Fundamental Tradeoffs in Learning with Prior Information

Anirudha Majumdar [1]

Abstract

We seek to understand fundamental tradeoffs between the accuracy of prior information that a learner has on a given problem and its learning performance. We introduce the notion of prioritized risk, which differs from traditional notions of minimax and Bayes risk by allowing us to study such fundamental tradeoffs in settings where reality does not necessarily conform to the learner's prior. We present a general reduction-based approach for extending classical minimax lower-bound techniques in order to lower bound the prioritized risk for statistical estimation problems. We also introduce a novel generalization of Fano's inequality (which may be of independent interest) for lower bounding the prioritized risk in more general settings involving unbounded losses. We illustrate the ability of our framework to provide insights into tradeoffs between prior information and learning performance for problems in estimation, regression, and reinforcement learning.

1. Introduction

We are motivated by the problem of understanding fundamental limits and tradeoffs in learning with prior information: how much prior knowledge does one require in order to learn quickly on a given task? In other words, for a given problem, is there a fundamental limit on the performance of a learner in the absence of sufficient prior information?

Fundamental limits in learning are typically formalized via the classical notions of minimax risk and Bayes risk (Berger, 2013). Consider the standard setting in statistical learning theory where a learner receives a dataset $x_1^n := \{x_i\}_{i=1}^n$ containing n i.i.d. observations of a random variable X. Suppose that the distribution P of X is defined by a parameter $\in$ which is a priori unknown to the learner.

[1]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA. Correspondence to: Anirudha Majumdar <ani.majumdar@princeton.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

The learner's task is to use the dataset to minimize risk (i.e., expected loss) for this unknown distribution. The minimax risk then corresponds to the lowest worst-case risk (over the family of distributions defined by $\in$) achievable by any learner (see Sec. 2 for a formal definition). In other words, the minimax risk corresponds to the minimax-optimal value of the following game: the learner first chooses a particular learning algorithm, and then an adversary chooses a distribution parameter $\in$; the learner's cost (negative payoff) is equal to its risk.

The minimax risk does not allow one to take into account prior information that a learner may have (beyond the weak prior knowledge that belongs to the set). As an alternative, one may assume that the learner is equipped with prior knowledge of the distribution that the parameter is drawn from. The Bayes risk then corresponds to the smallest possible average risk assuming that is drawn from the prior distribution (see Sec. 2 for a formal definition). However, the Bayes risk formulation assumes that the learner is equipped with nature's prior, i.e., the true distribution from which is drawn. The Bayes risk formulation thus does not capture settings where reality does not conform to the learner's prior.

Statement of contributions. We seek to address the challenges with using the minimax and Bayes risk for understanding fundamental limits of learning with prior information. To this end, we make the following contributions:

- We propose a quantity — which we refer to as prioritized risk — for analyzing settings where reality does not fully conform to the learner's prior. The key idea is to consider a version of the minimax risk where the risk associated with a given distribution P is weighted by the prior that the learner has on . We show that this quantity allows us to understand fundamental tradeoffs between the accuracy of prior information and learning performance (risk) on a given problem: a lower bound on the prioritized risk allows one to establish a fundamental limit on learning performance in the absence of sufficient prior information (Sec. 3).
- We provide a general reduction-based strategy for extending classical techniques for lower bounding minimax and Bayes risk — including the methods of LeCam, Assouad, and Fano (Tsybakov, 2008; Yang & Barron, 1999; Yu et al., 1997) — to obtain lower bounds on the

prioritized risk for estimation problems (Sec. 4).

We derive a novel generalized Fano inequality for obtaining lower bounds on the prioritized risk for general learning problems beyond estimation (Sec. 5). This inequality handles problems with unbounded losses (in contrast to prior generalized Fano inequalities (Chen et al., 2016; Gerchinovitz et al., 2020; Majumdar & Pacelli, 2022)), and may thus be of independent interest.

We illustrate the ability of the prioritized risk framework to provide insights into tradeoffs between prior information and learning performance for various problems including prior-informed versions of the following (Sec. 6): (i) Bernoulli mean estimation, (ii) logistic regression, and (iii) reinforcement learning (RL) with environments drawn from Zipfian distributions.

2. Background: Minimax and Bayes risk

We provide a brief introduction to the minimax and Bayes risks, and refer the reader to (Berger, 2013; Tsybakov, 2008) for a more thorough exposition. Consider a random variable X that takes values in a sample space $\mathcal{X}$. Suppose that the distribution P of X is defined by a parameter $\theta \in \Theta$ which is unknown to the learner. A learner $\pi : \mathcal{X}^n \to \mathcal{A}$ receives a dataset $x_1^n := \{x_i\}_{i=1}^n$ of n i.i.d. realizations of X, and must output an action $a \in \mathcal{A}$ that is evaluated according to a loss function $L : \mathcal{A} \to [0, \infty)$. This setting captures many problems of interest; in estimation problems, one can take $\mathcal{A} = \Theta$ and evaluate the learner using a loss $L(\theta, \pi(x_1^n)) = (\theta - \pi(x_1^n))^2$. More broadly, $\mathcal{A}$ can correspond to a hypothesis space $\mathcal{H}$; the learner may then be evaluated by its future expected performance $L(\theta, \pi(x_1^n)) = \mathbb{E}_{X \sim P}[l(X; \pi(x_1^n))]$ on data from P.

Minimax risk. The risk of a learner is defined as

$$R(\theta, \pi) = \mathbb{E}_{X_1^n \sim P^n} L(\theta, \pi(X_1^n)), \tag{1}$$

where the expectation is taken with respect to the dataset used by the learner. We can then define the minimax risk:

$$R_{\text{minimax}}(\Theta, L, \pi) =: \inf_{\theta \in \Theta} \sup R(\theta, \pi). \tag{2}$$

This can be interpreted as a game where the learner fixes π, and an adversary with knowledge of π then chooses $\theta \in \Theta$ in order to maximize the risk. The minimax risk thus corresponds to the lowest worst-case risk (over the family of distributions defined by $\theta \in \Theta$) of any learner.

Bayes risk. Since the minimax risk corresponds to the worst-case risk, it does not capture prior knowledge that a learner may have beyond the fact that $\theta \in \Theta$. One can instead consider the Bayes risk, which is the smallest possible average risk assuming that θ is drawn from a distribution π:

$$R_{\text{Bayes}}(\pi, L, \Theta) = \inf_{\pi} \int R(\theta, \pi)\, \pi(d\theta). \tag{3}$$

A learner that implements Bayesian inference with the prior π and observation x_1^n achieves the optimal Bayes risk (Berger, 2013, Ch. 4). However, the Bayes risk formulation makes a strong assumption on how π is chosen and the knowledge that the learner has, i.e., that the parameter θ is drawn from a "true" distribution π ("nature's prior") and that this distribution is known to the learner. The Bayes risk is thus not directly useful in analyzing settings where reality does not conform to the learner's prior.

3. Prioritized Risk

Motivated by the challenges associated with the notions of minimax and Bayes risk for analyzing learning algorithms with prior information, we propose a different quantity that aims to capture the relationship between prior knowledge and learning performance on a given task. We refer to this quantity as prioritized risk and discuss its interpretation below. Similar to the minimax risk, the prioritized risk operates in a setting where nature chooses a particular value $\theta \in \Theta$ (instead of randomly choosing θ from a distribution). However, similar to the Bayes risk, we allow learners to be equipped with prior information (which may not fully capture the true value of θ).

3.1. Definition

Let $\pi : \Theta \to \mathbb{R}_{>0}$ denote a function that captures prior information that a learner has about the learning problem (i.e., about the value of $\theta \in \Theta$ that the learner will encounter). If π is normalized such that $\int_\Theta \pi(\theta) d\theta = 1$, it may be interpreted as a density corresponding to a Bayesian prior. However, here we eschew this Bayesian interpretation and simply think of π as a way for the learner to encode all available prior knowledge or inductive bias it has on the problem (specified before the learner observes any data). In this sense, π is similar to the "luckiness function" (Shawe-Taylor et al., 1998) or the prior in PAC-Bayes approaches (McAllester, 1999). In cases where π does not integrate to 1, it may be interpreted as an energy-based model (LeCun et al., 2006) which encodes prior information. Our frame-work allows us to handle both normalized and unnormalized priors.

