Electronic Journal of Statistics
Vol. 17 (2023) 1996–2043
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2151

Adversarial meta-learning of Gamma-minimax estimators that leverage prior knowledge

Hongxiang Qiu[1] **and Alex Luedtke**[2]

[1] *Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA*
e-mail: qiuhongx@msu.edu

[2] *Department of Statistics, University of Washington, Seattle, WA, USA*
e-mail: aluedtke@uw.edu

Abstract: Bayes estimators are well known to provide a means to incorporate prior knowledge that can be expressed in terms of a single prior distribution. However, when this knowledge is too vague to express with a single prior, an alternative approach is needed. Gamma-minimax estimators provide such an approach. These estimators minimize the worst-case Bayes risk over a set Γ of prior distributions that are compatible with the available knowledge. Traditionally, Gamma-minimaxity is defined for parametric models. In this work, we define Gamma-minimax estimators for general models and propose adversarial meta-learning algorithms to compute them when the set of prior distributions is constrained by generalized moments. Accompanying convergence guarantees are also provided. We also introduce a neural network class that provides a rich, but finite-dimensional, class of estimators from which a Gamma-minimax estimator can be selected. We illustrate our method in two settings, namely entropy estimation and a prediction problem that arises in biodiversity studies.

Keywords and phrases: Gamma-minimax estimation, machine learning.

Received October 2022.

1. Introduction

A variety of principles can be used to guide the search for a suitable statistical estimator. Asymptotic efficiency (Pfanzagl, 1990), minimaxity (Wald, 1945) and Bayes optimality (Berger, 1985) are popular examples of such principles. Defining the performance criteria underlying these principles requires specifying a model space, that is, a collection of possible data-generating mechanisms known to contain the true, underlying distribution.

It is often desirable to incorporate prior information about the true data-generating mechanism into a statistical procedure. This might be done differently in different statistical paradigms. For frequentist methods, such as those based on the asymptotic efficiency or minimax principle, the primary way to incorporate this information is via the definition of the model space. In the Bayesian paradigm, such information may be represented by further specifying a prior distribution (or *prior* for short) over the model space and aiming for an estimator that minimizes the induced Bayes risk. However, in many cases, there may be several priors that are compatible with the available information, especially in the context of rich model spaces.

The Gamma-minimax paradigm, proposed by Robbins (1951), provides a principled means to overcome the challenge of specifying a single prior distribution. Under this paradigm, a statistician first specifies a set Γ of all priors that are consistent with the available prior information and subsequently seeks an estimator that minimizes the worst-case Bayes risk over this set of priors. The Gamma-minimax paradigm may be viewed as a robust version of the Bayesian paradigm that is less sensitive to misspecification of a prior distribution (Vidakovic, 2000). When it is infeasible to specify a prior due to the complexity of the model space, the Gamma-minimax paradigm may also be viewed as a feasible substitute for the Bayesian paradigm. The Gamma-minimax paradigm is closely related to Bayes and minimax paradigms: when the set of priors consists of a single prior, a Gamma-minimax estimator is Bayes with respect to that prior; when the set Γ of priors is the entire set of possible prior distributions, a Gamma-minimax estimator is also minimax.

Gamma-minimax estimators have been studied for a variety of problems. Some explicit forms of Gamma-minimax estimators have been obtained. For example, Olman and Shmundak (1985) studied Gamma-minimax estimation of the mean of a normal distribution for the set of symmetric and unimodal priors on an interval and obtained an explicit form when this interval is sufficiently small. Eichenauer-Herrmann (1990) generalized this result to more general parametric models and Eichenauer-Herrmann, Ickstadt and Weiß (1994) obtained a further generalization with the requirement of symmetry on the priors dropped. Chen, Eichenauer-Herrmann and Lehn (1988) studied Gamma-minimax estimation for multinomial distributions and the set of priors with bounded mean. Chen et al. (1991) studied Gamma-minimax estimation for one-parameter exponential families and the set of priors that place certain bounds on the first two moments. These results do not deal with general model spaces, such as semiparametric or nonparametric models, and general forms of the set of priors that may not directly impose bounds on prior moments of the parameters of interest. One reason for this lack of generality might be that, in the existing literature, Gamma-minimaxity is defined only for parametric models. However, an issue with parametric models is that they often fail to contain the true data-generating mechanism, in which case output from the aforementioned statistical procedures may no longer be interpretable. Another possible reason is that it is typically intractable to analytically derive Gamma-minimax estimators, even for parametric models.

To overcome this lack of analytical tractability, meta-learning algorithms to compute a minimax or Gamma-minimax estimator have been proposed. Still, most of these works focus on parametric models. For example, Nelson (1966) and Kempthorne (1987) each proposed an algorithm to compute a minimax estimator. Bryan et al. (2007) and Schafer and Stark (2009) proposed an algorithm to compute an approximate confidence region of optimal expected size in the minimax sense. Noubiap and Seidel (2001) proposed an iterative algorithm to compute a Gamma-minimax decision for the set of priors constrained by generalized moment conditions. More recent works explored computing estimators under more general models. For example, Luedtke et al. (2020) introduced an approach, termed Adversarial Monte Carlo meta-learning (AMC), for constructing minimax estimators. In the special case of prediction problems with mean-squared error, Luedtke, Chung and Sofrygin (2020) studied the invariance properties of the decision problem and their implications for AMC.

In this paper, we make the following contributions:

1. We propose iterative adversarial meta-learning algorithms for constructing Gamma-minimax estimators for a general model space and class of estimators. We further provide convergence guarantees for these algorithms.

To our best knowledge, this is the first algorithm to compute Gamma-minimax estimators under general models, including infinite-dimensional models. We also show that, for certain problems, there is a unique Gamma-minimax estimator and our computed estimator converges to this estimator as the number of iterations increases to infinity.

Like the approach proposed in Noubiap and Seidel (2001), we consider sets of priors characterized by (in)equality constraints on prior generalized moments and our proposed iterative algorithm involves solving a discretized Gamma-minimax optimization problem in each intermediate step. However, we explicitly describe algorithms to solve these minimax problems, which facilitates the use of our approach by practitioners. When the space of estimators can be parameterized by a Euclidean space and gradients are available, we propose to use a gradient-based algorithm or a stochastic variant thereof. When gradients are unavailable, we propose to instead use fictitious play (Brown, 1951; Robinson, 1951) to compute a stochastic estimator, which is a mixture of deterministic estimators belonging to some specified collection. We also provide a convergence result that is applicable even when this collection has infinite cardinality. This is in contrast to the results in Robinson (1951), which require that each player has only finitely many possible deterministic strategies.

2. We propose a Markov chain Monte Carlo (MCMC) method to construct the approximating grids defining the discretized Gamma-minimax problems used in our iterative scheme.

Like the approach proposed in Noubiap and Seidel (2001), our proposed iterative algorithm relies on increasingly fine finite grids over the model space. However, since we allow the model space to be high or even infinite-dimensional, randomly adding points to the grid may lead to unacceptably slow convergence. To overcome this challenge, we propose to use MCMC to efficiently construct such grids.

Our theoretical results allow for many different choices of classes of estimators. Our final contribution concerns the introduction of one such class:

3. We introduce a new neural network architecture that can be used to parameterize statistical estimators and argue that this class represents an appealing class to optimize over.

For this final point, we build on existing works in adversarial learning (e.g., Goodfellow et al., 2014; Luedtke et al., 2020; Luedtke, Chung and Sofrygin, 2020) and extreme learning machines (Huang, Zhu and Siew, 2006). Thanks to the universal approximation properties of neural networks (e.g., Hornik, 1991; Csáji, 2001) and extreme learning machines (Huang, Chen and Siew, 2006), we also show that both of these parameterizations can achieve good performance for sufficiently large networks. Furthermore, inspired by pre-training (e.g., Erhan et al., 2010) and transfer learning (e.g., Torrey and Shavlik, 2009), we recommend leveraging knowledge of existing estimators as inputs to the network in settings where this is possible. Under such choices of the space of estimators, we can expect to obtain a useful estimator even if the associated nonconvex-concave minimax problems prove to be difficult.

This paper is organized as follows. In Section 2, we introduce the framework of Gamma-minimax estimation and regularity conditions that we assume throughout the paper. In Section 3, we describe our proposed iterative adversarial meta-learning algorithms. In Section 4, we discuss considerations when choosing hyperparameters in the algorithms. In Section 5, we demonstrate our method in three simulation studies. We conclude with a discussion in Section 6. Proof sketches of key results are provided in the main text, and complete proofs can be found in the appendix. We also provide a table summarizing the frequently used symbols in Table 7 in the appendix. The code for our simulations is available on GitHub (Qiu, 2022).

2. Problem setup

Let $\mathcal{M}$ be a separable Hausdorff space of data-generating mechanisms that contains the truth P_0 and is equipped with a metric ρ. Under a data-generating mechanism $P \in \mathcal{M}$, let $\mathbf{X}^* \in \mathcal{X}^*$ denote the random data being generated, where $\mathcal{X}^*$ is the space of values that the random data takes. Let $\mathcal{C}$ denote a known coarsening mechanism such that the observed data $\mathbf{X} = \mathcal{C}(\mathbf{X}^*) \in \mathcal{X}$, where $\mathcal{X}$ is the space of observed data. In some cases, the coarsening mechanism will be the identity map, whereas in other settings, such as those in which missing, censored or truncated data is present, the coarsening mechanism will be nontrivial (e.g., Birmingham, Rotnitzky and Fitzmaurice, 2003; Gill, van der Laan and Robins, 1997; Heitjan and Rubin, 1991; Heitjan, 1993, 1994). Let $\mathcal{D}$ denote the space of estimators (or decision functions) equipped with a metric ϱ. In practice, for computational feasibility, we will mainly consider an estimator space $\mathcal{D}$ that contains estimators parameterized by a Euclidean space such as linear estimators or neural networks, and approximates a more general space $\mathcal{D}_0$, for example, the space of all estimators satisfying certain smoothness conditions. We discuss considerations concerning the choice of $\mathcal{D}$ in Section 4.2 and note that our proposed methods may be applied to broader estimator classes. We treat $\mathcal{D}$ as fixed throughout this paper. Let $R : \mathcal{D} \times \mathcal{M} \to \mathbb{R}$ denote a risk function that measures the performance of an estimator under a data-generating mechanism such that smaller risks are preferable. We suppose throughout that $\mathcal{M}$ and $\mathcal{D}$ are equipped with the topologies induced by ρ and ϱ, respectively.

