_{z \in T}\ell(h, z). \tag{5}$$

Our goal is to control the generalization gap $R(h) - \hat{R}(h)$ via stability properties of the algorithm. Classical analyses often assume uniform β-stability, which imposes bounded differences uniformly over all partitions. To accommodate heavy-tailed losses or unbounded domains, we instead adopt an (L_p, β)-Lipschitz stability notion.

Assumption 3.5 (Transductive (L_p, β)-Lipschitz stability)**.** Let A be a transductive learning algorithm. For a partition $\mathcal{X} \vdash (S, T)$, let h be the hypothesis returned by A, and let h' be the hypothesis returned for a modified partition $\mathcal{X} \vdash (S', T')$. We say A is uniformly L_p-Lipschitz stable with respect to the cost function ℓ if there exist nonnegative measurable functions H_i such that, whenever (S', T') is obtained from (S, T) by swapping exactly one point $x_i \in S$ with one point $x_{m+j} \in T$, then for all $x \in \mathcal{X}$,

$$|\ell(h, x) - \ell(h', x)| \leq \beta H(x_i, x_{m+j}),$$

and $\|H(x_i, x_{m+j})\|_p < \infty$ for some $p \geq 2$.

Remark 3.6. As with Assumption 3.1, we do not require a deterministic control. To see this, view $\ell(h, x) = f(x_1, \ldots, x_m, x)$ and $\ell(h', x) = f(x_1, \ldots, x_{i-1}, x_{m+j}, x_{i+1}, \ldots, x_m, x)$; the assumption above is then equivalent to the function $f : \mathcal{X}^{m+1} \to \mathbb{R}$ being (L_p, β)-Lipschitz Stable.

Assumption 3.7 (L_p-bounded hypothesis class). A hypothesis class $\mathcal{H}$ is L_p-bounded with respect to ℓ if there exists a measurable function G such that for all $h \in \mathcal{H}$ and all $x, x' \in \mathcal{X}$,

$$|\ell(h, x) - \ell(h, x')| \leq \beta' G(x, x'), \quad \beta' > 0,$$

where $\|G(x, x')\|_p < \infty$ for some $p \geq 2$, and the norm $\|\cdot\|_p$ is computed with respect to $x, x' \sim \mathcal{D}$.

A key technical challenge in transduction is that $S = \{x_1, \ldots, x_m\}$ is sampled without replacement from the finite population $\mathcal{X}$, which induces dependencies among the training points and precludes applying concentration inequalities for independent variables directly. Prior work addresses this using McDiarmid-type inequalities tailored to sampling without replacement, typically under strict bounded-differences assumptions. In our (L_p, β)-Lipschitz stability regime, we develop corresponding extensions of Theorem 2.2 that accommodate the transductive sampling dependence while requiring only finite L_p moments.

Theorem 3.8. *Let X be a finite set with $|X| = N = m + u$. Let $x_1^m = (x_1, \ldots, x_m)$ be sampled uniformly without replacement from X and let $\phi : X^m \to \mathbb{R}$ be a symmetric measurable function. Assume that for each $i \in [m]$ there exists a measurable function $H : X \times X \to \mathbb{R}$ such that for all $(x_1, \ldots, x_m) \in X^m$ and all $x_i' \in X$, where x_i' is an independent copy of x_i, defining $\varphi = \phi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m)$ and $\varphi_i = \phi(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)$, we have*

$$|\varphi - \varphi_i| \leq H(x_i, x_i').$$

Suppose that for some $p \geq 2$, $\|H(x, x')\|_p < \infty$ for all $i \in [m]$. Then for all $y > 0$,

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > y) \leq c_1 \frac{V_p}{y^p} + 2 \exp\left(- c_2 \frac{y^2}{V_2}\right),$$

where

$$V_p = \frac{u^p}{p-1}\left(\frac{1}{(u - \frac{1}{2})^{p-1}} - \frac{1}{(m + u - \frac{1}{2})^{p-1}}\right)\mathbb{E}|H(x, x')|^p,$$

$$V_2 = \frac{mu}{(m + u - 1/2)(1 - 1/(2\max\{m, u\}))}\mathbb{E}|H(x, x')|^2,$$

and $c_1 = 3(1 + 2/p)^p$ and $c_2 = 2/((p+2)^2 e^p)$ are some constants depending only on p.

We now apply Theorem 3.8 to study stability-based generalization for transductive regression. Our target is the generalization gap between the transductive risk and the training risk, which defined as $\phi(S) := R(S) - \hat{R}(S)$. By controlling $|\mathbb{E}\phi(S)|$ and the one-swap increment $|\phi(S) - \phi(S')|$, where S and S' differ in exactly one point, we can invoke Theorem 3.8 to conclude the following result.

Theorem 3.9. *Let $\mathcal{H}$ be an L_p-bounded hypothesis class and let A be a symmetric transductive (L_p, β)-Lipschitz stability algorithm satisfying Assumption 3.5. Let h be the hypothesis returned by A for a random partition $X \vdash (S, T)$. Then for any $y > 0$, we have*

$$\mathbb{P}\left(R(h) - \hat{R}(h) \geq y + \beta\mathbb{E}H(x_i, x_i')\right) \leq \frac{c_1 V_p'}{y^p} + \exp\left(- \frac{c_2 y^2}{V_2'}\right),$$

where

$$V_p' = \frac{u^p}{p-1}\left(\frac{1}{(u - 1/2)^{p-1}} - \frac{1}{(m + u - 1/2)^{p-1}}\right)\mathbb{E}\left|2\beta H(x, x') + \left|\frac{1}{u} - \frac{1}{m}\right|\beta' G(x, x')\right|^p,$$

$$V_2' = \frac{mu}{(m + u - 1/2)(1 - 1/(2\max\{m, u\}))}\mathbb{E}\left|2\beta H(x, x') + \left|\frac{1}{u} - \frac{1}{m}\right|\beta' G(x, x')\right|^2,$$

and the constants $c_1 = 2(1 + 2/p)^p$ and $c_2 = 2/((p+2)^2 e^p)$ depend only on p.

Remark 3.10. In Appendix § B.2, we extend Theorems 3.8 and 3.9 to allow coordinate-dependent control through functions $H_i(x_i, x_i')$ as in Assumption 3.5. Theorem 3.9 shows that, with probability at least $1 - \delta$,

$$R(h) - \hat{R}(h) \lesssim \sqrt{um}\Big(\beta\|H\|_2 + \frac{|u - m|}{um}\beta'\|G\|_2\Big)\sqrt{\log\frac{1}{\delta}}$$
$$+ (um)^{\frac{1}{p}}\Big(\beta\|H\|_p + \frac{|u - m|}{um}\beta'\|G\|_p\Big)\delta^{-\frac{1}{p}} + \beta\|H\|_1.$$

For comparison, under uniform stability with constant β and almost surely bounded loss $0 \le \ell \le L$, [16] show that with probability at least $1 - \delta$,

$$R(h) - \hat{R}(h) \lesssim \beta + \Big(\beta + \frac{L^2(m + u)}{mu}\Big)\sqrt{m}\sqrt{\log\frac{1}{\delta}}.$$

To achieve vanishing bounds, their result, with significantly stronger assumptions, requires $\beta \ll m^{-1/2}$, whereas ours requires faster decay $\beta \ll (um)^{-1/2}$ with $\|H\|_p = O(1)$.

3.3 Meta-Learning

Consider a (possibly randomized) meta-learning algorithm A acting on a meta-sample $\mathbb{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$, where each task dataset $\mathcal{S}_j = \{z_j^1, \ldots, z_j^n\}$ is drawn independently. Specifically, task distributions $\mathcal{D}_1, \ldots, \mathcal{D}_m$ are sampled i.i.d. from an unknown meta-distribution μ over a measurable space $\mathcal{Z}$, and for each j, the samples $z_j^1, \ldots, z_j^n$ are drawn i.i.d. from $\mathcal{D}_j$. Given $\mathbb{S}$, the meta-learner outputs a task-level learning algorithm $A(\mathbb{S})$, which, when trained on a new task dataset $\mathcal{S} \sim \mathcal{D}^n$ with $\mathcal{D} \sim \mu$, produces a model $A(\mathbb{S})(\mathcal{S}) \in \mathcal{P}$, where $\mathcal{P}$ denotes the model space. We evaluate a model $P \in \mathcal{P}$ on a test point $z \in \mathcal{Z}$ via a loss function $\ell : \mathcal{P} \times \mathcal{Z} \to \mathbb{R}_+$.

In the context of meta-learning, the empirical meta-risk of a meta-learning algorithm A, evaluated on the meta-sample $\mathbb{S}$, is given by

$$R(A(\mathbb{S}), \mathbb{S}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\ell(A(\mathbb{S})(\mathcal{S}_j), z_j^i), \tag{6}$$

and the population meta-risk is

$$R(A(\mathbb{S}), \mu) = \mathbb{E}_{\mathcal{D}\sim\mu}\mathbb{E}_{(\mathcal{S}, z)\sim\mathcal{D}^{n+1}}\ell(A(\mathbb{S})(\mathcal{S}), z). \tag{7}$$

In order to relate the empirical meta-risk (6) and the population meta-risk (7), we impose meta-stability assumptions. We first introduce neighboring datasets. For a task dataset $\mathcal{S} = \{z^1, \ldots, z^n\} \sim \mathcal{D}^n$ and index $i \in [n]$, let $\mathcal{S}^{(i)} = \{z^1, \ldots, z^{i-1}, z^{i'}, z^{i+1}, \ldots, z^n\}$, where $z^{i'} \sim \mathcal{D}$ is an independent copy of z^i. For a meta-sample $\mathbb{S}$ and indices $j \in [m], i \in [n]$, define the neighboring meta-sample $\mathbb{S}^{(j,i)} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{j-1}, \mathcal{S}_j^{(i)}, \mathcal{S}_{j+1}, \ldots, \mathcal{S}_m\}$, so that $\mathbb{S}$ and $\mathbb{S}^{(j,i)}$ differ only by replacing z_j^i with an i.i.d. copy $z_j^{i'} \sim \mathcal{D}_j$. We write $\mathcal{S}^{\backslash i} = \mathcal{S} \setminus \{z^i\}$ and $\mathbb{S}^{\backslash j} = \mathbb{S} \setminus \mathcal{S}_j$.

Assumption 3.11. The meta-learning algorithm A satisfies the following three stability conditions.

