A UNIFIED FRAMEWORK FOR CHANGE OF MEASURE INEQUALITIES: APPLICATIONS TO GENERALIZATION, MEMORIZATION AND PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Change of measure inequalities translate divergences between probability measures into explicit bounds on event probabilities, and play an important role in deriving probabilistic guarantees in learning theory, information theory, and statistics. We propose novel change of measure inequalities via a unified framework based on the data processing inequality, which is surprisingly elementary yet powerful enough to yield novel, tighter inequalities. We provide change of measure inequalities in terms of a broad family of information measures, including f-divergences (with Kullback-Leibler divergence and χ^2-divergence as special cases), Rényi divergence, and α-mutual information (with maximal leakage as a special case). We apply these results to generalization error analysis, PAC-Bayesian theory, differential privacy, and data memorization, obtaining stronger guarantees while recovering best-known results through simplified analyses.

1 INTRODUCTION

In information theory and learning theory, a recurring task is to bound the probability $P(E)$ of an event E in terms of $Q(E)$ and some divergence between P and Q. That is, for a measurable set $E \in \mathcal{F}$ and probability measures P, Q on a common measurable space $(\mathcal{X}, \mathcal{F})$ such that $P \ll Q$[1], we aim to establish an inequality of the form

$$P(E) \leq \xi\big(Q(E), \mathrm{d}P/\mathrm{d}Q\big), \tag{1}$$

which are referred to as "change of measure inequalities" and serve as a bridge between divergences and probabilistic guarantees. Their applications span a broad range of problems, including generalization error bounds of learning algorithms (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Chu & Raginsky, 2023), hypothesis testing (Polyanskiy & Wu, 2025), distributed detection (Chen & Wang, 2019), and Monte Carlo estimation (Ohnishi & Honorio, 2021). A classical example is the strong converse lemma (Polyanskiy & Wu, 2025):

$$P(E) \leq \gamma Q(E) + P\left(\mathrm{d}P/\mathrm{d}Q > \gamma\right), \quad \forall \gamma \geq 0, \tag{2}$$

which is well known in information theory and statistics (Cover, 1999; Polyanskiy & Wu, 2025).

Recent research has increasingly focused on change of measure inequalities and their applications, with several key studies establishing frameworks for deriving such inequalities. Notably, Hellström & Durisi (2020) applied the strong converse lemma (2) to derive a wide range of generalization error bounds. This line of work was further extended by (Ohnishi & Honorio, 2021; Picard-Weibel & Guedj, 2022), which utilized variational representations of f-divergences to obtain tighter change of measure inequalities. In (Esposito et al., 2021), change of measure inequalities were used to derive novel results on generalization error and privacy analyses.

While these contributions significantly tightened the resultant bounds, they remain largely isolated without a clear connective bridge, and the strong converse lemma used in Hellström & Durisi (2020) was not involved in the other preceding literature. The results in Esposito et al. (2021) are convenient

[1] $P \ll Q$ denotes that P is absolutely continuous with respect to Q, and $\mathrm{d}P/\mathrm{d}Q$ denotes the Radon-Nikodym derivative.

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.

to use, but not as tight as those in Picard-Weibel & Guedj (2022); Ohnishi & Honorio (2021), which, in some cases, rely on additional structural assumptions such as strict convexity. Though representing the sharpest known bounds, the results in Picard-Weibel & Guedj (2022) involve auxiliary parameters that are nontrivial to optimize. Most importantly, the literature currently lacks a unified framework that integrates existing methodologies to provide a comprehensive analysis.

In this work, we propose a unified framework for deriving change of measure inequalities via a single elementary tool: the *data processing inequality (DPI)* for f-divergences (Sason & Verdú, 2016). Although the DPI for f-divergences has been used in previous research, its optimality in bounding event probabilities via change of measure inequalities remains elusive. In this paper, we unify existing approaches via the application of DPI through an indicator channel, which is surprisingly elementary yet powerful enough to yield novel inequalities that are usually tighter than existing ones. We further establish the optimality of this DPI approach for deriving change of measure inequalities in the form of (1). Besides being convenient to use, one advantage of our approach is its flexibility in choosing the error event E in (1), which enables our framework to be applied in diverse problems, e.g., generalization analysis (Xu & Raginsky, 2017) and data memorization (Attias et al., 2024; Feldman et al., 2025). A more detailed literature review is in Appendix A.

2 Our Contribution

We first present the key technique of our paper, demonstrate its use by deriving an tighter change of measure inequality compared with (2), and then summarize our contributions.

According to DPI, given any "channel" T from $\mathcal{X}$ to $\mathcal{Y}$, applying the same T to probability measures P, Q on $\mathcal{X}$ cannot increase their f-divergence, that is,

$$D_f(T \circ P \| T \circ Q) \le D_f(P\|Q).$$

Now, if we apply DPI for the indicator channel $T = \mathbb{1}_E$, and let $p := P(E), q := Q(E)$, we obtain

$$D_f(P\|Q) \ge D_f(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) = D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) = q f\Big(\frac{p}{q}\Big) + \big(1-q\big) f\Big(\frac{1-p}{1-q}\Big). \quad (3)$$

Once f is specified, the upper bound on p relies on expressing the right hand side of (3) in terms of p. This single inequality is the main tool behind all of our results.

To illustrate the event-DPI calculation in its simplest form, consider the E_γ-divergence (5) (a.k.a. the hockey-stick divergence (Sason & Verdú, 2016)). The following inequality is a standard consequence of its variational definition; we state it separately because it will be used repeatedly and because the strictly-convex framework of Picard-Weibel & Guedj (2022) does not cover this case.

Proposition 2.1. *Fix probability measures P, Q on $\mathcal{X}$ such that $P \ll Q$. For all measurable E,*

$$P(E) \le \gamma Q(E) + E_\gamma(P\|Q), \quad \forall \gamma \ge 1. \quad (4)$$

Proof Sketch. E_γ is an f-divergence with $f(t) = [t-\gamma]_+$. Let $p := P(E), q := Q(E)$. By (3), $E_\gamma(P\|Q) \ge E_\gamma(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) = q \cdot [p/q - \gamma]_+ + (1-q) \cdot [(1-p)/(1-q) - \gamma]_+, \ \forall \gamma \ge 1.$ Since $(1-q)[\cdot]_+ \ge 0$, it follows that $E_\gamma(P\|Q) \ge q(p/q - \gamma)$, and rearrangement yields (4). □

For the same $\gamma \ge 1$, Proposition 2.1 is pointwise no worse than the strong converse lemma (2):

$$E_\gamma(P\|Q) = P(\mathrm{d}P/\mathrm{d}Q > \gamma) - \gamma Q(\mathrm{d}P/\mathrm{d}Q > \gamma) \le P(\mathrm{d}P/\mathrm{d}Q > \gamma).$$

The inequality is strict whenever $Q(\mathrm{d}P/\mathrm{d}Q > \gamma) > 0$. Proposition 2.1 itself is not a novelty claim: its variational form is standard, in particular in differential privacy. Its role here is to illustrate our DPI viewpoint and to provide a sharper replacement whenever the strong-converse relaxation is used.

We then summarize our contributions from various aspects as follows.

Change of Measure Inequalities. We employ the machinery of (3) together with other techniques to derive change of measure inequalities in terms of a large family of information measures. Our methods significantly simplify the analysis without sacrificing the tightness. Our results recover, and usually improve, existing inequalities in terms of the same information measure (see Table 1), which seems too good to be true at first glance. However, we can prove the *optimality* of our framework under reasonable assumptions, and we show how to recover existing methods (Picard-Weibel & Guedj, 2022; Esposito et al., 2021), not just their results, through the simplified routes in Section 4.1.

Applications The generality of our result yields strong bounds across a diverse spectrum of fundamental problems in learning theory and other areas. Our framework provides a unified machinery for these problems by specifying P, Q, and the event E according to the scenario. For example, when analyzing the generalization error of a learning algorithm $P_{W|S}$, we specify P as the joint distribution P_{SW}, Q as the product distribution $P_S P_W$, and E as the event that the error exceeds a threshold, which gives a path to translate change of measure inequalities into generalization bounds.

- **Generalization Analysis.** In Section 5, we translate the event bounds into high-probability and PAC-Bayesian guarantees. Some consequences recover known results, while retained-complement refinements give strictly smaller bounds in concrete regimes. We add an exactly enumerable Gibbs learner and a Gaussian PAC-Bayesian example; the latter also exhibits an asymptotic scaling separation when the posterior displacement grows with n.
- **Differential Privacy.** The equivalence between approximate DP and hockey-stick event inequalities is classical (Barthe et al., 2013; Balle et al., 2020; Kairouz et al., 2015). We use this established connection, together with approximate max-information (Rogers et al., 2016), to transfer approximate-DP control into our event-level generalization bound and state explicitly the regime in which the resulting guarantee is non-vacuous.
- **Data Memorization.** We apply the same event-level reduction to selector recovery (Sefidgaran et al., 2025). Optimizing the Bernoulli-KL step removes the $\Theta(1/n)$ Fano slack for bounded CMI and gives a strict rate improvement when the CMI vanishes; it also strengthens the resulting false-positive conclusion in Theorem 7.1.

Notation We use calligraphic letters (e.g., $\mathcal{W}$), capital letters (e.g., W), and lower-case letters (e.g., w) to denote sets, random variables, and instances, respectively. We assume logarithms are base e, $[x]_+ := \max\{x, 0\}$, $\mathbb{1}_{\{S\}}$ is an indicator of a statement S, and $1/\infty = 0$. When distributions P, Q are discussed together, we assume they are on a measurable space $(\mathcal{X}, \mathcal{F})$ where $\mathcal{F}$ is suppressed.

3 Information Measures

Information measures are used to quantify the dissimilarity between distributions or the correlation between random variables. One of the most general notion is the f-divergence:

Definition 3.1. Let $f : [0, \infty) \to \mathbb{R}$ be convex, $f(1) = 0$ and $f(0) = \lim_{t \downarrow 0} f(t)$. For $P \ll Q$, let

$$D_f(P\|Q) := \int f\left(\mathrm{d}P/\mathrm{d}Q\right) \mathrm{d}Q.$$

f-divergences (Csiszár, 1967; 1963; Ali & Silvey, 1966; Sason & Verdú, 2016) include most well-known information measures as special cases, e.g., KL divergence and TV distance, and possess desirable properties such as non-negativity, joint convexity, and the data processing inequality.

E_γ-divergence is of special interests due to Proposition 2.1 and its use across our paper. It is defined:

$$E_\gamma(P\|Q) := \sup_E \big(P(E) - \gamma Q(E)\big) = \int \left([\mathrm{d}P/\mathrm{d}Q - \gamma]_+\right) \mathrm{d}Q, \qquad \gamma \geq 1. \tag{5}$$

Definition 3.2. Let $\alpha \in (0, \infty) \setminus \{1\}$ and $P \ll Q$, *Rényi divergence of order* α is given by

$$D_\alpha(P\|Q) := 1/(\alpha - 1) \cdot \log \int \left(\mathrm{d}P/\mathrm{d}Q\right)^\alpha \mathrm{d}Q. \tag{6}$$

Note that power divergence $\mathcal{H}_\beta(P\|Q)$ has a one-to-one transformation with $D_\alpha(P\|Q)$, by

$$D_\alpha(P\|Q) = \log\big(1 + (\alpha - 1)\mathcal{H}_\alpha(P\|Q)\big)/(\alpha - 1). \tag{7}$$

Another important information measure is the Sibson α-*mutual information* $I_\alpha(S, W)$, initiated from the "information radius" (Sibson, 1969) and revisited by Verdú (2015). It is defined as follows.

Definition 3.3. Let $S, W \sim P_{SW}$ and Q_W be a probability measure on $\mathcal{W}$, α-*mutual information* is

$$I_\alpha(S, W) := \min_{Q_W} D_\alpha\big(P_{SW}\|P_S Q_W\big), \quad \alpha > 0. \tag{8}$$

Taking $\alpha \to 1$ recovers $I(S; W)$ and $\alpha \to \infty$ recovers *maximal leakage* (that will be used in (16)):

$$\mathcal{L}(S \to W) = \sup_{U - S - W - \hat{U}} \log\Big(\mathbf{P}(U = \hat{U}) \big/ \max_{u \in \mathcal{U}} \mathbf{P}_U(u)\Big), \tag{9}$$

where the supremum is over all U and $\hat{U}$ taking values in the same finite alphabet (Issa et al., 2016).

4 CHANGE OF MEASURE INEQUALITIES

We have briefly explained our main technique in Section 2, namely the *data processing inequality (DPI)* for f-divergences (Polyanskiy & Wu, 2025, Section 7.2) that is surprisingly elementary yet remarkably powerful for yielding strong inequalities. Here we present it again for completeness.

$$D_f(T \circ P \| T \circ Q) \le D_f(P\|Q).$$

To derive change of measure inequalities of form (1), we consider the deterministic indicator channel $T = \mathbb{1}_E$ (see Section 4.1 for the intuition behind). Denoting $p := P(E)$ and $q := Q(E)$, we observe

$$D_f(P\|Q) \ge D_f(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) = D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) = q f\Big(\frac{p}{q}\Big) + \big(1-q\big) f\Big(\frac{1-p}{1-q}\Big),$$

and the machinery of proof we use is simply to specify f and calculate the last term of (3).

We have used Proposition 2.1 to demonstrate how to use our machinery to derive change of measure inequalities of form (1). By the same pipeline, we derive change of measure inequalities in terms of various information measures in Theorem 4.1; the proof is in Appendix C.

Theorem 4.1. *For probability measures P, Q on $\mathcal{X}$ such that $P \ll Q$, for all measurable E,*

$$P(E) \le \big(D_{\mathrm{KL}}(P\|Q) + \log(1 + Q(E)(e^c - 1))\big)/c, \quad c > 0, \tag{10}$$

$$P(E) \le Q(E) + \sqrt{Q(E)(1-Q(E))\chi^2(P\|Q)}, \tag{11}$$

$$2\left(1 - \sqrt{P(E)Q(E)} - \sqrt{(1-P(E))(1-Q(E))}\right) \le H^2(P;Q), \tag{12}$$

$$P(E)^\beta Q(E)^{1-\beta} + (1-P(E))^\beta (1-Q(E))^{1-\beta} \le 1 + (\beta-1)\mathcal{H}_\beta(P\|Q). \tag{13}$$

Equations (10)–(11) are known: (10) appears in Issa et al. (2019), and (11) is the standard binary-DPI/Cauchy–Schwarz bound (see also Jiao et al. (2017)). The new content we emphasize is the sharp-envelope interpretation below and explicit refinements obtained by retaining the complementary-event term, e.g., (25), which recovers Esposito et al. (2021, Corollary 7) after that term is dropped. One can derive cleaner forms by further relaxations, see Appendix C. By (7), we derive inequalities in terms of the Rényi divergence (6). In Table 1, we list the results from our DPI approach, for typical f-divergences, and explicitly compare with (Picard-Weibel & Guedj, 2022) in Appendix L.

For a complete, σ-finite probability space $(\Omega, \mathcal{F}, \mu)$, for U that is $\mathcal{F}$-measurable, with respect to μ, the *Luxemburg norm* of U is defined as $\|U\|_\psi^\mu := \inf\big\{\sigma > 0 : \mathbf{E}_\mu\big[\psi\big(|U|/\sigma\big)\big] \le 1\big\}$ and the *Amemiya norm* of U is defined as $\|U\|_\psi^{A,\mu} := \inf\big\{\big(\mathbf{E}_\mu\big[\psi\big(t|U|\big)\big] + 1\big)/t : t > 0\big\}$. Recall for a convex ψ, its *conjugate* $\psi^\star : [0,\infty) \to \mathbb{R}$ is defined as $\psi^\star(t) = \sup_{\lambda>0} \lambda t - \psi(\lambda)$. An *Orlicz function* ψ is a convex function $\psi : [0,\infty) \to [0,\infty]$ that vanishes at zero and is not identically 0 or ∞ on $(0,\infty)$, and its generalized *inverse* is defined as $\psi^{-1}(s) := \inf\big\{t \ge 0 : \psi(t) \ge s\big\}$ for $s \ge 0$.

Theorem 4.2. *Fix P, Q such that $P \ll Q$. For all measurable E, fix an Orlicz function ψ,*

$$P(E) \le \gamma Q(E) + \frac{1}{\psi^{-1}(1/Q(E))} \cdot \left\| \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right]_+ \right\|_{\psi^\star}^{A,Q}, \quad \forall \gamma \in \mathbb{R}. \tag{14}$$

See Appendix D for a proof and discussions.

4.1 DISCUSSIONS

The idea behind the DPI scheme is: the discrepancy between $P(E)$ and $Q(E)$ depends on two ingredients: (i) the choice of event E; (ii) the statistical dissimilarity between P and Q. For example, if $E = \Omega$, then the event selection eliminates any observable difference. We interpret $P(E)$ and $Q(E)$ as the output probabilities induced by passing P and Q through the same kernel $T = \mathbb{1}_E$, and the difference between $P(E)$ and $Q(E)$ can then be quantified by information measures between the induced laws $P \circ T$ and $Q \circ T$, implying (3). Though the DPI was also used in a step of (Esposito et al., 2021, Theorem 3), given the explicit expression of f, one can provide a tighter bound on $f((1-p)/(1-q))$ than $-f^*(0)$ (see (15)), which is one of the motivations for Theorem 4.1.

