Generalization Bounds Under Heavy-Tailed Losses

Gholamali Aminian
The Alan Turing Institute

Abstract

The generalization error of a supervised statistical learning algorithm, defined as the difference between the population risk and the empirical risk, quantifies its ability to predict performance on previously unseen data. In this work, we analyze the generalization error under the heavy-tailed assumption on the loss function with respect to the data-generating distribution. Specifically, we derive uniform, information-theoretic, and PAC-Bayesian bounds on the generalization error under the assumption that the $(1+\epsilon)$-th moment of the loss function is bounded for $\epsilon \in (0, 1]$. The generalization error is shown to have a convergence rate of $O(n^{-\epsilon/(1+\epsilon)})$ where n is the number of training samples. Furthermore, we apply our results to study the generalization error of the Gibbs posterior and noisy iterative learning algorithms under the heavy-tailed assumption.

1 INTRODUCTION

A central concern in statistical learning theory is understanding the efficacy of a learning algorithm when applied to *test* data under Empirical risk minimization (ERM) as a popular framework in machine learning. This evaluation is typically carried out by investigating the *generalization error*, which quantifies the disparity between the performance of the algorithm on the training dataset and its performance on previously unseen data drawn from the same underlying distribution via a risk function.

The performances of the empirical risk and generalization error are affected when the data set is strongly imbalanced or contains outliers, which results in a heavy-tailed scenario where some moments of loss function under data-generating distribution are not bounded. Understanding the generalization behaviour of learning algorithms under heavy-tailed scenarios is one of the most important objectives in statistical learning theory. Unbounded loss functions present a significant challenge in studying the generalization error of learning algorithms. To address this, various assumptions, such as sub-Gaussian or sub-Exponential on data generating distributions, are often made to model the loss function. Under these assumptions, all moments of the loss function under the data-generating distribution are bounded.

However, in many real-world applications, such as financial modeling (Cont, 2001; Müller et al., 1998), network traffic analysis (Resnick, 2007), image recognition (Zhu et al., 2014; Wang et al., 2017; Van Horn et al., 2018), and language modeling (Zhang et al., 2022), data exhibits heavy-tailed behavior, where moments beyond a certain order may be unbounded or poorly controlled. Furthermore, empirical data distributions across diverse fields—including natural language processing, medicine, finance, and physics often follow power-law distributions as heavy-tailed distributions. Prominent examples include Zipf's law in language (Piantadosi, 2014) (where the frequency of words is roughly inversely proportional to their rank), Pareto's law in finance and economics (Bouchaud, 2001) (describing, for example, the heavy-tailed distribution of incomes or firm sizes), the Gutenberg–Richter law in seismology (Sornette and Sornette, 1999) (which relates earthquake magnitudes to their frequency), and scaling laws observed in physiological and physical systems (Marquet et al., 2005) (such as $1/f$ noise in heart rate variability and the inverse-square law in gravitational and electrostatic interactions). These distributions have been widely analyzed, as discussed in studies such as (Newman, 2005; Clauset et al., 2009) and related references. Various mechanisms contribute to the emergence of power-law distributions in both natural and artificial systems, each relevant to specific applications; see (Sornette, 2006, Chapter 14) and (Newman, 2005; Mitzenmacher, 2004). Among these, the most significant are growth with preferential

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

attachment (Yule's process) and critical phenomena (Newman, 2005). Additionally, the Generalized Central Limit Theorem states that the normalized sum of independent and identically distributed random variables with infinite variance converges only to a stable distribution; see, for instance, (Nolan, 2020). All stable distributions with infinite variance exhibit power-law tails, whereas the Gaussian distribution is the only stable distribution with finite variance (Samoradnitsky and Taqqu, 2017). This discrepancy highlights the need for more general theoretical frameworks that account for such heavy-tailed scenarios (Asadi, 2024).

Some recent works studied the generalization error under unbounded loss functions with heavy-tailed assumption via the PAC-Bayesian approach, specifically under bounded second-moment assumption (Kuzborskij and Szepesvári, 2019; Haddouche and Guedj, 2022). However, to the best of our knowledge, there is no unified framework for deriving generalization bounds that is applicable to uniform, information-theoretic, and PAC-Bayesian approaches, under the heavy-tailed assumption for bounded $(1+\epsilon)$-th moment of loss functions for $\epsilon \in (0, 1]$ with respect to data-generating distribution. This is the problem that we address in this paper, and our contributions here can be summarized as follows.

- We extend Bernstein's inequality for heavy-tailed random variables.
- We derive generalization bounds through uniform, information-theoretic, and PAC-Bayesian approaches for *unbounded loss* functions with bounded $(1 + \epsilon)$-th moment for $\epsilon \in (0, 1]$. We also derived a bound on absolute expected generalization error.
- As applications of our results, we use our bounds to bound the generalization error and excess risk of the Gibbs posterior[1] and the noisy iterative learning algorithms.

The paper is organized as follows: Section 2 introduces notation, the problem, and the risk functions used in this paper. Then, we discuss some related works in Section 3. The main theoretical tool and the extension of one-sided Bernstein's inequality are introduced in Section 4. We derive generalization bounds via uniform, PAC-Bayesian and information-theoretic approaches in Section 5. As applications of our results, we provide an upper bound on the expected generalization error of the Gibbs posterior under heavy-tailed assumption, an upper bound on the generalization error of noisy iterative learning algorithm in Section 6.

[1]It is also known as the Gibbs algorithm (Aminian et al., 2021a)

2 PRELIMINARIES

Notations: Upper-case letters denote random variables (e.g., Z), lower-case letters denote the realizations of random variables (e.g., z), and calligraphic letters denote sets (e.g., $\mathcal{Z}$). All logarithms are in the natural base. The set of probability distributions (measures) over a space $\mathcal{X}$ with finite variance is denoted by $\mathcal{P}(\mathcal{X})$. $\mathcal{N}(a, b)$ denotes the Gaussian distribution with mean a and variance of b.

Information measures: For two probability measures P and Q defined on the space $\mathcal{X}$, such that P is absolutely continuous with respect to Q, the *Kullback-Leibler* (KL) divergence between P and Q is $\mathrm{KL}(P\|Q) := \int_{\mathcal{X}} \log\left(dP/dQ\right) dP$. The mutual information between two random variables X and Y is defined as the KL divergence between the joint distribution and product-of-marginal distribution $I(X;Y) := \mathrm{KL}(P_{X,Y}\|P_X \otimes P_Y)$, or equivalently, the *conditional KL divergence* between $P_{Y|X}$ and P_Y over P_X, $\mathrm{KL}(P_{Y|X}\|P_Y|P_X) := \int_{\mathcal{X}} \mathrm{KL}(P_{Y|X=x}\|P_Y) dP_X(x)$. The *symmetrized KL information* between X and Y is given by $I_{\mathrm{SKL}}(X;Y) := I_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y)$, see (Aminian et al., 2015). For function $f : x \to \mathbb{R}$, we define $f'(x)$ as the derivative of function $f(x)$.

Heavy-tailed Random Variable: There are different notions of heavy-tailed random variables (Foss et al., 2011; Bakhshizadeh et al., 2023). In this work, we focus on the definition of a heavy-tailed scenario, where the $(1 + \epsilon)$-th moment of the loss function is bounded for some $\epsilon \in (0, 1]$. In the following, we define the heavy-tailed random variables.

Definition 2.1 (Heavy-tailed Random Variable)**.** A random variable X has heavy-tailed distribution, if there exists $\epsilon \in (0, 1]$, where $(1 + \epsilon)$-th moment of X is finite, i.e,

$$\mathbb{E}[|X|^{1+\epsilon}] < \infty, \tag{1}$$

and higher order moments, larger than $(1 + \epsilon)$, are unbounded.

2.1 Problem Formulation

Let $S = \{Z_i\}_{i=1}^n$ be the training set, where each sample $Z_i = (X_i, Y_i)$ belongs to the instance space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$; here $\mathcal{X}$ is the input (feature) space and $\mathcal{Y}$ is the output (label) space. We assume that Z_i are i.i.d. generated from the same data-generating distribution μ.

Here we consider the set of hypothesis $\mathcal{H}$ with elements $h : \mathcal{X} \mapsto \mathcal{Y} \in \mathcal{H}$. When $\mathcal{H}$ is finite, then its cardinality is denoted by $\mathrm{card}(\mathcal{H})$. In order to measure the performance of the hypothesis h, we consider a non-negative loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_0^+$.

We apply different methods to study the performance of our algorithms, including uniform and information-theoretic approaches. In uniform approaches, such as the VC-dimension and the Rademacher complexity approach (Vapnik, 1999; Bartlett and Mendelson, 2002), the hypothesis space is independent of the learning algorithm. Therefore, these methods are algorithm-independent; our results for these methods do not specify the learning algorithms.

For information-theoretic approaches in supervised learning, following (Xu and Raginsky, 2017), we consider learning algorithms that are characterized by a Markov kernel (a conditional distribution) $P_{H|S}$. Such a learning algorithm maps a data set S to a hypothesis in $\mathcal{H}$, which is chosen according to $P_{H|S}$. This concept thus includes randomized learning algorithms.

2.2 Risk Functions

The main quantity we are interested in is the *population risk*, defined by

$$\mathrm{R}(h,\mu) := \mathbb{E}_{\tilde{Z}\sim\mu}[\ell(h,\tilde{Z})], \quad h \in \mathcal{H}.$$

As the distribution μ is unknown, in classical statistical learning, the (true) population risk for $h \in \mathcal{H}$ is estimated by the (linear) *empirical risk*

$$\widehat{\mathrm{R}}(h,S) = \frac{1}{n}\sum_{i=1}^{n}\ell(h,Z_i). \tag{2}$$

The *generalization error* for the empirical risk is given by

$$\mathrm{gen}(h,S) := \mathrm{R}(h,\mu) - \widehat{\mathrm{R}}(h,S); \tag{3}$$

this is the difference between the true risk and the empirical risk.

Uniform Bounds: In learning theory, to obtain uniform bounds, most works focus on bounding the generalization error $\mathrm{gen}(h,S)$ from (3) such that under the distribution of the dataset S, with probability at least $(1-\delta)$, it holds that

$$\sup_{h\in\mathcal{H}} \mathrm{gen}(h,S) \leq g(\delta,n), \tag{4}$$

where g is a real function dependent on $\delta \in (0,1)$ and n is the number of data samples.