We now present three examples in which we formulate statistical decision problems in the above form. The first example is a general example of point estimation. We use this example to illustrate how the Gamma-minimax estimation framework naturally fits into many statistical problems. The other two examples are more concrete and we will study them in the simulations and data analyses.

Example 1 (Point estimation). Suppose that $\mathcal{M}$ is a statistical model, which may be parametric, semiparametric, or nonparametric (Bickel et al., 1993). The data $\mathbf{X}^*$ consists of n independently and identically distributed (iid) random variables O_i, $i = 1, \dots, n$, following the true distribution $P_0 \in \mathcal{M}$. We set $\mathcal{C}$ to be the identity function so that $\mathbf{X} = \mathbf{X}^*$. We wish to estimate an aspect $\Psi(P_0) \in \mathbb{R}$ of P_0. Then, we can consider $\mathcal{D}$ to be a set of $\mathcal{X} \to \mathbb{R}$ functions and the mean-squared error risk $R(d, P) = \mathbb{E}_P[\{d(\mathbf{X}) - \Psi(P)\}^2]$. Some specific examples of estimands include:

i) Mean: $\Psi(P) = \mathbb{E}_P[O_i]$;
ii) Cumulative distribution function at a point o: $\Psi(P) = \mathbb{P}_P(O_i \leq o)$;
iii) Correlation: with $O_i = (X_i, Y_i) \in \mathbb{R}^2$, $\Psi(P) = \mathbb{E}_P[X_i Y_i] - \mathbb{E}_P[X_i]\mathbb{E}_P[Y_i]$.

Example 2 (Predicting the expected number of novel categories to be observed in a new sample). Suppose that $\mathcal{M}$ consists of multinomial distributions with an unknown number of categories. Let an iid random sample of size n be generated from the true multinomial distribution, so that $\mathbf{X}^*$ is a multiset containing the number X_k of observations in each category k. Suppose that only categories with nonzero occurrences are observed, so that $\mathbf{X}$ is a left-truncated version of $\mathbf{X}^*$. In other words, $\mathbf{X}$ is the multiset $\mathcal{C}(\mathbf{X}^*) = \{X_k : X_k > 0\}$. Then, we may wish to predict the number of new categories that would be observed if a new sample of size m were collected. This problem has been extensively studied in the literature, with applications in microbiome data, species taxonomic surveys, and assessment of vocabulary size, among other areas (e.g., Shen, Chao and Lin, 2003; Bunge, Willis and Walsh, 2014; Orlitsky, Suresh and Wu, 2016). This prediction problem can be formulated in our framework. For each $P \in \mathcal{M}$, let p_k ($k = 1, \dots, K_P$) be the probability of category k, and $\Psi(P)(\mathbf{X}^*)$ be $\sum_{k=1}^{K_P} I(X_k = 0)(1 - (1 - p_k)^m)$, the expected number of new observed categories given the current full data $\mathbf{X}^*$. We consider $\mathcal{D}$ to be a set of $\mathcal{X} \to \mathbb{R}$ functions and set the risk to be the mean-squared error, that is, $R(d, P) = \mathbb{E}_P[\{d(\mathbf{X}) - \Psi(P)(\mathbf{X}^*)\}^2]$. This prediction problem is known to be intrinsically difficult when the future sample size m is greater than the observed sample size n, and we might expect prior information to substantially improve prediction.

Example 3 (Entropy estimation). Consider the same data-generating mechanism and observed data as in Example 2. We may wish to estimate Shannon entropy (Shannon, 1948) $\Psi(P) = -\sum_{k=1}^{K_P} p_k \log p_k$, a measure of diversity. We consider $\mathcal{D}$ to be a set of $\mathcal{X} \to \mathbb{R}$ functions and set the risk to be the mean-squared error, that is, $R(d, P) = \mathbb{E}_P[\{d(\mathbf{X}) - \Psi(P)\}^2]$. Jiao et al. (2015) proposed a rate-minimax estimator. Thus, in contrast to Example 2, this is an example of a statistical problem with a satisfactory solution. For these problems, we might not expect prior information to substantially improve estimation.

We now define Gamma-minimaxity within our decision-theoretic framework. We assume that $\mathcal{M}$ is equipped with the Borel σ-field $\mathcal{B}$ and let Π denote the set of all probability distributions on the measurable space $(\mathcal{M}, \mathcal{B})$. We also assume that, for any $d \in \mathcal{D}$ and any $\pi \in \Pi$, $P \mapsto R(d, P)$ is π-integrable. The Bayes risk corresponding to an estimator d and a prior π is defined as $r : (d, \pi) \mapsto \int R(d, P)\, \pi(\mathrm{d}P)$. Let $\Gamma \subseteq \Pi$ be the set of priors such that all $\pi \in \Gamma$ are consistent with the available prior information. An estimator is called a Γ-minimax estimator if it is in the set

$$\operatorname*{argmin}_{d \in \mathcal{D}} \sup_{\pi \in \Gamma} r(d, \pi). \tag{1}$$

Throughout the rest of this paper, we assume the existence of this solution set and other solution sets to minimax problems, and that $\sup_{\pi \in \Gamma} r(d, \pi)$ is finite for any $d \in \mathcal{D}$.

In this paper, we consider the case in which Γ is characterized by finitely many generalized moment conditions, that is,

$$\Gamma = \left\{ \pi \in \Pi : \Phi_k \in L^1(\pi), \int \Phi_k(P)\, \pi(\mathrm{d}P) \leq c_k, k = 1, \ldots, K \right\}$$

where each $\Phi_k : \mathcal{M} \to \mathbb{R}$ is a prespecified function that extracts an aspect of a data-generating mechanism and $c_k \in \mathbb{R}$ is a prespecified constant. The validity of our proposed template to find a Γ-minimax estimator in Section 3.1 does not require Γ to take this form, but our proposed algorithms in Sections 3.2 and 3.3 are computationally feasible for such constraints because these linear constraints lead to linear programs, which can be solved efficiently (e.g., Jiang et al., 2020). In principle, more general constraints can be handled by using suitable minimax problem solvers. Such constraints were considered in Noubiap and Seidel (2001) and can represent a variety of forms of prior information. For example, with $\Phi_k = \pm\Psi^\kappa$ for some $\kappa \geq 1$, Γ imposes bounds on prior moments of $\Psi(P)$; with $\Phi_k(P) = \pm\mathbb{1}(\Psi(P) \in I)$ for some known interval I, Γ imposes bounds on the prior probability of $\Psi(P)$ lying in I. Similar prior information on aspects of P_0 other than $\Psi(P_0)$ can also be represented. In addition, since an equality can be equivalently expressed by two inequalities, Γ may also impose equality constraints on prior generalized moments. Such information is commonly used to choose prior distributions in Bayesian settings (Sarma and Kay, 2020). Since we do not require specifying a parametric model or specifying an entire prior distribution for any finite-dimensional summary of P_0, specifying a set Γ of prior distributions in the above form is no more difficult — and often easier — than specifying a single prior distribution, as would be required in a Bayesian approach.

3. Proposed meta-learning algorithms to compute a Γ-minimax estimator

Since both the model space $\mathcal{M}$ and the estimator space $\mathcal{D}$ may be infinite, it is computationally infeasible to directly solve the minimax problem (1) defining

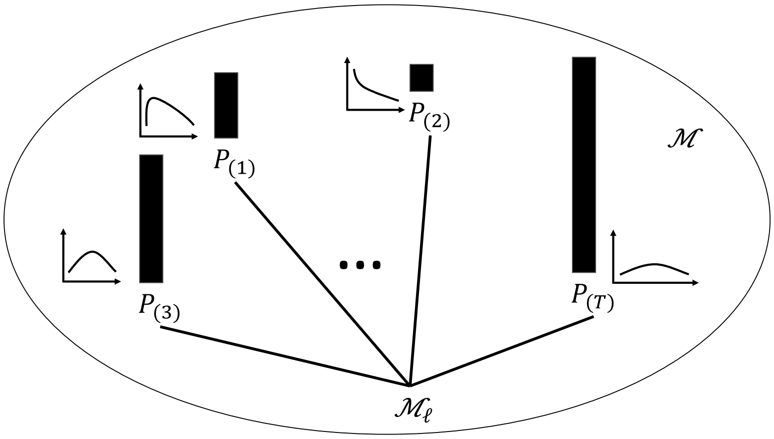

Fig 1. *Illustration of grid $\mathcal{M}_\ell = \{P_{(1)}, P_{(2)}, P_{(3)}, \ldots, P_{(T)}\} \subseteq \mathcal{M}$ approximating the entire model space $\mathcal{M}$. Examples of densities of distributions $P_{(t)}$ ($t = 1, \ldots, T$) in the grid are displayed. A prior distribution with support in $\mathcal{M}_\ell$ is parameterized by the probability mass at each distribution $P_{(t)}$. An example of a prior distribution is displayed as black bars with their heights being proportional to the probability masses.*

a Γ-minimax estimator. Similarly to Noubiap and Seidel (2001), our general strategy is to discretize $\mathcal{M}$ and thus consider prior distributions with discrete supports. Once the supports of prior distributions are discrete, the optimization over prior distributions only involves finitely many parameters, namely the probability masses at support points, and thus is computationally possible. We will show that, when the grid is sufficiently fine, a solution to the discretized minimax problem is close to a solution to the original minimax problem.