Using DPI to an event indicator is not entirely new. For example, Feldman & Steinke (2018, Lemma 3.14) use an indicator reduction in adaptive data analysis, and Esposito et al. (2021, Theorem 3) use f-divergence DPI for event probabilities. Our point is to identify the exact Bernoulli

Table 1: Our change of measure inequalities in terms of typical f-divergences via DPI, all of which are never worse, and usually tighter, than **best-known** inequalities in literature (Picard-Weibel & Guedj, 2022; Esposito et al., 2021; Ohnishi & Honorio, 2021).

f-div	$f(t)$	Change of Measure Inequalities via DPI	Tighter?
E_γ-div, $\gamma \geq 1$	$[t-\gamma]_+$	$p \leq \gamma\, q + E_\gamma(P\|Q)$	—
KL	$t \log t$	$p \leq \big(D_{\mathrm{KL}}(P\|Q) + \log\big(1 + q(e^c - 1)\big)\big)/c,\ c > 0$	same
χ^2-div	$t^2 - 1$	$p \leq q + \sqrt{q\big(1-q\big)\chi^2(P\|Q)}$	same
Power-β, $\beta > 1$	$(t^\beta - 1)/(\beta - 1)$	$p^\beta q^{1-\beta} + (1-p)^\beta (1-q)^{1-\beta} \leq 1 + (\beta-1)\mathcal{H}_\beta(P\|Q)$	✓
Squared Hellinger	$(1-\sqrt{t})^2$	$2\Big(1 - \sqrt{pq} - \sqrt{(1-p)(1-q)}\Big) \leq H^2(P;Q)$	✓
Reverse χ^2-div	$1/t - 1$	$(p-q)^2/(p(1-p)) \leq \chi^2(Q\|P)$	✓
Reverse-KL	$-\log t$	$q\log(\frac{q}{p}) + (1-q)\log(\frac{1-q}{1-p}) \leq D_{\mathrm{KL}}(Q\|P)$	✓
Jensen-Shannon	$\log \frac{2^{t+1}t^t}{(1+t)^{t+1}}$	$2h_2\left((p+q)/2\right) - h_2\big(p\big) - h_2\big(q\big) \leq \mathrm{JS}(P\|Q)$	✓
Vincze-Le Cam	$(2-2t)/(t+1)$	$2(p-q)^2/\big((p+q)\big(2-p-q\big)\big) \leq \mathrm{VC}(P;Q)$	✓

divergence as the central event-level object and to state precisely the scope in which it is optimal. When applied to bound $P(E)$ by $Q(E)$ and $D_f(P\|Q)$ of some measurable E, DPI is tighter than the relaxation of the Donsker-Varadhan representation

$$D_f(P\|Q) = \sup_{T \in \mathcal{T}} \left\{\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)]\right\},$$

and it is the best way to perform: To bound $P(E)$, T is restricted to $\mathcal{T}_E := \big\{T :\ T = a\mathbb{1}_E + b\mathbb{1}_{E^c} \text{ for } a, b \in \mathbb{R}\big\}$. In this case, for every measurable E,

$$\sup_{T \in \mathcal{T}_E} \left\{\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)]\right\} = D_f\big(\mathrm{Ber}(P(E))\|\mathrm{Ber}(Q(E))\big),$$

i.e., to compare P, Q via E, optimizing over the admissible test functions gives *exactly* Bernoulli divergence. Hence, DPI with $x \mapsto \mathbb{1}_E(x)$ gives the *best possible* result. The inequality here comes from the loss of information from the indicator mapping $\mathbb{1}(E)$ in measuring $P(E)$, instead of the relaxation of $\mathbf{E}_Q[f^*(T)]$, see Appendix M for details. We recover existing approaches as follows:

- **Recovering (Picard-Weibel & Guedj, 2022)**: the DPI (3) gives
$$G_f(p,q) := qf(p/q) + \big(1-q\big)f((1-p)/(1-q)) \leq D_f(P\|Q),$$
and applying Fenchel duality, $f(x) = \sup_{u\in\mathbb{R}}\{ux - f^*(u)\}$, to the two terms of $G_f(p,q)$:
$$G_f(p,q) = \sup_{u,v} \big\{p(u-v) + v - qf^*(u) - (1-q)f^*(v)\big\}.$$
Let $u = \lambda + c$ and $v = c$ for $\lambda > 0$ and $c \in \mathbb{R}$, we recover (Picard-Weibel & Guedj, 2022) on $\mathbb{1}_E$:
$$p \leq \inf_{u>v} \big(D_f(P\|Q) - v + qf^*(u) + (1-q)f^*(v)\big)/(u-v).$$
Comparing to (Picard-Weibel & Guedj, 2022), our scheme is simpler and $G_f(p,q)$ is easier to control, which gives us a better control over the inequalities and eliminates the parameters in (Picard-Weibel & Guedj, 2022, Table 1) that can be hard to optimized, hence gives tighter inequalities. Their f has to be strictly convex, hence they cannot recover our Proposition 2.1.
- **Recovering (Esposito et al., 2021)**: We rearrange (3) and recover (Esposito et al., 2021, Theorem 3)
$$\begin{aligned} f\big(p/q\big) &\leq \big(D_f(P\|Q) - (1-q)f\big((1-p)/(1-q)\big)\big)/q \\ &\leq \big(D_f(P\|Q) + (1-q)f^*(0)\big)/q, \end{aligned} \tag{15}$$
where the last inequality is by $f^*(0) = -\inf_{t\geq 0} f(t) \geq -f\big((1-p)/(1-q)\big)$.

Compared to (Guan et al., 2025) which uses DPI for PAC-Bayesian bounds only, our bounds are usually tighter than theirs (e.g., compare (25) to (Guan et al., 2025, Lemma 2)). Moreover, our results can be applied to σ-sub-Gaussian loss, instead of their bounded $[0,1]$-loss. Most importantly, the implication of our study is broader: we establish a unified framework that is applicable to generalization, privacy and memorization problems.

5 GENERALIZATION ERROR BOUNDS

5.1 GENERALIZATION ERROR BOUNDS VIA f-DIVERGENCE

Consider a stochastic learning algorithm $P_{W|S}$ as a probabilistic mapping from a training dataset $S := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, where $Z_1, \ldots, Z_n \sim P_Z$ i.i.d., to hypothesis $W \in \mathcal{W}$. With loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, the generalization performance of $P_{W|S}$ is measured by its *generalization error*

$$\mathrm{gen}(S, W) := \mathbf{E}_{P_Z}\left[\ell(Z, W)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, W),$$

i.e., the gap between the expected loss $\mathbf{E}_{P_Z}\left[\ell(Z, W)\right]$ and the empirical loss $1/n \cdot \sum_{i=1}^{n} \ell(Z_i, W)$.

Started from (Russo & Zou, 2016; Xu & Raginsky, 2017) which bounded the expected generalization error by the mutual information $I(S; W)$ between training data and hypothesis, various types of bounds were subsequently proposed (Asadi et al., 2018; Steinke & Zakynthinou, 2020; Sefidgaran et al., 2022; Chu & Raginsky, 2023; Aminian et al., 2024). We consider high-probability generalization bounds (Esposito et al., 2021; Hellström & Durisi, 2020), e.g.,

$$\mathbf{P}\left(\left|\mathrm{gen}(S, W)\right| \geq \eta\right) \leq 2 \cdot \exp\left(\mathcal{L}(S \to W) - n\eta^2/(2\sigma^2)\right), \tag{16}$$

where $\mathcal{L}(S \to W)$ is the maximal leakage. We assume, uniformly over $w \in \mathcal{W}$, $\ell(w, Z) - \mathbf{E}[\ell(w, Z)]$ is σ-sub-Gaussian under $Z \sim P_Z$; hence for fixed w, $\mathbf{P}_S(|\mathrm{gen}(S, w)| \geq \eta) \leq 2\exp(-n\eta^2/(2\sigma^2))$.

Our change of measure inequalities in Table 1 can be converted to this class of guarantees by taking

$$P := P_{SW}, \quad Q := P_W P_S, \quad E = \{W, S : |\mathrm{gen}(S, W)| \geq \epsilon\}. \tag{17}$$

Take some common f-divergences as examples, we derive the following generalization bounds.

Theorem 5.1. *Fix $\gamma \in \mathbb{R}$. Let $P \ll Q$, if convex $f : (0, \infty) \to \mathbb{R}$ satisfies $f(1) = 0$, $f'(\gamma) > f'(1)$ for $\gamma > 1$, then for all $\eta > 0$, let $P = P_{SW}$ and $Q = P_S P_W$, we have $\mathbf{P}\left(|\mathrm{gen}(S, W)| \geq \eta\right) \leq$*

$$\min\left\{\begin{array}{l} \gamma\zeta + E_\gamma(P\|Q) \\ \sqrt{\zeta(1-\zeta)\cdot\chi^2(P\|Q)} + \zeta \\ \sqrt{1-\zeta}\cdot\sqrt{1-\left(1-H^2(P;Q)/2\right)^2}^2 + \left(\left(1-H^2(P;Q)/2\right)\sqrt{\zeta}\right. \\ \left(\zeta^{\beta-1}\left[1+(\beta-1)\mathcal{H}_\beta(P\|Q) - (1-\zeta)^{1-\beta}(1-u_0)^\beta\right]_+\right)^{1/\beta} \end{array}\right\}. \tag{18}$$

where $u_0 := \min\left\{1,\ \left((1+(\beta-1)\mathcal{H}_\beta(P\|Q))\, q^{\beta-1}\right)^{1/\beta}\right\}$ and $\zeta := 2\exp\left(-n\eta^2/(2\sigma^2)\right)$.[2]

Theorem 5.1 intentionally contains both recoveries and refinements. One purpose of deriving generalization bounds in terms of different measures is that, in different applications, different information measures have desirable properties in different senses, e.g., $\chi^2(P\|Q)$ is sometimes significantly smaller than $\mathcal{L}(S \to W)$ since $\chi^2\left(P_{SW}\|P_S P_W\right) \leq \exp\left(\mathcal{L}(S \to W)\right) - 1$. Appendix I gives Gaussian/Laplace comparisons, and Section 5.3 below gives learning-algorithm examples.

Although we focus on high-probability bounds, we can also derive average generalization bounds via change of measure inequalities if they are of interest. For example, Proposition 2.1 recovers the average bound of (Xu & Raginsky, 2017, Theorem 1), up to a multiplicative constant:

$$\mathbf{E}\left[|\mathrm{gen}(S, W)|\right] \leq \left((2\sigma)/\sqrt{n}\right) \cdot \left(2\sqrt{I(S;W) + 2/e} + \sqrt{\pi}\right),$$

which is tighter than the one recovered by (Chu & Raginsky, 2023, Corollary 1); see Appendix J.

Generalization bounds via maximal leakage and α-mutual information can be found in Appendix E, while conditional mutual information bounds can be found in Appendix G.

5.2 PAC-BAYESIAN BOUNDS

Probably approximately correct (PAC)-Bayesian bounds (McAllester, 1998; 1999; Catoni, 2004; 2007) are another important class of generalization bounds, which have witnessed a surge of interest in recent years (Pérez-Ortiz et al., 2021; Guedj & Pujol, 2021; Biggs & Guedj, 2022; Guedj, 2019;

[2] We upper bound $Q(E)$ by the Hoeffding's inequality on the σ-sub-Gaussian loss function and hence have ζ.

Alquier, 2024). In (Hellström & Durisi, 2020; Ohnishi & Honorio, 2021; Picard-Weibel & Guedj, 2022; Guan et al., 2025), change of measure inequalities have been used to derive PAC-Bayesian bounds. In this setting, each time the algorithm is used, a new hypothesis W is drawn from $P_{W|S}$, and we aim to find bounds of the form $\mathbf{P}_{P_S}\big(\mathbf{E}_{P_{W|S}}\big[\mathrm{gen}(S,W)\big] \le \epsilon\big) \ge 1-\delta$.

We take the power-β relation (13) as an example. The resulting PAC-Bayesian bound is not uniformly better than prior bounds, but Section 5.3 identifies explicit finite-sample and asymptotic regimes where it is strictly tighter. The proof is in Appendix F.

Corollary 5.2. *For a learning algorithm $P_{W|S}$ with a σ-sub-Gaussian loss function, denoting $\mathsf{H}_\beta := \mathcal{H}_\beta(P_{W|S}\|P_W)$, the following holds with P_S-probability at least $1-\delta$:*

$$\mathbf{E}_{P_{W|S}}[\mathrm{gen}(S,W)] \le \sqrt{\frac{2\sigma^2}{n}}\left(\log 2 + \log\sqrt{\frac{\pi\beta}{\beta-1}} + \frac{\beta}{4(\beta-1)} + \log\left(\frac{((\beta-1)\mathsf{H}_\beta+1)^{1/\beta}}{\delta}\right)\right).$$

For fixed $\beta > 1$, the bound scales as $\mathcal{O}\big(\frac{\sigma}{\sqrt{n}}\big[1 + \log\frac{1}{\delta} + \frac{1}{\beta}\log\big(1+(\beta-1)\mathsf{H}_\beta\big)\big]\big)$. Thus its dependence on the power-divergence complexity is logarithmic. Section 5.3 below gives a same-divergence Gaussian comparison with Ohnishi & Honorio (2021) and shows that this difference yields a genuine asymptotic separation when the posterior displacement grows with n.

Moreover, our framework also applies to the conditional mutual information (CMI) framework for generalization (Steinke & Zakynthinou, 2020), which remains finite and usually provides more stable generalization guarantees. Novel CMI bounds are deferred to Appendix G due to space limitations.

5.3 Concrete Examples

We give two quantitative examples to show the improvements from our bounds are meaningful.

Example 1: Gibbs/exponential-weights learner. Let $Z_1,\dots,Z_n$ be i.i.d. Rademacher variables, let $\mathcal{W} = \{-1,-0.95,\dots,0.95,1\}$, and take $\ell(w,z) = (w-z)^2/4 \in [0,1]$. For $n = 50$, consider $P_{W|S}(w|s) \propto \exp\{-n\widehat{R}_s(w)\}$. Writing $\bar{Z} = n^{-1}\sum_i Z_i$ gives $\mathrm{gen}(S,w) = w\bar{Z}/2$; all event probabilities and divergences can be evaluated exactly by summing over the binomial statistic. For each threshold η, the hockey-stick bound $\inf_{\gamma\ge1}\{\gamma q_\eta + E_\gamma(P\|Q)\}$ is optimized separately.

At $\eta = 0.10, 0.15, 0.20, 0.25$, the optimized hockey-stick bound is respectively $28.1\%, 28.4\%, 28.2\%, 27.5\%$ smaller than the optimized strong-converse bound; the corresponding optimizers are $6.8431, 19.0091, 62.6511, 179.7890$. At $\eta = 0.20$, our exact χ^2 inversion is 42.5% smaller than the existing same-divergence bound, and the retained-complement power-1.5 bound improves Esposito et al. (2021, Corollary 7) by 71.7%.

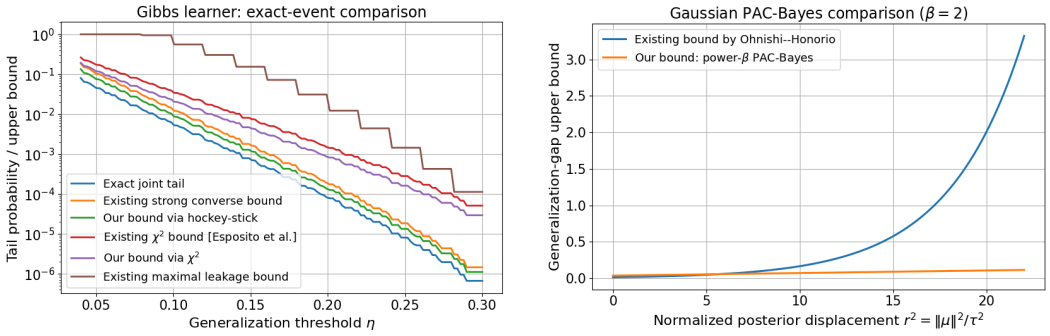
Example 2: Gaussian power-β PAC-Bayes. Take $P_W = \mathcal{N}(0,\tau^2 I), P_{W|S} = \mathcal{N}(\mu,\tau^2 I), r^2 := \|\mu\|^2/\tau^2$. Then we can calculate $D_\beta(P_{W|S}\|P_W) = \frac{\beta r^2}{2}$ and $1+(\beta-1)\mathsf{H}_\beta = \exp\big(\beta(\beta-1)r^2/2\big)$.

For $\beta = 2$, $n = 10000$, $\sigma = 0.5$, and $\delta = 0.05$, Corollary 5.2 crosses the power-divergence PAC-Bayes bound of Ohnishi & Honorio (2021) at $r^2 \simeq 5.68$. At $r^2 = 8, 10, 15, 20$, our bound is respectively $35.8\%, 56.8\%, 84.6\%, 94.7\%$ smaller.

Importantly, this example yields an asymptotic separation. For fixed $\beta > 1$ and δ, Corollary 5.2 gives $\sqrt{\frac{2\sigma^2}{n}}\big(C_{\beta,\delta} + \frac{\beta-1}{2}r_n^2\big)$ where $C_{\beta,\delta} := \log 2 + \frac{1}{2}\log\frac{\pi\beta}{\beta-1} + \frac{\beta}{4(\beta-1)} + \log(1/\delta)$. The corresponding sub-Gaussian power-divergence bound of Ohnishi & Honorio (2021), specialized to the same Gaussian pair, scales as $\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\,\exp\big(\frac{\beta-1}{4}r_n^2\big)$. So if $r_n^2 = \rho\log n$ for a fixed $\rho > 0$, our bound is of order $O(\log n/\sqrt{n})$ while theirs is of order $\Theta(n^{-1/2+(\beta-1)\rho/4})$.

6 Generalization and Differential Privacy

Differential privacy (Dwork et al., 2006; 2014) is one of the most celebrated privacy measures. An algorithm $\mathcal{A} := P_{W|S}$ is said to satisfy (ε,δ)-differential privacy if, for every pair of *neighboring* datasets $s, s' \in \mathcal{S}$ that differ in exactly one coordinate, and for every measurable set $\mathcal{V} \subseteq \mathcal{W}$, we have



(a) Gibbs learner. (b) Gaussian PAC-Bayes.

Figure 1: Concrete comparisons. Left: exact-event Gibbs comparison. Right: Gaussian PAC-Bayes comparison for $\beta = 2$, $n = 10{,}000$, $\sigma = 0.5$, and $\delta = 0.05$.

$$\Pr\left(\mathcal{A}(s) \in \mathcal{V}\right) \leq e^{\varepsilon} \Pr\left(\mathcal{A}(s') \in \mathcal{V}\right) + \delta.$$

Differential privacy (DP) is tightly related to generalization analysis since private algorithms leak little information, are relatively "stable" (Dwork et al., 2015a;b), and generalize well.

For the case of $\delta = 0$, it has been shown that the generalization error can be readily bounded (Dwork et al., 2015a;b; Rodríguez-Gálvez et al., 2021; Liu et al., 2026). It becomes more tricky and difficult for the case when $\delta > 0$. As a relaxation of max-information defined by Dwork et al. (2015a), *approximate max-information* is proposed by (Rogers et al., 2016) to capture the generalization.

Definition 6.1. The τ-approximate max-information between S and W, is defined as

$$I_{\infty}^{\tau}(S;W) = \log \sup_{\mathcal{V} \subseteq (\mathcal{S}\times\mathcal{W}), \mathbf{P}((S,W)\in\mathcal{V}) > \tau} \frac{\mathbf{P}((S,W) \in \mathcal{V}) - \tau}{\mathbf{P}(S \otimes W \in \mathcal{V})}.$$

Consider the generalization error bounds in Section 5 in terms of various information measures. Interestingly, for an algorithm $P_{W|S}$ that is (ε, δ)-DP with $\delta > 0$, denote $P := P_{WS}$ and $Q := P_W P_S$, the *only* information measure that admits an explicit, finite, and non-trivial upper bound in terms of ε and δ is $E_\gamma(P\|Q)$. In comparison, $\mathcal{H}_\beta(P\|Q)$ and $\mathcal{L}(S \to W)$ can be unbounded, and $\mathrm{TV}(P,Q)$ and $H^2(P,Q)$ only admit trivial bounds. Therefore, it is of interest to connect $E_\gamma(P\|Q)$ with $I_{\infty}^{\tau}(S;W)$. We thus derive the following connection, whose proof can be found in Apendix H.

