AN INFORMATION-THEORETIC FRAMEWORK FOR STABILITY-BASED GENERALIZATION BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We derive a family of novel bounds on the expected generalization error of learning algorithms. Our bounds simultaneously leverage two properties of the learning algorithm: 1) algorithmic stability and 2) information measures of the dependence between the training data and the learned hypothesis. Our approach establishes a general bound based on f-divergences and allows for various specializations under a series of new stability conditions. These conditions are weaker than established notions of stability from the literature which makes them applicable to a wider range of learning scenarios, especially when the loss function is unbounded. When evaluated with several concrete examples, our bounds strictly improve upon both the classical stability-based and information-theoretic bounds.

1 INTRODUCTION

We study generalization in the standard statistical learning setting (Shalev-Shwartz & Ben-David, 2014). Assume that we have access to a training dataset $Z_{[n]} := (Z_1, \dots, Z_n)$ which is composed of n observations sampled i.i.d. from an unknown data generating distribution P_Z over a known sample space $\mathcal{Z}$. Fed with the training samples $Z_{[n]}$, a (possibly randomized) learning algorithm $\mathcal{A} : \mathcal{Z}^n \times \mathcal{R} \to \mathcal{W}$ produces a hypothesis $W := \mathcal{A}(Z_{[n]}, R)$, where $R \in \mathcal{R}$ captures the randomness of the algorithm and R is independent of $Z_{[n]}$. For instance, in the case of stochastic gradient descent (SGD) which is commonly used to train neural networks, R may encode the random choice of the training samples used at each update. Formally, given a non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [0, +\infty)$, the quality of a hypothesis $w \in \mathcal{W}$ is evaluated through the population loss defined as $L(w, P_Z) := \mathbb{E}_{Z\sim P_Z}[\ell(w, Z)]$. In practice, however, it is intractable to evaluate the true loss $L(w, P_Z)$ since the underlying distribution P_Z is unknown. Instead, the available training samples $Z_{[n]}$ are exploited to compute an estimate of $L(w, P_Z)$, namely the empirical loss defined by $\hat{L}(w, Z_{[n]}) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$. Moreover, the difference between $L(w, P_Z)$ and $\hat{L}(w, Z_{[n]})$ is defined as the generalization error. In this paper, we analyze its expected value $\overline{\mathrm{gen}} := \mathbb{E}[L(W, P_Z) - \hat{L}(W, Z_{[n]})]$, where the expectation is with respect to $P_{Z_{[n]},R}$.

One established approach to studying the generalization error is based on algorithmic stability (Kearns & Ron, 1997; Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Mukherjee et al., 2006). The underlying intuition is that a *stable* learning algorithm, in the sense that it produces similar results when trained on only slightly perturbed data, tends to generalize better. To make this intuition precise, several specific notions of stability have been proposed to study the generalization performance. Amongst them, one of the most commonly used instances is the notion of *uniform stability* defined in Hardt et al. (2016, Definition 2.1): a (randomized) learning algorithm $\mathcal{A}$ is said to be β-uniformly-stable if

$$\sup_{z_{[n]}, z'_i, z} \mathbb{E}_R\left[\ell(\mathcal{A}(z_{[n]\sim i}, R), z) - \ell(\mathcal{A}(z_{[n]}, R), z)\right] \leq \beta, \tag{1}$$

where we introduce $z'_{[n]} := (z'_1, \dots, z'_n)$ as an independent copy of $z_{[n]}$ and $z_{[n]\sim i}$ denotes the replace-one training set obtained by replacing z_i with z'_i. Hardt et al. (2016, Theorem 2.2) introduces a fundamental property of uniform stability:

$$\overline{\mathrm{gen}} \leq \beta, \tag{2}$$

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

which, in plain words, implies that this uniform stability directly controls the generalization error in expectation.

Information theory provides a complementary approach for bounding the expected generalization error. Initiated in the seminal works of Russo & Zou (2016); Xu & Raginsky (2017), the very first information-theoretic bound hinges on the classical Shannon mutual information (MI) $I(W; Z_{[n]})$ which measures how much information the output of a learning algorithm $\mathcal{A}$ contains about its input training samples $Z_{[n]}$. The bound was further tightened and advanced to the well-known individual-sample mutual information (ISMI) bound (Bu et al., 2020) as follows:

$$\overline{\text{gen}} \leq \frac{1}{n} \sum_{i=1}^{n} \inf_{t>0} \left\{ \frac{1}{t} \left(I(W; Z_i) + \psi(t) \right) \right\}, \tag{3}$$

where ψ defined on $(0, a)$ is an upper bound on the cumulant generating function (CGF) of the centered loss $\ell(W, Z) - \mathbb{E}[\ell(W, Z))]$ under the product distribution $P_W P_Z$. For ξ^2-sub-Gaussian loss functions, we can choose $\psi(t) = \frac{\xi^2 t^2}{2}$, then (3) admits an analytical expression as

$$\overline{\text{gen}} \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\xi^2 I(W; Z_i)}. \tag{4}$$

Inspired by Wang & Mao (2023a), our work combines the notion of stability and information measures to develop bounds for *stable* learning algorithms. In general, the proposed bounds can be expressed in this form:

$$\overline{\text{gen}} \leq \frac{1}{n} \sum_{i=1}^{n} \inf_{t>0} \left\{ \frac{1}{t} \left(I_f(W; Z_i | W_i') + \tilde{\psi}(\beta', t) \right) \right\}, \tag{5}$$

where:

- $W_i' := \mathcal{A}(Z_{[n]\sim i}, R)$ is the *neighboring* hypothesis obtained by training on the replace-one training set $Z_{[n]\sim i}$ and the same randomness R shared with W, which is trained on $Z_{[n]}$.
- $I_f(W; Z_i \mid W_i')$ is an information measure (defined formally in Definition 2) which quantifies the dependence between the original hypothesis W and the i-th training sample Z_i given the neighboring hypothesis W_i'.
- β' is a stability parameter (see, e.g., formal Definition 6) incurred by the learning algorithm. The stability notions developed in this paper require less restrictive conditions than several existing notions which often do not hold for unbounded loss functions. In this sense, our stability definitions allow to expand the range of *stable* algorithms.
- $\tilde{\psi}(\beta', t)$ is a function of the stability parameter β' and an auxiliary parameter t which can be optimized. One specialization of this general bound leads to

$$\overline{\text{gen}} \leq \frac{\beta'}{n} \sum_{i=1}^{n} \sqrt{2 I_f(W; Z_i | W_i')}, \tag{6}$$

  which follows from the formal results in Theorem 1 and Corollary 4.

We show benefits of the proposed bounds through some simple but inspiring learning scenarios. We give below a brief illustrative example of how our bounds improve upon already existing ones.

Consider, for instance, the problem of estimating the Bernoulli mean. Specifically, suppose $Z_i \overset{i.i.d.}{\sim} \text{Bern}(\theta)$ and the task is to estimate the unknown mean θ. Using $\ell(w, z) = (w - z)^2$, the empirical risk minimization (ERM) algorithm outputs $W = \frac{1}{n} \sum_{i=1}^{n} Z_i$, and the expected generalization error $\overline{\text{gen}}$ is exactly $\frac{2\theta(1-\theta)}{n}$. In this case, the bound (2), based solely on stability, gives $\overline{\text{gen}} \leq \frac{2n-1}{n^2}$, while the existing information-theoretic bounds (including the ISMI and CMI bounds) are sub-optimal with order $O(\frac{1}{\sqrt{n}})$ (Lu et al., 2026). In contrast, our proposed bound (Corollary 3) gives $\overline{\text{gen}} \leq \frac{2\sqrt{2}\theta(1-\theta)}{n} + o(\frac{1}{n})$, which strictly improves upon both stability-based and information-theoretic bounds. Besides, we provide in Section 5 additional examples (e.g., linear regression)

where (2) is not applicable while the proposed bounds yield non-vacuous result and are even tighter than the fast-rate information-theoretic bounds (Wu et al., 2025).

We summarize below the main contribution of this paper.

- We first prove an information-theoretic bound (Theorem 1) using mutual information and a new notion of distribution-dependent stability. This stability notion is weaker than the *strong uniform stability*. The resultant bound bridges the information-theoretic and stability-based generalization bounds and therefore combines their advantages.
- We provide a general information-theoretic generalization bound (Theorem 2) using the f-divergence and generalized CGF. The general bound exhibits flexibility in terms of more choices of information measures beyond the mutual information. It also enables various specializations when invoking some new stability conditions. These stability notions impose less limiting requirements on the learning algorithm, thus endowing our new bounds with wider applicability.
- We demonstrate the usefulness of the proposed bounds via several examples in Section 5. In these examples, the existing bounds are either inapplicable or order-wise (in terms of training data size n) looser than the proposed bounds.

2 Related Works

Stability-Based Framework. Early analyses of algorithmic stability focused primarily on bounds that hold with high probability, whereas bounds in expectation often follow more directly from the stability definition. A framework comprising a family of stability notions was introduced by Bousquet & Elisseeff (2002) and used to develop high-probability bounds for deterministic learning algorithms. Subsequent works conceptually extended the stability family to cover almost stable algorithms (Kutin & Niyogi, 2002) and randomized learning algorithms (Elisseeff et al., 2005). Foster et al. (2019) introduced a family of stability relying on data-dependent hypothesis sets, thereby admitting the conventional notions of algorithmic stability as special cases. Other notions were developed specifically for particular learning scenarios, including k-fold cross-validation (Kale et al., 2011) and randomized iterative algorithms such as SGD (Hardt et al., 2016).

Information-Theoretic Framework. The information-theoretic approach for bounding generalization errors was pioneered by Russo & Zou (2016), and further developed by Xu & Raginsky (2017). Bu et al. (2020) sharpened the original information-theoretic bounds by quantifying mutual information on each individual training sample. The chaining methods in high dimensional probability can further exploit dependencies among hypotheses, although they require an auxiliary random process adapted to the hypothesis space (Asadi et al., 2018; Zhou et al., 2023; Clerico et al., 2022). Beyond mutual information, other information measures are further employed to obtain tighter bounds (Aminian et al., 2021b; Rodríguez Gálvez et al., 2021; Aminian et al., 2021a; Liu et al., 2025). Notably, Steinke & Zakynthinou (2020) introduced an adjusted information-theoretic framework, under the name of conditional mutual information (CMI). Using an augmented supersample construction, the resulting CMI-based bounds are bounded by nature. Combining the CMI framework with the individual-sample technique allows for the derivation sharper bounds (Zhou et al., 2022). Further extensions and refinements of the CMI framework have since been made in various ways (Haghifam et al., 2020; Harutyunyan et al., 2021; Hellström & Durisi, 2022; Rammal et al., 2022; Wang & Mao, 2023b; Lu et al., 2025; Dong et al., 2025).

Combination of Two Frameworks. It is of theoretical and practical interest to combine information-theoretic analysis of generalization with stability notions. Raginsky et al. (2016); Li et al. (2019); Steinke & Zakynthinou (2020) mainly views various information measures as stability notions. For instance, the erasure mutual information and CMI have both been interpreted as stability measures (Raginsky et al., 2016; Steinke & Zakynthinou, 2020). In contrast, some other literature derived structured high-probability bounds (London, 2017; Rivasplata et al., 2018) and average bounds (Wang & Mao, 2023a), which each comprises typically a stability parameter and an information measure term. In spirit to these works, our work presents a general f-divergence average bound with various specializations under several stability conditions, including mild conditions that remain meaningful when the previous stability notions (of, e.g., London (2017); Wang & Mao

(2023a)) do not. Besides, we show that the MI-based bounds in Wang & Mao (2023a, Section 3) do not hold as stated, with counterexamples and further discussions provided in Appendix O.

3 PRELIMINARIES

We use uppercase letters (e.g., X) for random variables, lowercase letters (e.g., x) for their realizations, and calligraphic letters (e.g., $\mathcal{X}$) for the corresponding alphabets. We write P_X for the distribution of X, and $P_{X|Y}$ for the conditional distribution of X given Y. When conditioning on a specific realization $Y = y$, we use $P_{X|Y=y}$ or simply $P_{X|y}$. The notation $X \perp\!\!\!\perp Y$ means that X and Y are independent. For a positive integer n, we use $[n]$ as a shorthand of the set $\{1, \dots, n\}$. We use Bern, Bino and $\mathcal{N}$ to represent Bernoulli, binomial and Gaussian distributions, respectively. For $X \sim \mathrm{Bern}(\theta)$, its binary entropy is $H_2(\theta) := -\theta \log \theta - (1-\theta)\log(1-\theta)$. All logarithms are to the base e.

Our analysis and results will be relying on the following definitions and lemmas.

Definition 1. *(f-Divergence (Polyanskiy & Wu, 2025)) Let $P, Q \in \mathcal{P}(\mathcal{X})$ satisfy $P \ll Q$, and let $f : \mathbb{R}^+ \to \mathbb{R}$ be convex with $f(1) = 0$. The f-divergence between P and Q is defined as $D_f(P\|Q) := \mathbb{E}_Q[f(\frac{\mathrm{d}P}{\mathrm{d}Q})]$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative.*

Definition 2. *(f-Information (Polyanskiy & Wu, 2025)) Let f be as defined above. The f-information between two random variables X and Y is defined as $I_f(X;Y) := D_f(P_{X,Y}\|P_X P_Y)$. When conditioned on $Z = z$, the disintegrated conditional f-information is defined as $I_f^z(X;Y) = I_f(X;Y|Z=z) := D_f(P_{X,Y|Z=z}\|P_{X|Z=z}P_{Y|Z=z})$. Averaging over Z gives the (integrated) conditional f-information $I_f(X;Y|Z) = \mathbb{E}_Z[I_f^Z(X;Y)]$.*

Definition 3 (Fenchel Dual (Borwein & Lewis, 2006))**.** *Let $\psi : \mathcal{I} \to \mathbb{R}$ be defined on $\mathcal{I} \subseteq \mathbb{R}$. Its Fenchel dual (a.k.a. convex conjugate) is defined as $\psi^*(v) := \sup_{u\in\mathcal{I}}\{uv - \psi(u)\}$.*

Definition 4. *(Generalized (Centered) CGF (Agrawal & Horel, 2021)) Let f be as defined above and introduce $\phi : u \to f(u+1)$. The generalized CGF of a random variable X with respect to f is defined as $\Lambda_{f;X}(t) := \inf_{\lambda\in\mathbb{R}} \mathbb{E}[\phi^*(tX + \lambda)]$. When conditioned on $Y = y$, the generalized CGF is $\Lambda_{f;X|Y=y}(t) := \inf_{\lambda\in\mathbb{R}} \mathbb{E}_{X|Y=y}[\phi^*(tX + \lambda)]$.*

Definition 5 (Generalized Inverse)**.** *For a non-decreasing function $\psi : [0,\infty) \to \mathbb{R}$, its generalized inverse function is defined by $\psi^{-1}(v) := \inf\{u \geq 0 : \psi(u) \geq v\}$.*

4 GENERALIZATION BOUNDS FOR STABLE ALGORITHMS

In this section, we quantify the *stability* of learning algorithms using a series of notions. These stability notions allow us to derive various information-theoretic bounds through a unified and general f-divergence framework.

