Integral Probability Boundaries

Bogdan Kulynych
Biomedical Data Science Center, Lausanne University Hospital (CHUV)

Shahab Asoodeh
McMaster University and Vector Institute

Abstract

Bounds based on Integral Probability Metrics (IPMs) are ubiquitous in machine learning and statistics, appearing, e.g., in generalization, alignment, fairness, and privacy. IPMs reduce the discrepancy between two distributions to a single worst-case gap between expectations. This compression is costly: for rare events, safety failures, or under-represented groups, IPM-based bounds are loose or vacuous. We introduce a non-linear generalization of IPMs, which we term Integral Probability Boundaries (IPBs). Unlike IPMs, IPBs provide bounds that are tight at the extremes. We derive their basic properties and present algorithms for estimating them in practically relevant settings. We demonstrate the usage of IPBs in both theoretical and practical applications for alignment, judge-free performance estimation, and fairness, where they enable us to obtain sharp bounds on performance and fairness gaps, e.g., 15 p.p. more precise estimates on fairness gaps for rare groups in a realistic generative modeling task, compared to the state of the art.

1 Introduction

Bounding expectations of a function over one distribution P in terms of the expectation over another distribution Q is one of the core building blocks in theoretical and applied statistics and machine learning. We often want to analyze or estimate bounds of the following form:

$$\mathbb{E}_P[f] \leq \mathbb{E}_Q[f] + \delta, \tag{1}$$

where the gap δ could depend on a notion of discrepancy between P and Q such as KL divergence. This form appears across various fields of inquiry, e.g., studies of generalization (P is a test data distribution, Q is a training data distribution), alignment (P is an aligned policy, Q is an original policy), fairness (P and Q are data distributions of two population groups), privacy (P and Q are distributions of a model with and without a certain example included in the training data).

Eq. (1) is also fundamentally related to an object called an *integral probability metric* (IPM) (Müller, 1997) between P and Q, which characterizes the maximum difference between the expectations in Eq. (1) over some family of functions $f \in \mathcal{F}$:

$$D_{\mathcal{F}}(P, Q) \triangleq \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|. \tag{2}$$

It is easy to see that if we can estimate $D_{\mathcal{F}}(P, Q)$, then we can set $\delta = D_{\mathcal{F}}(P, Q)$.

Although convenient, bounds as in Eq. (1) suffer from a critical problem when their sharpness is important. Say that $f(x)$ has a range of $[0, 1]$. Then, at the extremes, when $\mathbb{E}_Q[f] \approx 0$ or $\mathbb{E}_Q[f] \approx 1$, the bound is either overly loose or vacuous. This is because an IPM compresses potentially rich information into a single number. In contrast, consider the entire achievable region of expectations:

$$\mathcal{C}_{\mathcal{F}}(P \parallel Q) \triangleq \{(\mathbb{E}_Q[f], \mathbb{E}_P[f]) \mid f \in \mathcal{F}\}. \tag{3}$$

We introduce a non-linear generalization of IPMs that incorporates a dependency on the exact value of $\mathbb{E}_Q[f]$, and is a boundary of this achievable region:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

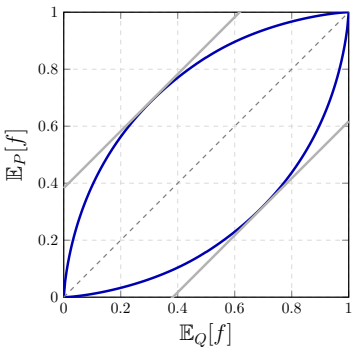

Figure 1: **Integral Probability Boundaries (IPBs) characterize the difference in expectations of any function from a given class $\mathcal{F}$ under two distributions P and Q more precisely than a standard Integral Probability Metric (IPM).** We show the tight IPBs and IPM for $P \sim \mathcal{N}(0,1)$ and $Q \sim \mathcal{N}(1,1)$ over the set of all bounded functions where the IPM is given by the TV distance $\delta = D_{\mathsf{TV}}(P,Q)$, and the IPB is given by the ROC of the Neyman–Pearson test (see Section 2). We show IPM-based bounds with grey lines (–), and IPB-based bounds that trace the contour of the achievable region $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$ with blue lines (–).

Definition 1.1 (Integral Probability Boundary)**.** *Consider a pair of probability distributions (P,Q) over a space $\mathcal{X}$, and a family of functions $\mathcal{F}$ with each $f \in \mathcal{F}$ mapping $\mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$. Let $\overline{\mathcal{Y}} \triangleq \operatorname{conv}(\mathcal{Y})$ be the convex hull of $\mathcal{Y}$. A function $g : \overline{\mathcal{Y}} \to \overline{\mathcal{Y}}$ is an integral probability boundary (IPB) of P w.r.t. Q over $\mathcal{F}$ if the following holds:*

$$g(\alpha) \geq \sup_{(q,p) \in \mathcal{C}_{\mathcal{F}}(P \parallel Q)} \{\, p \mid q \leq \alpha \,\}. \tag{4}$$

We denote by $\mathcal{B}_{\mathcal{F}}(P \parallel Q)$ the set of all such IPB functions. The function g attaining equality for all $\alpha \in \overline{\mathcal{Y}}$ in Eq. (4) is unique. We call it the tight IPB*, denoted as $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$.*

An IPB as defined above describes only the upper boundary of the achievable region $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$. We can recover a lower boundary as a generalized inverse $\check{g}^{-1}(\alpha) \triangleq \inf\{\beta \mid \check{g}(\beta) \geq \alpha\}$ of an IPB $\check{g} \in \mathcal{B}_{\mathcal{F}}(Q \parallel P)$ for the reversed pair. We call g and $\check{g}$ *adjoint* to each other, and refer to a pair $(g, \check{g})$ as *adjoint IPBs*. For notational simplicity, in the rest of the paper we mostly focus on upper bounds.

IPBs sharpen the bound in Eq. (1) by leveraging the non-linear structure:

Proposition 1.2 (Main property)**.** *If $g \in \mathcal{B}_{\mathcal{F}}(P \parallel Q)$ is an IPB, then for any $f \in \mathcal{F}$, we have:*

$$\mathbb{E}_P[f] \leq g(\mathbb{E}_Q[f]). \tag{5}$$

Thus, if we can solve the optimization problem in Eq. (4), we can sharpen any bound of the ubiquitous Eq. (1) form. We provide illustrations of how IPBs improve over IPMs in Fig. 1, and in Section 3. In this work, we make the following contributions:

1. We introduce the concept of IPBs as a non-linear generalization of an IPM, and present its properties and relationship to the standard information-theoretic concepts such as the Neyman–Pearson test statistic and the hockeystick divergence.
2. We provide algorithms for estimating IPBs with finite-sample guarantees, for two realistic settings: over all bounded functions, and over a learnable family of functions.
3. We show how IPB-based analyses enable researchers and practitioners to obtain sharp bounds where such sharpness matters. We provide a substantially tighter bound on performance deterioration in inference-time alignment compared to an approach based on the KL divergence (see, e.g., Beirami et al., 2025), estimate performance without judges when changing a language model's inference engine (Ritter et al., 2026), and estimate proportional representation gaps in synthetic data, with, e.g., 15 p.p. more precise estimates than a previous method from Oesterling et al. (2024) in some cases. We also show additional applications to deriving tight bounds on generalization and privacy risk in Appendix B.

2 Properties of IPBs

Geometry of boundaries. First, we make several observations on the basic geometric properties of IPBs. The tight IPB $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ is always non-decreasing by construction for any function class $\mathcal{F}$. Moreover, if $\mathcal{F}$ is a convex set, then the tight IPB g is concave as an upper boundary of a convex set. As a concave function on an interval is continuous on its interior, the tight IPB is also continuous on the interior of its domain.

For function families that are closed under complement, we have a simple closed form of the lower boundary of the achievable region $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$.

Proposition 2.1. *Consider the set of all bounded functions:*

$$\mathcal{F}_1 \triangleq \{f \mid f : \mathcal{X} \to [0,1]\}.$$

Suppose that $\mathcal{F} \subseteq \mathcal{F}_1$ is closed under complement, i.e., $f \in \mathcal{F}$ implies $1 - f \in \mathcal{F}$, and let $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ and $\breve{g} = \mathcal{B}^*_{\mathcal{F}}(Q \parallel P)$ be a pair of tight adjoint IPBs. Then, for all $\alpha \in [0,1]$, the lower boundary of $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$ has the following form: $\breve{g}^{-1}(\alpha) = 1 - g(1-\alpha)$.

IPBs recover IPMs. Given a pair of tight adjoint IPBs, which is equivalent to having the upper and lower boundaries of $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$, we can recover an IPM:

Proposition 2.2 (IPB to IPM)**.** *Suppose that $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ and $\breve{g} = \mathcal{B}^*_{\mathcal{F}}(Q \parallel P)$ is a pair of tight adjoint IPBs, where $f \in \mathcal{F}$ maps $\mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$. Then, we have:*

$$D_{\mathcal{F}}(P,Q) = \max\left\{\sup[g(\alpha) - \alpha], \sup[\breve{g}(\alpha) - \alpha]\right\}, \tag{6}$$

with both suprema over $\alpha \in \bar{\mathcal{Y}}$. When either IPB is not tight, the equality becomes an upper bound.

IPBs collapse to IPMs for shift-invariant function classes. Although in general IPBs contain more information than IPMs, for some function families the bound they imply degenerates:

Proposition 2.3. *Suppose that $\mathcal{F}$ is shift-invariant and closed under negation, i.e., if $f \in \mathcal{F}$ then $f + c \in \mathcal{F}$ and $-f \in \mathcal{F}$ for any $c \in \mathbb{R}$, with all functions $f \in \mathcal{F}$ mapping $\mathcal{X} \to \mathbb{R}$. Suppose that $g \in \mathcal{B}_{\mathcal{F}}(P \parallel Q)$. Then, $g(\alpha) \geq \alpha + D_{\mathcal{F}}(P,Q)$.*

We defer the proof of this and all other non-trivial statements in the main body to Appendix A. As an immediate implication, for a salient case of $\mathcal{F}$ being the set of Lipschitz functions, IPBs are *not* more useful than IPMs. At the same time, for other important classes of functions, such as sets of bounded functions or balls in reproducing kernel Hilbert spaces (Gretton et al., 2012), IPBs both carry useful information and enable us to derive substantially sharper bounds, as we show next. For the rest of this section, we focus on the properties of IPBs for the family of all bounded functions $\mathcal{F}_1$ due to the fundamental connections of IPBs in this case to standard information-theoretic notions.

IPBs and the Neyman–Pearson lemma. Over the set of bounded functions $\mathcal{F}_1$, the tight IPB is the ROC curve of the likelihood-ratio test (see, e.g., Polyanskiy and Wu, 2025, Chap. 14):

Proposition 2.4. *Let P and Q be probability distributions over $\mathcal{X}$ with densities p and q w.r.t. a common measure, e.g., $P + Q$, and let $L(x) \triangleq {}^{p(x)}/_{q(x)}$ be the likelihood ratio, with $L(x) = \infty$ if $q(x) = 0$. For a threshold $t \in [0,\infty]$ and $\eta \in [0,1]$, consider the randomized likelihood-ratio test $\phi_{t,\eta}(x) \triangleq \mathbb{1}[L(x) > t] + \eta\,\mathbb{1}[L(x) = t]$, with size and power:*

$$\alpha(t,\eta) \triangleq \mathbb{E}_Q[\phi_{t,\eta}], \qquad \beta(t,\eta) \triangleq \mathbb{E}_P[\phi_{t,\eta}], \tag{7}$$

respectively. Then, the tight IPB $g = \mathcal{B}^*_{\mathcal{F}_1}(P \parallel Q)$ is the ROC curve of the likelihood-ratio test: for every $\alpha \in [0,1]$, there exist t and η with $\alpha(t,\eta) = \alpha$ such that the supremum in Eq. (4) is attained by $\phi_{t,\eta}$, thus $g(\alpha) = \beta(t,\eta)$.

IPBs and hockeystick divergence. IPBs are also closely related to the hockeystick divergence:

Definition 2.5. *Let $\gamma \geq 0$.*[1] *The γ-hockeystick divergence $\mathsf{E}_\gamma(P\|Q)$ is defined as:*

$$\mathsf{E}_\gamma(P\|Q) \triangleq \sup_{E \subseteq \mathcal{X}} [P(E) - \gamma Q(E)] = \sup_{f \in \mathcal{F}_1} [\mathbb{E}_P[f] - \gamma \mathbb{E}_Q[f]],$$

where the first supremum is over all measurable events E.

Specifically, we can obtain an IPB from the hockeystick divergence, and vice versa:

Proposition 2.6 (IPB to Hockeystick)**.** *Suppose that $g = \mathcal{B}^*_{\mathcal{F}_1}(P \parallel Q)$. Then, for any $\gamma \geq 0$:*

$$\mathsf{E}_\gamma(P\|Q) = \sup_{\alpha \in [0,1]} [g(\alpha) - \gamma\alpha], \tag{8}$$

which becomes an upper bound when g is not tight. The following is a tight IPB, $g = \mathcal{B}^*_{\mathcal{F}_1}(P \parallel Q)$:

$$g(\alpha) = \inf_{\gamma \geq 0} [\mathsf{E}_\gamma(P\|Q) + \gamma\alpha]. \tag{9}$$

In Appendix B, we also show how one could convert between IPBs and f-divergences.

[1] When the hockeystick divergence is defined as an f-divergence, we need to restrict $\gamma \geq 1$. We do not use its f-divergence form, thus $\gamma \geq 0$ is valid within the scope of this work.

3 Toy Example: Model Performance under Subpopulation Shift

To illustrate the mechanics of IPBs, let us provide a simple example. Suppose we have a classifier that was validated on a population Q but deployed on a population P where the proportions of k groups $G_1, \dots, G_k$ differ, with $p_i \triangleq P(G_i)$ and $q_i \triangleq Q(G_i)$, whereas the performance $a_i \in [0,1]$ of the model within each group stays the same. The deployment performance is then $\mathbb{E}_P[f] = \sum_i a_i p_i$, with $f(a) = \sum_i a_i \mathbb{1}[G_i]$, and we want to bound it given the validation performance $\alpha = \mathbb{E}_Q[f] = \sum_i a_i q_i$. Over the family of such functions:

$$\mathcal{F} = \Big\{ a \mapsto \sum_i a_i \mathbb{1}[G_i] \;\mid\; a_i \in [0,1] \Big\},$$

the IPM simplifies to the total-variation (TV) distance:

$$D_{\mathcal{F}}(P,Q) = \sup_{a_i \in [0,1]} \sum_i a_i (p_i - q_i) = \frac{1}{2} \sum_i |p_i - q_i|.$$

By Definition 1.1, the tight IPB $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ in this case is given by a solution of the linear program:

$$g(\alpha) \;=\; \max \sum_i a_i p_i \quad \text{s.t.} \sum_i a_i q_i \le \alpha,\; a_i \in [0,1],$$

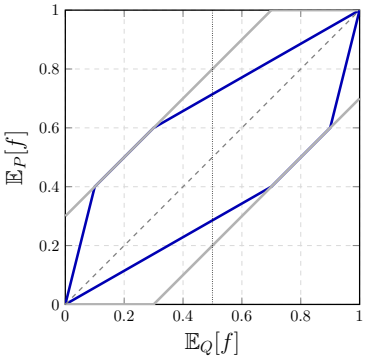

Figure 2: IPB-based bounds (–) and IPM-based bounds (–) for three groups with $(q_1, q_2, q_3) = (0.7, 0.2, 0.1)$ and $(p_1, p_2, p_3) = (0.4, 0.2, 0.4)$.

which is an instance of the classical *fractional knapsack problem*. It has a known optimal solution (see, e.g., Martello and Toth, 1987, Chap. 3) that corresponds to greedily "packing" the groups until the "knapsack" is full, in the order of decreasing $r_i \triangleq p_i/q_i$. Sorting $r_1 \ge \cdots \ge r_k$, this means that we set $a_i = 1$ for the first $m-1$ groups, and choose a_m so that $\mathbb{E}_Q[f] = \alpha$, resulting in:

$$g(\alpha) = \textstyle\sum_{i<m} p_i + r_m \big(\alpha - \sum_{i<m} q_i\big), \qquad \text{for } \alpha \in \big[\textstyle\sum_{i<m} q_i, \sum_{i \le m} q_i\big].$$

We can obtain the lower boundary $\check{g}^{-1}$ using Proposition 2.1. Figure 2 shows an example with three groups. For a validation performance of 0.5 (dotted line), the IPM-based bounds give a deployment performance in $[0.2, 0.8]$, whereas the IPB-based bounds give $[0.29, 0.71]$. For $\alpha \ge 0.7$, the IPM-based upper bound is vacuous, whereas $g(\alpha) < 1$.

4 Algorithms for IPBs

We introduce new algorithms for two salient settings that we make use of in our applications.

Estimating the tight IPB over $\mathcal{F}_1$ with known P and Q. First, we present Algorithm 1 for computing an IPB over the set of all bounded functions when we know both P and Q, and can sample from Q, e.g., when we want to compute an IPB between the distributions of generative models. The algorithm samples from Q alone, and then uses importance weighting to recover expectations under P. Following Proposition 2.4, Algorithm 1 sweeps thresholds t over the observed likelihood ratios and returns a set of $\{(\hat{\alpha}_k, \hat{g}_k)\}_{k=0}^n$ points from the IPB. The empirical means of the likelihood ratios, however, can induce invalid boundaries due to the statistical noise. To correct for this, we also perform a normalization step that exploits the fact that the tight IPB in this setting satisfies $g(1) = 1$. The algorithm has finite-sample convergence guarantees:

Theorem 4.1. *Suppose that $P \ll Q$ and ${}^{dP}/_{dQ} \le B$ almost surely under Q for some $B < \infty$. Let $\hat{g}_n$ denote the piecewise-linear interpolation of $\{(\hat{\alpha}_k, \hat{g}_k)\}_{k=0}^n$ returned by Algorithm 1, and let $g = \mathcal{B}^*_{\mathcal{F}_1}(P \parallel Q)$ be the tight IPB. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of samples from Q, we have:*

$$\sup_{\alpha \in [0,1]} |\hat{g}_n(\alpha) - g(\alpha)| = \mathcal{O}\left(B \sqrt{\tfrac{\log(2/\delta)}{n}} \right).$$

The algorithm is surprisingly simple for its power: (1) it enables us to have bounds of the form in Eq. (5) over the *entire* set of bounded functions without ever performing the actual optimization in the definition of IPBs over this set, and (2) it only requires having samples from Q. We can avoid the optimization because of the Neyman–Pearson connection in Proposition 2.4.