PAC-Bayesian Bounds: In the PAC-Bayesian approach, we fix a probability distribution over the hypothesis (parameter) space as a prior distribution, denoted as Q_h. Then, we are interested in the generalization performance under a data-dependent distribution over the hypothesis space, known as posterior distribution, denoted as ρ_h. Then, an upper bound on the expectation of generalization error with respect to ρ_h is derived under the distribution of the dataset S with probability at least $(1-\delta)$,

$$\mathbb{E}_{H\sim\rho_h}[\mathrm{gen}(H,S)] \leq g_p(\delta,n,Q_h), \tag{5}$$

where g_p is a real function dependent on $\delta \in (0,1)$, prior distribution Q_h and n as the number of data samples.

Information-theoretic Bounds: For the information-theoretic approach to generalization, we consider the hypothesis H to be a random variable under a learning algorithm as Markov kernel, i.e., $P_{H|S}$, and then we take expectations over the hypothesis H,

$$\overline{\mathrm{gen}}(H,S) := \mathbb{E}_{P_{H,S}}[\mathrm{gen}(H,S)]. \tag{6}$$

We provide upper bounds on the expected generalization error with respect to the joint distribution of S and H of the form

$$\overline{\mathrm{gen}}(H,S) \leq g_e(n),$$

where g_e is a real function. We are also interested in the expected excess risk of the learning algorithm $P_{H|S}$, defined as:

$$\mathcal{E}(H,\mu) := \mathbb{E}_{P_H}[\mathrm{R}(H,\mu)] - \inf_{h\in\mathcal{H}} \mathrm{R}(h,\mu). \tag{7}$$

3 RELATED WORKS

Different approaches have been applied to study the generalization error of general learning problems under empirical risk minimization, including uniform, PAC-Bayesian, and information-theoretic bounds. In this section, we discuss the related works on Uniform, PAC-Bayesian, and information-theoretic bounds. We also further discuss the generalization error of the Gibbs posterior.

Uniform Bounds: Uniform bounds (or VC bounds) are proposed by (Vapnik and Chervonenkis, 1971; Bartlett et al., 1998, 2019). For any class of functions $\mathcal{F}$ of VC dimension d, with probability at least $1-\delta$ the generalization error is $O\big((d+\log(1/\delta))^{1/2}n^{-1/2}\big)$. This bound depends solely on the VC dimension of the function class and on the sample size; in particular, it is independent of the learning algorithm.

PAC-Bayes bounds: First proposed by Shawe-Taylor and Williamson (1997); McAllester (1999) and (McAllester, 2003), PAC-Bayesian analysis provides high probability bounds on the generalization error in terms of the KL divergence between the data-dependent posterior induced by the learning algorithm and a data-free prior that can be chosen arbitrarily (Alquier et al.,

2024). There are multiple ways to generalize the standard PAC-Bayesian bounds, including using information measures other than KL divergence (Alquier and Guedj, 2018; Bégin et al., 2016; Hellström and Durisi, 2020; Aminian et al., 2021b) and considering data-dependent priors (Rivasplata et al., 2020; Catoni, 2007; Dziugaite and Roy, 2018; Ambroladze et al., 2007). There are also some works (Haddouche and Guedj, 2022; Kuzborskij et al., 2019) on using the PAC-Bayesian approach for deriving the generalization error bounds for unbounded loss function under more relaxed assumptions, e.g., bounded second moments.

Information-theoretic bounds: (Russo and Zou, 2019; Xu and Raginsky, 2017) propose using the mutual information between the input training set and the output hypothesis to upper bound the expected generalization error. Multiple approaches have been proposed to tighten the mutual information-based bound: (Bu et al., 2020) provide tighter bounds by considering the individual sample mutual information; (Asadi et al., 2018; Asadi and Abbe, 2020) propose using chaining mutual information; propose using rate-distortion approach (Nokleby et al., 2016; Masiha et al., 2023); and (Steinke and Zakynthinou, 2020; Hafez-Kolahi et al., 2020; Aminian et al., 2020, 2021b, 2022) provide different upper bounds on the expected generalization error based on the linear empirical risk framework.

Non-linear risk functions: Some recent works (Aminian et al., 2025; Mulumudi et al., 2026) have investigated the generalization error of nonlinear risk functions. In particular, Aminian et al. (2025) and Mulumudi et al. (2026) study the tilted empirical risk and conditional value-at-risk under heavy-tailed assumptions, respectively. By contrast, our focus here is on the linear risk function in the heavy-tailed assumption.

Other heavy-tailed assumptions: While our work focuses on deriving generalization bounds relying strictly on the boundedness of the $(1+\epsilon)$-th moment, yielding rates of $O(n^{-\epsilon/(1+\epsilon)})$, there is a significant body of work establishing *fast rates* for unbounded and heavy-tailed losses under stronger assumptions. (Dinh et al., 2016) derives fast learning rates by assuming a multi-scale Bernstein condition on the loss distribution. Similarly, (Grünwald and Mehta, 2020) establish fast rates for ERM and generalized Bayes by introducing the *central condition* (a generalization of the Bernstein condition) and witness conditions. In the context of Bayesian learning, (Ho et al., 2020) extends these results to show that the generalized posterior concentrates at a fast rate under similar Bernstein-type assumptions. In contrast to these works, our analysis does not rely on Bernstein or central conditions, which control the variance relative to the risk. Instead, we derive bounds based solely on the heavy-tailed moment properties. This provides theoretical guarantees in regimes where the structural conditions required for fast rates may not hold, albeit at the cost of a slower convergence rate. Furthermore, our results can be applied to a wide range of methods, including uniform, PAC-Bayesian and information-theoretical approaches.

Generalization error of the Gibbs posterior: (Raginsky et al., 2016) provide an information-theoretic upper bound with a convergence rate of $\mathcal{O}(1/n)$ for the Gibbs posterior with a bounded loss function. (Asadi and Abbe, 2020, Appendix D) provides an upper bound on the excess risk of the Gibbs posterior under the sub-Gaussian assumption. (Kuzborskij et al., 2019) focus on the excess risk of the Gibbs posterior and establish a similar generalization bound with a rate of $\mathcal{O}(1/n)$ under the sub-Gaussian assumption. (Aminian et al., 2021a) provide an exact characterization and an upper bound on the expected generalization error. Although these bounds are tight in terms of sample complexity, they rely on restrictive assumptions such as bounded or sub-Gaussian loss function.

4 EXTENSION OF BERNSTEIN'S INEQUALITY

In this section, we provide an extension of Bernstein's inequality where can be useful in deriving our main results. This result is based on an useful Lemma by Behnamnia et al. (2025). All proof details are deferred to Appendix B.

Lemma 4.1 (Lemma B.8, (Behnamnia et al., 2025))**.** *For $x > 0$, the following inequality holds for $\epsilon \in (0, 1]$,*

$$\exp(-x) \leq 1 - x + \frac{x^{(1+\epsilon)}}{1+\epsilon}. \tag{8}$$

Note that for $\epsilon = 1$, the inequality in Lemma 4.1 becomes the following known inequality:

$$\exp(-x) \leq 1 - x + \frac{x^2}{2},$$

which has been widely utilized in different applications.

Remark 4.2. Lemma 28 in (Lugosi and Neu, 2023) states that for $y < 0$ and $\varepsilon \in [0, 1]$,

$$e^y \leq 1 + y + |y|^{1+\varepsilon}.$$

Lemma 4.1 sharpens this inequality by replacing $|y|^{1+\varepsilon}$ with $|y|^{1+\varepsilon}/(1+\varepsilon)$, yielding a uniformly tighter bound (strictly tighter for $\varepsilon > 0$, with equality when $\varepsilon = 0$).

In the following, we extend the one-sided Bernstein's inequality (Wainwright, 2019, Proposition 2.14) to the heavy-tailed scenario using Lemma 4.1.

Theorem 4.3. *Suppose that $X \leq b$ and there exists $\epsilon \in (0,1]$ such that $\mathbb{E}[|X-b|^{1+\epsilon}] \leq \infty$. Then, we have for $\lambda > 0$,*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X])] \leq \exp\left(\frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}[|X-b|^{1+\epsilon}]\right), \quad (9)$$

Furthermore, given n i.i.d samples such that $X_i \leq b$ for all $i \in [n]$, with probability at least $1-\delta$ for $\delta \in (0,1)$, we have,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \leq \mathbb{E}\left[|X-b|^{1+\epsilon}\right]^{\frac{1}{1+\epsilon}} \times \left(\frac{\log(1/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \left(\frac{\epsilon+1}{\epsilon}\right)^{\frac{\epsilon}{\epsilon+1}}. \quad (10)$$

We can observe that when $b = 0$, if the $(1+\epsilon)$-th moment of $|X|$ is bounded for some $\epsilon \in (0,1]$, Theorem 4.3 gives a non-trivial bound, and for $\epsilon = 1$, it recovers Bernstein's inequality. In the next section, we use Lemma 4.1 to derive generalization bounds under the heavy-tailed assumption.

5 GENERALIZATION BOUNDS

In this section, we provide generalization error bounds via uniform, PAC-Bayesian, and information-theoretic approaches under heavy-tailed loss function assumption for linear empirical risk. We also compared our results with other works related to the heavy-tailed assumption in Table 1. More details are provided in the following sections. All proof details are deferred to Appendix C.

5.1 Assumptions

We first provide assumptions which are needed to derive generalization bounds, using uniform, PAC-Bayesian and information-theoretical approaches.

Assumption 5.1 (Uniform Heavy-tailed Loss Function)**.** There exists $\epsilon \in (0,1]$ and $\alpha_u \in \mathbb{R}^+$, where satisfies $\mathbb{E}_{\tilde{Z}\sim\mu}[\ell(h,z)^{(1+\epsilon)}] \leq \alpha_u$ uniformly for all $h \in \mathcal{H}$.

Assumption 5.1 can be relaxed in the following assumption which is made for the information theoretical analysis.

Assumption 5.2 (Expected Heavy-tailed Loss Function)**.** There exists $\epsilon \in (0,1]$ and $\alpha_\varepsilon \in \mathbb{R}^+$, such that the loss function $(H,Z) \mapsto \ell(H,Z)$ satisfies $\mathbb{E}_{P_H\otimes\mu}[\ell^{(1+\epsilon)}(H,Z)] \leq \alpha_\varepsilon$.

The assumption on $(1+\epsilon)$-th moment, Assumption 5.2, can be satisfied if the loss function is sub-Gaussian or sub-Exponential (Boucheron et al., 2013) under the distribution $\mu \otimes P_H$. Finally, we consider the following assumption for our analysis on absolute expected generalization error.