4.1 REPLACE-ONE SETTING AND NEIGHBORING HYPOTHESES

We start by describing the replace-one setting and neighboring hypotheses that will be useful in our analyses. Let $Z_{[n]}, Z_i', W, W_i'$ be defined as in Section 1, their joint probability is induced by

$$P_{W,W_i',Z_{[n]},Z_i'} := P_{Z_{[n]},Z_i'} P_{W,W_i'|Z_{[n]},Z_i'}. \tag{7}$$

Since W_i' has the same marginal distribution as W and $W_i' \perp\!\!\!\perp Z_i$, the expected generalization error $\overline{\mathrm{gen}}$ can be rewritten as

$$\begin{aligned}
\overline{\mathrm{gen}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_W \mathbb{E}_{Z_i}\left[\ell(W, Z_i)\right] - \mathbb{E}_{W,Z_i}\left[\ell(W, Z_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W_i'} \mathbb{E}_{Z_i}\left[\ell(W_i', Z_i)\right] - \mathbb{E}_{W,Z_i}\left[\ell(W, Z_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W_i',W,Z_i}\left[\ell(W_i', Z_i) - \ell(W, Z_i)\right],
\end{aligned} \tag{8}$$

where all distributions are the marginals of $P_{W,W_i',Z_{[n]},Z_i'}$ in (7). (8) motivates the definition of the loss difference

$$\Delta(w, w_i', z) := \ell(w_i', z) - \ell(w, z), \tag{9}$$

which underlies the stability notions introduced later. This quantity exhibits a useful property:

$$\mathbb{E}_{W,W_i'}\mathbb{E}_Z[\Delta(W, W_i', Z)] = 0, \tag{10}$$

that is, the loss difference is centered when averaged over the product distribution $P_{W,W_i'}P_Z$.

4.2 Bridging Stability and Information-Theoretic Bounds

To start with, we first recall a stronger version of *uniform stability* from Wang & Mao (2023a).

Definition 6 (Strong Uniform Stability (Bousquet & Elisseeff, 2002; Wang & Mao, 2023a))**.** *We say that a learning algorithm $\mathcal{A}$ is β-*strongly-uniformly-stable *if, for all $i \in [n]$, $z_{[n]} \in \mathcal{Z}^n$, $z_i' \in \mathcal{Z}$ and $z \in \mathcal{Z}$,*

$$\sup_{r\in\mathcal{R}} \left|\ell(\mathcal{A}(z_{[n]\sim i}, r), z) - \ell(\mathcal{A}(z_{[n]}, r), z)\right| \le \beta. \tag{11}$$

Note that this notion coincides with (1) for deterministic learning algorithms. We now introduce below a weaker, distribution-dependent stability condition.

Definition 7 (Sub-Gaussian Stability)**.** *We say that a learning algorithm $\mathcal{A}$ is β-*sub-Gaussian-stable *if, for all $i \in [n]$,*

$$\log \mathbb{E}_{Z_{[n]},Z_i',Z,R}\left[e^{t\left(\ell(\mathcal{A}(Z_{[n]\sim i},R),Z)-\ell(\mathcal{A}(Z_{[n]},R),Z)\right)}\right] \le \frac{t^2\beta^2}{2}, \quad \forall t \in \mathbb{R}. \tag{12}$$

The *strong uniform stability* (Definition 6) implies the *sub-Gaussian stability* by Hoeffding's lemma (a straightforward proof is included in Appendix B).

Lemma 1. *If a learning algorithm is β-*strongly-uniformly-stable*, then it is β-*sub-Gaussian-stable*.*

By definition (9), the *sub-Gaussian stability* in Definition 7 can be rewritten in its equivalent form

$$\log \mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[e^{t\Delta(W,W_i',Z)}\right] \le \frac{t^2\beta^2}{2}, \quad \forall t \in \mathbb{R}. \tag{13}$$

The remaining stability notions will also be stated in terms of Δ. We provide our first bound for *sub-Gaussian-stable* learning algorithms. The proof uses Theorem 2 and is deferred to Appendix E.

Theorem 1. *If the learning algorithm is β-*sub-Gaussian-stable*, then*

$$\overline{\mathrm{gen}} \le \frac{\beta}{n}\sum_{i=1}^{n}\sqrt{2I(W; Z_i|W_i')}. \tag{14}$$

For *uniformly-stable* deterministic algorithms, the result in Theorem 1 improves strictly upon (2) whenever $I(W; Z_i|W_i') < \frac{1}{2}$ for all $i \in [n]$. While the proposed sub-Gaussian stability is less restrictive as compared to the *strong uniform stability*, it does not necessarily hold in general, as shown in Section 5. This motivates the goal to find weaker stability notions that are sufficient to derive good generalization bounds.

4.3 A General f-Divergence Framework

In this subsection, we present a general framework to analyze the expected generalization error in (8) using the f-divergence (Definition 1) and generalized CGF (Definition 4), providing a unified view and derivation of the stability-based bounds introduced in this section. We first prove a useful lemma, with details deferred to Appendix C.

Lemma 2. *Let $\tilde{W}$ and $\tilde{Z}_i$ be independent copies of W and Z_i, respectively, such that $P_{\tilde{W},W_i',\tilde{Z}_i} = Q_{W,W_i'}P_{Z_i}$, where $Q_{W,W_i'}$ is a prior distribution. Let $a \in (0, +\infty]$ be a constant. Then,*

$$\overline{\mathrm{gen}} \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W_i'}\left[\inf_{0<t<a}\left\{\frac{1}{t}\left(\Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)|W_i'}(t) + D_f(P_{W,Z_i|W_i'}\|Q_{W|W_i'}P_{Z_i})\right)\right\}\right] \tag{15}$$

$$\le \frac{1}{n}\sum_{i=1}^{n}\inf_{0<t<a}\left\{\frac{1}{t}\left(\Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)}(t) + D_f(P_{W,W_i',Z_i}\|Q_{W,W_i'}P_{Z_i})\right)\right\}. \tag{16}$$

The bounds in Lemma 2 are useful when the CGFs can be controlled. The next theorem replaces the CGF $\Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)}$ by a tractable upper bound and specializes the integrated bound (16) to $Q_{W,W_i'} = P_{W,W_i'}$. It may be of independent interest to explore whether other priors can yield sharper bounds or weaker assumptions. We consider this as an potential future research direction.

Theorem 2. *Let $\tilde{Z}_i$ be an independent copy of Z_i such that $P_{W,W_i',\tilde{Z}_i} = P_{W,W_i'}P_{Z_i}$ and let $a \in (0,+\infty]$ be a constant. Assume there exists a function $\psi_f : [0,a) \to [0,\infty)$ satisfying*

1. *ψ_f is convex and continuously differentiable;*
2. *$\psi_f(0) = \psi_f'(0) = 0$;*
3. *$\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}(t) \le \psi_f(t)$ for all $t \in (0,a)$.*

Then,

$$\overline{\text{gen}} \le \frac{1}{n}\sum_{i=1}^{n}(\psi_f^*)^{-1}\left(I_f(W;Z_i|W_i')\right). \tag{17}$$

The proof can be found in Appendix D. We now discuss some aspects of Theorem 2.

- The incorporation of the f-information allows for flexible choices of f, resulting in more results which are potentially tighter than Theorem 1. In particular, choosing $f(x) = x\log x$ for the f-information yields the mutual information, and meanwhile the generalized CGF can be specialized to the classical CGF (Agrawal & Horel, 2021). In this case, Theorem 1 can be derived as a corollary (see proof in Appendix E). We omit the subscript f throughout when referring to mutual information or KL-divergence.
- We remark that the bound in Theorem 2 remains intractable unless we have an expression of ψ_f (upper bound of the CGF) which is amenable for calculation. We show in the subsequent subsections that the proposed stability notions (e.g., Definitions 8 and 9) offer feasible ψ_f, and consequently the bound (17) can admit analytical expressions. Moreover, in Section 5, we further show some examples where these stability assumptions are satisfied and more feasible convex functions ψ_f do exist.

4.4 Bounds Based on Mutual Information

In this subsection, we limit Theorem 2 to $f(x) = x\log x$ and hence to mutual information. Now, we are in a position to derive more bounds from Theorem 2, based on mild assumptions on Δ.

Definition 8 (Sub-Exponential/Gamma Stability)**.** *We say that a learning algorithm $\mathcal{A}$ is (β,b)-sub-exponential-stable if*

$$\log \mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[e^{t\Delta(W,W_i',Z)}\right] \le \frac{t^2\beta^2}{2}, \quad \forall|t| < \frac{1}{b}. \tag{18}$$

We say that a learning algorithm $\mathcal{A}$ is (β,c)-sub-gamma-stable if

$$\log \mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[e^{t\Delta(W,W_i',Z)}\right] \le \frac{t^2\beta^2}{2(1-ct)}, \quad \forall 0 < t < \frac{1}{c}. \tag{19}$$

The following implications are immediate from the definitions.

Lemma 3. *If a learning algorithm is β-sub-Gaussian-stable, then it is (β,b)-sub-exponential-stable for all $b > 0$. If a learning algorithm is (β,b)-sub-exponential-stable, then it is (β,b)-sub-gamma-stable.*

By invoking the stability notions in Definition 8, Theorem 2 yields the following bounds. The proofs appear in Appendices F and G

Corollary 1. *If the learning algorithm is (β,b)-sub-exponential-stable, then*

$$\overline{\text{gen}} \le \frac{1}{n}\sum_{i=1}^{n}\min\left\{\beta\sqrt{2I(W;Z_i|W_i')}, bI(W;Z_i|W_i') + \frac{\beta^2}{2b}\right\}. \tag{20}$$

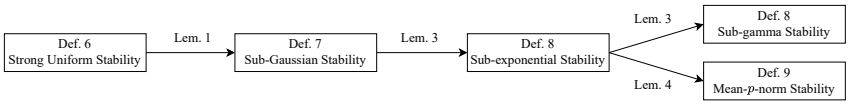

Figure 1: Implications between various stability notions.

Corollary 2. *If the learning algorithm is* (β, c)-sub-gamma-stable*, then*

$$\overline{\text{gen}} \leq \frac{1}{n}\sum_{i=1}^{n}\left(\beta\sqrt{2I(W;Z_i|W_i')} + cI(W;Z_i|W_i')\right). \tag{21}$$

In particular, Corollary 2 recovers Theorem 1 by setting $c = 0$, consistently with Lemma 3.

4.5 Bounds Based on f-Information

In this subsection, we discuss Theorem 2 under the other choices of f-divergence measures, resulting in new bounds that require weaker stability conditions. We define the following.

Definition 9 (Mean-p-Norm Stability)**.** *Let $p > 1$ be a constant. We say that a learning algorithm $\mathcal{A}$ is* β-mean-p-norm-stable *if, for all $i \in [n]$,*

$$\mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[|\Delta(W,W_i',Z)|^p\right] \leq \beta^p. \tag{22}$$

When $p = 2$, the proposed *mean-p-norm stability* essentially recovers the *mean-square stability* (Kale et al., 2011) and further extends it to randomized learning algorithms. Unlike the preceding stability notions which are required to hold for all possible t, the *mean-p-norm stability* imposes a less limiting condition on the p-th moment of Δ. Moreover, the following lemma establishes implications between stability notions (see proof in Appendix H).

Lemma 4. *For $p > 1$, if a learning algorithm is* (β, b)-sub-exponential-stable*, then it is* $\tilde{\beta}$-mean-p-norm-stable*, where*

$$\tilde{\beta} = \begin{cases} 2^{\frac{1}{p}}\sqrt{\frac{p}{e}}\beta, & \text{if } \beta \geq b\sqrt{p}, \\ 2^{\frac{1}{p}}\left(\frac{pb}{e}\right)e^{\frac{\beta^2}{2pb^2}}, & \text{if } \beta < b\sqrt{p}. \end{cases} \tag{23}$$

Moreover, for $p \in (1, 2]$ and every $b > 0$, if a learning algorithm is (β, b)-sub-exponential-stable*, then it is* β-mean-p-norm-stable*.*

Remark 1. *We remark that the Bernstein condition, which has been explored in a vast study of fast-rate PAC-Bayes bounds (Klochkov & Zhivotovskiy, 2021; Wu et al., 2025), implies the* mean-p-norm stability *in the case $p \in (1, 2]$, demonstrating that the latter stability with a small p can be a weaker notion. Detailed discussions are provided in Appendix K.*

Together, Lemmas 1, 3, and 4 give a hierarchy of stability notions, as depicted in Fig. 1. By using the χ^α-divergence[1] and the α-divergence[2] (see following for their generator f) for Theorem 2, we obtain the following bound. Its proof is given in Appendix I.

Corollary 3. *Define $\alpha := \frac{p}{p-1} > 1$. If the learning algorithm is* β-mean-p-norm-stable*, then*

$$\overline{\text{gen}} \leq \frac{\beta}{n}\sum_{i=1}^{n}\min\left\{I_{\chi^\alpha}(W;Z_i|W_i')^{\frac{1}{\alpha}}, (1+\alpha(\alpha-1)I_\alpha(W;Z_i|W_i'))^{\frac{1}{\alpha}}\right\}, \tag{24}$$

where I_{χ^α} specializes I_f with $f(x) = |x-1|^\alpha$, and I_α specializes I_f with $f(x) = \frac{x^\alpha - \alpha x + \alpha - 1}{\alpha(\alpha-1)}$.

In essence, the *strong uniform stability* also permits a common bound which allows to incorporate various f-information freely. For the f functions listed in Table 1, a quadratic upper bound ψ_f on the generalized CGF follows from Agrawal & Horel (2021, Corollary 92). A direct substitution into Theorem 2 yields the following corollary, as proved in Appendix J.

[1]The χ^α-divergence recovers the total variation for $\alpha = 1$, and the χ^2-divergence for $\alpha = 2$.

[2]Up to constants, the α-divergence recovers the H^2-divergence for $\alpha = \frac{1}{2}$, and the χ^2-divergence for $\alpha = 2$.

Corollary 4. *If the learning algorithm is* β-strongly-uniformly-stable*, then for all* f*-divergences listed in Table 1,*

$$\overline{\mathrm{gen}} \leq \frac{c_f \beta}{n} \sum_{i=1}^{n} \sqrt{2 I_f(W; Z_i | W_i')}, \tag{25}$$

where c_f *is given in Table 1.*

Table 2 concludes this section by providing an overview of each stability notion and the corresponding bound derived from Theorem 2.

Table 1: Value of c_f for various f-information in Corollary 4

$f(x)$	f-Divergence	c_f
$x \log x$	KL	1
$(x-1)^2$	χ^2	$\frac{\sqrt{2}}{2}$
$(\sqrt{x}-1)^2$	H^2	$\sqrt{2}$
$\frac{(x-1)^2}{x+1}$	Le Cam	1
$\frac{x^\alpha - \alpha x + \alpha - 1}{\alpha(\alpha-1)}$	$\alpha^\dagger$	1

† For $\alpha \in [-1, 2]$.

Table 2: Overview of the proposed bounds

	KL	$\chi^{\alpha\,\dagger}$	α	χ^2/H^2/Le Cam
Strong uniform	Cor. 4	–	Cor. 4‡	Cor. 4
Sub-Gaussian	Thm. 1	–	–	–
Sub-exponential	Cor. 1	–	–	–
Sub-gamma	Cor. 2	–	–	–
Mean-p-norm	–	Cor. 3	Cor. 3§	–

‡ For $\alpha \in [-1, 2]$.
§ For $\alpha \in (1, +\infty)$.

5 Usefulness of the Proposed Bounds

5.1 Estimating the Gaussian Mean

We first consider the problem of Gaussian mean estimation, a typical example for evaluating information-theoretic generalization bounds (Bu et al., 2020; Zhou et al., 2022; Wu et al., 2025).

Example 1 (Gaussian Mean Estimation)**.** *Suppose that* $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ *with unknown* $\mu \in \mathbb{R}$*. Using a quadratic loss function* $\ell(w, z) = (w-z)^2$*, the ERM algorithm will output* $W = \frac{1}{n}\sum_{i=1}^n Z_i$ *as an estimate of* μ*. The true generalization error is* $\overline{\mathrm{gen}} = \frac{2\sigma^2}{n}$ *(Bu et al., 2020).*

We show that several of the proposed stability conditions in this paper hold with parameters of order $\Theta(\frac{1}{n})$, and the corresponding bounds recover the exact generalization error up to constants. We summarize the results below for $n \geq 2$, with detailed calculations given in Appendix L.

First, the ERM algorithm is both (β, b)-*sub-exponential-stable* and (β, b)-*sub-gamma-stable*, with $\beta = b \in \Theta(\frac{1}{n})$. Moreover, the adopted mutual information can be exactly calculated as $I(W; Z_i | W_i') = \frac{1}{2}\log\frac{2n-1}{n-1}$. With this, Corollaries 1 and 2 therefore give bounds of order $\Theta(\frac{1}{n})$.