Our proposed algorithm consists of two main steps. The first step is to discretize the model space $\mathcal{M}$ and consider an approximating grid $\mathcal{M}_\ell$ instead of the original complicated model space $\mathcal{M}$. This discretization is illustrated in Fig. 1. We will describe $\mathcal{M}_\ell$ in more detail in Section 3.1. In the second step, we consider a set Γ_ℓ of priors with support contained $\mathcal{M}_\ell$ and compute a Γ_ℓ-minimax estimator. We will describe two classes of algorithms to solve this discretized minimax problem in Sections 3.2 and 3.3, respectively.

3.1. Grid-based approximation of Γ-minimax estimators

We first define the discretization of the model space $\mathcal{M}$ that we will use. Let $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ be an increasing sequence of finite subsets of $\mathcal{M}$ such that $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in $\mathcal{M}$. That is, $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ is an increasingly fine grid over $\mathcal{M}$. Since $\mathcal{M}$ is separable, such an $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ necessarily exists. Define

$$\Gamma_\ell := \{\pi \in \Gamma : \pi \text{ has support in } \mathcal{M}_\ell\} \qquad \text{and} \qquad r_{\sup}(d, \Gamma') := \sup_{\pi \in \Gamma'} r(d, \pi)$$

for any $d \in \mathcal{D}$ and $\Gamma' \subseteq \Pi$.

Algorithm 1 describes how the grids $\mathcal{M}_\ell$ are used to compute an approximately Γ-minimax estimator in our proposed algorithms. We will show that the approximation error decays to zero as ℓ grows to infinity. Here and in the rest of the algorithms in the paper, for any real-valued function f, when we assign $\operatorname{argmin}_x f(x)$ or $\operatorname{argmax}_x f(x)$ to a variable, we arbitrarily pick a minimizer or maximizer if there are multiple optimizers. In practice, the user may stop the iteration at some ℓ and use a Γ_ℓ-minimax estimator d^*_ℓ as the output estimator. We discuss the stopping criterion in more detail at the end of this section.

Algorithm 1 Iteratively approximate a Γ-minimax estimator over an increasingly fine grid.

```
1: for $\ell = 1, 2, \ldots$ do
2:     Construct a grid $\mathcal{M}_\ell \subseteq \mathcal{M}$ such that $\mathcal{M}_{\ell-1} \subsetneq \mathcal{M}_\ell$
3:     $d^*_\ell \leftarrow \operatorname{argmin}_{d \in \mathcal{D}} \sup_{\pi \in \Gamma_\ell} r(d, \pi)$
```

We note that the minimax problem in Line 3 of Algorithm 1 is nontrivial to solve, and therefore we propose two algorithms that can solve this minimax problem in Sections 3.2 and 3.3.

We assume that the following conditions hold throughout the rest of the paper.

Condition 1. There exists a limit point $d^* \in \mathcal{D}$ of the sequence $\{d^*_\ell\}_{\ell=1}^\infty$.

Condition 1 holds if the sequence $\{d^*_\ell\}_{\ell=1}^\infty$ eventually falls in a compact set. For example, if $\mathcal{D}$ is a space of neural networks and we take ϱ to be the Euclidean norm in the coefficient space, then we expect Condition 1 to hold if all coefficients are restricted to fall in a bounded set, which is a common restriction in theoretical analyses of neural networks (see, e.g., Goel et al., 2016; Zhang, Lee and Jordan, 2016; Eckle and Schmidt-Hieber, 2019) and often leads to desirable generalization bounds (see, e.g., Bartlett, 1997; Bartlett, Foster and Telgarsky, 2017; Neyshabur et al., 2017). Our theoretical results hold for any limit point d^* in Condition 1, even if there is more than one of them.

Condition 2. The mapping $d \mapsto R(d, P)$ is continuous at d^* for all $P \in \mathcal{M}$.

Condition 2 also often holds. For example, when parameterized using neural networks, all estimators are continuous functions of coefficients for common activation functions such as the sigmoid or the rectified linear unit (ReLU) (Glorot, Bordes and Bengio, 2011) function, and therefore $d \mapsto R(d, P)$ is continuous everywhere.

We next present a sufficient condition to ensure that d^* is Γ-minimax, so that d^*_ℓ is approximately Γ-minimax for sufficiently large ℓ.

Condition 3. We assume that there exists an increasing sequence $\{\Omega_\ell\}_{\ell=1}^\infty$ of subsets of $\mathcal{M}$ such that

1. $\bigcup_{\ell=1}^\infty \Omega_\ell = \mathcal{M}$;
2. for all $\ell = 1, 2, \ldots$ and all $d \in \mathcal{D}$, define $\tilde{\Gamma}_\ell := \{\pi \in \Gamma : \pi \text{ has support in } \Omega_\ell\}$ and $\tilde{\Gamma}_{i|\ell} := \{\pi \in \Gamma : \pi \text{ has support in } \mathcal{M}_i \bigcap \Omega_\ell\}$. For any $\pi \in \tilde{\Gamma}_\ell$ with a fi-

nite support, there exists a sequence $\pi_i \in \tilde{\Gamma}_{i|\ell}$ such that $r(d, \pi_i) \to r(d, \pi)$ as $i \to \infty$.

We note that, in contrast to $\mathcal{M}_\ell$, Ω_ℓ may be an infinite set. We may expect Condition 3 to hold in many cases, especially when $P \mapsto R(d, P)$ is continuous at each $d \in \mathcal{D}$ and the grid $\mathcal{M}_\ell$ contains a variety of distributions that are consistent with prior information represented by Γ. We illustrate this point with two counterexamples in Appendix A. We will check the plausibility of Condition 3 for Example 2 in our simulation and data analysis in Section 5.1 for exemplar prior information; an almost identical argument shows the plausibility of Condition 3 for Example 3.

We now present the theorem on Γ-minimaxity of d^*.

Theorem 1 (Validity of grid-based approximation)**.** *Under Conditions 1–3, d^* is Γ-minimax and*

$$r_{\sup}(d^*_\ell, \Gamma_\ell) \nearrow \min_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) \quad \textit{as} \quad \ell \to \infty.$$

To prove Theorem 1, we utilize a result in Pinelis (2016) to establish that $r_{\sup}(d, \Gamma)$ can be approximated arbitrarily well by a discrete prior in Γ for any $d \in \mathcal{D}$. This is a key ingredient in the proof of Lemma 1, which states that, for any $d \in \mathcal{D}$, $r_{\sup}(d, \tilde{\Gamma}_\ell)$ converges to $r_{\sup}(d, \Gamma)$. Then, we show that the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ is nondecreasing and upper bounded by $\inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$, which is less than or equal to the Γ-maximal Bayes risk $r_{\sup}(d^*, \Gamma)$ of the limit point d^* of $\{d^*_\ell\}_{\ell=1}^\infty$ in Condition 1. Therefore, $r_{\sup}(d^*_\ell, \Gamma_\ell)$ converges to a limit. We finally use a contradiction argument to prove that this limit is greater than or equal to $r_{\sup}(d^*, \Gamma)$, which implies Theorem 1.

We have the following corollary on the uniqueness of the Γ-minimax estimator and the convergence of $\{d^*_\ell\}_{\ell=1}^\infty$ for certain problems.

Corollary 1 (Convergence of Γ_ℓ-minimax estimator)**.** *Suppose that $\mathcal{D}$ is a convex subset of a vector space, $d \mapsto R(d, P)$ is strictly convex for each $P \in \mathcal{M}$, and $r_{\sup}(d, \Gamma)$ is attainable for each $d \in \mathcal{D}$ in the sense that, for all $d \in \mathcal{D}$, there exists a $\pi \in \Gamma$ such that $r(d, \pi) = r_{\sup}(d, \Gamma)$. Under Conditions 1–3, d^* is the unique Γ-minimax estimator and*

$$d^*_\ell \to d^* \quad \textit{as} \quad \ell \to \infty.$$

We prove Corollary 1 by establishing that $d \mapsto r_{\sup}(d, \Gamma)$ is strictly convex.

In practice, the user also needs to specify a stopping criterion for Algorithm 1. In Noubiap and Seidel (2001), the authors recommended computing or approximating $r_{\sup}(d^*_\ell, \Gamma)$ and stop if $r_{\sup}(d^*_\ell, \Gamma)$ is sufficiently close to $r_{\sup}(d^*_\ell, \Gamma_\ell)$. However, the procedure to approximate $r_{\sup}(d^*_\ell, \Gamma)$ in that work relies on the compactness of $\mathcal{M}$, but we do not want to assume this condition because it may restrict the applicability of the method. Therefore, we propose to use the following alternative criterion: stop if $r_{\sup}(d^*_\ell, \Gamma_{\ell+1}) - r_{\sup}(d^*_\ell, \Gamma_\ell) \leq \epsilon$ for a prespecified tolerance level $\epsilon > 0$. This criterion was proposed but not recommended in Noubiap and Seidel (2001) because it does not guarantee that $r_{\sup}(d^*_\ell, \Gamma_\ell)$ is close to $r_{\text{sup}}(d^*, \Gamma)$. For example, if $\mathcal{M}_{\ell+1} \setminus \mathcal{M}_\ell$ is small, it is even possible that $r_{\text{sup}}(d^*_\ell, \Gamma_{\ell+1}) - r_{\text{sup}}(d^*_\ell, \Gamma_\ell) = 0$, but d^*_ℓ is far from being Γ-minimax. In contrast, we recommend this criterion for our proposed methods because we allow more flexibility in model specification, that is, $\mathcal{M}$ need not be compact. We discuss this issue in more detail in Section 4.1.