Definition 3.1 (Prioritized risk). For a given family of distributions $\{P_\theta\}_{\theta \in \Theta}$, loss function L, and prior function $\pi : \Theta \to \mathbb{R}_{>0}$, the prioritized risk is defined as:

$$\begin{aligned} R_{\text{prior}}(\pi, L, \Theta) &= \inf_{\pi} \sup_{\theta \in \Theta} \pi(\theta) R(\theta, \pi) \\ &= \inf_{\pi} \sup_{\theta \in \Theta} \pi(\theta) \mathop{\mathbb{E}}_{X^n \sim P_\theta^n} L(\theta, \pi(X^n)). \end{aligned} \tag{4}$$

We will refer to the quantity:

$$R_{\text{prior}}(\pi, L, \Theta) = \sup_{\theta \in \Theta} \pi(\theta) R(\theta, \pi) \tag{5}$$

as the learner-specific prioritized risk.

As we discuss below, the prioritized risk can provide insights along two dimensions: (i) analyzing different learning algorithms, and (ii) analyzing different learning problems.

3.2. Implications for Learning Algorithms

In order to interpret the prioritized risk, consider a learner that achieves a small learner-specific prioritized risk:

$$()R(;) \quad ; \ 8 \ 2 : \tag{6}$$

We will say that a parameter 2 chosen by nature conforms to the learner's prior if () is high; conversely, we will say that does not conform to the learner's prior if () is low. Then, we have the following implication for a learner that satisfies (6): the more closely nature conforms to the learner's prior (i.e., the higher () is for the chosen), the lower the risk is guaranteed to be: $R(;) = ()$.

The prioritized risk also allows us to compare different learning algorithms. For a given learning problem and prior , consider two learners and ^{0} that have learner-specific prioritized risks and ^{0} respectively, with $^0 <$. Then,

$$R(;) = (); \ 8 \ 2 ; \qquad (7) \ R(^0;) \ ^0 = ()$$
$$< = (); \ 8 \ 2 : \tag{8}$$

In such a case, one should prefer the learner ^{0} since it affords a better tradeoff between learning performance (i.e., risk) and the accuracy of prior information (i.e., how much conforms to the prior).

3.3. Implications for Learning Problems

In this work, we will focus on lower bounds for the prioritized risk. Consider a problem where one has established:

$$R_{prior}(; L;) \ : \tag{9}$$

This implies[1] that no matter what learning algorithm one chooses (i.e., for any choice of),

$$9 \ 2 \ \text{such that} \ () \ \underbrace{R(;)}_{\text{Learning}} \ : \tag{10}$$
("Nature") ("Nurture")

This relationship between prior information ("nature") and learning ("nurture") takes the form of an uncertainty principle[2] and captures a fundamental tradeoff: it is impossible for both the risk and the prior to be low for all 2 . In other words, for any learning algorithm , there exists 2 such that if the learner achieves low risk, it must be the case that reality conforms to the learner's prior (i.e., () is large).

[1] Here we assume for simplicity that the supremum and infimum are achieved (i.e., sup = max, inf = min).

[2] Recall the form of uncertainty principles in quantum mechanics, e.g., $xp \ h=4$, which states that there is a fundamental tradeoff between knowing a particle's position and its momentum.

Example 3.3. As an example of the kind of implications one may derive from a lower bound (9) on the prioritized risk, consider a learning problem with an associated prior (with () 2 [0; 1]; 8 2). Consider a learner that achieves low risk for values of that have a high prior:

$$R(;) \ < \ ; \ 8 \text{ s.t. } _2 () \ 1:\frac{1}{} \tag{11}$$

Then, from (10), we see that there must exist a with low prior where the learner performs poorly:

$$9 \text{ s.t. } () < \ _2^1 ; \text{ where } R(;) \ _{()} > 2: \frac{}{} \tag{12}$$

A lower bound on the prioritized risk thus establishes a fundamental tradeoff for a given learning problem: any learner that performs extremely well for values of with high () must give up performance for a low value of (); the learner may thus perform poorly if reality does not conform to its prior (i.e., if () is low).

Remark 3.2 (Relationship to minimax and Bayes risk). We make the following straightforward observations relating the prioritized risk to the minimax and Bayes risks:

- The prioritized risk reduces to the minimax risk if () 1. If () 1 8, then $R_{prior} \ R_{minimax}$.
- In settings where is a countable set, the prioritized risk lower bounds the Bayes risk (both computed using a given prior that is normalized to be a valid probability distribution). This is because for any learner : $\sup_2 \ ()R(;) \ _{P \ 2} \ ()R(;)$. Since the Bayes risk lower bounds the minimax risk, we have for countable that: $R_{prior} \ R_{Bayes} \ R_{minimax}$.
- In the more general setting of uncountable , the prioritized risk may be larger than the Bayes risk (e.g., take to be the density function of a univariate Gaussian such that () > 1 for some , and let $R(;) \ 1$).

4. Lower Bounds on Prioritized Risk: Estimation Problems

We now describe techniques for obtaining lower bounds on the prioritized risk. In this section, we focus on estimation problems; here, a learner : X^n ! receives a dataset $x_1 = ^n f x_i g_{i=1}$ of n i.i.d. realizations from a distribution P and outputs an estimate = (x^n) of the underlying parameter . The learner is evaluated using a loss:

$$L(; (x_1))_n = \ ; (x_1); \ _n \tag{13}$$

where : ! R_+ is a (pseudo)metric, e.g., $k \ k_2$.

4.1. Reduction from Estimation to Testing

In this section, we describe a general strategy for extending classical techniques for lower bounding the minimax risk

(e.g., the methods of LeCam, Assouad, and Fano) in order to lower bound the prioritized risk for estimation problems. The standard starting point for proving lower bounds on the minimax risk is to reduce the problem of estimation to one of hypothesis testing (see, e.g., (Tsybakov, 2008) (Duchi, 2016, Ch. 7)). One can then use information-theoretic techniques to lower bound the Bayes risk for the testing problem, which yields a lower bound for the minimax risk. Here, we extend this classical reduction from estimation to testing in order to lower bound the prioritized risk.

Consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$, where $\mathcal{V}$ is a finite index set. We will refer to this set as a (α, τ)-packing if balls (as defined by the pseudometric ρ) of radius $\tau(v)/2$ centered around each v are non-overlapping. Now consider the following hypothesis testing problem. First, we define a random variable V corresponding to a uniform distribution over $\mathcal{V}$. Conditioned on a choice $V = v$, a dataset $X_1^n := \{X_i\}_{i=1}^n$ is drawn from P_v^n. Given this dataset, the hypothesis testing problem is to determine the underlying index $v \in \mathcal{V}$. A mapping $\psi : \mathcal{X}^n \to \mathcal{V}$ is referred to as a test function. Its associated probability of error is:

$$P(\psi(X_1^n) \neq V) := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v(\psi(X_1^n) \neq v | V = v).$$

We now establish the reduction from estimation (where a learner is evaluated according to the learner-specific prioritized risk) to hypothesis testing via the following argument:

1. Suppose there exists a learner $\hat{\theta} : \mathcal{X}^n \to \Theta$ with a low learner-specific prioritized risk $R_{\text{prior}}(\hat{\theta}; L; \pi)$ (Eq. (5)) for the original estimation problem.

2. Then, the following prior-weighted test function (which uses $\hat{\theta}$ as a subroutine) achieves a small error probability for the hypothesis testing problem:

$$\psi(x_1^n) := \operatorname*{argmin}_{v \in \mathcal{V}} \pi(v) \rho(v, \hat{\theta}(x^n)). \quad (14)$$

This reduction (stated formally below) allows us to turn a learner that achieves low prioritized risk (for the estimation problem) to a test function that achieves a small probability of error (for the hypothesis testing problem). The contrapositive then allows us to translate lower bounds on the testing problem (which can be obtained using standard information-theoretic techniques) to lower bounds on the estimation problem. Specifically, suppose we have a lower bound on the achievable probability of error for the testing problem. This lower bounds the probability of error for the prior-weighted test function (14) (since this is just a particular test function). The reduction above then provides a lower bound on the prioritized risk for the estimation problem.

Proposition 4.1 (Reduction from estimation to testing). Let $\hat{\theta} : \mathcal{X}^n \to \Theta$ be an estimator for a learning problem defined by a loss function of the form (13). Let $\{P_v\}_{v \in \mathcal{V}}$ form a (α, τ)-packing and define the prior-weighted test function as in (14). We then have the following bound:

$$R_{\text{prior}}(\hat{\theta}; L; \pi) \geq P(\psi(X_1^n) \neq V). \quad (15)$$

Hence, taking an infimum over estimators, we have:

$$R_{\text{prior}}(\hat{\theta}; L; \pi) \geq \inf_{\psi} P(\psi(X_1^n) \neq V). \quad (16)$$

Proof. The proof is presented in Appendix A. The primary distinctions of this reduction as compared to the standard reduction from estimation to testing are: (i) the use of the non-uniform (α, τ)-packing, and (ii) the use of the prior-weighted test function (14). If $\pi(\cdot) \equiv 1$, the reduction presented here reduces to the standard reduction. □

As we demonstrate below, this reduction allows us to extend classical techniques for lower bounding the minimax risk (Tsybakov, 2008; Yang & Barron, 1999; Yu et al., 1997) in order to lower bound the prioritized risk. Proofs of the results below are deferred to Appendix A.