(i) *Meta-stability across training tasks.* There exists a measurable function $H : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ with $\mathbb{E}_{\mathcal{D}\sim\mu}\mathbb{E}_{z,z' \overset{i.i.d.}{\sim} \mathcal{D}}[H(z, z')]^p < \infty$ for some $p \ge 2$ such that for all $j \in [m]$ and $i \in [n]$,

$$|\ell(A(\mathbb{S})(\mathcal{S}), z) - \ell(A(\mathbb{S}^{(j,i)})(\mathcal{S}), z)| \le \beta H(z_j^i, z_j^{i'}), \tag{8}$$

(ii) *Within-task stability.* There exists a measurable function $G : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ with

$$\mathbb{E}_{\mathcal{D} \sim \mu} \mathbb{E}_{z,z' \overset{i.i.d.}{\sim} \mathcal{D}} |G(z,z')|^{p'} < \infty$$

for some $p' > 2$ such that for all $i \in [n]$,

$$|\ell(A(\mathbb{S})(\mathcal{S}), z) - \ell(A(\mathbb{S})(\mathcal{S}^{(i)}), z)| \leq \beta' G(z^i, z^{i'}), \tag{9}$$

(iii) *Test-sample stability.* There exists a measurable function $\mathcal{M} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ with $\mathbb{E}_{\mathcal{D} \sim \mu} \mathbb{E}_{z,z' \overset{i.i.d.}{\sim} \mathcal{D}} |\mathcal{M}(z,z')|^{p''} < \infty$ for some $p'' > 2$ such that for all $i \in [n]$,

$$|\ell(A(\mathbb{S})(\mathcal{S}), z) - \ell(A(\mathbb{S})(\mathcal{S}), z')| \leq \beta'' \mathcal{M}(z, z'), \tag{10}$$

Without loss of generality, we take $p' = p'' = p$; otherwise, one can work with $p \wedge p' \wedge p''$.

Remark 3.12. We contrast Assumption 3.11 with the closely related stability assumptions of [41]. First, Assumption 3.11 replaces almost-sure bounded stability gaps by a finite L_p-moment requirement. More importantly, part (i) is strictly weaker: [41] perturbs the meta-sample by replacing an entire task with an i.i.d. copy, whereas we replace only a single within-task sample among the m tasks in $\mathbb{S}$. This smaller, more realistic perturbation yields a more verifiable stability condition, yet remains sufficient for sharp high-probability bounds.

Theorem 3.13. *Let μ denote the task distribution. Given a meta-sample $\mathbb{S}$ and a meta-algorithm A, recall the empirical and population meta-risks $R(A(\mathbb{S}), \mathbb{S})$ and $R(A(\mathbb{S}), \mu)$ from (6) and (7) respectively. Suppose A satisfies Assumption 3.11. Then, for all $y > 0$ it follows that*

$$\mathbb{P}\bigg(R(A(\mathbb{S}), \mu) - R(A(\mathbb{S}), \mathbb{S}) \geq y + \mathbb{E}[H + G] \bigg)$$

$$\leq c_1 \frac{mn \mathbb{E}[|\beta H|^p + nm^{-(p-1)} \mathbb{E}|\beta' G|^p + (mn)^{-(p-1)} \mathbb{E}|\beta'' \mathcal{M}|^p}{y^p}$$

$$+ \exp\bigg(- \frac{c_2 y^2}{mn \mathbb{E}|\beta H|^2 + nm^{-1} \mathbb{E}|\beta' G|^2 + (mn)^{-1} \mathbb{E}|\beta'' \mathcal{M}|^2} \bigg),$$

where $\mathbb{E}$ is taken with respect to $z, z' \overset{i.i.d.}{\sim} \mathcal{D}$, $\mathcal{D} \sim \mu$, and c_1, c_2 are constants solely depending on p.

Remark 3.14. Theorem 3.13 admits an equivalent high-probability form: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(A(\mathbb{S}), \mu) - R(A(\mathbb{S}), \mathbb{S}) \lesssim \mathbb{E}[\beta H + \beta' G] + \delta^{-\frac{1}{p}} \big((mn)^{\frac{1}{p}} \beta \|H\|_p + n^{\frac{1}{p}} m^{\frac{1}{p}-1} \beta' \|G\|_p + (mn)^{\frac{1}{p}-1} \|\mathcal{M}\|_p \big)$$
$$+ \sqrt{\log(1/\delta)} \big(\sqrt{mn} \beta \|H\|_2 n^{\frac{1}{2}} m^{-\frac{1}{2}} \beta' \|G\|_2 + (nm)^{-\frac{1}{2}} \|\mathcal{M}\|_2 \big).$$

For comparison, under uniform meta-stability with constant β, and almost surely bounded loss $0 \leq \ell \leq L$, [60] show that with probability at least $1 - \delta$,

$$R(A(\mathbb{S}), \mu) - R(A(\mathbb{S}), \mathbb{S}) \lesssim \beta \log(mn) \log(1/\delta) + L \sqrt{\log(1/\delta)/(mn)}.$$

To achieve vanishing bounds, their result requires $\beta \ll (\log(mn))^{-1}$, whereas ours requires faster decay $\|H\|_q = O(1), \|G\|_q = O(1), \|\mathcal{M}\|_q = O(1)$, $\beta \ll (mn)^{-1/2}$, $\beta' \ll n^{-1/q} m^{1-1/q}$, and $\beta'' \ll (mn)^{1-1/q}$ for $q = 2, p$. The qualitative message mirrors the supervised case: replacing bounded uniform stability with weaker (L_p, β)-Lipschitz stability relaxes tail assumptions, but requires stronger decay of the stability moments (in m, n and δ) for the generalization bound to vanish.

4 Numerical experiments

In this section, we provide an empirical study highlighting the tightness of our theoretical bounds. Consider i.i.d. observations $S := \{z_i = (x_i, y_i)\}_{i=1}^{m} \in \mathbb{R}^d \times \mathbb{R}$ from the linear model $Y = X\beta + \varepsilon$, where the errors satisfy $\varepsilon \overset{d}{=} U_1^{-1/\nu} - U_2^{-1/\nu}$ with $U_1, U_2 \overset{i.i.d.}{\sim} U[0,1]$. Note that here $p = \nu/2$. We set $d = 5$, $\beta = (1, 1, \ldots, 1)^\top$, and vary $m \in \{500, 1000\}$, and $\nu = \{2.2, 4.4\}$. If the bound in Theorem 3.3 is sharp, the ratio $p(y) := \frac{\mathbb{P}(|R - R_{emp}| > y)}{\mathbb{P}(|R - R_{emp}| > C_0 y)}$ should stabilize near $C_0^{\nu/2}$ for large y. We set $C_0 = 1.5$. To incorporate stability in our analysis, we use ridge regression for empirical risk minimization with regularization parameter $\lambda = 1.0$. Tail probabilities are empirically estimated via $50,000$ Monte Carlo draws. Figure 1 shows that $p(y)$ exhibits initial exponential
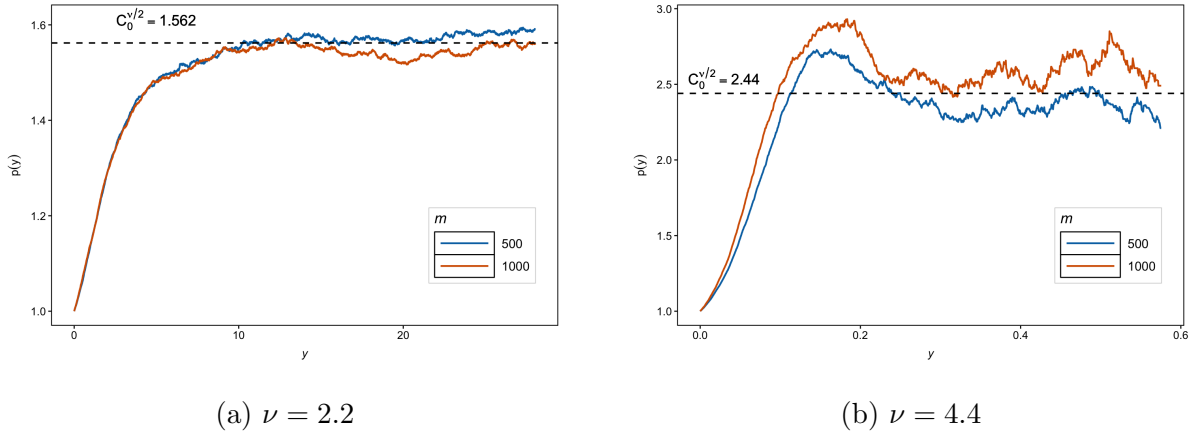

(a) $\nu = 2.2$　　　　　　　　　　　　　　　(b) $\nu = 4.4$

Figure 1: Plot of $p(y)$ versus y; both the curves stabilize around $C_0^{\nu/2}$ for large y.

growth before slowing down and stabilizing near $C_0^{\nu/2}$, further vindicating the importance of the polynomial-in-y term in Theorem 3.3. For larger ν (e.g., $\nu = 4.4$), the Gaussian tail dominates the polynomial tail more strongly at smaller values of y. Consequently, $p(y)$ may exceed the threshold $C_0^{\nu/2}$ initially, before stabilizing in the large-y regime. This behavior confirms the tightness of our results. Additional details and experiments for transductive learning appear in Appendix § C.

5 Conclusion and Future Works

In this work, we provide a systematic treatment of stability analysis under weakened L_p assumptions by establishing a new large-deviation inequality, which may be of independent interest. Theorems 2.2 and 2.5 broaden the scope of stability-based generalization by accommodating settings where the stability increment admits either higher-order finite moments ($p \geq 2$) or only lower-order moments ($p \in (1, 2)$). We develop three representative applications in Section 3, and provide supporting empirical evidence for the sharpness of our bounds in Section 4.

Our results also suggest several natural directions for future research: **(i)** extending the analysis to dependent data (e.g., Markovian sampling) [62, 59]; **(ii)** extending the framework to unsupervised learning problems [14]; and **(iii)** establishing matching lower bounds in specific settings to formally quantify when the polynomial correction term is unavoidable under finite-moment assumptions.

References

[1] K. Abou-Moustafa and C. Szepesvári. An exponential efron-stein inequality for l_q stable learning rules. In *Algorithmic Learning Theory*, pages 31–63. PMLR, 2019.

[2] M. Al-Shedivat, L. Li, E. Xing, and A. Talwalkar. On data efficiency of meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1369–1377. PMLR, 2021.

[3] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

[4] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, Mar. 2003.

[5] J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[6] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 567–580. Springer, 2003.

