PAC-Bayesian Bounds on Constrained f-Entropic Risk Measures

Hind Atbir[1,◇]
hind.atbir@univ-st-etienne.fr

Farah Cherfaoui[1,◇]
farah.cherfaoui@univ-st-etienne.fr

Guillaume Metzler[2]
guillaume.metzler@univ-lyon2.fr

Emilie Morvant[1]
emilie.morvant@univ-st-etienne.fr

Paul Viallard[3]
paul.viallard@inria.fr

[1] Université Jean Monnet, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, Inria◇, F-42023, Saint-Etienne, France
[2] Université Lumière Lyon 2, Universite Claude Bernard Lyon 1, ERIC, 69007, Lyon, France
[3] Univ Rennes, Inria, CNRS IRISA - UMR 6074, F35000 Rennes, France

Abstract

PAC generalization bounds on the risk, when expressed in terms of the expected loss, are often insufficient to capture imbalances between subgroups in the data. To tackle this limitation, we introduce a new family of risk measures, called *constrained f-entropic risk measures*, which enable finer control over distributional shifts and subgroup imbalances via f-divergences, and include the Conditional Value at Risk (CVaR), a well-known risk measure. We derive both classical and disintegrated PAC-Bayesian generalization bounds for this family of risks, providing the first *disintegrated* PAC-Bayesian guarantees beyond standard risks. Building on this theory, we design a self-bounding algorithm minimizing our bounds directly, yielding models with guarantees at the subgroup level. We empirically demonstrate the usefulness of our approach.

1 INTRODUCTION

A machine learning task is modeled by a fixed but unknown joint probability distribution over $\mathcal{X}\times\mathcal{Y}$ denoted by D, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. Given a family of hypotheses $\mathcal{H}$, consisting of predictors $h:\mathcal{X}\to\mathcal{Y}$, the learner aims to find the hypothesis $h\in\mathcal{H}$ that best captures the relationship between the input space $\mathcal{X}$ and the output space $\mathcal{Y}$. In other words, the learned hypothesis h must correspond to the one that minimizes the true risk defined by

$$L(h) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D} \ell(y, h(\mathbf{x})),$$

with $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$ a (measurable) loss function to assess the quality of h. Since D is unknown, the true risk cannot be computed, so we need tools to estimate it and to assess the quality of the selected hypothesis $h\in\mathcal{H}$. To do so, a learning algorithm relies on a learning set $\mathcal{S}$ composed of examples drawn *i.i.d.* from D, and minimizes the empirical risk defined by

$$\hat{L}(h) := \hat{L}_{\mathcal{S}}(h) := \tfrac{1}{|\mathcal{S}|}\textstyle\sum_{(x,y)\in\mathcal{S}} \ell(y, h(\mathbf{x})).$$

Thus, a central question in statistical learning theory is how well the empirical risk $\hat{L}(h)$ approximates the true risk $L(h)$. This is commonly captured by the generalization gap defined as a deviation between $L(h)$ and $\hat{L}(h)$, which can be upper-bounded with a Probably Approximately Correct (PAC) generalization bound (Valiant, 1984). Several theoretical frameworks have been developed to provide such bounds, notably uniform-convergence-based bounds (Bartlett and Mendelson, 2002; Vapnik and Chervonenkis, 1971). In this paper, we focus on the PAC-Bayesian framework (Shawe-Taylor and Williamson, 1997; McAllester, 1998), which is able to provide tight and often easily computable generalization bounds. As a consequence, a key feature of PAC-Bayesian bounds is that they can be optimized during the learning process, giving rise to self-bounding algorithms (Freund, 1998)[1]. Such algorithms not only return a model but also provide its own generalization guarantee: The bound is optimized.

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

[1] Self-bounding algorithms have recently regained interest in PAC-Bayes (see, *e.g.*, Rivasplata (2022); Viallard (2023)).

However, when the distribution D exhibits imbalances, for example, when subgroups of the population may be under (or over) represented, the classical generalization gap generally fails to capture these imbalances. This issue arises in many practical scenarios, including class imbalance. In fact, when the learning set $\mathcal{S}$ is sampled *i.i.d.* from D, the imbalances are likely to be replicated, resulting in learning a hypothesis with a high error rate for underrepresented subgroups or classes. A way to address such under-representation is to partition the data into subgroups and compute a re-weighted risk across the subgroups. We formalize this scenario as follows. Let $\mathcal{A}$ be an arbitrary partition of the data space $\mathcal{X}\times\mathcal{Y}$, then $D_{|\mathrm{A}}$ is the conditional distribution on a subset $\mathrm{A}\in\mathcal{A}$, and the associated true risk on A is

$$L_{\mathrm{A}}(h) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D_{|\mathrm{A}}} \ell(y, h(\mathbf{x})).$$

Here, we assume that the learning set is partitioned[2] as $\mathcal{S}=\{\mathcal{S}_{\mathrm{A}}\}_{\mathrm{A}\in\mathcal{A}}$. The empirical risk of a subgroup A is evaluated on $\mathcal{S}_{\mathrm{A}}$ of size m_{A} with

$$\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) := \frac{1}{m_{\mathrm{A}}} \sum_{(\mathbf{x},y)\in\mathcal{S}_{\mathrm{A}}} \ell(y, h(\mathbf{x})),$$

More precisely, we consider the following risk measure enabling the re-weighting of the subgroups' risks[3]:

$$\mathcal{R}(h) := \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h), \quad \text{with} \quad E \subseteq \mathcal{M}(\mathcal{A}), \qquad (1)$$

where $\mathcal{M}(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$. Here, ρ is a probability distribution over the subgroups, controlling the weight of each subgroup loss $L_{\mathrm{A}}(h)$, and E denotes a set of admissible distributions.

In this paper, we go beyond previous PAC-Bayesian generalization bounds by considering a new class of risk measures, which we call *constrained f-entropic risk measures*, that go beyond the traditional vanilla true/empirical risks. The key idea is to constrain the set E in Equation (1) to better control the subgroup imbalances while taking into account the distribution shifts thanks to a f-divergence. Our definition extends the Conditional Value at Risk (CVaR, see Rockafellar and Uryasev, 2000) while keeping the flexibility of f-entropic risk measures (Ahmadi-Javid, 2012). Then, we propose disintegrated (and classical) PAC-Bayesian generalization bounds for constrained f-entropic risk measures in two regimes: *(i)* when the set of subgroups can be smaller than the learning set, and *(ii)* when there is only one example per subgroup. Then, we design a self-bounding algorithm that minimizes our disintegrated PAC-Bayesian bound associated with each regime. Finally, we illustrate the effectiveness of our bounds and self-bounding algorithm in both regimes.

Organization of the paper. Section 2 introduces notations, recalls on PAC-Bayes, f-entropic risk measures, and related works. Section 3 defines our constrained f-entropic risk measures, and Section 4 derives our new PAC-Bayesian bounds. Section 5 presents the associated self-bounding algorithm, evaluated in Section 6.

2 PRELIMINARIES

2.1 Additional Notations[4]

We consider learning tasks modeled by an unknown distribution D on $\mathcal{X}\times\mathcal{Y}$. A learning algorithm is provided with a learning set $\mathcal{S}=\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ consisting of m examples $(\mathbf{x}_i, y_i)$ drawn *i.i.d.* from D; we denote by D^m the distribution of such a m-sample. We assume n subgroups, defining a partition $\mathcal{A}=\{\mathrm{A}_1, \ldots, \mathrm{A}_n\}$ of the data in D. To simplify the reading, A denotes the index of the subgroup in $\mathcal{A}$. Then, we assume that the learning set can be partitioned into subgroups $\mathcal{S}=\{\mathcal{S}_{\mathrm{A}}\}_{\mathrm{A}\in\mathcal{A}}$, such that $\forall \mathrm{A}\in\mathcal{A}$, we have $\mathcal{S}_{\mathrm{A}}=\{(\mathbf{x}_j, y_j)\}_{j=1}^{m_{\mathrm{A}}}$, and the size of $\mathcal{S}_{\mathrm{A}}$ is $m_{\mathrm{A}}\in\mathbb{N}^*$. Therefore, the learner's objective is to minimize the true risks $L_{\mathrm{A}}(h)$ of each subgroup aggregated with the risk $\mathcal{R}(h)$ as defined in Equation (1). The set E will be further specialized in Section 3.

2.2 PAC-Bayes in a Nutshell

We specifically work in the setting of the PAC-Bayesian theory. We assume a *prior* distribution P over the hypothesis space $\mathcal{H}$, which encodes an *a priori* belief about the hypotheses in $\mathcal{H}$ before observing any data. Then, given P and a learning set $\mathcal{S}\sim D^m$, the learner constructs a *posterior* distribution $Q_{\mathcal{S}}\in\mathcal{M}(\mathcal{H})$. We assume that $Q_{\mathcal{S}} \ll P$, *i.e.*, the posterior $Q_{\mathcal{S}}$ is absolutely continuous *w.r.t.* the prior P. In practice, this condition ensures that the corresponding densities have the same support. Depending on the interpretation, $Q_{\mathcal{S}}$ can be used in the two following ways.

In **classical PAC-Bayes**, $Q_{\mathcal{S}}$ defines a randomized predictor[5], which samples $h\sim Q_{\mathcal{S}}$ for each input $\mathbf{x}$, and then outputs $h(\mathbf{x})$. The generalization gap is then the deviation between the expected true risk $\mathbb{E}_{h\sim Q_{\mathcal{S}}} L(h)$ and the expected empirical risk $\mathbb{E}_{h\sim Q_{\mathcal{S}}} \hat{L}(h)$.

In **disintegrated (or derandomized) PAC-Bayes**, $Q_{\mathcal{S}}=\Phi(\mathcal{S}, P)$ is learned by a deterministic algorithm[6] $\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$. Then, a single deterministic hypothesis h drawn from $Q_{\mathcal{S}}$ is considered.

[2] We assume every subgroup in $\mathcal{A}$ is represented in $\mathcal{S}$.

[3] Note that Equation (1) is a distributionally robust optimization problem (Scarf, 1957; Delage and Ye, 2010).

[4] A summary table of notations is given in Appendix A.

[5] The randomized predictor is called the Gibbs classifier.

[6] More formally, $Q_{\mathcal{S}}$ can be seen as a Markov kernel.

This implies that the generalization gap measures the deviation between $L(h)$ and $\hat{L}(h)$ for this hypothesis h.

Historically, PAC-Bayes has focused on the randomized risk (Shawe-Taylor and Williamson, 1997; McAllester, 1998). A seminal result is the bound of McAllester (2003), improved by Maurer (2004), stating that with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, we have

$$\forall Q \in \mathcal{M}(\mathcal{H}),$$
$$\mathop{\mathbb{E}}_{h\sim Q} L(h) - \mathop{\mathbb{E}}_{h\sim Q} \hat{L}(h) \le \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}, \quad (2)$$

where $\mathrm{KL}(Q\|P) := \mathbb{E}_{h\sim Q} \ln\big(\frac{dQ}{dP}(h)\big)$, and $\frac{dQ}{dP}$ the Radon-Nikodym derivative. If $Q \ll P$, then $\mathrm{KL}(Q\|P)$ is the KL-divergence; otherwise $\mathrm{KL}(Q\|P) = +\infty$. While the randomized risk may be meaningful (*e.g.*, when studying randomized predictors (Dziugaite and Roy, 2017) or majority votes (Germain et al., 2009)), in practice, we often deploy a single deterministic model. To tackle this, disintegrated PAC-Bayes (Blanchard and Fleuret, 2007; Catoni, 2007; Viallard et al., 2024b,a) has been proposed, where generalization bounds apply directly to a single hypothesis $h\sim Q_{\mathcal{S}}$, after learning $Q_{\mathcal{S}}$. For instance, Rivasplata et al. (2020) derived bounds of the following form. With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ **and** $h\sim Q_{\mathcal{S}}$, we have

$$L(h) - \hat{L}(h) \le \sqrt{\frac{1}{2m}\left[\ln^{+}\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\frac{2\sqrt{m}}{\delta}\right]}, \quad (3)$$

where $Q_{\mathcal{S}} = \Phi(\mathcal{S}, P)$, and $\ln^{+}(\cdot) = \ln(\max(0,\cdot))$, and $\ln^{+}\big(\frac{dQ_{\mathcal{S}}}{dP}(h)\big)$ is the "disintegrated" KL-divergence. Such results are crucial when we seek guarantees for a single deployed model h.

In our work, we are not interested in upper-bounding the classical gap between $L(h)$ and $\hat{L}(h)$. We want to study the gap between the risk measures:

$$\mathcal{R}(h) = \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h) \quad\text{and}\quad \widehat{\mathcal{R}}_{\mathcal{S}}(h) = \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),$$
$$\text{with } E \subseteq E_\alpha = \left\{\rho \,\middle|\, \rho \ll \pi,\ \text{and}\ \frac{d\rho}{d\pi} \le \frac{1}{\alpha}\right\}, \quad (4)$$

with $\alpha\in(0,1]$, and π a reference[7] distribution on the subgroups $\mathrm{A}\in\mathcal{A}$. Intuitively, α constrains how much ρ can deviate from π. We derive in Section 4, classical and disintegrated PAC-Bayesian bounds; thus, we are interested in the true randomized risk measures

$$\mathop{\mathbb{E}}_{h\sim Q} \mathcal{R}(h) := \mathop{\mathbb{E}}_{h\sim Q} \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h), \quad (5)$$
$$\text{or}\qquad \mathcal{R}(Q) := \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h). \quad (6)$$

[7]To avoid any confusion with PAC-Bayes posterior/prior distributions, we call "reference distribution" the distribution π of the (constrained) f-entropic risk measures.

By Jensen's inequality, we have $\mathcal{R}(Q) \le \mathbb{E}_{h\sim Q}\mathcal{R}(h)$. Furthermore, the associated empirical counterparts are

$$\mathop{\mathbb{E}}_{h\sim Q} \widehat{\mathcal{R}}_{\mathcal{S}}(h) := \mathop{\mathbb{E}}_{h\sim Q} \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \quad (7)$$
$$\text{or}\qquad \widehat{\mathcal{R}}_{\mathcal{S}}(Q) := \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \mathop{\mathbb{E}}_{h\sim Q} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h). \quad (8)$$

2.3 f-Entropic Risk Measures in a Nutshell

In Equations (4) to (8), we have to define the right set E. For example, we can use f-divergences (Csiszár, 1963, 1967; Morimoto, 1963; Ali and Silvey, 1966) as follows.

Assumption 1. *Let f be a convex function with $f(1)=0$ and $f(0)=\lim_{t\to 0^+} f(t)$ such that $D_f(\rho\|\pi) := \mathbb{E}_{\mathrm{A}\sim\pi}\big[f\big(\frac{d\rho}{d\pi}(\mathrm{A})\big)\big]$ is a f-divergence. Let $\beta\ge 0$. We have*

$$E := E_{f,\beta} := \left\{\rho \,\middle|\, \rho \ll \pi,\ \text{and} \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right) \le \beta\right\},$$

where $\ll$ denotes absolute continuity and π is a reference distribution over $\mathcal{A}$.

Definition 1. *(Ahmadi-Javid, 2012) We say that $\mathcal{R}$ of Equation (1) is a f-entropic risk measure if E satisfies Assumption 1.*

The Conditional Value at Risk (CVaR, Rockafellar and Uryasev (2000)) is an example of a f-entropic risk measure. Let $\alpha\in(0,1]$ and $g_\alpha(x) := \iota\big[x\in[0,\frac{1}{\alpha}]\big]$ with $\iota[a]=0$ if a is true and $+\infty$ otherwise, CVaR is defined for

$$\begin{aligned} E = E_{g_\alpha,0} &:= \left\{\rho \,\middle|\, \rho \ll \pi,\ \text{and} \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} g_\alpha\left(\frac{d\rho}{d\pi}(\mathrm{A})\right) \le 0\right\} \\ &= \left\{\rho \,\middle|\, \rho \ll \pi,\ \text{and}\ D_{g_\alpha}(\rho\|\pi) \le 0\right\} \\ &= \left\{\rho \,\middle|\, \rho \ll \pi,\ \text{and}\ \frac{d\rho}{d\pi} \le \frac{1}{\alpha}\right\} = E_\alpha. \end{aligned} \quad (9)$$

Note that CVaR also belongs to another family of measures known as *Optimized Certainty Equivalents* (OCE, Ben-Tal and Teboulle, 1986, 2007).[8]

2.4 Related Work

There exist some generalization bounds related to ours. Unlike our setting, which allows partitioning $\mathcal{S}$ into n subgroups $\mathcal{A}$, these existing bounds hold for $|\mathcal{A}|=n=m=|\mathcal{S}|$, *i.e.*, there is only one example per subgroup.

Apart from PAC-Bayes bounds, generalization bounds that focus on the worst-case generalization gap,

$$\sup_{h\in\mathcal{H}} \big|\mathcal{R}(h) - \widehat{\mathcal{R}}_{\mathcal{S}}(h)\big|,$$

[8]The link between f-entropic risk measures and OCEs is detailed in Appendix B.

have been introduced. For example, Curi et al. (2020) derived an upper bound on the CVaR, relying on Brown (2007)'s concentration inequality. Their bound holds either for finite hypothesis sets or for infinite hypothesis sets, but with a bound depending on covering numbers or Pollard (1984)'s pseudo-dimension. Another example is Lee et al. (2020)'s generalization bound for OCEs, which relies on the Rademacher complexity associated with $\mathcal{H}$ (see, *e.g.*, Bartlett and Mendelson, 2002). In these examples, the bounds are not easy to manipulate in practice.

The bound that is most closely related to our bounds in Section 4 is the classical PAC-Bayes bound of Mhammedi et al. (2020) on the CVaR (recalled in Theorem 1). Their bound holds only with one example per subgroup with a uniform *reference* distribution π.

Theorem 1 (PAC-Bayesian Bound on CVaR (Mhammedi et al., 2020))**.** *For any distribution D over $\mathcal{X}\times\mathcal{Y}$, for any prior $P\in\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any $\alpha\in(0,1]$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, we have for all $Q\in\mathcal{M}(\mathcal{H})$,*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)\le\widehat{\mathcal{R}}_{\mathcal{S}}(Q)+\\
&2\,\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\left[\sqrt{\frac{1}{2\alpha m}\ln\frac{2\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}}+\frac{1}{3m\alpha}\ln\frac{2\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}\right]\\
&\qquad+\sqrt{\frac{27}{5\alpha m}\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\left[\mathrm{KL}(Q\|P)+\ln\frac{2\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}\right]}\\
&\qquad\qquad+\frac{27}{5\alpha m}\left[\mathrm{KL}(Q\|P)+\ln\frac{2\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}\right],
\end{aligned}$$

$$\begin{aligned}
&\textit{where}\quad \mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h):=\mathop{\mathbb{E}}_{h\sim Q}\sup_{\rho\in E}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\rho}\ell(y,h(\mathbf{x}))\\
&\textit{with}\quad E=\left\{\rho\;\middle|\;\rho\ll D,\ \textit{and}\ \frac{d\rho}{dD}\le\frac{1}{\alpha}\right\},\\
&\textit{and}\quad \widehat{\mathcal{R}}_{\mathcal{S}}(Q):=\sup_{\rho\in\widehat{E}}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho}\mathop{\mathbb{E}}_{h\sim Q}\ell(y_{\mathrm{A}},h(\mathbf{x}_{\mathrm{A}})),\\
&\textit{with}\quad \widehat{E}=\left\{\rho\;\middle|\;\rho\ll\pi,\ \textit{and}\ \frac{d\rho}{d\pi}\le\frac{1}{\alpha}\right\},\\
&\textit{where}\quad \pi(\mathrm{A})=\frac{1}{m}.
\end{aligned}$$

Theorem 1 upper-bounds the expected true CVaR by its empirical counterpart and terms that depend on the KL-divergence between *posterior* and *prior* over $\mathcal{H}$. Note that contrary to our bounds, Theorem 1 does not hold for other measures. This is due to the proof that involves concentration inequalities tailored for CVaR, making extensions to other measures hard to obtain.