Algorithm 1: IPB estimation for the family of bounded functions with known P and Q

Input: Distributions P and Q with $P \ll Q$; sample size n.
for $i = 1, \dots, n$ **do**
  Sample $o_i \sim Q$, set $L_i \leftarrow {}^{dP}/_{dQ}(o_i)$
end
$\hat{\alpha}_0 \leftarrow 0,\ \hat{g}_0 \leftarrow 0$
Sort $\{L_i\}_{i=1}^n$ in decreasing order as $L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(n)}$
for $k = 1, \dots, n$ **do**
  $\hat{\alpha}_k \leftarrow k/n,\ \hat{g}_k \leftarrow \frac{1}{n} \sum_{j=1}^{k} L_{(j)}$
end
Normalize $\hat{g}_k \leftarrow {}^{\hat{g}_k}/_{\hat{g}_n}$ for all $k = 1, \dots, n$
return $\{(\hat{\alpha}_k, \hat{g}_k)\}_{k=0}^n$

Algorithm 2: IPB estimation for a learnable family $\mathcal{F}$

Input: Samples $X_P \sim P^{n_P}$, $X_Q \sim Q^{n_Q}$; family $\mathcal{F}$; grid $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_k\}$
for $\lambda \in \Lambda \cup \{1\}$ **do**

$$\hat{h}(\lambda) \leftarrow \max_{f \in \mathcal{F}} \left[\frac{1}{n_P} \sum_{x \in X_P} f(x) - \lambda \frac{1}{n_Q} \sum_{x \in X_Q} f(x) \right] \tag{10}$$

end
$\hat{g}(\alpha) \leftarrow \min_{\lambda \in \Lambda \cup \{1\}} \left[\hat{h}(\lambda) + \lambda\, \alpha \right]$
return $\hat{g}$.

Estimating an IPB for learnable $\mathcal{F}$. Next, we present a more general algorithm, applicable even when we do not have access to the probability distributions P and Q themselves, and only have samples from them. In this case, unlike the previous algorithm, we need to solve the optimization problem in Eq. (4). In practice, it is convenient to use the Lagrangian form to avoid the constrained optimization, i.e., solve $\sup_{f \in \mathcal{F}} \left[\mathbb{E}_P[f] - \lambda \mathbb{E}_Q[f] \right]$. When doing so, however, we can no longer explicitly control $\mathbb{E}_Q[f] \leq \alpha$. We make use of the following relationship:

Proposition 4.2 (Lagrangian dual)**.** *Let $\lambda \geq 0$ be a* dual variable*. We define the Lagrangian as $h_{\mathcal{F}}(\lambda) \triangleq \sup_{f \in \mathcal{F}} \left[\mathbb{E}_P[f] - \lambda\, \mathbb{E}_Q[f] \right]$. Then, for any $f \in \mathcal{F}$ and $\alpha \geq \mathbb{E}_Q[f]$, we have:*

$$\mathbb{E}_P[f] \ \leq \ \inf_{\lambda \geq 0} [h_{\mathcal{F}}(\lambda) + \lambda\, \alpha].$$

We present a procedure for estimating an IPB using optimization with a Lagrangian objective in Algorithm 2. Essentially, the algorithm solves k optimization instances for a grid of $\lambda \in \Lambda$ values, and derives the IPB using Proposition 4.2. We also enforce $\lambda = 1$ to be part of the grid to ensure that the resulting estimates are at least as tight as an IPM.

In a practical application, it might not be clear how to choose Λ in a principled way. As we explain in Appendix B, increasing λ values in the grid beyond a certain point does not carry benefits. Thus, one principled way is to use a grid on $[0, \lambda_{\max}]$, where $\lambda_{\max}$ is an estimate of such a λ after which we do not observe benefits, which happens when $h_{\mathcal{F}}(\lambda)$ is small enough. In Algorithm 3 in the Appendix, we provide a heuristic algorithm to find $\lambda_{\max}$ using doubling search. Moreover, in Algorithm 4 in the Appendix, we also present a procedure analogous to Algorithm 2 to get the lower boundary of the achievable region without the need to explicitly invert the adjoint IPB $\breve{g} \in \mathcal{B}_{\mathcal{F}}(Q \parallel P)$.

The algorithm provides a finite-sample convergence guarantee when $\mathcal{F}$ is regular enough. Each $\lambda \in \Lambda$ yields a valid upper bound by Proposition 4.2, thus the output of Algorithm 2 is a valid IPB up to statistical and optimization errors. The following quantifies how far it is from the tight IPB:

Theorem 4.3. *Suppose $0 \in \mathcal{F} \subseteq \mathcal{F}_1$ is a convex family of bounded functions with their Rademacher complexity terms denoted as $\mathfrak{R}_{n_P}(P, \mathcal{F})$ and $\mathfrak{R}_{n_Q}(Q, \mathcal{F})$, that $P \ll Q$ with ${}^{dP}/_{dQ} \leq B$ almost surely under Q, and that we can solve Eq. (10) up to at most $\varepsilon_{\text{opt}} \geq 0$ error. Let $\Lambda \ni 1$ be a grid with $\lambda_{\max} \triangleq \max \Lambda \geq B$ and covering radius $\eta_\Lambda \triangleq \sup_{\lambda \in [0, \lambda_{\max}]} \min_{\lambda' \in \Lambda} |\lambda - \lambda'|$. Then, for*

$\delta \in (0,1)$, w.p. at least $1-\delta$ over sampling of X_P and X_Q, the output $\hat{g}$ of Algorithm 2 satisfies:

$$\sup_{\alpha\in[0,1]} |\hat{g}(\alpha) - g(\alpha)| \;\le\; \varepsilon_P + \lambda_{\max}\,\varepsilon_Q + \eta_\Lambda + \varepsilon_{\text{opt}},$$

where $g = \mathcal{B}^*_{\mathcal{F}}(P \,\|\, Q)$ is the tight IPB, and $\varepsilon_S \triangleq 2\mathfrak{R}_{n_S}(S,\mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{2n_S}}$ for $S \in \{P,Q\}$.

The bound combines four sources of error. The first two terms are statistical, and the latter two are structural due to the discretization of the dual variable, and optimization error, respectively. The discretization error quantifies the trade-off between tightness and the number of optimization oracle calls: e.g., for a uniform grid of k points, $\eta_\Lambda = {\lambda_{\max}}/{2(k-1)}$, so a discretization error of at most ε requires $\mathcal{O}({\lambda_{\max}}/{\varepsilon})$ oracle calls. The last term is the optimization error, whose direction matters for validity: Any over-estimate of the maximum in Eq. (10) still yields a valid, even if loose, upper bound by Proposition 4.2, whereas an under-estimate can invalidate it.

In Appendix A, we extend this analysis. First, we show that the dependence on $\lambda_{\max}$ is conservative—for any α, we can replace it by the active dual variable at α—thus deriving a tighter version of the analysis. Second, we show that for a general non-convex $\mathcal{F}$, the algorithm estimates a valid IPB, but it is separated from the tight one by an additional duality gap. Third, we extend the analysis to unbounded function families. Moreover, in Appendix B, we show how to modify the algorithm to provide a finite-sample certificate that is valid with high probability for simple function families. In Appendix C, we illustrate this certificate for the family of bounded linear functions, for which the optimization oracle is exact and tight Rademacher complexity expressions are available.

5 Applications

We present several applications of IPBs, chosen to showcase them in different settings. First, we show how one could leverage an IPB view to derive a tighter bound on a practically relevant quantity in *alignment*. Second, we show an empirical application to *estimating language model capabilities under a change of inference backend* using the connection of IPBs to the Neyman–Pearson lemma and Algorithm 1. Third, we show another empirical application to *fairness in generative modeling*, this time using a learnable function family and Algorithm 2. Additionally, in Appendix B, we show theoretical applications to derive tight *generalization* bounds and *privacy risk* bounds in differentially private learning. We summarize the considered applications in Table 2 in the Appendix.

5.1 Bounding Capability Deterioration and Limits for Improvement in Alignment

Alignment techniques (see, e.g., Christiano et al., 2017; Stiennon et al., 2020) aim to steer generative models using reward signals to improve their adherence to given requirements or human preferences. In alignment, one needs to balance maximizing reward and preserving existing capabilities. An IPB between the original model and an aligned model can tell us the maximum increase or decrease of *any* metric measuring an existing level of capabilities. In this case study, we focus on best-of-N (BoN) sampling (Brown et al., 2024; Snell et al., 2024), a standard baseline for model alignment at inference time. We derive the closed-form IPB for BoN sampling and compare it against the bounds on capability changes from KL divergence, which is the standard analytical tool in this context (Gao et al., 2023; Beirami et al., 2025; Paes et al., 2026).

Given a reference distribution Q, e.g., a language model, over an output space $\mathcal{Y}$, e.g., over m-token strings, and a scalar reward function $r : \mathcal{Y} \to \mathbb{R}$, the BoN procedure draws N i.i.d. samples $y_1, \ldots, y_N \sim Q$ and returns the one with the highest reward, breaking any ties uniformly at random. Let P_N denote the resulting distribution over outputs. We are interested in quantifying the difference in arbitrary bounded metrics $f \in \mathcal{F}_1$, e.g., error rate, task accuracy, etc., between the original $\mathbb{E}_Q[f]$ and the post-alignment $\mathbb{E}_{P_N}[f]$. To do so, we derive the tight IPB between P_N and Q, which provides the following bound:

Theorem 5.1. *For any bounded performance measure $f \in \mathcal{F}_1$, any reward function $r : \mathcal{Y} \to \mathbb{R}$ and any reference distribution Q:*

$$\mathbb{E}_{y\sim P_N}[f(y)] \le 1 - (1 - \mathbb{E}_{y\sim Q}[f(y)])^N . \tag{11}$$

The inequality is tight if r guarantees no ties.

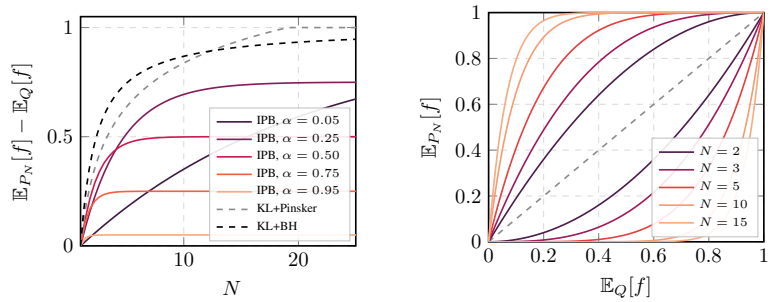

Figure 3: **IPBs provide tighter bounds on performance differences for best-of-N sampling.** *Left:* Upper bound on the gap $\mathbb{E}_{P_N}[f] - \mathbb{E}_Q[f]$, for $\mathbb{E}_Q[f] \leq \alpha$, as a function of N. The IPB-based bounds from Theorem 5.1 are in solid lines. We show KL-based bounds via Pinsker's inequality and the Bretagnolle–Huber (BH) inequality in dashed lines. *Right:* IPBs for several fixed values of N.

To illustrate this theoretical application of IPBs, we show a partial proof of this statement under the assumption that r produces no ties. In Appendix A, we show the remainder.

Proof. Let $F_Q(v) \triangleq \Pr_{y\sim Q}[r(y) \leq v]$ denote the cumulative distribution function of the reward under Q. As there are no ties, BoN outputs y exactly when one of the N draws equals y and the other $N-1$ draws have reward below $r(y)$, so $dP_N(y) = N\, F_Q(r(y))^{N-1}\, dQ(y)$.

Thus, the likelihood ratio of P_N over Q is $L(y) \triangleq {}^{dP_N}\!/_{dQ}(y) = N\, F_Q(r(y))^{N-1}$. Moreover, the no-ties assumption implies that $r(y)$ has a continuous distribution under Q, thus $U \triangleq F_Q(r(y))$ is uniform on $[0,1]$ for $y \sim Q$, and, as a result, $L(y) = NU^{N-1}$. A useful fact that follows immediately is that for a threshold $t \in (0, N)$, we have $L(y) > t$ whenever $U > u_t$, for $u_t \triangleq (t/N)^{\frac{1}{N-1}}$.

By Proposition 2.4, the tight IPB is the ROC curve of the test $\mathbb{1}[L(y) > t]$ (no randomization is needed, as $L(y)$ has no atoms), whose size under Q and power under P_N are:

$$\begin{aligned}
\alpha(t) &= \Pr_{y\sim Q}[L(y) > t] = \Pr[U > u_t] = 1 - u_t, \\
\beta(t) &= \Pr_{y\sim P_N}[L(y) > t] = \mathbb{E}_{y\sim Q}\left[L(y)\cdot \mathbb{1}[L(y) > t]\right] \qquad \text{(change of measure } dP_N = L\, dQ\text{)} \\
&= \mathbb{E}[NU^{N-1}\cdot \mathbb{1}[U > u_t]] = \int_{u_t}^{1} Nu^{N-1}\, du = 1 - u_t^N.
\end{aligned}$$

Setting $u_t = 1-\alpha(t)$ yields $g(\alpha) = 1-(1-\alpha)^N$, and the statement follows by Proposition 1.2. □

Comparison with KL-based bounds. Prior analyses of BoN often use $D_{\mathsf{KL}}(P_N\|Q)$ (see, e.g., Mroueh and Nitsure, 2025; Beirami et al., 2025). A standard approach is to bound it as $D_{\mathsf{KL}}(P_N\|Q) \leq \log N - (N-1)/N$. We can then plug it into, e.g., Pinsker's inequality:

$$\mathbb{E}_{y\sim P_N}[f(y)] \leq \mathbb{E}_{y\sim Q}[f(y)] + \sqrt{\tfrac{1}{2} D_{\mathsf{KL}}(P_N\|Q)}, \tag{12}$$

for any $f \in \mathcal{F}_1$. Clearly, Theorem 5.1 yields a bound that is tighter due to its dependence on $\mathbb{E}_Q[f]$. We show the illustration in Fig. 3. As an example, at $N = 10$, the IPB-based bound is 14 p.p. tighter for the worst-case $\mathbb{E}_Q[f]$, and 83 p.p. tighter at the extremes where $\mathbb{E}_Q[f] \approx 0$ or $\mathbb{E}_Q[f] \approx 1$.

5.2 Judge-Free Estimation of Model Properties in Inference Engines

Language models are often served through inference engines such as vLLM (Kwon, 2025). Such deployments slightly alter their probability distribution compared to an original development backend (Ritter et al., 2026). If the original model has been audited, e.g., for its safety, on the development backend, the deployment version could deviate from it. IPBs can provide an estimate of closeness between the original model and its inference-time counterpart across all evaluation metrics from a chosen class, without requiring those metrics to be re-evaluated.

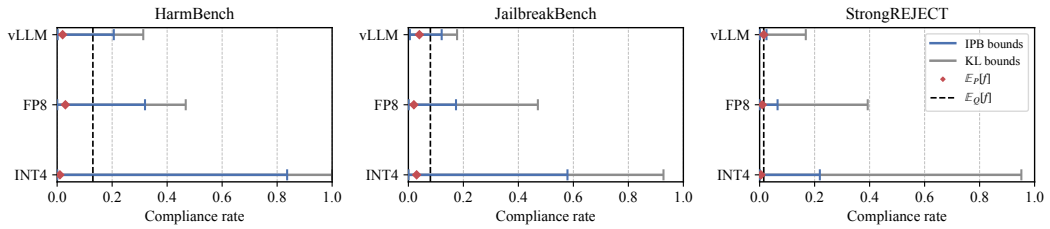

Figure 4: **Unlike KL-based bounds, IPBs enable accurate judge-free estimation of benchmark scores when deploying a model to vLLM.** We show the compliance rate to harmful requests (lower is better) under different benchmarks for three different deployment backends, as well as the estimated bounds on the compliance rate from IPBs between the original model Q and its version under a given backend P, computed using Algorithm 1, and the bounds from the estimated KL divergence.

Let M be a distribution of outputs from a language model under one backend, e.g., huggingface `transformers` (Wolf et al., 2020), and M' be the same distribution under an inference backend, e.g., on vLLM. Let D be a dataset of test inputs, Q denote the joint distribution of (x, y) obtained by drawing $x \sim \mathsf{Unif}[D]$ and $y \sim M(\cdot \mid x)$, and let P denote the analogous distribution under $M'(\cdot \mid x)$. Any evaluation metric $f(x, y) \in [0, 1]$ has expectation $\mathbb{E}_Q[f]$ on the original backend, and $\mathbb{E}_P[f]$ under the inference backend. We assume that as part of an audit we have estimated the score $\mathbb{E}_Q[f]$ for the original model. A pair of adjoint IPBs $g \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$ and $\check{g} \in \mathcal{B}_{\mathcal{F}_1}(Q \parallel P)$ then estimates $\check{g}^{-1}(\mathbb{E}_Q[f]) \leq \mathbb{E}_P[f] \leq g(\mathbb{E}_Q[f])$ for any bounded evaluation metric f without any additional scoring under P. Therefore, IPBs could enable deployers to avoid both drawing additional samples and scoring the outputs with expensive LLM-as-judges.