[7] Y. Bengio, S. Bengio, and J. Cloutier. *Learning a synaptic learning rule*, volume 2. Université de Montréal, Département d'informatique et de recherche . . . , 1990.

[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, Oct. 1989.

[9] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, Mar. 2002.

[10] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 610–626. PMLR, 09–12 Jul 2020.

[11] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[12] O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148 – 1185, 2012.

[13] J. Chen, X.-M. Wu, Y. Li, Q. Li, L.-M. Zhan, and F.-l. Chung. A closer look at the training strategy for modern meta-learning. *Advances in neural information processing systems*, 33:396–406, 2020.

[14] Q. Cheng et al. Algorithmic stability and generalization of an unsupervised feature selection algorithm. *Advances in neural information processing systems*, 34:19860–19875, 2021.

[15] C. Cortes and M. Mohri. On transductive regression. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[16] C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi. Stability of transductive regression algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 176–183, 2008.

[17] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

[18] A. Fallah, A. Mokhtari, and A. Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34:5469–5480, 2021.

[19] J. Fan and Y. Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

[20] V. Feldman and J. Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[21] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 25–28 Jun 2019.

[22] J. Guan, Y. Liu, and Z. Lu. Fine-grained analysis of stability and generalization for modern meta learning algorithms. *Advances in Neural Information Processing Systems*, 35:18487–18500, 2022.

[23] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR.

[24] M. Holland and K. Ikeda. Better generalization with less data using robust gradient descent. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2761–2770. PMLR, 09–15 Jun 2019.

[25] B. Hoskins and U. Fredriksson. *Learning to Learn: What is it and Can it Be Measured?* 2008.

[26] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17(1):543–582, Jan. 2016.

[27] K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

[28] Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

[29] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593 – 2656, 2006.

[30] A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International conference on machine learning*, pages 28–36. PMLR, 2014.

[31] S. Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, Tech. Rep. TR-2002-04*, 2002.

[32] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, page 275–282, 2002.

[33] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2815–2824. PMLR, 10–15 Jul 2018.

[34] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931, 2020.

[35] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5809–5819. PMLR, 13–18 Jul 2020.

[36] S. Li and Y. Liu. Distribution-dependent McDiarmid-type inequalities for functions of unbounded interaction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19789–19810. PMLR, 23–29 Jul 2023.

[37] S. Li, B. Zhu, and Y. Liu. Algorithmic stability unleashed: Generalization bounds with unbounded losses. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29489–29510. PMLR, 21–27 Jul 2024.

[38] T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

[39] Y. Liu, Q. Li, J. Tan, Y. Shi, L. Shen, and X. Cao. Understanding the stability-based generalization of personalized federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 2021.

[41] A. Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(33):967–994, 2005.

[42] A. Maurer and M. Pontil. Concentration inequalities under sub-gaussian and sub-exponential conditions. In *Advances in Neural Information Processing Systems*, volume 34, pages 7588–7597. Curran Associates, Inc., 2021.

[43] C. McDiarmid. *On the method of bounded differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

[44] S. Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 25–39, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[45] K. Mishchenko, S. Hanzely, and P. Richtárik. Convergence of first-order algorithms for meta-learning with moreau envelopes. *arXiv preprint arXiv:2301.06806*, 2023.

[46] S. V. Nagaev. Large deviations of sums of independent random variables. *Ann. Probab.*, 7(5):745–789, 1979.

[47] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[48] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, Dec. 2010.

[49] W. Shen, Z. Yang, Y. Ying, and X. Yuan. Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, 18(05):887–927, 2020.

[50] U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[51] M. J. A. Smith and S. Whiteson. Stability and generalisation in batch reinforcement learning, 2022.

[52] Z. Sun, X. Niu, and E. Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 676–684. PMLR, 02–04 May 2024.

[53] A. Tamar, D. Soudry, and E. Zisselman. Regularization guarantees generalization in bayesian reinforcement learning through algorithmic stability. *ArXiv*, abs/2109.11792, 2021.

[54] S. Thrun and L. Pratt, editors. *Learning to Learn.* Springer, New York, NY, 1 edition, 1998.

[55] V. Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[56] V. Vapnik. *Estimation of dependences based on empirical data.* Springer Science & Business Media, 2006.

[57] V. N. Vapnik. An overview of statistical learning theory. *Trans. Neur. Netw.*, 10(5):988–999, Sept. 1999.

[58] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[59] P. Wang, Y. Lei, Y. Ying, and D.-X. Zhou. Stability and generalization for markov chain stochastic gradient methods. *Advances in Neural Information Processing Systems*, 35:37735–37748, 2022.

[60] Y. Wang and R. Arora. On the stability and generalization of meta-learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[61] L. Warnke. On the method of typical bounded differences. *Combinatorics, Probability and Computing*, 25(2):269–299, 2016.

[62] Z. Yang, Y. Lei, P. Wang, T. Yang, and Y. Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021.

[63] X. Yuan and P. Li. l_2-uniform stability of randomized learning algorithms: Sharper generalization bounds and confidence boosting. *Advances in Neural Information Processing Systems*, 36:78580–78592, 2023.

[64] X. Yuan and P. Li. Exponential generalization bounds with near-optimal rates for l_q-stable algorithms. In *The Eleventh International Conference on Learning Representations*, 2023.

[65] P. Zhou, X. Yuan, H. Xu, S. Yan, and J. Feng. Efficient meta learning via minibatch proximal update. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] S. Zhou, Y. Lei, and A. Kabán. Toward better pac-bayes bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 36:29602–29614, 2023.

Appendices

In this appendix, we provide the proofs of our technical results and additional experimental evidence behind the tightness of our theoretical results. In particular, §A collects the proofs of our main technical results Theorem 2.2 and 2.5. §B.1, §B.2 and §B.3 contains the proofs of results in §3.1, §3.2 and §3.3 respectively. Finally, we provide some additional simulation results in §C.

A   Proofs of Theorems in Section 2

A.1   Technical lemmas

Lemma A.1. *Let $X \in \mathcal{X}$ be a random variable and a function $g : \mathcal{X} \to \mathbb{R}$ satisfying $\mathbb{P}(g(X) > b) = 0, b > 0$, and for any $p \geq 2$*

$$\|g(X)\|_p < \infty.$$

Then for any positive ϕ,

$$\log \mathbb{E} \exp\{\phi g(X)\} \leq \phi \mathbb{E} g(X) + \frac{e^p \phi^2}{2} \mathbb{E}[g(X)^2] + \frac{e^{\phi b} - 1 - \phi b}{b^p} \mathbb{E}[g(X)^p] \times \mathbb{I}\{\phi > p/b\}. \quad (11)$$

Proof of Lemma A.1. By Taylor expansion,

$$\mathbb{E} \exp\{\phi g(X)\} = 1 + \phi \mathbb{E} g(X) + \int_{g(X) \leq b} (e^{\phi g(X)} - 1 - \phi g(X)) \mathrm{d}F(g(X)).$$

Consider 2nd-order Taylor formula with integral remainder,

$$e^x = 1 + x + x^2 \int_0^1 (1 - s) e^{sx} \mathrm{d}s.$$

Then for any $x \leq t$, when $s \in [0, 1]$, $sx \leq x \leq t$, we have

$$e^x - 1 - x = x^2 \int_0^1 (1 - s) e^{sx} \mathrm{d}s \leq x^2 \int_0^1 (1 - s) e^t \mathrm{d}s = \frac{e^t x^2}{2}.$$

Therefore, suppose that $\phi \leq p/b$, then

$$\int_{g(X) \leq b} (e^{\phi g(X)} - 1 - \phi g(X)) \mathrm{d}F(g(X)) \leq \int_{g(X) \leq p/\phi} (e^{\phi g(X)} - 1 - \phi g(X)) \mathrm{d}F(g(X)) \leq \frac{e^p \phi^2}{2} \mathbb{E}[g(X)^2]. \tag{12}$$

Then, for $\phi > p/b$,

$$\int_{g(X) \leq b} (e^{\phi g(X)} - 1 - \phi g(X)) \mathrm{d}F(g(X))$$

$$\leq \int_{g(X) \leq p/\phi} (e^{\phi g(X)} - 1 - \phi g(X)) \mathrm{d}F(g(X)) + \int_{g(X) > p/\phi} \frac{e^{\phi g(X)} - 1 - \phi g(X)}{g(X)^p} g(X)^p \mathrm{d}F(g(X)) \tag{13}$$

Since $\frac{e^{\phi g(X)} - 1 - \phi g(X)}{g(X)^p}$ increase w.r.p to $g(X)$ for $g(X) > p/\phi$, thus

$$\int_{g(X) > p/\phi} \frac{e^{\phi g(X)} - 1 - \phi g(X)}{g(X)^p} g(X)^p \mathrm{d}F(g(X))$$

$$\leq \frac{e^{\phi b} - 1 - \phi b}{b^p} \int_{g(X) > p/\phi} g(X)^p \mathrm{d}F(g(X)) \leq \frac{e^{\phi b} - 1 - \phi b}{b^p} \mathbb{E}[g(X)^p].$$

From Equation (12) and Equation (13), with $\log x \leq x - 1$, we obtain the bound in Equation (11).

$\square$

A.2 Proof of Theorem 2.2

Proof of Theorem 2.2. Let $\mathcal{F}_k = (\ldots, X_{k-1}, X_k)$ and define the projection operator $\mathcal{P}_i(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_i) - \mathbb{E}(\cdot|\mathcal{F}_{i-1})$. For simplicity of notation, denote $\Delta_n = f(X_1, X_2, \ldots, X_n) - \mathbb{E}f(X_1, X_2, \ldots, X_n)$. Then, we can obtain the following decomposition,

$$|\Delta_n| = |\sum_{i \leq n} \mathcal{P}_i(g)| = |\sum_{i \leq n} \mathbb{E}(g - g_i|\mathcal{F}_i)|,$$

where $g = f(X_1, X_2, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)$ and $g_i = f(X_1, X_2, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$ is the coupled version of g obtained by replacing X_i by an i.i.d. copy X_i'.