4.2. LeCam's Method for Prioritized Risk

We first consider an extension of LeCam's method in order to lower bound the prioritized risk. This technique operates by constructing a packing $\{P_v\}_{v \in \mathcal{V}}$, where $\mathcal{V} = \{0, 1\}$, and utilizing lower bounds for binary hypothesis testing.

Theorem 4.2 (LeCam's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P_\theta\}_{\theta \in \Theta}$ with a loss function of the form (13). A learner has prior π and will receive n i.i.d. observations from a chosen distribution. Let $\{\theta_0, \theta_1\}$ form a (α, τ)-packing. We then have the following lower bound on the prioritized risk:

$$R_{\text{prior}}(\hat{\theta}; L; \pi) \geq \frac{\alpha(\tau)}{2}\left[1 - \|P_0^n - P_1^n\|_{TV}\right]. \quad (17)$$

In order to understand the dependence of R_{prior} on the number of samples n, we can find a $(\alpha(n), \tau)$-packing as a function of n such that the total variation distance is bounded:

$$\|P_0^n - P_1^n\|_{TV} \leq 1 - \frac{\alpha(n)}{\tau(n)}. \quad (18)$$

Theorem 4.2 then establishes that $R_{\text{prior}} \geq \tau(n)/2$.

4.3. Fano's Method for Prioritized Risk

Next, we extend Fano's method in order to lower bound the prioritized risk; this operates by lower bounding the testing error (16) via Fano's inequality (Cover & Thomas, 2012).

Theorem 4.3 (Fano's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P_\theta\}_{\theta \in \Theta}$ with a loss function of the form (13). A learner has prior π and will receive n i.i.d. observations from a chosen distribution. Let $\{P_v\}_{v \in \mathcal{V}}$ form a (α, τ)-packing. Define a

random variable V corresponding to a uniform distribution over $\mathcal{V}$. We can then lower bound the prioritized risk:

$$R_{\text{prior}}(\pi; L; \pi) \geq 1 - \left[\frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|} \right], \tag{19}$$

where $I(V; X_1^n)$ denotes the mutual information.

4.4. Assouad's Method for Prioritized Risk

We now extend Assouad's method in order to lower bound R_{prior}; instead of reducing the problem of estimation to a single hypothesis testing problem as in Prop. 4.1, Assouad's method proceeds via a reduction to multiple binary hypothesis testing problems. We first extend the notion of Hamming separation (Duchi, 2016, Ch. 7) to incorporate the prior π.

Definition 4.4 ($(2\tau, \alpha)$-Hamming separation). Consider a family of distributions given by $\{P_v\}_{v \in \mathcal{V}}$ indexed by the hypercube $\mathcal{V} = \{-1, 1\}^d$ (for some $d \in \mathbb{N}$). This family induces a $(2\tau, \alpha)$-Hamming separation if there exists $\psi : \mathcal{A} \to \{-1, 1\}^d$ such that $\forall v \in \mathcal{V}$, we have:

$$\ell(v; \cdot) \geq \tau \Phi(\alpha) \left(\sum_{i=1}^d \mathbb{1}\left\{ [\psi(\cdot)]_j = v_j \right\} \right), \quad \forall \cdot \in \mathcal{A}. \tag{20}$$

Let P_j denote the joint distribution over the (uniformly chosen) random index V and data X_1^n conditioned on the j-th coordinate $V_j = 1$.

Theorem 4.5 (Assouad's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P_\theta\}_{\theta \in \Theta}$ with a loss function of the form (13). A learner has prior π and will receive n i.i.d. observations from a chosen distribution. Let $\{P_v\}_{v \in \mathcal{V} = \{-1,1\}^d}$ form a $(2\tau, \alpha)$-Hamming separation with ψ. We then have:

$$R_{\text{prior}}(\pi; L; \pi) \geq \sum_{j=1}^d \inf_{\phi} \left[P_{+j}(\phi(X_1^n) = +1) + \cdots + P_{-j}(\phi(X_1^n) = -1) \right],$$

where the infimum is over tests $\phi : \mathcal{X}^n \to \{+1, -1\}$.

Combining this with the variational representation of the total variation distance, we see:

$$R_{\text{prior}}(\pi; L; \pi) \geq \sum_{j=1}^d \left[1 - \left\| P_{+j} - P_{-j} \right\|_{TV}^n \right], \tag{21}$$

where $P_{+j} := 2^{1-d} \sum_{v : v_j = 1} P_v$ (and similarly for P_{-j}). Thus, similar to LeCam's method, we can obtain lower bounds on the prioritized risk by finding an appropriate packing (forming a $(2\tau, \alpha)$-Hamming separation) and lower bounding the total variation distances in (21).

5. Generalized Fano Inequality for Lower Bounds on Prioritized Risk

We now describe techniques for lower bounding the prioritized risk for learning problems beyond estimation. We consider the general setting described in Sec. 2, where a learner $\pi : \mathcal{X}^n \to \mathcal{A}$ receives a dataset $x_1^n =: \{x_i\}_{i=1}^n$ of n i.i.d. realizations of X, and must output an action $a \in \mathcal{A}$ (e.g., a hypothesis) that is evaluated according to a loss function $L : \mathcal{A} \to [0, \infty)$.

The key technical tool we use to obtain lower bounds on prioritized risk in this setting is a generalized version of Fano's inequality. In its original form, Fano's inequality (Cover & Thomas, 2012) provides lower bounds on the achievable error of estimating a signal given a potentially corrupted observation of the signal. Recent generalized Fano inequalities (Chen et al., 2016; Gerchinovitz et al., 2020; Majumdar & Pacelli, 2022) allow one to establish lower bounds on Bayes and minimax risks for various learning problems beyond estimation. Here, we present a novel generalization of Fano's inequality, which allows us to handle learning problems with unbounded loss functions (in contrast to Chen et al. (2016); Gerchinovitz et al. (2020); Majumdar & Pacelli (2022), which assume that the loss is bounded within $[0, 1]$). This is particularly important in our setting for lower bounding the prioritized risk (4), where the product $\ell(\cdot)R(\cdot; \cdot)$ may not be bounded (even if the loss function L is bounded). We first present a general version of our result — which may be of independent interest — for lower bounding the Bayes risk (3), and then use this to lower bound the prioritized risk.

Theorem 5.1 (Generalized Fano inequality for unbounded losses). For any prior distribution p on $\mathcal{A}$, the following lower bound on the Bayes risk holds for all $\gamma > 0$:

$$R_{\text{Bayes}}(p; L; \cdot) =: \inf_{\cdot} \int_{\mathcal{X}^n} \mathbb{E}_{P_1^n} \left[L(\cdot; \cdot(X^n)) \right] p(d) \geq \Gamma_{\gamma, L}^{-1} \left[\frac{I(\cdot; X_1^n) + \log 2}{n} \right], \tag{22}$$

where $I(\cdot; X^n)$ is the mutual information, and

$$\Gamma_{\gamma, L}^{-1} =: \sup_a \log \int \exp\left(-\gamma L(\cdot; a) \right) p(d). \tag{23}$$

Proof. The proof is in App. A. In contrast to (Chen et al., 2016; Gerchinovitz et al., 2020; Majumdar & Pacelli, 2022), we use the Donsker-Varadhan change of measure inequality (Gray, 2011, Thm. 2.3.2) to handle unbounded losses. □

Consistent with intuition, this bound increases when the data X_1^n do not provide much information about the underlying (i.e., when the mutual information is small). The term $\Gamma_{\gamma, L}^{-1}$ is related to the best achievable (exponentiated) average loss when one simply chooses an action a without observing any data. This term may be easily computed when $\mathcal{A}$ and

are finite; here, the supremum over a and the expectation can both be computed exactly. In addition, one may also compute a bound when A is finite, but is not (by computing or bounding the expectation above with high probability via sampling and applying a concentration inequality; the supremum over a can then be computed by enumeration). We also note that $?_{;L}$ is similar to the "small-ball probability" that appears in previous Fano inequalities Chen et al. (2016); Gerchinovitz et al. (2020). While there is no general recipe for bounding the small-ball probability, this can be achieved on a case-by-case basis for particular examples (see Chen et al. (2016)). The quantity $?_{;L}$ similarly needs to be computed on a case-by-case basis in general.