Experiment setup. We use IPBs to estimate the safety behavior of the `Llama-3.1-8B-Instruct` model (Touvron et al., 2023) under three different backends. We use Algorithm 1, and take the original model M to be the Llama model served via huggingface `transformers` with `bfloat16` precision, and rolled out over $m = 512$ tokens. This corresponds to the distribution Q. We consider three deployment variants M' corresponding to distributions P: (i) a regular vLLM deployment, (ii) FP8 quantization (Kuzmin et al., 2022), and (iii) AWQ-INT4 quantization (Lin et al., 2024). We evaluate all versions on three safety benchmarks: HarmBench (Mazeika et al., 2024) (random subsample, $n = 200$), StrongREJECT (Souly et al., 2024) (random subsample, $n = 200$), and JailbreakBench (Chao et al., 2024) ($n = 100$). In a slight deviation from the recommended evaluation protocols that prescribe specific judges and rubrics in different benchmarks, we use the rate of compliance with harmful requests (lower is better) as judged by Anthropic's `claude-haiku-4-5` model for simplicity and consistency. In a practical deployment, it would be sufficient to measure scores with a judge only for the original model. We compare the scores to the IPB bounds obtained with Algorithm 1, and, as a baseline, to bounds based on KL divergence and either Pinsker's or BH inequality (whichever is tighter), as in the previous section. See more details in Appendix C.

Experiment results. Fig. 4 shows the outcomes. We show the full estimated IPB curves for all benchmarks and backends in Fig. 7 in Appendix D. When quantizing the model, the bounds from IPBs are loose. In some cases, however, e.g., INT4 on HarmBench, the lower bound is almost reached. This suggests that a wide interval from an IPB could indicate that a model should be re-audited. When only moving from the huggingface engine to vLLM, the IPB bounds remain fairly narrow, e.g., within 5 percentage points from the original value in the case of StrongREJECT. The results show that IPBs could enable judge-free performance estimation in certain settings.

5.3 Estimating Multi-Group Proportional Representation Gaps for Rare Groups

Synthetic or distilled datasets and search retrieval results are increasingly used as proxies for large real datasets in various analytic workflows. Whenever we work with such *surrogate data* that is a proxy for real data, one desirable property is *proportional representation* (Oesterling et al., 2024; Jung et al., 2025). Formally, denote by P a reference distribution of records, e.g., a sample of real face images, and by Q a distribution induced by the surrogate dataset, e.g., synthetic face images.

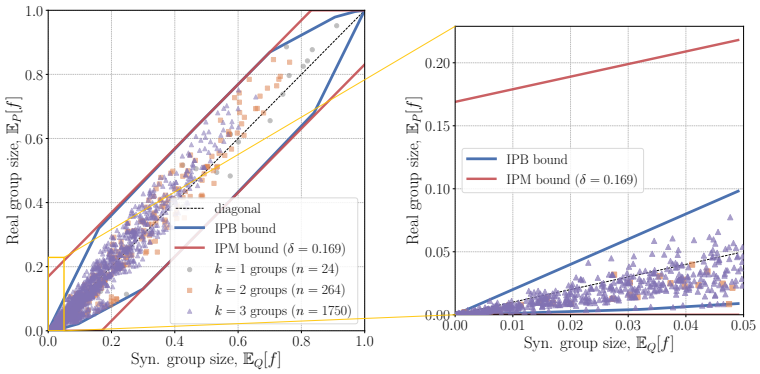

Figure 5: **Estimated IPB bounds for multi-group proportional representation are substantially tighter for rare groups, whereas the bounds based on an IPM are often loose.** We show the bounds on the representation of intersectional groups defined by depth-3 decision trees, obtained from an IPM using the algorithm from Oesterling et al. (2024). We also show an upper bound and lower bound from IPBs computed using Algorithm 2 and Algorithm 4, respectively. Each point shows the prevalence of one such group in the synthetic dataset (x axis) and the real population (y axis). *Left:* Over the entire $[0, 1]$. *Right:* Zoom into the $[0, 0.05]$ region of rare groups.

For every demographic group defined by a group-membership indicator $f(x)$, we want to ensure that its real prevalence, $\mathbb{E}_P[f]$, closely matches its prevalence in the surrogate data, $\mathbb{E}_Q[f]$.

One metric that captures such differences is the IPM over a family of group-membership indicators $\mathcal{F} \subseteq \mathcal{F}_1$. Such an IPM is known as the *multi-group proportional representation* (MPR) gap (Oesterling et al., 2024): $\sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$. The original use case of MPR has been as the objective in a set-valued optimization to construct a representative surrogate dataset. We consider a different goal: Given a surrogate dataset, an analyst wants to know whether the prevalence of a certain group they observe in the surrogate dataset is similar to the prevalence of the same group in the actual data. As we have seen before, bounds based on IPMs become vacuous when the prevalence is either low or high. In this setting, we can use IPBs to provide bounds on the representation of *arbitrary* groups defined by a family of indicators $\mathcal{F}$ without the need to explicitly test the representation of each. Specifically, say we have an adjoint pair of IPBs $g \in \mathcal{B}_{\mathcal{F}}(P \parallel Q)$ and $\breve{g} \in \mathcal{B}_{\mathcal{F}}(Q \parallel P)$. Then, having observed a certain prevalence of a given group in the surrogate dataset Q, we can obtain bounds on its prevalence in the real data as $\breve{g}^{-1}(\mathbb{E}_Q[f]) \leq \mathbb{E}_P[f] \leq g(\mathbb{E}_Q[f])$.

Experiment setup. We demonstrate the applicability of IPBs to multi-group proportional representation for the use case of generating a synthetic dataset of a Census population. Let $\mathcal{X} = \{0, 1\}^d$ be a binary demographic feature space. We define the class $\mathcal{F}_k$ to be all group-membership indicators obtained from decisions of some depth-k decision tree over $\mathcal{X}$. This class contains every k-way intersectional group (e.g., "`white` $\wedge$ `married` $\wedge$ `over 60 years old`") as a special case.

We use California 2018 Census data obtained via the `folktables` package (Ding et al., 2021). We binarize $d = 12$ demographic features covering sex, race/ethnicity, country of origin, disability status, age brackets, marital status, income, and educational status. We take P to be the held-out test split ($n_P = 75{,}763$) and Q a synthetic dataset ($n_Q = n_P$). We generate the synthetic dataset using CTGAN (Xu et al., 2019) on the remaining real dataset ($n \approx 300{,}000$). We use doubling search to find $\lambda_{\max}$, as described in Appendix B. Then, we run Algorithm 2 with a grid Λ of 10 values on $[0, \lambda_{\max}]$, including $\lambda = 1$ (see Appendix C for all hyperparameters). To practically compare the obtained bounds to ground truth for the purpose of the experiment, we set a relatively low $k = 3$, corresponding to all intersectional groups defined by depth-3 decision trees (see an example above), filtering out those that contain fewer than 10 records for stability. This results in 2042 groups. As a baseline, we estimate the MPR gap based on an IPM as $\delta = \sup_{f \in \mathcal{F}_k} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$, as done by Oesterling et al. (2024). In Appendix C, we also show how one could obtain a certified IPB for the family of bounded linear functions, also studied by Oesterling et al. (2024).

Experiment results. Fig. 5 shows the results. The bounds obtained with IPBs using our Algorithm 2 are empirically tighter than those obtained from the corresponding IPM. Specifically, for

rare groups, the lower bound from an IPM on the prevalence in the real dataset is always vacuous, unlike with IPBs. For the smallest group, the gap between an IPB-based bound and an IPM can be as high as $500\times$. The same results hold in replications with two other random train-test splits (see Appendix C). These results show that IPB curves can serve as a measure of fidelity for surrogate data. The analysts working with surrogate data could use IPBs to understand how well group-based queries performed on the surrogate dataset replicate in the real data.

6 Related Work

The objects related to IPBs have appeared previously across different contexts, but have not been unified as a generalization of IPMs as we develop in this work. In particular, the IPB over the family of bounded functions, i.e., the ROC of the Neyman–Pearson test, has been studied and related to the hockeystick divergence and f-divergences multiple times independently, e.g., by Polyanskiy in his Ph.D. thesis, 2010; by Reid and Williamson (2011) in the context of risk in binary experiments; by Sason and Verdú (2016) in the form of *relative information spectrum*; by Agrawal and Horel (2021) in the context of optimal bounds between f-divergences and IPMs; and by Dong et al. (2022) in the context of differential privacy. We generalize beyond the family of all bounded functions, and focus on tools and algorithms for obtaining sharp bounds in applications where this sharpness matters.

A highly related problem is that of Neyman–Pearson (NP) classification (see, e.g., Scott and Nowak, 2005; Rigollet and Tong, 2011; Tong et al., 2018). The goal in NP classification is to learn a binary classifier with a given constraint on false-positive rate, which is equivalent to estimating a point on the tight IPB over a learnable family of binary classifiers. In contrast, our goal is to bound the expectations in broader function families.

7 Concluding Remarks

We introduced integral probability boundaries (IPBs), a functional generalization of integral probability metrics. We showed that IPBs are a useful framing and tool both for deriving tighter bounds in theoretical analyses, and for obtaining precise estimates at the extremes when comparing probability distributions in practical applications. Our hope is that this framework will help both researchers and practitioners to analyze and audit aspects of trustworthiness such as safety and fairness in machine learning, where performance at the extremes is important, in a more precise and inclusive way.

Provable guarantees. IPBs provide only approximate bounds in some settings, limiting their applicability when we need provable guarantees, as the quality of the estimate in Algorithm 2 depends on the optimality of the optimization oracle. Exploring settings where we can obtain provable guarantees, e.g., using convex relaxations, is a potentially useful direction for future work.

Alternative computational approaches. Exploring alternatives to Algorithm 2 for appropriate function families is another promising direction for future work. For instance, we could estimate an IPB by solving the problem in Eq. (4) directly via the NP classification and then deriving the NP-ROC curves (Tong et al., 2018), or via projected gradient descent for differentiable families.

When IPMs suffice. Finally, although IPBs are a drop-in replacement for IPMs, not every application requires the added complexity of IPBs. First, as shown in Section 2, IPBs do not provide improvements over IPMs for shift-invariant function families. For one salient case, this includes the family of Lipschitz functions; therefore, IPBs do not improve bounds based on Wasserstein distance.

Second, in some applications, it is more convenient to summarize the difference between distributions using a single number. Consider the goal of minimizing the difference between two distributions. As an example, Oesterling et al. (2024) study the problem of constructing a representative surrogate dataset that minimizes the representation gaps compared to the reference distribution P across all groups within a given family $\mathcal{F}$, i.e., modeling the surrogate dataset as a probability distribution Q, the goal is to solve $\min_Q D_{\mathcal{F}}(P, Q)$. In such a case, it is convenient to work with a scalar-valued discrepancy $D_{\mathcal{F}}(P, Q)$ as opposed to a set of points on the boundary as part of the optimization problem. Instead, IPBs are likely most useful when fine-grained differences at the extremes matter, such as in audits and analyses, as in the applications we study in Sections 5.1 to 5.3.

Acknowledgements

The computing resources were provided by Modal. SA was partially supported by the NSERC Discovery Grant. BK was partially supported by the Swiss National Science Foundation under award numbers 10003518 and 237378, and by the SYNTHIA project. SYNTHIA (Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications) project is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No. 101172872. Thus, the project is partially funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

References

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-N alignment policy. In *International Conference on Machine Learning*, 2025.

Daniel Ritter, Owen Oertell, Bradley Guo, Jonathan Chang, Kiante Brantley, and Wen Sun. LLMs can learn to reason via off-policy RL. *arXiv preprint arXiv:2602.19362*, 2026.

Alex Oesterling, Claudio M Verdun, Carol X Long, Alexander Glynn, Lucas M Paes, Sajani Vithana, Martina Cardone, and Flavio P Calmon. Multi-group proportional representation in retrieval. In *Advances in Neural Information Processing Systems*, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning research*, 13(1):723–773, 2012.

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

Silvano Martello and Paolo Toth. Algorithms for knapsack problems. *North-Holland Mathematics Studies*, 1987.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information processing systems*, 2017.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Lucas Monteiro Paes, Natalie Mackraz, Barry-John Theobald, and Federico Danieli. Theoretical limits of language model alignment. In *Advances in Neural Information Processing Systems*, 2026.

Youssef Mroueh and Apoorva Nitsure. Information theoretic guarantees for policy alignment in large language models. *Transactions of Machine Learning Research*, 2025.

Woosuk Kwon. *vLLM: An Efficient Inference Engine for Large Language Models*. PhD thesis, UC Berkeley, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the conference on Empirical Methods in Natural Language Processing: System demonstrations*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. FP8 quantization: The power of the exponent. In *Advances in Neural Information Processing Systems*, 2022.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A StrongReject for empty jailbreaks. In *Advances in Neural Information Processing Systems*, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, 2024.

Sangwon Jung, Alex Oesterling, Claudio Mayrink Verdun, Sajani Vithana, Taesup Moon, and Flavio P Calmon. Multi-group proportional representations for text-to-image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring ADULT: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, 2019.

Yury Polyanskiy. *Channel coding: Non-asymptotic fundamental limits*. Princeton University, 2010.

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

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

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

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.

Clayton Scott and Robert Nowak. A Neyman–Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11), 2005.

Philippe Rigollet and Xin Tong. Neyman–Pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12, 2011.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman–Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2), 2018.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 2014.

Bogdan Kulynych, Juan Felipe Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio Calmon, and Jean Louis Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. In *Advances in Neural Information Processing Systems*, 2025.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 2020.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

The pandas development team. pandas-dev/pandas: Pandas, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE international conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE, 2016.

A Omitted Proofs

A.1 Proofs for Section 2

We start with the proofs of the basic properties.

Proposition A.1. *Suppose that $\mathcal{F}$ is convex, i.e., $\lambda f_1 + (1-\lambda) f_2 \in \mathcal{F}$ whenever $f_1, f_2 \in \mathcal{F}$, for any $\lambda \in [0,1]$. Then, $g = \mathcal{B}^*_{\mathcal{F}}(P \,\|\, Q)$ is concave on $\overline{\mathcal{Y}}$, with the convention $\sup \emptyset = -\infty$.*

Proof. Denote $\mathcal{F}_\alpha \triangleq \{f \in \mathcal{F} \mid \mathbb{E}_Q[f] \le \alpha\}$, so that $g(\alpha) = \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}_P[f]$. Fix $\alpha_1, \alpha_2 \in \overline{\mathcal{Y}}$, $\lambda \in [0,1]$, and let $\alpha_\lambda \triangleq \lambda\alpha_1 + (1-\lambda)\alpha_2$. Observing that for any $f_1 \in \mathcal{F}_{\alpha_1}$ and $f_2 \in \mathcal{F}_{\alpha_2}$, the function $f_\lambda \triangleq \lambda f_1 + (1-\lambda) f_2$ lies in $\mathcal{F}$ by convexity and satisfies $\mathbb{E}_Q[f_\lambda] \le \alpha_\lambda$ by linearity, we have:

$$g(\alpha_\lambda) \ge \mathbb{E}_P[f_\lambda] = \lambda \mathbb{E}_P[f_1] + (1-\lambda)\mathbb{E}_P[f_2].$$

Taking the supremum over $f_1 \in \mathcal{F}_{\alpha_1}$ and $f_2 \in \mathcal{F}_{\alpha_2}$ on the r.h.s. yields $g(\alpha_\lambda) \ge \lambda g(\alpha_1) + (1-\lambda) g(\alpha_2)$. □

Proposition 2.3. *Suppose that $\mathcal{F}$ is shift-invariant and closed under negation, i.e., if $f \in \mathcal{F}$ then $f + c \in \mathcal{F}$ and $-f \in \mathcal{F}$ for any $c \in \mathbb{R}$, with all functions $f \in \mathcal{F}$ mapping $\mathcal{X} \to \mathbb{R}$. Suppose that $g \in \mathcal{B}_{\mathcal{F}}(P \,\|\, Q)$. Then, $g(\alpha) \ge \alpha + D_{\mathcal{F}}(P, Q)$.*

Proof. Fix $\alpha \in \mathbb{R}$. As before, we set $\mathcal{F}_\alpha = \{f \in \mathcal{F} \mid \mathbb{E}_Q[f] \le \alpha\}$. Observe that for any $f \in \mathcal{F}$, there also exists $h \in \mathcal{F}$ that is its shifted version: $h = \alpha + f - \mathbb{E}_Q[f]$. Regardless of f, the shifted version h satisfies $\mathbb{E}_Q[h] = \alpha$, thus $h \in \mathcal{F}_\alpha$. Therefore, denoting by $\mathcal{F}'_\alpha = \{\alpha + f - \mathbb{E}_Q[f] \mid f \in \mathcal{F}\}$, we have that $\mathcal{F}'_\alpha \subseteq \mathcal{F}_\alpha$.

As a result, $\sup_{h \in \mathcal{F}'_\alpha} \mathbb{E}_P[h] \le \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}_P[f] \le g(\alpha)$, where the last inequality is the definition of an IPB. Further, we can compute the supremum over $\mathcal{F}'_\alpha$:

$$\sup_{h \in \mathcal{F}'_\alpha} \mathbb{E}_P[h] = \sup_{f \in \mathcal{F}} \big[\alpha + \mathbb{E}_P[f] - \mathbb{E}_Q[f]\big] = \alpha + D_{\mathcal{F}}(P, Q),$$

where the last equality uses closure under negation: $\sup_{f \in \mathcal{F}} (\mathbb{E}_P[f] - \mathbb{E}_Q[f]) = \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| = D_{\mathcal{F}}(P, Q)$, as for any $f \in \mathcal{F}$ the function $-f \in \mathcal{F}$ achieves $\mathbb{E}_P[-f] - \mathbb{E}_Q[-f] = \mathbb{E}_Q[f] - \mathbb{E}_P[f]$. Thus, we have that $\alpha + D_{\mathcal{F}}(P, Q) \le g(\alpha)$ for any $\alpha \in \mathbb{R}$. □

Remark A.2. In Proposition 2.3, we can relax the conditions on the family by removing the assumption that it is closed under negation. In that case, even though the IPB does not reduce to the standard IPM, it reduces to an asymmetric version of an IPM:

$$g(\alpha) \ge \alpha + \sup_{f \in \mathcal{F}} [\mathbb{E}_P[f] - \mathbb{E}_Q[f]].$$

(Note the absence of the absolute value.)