For any $y = pz/(p+2)$ and $i \in [n]$, define the truncated version of $\tilde{\mathcal{P}}_i(g) = \mathcal{P}_i(g)\mathbf{1}\{|\mathcal{P}_i(g)| \leq y\}$. Consequently, we can obtain the following decomposition

$$\mathbb{P}(|\Delta_n| > z) \leq \mathbb{P}(\sum_{i \leq n} |\tilde{\mathcal{P}}_i(\Delta_n)| \geq z) + \sum_{i=1}^n \mathbb{P}(|\mathcal{P}_i(\Delta_n)| \geq y) =: I_1 + I_2. \tag{14}$$

For I_2 in (14), by Markov inequality, one derives

$$I_2 \leq \frac{\sum_{i=1}^n \mathbb{E}|\mathcal{P}_i(g)|^p}{y^p} \leq \frac{\sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p}{y^p},$$

where the second inequality comes from

$$\mathbb{E}|\mathcal{P}_i(g)|^p = \mathbb{E}|\mathbb{E}(g - g_i|\mathcal{F}_i)|^p \leq \mathbb{E}[\mathbb{E}(|g - g_i|^p|\mathcal{F}_i)] = \mathbb{E}|g - g_i|^p \leq \mathbb{E}|H_i(X_i, X_i')|^p.$$

For the term I_1, we need to control the following moment generating function

$$M(t) = \mathbb{E}\exp\left\{t\sum_{h=1}^n \tilde{\mathcal{P}}_i(g)\right\}, \ t > 0.$$

To that end, observe that for each $h \in [n]$, $|\tilde{\mathcal{P}}_i(g)| \leq y$, by Lemma A.1, we have

$$\mathbb{E}\{\exp(t\tilde{\mathcal{P}}_i(g))|\mathcal{F}_{i-1}\} \leq 1 + \frac{e^p t^2}{2}\mathbb{E}[|\tilde{\mathcal{P}}_i(g)|^2|\mathcal{F}_{i-1}] + \frac{\exp(ty) - 1 - ty}{y^p}\mathbb{E}[|\tilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}] \times 1\{t > p/y\}$$

$$\leq 1 + \frac{e^p t^2}{2}\mathbb{E}|H_i(X_i, X_i')|^2 + \frac{\exp(ty) - 1 - ty}{y^p}\mathbb{E}|H_i(X_i, X_i')|^p \times 1\{t > p/y\},$$

where the upper bound above is independent of $\mathcal{F}_{i-1}$. Therefore, iterated from 1 to n, we can obtain that

$$M(t) \leq \prod_{i=1}^n \left(1 + \frac{e^p t^2}{2}\mathbb{E}|H_i(X_i, X_i')|^2 + \frac{\exp(ty) - 1 - ty}{y^p}\mathbb{E}|H_i(X_i, X_i')|^p \times 1\{t > p/y\}\right)$$

$$\leq \exp\left(\frac{1}{2}e^p t^2 \sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^2 + \frac{\exp(ty) - 1 - ty}{y^p} \sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p \times 1\{t > p/y\}\right).$$

By Chernoff's bound, it is obvious that

$$
\mathbb{P}(|\Delta_n| > z) \leq \frac{\sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^p}{y^p}
$$
$$
+ 2\exp\left(-tz + \frac{1}{2}e^p t^2 \sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^2 + \frac{\exp(ty) - 1 - ty}{y^p} \sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^p \times 1\{t > p/y\} \right)
$$
$$
:= \frac{\sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^p}{y^p} + 2\mathcal{I}_3.
$$

Then, for simplicity of representation, denote $M_2 = \sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^2$, $M_p = \sum_{i=1}^{n} \mathbb{E}|H_i(X_i, X_i')|^p$. Define

$$
h(t) = \frac{e^p t^2}{2} M_2 - tz,
$$
$$
h_1(t) = \frac{e^p t^2}{2} M_2 - \alpha tz, \quad 0 < \alpha < 1,
$$
$$
h_2(t) = \frac{\exp(ty) - 1 - ty}{y^p} M_p - \gamma tz, \quad \gamma = 1 - \alpha.
$$

By taking derivative of $h_1(t)$ and $h_2(t)$ w.r.p to t, then we can get

$$
t_1 = \alpha z / e^p\, M_2, \quad t_2 = \frac{1}{y}\log(\gamma z y^{p-1}/M_p + 1),
$$

where $h_1'(t_1) = 0$ and $h_2'(t_2) = 0$, which implies that $h_1(t)$ and $h_2(t)$ reaches their minimum at t_1 and t_2 separately.

First we examine the case when $t_1 \leq p/y$, we can conclude that

$$
\mathcal{I}_3 \leq \exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\}.
$$

Then, for $t_2 > t_1 \geq p/y$, plug-in $t = t_1$, $\mathcal{I}_3$ becomes

$$
\mathcal{I}_3 \leq \exp\{h_1(t_1) + h_1(t_2)\} \leq \exp\{h_1(t_1)\} = \exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\},
$$

where the second inequality comes from $h_2(t_2) < h_2(t_1) < h_2(0) < 0$ since h_2 is convex and t_2 is the minimizer of f_2.

Consider when $t_1 > t_2 > p/y$

$$
h_1(t_2) + h_2(t_2) < t_2\left(\frac{e^p M_2 t_1}{2} - z \right) + \frac{(e^{t_2 y} - 1)M_p}{y^p}
$$
$$
= \frac{\gamma z}{y} - t_2(1 - \alpha/2)z
$$
$$
= \frac{\gamma z}{y} - \frac{\alpha z t_2}{2} - \gamma z t_2
$$
$$
< \left(\gamma - \frac{p\alpha}{2} \right)\frac{z}{y} - \gamma z t_2,
$$

which lead to

$$\mathcal{I}_3 \leq \exp\left\{ \left(\gamma - \frac{t\alpha}{2}\right)\frac{z}{y} - \gamma\frac{z}{y}\log\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right) \right\}.$$

It now remains to examine when $t_1 > p/y \geq t_2$, we only need to consider the case when plug in $t = p/y$

$$
\begin{aligned}
h(p/y) &< \frac{e^p M_2}{2}\frac{p^2}{y^2} - \frac{p}{y}z \\
&< \frac{p}{y}\left(\frac{e^p M_2}{2}t_1 - z\right) \\
&= \frac{p}{y}\left(\frac{\alpha z}{2} - z\right) \\
&< -\gamma z\frac{p}{y} - \frac{\alpha z p}{2y} \\
&< -\gamma z t_2 - -\frac{\alpha z p}{2y},
\end{aligned}
$$

thus we get

$$\mathcal{I}_3 \leq \exp\left\{ \left(\gamma - \frac{t\alpha}{2}\right)\frac{z}{y} - \gamma\frac{z}{y}\log\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right) \right\}.$$

Combining all the result above, since either

$$\mathcal{I}_3 \leq \exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\},$$

or

$$\mathcal{I}_3 \leq \exp\left\{ \left(\gamma - \frac{t\alpha}{2}\right)\frac{z}{y} - \gamma\frac{z}{y}\log\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right) \right\},$$

holds for all $z > 0$, we have

$$
\begin{aligned}
\mathbb{P}(|\Delta_n| > z) &= \frac{\sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p}{z^p} + 2\mathcal{I}_3 \\
&\leq \frac{\sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p}{z^p} + 2\exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\} + 2\exp\left\{ \left(\gamma - \frac{t\alpha}{2}\right)\frac{z}{y} - \gamma\frac{z}{y}\log\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right) \right\}.
\end{aligned}
$$

By taking $\gamma = \frac{t\alpha}{2}$, above equation can be generalized to

$$
\begin{aligned}
\mathbb{P}(|\Delta_n| > z) &\leq \frac{\sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p}{z^p} + 2\exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\} + 2\exp\left\{ \left(\gamma - \frac{t\alpha}{2}\right)\frac{z}{y} - \gamma\frac{z}{y}\log\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right) \right\} \\
&\leq \frac{\sum_{i=1}^n \mathbb{E}|H_i(X_i, X_i')|^p}{z^p} + 2\exp\left\{ -\frac{\alpha^2 z^2}{2e^p M_2} \right\} + 2\left(\frac{\gamma z y^{p-1}}{M_p} + 1\right)^{-\gamma z/y}.
\end{aligned}
$$

Furthermore, taking $y = \gamma z$, the proof is completed. $\qquad\square$

A.3   Proof of Theorem 2.5

Proof of Theorem 2.5. We follow the same martingale–difference decomposition as in the proof of Theorem 2.2. Let

$$\mathcal{F}_k = \sigma(X_1, \ldots, X_k), \qquad \mathcal{P}_i(g) = \mathbb{E}(g \mid \mathcal{F}_i) - \mathbb{E}(g \mid \mathcal{F}_{i-1}),$$

and write

$$\Delta_n := f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) = \sum_{i=1}^{n} \mathcal{P}_i(g).$$

For a threshold $y > 0$ (to be chosen in terms of z and p below) set

$$\widetilde{\mathcal{P}}_i(g) := \mathcal{P}_i(g)\mathbf{1}\{|\mathcal{P}_i(g)| \leq y\}.$$

Then, as in Equation(14) of the proof of Theorem(2.2), we have the decomposition

$$\mathbb{P}(|\Delta_n| > z) \leq \mathbb{P}\Big(\sum_{i=1}^{n} |\widetilde{\mathcal{P}}_i(g)| \geq z\Big) + \sum_{i=1}^{n} \mathbb{P}\big(|\mathcal{P}_i(g)| > y\big) =: I_1 + I_2.$$

By Markov's inequality and the same conditional Jensen argument, we have

$$\mathbb{P}\big(|\mathcal{P}_i(g)| > y\big) \leq \mathbb{P}(|H_i(X_i, X_i')| > y).$$

Write

$$S_n := \sum_{i=1}^{n} \widetilde{\mathcal{P}}_i(g).$$

Each $\widetilde{\mathcal{P}}_i(g)$ is $\mathcal{F}_i$–measurable, centered ($\mathbb{E}[\widetilde{\mathcal{P}}_i(g) \mid \mathcal{F}_{i-1}] = 0$) and bounded by y. Then, defining $u := \widetilde{\mathcal{P}}_i(g)$,we can derive the following

$$\mathbb{E}[\exp(tu)|\mathcal{F}_{i-1}] \leq 1 + \int_{|u| \leq y} \frac{e^{tu} - 1 - tu}{u^2} u^2 \mathrm{d}F(u|\mathcal{F}_{i-1})$$

$$\leq 1 + \frac{e^{ty} - 1 - ty}{y^2} \int_{|u| \leq y} u^2 \mathrm{d}F(u|\mathcal{F}_{i-1})$$

$$\leq 1 + \frac{e^{ty} - 1 - ty}{y^p} \int_{|u| \leq y} |u|^p \mathrm{d}F(u|\mathcal{F}_{i-1}),$$

where the second inequality comes from the monotonicity of $\frac{e^{tu}-1-tu}{u^2}$ w.r.p to $u \leq y$ and the third inequality comes from monotonicity of $\frac{|u|^p}{y^p}$ w.r.p. to p for $u > 0$.