Corollary 5.2 (Prioritized risk lower bound via generalized Fano inequality). For any distribution p on , we have the following lower bound for all > 0:

$$R_{prior}(;L;) \quad _{;L} \quad \frac{1}{} \left[I(;?X_1) \; ; \qquad \right]_{n} \quad (24)$$

where $I(; X^n_1)$ is the mutual information (computed using p for), and

$$_{;L}? = \; : \sup_a \log \int \exp \; ()L(;a) \; p(d):$$

Proof. The result follows directly by combining Thm. 5.1 with the following (which follows from the fact that a supremum is lower bounded by an average): $R_{prior}(;L;) \; R_{Bayes}(p;L \; ;)$, where $L \; (;a) = ()L(;a)$. □

We note that the bound above holds for any choice of p and ; thus, we are free to choose p and judiciously in order to obtain a lower bound on the prioritized risk.

6. Examples

We illustrate the ability of the prioritized risk framework to provide insights into tradeoffs between prior information and learning performance for prior-informed versions of: (i) Bernoulli mean estimation, (ii) logistic regression, and (iii) RL with environments drawn from Zipfian distributions.

6.1. Bernoulli Mean Estimation with Priors

We start by considering the problem of estimating the mean of a Bernoulli distribution in the presence of prior information. Suppose we have a family of Bernoulli distributions P over $X \; 2 \; f0;1g$; the distributions are parameterized using the mean $2 \; = \; [0;1]$. Suppose a learner has a prior over . The learner's goal is to estimate given data from P. The learner $: X^n \; ! \;$ outputs an estimate $=^(x^n)$ and is evaluated using the loss $L(;) \; ^{=} \; : \; j \; j. \; ^$

In order to obtain numerical results, we choose a prior of a particular form. Let correspond to the density function of a Beta distribution Beta(= 1, = 2); this prior assigns higher weight to low values of the mean (specifically, is linear in with (0) = 2 and (1) = 0). We can then apply LeCam's method for lower bounding the prioritized risk (Thm. 4.2). By applying Pinsker's inequality and the tensorization property of the KL divergence, we obtain:

$$R_{prior}(;L;) \quad 1 \quad \frac{}{2} \quad \left[\frac{nKL(P_0 \, kP_1)}{2} \right]; \qquad (25)$$

where $f_0;_1g$ forms a (;)-packing. Since the KL divergence can be computed analytically for Bernoulli distributions, we can then find a value of that maximizes this bound for each n.

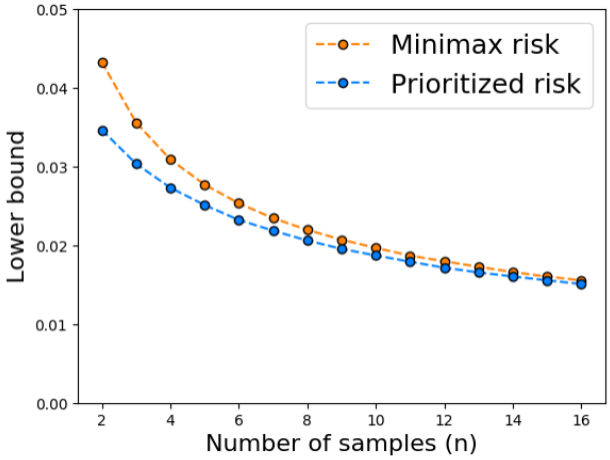

Figure 1. Lower bounds on $R_{prior}(;L;)$ and $R_{minimax}(L;)$ for Bernoulli mean estimation.

Fig. 1 plots the lower bound on the prioritized risk R_{prior} as a function of the number of samples n. As described in Sec. 3.3, this establishes a fundamental tradeoff for learners for this problem (for each value of n). We also compare the lower bound for the non-uniform prior Beta(= 1, = 2) with the lower bound for a uniform prior ((() = 1;8 2 [0;1]), which corresponds to the minimax risk.

In Appendix B, we present upper bounds on prioritized risk for this problem computed using (i) Bayesian inference with prior and (ii) a learner that does not exploit the prior (Bayesian inference with uniform prior). As expected, Bayesian inference with achieves a lower learner-specific prioritized risk. However, we also demonstrate that Bayesian inference with is not optimal in general from the perspective of the prioritized risk; we do this by constructing a different learner that achieves a lower learner-specific prioritized risk. This thus motivates the search for learning algorithms that achieve optimal prioritized risk.

6.2. Logistic Regression with Directional Priors

We now consider the problem of logistic regression with priors, and utilize Assouad's method to obtain lower bounds on the prioritized risk. Given a fixed set of regressors $fz_ig_{i=1}^n$,

where $z_i \in \mathbb{R}^d$, the logistic regression model assigns the following probability to the label $y \in \{-1, 1\}$:

$$P(Y_i = y | z_i; \theta) = \frac{1}{1 + \exp(-y z_i^T \theta)}. \tag{26}$$

The goal of the learner is to infer the unknown parameter $\theta \in \mathbb{R}^d$ given n observations of labels $\{y_i\}_{i=1}^n$. Here, we use the L_1-error $L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_1$ to evaluate the learner.

The learner has a prior π on the parameter θ. Let $V = \{-1, 1\}^d$ denote the vertices of the hybercube, and suppose that π is normalized such that $\pi(v_j) + \pi(-v_j) = 1$. Define:

$$\gamma_j = \frac{1}{4}\left[\frac{1}{\pi(v_j)} + \frac{1}{\pi(-v_j)}\right]. \tag{27}$$

This quantity captures directional biases imposed by the prior. Specifically, given the logistic model (26), we see that if $\pi(v_j)$ is higher than $\pi(-v_j)$, this implies confidence that the j-th component of z being positive will lead to the label y being positive (and that the j-th component of z being negative will lead to the label y being negative). Such asymmetries are captured by γ. Specifically, if the prior is symmetric such that $\pi(v_j) = \pi(-v_j) = 0.5$, then $\gamma_j = 1$. If the prior is asymmetric, then $\gamma_j > 1$ (with the distance from 1 capturing the degree of asymmetry).

Defining $[\psi(\theta)]_j = \text{sign}(\theta_j)$, we see that V forms a (γ, γ)-Hamming separation (Defn. 4.4). We can thus apply Assouad's method (Thm. 4.5) to lower bound the prioritized risk (see Appendix A for the proof):

$$R_{\text{prior}}(\gamma; L; \pi) \geq \frac{1}{16}\left(\frac{d}{\sum_{j=1}^d \gamma_j}\right)^2 \frac{d}{\sum_{i=1}^n \sum_{j=1}^d z_{ij}^2}. \tag{28}$$

As a special case, if $\gamma_j = \gamma (\forall j)$, we obtain:

$$R_{\text{prior}}(\gamma; L; \pi) \geq \frac{1}{16}\frac{1}{\gamma^2}\frac{d^2}{\|Z\|_{Fr}^2}, \tag{29}$$

where $\|Z\|_{Fr}$ is the Frobenius norm of the matrix $Z \in \mathbb{R}^{d \times n}$ consisting of the regressors z_i.

We observe that for a fixed dimension d, the bound (29) on R_{prior} is determined by the product between γ and $\|Z\|_{Fr}$ (and, similarly for (28), the bound depends on the products $\gamma_j z_{ij}$). Thus, for a given prior π, we see that regressors with a smaller L_2-norm induce a more stringent tradeoff between prior information and risk of a learner. As a concrete example, we build on the analysis in Example 3.3. Consider two learning problems corresponding to two sets of regressors with $\|Z\|_{Fr} \gg \|Z^0\|_{Fr}$. Suppose we have a learner θ^0 that achieves low risk for high values of the prior using Z:

$$R(\theta^0; \gamma; Z^0) < \theta^0 : \gamma = \frac{1}{16}\frac{\theta^0 d^2}{\|Z\|_{Fr}^2}, \forall \text{ s.t. } \gamma \geq (\theta^0)^{-1} > 1.$$

Now, suppose we have a learner θ that achieves the same level of performance as θ^0 for high values of the prior using regressors Z:

$$R(\theta; \gamma; Z) < \theta^0 = \frac{1}{16}\frac{\theta^0 d^2}{\|Z\|_{Fr}^2}, \forall \text{ s.t. } \gamma \geq (\theta^0)^{-1} > 1$$
$$\geq \frac{1}{16}\frac{\theta^0 d^2}{\|Z\|_{Fr}^2}. \tag{30}$$

Then, building on Example 3.3, we see that for the problem with regressors Z^0,

$$\exists \text{ s.t. } \gamma < \gamma_2, \text{ where } R(\theta^0; \gamma; Z^0) > \frac{1}{8}\frac{1}{\|Z\|_{Fr}^2}\frac{d^2}{c}. \tag{31}$$

whereas using regressors Z, we have:

$$\exists \text{ s.t. } \gamma < \gamma_2, \text{ where } R(\theta; \gamma; Z) > \frac{1}{8}\frac{1}{\|Z\|_{Fr}^2}\frac{d^2}{(32)^2}$$
$$\geq \frac{1}{8}\frac{1}{\|Z^0\|_{Fr}^2}\frac{d^2}{(32)^2}$$

The bounds on prioritized risk thus suggest that in order for θ to achieve the same level of performance (using Z) as θ^0 (using Z^0) for high values of the prior, θ must sacrifice a greater level of performance for low prior values.