Another related framework is *Group Distributionally Robust Optimization* (Group DRO, Sagawa et al., 2020), which considers a risk measure defined as the maximum of the expected loss on each subgroup. Sagawa et al. (2020) proposes a learning procedure based on the principle of structural risk minimization (Vapnik, 1991), where the regularization term estimates the generalization gap and mainly depends on a tunable hyperparameter. Our work differs in two ways: *(i)* we aggregate subgroup expected loss using the worst-case distribution in a ball defined by a f-entropic divergence and constrained by α, *(ii)* our learning procedure directly minimizes the generalization gap by optimizing PAC-Bayes bounds, yielding models with built-in generalization guarantees.

3 CONSTRAINED f-ENTROPIC RISK MEASURES

In this paper, we extend the definition of the CVaR to obtain more general PAC-Bayesian generalization bounds (in Section 4) for a larger class of risk measures, which we call *constrained f-entropic risk measures*. We construct our new class as a restricted subclass of f-entropic risk measures by preserving their flexibility (Assumption 1) while considering an additional constraint that controls how much the distribution ρ can deviate from a given reference π (Equation (9)). To do so, we assume the following restricted set E.

Assumption 2. *Let f be defined such that $D_f(\rho\|\pi)$ is a f-divergence. Let $\beta\ge 0$ and $\alpha>0$. We have*

$$E=\left\{\rho\;\middle|\;\rho\ll\pi\ \textit{and}\ \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)\le\beta\ \textit{and}\ \forall\mathrm{A}\in\mathcal{A},\ \frac{d\rho}{d\pi}(\mathrm{A})\le\frac{1}{\alpha}\right\},$$

with π a reference distribution over $\mathcal{A}$.

Put into words, E contains all distributions ρ that: *(i)* are absolutely continuous *w.r.t.* π; *(ii)* have a f-divergence with π bounded by β; *(iii)* satisfy a uniform upper bound on the density ratio $\frac{d\rho}{d\pi}(\mathrm{A})\le\frac{1}{\alpha}$. We now define the constrained f-entropic risk measures.

Definition 2. *We say that $\mathcal{R}$ or $\widehat{\mathcal{R}}_{\mathcal{S}}$ is a* constrained f-entropic risk measure *if E satisfies Assumption 2.*

A key observation is that a constrained f-entropic risk measure corresponds to a standard f-entropic risk measure with an augmented function $f+g_\alpha$ (with g_α as defined for Equation (9)). Indeed, E can be rewritten as

$$\begin{aligned}
E&=\left\{\rho\,\middle|\,\rho\ll\pi\text{ and }\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)\le\beta\ \text{and}\ \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}g_\alpha\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)\le 0\right\}\\
&=\left\{\rho\,\middle|\,\rho\ll\pi\text{ and }\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\left[f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)+g_\alpha\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)\right]\le\beta\right\}\\
&=\left\{\rho\,\middle|\,\rho\ll\pi\text{ and }D_{f+g_\alpha}(\rho\|\pi)\le\beta\right\}=E_{f+g_\alpha,\beta}\subseteq E_\alpha,
\end{aligned}$$

where $f+g_\alpha$ generates the divergence $D_{f+g_\alpha}(\rho\|\pi)$, since it is convex, and we have $f(1)+g_\alpha(1)=0$, and $\lim_{t\to 0^+} f(t)+g_\alpha(t) = f(0)+g_\alpha(0)$. Thanks to Definition 2, when $\beta\to+\infty$, the measure ρ becomes less constrained by $D_f(\rho\|\pi)$, implying that $\mathcal{R}(h)$ becomes the true CVaR. Moreover, when $\alpha\to 0$, the condition $\frac{d\rho}{d\pi}(\mathrm{A}) \le \frac{1}{\alpha}$ does not restrict the set E. In this case, $\mathcal{R}$ of Definition 2 becomes a f-entropic risk measure.

4 PAC-BAYESIAN BOUNDS ON CONSTRAINED f-ENTROPIC RISK MEASURES

We present our main contribution, *i.e.*, classical and disintegrated PAC-Bayesian bounds for constrained f-entropic risk measures, by distinguishing two regimes. In Section 4.1, we focus on the case where the number of subgroups is smaller than the learning set size, *i.e.*, $|\mathcal{A}| \le m$. For completeness, since the bound of Section 4.1 becomes vacuous when $|\mathcal{A}|=m$, we consider, in Section 4.2, the case where each subgroup contains only one example (more specifically, one loss), *i.e.*, $|\mathcal{A}|=m$.

4.1 When $|\mathcal{A}| \le m$

In Theorem 2, we present both classical and disintegrated *general* PAC-Bayesian bounds. As commonly done in PAC-Bayes (*e.g.*, Germain et al., 2009), these general results are flexible since they depend on a convex deviation function φ between true and empirical risks. Different choices of φ result in different instantiations of the bound, allowing us to capture the deviation in different ways. Our theorem below upper-bounds the deviations $\varphi\big(\widehat{\mathcal{R}}_{\mathcal{S}}(Q), \mathbb{E}_{h\sim Q}\,\mathcal{R}(h)\big)$ and $\varphi\big(\widehat{\mathcal{R}}_{\mathcal{S}}(h), \mathcal{R}(h)\big)$ for the classical and disintegrated settings, respectively.

Theorem 2. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any positive, jointly convex function $\varphi(a,b)$ that is non-increasing in a for any fixed b, for any finite set $\mathcal{A}$ of n subgroups, for any $\lambda_{\mathrm{A}} > 0$ for each $\mathrm{A}\in\mathcal{A}$, for any distribution π on $\mathcal{A}$, for any distribution $P\in\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any constrained f-entropic risk measure $\mathcal{R}$ satisfying Definition 2, for any $\delta\in(0,1]$, for any $\alpha\in(0,1]$, we have the following bounds.*

Classical PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, for all distributions $Q\in\mathcal{M}(\mathcal{H})$, we have*

$$\varphi\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(Q), \mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)\Big) \le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha\,\lambda_{\mathrm{A}}}\Bigg[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\Bigg]. \quad (10)$$

Disintegrated PAC-Bayes. *For any algorithm $\Phi : (\mathcal{X}\times\mathcal{Y})^m\times\mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ and $h\sim Q_{\mathcal{S}}$, we have*

$$\varphi\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(h), \mathcal{R}(h)\Big) \le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha\,\lambda_{\mathrm{A}}}\Bigg[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\Bigg], \quad (11)$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S}, P)$.

Proof. (Complete proofs are deferred in Appendix C.) To prove Equation (10), we start by using the proof technique of Germain et al. (2009) to obtain a general PAC-Bayes bound on subgroup risks that holds for any $\mathrm{A}\in\mathcal{A}$: With probability at least $1-\delta$ over the random choice of $\mathcal{S}\sim D^m$, we have $\forall Q\in\mathcal{M}(\mathcal{H})$,

$$\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \le \frac{1}{\lambda_{\mathrm{A}}}\Bigg[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\Bigg].$$

Then, we apply a union bound to combine these subgroup bounds with an expectation over $\mathcal{A}$ using the reference π (see Lemma 1). Finally, we use our class of measure of risks' constraint stating that $\forall\mathrm{A}$, $\frac{d\rho}{d\pi}(\mathrm{A}) \le \frac{1}{\alpha}$, and we perform a change of measure to prove that

$$\varphi\Big[\widehat{\mathcal{R}}_{\mathcal{S}}(Q), \mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)\Big] \le \frac{1}{\alpha}\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\varphi\Big[\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\Big],$$

leading to the desired result. The proof of Equation (11) follows the same steps after using Rivasplata et al. (2020)'s proof technique to get the first bound. □

As in Equations (2) and (3), the bounds in Equations (10) and (11) depend respectively on the KL-divergence and its disintegrated version between Q and P. Our bounds additionally involve the parameter λ_{A}, which varies *w.r.t.* the subgroup $\mathrm{A}\in\mathcal{A}$. Interestingly, since the Radon-Nikodym derivative is uniformly bounded by $\frac{1}{\alpha}$, our bounds depend only on the parameter α of the constrained f-entropic risk measure.

To make the result more concrete, we instantiate our disintegrated bound in Corollary 1 with two choices of deviation φ. For completeness, we report in Appendix (Corollary 2) the corresponding classical bounds. First, we use $\varphi(a,b)=\mathrm{kl}^+(a\|b)$ defined, for any $a,b\in[0,1]$, as

$$\mathrm{kl}^+(a\|b) \triangleq \begin{cases} \mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} & \text{if } a\le b,\\ 0 \quad \text{otherwise.}\end{cases}$$

This quantity corresponds to the KL-divergence between two Bernoulli distributions with parameters a and b (truncated to $a\le b$). Second, thanks to Pinsker's inequality, we have $2(a-b)^2 \le \mathrm{kl}^+(a\|b)$ for $a\le b$, which yields another (direct) bound with $\varphi(a,b)=2(a-b)^2$. Hence, we obtain the following corollary.

Corollary 1. *For any D on $\mathcal{X}\times\mathcal{Y}$, for any $\mathcal{A}$ of n subgroups, for any π over $\mathcal{A}$, for any $P\in\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any $\mathcal{R}$ satisfying Definition 2, for any $\delta\in(0,1]$, for any $\alpha\in(0,1]$, for any algorithm $\Phi:(\mathcal{X}\times\mathcal{Y})^m\times\mathcal{M}(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ and $h\sim Q_{\mathcal{S}}$, we have*

$$\mathrm{kl}^+\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(h)\Big\|\mathcal{R}(h)\Big)\leq\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{\ln^+\Big[\frac{dQ_{\mathcal{S}}}{dP}(h)\Big]+\ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{\alpha\, m_{\mathrm{A}}},\quad(12)$$

$$\text{and }\mathcal{R}(h)\leq\widehat{\mathcal{R}}_{\mathcal{S}}(h)+\sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{\ln^+\Big[\frac{dQ_{\mathcal{S}}}{dP}(h)\Big]+\ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{2\,\alpha\, m_{\mathrm{A}}}},\quad(13)$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S},P)$.

Proof. Deferred to Section E.1. □

Put into words, the larger the subgroup size m_{A}, the tighter the bound. Conversely, smaller values of α make the bound looser, due both to the multiplicative factor $\frac{1}{\alpha}$ and to the fact that smaller α values make the constrained f-entropic risk measures more pessimistic.

4.2 When $|\mathcal{A}|=m$

When each subgroup corresponds to a single example of $\mathcal{S}$, the bounds of Theorem 2 become vacuous (since $\forall\mathrm{A}\in\mathcal{A}$, $m_{\mathrm{A}}=1$). To obtain a non-vacuous bound in this context, we derive bounds that take a different form. Formally, for a learning set $\mathcal{S}=\{(\mathbf{x}_{\mathrm{A}},y_{\mathrm{A}})\}_{\mathrm{A}=1}^m\sim D^m$, we set the reference distribution π to be the uniform distribution over $\mathcal{S}$, we have

$$\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)=\ell(y_{\mathrm{A}},h(\mathbf{x}_{\mathrm{A}})),\quad\text{and}\quad\pi(\mathrm{A})=\frac{1}{m},\qquad(14)$$

and we constrain the distribution ρ with α, *i.e.*, for each $(\mathbf{x}_{\mathrm{A}},y_{\mathrm{A}})$. We obtain the following PAC-Bayes bounds.

Theorem 3. *For any D on $\mathcal{X}\times\mathcal{Y}$, for any $\lambda>0$, for any $P\in\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any constrained f-entropic risk measure $\widehat{\mathcal{R}}_{\mathcal{S}}$ satisfying Definition 2* and *Equation (14), for any algorithm $\Phi:(\mathcal{X}\times\mathcal{Y})^m\times\mathcal{M}(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, for any $\delta\in(0,1]$, we have the following bounds.*

Classical PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, we have*

$$\Big|\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}}\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\Big|\leq$$
$$\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\mathrm{KL}(Q_{\mathcal{S}}\|P)+\ln\left[\frac{2(\lambda+1)}{\delta}\right]+3.5\right)},\quad(15)$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S},P)$.

Disintegrated PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ and $h\sim Q_{\mathcal{S}}$, we have*

$$\Big|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\Big|\leq$$
$$\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\delta}\right]\right)},\quad(16)$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S},P)$.

Proof. (Complete proofs are deferred in Appendix F.) With McDiarmid's inequality, we prove the next concentration inequality for constrained f-entropic measures (see Lemma 5), holding for a fixed hypothesis $h\in\mathcal{H}$, with probability of at most δ on $\mathcal{S}\sim D^m$, we have

$$|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}}(h)|\geq\frac{1}{\alpha}\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

Then, we use Occam's hammer (Blanchard and Fleuret, 2007) to make the disintegrated KL and the sampling $h\sim Q_{\mathcal{S}}$ appear and obtain the disintegrated bound of Equation (16). The bound of Equation (15) then follows by plugging Equation (16) into the argument of Blanchard and Fleuret (2007) that recovers a classical PAC-Bayesian bound from a disintegrated one. □

The proof of Theorem 3 follows Blanchard and Fleuret (2007)'s technique for the classical generalization gap. Unlike Theorem 2, Theorem 3 is not a general PAC-Bayesian theorem (*i.e.*, it does not involve a deviation φ), but it is a *parametrized* PAC-Bayes bound with parameter λ which controls the trade-off between the concentration terms and the KL-divergence, and which is independent of the risk measure and the subgroups. Moreover, the classical PAC-Bayes bound of Equation (15) derives from the disintegrated one, so it holds only for the posterior $Q_{\mathcal{S}}$ learned from $\mathcal{S}$. Finally, we recall that Corollary 1 suffers from subgroup sizes m_{A} when some m_{A} are small, due to the $\frac{1}{m_{\mathrm{A}}}$ term. In contrast, the bounds of Theorem 3 only depend on the global sample size m with a $\frac{1}{m}$ term, as in standard PAC-Bayesian bounds.

Comparison with Theorem 1. We compare the two classical PAC-Bayes bounds, Equation (15) and the one of Mhammedi et al. (2020) (see Theorem 1). Even though the generalization gaps of the two bounds do not involve the same quantities, we can compare the rates. Interestingly, when $\widehat{\mathcal{R}}_{\mathcal{S}}(Q)>0$, which is a reasonable assumption in practice, our bound is asymptotically tighter, with a rate of $\mathcal{O}(\sqrt{1/m})$ compared to their $\mathcal{O}(\sqrt{(\ln\ln m)/m})$. Importantly, our work establishes the first disintegrated PAC-Bayesian bounds that are not the vanilla true/empirical risk $L(h)$ and $\hat{L}(h)$. This yields a key practical advantage: The empirical

CVaR becomes computable. In contrast, Theorem 1 relies on the computation of $\widehat{\mathcal{R}}_{\mathcal{S}}(Q)$, which can only be estimated and for which no standard concentration inequality (*e.g.*, Hoeffding's inequality) provides a non-vacuous bound. Additionally, although our bound can suffer from the $\frac{1}{\alpha^2}$ factor (larger than the $\frac{1}{\alpha}$ factor in Theorem 1), we observe in practice that our disintegrated bound remains at least comparable.

5 SELF-BOUNDING ALGORITHMS

Our bounds are general, as they do not impose any algorithm for learning the posterior. In the following, we have two objectives: *(i)* in this section, designing a self-bounding algorithm (Freund, 1998) to learn a model by directly minimizing our bounds, and *(ii)* in Section 6, showing the usefulness of our bounds on two types of subgroups (one class per group, one example per group). A self-bounding algorithm outputs a model together with its own non-vacuous generalization bound: the one optimized. For practical purposes, we focus on an algorithm for disintegrated bounds, since they apply to a deterministic model. Indeed, we recall that *(i)* classical PAC-Bayes bounds hold for a randomized model over the entire hypothesis space, which incurs additional computational cost, and *(ii)* the measure $\widehat{\mathcal{R}}_{\mathcal{S}}(Q)$ involved in the classical bounds (*e.g.*, Mhammedi et al. (2020)) is not directly computable, unlike $\widehat{\mathcal{R}}_{\mathcal{S}}(h)$ in our disintegrated bounds (we detail the objective functions associated with our bounds in Appendix G.1).

Algorithm 1 below summarizes the bound's minimization procedure[9]. We parametrize the posterior distribution denoted by Q_θ and we update the parameters θ by (a variant of) stochastic gradient descent as follows. For each epoch and mini-batch $U \subset \mathcal{S}$ (Lines 2-3), we draw a model $h_{\tilde{\theta}}$ from the current posterior distribution Q_θ (Line 4). Then, we compute the empirical risk $\widehat{\mathcal{R}}_U(h_{\tilde{\theta}})$ of $h_{\tilde{\theta}}$ on U (Line 5), which is used to compute the bound, denoted $\mathcal{B}$ (Line 6), and we update the parameters θ of the posterior distribution using the gradient $\nabla_\theta \mathcal{B}(\widehat{\mathcal{R}}_U(h_{\tilde{\theta}}), Q_\theta, h_{\tilde{\theta}})$ (Line 7). Finally, we return a model drawn from the learned Q_θ (Line 10).

On the prior distribution P. A key ingredient of PAC-Bayesian methods is the choice of P (which can be set to uniform by default). Here, we adopt a different, but classical, approach (*e.g.*, Ambroladze et al., 2006; Germain et al., 2009; Parrado-Hernández et al., 2012; Pérez-Ortiz et al., 2021; Dziugaite et al., 2021; Viallard et al., 2024b): The prior P is learned from an auxiliary set $\mathcal{S}_P$, disjoint from the learning set $\mathcal{S}$ (often obtained by a 50/50 split). Here, we learn the parameters θ_P of the prior distribution with a variant of Algorithm 1: We remove the bound computation (Line 6), replace the gradient in Line 7 by $\nabla_{\theta_P} \widehat{\mathcal{R}}_U(h_{\tilde{\theta}_P})$, and keep the rest unchanged. Concretely, for each mini-batch $U \subset \mathcal{S}_P$ (Lines 2-3), we sample $h_{\tilde{\theta}_P}$ from $P_\theta = \mathcal{N}_P(\theta_P, \sigma^2 I_d)$, evaluate $\widehat{\mathcal{R}}_U(h_{\tilde{\theta}_P})$ (Line 5), and update θ_P with the gradient $\nabla_{\theta_P} \widehat{\mathcal{R}}_U(h_{\tilde{\theta}_P})$. Instead of returning a model sampled from the final P_θ (Line 10), we output the prior P parametrized by the best-performing θ_P over the epochs and across the hyperparameter grid search.

Algorithm 1 Self-bounding algorithm for constrained f-entropic risk measures

Require: Set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, number of epochs T, variance σ^2, prior $P = \mathcal{N}(\theta_P, \sigma^2 I_d)$ where $\theta_P \in \mathbb{R}^d$, bound $\mathcal{B}$, reference π, parameters α, β

1: Initialize $\theta \leftarrow \theta_P$
2: **for** $t = 1$ **to** T **do**
3: **for all** mini-batches $U \subset \mathcal{S}$ drawn *w.r.t.* π **do**
4: Draw a model $h_{\tilde{\theta}}$ from $Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$
5: Compute the risk $\widehat{\mathcal{R}}_U(h_{\tilde{\theta}})$ on the mini-batch
6: Compute the bound $\mathcal{B}(\widehat{\mathcal{R}}_U(h_{\tilde{\theta}}), Q_\theta, \tilde{\theta})$
7: Update θ with gradient $\nabla_\theta \mathcal{B}(\widehat{\mathcal{R}}_U(h_{\tilde{\theta}}), Q_\theta, \tilde{\theta})$
8: **end for**
9: **end for**
10: Draw a model $h_{\hat{\theta}}$ from Q_θ
11: **return** $h_{\hat{\theta}}$

6 EXPERIMENTS[10]

We now illustrate the potential of our PAC-Bayes bounds for constrained f-entropic risk measures with the CVaR, focusing on imbalances in the classical class-imbalance setting. To do so, we study the behavior of our self-bounding algorithm with our bounds in Equation (13) (Corollary 1, with one group corresponding to a class, *i.e.*, $|\mathcal{A}| = |\mathcal{Y}| \leq m$), and Equation (15) (Theorem 3, with one example per group, *i.e.*, $|\mathcal{A}| = m$), with Mhammedi et al. (2020)'s bound (Theorem 1), and discuss their potential. Before analyzing our results, we present our general experimental setting (more details are given in Appendix G).