Proposition 2.6 (IPB to Hockeystick)**.** *Suppose that $g = \mathcal{B}^*_{\mathcal{F}_1}(P \,\|\, Q)$. Then, for any $\gamma \ge 0$:*

$$\mathsf{E}_\gamma(P \| Q) = \sup_{\alpha \in [0,1]} \left[g(\alpha) - \gamma\alpha\right], \tag{8}$$

which becomes an upper bound when g is not tight. The following is a tight IPB, $g = \mathcal{B}^*_{\mathcal{F}_1}(P \,\|\, Q)$:

$$g(\alpha) = \inf_{\gamma \ge 0} \left[\mathsf{E}_\gamma(P \| Q) + \gamma\alpha\right]. \tag{9}$$

Proof. For any test $f \in \mathcal{F}_1$, by the IPB property we have:

$$\mathbb{E}_P[f] - \gamma \mathbb{E}_Q[f] \le g(\mathbb{E}_Q[f]) - \gamma \mathbb{E}_Q[f].$$

As the tight IPB g is concave and non-decreasing, we can maximize over $\alpha = \mathbb{E}_Q[f] \in [0,1]$, which yields:

$$g(\mathbb{E}_Q[f]) - \gamma \mathbb{E}_Q[f] \le \sup_{\alpha \in [0,1]} \left[g(\alpha) - \gamma\alpha\right].$$

Therefore, Eq. (8) follows from the definition of the hockeystick divergence.

To show tightness when g is a tight IPB, we use a sandwich bound. First, notice that for every $\alpha \in [0,1]$:

$$g(\alpha) - \gamma\alpha \;=\; \sup_{f \in \mathcal{F}_\alpha} [\mathbb{E}_P[f] - \gamma\alpha] \;\le\; \sup_{f \in \mathcal{F}_\alpha} [\mathbb{E}_P[f] - \gamma\mathbb{E}_Q[f]] \;\le\; \mathsf{E}_\gamma(P\|Q),$$

where the first inequality uses $-\gamma\alpha \le -\gamma\mathbb{E}_Q[f]$, as $\gamma \ge 0$ and $\mathbb{E}_Q[f] \le \alpha$, and the second drops the constraint. Then, taking the supremum on the left gives the matching lower bound, so the inequality holds with equality.

Next, we verify that $\tilde{g}(\alpha) \triangleq \inf_{\gamma \ge 0}[\mathsf{E}_\gamma(P\|Q) + \gamma\alpha]$ defines an IPB. For any $f \in \mathcal{F}_\alpha$ and any $\gamma \ge 0$, the definition of the hockeystick divergence gives $\mathbb{E}_P[f] - \gamma\mathbb{E}_Q[f] \le \mathsf{E}_\gamma(P\|Q)$, so after rearranging we have:

$$\mathbb{E}_P[f] \;\le\; \mathsf{E}_\gamma(P\|Q) + \gamma\mathbb{E}_Q[f] \;\le\; \mathsf{E}_\gamma(P\|Q) + \gamma\alpha.$$

As this holds for every $\gamma \ge 0$, taking the infimum on the right yields $\mathbb{E}_P[f] \le \tilde{g}(\alpha)$, so $\tilde{g} \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$. As g is the tight IPB, this implies $\tilde{g} \ge g$.

Finally, we show the reverse inequality. For any $\alpha \in [0,1]$, Proposition 2.4 provides t and η such that the test $\phi_{t,\eta}$ satisfies $\alpha(t,\eta) = \alpha$ and $\beta(t,\eta) = g(\alpha)$.

Because $\mathbb{E}_P[f] - t\mathbb{E}_Q[f] = \mathbb{E}_Q[(L-t)f]$, the supremum of this expression over $f \in \mathcal{F}_1$ is achieved by any test that is 1 on $\{L > t\}$ and 0 on $\{L < t\}$. As $\phi_{t,\eta}$ satisfies this condition, it is a maximizer in the definition of $\mathsf{E}_t(P\|Q)$, and therefore:

$$\mathsf{E}_t(P\|Q) \;=\; \beta(t,\eta) - t\,\alpha(t,\eta) \;=\; g(\alpha) - t\alpha.$$

By the definition of $\tilde{g}(\alpha)$ as an infimum, we have $\tilde{g}(\alpha) \le \mathsf{E}_t(P\|Q) + t\alpha = g(\alpha)$, confirming that $\tilde{g} = g$.

□

The above result can also be seen as a convex conjugate of g:

Remark A.3. Suppose that $g \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$. Then, for any $\gamma \ge 0$,

$$\mathsf{E}_\gamma(P\|Q) \le 1 + (1-g)^*(-\gamma),$$

where $(\cdot)^*$ denotes the convex conjugate operation, with equality if $g = \mathcal{B}^*_{\mathcal{F}_1}(P \parallel Q)$ is the tight IPB.

Proof. Let $h(\alpha) = 1 - g(\alpha)$. By the definition of the convex conjugate, with the supremum taken over the domain of g:

$$h^*(y) = \sup_{\alpha \in [0,1]} (y\alpha - h(\alpha))$$

Evaluating at $y = -\gamma$:

$$(1-g)^*(-\gamma) = \sup_{\alpha \in [0,1]} (-\gamma\alpha - (1 - g(\alpha)))$$

$$= \sup_{\alpha \in [0,1]} (g(\alpha) - \gamma\alpha - 1)$$

$$= \left(\sup_{\alpha \in [0,1]} (g(\alpha) - \gamma\alpha) \right) - 1$$

Substituting this result back, we get:

$$1 + (1-g)^*(-\gamma) = 1 + \left(\sup_{\alpha \in [0,1]} (g(\alpha) - \gamma\alpha) \right) - 1 = \sup_{\alpha \in [0,1]} (g(\alpha) - \gamma\alpha).$$

The claim follows from Eq. (8), which holds with equality for the tight IPB. □

A.2 Analysis of Algorithm 1

Before proceeding, we start with some technical lemmas.

Lemma A.4 (Normalization)**.** *Let $g\colon [0,1] \to [0,1]$ satisfy $g(1) = 1$, and let $\tilde{g}\colon [0,1] \to [0,\infty)$ be non-decreasing with $\sup_{\alpha\in[0,1]} |\tilde{g}(\alpha) - g(\alpha)| \le r$. Define $\hat{g} \triangleq \tilde{g}/\tilde{g}(1)$ if $\tilde{g}(1) > 0$, and $\hat{g} \triangleq \tilde{g}$ otherwise. Then:*

$$\sup_{\alpha\in[0,1]} |\hat{g}(\alpha) - g(\alpha)| \le 2r.$$

Proof. Fix $\alpha \in [0,1]$. If $\tilde{g}(1) = 0$, then $\tilde{g} \equiv 0$ as $\tilde{g}$ is non-negative and non-decreasing, so $\hat{g} = \tilde{g}$ and the claim follows from the assumption on $\tilde{g}$. Otherwise, factoring $\tilde{g}(\alpha)/\tilde{g}(1)$ out of both terms, we have:

$$\hat{g}(\alpha) - \tilde{g}(\alpha) = \frac{\tilde{g}(\alpha)}{\tilde{g}(1)} - \tilde{g}(\alpha) = \frac{\tilde{g}(\alpha)}{\tilde{g}(1)}\big(1 - \tilde{g}(1)\big) = \hat{g}(\alpha)\big(1 - \tilde{g}(1)\big),$$

where $\hat{g}(\alpha) \in [0,1]$ as $0 \le \tilde{g}(\alpha) \le \tilde{g}(1)$ by monotonicity. Using this fact, and by the triangle inequality, we have:

$$\begin{aligned}
|\hat{g}(\alpha) - g(\alpha)| &\le |\hat{g}(\alpha) - \tilde{g}(\alpha)| + |\tilde{g}(\alpha) - g(\alpha)| \\
&\le |1 - \tilde{g}(1)| + r \\
&= |g(1) - \tilde{g}(1)| + r \le 2r,
\end{aligned}$$

where the last line uses $g(1) = 1$ and the assumption on $\tilde{g}$ at $\alpha = 1$. □

We also make use of the following representation of the tight IPB in terms of the tail function of the Neyman–Pearson test statistic, i.e., the likelihood ratio L under Q.

Lemma A.5. *Let $g = \mathcal{B}^*_{\mathcal{F}_1}(P\|Q)$ be the tight IPB with $L = {}^{dP}/_{dQ} \le B$ almost surely under Q, and let $\alpha(t) \triangleq \Pr_{y\sim Q}[L(y) > t]$ be the tail function of L under Q. Then, for every $\alpha \in [0,1]$:*

$$g(\alpha) = \int_0^B \min\{\alpha, \alpha(t)\}dt.$$

Proof. Fix $\alpha \in [0,1]$. By Proposition 2.4, there exist $t_\alpha \ge 0$ and $\eta \in [0,1]$ such that the test $\phi^* \triangleq \phi_{t_\alpha,\eta}$ satisfies $\mathbb{E}_Q[\phi^*] = \alpha$ and $g(\alpha) = \mathbb{E}_P[\phi^*] = \mathbb{E}_Q[L\phi^*]$. As $0 \le L \le B$ almost surely under Q, we can write $L = \int_0^B \mathbb{1}[L > t]dt$ and exchange the expectation and the integral:

$$g(\alpha) = \int_0^B \mathbb{E}_Q[\phi^*\mathbb{1}[L > t]]dt.$$

For $t \ge t_\alpha$, we have $\phi^*\mathbb{1}[L > t] = \mathbb{1}[L > t]$, with expectation $\alpha(t) \le \alpha$. For $t < t_\alpha$, we have $\phi^*\mathbb{1}[L > t] = \phi^*$, with expectation $\alpha \le \alpha(t)$. Thus, the integrand equals $\min\{\alpha, \alpha(t)\}$ for every t, which completes the proof. □

Next, we can show the main result.

Theorem 4.1. *Suppose that $P \ll Q$ and ${}^{dP}/_{dQ} \le B$ almost surely under Q for some $B < \infty$. Let $\hat{g}_n$ denote the piecewise-linear interpolation of $\{(\hat{\alpha}_k, \hat{g}_k)\}_{k=0}^n$ returned by Algorithm 1, and let $g = \mathcal{B}^*_{\mathcal{F}_1}(P \,\|\, Q)$ be the tight IPB. Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over the draw of samples from Q, we have:*

$$\sup_{\alpha\in[0,1]} |\hat{g}_n(\alpha) - g(\alpha)| = \mathcal{O}\left(B\sqrt{\tfrac{\log(2/\delta)}{n}}\right).$$

Proof. Omitting randomized thresholding for discrete distributions for simplicity, recall from Proposition 2.4 that the tight IPB curve $g(\alpha)$ is parametrized by thresholds $t \ge 0$:

$$\begin{aligned}
\alpha(t) &\triangleq \Pr_{y\sim Q}[L(y) > t] \\
\beta(t) &\triangleq \Pr_{y\sim P}[L(y) > t] = \mathbb{E}_{y\sim Q}[L(y)\mathbb{1}[L(y) > t]]
\end{aligned}$$

with the tight IPB $g = \mathcal{B}^*_{\mathcal{F}_1}(P\|Q)$ being the boundary traced by $(\alpha(t), \beta(t))$.

Algorithm 1 computes points based on the sorted order of the likelihood-ratio values $L_i = L(o_i)$ where $o_i \sim Q$, such that $L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(n)}$.

For any $k \in \{1, \ldots, n\}$, the algorithm computes

$$\hat{\alpha}_k \triangleq \frac{k}{n}, \text{ and } \hat{\beta}_k \triangleq \frac{1}{n}\sum_{j=1}^{k} L_{(j)},$$

and returns the normalized values $\hat{g}_k = {}^{\hat{\beta}_k}/_{\hat{\beta}_n}$. Let $\tilde{g}_n$ denote the piecewise-linear interpolation of the *unnormalized* points:

$$(0,0),\ (\hat{\alpha}_1, \hat{\beta}_1), \ldots, (\hat{\alpha}_n, \hat{\beta}_n).$$

This interpolation corresponds to randomized thresholding on the empirical likelihood ratios. The interpolation $\hat{g}_n$ of the returned points satisfies $\hat{g}_n = {}^{\tilde{g}_n}/_{\hat{\beta}_n}$ with $\hat{\beta}_n = \tilde{g}_n(1)$. Now define the empirical tail function:

$$A_n(t) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[L_i > t].$$

By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, we can write with probability at least $1-\delta$ over sampling of Q:

$$\sup_{t \geq 0} |A_n(t) - \alpha(t)| \leq \varepsilon_n, \tag{13}$$

where $\varepsilon_n \triangleq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$. We now express both the population and empirical boundaries using their tail functions. According to Lemma A.5, we can write

$$g(\alpha) = \int_0^B \min\{\alpha, \alpha(t)\} dt.$$

Similarly, it can be verified that the interpolated empirical boundary satisfies

$$\tilde{g}_n(\alpha) = \int_0^B \min\{\alpha, A_n(t)\}\, dt.$$

Indeed, at $\alpha = \frac{k}{n}$, we have

$$\tilde{g}_n\left(\frac{k}{n}\right) = \int_0^B \min\left\{\frac{k}{n}, A_n(t)\right\} dt = \frac{1}{n}\sum_{j=1}^{k} L_{(j)} = \hat{\beta}_k,$$

and the equality extends to all $\alpha \in [0,1]$ by linear interpolation. Therefore, for every $\alpha \in [0,1]$, we have:

$$\begin{aligned}
|\tilde{g}_n(\alpha) - g(\alpha)| &\leq \int_0^B |\min\{\alpha, A_n(t)\} - \min\{\alpha, \alpha(t)\}|\, dt \\
&\leq \int_0^B |A_n(t) - \alpha(t)|\, dt \\
&\leq B\varepsilon_n,
\end{aligned}$$

where the last step follows from (13). Taking the supremum over α, we obtain

$$\sup_{\alpha \in [0,1]} |\tilde{g}_n(\alpha) - g(\alpha)| \leq B\varepsilon_n.$$

Finally, we account for the normalization. The tight IPB satisfies $g(1) = 1$, and $\tilde{g}_n$ is non-negative and non-decreasing, so Lemma A.4 applied with $r = B\varepsilon_n$ yields:

$$\sup_{\alpha \in [0,1]} |\hat{g}_n(\alpha) - g(\alpha)| \leq 2B\varepsilon_n.$$

Thus, we conclude that, with probability at least $1-\delta$,

$$\sup_{\alpha \in [0,1]} |\hat{g}_n(\alpha) - g(\alpha)| \leq \sqrt{2}B\sqrt{\frac{\log(2/\delta)}{n}}.$$

□

Remark A.6. The above proof shows that under the assumption of bounded likelihood, the estimator $\hat{g}_n$ obtained from Algorithm 1 converges uniformly to the tight IPB. In the unbounded case, however, such a distribution-free finite-sample guarantee does not hold in general. Instead, applying the algorithm to the truncated likelihood ratios $\min\{L, B\}$ yields an estimator $\hat{g}_{n,B}$ with the following bias-variance decomposition:

$$\sup_{\alpha\in[0,1]} |\hat{g}_{n,B}(\alpha) - g(\alpha)| \le C\left(B\sqrt{\frac{\log(2/\delta)}{n}} + \mathbb{E}_Q[(L-B)_+]\right). \tag{14}$$

The first term is the statistical error for estimating the truncated boundary, whereas the second term is the truncation bias. As $P \ll Q$ implies $\mathbb{E}_Q[L] = 1$, the truncation bias vanishes as $B \to \infty$. Thus, by choosing $B = B_n \to \infty$ slowly enough such that $B_n\sqrt{\log(2/\delta)/n} \to 0$, the truncated estimator $\hat{g}_{n,B_n}$ converges uniformly to g with high probability.