By $\log x \leq x - 1$, above result can be expressed as

$$\log \mathbb{E} \exp(t\widetilde{\mathcal{P}}_i(g)) \leq \frac{e^{ty} - 1 - ty}{y^p} \mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}].$$

Taking the sum over i, we can further obtain

$$e^{-tz}\mathbb{E}\exp(\sum_{i=1}^{n}\widetilde{\mathcal{P}}_i(g)) \le \exp\left\{\frac{e^{ty}-1-ty}{y^p}\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|\mathcal{F}_{i-1}] - tz\right\}.$$

Setting

$$t = \frac{1}{y}\log\left\{\frac{zy^{p-1}}{\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}]} + 1\right\},$$

the right hand side becomes

$$e^{-tz}\mathbb{E}\exp(\sum_{i=1}^{n}\widetilde{\mathcal{P}}_i(g)) \le \exp\left\{\frac{z}{y} - \left(\frac{z}{y} + \frac{\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}]}{y^p}\right)\log\left(\frac{zy^{p-1}}{\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}]} + 1\right)\right\}$$

$$\le \exp\left\{\frac{z}{y}\left(-\log\left(\frac{zy^{p-1}}{\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}]}\right)\right)\right\}$$

$$= \left(\frac{e\sum_{i=1}^{n}\mathbb{E}[|\widetilde{\mathcal{P}}_i(g)|^p|\mathcal{F}_{i-1}]}{zy^{p-1}}\right)^{z/y}.$$

Combining the bounds for I_1 and I_2 and plug in $y = \frac{z}{Q}$ for any $Q > 1$, the proof is completed. $\square$

B Proofs of Theorems in Section 3

B.1 Proof of Results in Section 3.1: Empirical Risk Minimization

Proof of Theorem 3.3. Clearly,

$$|R - R^i| \le |\mathbb{E}_z[\ell(A_S, z) - \ell(A_{S^i}, z)]| \le \beta H(z_i, z_i'). \tag{15}$$

On the other hand,

$$|R_{loo} - R_{loo}^i| \le \frac{1}{m}\sum_{j\ne i}|\ell(A_{S\backslash j}, z_j) - \ell(A_{S^i,\backslash j}, z_j)| + \frac{1}{m}|\ell(A_{S\backslash i}, z_i) - \ell(A_{S\backslash i}, z_i')|$$

$$\le \frac{m-1}{m}\beta H(z_i, z_i') + \frac{1}{m}\beta' G(z_i, z_i'). \tag{16}$$

Therefore, with $\phi = R - R_{loo}$ and $\phi^i = R^i - R_{loo}^i$, (15) and (16) yields

$$|\phi - \phi^i| \le \frac{2m-1}{m}\beta H(z_i, z_i') + \frac{1}{m}\beta' G(z_i, z_i').$$

Moreover,

$$\mathbb{E}_S[R - R_{loo}] = R_n - \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_S[\ell(A_{S\backslash j}, z_j)] = R_n - R_{n-1}.$$

Applying Theorem 2.2, we can get

$$\mathbb{P}(|R-R_{loo}| > y+2\beta\mathbb{E}H(z_i)) \le c_1 \frac{m\mathbb{E}|\frac{2m-1}{m}\beta H(z,z') + \beta'\frac{G(z,z')}{m}|^p}{y^p} + 2\exp\{-c_2 \frac{y^2}{m\mathbb{E}|\frac{2m-1}{m}\beta H(z,z') + \beta'\frac{G(z,z')}{m}|^2}\}.$$

Note that by Hölder's inequality, we can have for $q \in \{2,p\}$

$$\mathbb{E}\left|\frac{2m-1}{m}\beta H(z_i) + \frac{G}{m}\right|^q \le 2^q\mathbb{E}\left[\left(\frac{2m-1}{m}\beta\right)^q|H(z_i)|^q + \frac{1}{m^q}|\beta'G|^q\right]$$
$$\le 2^q\left((2\beta)^q\mathbb{E}|H(z_i)|^q + \frac{(\beta')^q}{m^q}\mathbb{E}|G|^q\right),$$

thus taking sum over 1 to m, we can obtain Equation (5).

For the R_{emp}, we proceed similarly. We have

$$|R_{emp} - R_{emp}^i| \le \frac{1}{m}\sum_{j\ne i}|\ell(A_S,z_j) - \ell(A_{S^i},z_j)| + \frac{1}{m}|\ell(A_S,z_i) - \ell(A_{S^i},z_i)| + \frac{1}{m}|\ell(A_{S^i},z_i) - \ell(A_{S^i},z_i')|$$
$$\le \beta H(z_i,z_i') + \beta'\frac{G(z_i,z_i')}{m}$$

and

$$\mathbb{E}_S[R - R_{emp}] = \mathbb{E}_{S,z_i'}[\ell(A_S,z_i')] - \frac{1}{m}\sum_{j=1}^m \mathbb{E}_S[\ell(A_S,z_j)]$$
$$\le \frac{1}{m}\sum_{j=1}^m (\mathbb{E}_{S,z_i'}[\ell(A_{S^j},z_j)] - \mathbb{E}_{S,z_i'}[\ell(A_S,z_j)]) \quad \text{(where } S^j = (z_1,\ldots,z_{j-1},z_i',z_{j+1,\ldots,z_n}))$$
$$\le \beta\mathbb{E}[H(z,z')].$$

So the Theorem 2.2 can again be applied and the proof is completed.

$\square$

B.2　Proof of Results in Section 3.2: Transductive Regression Algorithms

Proof of Theorem 3.8. Let $\mathcal{F}_i := \sigma(x_1,\ldots,x_i)$ and define the Doob martingale

$$M_i := \mathbb{E}[\varphi \mid \mathcal{F}_i], \qquad i = 0,1,\ldots,m,$$

so that $M_0 = \mathbb{E}\varphi$ and $M_m = \varphi$. Let the martingale differences be

$$D_i := M_i - M_{i-1}, \qquad i = 1,\ldots,m,$$

so $\varphi - \mathbb{E}\varphi = \sum_{i=1}^m D_i$ and $\mathbb{E}[D_i \mid \mathcal{F}_{i-1}] = 0$.

For fixed i, consider $x_i \in X$,
$$G_i(x_i) := \mathbb{E}[\varphi \mid \mathcal{F}_{i-1}, x_i].$$
Then $M_i = G_i(x_i)$ and $M_{i-1} = \mathbb{E}[G_i(x_i) \mid \mathcal{F}_{i-1}]$, hence

$$D_i = G_i(x_i) - \mathbb{E}[G_i(x_i) \mid \mathcal{F}_{i-1}].$$

Then, Consider the sequence of x_i^m , since the function ϕ is symmetric, any permutations containing same elements may lead to same value. Thus, we only need to consider for the case that x_i is not included in sequence x_{i+1}^m, and the number of those cases are $(m-i)!\binom{N-i-1}{m-i}$

Thus, we have that for all $x_i, x_i' \in R_{i-1}$,

$$|D_i| \leq \frac{u!}{(N-i)!}(m-i)!\binom{N-i-1}{m-i}|\varphi(\mathrm{x}_1^{i-1},\mathrm{x}_i,\mathrm{x}_{i+1}^m) - \varphi(\mathrm{x}_1^{i-1},\mathrm{x}_i',\mathrm{x}_{i+1}^m)| \leq \frac{u}{N-i}H(x_i,x_i')$$

Then, by similar proof as Theorem 2.2, we have

$$\mathbb{P}\bigg(\sum_{i=1}^m D_i > z\bigg) \leq 3\left(1+\frac{2}{p}\right)^p\frac{S_p}{z^p} + \exp\left(-\frac{2}{(p+2)^2 e^p}\frac{z^2}{S_2}\right),$$

where

$$S_p = \mathbb{E}|H(x_i,x_i')|_p^p\sum_{i=1}^m\left(\frac{u}{N-i}\right)^p.$$

Then, consider for $p=2$,

$$\sum_{i=1}^m\left(\frac{u}{N-i}\right)^2 \leq \frac{mu}{(N-1/2)(1-1/2\max\{m,u\})},$$

for $p > 2$,

$$\sum_{i=1}^m\left(\frac{u}{m+n-i}\right)^p \leq \frac{u^p}{p-1}\left(\frac{1}{(u-1/2)^{p-1}} - \frac{1}{(m+u-1/2)^{p-1}}\right)$$

The proof is complete.

$\square$

Before concluding the proof of Theorem 3.9, we bound $|\mathbb{E}\phi(S)|$ and the one-swap increment $|\phi(S) - \phi(S')|$, where S and S' differ in exactly one point.

Lemma B.1. *Let $\mathcal{H}$ be an L_p-bounded hypothesis class and let A be an L_p-stable algorithm. Let h and h' be the hypotheses returned by A when trained on $S = \{x_1,\ldots,x_i,\ldots,x_m\}$ and $S' = \{x_1,\ldots,x_i',\ldots,x_m\}$, which differ in exactly one point, where x_i' is an independent copy of x_i for any $i \in [1,m]$. Define $\phi(S) = R(h) - \hat{R}(h)$. Then, for any $p \geq 2$,*

$$|\phi(S) - \phi(S')| \leq 2\beta H(x,x') + \left|\frac{1}{u} - \frac{1}{m}\right|\beta'G(x,x').$$

Lemma B.2. *Let h be the hypothesis returned by an L_p-stable algorithm A trained on $S = \{x_1,\ldots,x_m\}$. Then,*

$$|\mathbb{E}_S[\phi(S)]| \leq \frac{1}{m}\sum_{i=1}^m\mathbb{E}H(x_i,x_i').$$

Proof of Lemma B.1. WLOG, we assume $x_i' = x_{m+j}$ for $j \in [1, u]$. By definition and assumption, we have

$$|\phi(S) - \phi(S')|_p = \left| \frac{1}{u} \sum_{k \in [1,u], k \neq j} \left(\ell(h(S), x_{m+k}) - \ell(h(S'), x_{m+k}) \right) - \frac{1}{m} \sum_{\ell \in [1,m], \ell \neq i} \left(\ell(h(S), x_\ell) - \ell(h(S'), x_\ell) \right) \right.$$

$$\left. + \frac{1}{u} \left(\ell(h(S), x_{m+j}) - \ell(h(S'), x_i) \right) - \frac{1}{m} \left(\ell(h(S), x_i) - \ell(h(S'), x_{m+j}) \right) \right|$$

$$\leq \frac{u-1}{u} H(x, x') + \frac{m-1}{m} H(x, x')$$

$$+ \min\{\frac{1}{u}, \frac{1}{m}\} \left(\left| \ell(h(S), x_{m+j}) - \ell(h(S'), x_{m+j}) \right| + \left| \ell(h(S), x_i) - \ell(h(S'), x) \right| \right)$$

$$+ \left| \frac{1}{u} - \frac{1}{m} \right| \times \left(\left| \ell(h(S), x_{m+j}) - \ell(h(S), x_i) \right| + \left| \ell(h(S), x_i) - \ell(h(S'), x_i) \right| \right)$$

$$\leq 2\beta H(x, x') + \left| \frac{1}{u} - \frac{1}{m} \right| \beta' G(x, x')$$