Datasets. We report results for the 4 most imbalanced datasets we considered (taken from OpenML, Vanschoren et al., 2013): *Oilspill* (class ratio .96/.04) (Kubat et al., 1998), *Mammography* (.98/.02), *Balance* (.08/.46/.46) (Siegler, 1976), and *Pageblocks* (.90/.06/.01/.02/.02) (Malerba, 1994). Each dataset is split into a training set ($\mathcal{S}'$) and a test set ($\mathcal{T}$) with a 80%/20% ratio. Following our PAC-Bayesian Algorithm 1, we split $\mathcal{S}'$ into two disjoint sets $\mathcal{S}$ and $\mathcal{S}_P$

[9] Algorithm 1 follows a quite standard procedure to minimize a bound, but is specialized to our setting.

[10] The code is available at https://gitlab.com/hatbir/aistats26-pb-cferm.

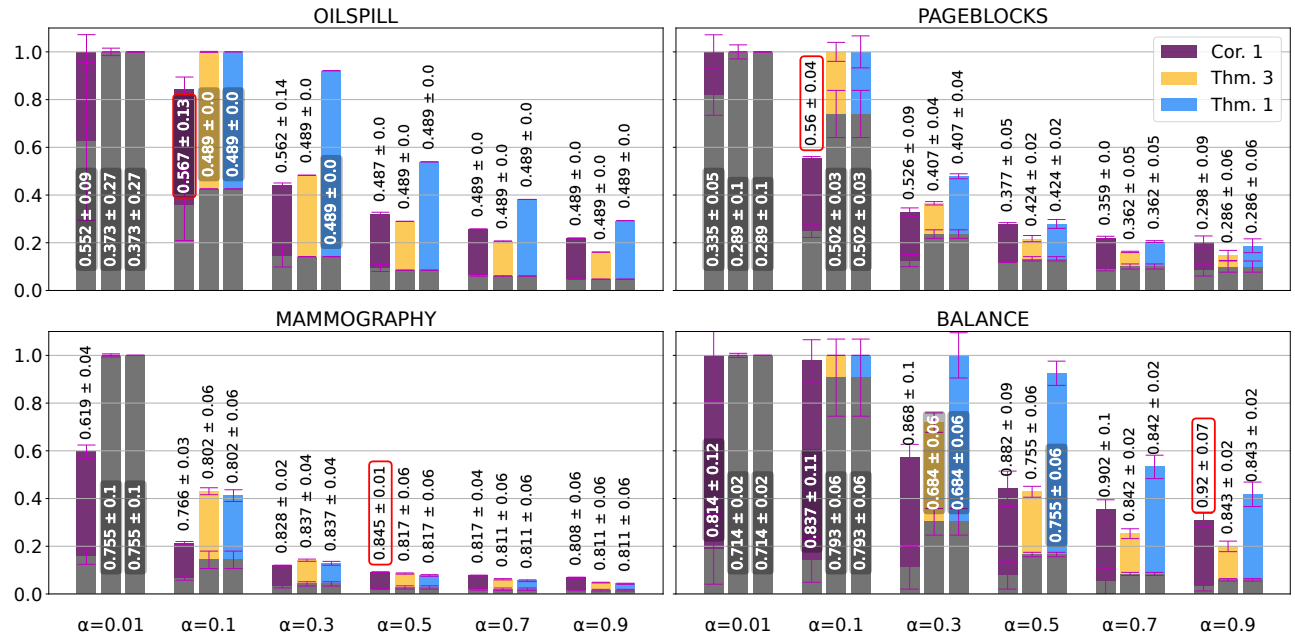


Figure 1: Bound values (in color), test risk $\mathcal{R}_{\mathcal{T}}$ (in grey), and F-score value on $\mathcal{T}$ (with their standard deviations) for Theorem 3, Corollary 1, and Theorem 1, as a function of α (on the x-axis). The y-axis corresponds to the value of the bounds and test risks. The highest F-score for each dataset is emphasized with a red frame.

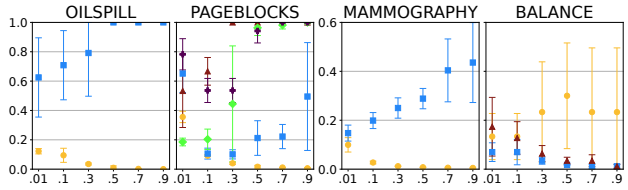


Figure 2: Evolution of the class-wise error rates and standard deviation on the set $\mathcal{T}$ (y-axis) as a function of the parameter α (x-axis) with Corollary 1. Each class is represented by different markers and colors.

with a 50%/50% ratio; $\mathcal{S}$ is used to learn the posterior Q_θ and $\mathcal{S}_P$ to learn the prior P. All the splits preserve the original class ratio. Note that each experiment is repeated 3 times with random splits.

Models & distributions. We consider neural networks with 2 hidden layers of size 128 (a 2-hidden-layer multilayer perceptron), with leaky ReLUs activations. To learn the prior $P = \mathcal{N}(\theta_P, \sigma^2 I_d)$, *i.e.*, θ_P, we initialize the parameters with a Xavier uniform distribution (Glorot and Bengio, 2010), then, to learn the posterior distribution $Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, the parameters are initialized with θ_P (Line 1 of Algorithm 1), and $\sigma^2 = 10^{-6}$.

Risk. We recall that we compare two regimes with the CVaR as the risk measure: *(i)* for Corollary 1 when $\mathcal{A} \leq m$ with $\mathcal{A}$ defined by classes, *i.e.*, for all $y \in \mathcal{Y}$, we have a subgroup $\mathcal{S}_{\mathrm{A}} = \{(\mathbf{x}_j, y)\}_{j=1}^{m_{\mathrm{A}}}$, with the reference π set to the class ratio, and *(ii)* for Theorem 3 and Theorem 1 when $\mathcal{A} = m$ where each subgroup is a single example, *i.e.*, $\mathcal{A} = \mathcal{S} = \{(\mathbf{x}_{\mathrm{A}}, y_{\mathrm{A}})\}_{\mathrm{A}=1}^{m}$ with π set to the uniform distribution. The CVaR is computed with bounded cross-entropy of Dziugaite and Roy (2018) as the loss, with parameter $\ell_{\max} = 4$ (the loss is rescaled to $[0, 1]$ to align with the theoretical assumptions). To solve the maximization problem associated with Equation (8), we use the python library *cvxpylayers* (Agrawal et al., 2019) that creates differentiable convex optimization layers. This layer is built on top of *CVXPY* (Diamond and Boyd, 2016); We use the optimizer SCS (O'Donoghue et al., 2023) under the hood, with $\varepsilon = 10^{-5}$ and a maximum of 100000 iterations. In additional experiments, in Appendix H, we provide results with π as the uniform distribution, and for another constrained f-entropic risk measure (a constrained version of the EVaR Ahmadi-Javid (2012)).

Bound. We compare our disintegrated bounds of Corollary 1 and Theorem 3 with an estimate of Mhammedi et al.'s bound (Theorem 1), obtained by sampling a single model from the posterior Q_θ. We think this estimation is reasonable, since our bounds also rely on a single model sampled from Q_θ, and since Theorem 1's bound is harder to estimate, as it requires sampling and evaluating a large number of models to estimate the expectation over Q_θ. For all bounds, we fix $\delta = 0.05$ and for Theorem 3 we fix $\lambda = 1$. The details of the bounds' evaluation are deferred in Appendix G.1.

Optimization. We use the Adam optimizer (Kingma

and Ba, 2015). We set the parameters β_1 and β_2 to their default values in PyTorch. For each experiment, we learn 3 prior distributions with $\mathcal{S}_P$ using learning rates in $\{0.1, 0.01, 0.001\}$, with 20 epochs. We select the best-performing prior (according to the same loss used for optimization) on $\mathcal{S}$ to compute the bound. To learn the posterior on $\mathcal{S}$ we set the learning rate to 10^{-8}, and the number of epochs to 10. We fix the batch size to 256.

Analysis. Figure 1 exhibits the bounds values computed on $\mathcal{S}$, along with the CVaR computed on the test set $\mathcal{T}$, highlighting the tightness of the bounds as a function of $\alpha \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9\}$. To give additional information on the performance of the models, and since the CVaR is not necessarily easy to interpret on its own, we report the F-score on $\mathcal{T}$.

First of all, as expected, Figure 1 shows that α strongly influences the tightness of the bounds: the higher α, the tighter the bounds. This is not only due to the factor $\frac{1}{\alpha}$ or $\frac{1}{\alpha^2}$ in the bounds, but also because a larger α makes the CVaR tighter. Indeed, when $\alpha \to 0$, the CVaR acts as a supremum over the subgroups and puts all the weights on the subgroup that has the highest risk, while when $\alpha \to 1$ the weights are equal to the reference. This highlights that α plays an important role in the behavior of the algorithm and needs to be chosen to balance the predictive performance and the theoretical guarantee. To confirm this, Figure 2 reports class-wise error rates on the test set $\mathcal{T}$ as a function of α when optimizing Corollary 1's bound (since it provides the best F-score). We observe that depending on the dataset and on the value of α, the class-wise error rates move closer or farther apart. The main issue with the choice of α is that we cannot select the one associated with the lowest bound, since *(i)* the tightest bound is obtained when $\alpha = 1$, and *(ii)* we observe on Figure 1 that the value of α leading to the tightest bounds does not yield the best F-score.

If we compare Theorems 1 and 3 (which uses the same subgroups defined by one example), as expected, our bound is generally tighter (or very close for *mammography*), for all values of α. Remarkably, when $\alpha \in \{0.01, 0.1, 0.3\}$, Corollary 1 gives the smallest bound, and it continues to give non-vacuous and competitive bounds as long as α remains relatively high despite the $\frac{1}{\alpha m_{\mathrm{A}}}$ term in the bound. Moreover, as mentioned previously, Corollary 1 gives the best F-score, confirming the interest of capturing the subgroups in $\mathcal{S}$ with our constrained f-entropic risk measures to tackle the imbalance better.

Synthetic experiments. We report in Figure 3 additional experiments on synthetic data, to show the convergence of the bounds of Theorems 1 and 3, and Corollary 1 to the associated empirical risk $\widehat{\mathcal{R}}_{\mathcal{S}}$ as a function of the number m of examples in $\mathcal{S}$ from 100 to 100,000 (with a test set size of m). We consider two settings (with 3 random seeds) *(on the top figure)* when the classes are perfectly balanced *(on the bottom figures)* when one class has a fixed size of 50 and the other one varies. We observe that for any value of α and even in imbalanced settings, the bound values tend to the empirical risk when m increases.

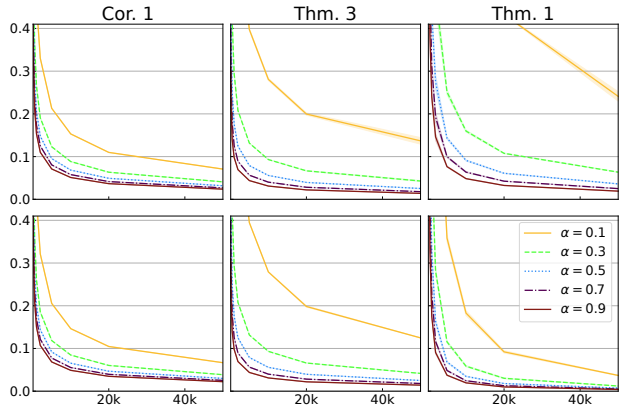


Figure 3: Evolution of the difference between the bound values and the empirical risks as a function of m for Theorems 1 and 3. On the top figures the classes are balanced when m varies. On the bottom figures one class size is fixed to 50 when the other varies.

7 CONCLUSION

In this paper, we introduce classical and disintegrated PAC-Bayesian generalization bounds for a broad new family of risks, namely the constrained f-entropic risk measures. We show that the computable terms of the disintegrated bounds can be minimized with a self-bounding algorithm, leading to models equipped with tight PAC-Bayesian generalization guarantees.

As direct future work, we plan to extend our algorithm to different frameworks with subgroup structure (*e.g.*, groups defined by populations in fairness settings or by tasks in multitask learning). Another direction is to explore the integration of alternative risk measures into self-bounding algorithms. One could replace the f-divergence with Integral Probability Metrics, such as Wasserstein distance, or consider related risk measures such as group DRO or OCEs. Finally, we believe that our work opens the door to studying the generalization properties of other measures. For example, we could design an extension where α varies across subgroups $\mathrm{A} \in \mathcal{A}$, which can be relevant, *e.g.*, in cost-sensitive learning, or adapt (and potentially learn) α dynamically to better handle harder-to-learn subgroups.

Acknowledgments

We would like to sincerely thank the reviewers for their valuable feedback. We are also grateful to Rémi Eyraud for his support during this work and to Rémi Emonet for valuable feedback on a draft of the manuscript. This work was supported in part by the French Project FAMOUS ANR-23-CE23-0019. Paul Viallard is partially funded through Inria with the associate team PACTOL and the exploratory action HYPE.

References

Agrawal, A., Amos, B., Barratt, S. T., Boyd, S. P., Diamond, S., and Kolter, J. Z. (2019). Differentiable Convex Optimization Layers. In *Advances in Neural Information Processing Systems*.

Ahmadi-Javid, A. (2012). Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*.

Ali, S. M. and Silvey, S. D. (1966). A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Ambroladze, A., Parrado-Hernández, E., and Shawe-Taylor, J. (2006). Tighter PAC-Bayes Bounds. In *Advances in Neural Information Processing Systems*.

Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*.

Ben-Tal, A. and Teboulle, M. (1986). Expected Utility, Penalty Functions, and Duality in Stochastic Nonlinear Programming. *Management Science*.

Ben-Tal, A. and Teboulle, M. (2007). An Old-New Concept Of Convex Risk Measures: The Optimized Certainty Equivalent. *Mathematical Finance*.

Blanchard, G. and Fleuret, F. (2007). Occam's Hammer. In *Conference on Learning Theory*.

Brown, D. B. (2007). Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*.

Catoni, O. (2007). PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv*, abs/0712.0248.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*.

Csiszár, I. (1967). Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica*.

Curi, S., Levy, K. Y., Jegelka, S., and Krause, A. (2020). Adaptive Sampling for Stochastic Risk-Averse Learning. In *Advances in Neural Information Processing Systems*.

Delage, E. and Ye, Y. (2010). Distributionally Robust Optimization under Moment Uncertainty with Application to Data-Driven Problems. *Operations research*.

Diamond, S. and Boyd, S. P. (2016). CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*.

Dziugaite, G. K., Hsu, K., Gharbieh, W., Arpino, G., and Roy, D. M. (2021). On the role of data in PAC-Bayes. In *International Conference on Artificial Intelligence and Statistics*.

Dziugaite, G. K. and Roy, D. M. (2017). Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence*.

Dziugaite, G. K. and Roy, D. M. (2018). Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*.

Freund, Y. (1998). Self Bounding Learning Algorithms. In *Conference on Computational Learning Theory*.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kubat, M., Holte, R. C., and Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*.

Lee, J., Park, S., and Shin, J. (2020). Learning Bounds for Risk-sensitive Learning. In *Advances in Neural Information Processing Systems*.

Malerba, D. (1994). Page Blocks Classification. UCI Machine Learning Repository.

Maurer, A. (2004). A Note on the PAC Bayesian Theorem. *arXiv*, cs.LG/0411099.

McAllester, D. A. (1998). Some PAC-Bayesian Theorems. In *Conference on Computational Learning Theory*.

McAllester, D. A. (2003). PAC-Bayesian Stochastic Model Selection. *Machine Learning*.

Mhammedi, Z., Guedj, B., and Williamson, R. C. (2020). PAC-Bayesian Bound for the Conditional Value at Risk. In *Advances in Neural Information Processing Systems*.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2023). SCS: Splitting Conic Solver.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. (2012). PAC-bayes bounds with data dependent priors. *Journal of Machine Learning Research*.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. (2021). Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer New York.

Rivasplata, O. (2022). *PAC-Bayesian Computation*. PhD thesis, University College London, United Kingdom.

Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. (2020). PAC-Bayes Analysis Beyond the Usual Bounds. In *Advances in Neural Information Processing Systems*.

Rockafellar, R. T. and Uryasev, S. (2000). Optimization of Conditional Value-at-Risk. *Journal of risk*.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*.

Scarf, H. E. (1957). A min-max solution of an inventory problem. Technical report, Rand Corporation.

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC Analysis of a Bayesian Estimator. In *Conference on Computational Learning Theory*.

Siegler, R. (1976). Balance Scale. UCI Machine Learning Repository.

Valiant, L. (1984). A Theory of the Learnable. *Communications of the ACM*.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). OpenML: networked science in machine learning. *SIGKDD Explorations*.

Vapnik, V. (1991). Principles of Risk Minimization for Learning Theory. In *Advances in Neural Information Processing Systems*.

Vapnik, V. and Chervonenkis, A. (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*.

Viallard, P. (2023). *PAC-Bayesian Bounds and Beyond: Self-Bounding Algorithms and New Perspectives on Generalization in Machine Learning*. PhD thesis, University Jean Monnet Saint-Etienne, France.

Viallard, P., Emonet, R., Habrard, A., Morvant, E., and Zantedeschi, V. (2024a). Leveraging PAC-Bayes Theory and Gibbs Distributions for Generalization Bounds with Complexity Measures. In *International Conference on Artificial Intelligence and Statistics*.

Viallard, P., Germain, P., Habrard, A., and Morvant, E. (2024b). A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **No**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**, *The code is available* here.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**, *the assumptions are recalled in the statements of the theorems and corollaries.*
   (b) Complete proofs of all theoretical results. **Yes**, *for the sake of readability, the proofs are deferred in the Appendix.*
   (c) Clear explanations of any assumptions. **Yes**, *all the assumptions are related to notations and mathematical settings that have been introduced.*

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**, *as supplementary material.*

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**, *the main details are in Section 6 (additional details are given in Appendix).* **Yes**, *for the sake of readability, note that the main details are given in Section 6 (and the complete details are given in the Appendix).*

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **No**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. **Yes**, *in Section 6.*

   (b) The license information of the assets, if applicable. **Not Applicable**

   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes**, *our code is in the supplementary material.*

   (d) Information about consent from data providers/curators. **Not Applicable**, *the data used are publicly available from OpenML (Vanschoren et al., 2013).*

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. **Not Applicable**

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Supplementary Materials of PAC-Bayesian Bounds on Constrained f-Entropic Risk Measures

The supplementary materials are organized as follows.