We finally remark that $r_{\text{sup}}(d, \Gamma_\ell)$ may be difficult to evaluate exactly. Since the risk is often an expectation, we recommend approximating $r_{\text{sup}}(d, \Gamma_\ell)$ for any given d via Monte Carlo as follows: first, estimate risks $R(d, P)$ for all $P \in \mathcal{M}_\ell$ with a large number of Monte Carlo runs; second, estimate the corresponding least favorable prior $\pi_{d,\ell} \in \operatorname{argmax}_{\pi \in \Gamma_\ell} r(d, \pi)$ using the estimated risks; third, estimate the risks $R(d, P)$ $(P \in \mathcal{M}_\ell)$ again with independent Monte Carlo runs, and, finally, calculate $r(d, \pi_{d,\ell})$ with the estimated risks and the estimated least favorable prior. Using two independent estimates of the risk can remove the positive bias that would otherwise arise due to using the same data to estimate the risks and the least favorable prior.

3.2. Computation of an estimator on a grid via stochastic gradient descent with max-oracle

In this section, we present methods to compute a Γ_ℓ-minimax estimator, which corresponds to Line 3 in Algorithm 1. Gradient descent with max-oracle (GDmax) and its stochastic variant (SGDmax), which were presented in Lin, Jin and Jordan (2020), can be used to solve general minimax problems in Euclidean spaces. We focus on SGDmax in the main text and present GDmax in Appendix B. To apply these algorithms to find a Γ_ℓ-m inimax estimator, we need to assume that $\mathcal{D}$ can be parameterized by a subset of a Euclidean space, that is, that for any $d \in \mathcal{D}$, there exists a real vector-valued coefficient $\beta \in \mathbb{R}^D$ such that d may be written as $d(\beta)$. For example, $\mathcal{D}$ may be a neural network class. More discussions on the parameterization of $\mathcal{D}$ can be found in Section 4.2. In this section, in a slight abuse of notation, we define $R(\beta, P) := R(d(\beta), P)$, $r(\beta, \pi) := r(d(\beta), \pi)$ and $r_{\text{sup}}(\beta, \Gamma_\ell) := r_{\text{sup}}(d(\beta), \Gamma_\ell)$ for a coefficient $\beta \in \mathbb{R}^D$, a data-generating mechanism $P \in \mathcal{M}$ and a prior $\pi \in \Gamma$. We assume that $\beta \mapsto R(\beta, P)$ is differentiable for all $P \in \mathcal{M}$, and hence so is $\beta \mapsto r(\beta, \pi)$ for all $\pi \in \Gamma$.

It is often the case that $R(\beta, P)$ is expressed as an expectation. In this case, $R(\beta, P)$ may instead be approximated using Monte Carlo techniques. With ξ being an exogenous source of randomness according to law Ξ, let $\hat{R}(\beta, P, \xi)$ be an unbiased approximation of $R(\beta, P)$ with $\mathbb{E}[\|\nabla_\beta\{\hat{R}(\beta, P, \xi) - R(\beta, P)\}\|^2] \leq \sigma^2 < \infty$, where $\|\cdot\|$ denotes the ℓ_2-norm in Euclidean spaces. Let $\hat{r}(\beta, \pi, \xi) := \int \hat{R}(\beta, P, \xi)\, \pi(\mathrm{d}P)$ for $\pi \in \Gamma_\ell$. In this case, SGDmax (Algorithm 2) may be used to find a (locally) Γ_ℓ-minimax estimator. Note that Algorithm 2 represents a generalization of the nested minimax AMC strategy in Luedtke et al. (2020) to Γ_ℓ-minimax problems.

We next present two conditions needed for the validity of Algorithm 2.

Condition 4. For each $\ell = 1, 2, \ldots$ and all $\beta \in \mathbb{R}^D$, $\beta \mapsto R(\beta, P)$ is Lipschitz continuous with a universal Lipschitz constant L_1 independent of $P \in \mathcal{M}_\ell$.

Algorithm 2 Stochastic gradient descent with max-oracle (SGDmax) to compute a Γ_ℓ-minimax estimator

```
Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$, max-oracle accuracy $\zeta > 0$ and batch size $J$.
for $t = 1, 2, \ldots$ do
    Stochastic maximization: use a stochastic procedure to find $\pi_{(t)} \in \Gamma_\ell$ such that $\mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq \max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \zeta$, where the expectation is over the randomness in stochastic maximization (e.g., variants of stochastic gradient ascent).
    Generate iid copies $\xi_1, \ldots, \xi_J$ of $\xi$.
    Stochastic gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \frac{\eta}{J} \sum_{j=1}^{J} \nabla_\beta \hat{r}(\beta, \pi_{(t)}, \xi_j)|_{\beta=\beta_{(t-1)}}$.
```

Note that Condition 4 differs from Condition 2 in that the former relies on the parameterization of $\mathcal{D}$ in a Euclidean space equipped with the Euclidean norm, while the latter may rely on a different metric on $\mathcal{D}$ such as an L^2-distance.

Condition 5. For each $\ell = 1, 2, \ldots$ and all $\beta \in \mathbb{R}^D$, $\nabla_\beta R(\beta, P)$ is bounded; $\beta \mapsto \nabla_\beta R(\beta, P)$ is Lipschitz continuous with a universal Lipschitz constant L_2 independent of $P \in \mathcal{M}_\ell$.

Under these conditions, using the results in Lin, Jin and Jordan (2020), we can show that SGDmax yields an approximation to a local minimum of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ when the algorithms' hyperparameters are suitably chosen. Before we formally present the theorem, we introduce some definitions related to the local optimality of a potentially nondifferentiable and nonconvex function. A real-valued function f is called q-weakly convex if $x \mapsto f(x) + (q/2)\|x\|^2$ is convex ($q > 0$). The Moreau envelope of a real-valued function f with parameter $q > 0$ is $f_q : x \mapsto \min_{x'} f(x') + \|x' - x\|^2/(2q)$. A point x is an ϵ-stationary point ($\epsilon \geq 0$) of a q-weakly convex function f if $\|\nabla f_{1/(2q)}(x)\| \leq \epsilon$. Similarly, a random point x is an ϵ-stationary point ($\epsilon \geq 0$) of a q-weakly convex function f in expectation if $\mathbb{E}[\|\nabla f_{1/(2q)}(x)\|] \leq \epsilon$. If x is an ϵ-stationary point in expectation, we may conclude that it is an ϵ-stationary point with high probability by Markov's inequality. Lemma 3.8 in Lin, Jin and Jordan (2020) shows that an ϵ-stationary point of f is close to a point x' at which f has at least one small subgradient for small ϵ, so that $f(x')$ is close to a local minimum. In other words, if an algorithm outputs an estimator $\hat{d} = d(\hat{\beta})$ such that $\hat{\beta}$ is an ϵ-stationary point of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$, then we know that $r_{\sup}(\hat{\beta}, \Gamma_\ell)$ is close to a local minimum of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$.

We next present the validity result for Algorithm 2.

Theorem 2 (Validity of SGDmax (Algorithm 2))**.** *Suppose that Conditions 1–2 and 4–5 hold. Let $\epsilon > 0$ be fixed and define $\Delta := (r_{\sup})_{1/(2L_1)}(\beta_{(0)}) - \min_{\beta \in \mathbb{R}^D} (r_{\sup})_{1/(2L_1)}(\beta)$, where we recall that $(r_{\sup})_{1/(2L_1)}$ is the Moreau envelope of $r_{\sup}$ with parameter $1/(2L_1)$. In Algorithm 2, with $\eta = \epsilon^2/[L_1(L_2^2 + \sigma^2)]$, $\zeta = \epsilon^2/(24L_1)$ and $J = 1$, $\beta_{(t)}$ is an ϵ-stationary point of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ in expectation for $t = O(L_1(L_2^2 + \sigma^2)\Delta/\epsilon^4)$, and is thus close to a local minimum of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ with high probability.*

The assumption that the batch size $J = 1$ is purely for convenience since increasing J corresponds to decreasing variance σ^2. To run Algorithm 2 in practice, the user only needs to specify tuning parameters in Line 1 and all other constants in Theorem 2 need not be known. In general, a small learning rate η, a stringent accuracy ζ, and a large batch size J make Algorithm 2 likely to eventually reach an approximation of a local minimum of $\beta \mapsto r_{\text{sup}}(\beta, \Gamma_\ell)$, but computation time might increase. Similar to most numeric optimization algorithms, fine-tuning is needed to achieve a balance between convergence guarantee and computation time, but a conservative choice of tuning parameters would typically result in convergence at the cost of computation time.

We note that Line 3 in Algorithm 2 may be inconvenient to implement because linear program solvers often do not use stochastic optimization. Therefore, we propose to use a convenient variant (Algorithm 6 in Appendix B), where the stochastic maximization step (Line 3 in Algorithm 2) is replaced by solving a linear program where the objective is approximated via Monte Carlo. This variant has similar validity under similar conditions. We also note that the two uniform Lipschitz continuity conditions (4 and 5) heavily rely on the fact that $\mathcal{M}_\ell$ is finite and the compactness of a set containing the coefficients. Nevertheless, the latter compactness restriction is common in theoretical analyses of neural networks (see, e.g., Goel et al., 2016; Zhang, Lee and Jordan, 2016; Eckle and Schmidt-Hieber, 2019). Moreover, these two conditions are sufficient conditions for the validity of the gradient-based methods, namely SGDmax, our variant of SGDmax and GDmax; a guarantee similar to these validity results might hold when two conditions are violated.

We finally remark that other algorithms similar to SGDmax can be applied, for example, (stochastic) gradient descent ascent with projection (Lin, Jin and Jordan, 2020), (stochastic) mirror descent ascent, or accelerated (stochastic) mirror descent ascent (Huang, Wu and Huang, 2021). It is of future research interest to develop gradient-based methods to solve minimax problems with convergence guarantees under weaker conditions.