$\square$

Proof of Lemma B.2. By definition of $\mathbb{E}[\phi(S)]$,

$$\mathbb{E}[\phi(S)] = \mathbb{E}\left[\frac{1}{u} \sum_{j=1}^{u} \ell(h_S, x_{m+j}) \right] - \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \ell(h_S, x_i) \right]$$

$$= \frac{1}{u} \sum_{j=1}^{u} \mathbb{E}[\ell(h_S, x_{m+j})] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[\ell(h_S, x_i)]$$

$$= \mathbb{E}[\ell(h_S, x_{m+j})] - \mathbb{E}[\ell(h_S, x_i)]$$

$$= \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \ell(h_{S^{(j)}}, x_i) \right] - \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \ell(h_S, x_i) \right]$$

$$= \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[\ell(h_{S^{(j)}}, x_i) - \ell(h_S, x_i)]$$

$$\leq \frac{\beta}{m} \sum_{j=1}^{m} \mathbb{E} H(x_i, x_i')$$

$\square$

Proof of Theorem 3.9. The result follows directly from Theorem 3.8 and Lemmas B.1 and B.2. $\square$

Here, we give the a more general result of Transductive regression algorithm. Instead of assuming a single universal bound $H(x_i, x_i')$ as in Assumption 3.5, we allow a weaker, potentially coordinate-dependent control through functions $H_i(x_i, x_i')$.

Assumption B.3 (Transductive L_p-Lipschitz stability). Let A be a transductive learning algorithm. For a partition $\mathcal{X} \vdash (S, T)$, let h be the hypothesis returned by A, and let h' be the hypothesis returned for a modified partition $\mathcal{X} \vdash (S', T')$. We say A is uniformly L_p-stable with respect to the cost function ℓ if there exist nonnegative measurable functions H_i such that, whenever (S', T') is obtained from (S, T) by swapping exactly one point $x_i \in S$ with one point $x_{m+j} \in T$, then for all $x \in \mathcal{X}$,

$$|\ell(h, x) - \ell(h', x)| \leq H_i(x_i, x_{m+j}),$$

and $\|H_i(x_i, x_{m+j})\|_p < \infty$ for some $p \geq 2$.

Under this transductive L_p stability notion, Theorems 3.8 and 3.9 extend to Theorems B.4 and B.5.

Theorem B.4. *Let X be a finite set with $|X| = N = m + u$. Let $x_1^m = (x_1, \ldots, x_m)$ be sampled uniformly without replacement from X and let $\phi : X^m \to \mathbb{R}$ be a measurable function. Assume that for each $i \in [m]$ there exists a measurable function $H_i : X \times X \to \mathbb{R}$ such that for all $(x_1, \ldots, x_m) \in X^m$ and all $x_i' \in X$, where x_i' is an independent copy of x_i, defining $\varphi = \phi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m)$ and $\varphi_i = \phi(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)$, we have*

$$|\varphi - \varphi_i| \leq H_i(x_i, x_i')$$

Suppose that for some $p \geq 2$, $\|H_i(x_i, x_i')\|_p < \infty$ for all $i \in [m]$. Then for all $y > 0$,

$$\mathbb{P}\Big(|\varphi - \mathbb{E}\varphi| > y\Big) \leq c_1 \frac{V_p}{y^p} + 2\exp\Big(-c_2 \frac{y^2}{V_2}\Big),$$

where for $q \in \{2, p\}$,

$$V_q := \Big(1 + \frac{m}{N}\Big)^{q-1}\Big(1 + \log\Big(\frac{N}{u}\Big)\Big) \sum_{i=1}^m \|H_i(x_i, x_i')\|_q^q,$$

and $c_1 = 3(1 + 2/p)^p$ and $c_2 = 2/((p+2)^2 e^p)$ are some constants depending only on p.

Theorem B.5. *Let $\mathcal{H}$ be a L_p-bounded hypothesis class and let A be a transductive L_p stability algorithm satisfying Assumption B.3. Let h be the hypothesis returned by A for a random partition $X \vdash (S, T)$. Then for any $y > 0$, we have*

$$\mathbb{P}\Big(R(h) - \hat{R}(h) \geq y + \frac{1}{m}\sum_{i=1}^m \mathbb{E}H_i(x_i, x_i')\Big)$$

$$\leq c_1 \frac{V_p}{y^p} + \exp\Big(-c_2 \frac{y^2}{V_2}\Big),$$

where for $q \in \{2, p\}$,

$$V_q = \Big(1 + \frac{m}{N}\Big)^{q-1}\Big(1 + \log\Big(\frac{N}{u}\Big)\Big)$$

$$\times \sum_{i=1}^m \mathbb{E}\Big|2H_i(x_i, x_i') + \Big|\frac{1}{m} - \frac{1}{u}\Big| G(x_i, x_i')\Big|^q,$$

and the constants $c_1 = 2(1 + 2/p)^p$ and $c_2 = 2/((p+2)^2 e^p)$ depend only on p.

Proof of Theorem B.4. We do the similar decomposition as in Proof of Theorem B.4

$$G_i(x_i) := \mathbb{E}[\varphi \mid \mathcal{F}_{i-1}, x_i],$$

and

$$D_i = G_i(x_i) - \mathbb{E}[G_i(x_i) \mid \mathcal{F}_{i-1}].$$

Let $R_{i-1} := X \setminus \{x_1, \ldots, x_{i-1}\}$ denote the remaining pool after the first $i - 1$ draws, so $|R_{i-1}| = N - i + 1$. Conditional on $\mathcal{F}_{i-1}$, x_i is uniform on R_{i-1}. Denote $\mathrm{x}_{i=k}^\ell = \{x_k, x_{k+1}, \ldots, x_\ell\}$

and S_1^m a sequence of random variables $S_1, \ldots, s_m$. We write $S_i^m = \mathrm{x}_i^m$ as a shorthand for the m equalities $S_i = x_i, i = 1, \ldots, m$ and $\mathbb{P}(\mathrm{x}_{i+1}^m | \mathrm{x}_1^{i-1}, x_i) = \mathbb{P}(S_{i+1}^m = \mathrm{x}_{i+1}^m | S_1^{i-1} = \mathrm{x}_1^{i-1}, S_i = x_i)$. Let x_i' be an independent copy of x_i conditional on $\mathcal{F}_{i-1}$, which means conditional on $\mathcal{F}_{i-1}$, x_i' is also uniform on R_{i-1} and independent of x_i. Then $\mathbb{E}[G_i(x_i) \mid \mathcal{F}_{i-1}] = \mathbb{E}[G_i(x_i') \mid \mathcal{F}_{i-1}]$. Conditioning further on $(\mathcal{F}_{i-1}, x_i)$ gives the key identity

$$D_i = \mathbb{E}\left[G_i(x_i) - G_i(x_i') \mid \mathcal{F}_{i-1}, x_i\right].$$

Then, consider fix $\mathcal{F}_{i-1}$ and fix $x_i, x_i' \in R_{i-1}$. Under $\mathcal{F}_{i-1}$ and $S_i = x_i$, the remaining coordinates $(S_{i+1}, \ldots, S_m)$ are sampled uniformly without replacement from $R_{i-1} \setminus \{x_i\}$; similarly under $S_i = x_i'$ they are sampled from $R_{i-1} \setminus \{x_i'\}$. By definition of conditional expectation, we have

$$D_i = \sum_{\mathrm{x}_{i+1}^m} \phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^m) \mathbb{P}(\mathrm{x}_{i+1}^m | \mathrm{x}_1^{i-1}, x_i) - \sum_{\mathrm{x}_{i+1}'^m} \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}'^m) \mathbb{P}(\mathrm{x}_{i+1}'^m | \mathrm{x}_1^{i-1}, x_i').$$

Note that

$$\mathbb{P}(\mathrm{x}_{i+1}^m | \mathrm{x}_1^{i-1}, x_i) = \prod_{k=i}^{m-1} \frac{1}{N-i} = \frac{u!}{(N-i)!} = \mathbb{P}(\mathrm{x}_{i+1}'^m | \mathrm{x}_1^{i-1}, x_i'),$$

we can obtain

$$D_i = \frac{u!}{(N-i)!} \left(\sum_{\mathrm{x}_{i+1}^m} \phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^m) - \sum_{\mathrm{x}_{i+1}'^m} \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}^m) \right).$$

Consider the sequence of x_{i+1}^m and $\mathrm{x}_{i+1}'^m$, We can find bijection relationship among the two sequence: i) $\mathrm{x}_j = x_i', \mathrm{x}_j' = x_i$ and $\mathrm{x}_{i+1}^m \setminus \mathrm{x}_j = \mathrm{x}_{i+1}'^m \setminus \mathrm{x}_j'$; ii) $x_i' \notin \mathrm{x}_{i+1}^m, x_i \notin \mathrm{x}_{i+1}'^m$ and $\mathrm{x}_{i+1}^m = \mathrm{x}_{i+1}'^m$.