- Section A recalls the list of the main notations of the paper;
- Section B discusses the relationship between (constrained) f-entropic risk measures and OCE measures;
- Sections C to F contains all the proofs of our statements;
- Section G gives more details about our method and experimental setting;
- Section H reports the associated additional empirical results;

A TABLES OF NOTATION

Probability theory

$\mathbb{E}_{z\sim\mathcal{Z}}$	Expectation *w.r.t.* the random variable $z \sim \mathcal{Z}$
$\mathbb{P}_{z\sim\mathcal{Z}}$	Probability *w.r.t.* the random variable $z \sim \mathcal{Z}$
$\rho \ll \pi$	ρ is absolutely continuous *w.r.t.* π
$\dfrac{d\rho}{d\pi}$	Radon-Nikodym derivative
$\mathrm{KL}(\cdot\|\cdot)$	Kullback-Leibler (KL) divergence
$\mathrm{kl}^+(a\|b)$	KL divergence between 2 Bernoulli distributions with param. a and b (truncated to $a \leq b$)
$\mathcal{M}(\mathcal{H})$	Set of probability measures / distributions
$\mathcal{N}(\theta, \sigma^2)$	Normal distribution with mean θ and variance σ^2

Main notations

$\mathcal{X}$	Input space
$\mathcal{Y}$	Output/label space
D	Data distribution over $\mathcal{X}\times\mathcal{Y}$
D^m	Distribution of a m-sample
$\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim D^m$	Learning set of m examples drawn *i.i.d.* from D
$\mathcal{A} = \{\mathrm{A}_1, \ldots, \mathrm{A}_n\}$	Partition of the data from D into n subgroups
$\mathcal{S} = \{\mathcal{S}_\mathrm{A}\}_{\mathrm{A}\in\mathcal{A}}$	Partition of $\mathcal{S}$ into n subgroups
$\forall \mathrm{A} \in \mathcal{A},\ \mathcal{S}_\mathrm{A} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m_\mathrm{A}}$	A subgroup $\mathcal{S}_\mathrm{A}$ consists of m_A examples
$D_{\|\mathrm{A}}$	Conditional distribution on $\mathrm{A} \in \mathcal{A}$
π	Reference distribution over $\mathcal{A}$
ρ	Distribution over $\mathcal{A}$
$\mathcal{H}$	Hypothesis space of predictors $h : \mathcal{X} \to \mathcal{Y}$
P	(PAC-Bayesian) prior distribution over $\mathcal{H}$
Q or $Q_\mathcal{S}$	(PAC-Bayesian) posterior distribution over $\mathcal{H}$
$\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$	Deterministic algorithm to learn $Q_\mathcal{S} = \Phi(\mathcal{S}, P)$

Risk measures

$\ell(\cdot,\cdot)$	Loss function $\mathcal{Y}\times\mathcal{Y}\to[0,1]$
$L(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D} \ell(y, h(\mathbf{x}))$	Classical true risk of h
$\hat{L}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(\mathbf{x}_i))$	Classical empirical risk of h
$L_{\mathrm{A}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D_{\mid \mathrm{A}}} \ell(y, h(\mathbf{x}))$	Classical true risk of h on subgroup A
$\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) = \frac{1}{m_{\mathrm{A}}}\sum_{j=1}^{m_{\mathrm{A}}} \ell(y_j, h(\mathbf{x}_j))$	Classical empirical risk of h on subgroup $\mathcal{S}_{\mathrm{A}}$ of size m_{A}
$\mathcal{R}(h) = \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h)$	True risk measure
$\widehat{\mathcal{R}}_{\mathcal{S}}(h) = \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)$	Empirical risk measure
with $E = E_{f,\beta} := \left\{\rho \;\middle|\; \rho \ll \pi \text{ and } \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right) \le \beta\right\}$	f-entropic risk measure
with $E = E_{\alpha} = \left\{\rho \;\middle|\; \rho \ll \pi \text{ and } \frac{d\rho}{d\pi} \le \frac{1}{\alpha}\right\}$	Conditional Value at Risk (CVaR)
with $E = \left\{\rho \;\middle|\; \rho \ll \pi \text{ and } \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right) \le \beta \text{ and } \forall \mathrm{A}\in\mathcal{A},\ \frac{d\rho}{d\pi}(\mathrm{A}) \le \frac{1}{\alpha}\right\}$	Constrained f-entropic risk measure
$\mathcal{R}(Q) := \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)$	Randomized risk measure
$\mathop{\mathbb{E}}_{h\sim Q} \mathcal{R}(h) := \mathop{\mathbb{E}}_{h\sim Q} \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h)$	we have $\mathcal{R}(Q) \le \mathbb{E}_{h\sim Q}\, \mathcal{R}(h)$

Specific notations of Section 5, *i.e.*, for the self-bounding algorithm

$\mathcal{S}$	Learning set for the posterior
$\mathcal{S}_P$	Learning set for the prior (independent from $\mathcal{S}$)
$\mathcal{T}$	Test set
$U \subset \mathcal{S}$	A mini-batch
$P = \mathcal{N}(\theta_P, \sigma^2 I_d)$	Prior parametrized by θ_P
$Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$	Posterior parametrized by θ
θ	Parameters of Q
$h_{\tilde{\theta}}$	Model drawn from the current Q_θ at each iteration
$\widehat{\mathcal{R}}_U(h_{\tilde{\theta}})$	Risk measure evaluated on the mini-batch U
$\mathcal{B}(\cdot)$	Objective function associated with the bound
$h_{\hat{\theta}}$	The final model drawn from the final Q_θ

B ABOUT THE LINK BETWEEN (CONSTRAINED) f-ENTROPIC RISK MEASURES AND OCES

In order to compare more precisely the (constrained) f-entropic risk measure and the Optimized Certainty Equivalents (OCE), we first present another formulation of the f-entropic risk measure, and the definition of an OCE.

(Constrained) f-entropic risk measure. Let $\beta \geq 0$, recall from Assumption 1 and Definition 1 that true and empirical f-entropic risk measures are defined by

$$\mathcal{R}(h) = \sup_{\rho \in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h) \quad \text{and} \quad \widehat{\mathcal{R}}_{\mathcal{S}}(h) = \sup_{\rho \in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),$$
$$\text{with } E = E_{f,\beta} := \left\{ \rho \;\middle|\; \rho \ll \pi, \text{ and } \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right) \leq \beta \right\},$$

where f is defined such that $D_f(\rho\|\pi) := \mathbb{E}_{\mathrm{A}\sim\pi}\left[f\left(\frac{d\rho}{d\pi}(\mathrm{A})\right)\right]$ is a f-divergence.
From Ahmadi-Javid (2012, Theorem 5.1), we have the following equalities:

$$\mathcal{R}(h) = \inf_{t>0,\mu\in\mathbb{R}} \left\{ t \left[\mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(\frac{L_{\mathrm{A}}(h)}{t} - \mu + \beta\right)\right]\right\}, \text{ and } \widehat{\mathcal{R}}_{\mathcal{S}}(h) = \inf_{t>0,\mu\in\mathbb{R}} \left\{ t \left[\mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(\frac{\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)}{t} - \mu + \beta\right)\right]\right\}, \tag{17}$$

where f^* is the convex conjugate of f. Note that these results hold also for the constrained f-entropic risk measure since it is a f-entropic risk measure as we use the divergence $f + g_\alpha$ instead of f; see Section 3.

OCE Risk Measure. According to Ben-Tal and Teboulle (1986, 2007), an OCE is defined by

$$\mathcal{R}^{\mathsf{oce}}(h) := \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(L_{\mathrm{A}}(h) - \mu\right)\right\} \text{ and } \widehat{\mathcal{R}}^{\mathsf{oce}}_{\mathcal{S}}(h) := \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) - \mu\right)\right\}. \tag{18}$$

Comparison. By comparing Equation (17) and Equation (18), we can remark that in Equation (18), we have $t = 1$ and $\beta = 0$. Following the proof of Theorem 5.1 in Ahmadi-Javid (2012) (with $t = 1$ and $\beta = 0$), we deduce that

$$\mathcal{R}^{\mathsf{oce}}(h) := \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(L_{\mathrm{A}}(h) - \mu\right)\right\} = \sup_{\rho\ll\pi} \left\{ \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h) - D_f(\rho\|\pi)\right\},$$
$$\text{and} \quad \widehat{\mathcal{R}}^{\mathsf{oce}}_{\mathcal{S}}(h) := \inf_{\mu\in\mathbb{R}} \left\{ \mu + \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} f^*\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) - \mu\right)\right\} = \sup_{\rho\ll\pi} \left\{ \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) - D_f(\rho\|\pi)\right\}.$$

Hence, as we can remark, the OCE corresponds to a different optimization problem from the (constrained) f-entropic risk measures. Indeed, the OCE finds the distribution ρ that maximizes $\mathbb{E}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h) - D_f(\rho\|\pi)$ or $\mathbb{E}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h) - D_f(\rho\|\pi)$. The (constrained) f-entropic risk maximizes the risk $\mathbb{E}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h)$ or $\mathbb{E}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)$ while keeping $D_f(\rho\|\pi) \leq \beta$.

C PROOF OF THEOREM 2

In this section, we give the proof of the following theorem.

Theorem 2. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any positive, jointly convex function $\varphi(a,b)$ that is non-increasing in a for any fixed b, for any finite set $\mathcal{A}$ of n subgroups, for any $\lambda_{\mathrm{A}} > 0$ for each $\mathrm{A}\in\mathcal{A}$, for any distribution π on $\mathcal{A}$, for any distribution $P\in\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any constrained f-entropic risk measure $\mathcal{R}$ satisfying Definition 2, for any $\delta\in(0,1]$, for any $\alpha\in(0,1]$, we have the following bounds.*

Classical PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, for all distributions $Q\in\mathcal{M}(\mathcal{H})$, we have*

$$\varphi\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(Q), \mathop{\mathbb{E}}_{h\sim Q} \mathcal{R}(h)\Big) \leq \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha\,\lambda_{\mathrm{A}}} \left[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right]. \tag{10}$$

Disintegrated PAC-Bayes. *For any algorithm $\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$ and $h \sim Q_{\mathcal{S}}$, we have*

$$\varphi\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(h), \mathcal{R}(h)\Big) \leq \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha\,\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right], \tag{11}$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S}, P)$.

We prove Equation (10) in Section C.1, and Equation (11) in Section C.2.

C.1 Proof of Equation (10)

To prove Equation (10), we first prove Lemma 1, which follows the steps of the general proof of the PAC-Bayesian theorem by Germain et al. (2009) and a union bound.

Lemma 1. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any positive, jointly convex function $\varphi(a,b)$, for any finite set $\mathcal{A}$ of n subgroups, for any $\lambda_{\mathrm{A}} > 0$ for each $\mathrm{A}\in\mathcal{A}$, for any distribution π over $\mathcal{A}$, for any distribution $P \in \mathcal{M}(\mathcal{H})$, for any loss function $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$, for any $\delta \in (0,1]$, for any $\alpha \in (0,1)$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, for all distributions $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathrm{A}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \leq \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right].$$

Proof. First of all, our goal is to upper-bound $\lambda_{\mathrm{A}}\varphi\left(\mathbb{E}_{h\sim Q}\,\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathbb{E}_{h\sim Q}\,L_{\mathrm{A}}(h)\right)$ for each $\mathrm{A}\in\mathcal{A}$. To do so, we follow the steps of Germain et al. (2009). From the Donsker-Varadhan representation of the KL divergence, we have

$$\lambda_{\mathrm{A}}\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \leq \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)}\right). \tag{19}$$

Now, we apply Markov's inequality to $\mathbb{E}_{h\sim P}\, e^{\lambda_{\mathrm{A}}\varphi(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h))}$, which is positive. We have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)} \leq \frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right] \geq 1-\frac{\delta}{n} \\
\iff\quad &\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\ln\left(\mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)}\right) \leq \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right] \geq 1-\frac{\delta}{n}.
\end{aligned} \tag{20}$$

Hence, by combining Equation (19) and Equation (20), we have for any $\mathrm{A}\in\mathcal{A}$,

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall Q \in \mathcal{M}(\mathcal{H}), \\ \varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \leq \frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right] \geq 1-\frac{\delta}{n}.$$

As $\mathcal{A}$ is finite with $|\mathcal{A}| = n$, we apply a union bound argument to obtain

$$\iff \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall \mathrm{A}\in\mathcal{A},\ \forall Q \in \mathcal{M}(\mathcal{H}), \\ \varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathrm{A}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \\ \qquad \leq \frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right] \geq 1-\delta \tag{21}$$

$$\iff \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall \mathrm{A}\in\mathcal{A},\ \forall Q \in \mathcal{M}(\mathcal{H}), \\ \pi(\mathrm{A})\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathrm{A}}(h), \mathop{\mathbb{E}}_{h\sim Q} L_{\mathrm{A}}(h)\right) \\ \qquad \leq \pi(\mathrm{A})\frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right] \geq 1-\delta \tag{22}$$

$$\implies \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall Q\in\mathcal{M}(\mathcal{H}),\\ \displaystyle\sum_{\mathrm{A}\in\mathcal{A}}\pi(\mathrm{A})\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathrm{A}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)\\ \qquad\displaystyle\le\sum_{\mathrm{A}\in\mathcal{A}}\pi(\mathrm{A})\frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P)+\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P}e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'),L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right]\ge 1-\delta \tag{23}$$

$$\iff \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall Q\in\mathcal{M}(\mathcal{H}),\\ \displaystyle\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathrm{A}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)\\ \qquad\displaystyle\le\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{1}{\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P)+\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P}e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'),L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right]\ge 1-\delta, \tag{24}$$

which is the desired result. □

Thanks to Lemma 1, we are now ready to prove Equation (10) of Theorem 2.

Proof. For any $\rho^*\in E$, we can define $\varepsilon_{\rho^*}\ge 0$ such that we have

$$\mathcal{R}(h)=\sup_{\rho\in E}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho}L_{\mathrm{A}}(h)=\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}L_{\mathrm{A}}(h)+\varepsilon_{\rho^*}.$$

Therefore, we have for all $\rho^*\in E$

$$\begin{aligned}\varphi\left(\widehat{\mathcal{R}}_{\mathcal{S}}(Q),\mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)-\varepsilon_{\rho^*}\right)&=\varphi\left(\sup_{\rho\in E}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho}\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)\,,\mathop{\mathbb{E}}_{h\sim Q}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}L_{\mathrm{A}}(h)\right)\\&\le\varphi\left(\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)\\&\le\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right),\end{aligned} \tag{25}$$

where the first inequality comes from the fact that $\rho^*\in E$ and φ is non-increasing with respect to its first argument, and we used, for the second inequality, Jensen's inequality (since φ is jointly convex). Moreover, as φ is positive and since $\frac{d\rho^*}{d\pi}(\mathrm{A})\le\frac{1}{\alpha}$ for all $\mathrm{A}\in\mathcal{A}$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*}\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)&=\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{d\rho^*}{d\pi}(\mathrm{A})\,\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)\\&\le\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{1}{\alpha}\,\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right)\\&=\frac{1}{\alpha}\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\varphi\left(\mathop{\mathbb{E}}_{h\sim Q}\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h),\mathop{\mathbb{E}}_{h\sim Q}L_{\mathrm{A}}(h)\right).\end{aligned} \tag{26}$$

By combining Equations (25) and (26) and Lemma 1 we get

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall Q\in\mathcal{M}(\mathcal{H}),\forall\rho^*\in E,\\ \displaystyle\varphi\left(\widehat{\mathcal{R}}_{\mathcal{S}}(Q),\mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)-\varepsilon_{\rho^*}\right)\le\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{1}{\alpha\,\lambda_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P)+\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P}e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'),L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right]\ge 1-\delta\,. \tag{27}$$

Finally, since the bound holds for all $\rho^*\in E$, we can let $\varepsilon_{\rho^*}\to 0$ to get the desired result. □

C.2 Proof of Equation (11)

To prove Equation (11), we first prove Lemma 2; the proof essentially follows the steps of Rivasplata et al. (2020) before applying a union bound.

Lemma 2. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any positive, jointly convex function $\varphi(a,b)$, for any finite set $\mathcal{A}$ of n subgroups, for any $\lambda_{\mathrm{A}} > 0$ for each $\mathrm{A} \in \mathcal{A}$, for any distribution π over $\mathcal{A}$, for any distribution $P \in \mathcal{M}(\mathcal{H})$, for any loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, for any $\delta \in (0,1]$, for any $\alpha \in (0,1)$, for any algorithm $\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$ and $h \sim Q_{\mathcal{S}}$, we have*

$$\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \varphi\left(\hat{L}_{\mathrm{A}}(h), L_{\mathrm{A}}(h)\right) \leq \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right].$$

Proof. We apply Markov's inequality on $e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}}}{dP}(h)}$. Indeed, we have, with probability at least $1-\delta/n$ over $\mathcal{S} \sim D^m$ and $h \sim Q_{\mathcal{S}}$

$$\begin{aligned}
& e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}}}{dP}(h)} \leq \frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}'}} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}'}}{dP}(h)} \\
\iff \quad & \ln\left(e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}}}{dP}(h)}\right) \leq \ln\frac{n}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}'}} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}'}}{dP}(h)}\right) \\
\iff \quad & \lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}}}{dP}(h) \leq \ln\frac{n}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}'}} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}'}}{dP}(h)}\right) \\
\iff & \lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) - \ln\frac{dQ_{\mathcal{S}}}{dP}(h) \leq \ln\frac{n}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)}\right) \\
\iff & \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) \leq \frac{1}{\lambda_{\mathrm{A}}}\left[\ln\frac{dQ_{\mathcal{S}}}{dP}(h) + \ln\frac{n}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)}\right)\right].
\end{aligned}$$

Furthermore, since $\ln(\cdot) \leq \ln^+(\cdot)$, with probability at least $1-\delta/n$ over $\mathcal{S} \sim D^m$ and $h \sim Q_{\mathcal{S}}$, we have

$$\varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) \leq \frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\frac{n}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)}\right)\right].$$

As $\mathcal{A}$ is finite with $|\mathcal{A}| = n$, we apply the union bound argument to obtain

$$\begin{aligned}
\iff & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}}\left[\begin{array}{l} \forall \mathrm{A} \in \mathcal{A}, \\ \varphi\left(\hat{L}_{\mathrm{A}}(h), L_{\mathrm{A}}(h)\right) \\ \qquad \leq \frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right] \end{array}\right] \geq 1-\delta \\
\iff & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}}\left[\begin{array}{l} \forall \mathrm{A} \in \mathcal{A}, \\ \pi(\mathrm{A})\varphi\left(\hat{L}_{\mathrm{A}}(h), L_{\mathrm{A}}(h)\right) \\ \qquad \leq \pi(\mathrm{A})\frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right] \end{array}\right] \geq 1-\delta \\
\implies & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}}\left[\begin{array}{l} \sum_{\mathrm{A}\in\mathcal{A}} \pi(\mathrm{A})\varphi\left(\hat{L}_{\mathrm{A}}(h), L_{\mathrm{A}}(h)\right) \\ \qquad \leq \sum_{\mathrm{A}\in\mathcal{A}} \pi(\mathrm{A})\frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right] \end{array}\right] \geq 1-\delta \\
\iff & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}}\left[\begin{array}{l} \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \varphi\left(\hat{L}_{\mathrm{A}}(h), L_{\mathrm{A}}(h)\right) \\ \qquad \leq \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right] \end{array}\right] \geq 1-\delta,
\end{aligned}$$

which is the desired result. □

We are now ready to prove Equation (11) of Theorem 2.

Proof. For any $\rho^* \in E$, we can define $\varepsilon_{\rho^*} \geq 0$ such that we have

$$\mathcal{R}(h) = \sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} L_{\mathrm{A}}(h) = \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} L_{\mathrm{A}}(h) + \varepsilon_{\rho^*}.$$

Therefore, we have for all $\rho^* \in E$

$$\begin{aligned}\varphi\left(\widehat{\mathcal{R}}_{\mathcal{S}}(h), \mathcal{R}(h) - \varepsilon_{\rho^*}\right) &= \varphi\left(\sup_{\rho\in E} \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h)\,,\ \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} L_{\mathrm{A}}(h)\right)\\ &\le \varphi\left(\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} \hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} L_{\mathrm{A}}(h)\right)\\ &\le \mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right), \end{aligned} \tag{28}$$

where the first inequality comes from the fact that $\rho^* \in E$ and φ is non-increasing with respect to its first argument, and we used, for the second inequality, Jensen's inequality (since φ is jointly convex).
Moreover, as φ is positive and since $\frac{d\rho^*}{d\pi}(\mathrm{A}) \le \frac{1}{\alpha}$ for all $\mathrm{A} \in \mathcal{A}$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{A}\sim\rho^*} \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right) &= \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{d\rho^*}{d\pi}(\mathrm{A})\ \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)\\ &\le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha}\ \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right)\\ &= \frac{1}{\alpha} \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \varphi\left(\hat{L}_{\mathcal{S}_{\mathrm{A}}}(h), L_{\mathrm{A}}(h)\right). \end{aligned} \tag{29}$$

By combining Equations (28) and (29) and Lemma 2 we get

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\forall \rho^* \in E,\\ \varphi\left(\widehat{\mathcal{R}}_{\mathcal{S}}(h), \mathcal{R}(h) - \varepsilon_{\rho^*}\right) \le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{\alpha\,\lambda_{\mathrm{A}}}\left[\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \mathop{\mathbb{E}}_{h'\sim P} e^{\lambda_{\mathrm{A}}\varphi\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h'), L_{\mathrm{A}}(h')\right)}\right)\right]\end{array}\right] \ge 1-\delta\,. \tag{30}$$

Finally, since the bound holds for all $\rho^* \in E$, we can have $\varepsilon_{\rho^*} \to 0$ to get the desired result. □

D ABOUT THE kl^+

In this section, we prove two properties of kl^+ that are useful in Section E.