Lemma 6.2. *Let P, Q be probability measures on $\mathcal{X}$. Fix $\tau \in [0,1)$. For any $\gamma \geq 1$, $E_\gamma(P\|Q) \leq \tau$ if and only if $I_{\infty}^{\tau}(P\|Q) \leq \log \gamma$.*

Similar to (Rogers et al., 2016), we assume the dataset S is drawn i.i.d. from a product distribution. For the remainder of this section, set $P := P_{SW}$ and $Q := P_S P_W$. Combining Lemma 6.2 and (Rogers et al., 2016, Theorem III.1), we derive the following result.

Proposition 6.3. *Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ be (ϵ, δ)-DP with $\epsilon \in (0, \frac{1}{2}]$ and $\delta \in (0, \epsilon)$, there exist constants $c_1, c_2 > 0$ such that by taking $\tau = e^{-\epsilon^2 n} + c_1 n\sqrt{\delta/\epsilon}$ and $k = c_2(\epsilon^2 n + n\sqrt{\delta/\epsilon})$, we have*

$$E_{e^k}(P\|Q) \leq \tau.$$

We hence derive the following generalization bound by combining Proposition 6.3 with Proposition 2.1: if $P_{W|S}$ is (ϵ, δ)-DP and satisfies the conditions in Proposition 6.3, then for any $\eta \in (0,1)$,

$$\mathbf{P}\left(\left|\mathrm{gen}(S,W)\right| \geq \eta\right) \leq 2\exp\left(c_2(\epsilon^2 n + n\sqrt{\delta/\epsilon}) - n\eta^2/(2\sigma^2)\right) + e^{-\epsilon^2 n} + c_1 n\sqrt{\delta/\epsilon}.$$

For fixed positive δ, the additive term $c_1 n\sqrt{\delta/\epsilon}$ eventually becomes vacuous. In contrast, if $\delta = \delta_n$ decreases with n, the bound can vanish. For example, for fixed $\epsilon, \eta > 0$, taking $\delta_n = \epsilon n^{-4}$ gives $n\sqrt{\delta_n/\epsilon} = 1/n$, and the bound tends to zero whenever $\eta^2/(2\sigma^2) > c_2\epsilon^2$.

7 DATA MEMORIZATION

In machine learning, a critical question is *data memorization*, where the model output contains specific training information beyond what is needed to learn the underlying data distribution. Unlike privacy, which asks whether an algorithm's output reveals sensitive information about the training data and is usually a guarantee on the algorithm itself, memorization focuses on whether the trained model has actually retained specific training examples and can reveal or exploit them later.

In this section, by exploiting the flexibility of the event choice, we apply our scheme to strengthen the memorization guarantee of Sefidgaran et al. (2025, Theorem 5). We use the standard supersample/selector setup detailed in Appendix N: $\tilde{Z}$ is a $2n$-sample supersample, $J \in \{0,1\}^n$ is the uniform selector, and $I_n := I(W; J \mid \tilde{Z})$. For $t \in [0, 1/2)$, define

$$q_t := 2^{-n} \sum_{k=0}^{\lfloor nt \rfloor} \binom{n}{k}, \qquad \Psi_n^\star(t) := \sup\{p \in [q_t, 1] : D_{\mathrm{KL}}(p \| q_t) \le I_n\},$$

More details and explanations can be found in Appendix N.

Theorem 7.1. *Let $Q : \mathcal{W} \times \mathcal{Z} \times \mathcal{P}(\mathcal{Z}) \to \{0,1\}$ be a membership query, i.e., $Q(W, z, \mu) = 1$ if the adversary says z is a training sample. By relabeling, let $Z_{i,1}$ and $Z_{i,0}$ be the training and fresh sample, respectively, and define $T_n := \sum_{i\in[n]} Q(W, Z_{i,1}, \mu)$ and $F_n := \sum_{i\in[n]} Q(W, Z_{i,0}, \mu)$. If $\mathbf{P}(T_n \ge \alpha n) \ge q$ for some $\alpha \in (0,1)$, then for every $\beta \in [0, \alpha)$ and every $\eta \in (0, (\alpha-\beta)/2)$,*

$$\mathbf{P}(F_n > \beta n) \ge q - \Psi_n^\star\big(\frac{1-\alpha+\beta}{2} + \eta\big) \Big/ \Big(1 - \exp\big(-\frac{2\eta^2}{1-\alpha+\beta} n\big)\Big).$$

Hence if $I(W; J \mid \tilde{Z}) = o(n)$, then for every fixed $\beta < \alpha$,

$$\mathbf{P}(F_n > \beta n) \ge q - o(1).$$

The details and proof are deferred to Appendix N. Theorem 7.1 shows that under sublinear CMI, an adversary cannot maintain a persistent linear separation between training and fresh samples declared "in": if it declares at least αn training samples as "in" with probability at least q, then for every fixed $\beta < \alpha$ it must also declare more than βn fresh samples as "in" with probability at least $q - o(1)$. Equivalently, a persistent linear membership gap of this type rules out the sublinear-CMI regime.

Our result strengthens Sefidgaran et al. (2025, Theorem 5(ii)), which gives $\mathbf{P}(F_n \ge m_\epsilon) \ge (\alpha - \epsilon)q$ for $m_\epsilon = \frac{\epsilon}{1/q+\epsilon-\alpha} n - o(n)$. Since the leading coefficient of m_ϵ is smaller than α, one may choose β between that coefficient and α; Theorem 7.1 then implies $\mathbf{P}(F_n \ge m_\epsilon) \ge q - o(1)$ for all sufficiently large n.

8 CONCLUDING REMARKS AND LIMITATIONS

We have studied a unified framewor, using DPI as the organizing principle for change-of-measure inequalities, and characterized precisely the sharp information retained by the resulting Bernoulli problem. It is of interest to explore whether our framework can be connected to other unified approaches, e.g., compressibility bounds (Sefidgaran et al., 2022) and a decorrelation lemma (Chu & Raginsky, 2023). Moreover, the strong DPI has recently found applications in DNN generalization (He & Goldfeld, 2025) and data memorization (Feldman et al., 2025). Incorporating strong DPI into our framework is yet another future direction.

Limitations. Our approach requires an appropriate choice of P, Q, and E, which is not always straightforward across applications. While our DPI framework is often tight at the change-of-measure level, translating it into explicit generalization or memorization guarantees may still involve problem-dependent relaxations and require substantial effort. Our examples identify concrete regimes where the refinements are substantial, including an exactly enumerable Gibbs learner and a Gaussian PAC-Bayesian asymptotic separation. Nevertheless, broader empirical evaluation on practical learning algorithms remains an important direction.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing, including grammar correction, and with developing Python code for numerical simulations and figure generation. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28 (1):131–142, 1966.

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

Gholamali Aminian, Yuheng Bu, Gregory W Wornell, and Miguel RD Rodrigues. Tighter expected generalization error bounds via convexity of information measures. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2481–2486. IEEE, 2022.

Gholamali Aminian, Saeed Masiha, Laura Toni, and Miguel RD Rodrigues. Learning algorithm generalization error bounds via auxiliary distributions. *IEEE Journal on Selected Areas in Information Theory*, 5:273–284, 2024.

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

Shahab Asoodeh, Mario Diaz, and Flavio P Calmon. Contraction of E_γ-divergence and its applications to privacy. *arXiv preprint arXiv:2012.11035*, 2020.

Shahab Asoodeh, Maryam Aliakbarpour, and Flavio P Calmon. Local differential privacy is equivalent to contraction of an f-divergence. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 545–550. IEEE, 2021.

Idan Attias, Gintare Karolina Dziugaite, Mahdi Haghifam, Roi Livni, and Daniel M. Roy. Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing. In *International Conference on Machine Learning (ICML)*, pp. 2035–2068. PMLR, 2024.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020. doi: 10.29012/jpc.726.

Pradeep Kr Banerjee and Guido Montúfar. Information complexity and generalization bounds. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 676–681. IEEE, 2021.

Gilles Barthe, Boris Köpf, Federico Olmedo, and Santiago Zanella-Béguelin. Probabilistic relational reasoning for differential privacy. *ACM Transactions on Programming Languages and Systems*, 35 (3), 2013.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pp. 25–55. PMLR, 2018.

Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1963–1981. PMLR, 2022.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *ACM Symposium on Theory of Computing*, pp. 123–132, 2021.

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.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, pp. 267–284. USENIX Association, 2019.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

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

Wei-Ning Chen and I-Hsiang Wang. Anonymous heterogeneous distributed detection: Optimal decision rules, error exponents, and the price of anonymity. *IEEE Transactions on Information Theory*, 65(11):7390–7406, 2019.

Yifeng Chu and Maxim Raginsky. A unified framework for information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 36:79260–79278, 2023.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1-2):85–108, 1963.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pp. 146–158, 1975.

Imre Csiszár. Generalized cutoff rates and Rényi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, 2002.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Monroe Donsker and S. R. Srinivasa Varadhan. Large deviations for markov processes and the asymptotic evaluation of certain markov process expectations for large times. In *Probabilistic Methods in Differential Equations: Proceedings of the Conference Held at the University of Victoria, August 19–20, 1974*, pp. 82–88. Springer, 1974.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pp. 265–284, 2006.

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

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems*, 28, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *ACM Symposium on Theory of Computing*, pp. 117–126, 2015b.

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

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Sibson α-mutual information and its variational representations. *IEEE Transactions on Information Theory*, 72(7):1–36, 2025.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959. Association for Computing Machinery, 2020.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 535–544. PMLR, 2018.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2881–2891, 2020.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. In *Conference on Learning Theory (COLT)*, pp. 1935–1973. PMLR, 2025.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

Muhan Guan, Farhad Farokhi, and Jingge Zhu. A DPI-PAC-Bayesian framework for generalization bounds. In *IEEE Information Theory Workshop (ITW) 2025*, pp. 1–6, Australia, 2025.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the 2nd Congress of the Société Mathématique de France*, volume 33, pp. 391–414. 2019.

Benjamin Guedj and Louis Pujol. Still no free lunches: the price to pay for tighter PAC-bayes bounds. *Entropy*, 23(11):1529, 2021.

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.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22911–22924, 2022.

Peter Harremoës. Interpretations of rényi entropies and divergences. *Physica A: Statistical Mechanics and its Applications*, 365(1):57–62, 2006.

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.

Haiyun He and Ziv Goldfeld. Information-theoretic generalization bounds for deep neural networks. *IEEE Transactions on Information Theory*, 71(8):6227–6247, 2025.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020.

Henryk Hudzik and Lech Maligranda. Amemiya norm equals orlicz norm in general. *Indagationes Mathematicae*, 11(4):573–585, 2000.

Ibrahim Issa, Sudeep Kamath, and Aaron B Wagner. An operational measure of information leakage. In *Conference on Information Science and Systems (CISS)*, pp. 234–239. IEEE, 2016.

Ibrahim Issa, Amedeo Roberto Esposito, and Michael Gastpar. Strengthened information-theoretic bounds on the generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 582–586, 2019. doi: 10.1109/ISIT.2019.8849834.

Ibrahim Issa, Amedeo Roberto Esposito, and Michael Gastpar. Generalization error bounds for noisy, iterative algorithms via maximal leakage. In *Conference on Learning Theory (COLT)*, pp. 4952–4976. PMLR, 2023.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1475–1479. IEEE, 2017.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1376–1385. PMLR, 2015.

Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Scalable information inequalities for uncertainty quantification. *Journal of Computational Physics*, 336:513–545, 2017.

Jiachun Liao, Lalitha Sankar, Flavio P Calmon, and Vincent YF Tan. Hypothesis testing under maximal leakage privacy constraints. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 779–783. IEEE, 2017.

Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. A tunable measure for information leakage. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 701–705. IEEE, 2018.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

Yanxiao Liu, Chun Hei Michael Shiu, Lele Wang, and Deniz Gündüz. On the generalization error of differentially private algorithms via typicality. *arXiv preprint arXiv:2601.08386*, 2026.

Saeed Masiha, Amin Gohari, and Mohammad Hossein Yassaee. f-divergences and their applications in lossy compression and bounding generalization error. *IEEE Transactions on Information Theory*, 69(12):7538–7564, 2023.

David A McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory*, pp. 230–234, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Conference on Learning Theory (COLT)*, pp. 164–170. PMLR, 1999.

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

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

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021.

Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f-divergences. *arXiv preprint arXiv:2202.05568*, 2022.

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

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Borja Rodríguez-Gálvez, Germán Bassi, and Mikael Skoglund. Upper bounds on the generalization error of private algorithms for discrete data. *IEEE Transactions on Information Theory*, 67(11): 7362–7379, 2021.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494. IEEE, 2016.

Kjetil Røysland. A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17(3):895–915, August 2011. ISSN 1350-7265. doi: 10.3150/10-BEJ303.

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

Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Pointwise maximal leakage. *IEEE Transactions on Information Theory*, 69(12):8054–8080, 2023.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Milad Sefidgaran and Abdellatif Zaidi. Data-dependent generalization bounds via variable-size compressibility. *IEEE Transactions on Information Theory*, 70(9):6572–6595, 2024.

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

Milad Sefidgaran, Kimia Nadjahi, and Abdellatif Zaidi. Tighter CMI-based generalization bounds via stochastic projection and quantization. *Advances in Neural Information Processing Systems*, 38, 2025.

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

Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.

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

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pp. 11–30. Springer, 2015.

Vladimir Naumovich Vapnik, Vlamimir Vapnik, et al. Statistical learning theory. 1998.

Sergio Verdú. α-mutual information. In *Information Theory and Applications Workshop (ITA)*, pp. 1–6. IEEE, 2015.

Sasha Voitovych, Mahdi Haghifam, Idan Attias, Gintare Karolina Dziugaite, Roi Livni, and Daniel M. Roy. On traceability in ℓ_p stochastic convex optimization. *CoRR*, abs/2502.17384, 2025.

Ziqiao Wang and Yongyi Mao. Generalization bounds via conditional f-information. *Advances in Neural Information Processing Systems*, 37:52159–52188, 2024.

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

Moshe Zakai and Jacob Ziv. A generalization of the rate-distortion theory and applications. In *Information Theory New Trends and Open Problems*, pp. 87–123. Springer, 1975.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

A More on Related Work

We provide a more detailed discussion of related works as follows.

Change of Measure Inequalities. Change of measure inequalities have been studied in learning theory since Csiszár (1975); Donsker & Varadhan (1974) for PAC-Bayesian bounds. In particular, Hellström & Durisi (2020) used a special change of measure inequality, namely the strong converse lemma (2) (see also (Polyanskiy & Wu, 2025)), to obtain various types of generalization error bounds. This lemma (2) has also been used widely for large-deviation analysis and hypothesis testing (Polyanskiy & Wu, 2025). See also the use of change of measure inequalities in marginal structural models (Røysland, 2011) and robust uncertainty quantification bounds for statistical estimators (Katsoulakis et al., 2017). Among change of measure inequalities, the two works that are most related to our results are Picard-Weibel & Guedj (2022) and Ohnishi & Honorio (2021), which used the Young-Fenchel inequality and the variational representation of f-divergences, respectively, to derive new change of measure inequalities and hence new PAC-Bayesian generalization bounds; see also the references therein.

Generalization Error Bounds. Generalization error analysis has been one of the most important problems in machine learning in the past decades. It measures how a stochastic learning algorithm performs on data that are outside the training dataset, and there has been a vast array of work characterizing it. Early discussions include sample compression schemes (Littlestone & Warmuth, 1986) (the output has to be "compressible"; see also (Vapnik et al., 1998; Boucheron et al., 2005)), model over-parameterization (Shalev-Shwartz & Ben-David, 2014) (over-parameterized models could overfit, although counterexamples were found by Zhang et al. (2017)), uniform stability (Bousquet & Elisseeff, 2002), and the theory of uniform convergence (Vapnik & Chervonenkis, 2015) (the output has to be "sufficiently simple"; see also (Blumer et al., 1987)). However, it has been argued by Zhang et al. (2017) that algorithm-independent generalization bounds fall short of explaining the surprising success of modern artificial intelligence, and hence algorithm-dependent generalization bounds through different lenses have been explored.

In this direction, information-theoretic measures have been used to characterize generalization performance in the past decade. This line of work was initiated by Russo & Zou (2016); Xu & Raginsky (2017), who connected the generalization error to the *mutual information* between the training dataset S and the algorithm output W, indicating that algorithms that leak little information about the dataset generalize well (see also (Banerjee & Montúfar, 2021; Bassily et al., 2018)). These bounds were further tightened using chaining (Asadi et al., 2018) and other techniques (Harutyunyan et al., 2021; Bu et al., 2020). Since mutual information can easily be infinite, *conditional* mutual information has been used by Steinke & Zakynthinou (2020); Haghifam et al. (2020; 2021). Viewing mutual information as a KL divergence, it was further generalized to more general information measures through f-divergences (Esposito et al., 2021; Masiha et al., 2023; Wang & Mao, 2024). Various unified frameworks that cover multiple types of information measures have also been proposed, e.g., via information density (Hellström & Durisi, 2020), rate-distortion theory (Sefidgaran et al., 2022; Sefidgaran & Zaidi, 2024), convexity of the information measures (Aminian et al.,

2022), auxiliary distributions (Aminian et al., 2024), and a probabilistic decorrelation lemma (Chu & Raginsky, 2023).

Information Measures. In this paper, we explore the use of a wide range of information measures to characterize the generalization of stochastic algorithms. Our refinements and applications are mainly based on f-divergences, which were introduced by Ali & Silvey (1966); Csiszár (1963; 1967) as a generalization of the relative entropy that preserves useful properties, e.g., the data processing inequality (Zakai & Ziv, 1975). See Sason & Verdú (2016) for a comprehensive study. A measure that is closely related to f-divergences is the Rényi divergence (Rényi, 1961), which has several useful operational interpretations, e.g., the number of bits by which a mixture of two codes can be compressed (Harremoës, 2006; Grünwald, 2007) and the cutoff rate in block coding and hypothesis testing (Csiszár, 2002). See also (Van Erven & Harremos, 2014) for a comprehensive study.

Besides f-divergence and Rényi divergence, another important class of information measures is the Sibson α-mutual information (Sibson, 1969); see also Verdú (2015) for a revisit. It generalizes both mutual information and maximal leakage (Issa et al., 2016). Maximal leakage, in particular, was also proposed as an information leakage measure, and has found recent applications in security; see (Liao et al., 2017; 2018; Saeidian et al., 2023). Variational representations of the Sibson α-mutual information have been studied by Esposito et al. (2025). In Esposito et al. (2021), high-probability generalization bounds in terms of maximal leakage and Sibson α-mutual information were proposed. Also see the study of the generalization behavior of iterative, noisy learning algorithms (Pensia et al., 2018) in terms of maximal leakage by Issa et al. (2023).

Privacy. Information leakage, or privacy, is tightly related to the generalization of stochastic algorithms, in the sense that private algorithms leak less information and are more "stable" (Dwork et al., 2015a;b). Differential privacy (Dwork et al., 2006; 2014) has been one of the most popular privacy measures in the past decade. Typicality, as a celebrated tool in information theory, has been used to derive generalization bounds for differentially private (and its variants (Dong et al., 2022)) algorithms that are easy to compute (Rodríguez-Gálvez et al., 2021; Liu et al., 2026).