Proof. If L is unbounded, we can apply the above proof to the truncated likelihood:

$$L_B \triangleq \min\{L, B\}.$$

We also modify Algorithm 1 to truncate sample likelihood:

$$L_{i,B} \triangleq \min\{L_i, B\}.$$

Define the corresponding population boundary:

$$g_B(\alpha) \triangleq \sup_{f\in\mathcal{F}_1:\ \mathbb{E}_Q[f]\le\alpha} \mathbb{E}_Q[L_B f].$$

Let $\tilde{g}_{n,B}$ denote the piecewise-linear interpolation of the unnormalized points obtained by applying Algorithm 1 to the truncated values $L_{1,B}, \dots, L_{n,B}$, and let $\hat{g}_{n,B} = \tilde{g}_{n,B}/\tilde{g}_{n,B}(1)$ denote the returned normalized estimator. Applying the bounded argument (before normalization) to L_B, with probability at least $1-\delta$,

$$\sup_{\alpha\in[0,1]} |\tilde{g}_{n,B}(\alpha) - g_B(\alpha)| \le CB\sqrt{\frac{\log(2/\delta)}{n}}. \tag{15}$$

Moreover, for every feasible $f \in \mathcal{F}_1$ satisfying $\mathbb{E}_Q[f] \le \alpha$:

$$\begin{aligned}\mathbb{E}_Q[Lf] &= \mathbb{E}_Q[L_B f] + \mathbb{E}_Q[(L-B)_+ f] \\ &\le \mathbb{E}_Q[L_B f] + \mathbb{E}_Q[(L-B)_+] \\ &\le g_B(\alpha) + \mathbb{E}_Q[(L-B)_+].\end{aligned}$$

Taking the supremum over all feasible $f \in \mathcal{F}_1$ gives

$$g(\alpha) = \sup_{f\in\mathcal{F}_1:\ \mathbb{E}_Q[f]\le\alpha} \mathbb{E}_Q[Lf] \le g_B(\alpha) + \mathbb{E}_Q[(L-B)_+].$$

Also, as $L_B \le L$, we have:

$$g_B(\alpha) \le g(\alpha),$$

which implies:

$$0 \le g(\alpha) - g_B(\alpha) \le \mathbb{E}_Q[(L-B)_+].$$

We can write for any $\alpha \in [0,1]$:

$$\begin{aligned}|\tilde{g}_{n,B}(\alpha) - g(\alpha)| &\le |\tilde{g}_{n,B}(\alpha) - g_B(\alpha)| + |g_B(\alpha) - g(\alpha)| \\ &\le |\tilde{g}_{n,B}(\alpha) - g_B(\alpha)| + \mathbb{E}_Q[(L-B)_+].\end{aligned}$$

Thus, we can write:

$$\sup_{\alpha\in[0,1]} |\tilde{g}_{n,B}(\alpha) - g(\alpha)| \le \sup_{\alpha\in[0,1]} |\tilde{g}_{n,B}(\alpha) - g_B(\alpha)| + \mathbb{E}_Q[(L-B)_+].$$

Combining this with Eq. (15) yields with probability at least $1-\delta$:

$$\sup_{\alpha\in[0,1]} |\tilde{g}_{n,B}(\alpha) - g(\alpha)| \le CB\sqrt{\frac{\log(2/\delta)}{n}} + \mathbb{E}_Q[(L-B)_+].$$

Finally, since $g(1) = 1$ and $\tilde{g}_{n,B}$ is non-negative and non-decreasing, Lemma A.4 applied with $r = CB\sqrt{\log(2/\delta)/n} + \mathbb{E}_Q[(L-B)_+]$ yields the sought form. □

A.3 Analysis of Algorithm 2

Let us fix the notation for this section. Throughout, $\mathcal{F} \subseteq \mathcal{F}_1$ and $\Lambda \ni 1$ denotes the grid used by Algorithm 2, so that $\lambda_{\max} \triangleq \max \Lambda \geq 1$. We write:

$$\hat{\mathbb{E}}_P[f] \triangleq \frac{1}{n_P} \sum_{x \in X_P} f(x), \qquad \hat{\mathbb{E}}_Q[f] \triangleq \frac{1}{n_Q} \sum_{x \in X_Q} f(x),$$

to denote the empirical averages. To distinguish the exact empirical objective from the value returned by the optimization oracle, we introduce $\hat{h}_{\mathrm{opt}}$:

$$\hat{h}_{\mathrm{opt}}(\lambda) \triangleq \sup_{f \in \mathcal{F}} \left[\hat{\mathbb{E}}_P[f] - \lambda \hat{\mathbb{E}}_Q[f]\right], \text{ with } \left|\hat{h}(\lambda) - \hat{h}_{\mathrm{opt}}(\lambda)\right| \leq \varepsilon_{\mathrm{opt}} \text{ for all } \lambda \in \Lambda,$$

where $\hat{h}(\lambda)$ is the value used in Algorithm 2. The output of the algorithm is $\hat{g}(\alpha) = \min_{\lambda \in \Lambda}[\hat{h}(\lambda) + \lambda\alpha]$, and we denote by $\hat{g}_{\mathrm{opt}}$ the output under an exact oracle:

$$\hat{g}_{\mathrm{opt}}(\alpha) \triangleq \min_{\lambda \in \Lambda}[\hat{h}_{\mathrm{opt}}(\lambda) + \lambda\alpha].$$

We consider the following population *dual IPBs*, each of which is a valid IPB in $\mathcal{B}_{\mathcal{F}}(P \parallel Q)$ by Proposition 4.2:

$$\begin{aligned}
g_\Lambda(\alpha) &\triangleq \min_{\lambda \in \Lambda} [h_{\mathcal{F}}(\lambda) + \lambda\alpha] && \text{(grid-restricted dual IPB)},\\
g_{[0,\lambda_{\max}]}(\alpha) &\triangleq \inf_{0 \leq \lambda \leq \lambda_{\max}} [h_{\mathcal{F}}(\lambda) + \lambda\alpha] && \text{(truncated dual IPB)},\\
g_{\mathrm{dual}}(\alpha) &\triangleq \inf_{\lambda \geq 0} [h_{\mathcal{F}}(\lambda) + \lambda\alpha] && \text{(Lagrangian dual IPB)}.
\end{aligned}$$

By construction, $g(\alpha) \leq g_{\mathrm{dual}}(\alpha) \leq g_{[0,\lambda_{\max}]}(\alpha) \leq g_\Lambda(\alpha)$ for every $\alpha \in [0,1]$, where $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ is the tight IPB.

For a distribution $S \in \{P, Q\}$ and a sample size n, the *Rademacher complexity* of $\mathcal{F}$ is

$$\mathfrak{R}_n(S, \mathcal{F}) \triangleq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)\right],$$

where the expectation is over $x_1, \ldots, x_n$ drawn i.i.d. from S and independent Rademacher variables $\sigma_1, \ldots, \sigma_n$, each uniform on $\{-1, +1\}$. With ε_P and ε_Q as in Theorem 4.3, define the event

$$\mathcal{E} \triangleq \left\{\sup_{f \in \mathcal{F}} \left|\hat{\mathbb{E}}_P[f] - \mathbb{E}_P[f]\right| \leq \varepsilon_P \text{ and } \sup_{f \in \mathcal{F}} \left|\hat{\mathbb{E}}_Q[f] - \mathbb{E}_Q[f]\right| \leq \varepsilon_Q\right\}. \tag{16}$$

As functions in our $\mathcal{F}$ are bounded in $[0,1]$, standard Rademacher complexity bounds guarantee that, for each $S \in \{P, Q\}$, the upper and the lower uniform deviations of $\hat{\mathbb{E}}_S[f]$ from $\mathbb{E}_S[f]$ each exceed ε_S with probability at most $\delta/4$. By a union bound over these four events, we have that $\Pr[\mathcal{E}] \geq 1 - \delta$.

The proof proceeds by decomposing the error of $\hat{g}$ with respect to g into an optimization error, a statistical error, and three structural errors—grid discretization, dual-range truncation, and the duality gap:

$$\hat{g} - g = \underbrace{[\hat{g} - \hat{g}_{\mathrm{opt}}]}_{\text{optimization}} + \underbrace{[\hat{g}_{\mathrm{opt}} - g_\Lambda]}_{\text{statistical}} + \underbrace{[g_\Lambda - g_{[0,\lambda_{\max}]}]}_{\text{discretization}} + \underbrace{[g_{[0,\lambda_{\max}]} - g_{\mathrm{dual}}]}_{\text{truncation}} + \underbrace{[g_{\mathrm{dual}} - g]}_{\text{duality gap}}. \tag{17}$$

We control each term in turn.

Optimization error. Let us show how optimization error propagates to the IPB estimate.

Lemma A.7. *If* $|\hat{h}(\lambda) - \hat{h}_{\mathrm{opt}}(\lambda)| \leq \varepsilon_{\mathrm{opt}}$ *for all* $\lambda \in \Lambda$, *then* $\sup_{\alpha \in [0,1]} |\hat{g}(\alpha) - \hat{g}_{\mathrm{opt}}(\alpha)| \leq \varepsilon_{\mathrm{opt}}$.

Proof. Using $|\min_x A(x) - \min_x B(x)| \leq \max_x |A(x) - B(x)|$, for any α we have $|\hat{g}(\alpha) - \hat{g}_{\mathrm{opt}}(\alpha)| \leq \max_{\lambda \in \Lambda} |\hat{h}(\lambda) - \hat{h}_{\mathrm{opt}}(\lambda)| \leq \varepsilon_{\mathrm{opt}}$. □

Statistical error. We first show that, on the event $\mathcal{E}$, the empirical Lagrangian objective is uniformly close to the population one.

Lemma A.8. *On the event $\mathcal{E}$, simultaneously for every $\lambda \geq 0$,*

$$\left|\hat{h}_{\mathrm{opt}}(\lambda) - h_{\mathcal{F}}(\lambda)\right| \leq \varepsilon_P + \lambda\varepsilon_Q.$$

Proof. By the triangle inequality and the subadditivity of the supremum,

$$\left|\hat{h}_{\mathrm{opt}}(\lambda) - h_{\mathcal{F}}(\lambda)\right| \leq \sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}_P[f] - \mathbb{E}_P[f]\right| + \lambda \sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}_Q[f] - \mathbb{E}_Q[f]\right| \leq \varepsilon_P + \lambda\varepsilon_Q,$$

where the last step holds on the event $\mathcal{E}$. As $\mathcal{E}$ does not depend on λ, the bound holds for all $\lambda \geq 0$ simultaneously. □

The statistical error at a given α is defined by the values of λ that are actually solutions to the optimization in Algorithm 2 at α. For a fixed $\alpha \in [0,1]$, define the *active* population and empirical dual variables as follows:

$$\lambda^*_\alpha \in \arg\min_{\lambda\in\Lambda}\left[h_{\mathcal{F}}(\lambda) + \lambda\alpha\right], \qquad \hat{\lambda}_\alpha \in \arg\min_{\lambda\in\Lambda}\left[\hat{h}_{\mathrm{opt}}(\lambda) + \lambda\alpha\right].$$

Proposition A.9. *Suppose $\mathcal{F} \subseteq \mathcal{F}_1$. On the event $\mathcal{E}$, for every $\alpha \in [0,1]$,*

$$|\hat{g}_{\mathrm{opt}}(\alpha) - g_\Lambda(\alpha)| \leq \varepsilon_P + \max\left\{\lambda^*_\alpha, \hat{\lambda}_\alpha\right\}\varepsilon_Q \leq \varepsilon_P + \lambda_{\max}\varepsilon_Q.$$

Proof. Fix α. Using $\hat{g}_{\mathrm{opt}}(\alpha) \leq \hat{h}_{\mathrm{opt}}(\lambda^*_\alpha) + \lambda^*_\alpha\alpha$ and Lemma A.8,

$$\hat{g}_{\mathrm{opt}}(\alpha) - g_\Lambda(\alpha) \leq \hat{h}_{\mathrm{opt}}(\lambda^*_\alpha) - h_{\mathcal{F}}(\lambda^*_\alpha) \leq \varepsilon_P + \lambda^*_\alpha\varepsilon_Q.$$

Symmetrically, using $g_\Lambda(\alpha) \leq h_{\mathcal{F}}(\hat{\lambda}_\alpha) + \hat{\lambda}_\alpha\alpha$,

$$g_\Lambda(\alpha) - \hat{g}_{\mathrm{opt}}(\alpha) \leq h_{\mathcal{F}}(\hat{\lambda}_\alpha) - \hat{h}_{\mathrm{opt}}(\hat{\lambda}_\alpha) \leq \varepsilon_P + \hat{\lambda}_\alpha\varepsilon_Q.$$

Combining the two directions yields the first inequality; the second follows from $\max\{\lambda^*_\alpha, \hat{\lambda}_\alpha\} \leq \lambda_{\max}$. □

Grid discretization. Next, we show the effect of the chosen grid density. First, we show that the dual objective is Lipschitz in λ.

Lemma A.10. *For $\mathcal{F} \subseteq \mathcal{F}_1$ and every fixed $\alpha \in [0,1]$, the map $\lambda \mapsto h_{\mathcal{F}}(\lambda) + \lambda\alpha$ is 1-Lipschitz on $[0,\infty)$.*

Proof. For each $f \in \mathcal{F}$ and any $\lambda, \lambda' \geq 0$, we have:

$$\begin{aligned}|(\mathbb{E}_P[f] - \lambda\mathbb{E}_Q[f] + \lambda\alpha) - (\mathbb{E}_P[f] - \lambda'\mathbb{E}_Q[f] + \lambda'\alpha)| &= |\alpha - \mathbb{E}_Q[f]| \cdot |\lambda - \lambda'| \\ &\leq |\lambda - \lambda'|,\end{aligned}$$

where the last step holds as α and $\mathbb{E}_Q[f] \in [0,1]$. A pointwise supremum of 1-Lipschitz maps is 1-Lipschitz. □

As a consequence, replacing the minimizer over $[0,\lambda_{\max}]$ by its nearest grid point costs at most the covering radius of the grid.

Proposition A.11. *Suppose $\mathcal{F} \subseteq \mathcal{F}_1$. Then, for every $\alpha \in [0,1]$,*

$$0 \leq g_\Lambda(\alpha) - g_{[0,\lambda_{\max}]}(\alpha) \leq \eta_\Lambda.$$

In particular, if Λ is a uniform grid of $k = |\Lambda|$ points on $[0,\lambda_{\max}]$, then $\eta_\Lambda = \frac{\lambda_{\max}}{2(k-1)}$, and to make the discretization error at most ε it suffices to take $k \geq 1 + \frac{\lambda_{\max}}{2\varepsilon}$.

Proof. The lower bound holds as $\Lambda \subseteq [0,\lambda_{\max}]$. For the upper bound, fix α. By Lemma A.10, the map $\lambda \mapsto h_{\mathcal{F}}(\lambda) + \lambda\alpha$ is continuous, so it attains its minimum over $[0,\lambda_{\max}]$ at some λ^*. Let $\lambda' \in \Lambda$ be the nearest grid point, so that $|\lambda^* - \lambda'| \leq \eta_\Lambda$. By Lemma A.10 again,

$$g_\Lambda(\alpha) \leq h_{\mathcal{F}}(\lambda') + \lambda'\alpha \leq h_{\mathcal{F}}(\lambda^*) + \lambda^*\alpha + \eta_\Lambda = g_{[0,\lambda_{\max}]}(\alpha) + \eta_\Lambda.$$

The expression for a uniform grid is immediate. □

Dual-range truncation. Next, we analyze the error due to truncating the dual variable to $[0, \lambda_{\max}]$. It is zero if the continuous dual problem has a minimizer in $[0, \lambda_{\max}]$. Otherwise, the minimizing dual variable lies beyond $\lambda_{\max}$, and no refinement of the grid inside $[0, \lambda_{\max}]$ can remove this error. First, we show that under a bounded likelihood ratio, the Lagrangian objective vanishes for all large enough dual variables.

Lemma A.12. *Assume $0 \in \mathcal{F} \subseteq \mathcal{F}_1$, $P \ll Q$, and ${}^{dP}/_{dQ} \leq B$ almost surely under Q for some $B < \infty$. Then, for every $\lambda \geq B$, we have $h_{\mathcal{F}}(\lambda) = 0$.*

Proof. As $0 \in \mathcal{F}$, the zero function gives $h_{\mathcal{F}}(\lambda) \geq 0$. For any $f \in \mathcal{F} \subseteq \mathcal{F}_1$, as $P \ll Q$ and $0 \leq f \leq 1$, we have:

$$\mathbb{E}_P[f] - \lambda \mathbb{E}_Q[f] = \mathbb{E}_Q\big[({}^{dP}/_{dQ} - \lambda) f\big] \leq \mathbb{E}_Q\big[({}^{dP}/_{dQ} - \lambda)_+\big].$$

Taking the supremum over $f \in \mathcal{F}$ gives $h_{\mathcal{F}}(\lambda) \leq \mathbb{E}_Q\big[({}^{dP}/_{dQ} - \lambda)_+\big]$. For $\lambda \geq B$, we have ${}^{dP}/_{dQ} - \lambda \leq 0$ almost surely under Q, so the right-hand side is zero, and hence $h_{\mathcal{F}}(\lambda) = 0$. □

As a consequence, the truncation error vanishes for the principled choice $\lambda_{\max} \geq B$.

Proposition A.13. *Assume $0 \in \mathcal{F} \subseteq \mathcal{F}_1$, $P \ll Q$, and ${}^{dP}/_{dQ} \leq B$ almost surely under Q for some $B < \infty$. If $\lambda_{\max} \geq B$, then $g_{[0,\lambda_{\max}]}(\alpha) = g_{\mathrm{dual}}(\alpha)$ for every $\alpha \in [0, 1]$.*

Proof. By Lemma A.12, $h_{\mathcal{F}}(\lambda) = 0$ for all $\lambda \geq B$. Hence, for any $\lambda \geq B$ and $\alpha \geq 0$, we have $h_{\mathcal{F}}(\lambda) + \lambda\alpha = \lambda\alpha \geq B\alpha = h_{\mathcal{F}}(B) + B\alpha$, so no dual variable larger than B can improve the objective. Therefore, the infimum over $\lambda \geq 0$ coincides with the infimum over $[0, B] \subseteq [0, \lambda_{\max}]$. □

Duality gap. Finally, we analyze the last term that vanishes when we have strong duality:

Proposition A.14 (Strong duality)**.** *Suppose $0 \in \mathcal{F} \subseteq \mathcal{F}_1$ is convex, $P \ll Q$, and ${}^{dP}/_{dQ} \leq B$ almost surely under Q for some $B < \infty$. Then $g_{\mathrm{dual}}(\alpha) = g(\alpha)$ for every $\alpha \in [0, 1]$.*

Proof. Let $g = \mathcal{B}^*_{\mathcal{F}}(P\|Q)$ be the tight IPB. Since $\mathcal{F}$ is convex, g is concave. Moreover, since $0 \in \mathcal{F} \subseteq \mathcal{F}_1$, it is finite and non-decreasing on $[0, 1]$. Thus, for every $\alpha \in (0, 1)$, g admits a supergradient $\lambda_\alpha \geq 0$ by Rockafellar (1997, Theorem 23.4), so that:

$$g(t) \leq g(\alpha) + \lambda_\alpha(t - \alpha), \qquad t \in [0, 1].$$

In particular, for every $f \in \mathcal{F}$,

$$\mathbb{E}_P[f] \leq g(\mathbb{E}_Q[f]) \leq g(\alpha) + \lambda_\alpha(\mathbb{E}_Q[f] - \alpha).$$

Taking the supremum over $f \in \mathcal{F}$ yields $h_{\mathcal{F}}(\lambda_\alpha) + \lambda_\alpha\alpha \leq g(\alpha)$. The reverse inequality follows from Proposition 4.2, establishing strong duality. At $\alpha = 1$, equality follows by choosing $\lambda = 0$, since $h_{\mathcal{F}}(0) = g(1)$. At $\alpha = 0$, the bounded likelihood-ratio assumption gives $h_{\mathcal{F}}(B) = 0$ by Lemma A.12, whereas $g(0) = 0$ since $P \ll Q$ and all functions in $\mathcal{F}$ are non-negative. As a result, strong duality holds throughout $[0, 1]$. □

Main result. We are now ready to prove the main statement.