Lemma 3 (Useful properties on kl^+)**.** *For any $a, b \in [0,1]$ we have*

$$\mathrm{kl}(a\|b) \triangleq a \ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} \quad \textit{and} \quad \mathrm{kl}^+(a\|b) \triangleq \begin{cases}\mathrm{kl}(a\|b) & \textit{if } a \le b,\\ 0 & \textit{otherwise.}\end{cases}$$

1. $\mathrm{kl}^+(a\|b)$ *is non-increasing in a for any fixed b.*
2. $\mathrm{kl}^+(a\|b) \le \mathrm{kl}(a\|b)$.

Proof of 1. If $a > b$, by definition, $\mathrm{kl}^+(a\|b) = 0$, which is constant. Otherwise, if $a \le b$, we compute the derivative of $\mathrm{kl}(a\|b)$ with respect to a. We have

$$\begin{aligned}\frac{d}{da}\mathrm{kl}(a\|b) &= \frac{d}{da}\left[a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}\right]\\ &= \ln\frac{a}{b} - \ln\frac{1-a}{1-b}.\\ &= \ln\left(\frac{a(1-b)}{b(1-a)}\right).\end{aligned}$$

For $a \le b$, we have $\frac{a(1-b)}{b(1-a)} \le 1$, so its logarithm is non-positive, meaning $\frac{d}{da}\mathrm{kl}(a\|b) \le 0$. Thus, $\mathrm{kl}(a\|b)$ is non-increasing in a when $a \le b$. □

Proof of 2. If $a \le b$, $\mathrm{kl}^+(a\|b) = \mathrm{kl}(a\|b)$. Otherwise, $a > b$, $\mathrm{kl}^+(a\|b) = 0 \le \mathrm{kl}(a\|b)$ as $\mathrm{kl}(a\|b) \ge 0$. □

Lemma 4 (Pinsker's inequality for kl^+)**.** *For any $a, b \in [0,1]$,*

$$b - a \leq \sqrt{\frac{1}{2}\,\mathrm{kl}^+\left(a \| b\right)}$$

Proof. If $a \leq b$, $\mathrm{kl}^+ = \mathrm{kl}$, we apply Pinsker's inequality. Otherwise, $a > b$, meaning $b - a < 0$, and $\sqrt{\frac{1}{2}\,\mathrm{kl}^+\left(a \| b\right)} = 0$, so the inequality holds. □

E COROLLARIES OF THEOREM 2

E.1 Corollary 1

Corollary 1. *For any D on $\mathcal{X} \times \mathcal{Y}$, for any $\mathcal{A}$ of n subgroups, for any π over $\mathcal{A}$, for any $P \in \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, for any $\mathcal{R}$ satisfying Definition 2, for any $\delta \in (0,1]$, for any $\alpha \in (0,1]$, for any algorithm $\Phi : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$ and $h \sim Q_{\mathcal{S}}$, we have*

$$\mathrm{kl}^+\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(h) \Big\| \mathcal{R}(h)\Big) \leq \mathop{\mathbb{E}}_{\mathrm{A} \sim \pi} \frac{\ln^+\Big[\frac{dQ_{\mathcal{S}}}{dP}(h)\Big] + \ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{\alpha\, m_{\mathrm{A}}}, \tag{12}$$

$$\text{and } \mathcal{R}(h) \leq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A} \sim \pi} \frac{\ln^+\Big[\frac{dQ_{\mathcal{S}}}{dP}(h)\Big] + \ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{2\,\alpha\, m_{\mathrm{A}}}}, \tag{13}$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S}, P)$.

Proof of Equation (12). As $\mathrm{kl}^+(a,b)$ is positive and non-increasing in a (Lemma 3) we can apply Theorem 2 with $\lambda_{\mathrm{A}} = m_{\mathrm{A}}$ for any $\mathrm{A} \in A$ and the function kl^+. We have with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$, $\forall Q \in \mathcal{M}(\mathcal{H})$,

$$\mathrm{kl}^+\left(\widehat{\mathcal{R}}_{\mathcal{S}}\left(Q\right) \Big\| \mathcal{R}\left(h\right)\right) \leq \mathop{\mathbb{E}}_{\mathrm{A} \sim \pi} \frac{1}{\alpha\, m_{\mathrm{A}}} \left[\mathrm{KL}(Q \| P) + \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} \mathop{\mathbb{E}}_{h' \sim P} e^{m_{\mathrm{A}} \mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h') \| L_{\mathrm{A}}(h')\right)}\right)\right]. \tag{31}$$

Since P does not depend on $\mathcal{S}'$, we have for any $\mathrm{A} \in \mathcal{A}$,

$$\ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} \mathop{\mathbb{E}}_{h' \sim P} e^{m_{\mathrm{A}}\, \mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h') \| L_{\mathrm{A}}(h')\right)}\right) = \ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{h' \sim P} \mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} e^{m_{\mathrm{A}}\, \mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h') \| L_{\mathrm{A}}(h')\right)}\right).$$

Thanks to Maurer (2004), for any $\mathrm{A} \in \mathcal{A}$ for any $h \in \mathcal{H}$, we have

$$\mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} e^{m_{\mathrm{A}}\, \mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h) \,\|\, L_{\mathrm{A}}(h)\right)} \leq \mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} e^{m_{\mathrm{A}}\, \mathrm{kl}\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h) \,\|\, L_{\mathrm{A}}(h)\right)} \leq 2\sqrt{m_{\mathrm{A}}},$$

Where the first inequality comes from the fact that $\mathrm{kl}^+ \leq \mathrm{kl}$ (see Lemma 3).

Therefore, we have

$$\ln\left(\frac{n}{\delta} \mathop{\mathbb{E}}_{h' \sim P} \mathop{\mathbb{E}}_{\mathcal{S}' \sim D^m} e^{m_{\mathrm{A}}\, \mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h') \| L_{\mathrm{A}}(h')\right)}\right) \leq \ln\left(\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}\right). \tag{32}$$

We get the desired result by combining Equation (31) and Equation (32) □

Proof of Equation (13). We apply Lemma 4 on Equation (12) and rearrange the terms. □

E.2 Corollary 2

Corollary 2. *For any* finite *set of n subgroups $\mathcal{A}$, for any distribution π over $\mathcal{A}$, for any distribution D over $\mathcal{X}\times\mathcal{Y}$, for any distribution $P \in \mathcal{M}(\mathcal{H})$, for any loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, for any $\delta \in (0,1]$, for any $\alpha \in (0,1)$ with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, for all distributions $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\mathrm{kl}^+\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\Big\|\mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h)\Big) \le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{\mathrm{KL}(Q\|P)+\ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{\alpha\, m_{\mathrm{A}}}, \tag{33}$$

$$\text{and}\quad \mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h) \le \widehat{\mathcal{R}}_{\mathcal{S}}(Q) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{\mathrm{KL}(Q\|P)+\ln\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}}{2\,\alpha\, m_{\mathrm{A}}}}. \tag{34}$$

Proof of Equation (33). As $\mathrm{kl}^+(a,b)$ is positive and non-increasing in a (Lemma 3) we can apply of Theorem 2 with $\lambda_{\mathrm{A}} = m_{\mathrm{A}}$ for any $\mathrm{A} \in A$ and the function kl^+. We have with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$, $\forall Q \in \mathcal{M}(\mathcal{H})$,

$$\mathrm{kl}^+\left(\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\Big\|\mathcal{R}(h)\right) \le \mathop{\mathbb{E}}_{\mathrm{A}\sim\pi}\frac{1}{\alpha\, m_{\mathrm{A}}}\left[\mathrm{KL}(Q\|P)+\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{m_{\mathrm{A}}\mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h')\|L_{\mathrm{A}}(h')\right)}\right)\right]. \tag{35}$$

Since P does not depend on $\mathcal{S}'$ we have for any $\mathrm{A} \in \mathcal{A}$,

$$\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\mathop{\mathbb{E}}_{h'\sim P} e^{m_{\mathrm{A}}\,\mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h')\|L_{\mathrm{A}}(h')\right)}\right) = \ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{h'\sim P}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} e^{m_{\mathrm{A}}\,\mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h')\|L_{\mathrm{A}}(h')\right)}\right).$$

Thanks to Maurer (2004), for any $\mathrm{A} \in \mathcal{A}$ for any $h \in \mathcal{H}$, we have

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} e^{m_{\mathrm{A}}\,\mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h)\,\|\,L_{\mathrm{A}}(h)\right)} \le \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} e^{m_{\mathrm{A}}\,\mathrm{kl}\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h)\,\|\,L_{\mathrm{A}}(h)\right)} \le 2\sqrt{m_{\mathrm{A}}},$$

where the first inequality comes from the fact that $\mathrm{kl}^+ \le \mathrm{kl}$ (see Lemma 3). Therefore, we have

$$\ln\left(\frac{n}{\delta}\mathop{\mathbb{E}}_{h'\sim P}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} e^{m_{\mathrm{A}}\,\mathrm{kl}^+\left(\hat{L}_{\mathcal{S}'_{\mathrm{A}}}(h')\|L_{\mathrm{A}}(h')\right)}\right) \le \ln\left(\frac{2n\sqrt{m_{\mathrm{A}}}}{\delta}\right). \tag{36}$$

We get the desired result by combining Equation (35) and Equation (36) □

Proof of Equation (34). We apply Lemma 4 on Equation (33) and rearrange the terms. □

F THEOREM 3

Theorem 3. *For any D on $\mathcal{X}\times\mathcal{Y}$, for any $\lambda > 0$, for any $P \in \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, for any constrained f-entropic risk measure $\widehat{\mathcal{R}}_{\mathcal{S}}$ satisfying Definition 2* and *Equation* (14), *for any algorithm $\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any $\delta \in (0,1]$, we have the following bounds.*

Classical PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, we have*

$$\left|\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}}\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \le \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\mathrm{KL}(Q_{\mathcal{S}}\|P)+\ln\left[\frac{2(\lambda+1)}{\delta}\right]+3.5\right)}, \tag{15}$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S}, P)$.

Disintegrated PAC-Bayes. *With probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ and $h\sim Q_{\mathcal{S}}$, we have*

$$\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \le \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\delta}\right]\right)}, \tag{16}$$

where $Q_{\mathcal{S}}$ is the posterior learned with $\Phi(\mathcal{S}, P)$.

In the following, we first start by proving Equation (16) and then we prove Equation (15).

F.1 Proof of Equation (16)

To prove Theorem 3, we first prove the following lemma.

Lemma 5. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any constrained f-entropic risk measure $\widehat{\mathcal{R}}_{\mathcal{S}}$ satisfying Definition 2* and *Equation (14), for any hypothesis $h\in\mathcal{H}$, for any $\delta\in(0,1]$, for any $\alpha\in(0,1]$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|\geq\frac{1}{\alpha}\sqrt{\frac{\ln(2/\delta)}{2m}}\right]\leq\delta.$$

Proof. To prove the result, we aim to apply McDiarmid's inequality. To do so, we need to find an upper-bound of $\sup_{(x'_j,y'_j)\in\mathcal{X}\times\mathcal{Y}}\sup_{\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^m}|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\widehat{\mathcal{R}}_{\mathcal{S}'_j}(h)|$, where $\mathcal{S}$ and $\mathcal{S}'_j$ differ from the j-th example. For any $h\in\mathcal{H}$, any $(x'_j,y'_j)\in\mathcal{X}\times\mathcal{Y}$ and $\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^m$, we have

$$\begin{aligned}
\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\widehat{\mathcal{R}}_{\mathcal{S}'_j}(h)&=\sup_{\rho\in\widehat{E}}\left\{\sum_{i=1}^m\rho(i)\cdot\ell(y_i,h(\mathbf{x}_i))\right\}-\sup_{\rho\in\widehat{E}}\left\{\sum_{i=1}^m\rho(i)\cdot\ell(y'_i,h(\mathbf{x}'_i))\right\}\\
&\leq\sup_{\rho\in\widehat{E}}\left\{\sum_{i=1}^m\rho(i)\cdot\ell(y_i,h(\mathbf{x}_i))-\sum_{i=1}^m\rho(i)\cdot\ell(y'_i,h(\mathbf{x}'_i))\right\}\\
&=\sup_{\rho\in\widehat{E}}\left\{\sum_{i=1}^m\rho(i)\cdot(\ell(y_i,h(\mathbf{x}_i))-\ell(y'_i,h(\mathbf{x}'_i)))\right\}\\
&\leq\sup_{\rho\in\widehat{E}}\left\{\sum_{i=1}^m\rho(i)\cdot|\ell(y_i,h(\mathbf{x}_i))-\ell(y'_i,h(\mathbf{x}'_i))|\right\}\\
&=\sup_{\rho\in\widehat{E}}\left\{\rho(j)\cdot\left|\ell(y_j,h(\mathbf{x}_j))-\ell(y'_j,h(\mathbf{x}'_j))\right|\right\}\\
&\leq\sup_{\rho\in\widehat{E}}\{\rho(j)\}\\
&\leq\sup_{\rho\in\widehat{E}}\left\{\frac{1}{\alpha}\pi(j)\right\}\;=\;\frac{1}{m\alpha}\,.
\end{aligned}$$

Moreover, by doing the same steps, for $\widehat{\mathcal{R}}_{\mathcal{S}'_j}(h)-\widehat{\mathcal{R}}_{\mathcal{S}}(h)$, we obtain: $\widehat{\mathcal{R}}_{\mathcal{S}'_j}(h)-\widehat{\mathcal{R}}_{\mathcal{S}}(h)\leq\dfrac{1}{m\alpha}$.

Finally, we get the desired result by applying McDiarmid's inequality. □

Now we recall Occam's hammer[11] (Theorem 2.4 of Blanchard and Fleuret (2007)) that we use along with Lemma 5 to prove Equation (16).

Lemma 6 (Occam's hammer)**.** *We assume that*

1. *we have*

$$\forall h\in\mathcal{H},\ \forall\delta\in[0,1],\qquad\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}[\mathcal{S}\in\mathcal{B}(h,\delta)]\leq\delta,$$

   where $\mathcal{B}(h,\delta)$ is a set of bad events at level δ for h;

2. *the function $(\mathcal{S},h,\delta)\in\mathcal{Z}^m\times\mathcal{H}\times[0,1]\to\mathbf{1}_{\{\mathcal{S}\in\mathcal{B}(h,\delta)\}}$ is jointly measurable in its three variables;*

3. *for any $h\in H$, we have $\mathcal{B}(h,0)=\emptyset$;*

4. *for any $h\in\mathcal{H}$, $\mathcal{B}(h,\delta)$ is a nondecreasing sequence of sets: for $\delta\leq\delta'$, we have $\mathcal{B}(h,\delta)\subseteq\mathcal{B}(h,\delta')$.*

[11] Lemma 6 is a simpler version than Occam's hammer presented in Blanchard and Fleuret (2007).

Then, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D,\ h\sim Q_{\mathcal{S}}}\left[\mathcal{S}\in\mathcal{B}\left(h,\Delta\left(h,\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)\right)\right]\leq\delta,$$

where $\Delta(h,u) := \min(\delta\beta(u),1)$*, with* Γ *be a probability distribution on* $(0,+\infty)$ *and* $\beta(x) = \int_0^x u d\Gamma(u)$ *for* $x\in(0,+\infty)$*.*

We are now ready to prove Equation (16) based on Lemma 5 and Lemma 6.

Proof. Thanks to Lemma 5 we define for any $h\in\mathcal{H}$, any $\delta\in[0,1]$,

$$\mathcal{B}(h,\delta)=\left\{\mathcal{S}\in\mathcal{Z}^m\ \middle|\ \left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{\ln(2/\delta)}{2m}}\right\}. \tag{37}$$

Now we apply Lemma 6 to our set of Equation (37). As in the proof of Proposition 3.1 in Blanchard and Fleuret (2007), we set Γ as the probability distribution on $[0,1]$ having density $\Gamma(u)=\frac{1}{k}u^{-1+\frac{1}{k}}$ for any $k>0$. Then we can compute $\beta(x)$. For the sake of completeness, we compute β. We consider two cases.

- For $x\leq 1$, we have

$$\begin{aligned}
\beta(x) &= \int_0^x u d\Gamma(u)\\
&= \int_0^x u\frac{1}{k}u^{-1+\frac{1}{k}}du\\
&= \frac{1}{k}\int_0^x u^{\frac{1}{k}}du\\
&= \frac{1}{k}\left[\frac{1}{\frac{1}{k}+1}u^{\frac{1}{k}+1}\right]_0^x\\
&= \frac{1}{k}\left[\frac{k}{k+1}u^{\frac{1}{k}+1}\right]_0^x\\
&= \frac{1}{k}\frac{k}{k+1}x^{\frac{1}{k}+1} = \frac{1}{k+1}x^{\frac{1}{k}+1}.
\end{aligned}$$

- For $x>1$, we have

$$\begin{aligned}
\beta(x) &= \int_0^x u d\Gamma(u)\\
&= \int_0^1 u\frac{1}{k}u^{-1+\frac{1}{k}}du + \int_1^x 0du\\
&= \frac{1}{k}\int_0^1 u^{\frac{1}{k}}du + 0\\
&= \frac{1}{k}\left[\frac{1}{\frac{1}{k}+1}u^{\frac{1}{k}+1}\right]_0^1\\
&= \frac{1}{k}\left[\frac{k}{k+1}u^{\frac{1}{k}+1}\right]_0^1\\
&= \frac{1}{k}\frac{k}{k+1}1^{\frac{1}{k}+1}\\
&= \frac{1}{k+1}.
\end{aligned}$$

Therefore, we can deduce that $\beta(x) = \frac{1}{k+1}\min(x^{1+\frac{1}{k}}, 1)$. Then, by applying Lemma 6, we have with probability at least $1-\delta$ over $\mathcal{S} \sim D^m,\ h \sim Q_{\mathcal{S}}$

$$
\begin{aligned}
&\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2}{\Delta\left(h, \left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)}\right)\right]} \\
\iff &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2}{\min\left(\delta\beta\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right), 1\right)}\right)\right]} \\
\iff &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(2\max\left(\frac{1}{\delta\beta\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)}, 1\right)\right)\right]} \\
\iff &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln(2) + \ln\left(\max\left(\frac{1}{\delta\beta\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)}, 1\right)\right)\right]} \\
\implies &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln(2) + \ln^+\left(\frac{1}{\delta\beta\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)}\right)\right]} \\
\iff &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln(2) + \ln^+\left(\frac{1}{\delta\frac{1}{k+1}\min\left(\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)^{1+\frac{1}{k}}, 1\right)}\right)\right]} \\
\implies &\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln(2) + \ln\left(\frac{k+1}{\delta}\right) + \ln^+\left(\frac{1}{\min\left(\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)^{1+\frac{1}{k}}, 1\right)}\right)\right]}.
\end{aligned}
$$

The last implication is due to the fact that $\frac{k+1}{\delta} \geq 1$.

Let $x, y \in \mathbb{R}_+$ such that $x \geq 1$, we have

$$
\begin{aligned}
\ln^+(xy) &= \max(\ln(xy), 0) \\
&= \max(\ln(x) + \ln(y), 0) \\
&\leq \max(\ln(x), 0) + \max(\ln(y), 0) \\
&= \ln(x) + \max(\ln(y), 0) \\
&= \ln(x) + \ln^+(y)\,,
\end{aligned}
$$

where the inequality is due to the sub-additivity of max.