For the first case, for any fixed $j \in [i+1, m]$, the number of the permutations are $(m-i-1)!\binom{N-i-1}{m-i-1}$, thus we have

$$\left| \sum_{j=i+1}^{m} \left(\sum_{\mathrm{x}_{i+1}^m \setminus \mathrm{x}_j = \mathrm{x}_{i+1}'^m \setminus \mathrm{x}_j'} \phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^{j-1}, x_i', \mathrm{x}_{j+1}^m) - \sum_{\mathrm{x}_{i+1}^m \setminus \mathrm{x}_j = \mathrm{x}_{i+1}'^m \setminus \mathrm{x}_j'} \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}'^{j-1}, x_i, \mathrm{x}_{j+1}'^m) \right) \right|$$

$$\leq \sum_{j=i+1}^{m} (m-i-1)!\binom{N-i-1}{m-i-1} \left| \phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^{j-1}, x_1', \mathrm{x}_{j+1}^m) - \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}'^{j-1}, x_i, \mathrm{x}_{j+1}'^m) \right|$$

$$\leq \sum_{j=i+1}^{m} (m-i-1)!\binom{N-i-1}{m-i-1} \left(H_i(x_i, x_i') + H_j(x_i, x_i') \right).$$

For the second case, the number of those permutations are $(m-i)!\binom{N-i-1}{m-i}$, thus we can claim that,

$$\sum_{\mathrm{x}_{i+1}^m : x_i' \notin \mathrm{x}_{i+1}^m} \phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^m) - \sum_{\mathrm{x}_{i+1}'^m : x_i \notin \mathrm{x}_{i+1}'^m} \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}^m)$$

$$= (m-i)!\binom{N-i-1}{m-i} |\phi(\mathrm{x}_1^{i-1}, x_i, \mathrm{x}_{i+1}^m) - \phi(\mathrm{x}_1^{i-1}, x_i', \mathrm{x}_{i+1}^m)|$$

$$\leq (m-i)!\binom{N-i-1}{m-i} H_i(x_i, x_i').$$

Combining above result, we can derive

$$|D_i| \leq \frac{u!}{(N-i)!} \left[(m-i)! \binom{N-i-1}{m-i} H_i(x_i, x_i') + \sum_{j=i+1}^{m} (m-i-1)! \binom{N-i-1}{m-i-1} \left(H_i(x_i, x_i') + H_j(x_i, x_i') \right) \right]$$

$$\leq \frac{u}{N-i} H_i(x_i, x_i') + \sum_{j=i+1}^{m} \frac{1}{N-i} \left(H_i(x_i, x_i') + H_j(x_i, x_i') \right)$$

$$= H_i(x_i, x_i') + \sum_{j=i+1}^{m} \frac{1}{N-i} H_j(x_i, x_i')$$

Then, by similar proof as Theorem 2.2, we have

$$\mathbb{P}\left(\sum_{i=1}^{m} D_i > z \right) \leq 3 \left(1 + \frac{2}{p} \right)^p \frac{S_p}{z^p} + \exp\left(-\frac{2}{(p+2)^2 e^p} \frac{z^2}{S_2} \right)$$

where

$$S_p = \sum_{i=1}^{m} \mathbb{E} \left| H_i(x_i, x_i') + \sum_{j=i+1}^{m} \frac{1}{N-i} H_j(x_i, x_i') \right|^p.$$

By Hölder's inequality, we can obtain

$$\mathbb{E} \left| H_i(x_i, x_i') + \sum_{j=i+1}^{m} \frac{1}{N-i} H_j(x_i, x_i') \right|^p \leq \left(1 + \sum_{j=i+1}^{m} \frac{1}{N-i} \right)^{p-1} \left[\mathbb{E}|H_i(x_i, x_i')|^p + \left(\frac{1}{N-i} \right) \sum_{j=i+1}^{m} \mathbb{E}|H_j(x_i, x_i')|^p \right]$$

$$\leq \left(1 + \frac{m}{N} \right)^{p-1} \left[\mathbb{E}|H_i(x_i, x_i')|^p + \left(\frac{1}{N-i} \right) \sum_{j=i+1}^{m} \mathbb{E}|H_j(x_i, x_i')|^p \right].$$

By taking sum over 1 to m,

$$S_p \leq \sum_{i=1}^{m} \left(1 + \frac{m}{N} \right)^{p-1} \left[\mathbb{E}|H_i(x_i, x_i')|^p + \left(\frac{1}{N-i} \right) \sum_{j=i+1}^{m} \mathbb{E}|H_j(x_i, x_i')|^p \right]$$

$$\leq \left(1 + \frac{m}{N} \right)^{p-1} \sum_{i=1}^{m} \left[\mathbb{E}|H_i(x_i, x_i)|^p + \sum_{j=1}^{m} \mathbb{E}|H_j(x_j, x_j')|^p \left(\sum_{i=j-1}^{m} \frac{1}{N-i} \right) \right]$$

$$\leq \left(1 + \frac{m}{N} \right)^{p-1} \left[\sum_{i=1}^{m} \mathbb{E}|H_i(x_i, x_i)|^p + \sum_{j=1}^{m} \mathbb{E}|H_j(x_j, x_j')|^p \times \log\left(\frac{N}{u} \right) \right]$$

$$\leq \left(1 + \frac{m}{N} \right)^{p-1} \left(1 + \log\left(\frac{N}{u} \right) \right) \sum_{i=1}^{m} \mathbb{E}|H_i(x_i, x_i')|^p$$

The proof is completed.

$$\square$$

Proof of Theorem B.5. The result follows directly from Theorem B.4 and similar versions of Lemmas B.1 and B.2. $\square$

B.3 Proof of Results in Section 3.3: Meta-Learning

Proof of Theorem 3.13. In the following, we suppress β, β' and β'' inside the functions H, G and $\mathcal{M}$. We follow a similar, but slightly more convoluted strategy compared to that of Theorem 3.3. Observe that, for $k \in [m], l \in [n]$, it holds almost surely that

$$mn\big|R(A(\mathbb{S}), \mathbb{S}) - R(A(\mathbb{S}^{(k,l)}), \mathbb{S}^{(k,l)})\big|$$

$$\leq \sum_{j \neq k}^{m} \sum_{i=1}^{n} |\ell(A(\mathbb{S})(\mathcal{S}_j), z_j^i) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_j), z_j^i)| + \sum_{i \neq l}^{n} |\ell(A(\mathbb{S})(\mathcal{S}_k), z_k^i) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)}), z_k^i)|$$

$$+ |\ell(A(\mathbb{S})(\mathcal{S}_k), z_k^l) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)}), z_k^{l'})|$$

$$\overset{(a)}{\leq} (m-1)n \; H(z_k^l, z_k^{l'}) + \sum_{i \neq l}^{n} \Big(|\ell(A(\mathbb{S})(\mathcal{S}_k), z_k^i) - \ell(A(\mathbb{S})(\mathcal{S}_k^{(l)}), z_k^i)| + |\ell(A(\mathbb{S})(\mathcal{S}_k^{(l)}), z_k^i) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)}), z_j^i)| \Big)$$

$$+ |\ell(A(\mathbb{S})(\mathcal{S}_k), z_k^l) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k), z_k^l)| + |\ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k), z_k^l) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)})), z_k^l)|$$

$$+ |\ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)}), z_k^l) - \ell(A(\mathbb{S}^{(k,l)})(\mathcal{S}_k^{(l)})), z_k^{l'})|$$

$$\overset{(b)}{\leq} mn \; H(z_k^l, z_k^{l'}) + n \; G(z_k^l, z_k^{l'}) + \mathcal{M}(z_k^l, z_k^{l'}), \tag{17}$$

where (a) follows from (8), and (b) follows from (9) and (10) . On the other hand, it is trivial to note that

$$|R(A(\mathbb{S}), \mu) - R(A(\mathbb{S}^{(k,l)}), \mu)| \leq H(z_k^l, z_k^{l'}) \text{ almost surely.} \tag{18}$$

Let $g \equiv R(A(\mathbb{S}), \mathbb{S}) - R(A(\mathbb{S}), \mu)$, and denote by $g_{k,l} \equiv R(A(\mathbb{S}^{(k,l)}), \mathbb{S}^{(k,l)}) - R(A(\mathbb{S}^{(k,l)}), \mu)$. Combining (17)-(18) provides that

$$|g - g_{k,l}| \leq 2 \; H(z_k^l, z_k^{l'}) + \frac{1}{m} \; G(z_k^l, z_k^{l'}) + \frac{1}{mn} \mathcal{M}(z_k^l, z_k^{l'}).$$

To deliver the coup de grâce via an application of Theorem 2.2, we are also required to control $\mathbb{E}_{\mathbb{S}}[g]$. To that end, we proceed as follows. Let $\mathbb{S}' = \{\mathcal{S}_1', \ldots, \mathcal{S}_m'\}$ be an i.i.d. copy of $\mathbb{S}$, and $\mathbb{S}^{(j)} := \{\mathcal{S}_1, \ldots, \mathcal{S}_{j-1}, \mathcal{S}_j', \mathcal{S}_{j+1}, \ldots, \mathcal{S}_m\}$. For each i, j, let $\mathbb{S}'' := \{z_j^{i''}\}$ be an i.i.d. copy of $\{z_j^i\}$, independent of $\mathbb{S}$ and $\mathbb{S}'$. Let $\mathcal{S}_j^{(i)} = \{z_j^1, \ldots, z_j^{i-1}, z_j^{i''}, z_j^{i+1}, \ldots, z_j^n\}$.

$$\mathbb{E}_{\mathbb{S}, \mathcal{S}, z}\Big[\frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \big(\ell(A(\mathbb{S})(\mathcal{S}_j), z_j^i) - \ell(A(\mathbb{S})(\mathcal{S}), z)\big)\Big]$$

$$= \mathbb{E}_{\mathbb{S}, \mathbb{S}', \mathbb{S}''}\Big[\frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \big(\ell(A(\mathbb{S})(\mathcal{S}_j), z_j^i) - \ell(A(\mathbb{S}^{(j)})(\mathcal{S}_j^{(i)}), z_j^i)\big)\Big]$$

$$= \mathbb{E}_{\mathbb{S}, \mathbb{S}', \mathbb{S}''}\Big[\frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \big(\ell(A(\mathbb{S})(\mathcal{S}_j), z_j^i) - \ell(A(\mathbb{S}^{(j)})(\mathcal{S}_j), z_j^i) + \ell(A(\mathbb{S}^{(j)})(\mathcal{S}_j), z_j^i) - \ell(A(\mathbb{S}^{(j)})(\mathcal{S}_j^{(i)}), z_j^i)\big)\Big]$$

$$\leq \mathbb{E}_{\mathbb{S}, \mathbb{S}', \mathbb{S}''}\Big[\frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} H(z_j^i, z_j^{i'}) + \frac{1}{n} \sum_{i=1}^{n} G(z_j^i, z_j^{i''})\Big]$$

$$= \mathbb{E}_{z, z' \sim \mathcal{D}, \mathcal{D} \sim \mu}[H(z, z') + G(z, z')].$$