3.3. Computation of an estimator on a grid via fictitious play

The algorithms in Section 3.2 may be convenient in many cases, but the requirements such as parameterization of the space $\mathcal{D}$ of estimators in a Euclidean space, differentiability of the risk function R with respect to the coefficients β, and uniform Lipschitz continuity may be restrictive for certain problems. In this section, we propose an alternative algorithm, fictitious play, that avoids these requirements. We also present its convergence results.

Brown (1951) introduced fictitious play as a means to find the value of a zero-sum game, that is, the optimal mixed strategy for both players and their expected gains. Robinson (1951) then proved that fictitious play can be used to iteratively solve a two-player zero-sum game for a saddle point that is a pair of mixed strategies where both players have finitely many pure strategies. Our problem of finding a Γ-minimax estimator may also be viewed as a two-player zero-sum game where one player chooses a prior from Γ and the other player chooses an estimator from $\mathcal{D}$. If we assume that, for the Γ-minimax problem at hand, the pair of both players' optimal strategies is a saddle point, which holds in many minimax problems (e.g., v. Neumann, 1928; Fan, 1953; Sion, 1958), then fictitious play may also be used to find a Γ-minimax estimator. Since Γ may be too rich to allow for feasible implementation of fictitious play, we propose to use this algorithm to find a Γ_ℓ-minimax estimator.

In the fictitious play algorithm in Robinson (1951), the two players take turns to play the best pure strategy against the mixture of the opponent's historic pure strategies, and the final output is a pair of mixtures of the two players' historic pure strategies. Since this algorithm aims to find minimax mixed strategies, we consider stochastic estimators. That is, consider the Borel σ-field $\mathcal{F}$ over $\mathcal{D}$ and let $\mathit{\Pi}$ denote the set of all probability distributions on the measurable space $(\mathcal{D}, \mathcal{F})$. We define $\overline{\mathcal{D}}$ to be the space of stochastic estimators with each element taking the following form: first draw an estimator from $\mathcal{D}$ according to a distribution $\varpi \in \mathit{\Pi}$ with an exogenous random mechanism and then use the estimator to obtain an estimate based on the data. Note that we may write any $\overline{d} \in \overline{\mathcal{D}}$ as $\overline{d}(\varpi)$ for some $\varpi \in \mathit{\Pi}$. We consider estimators in $\overline{\mathcal{D}}$ throughout this section, with the definition of Γ-minimaxity extended in the natural way, so that $\overline{d}^* = \overline{d}(\varpi^*) \in \overline{\mathcal{D}}$ is Γ-minimax if $r_{\sup}(\overline{d}^*, \Gamma) = \min_{\overline{d}\in\overline{\mathcal{D}}} r_{\sup}(\overline{d}, \Gamma)$; we similarly extend all other definitions from Section 2. We assume that there exists $\pi_\ell^* \in \Gamma_\ell$ $(\ell = 1, 2, \ldots)$ such that

$$r(\overline{d}^*, \pi_\ell^*) = \sup_{\pi\in\Gamma_\ell} \inf_{\overline{d}\in\overline{\mathcal{D}}} r(\overline{d}, \pi) = \inf_{\overline{d}\in\overline{\mathcal{D}}} \sup_{\pi\in\Gamma_\ell} r(\overline{d}, \pi). \tag{2}$$

In other words, $(\overline{d}^*, \pi_\ell^*)$ is a saddle point of r in $\overline{\mathcal{D}} \times \Gamma_\ell$. Under this condition and the further conditions that $\mathcal{D}$ is convex and $d \mapsto R(d, P)$ is convex for all $P \in \mathcal{M}$, it is possible to use a Γ-minimax estimator over the richer class $\overline{\mathcal{D}}$ of stochastic estimators to derive a Γ-minimax estimator over the original class $\mathcal{D}$. Indeed, for any $\overline{d}(\varpi) \in \overline{\mathcal{D}}$ and $P \in \mathcal{M}$, by Jensen's inequality, $R(\overline{d}(\varpi), P) = \int R(d, P)\, \varpi(\mathrm{d}d) \geq R(\underline{d}(\varpi), P)$ where $\underline{d}(\varpi) := \int d\, \varpi(\mathrm{d}d) \in \mathcal{D}$ is the average of the stochastic estimator $\overline{d}(\varpi)$; that is, the risk of $\underline{d}(\varpi)$ is never greater than that of $\overline{d}(\varpi)$. Therefore, we may use the fictitious play algorithm to compute $\overline{d}(\varpi_\ell^*)$ for each ℓ and further apply Algorithm 1 to compute $\overline{d}(\varpi^*)$. After that, we may take $\underline{d}(\varpi^*)$ as the final output deterministic estimator.

Algorithm 3 presents the fictitious play algorithm for finding a Γ_ℓ-minimax estimator in $\overline{\mathcal{D}}$. Note that Γ_ℓ is convex, and hence π always lies in Γ_ℓ throughout the iterations. In practice, we may initialize ϖ as a point mass at an initial estimator in $\mathcal{D}$. In addition, similarly to Robinson (1951), we may replace Line 5 with $d_{(t)}^\dagger \leftarrow \mathrm{argmin}_{d\in\mathcal{D}}\, r(d, \pi_{(t)})$, that is, minimizing the Bayes risk with the most recently updated prior rather than with the previous prior.

We next present a convergence result for this algorithm.

Theorem 3 (Validity of fictitious play (Algorithm 3))**.** *Assume that there exists a compact subset $\bar{\mathcal{D}}$ of $\mathcal{D}$ that contains all $d_{(t)}^\dagger$ $(t = 1, 2, \ldots)$. Under Conditions 1–2, it holds that*

$$r(d_{(t)}^\dagger, \pi_{(t-1)}) \leq r(\overline{d}(\varpi_\ell^*), \pi_\ell^*) \leq r(\overline{d}(\varpi_{(t-1)}), \pi_{(t)}^\dagger)$$

Algorithm 3 Fictitious play to compute a Γ_ℓ-minimax stochastic estimator

```
1: Initialize $\varpi_{(0)} \in \mathit{\Pi}$ and $\pi_{(0)} \in \Gamma_\ell$.
2: for t=1,2,... do
3:     $\pi^\dagger_{(t)} \leftarrow \operatorname{argmax}_{\pi \in \Gamma_\ell} r(\overline{d}(\varpi_{(t-1)}), \pi)$
4:     $\pi_{(t)} \leftarrow \frac{t-1}{t}\pi_{(t-1)} + \frac{1}{t}\pi^\dagger_{(t)}$
5:     $d^\dagger_{(t)} \leftarrow \operatorname{argmin}_{d \in \mathcal{D}} r(d, \pi_{(t-1)})$
6:     $\varpi_{(t)} \leftarrow \frac{t-1}{t}\varpi_{(t-1)} + \frac{1}{t}\delta(d^\dagger_{(t)})$, where $\delta(d)$ denotes a point mass at $d \in \mathcal{D}$.
```

for all t and

$$\lim_{t\to\infty} \left[r(\overline{d}(\varpi_{(t-1)}), \pi^\dagger_{(t)}) - r(d^\dagger_{(t)}, \pi_{(t-1)}) \right] = 0.$$

Consequently, the Γ_ℓ-maximal risk of $\overline{d}(\varpi_{(t)})$ converges to the Γ_ℓ-minimax risk, that is,

$$r_{\sup}(\overline{d}(\varpi_{(t-1)}), \Gamma_\ell) \to r_{\sup}(\overline{d}(\varpi^*_\ell), \Gamma_\ell) \quad \textit{as} \quad t \to \infty.$$

Robinson (1951) proved a similar case for two-player zero-sum games where each player has finitely many pure strategies. In contrast, in our problem, each player may have infinitely many pure strategies. A natural attempt to prove Theorem 3 would be to consider finite covers of $\bar{\mathcal{D}}$ and Γ_ℓ, that is, $\bar{\mathcal{D}} = \bigcup_{i=1}^{I} \mathcal{D}_i$ and $\Gamma_\ell = \bigcup_{j=1}^{J} \Pi_j$, such that the range of $r(d, \pi)$ in each $\mathcal{D}_i$ and Π_j is small (say less than ϵ), bin pure strategies into these subsets, and then apply the argument in Robinson (1951) to these bins. The collection of $\mathcal{D}_i$ and Π_j may be viewed as finitely many approximated pure strategies to Γ_ℓ and $\bar{\mathcal{D}}$ up to accuracy ϵ, respectively. Unfortunately, we found that this approach fails. The problem arises because Robinson (1951) inducted on I and J, and, after each induction step, the corresponding upper bound becomes twice as large. Unlike the case with finitely many pure strategies that was considered in Brown (1951) and Robinson (1951), as the desired approximation accuracy ϵ approaches zero, the numbers of approximated pure strategies, I and J, may diverge to infinity, and so does the number of induction steps. Therefore, the resulting final upper bound is of order $2^{I+J}\epsilon$ and generally does not converge to zero as ϵ tends to zero. To overcome this challenge, we instead control the increase in the relevant upper bound after each induction step more carefully so that the final upper bound converges to zero as ϵ decreases to zero, despite the fact that I and J may diverge to infinity.

We remark that, because Line 5 of Algorithm 3 typically involves another layer of iteration in addition to that over t, this algorithm will often be more computationally intensive than SGDmax. Nevertheless, Algorithm 3 provides an approach to construct Γ_ℓ-minimax estimators in cases where these other algorithms cannot be applied, for example, in settings where the risk is not differentiable in the parameters indexing the estimator or uniform Lipschitz conditions fail. In our numerical experiments, we have implemented this algorithm in the context of mean estimation (Appendix C).