The E_γ-divergence is tightly related to differential privacy. In particular, for $\gamma = e^\varepsilon$, the usual (ε, δ)-DP event inequality is exactly a hockey-stick divergence bound; this characterization appears in earlier privacy work (Barthe et al., 2013; Balle et al., 2020), and Kairouz et al. (2015) give the closely related hypothesis-testing/data-processing interpretation. The privacy-profile formulation of Balle et al. (2020) explicitly takes a supremum of hockey-stick divergences over neighboring datasets. Beyond this classical characterization, contraction coefficients for E_γ were studied by Asoodeh et al. (2020), and Asoodeh et al. (2021) characterize local differential privacy through contraction of an f-divergence. We build on, rather than claim novelty for, these established privacy connections.

For $\gamma = e^\varepsilon$, the eventwise (ε, δ) inequality is equivalent to $E_{e^\varepsilon}(\mathcal{A}(s)\|\mathcal{A}(s')) \leq \delta$ for neighboring datasets (with both directions covered by the symmetric neighboring relation); see, e.g., Barthe et al. (2013); Balle et al. (2020) and the related hypothesis-testing/DPI view of Kairouz et al. (2015). Our use of E_γ below builds on this established characterization.

Data Memorization. Data memorization has received substantial recent attention, both as a possible explanation of good *generalization* in hard regimes and as a *privacy* risk. Modern machine learning models can leak verbatim or near-verbatim information about their training sets, and this behavior has been connected to membership inference, extraction, and reconstruction attacks (Carlini et al., 2019; 2021; 2023; Haim et al., 2022). On the theoretical side, Feldman (2020) argued that long-tailed distributions can make memorization statistically useful, while Brown et al. (2021) identified natural prediction problems in which every sufficiently accurate learner must encode essentially all the information contained in a large subset of its training examples, including portions that are irrelevant to the target task. Related empirical studies further investigated which samples are memorized and how such memorization may decay during training (Feldman & Zhang, 2020; Jagielski et al., 2023).

In the context of stochastic convex optimization (SCO) (Attias et al., 2024), memorization was quantified by conditional mutual information (CMI), and a trade-off between accuracy and information leakage was provided. A broader lower-bound framework was later developed by Feldman et al. (2025), who introduced *excess data memorization* and related its lower bounds to strong data processing inequalities (Polyanskiy & Wu, 2025). This yields sample-size/memorization trade-offs

for binary prediction problems with latent structure, extending (Brown et al., 2021). Because CMI is always at most n, these SDPI-based lower bounds are conceptually distinct from the CMI lower bounds of Attias et al. (2024), and they emphasize a complementary phenomenon: additional data can gradually reduce the memorization burden. See also the follow-up work of Voitovych et al. (2025). Meanwhile, Sefidgaran et al. (2025) showed that the necessity of memorization should not be interpreted too broadly. By combining stochastic projection with lossy compression and quantization, they derived tighter CMI-based generalization bounds that remain non-vacuous on the SCO counterexamples where standard MI/CMI bounds fail. These works suggest that data memorization is highly notion- and representation-dependent: it can be unavoidable for certain outputs or learner classes, yet avoidable after an appropriate randomized compression of the learned model.

B Proofs of Proposition 2.1

A proof sketch has already been presented following Proposition 2.1 in Section 4.

In this section, we present two complete proofs of Proposition 2.1. The first is a more detailed, complete version of the proof sketch that used (3), highlighting the usefulness of our DPI approach. The second follows a more standard route and helps clarify why Proposition 2.1 is tighter than he strong converse lemma (2).

B.1 Proof 1 of Proposition 2.1

Consider $f(t) := [t-\gamma]_+$ that is convex on $(0,\infty)$ for any $\gamma \geq 1$. Recall $P \ll Q$.

If $q = 0$, then by absolute continuity, $P(E) = 0$ and (4) holds trivially.

If $q = 1$, then $p = 1$ and we have $E_\gamma(P\|Q) \geq [1-\gamma]_+$, hence

$$\gamma + E_\gamma(P\|Q) \geq \gamma + [1-\gamma]_+ \geq 1 = P(E).$$

If $q \in (0,1)$, by data processing applied to the indicator map $\mathbb{1}_E$,

$$E_\gamma(P\|Q) = D_f(P\|Q) \geq D_f(\mathbb{1}_E \circ P \| \mathbb{1}_E \circ Q) = D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)).$$

With $f(t) := [t-\gamma]_+$, we calculate

$$D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) = qf\Big(\frac{p}{q}\Big) + (1-q)f\Big(\frac{1-p}{1-q}\Big) = q\Big[\frac{p}{q} - \gamma\Big]_+ + (1-q)\Big[\frac{1-p}{1-q} - \gamma\Big]_+.$$

Using $[x]_+ \geq x$ for all x and $(1-q)[\cdot]_+ \geq 0$, we get

$$E_\gamma(P\|Q) \geq q\Big(\frac{p}{q} - \gamma\Big) = p - \gamma q,$$

and rearrangement yields $p \leq \gamma q + E_\gamma(P\|Q)$, i.e. (4).

B.2 Proof 2 of Proposition 2.1

For probability distributions P and Q such that $P \ll Q$, for any measurable set E and any $\gamma \geq 1$, we have

$$\begin{aligned}
P(E) &= \int \mathbb{1}_E \mathrm{d}P \\
&= \int \mathbb{1}_E \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}Q \\
&= \int \mathbb{1}_E \left(\min\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q}, \gamma \right\} + \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma \right]_+ \right) \mathrm{d}Q \\
&= \int \mathbb{1}_E \left(\min\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q}, \gamma \right\} \right) \mathrm{d}Q + \int \mathbb{1}_E \left(\left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma \right]_+ \right) \mathrm{d}Q \\
&\leq \gamma Q(E) + \int \mathbb{1}_E \left(\left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma \right]_+ \right) \mathrm{d}Q \\
&\leq \gamma Q(E) + \int \left(\left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma \right]_+ \right) \mathrm{d}Q \\
&= \gamma Q(E) + E_\gamma(P \| Q).
\end{aligned}$$

B.3 Discussions on Strong Converse Lemma

In this subsection, we discuss the strong converse lemma (Polyanskiy & Wu, 2025, Chapter 14), and in particular why Proposition 2.1 turns out to be tighter.

We first review the strong converse lemma (2) together with its proof, which assist our discussion later.

Lemma B.1. *For probability distributions P and Q such that $P \ll Q$, for any measurable set E and any $\gamma \geq 0$, we have*

$$P(E) \leq \gamma Q(E) + P\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right).$$

Proof. For probability distributions P and Q such that P is absolutely continuous with respect to Q, for a measurable set E and any $\gamma \in \mathbb{R}$, we have

$$\begin{aligned}
P(E) &= \int \mathbb{1}_E \mathrm{d}P \\
&= \int \mathbb{1}_E \mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \leq \gamma \right) \left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right) \mathrm{d}Q + \int \mathbb{1}_E \mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right) \mathrm{d}P \\
&\leq \int \mathbb{1}_E \gamma \mathrm{d}Q + \int \mathbb{1}_E \mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right) \mathrm{d}P \\
&= \gamma Q(E) + \int \mathbb{1}_E \mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right) \mathrm{d}P \\
&\leq \gamma Q(E) + \int \mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right) \mathrm{d}P \\
&= \gamma Q(E) + P\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right).
\end{aligned}$$

□

For the same $\gamma \geq 1$, Proposition 2.1 is no worse than Lemma B.1; it recovers the latter by the relaxation

$$E_\gamma(P \| Q) \leq \int_{\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right\}} \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}Q = P\left(\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma \right\} \right).$$

It is easy to compare the second proof of Proposition 2.1, shown in Section B.2, to the proof of the strong converse lemma as shown above. The advantage of Proposition 2.1 can be understood as follows. The event $\{\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\}$ is the region where the event on P is much heavier than that on Q; to bound $P(E)$ we essentially need a pointwise inequality such that

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) \le \gamma + r(\omega) \tag{19}$$

and then integrate r on Q. To prove Proposition 2.1, we employ

$$r(\omega) = \left[\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) - \gamma\right]_+,$$

which is the smallest nonnegative function satisfying (19), while the choice of $r(\omega)$ in the proof of Lemma B.1 is

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega)\,\mathbb{1}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) > \gamma\right),$$

which is lower bounded by $[\mathrm{d}P/\mathrm{d}Q(w) - \gamma]_+$.

From another perspective, the proof of Lemma B.1 effectively discards a conditional term of the form $P(A|E)$, thereby losing information about how the event E is positioned relative to the "high-information" region $\{\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\}$. In contrast, the E_γ-divergence captures both *how often* $\frac{\mathrm{d}P}{\mathrm{d}Q}$ exceeds γ and *how far above* γ it typically lies. Consequently, $E_\gamma(P\|Q)$ retains finer information about "where E sits" within the high-information regime, partially compensating for what is lost by dropping $P(A|E)$.

C Proof of Theorem 4.1 and Discussions

In this section, we present the proof of Theorem 4.1 case by case, all of which are based on (3). Some cases require further fine relaxations, which will also be discussed in Section C.7 together with bound comparisons.

C.1 Proof of (11)

Consider the χ^2 distance is an f-divergence with $f = t^2 - 1$.

By (3) we can calculate

$$\chi^2(P\|Q) \ge Q(E)\cdot\left(\left(\frac{P(E)}{Q(E)}\right)^2 - 1\right) + (1-Q(E))\cdot\left(\left(\frac{1-P(E)}{1-Q(E)}\right)^2 - 1\right),$$

and rearrangement yields (11).

C.2 Proof of (10)

Consider the KL divergence is an f-divergence with $f(t) = t\log t$, at first it is easy to find

$$\begin{aligned}
\mathrm{KL}(P\|Q) &\ge P(E)\log\frac{P(E)}{Q(E)} + (1-P(E))\log\frac{1-P(E)}{1-Q(E)} \\
&= P(E)\log\frac{1}{Q(E)} + (1-P(E))\log\frac{1}{1-Q(E)} - h_2(P(E)) \\
&\ge P(E)\log\frac{1}{Q(E)} - \log 2.
\end{aligned}$$

However, the last inequality can be a bit crude so that it is not always tighter than (Picard-Weibel & Guedj, 2022). To improve the bound, we can perform a finer analysis as follows.

Let $p := P(E)$ and $q := Q(E) \in (0,1)$, we can follow the same procedure until

$$\mathrm{KL}(P\|Q) \ge p\log\frac{1}{q} + (1-p)\log\frac{1}{1-q} - h_2(p),$$

and then we instead use the Fenchel inequality for h_2: for every $t \in \mathbb{R}$,

$$h_2(p) \le \log\left(1 + e^t\right) - pt. \tag{20}$$

Substituting (20) into the previous display yields, for every $t \in \mathbb{R}$,

$$\begin{aligned}\mathrm{KL}(P\|Q) &\ge p \log \frac{1}{q} + (1-p)\log\frac{1}{1-q} + pt - \log(1+e^t) \\ &= p\Big(t + \log\frac{1-q}{q}\Big) - \log(1-q) - \log(1+e^t).\end{aligned}$$

If we take $c := t + \log\frac{1-q}{q}$, we have

$$\log(1+e^t) = \log\Big(1 + e^c \frac{q}{1-q}\Big) = \log(1 - q + qe^c) - \log(1-q),$$

and hence, for every $c \in \mathbb{R}$,

$$\mathrm{KL}(P\|Q) \ge pc - \log(1 - q + qe^c). \tag{21}$$

By rearrangement we have for $c > 0$,

$$P(E) \le \frac{\mathrm{KL}(P\|Q) + \log\left(1 + q(e^c - 1)\right)}{c}. \tag{22}$$

C.3 Proof of (12)

Consider the Hellinger squared distance is an f-divergence with $f = (\sqrt{t} - 1)^2$.

By (3) we can calculate

$$\begin{aligned}H^2(P;Q) &\ge Q(E) \cdot \left(\sqrt{\frac{P(E)}{Q(E)}} - 1\right)^2 + (1 - Q(E)) \cdot \left(\sqrt{\frac{1-P(E)}{1-Q(E)}} - 1\right)^2 \\ &= 2\left(1 - \sqrt{P(E)Q(E)} - \sqrt{(1-P(E))(1-Q(E))}\right),\end{aligned}$$

which results in (12).

C.4 Proof of (13)

Consider the Power divergence of order β is an f-divergence with $f = (t^\beta - 1)/(\beta - 1)$.

By (3) we can calculate

$$\mathcal{H}_\beta(P\|Q) \ge \frac{1}{\beta - 1} \cdot \left(Q(E)\left(\left(\frac{P(E)}{Q(E)}\right)^\beta - 1\right) + (1 - Q(E))\left(\left(\frac{1-P(E)}{1-Q(E)}\right)^\beta - 1\right)\right),$$

and rearrangement yields (13).

C.5 Further Relaxations

To convert (12) and (13) to the form of (1), we can relax them as follows:

$$P(E) \le \left(\sqrt{Q(E)\left(1 - H^2(P;Q)/2\right)^2} + \sqrt{(1-Q(E))\left(1 - (1 - H^2(P;Q)/2)^2\right)}\ \right)^2, \tag{23}$$

$$P(E) \le \left((\beta-1)\mathcal{H}_\beta(P\|Q) + 1\right)^{1/\beta} \Big/ \left(Q(E)^{1-\beta} + M^\beta (1 - Q(E))^{1-\beta}\right)^{1/\beta}, \tag{24}$$

where $M \le 1/P(E) - 1$ in (24), which is a valid upper bound on $P(E)$ as long as $Q(E)$ is bounded by $Q_{\max} < 1$. For such a case with fixed β, we let $M = Q_{\max}^{(1-\beta)/\beta} \cdot ((\beta-1)\mathcal{H}_\beta(P\|Q) + 1)^{-1/\beta} - 1$.

We can relax (13) to derive a bound that is even tighter than (24) when $Q(E)$ is small, precisely the regime relevant for generalization bounds where $Q(E)$ decays exponentially with sample size[3]:

$$P(E) \le \left((Q(E))^{\beta-1}\left[1 + (\beta-1)\mathcal{H}_\beta(P\|Q) - (1-Q(E))^{1-\beta}(1-u_0)^\beta\right]_+\right)^{1/\beta}, \tag{25}$$

[3]If $X \sim P_X$ is σ-sub-Gaussian, Hoeffding's inequality gives $\forall \eta > 0$, $P_X\left(|X - \mathbf{E}[X]| \ge \eta\right) \le 2\exp(-\eta^2/2\sigma^2)$.

where $u_0 := \min\left\{1,\ \left((1+(\beta-1)\mathcal{H}_\beta(P\|Q))\, q^{\beta-1}\right)^{1/\beta}\right\}$. See Appendix C.6 for a proof.

Based on (7), one can also convert (24) and (25) to inequalities in terms of Rényi divergence.

We then compare our bounds with existing results. It turns out that, in almost all cases, our bounds are tighter than the existing results (Picard-Weibel & Guedj, 2022; Esposito et al., 2021), while (11) and (10) recover the results in (Picard-Weibel & Guedj, 2022); we conjecture that they may already be optimal. The bound in (11) improves upon (Esposito et al., 2021, Corollary 7) whenever $\chi^2(P\|Q) \lesssim 1/(4Q(E))$ and (10) is tighter than (Bassily et al., 2018, Lemma 9). The bound in (12) is tighter than (Esposito et al., 2021, Corollary 9) and does not require $P(E) \geq Q(E)$ as they do. The bound in (13) recovers (Esposito et al., 2021, Corollary 7) by dropping the second term on the left-hand side. Finally, the bound in (25) is tighter than (Esposito et al., 2021, Corollary 7) whenever $Q(E)$ is small. Comparing (25) to (Picard-Weibel & Guedj, 2022), neither is stronger than another, but one advantage of (25) it does not need to optimize over $s \in \mathbb{R}$.

C.6 Proof of (25)

We let $p := P(E)$ and $q := Q(E)$. Recall $u_0 := \min\left\{1,\ \left((1+(\beta-1)\mathcal{H}_\beta(P\|Q))\, q^{\beta-1}\right)^{1/\beta}\right\}$, which is less than 1 when $Q(E)$ is small.

From (13) and $(1-p)^\beta(1-q)^{1-\beta} \geq 0$, we get $p^\beta q^{1-\beta} \leq 1+(\beta-1)\mathcal{H}_\beta(P\|Q)$, hence $p \leq u_0$. Therefore $1-p \geq 1-u_0$, which implies $(1-p)^\beta \geq (1-u_0)^\beta$. Plugging this into (13) yields

$$p^\beta q^{1-\beta} \leq 1+(\beta-1)\mathcal{H}_\beta(P\|Q) - (1-u_0)^\beta(1-q)^{1-\beta}.$$

Multiplying $q^{\beta-1}$ and then taking the $1/\beta$ power on both sides give the desired result.

C.7 Discussions and Comparison

We then discuss our inequalities and compare them with existing results.

KL divergence. From (Bassily et al., 2018, Lemma 9) we have

$$P(E) \leq \frac{\mathrm{KL}(P\|Q)+\log 2}{\log(1/Q(E))}. \tag{26}$$

It is easy to see (22) implies (26) by

$$\log\left(1+q(e^c-1)\right) \leq \log(1+e^c) \leq c+\log 2,$$

and in particular, choosing $c = \log(1/q) > 0$ in (22) yields

$$P(E) \leq \frac{\mathrm{KL}(P\|Q)+\log(2-q)}{\log(1/q)} < \frac{\mathrm{KL}(P\|Q)+\log 2}{\log(1/q)}, \tag{27}$$

which is strict since $2-q<2$ for every $q \in (0,1)$.

Power divergence. The power divergence bound (13) is equivalent to

$$\mathcal{H}_\beta(P\|Q) \geq \frac{P(E)^\beta Q(E)^{1-\beta} + (1-P(E))^\beta (1-Q(E))^{1-\beta} - 1}{\beta-1}.$$

One crude but clean relaxation is

$$\mathcal{H}_\beta(P\|Q) \geq \frac{P(E)^\beta Q(E)^{1-\beta} - 1}{\beta-1},$$

implying the result

$$P(E) \leq Q(E)^{\frac{\beta-1}{\beta}} \left((\beta-1)\mathcal{H}_\beta(P\|Q)+1\right)^{\frac{1}{\beta}},$$

which exactly recovers (Esposito et al., 2021, Corollary 7).