6.3. Reinforcement Learning in Zipfian Environments

In our final example, we consider a reinforcement learning (RL) setting where an agent interacts with environments that are drawn from a Zipfian (i.e., discrete power law) distribution. Specifically, we build on the Zipfian Gridworld environments from Chan et al. (2022), where an agent must find objects using visual feedback. Each environment (Fig. 2 left) consists of a grid-world with four rooms containing 20 objects. There are a fixed set of 400 environments that the agent may be deployed in; the agent's start location, target object, as well as the other object shapes, colors, and locations are fixed within each environment. In each episode, the agent receives a top-down camera view of its immediate surroundings along with a visual depiction of the target object. The agent receives a loss of 0 for the episode if it reaches the target; the episode ends if the agent touches any other object, and a loss equal to the number of steps taken is then assigned.

Let $X = \{x\}_{x=1}^{400}$ denote the set of possible environments. The distribution P over environments is defined by a discrete power law (Zipfian distribution) $p(x) \propto \frac{1}{x^\gamma}$, where the exponent γ determines how skewed the distribution is. Such a distribution captures the heavy-tailed nature of many real-world environments (Chan et al., 2022). In our setting, γ is chosen from a set of 50 possible exponents between 0 and 5. Here, we use the prioritized risk to study fundamen-tal limits on the agent's learning performance when γ (i.e.,

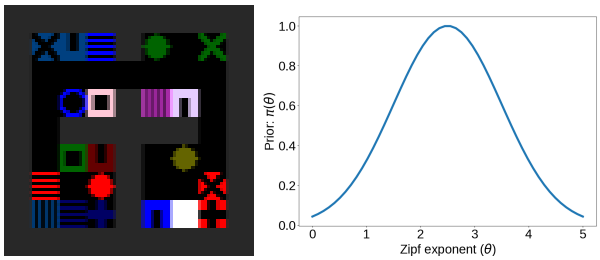

Figure 2. L: A Zipfian Gridworld environment. R: Prior ().

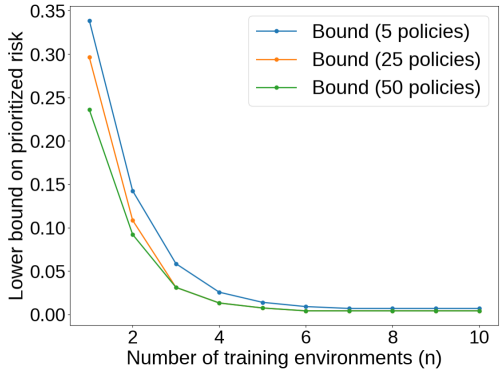

Figure 3. Lower bounds on prioritized risk (Zipfian environments).

how heavy-tailed the distribution is) does not conform to the learner's prior. The agent does not have a priori knowledge of , but has a prior () = exp((2:5)^2), which cap-tures the prior knowledge that real-world distributions are likely to have Zipf exponents close to 2.5 (Fig. 2 right). For training, the agent receives n environments from P. Based on this training set, it must choose a policy (i.e., a mapping from images to actions) from a set A. Here, we choose a (pre-computed) set A of policies, each trained on a given Zipfian exponent using IMPALA (Espeholt et al., 2018).

We use the generalized Fano inequality to compute lower bounds on the prioritized risk (Cor. 5.2); since and A are finite, we can compute all quantities in the bound exactly. Fig. 3 plots the resulting bounds for different sizes of A. We see that a learning agent that has access to a smaller set A faces a more stringent tradeoff between prior information and expected cost. Analogous to the analysis in Sec. 6.2, the bound suggests that for an agent $_{\mathrm{jAj=5}}$ (with access to A of size 5) to achieve the same level of performance as $_{\mathrm{jAj=50}}$ (with access to A of larger size) for high values of the prior (i.e., close to 2.5), $_{\mathrm{jAj=5}}$ must sacrifice a greater level of performance for less likely values of .

7. Related Work

No-free-lunch theorems and minimax lower bounds. The no-free-lunch (NFL) theorem (Wolpert, 1996)(Shalev-Shwartz & Ben-David, 2014, Ch. 5) establishes fundamental limits on learners that have no prior information. One statement of the NFL theorem is via the minimax risk: if one considers binary classification tasks and allows for any distribution whatsoever over samples, then the minimax is lower bounded by a constant (for all sample sizes n). In general, lower bounds on the minimax risk for a given learning problem allow us to formalize the fundamental limits of learning on that problem. Minimax lower bounds have been established for many learning problems of practical interest, e.g., sparse linear regression (Raskutti et al., 2011), nonparametric classification (Yang, 1999), crowdsourcing (Zhang et al., 2014), differentially private learning (Duchi et al., 2013), and inverse reinforcement learning (Komanduru & Honorio, 2021). Such lower bounds are typically proved using information-theoretic techniques, e.g., LeCam's method, Fano's inequality, or Assouad's method (Tsybakov, 2008; Yang & Barron, 1999; Yu et al., 1997). As highlighted in Sec. 2, the minimax risk does not reason about prior knowledge beyond the weak knowledge that the data-generating distribution belongs to a certain set.

Lower bounds on Bayes risk. The Bayes risk (Eq. (3)) considers the average risk relative to a prior over data-generating distributions (i.e., over). Bayes risk lower bounds are known for a variety of problems including finite-dimensional estimation problems with quadratic losses (Van Trees, 1968; Brown & Gajek, 1990; Gill et al., 1995), estimation problems involving generalized linear models (Chen et al., 2016), and high-dimensional sparse linear regression (Chen et al., 2016). As highlighted in Sec. 2, the Bayes risk framework assumes that the learner has access to the "true" distribution over , and thus does not capture settings where reality does not conform to the prior.

More refined versions of the Bayes risk include "local" versions (Zhang, 2006) where one partitions the parameter space into disjoint balls (of small size) and computes a lower bound on the Bayes risk defined relative to a local prior. This quantity is distinct from the one considered here. The work presented by Haussler et al. (1994) considers the tradeoff between prior knowledge and sample complexity by analyzing the impact of a misspecified prior on the Bayes risk. However, Haussler et al. (1994) only provide upper bounds on sample complexity in the setting with misspecified priors; lower bounds are only provided for the case where the learner has access to nature's prior. More recently, Chollet (2019) considers a formalism based on algorithmic information theory for capturing tradeoffs between prior knowledge and learning efficiency (but computing the relevant quantities is generally computationally intractable).

PAC-Bayes generalization bounds. One motivation for studying fundamental limits on learners with prior information comes from PAC-Bayes theory (McAllester, 1999), which provides upper bounds on risk for learners equipped with prior information. In PAC-Bayes learning, one defines a prior distribution over the space of hypotheses, and then obtains a bound on the risk that holds for any choice of posterior. This bound can be operationalized by fixing a

data-independent prior and then finding a data-dependent posterior that minimizes the PAC-Bayes bound (Dziugaite & Roy, 2017). The idea of using such prior information to improve generalization bounds is also a key component of the "luckiness" framework (Shawe-Taylor et al., 1998), Occam's razor bounds (Blumer et al., 1987; Langford, 2005), and the minimum description length principle (Rissanen, 1989). PAC-Bayes provides some of the tightest known generalization bounds for neural networks for supervised learning (Dziugaite & Roy, 2017; Neyshabur et al., 2017a;b; Bartlett et al., 2017; Arora et al., 2018; Rivasplata et al., 2019; Perez-Ortiz et al., 2020; Jiang et al., 2020; Lotfi et al., 2022) and policy learning (Fard et al., 2012; Majumdar et al., 2021; Veer & Majumdar, 2020; Ren et al., 2021).