The algorithm is also β-*mean-2-norm-stable*, with $\beta \in \Theta(\frac{1}{n})$. In addition, using $I_{\chi^2}(W; Z_i | W_i') = \frac{n}{n-1}$, Corollary 3 with $p = 2$ recovers the optimal rate again.

We also remark that a direct application of Theorem 2 can give a tighter constant, as justified in Appendix L.3. Besides, the disintegrated bound (15), although unexplored in this paper, may yield a further improvement, which we leave for future work. For Example 1, Appendix L.4 also gives a lower bound of order $\Theta(\frac{1}{n})$, matching the rate of those upper bounds.

For comparison, the previous stability bounds, including (2) and those from Wang & Mao (2023a), are not applicable in this problem since their stability parameters are unbounded. As for the existing information-theoretic bounds, the chaining bound of Zhou et al. (2023) recovers the optimal rate $O(\frac{1}{n})$ using a carefully constructed auxiliary random process. The approach of Sefidgaran et al. (2022) also yields the optimal rate but similarly requires an auxiliary hypothesis satisfying a distortion constraint. Our bounds are easier to evaluate since, by construction, the auxiliary hypothesis W_i' can be naturally obtained by applying the original algorithm to a replace-one training sample. Additionally, the proposed bound gives $\overline{\mathrm{gen}} \leq \frac{2.77\sigma^2}{n} + o(\frac{1}{n})$ (see (93)), which is tighter than the chaining bound ($\overline{\mathrm{gen}} \leq \frac{13\sigma^2}{n}$) of Zhou et al. (2023) and fast-rate bound ($\overline{\mathrm{gen}} \leq \frac{2.82\sigma^2}{n} + o(\frac{1}{n})$) of Wu et al. (2025) by constants.

5.2 Estimating the Bernoulli Mean

We now explore another example in which the existing stability bound (2) is applicable and hence comparable to the proposed bounds. All calculations are provided in Appendix M.

Example 2 (Bernoulli Mean Estimation)**.** *Suppose that* $Z_i \overset{i.i.d.}{\sim} \mathrm{Bern}(\theta)$ *with unknown* $\theta \in (0,1)$*. Consider a quadratic loss function* $\ell(w,z) = (w-z)^2$*. The ERM algorithm will output* $W = \frac{1}{n}\sum_{i=1}^n Z_i$ *and the true generalization error is* $\overline{\mathrm{gen}} = \frac{2\theta(1-\theta)}{n}$ *(Lu et al., 2026).*

In this example, the learning algorithm is $\frac{2n-1}{n^2}$*-strongly-uniformly-stable*, giving $\overline{\mathrm{gen}} \le \frac{2n-1}{n^2}$ by (2). Additionally, the learning algorithm is also β*-mean-2-norm-stable* with $\beta \in \Theta(\frac{1}{n})$, and Corollary 3 yields $\overline{\mathrm{gen}} \le \frac{2.83\theta(1-\theta)}{n} + o(\frac{1}{n})$ which is tighter than (2) for every $\theta \in (0,1)$ and sufficiently large n. Figure 2 shows that all the bounds established in Corollary 4 decay with the optimal rate $O(\frac{1}{n})$ and differ by a constant. Besides, the proposed KL, χ^2, H^2 and α-divergence bounds are tighter than (2).

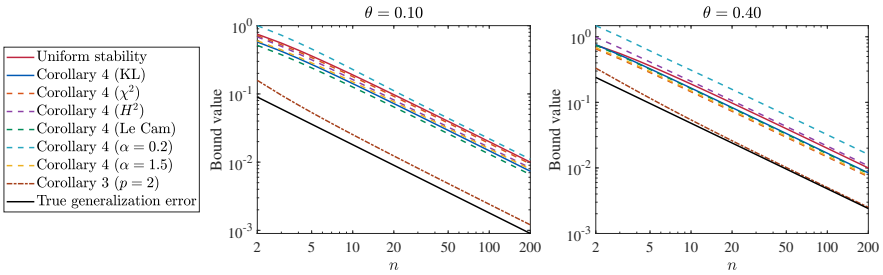

Figure 2: Comparison of the uniform stability bound (2), Corollary 3 ($p = 2$) and Corollary 4 bounds for KL, χ^2, H^2, Le Cam, and α-divergences ($\alpha = 0.2, 1.5$).

5.3 Linear Regression

Example 3 (Linear Regression[3])**.** *Let the data space be* $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$*, where* $\mathcal{X} \subseteq \mathbb{R}$ *and* $\mathcal{Y} \subseteq \mathbb{R}$*. Let* w^* *be the underlying (but unknown) hypothesis, by which we draw* n *i.i.d. observations as* $Y_i = w^* X_i + N_i$*, where* $N_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$*. Using* $\ell(w,z) = (y - wx)^2$*, the ERM algorithm outputs* $W = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = w^* + \frac{\sum_{i=1}^n X_i N_i}{\sum_{i=1}^n X_i^2}$*. For simplicity, we assume that* x_i *is a constant independent of* n *and* $x_i \neq 0$ *for every* $i \in [n]$*. The exact generalization error is* $\overline{\mathrm{gen}} = \frac{2\sigma^2}{n}$ *(Wu et al., 2025).*

We now evaluate Corollary 3 with $p = 2$, and Theorem 2 with $f(x) = x\log x$. The details are deferred to Appendix N. In particular, it is verified that the learning algorithm is β*-mean-2-norm-stable*, with $\beta \in \Theta(\frac{1}{n})$, and therefore Corollary 3 achieves the optimal rate $\Theta(\frac{1}{n})$. Furthermore, Theorem 2 with $f(x) = x\log x$ can give a slightly tighter bound $\overline{\mathrm{gen}} \le \frac{3.40\sigma^2}{n} + o(\frac{1}{n})$, which improves the fast-rate bound $\overline{\mathrm{gen}} \le \frac{4\sigma^2}{n} + o(\frac{1}{n})$ of Wu et al. (2025).

6 Limitations and Future Works

Our bounds share a common limitation of information-theoretic bounds: they can be difficult to evaluate in practice because the information measures depend on the unknown data distribution. One way to address this limitation is to further upper bound these measures using properties of the data and the learning algorithm, as explored by Negrea et al. (2019); Steinke & Zakynthinou (2020). Analyzing the generalization of SGD within our framework may also be of independent interest, although separately quantifying the stability parameters and information measures could be challenging. In such cases, Lemma 2 may be useful with a suitable choice of prior $Q_{W,W_i'}$. We leave these directions for future work.

[3]Example 1 with $\mu = 0$ can be viewed as a special case of Example 3, by letting $x_i = 1$ for all $i \in [n]$.

AI USE STATEMENT

Generative AI tools were used for calculating the considered bounds under Example 2 in Section 5.2. Additionally, generative AI tools were used to engineer Example 4 for the comparison with the prior literature. We have not used generative AI tools for development of all theoretical proofs in the paper. We have carefully checked all AI-assisted works, including the calculations in Appendix M and the validity of Example 4. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 22(128):1–59, 2021.

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)*, pp. 682–687. IEEE, 2021a.

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)*, pp. 1–5. IEEE, 2021b.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of machine learning research*, 23(39):1–70, 2022.

Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization: Theoryand Examples*. Springer, 2006.

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. *Journal of machine learning research*, 2(Mar):499–526, 2002.

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.

Eugenio Clerico, Amitis Shidani, George Deligiannidis, and Arnaud Doucet. Chained generalisation bounds. In *Conference on Learning Theory*, pp. 4212–4257. PMLR, 2022.

Yuxin Dong, Haoran Guo, Tieliang Gong, Wen Wen, and Chen Li. Exactly tight information-theoretic generalization bounds via binary jensen-shannon divergence. In *Forty-second International Conference on Machine Learning*, 2025.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated cmi. *Advances in Neural Information Processing Systems*, 35:10108–10121, 2022.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Proceedings of the 2nd Symposium on Innovations in Computer Science (ICS)*, pp. 182–195, 2011.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pp. 152–162, 1997.

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

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 275–282, 2002.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019.

Wenliang Liu, Guanding Yu, Lele Wang, and Renjie Liao. An information-theoretic framework for out-of-distribution generalization with applications to stochastic gradient langevin dynamics. *IEEE Transactions on Information Theory*, 2025.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Yang Lu, Margreta Kuijper, and Jingge Zhu. Towards unified and sharpened cmi bounds for generalization errors. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2025. doi: 10.1109/ISIT63088.2025.11195634.

Yang Lu, Matthias Frey, Margreta Kuijper, and Jingge Zhu. On unified and sharpened cmi bounds for generalization errors. *arXiv preprint arXiv:2605.21056*, 2026.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

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.

Leland Gerson Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

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)*, pp. 26–30. IEEE, 2016.

Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, and Stefano Soatto. On leave-one-out conditional mutual information for generalization. *Advances in Neural Information Processing Systems*, 35:10179–10190, 2022.

Omar Rivasplata, Emilio Parrado-Hernández, John S Shawe-Taylor, Shiliang Sun, and Csaba Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*, 31, 2018.

Borja Rodríguez Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via wasserstein distance. *Advances in Neural Information Processing Systems*, 34:19109–19121, 2021.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240. PMLR, 2016.

Milad Sefidgaran, Amin Gohari, Gaël Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Conference on Learning Theory*, pp. 4416–4463. PMLR, 2022.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Ziqiao Wang and Yongyi Mao. Sample-conditioned hypothesis stability sharpens information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 36:49513–49541, 2023a.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In *International Conference on Machine Learning*, pp. 36111–36137. PMLR, 2023b.

Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. Fast rate information-theoretic bounds on generalization errors. *IEEE Transactions on Information Theory*, 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. *IEEE Transactions on Information Theory*, 68(5):3304–3316, 2022.

Ruida Zhou, Chao Tian, and Tie Liu. Stochastic chaining and strengthened information-theoretic generalization bounds. *Journal of the Franklin Institute*, 360(6):4114–4134, 2023.

A Prerequisite Definitions and Lemmas

Lemma 5. *(Variational Representation of f-Divergence (Birrell et al., 2022)) Let f be as defined above and $\mathcal{G} := \mathcal{M}(\mathcal{X})$ be the set of all measurable functions on $\mathcal{X}$. Then,*

$$D_f(P\|Q) = \sup_{g\in\mathcal{G}} \left\{ \mathbb{E}_{X\sim P}\left[g(X)\right] - \inf_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}_{X\sim Q}\left[f^*(g(X)+\lambda)\right] - \lambda \right\} \right\}, \tag{26}$$

where the equality is attained when $g = f'(\frac{dP}{dQ})$.

Lemma 6. *(Boucheron et al. (2013, Lemma 2.4)) Let $\psi : [0, a) \to \mathbb{R}$ be convex and continuously differentiable, where $a \in (0, +\infty]$. Assume that $\psi(0) = \psi'(0) = 0$, then the Fenchel dual ψ^* is non-negative, convex and non-decreasing on $[0, +\infty)$. Moreover, the generalized inverse function $(\psi^*)^{-1}$ can be written as*

$$(\psi^*)^{-1}(v) = \inf_{t\in(0,a)} \frac{v + \psi(t)}{t}. \tag{27}$$

which is concave and non-decreasing.

Lemma 7. *(Borwein & Lewis (2006, Table 3.2)) Let ψ be as defined in Definition 3 and define $\phi(u) := \psi(u + c)$, where $c \in \mathbb{R}$, then $\phi^*(v) = \psi^*(v) - cv$.*

Remark 2. *Recall that $\phi(u) = f(u + 1)$ (see Definition 4), Lemma 7 implies that the generalized CGF $\Lambda_{f;X}$ can be written as*

$$\Lambda_{f;X}(t) = \inf_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}\left[f^*(tX+\lambda) - tX\right] - \lambda \right\}. \tag{28}$$

Similarly, when conditioned on $Y = y$,

$$\Lambda_{f;X|Y=y}(t) = \inf_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}_{X|Y=y}\left[f^*(tX+\lambda) - tX\right] - \lambda \right\}. \tag{29}$$

Lemma 8. *Let $\Lambda_{f;X}$ and $\Lambda_{f;X|Y}$ be defined in Definition 4, then for every $t \in \mathbb{R}$, we have $\mathbb{E}_Y[\Lambda_{f;X|Y}(t)] \le \Lambda_{f;X}(t)$.*

Proof.

$$\begin{aligned} \mathbb{E}_Y\left[\Lambda_{f;X|Y}(t)\right] &= \mathbb{E}_Y\left[\inf_{\lambda\in\mathbb{R}} \mathbb{E}_{X|Y}\left[\phi^*(tX+\lambda)\right]\right] \\ &\le \inf_{\lambda\in\mathbb{R}} \mathbb{E}_Y \mathbb{E}_{X|Y}\left[\phi^*(tX+\lambda)\right] \\ &= \Lambda_{f;X}(t). \end{aligned} \tag{30}$$

□

B Proof of Lemma 1

It holds by Definition 6 that $|\Delta(W, W_i', Z)| \le \beta$ almost surely, an application of Hoeffding's lemma yields

$$\log \mathbb{E}\left[e^{t\left(\Delta(W,W_i',Z) - \mathbb{E}\left[\Delta(W,W_i',Z)\right]\right)}\right] \le \frac{\beta^2 t^2}{2}, \quad \forall t \in \mathbb{R}. \tag{31}$$

Combining this with (10) recovers Definition 7.

C Proof of Lemma 2

Let us condition on $W_i' = w_i'$. By using the variational representation of f-divergence in Lemma 5, for any $t \in \mathbb{R}$, we have

$$\begin{aligned} &D_f(P_{W,Z_i|w_i'}\|Q_{W|w_i'}P_{Z_i}) \\ &= D_f(P_{W,Z_i|w_i'}\|P_{\tilde{W},\tilde{Z}_i|w_i'}) \\ &\ge \mathbb{E}_{W,Z_i|w_i'}\left[t\Delta(W, w_i', Z_i)\right] - \inf_{\lambda\in\mathbb{R}} \left\{ \mathbb{E}_{\tilde{W},\tilde{Z}_i|w_i'}\left[f^*(t\Delta(\tilde{W}, w_i', \tilde{Z}_i) + \lambda)\right] - \lambda \right\} \\ &\overset{(a)}{=} \mathbb{E}_{W,Z_i|w_i'}\left[t\Delta(W, w_i', Z_i)\right] - \Lambda_{f;\Delta(\tilde{W},w_i',\tilde{Z}_i)}(t) - \mathbb{E}_{\tilde{W},\tilde{Z}_i|w_i'}\left[t\Delta(\tilde{W}, w_i', \tilde{Z}_i)\right], \end{aligned} \tag{32}$$

where (a) uses Remark 2. Since this inequality holds for all $t \in \mathbb{R}$, taking the infimum over $t \in (0, a)$ preserves the inequality. After rearranging terms,

$$\begin{aligned}&\mathbb{E}_{W,Z_i|w_i'}\left[\Delta(W, w_i', Z_i)\right] - \mathbb{E}_{\tilde{W},\tilde{Z}_i|w_i'}\left[\Delta(\tilde{W}, w_i', \tilde{Z}_i)\right] \\ &\leq \inf_{0<t<a}\left\{\frac{1}{t}\left(D_f(P_{W,Z_i|w_i'}\|Q_{W|w_i'}P_{Z_i}) + \Lambda_{f;\Delta(\tilde{W},w_i',\tilde{Z}_i)}(t)\right)\right\},\end{aligned} \tag{33}$$

which is measurable as justified at the end of this proof. Moreover, since $\tilde{W}$ and W_i' have the same marginal distribution, following a similar argument in (10), we can see that $\Delta(\tilde{W}, W_i', \tilde{Z}_i)$ is also centered when taken expectation, i.e.,

$$\mathbb{E}_{\tilde{W},W_i',\tilde{Z}_i}\left[\Delta(\tilde{W}, W_i', \tilde{Z}_i)\right] = 0. \tag{34}$$

Then, it follows from (8) that

$$\begin{aligned}\overline{\text{gen}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell(W_i', Z_i) - \ell(W, Z_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W_i'}\left[\mathbb{E}_{W,Z_i|W_i'}\left[\Delta(W, W_i', Z_i)\right]\right] \\ &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W_i'}\left[\inf_{0<t<a}\left\{\frac{1}{t}\left(D_f(P_{W,Z_i|W_i'}\|Q_{W|W_i'}P_{Z_i}) + \Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)|W_i'}(t)\right)\right\}\right] \\ &\quad + \mathbb{E}_{W_i'}\left[\mathbb{E}_{\tilde{W},\tilde{Z}_i|W_i'}\left[\Delta(\tilde{W}, W_i', \tilde{Z}_i)\right]\right] \\ &\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W_i'}\left[\inf_{0<t<a}\left\{\frac{1}{t}\left(D_f(P_{W,Z_i|W_i'}\|Q_{W|W_i'}P_{Z_i}) + \Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)|W_i'}(t)\right)\right\}\right],\end{aligned} \tag{35}$$

where (a) is by (33) and (b) uses (34). This proves (15). By exchanging the order of expectation and infimum in (35), we obtain

$$\begin{aligned}\overline{\text{gen}} &\leq \frac{1}{n}\sum_{i=1}^{n}\inf_{0<t<a}\left\{\frac{1}{t}\left(D_f(P_{W,W_i',Z_i}\|Q_{W,W_i'}P_{Z_i}) + \mathbb{E}_{W_i'}\left[\Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)|W_i'}(t)\right]\right)\right\} \\ &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\inf_{0<t<a}\left\{\frac{1}{t}\left(D_f(P_{W,W_i',Z_i}\|Q_{W,W_i'}P_{Z_i}) + \Lambda_{f;\Delta(\tilde{W},W_i',\tilde{Z}_i)}(t)\right)\right\}.\end{aligned} \tag{36}$$

where (a) is due to Lemma 8. This completes the proof of (16).