Therefore, from Theorem 2.2, it follows that

$$\mathbb{P}\Big(R(A(\mathbb{S}), \mu) - R(A(\mathbb{S}), \mathbb{S}) \geq y + \mathbb{E}_{z, z' \overset{i.i.d.}{\sim} \mathcal{D}, \mathcal{D} \sim \mu}[H(z, z') + G(z, z')]\Big) \leq c_1 \frac{L_{p1}}{y^p} + 2 \exp\big(-c_2 \frac{y^2}{L_{p2}}\big), \tag{19}$$

where

$$L_{p1} := \sum_{k=1}^{m} \sum_{l=1}^{n} \mathbb{E}[(2\ H(z_k^l, z_k^{l'}) + \frac{1}{m}\ G(z_k^l, z_k^{l'}) + \frac{1}{mn}\mathcal{M}(z_k^l, z_k^{l'}))^p], \text{ and,}$$

$$L_{p2} := \sum_{k=1}^{m} \sum_{l=1}^{n} \mathbb{E}[(2\ H(z_k^l, z_k^{l'}) + \frac{1}{m}\ G(z_k^l, z_k^{l'}) + \frac{1}{mn}\mathcal{M}(z_k^l, z_k^{l'}))^2].$$

Evidently, using Hölder's inequality, it follows that

$$L_{p1} \leq C_p(mn\mathbb{E}[H^p] + \frac{n}{m^{p-1}}\mathbb{E}[G^p] + \frac{1}{(mn)^{p-1}}\mathbb{E}[\mathcal{M}^p])$$

$$L_{p2} \leq C_2(mn\mathbb{E}[H^2] + \frac{n}{m}\mathbb{E}[G^2] + \frac{1}{mn}\mathbb{E}[\mathcal{M}^2]),$$

which completes the proof in light of (19). $\qquad\square$

C Numerical Experiments

In this section, we provide empirical studies highlighting the tightness of our theoretical bounds in Sections 3.1 and 3.2. It is imperative we describe the general recipe of our experiments before going into the details. Since the applications in Section 3 stem from the key theoretical result Theorem 2.2, let us presume that the bounds in (3) are tight upto some constant. Therefore, when z is large, the polynomial term z^{-p} dominates in the decay, and from (3), it follows that

$$\frac{\mathbb{P}(|f(x_1, x_2, \ldots, x_n)| - \mathbb{E}[f(x_1, x_2, \ldots, x_n)]| > z)}{\mathbb{P}(|f(x_1, x_2, \ldots, x_n)| - \mathbb{E}[f(x_1, x_2, \ldots, x_n)]| > C_0 z)} \approx C_0^p. \tag{20}$$

On the other hand, in the absence of the polynomial term z^{-p} or when z is small, the sub-gaussian term dominates the tail probability, and consequently we should expect

$$\frac{\mathbb{P}(|f(x_1, x_2, \ldots, x_n)| - \mathbb{E}[f(x_1, x_2, \ldots, x_n)]| > z)}{\mathbb{P}(|f(x_1, x_2, \ldots, x_n)| - \mathbb{E}[f(x_1, x_2, \ldots, x_n)]| > C_0 z)} \approx 2\exp(C_1 z^2), \tag{21}$$

where C_1 is some constant depending on p, $\mathbb{E}[H_i]$ and C_0. The twin observations (20) and (21) informs our subsequent discussion, whereby any deviations from the expected behaviors would indicate our bounds are not sharp. In particular, we analyze the setting corresponding to Sections 3.1 and 3.2 in the following Sections C.1 and C.2, respectively.

C.1 Simulations for empirical risk minimization

Consider i.i.d. observations $S := \{z_i = (x_i, y_i)\}_{i=1}^{m} \in \mathbb{R}^d \times \mathbb{R}$ from a linear model $Y = X\beta + \varepsilon$, where the errors $\varepsilon \overset{d}{=} U_1^{-1/\nu} - U_2^{-1/\nu}$, $U_1, U_2 \overset{i.i.d.}{\sim} U[0,1]$, is drawn from the distribution of the difference of two independent Pareto (type I) random variables. Note that here $p = \nu/2$. For our analysis, we take $d = 5$, $\beta = (1, 1, \ldots, 1)^\top$, and vary $m \in \{500, 1000\}$, and $\mu = \{2.2, 4.4\}$. For the scaling parameter C_0 in (20), we choose $C_0 = 1.5$.

To incorporate stability in our analysis, we resort to a ridge-regression with regularization parameter $\lambda = 1.0$; formally, let

$$\hat{\beta} = (X^\top X + \lambda I_d)^{-1} X^\top Y, \ X = (x_1 : \ldots : x_m)^\top, Y = (y_1, \ldots, y_m).$$

Note that $\hat{\beta}$ plays the role of A_S in Section 3.1. Here we consider $(x_i)_{i=1}^m \overset{i.i.d.}{\sim} N(0, \Sigma)$, $\Sigma_{ij} = 0.3^{|i-j|}$. We consider ℓ to be the squared error loss, so that

$$R(A, S) = \mathbb{E}_{x,y}[(y - x^\top \hat{\beta})^2], \quad R_{emp}(A, S) = \frac{1}{m} \sum_{i=1}^m (y_i - x_i^\top \hat{\beta})^2.$$

Elementary calculations show that Assumption 3.1 is satisfied with probability approaching 1 for $H(z_i, z_i') = |x_i y_i - x_i' y_i'|$ and $\beta = \frac{\log m}{m}$. Since β is extremely small for large values of m, we may be excused for comparing $p(y) := \frac{\mathbb{P}(|R - R_{emp}| > y)}{\mathbb{P}(|R - R_{emp}| > C_0 y)}$ against $C_0^{\nu/2}$. The tail probabilities are empirically estimated via $50,000$ Monte Carlo draws. In Figure 2, the ratio $p(y)$ exhibits an
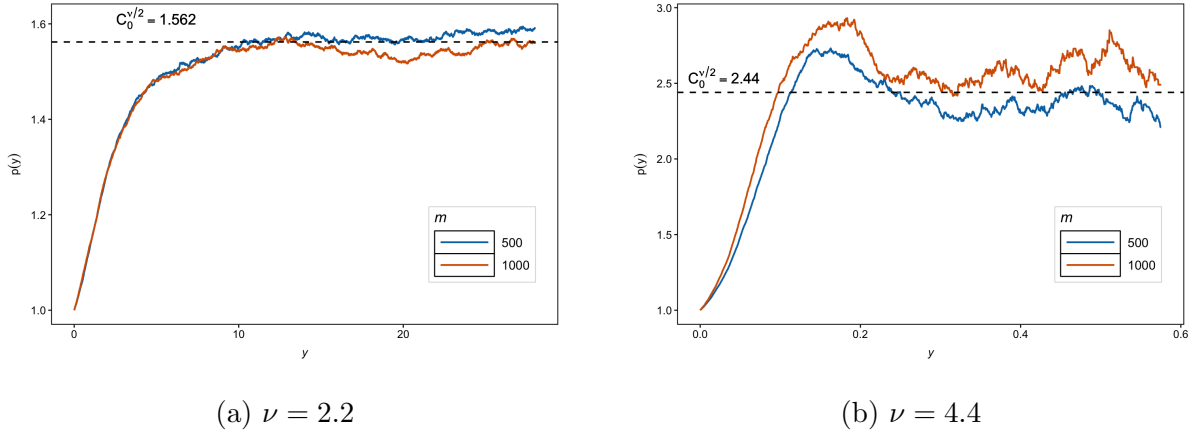

(a) $\nu = 2.2$ (b) $\nu = 4.4$

Figure 2: Plot of $p(y)$ versus y; both the worms stabilize around $C_0^{\nu/2}$ for large y.

initial exponential growth before slowing down and stabilizing near $C_0^{\nu/2}$, further vindicating the behavior typified in (20) in light of Theorem 3.3. For larger ν (e.g., $\nu = 4.4$), the Gaussian tail dominates the polynomial tail more strongly at smaller values of y. Consequently, $p(y)$ may exceed the threshold $C_0^{\nu/2}$ initially, before stabilizing in the large-y regime. This exhibits the tightness of our results.

C.2   Simulations for transductive regression algorithms

For the numerical studies involving transductive regression, we maintain the experimental set-up of ridge regression from Section C.1. Borrowing the notations of Section C.2, let S and T denote the training and test set respectively, with $|S| = m$, and $|T| = u$. For the purpose of this experiment, we take $m = u$, and vary $m \in \{500, 1000\}$. Similar to Theorem B.5, we are concerned with

$$\hat{R}(h) = \frac{1}{m} \sum_{i \in S} (y_i - x_i^\top \hat{\beta})^2, \qquad R(h) = \frac{1}{m} \sum_{i \in T} (y_i - x_i^\top \hat{\beta})^2.$$

Note that here $\hat{\beta}$ is trained via ridge regression with $\lambda = 1.0$ *only* on the training sample S. Similar to Section C.1, tail probabilities are empirically estimated via $50,000$ Monte Carlo draws. Figure 3 showcases the corresponding line plots of $p(y)$ versus y, where $p(y) := \frac{\mathbb{P}(|R - \hat{R}| > y)}{\mathbb{P}(|R - \hat{R}| > C_0 y)}$. The stabilization of the curve $p(y)$ around the threshold $C_0^{\nu/2}$ is evident for this example as well, exhibiting the sharpness of our bounds.
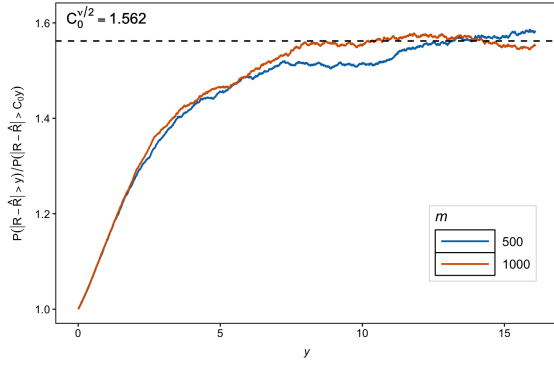
(a) $\nu = 2.2$
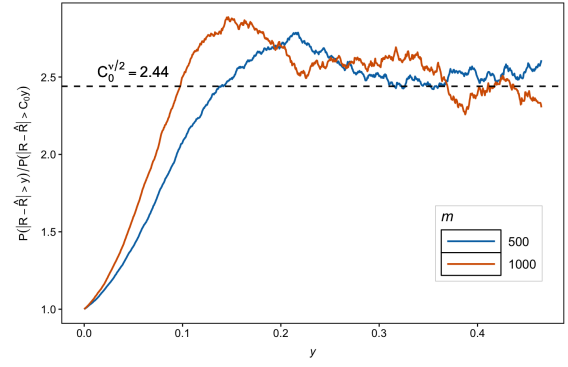
(b) $\nu = 4.4$

Figure 3: Plot of $p(y)$ versus y for the transductive regression problem in Section 3.2.