If we assume $P(E) \leq \frac{1}{M+1}$ for some $M > 0$, which is reasonable in generalization error analysis since the error is usually not large, then we have $1 - P(E) \geq MP(E)$, and hence obtain (24):

$$P(E) \leq \frac{((\beta-1)\mathcal{H}_\beta(P\|Q)+1)^{\frac{1}{\beta}}}{(Q(E)^{1-\beta} + M^\beta(1-Q(E))^{1-\beta})^{\frac{1}{\beta}}}.$$

We here verify the validity of the assumption on $P(E)$. By fixing β and assuming that $Q(E)$ is upper bounded by some $Q_{\max} < 1$, we can first upper bound $P(E)$ by

$$P(E) \leq Q_{\max}^{\frac{\beta-1}{\beta}} \cdot ((\beta-1)\mathcal{H}_\beta(P\|Q)+1)^{1/\beta} = \frac{1}{M+1},$$

which is equivalent to

$$M = Q_{\max}^{\frac{1-\beta}{\beta}} \cdot ((\beta-1)\mathcal{H}_\beta(P\|Q)+1)^{-1/\beta} - 1.$$

However, we note that in the regime where generalization error analysis is performed, i.e., $Q(E)$ is small due to Hoeffding's inequality, above is usually not as tight as (25).

Hellinger Squared Distance. The Hellinger squared distance bound (12) can easily recover the same functional form as (Esposito et al., 2021, Corollary 9), without requiring the condition $P_{SW}(E) \geq P_S P_W(E)$, which is assumed in (Esposito et al., 2021, Corollary 9). By utilizing a better relaxation, (24) is a tighter bound.

By (Picard-Weibel & Guedj, 2022), we have

$$P(E) \leq 1 + c - \frac{c(1+c)(1-H^2(P;Q))^2}{Q(E)+c}$$

for any $c > 0$. The optimal choice of c is for them is

$$c^* = -Q(E) + \sqrt{\frac{(1-H^2(P;Q))^2 Q(E)(1-Q(E))}{1-(1-H^2(P;Q))^2}},$$

which gives

$$P(E) \leq \left(\sqrt{\left(1-(1-H^2(P;Q))^2\right)\left(1-Q(E)\right)} + \left(1-H^2(P;Q)\right)\sqrt{Q(E)}\right)^2. \tag{28}$$

For our bound shown in (12), with $f(t) = (\sqrt{t}-1)^2$, we have

$$\begin{aligned} H^2(P;Q) &\geq Q(E) \cdot \left(\sqrt{\frac{P(E)}{Q(E)}} - 1\right)^2 + (1-Q(E)) \cdot \left(\sqrt{\frac{1-P(E)}{1-Q(E)}} - 1\right)^2 \\ &= 2\left(1 - \sqrt{P(E)Q(E)} - \sqrt{(1-P(E))(1-Q(E))}\right), \end{aligned}$$

which is equivalent to

$$1 - \frac{H^2(P;Q)}{2} \leq \sqrt{P(E)Q(E)} + \sqrt{(1-P(E))(1-Q(E))}.$$

Then we can solve that

$$P(E) \leq \left(\sqrt{Q(E)\left(1-\frac{H^2(P;Q)}{2}\right)^2} + \sqrt{(1-Q(E))\left(1-\left(1-\frac{H^2(P;Q)}{2}\right)^2\right)}\right)^2,$$

which is (23). Observe that both the above result and the optimal result in (Picard-Weibel & Guedj, 2022) can be written in the form of

$$F(x) = \left(\sqrt{Q(E)x^2} + \sqrt{(1-Q(E))(1-x^2)}\right)^2.$$

This function $F(x)$ is increasing on $x \in (0, \sqrt{Q(E)})$ and decreasing on $x \in [\sqrt{Q(E)}, 1)$. Since (28) is $F(1 - H^2(P;Q))$ and our result (23) is $F(1 - H^2(P;Q)/2)$. For

$$\sqrt{Q(E)} \leq 1 - H^2(P;Q) \leq 1 - \frac{H^2(P;Q)}{2},$$

we have

$$F(1 - H^2(P;Q)) \geq F(1 - H^2(P;Q)/2),$$

implying that our result (23) is tighter than (28) given in (Picard-Weibel & Guedj, 2022).

Then we compare (23) with the result in (Esposito et al., 2021). Observe that (23) is equivalent to

$$P(E) \leq \left(\sqrt{Q(E)}\Big(1 - \frac{H^2(P;Q)}{2}\Big) + \sqrt{1 - Q(E)} \cdot \frac{\sqrt{H^2(P;Q)\big(4 - H^2(P;Q)\big)}}{2} \right)^2.$$

Using $\sqrt{1 - (1 - H^2(P;Q)/2)^2} = \sqrt{H^2(P;Q) - (H^2(P;Q))^2/4} \leq \sqrt{H^2(P;Q)}$ we can relax above to

$$P(E) \leq \left(\sqrt{Q(E)} + \sqrt{H^2(P;Q)} \right)^2.$$

D Background and Proof of Theorem 4.2

In this section, we provide technical background of Theorem 4.2, as well as its proof.

D.1 Technical Background

We first present necessary background and technical details for Theorem 4.2. For more discussions, we refer the readers to (Hudzik & Maligranda, 2000; Jiao et al., 2017; Esposito et al., 2021).

We say a function ψ is an Orlicz function if it is a convex function $\psi : [0, \infty) \to [0, \infty]$ that vanishes at zero and is not identically 0 or ∞ on $(0, \infty)$. Given a convex function $\psi : [0, \infty) \to \mathbb{R}$, define its conjugate $\psi^\star : [0, \infty) \to \mathbb{R}$ as

$$\psi^\star(t) = \sup_{\lambda > 0} \lambda t - \psi(\lambda).$$

The generalized inverse of ψ is defined as

$$\psi^{-1}(s) := \inf \big\{ t \geq 0 : \psi(t) \geq s \big\}$$

for $s \geq 0$ and with the convention that $1/0 = \infty$ and $\psi^{-1}(\infty) = \infty$.

Consider a complete and σ-finite probability space $(\Omega, \mathcal{F}, \mu)$. Let $L^0(\mu)$ denote the space of all the $\mathcal{F}$-measurable and real valued functions on Ω. We can define a functional $I_\psi : L^0(\mu) \to [0, \infty]$ as

$$I_\psi(x) = \int_\Omega \psi(|x(t)|) \mathrm{d}\mu(t)$$

and then an Orlicz space can be defined as (Hudzik & Maligranda, 2000)

$$L_\psi(\mu) = \big\{ x \in L^0(\mu) : I_\psi(\lambda x) < \infty \text{ for some } \lambda > 0 \big\}.$$

Note the Orlicz space is a Banach space and can be endowed with Luxemburg, Orlicz and Amemiya norms. It has been shown in (Hudzik & Maligranda, 2000) that in general memiya norms are equivalent to Orlicz norms. Same to (Esposito et al., 2021), in this paper we restrict our discussions to probability spaces and define the corresponding norms with respect to random variables and the expectation operator.

For other uses of duality in Orlicz spaces in the context of generalization bounds, we refer the readers to (Esposito et al., 2021; Chu & Raginsky, 2023).

Let U be an $\mathcal{F}$-measurable random variable. then the Luxemburg norm of U with respect to μ is defined as

$$\|U\|_\psi^\mu = \inf \left\{ \sigma > 0 : \mathbf{E}_\mu \left[\psi\Big(\frac{|U|}{\sigma} \Big) \right] \leq 1 \right\},$$

and the Amemiya norm of U with respect to μ is defined as

$$\|U\|_{\psi}^{A,\mu} = \inf\left\{\frac{\mathbf{E}_{\mu}\left[\psi\big(t|U|\big)\right]+1}{t} : t>0\right\}.$$

Then we introduce the following lemma:

Lemma D.1 ((Hudzik & Maligranda, 2000)). *Let ψ be an Orlicz function and $\psi^{\star}$ denote its conjugate, then for every couple of random variable U, V, we have*

$$\mathbf{E}[UV] \leq \|U\|_{\psi}\|V\|_{\psi^{\star}}^{A}.$$

Note this Lemma D.1 recovers the Hölder's inequality by taking $\psi(t) = t^{\alpha}/\alpha$ (which consequently gives $\psi^{\star}(t) = t^{\beta}/\beta$ with $1/\alpha+1/\beta = 1$). One proof can be found in (Esposito et al., 2021, Appendix A).

D.2 Proof of our Theorem 4.2

We then provide formal proof of our Theorem 4.2.

For probability distributions P and Q such that $P \ll Q$, for a measurable set E, fix $\gamma \in \mathbb{R}$, and let ψ be an Orlicz function with convex conjugate $\psi^{\star}$. Recall we denote by $\|\cdot\|_{\psi}^{Q}$ the Luxemburg norm and by $\|\cdot\|_{\psi^{\star}}^{A,Q}$ the Amemiya norm.

For a measurable set E, we have

$$\begin{aligned}
P(E) &= \int \mathbb{1}_E \,\mathrm{d}P \\
&= \int \mathbb{1}_E \,\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q \\
&= \int \mathbb{1}_E \left(\min\left\{\frac{\mathrm{d}P}{\mathrm{d}Q},\gamma\right\} + \left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}\right)\mathrm{d}Q \\
&= \int \mathbb{1}_E \min\left\{\frac{\mathrm{d}P}{\mathrm{d}Q},\gamma\right\}\mathrm{d}Q + \int \mathbb{1}_E \left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}\mathrm{d}Q \\
&\overset{(a)}{\leq} \gamma\, Q(E) + \int \mathbb{1}_E \left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}\mathrm{d}Q \\
&\overset{(b)}{\leq} \gamma\, Q(E) + \left\|\mathbb{1}_E\right\|_{\psi}^{Q}\left\|\left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}\right\|_{\psi^{\star}}^{A,Q},
\end{aligned}$$

where (a) is because for every $s \in \mathcal{S}$, $0 \leq \mathbb{1}_E(x)\min\{\frac{\mathrm{d}P}{\mathrm{d}Q}(x),\gamma\} \leq \mathbb{1}_E(x)\gamma$, and integrating both sides w.r.t Q gives $\int \mathbb{1}_E \min\{\frac{\mathrm{d}P}{\mathrm{d}Q},\gamma\}\mathrm{d}Q \leq \int \mathbb{1}_E(x)\gamma \mathrm{d}Q = \gamma Q(E)$; (b) is by the generalized Hölder's inequality in Lemma D.1, which is applied to $U = \mathbb{1}_E$ and $V = \left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}$.

This proves the Orlicz form

$$P(E) \leq \gamma Q(E) + \left\|\mathbb{1}_E\right\|_{\psi}^{Q}\left\|\left[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\right]_{+}\right\|_{\psi^{\star}}^{A,Q}. \tag{29}$$

Then we note that the Luxemburg norm of the indicator has the explicit form

$$\|\mathbb{1}_E\|_{\psi}^{Q} = \frac{1}{\psi^{-1}(1/Q(E))}$$

where the generalized inverse is

$$\psi^{-1}(s) := \inf\left\{t \geq 0 : \psi(t) \geq s\right\}$$

for $s \geq 0$ and with the convention that $1/0 = \infty$ and $\psi^{-1}(\infty) = \infty$.

D.3 Further Generalizations

Moreover, if we employ Hölder's inequality, which is also used by (Chu & Raginsky, 2023; Esposito et al., 2021), we can further generalize Proposition 2.1. Theorem 4.2 implies the following theorem.

Theorem D.2. *Assume $P_{SW} \ll P_S P_W$. Given $E \in \mathcal{F}$ and two Orlicz functions ψ, φ, for any $\gamma \in \mathbb{R}$,*

$$P_{SW}(E) \le \gamma\, P_S P_W(E) + \left\| \left\| \mathbb{1}_{W \in E_W} \right\|_{\varphi}^{P_S} \right\|_{\psi}^{P_W} \cdot \left\| \left\| \left[\frac{\mathrm{d}P_{SW}}{\mathrm{d}P_S P_W} - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S} \right\|_{\psi^\star}^{A,P_W}, \tag{30}$$

where for each $w \in \mathcal{W}$, $E_w := \{s : (s,w) \in E\}$ (i.e., the "fiber" of E with respect to w).

The proof of Theorem D.2 can be found in Appendix D.4. It strictly generalizes (Esposito et al., 2021, Theorem 1) (and recovers it by taking $\gamma = 0$), which in turn implied various generalization error bounds. We strictly improve them when the reduction of the Amemiya term outweighs the additive $\gamma\, P_S P_W(E)$, usually in the scenarios where $P_S P_W(E)$ is tiny but $\mathrm{d}P_{SW}/\mathrm{d}P_S P_W$ has heavy tails.

D.4 Proof of Theorem D.2

Proof. Fix $\gamma \ge 0$, assume $P_{SW} \ll P_S P_W$ and denote $L(s,w) := \frac{\mathrm{d}P_{SW}}{\mathrm{d}(P_S P_W)}(s,w)$. By Tonelli's theorem,

$$\int_{\mathcal{S}} L(s,w)\, \mathrm{d}P_S(s) = 1 \quad \text{for } P_W\text{-a.e. } w.$$

For such w, define probability measure P_w on $(\mathcal{S}, \mathcal{F}_{\mathcal{S}})$ by

$$\mathrm{d}P_w(s) := L(s,w)\, \mathrm{d}P_S(s)$$

so that $\frac{\mathrm{d}P_w}{\mathrm{d}P_S}(s) = L(s,w)$. Given measurable $E \subseteq \mathcal{S} \times \mathcal{W}$, denote $E_w := \{s \in \mathcal{S} : (s,w) \in E\}$.

We apply Theorem 4.2 on the space $(\mathcal{S}, P_S)$ with Orlicz function φ, taking

$$P = P_w, \quad Q = P_S, \quad E = E_w,$$

to get, for P_W-a.e. w,

$$P_w(E_w) \le \gamma\, P_S(E_w) + \left\| \mathbb{1}_{E_w} \right\|_{\varphi}^{P_S} \left\| \left[L(\cdot, w) - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S}. \tag{31}$$

Then we integrate over w by using $P_{SW} = L \cdot (P_S P_W)$ and the definition of P_w, and get

$$P_{SW}(E) = \int_{\mathcal{W}} \int_{\mathcal{S}} \mathbb{1}_{E_w}(s)\, L(s,w)\, \mathrm{d}P_S(s)\, \mathrm{d}P_W(w) = \int_{\mathcal{W}} P_w(E_w)\, \mathrm{d}P_W(w).$$

Integrating (31) over w yields

$$P_{SW}(E) \le \gamma \int_{\mathcal{W}} P_S(E_w)\, \mathrm{d}P_W(w) + \int_{\mathcal{W}} \left\| \mathbb{1}_{E_w} \right\|_{\varphi}^{P_S} \left\| \left[L(\cdot, w) - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S} \mathrm{d}P_W(w). \tag{32}$$

By Fubini's theorem,

$$\int_{\mathcal{W}} P_S(E_w)\, \mathrm{d}P_W(w) = \int_{\mathcal{S} \times \mathcal{W}} \mathbb{1}_E(s,w)\, \mathrm{d}(P_S P_W)(s,w) = P_S P_W(E).$$

Substitute this into (32).

We then apply the generalized Hölder inequality on $(\mathcal{W}, P_W)$ with Orlicz function ψ to

$$w \mapsto \left\| \mathbb{1}_{E_w} \right\|_{\varphi}^{P_S}, \qquad w \mapsto \left\| \left[L(\cdot, w) - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S},$$

to obtain

$$\int_{\mathcal{W}} \left\| \mathbb{1}_{E_w} \right\|_{\varphi}^{P_S} \left\| \left[L(\cdot, w) - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S} \mathrm{d}P_W(w) \le \left\| \left\| \mathbb{1}_{S \in E_W} \right\|_{\varphi}^{P_S} \right\|_{\psi}^{P_W} \left\| \left\| \left[L - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S} \right\|_{\psi^\star}^{A,P_W}.$$

Combining the last two displays gives

$$P_{SW}(E) \le \gamma P_S P_W(E) + \left\| \left\| \mathbb{1}_{S \in E_W} \right\|_{\varphi}^{P_S} \right\|_{\psi}^{P_W} \left\| \left\| \left[L - \gamma \right]_+ \right\|_{\varphi^\star}^{A,P_S} \right\|_{\psi^\star}^{A,P_W},$$

and taking infimum over $\gamma \ge 0$ concludes the proof. □

E Results on Maximal Leakage and α-Mutual

E.1 Generalization Bounds via Maximal Leakage and α-Mutual Information

As shown in (16), maximal leakage (Issa et al., 2016), as a special case of α-mutual information (Verdú, 2015), can be used to bound the generalization error (Esposito et al., 2021). This provides desirable properties, including an *exponentially* decaying probability of generalization error and independence from the sample distribution, which is useful in analyzing additive-noise settings.

Directly employing our Proposition 2.1 recovers the maximal leakage bound (16) *exactly*, with simplified analysis, while Hellström & Durisi (2020) can only recover it up to a logarithmic term.

Corollary E.1. *For $P_{SW} \ll P_S P_W$, E measurable, $w \in \mathcal{W}$ and $M(w) = \operatorname{ess\,sup}_{s\sim P_S} \frac{\mathrm{d}P_{S|W=w}}{\mathrm{d}P_S}(s)$,*

$$P_{SW}(E) \le \mathbf{E}_{P_W}\big[M(W)P_S(E_W)\big] \le \Big(\operatorname*{ess\,sup}_{w\sim P_W} P_S(E_w)\Big)\,\exp\big(\mathcal{L}(S\to W)\big). \tag{33}$$

The proof is given in Appendix E. The second inequality in (33) is the known maximal-leakage bound of Esposito et al. (2021, Corollary 3); the first inequality is the sharper event-dependent intermediate form before taking the worst-case fiber. Interestingly, similar to f-divergences, α-mutual information also satisfies the data processing inequality. We can use a similar DPI approach to recover the α-mutual information bound (Esposito et al., 2021, Corollary 1) with a simple analysis, and hence also recover (16). Suppose P_{Y_α} is the minimizer of $I_\alpha(X,Y)$. Denoting $p := P_{XY}(E)$ and $q_\alpha := (P_X P_{Y_\alpha})(E)$, the DPI gives

$$I_\alpha(X,Y) \ge D_\alpha\big(\mathrm{Ber}(p)\|\mathrm{Ber}(q_\alpha)\big) = 1/(\alpha-1)\cdot\log\big(p^\alpha q_\alpha^{1-\alpha} + (1-p)^\alpha(1-q_\alpha)^{1-\alpha}\big), \tag{34}$$

and $\alpha\to\infty$ gives $I_\infty(X;Y) \ge \log(p/q_\infty)$, which recovers Corollary E.1 (see Appendix E.5).

E.2 Proof of Corollary E.1

Proof. Fix $w \in \mathcal{W}$. Apply Proposition 2.1 with $P = P_{S|W=w}$, $Q = P_S$ and $E = E_w$. For any $\gamma \in \mathbb{R}$,

$$P_{S|W=w}(E_w) \le \gamma\, P_S(E_w) + E_\gamma(P_{S|W=w}\|P_S). \tag{35}$$

Let $\gamma = M(w)$, we have $E_{M(w)}(P_{S|W=w}\|P_S) = 0$ and (35) yields $P_{S|W=w}(E_w) \le M(w)\,P_S(E_w)$. Taking expectation over W gives the first inequality of (33) and $P_S(E_W) \le \operatorname{ess\,sup}_w P_S(E_w)$ almost surely yields the second inequality, since $\mathbf{E}_{P_W}[M(W)] = \exp(\mathcal{L}(S\to W))$. □

E.3 Results on α-Mutual Information

Since maximal leakage is a special α-mutual information (Verdú, 2015) when $\alpha\to\infty$ (defined in (8)), we can use Theorem 4.2 to recover the α-mutual information bound (Esposito et al., 2021, Corollary 1) with a simple analysis, as shown in Corollary E.2. The proof is similar to that of Corollary E.1 and is in Appendix E.4. Note that (36) recovers $M(w)$ in Corollary E.1 when $\alpha\to\infty$.