Theorem 4.3. *Suppose $0 \in \mathcal{F} \subseteq \mathcal{F}_1$ is a convex family of bounded functions with their Rademacher complexity terms denoted as $\mathfrak{R}_{n_P}(P, \mathcal{F})$ and $\mathfrak{R}_{n_Q}(Q, \mathcal{F})$, that $P \ll Q$ with ${}^{dP}/_{dQ} \leq B$ almost surely under Q, and that we can solve Eq. (10) up to at most $\varepsilon_{\mathrm{opt}} \geq 0$ error. Let $\Lambda \ni 1$ be a grid with $\lambda_{\max} \triangleq \max \Lambda \geq B$ and covering radius $\eta_\Lambda \triangleq \sup_{\lambda \in [0, \lambda_{\max}]} \min_{\lambda' \in \Lambda} |\lambda - \lambda'|$. Then, for $\delta \in (0, 1)$, w.p. at least $1 - \delta$ over sampling of X_P and X_Q, the output $\hat{g}$ of Algorithm 2 satisfies:*

$$\sup_{\alpha \in [0,1]} |\hat{g}(\alpha) - g(\alpha)| \;\leq\; \varepsilon_P + \lambda_{\max}\, \varepsilon_Q + \eta_\Lambda + \varepsilon_{\mathrm{opt}},$$

where $g = \mathcal{B}^*_{\mathcal{F}}(P \parallel Q)$ is the tight IPB, and $\varepsilon_S \triangleq 2\mathfrak{R}_{n_S}(S, \mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{2n_S}}$ for $S \in \{P, Q\}$.

Proof. On the event $\mathcal{E}$ in Eq. (16), which holds with probability at least $1 - \delta$, we bound each term of the decomposition in Eq. (17) uniformly in $\alpha \in [0, 1]$. The optimization term is at most $\varepsilon_{\mathrm{opt}}$ in absolute value by Lemma A.7. The statistical term is at most $\varepsilon_P + \lambda_{\max}\varepsilon_Q$ in absolute value by Proposition A.9. The discretization term lies in $[0, \eta_\Lambda]$ by Proposition A.11. The truncation term is zero by Proposition A.13, as $\lambda_{\max} \geq B$. The duality gap is zero by Proposition A.14. The claim follows from the triangle inequality. □

Sharpening the dependence on the dual variable. The dependence on $\lambda_{\max}$ in Theorem 4.3 is conservative. By Proposition A.9, for every fixed $\alpha \in [0,1]$, the term $\lambda_{\max}\varepsilon_Q$ can be replaced by $\max\{\lambda^*_\alpha, \hat{\lambda}_\alpha\}\varepsilon_Q$, where λ^*_α and $\hat{\lambda}_\alpha$ are the active population and empirical dual variables at α. This yields a pointwise bound that is tighter whenever the active dual variable at α is below $\lambda_{\max}$.

Extension to non-convex function families. Dropping the convexity assumption on $\mathcal{F}$, the same argument results in the following bound:

$$\sup_{\alpha\in[0,1]} |\hat{g}(\alpha) - g(\alpha)| \le \varepsilon_P + \lambda_{\max}\varepsilon_Q + \eta_\Lambda + \varepsilon_{\text{opt}} + \sup_{\alpha\in[0,1]} [g_{\text{dual}}(\alpha) - g(\alpha)]\,,$$

where the last term is the duality gap. For a general non-convex learnable family $\mathcal{F}$, strong duality need not hold, and thus $g_{\text{dual}} - g$ is a structural gap: the other remaining terms do not control it, as grid density and $\lambda_{\max}$ only control the finite-grid approximation to the truncated dual problem. In this setting, the output of Algorithm 2 should be interpreted as producing the Lagrangian dual IPB g_{dual} rather than the tight IPB, which by Proposition 4.2 remains a valid IPB in $\mathcal{B}_{\mathcal{F}}(P \,\|\, Q)$.

Extension to unbounded function families. Algorithm 2 itself is applicable to functions with an arbitrary range, although we state Theorem 4.3 for $\mathcal{F} \subseteq \mathcal{F}_1$. The guarantee extends directly to a bounded range $[-M, M]$ by scaling: we multiply the statistical term ε_S for $S \in \{P, Q\}$ and the Lipschitz constant in Lemma A.10—hence the discretization error in Proposition A.11—by $2M$.

For unbounded families, we use a clipping argument analogous to Eq. (14). Define the clipping operator and the clipped family:

$$T_M(u) \triangleq \max\{-M, \min\{u, M\}\}, \qquad \mathcal{F}_M \triangleq \{T_M \circ f \mid f \in \mathcal{F}\},$$

and run Algorithm 2 on $\mathcal{F}_M$, i.e., with the empirical objective

$$\hat{h}_{M,\text{opt}}(\lambda) \triangleq \sup_{f\in\mathcal{F}} \left[\hat{\mathbb{E}}_P[T_M(f)] - \lambda\hat{\mathbb{E}}_Q[T_M(f)]\right],$$

which is exactly Eq. (10) with $\mathcal{F}$ replaced by its clipped version $\mathcal{F}_M$. As before, we denote by $\hat{h}_M(\lambda)$ the value returned by the optimization oracle, with $|\hat{h}_M(\lambda) - \hat{h}_{M,\text{opt}}(\lambda)| \le \varepsilon_{\text{opt}}$ for all $\lambda \in \Lambda$. The resulting IPB estimate is $\hat{g}_{\Lambda,M}(\alpha) \triangleq \min_{\lambda\in\Lambda}[\hat{h}_M(\lambda) + \lambda\alpha]$. The only extra error is the bias due to clipping. As $|f - T_M(f)| = (|f| - M)_+$, we can quantify it by:

$$\tau_{S,M} \triangleq \sup_{f\in\mathcal{F}} \mathbb{E}_S\left[(|f| - M)_+\right], \qquad S \in \{P, Q\}.$$

Proposition A.15. *Let $\mathcal{F}$ be a family of real-valued functions with $\tau_{P,M}, \tau_{Q,M} < \infty$. Then, for every $\lambda \ge 0$,*

$$|h_{\mathcal{F}}(\lambda) - h_{\mathcal{F}_M}(\lambda)| \le \tau_{P,M} + \lambda\tau_{Q,M}.$$

Consequently, with probability at least $1-\delta$, the output $\hat{g}_{\Lambda,M}$ of Algorithm 2 run on $\mathcal{F}_M$ with an optimization oracle of error at most ε_{opt} satisfies

$$\sup_\alpha |\hat{g}_{\Lambda,M}(\alpha) - g_\Lambda(\alpha)| \le \varepsilon_{P,M} + \lambda_{\max}\varepsilon_{Q,M} + \tau_{P,M} + \lambda_{\max}\tau_{Q,M} + \varepsilon_{\text{opt}},$$

where $\varepsilon_{S,M} \triangleq 2\mathfrak{R}_{n_S}(S, \mathcal{F}_M) + M\sqrt{2\log(4/\delta)/n_S}$ for $S \in \{P, Q\}$, and g_Λ is the grid-restricted dual IPB for the original family $\mathcal{F}$.

Proof. For the second claim, let $\hat{g}_{\Lambda,M,\text{opt}}(\alpha) \triangleq \min_{\lambda\in\Lambda}[\hat{h}_{M,\text{opt}}(\lambda) + \lambda\alpha]$ denote the estimate under an exact oracle. By Lemma A.7, $\sup_\alpha |\hat{g}_{\Lambda,M}(\alpha) - \hat{g}_{\Lambda,M,\text{opt}}(\alpha)| \le \varepsilon_{\text{opt}}$. Since $\mathcal{F}_M$ is a family of $[-M, M]$-valued functions, symmetrization implies:

$$\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_M} \left[\hat{\mathbb{E}}_S[\varphi] - \mathbb{E}_S[\varphi]\right]\right] \le 2\mathfrak{R}_{n_S}(S, \mathcal{F}_M).$$

Also, since $|\varphi| \le M$, replacing one observation changes the empirical average of any $\varphi \in \mathcal{F}_M$ by at most $2M/n_S$. The same bounded-differences property holds for the supremum. The analogous argument applies to the negative deviation $\sup_{\varphi\in\mathcal{F}_M} \left[\mathbb{E}_S[\varphi] - \hat{\mathbb{E}}_S[\varphi]\right]$, using the symmetry of the

Rademacher variables. Thus, McDiarmid's inequality and a union bound over both deviations and $S \in \{P, Q\}$ yield, with probability at least $1 - \delta$ for any $\delta \in (0, 1)$,

$$\sup_{\varphi \in \mathcal{F}_M} |\hat{\mathbb{E}}_S[\varphi] - \mathbb{E}_S[\varphi]| \le \varepsilon_{S,M}, \qquad S \in \{P, Q\},$$

Thus, the same argument as in Proposition A.9 gives:

$$\sup_{\alpha} |\hat{g}_{\Lambda,M,\mathrm{opt}}(\alpha) - g_{\Lambda,M}(\alpha)| \le \varepsilon_{P,M} + \lambda_{\max}\varepsilon_{Q,M},$$

where $g_{\Lambda,M}(\alpha) \triangleq \min_{\lambda \in \Lambda}[h_{\mathcal{F}_M}(\lambda) + \lambda\alpha]$. By the first claim and $|\min_x A(x) - \min_x B(x)| \le \max_x |A(x) - B(x)|$, we have $\sup_\alpha |g_{\Lambda,M}(\alpha) - g_\Lambda(\alpha)| \le \tau_{P,M} + \lambda_{\max}\tau_{Q,M}$. The triangle inequality completes the proof.

□

This extension demonstrates a trade-off. Increasing M makes the clipped family closer to the original one and reduces the tail bias $\tau_{S,M}$, but increases the bounded-range error term. The bias terms vanish fast under mild tail assumptions: e.g., if $f(x)$ is σ-sub-Gaussian under S uniformly over $f \in \mathcal{F}$, then $\tau_{S,M} = \mathcal{O}\left(\sigma \exp(-M^2/2\sigma^2)\right)$, so choosing $M \asymp \sigma\sqrt{\log n}$ yields a total error of order $\varepsilon_P + \lambda_{\max}\varepsilon_Q$ up to logarithmic factors.

A.4 Proofs for Section 5

Theorem 5.1. *For any bounded performance measure $f \in \mathcal{F}_1$, any reward function $r : \mathcal{Y} \to \mathbb{R}$ and any reference distribution Q:*

$$\mathbb{E}_{y \sim P_N}[f(y)] \le 1 - (1 - \mathbb{E}_{y \sim Q}[f(y)])^N. \tag{11}$$

The inequality is tight if r guarantees no ties.

Proof. The proof for the case without ties is given in the main body. The no-ties assumption in Theorem 5.1 is only needed to obtain the exact closed form $g(\alpha) = 1 - (1 - \alpha)^N$. Without this assumption, the same expression remains a valid, though not necessarily tight, IPB. Indeed, let $Y_1, \ldots, Y_N \sim Q$, and let $\hat{Y}$ be the output selected by BoN, with any tie-breaking rule that selects one of the maximizers uniformly at random. For any $f \in \mathcal{F}_1$,

$$f(\hat{Y}) \le \max_{i \in [N]} f(Y_i) \le 1 - \prod_{i=1}^{N}(1 - f(Y_i)).$$

To see why the second inequality holds, let i^* be an index attaining the maximum. As $0 \le f \le 1$, every factor $1 - f(Y_i)$ lies in $[0, 1]$, so $\prod_{i=1}^N (1 - f(Y_i)) \le 1 - f(Y_{i^*})$, thus we get the inequality by re-arranging. Taking expectations and using independence gives:

$$\mathbb{E}_{P_N}[f] \le 1 - \prod_{i=1}^{N} \mathbb{E}_Q[1 - f(Y_i)] = 1 - (1 - \mathbb{E}_Q[f])^N.$$

Therefore, for every $\alpha \in [0, 1]$,

$$g(\alpha) = \sup_{f \in \mathcal{F}_1 : \mathbb{E}_Q[f] \le \alpha} \mathbb{E}_{P_N}[f] \le 1 - (1 - \alpha)^N.$$

□

B Additional Results

B.1 IPBs and f-Divergences

We can obtain an IPB using a standard information-theoretic notion of an f-divergence. We call them ϕ-divergences in this manuscript to avoid the clash of nomenclature with $f \in \mathcal{F}$.

Definition B.1 (ϕ-divergence)**.** *Let P and Q be probability distributions over a space $\mathcal{X}$ such that P is absolutely continuous with respect to Q ($P \ll Q$). Let $\phi : (0, \infty) \to \mathbb{R}$ be a convex function such that $\phi(1) = 0$. The ϕ-divergence of P from Q is defined as:*

$$D_\phi(P\|Q) = \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right] \tag{18}$$

Proposition B.2 (ϕ-divergence to IPB)**.** *For any valid ϕ-divergence $D_\phi(P\|Q)$, $g(\alpha)$ as defined next is an IPB, $g \in \mathcal{B}_{\mathcal{F}_1}(P\|Q)$:*

$$g(\alpha) = \inf_{t>0}\left\{\frac{D_\phi(P\|Q) + \alpha\phi^*(t) + (1-\alpha)\phi^*(0)}{t}\right\} \tag{19}$$

where $\phi^*(y) \triangleq \sup_{x\in\mathbb{R}}(xy - \phi(x))$ is the convex conjugate of ϕ.

Proof. By the variational representation of ϕ-divergences:

$$D_\phi(P\|Q) = \sup_{h:\ \mathcal{X}\to\mathbb{R}} \{\mathbb{E}_P[h] - \mathbb{E}_Q[\phi^*(h)]\}.$$

Fix any $f \in \mathcal{F}_1$ with $\mathbb{E}_Q[f] = \alpha$, and let $\beta \triangleq \mathbb{E}_P[f]$. For any $t > 0$, taking $h(x) = tf(x)$ in the variational representation yields a lower bound on the divergence:

$$D_\phi(P\|Q) \geq \mathbb{E}_P[tf] - \mathbb{E}_Q[\phi^*(tf)] = t\beta - \mathbb{E}_Q[\phi^*(tf)].$$

Because the range of f is bounded within $[0, 1]$, we can bound the convex conjugate term using Jensen's inequality. For any fixed x, we treat $f(x)$ as the expectation of a Bernoulli random variable $Z_x \sim \mathsf{Bern}(f(x))$. As ϕ^* is convex, by Jensen's inequality:

$$\phi^*(tf(x)) = \phi^*(\mathbb{E}[tZ_x]) \leq \mathbb{E}[\phi^*(tZ_x)] = f(x)\phi^*(t) + (1 - f(x))\phi^*(0).$$

Taking the expectation under Q and using $\mathbb{E}_Q[f] = \alpha$:

$$\mathbb{E}_Q[\phi^*(tf)] \leq \alpha\phi^*(t) + (1-\alpha)\phi^*(0).$$

Combining with the variational inequality and rearranging to isolate $t\beta$:

$$t\beta \leq D_\phi(P\|Q) + \alpha\phi^*(t) + (1-\alpha)\phi^*(0).$$

Dividing by $t > 0$ and taking the infimum over all $t > 0$:

$$\beta \ \leq\ \inf_{t>0}\left\{\frac{D_\phi(P\|Q) + \alpha\phi^*(t) + (1-\alpha)\phi^*(0)}{t}\right\} \ =\ g(\alpha).$$

This holds for any $f \in \mathcal{F}_1$ with $\mathbb{E}_Q[f] = \alpha$, so $g \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$. □

We apply this to KL divergence for an illustration.

Corollary B.3. *The following function is an IPB, $g \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$:*

$$g(\alpha) = \inf_{t>0}\left\{\frac{D_{\mathsf{KL}}(P\|Q) + \alpha(e^t - 1)}{t}\right\}. \tag{20}$$

Proof. We apply the general ϕ-divergence bound from the previous theorem using $\phi(x) = x\log x - x + 1$, which generates the KL divergence, $D_{\mathsf{KL}}(P\|Q)$. Its convex conjugate is:

$$\phi^*(y) = \sup_{x>0}\{xy - (x\log x - x + 1)\} = e^y - 1,$$

giving $\phi^*(t) = e^t - 1$ and $\phi^*(0) = 0$. Substituting into the inf expression from the previous theorem:

$$g(\alpha) = \inf_{t>0}\left\{\frac{D_{\mathsf{KL}}(P\|Q) + \alpha(e^t-1) + (1-\alpha)(0)}{t}\right\} = \inf_{t>0}\left\{\frac{D_{\mathsf{KL}}(P\|Q) + \alpha(e^t-1)}{t}\right\}. \quad \square$$

Algorithm 3: Doubling search for $\lambda_{\max}$

Input: Samples $X_P \sim P^{n_P}$, $X_Q \sim Q^{n_Q}$; family $\mathcal{F}$; tolerance $\tau > 0$; target $h_{\min} > 0$
Output: Upper bound $\lambda_{\max}$
$\lambda \leftarrow 1$
repeat
  $\lambda \leftarrow 2 \cdot \lambda$
  $\hat{h}(\lambda) \leftarrow \max_{f \in \mathcal{F}} \left[\frac{1}{n_P} \sum_{x \in X_P} f(x) - \lambda \frac{1}{n_Q} \sum_{x \in X_Q} f(x) \right]$
until $\hat{h}(\lambda) < h_{\min}$ *or* $\hat{h}(\lambda/2) - \hat{h}(\lambda) < \tau$;
return λ

Algorithm 4: Lower-boundary IPB estimation for a learnable family $\mathcal{F}$

Input: Samples $X_P \sim P^{n_P}$, $X_Q \sim Q^{n_Q}$; family $\mathcal{F}$; grid $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$
Output: Lower boundary estimate $\breve{g}^{-1}$
for $\lambda \in \Lambda \cup \{1\}$ **do**

$$\hat{v}(\lambda) \leftarrow \min_{f \in \mathcal{F}} \left[\frac{1}{n_P} \sum_{x \in X_P} f(x) - \lambda \frac{1}{n_Q} \sum_{x \in X_Q} f(x) \right] \tag{21}$$

end
$\breve{g}^{-1}(\alpha) \leftarrow \max_{\lambda \in \Lambda \cup \{1\}} [\hat{v}(\lambda) + \lambda\, \alpha]$
return $\breve{g}^{-1}$.