4. Considerations in implementation

4.1. Considerations when constructing the grid over the model space

By Theorem 1, $r_{\sup}(d^*_\ell, \Gamma_\ell) \nearrow \min_{d\in\mathcal{D}} r_{\sup}(d, \Gamma)$ whenever Conditions 1–3 hold and the increasing sequence $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ is such that $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in $\mathcal{M}$. Though this guarantee holds for all such sequences $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$, in practice, judiciously choosing this sequence of grids of distributions can lead to faster convergence. In particular, it is desirable that the least favorable prior Γ_ℓ puts mass on some of the distributions in $\mathcal{M}_\ell \backslash \mathcal{M}_{\ell-1}$ since, if this is not the case, then d^*_ℓ will be the same as $d^*_{\ell-1}$. While we may try to arrange for this to occur by adding many new points when enlarging $\mathcal{M}_{\ell-1}$ to $\mathcal{M}_\ell$, it may not be likely that any of these points will actually modify the least favorable prior unless they are carefully chosen.

To better address this issue, we propose to add grid points using a Markov chain Monte Carlo (MCMC) method. Our intuition is that, given an estimator d, the maximal Bayes risk is likely to significantly increase if we add distributions that (i) have a high risk for d, and (ii) are consistent with prior information so that there exists some prior such that these distributions lie in a high-probability region. We propose to use the MCMC algorithm to bias the selection of distributions in favor of those with the above characteristics. Let $\tau : \mathcal{M} \to [0, \infty)$ denote a function such that $\tau(P) > \tau(P')$ if P is more consistent with prior information than P'. For example, given a prior mean μ of some real-valued summary $\Psi(P)$ of P and an interval I that contains $\Psi(P)$ with prior probability at least 95%, we may choose $\tau : P \mapsto \phi(\Psi(P))$, where ϕ is the density of a normal distribution that has mean μ and places 95% of its probability mass in I. We call τ a pseudo-prior. Then, with the current estimator being d, we wish to select distributions P for which $R(d, P)\tau(P)$ is large. We may use the Metropolis-Hastings-Green algorithm (Metropolis et al., 1953; Hastings, 1970; Green, 1995) to draw samples from a density proportional to $P \mapsto R(d, P)\tau(P)$. We then let $\mathcal{M}_\ell$ be equal to the union of $\mathcal{M}_{\ell-1}$ and the set containing all unique distributions in this sample.

Details of the proposed scheme are provided in Algorithm 4. To use this proposed algorithm, we rely on it being possible to define a sequence of parametric models $\{\tilde{\Omega}_\ell\}_{\ell=1}^\infty$ such that $\tilde{\mathcal{M}} := \cup_{\ell=1}^\infty \tilde{\Omega}_\ell$ is dense in $\mathcal{M}_\ell$—this is possible in many interesting examples (see, e.g., Chen, 2007). When combined with separability of $\mathcal{M}$, this condition enables the definition of an increasing sequence of grids of distributions $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ such that, for each ℓ, $\mathcal{M}_\ell \subseteq \tilde{\mathcal{M}}$.

The following theorem on distributional convergence follows from that for the Metropolis-Hastings-Green algorithm (see Section 3.2 and 3.3 of Green, 1995).

Theorem 4 (Validity of MCMC algorithm (Algorithm 4))**.** *Suppose that $P \mapsto R(d^*_{\ell-1}, P)\tau(P)$ is bounded and integrable with respect to some measure μ on $\tilde{\mathcal{M}}$ and let $\mathscr{L}$ denote the probability law on $\tilde{\mathcal{M}}$ whose density function with respect to μ is proportional to this function. Suppose that the MCMC is constructed such that the Markov chain is irreducible and aperiodic. Then, $P_{(t)}$ converges*

Algorithm 4 MCMC algorithm to construct $\mathcal{M}_\ell$

```
Require: Previous grid $\mathcal{M}_{\ell-1}$, current estimator $d^*_{\ell-1}$ and number $T$ of iterations. We define
    $\mathcal{M}_{-1} := \emptyset$. An initial estimator $d^*_0$ must be available if $\ell = 1$.
Initialize $P_{(0)} \in \tilde{\mathcal{M}}$.
for $t = 1, 2, \ldots, T$ do
    Propose a distribution $P' \in \tilde{\mathcal{M}}$ from $P_{(t-1)}$
    Calculate the MCMC acceptance probability $p_{\text{accept}}$ of $P'$ for target density $P \mapsto R(d^*_{\ell-1}, P)\tau(P)$
    With probability $p_{\text{accept}}$, accept $P'$ and $P_{(t)} \leftarrow P'$
    if $P'$ is not accepted then
        $P_{(t)} \leftarrow P_{(t-1)}$
$\mathcal{M}_\ell \leftarrow$ unique elements of the multiset $\mathcal{M}_{\ell-1} \bigcup \{P_{(1)}, P_{(2)}, \ldots, P_{(T)}\}$
```

weakly to $\mathscr{L}$ as t $\to \infty$.

Therefore, if $\mathscr{L}$ corresponds to a continuous distribution with nonzero density over the parameter space of $\tilde{\mathcal{M}}$, then Theorem 4 implies that $\bigcup_{\ell=1}^{\infty} \mathcal{M}_\ell$ is dense in $\mathcal{M}$, as required by Algorithm 1.

Implementing Algorithm 4 relies on the user making several decisions. These decisions include the choice of the pseudo-prior τ and the technique used to approximate the risk $R(d, P)$ to a reasonable accuracy. Fortunately, regardless of the decisions made, Theorem 1 suggests that $r_{\text{sup}}(d^*_\ell, \Gamma_\ell) \nearrow \min_{d \in \mathcal{D}} r_{\text{sup}}(d, \Gamma)$ for a wide range of sequences $\{\mathcal{M}_\ell\}_{\ell=1}^{\infty}$. Indeed, all that theorem requires on this sequence is that the grid $\mathcal{M}_\ell$ becomes arbitrarily fine as ℓ increases. Though the final decisions made are not important when ℓ is large, we still comment briefly on the decisions that we have made in our experiments, First, we have found it effective to approximate $R(d, P)$ via a large number of Monte Carlo draws. Second, in a variety of settings, we have also identified, via numerical experiments, candidate pseudo-priors that balance high risk and consistency with prior information (see Sections 5.1 and 5.2 for details).

4.2. Considerations when choosing the space of estimators

It is desirable to consider a rich space $\mathcal{D}_0$ of estimators to obtain an estimator with low maximal Bayes risk, and thus good general performance. However, to make numerically constructing these estimators computationally feasible, we usually have to consider a restricted space $\mathcal{D}$ of estimators. This approximation is justified because, if estimators in $\mathcal{D}$ can approximate the *Gamma*-minimax estimator in $\mathcal{D}_0$ well, then we expect the resulting excess maximal Bayes risk is small.

Feedforward neural networks (or neural networks for short) are natural options for the space of estimators because of their universal approximation property (e.g., Hornik, 1991; Csáji, 2001; Hanin and Sellke, 2017; Kidger and Lyons, 2020). However, training commonly used neural networks can be computationally intensive. Moreover, a space of neural networks is typically nonconvex, and hence it may be difficult to find a global minimizer of the maximal Bayes risk

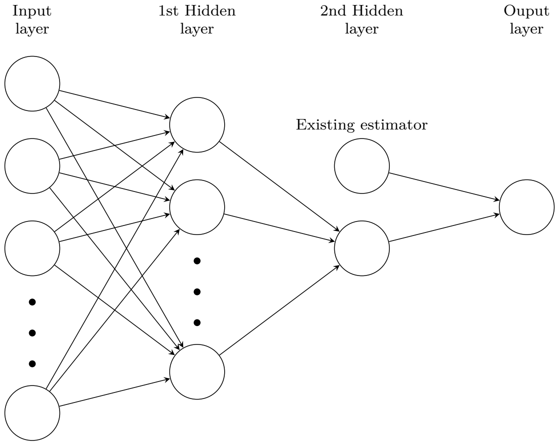

Fig 2. *Example of neural network estimator architecture utilizing an existing estimator. The arrows from the input nodes to the existing estimator are omitted from this graph.*

even if the risk is convex in the estimator. Therefore, the learned estimator might not perform well.

To help overcome this challenge, we advocate for utilizing available statistical knowledge when designing the space of estimators. We call estimators that take this form *statistical knowledge networks*. In particular, if a simple estimator is already available, we propose to use neural networks with such an estimator as a node connected to the output node. An example of such an architecture is presented in Fig. 2. In this sample architecture, each node is an activation function such as the sigmoid or the rectified linear unit (ReLU) (Glorot, Bordes and Bengio, 2011) function applied to an affine transformation of the vector containing the ancestors of the node. The only exception is the output node, which is again an affine transformation of its ancestors but uses the identity activation function. When training the neural network, we may initialize the affine transformation in the output layer to only give weight to the simple estimator. Under this approach, the space of estimators is a set of perturbations of an existing simple estimator. Although we may still face the challenge of nonconvexity and local optimality, we can at least expect to improve the initial simple estimator.