Moreover, we have with probability at least $1-\delta$ over $\mathcal{S}\sim D^m,\ h\sim Q_{\mathcal{S}}$

$$\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|\le\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2\,(k+1)}{\delta}\right)+\ln^+\left(\frac{1}{\min\left(\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)^{1+\frac{1}{k}},1\right)}\right)\right]}$$

$$\iff\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|\le\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2\,(k+1)}{\delta}\right)+\ln^+\left(\max\left(\frac{1}{\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)^{1+\frac{1}{k}}},1\right)\right)\right]}$$

$$\iff\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|\le\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2\,(k+1)}{\delta}\right)+\ln^+\left(\frac{1}{\left(\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]^{-1}\right)^{1+\frac{1}{k}}}\right)\right]}$$

$$\iff\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|\le\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left[\ln\left(\frac{2\,(k+1)}{\delta}\right)+\left(1+\frac{1}{k}\right)\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right)\right]},$$

which is the desired result. □

F.2 Proof of Equation (15)

This proof comes from Blanchard and Fleuret (2007, Corollary 3.2).

Proof. From Equation (16), we can deduce that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m,h\sim Q_{\mathcal{S}}}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\gamma\delta}\right]\right)}\right]\le\delta\gamma.$$

Moreover, from Markov's inequality, we can deduce that we have

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\gamma\delta}\right]\right)}\right]>\gamma\right]\\ \le&\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\gamma\delta}\right]\right)}\right]\ge\gamma\right]\\ \le&\frac{1}{\gamma}\mathop{\mathbb{E}}_{\mathcal{S}\sim D^m}\mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\gamma\delta}\right]\right)}\right]\le\delta.\end{aligned}\tag{38}$$

For any $i\in\mathbb{N}$, we consider $\delta_i=\delta 2^{-i}$ and $\gamma_i=2^{-i}$ in Equation (38) (instead of δ and γ). Concerning $i=0$, we have a special case: We know that $\delta=0$ since we have $\gamma_0=2^0=1$. Hence, we perform a union bound on δ_i where $i\in\mathbb{N}$; we have $\sum_{i\in\mathbb{N}}\delta_i=\delta_0+\sum_{i\in\mathbb{N},i>0}\delta_i=\sum_{i\in\mathbb{N},i>0}\delta_i=\delta$ and

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\begin{array}{l}\exists i\ge 0\,,\\ \mathbb{P}_{h\sim Q_{\mathcal{S}}}\left[\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathbb{E}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|>\frac{1}{\alpha}\sqrt{\frac{1}{2m}\left(\left[1+\frac{1}{\lambda}\right]\ln^+\left[\frac{dQ_{\mathcal{S}}}{dP}(h)\right]+\ln\left[\frac{2(\lambda+1)}{\delta 2^{-2i}}\right]\right)}\right]>2^{-i}\end{array}\right]\le\delta.$$

Moreover, let

$$\phi(h,\mathcal{S})=2m\alpha^2\left|\widehat{\mathcal{R}}_{\mathcal{S}}(h)-\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m}\widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|^2-\left(1+\frac{1}{\lambda}\right)\ln^+\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right)-\ln\left(\frac{2(\lambda+1)}{\delta}\right),\tag{39}$$

and we have

$$\begin{aligned} & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\exists i \geq 0\,,\ \mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}\left[\phi(h,\mathcal{S}) > 2i\ln(2)\right] > 2^{-i}\right] \leq \delta \\ \iff & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\forall i \geq 0\,,\ \mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}\left[\phi(h,\mathcal{S}) > 2i\ln(2)\right] \leq 2^{-i}\right] \geq 1-\delta. \end{aligned} \tag{40}$$

Moreover, note that we have

$$\begin{aligned} \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S})] &\leq \int_{t\geq 0} \mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S}) > t]\,dt \\ &= \sum_{i\in\mathbb{N}} \int_{2i\ln(2)}^{2(i+1)\ln(2)} \left[\mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S}) > t]\right] dt \\ &\leq \sum_{i\in\mathbb{N}} \int_{2i\ln(2)}^{2(i+1)\ln(2)} \left[\mathop{\mathbb{P}}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S}) > 2i\ln(2)]\right] dt \\ &\leq \sum_{i\in\mathbb{N}} \int_{2i\ln(2)}^{2(i+1)\ln(2)} 2^{-i} dt \\ &= 2\ln(2)\sum_{i\in\mathbb{N}} 2^{-i} dt \\ &= 4\ln(2) \ \leq\ 3. \end{aligned}$$

Put into words, having $\forall i \geq 0\,,\ \mathbb{P}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S}) > 2i\ln(2)] \leq 2^{-i}$ implies that $\mathbb{E}_{h\sim Q_{\mathcal{S}}}[\phi(h,\mathcal{S})] \leq 3$.

Hence, thanks to this implication and Equation (40), we can deduce that we have

$$\begin{aligned} & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} 2m\alpha^2 \left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|^2 - \left(1+\frac{1}{\lambda}\right) \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) - \ln\left(\frac{2(\lambda+1)}{\delta}\right) \leq 3\right] \geq 1-\delta \\ \iff & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\sqrt{\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} \left|\widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right|^2} \leq \frac{1}{\alpha}\sqrt{\frac{\left(1+\frac{1}{\lambda}\right)\mathbb{E}_{h\sim Q_{\mathcal{S}}} \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{2(\lambda+1)}{\delta}\right) + 3}{2m}}\right] \geq 1-\delta \\ \implies & \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m}\left[\left|\mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} \widehat{\mathcal{R}}_{\mathcal{S}}(h) - \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} \mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h)\right| \leq \frac{1}{\alpha}\sqrt{\frac{\left(1+\frac{1}{\lambda}\right)\mathbb{E}_{h\sim Q_{\mathcal{S}}} \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) + \ln\left(\frac{2(\lambda+1)}{\delta}\right) + 3}{2m}}\right] \geq 1-\delta, \end{aligned} \tag{41}$$

where the last implication comes from Jensen's inequality as $\sqrt{\cdot}$ is concave and $|\cdot|$ is convex. Finally, we have,

$$\begin{aligned} \mathop{\mathbb{E}}_{h\sim Q_{\mathcal{S}}} \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) &= \mathop{\mathbb{E}}_{h\sim P} \frac{dQ_{\mathcal{S}}}{dP}(h) \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) \\ &\leq \mathop{\mathbb{E}}_{h\sim P} \frac{dQ_{\mathcal{S}}}{dP}(h) \ln^+\!\left(\frac{dQ_{\mathcal{S}}}{dP}(h)\right) - \min_{0\leq x<1} x\log x \\ &= \mathrm{KL}(Q_{\mathcal{S}}\|P) + e^{-1} \end{aligned} \tag{42}$$

Combining Equation (41) and Equation (42) and bounding e^{-1} by $\frac{1}{2}$ gives the desired result. □

G DETAILS ABOUT THE EXPERIMENTS

G.1 Bounds in practice

G.1.1 Batch sampling

We follow a mini-batch sampling strategy where batches are constructed *w.r.t.* the reference distribution π on the classes in $\mathcal{A}$. In this setting, examples belonging to subgroups that are less represented in the data might

be present in different batches. However, for each batch, we ensure that the data is not redundant and that all subgroups are represented by at least one example.

G.1.2 Prior learning algorithm

Algorithm 1 requires a prior distribution P. In practice, we propose to learn this prior by running Algorithm 2 below (as described in Section 5).

Algorithm 2 Learning a prior distribution for constrained f-entropic risk measures

Require: Prior learning set $\mathcal{S}_P$, posterior learning set $\mathcal{S}$, number of epochs T, variance σ^2, reference π, set of hyperparameter configurations $\mathcal{C}$ of size K, parameters α, β

```
Initialize the set of prior distributions: P ← ∅
for all c ∈ C do
    Initialize θ_P
    for t = 1 to T do
        for all mini-batches U ⊂ S_P drawn w.r.t. π do
            Draw a model h_θ̃_P from P_θ = N(θ_P, σ² I_d)        ▷ where d is the size of θ_P
            Compute the risk R̂_U(h_θ̃_P) on the mini-batch
            Update θ_P with gradient ∇_θP R̂_U(h_θ̃_P)
        end for
        Add P_θ to set of prior distributions: P ← P ∪ {P_θ}
    end for
end for
return P = arg min_{P_θ∈P} { R̂_S(h_θ̃_P), with h_θ̃_P ∼ P_θ }
```

Across the T epochs and the hyperparameter configurations considered, we get $T \times K$ prior distributions on $\mathcal{S}_P$ stored in the set $\mathcal{P}$. In the end, the prior P selected to learn the posterior distribution with Algorithm 1, is the prior that minimizes the risk on the learning set $\mathcal{S}$.

G.1.3 Objective functions for learning the posterior with Algorithm 1

Note that the bounds of Corollary 1, Theorems 1 and 3 do not hold "directly" for the above choice of P as it depends on the posterior set $\mathcal{S}$. To tackle this issue in practice, we adapt and instantiate below the bounds to our practical setting. We respectively obtain Corollaries 3 to 5, which hold for any prior $P_t \in \mathcal{P}$ after drawing $\mathcal{S} \sim D^m$, then, they hold for the prior that minimizes the empirical risk on $\mathcal{S}$. In consequence, the bounds hold for a prior learned by Algorithm 2, and we can deduce the objective functions to minimize.

Instantiation of Corollary 1, and the objective function. The objective function associated to the minimization of Corollary 1 is

$$\boxed{\mathcal{B}_{\text{cor1}}(\widehat{\mathcal{R}}_{\mathcal{S}}(h_{\tilde{\theta}}), Q_\theta, \tilde{\theta}) = \sup_{\substack{\rho \in \mathbb{R}_+^n \\ \frac{\rho_{\mathrm{A}}}{\pi_{\mathrm{A}}} \le \frac{1}{\alpha}}} \sum_{\mathrm{A} \in \mathcal{A}} \rho_{\mathrm{A}} \sum_{i=1}^{m_{\mathrm{A}}} \frac{1}{m_{\mathrm{A}}} \ell(y_i, h_{\tilde{\theta}}(\mathbf{x}_i)) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A} \sim \pi} \frac{1}{2\,\alpha\, m_{\mathrm{A}}} \left[\frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln \frac{2nTK\sqrt{m_{\mathrm{A}}}}{\delta} \right]}}$$

with $Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, and $h_{\tilde{\theta}} \sim Q_\theta$ with parameters $\tilde{\theta}$, and $P = \mathcal{N}(\theta_P, \sigma^2 I_d)$, and $\sigma \in [0,1]$, and $\lambda > 0$, and $\alpha \in (0,1]$, and $\delta \in [0,1]$.

The definition of $\mathcal{B}_{\text{cor1}}$ comes from the following corollary of Corollary 1.

Corollary 3. *For any* finite *set of n subgroups $\mathcal{A}$, for any distribution π over $\mathcal{A}$, for any distribution D over $\mathcal{X} \times \mathcal{Y}$, for any number of epochs T, for any number of hyperparameter configurations K, for any set of distributions $\mathcal{P} \in \{P_1, ..., P_{T \times K}\}$, for any loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, for any $\delta \in (0,1]$, for any $\alpha \in (0,1)$, for any algorithm $\Phi : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any $\sigma \in [0,1]$, with probability at least $1-\delta$ over $\mathcal{S} \sim D^m$ and*

$h \sim Q_{\mathcal{S}}$, we have $\forall P_t = \mathcal{N}(\theta_P, \sigma^2 I_d) \in \mathcal{P}$,

$$\mathcal{R}(h) \leq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{2\,\alpha\, m_{\mathrm{A}}} \left[\frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln \frac{2nTK\sqrt{m_{\mathrm{A}}}}{\delta} \right]}, \tag{43}$$

with $Q_{\mathcal{S}} = \mathcal{N}(\theta, \sigma^2 I_d)$ the posterior distribution.

Proof. As $\frac{\delta}{TK} \in [0,1]$, we have from Corollary 2, for any $P_t \in \mathcal{P}$,

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}} \left[\mathcal{R}(h) \geq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{2\,\alpha\, m_{\mathrm{A}}} \left[\frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln \frac{2nTK\sqrt{m_{\mathrm{A}}}}{\delta} \right]} \right] \leq \frac{\delta}{TK}, \\
\implies &\sum_{t=1}^{TK} \mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}} \left[\mathcal{R}(h) \geq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{2\,\alpha\, m_{\mathrm{A}}} \left[\frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln \frac{2nTK\sqrt{m_{\mathrm{A}}}}{\delta} \right]} \right] \leq \sum_{t=1}^{TK} \frac{\delta}{TK}, \\
\implies &\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim Q_{\mathcal{S}}} \left[\forall \theta_P, \quad \mathcal{R}(h) \geq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\mathop{\mathbb{E}}_{\mathrm{A}\sim\pi} \frac{1}{2\,\alpha\, m_{\mathrm{A}}} \left[\frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln \frac{2nTK\sqrt{m_{\mathrm{A}}}}{\delta} \right]} \right] \leq \delta,
\end{aligned}$$

where the last inequality follows from the union bound.

Instantiation of Theorem 3, and the objective function. The objective function associated to the minimization of Theorem 3 is

$$\boxed{\mathcal{B}_{\mathrm{th3}}(\widehat{\mathcal{R}}_{\mathcal{S}}(h_{\tilde{\theta}}), Q_\theta, \tilde{\theta}) = \sup_{\substack{\rho\in\mathbb{R}_+^n \\ m\rho_{\mathrm{A}} \leq \frac{1}{\alpha}}} \sum_{\mathrm{A}=1}^{m} \rho_{\mathrm{A}} \ell(y_{\mathrm{A}}, h_{\tilde{\theta}}(\mathbf{x}_{\mathrm{A}})) + \sqrt{\frac{1}{2\,\alpha^2\, m} \left[\left(1 + \frac{1}{\lambda}\right) \frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln\left(\frac{2TK(\lambda+1)}{\delta}\right) \right]}}$$

with $Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, and $h_{\tilde{\theta}} \sim Q_\theta$ with parameters $\tilde{\theta}$, and $P = \mathcal{N}(\theta_P, \sigma^2 I_d)$, and $\sigma \in [0,1]$, and $\lambda > 0$, and $\alpha \in (0,1]$, and $\delta \in [0,1]$.

The definition of $\mathcal{B}_{\mathrm{th3}}$ comes from the following corollary of Theorem 3.

Corollary 4. *For any distribution D over $\mathcal{X}\times\mathcal{Y}$, for any $\lambda > 0$, for any number of epochs T, for any number of hyperparameter configuration K, for any set of distributions $\mathcal{P} \in \{P_1, ..., P_{T\times K}\}$, for any loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, for any $\delta \in (0,1]$, for any $\alpha \in (0,1)$, for any algorithm $\Phi : (\mathcal{X}\times\mathcal{Y})^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any $\sigma \in [0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$ and $h \sim Q_{\mathcal{S}}$, we have $\forall P_t = \mathcal{N}(\theta_P, \sigma^2 I_d) \in \mathcal{P}$,*

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim D^m} \widehat{\mathcal{R}}_{\mathcal{S}'}(h) \leq \widehat{\mathcal{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2\,\alpha^2\, m} \left[\left(1 + \frac{1}{\lambda}\right) \frac{\|\tilde{\theta} - \theta_P\|_2^2 - \|\tilde{\theta} - \theta\|_2^2}{2\sigma^2} + \ln\left(\frac{2TK(\lambda+1)}{\delta}\right) \right]}.$$

Proof. The proof follows the same steps as the proof of Corollary 3. □

Instantiation of Theorem 1, and the objective function. We recall that, in practice, we compute an estimation of the bound of Theorem 1 obtained by sampling a single model from the posterior Q_θ (since we deal with disintegrated bounds). The associated objective function is

$$\mathcal{B}_{\mathrm{th1}}(\widehat{\mathcal{R}}_{\mathcal{S}}(h_{\tilde{\theta}}), Q_\theta, \tilde{\theta}) = \sup_{\substack{\rho\in\mathbb{R}_+^n \\ m\rho_{\mathrm{A}}\le\frac{1}{\alpha}}} \sum_{\mathrm{A}=1}^{m} \rho_{\mathrm{A}}\ell(y_{\mathrm{A}}, h_{\tilde{\theta}}(\mathbf{x}_{\mathrm{A}}))$$
$$+\, 2 \sup_{\substack{\rho\in\mathbb{R}_+^n \\ m\rho_{\mathrm{A}}\le\frac{1}{\alpha}}} \sum_{\mathrm{A}=1}^{m} \rho_{\mathrm{A}}\ell(y_{\mathrm{A}}, h_{\tilde{\theta}}(\mathbf{x}_{\mathrm{A}})) \left[\left(\sqrt{\frac{\ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}}{2\,m\,\alpha}} + \frac{\ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}}{3\,m\,\alpha}\right)\right]$$
$$+ \sqrt{\frac{27}{5\,m\,\alpha} \sup_{\substack{\rho\in\mathbb{R}_+^n \\ m\rho_{\mathrm{A}}\le\frac{1}{\alpha}}} \sum_{\mathrm{A}=1}^{m} \rho_{\mathrm{A}}\ell(y_{\mathrm{A}}, h_{\tilde{\theta}}(\mathbf{x}_{\mathrm{A}})) \left[\frac{\|\theta-\theta_P\|_2^2}{2\sigma^2} + \ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}\right]}$$
$$+ \frac{27}{5\,m\,\alpha}\left[\frac{\|\theta-\theta_P\|_2^2}{2\sigma^2} + \ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}\right]$$

with $Q_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, with $h_{\tilde{\theta}} \sim Q_\theta$ with parameters $\tilde{\theta}$, and $P = \mathcal{N}(\theta_P, \sigma^2 I_d)$, and $\sigma \in [0,1]$, and $\lambda > 0$, and $\alpha \in (0,1]$, and $\delta \in [0,1]$.

The definition of $\mathcal{B}_{\mathrm{th1}}$ comes from the following corollary of Theorem 1.

Corollary 5. *For any distribution D over $\mathcal{X}\times\mathcal{Y}$, for any prior $P \in \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, for any $\alpha \in (0,1]$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim D^m$, we have $\forall Q = \mathcal{N}(\theta, \sigma^2 I_d)$ and $\forall P_t = \mathcal{N}(\theta_P, \sigma^2 I_d) \in \mathcal{P}$,*

$$\mathop{\mathbb{E}}_{h\sim Q} \mathcal{R}(h) \le \widehat{\mathcal{R}}_{\mathcal{S}}(Q) + 2\,\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\left[\sqrt{\frac{1}{2\alpha m}\ln\frac{2TK\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}} + \frac{1}{3m\alpha}\ln\frac{2TK\lceil\log_2[\frac{m}{\alpha}]\rceil}{\delta}\right]$$
$$+ \sqrt{\frac{27}{5\alpha m}\widehat{\mathcal{R}}_{\mathcal{S}}(Q)\left[\frac{\|\theta-\theta_P\|_2^2}{2\sigma^2} + \ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}\right]} + \frac{27}{5\alpha m}\left[\frac{\|\theta-\theta_P\|_2^2}{2\sigma^2} + \ln\frac{2TK\lceil\log_2(\frac{m}{\alpha})\rceil}{\delta}\right],$$

where $\mathop{\mathbb{E}}_{h\sim Q}\mathcal{R}(h) := \mathop{\mathbb{E}}_{h\sim Q} \sup_{\rho\in E} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\rho} \ell(y, h(\mathbf{x}))$, *with* $E = \left\{\rho \mid \rho \ll D,\ \text{and}\ \frac{d\rho}{dD} \le \frac{1}{\alpha}\right\}$,

and $\widehat{\mathcal{R}}_{\mathcal{S}}(Q) := \sup_{\rho\in\widehat{E}} \sum_{i=1}^{m} \rho_{\mathrm{A}} \mathop{\mathbb{E}}_{h\sim Q} \ell(y_{\mathrm{A}}, h(\mathbf{x}_{\mathrm{A}}))$, *with* $\widehat{E} = \left\{\rho \mid \forall \mathrm{A}\in\mathcal{A},\ \frac{d\rho_{\mathrm{A}}}{d\pi_{\mathrm{A}}} \le \frac{1}{\alpha}\right\}$ *and* $\pi_{\mathrm{A}} = \frac{1}{m}$.