Measurability of (33)**.** For brevity, let abbreviate $\Lambda_{\Delta(\tilde{W},W_i',\tilde{Z}_i)|W_i'=w}(t)$ as $K_w(t)$, and $D_f(P_{W,Z_i|W_i'=w}\|Q_{W|W_i'=w}P_{Z_i})$ as $D(w)$, respectively. We further let $M_w(t) = \frac{1}{t}(K_w(t) + D(w))$ and define $Q_a := \mathbb{Q} \cap (0, a)$ as the set comprising all rational numbers within the interval $(0, a)$. We are going to prove $\inf_{t\in(0,a)} M_w(t) = \inf_{t\in Q_a} M_w(t)$. Since $\inf_{t\in(0,a)} M_w(t) \leq \inf_{t\in Q_a} M_w(t)$ is immediate from the fact $Q_a \subset (0, a)$, it suffices to prove the reverse inequality, i.e.,

$$\inf_{t\in Q_a} M_w(t) \leq \inf_{t\in(0,a)} M_w(t). \tag{37}$$

We consider a fixed $t \in (0, a)$. By choosing a sequence of rational numbers q_i such that, for every i, $q_i \in Q_a \cap (0, t)$ and $\lim_{i\to+\infty} q_i = t$, convexity of the CGF $K_w(t)$ (see Agrawal & Horel (2021, Proposition 39)) implies that

$$\begin{aligned}K_w(q_i) &\leq \frac{q_i}{t}K_w(t) + \left(1 - \frac{q_i}{t}\right)K_w(0) \\ &= \frac{q_i}{t}K_w(t),\end{aligned} \tag{38}$$

where for the equation we use the fact $K_w(0) = 0$ (also introduced in Agrawal & Horel (2021, Proposition 39)). Then, it follows that

$$\begin{aligned}\limsup_{i\to+\infty} M_w(q_i) &= \limsup_{i\to+\infty}\left\{\frac{K_w(q_i)}{q_i} + \frac{D(w)}{q_i}\right\}\\ &\le \frac{K_w(t)}{t} + \limsup_{i\to+\infty}\frac{D(w)}{q_i}\\ &= \frac{K_w(t)}{t} + \frac{D(w)}{t}\\ &= M_w(t).\end{aligned} \tag{39}$$

Since $q_i \in Q_a$, thus for every $t \in (0, a)$,

$$\inf_{t'\in Q_a} M_w(t') \le \limsup_{i\to+\infty} M_w(q_i) \le M_w(t). \tag{40}$$

This inequality preserves when taking infimum over $t \in (0, a)$ on the RHS, leading to (37). The LHS of (37) is a countable infimum of measurable functions, thus proving that (33) is measurable.

D Proof of Theorem 2

Setting $Q_{W,W_i'} = P_{W,W_i'}$, (16) is specialized to

$$\overline{\mathrm{gen}} \le \frac{1}{n}\sum_{i=1}^{n}\inf_{0<t<a}\left\{\frac{1}{t}\left(I_f(W;Z_i|W_i') + \Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}(t)\right)\right\}, \tag{41}$$

where $\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}$ is w.r.t. $P_{W,W_i',\tilde{Z}_i} = P_{W,W_i'}P_{Z_i}$. For each $i \in [n]$ and $t \in (0, a)$, we use $\psi(t)$ to upper bound $\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}(t)$ and obtain

$$\begin{aligned}\overline{\mathrm{gen}} &\le \frac{1}{n}\sum_{i=1}^{n}\inf_{0<t<a}\left\{\frac{1}{t}\left(I_f(W;Z_i|W_i') + \psi(t)\right)\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}(\psi_f^*)^{-1}\left(I_f(W;Z_i|W_i')\right),\end{aligned} \tag{42}$$

where the last step is by Lemma 6.

E Proof of Theorem 1

We prove Theorem 1 as a corollary of Theorem 2. By setting $f(x) = x\log x - (x-1)$ for (17), the f-information $I_f(W, Z_i|W_i')$ recovers the mutual information $I(W, Z_i|W_i')$ and the generalized CGF $\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}$ recovers the classical CGF (Agrawal & Horel, 2021) (one can also obtain it through a direct calculation by Definition 4), i.e.,

$$\begin{aligned}\Lambda_{\Delta(W,W_i',\tilde{Z}_i)} &= \log\mathbb{E}\left[e^{t\left(\Delta(W,W_i',\tilde{Z}_i) - \mathbb{E}\left[\Delta(W,W_i',\tilde{Z}_i)\right]\right)}\right]\\ &= \log\mathbb{E}\left[e^{t\Delta(W,W_i',\tilde{Z}_i)}\right],\end{aligned} \tag{43}$$

where we drop subscript f and the last equation is by (10). According to Definition 7, we choose $\psi(t) = \frac{t^2\beta^2}{2}$ and then obtain $\psi^*(v) = \frac{v^2}{2\beta^2}$ by Definition 3. It therefore follows by Definition 5 that $(\psi^*)^{-1}(v) = \sqrt{2\beta^2 v}$. Thus, (17) recovers the bound (14).

F Proof of Corollary 1

By (18) in Definition 8, we choose $\psi(t) = \frac{\beta^2 t^2}{2}$ with $t \in (0, \frac{1}{b})$, then $(\psi^*)^{-1}(v) = \inf_{t\in(0,\frac{1}{b})}\left\{\frac{v}{t} + \frac{\beta^2 t}{2}\right\}$. The infimum is attained at $t = \frac{1}{\beta}\sqrt{2v}$ when $v < \frac{\beta^2}{2b^2}$, and otherwise at

$t = \frac{1}{b}$ due to the decreasing monotonicity within $t \in (0, \frac{1}{b})$. Thus, a direct calculation yields

$$(\psi^*)^{-1}(v) = \begin{cases} \beta\sqrt{2v}, & \text{if } v < \frac{\beta^2}{2b^2}, \\ \frac{\beta^2}{2b} + bv, & \text{otherwise.} \end{cases} \tag{44}$$

Applying (17) of Theorem 2 leads to the desired result.

G Proof of Corollary 2

By (19) in Definition 8, we choose $\psi(t) = \frac{\beta^2 t^2}{2(1-ct)}$ with $t \in (0, \frac{1}{c})$, then

$$\begin{aligned} (\psi^*)^{-1}(v) &= \inf_{t\in(0,\frac{1}{c})} \left\{ \frac{v}{t} + \frac{\beta^2 t}{2(1-ct)} \right\} \\ &\overset{(a)}{=} \inf_{\lambda>0} \left\{ v(\lambda + c) + \frac{\beta^2}{2\lambda} \right\} \\ &\overset{(b)}{=} \sqrt{2v\beta^2} + vc, \end{aligned} \tag{45}$$

where in (a) we substitute $t = \frac{1}{\lambda+c}$ and in (b) the infimum is attained at $\lambda = \beta\sqrt{\frac{1}{2v}}$. Applying (17) of Theorem 2 leads to the desired result.

H Proof of Lemma 4

The proof is adapted from Vershynin (2019, Proposition 2.7.1). One can verify the following inequality holds for all $x > 0$ and $p > 1$:

$$x^p \le \left(\frac{p}{e}\right)^p e^x. \tag{46}$$

Define $\cosh(x) := \frac{e^x + e^{-x}}{2}$, then it follows that

$$|x|^p \le 2\left(\frac{p}{e}\right)^p \cosh(x). \tag{47}$$

Substituting $x = t\Delta(W, W_i', Z)$ and taking expectations, we get

$$\begin{aligned} \mathbb{E}\left[|t\Delta(W, W_i', Z)|^p\right] &\le 2\left(\frac{p}{e}\right)^p \mathbb{E}\left[\cosh(t\Delta(W, W_i', Z))\right] \\ &= \left(\frac{p}{e}\right)^p \left(\mathbb{E}\left[e^{t\Delta(W,W_i',Z)}\right] + \mathbb{E}\left[e^{-t\Delta(W,W_i',Z)}\right]\right) \\ &\overset{(a)}{\le} 2\left(\frac{p}{e}\right)^p e^{\frac{1}{2}\beta^2 t^2}, \quad \forall |t| < \frac{1}{b} \end{aligned} \tag{48}$$

where (a) is by (18) in Definition 8. Minimizing for t gives

$$\mathbb{E}\left[|\Delta(W, W_i', Z)|^p\right] \le \begin{cases} 2\left(\frac{p}{e}\right)^{\frac{p}{2}} \beta^p, & \text{if } \beta \ge b\sqrt{p}, \\ 2\left(\frac{pb}{e}\right)^p e^{\frac{\beta^2}{2b^2}}, & \text{if } \beta < b\sqrt{p}. \end{cases} \tag{49}$$

which proves the result.

A sharper bound for $p \in (1, 2]$. Consider $p \in (1, 2]$, we have

$$\begin{aligned} \mathbb{E}\left[|\Delta(W, W_i', Z)|^p\right] &= \mathbb{E}\left[\left((\Delta(W, W_i', Z))^2\right)^{\frac{p}{2}}\right] \\ &\le \left(\mathbb{E}\left[(\Delta(W, W_i', Z))^2\right]\right)^{\frac{p}{2}}, \end{aligned} \tag{50}$$

where the inequality is an application of Jensen's inequality on the concave function $x^{\frac{p}{2}}$ for $x \ge 0$. Next, we will derive an upper bound for $\mathbb{E}[(\Delta(W, W_i', Z))^2]$.

By (18) in Definition 8, for all $|t| < \frac{1}{b}$ we have

$$\mathbb{E}\left[\cosh(t\Delta(W, W_i', Z))\right] \leq e^{\frac{1}{2}\beta^2 t^2}. \tag{51}$$

Using the inequality $\cosh(x) \geq 1 + \frac{x^2}{2}$, it follows that

$$\mathbb{E}\left[\cosh(t\Delta(W, W_i', Z))\right] \geq 1 + \frac{t^2}{2}\mathbb{E}\left[\left(\Delta(W, W_i', Z)\right)^2\right]. \tag{52}$$

By combining the two inequalities above and rearranging terms, we get

$$\mathbb{E}\left[\left(\Delta(W, W_i', Z)\right)^2\right] \leq \frac{2}{t^2}\left(e^{\frac{1}{2}\beta^2 t^2} - 1\right), \quad \forall |t| < \frac{1}{b}. \tag{53}$$

Since this inequality holds for all $|t| < \frac{1}{b}$, taking the limit $t \to 0$ preserves the inequality, i.e.,

$$\begin{aligned}\mathbb{E}\left[\left(\Delta(W, W_i', Z)\right)^2\right] &\leq \lim_{t\to 0} \frac{2}{t^2}\left(e^{\frac{1}{2}\beta^2 t^2} - 1\right) \\ &= \beta^2,\end{aligned} \tag{54}$$

which holds for any $b > 0$. Plugging this bound into (50) yields

$$\mathbb{E}\left[\left(\Delta(W, W_i', Z)\right)^p\right] \leq \beta^p, \quad \forall p \in (1, 2], \tag{55}$$

which therefore completes the proof.

I Proof of Corollary 3

Case of χ^α-divergence. We let $f(x) = |x-1|^\alpha$, where $x > 0$ and $\alpha > 1$. Therefore, we have $\phi(u) = |u|^\alpha$ for $u > -1$ by Definition 4. Through direct calculations, we can solve the supremum in Definition 3 and obtain

$$\phi^*(v) = \begin{cases}(\alpha-1)\alpha^{-p}|v|^p, & v > -\alpha, \\ -v - 1, & v \leq -\alpha.\end{cases} \tag{56}$$

One can verify that $-v - 1 \leq (\alpha-1)\alpha^{-p}|v|^p$ for $v \leq -\alpha$, and thus $\phi^*(v) \leq (\alpha-1)\alpha^{-p}|v|^p$ for all $v \in \mathbb{R}$. With this, $\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}$ is upper bounded by

$$\begin{aligned}\Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}(t) &= \inf_{\lambda\in\mathbb{R}} \mathbb{E}\left[\phi^*(t\Delta(W, W_i', \tilde{Z}_i) + \lambda)\right] \\ &\overset{(a)}{\leq} (\alpha-1)\alpha^{-p} \inf_{\lambda\in\mathbb{R}} \mathbb{E}\left[\left|t\Delta(W, W_i', \tilde{Z}_i) + \lambda\right|^p\right] \\ &\overset{(b)}{\leq} (\alpha-1)\alpha^{-p}\mathbb{E}\left[\left|\Delta(W, W_i', \tilde{Z}_i)\right|^p\right]|t|^p \\ &\overset{(c)}{\leq} (\alpha-1)\alpha^{-p}\beta^p|t|^p\end{aligned} \tag{57}$$

where (a) uses $\phi^*(v) \leq (\alpha-1)\alpha^{-p}|v|^p$ for all $v \in \mathbb{R}$, (b) is obtained by setting $\lambda = 0$, and (c) uses Definition 9. By choosing $\psi_f(t) = (\alpha-1)\alpha^{-p}\beta^p t^p$ with $t > 0$, it follows from Lemma 6 that

$$\begin{aligned}(\psi^*)^{-1}(v) &= \inf_{t>0}\left\{\frac{v}{t} + (\alpha-1)\alpha^{-p}\beta^p t^{p-1}\right\} \\ &= \beta v^{\frac{1}{\alpha}},\end{aligned} \tag{58}$$

where the infimum is attained at $t = \frac{\alpha}{\beta}v^{\frac{1}{p}}$. Substitution into (17) yields

$$\overline{\text{gen}} \leq \frac{\beta}{n}\sum_{i=1}^n I_{\chi^\alpha}(W; Z_i|W_i')^{\frac{1}{\alpha}}. \tag{59}$$

Case of α-divergence. We let $f(x) = \frac{x^\alpha - \alpha x + \alpha - 1}{\alpha(\alpha-1)}$, where $x > 0$ and $\alpha > 1$. It follows that $\phi(u) = \frac{(u+1)^\alpha - \alpha u - 1}{\alpha(\alpha-1)}$ for $u > -1$ by Definition 4. Further calculations gives

$$\phi^*(v) = \begin{cases} \frac{((\alpha-1)v+1)^p - 1}{\alpha} - v, & v > -\frac{1}{\alpha-1}, \\ -v - \frac{1}{\alpha}, & v \le -\frac{1}{\alpha-1}. \end{cases}. \tag{60}$$