In the simulation we describe in Appendix C, we compared the empirical performance of several spaces of estimators. This simulation concerns the simple problem of estimating the mean of a true distribution whose support has known bounds (Example 1), and the existing simple estimator we use in the statistical neural network is the sample mean. Fig. 3 presents the trajectory of estimated Bayes risks. As shown in subfigures (b)–(d), using the statistical knowledge network, the estimator is almost Γ-minimax after a few iterations; on the other hand, it took about 1000 iterations for the feedforward neural network to reach an approximately Γ-minimax estimator. Therefore, in this simple problem where the true Γ-minimax estimator is a shifted and scaled sample mean, statistical

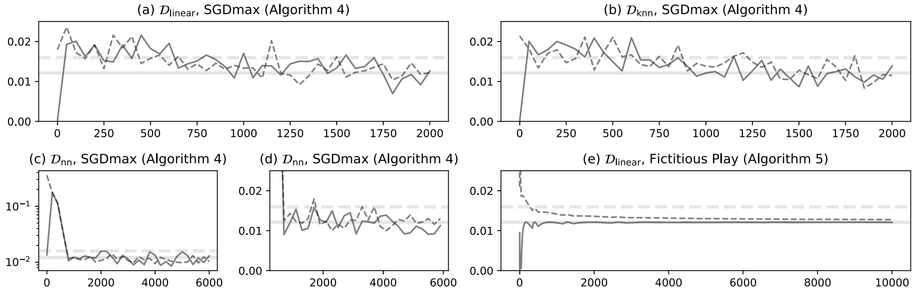

FIG 3. *Estimated Bayes risks of the estimator over iterations when computing a* Γ_1*-minimax estimator. The lines are the current Bayes risks (y-axis) over iterations (x-axis) (unbiased estimates with 50 Monte Carlo runs for Algorithm 6; exact values for Algorithm 3). The solid lines are the Bayes risks after an update in the estimator to decrease the Bayes risk. The dashed lines are the Bayes risks after an update in the prior to increase the Bayes risk. The two horizontal lines are the Bayes risk of the sample mean (dashed) and* d^* *(solid), respectively, for* π^**. For ease of visualization, in subfigures (a) and (b), the Bayes risks are plotted every 50 iterations; in subfigures (c) and (d), the Bayes risks are plotted every 200 iterations; subfigure (d) contains the part in subfigure (c) after 500 iterations.*

knowledge substantially reduced the number of iterations required to obtain an approximately Γ-minimax estimator. For more complicated problems, we expect statistical knowledge to further help improve the performance of the computed estimator.

We note that we might overcome the challenge of nonconvexity and local optimality by using an extreme learning machine (ELM) (Huang, Zhu and Siew, 2006) to parameterize the estimator. ELMs are neural networks for which the weights in hidden nodes are randomly generated and are held fixed, and only the weights in the output layer are trained. Thus, the space of ELMs with a fixed architecture and fixed hidden layer weights is convex. Like traditional neural networks, ELMs have the universal approximation property (Huang, Chen and Siew, 2006). In addition, Corollary 1 may be applied to an ELM so that the Γ_ℓ-minimax estimator may converge to the Γ-minimax estimator. As for traditional neural networks, we may incorporate knowledge of existing statistical estimators into an ELM.

We finally remark that, besides computational intensity when constructing (i.e., learning) a Γ-minimax estimator, another important factor to be considered when choosing $\mathcal{D}$ is the computational intensity to evaluate the learned estimator at the observed dataset. This is another reason for our choosing neural networks or ELMs as the space of estimators. Indeed, existing software packages (e.g., Paszke et al., 2019) make it easy to leverage graphics processing units to efficiently evaluate the output of neural networks for any given input. Therefore, if the existing estimator being used is not too difficult to compute, then estimators parameterized using similar architectures to that displayed in Figure 2 will be able to be computed efficiently in practice. This efficiency may be especially important in settings where the estimator will be applied to many datasets, so that the cost of learning the estimator is amortized and the main computational expense is evaluating the learned estimator.

5. Simulations and data analyses

We illustrate our methods in Examples 1–3. A toy example of Example 1 is presented in Appendix C. We focus on the more complex Examples 2 and 3 in this section.

5.1. *Prediction of the expected number of new categories*

We apply our proposed method to Example 2. In the simulation, we set the true population to be an infinite population with the same categories and same proportions as the sample studied in Miller and Wiegert (1989), which consists of 1088 observations in 188 categories. This setting is the same as the simulation setting in Shen, Chao and Lin (2003). We set the sample size to be $n = 100$ and the size of the new sample to be $m = 200$. In this setting, the expected number of new categories in the new sample unconditionally on the observed sample, namely $\Phi(P_0) := \mathbb{E}_{P_0}[\Psi(P_0)(\mathbf{X}^*)]$, can be analytically computed and equals 48.02. We note that this quantity can also be computed via simulation: (i) sample n and m individuals with replacement from the dataset in Miller and Wiegert (1989), (ii) count the number of new categories in the second sample, and (iii) repeat steps (i) and (ii) many times and compute the average.

It is well known that this prediction problem is difficult when $m > n$, and we run this simulation to investigate the potential gain from leveraging prior information by computing a Gamma-minimax estimator for such difficult or even ill-posed problems. We consider three sets of prior information:

1. strongly informative: prior mean of $\Phi(P)$ in $[45, 50]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[40, 55]$;
2. weakly informative: prior mean of $\Phi(P)$ in $[40, 55]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[30, 65]$; and
3. almost noninformative: prior mean of $\Phi(P)$ in $[35, 60]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[20, 75]$.

We note that a traditional Bayesian approach would require specifying a prior on $\mathcal{M}$, including the total number of categories and the proportion of each category, which may be difficult in practice.

We check the plausibility of Condition 3 in this context. We take the strongly informative prior information as an example. Take Ω_ℓ to be the collection of multinomial distributions with at most ℓ categories. It is obvious that $\bigcup_{\ell=1}^\infty \Omega_\ell = \mathcal{M}$. Let $d \in \mathcal{D}$ be fixed and $\pi \in \tilde{\Gamma}_\ell$ be a fixed prior with finite support, that is, $\pi = \sum_{j=1}^J q_j \delta(Q_j)$ where $\delta(\cdot)$ denotes the point mass distribution, $Q_j \in \Omega_\ell$, $q_j > 0$ and $\sum_{j=1}^J q_j = 1$. Let $\epsilon > 0$ be an arbitrary small number such that $\sum_{j=1}^J q_j \Phi(Q_j) \leq 50 - \epsilon$ or $\sum_{j=1}^J q_j \Phi(Q_j) \geq 45 + \epsilon$. Since $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in

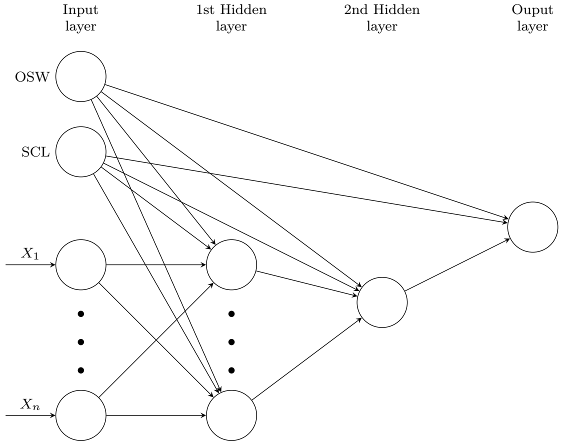

FIG 4. *Architecture of the neural network estimator of the expected number of new categories. X_k: number of categories with k observations; OSW: the estimator proposed in Orlitsky, Suresh and Wu (2016); SCL: the estimator proposed in Shen, Chao and Lin (2003). The arrows from data $(X_1, \dots, X_n)$ to the OSW and SCL estimators are omitted from this graph.*

$\mathcal{M}$ and Φ is continuous, there exists a sufficiently large i such that, for every distribution Q_j, there exists $P_j \in \mathcal{M}_i \cap \Omega_\ell$ satisfying the following:

- $|\Phi(P_j) - \Phi(Q_j)| \leq \epsilon$;
- if $\Phi(Q_j) \in [40, 55]$, then $\Phi(P_j) \in [40, 55]$;
- $|R(d, P_j) - R(d, Q_j)| \leq \epsilon$.

Take π_i to be $\sum_{j=1}^{J} q_j \delta(P_j)$. Then it is easy to verify that $|\sum_{j=1}^{J} q_j \Phi(P_j) - \sum_{j=1}^{J} q_j \Phi(Q_j)| \leq \epsilon$ and thus $\sum_{j=1}^{J} q_j \Phi(P_j) \in [45, 50]$; moreover, $\Phi(Q_j) \in [40, 55]$ implies that $\Phi(P_j) \in [40, 55]$ and therefore $\sum_{j=1}^{J} q_j \mathbb{1}(\Phi(P_j) \in [40, 55]) \geq \sum_{j=1}^{J} q_j \mathbb{1}(\Phi(Q_j) \in [40, 55]) \geq 95\%$. Thus, $\pi_i \in \tilde{\Gamma}_{i|\ell}$. Moreover, $|r(d, \pi) - r(d, \pi_i)| \leq \epsilon$. Therefore, $r(d, \pi_i) \to r(d, \pi)$ as $i \to \infty$ and Condition 3 holds.

We design the architecture of the neural network estimator as in Fig. 4. We choose two existing estimators (referred to as the OSW and SCL estimators, respectively) proposed by Orlitsky, Suresh and Wu (2016) and Shen, Chao and Lin (2003) as human knowledge inputs to the architecture. We use the ReLU activation function. There are 50 hidden nodes in the first hidden layer. We initialize the neural network that we train to output the average of these two existing estimators.

We use Algorithm 4 to construct $\mathcal{M}_\ell$. There are 2000 grid points in $\mathcal{M}_1$, and we add 1000 grid points each time we enlarge the grid. When generating $\mathcal{M}_1$, we chose the starting point to be a distribution $P_{(0)}$ with 146 categories and $\Phi(P_{(0)}) = 49.9$. The choice of this starting point $P_{(0)}$ was quite arbitrary. We first generated a sample from P_0 and treated it as data from a pilot study. We then came up with a distribution $P_{(0)}$ such that five random samples generated from $P_{(0)}$ all appear qualitatively similar to the pilot data. In practice, this starting point can be chosen based on prior knowledge. Our chosen grid sizes for Algorithm 4 were quite arbitrary. For $\mathcal{M}_1$, the generated distributions $P_{(t)}$ appear similar for all t, and thus the initial grid size 2000 and the increment size 1000 appeared sufficient. Smaller grid sizes would simply lead to more iterations in Algorithm 1, which effectively increases the grid size. We selected the log pseudo-prior as a weighted sum of two log density functions: (i) a normal distribution with the mean being the midpoint of the interval constraint on the prior mean of $\Phi(P)$ and central 95% probability interval being the interval with at least 95% prior probability, (ii) a negative-binomial distribution of the total number of categories with success probability 0.995 and 2 failures until the Bernoulli trial is stopped so that the mode and the variance are approximately 200 and 8×10^4, respectively. These log-densities are provided weights 30 and 10, respectively. We selected the weights based on the empirical observation that distributions with only a few categories tend to have high risks, but these distributions are relatively inconsistent with prior information and may well be given almost negligible probability weight in a computed least favorable prior, thus contributing little to computing a Γ-minimax estimator. We chose the aforementioned weights so that Algorithm 4 can explore a fairly large range of distributions and does not generate too many distributions with too few categories.