These results provide a motivation for the question we consider here: how can we establish fundamental limits on learners with imperfect prior information? Empirically, one often observes that the strength of PAC-Bayes bounds depends significantly on how good the prior is. If the data-generating distribution conforms to the PAC-Bayes prior (e.g., if the prior achieves low expected loss on the data-generating distribution), then the resulting PAC-Bayes bound can be strong (i.e., the upper bound on the expected loss can be low). However, if the prior is not well-matched with the data-generating distribution, the posterior may have to deviate significantly from the prior, resulting in a poor upper bound. Thus, one observes a tradeoff with PAC-Bayes bounds: if the data-generating distribution conforms to the learner's prior, one can achieve low risk; however, if the data-generating distribution does not conform to the learner's prior, the PAC-Bayes bound on the risk is high. This tradeoff is very similar to the one the prioritized risk attempts to capture. However, while the PAC-Bayes framework provides upper bounds on risk, we focus on lower bounds. We are thus motivated by the goal of establishing fundamental tradeoffs (i.e., ones that hold for any learning algorithm) between the accuracy of prior information and learning performance, while the PAC-Bayes approach provides a particular such tradeoff.

8. Discussion and Conclusions

We have introduced the notion of prioritized risk, which differs from classical notions of minimax and Bayes risk by allowing us to study fundamental tradeoffs in settings where reality does not conform to the learner's prior. Specifically, lower bounds on the prioritized risk for a given problem establish that it is impossible for both the risk of a learner and the prior to be low for all distributions. We have extended classical techniques based on the methods of LeCam, Assouad, and Fano for obtaining lower bounds on the prioritized risk for estimation. We also presented a technique for obtaining lower bounds in more general settings via a novel generalized Fano inequality (which may be of independent interest for lower bounding Bayes risk in settings with unbounded loss functions).

8.1. Future Work

There are a number of exciting directions for future work. First, developing prior-informed learning algorithms that are optimal from the perspective of the prioritized risk would be of practical interest (e.g., based on variants of PAC-Bayes bounds). Empirically, one often observes tradeoffs between accuracy of prior information and learning performance: for some problems, the choice of neural network architecture or regularization technique (which can both be interpreted as forms of inductive bias / prior knowledge) seem to have little impact on learning performance, while for other problems these choices can have significant impact. This raises the following important question: do the empirical observations reflect fundamental tradeoffs (as formalized by the prioritized risk), or are they artifacts of the specific learning algorithms we happen to be using? If the latter, this motivates the search for different learning algorithms for our problems of interest.

Second, an interesting theoretical direction is to explore if there are settings where the optimal asymptotic dependence on n (number of examples) for prioritized risk and minimax/Bayes risk are different (analogous to the difference between universal learning (Bousquet et al., 2021) and PAC learning). To expand on this, consider a learner $_{\text{minimax}}$ that achieves the optimal asymptotic rate for the minimax risk. There are then two possibilities. First, it may be the case that $_{\text{minimax}}$ does not achieve the optimal asymptotic rate for the prioritized risk. The practical implication is that one should use a different learner based on whether or not one has a uniform or non-uniform prior (even for large amounts of data). Second, it may be the case that $_{\text{minimax}}$ does achieve the optimal asymptotic rate for the prioritized risk. This implies that the same learner achieves a different asymptotic tradeoff between prior knowledge and learning performance depending on whether or not one has a uniform prior. Each possibility is interesting in its own right, and provides insights into the tradeoffs between prior knowledge and learning performance for the problem under consideration.

Finally, we are interested in using the prioritized risk framework to understand how much prior information ("nature") one needs in order to achieve a certain level of learning performance ("nurture") in a broader set of applications of practical interest (e.g., RL in robotics).

Acknowledgements

The author would like to thank Olivier Bousquet for a helpful discussion on this work, and Asher Hancock, Nate Simon, David Snyder, and Vince Pacelli for providing feedback on an early draft. This work was partially supported by the NSF CAREER Award [#2044149] and the Office of Naval Research [N00014-23-1-2148].

References

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger Generalization Bounds for Deep Nets via a Compression Approach. arXiv preprint arXiv:1802.05296, 2018.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-Normalized Margin Bounds for Neural Networks. In Advances in Neural Information Processing Systems, pp. 6240–6249, 2017.

Berger, J. O. Statistical Decision Theory and Bayesian Analysis. Springer Science & Business Media, 2013.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Occam's Razor. Information Processing Letters, 24(6):377–380, 1987.

Bousquet, O., Hanneke, S., Moran, S., Van Handel, R., and Yehudayoff, A. A theory of universal learning. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pp. 532–541, 2021.

Brown, L. D. and Gajek, L. Information inequalities for the Bayes risk. The Annals of Statistics, pp. 1578–1594, 1990.

Chan, S. C., Lampinen, A. K., Richemond, P. H., and Hill, F. Zipfian environments for reinforcement learning. In Conference on Lifelong Learning Agents, pp. 406–429. PMLR, 2022.

Chen, X., Guntuboyina, A., and Zhang, Y. On Bayes risk lower bounds. The Journal of Machine Learning Research, 17(1):7687–7744, 2016.

Chollet, F. On the measure of intelligence. arXiv preprint arXiv:1911.01547, 2019.

Cover, T. M. and Thomas, J. A. Elements of Information Theory. John Wiley & Sons, 2012.

Duchi, J. Lecture notes for Statistics 311. URL: https://web.stanford.edu/class/stats311/lecture-notes.pdf, 2016.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pp. 429–438. IEEE, 2013.

Dziugaite, G. K. and Roy, D. M. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence, 2017.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In International Conference on Machine Learning, pp. 1407–1416. PMLR, 2018.

Fard, M. M., Pineau, J., and Szepesvári, C. PAC-Bayesian policy evaluation for reinforcement learning. arXiv preprint arXiv:1202.3717, 2012.

Gerchinovitz, S., Ménard, P., Stoltz, G., et al. Fano's inequality for random variables. Statistical Science, 35(2): 178–201, 2020.

Gill, R. D., Levit, B. Y., et al. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. Bernoulli, 1 (1-2):59–79, 1995.

Gray, R. M. Entropy and Information Theory. Springer Science & Business Media, 2nd edition, 2011.

Haussler, D., Kearns, M., and Schapire, R. E. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. Machine learning, 14(1):83–113, 1994.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic Generalization Measures and Where to Find Them. Proceedings of the 8th International Conference on Learning Representations, 2020.

Komanduru, A. and Honorio, J. A lower bound for the sample complexity of inverse reinforcement learning. arXiv preprint arXiv:2103.04446, 2021.

Langford, J. Tutorial on Practical Prediction Theory for Classification. Journal of Machine Learning Research, 6 (Mar):273–306, 2005.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006.

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. PAC-Bayes compression bounds so tight that they can explain generalization. Advances in Neural Information Processing Systems, 35:31459–31473, 2022.

Majumdar, A. and Pacelli, V. Fundamental performance limits for sensor-based robot control and policy learning. Proceedings of Robotics: Science and Systems (RSS), 2022.

Majumdar, A., Farid, A., and Sonar, A. PAC-Bayes control: learning policies that provably generalize to novel environments. The International Journal of Robotics Research (IJRR), 40(2-3):574–593, 2021.

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

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. preprint arXiv:1707.09564, 2017a.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring Generalization in Deep Learning. In Advances in Neural Information Processing Systems, pp. 5949–5958, 2017b.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. Tighter Risk Certificates for Neural Networks. arXiv preprint arXiv:2007.12911, 2020.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ‘_q-balls. IEEE Transactions on Information Theory, 57 (10):6976–6994, 2011.

Ren, A., Veer, S., and Majumdar, A. Generalization guarantees for imitation learning. In Conference on Robot Learning, pp. 1426–1442. PMLR, 2021.

Rissanen, J. Stochastic Complexity in Statistical Inquiry. World Scientific, 1989.

Rivasplata, O., Tankasali, V. M., and Szepesvari, C. PAC-Bayes with Backprop. arXiv preprint arXiv:1908.07380, 2019.

Shalev-Shwartz, S. and Ben-David, S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory, 44(5):1926–1940, 1998.

Tsybakov, A. B. Introduction to Nonparametric Estimation. Springer Science & Business Media, 2008.

Van Trees, H. L. Detection, Estimation and Modulation Theory. Wiley, 1968.

Veer, S. and Majumdar, A. Probably approximately correct vision-based planning using motion primitives. In Conference on Robot Learning, pp. 1001–1014. PMLR, 2020.

Wolpert, D. H. The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7):1341–1390, 1996.

Yang, Y. Minimax nonparametric classification. I. Rates of convergence. IEEE Transactions on Information Theory, 45(7):2271–2284, 1999.

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. Annals of Statistics, pp. 1564–1599, 1999.

Yu, B., Assouad, F., and Le Cam, L. Festschrift for Lucien Le Cam, 1997.

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

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. arXiv preprint arXiv:1406.3824, 2014.