Assumption 5.3. There exists $\epsilon \in (0,1]$ and $\alpha_\varepsilon \in \mathbb{R}^+$, such that the loss function $(H,Z) \mapsto \ell(H,Z)$ satisfies $\max\left(\mathbb{E}_{P_H\otimes\mu}[\ell^{(1+\epsilon)}(H,Z)], \mathbb{E}_{P_{H,Z}}[\ell^{(1+\epsilon)}(H,Z)]\right) \leq \beta_\varepsilon$.

5.2 Uniform Bounds

Here, we provide uniform bound on generalization error under Assumption 5.1. Using Theorem 4.3 with union bound over different hypotheses, we have the following uniform generalization error bound.

Theorem 5.4 (Uniform Bound)**.** *Under Assumption 5.1, with probability at least $(1-\delta)$, and a finite hypothesis space, the supremum generalization error satisfies,*

$$\sup_{h\in\mathcal{H}} \text{gen}(h,S) \leq \alpha_u^{1/(1+\epsilon)}\left(\frac{\epsilon+1}{\epsilon}\right)^{\frac{\epsilon}{\epsilon+1}} \times \left(\frac{\log\left(\text{card}(\mathcal{H})\right) + \log(1/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}. \quad (11)$$

Theorem 5.4 assumed that the hypothesis space is finite; this is, for example, the case in classification problems with a finite number of classes. If this assumption is violated, we can apply the growth function technique from (Bousquet et al., 2003; Vapnik, 1999). Furthermore, the growth function can be bounded by VC-dimension in binary classification (Vapnik, 1999) or Natarajan dimension (Holden and Niranjan, 1995) for multi-class classification scenarios. Note that the VC-dimension and Rademacher complexity bounds are uniform bounds and are independent of the learning algorithms.

Comparison with existing uniform bounds: A proof of Bernstein's inequality via the inequality $\exp(x) \leq 1 + x + x^2/2$ for $x < 0$ is proposed by Maurer et al. (2003). We extend Bernstein's inequality via Lemma 4.1. An upper bound on generalization error via VC-dimension and growth function under bounded $(1+\epsilon)$-th moment for $\epsilon \in (0,1]$ is proposed in (Cortes et al., 2019, Corollary 12) which is motivated by relative deviation generalization bounds in binary classification. Furthermore, the final convergence rate for unbounded loss is $O(\log(n)n^{\frac{-\epsilon}{1+\epsilon}})$ based on (Cortes et al., 2019, Corollary 12). In contrast, we derive the results for a multi-classification scenario with a better convergence rate of $O(n^{\frac{-\epsilon}{1+\epsilon}})$.

5.3 PAC-Bayesian Bounds

Inspired by previous works on PAC-Bayesian theory (Alquier et al., 2024; Catoni, 2004), we derive a high probability bound on the generalization error with respect to the posterior distribution over the hy-

Table 1: Comparison of our work with existing literature for heavy-tailed losses: Key features of our results include support for detailed assumption with a focus on bounded second moments or heavy-tailed assumptions, methodological approaches along convergence rate with respect to number of samples n.

Work	Assumption Details	Approach	Convergence Rate
(Cortes et al., 2019)	Bounded $(1+\epsilon)$-th moment of loss for $\epsilon \in (0,1]$	Uniform	$O(\log(n)n^{-\epsilon/(1+\epsilon)})$
(Alquier and Guedj, 2018)	Bounded second moment	PAC-Bayesian	$O(n^{-1/2})$
(Kuzborskij and Szepesvári, 2019)	Bounded second moment	PAC-Bayesian	$O(n^{-1/2})$
(Haddouche and Guedj, 2022)	Bounded second moment with parameter selection	PAC-Bayesian	$O(n^{-1/2})$
(Zhang et al., 2024)	Bounded second moment with exponential moment on finite intervals	PAC-Bayesian	$O(n^{-1/2})$
(Holland, 2019)	Bounded second and third moments	PAC-Bayesian	$O(n^{-1/2})$
(Lugosi and Neu, 2022)	Bounded worst-case $(1+\epsilon)$-moment for centered loss for $\epsilon \in (0,1]$	Information-theoretic	$O(n^{-\epsilon/(1+\epsilon)})$,
Our Work	Bounded $(1+\epsilon)$-th moment of loss function for $\epsilon \in (0,1]$	Uniform, Information-theoretic, PAC-Bayesian	$O(n^{-\epsilon/(1+\epsilon)})$,

pothesis space. We have the following PAC-Bayesian upper bound,

Theorem 5.5 (PAC-Bayesian Bound)**.** *Under Assumption 5.1, with probability at least* $(1-\delta)$ *under distribution* P_S*, we have for any* $\eta > 0$*,*

$$\mathbb{E}_{\rho_h}[\mathrm{gen}(H,S)] \leq \frac{\eta^{\epsilon}\alpha_u}{n^{\epsilon}(1+\epsilon)} + \frac{\mathrm{KL}(\rho_h\|Q_h) + \log(1/\delta)}{\eta}. \tag{12}$$

Remark 5.6. Choosing η such that $\eta^{-1} \asymp n^{\frac{-\epsilon}{1+\epsilon}}$ results in a theoretical guarantee on the convergence rate of $O(n^{\frac{-\epsilon}{1+\epsilon}})$.

Comparison with existing PAC-Bayesian bounds: Some works studied unbounded loss function via the PAC-Bayesian approach. Heavy-tailed loss functions under data-generating distribution are studied by (Alquier and Guedj, 2018, Proposition 4) where probability bounds (non-high probability) are developed under bounded second-moment assumptions. (Kuzborskij and Szepesvári, 2019; Viallard et al., 2024) and (Haddouche and Guedj, 2022) also provide bounds for losses with a bounded second moment using the PAC-Bayesian approach. The bounds in (Haddouche and Guedj, 2022) rely on a parameter that must be selected before the training data is drawn. In contrast, we provide for a general case under bounded $(1+\epsilon)$-th moment for $\epsilon \in (0,1]$. In contrast, our bounds are free of this parameter selection. Recently, (Zhang et al., 2024) proposed an upper bound on generalization error using the PAC-Bayesian approach and defining exponential moment on finite intervals, which holds for bounded second-moment condition. However, it is not shown that it also holds for $(1+\epsilon)$-th moment with $\epsilon < 1$. Using a different estimator than empirical risk, PAC-Bayes bounds for losses with bounded second and third moments are developed by Holland (2019). Notably, their bounds include a term that can increase with the number of samples n.

5.4 Information-Theoretic Bounds

In the following, we present the results based on information-theoretic approaches.

Theorem 5.7. *Suppose that Assumption 5.2 holds. Then,*

$$\overline{\mathrm{gen}}(H,S) \leq \frac{1}{n}\sum_{i=1}^{n} \frac{2}{1+\epsilon}\big(I(H;Z_i)\big)^{\epsilon/(1+\epsilon)}\alpha_{\varepsilon}^{1/(1+\epsilon)}. \tag{13}$$

Sketch of proof. We first provide an upper bound on $\mathbb{E}_{\mu\otimes P_H}[\exp\big(\lambda\ell(H,Z_i)\big)]$ via Lemma 4.1 for $\lambda < 0$. We then apply Donsker's representation of KL divergence to derive the final result. □

Corollary 5.8. *Assuming the same assumption in Theorem 5.7 and bounded* $I(H;S)$*, then we have the convergence rate of* $O(n^{\frac{-\epsilon}{1+\epsilon}})$.

5.5 Absolute Expected Generalization Error

Next, we provide an upper bound on the absolute value of expected generalization error with a convergence rate of $O(n^{\zeta-1})$ for $1>\zeta>0$ under the heavy-tailed assumption. For this purpose, we utilize the following lemma.

Lemma 5.9 (Lemma 5.2 in (Behnamnia et al., 2025))**.** *Suppose that $X>0$ and $\gamma<0$, then we have for $\epsilon\in(0,1]$,*

$$\mathrm{Var}(\exp(\gamma X))\leq|\gamma|^{1+\epsilon}\mathbb{E}[X^{1+\epsilon}]. \tag{14}$$

We can also assume that the $(1+\epsilon)$-th moment of the loss function is uniformly bounded over the hypothesis space, i.e.,

$$\mathbb{E}_{Z\sim\mu}\Big[\sup_{h\in\mathcal{H}}\ell^{1+\epsilon}(h,Z)\Big]\leq\beta_\varepsilon. \tag{15}$$

However, we have

$$\max\Big(\mathbb{E}_{P_H\otimes\mu}[\ell^{(1+\epsilon)}(H,Z)],\mathbb{E}_{P_{H,Z}}[\ell^{(1+\epsilon)}(H,Z)]\Big)\leq\mathbb{E}_Z[\sup_{h\in\mathcal{H}}\ell^{1+\epsilon}].$$

Therefore, Assumption 5.3 is more relaxed in comparison with (15).

In the following, we provide an upper bound on the absolute of expected generalization error.

Theorem 5.10. *Under Assumption 5.3, we can derive the following upper bound on absolute expected generalization error if $\frac{I(H;S)}{n}\leq\frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2}$ holds, then*

$$\Big|\overline{\mathrm{gen}}(H,S)\Big|\leq\frac{1}{|\lambda|}\sqrt{2|\lambda|^{1+\epsilon}\beta_\varepsilon\frac{I(H;S)}{n}}+\frac{|\lambda|^{\epsilon}\beta_\varepsilon}{1+\epsilon} \tag{16}$$

and if $\frac{I(H;S)}{n}>\frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2}$ holds, then we have

$$\Big|\overline{\mathrm{gen}}(H,S)\Big|\leq\frac{I(H;S)}{|\lambda|n}+\frac{2|\lambda|^{\epsilon}\beta_\varepsilon}{1+\epsilon}. \tag{17}$$

Sketch of proof. Leveraging the sup-exponential upper bounds on expected generalization error from (Aminian et al., 2021a), we derive an upper bound on

$$\Big|\mathbb{E}_{P_H\otimes\mu}[\exp(\lambda\ell(H,\tilde{Z}))]-\mathbb{E}_{P_{H,S}}[\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda\ell(H,Z_i))]\Big|,$$

in terms of $\mathrm{Var}(\exp(\lambda\ell(H,\tilde{Z})))$. We then apply Lemma 5.9 to further refine the bound. Finally, using Lemma 4.1 and the inequality $1+x\leq\exp(x)$, we complete the proof. □

Remark 5.11. Assuming $|\lambda|^{-1}=n^{1/(1+\epsilon)}$, we have the convergence rate of $O(n^{-\epsilon/(1+\epsilon)})$. Furthermore, assuming negligible $\beta_\varepsilon\to0$, we can achieve convergence rate of $O(n^{\zeta-1})$ for $1>\zeta>0$ by choosing $|\lambda|=n^{-\zeta}$. For example, choosing $\zeta=1-\epsilon$, we have the convergence rate of $O(n^{-\epsilon})$.