By setting $\lambda = -\frac{1}{\alpha-1}$, we have the generalized CGF $\Lambda_{f;X}$ upper bounded by

$$\begin{aligned} \Lambda_{f;X}(t) &= \inf_{\lambda \in \mathbb{R}} \mathbb{E}\left[\phi^*(tX + \lambda)\right] \\ &\le \mathbb{E}\left[\phi^*\left(tX - \frac{1}{\alpha - 1}\right)\right]. \end{aligned} \tag{61}$$

In particular, we calculate

$$\begin{aligned} \phi^*\left(tX - \frac{1}{\alpha-1}\right) &= \begin{cases} \frac{((\alpha-1)tX)^p}{\alpha} - tX + \frac{1}{\alpha(\alpha-1)}, & tX > 0, \\ -tX + \frac{1}{\alpha(\alpha-1)}, & tX \le 0. \end{cases} \\ &\le \frac{((\alpha-1)|tX|)^p}{\alpha} - tX + \frac{1}{\alpha(\alpha-1)}, \quad \forall tX \in \mathbb{R}, \end{aligned} \tag{62}$$

where the inequality holds with $q > 1$. This way, we can further upper bound (61) as

$$\Lambda_{f;X}(t) \le \mathbb{E}\left[\frac{((\alpha-1)|tX|)^p}{\alpha} - tX + \frac{1}{\alpha(\alpha-1)}\right]. \tag{63}$$

Setting $X = \Delta(W, W_i', \tilde{Z}_i)$ gives

$$\begin{aligned} \Lambda_{f;\Delta(W,W_i',\tilde{Z}_i)}(t) &\le \frac{(\alpha-1)^p |t|^p}{\alpha} \mathbb{E}\left[\left|\Delta(W, W_i', \tilde{Z}_i)\right|^p\right] - \mathbb{E}\left[t\Delta(W, W_i', \tilde{Z}_i)\right] + \frac{1}{\alpha(\alpha-1)} \\ &\overset{(a)}{\le} \frac{(\alpha-1)^p}{\alpha} \beta^p |t|^p + \frac{1}{\alpha(\alpha-1)}, \end{aligned} \tag{64}$$

where (a) uses Definition 9 and (10). By choosing $\psi_f(t) = \frac{(\alpha-1)^p}{\alpha}\beta^p t^p + \frac{1}{\alpha(\alpha-1)}$ with $t > 0$, we calculate from Lemma 6 that

$$\begin{aligned} (\psi^*)^{-1}(v) &= \inf_{t>0} \left\{ \frac{1}{t}\left(v + \frac{1}{\alpha(\alpha-1)}\right) + \frac{(\alpha-1)^p}{\alpha}\beta^p t^{p-1} \right\} \\ &= \beta\left(\alpha(\alpha-1)v + 1\right)^{\frac{1}{\alpha}}, \end{aligned} \tag{65}$$

where the infimum is attained at $t = \frac{(\alpha(\alpha-1)v+1)^{\frac{1}{p}}}{(\alpha-1)\beta}$. Substitution into (17) yields

$$\overline{\text{gen}} \le \frac{\beta}{n} \sum_{i=1}^n \left(1 + \alpha(\alpha-1) I_\alpha(W; Z_i | W_i')\right)^{\frac{1}{\alpha}}. \tag{66}$$

The result now follows by combining two bounds (59) and (66).

Specially, when $p = \alpha = 2$, we have $2I_2(W; Z_i|W_i') = I_{\chi^2}(W; Z_i|W_i')$, and thus (24) simplifies to the χ^2-information bound

$$\overline{\text{gen}} \le \frac{\beta}{n} \sum_{i=1}^n I_{\chi^\alpha}(W; Z_i|W_i')^{\frac{1}{\alpha}}. \tag{67}$$

J Proof of Corollary 4

The proof relies on the following lemma, which directly follows from Lemma 87 and Corollary 92 of Agrawal & Horel (2021).

Lemma 9. *(Agrawal & Horel (2021, Lemma 87 and Corollary 92)) Let X be a random variable bounded within $[a, b]$. Suppose f is strictly convex and twice differentiable on its domain, thrice differentiable at 1, and*

$$\frac{27f''(1)}{(3 - xf'''(1)/f''(1))^3} \leq f''(x+1), \quad \forall x \geq -1, \tag{68}$$

then $\Lambda_{f;X}(t) \leq \frac{(b-a)^2}{8f''(1)}t^2$.

One can verify that all of the f-divergence measures in Table 1 satisfy the condition (68), and thus it is feasible to choose $\psi_f(t) = \frac{\beta^2}{2f''(1)}t^2$. After applying Theorem 2, we obtain

$$\overline{\text{gen}} \leq \frac{\beta}{n}\sum_{i=1}^{n}\sqrt{\frac{2}{f''(1)}I_f(W; Z_i|W_i')}, \tag{69}$$

which is the desired result.

K Detailed Arguments for Remark 1

Definition 10 (Bernstein Condition)**.** *Let $w^* := \arg\min_{w\in\mathcal{W}} L(w, P_Z)$ be the optimal hypothesis and define $L_{\text{excess}}(w, P_Z) := L(w, P_Z) - L(w^*, P_Z)$ as the excess loss with respect to w. We say that a learning algorithm satisfies the Bernstein condition with some $B > 0$ and $s \in (0, 1]$ if*

$$\mathbb{E}_Z\left[(\ell(w, Z) - \ell(w^*, Z))^2\right] \leq B\left(L_{\text{excess}}(w, P_Z)\right)^s \tag{70}$$

The following lemma shows that the Bernstein condition implies the proposed mean-p-norm stability when $p \in (1, 2]$.

Lemma 10. *Given $p \in (1, 2]$. If a learning algorithm satisfies the Bernstein condition with some B and s, then it is β-mean-p-norm stable, where*

$$\beta = 2\left(B\mathbb{E}_W\left[\left(L_{\text{excess}}(W, P_Z)\right)^s\right]\right)^{\frac{1}{2}}. \tag{71}$$

Proof. We first prove an upper bound for $\mathbb{E}[(\Delta(W, W_i', Z))^2]$:

$$\begin{aligned}
\mathbb{E}\left[(\Delta(W, W_i', Z))^2\right] &= \mathbb{E}\left[(\ell(W_i', Z) - \ell(W, Z))^2\right] \\
&= \mathbb{E}\left[((\ell(W_i', Z) - \ell(w^*, Z)) - (\ell(W, Z) - \ell(w^*, Z)))^2\right] \\
&\overset{(a)}{\leq} 2\mathbb{E}\left[(\ell(W_i', Z) - \ell(w^*, Z))^2 + (\ell(W, Z) - \ell(w^*, Z))^2\right] \\
&= 2\mathbb{E}\left[(\ell(W_i', Z) - \ell(w^*, Z))^2\right] + 2\mathbb{E}\left[(\ell(W, Z) - \ell(w^*, Z))^2\right] \\
&\overset{(b)}{\leq} 2B\mathbb{E}_{W_i'}\left[\left(L_{\text{excess}}(W_i', P_Z)\right)^s\right] + 2B\mathbb{E}_W\left[\left(L_{\text{excess}}(W, P_Z)\right)^s\right] \\
&\overset{(c)}{=} 4B\mathbb{E}_W\left[\left(L_{\text{excess}}(W, P_Z)\right)^s\right],
\end{aligned} \tag{72}$$

where (a) is by the inequality $(x-y)^2 \leq 2(x^2+y^2)$, (b) is by the Definition 10, and (c) is because W_i' and W have the same marginal distribution. Since $p \in (1, 2]$, it holds by Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\left[(\Delta(W, W_i', Z))^p\right] &\leq \left(\mathbb{E}\left[(\Delta(W, W_i', Z))^2\right]\right)^{\frac{p}{2}} \\
&\overset{(72)}{\leq} 2^p\left(B\mathbb{E}_W\left[\left(L_{\text{excess}}(W, P_Z)\right)^s\right]\right)^{\frac{p}{2}},
\end{aligned} \tag{73}$$

which completes the proof. □

L CALCULATIONS FOR EXAMPLE 1

L.1 VERIFYING VARIOUS STABILITY CONDITIONS

Exact CGF $\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}$. By definition, we first calculate exactly the CGF of $\Delta(W, W_i', \tilde{Z}_i)$ as

$$\begin{aligned}\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}(t) &= \log \mathbb{E}_{W,W_i'}\mathbb{E}_{\tilde{Z}_i}\left[e^{t\left(\ell(W_i',\tilde{Z}_i)-\ell(W,\tilde{Z}_i)\right)}\right] \\ &= \log \mathbb{E}_{W,W_i'}\mathbb{E}_{\tilde{Z}_i}\left[e^{t\left((W_i'-W)(W_i'+W-2\tilde{Z}_i)\right)}\right].\end{aligned} \tag{74}$$

Let $W_- = W_i' - W$ and $W_+ = W_i' + W - 2\tilde{Z}_i$. It can be verified that $W_- \sim \mathcal{N}(0, \frac{2\sigma^2}{n^2})$ and $W_+ \sim \mathcal{N}(0, \frac{2(2n^2+2n-1)}{n^2}\sigma^2)$. Furthermore, we calculate

$$\begin{aligned}\mathbb{E}\left[(W_- - \mathbb{E}[W_-])(W_+ - \mathbb{E}[W_+])\right] &= \mathbb{E}\left[W_-W_+\right] \\ &= \mathbb{E}\left[W_i'^2 - W^2 - 2W_-\tilde{Z}_i\right] \\ &= 0,\end{aligned} \tag{75}$$

which implies W_- and W_+ are independent. By using the well-known moment generating function (MGF) of the Gaussian r.v. W_-, we further calculate

$$\begin{aligned}\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}(t) &= \log \mathbb{E}_{W_+}\mathbb{E}_{W_-}\left[e^{t(W_-W_+)}\right] \\ &= \log \mathbb{E}_{W_+}\left[\exp\left(\frac{\sigma^2}{n^2}W_+^2t^2\right)\right].\end{aligned} \tag{76}$$

Here, W_+^2 follows the χ^2 distribution $\frac{2(2n^2+2n-1)}{n^2}\sigma^2\chi^2$ (with 1 degree of freedom), where the CGF of χ^2 is $-\frac{1}{2}\log(1-2t')$ for $t' < \frac{1}{2}$. With this, we find that (76) converges to

$$\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}(t) = -\frac{1}{2}\log\left(1 - \frac{4(2n^2+2n-1)}{n^4}t^2\sigma^4\right), \tag{77}$$

when

$$\frac{4(2n^2+2n-1)}{n^4}t^2\sigma^4 < 1.$$

Let abbreviate $\frac{\sqrt{4(2n^2+2n-1)}}{n^2}\sigma^2 = c_n \sim \frac{2\sqrt{2}\sigma^2}{n}$, and the above expression is briefly written as

$$\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}(t) = -\frac{1}{2}\log\left(1 - c_n^2t^2\right), \quad \forall t \in \left(-\frac{1}{c_n}, \frac{1}{c_n}\right). \tag{78}$$

Sub-Gaussian stability. We observe that (78) only holds within the interval $t \in (-\frac{1}{c_n}, \frac{1}{c_n})$, while the *sub-Gaussian stability* in Definition 7 requires that $t \in \mathbb{R}$. This contradiction implies that the *sub-Gaussian stability* is not satisfied.

Sub-exponential stability. Using the inequality $-\log(1-x) \leq 2x$ for $x \in [0, \frac{1}{2}]$ and setting $x = c_n^2t^2$, (78) can be upper bounded as

$$\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}(t) \leq c_n^2t^2, \tag{79}$$

which is valid for all

$$t \in \left[-\frac{1}{\sqrt{2}c_n}, \frac{1}{\sqrt{2}c_n}\right].$$

This directly implies the (β, b)-*sub-exponential stability*, where

$$\beta = \sqrt{2}c_n, \quad b = \sqrt{2}c_n.$$

Sub-gamma stability. We can infer from Lemma 3 that the learning algorithm is (β, b)-*sub-gamma-stable* as it is (β, b)-*sub-exponential-stable*, where β and b are given as above.

Mean-p-norm stability with $p = 2$. According to the exact CGF (78), the corresponding moment generating function is

$$M_{\Delta(W,W_i',\tilde{Z}_i)}(t) = \frac{1}{\sqrt{1 - c_n^2 t^2}}. \tag{80}$$

We calculate its second derivative as

$$M''_{\Delta(W,W_i',\tilde{Z}_i)}(t) = c_n^2 \left(1 - c_n^2 t^2\right)^{-\frac{3}{2}} + 3c_n^4 t^2 \left(1 - c_n^2 t^2\right)^{-\frac{5}{2}}, \tag{81}$$

for which the second moment of $\Delta(W, W_i', \tilde{Z}_i)$ is attained at $t = 0$, i.e.,

$$\begin{aligned} \mathbb{E}\left[\left(\Delta(W, W_i', \tilde{Z}_i)\right)^2\right] &= M''_{\Delta(W,W_i',\tilde{Z}_i)}(0) \\ &= c_n^2. \end{aligned} \tag{82}$$

Thus, we have $\beta = c_n$ and the learning algorithm is c_n-mean-square-stable.

L.2 Calculations of Information Measures

By definition, we have $W = \frac{1}{n}\sum_{j=1}^n Z_j$, $W_i' = \frac{1}{n}(\sum_{j=1}^n Z_j - Z_i + Z_i')$. Since Z_j's and Z_i' are i.i.d. Gaussian distributed and W and W_i' are affine transformation functions of $(Z_1, \ldots, Z_n, Z_i')$, then it follows that W, W_i' Z_i and Z_i' are jointly Gaussian distributed, and therefore we can obtain

$$P_{W,W_i',Z_i} = \mathcal{N}\left(\begin{bmatrix}\mu\\ \mu\\ \mu\end{bmatrix}, \begin{bmatrix}\frac{\sigma^2}{n}, & \frac{(n-1)\sigma^2}{n^2}, & \frac{\sigma^2}{n}\\ \frac{(n-1)\sigma^2}{n^2}, & \frac{\sigma^2}{n}, & 0\\ \frac{\sigma^2}{n}, & 0, & \sigma^2\end{bmatrix}\right), \tag{83}$$

$$P_{W,W_i'}P_{Z_i} = \mathcal{N}\left(\begin{bmatrix}\mu\\ \mu\\ \mu\end{bmatrix}, \begin{bmatrix}\frac{\sigma^2}{n}, & \frac{(n-1)\sigma^2}{n^2}, & 0\\ \frac{(n-1)\sigma^2}{n^2}, & \frac{\sigma^2}{n}, & 0\\ 0, & 0, & \sigma^2\end{bmatrix}\right). \tag{84}$$

Due to the fact $W_i' \perp\!\!\!\perp Z_i$, it is immediate from the definition that $I_f(W; Z_i|W_i') = I_f(W, W_i'; Z_i)$. The subsequent calculations of the latter information measure will be relying on (83) and (84).