Corollary E.2. *Let $P_{SW} \ll P_S P_W$. Fix $\alpha > 1$ and $\tau := \frac{\alpha}{\alpha-1}$, for any $w \in \mathcal{W}$ and measurable E,*

$$P_{SW}(E) \le \mathbf{E}_{P_W}\big[M_\alpha(W)\,P_S(E_W)^{1/\tau}\big] \tag{36}$$

$$\le \Big(\operatorname*{ess\,sup}_{w\sim P_W} P_S(E_w)\Big)^{1/\tau}\,\exp\big((\alpha-1)/\alpha\cdot I_\alpha(S,W)\big), \tag{37}$$

where $M_\alpha(w) := \big(\mathbf{E}_{P_S}\big[\big(\mathrm{d}P_{S|W=w}/\mathrm{d}P_S(s)\big)^\alpha\big]\big)^{1/\alpha}$.

E.4 Proof of Corollary E.2

In this section we show how to use Theorem D.2 to prove (Esposito et al., 2021, Corollary 1), with a simple analysis. The idea is similar to the proof of Corollary E.1.

Proof. Fix $w \in \mathcal{W}$. Apply Theorem 4.2 with $P = P_{S|W=w}, Q = P_S, E = E_w, \gamma = 0$, and choose the Orlicz function $\psi(t) := \frac{t^\kappa}{\kappa}$ (so that $\psi^\star(u) = \frac{u^\alpha}{\alpha}$). Then for this fixed w, Theorem 4.2 gives

$$P_{S|W=w}(E_w) \le \frac{1}{\psi^{-1}(1/P_S(E_w))} \left\| \frac{\mathrm{d}P_{S|W=w}}{\mathrm{d}P_S} \right\|_{\psi^\star}^{A,P_S}. \tag{38}$$

We can evaluate and derive

$$\frac{1}{\psi^{-1}(1/P_S(E_w))} = P_S(E_w)^{1/\kappa}\, \kappa^{-1/\kappa},$$
$$\left\| \frac{\mathrm{d}P_{S|W=w}}{\mathrm{d}P_S} \right\|_{\psi^\star}^{A,P_S} = \kappa^{1/\kappa} \left(\mathbf{E}_{P_S}\Big[\Big(\frac{\mathrm{d}P_{S|W=w}}{\mathrm{d}P_S}(S) \Big)^\alpha \Big] \right)^{1/\alpha} = \kappa^{1/\kappa} M_\alpha(w).$$

Plugging these two identities into (38), we obtain (36):

$$P_{S|W=w}(E_w) \le M_\alpha(w)\, P_S(E_w)^{1/\kappa}.$$

Taking expectation over W yields

$$P_{SW}(E) = \mathbf{E}_{P_W}\big[P_{S|W}(E_W)\big] \le \mathbf{E}_{P_W}\Big[M_\alpha(W)\, P_S(E_W)^{1/\kappa}\Big].$$

To prove (37), note that $P_S(E_W)^{1/\kappa} \le \big(\operatorname{ess\,sup}_{w\sim P_W} P_S(E_w)\big)^{1/\kappa}$ almost surely, hence

$$P_{SW}(E) \le \Big(\operatorname*{ess\,sup}_{w\sim P_W} P_S(E_w)\Big)^{1/\kappa} \mathbf{E}_{P_W}[M_\alpha(W)],$$

and by observing

$$\mathbf{E}_{P_W}[M_\alpha(W)] = \exp\Big(\frac{\alpha-1}{\alpha} I_\alpha(S,W)\Big),$$

we derive (37). □

E.5 Corollary E.1 by Data Processing Inequality of α-Mutual Information

Let P_{W_∞} be a minimizer of $Q_W \mapsto D_\infty(P_{SW}\|P_S Q_W)$, so that

$$I_\infty(S;W) = D_\infty(P_{SW}\|P_S P_{W_\infty}).$$

For any event E, with $p := P_{SW}(E)$ and $q_\infty := (P_S P_{W_\infty})(E)$, (34) gives

$$I_\infty(S;W) \ge D_\infty(\mathrm{Ber}(p)\|\mathrm{Ber}(q_\infty)) = \log\max\left\{\frac{p}{q_\infty}, \frac{1-p}{1-q_\infty}\right\}, \tag{39}$$

and in particular

$$p \le e^{I_\infty(S;W)}\, q_\infty. \tag{40}$$

Define $m(w) := \operatorname{ess\,sup}_{s\sim P_S} \frac{\mathrm{d}P_{W|S=s}}{\mathrm{d}P_W}(w)$. It is standard that $I_\infty(S;W) = \mathcal{L}(S\to W)$ and that a minimizer P_{W_∞} is given by

$$\frac{\mathrm{d}P_{W_\infty}}{\mathrm{d}P_W}(w) = \frac{m(w)}{\mathbf{E}_{P_W}[m(W)]},$$

and hence $P_{W_\infty} \ll P_W$. Apply (40) with $I_\infty(S;W) = \mathcal{L}(S\to W)$ to obtain

$$P_{SW}(E) \le e^{\mathcal{L}(S\to W)}\, (P_S P_{W_\infty})(E) = e^{\mathcal{L}(S\to W)} \int P_S(E_w)\, \mathrm{d}P_{W_\infty}(w).$$

Since $P_{W_\infty} \ll P_W$, the bound $P_S(E_w) \le \operatorname{ess\,sup}_{w\sim P_W} P_S(E_w)$ holds P_{W_∞}-a.s., so

$$\int P_S(E_w)\, \mathrm{d}P_{W_\infty}(w) \le \operatorname*{ess\,sup}_{w\sim P_W} P_S(E_w).$$

Combining the last two displays yields the desired result.

F Proof and Discussions of Corollary 5.2

In this section, we prove Corollary 5.2 by using the change of measure inequality from Section 4 as follows:

$$P(E)^\beta Q(E)^{1-\beta} + (1-P(E))^\beta (1-Q(E))^{1-\beta} \le 1 + (\beta-1)\mathcal{H}_\beta(P\|Q). \tag{41}$$

For any fixed $S = s$, let

$$a := \sqrt{\frac{n}{2\sigma^2}}, \qquad \alpha := 1 - \frac{1}{\beta} = \frac{\beta-1}{\beta},$$

and define

$$A(s) := 1 + (\beta-1)\mathcal{H}_\beta(P_{W|S=s}\|P_W).$$

For each $y \in \mathbb{R}$, let

$$E_y(s) := \{w : a\,\mathrm{gen}(s,w) > y\}.$$

Taking $P = P_{W|S=s}$, $Q = P_W$, and $E = E_y(s)$ in (41), we obtain

$$P_{W|S=s}(E_y(s))^\beta P_W(E_y(s))^{1-\beta} \le A(s).$$

Hence,

$$P_{W|S=s}(E_y(s)) \le A(s)^{1/\beta} P_W(E_y(s))^{1-1/\beta}. \tag{42}$$

Using the tail integral representation, for any fixed $S = s$, we have

$$\begin{aligned}
&\frac{\mathbf{E}_{P_{W|S=s}}\left[\exp\left(a\,\mathrm{gen}(s,W)\right)\right]}{A(s)^{1/\beta}} \\
&= \int_{-\infty}^{\infty} \frac{P_{W|S=s}\left(a\,\mathrm{gen}(s,W) > y\right)}{A(s)^{1/\beta}} e^y \,\mathrm{d}y \\
&\le \int_{-\infty}^{\infty} P_W\left(a\,\mathrm{gen}(s,W) > y\right)^{1-1/\beta} e^y \,\mathrm{d}y,
\end{aligned} \tag{43}$$

where the last step follows from (42).

By Jensen's inequality, we also have

$$\exp\left(\mathbf{E}_{P_{W|S=s}}\left[a\,\mathrm{gen}(s,W)\right]\right) \le \mathbf{E}_{P_{W|S=s}}\left[\exp\left(a\,\mathrm{gen}(s,W)\right)\right].$$

Combining this with (43) gives

$$\begin{aligned}
&\frac{\exp\left(\mathbf{E}_{P_{W|S=s}}\left[a\,\mathrm{gen}(s,W)\right]\right)}{A(s)^{1/\beta}} \\
&\le \int_{-\infty}^{\infty} P_W\left(a\,\mathrm{gen}(s,W) > y\right)^{1-1/\beta} e^y \,\mathrm{d}y.
\end{aligned} \tag{44}$$

Take expectation with respect to $S \sim P_S$. For $y \ge 0$, by the σ-sub-Gaussian assumption,

$$\mathbf{E}_{P_S}\left[P_W\left(a\,\mathrm{gen}(S,W) > y\right)\right] = P_S P_W\left(a\,\mathrm{gen}(S,W) > y\right) \le e^{-y^2}.$$

Since $x \mapsto x^{1-1/\beta}$ is concave on $[0,\infty)$, we have

$$\mathbf{E}_{P_S}\left[P_W\left(a\,\mathrm{gen}(S,W) > y\right)^{1-1/\beta}\right] \le e^{-\left(1-\frac{1}{\beta}\right)y^2}, \qquad y \ge 0.$$

For $y < 0$, we simply use

$$P_W\left(a\,\mathrm{gen}(S,W) > y\right)^{1-1/\beta} \le 1.$$

Therefore, from (44),

$$\begin{aligned}
&\mathbf{E}_{P_S}\left[\frac{\exp\left(\mathbf{E}_{P_{W|S}}\left[a\,\mathrm{gen}(S,W)\right]\right)}{\left(1+(\beta-1)\mathcal{H}_\beta(P_{W|S}\|P_W)\right)^{1/\beta}}\right] \\
&\le \int_{-\infty}^{0} e^y \,\mathrm{d}y + \int_0^{\infty} \exp\left(-\left(1-\frac{1}{\beta}\right)y^2 + y\right)\mathrm{d}y \\
&= 1 + \int_0^{\infty} \exp\left(-\frac{\beta-1}{\beta}y^2 + y\right)\mathrm{d}y.
\end{aligned} \tag{45}$$

Furthermore,

$$1+\int_0^\infty \exp\left(-\frac{\beta-1}{\beta}y^2+y\right)\mathrm{d}y \le 1+\int_{-\infty}^\infty \exp\left(-\frac{\beta-1}{\beta}y^2+y\right)\mathrm{d}y$$
$$= 1+\sqrt{\frac{\pi\beta}{\beta-1}}\exp\left(\frac{\beta}{4(\beta-1)}\right)$$
$$\le 2\sqrt{\frac{\pi\beta}{\beta-1}}\exp\left(\frac{\beta}{4(\beta-1)}\right). \tag{46}$$

Combining (45) and (46), we obtain

$$\mathbf{E}_{P_S}\left[\frac{\exp\left(\mathbf{E}_{P_{W|S}}\left[a\,\mathrm{gen}(S,W)\right]\right)}{\left(1+(\beta-1)\mathcal{H}_\beta(P_{W|S}\|P_W)\right)^{1/\beta}}\right]$$
$$\le 2\sqrt{\frac{\pi\beta}{\beta-1}}\exp\left(\frac{\beta}{4(\beta-1)}\right). \tag{47}$$

Applying Markov's inequality to the nonnegative random variable $\frac{\exp\left(\mathbf{E}_{P_{W|S}}[a\,\mathrm{gen}(S,W)]\right)}{\left(1+(\beta-1)\mathcal{H}_\beta(P_{W|S}\|P_W)\right)^{1/\beta}}$, we have that, with probability at least $1-\delta$,

$$\exp\left(\mathbf{E}_{P_{W|S}}\left[a\,\mathrm{gen}(S,W)\right]\right) \le \frac{2}{\delta}\left(1+(\beta-1)\mathcal{H}_\beta(P_{W|S}\|P_W)\right)^{1/\beta}\sqrt{\frac{\pi\beta}{\beta-1}}\exp\left(\frac{\beta}{4(\beta-1)}\right).$$

Taking logarithm on both sides and recalling that $a=\sqrt{n/(2\sigma^2)}$, we obtain

$$\mathbf{E}_{P_{W|S}}[\mathrm{gen}(S,W)]$$
$$\le \sqrt{\frac{2\sigma^2}{n}}\left(\log 2+\log\sqrt{\frac{\pi\beta}{\beta-1}}+\frac{\beta}{4(\beta-1)}+\log\left(\frac{\left(1+(\beta-1)\mathcal{H}_\beta(P_{W|S}\|P_W)\right)^{1/\beta}}{\delta}\right)\right).$$

G Generalization Error Bounds via Conditional Mutual Information

Conditional Mutual Information (CMI) framework was proposed by Steinke & Zakynthinou (2020) for generalization error analysis (also see recent works (Wang & Mao, 2024; Sefidgaran et al., 2025)). Unlike mutual information, which can easily be infinite even in settings where generalization is easy to prove, conditional mutual information is always finite.

In this section, we employ a strategy that is similar to (Hellström & Durisi, 2020), and derive CMI bounds from our change of measure inequalities. We make the same assumption as (Hellström & Durisi, 2020) that the loss function $\ell(\cdot,\cdot)$ is bounded on $[a,b]$ instead of in this subsection.

Let $\tilde{Z}=(Z_1,\ldots,Z_{2n})$ be the super-sample containing $2n$ i.i.d. tranining samples generated from P_Z, $S=(S_1,\ldots,S_n)\in\{0,1\}^n$ be a random selection vector that is independent of $\tilde{Z}$, and W be the output of a learning algorithm that may depend on $\tilde{Z}$. Use $Z(S)$ to denote the subset of $\tilde{Z}$ obtained from S by taking $Z_i(S_i)=\tilde{Z}_{i+S_i n}$ for $i=1,\ldots,n$. We then have $P_{W|Z(S)}$ as a stochastic learning algorithm, where W and $(\tilde{Z},S)$ are conditionally independent given $Z(S)$.

Define

$$\widehat{\mathrm{gen}}(W,\tilde{Z},S) := \frac{1}{n}\sum_{i=1}^n\left(\ell(W,Z_i(\bar{S}_i))-\ell(W,Z_i(S_i))\right), \tag{48}$$

where $\bar{S}$ is a vector whose entries are modulo-2 complements of the entries of S. As a result, $Z(\bar{S})$ contains all the elements of $\tilde{Z}$ that are not included in $Z(S)$. One can observe that the quantities like (48) are what being empirically calculated in the generalization performance assessment of an

algorithm in practice. Hellström & Durisi (2020, Theorem 3), provides a way to convert the bound based on (48) to a generalization bound: under $P_{W\tilde{Z}S}$, if with probability at least $1-\delta$ we

$$\widehat{\mathrm{gen}}(W,\tilde{Z},S) \le \epsilon(\delta)$$

holds with probability at least $1-\delta$, then also with probability at least $1-\delta$ we have

$$\big|\mathrm{gen}(W,Z(S))\big| \le \epsilon\left(\delta/2\right) + \sqrt{((b-a)^2/(2n))\cdot\log(4/\delta)}.$$

We then present bounds for $\widehat{\mathrm{gen}}(W,\tilde{Z},S)$ using Theorem 4.2. Note $P := P_{W\tilde{Z}S}$, $Q := P_{W|\tilde{Z}}P_{\tilde{Z}S}$.

Theorem G.1. *Let ψ be an Orlicz function and $\psi^\star$ its conjugate. For any $\gamma\in\mathbb{R}$ and $\eta>0$,*

$$\mathbf{P}_{P_{W\tilde{Z}S}}\big(\big|\widehat{\mathrm{gen}}(W,\tilde{Z},S)\big|\ge\eta\big) \le \frac{1}{\psi^{-1}\big(\frac{1}{2}\exp\big(\frac{n\eta^2}{2(b-a)^2}\big)\big)}\Big\|\Big[\frac{\mathrm{d}P}{\mathrm{d}Q}-\gamma\Big]_+\Big\|_{\psi^\star}^{A,Q} + 2[\gamma]_+\exp\Big(\frac{-n\eta^2}{2(b-a)^2}\Big).$$

The proof of Theorem G.1 is in Appendix G.1, and it can be specialized by Proposition 2.1 as follows.

Corollary G.2. *Fix any $\gamma\in\mathbb{R}$. Then for every $\eta>0$,*

$$\mathbf{P}_{P_{W\tilde{Z}S}}\Big(\big|\widehat{\mathrm{gen}}(W,\tilde{Z},S)\big|\ge\eta\Big) \le E_\gamma\big(P\big\|Q\big) + 2[\gamma]_+\exp\big(-n\eta^2/(2(b-a)^2)\big).$$

Similar to the use of Theorem 4.2, we can also employ other change of measure inequalities in Section 4 to derive CMI bounds in terms of other measures. They are omitted due to limited space.