Proof. The proof follows the same steps as the proof of Corollary 3. □

G.1.4 Additional parameters studied during our experiments

In Appendix H, we present the complete results of our experiments with CVaR, and an additional constrained f-entropic risk measure, EVaR defined by Definition 2 with the function $f(x) = x\ln x$ extended by continuity at $x = 0$ with $f(0) = 0$, and $\beta = -\ln\alpha$.

The different settings we considered are (the rest of the setting follows Section 6):

- Two model architectures: a 2-hidden-layer multilayer perceptron and a perceptron.
- When $|\mathcal{A}| \le m$ (a subgroup corresponds to a class), for Corollary 1:
  - Two reference distributions π: The class ratio, and the uniform distribution,
  - Two risks: CVaR and EVaR.
- When $|\mathcal{A}| = m$ (a subgroup corresponds to a single example), for Theorem 3:
  - One reference distribution: The uniform distribution,
  - Two risks: CVaR and EVaR,
  - Two values of parameter λ: $\lambda = 1$ and $\lambda = m$.
- When $|\mathcal{A}| = m$ (a subgroup corresponds to a single example), for Theorem 1:
  - One reference distribution: The uniform distribution,
  - One risk: CVaR (since Theorem 1 is only defined for CVaR).

G.2 Datasets

We perform our experiments on 19 datasets taken from OpenML (Vanschoren et al., 2013). Their main characteristics are summarized in Table 1.

Table 1: Main characteristics of the datasets (* means that the classes are uniformly distributed).

dataset	n examples	n features	n classes	class ratio
australian	690	14	2	0.56/0.44
balance	625	4	3	0.08/0.46/0.46
german	1,000	20	2	0.3/0.7
heart	270	13	2	0.56/0.44
iono	351	34	2	0.36/0.64
letter	20,000	16	26	0.04*
mammography	11,183	6	2	0.98/0.02
newthyroid	215	5	3	0.7/0.16/0.14
oilspill	937	49	2	0.96/0.04
pageblocks	5473	10	5	0.9/0.06/0.01/0.02/0.02
pendigits	10,992	16	10	0.1*
phishing	11,055	68	2	0.44/0.56
prima	768	8	2	0.65/0.35
satimage	6,430	36	6	0.24/0.11/0.21/0.1/0.11/0.23
segment	2,310	19	7	0.14*
spambase	4,601	57	2	0.61/0.39
spectfheart	267	44	2	0.21/0.79
splice	3,190	287	3	0.24/0.24/0.52
wdbc	569	30	2	0.63/0.37

H RESULTS OF THE ADDITIONAL EXPERIMENTS

In the main paper, we reported the main behaviors we observed on a representative subset of our experiments (on the four most imbalanced datasets). For completeness, the following pages provide all figures for every parameter setting and dataset, as described in Appendices G.1.4 and G.2. Below, we summarize the main trends across all experiments.

Results in a nutshell.

On the role of α. On the one hand, across all bar plots (Figures 6, 7, 10, 11), we observe that α strongly influences the tightness of all the bounds: higher values of α imply tighter bounds. As discussed in Section 6, this is not only due to the factor $\frac{1}{\alpha}$ or $\frac{1}{\alpha^2}$ in the bounds but also because a larger α makes the CVaR tighter. In consequence, the tightest bound values, which always correspond to the highest $\alpha = 0.9$, do not lead to the best-performing models across the subgroups (in terms of F-score or in terms in class-wise error rates).

On the other hand, α also plays an important role on the performance across the subgroups. Indeed, as we can see across all the bar plots, the best F-scores rarely coincide with the tightest bound values (68 times over 76), and Figures 6, 7, 10, and 11 show that the class-wise error rates evolve with α, showing that adjusting α can help to balance the performances across the subgroups.

On the comparison with Mhammedi et al. (2020) (one example per group setting). As expected, when comparing Theorems 1 and 3 (which rely on the same subgroups), our bound of Th. 3 is generally tighter (or very close) for all values of α. Note that we can observe that $\lambda = m$ leads always to bounds that are slightly higher than those of $\lambda = 1$, but although this has a slight impact on the tightness of the bound, it does not change the overall behavior.

On the role of π for Corollary 1. The reference distribution π also plays a role in the tightness of the bound. Except for the most balanced datasets (*australian*, *heart*, *letter*, *pendigits*, *phishing*, *segment*), where using a uniform π or the class ratio π yields similar results as expected, we observe that bounds computed with a uniform π are generally (and sometimes significantly) looser than those computed with π set to the class ratio. Remarkably, for $\alpha \in \{0.01, 0.1, 0.3\}$, Corollary 1 with π set to the class ratio continues to give non-vacuous and

competitive bounds, even when α is relatively high, despite the $\frac{1}{\alpha m_{\mathrm{A}}}$ term in the bound. This suggests that choosing a reference π that reflects the imbalance in the data can lead to better capturing the under-representation in the data while keeping guarantees

On the performances. Interestingly, the bound of Corollary 1 always leads to the best results in terms of F-score on the most imbalanced datasets (*oilspill*, *mammography*, *balance*, *pageblocks*, illustrating the usefulness of our subgroup-based approach. For the 15 more balanced datasets, the bound of Corollary 1 is always competitive, achieving the best performance in 9 cases (for each set of experiments), while the bound of Theorem 3 performs best 6 times.

A note on the EVaR. The results obtained with EVaR are similar to the one observed with the CVaR. This confirms that our bounds can be effectively applied to other constrained f-entropic risk measures beyond the CVaR.

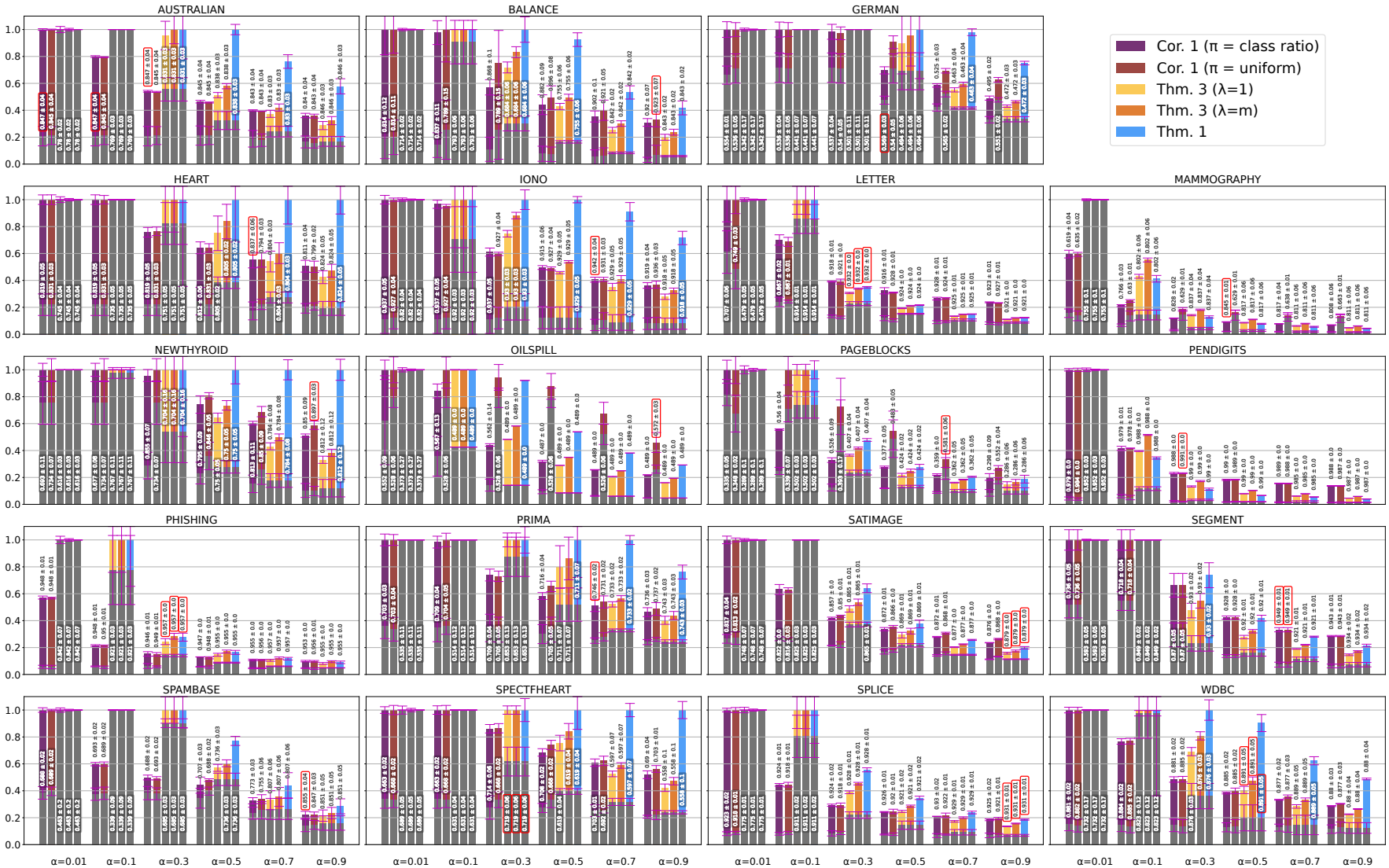


Figure 4: **2-hidden layer MLP with CVaR.** Bound values (in color), test risk $\mathcal{R}_{\mathcal{T}}$ (in grey), and F-score value on $\mathcal{T}$ (with their standard deviations) for Theorem 3, Corollary 1, and Theorem 1, as a function of α (on the x-axis). The y-axis corresponds to the value of the bounds and test risks. The highest F-score for each dataset is emphasized with a red frame.

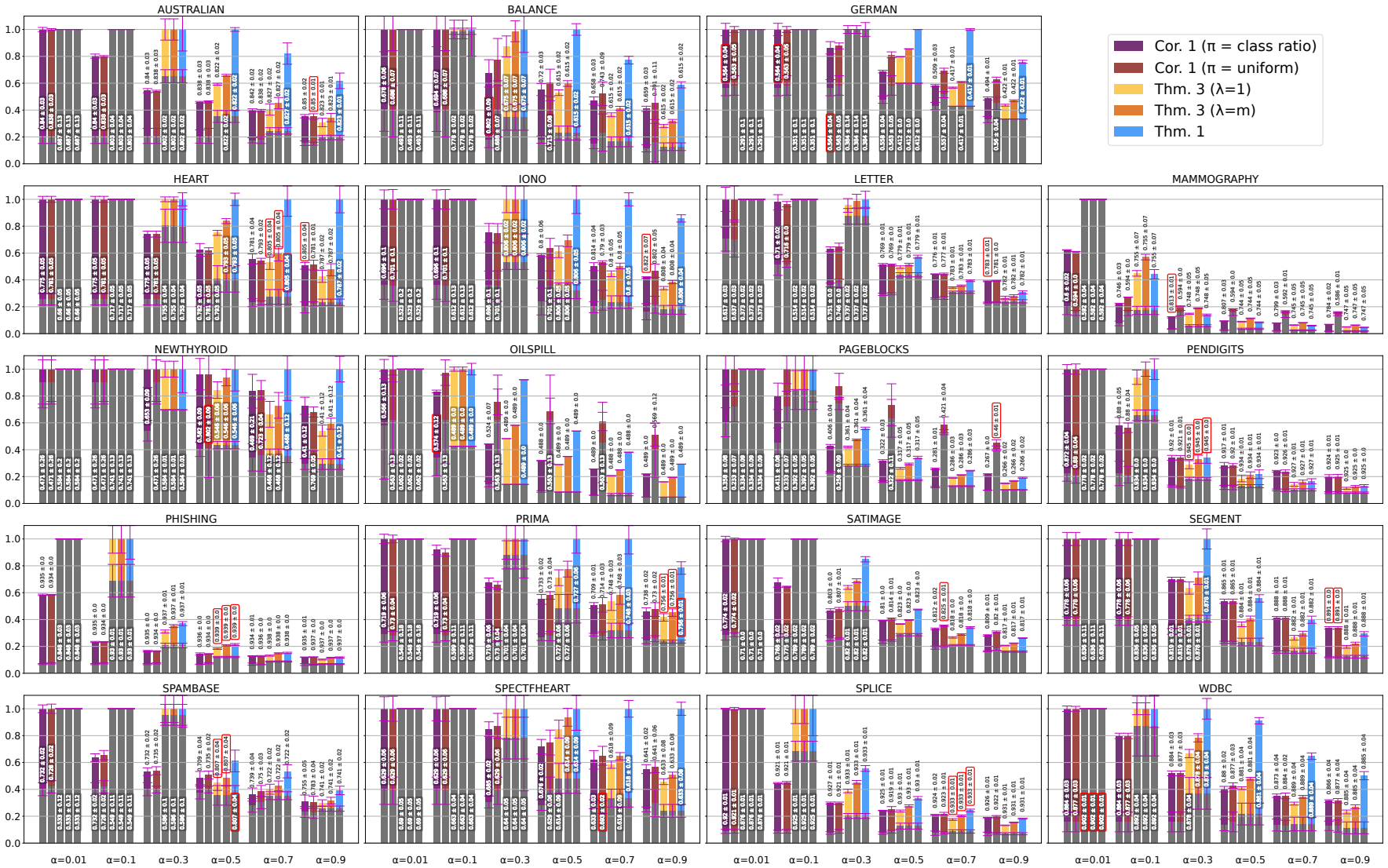
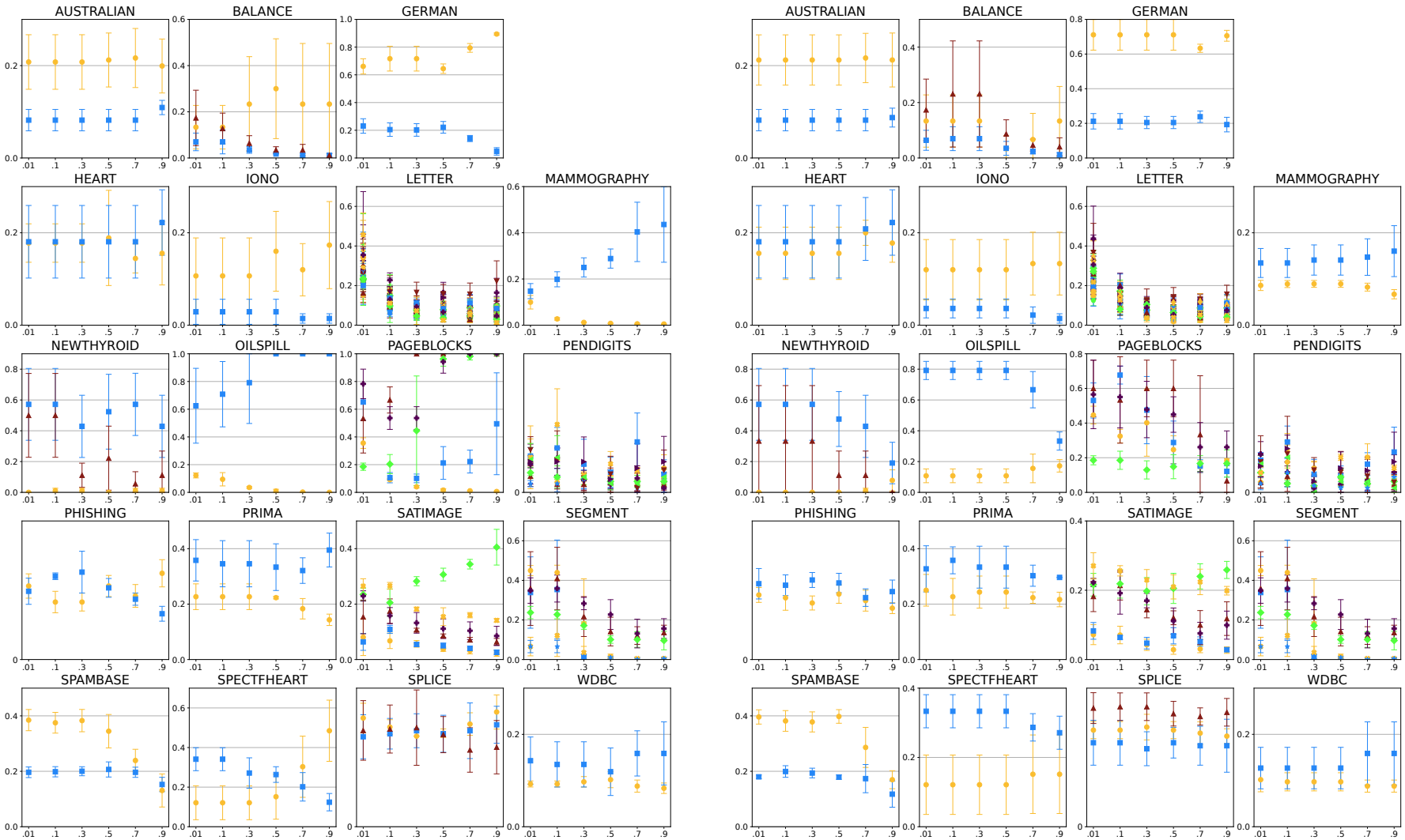


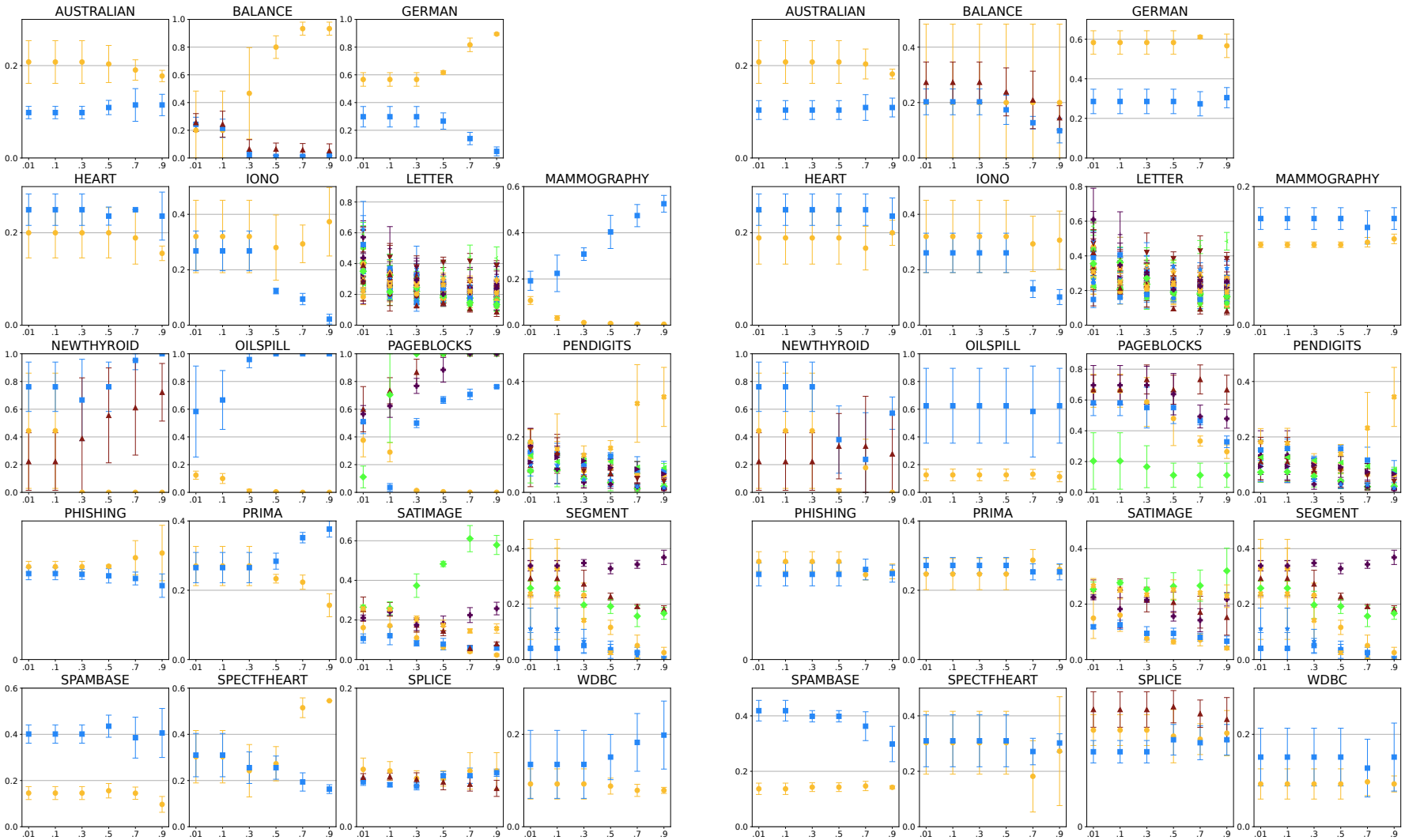
Figure 5: **Perceptron with CVaR.** Bound values (in color), test risk $\mathcal{R}_{\mathcal{T}}$ (in grey), and F-score value on $\mathcal{T}$ (with their standard deviations) for Theorem 3, Corollary 1, and Theorem 1, as a function of α (on the x-axis). The y-axis corresponds to the value of the bounds and test risks. The highest F-score for each dataset is emphasized with a red frame.