A. Proofs

Proposition A.1 (Reduction from estimation to testing). Let $: X^n \to$ be an estimator for a learning problem defined by a loss function of the form (13). Let $\{_v\}_{v \in V}$ form a $(;)$-packing and define the prior-weighted test function as in (14). We then have the following bound:

$$R_{prior}(; L;) \quad P(\ (X_1) = V^n): \tag{15}$$

Hence, taking an infimum over estimators, we have:

$$R_{prior}(; L;) \quad \inf P(\ (X_1) = V^n): \tag{16}$$

Proof. Let $: X^n \to$ be an estimator. We first observe:

$$R_{prior}(; L;) = \sup_{} ()_{X^n} E_{P^n}(; (X^n)) \tag{33}$$

$$= \sup_{} {}_{X^n} E_{P^n} ()(; (X^n)) \tag{34}$$

$$\sup_{} E_{X^n P^n} 1 ()(; (X^n)) \tag{35}$$

$$= \sup_{} P ()(; (X^n)) : \tag{36}$$

Now consider a family of distributions $\{_v\}_{v \in V}$ that forms a $(;)$-packing, and define the prior-weighted test function:

$$(x_1^n) := \operatorname{argmin}_{v \in V} (_v)(_v; (x^n)): \tag{37}$$

Now suppose that $(_v)(_v; (x_1)) < .$ Then, we claim that $(x_1) = v$. To see this, suppose not. Then,

$$9v^0 \text{ s.t. } (_{v^0})(_{v^0}; (x_1)) < (_v)(_v; (x_1)) \tag{38}$$

$$=) \quad (\ _{v^0}; (x \ _1)) < \frac{(_v)(_v; (x^n))}{(_{v^0})} < \frac{}{(_{v^0})}: \tag{39}$$

Thus, (x^n) is in the -ball of radius $=(_{v^0})$ around $_{v^0}$. But, we said that (x^n) is in the -ball of radius $=(_v)$ around $_v$, and that the balls are non-overlapping (since we have a $(;)$-packing). Thus, $(x_1) = v$.

Considering the contrapositive, we see that if $(x_1^n) = v$, then $(_v; (x_1)) \geq (_v)$. Averaging over V, we see:

$$\sup P (; (X^n)) \quad \frac{1}{jVj} \sum_{v \in V} P \frac{}{(_v; (X^n))} \quad V = v \quad \frac{}{} \tag{40}$$

$$\frac{1}{jVj} \sum_{v \in V} P \quad (X_1) \quad n = v \quad V = v \tag{41}$$

$$=: P(\ (X_1^n) = V): \tag{42}$$

Combining this with (36) establishes the claim in the proposition.

Theorem 4.2 (LeCam's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P\}_{2}$ with a loss function of the form (13). A learner has prior and will receive n i.i.d. observations from a chosen distribution. Let $\square$ $\{_0; _1\}$ form a $(;)$-packing. We then have the following lower bound on the prioritized risk:

$$R_{prior}(; L;) \quad \frac{1}{2} \left[1 - kP^n \quad P_0^n k_{TV} \right]; \tag{17}$$

Proof. The proof follows directly by combining Proposition 4.1 with the well-known variational representation of the total variation distance (see, e.g., (Duchi, 2016, Proposition 2.17)): for distributions P_0 and P_1 on sample space X, we have

$$\inf \{P_0(\ (X) = 0) + P_1(\ (X) = 1)\} = 1 \quad kP_0 \quad P_1 k_{TV};$$

where the infimum is over test functions $: X \to \{0; 1\}$. $\square$

Theorem 4.3 (Fano's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P_\theta\}_{\theta\in\Theta}$ with a loss function of the form (13). A learner has prior π and will receive n i.i.d. observations from a chosen distribution. Let $\{\theta_v\}_{v\in\mathcal{V}}$ form a (α,β)-packing. Define a random variable V corresponding to a uniform distribution over $\mathcal{V}$. We can then lower bound the prioritized risk:

$$R_{\text{prior}}(\pi;L;n) \geq \frac{\alpha}{2}\left[1 - \frac{I(V;X_1^n) + \log 2}{\log|\mathcal{V}|}\right], \tag{19}$$

where $I(V;X_1^n)$ denotes the mutual information.

Proof. The proof follows directly by combining Proposition 4.1 with Fano's inequality (Cover & Thomas, 2012):

$$\inf_{\psi} P(\psi(X_1^n) \neq V) \geq 1 - \frac{I(V;X_1^n) + \log 2}{\log|\mathcal{V}|}. \tag{43}$$

□

Theorem 4.5 (Assouad's method for prioritized risk). Consider an estimation problem over a family of distributions $\{P_\theta\}_{\theta\in\Theta}$ with a loss function of the form (13). A learner has prior π and will receive n i.i.d. observations from a chosen distribution. Let $\{\theta_v\}_{v\in\mathcal{V}=\{-1,1\}^d}$ form a $(2\alpha,\beta)$-Hamming separation with θ_0. We then have:

$$R_{\text{prior}}(\pi;L;n) \geq \frac{\alpha}{2}\sum_{j=1}^{d}\left[\inf_{\psi} P_{+j}(\psi(X_1^n) \neq +1) + P_{-j}(\psi(X_1^n) \neq -1)\right],$$

where the infimum is over tests $\psi : \mathcal{X}^n \to \{+1,-1\}$.

Proof. Fix an arbitrary estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$. We then have:

$$R_{\text{prior}}(\pi;L;n) = \sup_{\theta\in\Theta} \mathbb{E}_{X^n\sim P_\theta^n}[L(\theta,\hat{\theta}(X^n))] \geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{X^n\sim P_v^n}[L(\theta_v,\hat{\theta}(X^n))] \tag{44}$$

$$= \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{X^n\sim P_v^n}\left[\sum_{j=1}^{d} 2\alpha\, 1[\psi(\hat{\theta}(X^n))_j \neq v_j]\right] \quad \text{(By } (2\alpha,\beta)\text{-Hamming separation.)} \tag{46}$$

$$= \sum_{j=1}^{d}\frac{2\alpha}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v[\psi(\hat{\theta}(X^n))_j \neq v_j] \tag{47}$$

$$= \sum_{j=1}^{d}\frac{2\alpha}{|\mathcal{V}|}\left[\sum_{v:v_j=+1} P_v[\psi(\hat{\theta}(X^n))_j \neq v_j] + \sum_{v:v_j=-1} P_v[\psi(\hat{\theta}(X^n))_j \neq v_j]\right] \tag{48}$$

$$= \sum_{j=1}^{d}\left[P_{+j}[\psi(\hat{\theta}(X^n))_j \neq v_j] + P_{-j}[\psi(\hat{\theta}(X^n))_j \neq v_j]\right]. \tag{49}$$

Taking an infimum over estimators (on the LHS) and test functions $\psi : \mathcal{X}^n \to \{+1,-1\}$ (on the RHS) establishes the desired result. □

Theorem 5.1 (Generalized Fano inequality for unbounded losses). For any prior distribution p on Θ, the following lower bound on the Bayes risk holds for all $\gamma > 0$:

$$R_{\text{Bayes}}(p;L;n) := \inf_{\hat{\theta}} \int_{\Theta} \mathbb{E}_{X^n\sim P_\theta^n}[L(\theta,\hat{\theta}(X^n))]\, p(d\theta) \geq \frac{1}{\gamma}\left[I(\theta;X^n) - \log 2\right], \tag{22}$$

where $I(\theta; X_1^n)$ is the mutual information, and

$$\Gamma_{\pi,L}^{-1} = \sup_a \log \int \exp\left(L(\theta, a)\right) p(d\theta). \tag{23}$$

Proof. The Donsker-Varadhan change of measure inequality (Gray, 2011, Theorem 2.3.2) states that for any random variable Z, we have the following inequality for all distributions P and Q:

$$-\log_Z E_P[Z] \leq KL(P\|Q) + \log_Z E_Q \exp(Z). \tag{50}$$

Choosing $Z = L(\theta, (X_1^n))$, we then have:

$$-\inf_{p(\theta, X_1^n)} E\left[L(\theta, (X^n))\right] \leq KL(p(\theta, X^n)\|p(\theta)q(X^n)) - \log E_{p(\theta)q(X_1^n)}\left[\exp\left(L(\theta, (X^n))\right)\right], \tag{51}$$

where $p(\theta, X_1^n)$ is the joint distribution defined by p and P, and q is any arbitrary distribution on X^n. Taking an infimum over θ on both sides, we have:

$$\inf_{p(\theta, X_1^n)} E\left[L(\theta, (X^n))\right] \leq KL(p(\theta, X^n)\|p(\theta)q(X^n)) - \sup_q \log E_{p(\theta)q(X_1^n)}\left[\exp\left(L(\theta, (X^n))\right)\right] \tag{52}$$

$$= KL(p(\theta, X^n)\|p(\theta)q(X^n)) - \log \sup_q E_{p(\theta)q(X_1^n)}\left[\exp\left(L(\theta, (X^n))\right)\right]. \tag{53}$$