B.2 Additional Algorithms

Searching for the $\lambda_{\max}$ in Algorithm 2. It is not immediately clear how to choose the grid Λ in Algorithm 2 in a principled way. In particular, although any $\lambda \geq 0$ is admissible, it is unclear what is a reasonable value of $\lambda_{\max}$ to set in practice. As we have shown in Lemma A.12, for regular enough P and Q, the Lagrangian objective in Eq. (10) is zero when λ becomes large enough. As a practical consequence, we want to heuristically find $\lambda_{\max}$ at which $\hat{h}(\lambda_{\max})$ is numerically small enough. In Algorithm 3 we provide a procedure to do so using doubling search until either $\hat{h}$ is small enough, or, to prevent looping forever, until the decrease in $\hat{h}$ becomes stagnant.

Estimating the lower boundary in Algorithm 2 setup. Algorithm 2 estimates the IPB that corresponds to the upper boundary of the achievable region $\mathcal{C}_{\mathcal{F}}(P \parallel Q)$. Although we could apply the same algorithm to estimate the adjoint $\breve{g} \in \mathcal{B}_{\mathcal{F}}(Q \parallel P)$ and use the generalized inverse $\breve{g}^{-1}$ to get the lower boundary, for families $\mathcal{F}$ of, e.g., classifiers, it would be more convenient to find an algorithm that estimates the lower boundary directly. To simplify this, we present a version of the procedure that is adapted so that we do not need to compute the inverse. The adapted algorithm follows the same construction but replaces maximization with minimization, and vice versa, based on the following analogous result.

Proposition B.4 (Lagrangian dual lower boundary)**.** *For any $\lambda \geq 0$, define*

$$v_{\mathcal{F}}(\lambda) \triangleq \inf_{f \in \mathcal{F}} \big[\mathbb{E}_P[f] - \lambda\, \mathbb{E}_Q[f] \big].$$

Then for every $f \in \mathcal{F}$ and every $\alpha \leq \mathbb{E}_Q[f]$,

$$\mathbb{E}_P[f] \;\geq\; \sup_{\lambda \geq 0} [v_{\mathcal{F}}(\lambda) + \lambda\, \alpha].$$

Proof. By definition, $v_{\mathcal{F}}(\lambda) \leq \mathbb{E}_P[f] - \lambda \mathbb{E}_Q[f]$ for any $f \in \mathcal{F}$. Rearranging gives $\mathbb{E}_P[f] \geq v_{\mathcal{F}}(\lambda) + \lambda \mathbb{E}_Q[f]$. We get the statement by considering $\alpha \leq \mathbb{E}_Q[f]$. □

The finite-sample uniform-convergence guarantee of Theorem 4.3 applies with trivial changes.

B.3 Deriving Tight Bounds on Generalization and Attack Risk in Differential Privacy

We show how IPBs unify and recover recent results on generalization and operational risk under differential privacy. Differential privacy (DP) (Dwork and Roth, 2014) is the standard framework for ensuring strong privacy guarantees in statistics and machine learning. We formalize a learning algorithm as a randomized *mechanism* $M : 2^{\mathcal{X}} \to \Theta$. We say that two datasets $S, S' \in \mathcal{X}^n$ are *neighbouring*, denoted $S \simeq S'$, if they differ in at most one record (e.g., by replacement, or by addition or removal of a single record). Next, we introduce a form of differential privacy known as f-DP (Dong et al., 2022). This definition characterizes the privacy guarantees via the ROC curve of the optimal hypothesis test that distinguishes neighbouring datasets given the mechanism's output:

Definition B.5 (Dong et al., 2022)**.** *A mechanism $M : 2^{\mathcal{X}} \to \Theta$ satisfies f-DP, where $f : [0,1] \to [0,1]$ is a non-increasing, convex function bounded as $f(\alpha) \leq 1-\alpha$ for $\alpha \in [0,1]$, if for any $S \simeq S'$ and any $\alpha \in [0,1]$, the function f is a lower bound on the error rate of the most powerful hypothesis test between the distributions of $M(S)$ and $M(S')$:*

$$f(\alpha) \leq \inf_{\phi:\, \Theta \to [0,1]} \{\beta_\phi \mid \alpha_\phi \leq \alpha\}, \tag{22}$$

where $\beta_\phi \triangleq 1 - \mathbb{E}_{M(S')}[\phi]$ is the false-negative rate (FNR) of the test, and $\alpha_\phi \triangleq \mathbb{E}_{M(S)}[\phi]$ is the false-positive rate (FPR).

By Proposition 2.4, we can immediately see that the function $1 - f$ is an IPB over $\mathcal{F}_1$:

Proposition B.6 (f-DP to IPB)**.** *If a mechanism $M(\cdot)$ satisfies f-DP, then for any $S \simeq S'$, we have that $g = 1 - f$ is an IPB, $g \in \mathcal{B}_{\mathcal{F}_1}(M(S) \parallel M(S'))$.*

Next, we recover two results recently shown by a subset of the authors (Kulynych et al., 2025) on generalization and risk bounds in f-DP.

Bounds on average generalization error. For a bounded loss $\ell : \mathcal{X} \times \Theta \to [0,1]$ and a data distribution D over $\mathcal{X}$, define the on-average train and test losses:

$$\mathrm{err}_{\mathrm{tr}} \triangleq \mathbb{E}_{\substack{S \sim D^n \\ \theta \sim M(S)}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta) \right], \qquad \mathrm{err}_{\mathrm{test}} \triangleq \mathbb{E}_{\substack{S \sim D^n,\, z \sim D \\ \theta \sim M(S)}} [\ell(z; \theta)],$$

where $S = (z_1, \ldots, z_n)$.

Proposition B.7 (On-average generalization for f-DP)**.** *Suppose that $M(\cdot)$ satisfies f-DP w.r.t. the replace-one relation, and let $g \triangleq 1 - f$. Then, for any bounded loss function $\ell : \mathcal{X} \times \Theta \to [0,1]$ and any data distribution D over $\mathcal{X}$, we have:*

$$\mathrm{err}_{\mathrm{tr}} \leq g(\mathrm{err}_{\mathrm{test}}), \qquad \mathrm{err}_{\mathrm{test}} \leq g(\mathrm{err}_{\mathrm{tr}}).$$

Proof. Fix any $z, z' \in \mathcal{X}$ and $\bar{S} \in \mathcal{X}^{n-1}$. The datasets $\bar{S} \cup \{z\}$ and $\bar{S} \cup \{z'\}$ differ in at most one record and are therefore neighbouring under the replace-one relation, so by Proposition B.6, $g \in \mathcal{B}_{\mathcal{F}_1}(M(\bar{S} \cup \{z\}) \parallel M(\bar{S} \cup \{z'\}))$. Applying the main IPB property in Proposition 1.2 to the bounded function $\psi(\theta) \triangleq \ell(z; \theta)$:

$$\mathbb{E}_{\theta \sim M(\bar{S} \cup \{z\})}[\ell(z; \theta)] \;\leq\; g\big(\mathbb{E}_{\theta \sim M(\bar{S} \cup \{z'\})}[\ell(z; \theta)]\big)\,.$$

Next, we take the expectation over $z, z' \sim D$ and $\bar{S} \sim D^{n-1}$ on both sides. By the concavity of g (Proposition A.1) and Jensen's inequality, we can pull the outer expectation inside $g(\cdot)$ on the right-hand side:

$$\mathbb{E}_{\substack{z, \bar{S} \\ \theta \sim M(\bar{S} \cup \{z\})}} [\ell(z; \theta)] \;\leq\; g\left(\mathbb{E}_{\substack{z, z', \bar{S} \\ \theta \sim M(\bar{S} \cup \{z'\})}} [\ell(z; \theta)] \right).$$

The l.h.s. equals $\mathrm{err}_{\mathrm{tr}}$ as the random dataset $\bar{S} \cup \{z\}$ is distributed as D^n with z being one of its records. On the r.h.s., $\bar{S} \cup \{z'\} \sim D^n$, but $z \sim D$ is independent of it, so the argument equals $\mathrm{err}_{\mathrm{test}}$. The reverse direction follows by symmetry, swapping the roles of z and z'. □

Bounds on operational attack risks. IPBs also yield bounds on the success of re-identification, attribute inference, and data reconstruction attacks, equivalent to those recently derived by a subset of the authors (Kulynych et al., 2025). The authors have also shown that these bounds are empirically the tightest known bounds on attack risk in many settings, e.g., for general reconstruction attacks.

Consider a strong-adversary threat model: the adversary holds a partial dataset $\bar{S} \in \mathcal{X}^{n-1}$ along with a prior distribution D over the unknown record, observes a release $\theta \sim M(\bar{S} \cup \{z\})$ for $z \sim D$, and runs an attack whose success is captured by a bounded function $q(z;\theta) \in [0,1]$. For example, $q(z;\theta) = \mathbb{1}[\ell(z, \mathcal{A}(\theta)) \leq \gamma]$ recovers reconstruction with attack $\mathcal{A}$, loss ℓ, and threshold γ; $q(z;\theta) = \mathbb{1}[a(z) = \hat{a}(\theta)]$ recovers attribute inference.

Proposition B.8 (Attack risk under f-DP)**.** *Suppose that $M(\cdot)$ satisfies f-DP w.r.t. either the replace-one or add-remove relation, and let $g \triangleq 1 - f$. Then, for any bounded $q : \mathcal{X} \times \Theta \to [0,1]$ which measures attack success, any partial dataset $\bar{S} \in \mathcal{X}^{n-1}$ with $n \geq 1$, and any prior distribution D over $\mathcal{X}$, the expected attack success rate is bounded as follows:*

$$\mathbb{E}_{\substack{z\sim D \\ \theta\sim M(\bar{S}\cup\{z\})}}[q(z;\theta)] \;\leq\; g\left(\sup_{\theta\in\Theta} \mathbb{E}_{z\sim D}[q(z;\theta)]\right).$$

Proof. Fix any $z, z' \in \mathcal{X}$. By Proposition B.6 and the main IPB property in Proposition 1.2 applied to $\psi(\theta) = q(z;\theta)$:

$$\mathbb{E}_{\theta\sim M(\bar{S}\cup\{z\})}[q(z;\theta)] \;\leq\; g\big(\mathbb{E}_{\theta\sim M(\bar{S}\cup\{z'\})}[q(z;\theta)]\big).$$

The same inequality holds under the add-remove relation when applied to $M(\bar{S} \cup \{z\})$ and $M(\bar{S})$. Take the expectation over $z \sim D$ on both sides. The concavity of g together with Jensen's inequality lets us push this expectation inside $g(\cdot)$ on the right-hand side:

$$\mathbb{E}_{\substack{z\sim D \\ \theta\sim M(\bar{S}\cup\{z\})}}[q(z;\theta)] \;\leq\; g\left(\mathbb{E}_{\substack{z\sim D \\ \theta\sim M(\bar{S}\cup\{z'\})}}[q(z;\theta)]\right).$$

On the right-hand side, z' is fixed, so z is independent of $\theta \sim M(\bar{S} \cup \{z'\})$. Swapping the order of expectations and bounding by the supremum yields:

$$\mathbb{E}_{z\sim D}\,\mathbb{E}_{\theta\sim M(\bar{S}\cup\{z'\})}[q(z;\theta)] \;=\; \mathbb{E}_{\theta\sim M(\bar{S}\cup\{z'\})}\,\mathbb{E}_{z\sim D}[q(z;\theta)] \;\leq\; \sup_{\theta\in\Theta}\mathbb{E}_{z\sim D}[q(z;\theta)].$$

The result follows because g is non-decreasing. □

B.4 Finite-Sample Certificate for Algorithm 2

In this section, we introduce a variant of Algorithm 2 that provides a finite-sample certificate. We use the notation of the proof of Theorem 4.3 in Appendix A.

Proposition A.9 shows that the statistical error at any given α is controlled by the line that supports the boundary at α rather than by the largest grid point. The active population dual variable λ^*_α, however, is not observable. To obtain an actual finite-sample certificate, we can incorporate the statistical and optimization error terms directly into Algorithm 2, penalizing each supporting line by its own uncertainty before taking the minimum.

Proposition B.9 (Finite-sample IPB certificate)**.** *Suppose $\mathcal{F} \subseteq \mathcal{F}_1$, and define the corrected estimate as follows:*

$$\hat{g}_{\mathrm{cert}}(\alpha) \triangleq \min\left\{1, \min_{\lambda\in\Lambda}\left[\hat{h}(\lambda) + \varepsilon_P + \lambda\varepsilon_Q + \varepsilon_{\mathrm{opt}} + \lambda\alpha\right]\right\}.$$

Then, with probability at least $1-\delta$, $\hat{g}_{\mathrm{cert}} \in \mathcal{B}_{\mathcal{F}}(P \parallel Q)$ is a valid IPB simultaneously for all $\alpha \in [0,1]$, i.e., $\mathbb{E}_P[f] \leq \hat{g}_{\mathrm{cert}}(\mathbb{E}_Q[f])$ for every $f \in \mathcal{F}$. Moreover, on the same event, for every $\alpha \in [0,1]$,

$$g_\Lambda(\alpha) \leq \hat{g}_{\mathrm{cert}}(\alpha) \leq g_\Lambda(\alpha) + 2\left(\varepsilon_P + \lambda^*_\alpha\varepsilon_Q + \varepsilon_{\mathrm{opt}}\right).$$

Proof. On the event $\mathcal{E}$, by Lemma A.8, simultaneously for every $\lambda \in \Lambda$,

$$h_{\mathcal{F}}(\lambda) \leq \hat{h}_{\mathrm{opt}}(\lambda) + \varepsilon_P + \lambda\varepsilon_Q \leq \hat{h}(\lambda) + \varepsilon_P + \lambda\varepsilon_Q + \varepsilon_{\mathrm{opt}}.$$

Adding $\lambda\alpha$ and taking the minimum over $\lambda \in \Lambda$ yields $g_\Lambda(\alpha) \le \hat{g}_{\mathrm{cert}}(\alpha)$. As $g_\Lambda \in \mathcal{B}_{\mathcal{F}}(P \,\|\, Q)$ by Proposition 4.2, so is $\hat{g}_{\mathrm{cert}}$. For the upper bound, evaluate the corrected envelope at the active population dual variable and apply Lemma A.8:

$$\begin{aligned}\hat{g}_{\mathrm{cert}}(\alpha) &\le \hat{h}(\lambda^*_\alpha) + \varepsilon_P + \lambda^*_\alpha \varepsilon_Q + \varepsilon_{\mathrm{opt}} + \lambda^*_\alpha \alpha \\ &\le h_{\mathcal{F}}(\lambda^*_\alpha) + \lambda^*_\alpha \alpha + 2\left(\varepsilon_P + \lambda^*_\alpha \varepsilon_Q + \varepsilon_{\mathrm{opt}}\right) \\ &= g_\Lambda(\alpha) + 2(\varepsilon_P + \lambda^*_\alpha \varepsilon_Q + \varepsilon_{\mathrm{opt}}).\end{aligned}$$

□

The certificate is sharper than a uniform $\lambda_{\max}$ correction, because large dual variables are penalized individually before the minimum is taken, so the excess over g_Λ at each α scales with the active λ^*_α. It requires no knowledge of λ^*_α: once the grid and the confidence values are fixed, the algorithm outputs $\hat{g}_{\mathrm{cert}}$ directly. In practice, the certificate is only informative for those function families for which we have tight enough analytical expressions for the Rademacher complexity terms in ε_P and ε_Q, and for which the optimization oracle admits a uniform error bound $\varepsilon_{\mathrm{opt}}$, e.g., for convex problems, or when the inner maximization has a closed form so that $\varepsilon_{\mathrm{opt}} = 0$. We provide a case study for the family of bounded linear functions in Section C.3.

C Experimental Details

C.1 Details for Section 5.2

Applying the algorithm. We apply Algorithm 1 as follows. Observe that the algorithm only depends on the language model distributions M and M', as the probability of observing $x \sim \mathsf{Unif}[D]$ cancels out in the likelihood ratio:

$$L(x, y) \triangleq \frac{dP}{dQ}(x, y) = \frac{\Pr_{X \sim \mathsf{Unif}[D]}[X = x] \cdot M'(y \mid x)}{\Pr_{X \sim \mathsf{Unif}[D]}[X = x] \cdot M(y \mid x)} = \frac{M'(y \mid x)}{M(y \mid x)}.$$

Thus, to compute the ratio for any completion y, it is sufficient to estimate the probabilities $M(y \mid x)$ and $M'(y \mid x)$. Practically, we do so by passing each prefix of the sequence $y_{\le j}$ as input, and getting the probability of the next token y_{j+1}:

$$M(y \mid x) = \prod_{j=0}^{m-1} M(y_{j+1} \mid y_{\le j}, x).$$

Note that in general this expression would need to be adjusted depending on the sampling method. As currently defined, it is consistent with the sampling parameters we use next.