(a) π = Class ratio

(b) π = Uniform

Figure 6: **2-hidden layer MLP with CVaR.** Evolution of the class-wise error rates and standard deviation on the set $\mathcal{T}$ (y-axis) as a function of the parameter α (x-axis) with Corollary 1. Each class is represented by different markers and colors.



(a) π = Class ratio

(b) π = Uniform

Figure 7: **Perceptron with CVaR.** Evolution of the class-wise error rates and standard deviation on the set $\mathcal{T}$ (y-axis) as a function of the parameter α (x-axis) with Corollary 1. Each class is represented by different markers and colors.

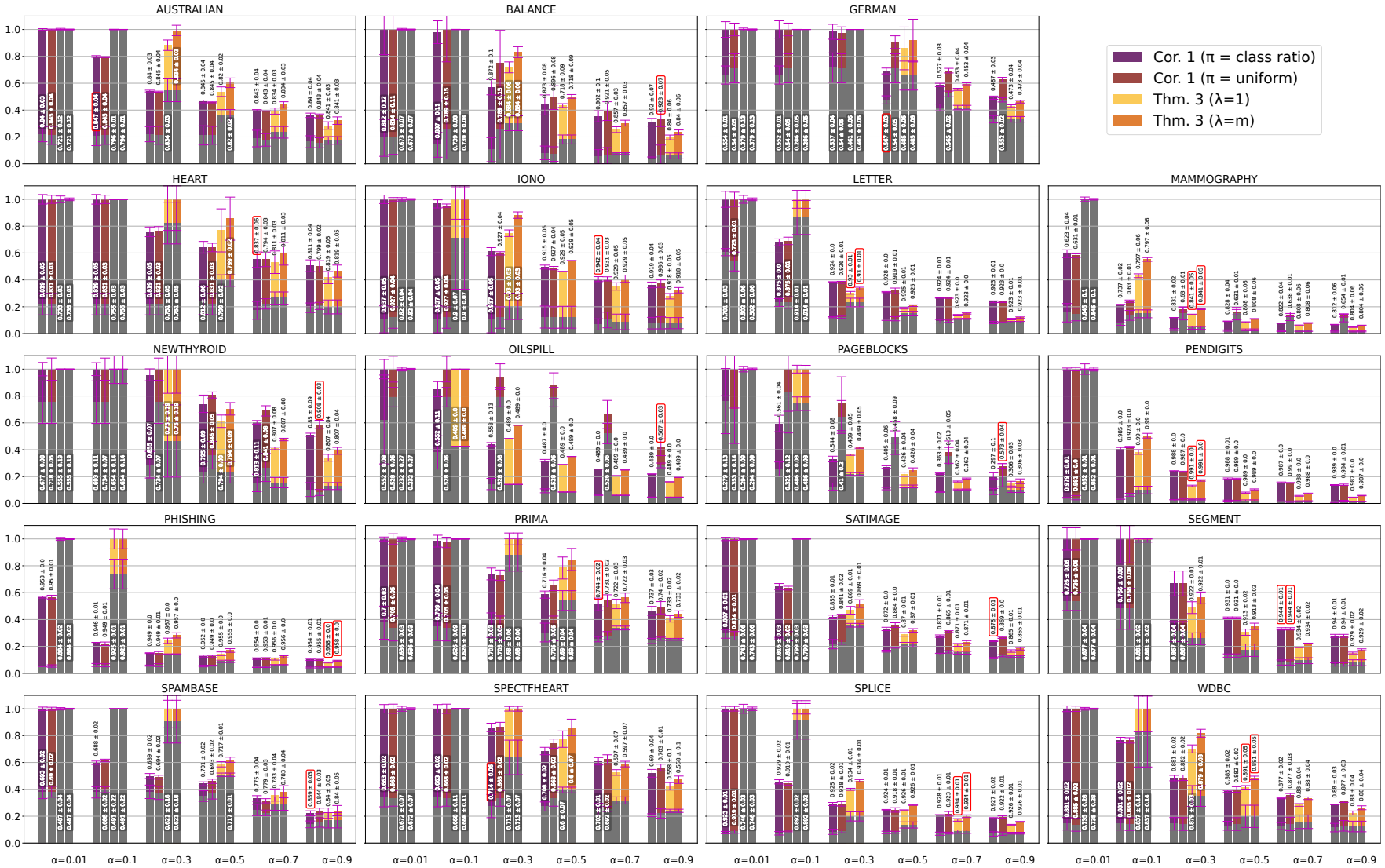


Figure 8: **2-hidden layer MLP with EVaR.** Bound values (in color), test risk $\mathcal{R}_{\mathcal{T}}$ (in grey), and F-score value on $\mathcal{T}$ (with their standard deviations) for Theorem 3, Corollary 1, and Theorem 1, as a function of α (on the x-axis). The y-axis corresponds to the value of the bounds and test risks. The highest F-score for each dataset is emphasized with a red frame.

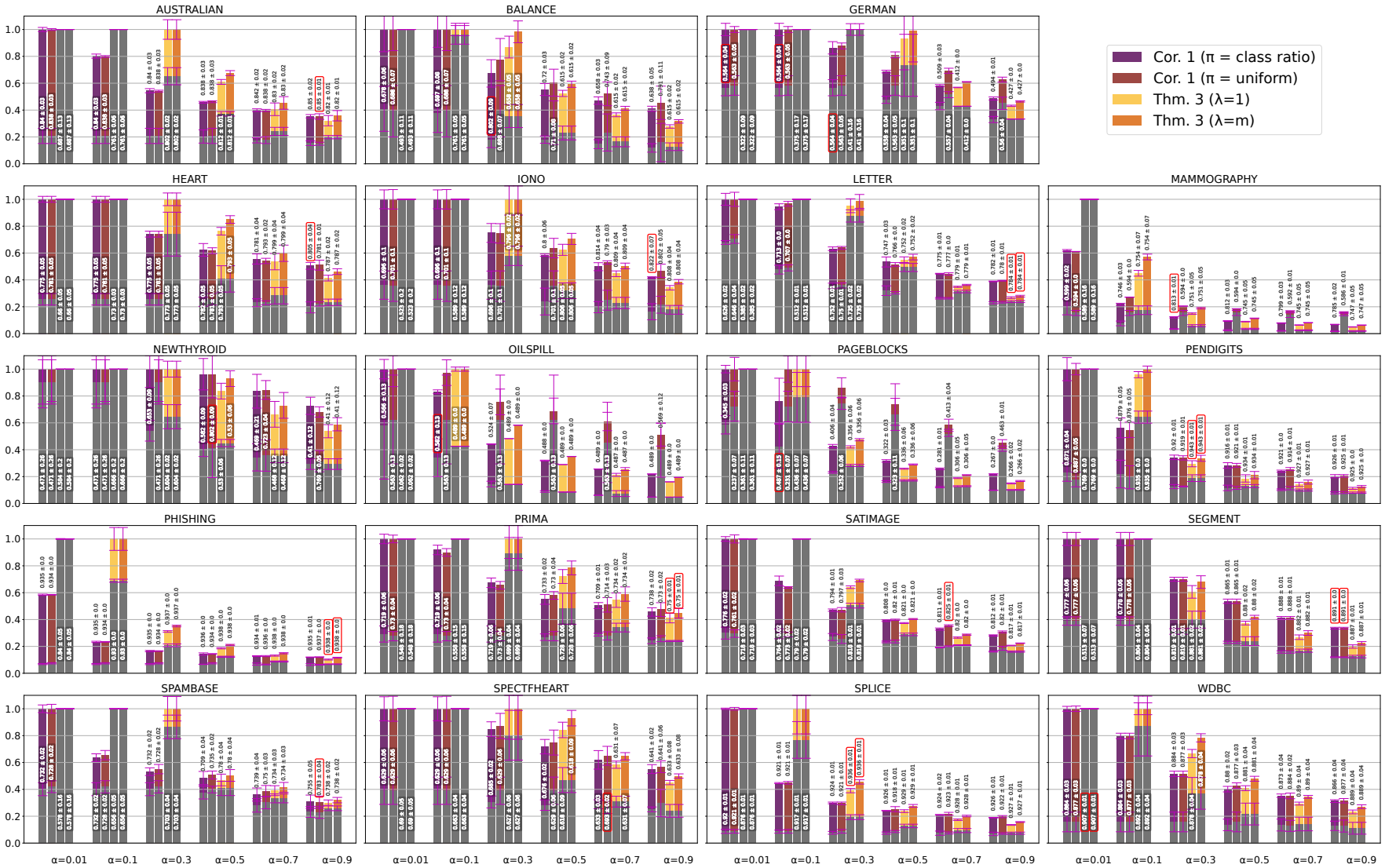
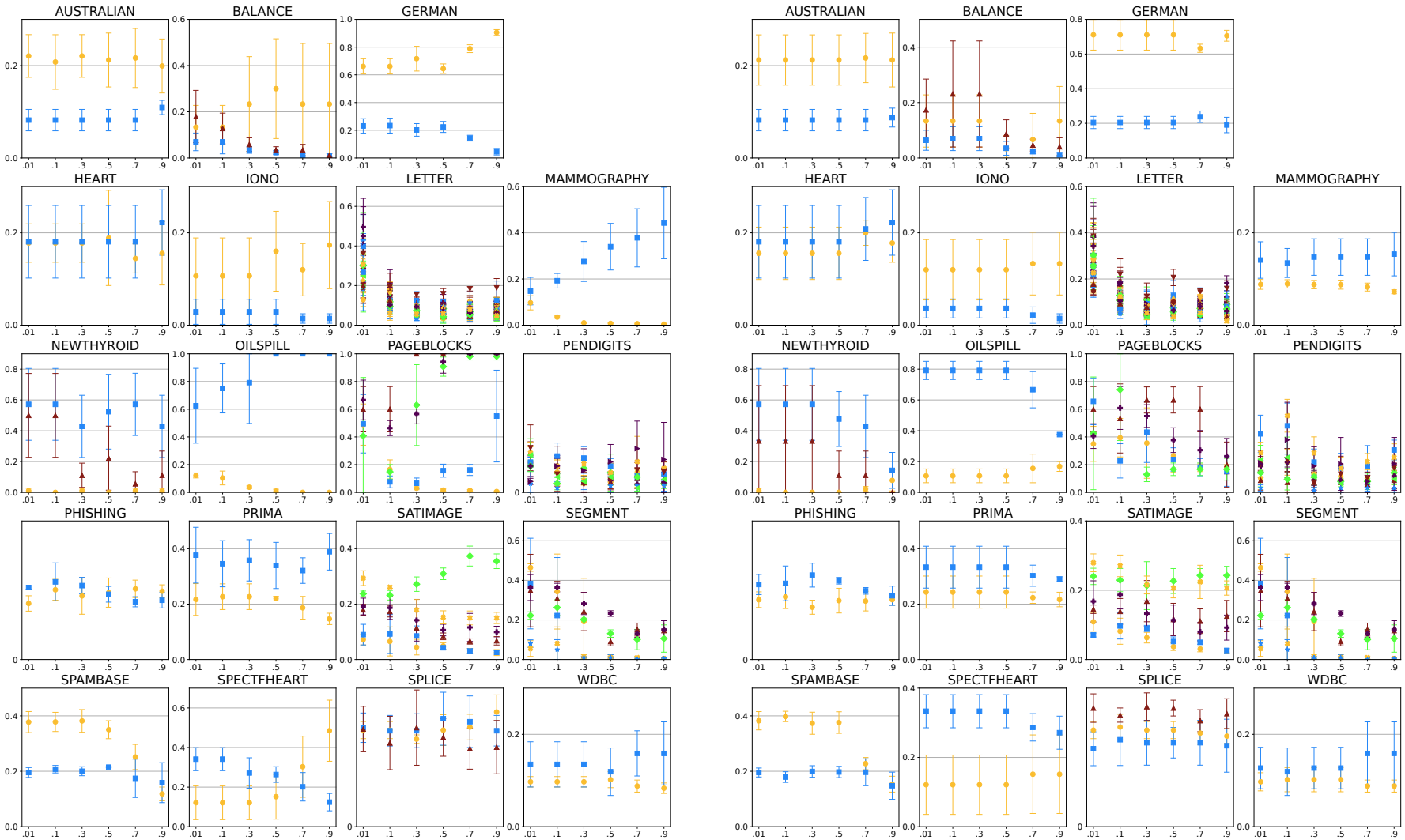


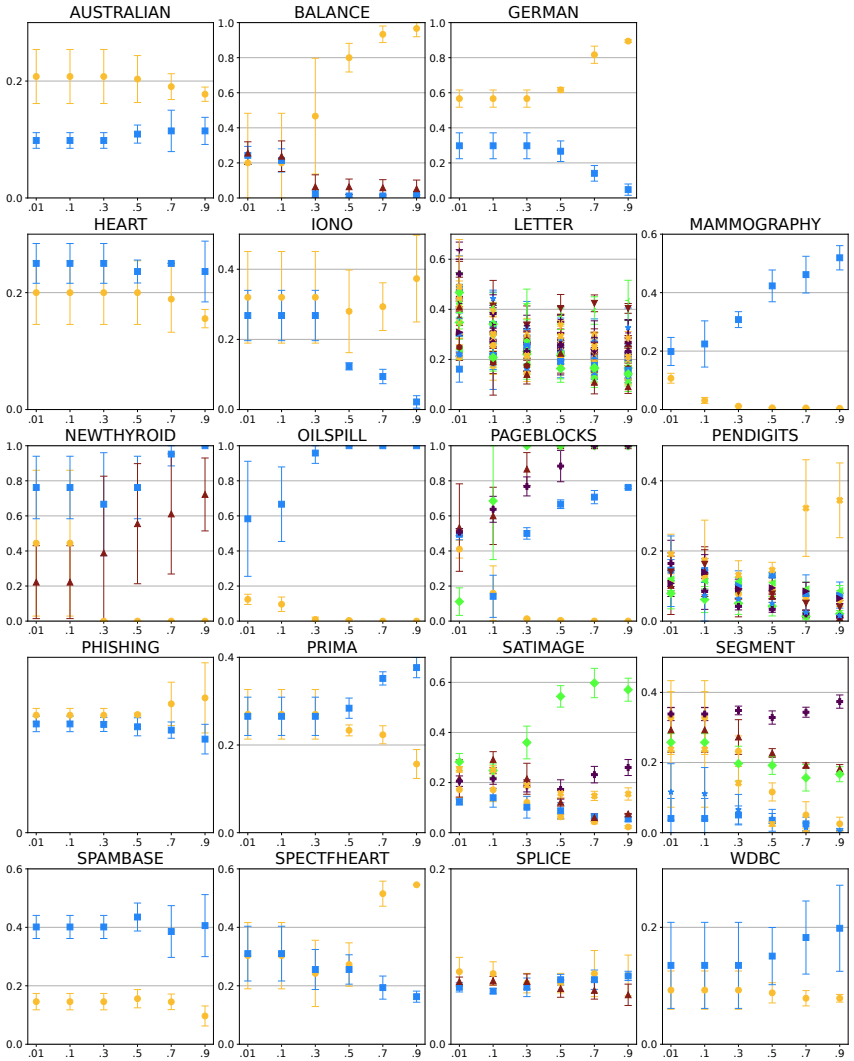
Figure 9: **Perceptron with EVaR.** Bound values (in color), test risk $\mathcal{R}_{\mathcal{T}}$ (in grey), and F-score value on $\mathcal{T}$ (with their standard deviations) for Theorem 3, Corollary 1, and Theorem 1, as a function of α (on the x-axis). The y-axis corresponds to the value of the bounds and test risks. The highest F-score for each dataset is emphasized with a red frame.



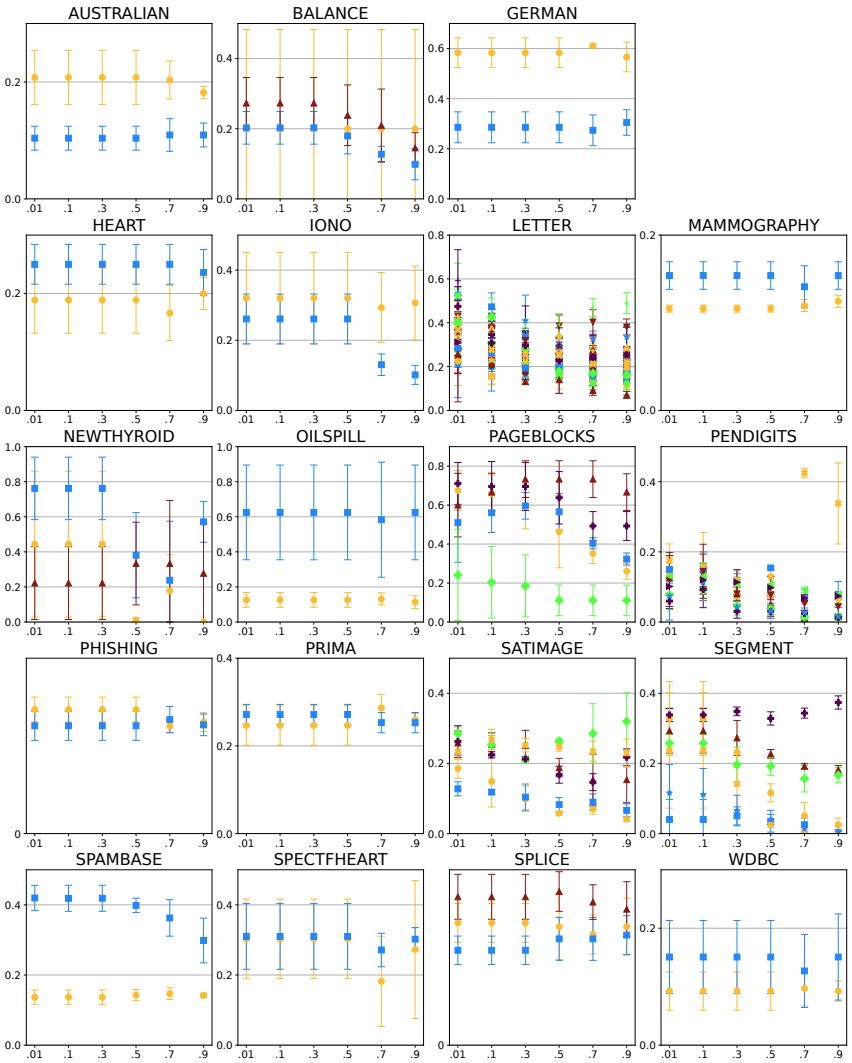
(a) π = class ratio

(b) π = uniform

Figure 10: **2-hidden layer MLP with EVaR.** Evolution of the class-wise error rates and standard deviation on the set $\mathcal{T}$ (y-axis) as a function of the parameter α (x-axis) with Corollary 1. Each class is represented by different markers and colors.



(a) π = class ratio



(b) π = uniform

Figure 11: **Perceptron MLP with EVaR.** Evolution of the class-wise error rates and standard deviation on the set $\mathcal{T}$ (y-axis) as a function of the parameter α (x-axis) with Corollary 1. Each class is represented by different markers and colors.