The equality above follows from the monotonicity of the log function. Now, via the Fubini-Tonelli theorem, we have:

$$\log \sup_q E_{p(\theta)q(X_1^n)}\left[\exp\left(L(\theta, (X^n))\right)\right] = \log \sup_q E_{q(X_1^n)} E_{p(\theta)}\left[\exp\left(L(\theta, (X^n))\right)\right] \tag{54}$$

$$\leq \log E_{q(X_1^n)} E_{p(\theta)}\left[\sup_q \exp\left(L(\theta, (X^n))\right)\right] \tag{55}$$

$$= \log E_{q(X_1^n)} E_{p(\theta)}\left[\sup_a \exp\left(L(\theta, a)\right)\right] \tag{56}$$

$$= \sup_a \log E_{p(\theta)}\left[\exp\left(L(\theta, a)\right)\right] \tag{57}$$

$$= : \Gamma_{\pi,L}^{-1}. \tag{58}$$

We thus have:

$$\inf_{p(\theta, X_1^n)} E\left[L(\theta, (X_1))\right] - nKL(p(\theta, X_1)\|p(\theta)q(X_1)) + \Gamma_{\pi,L}^{-1} \leq n \tag{59}$$

Noting that this inequality holds for any choice of q, we can supremize over q to obtain the tightest bound:

$$\sup_q KL(p(\theta, X_1^n)\|p(\theta)q(X_1^n)) = \inf_q KL(p(\theta, X_1)\|p(\theta)q(X_1)) = I(\theta; X_1). \tag{60}$$

We thus obtain the desired result:

$$R_{Bayes}(p; L; \pi) = \inf_{p(\theta, X_1^n)} E\left[L(\theta, (X_1))\right] \tag{61}$$

$$\geq \frac{1}{\Gamma_{\pi,L}}\left[I(\theta; X_1) - n\right]. \tag{62}$$

□

Lower bound for logistic regression (Sec. 6.2). We prove the following lower bound for the logistic regression problem described in Sec. 6.2:

$$R_{\text{prior}}(;L;) \quad \frac{1}{16}\sqrt{\frac{d^{\frac{3}{2}}}{\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2}}: \tag{63}$$

Proof. Using Assouad's method (Thm. 4.5) and (21) we see:

$$R_{\text{prior}}(;L;) \quad \sum \left[\frac{1}{2}\sum_{j=1}^{d} P_{+j} \quad P^{n}{}_{jTV}\right]^{n} \tag{64}$$

$$\geq \frac{d}{2}\left[1 - \frac{1}{d}\sum_{j=1}^{d}\frac{1}{2^d}\sum_{v\in V}\|P_{v;+j} \quad P_{v;\ j}\|_{TV}^2\right]^{\frac{1}{2}}; \tag{65}$$

where $P_{v;j}$ is defined as the distribution P_v where coordinate j takes the value $v_j = 1$. The inequality above follows from the Cauchy-Schwarz inequality and convexity of the total variation distance (see (Duchi, 2016, p.165)).

Define:

$$p_v(z) := \frac{1}{1+\exp(_{(v)}z^T v)}; \tag{66}$$

and let $KL(p\|q)$ denote the binary KL-divergence between Bernoulli(p) and Bernoulli(q). By Pinsker's inequality, we have for any $v; v^0$:

$$\|P^{n}_{v} \quad P^{n}_{v^0}\|_{TV}^2 \quad \frac{1}{4}\left[KL(P^{n}_{v}\|P^{n}_{v^0}) + KL(P_{v^0}{}^{n}\|P_{v})^{n}\right] = \frac{1}{4}\sum_{i=1}^{n}\left[KL(p_v(z_i)\|p_{v^0}(z_i)) + KL(p_{v^0}(z_i)\|p_v(z_i))\right]: \tag{67}$$

Letting $p_a := 1/(1+e^a)$ and $p_b := 1/(1+e^b)$, we have (see (Duchi, 2016, p.167)):

$$KL(p_a\|p_b) + KL(p_b\|p_a) \quad (a \quad b)^2: \tag{68}$$

This implies:

$$\|P^{n}_{v} \quad P^{n}_{0}\|_{TV}^2 \quad \frac{1}{4}\sum_{i=1}^{n}\left[\frac{}{(w)}z^T v^{i} \quad \frac{}{(v^0)}z_i^{T}v^{C}\right]^2 \tag{69}$$

$$= \frac{2}{4}\sum_{i=1}^{n} z_i^T\left[\frac{v}{(v)} \quad \frac{v^0}{(v^0)}\right]^2: \tag{70}$$

In order to lower bound (65), we use the preceding bound to note:

$$\frac{1}{2^d d}\sum_{j=1}^{d}\sum_{v\in V}\|P_{v;+j} \quad P_{v;\ j}\|_{TV}^2 \quad \frac{2}{4d2^d}\sum_{v\in V}\sum_{j=1}^{d}\left[\sum_{i=1}^{n} z_{ij}\right]^2\left[\frac{1}{(v_j)} + \frac{1}{(v_j)}\right] \tag{71}$$

$$= \frac{4}{d2^d}\sum_{v\in V}\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2 \tag{72}$$

$$= \frac{4^2}{d}\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2: \tag{73}$$

Thus, using (65), we have:

$$R_{\text{prior}}(;L;) \quad \frac{d}{2}\left[1 \quad \frac{4^2}{d}\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2\right]^{\frac{1}{2}}: \tag{74}$$

Setting

$$^2 = \frac{}{16\sqrt{\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2}}; \quad (75)$$

we obtain the desired result:

$$R_{prior}(;L;) \quad _4 = \frac{d}{16}\frac{1}{\sqrt{\sum_{j=1}^{d}\sum_{i=1}^{n} z_{ij}^2}}\frac{d^{\frac{3}{2}}}{}: \quad (76)$$

□

B. Bernoulli Mean Estimation: Upper Bounds

Fig. 4 presents upper bounds on prioritized risk computed using Bayesian inference (with prior) for the problem of Bernoulli mean estimation. Specifically, we present the learner-specific prioritized risk where the learner outputs the mean of the posterior distribution computed using the prior and a dataset of size n; since is chosen to be a Beta distribution (Beta(= 1, = 2)) and the underlying random variable has a Bernoulli distribution, one can analytically perform Bayesian inference in this setting. We estimate the expectation over datasets by averaging the loss over 10,000 datasets. We compare the results with two other learners corresponding to performing (i) performing Bayesian inference with a uniform prior, and (ii) Bayesian inference with a different prior (specifically, we use Beta(= 1, = 4), which concentrates the prior towards values of where is higher). As the figure indicates, Bayesian inference with achieves a lower learner-specific prioritized risk than Bayesian inference with a uniform prior. However, the figure also shows that the custom inference algorithm (Bayesian inference with a more concentrated prior) achieves a lower learner-specific prioritized risk compared to Bayesian inference with . Thus, Bayesian inference with prior is not necessarily optimal from the perspective of the prioritized risk. This observation thus leaves open the interesting direction for future work of identifying algorithms that achieve optimal prioritized risk.

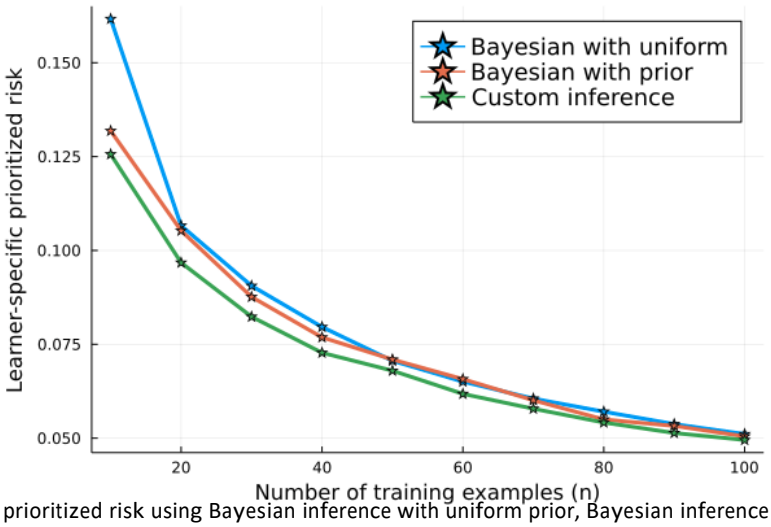

Figure 4. Upper bounds on prioritized risk using Bayesian inference with uniform prior, Bayesian inference with prior , and a custom inference algorithm (Bayesian inference with a modified prior).