Mutual information with $f(x) = x \log x$ For d-dimensional multivariate Gaussian random variables $P = \mathcal{N}(\mu_P, \Sigma_P)$ and $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$, where μ_P (resp., μ_Q) denotes the mean vector of P (resp., Q) and Σ_P (resp., Σ_Q) denotes covariance matrix of P (resp., Q), the KL-divergence between P and Q is (Liu et al., 2025)

$$D(P\|Q) = \frac{1}{2}\left(\log\frac{\det(\Sigma_Q)}{\det(\Sigma_P)} - d + \mathrm{tr}(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^\top \Sigma_Q^{-1}(\mu_Q - \mu_P)\right). \tag{85}$$

Setting $P = P_{W,W_i',Z_i}$ and $Q = P_{W,W_i'}P_{Z_i}$, direct calculations show that

$$\begin{aligned} I(W; Z_i|W_i') &= D(P_{W,W_i',Z_i}\|P_{W,W_i'}P_{Z_i}) \\ &= \frac{1}{2}\log\frac{\det(\Sigma_Q)}{\det(\Sigma_P)} \\ &= \frac{1}{2}\log\frac{2n-1}{n-1}. \end{aligned} \tag{86}$$

χ^2-information with $f(x) = (x-1)^2$. For $P = \mathcal{N}(\mu_P, \Sigma_P)$ and $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$, the χ^2-divergence between P and Q is (Liu et al., 2025)

$$D_{\chi^2}(P\|Q) = \frac{\det(\Sigma_Q)}{\sqrt{\det(\Sigma_P)\det(2\Sigma_Q - \Sigma_P)}} \exp\left((\mu_P - \mu_Q)^\top (2\Sigma_Q - \Sigma_P)^{-1}(\mu_P - \mu_Q)\right) - 1 \tag{87}$$

Setting $P = P_{W,W_i',Z_i}$ and $Q = P_{W,W_i'}P_{Z_i}$, direct calculations show that

$$\begin{aligned} I_{\chi^2}(W; Z_i|W_i') &= D_{\chi^2}(P_{W,W_i',Z_i}\|P_{W,W_i'}P_{Z_i}) \\ &= \frac{\det(\Sigma_Q)}{\sqrt{\det(\Sigma_P)\det(2\Sigma_Q - \Sigma_P)}} - 1 \\ &= \frac{n}{n-1}. \end{aligned} \tag{88}$$

L.3 Evaluation of the Proposed Bounds

Evaluation of Corollary 1. Using the stability parameters (β, b) (given in Appendix L.1) and mutual information $I(W; Z_i|W_i')$ (given in (86)), we find that $I(W; Z_i|W_i') \leq \frac{\beta^2}{2b^2}$ for $n \geq 3$. Hence, (20) specializes to

$$\overline{\text{gen}} \leq c_n \sqrt{2 \log \frac{2n-1}{n-1}} \sim \frac{4\sqrt{\log 2}\sigma^2}{n} \approx \frac{3.33\sigma^2}{n}, \quad \forall n \geq 3, \tag{89}$$

which is of order $\Theta(\frac{1}{n})$.

Evaluation of Corollary 2. By plugging the stability parameters (β, c) (given in Appendix L.1) and mutual information $I(W; Z_i|W_i')$ (given in (86)) into (21), we get

$$\overline{\text{gen}} \leq c_n \left(\sqrt{2 \log \frac{2n-1}{n-1}} + \frac{\sqrt{2}}{2} \log \frac{2n-1}{n-1} \right) \sim \frac{(4\sqrt{\log 2} + 2\log 2)\sigma^2}{n} \approx \frac{4.71\sigma^2}{n}, \tag{90}$$

which is slightly looser than (89) up to a constant.

Evaluation of Corollary 3. By setting $p = \alpha = 2$ and using the stability parameter β (given in Appendix L.1), χ^2-divergence $I_{\chi^2}(W; Z_i|W_i')$ (given in (88)), we have (67) specialized to

$$\overline{\text{gen}} \leq c_n \sqrt{\frac{n}{n-1}} \sim \frac{2\sqrt{2}\sigma^2}{n} \approx \frac{2.83\sigma^2}{n}, \tag{91}$$

which is tighter than both (89) and (90).

Direct evaluation of Theorem 2 We limit Theorem 2 within the choice of mutual information. In particular, we retain (78) as $\psi(t)$. Together with (86), the generalization bound (17) is calculated as

$$\begin{aligned} \overline{\text{gen}} &\leq \frac{1}{n} \sum_{i=1}^{n} \inf_{t>0} \frac{1}{t} \left(\Lambda_{\Delta(W, W_i', \tilde{Z}_i)}(t) + I(W; Z_i|W_i') \right) \\ &= \inf_{t \in (0, \frac{1}{c_n})} \frac{1}{2t} \left(-\log\left(1 - c_n^2 t^2\right) + \log \frac{2n-1}{n-1} \right) \\ &= \inf_{t \in (0,1)} \frac{c_n}{2t} \left(-\log\left(1 - t^2\right) + \log \frac{2n-1}{n-1} \right). \end{aligned} \tag{92}$$

Taking $t = \frac{1}{\sqrt{2}}$ yields

$$\overline{\text{gen}} \leq \frac{c_n\sqrt{2}}{2} \log \frac{2(2n-1)}{n-1} \sim \frac{(4\log 2)\sigma^2}{n} \approx \frac{2.77\sigma^2}{n}, \tag{93}$$

which strictly improves upon (91) for all n.

L.4 Justifying the Tightness of Bound (17)

We show that the mutual information $I(W; Z_i|W_i')$ also gives a lower bound for Example 1.

Proposition 1. *(A Lower Bound for Example 1) Consider Example 1, then the following lower bound holds for* $\overline{\text{gen}}$*:*

$$\overline{\text{gen}} \geq \frac{2\sigma^2(n-1)}{n^3} \sum_{i=1}^{n} I(W; Z_i|W_i'). \tag{94}$$

Proof. With $f(x) = x \log x$, Lemma 5 specializes to the well-known Donsker-Varadhan representation of the KL-divergence. In this case, the optimal function is attained at $g^* = f'(\frac{\mathrm{d}P}{\mathrm{d}Q}) =$

$1+\log\frac{\mathrm{d}P}{\mathrm{d}Q}$. Under Example 1, with $P=P_{W,W_i',Z_i}$ and $Q=P_{W,W_i'}P_{Z_i}$ provided in (83) and (84), we calculate the optimum as

$$\begin{aligned} g^*(W,W_i',Z_i) =& \overbrace{1+\frac{1}{2}\log\frac{2n-1}{n-1}+\frac{n^2}{2\sigma^2}\left(\frac{1}{2n-1}\bar{W}^2-\frac{n^2+n-1}{n(2n-1)}\bar{W}_i'^2+\frac{2n}{2n-1}\bar{W}\bar{W}_i'\right)}^{\tilde{g}(W,W_i')} \\ &+\frac{n}{2\sigma^2}\left((W_i'-Z_i)^2-\frac{n}{n-1}(W-Z_i)^2\right), \end{aligned} \tag{95}$$

where $\bar{W}=W-\mu, \bar{W}_i'=W_i'-\mu, \bar{Z}_i=Z_i-\mu$. Let define $\tilde{\Delta}(w,w_i',z)=\ell(w_i',z)-\frac{n}{n-1}\ell(w,z)$, then we briefly rewrite

$$g^*(W,W_i',Z_i)=\frac{n}{2\sigma^2}\tilde{\Delta}(W,W_i',Z)+\tilde{g}(W,W_i'). \tag{96}$$

The Donsker-Varadhan formula gives that

$$\begin{aligned} I(W;Z_i|W_i') &= \mathbb{E}_{W,W_i',Z_i}\left[g^*(W,W_i',Z_i)\right]-\log\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[e^{g^*(W,W_i',Z_i)}\right] \\ &\le \mathbb{E}_{W,W_i',Z_i}\left[g^*(W,W_i',Z_i)\right]-\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[g^*(W,W_i',Z_i)\right], \end{aligned} \tag{97}$$

where the inequality follows by Jensen's inequality. After plugging (96) and rearranging terms, it follows that

$$\begin{aligned} I(W;Z_i|W_i') \le& \mathbb{E}_{W,W_i',Z_i}\left[\frac{n}{2\sigma^2}\tilde{\Delta}(W,W_i',Z_i)\right]-\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[\frac{n}{2\sigma^2}\tilde{\Delta}(W,W_i',Z_i)\right] \\ &+\mathbb{E}_{W,W_i',Z_i}\left[\tilde{g}(W,W_i')\right]-\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[\tilde{g}(W,W_i')\right] \\ =&\frac{n}{2\sigma^2}\mathbb{E}_{W,W_i',Z_i}\left[\ell(W_i',Z_i)-\frac{n}{n-1}\ell(W,Z_i)\right] \\ &-\frac{n}{2\sigma^2}\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[\ell(W_i',Z_i)-\frac{n}{n-1}\ell(W,Z_i)\right] \\ =&\frac{n^2}{2\sigma^2(n-1)}\left(\mathbb{E}_{W,W_i'}\mathbb{E}_{Z_i}\left[\ell(W,Z_i)\right]-\mathbb{E}_{W,W_i',Z_i}\left[\ell(W,Z_i)\right]\right). \end{aligned} \tag{98}$$

Averaging over $i\in[n]$ on both sides completes the proof. □

M Calculations for Example 2

For brevity, we let $B_{n,k}(\theta):=\binom{n}{k}\theta^k(1-\theta)^{n-k}$ denote the Bernstein polynomial. Besides, let $S_i:=\sum_{j\neq i}Z_j$ and thus $S_i\sim\mathrm{Bino}(n-1,\theta)$. With this, we simplify Δ as

$$\begin{aligned} \Delta(W,W_i',Z) &= (W_i'-Z)^2-(W-Z)^2 \\ &= \frac{1}{n}(Z_i'-Z_i)\left(\frac{2}{n}S_i+\frac{1}{n}Z_i+\frac{1}{n}Z_i'-2Z\right). \end{aligned} \tag{99}$$

M.1 Verifying Various Stability Conditions

Strong uniform stability. For all $i\in[n]$,

$$\begin{aligned} \beta &= \sup_{z_{[n]},z_i',z}|\Delta(w,w_i',z)| \\ &= \frac{1}{n}\sup_{s_i,z_i,z_i',z}|z_i'-z_i|\left|\frac{2}{n}s_i+\frac{1}{n}z_i+\frac{1}{n}z_i'-2z\right| \\ &= \frac{2n-1}{n^2}. \end{aligned} \tag{100}$$

Mean-p-norm stability with $p \in (1,2]$. We can see that (99) is 0 unless $Z_i + Z_i' = 1$, thus the exact mean-p-norm stability parameter is

$$\begin{aligned}\beta^p &= \mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[|\Delta(W,W_i',Z)|^p\right] \\ &= \mathbb{E}_{W,W_i'}\mathbb{E}_Z\left[|\Delta(W,W_i',Z)|^p \,|Z_i + Z_i' = 1\right]\Pr[Z_i + Z_i' = 1] \\ &= 2\theta(1-\theta)\mathbb{E}_{S_i}\mathbb{E}_Z\left[\left|\frac{2}{n}S_i + \frac{1}{n} - 2Z\right|^p\right] \\ &= \frac{2\theta(1-\theta)}{n^{2p}}\mathbb{E}_{S_i}\left[(1-\theta)(2S_i+1)^p + \theta(2n-2S_i-1)^p\right].\end{aligned} \tag{101}$$

where $S_i \sim \mathrm{Bino}(n-1,\theta)$. In particular, for $p=2$, the above expression can be further calculated as

$$\beta^2 = \frac{2\theta(1-\theta)}{n^4}\left(1 + 4(n^2+n-2)\theta(1-\theta)\right). \tag{102}$$

M.2 Calculations of Information Measures

Calculation of the common Radon-Nikodym derivative. To evaluate various information measures, we first calculate the common derivative x used in each generator f, where

$$\begin{aligned}x &= \frac{\mathrm{d}P_{W,W_i',Z_i}}{\mathrm{d}P_{W,W_i'}P_{Z_i}}(w,w_i',z_i) \\ &= \frac{\mathrm{d}P_{Z_i|W,W_i'}}{\mathrm{d}P_{Z_i}}(w,w_i',z_i).\end{aligned} \tag{103}$$

In the following, we calculate $P_{Z_i|W,W_i'}$.

By definition, we have $W_i' - W = \frac{1}{n}(Z_i' - Z_i)$. This implies that $n|W_i' - W| = 0$ or $n|W_i' - W| = 1$, otherwise $P_{Z_i|W,W_i'} = 0$. We first consider $n|W_i' - W| = 1$, in which case Z_i can be immediately determined, i.e.,

$$\begin{aligned}P_{Z_i|W,W_i'}\left(0\,\middle|\,\frac{k}{n},\frac{k+1}{n}\right) &= 1, \\ P_{Z_i|W,W_i'}\left(1\,\middle|\,\frac{k}{n},\frac{k+1}{n}\right) &= 1,\end{aligned} \tag{104}$$

where $0 \le k \le n-1$. Next, we consider the case $W = W_i' = \frac{k}{n}$. It follows that

$$\begin{aligned}P_{Z_i,W,W_i'}\left(1,\frac{k}{n},\frac{k}{n}\right) &= \Pr\left[Z_i = 1, Z_i' = 1, S_i = k-1\right] \\ &= \theta^2 B_{n-1,k-1}(\theta), \\ P_{Z_i,W,W_i'}\left(0,\frac{k}{n},\frac{k}{n}\right) &= \Pr\left[Z_i = 0, Z_i' = 0, S_i = k\right] \\ &= (1-\theta)^2 B_{n-1,k}(\theta),\end{aligned} \tag{105}$$

and furthermore,

$$\begin{aligned}P_{W,W_i'}\left(\frac{k}{n},\frac{k}{n}\right) &= P_{Z_i,W,W_i'}\left(1,\frac{k}{n},\frac{k}{n}\right) + P_{Z_i,W,W_i'}\left(0,\frac{k}{n},\frac{k}{n}\right) \\ &= \theta^2 B_{n-1,k-1}(\theta) + (1-\theta)^2 B_{n-1,k}(\theta) \\ &= \frac{k\theta + (n-k)(1-\theta)}{n} B_{n,k}(\theta).\end{aligned} \tag{106}$$

Thus, it holds by Bayes' rule that

$$\begin{aligned}P_{Z_i|W,W_i'}\left(1\,\middle|\,\frac{k}{n},\frac{k}{n}\right) &= \frac{P_{Z_i,W,W_i'}\left(1,\frac{k}{n},\frac{k}{n}\right)}{P_{W,W_i'}\left(\frac{k}{n},\frac{k}{n}\right)} \\ &= \frac{k\theta}{k\theta + (n-k)(1-\theta)},\end{aligned} \tag{107}$$

and hence

$$\begin{aligned} P_{Z_i|W,W_i'}\left(0\middle|\frac{k}{n},\frac{k}{n}\right) &= 1 - P_{Z_i|W,W_i'}\left(1\middle|\frac{k}{n},\frac{k}{n}\right) \\ &= \frac{(n-k)(1-\theta)}{k\theta+(n-k)(1-\theta)}. \end{aligned} \tag{108}$$

By combining all the cases together, we have

$$x = \begin{cases} \frac{n-k}{k\theta+(n-k)(1-\theta)}, & (w,w_i',z_i) = \left(\frac{k}{n},\frac{k}{n},0\right), \quad 0\le k\le n, \\ \frac{k}{k\theta+(n-k)(1-\theta)}, & (w,w_i',z_i) = \left(\frac{k}{n},\frac{k}{n},1\right), \quad 0\le k\le n, \\ \frac{1}{1-\theta}, & (w,w_i',z_i) = \left(\frac{k}{n},\frac{k+1}{n},0\right), \quad 0\le k< n, \\ \frac{1}{\theta}, & (w,w_i',z_i) = \left(\frac{k+1}{n},\frac{k}{n},1\right), \quad 0\le k< n, \\ 0, & \text{otherwise.} \end{cases} \tag{109}$$

For any generator f and assume $f(0) = \lim_{x\to 0} f(x)$, substitution gives

$$\begin{aligned} I_f(W;Z_i|W_i') =& \theta(1-\theta)\left[(1-\theta)f\left(\frac{1}{1-\theta}\right) + \theta f\left(\frac{1}{\theta}\right) + f(0)\right] \\ &+ \mathbb{E}\left[\frac{\theta X+(1-\theta)(n-X)}{n}(1-\theta)f\left(\frac{n-X}{\theta X+(1-\theta)(n-X)}\right)\right] \\ &+ \mathbb{E}\left[\frac{\theta X+(1-\theta)(n-X)}{n}\theta f\left(\frac{X}{\theta X+(1-\theta)(n-X)}\right)\right]. \end{aligned} \tag{110}$$

For the subsequent calculations, we let $X \sim \mathrm{Bino}(n,\theta)$ for brevity.