Comparison with Existing Information-Theoretic Bounds: For $\epsilon=1$, Theorem 5.7 achieves the same convergence rate as (Bu et al., 2020), while requiring only the second-moment assumption, which is weaker than their sub-Gaussian assumption. (Steinke and Zakynthinou, 2020) derived an upper bound on the expected generalization error under the condition that the worst-case second moment of the loss function is bounded ($\mathbb{E}_{Z\sim\mu}[\sup_{h\in\mathcal{H}}\ell^2(h,Z)]<\infty$). In contrast, our result has two key advantages: it holds for $\epsilon<1$ and requires a more relaxed second-moment assumption, namely boundedness under the product measure $\mu\otimes P_H$. (Lugosi and Neu, 2022) proposed an approach using the convexity of information measures and derived an upper bound on the expected generalization error by assuming the bounded second moment and in terms of mutual information. In contrast to (Lugosi and Neu, 2022), our second-moment assumption is more relaxed, being based on the expected version with respect to the distribution over the hypothesis set and the data-generating distribution. Furthermore, our information-theoretic upper bound in Theorem 5.7 holds for a more relaxed assumption where $(1+\epsilon)$-th moment is needed to be bounded. In particular, (Lugosi and Neu, 2022, Corollary 5) is based on α-divergence for $\epsilon<1$, and the bounded moment assumption holds for centred loss function, $\sup_{h\in\mathcal{H}}|\ell(h,Z)-\mathbb{E}[\ell(h,\tilde{Z})]|$. In addition, we can extend our result to improve the results in (Aminian et al., 2021a) and derive the upper bound on the Gibbs posterior under a more relaxed assumption (see Section 6).

6 APPLICATIONS

We now provide some applications of our main results in KL-regularized empirical risk minimization and noisy iterative learning algorithms. All proof details are deferred to Appendix D.

6.1 KL-Regularized Empirical Risk Minimization

Next, we study the upper bound on the generalization error under the Gibbs posterior (Xu and Raginsky, 2017; Zhang, 2006). As outlined in (Aminian et al., 2021a), the Gibbs posterior is motivated by various scenarios, including information risk minimization and distribution over hypothesis due to the SGLD algorithm. The solution to the regularized ERM problem,

$$\inf_{P_{H|S}}\Big\{\mathbb{E}_{P_{H,S}}[\widehat{\mathrm{R}}(H,S)]+\frac{1}{\gamma}\mathrm{KL}(P_{H|S}\|\pi(H)|P_S)\Big\}, \tag{18}$$

corresponds to the Gibbs posterior, which is defined as:

$$P_{H|S}^{\gamma} \triangleq \frac{\pi(H)\, e^{-\gamma \widehat{\mathrm{R}}(H,S)}}{V(S)}, \quad \gamma \geq 0, \tag{19}$$

where γ is also called the inverse temperature and $V(S)$ is the normalization factor. In the following, we derive an upper bound on its expected generalization error.

Theorem 6.1. *Under Assumption 5.2, the following upper bound holds on the expected generalization error of the Gibbs posterior,*

$$0 \leq \overline{\mathrm{gen}}(H,S) \leq \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1} \alpha_{\varepsilon} \Big(\frac{\gamma}{n}\Big)^{\epsilon}. \tag{20}$$

Sketch of proof. In (Aminian et al., 2021a, Theorem 1), an exact characterization of the Gibbs posterior is derived for the general loss function in terms of $I_{\mathrm{SKL}}(H;S)$. Using the fact that $I(H;S) \leq I_{\mathrm{SKL}}(H;S)$ and combining with Theorem 5.7 complete the proof. □

Inspired by (Xu and Raginsky, 2017, Corollary 3), we can derive an upper bound on the excess risk of the Gibbs posterior under the heavy-tailed assumption.

Proposition 6.2. *Assume that the loss function is L-Lipschitz for all $z \in \mathcal{Z}$ and let $\mathcal{H} \subseteq \mathbb{R}^d$. Then, under the Gibbs posterior, we have,*

$$\begin{aligned} \mathcal{E}(H,\mu) \leq & \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1} \alpha_{\varepsilon} \Big(\frac{\gamma}{n}\Big)^{\epsilon} \\ & + \frac{1}{\gamma} \mathrm{KL}\Big(\mathcal{N}\big(h^{\star}, \beta^2 \mathbf{I}_d\big) \| \pi(H)\Big) + L\beta\sqrt{d}, \end{aligned} \tag{21}$$

where $h^{\star} = \arg\inf_{h\in\mathcal{H}} R(h,\mu)$ and we assumed that $\mathrm{KL}\Big(\mathcal{N}\big(h^{\star}, \beta^2 \mathbf{I}_d\big) \| \pi(H)\Big) < \infty$ for given $\pi(H)$.

Remark 6.3. Assuming $\pi(H) = \mathcal{N}(h_Q, \beta^2 \mathbf{I}_d)$ and choosing γ and β such that $\gamma \asymp n^{\frac{3\epsilon}{2(1+\epsilon)}}$ and $\beta \asymp n^{\frac{-\epsilon}{2(1+\epsilon)}}$, results in a convergence rate of $O\Big(\max\Big(n^{\frac{-\epsilon}{2(1+\epsilon)}}, n^{\frac{\epsilon(\epsilon-2)}{2(1+\epsilon)}}\Big)\Big)$. For $\epsilon = 1$, we recover the convergence rate of $O(n^{-1/4})$ in (Xu and Raginsky, 2017) under bounded second-moment assumption which is more relaxed in comparison with sub-Gaussian assumption.

Gibbs Posterior Comparison: In (Aminian et al., 2021a, Theorem 3), authors derive an upper bound on the expected generalization error of the Gibbs posterior under the assumption of a sub-Gaussian loss function. In contrast, we achieve similar convergence with $\epsilon = 1$ under a weaker assumption - requiring only bounded second moments rather than sub-Gaussianity. Furthermore, our bound in Theorem 6.1 also holds for $\epsilon < 1$ which is novel. Furthermore, we derive an upper bound on excess risk of the Gibbs posterior under heavy-tailed assumption.

6.2 Noisy Iterative Learning Algorithms

In this section, we leverage our information-theoretic result to derive an upper bound on the generalization error for the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, building on the results presented in (Pensia et al., 2018). For this aim, we adopt the notation introduced in (Pensia et al., 2018) to formalize the behaviour of noisy iterative learning algorithms.

Let the parameter vector at iteration t be represented by $H_t \in \mathbb{R}^d$, with $H_0 \in \mathcal{H}$ denoting an arbitrary initialization. At each iteration $t \geq 1$, a data point $Z_t \subseteq S$ is sampled, and a direction $F(H_{t-1}, Z_t) \in \mathbb{R}^d$ is computed, where for SGLD we have $F(H_{t-1}, Z_t) = \nabla_h \ell(H_{t-1}, Z_t)$. The update step involves scaling this direction by a stepsize η_t and adding isotropic Gaussian noise $\xi_t \sim \mathcal{N}(0, \eta_t I_d)$ where the variance of noise is proportional to stepsize η, resulting in the following iterative update:

$$H_t = H_{t-1} - \eta_t \nabla_h \ell(H_{t-1}, Z_t) + \xi_t, \quad \forall t \geq 1. \tag{22}$$

We also made the following assumptions.

Assumption 6.4. The derivative of loss function is bounded, $\sup_{h\in\mathcal{H}, z\in\mathcal{Z}} \|\nabla_h \ell(h,z)\|_2 \leq L$, for some $L > 0$.

Assumption 6.5. The sampling strategy is independent from the previous iterates of the parameter vectors:

$$P(Z_{t+1} \mid Z^{(t)}, H^{(t)}, S) = P(Z_{t+1} \mid Z^{(t)}, S).$$

where $Z^{(t)} = \{Z_1, \cdots, Z_t\}$ and $H^{(t)} = \{H_1, \cdots, H_t\}$.

Note that the bounded gradient can hold due to the clipped gradient approach in practice. Combining (Pensia et al., 2018, Theorem 1) with Theorem 5.7, we can derive the following upper bound on expected generalization error under the SGLD algorithm.

Theorem 6.6. *Under Assumptions 5.2, 6.4 and 6.5, we can derive the following upper bound on the expected generalization error,*

$$\begin{aligned} \overline{\mathrm{gen}}(H,S) \leq & \frac{2}{(1+\epsilon) n^{\epsilon/(1+\epsilon)}} \alpha_{\varepsilon}^{1/(1+\epsilon)} \\ & \times \Big(\frac{L^2}{2} \sum_{t=1}^{T} \eta_t\Big)^{\epsilon/(1+\epsilon)}. \end{aligned} \tag{23}$$

The convergence rate of the expected generalization error for SGLD under a heavy-tailed loss function is observed to be $O(n^{\frac{-\epsilon}{1+\epsilon}})$.

7 FUTURE WORKS

Since our theoretical tools are novel, they can be applied to other theoretical results and extended to heavy-tailed

scenarios. Below, we outline some potential extensions of our work.

Upper bound based on conditional mutual information: We aim to extend our approach to derive an upper bound on the expected generalization error in terms of conditional mutual information (Steinke and Zakynthinou, 2020) under heavy-tailed assumptions.

Noisy Iterative learning algorithms: Our results can be applied to establish upper bounds for SGLD using an information-theoretic approach based on data-dependent estimates, as proposed by Negrea et al. (2019), as well as to derive an information-theoretic bound for SGD based on (Neu et al., 2021). Furthermore, we can apply our results to perturbed SGD, noisy momentum and accelerated gradient descent as discussed in (Pensia et al., 2018).

Other methods: Leveraging our theoretical tools, it is possible to derive an upper bound using Rademacher complexity (Bartlett and Mendelson, 2002; Golowich et al., 2018) and stability approaches (Bousquet and Elisseeff, 2002; Bousquet et al., 2020; Mou et al., 2017; Chen et al., 2018; Aminian et al., 2023). For these methods, we need to extend McDiarmid's inequality under the heavy-tailed assumption.