G.1 Proof of Theorem G.1

Proof. We apply Theorem 4.2 with

$$P := P_{W\tilde{Z}S},\quad Q := P_{W|\tilde{Z}}P_{\tilde{Z}S},\quad E := \{(W,\tilde{Z},S) : \big|\widehat{\mathrm{gen}}(W,\tilde{Z},S)\big|\ge\eta\}.$$

Similar to (Hellström & Durisi, 2020), define the *fiber* of E with respect to $(W,\tilde{Z})$ by

$$E_{W\tilde{Z}} := \{s\in\mathcal{S} : (W,\tilde{Z},s)\in E\}.$$

Since under $Q = P_{W|\tilde{Z}}P_{\tilde{Z}S}$ we have $W\perp S|\tilde{Z}$, it follows that

$$\begin{aligned} Q(E) &= \mathbf{E}_Q\big[\mathbb{1}\{(W,\tilde{Z},S)\in E\}\big] \\ &= \mathbf{E}_{P_{\tilde{Z}}}\Big[\mathbf{E}_{P_{W|\tilde{Z}}}\big[P_{S|\tilde{Z}}(E_{W\tilde{Z}}|\tilde{Z})\big]\Big] = \mathbf{E}_{P_{W\tilde{Z}}}\big[P_{S|\tilde{Z}}(E_{W\tilde{Z}}|\tilde{Z})\big]. \end{aligned} \tag{49}$$

Since S is independent of $\tilde{Z}$, under $S\sim P_{S|\tilde{Z}=\tilde{z}}$ the coordinates $S_1,\dots,S_n$ are i.i.d. Bernoulli$(1/2)$. For each fixed $(w,\tilde{Z})$, define

$$\Delta_i(w,\tilde{Z}) := \ell(w,\tilde{z}_{i+n}) - \ell(w,\tilde{z}_i),\qquad i=1,\dots,n.$$

Then

$$\widehat{\mathrm{gen}}(w,\tilde{z},S) = \frac{1}{n}\sum_{i=1}^n(1-2S_i)\Delta_i(w,\tilde{Z}).$$

Hence $\mathbf{E}_{S\sim P_{S|\tilde{Z}=\tilde{z}}}\big[\widehat{\mathrm{gen}}(w,\tilde{z},S)\big] = 0$, and each summand belongs to $\big[-\frac{b-a}{n},\frac{b-a}{n}\big]$ since $|\Delta_i(w,\tilde{Z})|\le b-a$. By Hoeffding's lemma, $\widehat{\mathrm{gen}}(w,\tilde{z},S)$ is $(b-a)/\sqrt{n}$-sub-Gaussian under $S\sim P_{S|\tilde{Z}=\tilde{z}}$. Therefore, for all $(w,\tilde{Z})$,

$$P_{S|\tilde{Z}=\tilde{z}}\big(\,|\widehat{\mathrm{gen}}(w,\tilde{z},S)|\ge\eta\,\big) \le 2\exp\left(-\frac{n\eta^2}{2(b-a)^2}\right).$$

Combining this uniform bound with (49) yields

$$Q(E) \le 2\exp\left(-\frac{n\eta^2}{2(b-a)^2}\right).$$

We now apply Theorem 4.2 with the above P, Q, E, and derive

$$\begin{aligned} P_{W\tilde{Z}S}(E) &\le \gamma\, Q(E) + \frac{1}{\psi^{-1}\big(1/Q(E)\big)} \left\| \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right]_+ \right\|_{\psi^\star}^{A,Q} \\ &\le [\gamma]_+\, Q(E) + \frac{1}{\psi^{-1}\big(1/Q(E)\big)} \left\| \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right]_+ \right\|_{\psi^\star}^{A,Q} \\ &\le 2[\gamma]_+ \exp\left(-\frac{n\eta^2}{2(b-a)^2}\right) + \frac{1}{\psi^{-1}\left(\frac{1}{2}\exp\left(\frac{n\eta^2}{2(b-a)^2}\right)\right)} \left\| \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right]_+ \right\|_{\psi^\star}^{A,Q}. \end{aligned}$$

This proves the desired result. □

H Proof of Lemma 6.2

Proof. Recall that

$$E_\gamma(P\|Q) = \sup_E \big(P(E) - \gamma Q(E)\big).$$

($\Rightarrow$) Assume $E_\gamma(P\|Q) \le \tau$. Then for every measurable event E,

$$P(E) - \gamma Q(E) \le \tau.$$

Fix any E such that $P(E) > \tau$. If $Q(E) = 0$, then the above inequality would give $P(E) \le \tau$, a contradiction. Hence $Q(E) > 0$ and we may divide to obtain

$$\frac{P(E) - \tau}{Q(E)} \le \gamma.$$

Taking the supremum over all E with $P(E) > \tau$ yields

$$\sup_{E: P(E) > \tau} \frac{P(E) - \tau}{Q(E)} \le \gamma,$$

which is equivalent to $I_\infty^\tau(P\|Q) \le \log\gamma$.

($\Leftarrow$) Assume $I_\infty^\tau(P\|Q) \le \log\gamma$, i.e.,

$$\sup_{E: P(E) > \tau} \frac{P(E) - \tau}{Q(E)} \le \gamma.$$

Fix any measurable event E. If $P(E) \le \tau$, then trivially $P(E) - \gamma Q(E) \le P(E) \le \tau$. If $P(E) > \tau$, then (by the definition of the supremum) we have $Q(E) > 0$ and

$$\frac{P(E) - \tau}{Q(E)} \le \gamma,$$

hence $P(E) \le \gamma Q(E) + \tau$, i.e. $P(E) - \gamma Q(E) \le \tau$. Since this holds for all E, taking the supremum over E gives $E_\gamma(P\|Q) \le \tau$. □

I Generalization Error Bounds and Differential Privacy

We compare our generalization error bounds in terms of $\chi^2(P\|Q)$, as displayed in Theorem 5.1, with the best-known generalization bound in terms of maximal leakage (Esposito et al., 2021), as follows.

J Average Generalization Error Bound

By Xu & Raginsky (2017), an average generalization bound in terms of $I(S;W)$ has been proved:

$$\mathbf{E}\left[\mathrm{gen}(S,W)\right] \le \sqrt{(2\sigma^2/n) I(S;W)}, \tag{50}$$

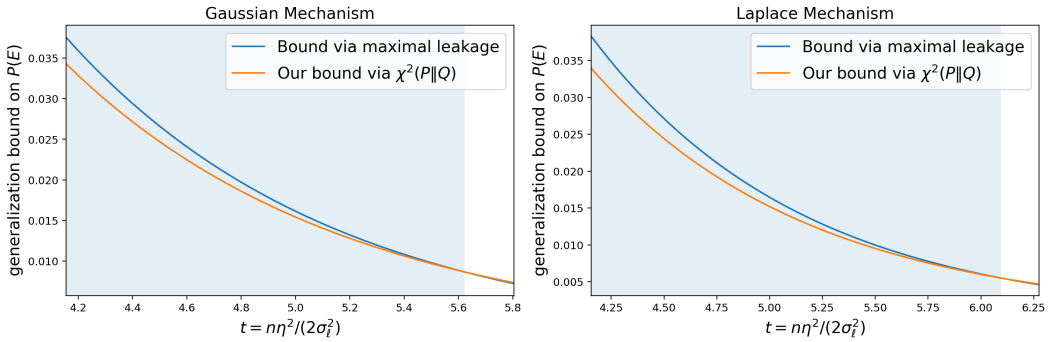


Figure 2: Comparison of generalization bounds of differential privacy algorithms, with binary skewed source distribution $\Pr(S = \Delta) = 10^{-3}$ and output $W = S + N$. Let $P = P_{SW}$, $Q = P_S P_W$, $Q(E) \le q_t = 2e^{-t}$ and $t = n\eta^2/(2\sigma_t^2)$. For both Gaussian ($N \sim \mathcal{N}(0, \tau^2)$, $\Delta/(2\tau) = 0.25$) and Laplace ($N \sim \mathrm{Lap}(b)$, $\Delta/(2b) = 0.25$) noises, bounds via $\chi^2(P\|Q)$ are **tighter** than the bounds via maximal leakage (Esposito et al., 2021).

In this section, we show that our change of measure inequalities can also be used to derive average generalization bounds, by deriving, as an example, a bound in terms of $I(S;W)$ that is close to (50) (up to a multiplicative constant) by using our Proposition 2.1, a simple change of measure inequality in terms of $E-\gamma$-divergence. We also show that our bound is strictly tighter than the one recovered by (Chu & Raginsky, 2023, Corollary 1).

If, instead of using Proposition 2.1, we use our other results (e.g., Theorem 4.2), then average generalization bounds in terms of other information measures can also be derived.

J.1 An Average Generalization Bound via $I(S;W)$

Corollary J.1. *For a learning algorithm $P_{W|S}$ with a σ-sub-Gaussian loss function,*

$$\mathbf{E}\left[|\mathrm{gen}(S,W)|\right] \le \frac{2\sigma}{\sqrt{n}}\left(2\sqrt{I(S;W)+2/e}+\sqrt{\pi}\right).$$

Proof. Take $P = P_{SW}$ and $Q = P_S P_W$, and $E = \{W, S : |\mathrm{gen}(S,W)| \ge \epsilon\}$, and we consider the σ-sub-Gaussianity of the loss function. By Hoeffding's inequality we know

$$Q(E) \le 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

Apply Proposition 2.1, for $\gamma > 0$, we have

$$\mathbf{P}\left(|\mathrm{gen}(S,W)| \ge \epsilon\right) \le 2\gamma\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) + E_\gamma(P_{SW}\|P_S P_W).$$

By (Sason & Verdú, 2016, Theorem 29) we know

$$E_\gamma(P_{SW}\|P_S P_W) \le \frac{I(S;W)+2/e}{\ln\gamma},$$

and therefore

$$\mathbf{P}\left(|\mathrm{gen}(S,W)| \ge \epsilon\right) \le 2\gamma\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) + \frac{I(S;W)+2/e}{\ln\gamma}. \tag{51}$$

We now need to convert the tail bound into an expectation bound.

Consider for nonnegative X we know

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X \ge \epsilon)\mathrm{d}\epsilon.$$

Take $X := |\mathrm{gen}(S, W)|$. To make (51) integrable, we choose γ depending on ϵ:

$$\gamma_\epsilon = \exp\left(\frac{n\epsilon^2}{4\sigma^2}\right),$$

and then we get from (51) that

$$\mathbf{P}\left(|\mathrm{gen}(S, W)| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{4\sigma^2}\right) + \frac{4\sigma^2}{n\epsilon^2}\left(I(S; W) + \frac{2}{e}\right). \tag{52}$$

We then integrate (52) as follows,

$$\begin{aligned}
&\mathbf{E}\left[|\mathrm{gen}(S, W)|\right] \\
&= \int_0^{\epsilon_0} \mathbf{P}\big(\,|\mathrm{gen}(S, W)| \geq \epsilon\big)\mathrm{d}\epsilon + \int_{\epsilon_0}^{\infty} \mathbf{P}\big(\,|\mathrm{gen}(S, W)| \geq \epsilon\big)\mathrm{d}\epsilon \\
&\leq \epsilon_0 + 2\int_0^{\infty} \exp\left(-\frac{n\epsilon^2}{4\sigma^2}\right)\mathrm{d}\epsilon + \frac{4\sigma^2}{n}\left(I(S; W) + \frac{2}{e}\right)\int_{\epsilon_0}^{\infty} \frac{1}{\epsilon^2}\mathrm{d}\epsilon.
\end{aligned} \tag{53}$$

We can then compute

$$2\int_0^{\infty} \exp\left(-\frac{n\epsilon^2}{4\sigma^2}\right)\mathrm{d}\epsilon = \frac{2\sigma\sqrt{\pi}}{\sqrt{n}},$$
$$\int_{\epsilon_0}^{\infty} \frac{1}{\epsilon^2}\mathrm{d}\epsilon = \frac{1}{\epsilon_0},$$

and therefore

$$\mathbf{E}\left[|\mathrm{gen}(S, W)|\right] \leq \epsilon_0 + \frac{2\sigma\sqrt{\pi}}{\sqrt{n}} + \frac{4\sigma^2}{n\epsilon_0}\left(I(S; W) + \frac{2}{e}\right), \tag{54}$$

on which we optimize the right-hand side over ϵ_0 by balancing the first and third terms and find

$$\epsilon_0^\star = 2\sigma\sqrt{\frac{(I(S; W) + 2/e}{n}}.$$

Plug above into (54) we get

$$\begin{aligned}
\mathbf{E}\left[|\mathrm{gen}(S, W)|\right] &\leq 4\sigma\sqrt{\frac{I(S; W) + 2/e}{n}} + \frac{2\sigma\sqrt{\pi}}{\sqrt{n}} \\
&= \frac{2\sigma}{\sqrt{n}}\left(2\sqrt{I(S; W) + \frac{2}{e}} + \sqrt{\pi}\right).
\end{aligned}$$

□

J.2 Comparison and Discussion

Note in the proof of Corollary J.1, for step (53) we calculated $\int_0^{\infty} \exp\left(-\frac{n\epsilon^2}{4\sigma^2}\right)\mathrm{d}\epsilon$ instead of calculating $\int_{\epsilon_0}^{\infty} \exp\left(-\frac{n\epsilon^2}{4\sigma^2}\right)\mathrm{d}\epsilon$; if we stick to the former, and keep the rest of the proof the same, result that is tighter but in a more complicated form can be derived:

$$\mathbf{E}\big[\,|\mathrm{gen}(S, W)|\,\big] \leq \frac{2\sigma}{\sqrt{n}}\left(2t^\star\big(1 - e^{-(t^\star)^2}\big) + \sqrt{\pi}\,\mathrm{erfc}(t^\star)\right),$$

where $t^\star$ is the unique solution of

$$(t^\star)^2\Big(1 - 2e^{-(t^\star)^2}\Big) = I(S; W) + \frac{2}{e}.$$

Comparing to what can be recovered by Chu & Raginsky (2023), Corollary J.1 is strictly tighter, and both of which are weaker than the original from by Xu & Raginsky (2017) up to some multiplicative constant factor, but still in the same order. Now we provide a detailed comparison.

In (Chu & Raginsky, 2023, Corollary 1) the following bound is recovered:

$$\mathbf{E}\left[|\mathrm{gen}(S,W)|\right] \le \frac{2\sigma}{\sqrt{n}}\sqrt{6\big(I(S;W)+4\big)}.$$

Since both share the factor $\sigma/\sqrt{n}$, it suffices to compare two right-hand sides in reduced forms

$$4\sqrt{I(S;W)+\frac{2}{e}}+2\sqrt{\pi} \quad \text{and} \quad \sqrt{24\big(I(S;W)+4\big)}.$$

Define their difference

$$\Delta := \sqrt{24\big(I(S;W)+4\big)} - 4\sqrt{I(S;W)+\frac{2}{e}} - 2\sqrt{\pi},$$

by taking derivative with respect to $I(S;W)$ and finding the global minimizer, we find at the global minimizer

$$\Delta = 2\Big(\sqrt{8-\frac{4}{e}}-\sqrt{\pi}\Big) \approx 1.5653 > 0,$$

and hence our Corollary J.1 is strictly tighter and the gap is uniformly around $1.5653\frac{\sigma}{\sqrt{n}}$. See Figure 3 for comparison.

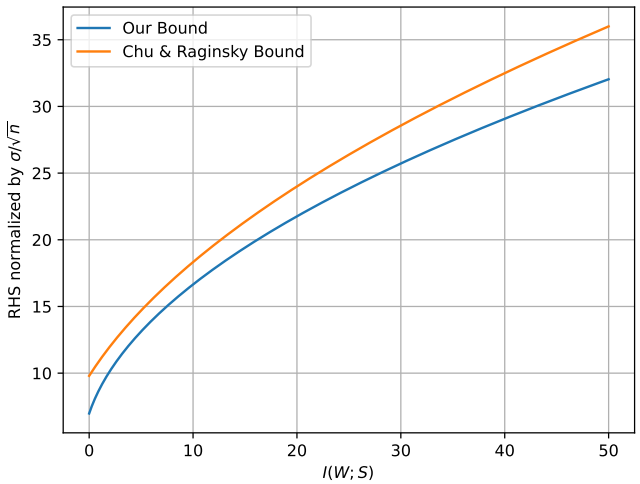


Figure 3: Comparison between Corollary J.1 and (Chu & Raginsky, 2023, Corollary 1).

K Details of Table 1

The proofs in this section are based on (3). We also discuss some details for the comparison between our bounds and the results in Section C.7.

K.1 Reverse Pearson $f(t) = t^{-1} - 1$

Consider the reverse Pearson divergence is an f-divergence with $f(t) = t^{-1} - 1$, which satisfies

$$D_f(P\|Q) = \int \left(\left(\frac{dP}{dQ}\right)^{-1} - 1\right) dQ = \chi^2(Q\|P).$$

Let $r := \chi^2(Q\|P)$, we have

$$r \ge q\Big(\frac{q}{p}-1\Big) + (1-q)\Big(\frac{1-q}{1-p}-1\Big) = \frac{q^2}{p} + \frac{(1-q)^2}{1-p} - 1.$$

Equivalently,

$$\frac{q^2}{p} + \frac{(1-q)^2}{1-p} \le 1 + r.$$

Multiplying by $p(1-p)$ gives

$$q^2 + (1-2q)p \le (1+r)p - (1+r)p^2,$$

i.e.

$$(1+r)p^2 - (r+2q)p + q^2 \le 0.$$

Thus p lies between the two roots of this quadratic, and in particular

$$p = P(E) \le \frac{r + 2q + \sqrt{r^2 + 4r\,q(1-q)}}{2(1+r)}, \qquad r = \chi^2(Q\|P). \tag{55}$$

K.2 Reverse KL $f(t) = -\log t$

Consider the reverse KL divergence is an f-divergence with $f(t) = -\log t$, so that

$$D_f(P\|Q) = D(Q\|P).$$

We have

$$D(Q\|P) \ge q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} =: \mathrm{kl}(q,p). \tag{56}$$

For fixed $q \in (0,1)$, the map $p \mapsto \mathrm{kl}(q,p)$ is strictly increasing on $[q,1)$. Hence, defining $p_+(q,d)$ as the unique solution in $[q,1)$ of $\mathrm{kl}(q,p) = d$, (56) implies the sharp inversion

$$p = P(E) \le p_+\big(q,\, D(Q\|P)\big), \qquad \text{where } p_+(q,d) \text{ solves } \mathrm{kl}(q,p) = d. \tag{57}$$

If one prefers an explicit (but one-sided) bound, dropping the first term in (56) yields

$$p \le 1 - (1-q)\exp\Big(-\frac{D(Q\|P)}{1-q}\Big). \tag{58}$$

K.3 Vincze-Le Cam divergence and the optimized quadratic bound

Consider the Vincze-Le Cam divergence, which is an f-divergence with $f(t) = (2-2t)/(t+1)$. Similar to above, we denote $p := P(E)$ and $q := Q(E)$ and calculate by the data-processing inequality:

$$\begin{aligned}
\mathrm{VC}(P\|Q) &\ge D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) \\
&= q\, f\Big(\frac{p}{q}\Big) + (1-q)\, f\Big(\frac{1-p}{1-q}\Big) \\
&= -2 + \frac{4q^2}{p+q} + \frac{4(1-q)^2}{2-p-q}, \\
&= \frac{2(p-q)^2}{(p+q)(2-p-q)}.
\end{aligned} \tag{59}$$

Let $V := \mathrm{VC}(P\|Q) \ge 0$. Then (59) gives

$$V(p+q)(2-p-q) \ge 2(p-q)^2.$$

Expanding and rearranging yields

$$(V+2)p^2 - 2(Vq+2q)p + Vq^2 + 2q^2 - 2Vq \le 0. \tag{60}$$

Thus p lies between the two roots of (60), and in particular

$$p \le \frac{V(1-q) + 2q + \sqrt{V\big(V + 8q(1-q)\big)}}{V+2}. \tag{61}$$

We now compare (61) with the following result from (Picard-Weibel & Guedj, 2022).

For $c > 0$, we first define the bound

$$p \le 2(1+c) - q - \frac{4\Big(q\sqrt{c} + (1-q)\sqrt{1+c}\Big)^2}{V+2} =: A(c). \tag{62}$$

We let $r := \sqrt{c/(1+c)}$ and hence (62) becomes

$$A(c) = h(r) := \frac{2 - \frac{4}{V+2}\big(1 - q + qr\big)^2}{1-r^2} - q, \qquad r \in (0,1). \tag{63}$$

By differentiating (63) one can find

$$r^\star = \frac{V + 4q(1-q) - \sqrt{V\big(V + 8q(1-q)\big)}}{4q(1-q)}. \tag{64}$$

and one can verify $r^\star$ is the unique global minimizer of h on $(0,1)$. Let $c^\star := \frac{(r^\star)^2}{1-(r^\star)^2}$ be the corresponding optimizer in (62). Then for all $c > 0$,

$$A(c) \ge A(c^\star) = h(r^\star). \tag{65}$$

Finally, evaluating $h(r^\star)$ using the defining quadratic relation of $r^\star$ yields

$$\min_{c>0} A(c) = h(r^\star) = \frac{V(1-q) + 2q + \sqrt{V\big(V + 8q(1-q)\big)}}{V+2}. \tag{66}$$

Hence we know that (61) is exactly the best bound obtainable from the family (62) by optimizing over $c > 0$.