Mutual information with $f(x) = x\log x$. With the convention $0\log 0 = 0$, (110) simplifies to

$$I(W;Z_i|W_i') = H_2(\theta) - \mathbb{E}\left[\frac{\theta X+(1-\theta)(n-X)}{n} H_2\left(\frac{\theta X}{\theta X+(1-\theta)(n-X)}\right)\right]. \tag{111}$$

χ^α-information with $f(x) = |x-1|^\alpha$, $\alpha \ge 2$. Substitution gives

$$\begin{aligned} I_{\chi^\alpha}(W;Z_i|W_i') =& \theta(1-\theta)\left(1 + \frac{\theta^\alpha}{(1-\theta)^{\alpha-1}} + \frac{(1-\theta)^\alpha}{\theta^{\alpha-1}}\right) \\ &+ \mathbb{E}\left[\frac{\theta X+(1-\theta)(n-X)}{n}\left((1-\theta)\left|\frac{n-X}{\theta X+(1-\theta)(n-X)} - 1\right|^\alpha\right.\right. \\ &\left.\left. +\theta\left|\frac{X}{\theta X+(1-\theta)(n-X)} - 1\right|^\alpha\right)\right]. \end{aligned} \tag{112}$$

For $\alpha = 2$, this reduces to

$$I_{\chi^2}(W;Z_i|W_i') = 1 - \mathbb{E}\left[\frac{X(n-X)}{n\left(\theta X+(1-\theta)(n-X)\right)}\right]. \tag{113}$$

α-information with $f(x) = \frac{x^\alpha-\alpha x+\alpha-1}{\alpha(\alpha-1)}$. For $\alpha > 0$ and $\alpha \ne 1$,

$$\begin{aligned} I_\alpha(W;Z_i|W_i') =& \frac{\theta(1-\theta)}{\alpha(\alpha-1)}\left((1-\theta)^{1-\alpha} + \theta^{1-\alpha}\right) - \frac{1}{\alpha(\alpha-1)} \\ &+ \frac{1}{\alpha(\alpha-1)n}\mathbb{E}\left[\frac{(1-\theta)(n-X)^\alpha + \theta X^\alpha}{\left(\theta X+(1-\theta)(n-X)\right)^{\alpha-1}}\right]. \end{aligned} \tag{114}$$

H^2-information with $f(x) = (\sqrt{x}-1)^2$. Substitution gives

$$\begin{aligned} I_{H^2}(W;Z_i|W_i') =& 2 - 2\theta(1-\theta)(\sqrt{1-\theta}+\sqrt{\theta}) \\ &- \frac{2}{n}\mathbb{E}\left[\sqrt{\theta X+(1-\theta)(n-X)}\left((1-\theta)\sqrt{n-X} + \theta\sqrt{X}\right)\right]. \end{aligned} \tag{115}$$

M.3 Evaluation of the Proposed Bounds

Evaluation of (2). Using (100) directly gives

$$\overline{\text{gen}} \leq \frac{2n-1}{n^2}. \tag{116}$$

Evaluation of Corollary 3. For every $p \in (1, 2]$, we substitute (101), (112) and (114) into (24). In particular, for $p = \alpha = 2$, we can theoretically justify that at the resulting χ^2-information bound (i.e., (67)) decays with $\Theta(\frac{1}{n})$. To see this, we evaluate separately the mean-2-norm stability parameter β (given in (102)) and the χ^2-information (given in (113)). It is easy to check that

$$\beta = \sqrt{\frac{2\theta(1-\theta)}{n^4}\left(1 + 4(n^2+n-2)\theta(1-\theta)\right)} \sim \frac{2\sqrt{2}\theta(1-\theta)}{n} \tag{117}$$

Besides, with $\theta \in (0, 1)$, Jensen's inequality implies $\frac{1}{\theta X + (1-\theta)(n-X)} \leq \frac{\theta}{X} + \frac{1-\theta}{n-X}$. With this, we can lower bound the χ^2-information as

$$\begin{aligned} I_{\chi^\alpha}(W; Z_i|W_i') &= 1 - \mathbb{E}\left[\frac{X(n-X)}{n\left(\theta X + (1-\theta)(n-X)\right)}\right] \\ &\geq 1 - \mathbb{E}\left[X(n-X)\left(\frac{\theta}{X} + \frac{1-\theta}{n-X}\right)\right] \\ &= 1 - \mathbb{E}\left[\frac{\theta(n-X) + (1-\theta)X}{n}\right] \\ &= 1 - 2\theta(1-\theta) \\ &\geq \frac{1}{2}. \end{aligned} \tag{118}$$

Combining this with the trivial upper bound $I_{\chi^\alpha}(W; Z_i|W_i') \leq 1$, we see that $I_{\chi^\alpha}(W; Z_i|W_i') \in \Theta(1)$. Using these facts, we conclude that (67) is

$$\begin{aligned} \overline{\text{gen}} &\leq \frac{\beta}{n}\sum_{i=1}^{n}\sqrt{I_{\chi^\alpha}(W; Z_i|W_i')} \\ &\leq \beta \sim \frac{2\sqrt{2}\theta(1-\theta)}{n} \approx \frac{2.83\theta(1-\theta)}{n}. \end{aligned} \tag{119}$$

Evaluation of Corollary 4. Substituting the strong uniform stability parameter from (100) and the information measures from Appendix M.2 gives the desired result.

N Calculations for Example 3

For brevity, we let $s := \sum_{j=1}^{n} x_j^2$ and let $C := \frac{\max_{j\in[n]}|x_j|}{\min_{j\in[n]}|x_j|}$.

N.1 Verifying Various Stability Conditions

Exact CGF $\Lambda_{\Delta(W,W_i',\tilde{Z}_i)}$. According to the replace-one setting, Z_i is replaced by $Z_i' = (x_i, w^* x_i + N_i')$, where $N_i' \sim \mathcal{N}(0, \sigma^2)$ is an independent Gaussian noise, we therefore have

$$W = w^* + \frac{\sum_{j=1}^{n} x_j N_j}{s}, \qquad W_i' = w^* + \frac{x_i N_i' + \sum_{j\neq i} x_j N_j}{s}, \tag{120}$$

. Besides, we have $\tilde{Z}_i = (x_i, w^* x_i + \tilde{N}_i)$, where $\tilde{N}_i$ is another noise drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$. With these, $\Delta(W, W_i', \tilde{Z}_i)$ is calculated as

$$\begin{aligned}\Delta(W, W_i', \tilde{Z}_i) &= \ell(W_i', \tilde{Z}_i) - \ell(W, \tilde{Z}_i) \\ &= \left(w^* x_i + \tilde{N}_i - W_i' x_i\right)^2 - \left(w^* x_i + \tilde{N}_i - W x_i\right)^2 \\ &= x_i \left(W - W_i'\right)\left(2w^* x_i + 2\tilde{N}_i - W_i' x_i - W x_i\right) \\ &= \frac{x_i^2 (N_i - N_i')}{s}\left(2\tilde{N}_i - \frac{x_i\left(x_i N_i + x_i N_i' + 2\sum_{j\neq i} x_j N_j\right)}{s}\right).\end{aligned} \tag{121}$$

Let $N_- = \frac{x_i^2(N_i - N_i')}{s}$ and $N_+ = 2\tilde{N}_i - \frac{x_i\left(x_i N_i + x_i N_i' + 2\sum_{j\neq i} x_j N_j\right)}{s}$. We verify that $N_- \sim \mathcal{N}(0, \frac{2x_i^4}{s^2}\sigma^2)$ and $N_+ \sim \mathcal{N}(0, 4\sigma^2 + \frac{2x_i^2(2s - x_i^2)}{s^2}\sigma^2)$ and furthermore,

$$\begin{aligned}&\mathbb{E}\left[\left(N_- - \mathbb{E}[N_-]\right)\left(N_+ - \mathbb{E}[N_+]\right)\right] \\ &= \mathbb{E}\left[N_- N_+\right] \\ &= \mathbb{E}\left[\frac{2x_i^2(N_i - N_i')\tilde{N}_i}{s} - \frac{x_i^3(N_i - N_i')\left(x_i N_i + x_i N_i' + 2\sum_{j\neq i} x_j N_j\right)}{s^2}\right] \\ &= \frac{2x_i^2}{s}\mathbb{E}\left[(N_i - N_i')\tilde{N}_i\right] - \frac{x_i^3}{s^2}\left(x_i \mathbb{E}\left[N_i^2 - N_i'^2\right] + 2\sum_{j\neq i} x_j \mathbb{E}\left[N_j(N_i - N_i')\right]\right) \\ &= 0.\end{aligned} \tag{122}$$

which implies $N_- \perp\!\!\!\perp N_+$. Similarly to Appendix L.1, we can exactly calculate the CGF of $\Delta(W, W_i', \tilde{Z}_i)$ as

$$\begin{aligned}\Lambda_{\Delta(W, W_i', \tilde{Z}_i)}(t_i) &= \log \mathbb{E}\left[\exp\left(t_i \Delta(W, W_i', \tilde{Z}_i)\right)\right] \\ &= \log \mathbb{E}\left[\exp\left(t_i N_- N_+\right)\right] \\ &\overset{(a)}{=} \log \mathbb{E}\left[\exp\left(\frac{t_i^2 x_i^4 \sigma^2}{s^2} N_+^2\right)\right] \\ &= -\frac{1}{2}\log\left(1 - \frac{4x_i^4}{s^2}\left(3 - \left(\frac{x_i^2}{s} - 1\right)^2\right)\sigma^4 t_i^2\right) \\ &\overset{(b)}{=} -\frac{1}{2}\log\left(1 - c_i^2 t_i^2\right), \qquad \forall t_i \in \left(-\frac{1}{c_i}, \frac{1}{c_i}\right)\end{aligned} \tag{123}$$

where (a) uses the fact that $N_- \perp\!\!\!\perp N_+$ and (b) uses the definition $c_i := \frac{2x_i^2\sigma^2}{s}\sqrt{3 - \left(\frac{x_i^2}{s} - 1\right)^2}$. Specially, we can see that

$$c_i \leq \frac{2\sqrt{3}x_i^2\sigma^2}{s} \tag{124}$$

$$\leq \frac{2\sqrt{3}C^2\sigma^2}{C^2 + n - 1}, \tag{125}$$

implying c_i is of $O(\frac{1}{n})$.

Mean-p-norm stability with $p = 2$. Since $\Delta(W, W_i', \tilde{Z}_i) = N_- N_+$ and $N_- \perp\!\!\!\perp N_+$, we calculate

$$\begin{aligned}\mathbb{E}\left[\Delta(W, W_i', \tilde{Z}_i)^2\right] &= \mathbb{E}[N_-^2]\mathbb{E}[N_+^2] \\ &= \frac{4x_i^4\sigma^4}{s^2}\left(3 - \left(\frac{x_i^2}{s} - 1\right)^2\right) \\ &= c_i^2.\end{aligned} \tag{126}$$

Thus, the common parameter required by Definition 9 is

$$\beta = \max_{i\in[n]} c_i, \tag{127}$$

which is $\Theta(\frac{1}{n})$. To obtain a tighter result, for each $i \in [n]$ we choose a point-wise stability parameter β_i via $\mathbb{E}[\Delta(W, W_i', \tilde{Z}_i)^2] \le \beta_i^2$. In this case, we simply have $\beta_i = c_i$ by (126).

N.2 Calculations of Information Measures

Since x_i's are fixed, the randomness of $Z_i = (x_i, Y_i)$ comes exactly from $Y_i = w^* x_i + N_i$. Consequently, we can omit x_i and abbreviate $P_{W,W_i',Y_i} = P$ and $P_{W,W_i'}P_{Y_i} = Q$. Recall (120), we can see that P and Q are both multivariate Gaussian distributions, i.e.,

$$P_{W,W_i',Z_i} = \mathcal{N}\left(\begin{bmatrix} w^* \\ w^* \\ w^* x_i \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{s}, & \frac{(s-x_i^2)\sigma^2}{s^2}, & \frac{x_i\sigma^2}{s} \\ \frac{(s-x_i^2)\sigma^2}{s^2}, & \frac{\sigma^2}{s}, & 0 \\ \frac{x_i\sigma^2}{s}, & 0, & \sigma^2 \end{bmatrix}\right), \tag{128}$$

$$P_{W,W_i'}P_{Z_i} = \mathcal{N}\left(\begin{bmatrix} w^* \\ w^* \\ w^* x_i \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{s}, & \frac{(s-x_i^2)\sigma^2}{s^2}, & 0 \\ \frac{(s-x_i^2)\sigma^2}{s^2}, & \frac{\sigma^2}{s}, & 0 \\ 0, & 0, & \sigma^2 \end{bmatrix}\right). \tag{129}$$

Direct calculation gives

$$\det(\Sigma_P) = \frac{\sigma^6 x_i^2 (s - x_i^2)}{s^4}, \qquad \det(\Sigma_Q) = \frac{\sigma^6 x_i^2 (2s - x_i^2)}{s^4}. \tag{130}$$

Mutual information with $f(x) = x\log x$. Applying (85) with (130), we obtain

$$\begin{aligned} I(W; Z_i|W_i') &= I(W, W_i'; Z_i) \\ &= \frac{1}{2}\log\frac{\det(\Sigma_Q)}{\det(\Sigma_P)} \\ &= \frac{1}{2}\log\frac{2s - x_i^2}{s - x_i^2}. \end{aligned} \tag{131}$$

χ^2-information with $f(x) = (x-1)^2$. Applying (87) with (130), we obtain

$$\begin{aligned} I_{\chi^2}(W; Z_i|W_i') &= I_{\chi^2}(W, W_i'; Z_i) \\ &= \frac{\det(\Sigma_Q)}{\sqrt{\det(\Sigma_P)\det(2\Sigma_Q - \Sigma_P)}} - 1 \\ &= \frac{s}{s - x_i^2}. \end{aligned} \tag{132}$$

N.3 Evaluation of the Proposed Bounds

Evaluation of Corollary 3 We consider $p = \alpha = 2$. Using the point-wise stability parameter β_i (given in (127)) and the χ^2-divergence $I_{\chi^2}(W; Z_i|W_i')$ (given in (132)), we have (67) specialized to

$$\overline{\text{gen}} \le \frac{1}{n}\sum_{i=1}^n c_i \sqrt{1 + \frac{x_i^2}{s - x_i^2}}. \tag{133}$$

Using (124) and $\frac{x_i^2}{s-x_i^2} \le \frac{C^2}{n-1}$, we further calculate

$$\begin{aligned} \overline{\text{gen}} &\le \frac{1}{n}\sum_{i=1}^n \frac{2\sqrt{3}x_i^2\sigma^2}{s}\sqrt{1 + \frac{C^2}{n-1}} \\ &= \frac{2\sqrt{3}\sigma^2}{n} + o\left(\frac{1}{n}\right) \\ &\approx \frac{3.46\sigma^2}{n} + o\left(\frac{1}{n}\right), \end{aligned} \tag{134}$$

which is of $\Theta(\frac{1}{n})$.