Bernstein's Condition: For a faster convergence rate in PAC-Bayesian, one approach is to assume Bernstein condition (Alquier et al., 2024). Applying our approach under a modified version of Bernstein's condition for a heavy-tailed scenario is an interesting direction.

Two-sided Bounds: The uniform bound in Theorem 5.4, the PAC-Bayesian bound in Theorem 5.5, and the information-theoretic bound in Theorem 5.7 are all one-sided. By contrast, Theorem 5.10 provides an upper bound on the absolute expected generalization error which is two-sided bound on generalization error. Extending the uniform and PAC-Bayesian frameworks to derive bounds on the absolute generalization error remains an interesting direction for future research.

8 CONCLUSION

In this paper, we extended some previous results on generalization error of learning algorithms and concentration inequality to support the heavy-tailed scenario. We first extended Bernstein's inequality to support the heavy-tailed random variables, providing a critical tool for analyzing learning systems under less restrictive assumptions. Then, we investigated the generalization error in the presence of heavy-tailed data distributions. Specifically, we derived bounds on the generalization error using uniform, PAC-Bayesian, and information-theoretic frameworks. These bounds establish theoretical guarantees with a convergence rate of $O(n^{\frac{-\epsilon}{(1+\epsilon)}})$ under the assumption of a heavy-tailed loss function, characterized by a bounded $(1+\epsilon)$-th moment for some $\epsilon \in (0, 1]$. Additionally, we presented an upper bound on the expected generalization error and excess risk of the Gibbs posterior in the context of heavy-tailed loss functions. Our analysis is further extended to noisy iterative learning algorithms, where we derived bounds on their expected generalization error under similar heavy-tailed assumptions with additional assumptions on noisy iterative algorithms. These results provide valuable insights into the behavior of learning algorithms in non-ideal conditions, contributing to a deeper understanding of generalization in the presence of heavy-tailed distributions.

ACKNOWLEDGMENT

The author would like to thank Amirreza Asadi, Tian Li, Ahmad Beirami and Gesine Reinert for their valuable comments and insightful feedback on first version of this work. Gholamali Aminian acknowledges the support of the UKRI Prosperity Partnership Scheme (FAIR) under EPSRC Grant EP/V056883/1 and the Alan Turing Institute.

References

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Pierre Alquier et al. User-friendly introduction to PAC-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. *Advances in neural information processing systems*, 19: 9, 2007.

Gholamali Aminian, Hamidreza Arjmandi, Amin Gohari, Masoumeh Nasiri-Kenari, and Urbashi Mitra. Capacity of diffusion-based molecular communication networks over LTI-Poisson channels. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 1(2):188–201, 2015.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *2020 IEEE Information Theory Workshop (ITW)*. IEEE, 2020.

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, and Gregory Wornell. An exact characterization of the generalization error for the gibbs algorithm. *Advances in Neural Information Processing Systems*, 34:8106–8118, 2021a.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 682–687. IEEE, 2021b.

Gholamali Aminian, Saeed Masiha, Laura Toni, and Miguel RD Rodrigues. Learning algorithm generalization error bounds via auxiliary distributions. *arXiv preprint arXiv:2210.00483*, 2022.

Gholamali Aminian, Samuel N Cohen, and Łukasz Szpruch. Mean-field analysis of generalization errors. *arXiv preprint arXiv:2306.11623*, 2023.

Gholamali Aminian, Amir R Asadi, Tian Li, Ahmad Beirami, Gesine Reinert, and Samuel N Cohen. Generalization and robustness of the tilted empirical risk. In *Forty-second International Conference on Machine Learning*, 2025.

Amir R. Asadi. An entropy-based model for hierarchical learning. *Journal of Machine Learning Research*, 25 (187):1–45, 2024.

Amir R. Asadi and Emmanuel Abbe. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research*, 21(139):1–32, 2020.

Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

Milad Bakhshizadeh, Arian Maleki, and Victor H De La Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3):1655–1685, 2023.

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

Peter L. Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Comput.*, 10(8):2159–2173, 1998. doi: 10.1162/089976698300017016. URL https://doi.org/10.1162/089976698300017016.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

Armin Behnamnia, Gholamali Aminian, Alireza Aghaei, Chengchun Shi, Vincent YF Tan, and Hamid R Rabiee. Log-sum-exponential estimator for off-policy evaluation and learning. In *Forty-second International Conference on Machine Learning*, 2025.

Jean-Philippe Bouchaud. Power laws in economics and finance: some ideas fromphysics. *Quantitative finance*, 1(1):105, 2001.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL https://doi.org/10.1162/153244302760200704.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

Olivier Catoni. A PAC-bayesian approach to adaptive classification. 2004. URL https://api.semanticscholar.org/CorpusID:1789271.

Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. ISSN 00361445, 10957200.

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85:45–70, 2019.

Vu C Dinh, Lam S Ho, Binh Nguyen, and Duy Nguyen. Fast learning rates with heavy-tailed losses. *Advances in neural information processing systems*, 29, 2016.

Gintare Karolina Dziugaite and Daniel Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pages 1377–1386. PMLR, 2018.

Sergey Foss, Dmitry Korshunov, Stan Zachary, et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/golowich18a.html.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020.

Maxime Haddouche and Benjamin Guedj. PAC-bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2022.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Lam Si Tung Ho, Binh T Nguyen, Vu Dinh, and Duy Nguyen. Posterior concentration and fast convergence rates for generalized bayesian learning. *Information Sciences*, 538:372–383, 2020.

Sean B Holden and Mahesan Niranjan. On the practical applicability of VC dimension bounds. *Neural Computation*, 7(6):1265–1288, 1995.

Matthew Holland. Pac-bayes under potentially heavy tails. *Advances in Neural Information Processing Systems*, 32, 2019.

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of Gibbs-ERM principle. In *Conference on Learning Theory*, pages 2028–2054, 2019.

Gábor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546. PMLR, 2022.

Gábor Lugosi and Gergely Neu. Online-to-PAC conversions: Generalization bounds via regret analysis. *arXiv preprint arXiv:2305.19674*, 2023.

Pablo A Marquet, Renato A Quiñones, Sebastian Abades, Fabio Labra, Marcelo Tognelli, Matias Arim, and Marcelo Rivadeneira. Scaling and power-laws in ecological systems. *Journal of Experimental Biology*, 208(9):1749–1769, 2005.

Saeed Masiha, Amin Gohari, and Mohammad Hossein Yassaee. f-divergences and their applications in lossy compression and bounding generalization error. *IEEE Transactions on Information Theory*, 2023.

Andreas Maurer et al. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1): 15, 2003.

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

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2004.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.

Ulrich A Müller, Michel M Dacorogna, and Olivier V Pictet. Heavy tails in high-frequency financial data. *A practical guide to heavy tails: Statistical techniques and applications*, pages 55–78, 1998.

Dinesh Karthik Mulumudi, Piyushi Manupriya, Gholamali Aminian, and Anant Raj. On the generalization and robustness in conditional value-at-risk. *arXiv preprint arXiv:2602.18053*, 2026.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

Matthew Nokleby, Ahmad Beirami, and Robert Calderbank. Rate-distortion bounds on bayes risk in supervised learning. In *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

John P. Nolan. *Univariate Stable Distributions*. Springer, 2020.

Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2018.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.

Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*, volume 10. Springer Science & Business Media, 2007.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems*, 33:16833–16845, 2020.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Gennady Samoradnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge, 2017.

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

Didier Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Springer Science & Business Media, 2006.

Didier Sornette and Anne Sornette. General theory of the modified gutenberg-richter law for large seismic moments. *Bulletin of the Seismological Society of America*, 89(4):1121–1130, 1999.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10 (5):988–999, 1999.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Theory of probability and its applications*, pages 11–30. Springer, 1971.

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via wasserstein-based high probability generalisation bounds. *Advances in Neural Information Processing Systems*, 36, 2024.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Chen Zhang, Lei Ren, Jingang Wang, Wei Wu, and Dawei Song. Making pretrained language models good long-tailed learners. *arXiv preprint arXiv:2205.05461*, 2022.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Xitong Zhang, Avrajit Ghosh, Guangliang Liu, and Rongrong Wang. Improving generalization of complex models under unbounded loss using PAC-bayes bounds. *Transactions on Machine Learning Research*, 2024.

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] (Problem setup, heavy-tailed definition, risks, and learning kernel $P_{H|S}$ in Secs. 2–2.2.)
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] (No novel algorithm; applications analyze Gibbs posterior/SGLD theoretically without runtime analysis.)
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable] (Purely theoretical paper; no code is released.)

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] (See Assumptions 5.1, 5.5, 5.9.)
   (b) Complete proofs of all theoretical results. [Yes] (Proofs in Appendices B–D for Secs. 4–6.)
   (c) Clear explanations of any assumptions. [Yes] (Heavy-tailed notion and learning setting explained in Sec. 2; discussion around Thm. 5.3 and related remarks.)

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable] (No experiments are included.)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable] (No experiments are included.)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable] (No experiments are included.)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable] (No experiments are included.)

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable] (No external assets beyond literature citations.)
   (b) The license information of the assets, if applicable. [Not Applicable] (No external assets are used.)
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable] (No new assets are released.)
   (d) Information about consent from data providers/curators. [Not Applicable] (No data is used.)
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] (No data is used.)

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable] (No human-subjects research.)
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable] (No human-subjects research.)
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable] (No human-subjects research.)

Instructions for Paper Submissions to AISTATS 2026: Supplementary Materials

Contents

A Technical Tools

Lemma A.1 (Bernstein's Inequality, Proposition 2.10 in Wainwright, 2019)**.** *Suppose that $S = \{Z_i\}_{i=1}^n$ are i.i.d. random variable such that $|Z_i - \mathbb{E}[Z]| \leq R$ almost surely for all i, and $\mathrm{Var}(Z) = \sigma^2$. Then the following inequality holds with probability at least $(1-\delta)$ under P_S,*

$$\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| \leq \sqrt{\frac{4\sigma^2 \log(2/\delta)}{n}} + \frac{4R\log(2/\delta)}{3n}. \tag{24}$$

Lemma A.2 (Donsker's representation of KL divergence, Theorem 4.6 in Polyanskiy and Wu, 2022)**.** *Let us consider the variational representation of the KL divergence between two probability distributions P and Q on a common space $\mathcal{X}$ given by:*

$$\mathrm{KL}(P||Q) = \sup_f \Big[\int_{\mathcal{X}} f(x)P(\mathrm{d}x) - \log \int_{\mathcal{X}} e^{f(x)}Q(\mathrm{d}x)\Big], \tag{25}$$

where $f \in \mathcal{G} = \{g : \mathcal{X} \to \mathbb{R},\ s.t.\ \mathbb{E}_{X\sim Q}[e^{g(X)}] < \infty\}$.