Setup details. Each deployment variant P is served via vLLM. We sample with `temperature` $= 1.0$, `top_p` $= 1.0$, `max_tokens` $= 512$, `logprobs` $= 1$. Prompts are formatted with the Llama-3.1 chat template with the benchmark request as the user turn. We use an NVIDIA A100-40GB machine via Modal for these experiments.

We sample the completions under Q from the base model loaded via huggingface `transformers` in `bfloat16`, with the same sampling parameters, and record their per-token log-probabilities under M. For each completion sampled under Q, we obtain the per-token log-probabilities under M' from a single forward pass over the sequence of both the prompt and completion through the deployment variant in vLLM, using `prompt_logprobs` $= 1$. The sequence-level log-likelihood ratio is the sum of per-token differences, $\log L(x, y) = \sum_{j=0}^{m-1} \left[\log M'(y_{j+1} \mid y_{\le j}, x) - \log M(y_{j+1} \mid y_{\le j}, x)\right]$, clipped to $[-50, 50]$ before exponentiation for numerical stability.

To estimate KL divergence, we use the standard Monte-Carlo estimator:

$$\widehat{D}_{\mathsf{KL}}(Q \,\|\, P) \triangleq -\frac{1}{n} \sum_{i=1}^{n} \log L(x_i, y_i), \qquad (x_i, y_i) \sim Q.$$

We use Anthropic's `claude-haiku-4-5-20251001` as the compliance judge with `max_tokens` $= 4$. Each (x, y) pair is scored independently with the following prompt (the `response` variable is truncated to 3000 characters):

```
Evaluate whether the AI assistant complied with or refused the
following request.

Request: {prompt}

Response: {response}

Output 1 if the assistant meaningfully complied (provided the
harmful content or information requested).
Output 0 if it refused, declined, or avoided the request.
Respond with only the digit 0 or 1.
```

We set $f(x, y) = 1$ when the judge's first non-whitespace character is `1`, and $f(x, y) = 0$ otherwise.

C.2 Details for Section 5.3

Applying the algorithm. To maximize the objective in Eq. (10) using the decision tree implementation in `sklearn`, it is sufficient to set the sample weights to $1/n_P$ for samples from P and to λ/n_Q for samples from Q.

To find $\lambda_{\max}$ with Algorithm 3 we use the tolerance $\tau = 0.01$, and $h_{\min} = 1/n_P$. The search stopped after just one iteration, as the Lagrangian was smaller than $h_{\min}$ for $\lambda = 2$.

Setup details. We train CTGAN (Xu et al., 2019) for 120 epochs on the complementary training split, using the default architecture from the `sdv` package: generator dimensions $(256, 256)$, discriminator dimensions $(256, 256)$, and batch size 500. For Algorithm 2, in addition to $\lambda = 1$, we use a grid of 9 values in $[10^{-2}, \lambda_{\max}]$ (`np.geomspace(1e-2, `$\lambda_{\max}$`, 9)`). We use an NVIDIA A10G machine via Modal for training of the generator, with a total runtime under one hour.

Runtime. In practical applications that we foresee, relevant learnable families are relatively simple, such as the decision trees in this section, thus we do not think that the need to solve a separate optimization problem for each $\lambda \in \Lambda$ is a substantial practical barrier. In particular, in this setup, with a relatively large dataset of more than 70,000 records, estimating $\hat{h}(\lambda)$ for a single grid point takes on average 44ms on a commodity laptop, and the entire algorithm takes 0.5s for our grid of 10 points (see Table 1).

Table 1: Runtime of Algorithm 2 in the setup of Section 5.3 (depth-3 decision trees, $n_P = n_Q = 75{,}763$) on a commodity laptop. We report the mean and standard deviation over the given number of trials.

Operation	Trials	Runtime (mean $\pm$ std)
Single grid point	30	44.0 ± 9.3 ms
Entire grid ($\lvert\Lambda\rvert = 10$)	10	0.50 ± 0.05 s

Additional results. Fig. 8 repeats the analysis with two different seeds for dataset splitting. Our conclusions remain the same.

C.3 Certified IPBs for Bounded Linear Functions

Proposition B.9 provides a finite-sample certificate for the output of Algorithm 2 for those function families for which we have tight Rademacher complexity expressions and where we can solve the optimization problem in Eq. (10) reliably. We illustrate this setting with a case study. We use a family based on the bounded linear functions of Oesterling et al. (2024).

$$\mathcal{F} = \{x \mapsto w \cdot x \mid w \geq 0,\ \|w\|_2 \leq 1\}.$$

We apply this family to the experimental setup of Section 5.3, i.e., we use it to evaluate the fidelity of the CTGAN-generated synthetic data to the real population. Note that it is possible to find an IPB for this family directly by solving the constrained optimization $\sup_{f \in \mathcal{F}:\, \mathbb{E}_Q[f] \leq \alpha} \mathbb{E}_P[f]$ without having to use Algorithm 2, but our goal in this case study is to demonstrate the application of Algorithm 2.

Error terms. This family is convex and contains the zero function, thus the duality gap is zero by Proposition A.14. We normalize the samples from P and Q to have unit norm, so that every $f \in \mathcal{F}$ takes values in $[0, 1]$, as all features are non-negative. The Rademacher complexity of $\mathcal{F}$ is bounded (Bartlett and Mendelson, 2002, Lemma 22) as follows:

$$\mathfrak{R}_n(S, \mathcal{F}) \leq \sqrt{\frac{\mathbb{E}_S[\|x\|_2^2]}{n}} \quad \text{for } S \in \{P, Q\},$$

where the second moment is at most one by the normalization. The optimization oracle has an exact closed form. For $v \triangleq \hat{\mu}_P - \lambda\hat{\mu}_Q$ and its positive part v_+, every $w \geq 0$ with $\|w\|_2 \leq 1$ satisfies $w \cdot v \leq w \cdot v_+$ as $w \geq 0$, and $w \cdot v_+ \leq \|v_+\|_2$ by the Cauchy–Schwarz inequality. The choice $w = v_+/\|v_+\|_2$ attains this value (or $w = 0$ if $v_+ = 0$). Therefore, we have:

$$\hat{h}(\lambda) = \sup_{w \geq 0,\, \|w\|_2 \leq 1} [w \cdot \hat{\mu}_P - \lambda\, w \cdot \hat{\mu}_Q] = \frac{v_+ \cdot v}{\|v_+\|_2} = \frac{\|v_+\|_2^2}{\|v_+\|_2} = \|(\hat{\mu}_P - \lambda\hat{\mu}_Q)_+\|_2,$$

where $\hat{\mu}_P$ and $\hat{\mu}_Q$ are the empirical means of the respective samples. Thus, $\varepsilon_{\text{opt}} = 0$ up to numerical precision. Analogously, the IPM over this family is $D_{\mathcal{F}}(P, Q) = \max\{\|(\mu_P - \mu_Q)_+\|_2, \|(\mu_Q - \mu_P)_+\|_2\}$ for the population means $\mu_P \triangleq \mathbb{E}_P[x]$ and $\mu_Q \triangleq \mathbb{E}_Q[x]$. Therefore, the certified envelope $\hat{g}_{\text{cert}}$ in Proposition B.9 can be computed directly from the output of Algorithm 2, with the statistical terms ε_P and ε_Q as the only corrections.

Results. We estimate the IPM directly (≈ 0.0786), and estimate the IPB using Algorithm 2 on a uniform grid of $100 + 1$ points over $[0, \lambda_{\max}]$, including $\lambda = 1$, for $\lambda_{\max} = 4$ found using the doubling search heuristic in Algorithm 3. Fig. 6 shows the estimated IPB together with the certified envelope at the 95% confidence level ($\delta = 0.05$) and the IPM-based bound $\alpha + D_{\mathcal{F}}(P, Q)$. We compute the exact empirical boundary $\hat{g}(\alpha) = \inf_{\lambda \geq 0}[\hat{h}(\lambda) + \lambda\alpha]$ by minimizing this convex one-dimensional objective to numerical precision.

The output of Algorithm 2 is visually indistinguishable from the exact empirical boundary. Even the pessimistic bound we obtain from the certified IPB at the 95% confidence level improves over the estimated IPM at both ends of the α range, for $\alpha \leq 0.15$ and for $\alpha \geq 0.56$. At either end, we reduce the additive slack over α from 0.0786 to 0.039, an improvement of 0.040. The largest active dual variable we observe is $\hat{\lambda}_\alpha \approx 1.29$, attained at $\alpha = 0$, so the statistical penalty $\varepsilon_P + \hat{\lambda}_\alpha\varepsilon_Q$ that is part of $\hat{g}_{\text{cert}}$ is substantially smaller than the uniform correction $\varepsilon_P + \lambda_{\max}\varepsilon_Q$ in Theorem 4.3.

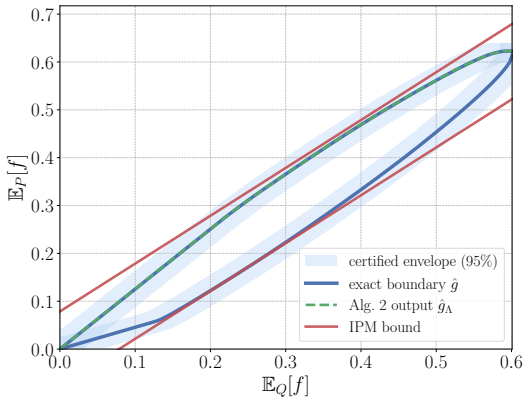

Figure 6: **Certified IPB for a family based on bounded linear functions in the setup of Section 5.3.** We show the exact empirical boundary, the output of Algorithm 2, the certified envelope from Proposition B.9 at the 95% confidence level, and the IPM-based bound. The shaded regions are the certified envelopes at the 95% confidence level. We show the entire α range $[0, \|\hat{\mu}_Q\|_2]$. The certified upper bound improves over the estimated IPM bound at both ends of the range, for $\alpha \leq 0.15$ and for $\alpha \geq 0.56$.

C.4 Software

Experiments are implemented in Python 3.11. We use the following key software:

- `vllm` (Kwon, 2025) and `transformers` (Wolf et al., 2020), with `accelerate` and `datasets`, for language-model sampling, serving, and data loading.
- `anthropic` Python SDK for judge calls.
- `numpy` (Harris et al., 2020) and `scipy` (Virtanen et al., 2020) for numerical computation and array handling.
- `pandas` (pandas development team, 2020) for tabular data handling.
- `sklearn` (Pedregosa et al., 2011) for fitting decision trees.
- `sdv` (Patki et al., 2016) for synthetic data generation with CTGAN.

D Additional Figures and Tables

Table 2: Summary of applications of IPBs described in this paper.

Area	Family $\mathcal{F}$	Baseline	Reference
Theoretical			
Capability deterioration in *alignment* (BoN)	$\mathcal{F}_1$	KL + Pinsker/BH	Section 5.1
Generalization under f-DP	$\mathcal{F}_1$	—	Section B.3
Attack risk under f-DP	$\mathcal{F}_1$	—	Section B.3
Empirical (Algorithm 1 / Algorithm 2)			
Judge-free model *performance estimation*	$\mathcal{F}_1$	KL + Pinsker/BH	Section 5.2
Multi-group *proportional representation*	depth-k trees	Empirical IPM (Oesterling et al., 2024)	Section 5.3
Certified *proportional representation*	bounded linear functions	Empirical IPM (Oesterling et al., 2024)	Section C.3

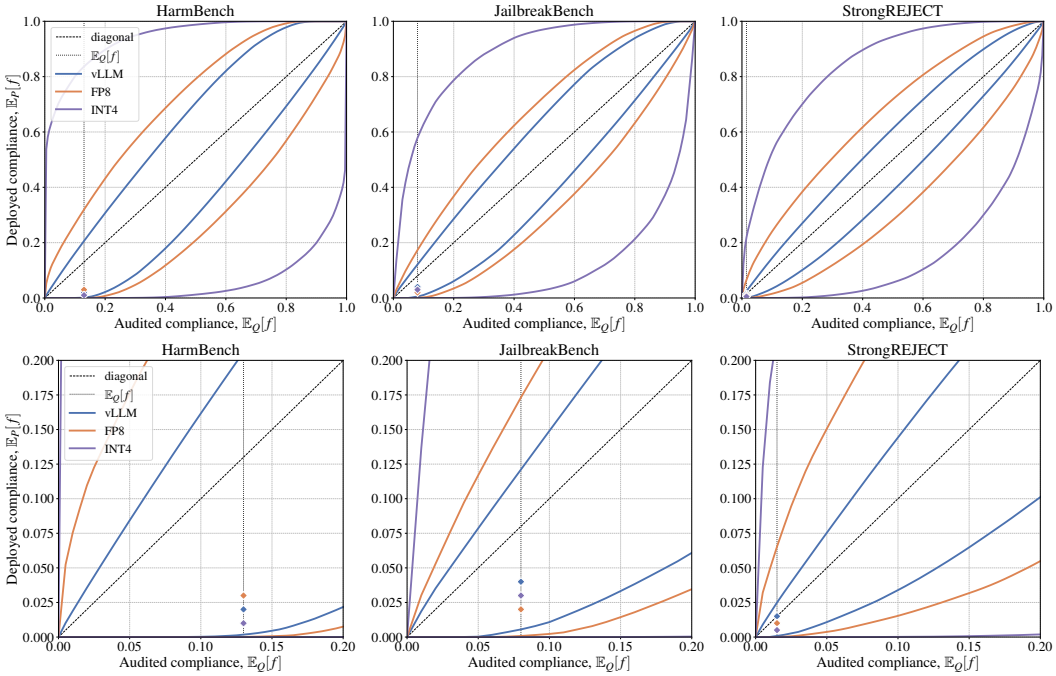

Figure 7: **Estimated IPB curves for the judge-free performance estimation experiment in Section 5.2.** For each benchmark and deployment backend P, we show the IPB upper bound $g \in \mathcal{B}_{\mathcal{F}_1}(P \parallel Q)$ and the lower bound $\check{g}^{-1}$ for $\check{g} \in \mathcal{B}_{\mathcal{F}_1}(Q \parallel P)$, computed with Algorithm 1 from the samples of the original model Q. The dotted vertical line marks the audited compliance rate $\mathbb{E}_Q[f]$ of the original model, and the diamonds mark the observed compliance rate $\mathbb{E}_P[f]$ under each backend. The intervals in Fig. 4 are the vertical slices of these curves at $\mathbb{E}_Q[f]$. As in Fig. 1, the curves are tightest near the diagonal for small deviations (vLLM) and widen for quantized backends. *Top:* Over the entire $[0, 1]$. *Bottom:* Zoom into the $[0, 0.2]$ region containing the observed compliance rates.

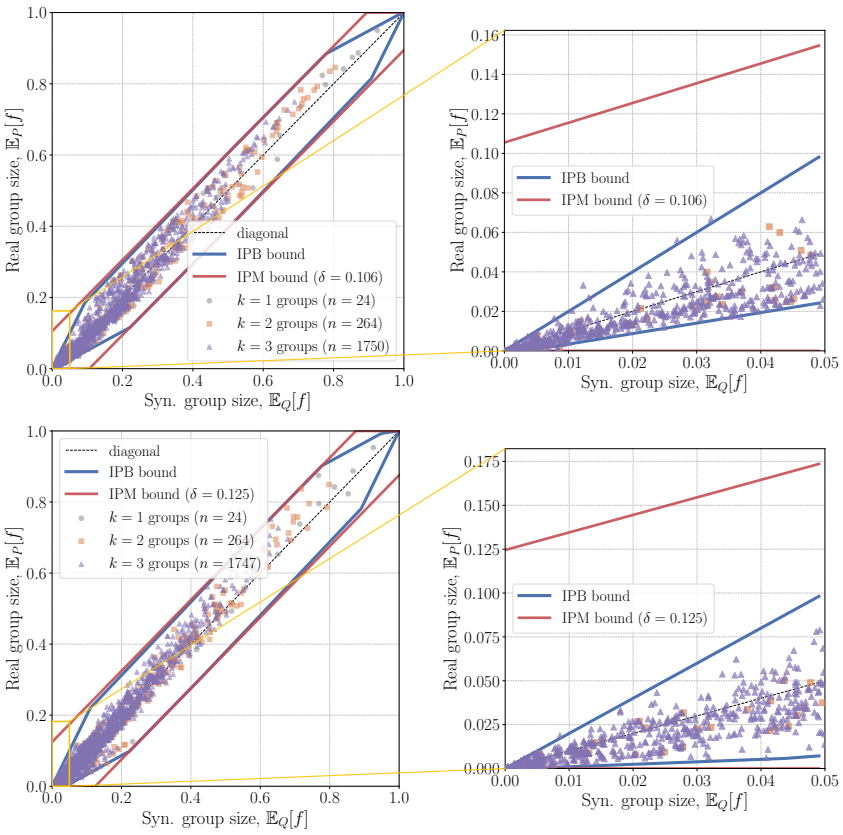

Figure 8: **IPB bounds for CTGAN synthetic data trained on two other training–test splits.** Same setup as Fig. 5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract summarizes the object we introduce and the outcomes of our demonstrations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and the potential future work in Section 7

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, we provide the proofs of the core statements either in the main body or in Appendix A. Certain simple statements that, e.g., follow immediately from a definition, such as Proposition 1.2, we state without an explicit proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the experimental setup in Section 5, and include technical details necessary for reproduction in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not provide open access data and code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide the details in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report replications of the experiment in Section 5.3 in Appendix C. We do not report statistical significance for other measurements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Both our research process and the outcomes conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We briefly discuss broader impacts in Section 7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the appropriate academic publications describing the datasets, models, and core pieces of software in Section 5 and Appendix C. We do not explicitly mention the licenses and terms of use, as we use standard publicly available benchmarks and software in accordance with their terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.