Direct evaluation of Theorem 2 First, we show (131) can be upper bounded by

$$\begin{aligned} I(W, W_i'; Z_i) &= \frac{1}{2} \log \left(2 + \frac{x_i^2}{s - x_i^2} \right) \\ &\leq \frac{1}{2} \log \left(2 + \frac{C^2}{n-1} \right) \\ &\leq \frac{1}{2} \log 2 + \frac{C^2}{4(n-1)}. \end{aligned} \tag{135}$$

where the last inequality uses $\log(1+x) \leq x$ for $x > 0$. By combining the results in (135) and (123), (17) specializes to

$$\begin{aligned} \overline{\text{gen}} &\leq \frac{1}{n} \sum_{i=1}^{n} \inf_{t_i > 0} \frac{1}{t_i} \left(\Lambda_{\Delta(W, W_i', \tilde{Z}_i)}(t_i) + I(W; Z_i | W_i') \right) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \inf_{t_i \in (0, \frac{1}{c_i})} \frac{1}{t_i} \left(-\frac{1}{2} \log \left(1 - c_i^2 t_i^2 \right) + \frac{1}{2} \log 2 + \frac{C^2}{4(n-1)} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \inf_{t_i \in (0,1)} \frac{c_i}{t_i} \left(-\frac{1}{2} \log \left(1 - t_i^2 \right) + \frac{1}{2} \log 2 + \frac{C^2}{4(n-1)} \right). \end{aligned} \tag{136}$$

Setting $t_i = \frac{1}{\sqrt{2}}$ for each $i \in [n]$, we have

$$\begin{aligned} \overline{\text{gen}} &\leq \frac{1}{n} \sum_{i=1}^{n} c_i \sqrt{2} \left(\log 2 + \frac{C^2}{4(n-1)} \right) \\ &\overset{(a)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \frac{2\sqrt{6} x_i^2 \sigma^2}{s} \left(\log 2 + \frac{C^2}{4(n-1)} \right) \\ &= \frac{(2\sqrt{6} \log 2) \sigma^2}{n} + o\left(\frac{1}{n} \right) \\ &\approx \frac{3.40 \sigma^2}{n} + o\left(\frac{1}{n} \right), \end{aligned} \tag{137}$$

where (a) uses (124). We find that (137) achieves the optimal rate $\Theta(\frac{1}{n})$ and is tighter than (134) by a factor 0.98.

O Mutual Information Bounds Based on Sample-Conditioned Hypothesis (SCH) Stability

Our results in Theorem 1 and Corollary 4 are based on the strong uniform stability in Definition 6 and hence directly comparable to Wang & Mao (2023a, Corollary 3.1). Similarly, our Example 2 (which we use for numerical evaluations of Corollary 4) satisfies the SCH-A and SCH-B stability notions from Wang & Mao (2023a), and consequently the bounds in Wang & Mao (2023a, Section 3) (including Wang & Mao (2023a, Theorem 3.1, 3.2, 3.3 and Corollary 3.1)) would be applicable to this example in principle. However, we believe that these prior results do not hold as they are stated in Wang & Mao (2023a). The reason for this is that in either case of Example 2 or Example 4 (which we propose below), the exact generalization error can be exactly calculated and each of the bounds from Wang & Mao (2023a, Section 3) gives a *lower* bound than the true generalization error, leading to a contradiction.

In this appendix, we argue in detail why the results in Wang & Mao (2023a, Section 3) imply contradictions with our calculations for Example 2. Additionally, we have identified that all these results are established on a common proof step, which, in our view, does not hold in general and thus is a possible cause for these contradictions. This leads us to the conclusion that the results in Wang & Mao (2023a, Theorem 3.1, 3.2, 3.3 and Corollary 3.1) do not hold as they are stated and therefore do not need to be compared to our results.

Evaluation of Wang & Mao (2023a, Corollary 3.1). We start with Wang & Mao (2023a, Corollary 3.1) which also hinges on the same strong-uniform stability as used in our work. For a β-strongly-uniformly-stable algorithm, their stated bound is

$$\overline{\text{gen}} \leq \frac{\beta}{n} \sum_{i=1}^{n} \sqrt{2I(W; Z_i)}. \tag{138}$$

However, we find Example 2 is a counterexample that this result does not hold. In Example 2, the specific mutual information $I(W; Z_i)$ can be exactly calculated as (for details c.f. Lu et al. (2026, Appendix V.C))

$$I(W; Z_i) = H_2(\theta) - \log n + \theta \mathbb{E}\left[\log(X+1)\right] + (1-\theta)\mathbb{E}\left[\log(n-X)\right], \tag{139}$$

where $X \sim \text{Bino}(n-1, \theta)$. This quantity is $O(\frac{1}{n})$ (Lu et al., 2026, Appendix V.C), whereas Appendix M.1 gives the admissible stability parameter $\beta = \frac{2n-1}{n^2}$. Thus, the RHS of (138) decays with order $O(n^{-\frac{3}{2}})$, while the LHS, i.e., the exact generalization error, is $\frac{2\theta(1-\theta)}{n}$. Consequently, the inequality fails for all sufficiently large n. This contradiction implies that (138) does not hold as stated.

In the following, we evaluate the remaining bounds from Wang & Mao (2023a, Section 3).

Verifying the SCH stability conditions for Example 2. We calculate the SCH stability parameters γ_1 and γ_2 in Wang & Mao (2023a, Definition 2.1 (a) and (b)). Let

$$g(w, z) := \mathbb{E}_{W_i'|W=w}[\ell(W_i', z)] - \ell(w, z). \tag{140}$$

First, we need to calculate $P_{W_i'|W}$. For $w = \frac{k}{n}$ with $0 \leq k \leq n$, it holds by Bayes' rule that

$$\begin{aligned} P_{W_i'|W}\left(w' \middle| \frac{k}{n}\right) &= \frac{P_{W_i',W}\left(w', \frac{k}{n}\right)}{P_W(\frac{k}{n})} \\ &= \frac{P_{W_i',W}\left(w', \frac{k}{n}\right)}{B_{n,k}(\theta)}. \end{aligned} \tag{141}$$

Next, we discuss $P_{W_i',W}(w', \frac{k}{n})$ for different values of w'. When $w' = \frac{k-1}{n}$ with $1 \leq k \leq n$, it means $S_i = k-1$, $Z_i = 1$ and $Z_i' = 0$, where we recall $S_i := \sum_{j \neq i} Z_j$, then

$$\begin{aligned} P_{W_i',W}\left(\frac{k-1}{n}, \frac{k}{n}\right) &= \Pr\left[S_i = k-1, Z_i = 1, Z_i' = 0\right] \\ &= B_{n-1,k-1}(\theta)\theta(1-\theta). \end{aligned} \tag{142}$$

Similarly, when $w' = \frac{k+1}{n}$ with $0 \leq k \leq n-1$,

$$\begin{aligned} P_{W_i',W}\left(\frac{k+1}{n}, \frac{k}{n}\right) &= \Pr\left[S_i = k, Z_i = 0, Z_i' = 1\right] \\ &= B_{n-1,k}(\theta)\theta(1-\theta). \end{aligned} \tag{143}$$

Lastly, when $w' = \frac{k}{n}$ with $0 \leq k \leq n$,

$$\begin{aligned} P_{W_i',W}\left(\frac{k}{n}, \frac{k}{n}\right) &= \Pr\left[S_i = k, Z_i = 0, Z_i' = 0\right] + \Pr\left[S_i = k-1, Z_i = 1, Z_i' = 1\right] \\ &= B_{n-1,k}(\theta)(1-\theta)^2 + B_{n-1,k-1}(\theta)\theta^2. \end{aligned} \tag{144}$$

Combining these cases with (141) gives

$$\Pr\left[W_i' = w' | W = w\right] = \begin{cases} \theta w + (1-\theta)(1-w), & \text{if } w' = w, \\ \theta(1-w), & \text{if } w' = w + \frac{1}{n}, , \\ (1-\theta)w, & \text{if } w' = w - \frac{1}{n}. \end{cases} \tag{145}$$

substituting which into (140) yields

$$g(w, z) = \frac{2(w-z)(\theta-w)}{n} + \frac{\theta + w - 2\theta w}{n^2}. \tag{146}$$

Since $w \in [0,1]$ and $z \in \{0,1\}$, an admissible SCH-A stability parameter γ_1 is

$$\begin{aligned}\gamma_1 &= \sup_{w,z} |g(w,z)| \\ &\le \sup_{w,z} \left| \frac{2(w-z)(\theta-w)}{n} \right| + \sup_{w} \left| \frac{\theta + w - 2\theta w}{n^2} \right| \\ &= \frac{2\bar{\theta}}{n} + \frac{\bar{\theta}}{n^2},\end{aligned} \tag{147}$$

where $\bar{\theta} := \max\{\theta, 1-\theta\}$. Furthermore, the SCH-B stability parameter γ_2 can be exactly calculated as

$$\begin{aligned}\gamma_2^2 &= \mathbb{E}_W \mathbb{E}_Z[g(W,Z)^2] \\ &= \frac{4\theta^2(1-\theta)^2}{n^3} + \frac{8\theta^2(1-\theta)^2}{n^4} + \frac{\theta(1-\theta)(1-12\theta(1-\theta))}{n^5}.\end{aligned} \tag{148}$$

Thus, both stability conditions hold for Example 2, with $\gamma_1 \in \Theta(\frac{1}{n})$ and $\gamma_2 \in \Theta(n^{-\frac{3}{2}})$.

Evaluation of Wang & Mao (2023a, Theorem 3.1). Wang & Mao (2023a, Theorem 3.1) uses the γ_1-SCH-A stability as opposed to the β-strongly-uniform stability. The theorem states that

$$\overline{\text{gen}} \le \frac{\gamma_1}{n} \sum_{i=1}^{n} \sqrt{2I(W;Z_i)} \tag{149}$$

$$\overline{\text{gen}} \le \frac{\gamma_1}{n} \sum_{i=1}^{n} \sqrt{2I(W;Z_i|W_i')}, \tag{150}$$

where (149) uses the ISMI $I(W;Z_i)$, while (150) uses the CMI $I(W;Z_i|W_i')$ as in our work.

We now show that Example 2 provides a counterexample to (149). It was argued in Wang & Mao (2023a, Remark 2.2) that $\gamma_1 \le \beta$, and consequently (149) is tighter than (138). This implies (149) decays faster than $O(n^{-\frac{3}{2}})$, contradicting the exact order $\Theta(\frac{1}{n})$. Also, one can find the failure of (149) explicitly when directly substituting (147), which leads to an impossible rate $O(n^{-\frac{3}{2}})$. Either approach holds out the validity of (149). The failure of (149) implies that the looseness of the adopted information measure may be an essential cost to compensate for n-decaying stability parameters.

We further evaluate (150) under Example 2. By applying (147) and (111), we observe from numerical simulations that (150) holds true for all n. However, we engineer the following example where (150) fails to hold.

Example 4 (A memorization learning example)**.** *Define $\mathcal{Z} := [d]$ for an integer $d \ge 2$ and draw n samples i.i.d. from $\mathcal{Z}$, i.e., $Z_i \sim \text{Unif}(\mathcal{Z})$. Let the loss function be $\ell(w,z) = \mathbb{1}\{W(z) = 0\}$ and the learning algorithm output the count vector $W = (W(1), \dots, W(d))$, where $W(k) = \sum_{j=1}^{n} \mathbb{1}\{Z_j = k\}$ counts the number of each symbol $k \in \mathcal{Z}$.*

This memorization algorithm is an interpolating algorithm, in which case the empirical risk is exactly 0. A test sample Z incurs a unit loss precisely when it is absent in the training dataset, thus

$$\begin{aligned}\overline{\text{gen}} &= \mathbb{E}_W \mathbb{E}_Z \left[\ell(W,Z)\right] \\ &= \left(\frac{d-1}{d}\right)^n.\end{aligned} \tag{151}$$

In this example, we first calculate by the definition (140):

$$\begin{aligned}g(w,z) &= \mathbb{E}_{W_i'|W=w}[\ell(W_i',z)] - \ell(w,z) \\ &= \mathbb{E}_{W_i'|W=w}[\mathbb{1}\{W_i'(z) = 0\}] - \mathbb{1}\{w(z) = 0\} \\ &= \Pr\left[W_i'(z) = 0 | W = w\right] - \mathbb{1}\{w(z) = 0\}.\end{aligned} \tag{152}$$

Given (w,z), we discuss various cases:

- When $w(z) = 0$, we have

$$\begin{aligned} g(w,z) &= \Pr\left[Z_i' \neq z | W = w\right] - 1 \\ &= \frac{d-1}{d} - 1. \end{aligned} \tag{153}$$

- When $w(z) = 1$, we have

$$\begin{aligned} g(w,z) &= \Pr\left[Z_i' \neq z, Z_i = z | W = w\right] - 0 \\ &= \frac{d-1}{nd}. \end{aligned} \tag{154}$$

- When $w(z) \geq 2$, it is impossible to have $W_i'(z) = 0$, then $g(w,z) = 0$.

Together, we obtain

$$g(w,z) = \begin{cases} -\frac{1}{d}, & w(z) = 0, \\ \frac{d-1}{nd}, & w(z) = 1, \\ 0, & w(z) \geq 2, \end{cases} . \tag{155}$$

The SCH-A parameter is therefore given by

$$\begin{aligned} \gamma_1 &= \sup_{w,z} |g(w,z)| \\ &= \frac{1}{d} \max\left\{1, \frac{d-1}{n}\right\}. \end{aligned} \tag{156}$$

Due to the fact $I(W; Z_i | W_i') \leq H(Z_i) = \log d$, the claimed bound (150) would imply

$$\left(\frac{d-1}{d}\right)^n \leq \frac{1}{d} \max\left\{1, \frac{d-1}{n}\right\} \sqrt{2 \log d}. \tag{157}$$

Set $d = n + 1$ and let $n \to \infty$, then the LHS of (157) becomes

$$\lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right)^n = \frac{1}{e}, \tag{158}$$

whereas the RHS converges to 0, i.e.,

$$\lim_{n\to\infty} \frac{1}{n+1} \sqrt{2 \log(n+1)} = 0. \tag{159}$$

This contradiction proves that (150) also fails in general.

Evaluation of Wang & Mao (2023a, Theorem 3.2). For a deterministic algorithm, the stated bound reduces to Wang & Mao (2023a, Corollary 3.1). The Bernoulli estimator in Example 2 is deterministic, hence Example 2 provides a counterexample again and Wang & Mao (2023a, Theorem 3.2) does not hold.

Evaluation of Wang & Mao (2023a, Theorem 3.3). It is stated that

$$\overline{\text{gen}} \leq \frac{\gamma_1}{n} \sum_{i=1}^{n} I(W; Z_i) + 0.72 \frac{\gamma_2^2}{\gamma_1}. \tag{160}$$

For Example 2, one can verify that this bound decays with $\Theta(\frac{1}{n^2})$ by using the SCH stability parameters calculated in (147) and (148). This implies Wang & Mao (2023a, Theorem 3.3) does not hold.

Discussions of the erroneous proof. The incorrectness arises when the proofs from Wang & Mao (2023a, Section 3) identify the expected generalization error $\overline{\text{gen}}$ as $\mathbb{E}_{W,Z_i}[g(W, Z_i)]$. By the definition of g in (140), this identification requires

$$\mathbb{E}_{W,Z_i}\left[\mathbb{E}_{W_i'|W}\left[\ell(W_i', Z_i)\right]\right] = \mathbb{E}_W \mathbb{E}_{Z_i}\left[\ell(W, Z_i)\right]. \tag{161}$$

Although W_i' and Z_i are independent, the inner expectation on the left uses $P_{W_i'|W}$ rather than $P_{W_i'|W,Z_i}$. These conditional laws need not agree.

The Markov condition $W_i' - W - Z_i$ would suffice to justify the substitution, but it is not in general guaranteed by the replace-one construction in Section 4.1. To see this, let $Z_{[n]\setminus i} := (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$ denote the leave-one-out training set and consider a deterministic algorithm for simplicity. Then, the dependence structure is represented by the following Bayesian network:

$$Z_i \to W \leftarrow Z_{[n]\setminus i} \to W_i' \leftarrow Z_i' \tag{162}$$

One can verify via the *d-separation technique* (Neuberg, 2003) that W_i' and Z_i are not independent when conditioned on W. This implies the Markov chain $W_i' - W - Z_i$ does not hold as desired.

Example 2 makes the incorrectness in (161) explicit. Specifically, direct calculations follow from (146), yielding

$$\mathbb{E}_{W,Z_i}\left[g(W, Z_i)\right] = \frac{2\theta(1-\theta)}{n^2}, \qquad \overline{\mathrm{gen}} = \frac{2\theta(1-\theta)}{n}. \tag{163}$$

It is therefore observed that (161) does not hold. The proofs from Wang & Mao (2023a, Section 3) inherit this incorrectness which, in our view, is possibly why Wang & Mao (2023a, Theorem 3.1, 3.2, 3.3 and Corollary 3.1) do not hold as stated.