Lemma A.3 (Pardo, 2018)**.** *Consider $P = \mathcal{N}(h_p, a_p\mathbf{I}_d)$ and $Q = \mathcal{N}(h_q, a_q\mathbf{I}_d)$, where, $h_p, h_q \in \mathbb{R}^d$. Then, we have,*

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\Big(d\frac{a_P}{a_Q} + \frac{\|h_q - h_p\|_2^2}{a_q} - d + d\log\left(\frac{a_q}{a_p}\right)\Big). \tag{26}$$

Lemma A.4. *For all $x \in \mathbb{R}$, we have $1 + x \leq \exp(x)$.*

Lemma A.5. *Suppose that $X > 0$ and $\mathbb{E}[X^{1+\epsilon}] \leq U$. Then, we have for all $\lambda < 0$,*

$$\mathbb{E}\Big[e^{\lambda(X-\mathbb{E}[X])}\Big] \leq e^{\frac{|\lambda|^{1+\epsilon}U}{1+\epsilon}}. \tag{27}$$

Proof. From Lemma 4.1, we have,

$$\begin{aligned}\mathbb{E}\Big[e^{\lambda X}\Big] &\leq 1 + \lambda\mathbb{E}[X] + \frac{|\lambda|^{1+\epsilon}\mathbb{E}[X^{1+\epsilon}]}{1+\epsilon} \\ &\leq e^{\lambda\mathbb{E}[X] + \frac{|\lambda|^{1+\epsilon}\mathbb{E}[X^{1+\epsilon}]}{1+\epsilon}}.\end{aligned} \tag{28}$$

The final result follows from rearranging terms. □

Lemma A.6 (Chernoff Bound Section 2.1.1. in (Wainwright, 2019))**.** *Let X be a random variable with moment-generating function $M_X(t) = \mathbb{E}[e^{tX}]$. Then, for any $t > 0$, the probability that X exceeds a threshold s satisfies*

$$\Pr(X \geq s) \leq \inf_{t>0} \frac{M_X(t)}{e^{ts}}.$$

Similarly, for any $t < 0$,

$$\Pr(X \leq s) \leq \inf_{t<0} \frac{M_X(t)}{e^{ts}}.$$

B Proofs and Details of Section 4

Theorem 4.3. ***(restated)*** *Suppose that $X \leq b$ and there exists $\epsilon \in (0, 1]$ such that $\mathbb{E}[|X - b|^{1+\epsilon}] \leq \infty$. Then, we have,*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X])] \leq \exp\Big(\frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}[|X-b|^{1+\epsilon}]\Big), \tag{29}$$

Furthermore, given n i.i.d samples such that $X_i \leq b$ for all $i \in [n]$, with probability at least $1 - \delta$ for $\delta \in (0, 1)$, we have,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \leq \mathbb{E}[|X-b|^{1+\epsilon}]^{1/(1+\epsilon)}\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{\epsilon}{1+\epsilon}}\big(\frac{\epsilon+1}{\epsilon}\big)^{\left(\frac{\epsilon}{\epsilon+1}\right)}. \tag{30}$$

Proof. If $X \leq b$ and $\lambda > 0$, then we have $\lambda(X-b) \leq 0$, using Lemma 4.1 applied to $-\lambda(X-b)$, we obtain

$$\mathbb{E}[\exp(\lambda(X-b))] \leq 1 + \lambda\mathbb{E}[X-b] + \frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}[|X-b|^{1+\epsilon}]. \tag{31}$$

Consequently, multiplying both sides of (31) by $\exp(-\lambda(\mathbb{E}[X]-b))$, we obtain

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X-\mathbb{E}[X])] &\leq \exp(-\lambda(\mathbb{E}[X]-b))\Big(1+\lambda\mathbb{E}[X-b] + \frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}\big[|X-b|^{1+\epsilon}\big]\Big) \\ &\leq \exp(-\lambda(\mathbb{E}[X]-b))\Big(\exp\Big(\lambda\mathbb{E}[X-b] + \frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}\big[|X-b|^{1+\epsilon}\big]\Big)\Big) \\ &= \exp\Big(\frac{\lambda^{1+\epsilon}}{1+\epsilon}\mathbb{E}[|X-b|^{1+\epsilon}]\Big),\end{aligned} \tag{32}$$

where the last inequality follows from Lemma A.4. The proof is completed by using the Chernoff bound in Lemma A.6. □

C Proofs and Details of Section 5

Theorem 5.4. ***(restated)*** *Under Assumption 5.1, with probability at least $(1-\delta)$, and a finite hypothesis space, the supremum generalization error satisfies,*

$$\sup_{h\in\mathcal{H}} \operatorname{gen}(h,S) \leq \alpha_u^{1/(1+\epsilon)}\Big[\log\big(\operatorname{card}(\mathcal{H})\big) + \log(1/\delta)\Big]^{\frac{\epsilon}{1+\epsilon}}\big(\frac{\epsilon+1}{\epsilon}\big)^{\left(\frac{\epsilon}{\epsilon+1}\right)}. \tag{33}$$

Proof. From Theorem 4.3, by choosing $b=0$ and having $X_i \leq 0$, we have,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \leq \mathbb{E}[|X|^{1+\epsilon}]^{1/(1+\epsilon)}\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{\epsilon}{1+\epsilon}}\big(\frac{\epsilon+1}{\epsilon}\big)^{\left(\frac{\epsilon}{\epsilon+1}\right)}. \tag{34}$$

Assuming the change of variable $Y_i = -X_i$, we have $Y_i \geq 0$ and,

$$\mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^{n} Y_i \leq \mathbb{E}[|Y|^{1+\epsilon}]^{1/(1+\epsilon)}\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{\epsilon}{1+\epsilon}}\big(\frac{\epsilon+1}{\epsilon}\big)^{\left(\frac{\epsilon}{\epsilon+1}\right)}. \tag{35}$$

Assuming $Y_i = \ell(h, Z_i)$, the final results hold by applying the union bound. □

Theorem 5.5. ***(restated)*** *Under Assumption 5.1 under distribution P_S, with probability at least $(1-\delta)$, we have for any $\eta > 0$,*

$$\mathbb{E}_{\rho_h}[\operatorname{gen}(H,S)] \leq \frac{\eta^{\epsilon}\alpha_u}{n^{\epsilon}(1+\epsilon)} + \frac{\mathrm{KL}(\rho_h\|Q_h) + \log(1/\delta)}{\eta}. \tag{36}$$

Proof. Using Lemma A.5, we have for $\lambda < 0$,

$$\mathbb{E}\left[e^{|\lambda|\left(\mathbb{E}_{\tilde{Z}\sim\mu}[\frac{\ell(h,\tilde{Z})}{n}] - \frac{\ell(h,Z_i)}{n}\right)}\right] \leq e^{\frac{|\lambda|^{1+\epsilon}\alpha_u}{n^{1+\epsilon}(1+\epsilon)}} \tag{37}$$

Using i.i.d. assumption with (37), we have,

$$\begin{aligned}\mathbb{E}\left[e^{|\lambda|\left(\mathbb{E}_{\tilde{Z}\sim\mu}[\ell(h,\tilde{Z})] - \widehat{\mathrm{R}}(h,S)\right)}\right] &= \Pi_{i=1}^{n}\mathbb{E}\left[e^{|\lambda|\left(\mathbb{E}_{\tilde{Z}\sim\mu}[\frac{\ell(h,\tilde{Z})}{n}] - \frac{\ell(h,Z_i)}{n}\right)}\right] \\ &\leq e^{\frac{|\lambda|^{1+\epsilon}\alpha_u}{n^{\epsilon}(1+\epsilon)}}.\end{aligned} \tag{38}$$

Using Donsker and Varadhan's variational formula, we have $\forall \rho_h \in \mathcal{P}(\mathcal{H})$,

$$\mathbb{E}\left[e^{|\lambda|\mathbb{E}_{\rho_h}[\text{gen}(H,S)]-\frac{|\lambda|^{1+\epsilon}\alpha_u}{n^\epsilon(1+\epsilon)}-\text{KL}(\rho_h\|Q_h)}\right] \leq 1. \tag{39}$$

Then, using Chernoff bound, Lemma A.6 for $t=1$ and $s>0$, we have,

$$\begin{aligned}
&\mathbb{P}_S\Big(|\lambda|\mathbb{E}_{\rho_h}[\text{gen}(H,S)] - \frac{|\lambda|^{1+\epsilon}\alpha_u}{n^\epsilon(1+\epsilon)} - \text{KL}(\rho_h\|Q_h) > s\Big) \\
&\leq \mathbb{E}\left[e^{|\lambda|\mathbb{E}_{\rho_h}[\text{gen}(H,S)]-\frac{|\lambda|^{1+\epsilon}\alpha_u}{n^\epsilon(1+\epsilon)}-\text{KL}(\rho_h\|Q_h)}\right] e^{-s} \\
&\leq e^{-s}.
\end{aligned} \tag{40}$$

Now setting $e^{-s}=\delta$ and $s=\log(1/\delta)$, we obtain the following:

$$\mathbb{P}_S\Big(|\lambda|\mathbb{E}_{\rho_h}[\text{gen}(H,S)] - \frac{|\lambda|^{1+\epsilon}\alpha_u}{n^\epsilon(1+\epsilon)} - \text{KL}(\rho_h\|Q_h) > \log(1/\delta)\Big) \leq \delta. \tag{41}$$

This in turn implies

$$\mathbb{P}_S\Big(\mathbb{E}_{\rho_h}[\text{gen}(H,S)] > \frac{|\lambda|^{\epsilon}\alpha_u}{n^\epsilon(1+\epsilon)} + \frac{\text{KL}(\rho_h\|Q_h)+\log(1/\delta)}{|\lambda|}\Big) \leq \delta. \tag{42}$$

Taking the complement completes the proof. □

Remark 5.6 Discussion:

Let $C(H) = KL(\rho_h||Q_h) + \log(1/\delta)$. The bound in Theorem 5.5 is given by:

$$g(\eta) = \frac{\alpha_u \eta^\epsilon}{n^\epsilon(1+\epsilon)} + \frac{C(H)}{\eta}.$$