L Comparison Table

In this section, we compare the exact Bernoulli event-DPI relations with the corresponding indicator-specialized bounds of Picard-Weibel & Guedj (2022). The purpose is structural comparison rather than a claim that every displayed Bernoulli identity is new or uniformly tighter. See Table 2.

M DPI and Optimality

We first record a standard pushforward/restricted-variational identity. We include it only to isolate the precise sharpness statement used in Section 4.1; the identity itself is not claimed as new.

Theorem M.1. *Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces, let P, Q be probability measures on $(\mathcal{X}, \mathcal{F})$, and let $\phi : \mathcal{X} \to \mathcal{Y}$ be measurable. Define the pushforwards*

$$P_\phi := P \circ \phi^{-1}, \qquad Q_\phi := Q \circ \phi^{-1}.$$

Define the restricted class

$$\mathcal{T}_\phi := \{\, T : \mathcal{X} \to \mathbb{R} : \; T = g \circ \phi \text{ for some measurable } g : \mathcal{Y} \to \mathbb{R} \,\}.$$

For $f : (0,\infty) \to \mathbb{R} \cup \{+\infty\}$ being convex with $f(1) = 0$ with its convex conjugate $f^*(t)$, we have

$$\sup_{T \in \mathcal{T}_\phi} \Big\{ \mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)] \Big\} = D_f(P_\phi \| Q_\phi).$$

Proof. Take any $T \in \mathcal{T}_\phi$. Then $T = g \circ \phi$ for some measurable $g : \mathcal{Y} \to \mathbb{R}$. Therefore

$$\mathbf{E}_P[T] = \mathbf{E}_P[g(\phi(X))] = \mathbf{E}_{P_\phi}[g].$$

Also,

$$\mathbf{E}_Q[f^*(T)] = \mathbf{E}_Q[f^*(g(\phi(X)))] = \mathbf{E}_{Q_\phi}[f^*(g)].$$

Table 2: Comparison between the exact Bernoulli event-DPI relations and the indicator-specialized bounds of Picard-Weibel & Guedj (2022) (PWG). The E_γ inequality is a standard variational consequence and is included to illustrate that the strict-convexity assumption in PWG does not cover the hockey-stick generator. Known rows and refinements are identified in Table 1 and Appendix C.7. Let $h_2(\cdot)$ be binary entropy, $c > 0$, $s \in \mathbb{R}$, and $q_\beta := \beta/(\beta-1)$. We sometimes omit $(P\|Q)$ or $(P;Q)$ after divergences.

f-div	$f(t)$	Change of Measure Inequalities
E_γ-div, $\gamma \geq 1$	$[t-\gamma]_+$	$p \leq \gamma\, q + E_\gamma(P\|Q)$
KL	$t \log t$	$p \leq \big(D_{\mathrm{KL}}(P\|Q) + \log\big(1 + q(e^c - 1)\big)\big)/c$
χ^2-div	$t^2 - 1$	$p \leq q + \sqrt{q\big(1-q\big)\chi^2}$
Power-β, $\beta > 1$	$(t^\beta - 1)/(\beta - 1)$	$p^\beta q^{1-\beta} + (1-p)^\beta (1-q)^{1-\beta} \leq 1 + (\beta-1)\mathcal{H}_\beta$
		PWG: $p \leq s + \big(1 + (\beta-1)\mathcal{H}_\beta\big)^{1/\beta}$ $\times \big(q(1-s)_+^{q_\beta} + (1-q)(-s)_+^{q_\beta}\big)^{1/q_\beta}$,
Squared Hellinger	$(1-\sqrt{t})^2$	$2\Big(1 - \sqrt{pq} - \sqrt{(1-p)(1-q)}\Big) \leq H^2$
		PWG: $p \leq 1 + c - \big(1 - H^2/2\big)^2 \Big(q/c + (1-q)/(1+c)\Big)^{-1}$
Reverse χ^2-div	$1/t - 1$	$(p-q)^2/(p(1-p)) \leq \chi^2(Q\|P)$
		PWG: $p \leq 1 + c - \dfrac{\Big(q\sqrt{c} + (1-q)\sqrt{1+c}\Big)^2}{1 + \chi^2(Q\|P)}, c > 0$
Reverse-KL	$-\log t$	$q \log(\frac{q}{p}) + (1-q)\log(\frac{1-q}{1-p}) \leq D_{\mathrm{KL}}(Q\|P)$
		PWG: $p \leq 1 + c - \exp\Big(q \log c + (1-q)\log(1+c) - D_{\mathrm{KL}}(Q\|P)\Big)$
Jensen-Shannon	$t \log \dfrac{2t}{1+t} + \log \dfrac{2}{1+t}$	$2h_2\left((p+q)/2\right) - h_2\big(p\big) - h_2\big(q\big) \leq \mathrm{JS}(P\|Q)$
		PWG: $p \leq 1 + c + \mathrm{JS}/(2\lambda) - \big(q\log(1 - e^{-2\lambda c})$ $+ (1-q)\log(1 - e^{-2\lambda(1+c)})\big)/(2\lambda)$
Vincze-Le Cam	$(2-2t)/(t+1)$	$2(p-q)^2/\big((p+q)\big(2-p-q\big)\big) \leq \mathrm{VC}$
		PWG: $p \leq 2(1+c) - q - 4\Big(q\sqrt{c} + (1-q)\sqrt{1+c}\Big)^2/\big(2 + \mathrm{VC}\big)$

Hence,

$$\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)] = \mathbf{E}_{P_\phi}[g] - \mathbf{E}_{Q_\phi}[f^*(g)].$$

Taking the supremum over all $T \in \mathcal{T}_\phi$ is therefore equivalent to taking the supremum over all measurable g on $\mathcal{Y}$. Thus,

$$\sup_{T\in\mathcal{T}_\phi}\Big\{\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)]\Big\} = \sup_g\Big\{\mathbf{E}_{P_\phi}[g] - \mathbf{E}_{Q_\phi}[f^*(g)]\Big\} = D_f(P_\phi\|Q_\phi).$$

Since $\mathcal{T}_\phi$ is a subclass of all measurable test functions on $\mathcal{X}$, we also have

$$D_f(P\|Q) = \sup_T\Big\{\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)]\Big\} \geq \sup_{T\in\mathcal{T}_\phi}\Big\{\mathbf{E}_P[T] - \mathbf{E}_Q[f^*(T)]\Big\}.$$

Combining the two displays yields

$$D_f(P\|Q) \geq D_f(P_\phi\|Q_\phi).$$

□

This theorem induces the following corollary.

Corollary M.2. *Let $\{A_1, \ldots, A_n\}$ be a measurable partition of $\mathcal{X}$. Restrict the variational class to functions of the form*

$$T = \sum_{i=1}^n a_i \mathbb{1}_{A_i}.$$

Then

$$\sup_{a_1,\ldots,a_n}\left\{\sum_{i=1}^n a_i P(A_i) - \sum_{i=1}^n Q(A_i) f^*(a_i)\right\} = D_f\big((P(A_i))_{i=1}^n \| (Q(A_i))_{i=1}^n\big).$$

Consequently,

$$D_f(P\|Q) \geq D_f\big((P(A_i))_{i=1}^n \| (Q(A_i))_{i=1}^n\big).$$

For the two-cell partition $\{E, E^c\}$, the preceding corollary gives exactly $D_f(\mathrm{Ber}(P(E))\|\mathrm{Ber}(Q(E)))$.

N Details on Data Memorization

N.1 Setup and Selector-Recovery Bounds

The framework is similar to the CMI setting in Appendix G. Let $\tilde{Z} = (Z_1, \ldots, Z_{2n})$ be i.i.d. from P_Z, and let $J = (J_1, \ldots, J_n)$ have i.i.d. $\mathrm{Bern}(1/2)$ entries, independent of $\tilde{Z}$. For $j \in \{0, 1\}$ and $i \in [n]$, define

$$Z_i(j) := \tilde{Z}_{i+jn}, \qquad Z(J) := (Z_1(J_1), \ldots, Z_n(J_n)).$$

The algorithm induces $P_{W|Z(J)}$. Let $Y = (W, \tilde{Z}, U)$, where U is independent auxiliary randomness, and let $\hat{J} = \phi(Y) \in \{0, 1\}^n$ be any estimator of J. For $b \in \{0, \ldots, n\}$, define

$$p_b := \mathbf{P}\big(d_H(\hat{J}, J) \leq b\big), \qquad q_b := 2^{-n}\sum_{k=0}^b \binom{n}{k}.$$

Let $P_0 := P_{W\tilde{Z}J}$ and $Q_0 := P_{W|\tilde{Z}}P_{\tilde{Z}J}$. Then $D(P_0\|Q_0) = I(W; J \mid \tilde{Z})$, and DPI applied to the event $\{d_H(\hat{J}, J) \leq b\}$ gives

$$D_f(P_0\|Q_0) \geq D_f\big(\mathrm{Ber}(p_b)\|\mathrm{Ber}(q_b)\big). \tag{67}$$

Specializing (67) to KL divergence, for every $c > 0$,

$$p_b \leq \big(I(W; J \mid \tilde{Z}) + \log\big(1 + q_b(e^c - 1)\big)\big)/c. \tag{68}$$

This recovers the Fano step of Sefidgaran et al. (2025, Theorem 5) by taking $c = \log(1/q_b)$, but the exact Bernoulli-KL expression permits optimization over c. For fixed $t \in [0, 1/2)$, let

$$b_t := \lfloor nt \rfloor, \qquad q_t := 2^{-n} \sum_{k=0}^{b_t} \binom{n}{k}, \qquad L_t := \log(1/q_t), \qquad I_n := I(W; J \mid \tilde{Z}),$$

and define

$$\Psi_n^\star(t) := \inf_{c>0} \frac{I_n + \log(1 + q_t(e^c - 1))}{c}. \tag{69}$$

Equivalently, by the binary-KL variational identity,

$$\Psi_n^\star(t) = \sup\{p \in [q_t, 1] : \mathrm{kl}(p\|q_t) \le I_n\}.$$

For fixed $t < 1/2$, $L_t = n(\log 2 - h_2(t)) + O(\log n)$. If $I_n = \Theta(1)$, then

$$\Psi_n^\star(t) = \frac{I_n}{L_t} + O\left(\frac{\log n}{n^2}\right), \tag{70}$$

whereas the Fano relaxation $(I_n + \log 2)/L_t$ contains an additional $\Theta(1/n)$ term. More strongly, if $I_n = \Theta(n^{-a})$ for any fixed $a > 0$, then

$$\Psi_n^\star(t) = (1 + o(1))\frac{I_n}{L_t} = \Theta(n^{-1-a}), \tag{71}$$

while the Fano relaxation remains $\Theta(1/n)$. The derivation is given below.

N.2 Discussions and Comparison on (Sefidgaran et al., 2025, Theorem 5)

We first justify the asymptotics above. The identity

$$\Psi_n^\star(t) = \inf_{c>0} \frac{I_n + \log(1 + q_t(e^c - 1))}{c} = \sup\{p \in [q_t, 1] : \mathrm{kl}(p\|q_t) \le I_n\}$$

is the exact binary-KL variational inversion. A Lambert-W choice can be used as a convenient feasible asymptotic choice, but it is not in general the exact optimizer.

For $p = o(1)$ and $q_t = e^{-L_t}$,

$$\mathrm{kl}(p\|q_t) = pL_t + p\log p - p + O(p^2 + q_t). \tag{72}$$

Together with $L_t = n(\log 2 - h_2(t)) + O(\log n)$, this gives (70) when $I_n = \Theta(1)$. If $I_n = \Theta(n^{-a})$ for a fixed $a > 0$, substituting $p = \kappa I_n/L_t$ into (72) gives $\mathrm{kl}(p\|q_t) = \kappa I_n(1 + o(1))$, which proves (71).

We next compare Theorem 7.1 with Sefidgaran et al. (2025, Theorem 5(ii)). Their result gives, under the same low-CMI condition and $\mathbf{P}(T_n \ge \alpha n) \ge q$,

$$\mathbf{P}(F_n \ge m_\epsilon) \ge (\alpha - \epsilon)q, \qquad m_\epsilon = \frac{\epsilon}{1/q + \epsilon - \alpha} n - o(n).$$

Our Theorem 7.1 gives, for every fixed $\beta < \alpha$,

$$\mathbf{P}(F_n > \beta n) \ge q - o(1).$$

Thus the probability level improves from $(\alpha - \epsilon)q$ to $q - o(1)$, and the false-positive threshold can be taken to be any linear coefficient $\beta < \alpha$. Since the leading coefficient of m_ϵ is strictly smaller than α, choosing

$$\frac{\epsilon}{1/q + \epsilon - \alpha} < \beta < \alpha$$

shows that our result also implies $\mathbf{P}(F_n \ge m_\epsilon) \ge q - o(1)$ for all sufficiently large n.

N.3 Proof of Theorem 7.1

Proof. Let

$$A := \{T_n \geq \alpha n,\ F_n \leq \beta n\}.$$

For each supersample pair, compare the two membership-query outputs after relabeling the selected training sample as $Z_{i,1}$ and the fresh sample as $Z_{i,0}$. Let a be the number of coordinates with outputs $(1,0)$, b the number with outputs $(0,1)$, and m the number of ties, i.e., outputs $(0,0)$ or $(1,1)$. Then

$$a + b + m = n, \qquad T_n - F_n = a - b.$$

Construct an estimator $\hat{J}$ that selects the unique sample declared "in" whenever the two outputs disagree and flips an independent fair coin on a tie. Conditional on the query-output configuration, its Hamming error $H := d_H(\hat{J}, J)$ satisfies

$$H = b + \mathrm{Bin}(m, 1/2).$$

On A, $a - b \geq (\alpha - \beta)n$. Therefore

$$\mathbf{E}[H \mid \text{configuration}] = b + \frac{m}{2} = \frac{n - (a-b)}{2} \leq \frac{1 - \alpha + \beta}{2} n,$$

and also $m \leq (1 - \alpha + \beta)n$. Put

$$t := \frac{1 - \alpha + \beta}{2} + \eta.$$

Hoeffding's inequality for the fair binomial tie-breaking term gives, uniformly over configurations in A,

$$\mathbf{P}(H \leq nt \mid \text{configuration}) \geq 1 - \exp\left(-\frac{2\eta^2}{1 - \alpha + \beta} n\right).$$

By definition of $\Psi_n^\star(t)$ and (68), every estimator satisfies

$$\mathbf{P}(H \leq nt) \leq \Psi_n^\star(t).$$

Consequently,

$$\Psi_n^\star(t) \geq \mathbf{P}(A)\left[1 - \exp\left(-\frac{2\eta^2}{1 - \alpha + \beta} n\right)\right].$$

Finally,

$$\mathbf{P}(A) \geq \mathbf{P}(T_n \geq \alpha n) - \mathbf{P}(F_n > \beta n) \geq q - \mathbf{P}(F_n > \beta n).$$

Rearranging proves the finite-sample claim. If $I_n = o(n)$ and $t < 1/2$, taking $c = L_t$ in (69) gives $\Psi_n^\star(t) \to 0$, which proves the asymptotic statement. □

N.4 Rényi-Divergence Extension

Corollary N.1. *Fix $\tau \in [0, 1/2)$, $b_n := \lfloor \tau n \rfloor$, and define $C_\tau := \log 2 - h_2(\tau) > 0$. For every $\beta > 1$,*

$$p_{b_n} \leq q_{b_n}^{\frac{\beta-1}{\beta}} \exp\big(\big((\beta-1)/\beta\big) D_\beta(P_0 \| Q_0)\big). \tag{73}$$

Since $q_{b_n} \leq e^{-nC_\tau}$, the bound decays exponentially whenever $D_\beta(P_0\|Q_0) \leq (C_\tau - \kappa)n$ for some fixed $\kappa > 0$; in particular, this holds when $D_\beta(P_0\|Q_0) = o(n)$.

N.5 Proof of Corollary N.1

Proof. Using DPI with the event

$$E_{b_n} := \{(\hat{J}, j) : d_H(\hat{J}, j) \leq b_n\}$$

and the f corresponding to power-β divergence, then also using (13), we have

$$p_{b_n}^\beta q_{b_n}^{1-\beta} + (1 - p_{b_n})^\beta (1 - q_{b_n})^{1-\beta} \leq 1 + (\beta - 1)\mathcal{H}_\beta(P_0\|Q_0).$$

Dropping the nonnegative second term yields

$$p_{b_n} \leq q_{b_n}^{\frac{\beta-1}{\beta}} \big(1 + (\beta-1)\mathcal{H}_\beta(P_0\|Q_0)\big)^{1/\beta}.$$

Using (7),

$$1 + (\beta-1)\mathcal{H}_\beta(P_0\|Q_0) = \exp\big((\beta-1)D_\beta(P_0\|Q_0)\big),$$

which proves (73). The remaining claims follow from $q_{b_n} \leq e^{-nC_\tau}$. □

N.6 Approximate-Max-Information Extension

By Lemma 6.2, the same setup also gives a selector-recovery bound in terms of approximate max-information.

Proposition N.2. *If $I_\infty^\tau(J;Y) \le \log\gamma$ for $\gamma \ge 1$ and $\tau \in [0,1)$, then for every $\hat{J} = \phi(Y)$ and $b \in \mathbb{N}_{\ge 0}$,*

$$p_b \le \gamma q_b + \tau. \tag{74}$$

This is an immediate application of Lemma 6.2 and Proposition 2.1; its role is to transfer approximate-max-information control to selector recovery, providing a complementary perspective to CMI-based anti-memorization results.

N.7 Proof of Proposition N.2

Proof. Let

$$P := P_{JY}, \qquad Q := P_J P_Y.$$

By Lemma 6.2,

$$I_\infty^\tau(J;Y) \le \log\gamma \quad \Longleftrightarrow \quad E_\gamma(P\|Q) \le \tau.$$

Apply Proposition 2.1 to the event

$$E_b := \{(j,y) : d_H(\phi(y), j) \le b\}.$$

Since $Q(E_b) = q_b$, we obtain

$$p_b \le \gamma q_b + E_\gamma(P\|Q) \le \gamma q_b + \tau,$$

which proves (74).

□

Consequently, if there exist sequences (γ_n) and (τ_n) such that

$$I_\infty^{\tau_n}(J;Y) \le \log\gamma_n, \qquad \tau_n \to 0, \qquad \log\gamma_n = o(n),$$

then for every fixed $\rho \in [0, 1/2)$, we have $\mathbf{P}\big(\frac{1}{n} d_H(\hat{J}, J) \le \rho\big) \to 0$.