To find the optimal η, we take the derivative with respect to η and set it to zero:

$$\frac{\partial g(\eta)}{\partial \eta} = \frac{\epsilon\alpha_u\eta^{\epsilon-1}}{n^\epsilon(1+\epsilon)} - \frac{C(H)}{\eta^2} = 0.$$

Solving for $\eta^\star$, we obtain:

$$\eta^\star = \left(\frac{n^\epsilon(1+\epsilon)C(H)}{\epsilon\alpha_u}\right)^{\frac{1}{1+\epsilon}} = n^{\frac{\epsilon}{1+\epsilon}}\left(\frac{(1+\epsilon)C(H)}{\epsilon\alpha_u}\right)^{\frac{1}{1+\epsilon}}.$$

Substituting η^* back into $g(\eta)$, we recover the convergence rate of $O(n^{-\frac{\epsilon}{1+\epsilon}})$ with the explicit constant:

$$g(\eta^\star) = \frac{1}{n^{\frac{\epsilon}{1+\epsilon}}} \cdot \alpha_u^{\frac{1}{1+\epsilon}} C(H)^{\frac{\epsilon}{1+\epsilon}} \left(\frac{1+\epsilon}{\epsilon}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

Theorem 5.7. ***(restated)*** *Suppose that Assumption 5.2 holds. Then,*

$$\overline{\text{gen}}(H,S) \leq \frac{1}{n}\sum_{i=1}^{n} \frac{2}{1+\epsilon}\big(D(P_{Z_i,H}\|\mu\otimes P_H)\big)^{\epsilon/(1+\epsilon)}\mathbb{E}_{\mu\otimes P_H}\big[\ell(H,Z_i)^{(1+\epsilon)}\big]^{1/(1+\epsilon)}. \tag{43}$$

Proof. Note that we have

$$\overline{\text{gen}}(H,S) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_H\otimes\mu}[\ell(h,Z_i)] - \mathbb{E}_{P_{H,Z_i}}[\ell(H,Z_i)]. \tag{44}$$

We first provide an upper bound on $\mathbb{E}_{P_H\otimes\mu}[\ell(h,Z_i)] - \mathbb{E}_{P_{H,Z_i}}[\ell(H,Z_i)]$ using Lemma A.2. From Lemma A.2, we know that for any $\lambda \in \mathbb{R}_-$,

$$D(P_{Z_i,H}\|\mu\otimes P_H) \geq \mathbb{E}_{P_{Z_i,H}}[\lambda\ell(H,Z_i)] - \log \mathbb{E}_{\mu\otimes P_H}[e^{\lambda\ell(H,Z_i)}]. \tag{45}$$

For $\lambda < 0$ and using Lemma 4.1 and Lemma A.4, we have

$$\begin{aligned}\mathbb{E}_{\mu\otimes P_H}[\exp\left(\lambda\ell(H,Z_i)\right)] &\leq 1 + \lambda\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)] + \frac{|\lambda|^{(1+\epsilon)}}{1+\epsilon}\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)^{(1+\epsilon)}] \\ &\leq \exp\left(\lambda\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)] + \frac{|\lambda|^{(1+\epsilon)}}{1+\epsilon}\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)^{(1+\epsilon)}]\right).\end{aligned} \tag{46}$$

This implies

$$\log\left(\mathbb{E}_{\mu\otimes P_H}[\exp\left(\lambda\ell(H,Z_i)\right)]\right) \leq \lambda\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)] + \frac{|\lambda|^{(1+\epsilon)}}{1+\epsilon}\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)^{(1+\epsilon)}]. \tag{47}$$

Combining (47) with (45) gives

$$\begin{aligned}D(P_{Z_i,H}\|\mu\otimes P_H) &\geq \mathbb{E}_{P_{Z_i,H}}[\lambda\ell(H,Z_i)] - \log \mathbb{E}_{\mu\otimes P_H}[e^{\lambda\ell(H,Z_i)}] \\ &\geq \lambda\Big(\mathbb{E}_{P_{Z_i,H}}[\ell(H,Z_i)] - \mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)]\Big) - \frac{|\lambda|^{(1+\epsilon)}}{1+\epsilon}\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)^{(1+\epsilon)}].\end{aligned} \tag{48}$$

Rearranging (48), we have for $\lambda < 0$,

$$\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)] - \mathbb{E}_{P_{Z_i,H}}[\ell(H,Z_i)] \leq \frac{D(P_{Z_i,H}\|\mu\otimes P_H)}{-\lambda} + \frac{|\lambda|^{\epsilon}}{1+\epsilon}\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)^{(1+\epsilon)}]. \tag{49}$$

Choosing $\lambda = -\frac{(1+\epsilon)\left(D(P_{Z_i,H}\|\mu\otimes P_H)\right)^{1/(1+\epsilon)}}{\mathbb{E}_{\mu\otimes P_H}\left[\ell(H,Z_i)^{(1+\epsilon)}\right]^{1/(1+\epsilon)}}$, we have,

$$\begin{aligned}&\mathbb{E}_{\mu\otimes P_H}[\ell(H,Z_i)] - \mathbb{E}_{P_{Z_i,H}}[\ell(H,Z_i)] \\ &\quad\leq \frac{2}{1+\epsilon}\big(D(P_{Z_i,H}\|\mu\otimes P_H)\big)^{\epsilon/(1+\epsilon)}\mathbb{E}_{\mu\otimes P_H}\left[\ell(H,Z_i)^{(1+\epsilon)}\right]^{1/(1+\epsilon)}.\end{aligned} \tag{50}$$

The result follows by noting that $I(Z_i;H) = D(P_{Z_i,H}\|\mu\otimes P_H)$. □

Corollary 5.8. ***(restated)*** *Assuming bounded $I(H;S)$, then we have the convergence rate of $O(n^{\frac{-\epsilon}{1+\epsilon}})$ for the upper bound in Theorem 5.7.*

Proof. Due to the i.i.d. assumption, we have,

$$\begin{aligned}\overline{\text{gen}}(H,S) &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{2}{1+\epsilon}\big(I(H;Z_i)\big)^{\epsilon/(1+\epsilon)}\alpha_\varepsilon^{1/(1+\epsilon)} \\ &\leq \Big[\frac{1}{n}\sum_{i=1}^{n}I(H;Z_i)\Big]^{\frac{\epsilon}{1+\epsilon}}\alpha_\varepsilon^{1/(1+\epsilon)} \\ &\leq \Big[\frac{I(H;S)}{n}\Big]^{\frac{\epsilon}{1+\epsilon}}\alpha_\varepsilon^{1/(1+\epsilon)},\end{aligned} \tag{51}$$

where the last inequality follows from chain rule property of mutual information. □

D Proofs and Details of Section 6

Theorem 6.1. ***(restated)*** *Assume that Assumption 5.2 holds. Then, the following upper bound holds on the expected generalization error of the Gibbs posterior,*

$$0 \leq \overline{\text{gen}}(H,S) \leq \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1}\alpha_{\varepsilon}\Big(\frac{\gamma}{n}\Big)^{\epsilon}. \tag{52}$$

Proof. From (Aminian et al., 2021a, Theorem 1), we know that for all loss functions we have,

$$\overline{\text{gen}}(H,S) = \frac{I_{\text{SKL}}(H;S)}{\gamma}, \tag{53}$$

where $I_{\text{SKL}}(H;S)$ is the symmetrized KL information as defined in (Aminian et al., 2015). Note that, we have $I(H;S) \leq I_{\text{SKL}}(H;S)$. Using Theorem 5.7, we have

$$\frac{I(H;S)}{\gamma} \leq \overline{\text{gen}}(H,S) \leq \Big(\frac{I(H;S)}{n}\Big)^{\frac{\epsilon}{1+\epsilon}}\alpha_{\varepsilon}^{1/(1+\epsilon)}. \tag{54}$$

Therefore, we have,

$$\frac{I(H;S)}{\gamma} \leq \Big(\frac{I(H;S)}{n}\Big)^{\frac{\epsilon}{1+\epsilon}}\alpha_{\varepsilon}^{1/(1+\epsilon)}, \tag{55}$$

where results in,

$$I(H;S)^{\frac{\epsilon}{1+\epsilon}} \leq \frac{\gamma^{\epsilon}}{n^{\frac{\epsilon^2}{1+\epsilon}}}\alpha_{\varepsilon}^{\epsilon/(1+\epsilon)}. \tag{56}$$

The final result holds by combining (56) with the upper bound in (51). □

Proposition 6.2. ***(restated)*** *Assume that the loss function is L-Lipschitz for all $z \in \mathcal{Z}$ and let $\mathcal{H} \in \mathbb{R}^d$. Then, under the Gibbs posterior, we have,*

$$\mathcal{E}(H,\mu) \leq \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1}\alpha_{\varepsilon}\Big(\frac{\gamma}{n}\Big)^{\epsilon} + \frac{1}{\gamma}\text{KL}\Big(\mathcal{N}\big(h^{\star},\beta^2\mathbf{I}_d\big)\|\pi(H)\Big) + L\beta\sqrt{d}, \tag{57}$$

where $h^{\star} = \arg\inf_{h\in\mathcal{H}} R(h,\mu)$ and we assumed that $\text{KL}\Big(\mathcal{N}\big(h^{\star},\frac{\mathbf{I}_d}{\sqrt{\gamma}}\big),\pi(H)\Big) < \infty$ for given $\pi(H)$.

Proof. We consider the following decomposition of the excess risk,

$$\mathcal{E}(H,\mu) = \overline{\text{gen}}(H,S) + \mathbb{E}_{P_{H,S}}[\widehat{\text{R}}(H,S)] - \inf_{h\in\mathcal{H}}\text{R}(h,\mu). \tag{58}$$

Then, for bounding $\overline{\text{gen}}(H,S)$ we can apply Theorem 6.1. Note that under the Gibbs posterior, we have

$$\begin{aligned}
\mathbb{E}_{P^{\gamma}_{H|S}P_S}[\widehat{\text{R}}(H,S)] &\leq \mathbb{E}_{P^{\gamma}_{H|S}P_S}[\widehat{\text{R}}(H,S)] + \frac{1}{\gamma}\text{KL}(P^{\gamma}_{H|S}\|\pi(H)|P_S) \\
&\leq \mathbb{E}_{Q_H P_S}[\widehat{\text{R}}(H,S)] + \frac{1}{\gamma}\text{KL}(Q_H\|\pi(H)|P_S) \\
&= \mathbb{E}_{Q_H}[\text{R}(H,\mu)] + \frac{1}{\gamma}\text{KL}(Q_H\|\pi(H)|P_S),
\end{aligned} \tag{59}$$

where $Q_H = \mathcal{N}\big(h^{\star},\beta^2\mathbf{I}_d\big)$. Thus,

$$\begin{aligned}
\mathbb{E}_{Q_H}[\text{R}(H,\mu)] &= \mathbb{E}_{Q_H}[\text{R}(h^{\star},\mu) + \text{R}(H,\mu) - \text{R}(h^{\star},\mu)] \\
&\leq \text{R}(h^{\star},\mu) + L\mathbb{E}_{Q_H}[\|H - h^{\star}\|_2] \\
&= \beta L\sqrt{d}.
\end{aligned} \tag{60}$$

It completes the proof. □

Remark 6.3 Discussion: From Lemma A.3, we have,

$$\mathrm{KL}(\mathcal{N}(h^\star, \beta^2\mathbf{I}_d)\|\mathcal{N}(h_Q, \beta^2\mathbf{I}_d)) = \frac{\|h^\star - h_Q\|_2^2}{2\beta^2}. \tag{61}$$

Using (61) in Proposition 6.2, we have,

$$\mathcal{E}(H,\mu) \leq \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1}\alpha_\varepsilon\Big(\frac{\gamma}{n}\Big)^{\epsilon} + \frac{\|h^\star - h_Q\|_2^2}{2\gamma\beta^2} + L\beta\sqrt{d}. \tag{62}$$

Choosing $\gamma \asymp n^{\frac{3\epsilon}{2(1+\epsilon)}}$ and $\beta \asymp n^{\frac{-\epsilon}{2(1+\epsilon)}}$, we have,

$$\mathcal{E}(H,\mu) \leq \Big(\frac{2}{1+\epsilon}\Big)^{\epsilon+1}\alpha_\varepsilon\frac{1}{n^{\frac{\epsilon(2-\epsilon)}{2(1+\epsilon)}}} + \frac{\|h^\star - h_Q\|_2^2}{2n^{\frac{\epsilon}{2(1+\epsilon)}}} + \frac{L\sqrt{d}}{n^{\frac{\epsilon}{2(1+\epsilon)}}}. \tag{63}$$

Therefore, the final convergence rate is $O\Big(\max\Big(n^{\frac{-\epsilon}{2(1+\epsilon)}}, n^{\frac{\epsilon(\epsilon-2)}{2(1+\epsilon)}}\Big)\Big)$.

Theorem 6.6. ***(restated)*** *Under Assumptions 5.2, 6.4 and 6.5, we can derive the following upper bound on the expected generalization error,*

$$\overline{\mathrm{gen}}(H,S) \leq \frac{2}{(1+\epsilon)n^{\epsilon/(1+\epsilon)}}\Big(\frac{L^2}{2}\sum_{t=1}^{T}\eta_t\Big)^{\epsilon/(1+\epsilon)}\alpha_\varepsilon^{1/(1+\epsilon)}. \tag{64}$$

Proof. From (Pensia et al., 2018, Theorem 1), under Assumptions 6.4 and 6.5, we have,

$$I(H^{(T)};S) \leq \frac{L^2}{2}\sum_{t=1}^{T}\eta_t. \tag{65}$$

From (51), we have,

$$\overline{\mathrm{gen}}(H,S) \leq \Big[\frac{I(H^{(T)};S)}{n}\Big]^{\frac{\epsilon}{1+\epsilon}}\alpha_\varepsilon^{1/(1+\epsilon)}. \tag{66}$$

The final result follows from combining (65) with (66). □

Lemma 5.9. *Suppose that $X > 0$ and $\gamma < 0$, then we have for $\epsilon \in (0,1]$,*

$$\mathrm{Var}(\exp(\gamma X)) \leq |\gamma|^{1+\epsilon}\mathbb{E}[X^{1+\epsilon}]. \tag{67}$$

We mention the similar proof in (Behnamnia et al., 2025).

Proof. By the mean value theorem, for each realisation $X(\omega)$ of X for an element ω of its underlying probability space there is a value $c(\omega)$ in the interval between $X(\omega)$ and $C \in \mathbb{R}^+$ such that

$$\exp(\gamma X(\omega)) - \exp(\gamma C) = \gamma(X - C)\exp(\gamma c(\omega)).$$

As $X > 0$ we have $c(\omega) > 0$. Moreover,

$$\begin{aligned}
\mathrm{Var}(\exp(\gamma X)) &= \mathbb{E}[(\exp(\gamma X) - \mathbb{E}[\exp(\gamma X)])^2] \\
&\overset{(a)}{=} \min_{C\in\mathbb{R}^+} \mathbb{E}[(\exp(\gamma X) - \exp(\gamma C))^2] \\
&= \min_{C\in\mathbb{R}^+} \mathbb{E}[|\exp(\gamma X) - \exp(\gamma C)|^{1+\epsilon} |\exp(\gamma X) - \exp(\gamma C)|^{1-\epsilon}] \\
&\overset{(b)}{\leq} \min_{C\in\mathbb{R}^+} \mathbb{E}[|\gamma|^{1+\epsilon} \exp((1+\epsilon)\gamma c)|X - C|^{1+\epsilon} |\exp(\gamma X) - \exp(\gamma C)|^{1-\epsilon}] \\
&\overset{(c)}{\leq} \min_{C\in\mathbb{R}^+} \mathbb{E}[|\gamma|^{1+\epsilon} \exp((1+\epsilon)\gamma c)|X - C|^{1+\epsilon}] \\
&\overset{(d)}{\leq} |\gamma|^{1+\epsilon} \mathbb{E}[X^{1+\epsilon}],
\end{aligned}$$

where (a) follows from the minimum mean square representation, (b) follows from the mean-value theorem, and (c) follows from the fact that $|\exp(\gamma X) - \exp(\gamma C)| \leq 1$. (d) follows from choosing $C = 0$ and the fact that the γX is negative and we have $\exp((1+\epsilon)\gamma c) \leq 1$. □

Theorem 5.10. *Under Assumption 5.3, we can derive the following upper bound on absolute expected generalization error,*

$$\left|\overline{\mathrm{gen}}(H,S)\right| \leq \begin{cases} \frac{1}{|\lambda|}\sqrt{2|\lambda|^{1+\epsilon}\beta_\varepsilon \frac{I(H;S)}{n}} + \frac{|\lambda|^\epsilon \beta_\varepsilon}{1+\epsilon} & \text{if } \frac{I(H;S)}{n} \leq \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \\ \frac{I(H;S)}{|\lambda| n} + \frac{2|\lambda|^\epsilon \beta_\varepsilon}{1+\epsilon} & \text{if } \frac{I(H;S)}{n} > \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \end{cases}. \tag{68}$$

Proof. Using Lemma 5.9 under Assumption 5.3, we have,

$$\mathrm{Var}(\exp(\lambda \ell(H,\tilde{Z}))) \leq |\lambda|^{1+\epsilon}\mathbb{E}[\ell(H,\tilde{Z})^{1+\epsilon}] \leq |\lambda|^{1+\epsilon}\beta_\varepsilon.$$

Furthermore, note that, for $\lambda < 0$, we have

$$0 \leq \exp(\lambda \ell(H,\tilde{Z})) \leq 1,$$

therefore, the variable $\exp(\lambda \ell(H,\tilde{Z}))$ is sub-exponential with parameters $(|\lambda|^{1+\epsilon}\beta_\varepsilon, 1)$ under the distribution $P_H \otimes \mu$. Using the approach in (Bu et al., 2020; Aminian et al., 2021a) for the sub-exponential case, we have

$$\begin{aligned}
&\left|\mathbb{E}_{P_H\otimes\mu}[\exp(\lambda \ell(H,\tilde{Z}))] - \mathbb{E}_{P_{H,S}}[\frac{1}{n}\sum_{i=1}^n \exp(\lambda \ell(H,Z_i))]\right| \\
&\leq \begin{cases} \sqrt{2|\lambda|^{1+\epsilon}\beta_\varepsilon \frac{I(H;S)}{n}} & \text{if } \frac{I(H;S)}{n} \leq \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \\ \frac{I(H;S)}{n} + \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} & \text{if } \frac{I(H;S)}{n} > \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \end{cases}.
\end{aligned} \tag{69}$$

Note that, for $\lambda < 0$ and using Lemma 4.1 and Lemma A.4, we have,

$$\begin{aligned}
1 + \mathbb{E}_{P_H\otimes\mu}[\lambda \ell(H,\tilde{Z})] \leq \mathbb{E}_{P_H\otimes\mu}[\exp(\lambda \ell(H,\tilde{Z}))] \leq 1 + \lambda \mathbb{E}_{P_H\otimes\mu}[\ell(H,\tilde{Z})] + \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{1+\epsilon}, \\
1 + \mathbb{E}_{P_{H,Z}}[\lambda \ell(H,Z)] \leq \mathbb{E}_{P_{H,Z}}[\exp(\lambda \ell(H,Z))] \leq 1 + \lambda \mathbb{E}_{P_{H,Z}}[\ell(H,Z)] + \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{1+\epsilon}.
\end{aligned} \tag{70}$$

Therefore,

$$\begin{aligned}
-|\lambda|\overline{\mathrm{gen}}(H,S) - \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{1+\epsilon} &\leq \mathbb{E}_{P_H\otimes\mu}[\exp(\lambda \ell(H,\tilde{Z}))] - \mathbb{E}_{P_{H,Z}}[\exp(\lambda \ell(H,Z))] \\
&\leq |\lambda|\overline{\mathrm{gen}}(H,S) + \frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{1+\epsilon}.
\end{aligned} \tag{71}$$

Thus,

$$\begin{aligned}&\left|\mathbb{E}_{P_H\otimes\mu}[\ell(H,\tilde{Z})]-\mathbb{E}_{P_{H,Z}}[\ell(H,Z)]\right| \\ &\leq \begin{cases} \frac{1}{|\lambda|}\sqrt{2|\lambda|^{1+\epsilon}\beta_\varepsilon \frac{I(H;S)}{n}}+\frac{|\lambda|^\epsilon\beta_\varepsilon}{1+\epsilon} & \text{if } \frac{I(H;S)}{n}\leq\frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \\ \frac{I(H;S)}{|\lambda|n}+\frac{2|\lambda|^\epsilon\beta_\varepsilon}{1+\epsilon} & \text{if } \frac{I(H;S)}{n}>\frac{|\lambda|^{1+\epsilon}\beta_\varepsilon}{2} \end{cases},\end{aligned} \tag{72}$$

which completes the proof. □