We use Algorithm 6 with learning rate $\eta = 0.005$ and batch size $J = 30$ to compute Γ_ℓ-minimax estimators. The number of iterations is 4,000 for Γ_1 and 200 for Γ_ℓ ($\ell > 1$). The stopping criterion in Algorithm 1 is that the estimated maximal Bayes risk with 2000 Monte Carlo runs does not relatively increase by more than 2% or absolutely increase by more than 0.0001. We chose the aforementioned tuning parameters based on the prior belief that at least one of OSW and SCL estimators should perform reasonably well, but the performance of SGDmax (Algorithm 6) and Algorithm 4 might be sensitive to tuning parameters. Thus, the network we used is neither deep nor wide. We chose a moderately small learning rate and a large number of iterations for SGDmax. Our chosen learning rate and chosen number of iterations led to a trajectory of estimated Bayes risks that approximately reached a plateau with small fluctuations, suggesting that the obtained estimator is approximately Γ_1-minimax (see Fig. 5). In practice, such trajectory plots may help tune the learning rate and the number of iterations.

We also run additional simulations to investigate the sensitivity of our methods to tuning parameter selections. We present these simulations in Appendix D. The results suggest that our methods may be insensitive to tuning parameter selections.

We examine the performance of the OSW estimator, the SCL estimator and our trained Γ-minimax estimator by comparing their risks under our set data-generating mechanism computed with 20000 Monte Carlo runs. We also compare their Bayes risks under the computed prior from Algorithm 6 using the last and finest grid in the computation with 20000 Monte Carlo runs. We present the results in Table 1. In this simulation experiment, our Γ-minimax estimator substantially reduces the risk compared to two existing estimators. The Γ-minimax

TABLE 1

Risks and Bayes risks of estimators. $R(d, P_0)$: risk of the estimator under the true data-generating mechanism P_0. $r(d, \hat{\pi}^*)$: Bayes risk under prior $\hat{\pi}^*$, the computed prior from Algorithm 6 in the last and finest grid in the computation.

Strength of prior	Estimator	$R(d, P_0)$	$r(d, \hat{\pi}^*)$
strong	OSW	265	303
	SCL	146	159
	Γ-minimax	18	35
weak	OSW	265	328
	SCL	146	184
	Γ-minimax	17	61
almost none	OSW	265	293
	SCL	146	124
	Γ-minimax	24	81

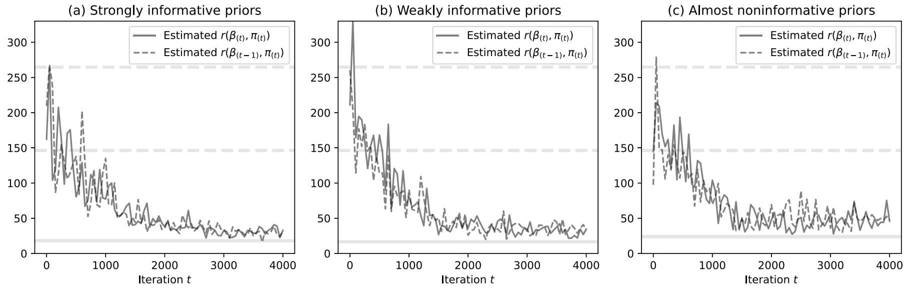

FIG 5. *Estimated Bayes risks of the estimator over iterations when computing a Γ_1-minimax estimator. The lines are unbiased estimates of the current Bayes risks (y-axis) with 30 Monte Carlo runs over iterations (x-axis). The two dashed horizontal lines are the risks of the OSW (upper) and the SCL (lower) estimators, respectively, under P_0 in the simulation. The solid horizontal line is the risk of the computed Γ-minimax estimator under P_0. For clearness of visualization, the estimated Bayes risks are plotted every 50 iterations.*

estimator also has the lowest Bayes risk in all cases. Therefore, incorporating fairly informative prior knowledge into the estimator may lead to a significant improvement in predicting the number of new categories. We expect similar substantial improvement for difficult or even ill-posed statistical problems by incorporating prior knowledge.

Fig. 5 presents the unbiased estimator of Bayes risks over iterations when computing a Γ_1-minimax estimator. The Bayes risks appear to have a decreasing trend and to approach a liming value. Over iterations, the Bayes risks decrease by a considerable amount. The limiting value of the Bayes risks appears to be slightly higher than the risk of the computed Γ-minimax estimator under P_0. This might indicate that P_0 is not an extreme distribution that yields a high risk.

We also apply the above methods to analyze the dataset studied in Miller and Wiegert (1989), which is used as the true population in the simulation. Based on this sample consisting of $n = 1088$ observations in 188 categories, we use var-

Table 2

Predicted number of new categories (rounded to the nearest integer) in a new sample with size 2000 based on the sample with size 1088 studied in Miller and Wiegert (1989). The strength of prior information in Γ-minimax estimators is shown in brackets.

Estimator	Predicted # new categories
OSW	72
SCL	51
Γ-minimax (strong)	57
Γ-minimax (weak)	57
Γ-minimax (almost none)	58

ious methods to predict the number of new categories that would be observed if another $m = 2000$ observations were to be collected. We train Gamma-minimax estimators using exactly the same tuning parameters as those in the above simulation, except that the starting point in Algorithm 4 has more categories. The predictions of all methods are presented in Table 2. The Γ-minimax estimator outputs a more similar prediction to the SCL estimator. This similarity appears different from our observation in the simulation, but can be explained by the fact that having more observations ($n = 1088$ vs $n = 100$; $m = 2000$ vs $m = 200$) decreases the variance of the number of new observed categories and thus lowers discrepancies between predictions from these methods. Since the SCL estimator outperforms the OSW estimator in the above simulation where this dataset is the true population, we expect the SCL estimator to achieve reasonably good performance in this application. Moreover, given that the Γ-minimax estimators outperform the SCL estimator in the above simulation, we expect that 57 or 58 represents an improved prediction of the number of new categories as compared to the SCL prediction of 51 in the case where there is limited prior information available.

The computation time to compute an approximated Γ-minimax estimator was about five to seven hours on an AWS EC2 instance (Amazon, 2019) with at least 4 vCPUs and at least 8 GiB of memory, depending on the number of times the grid was enlarged. As shown in Fig. 5, far few iterations are needed for SGDmax to output a good approximation of a Γ_1-minimax estimator, which is itself quite close to Γ-minimax. Therefore, with suitably less conservative tuning parameters or more adaptive minimax problem solvers, the computation time might drastically decrease. Moreover, the computation time needed to evaluate the computed Γ-minimax estimator at any sample is almost zero.

5.2. *Estimation of the entropy*

We also apply our method to estimate the entropy of a multinomial distribution (Example 3). The data-generating mechanism is the same as that described in Example 2, and the estimand of interest is Shannon entropy (Shannon, 1948), that is, $\Psi(P_0) = -\sum_{k=1}^{K} p_k \log p_k$. In the simulation, we choose the same true population and the same sample size $n = 100$ as in Section 5.1. The true entropy $\Psi(P_0)$ is 4.57. As a reference, the entropy of the uniform distribution with the same number of categories—which corresponds to the maximum entropy of multinomial distributions with the same total number of categories—is 5.24.

Jiao et al. (2015) developed a minimax rate optimal estimator of the Shannon entropy, and we run this simulation to investigate the potential gain of computing a Gamma-minimax estimator in well-posed problems with satisfactory solutions. As in Section 5.1, we consider three sets of prior information:

1. Strongly informative: Prior mean of $\Psi(P)$ in $[4.3, 4.7]$, $\geq 95\%$ probability that $\Psi(P)$ lies in $[4, 5]$;
2. Weakly informative: Prior mean of $\Psi(P)$ in $[4, 5]$, $\geq 95\%$ probability that $\Psi(P)$ lies in $[3.5, 5.5]$;
3. Almost noninformative: Prior mean of $\Psi(P)$ in $[3.7, 5.3]$, $\geq 95\%$ probability that $\Psi(P)$ lies in $[3, 6]$.

The architecture of our neural network estimator is almost identical to that in Section 5.1 except that the existing estimator being used is the one proposed in Jiao et al. (2015) (referred to as the JVHW estimator), and we initialize the network to return the JVHW estimator. We use Algorithm 4 to construct $\mathcal{M}_\ell$ and Algorithm 6 to compute a Γ_ℓ-minimax estimator. The tuning parameters in the algorithms are identical to those used in Section 5.1 except that, in Algorithm 6, (i) the learning rate is $\eta = 0.001$, and (ii) the number of iterations is 6,000 for Γ_1. We change these tuning parameters because the JVHW estimator is already minimax in terms of its convergence rate (Jiao et al., 2015), and we may need to update the estimator in a more cautious manner in Algorithm 6 to obtain any possible improvement. The trajectories of the estimated Bayes risks (Fig. 6) all appear to approximately reach a plateau, suggesting that the obtained estimator approximately Γ_1-minimax and that our choice of a smaller learning rate and a larger number of iterations is valid. Because of the additional complexity of the JVHW estimator, we ran our simulations on an AWS EC2 instance (Amazon, 2019) with 4 vCPUs and 32 GiB of memory. The computation time was ten to seventeen hours, depending on the number of times the grid was enlarged. The longer computation time than that described in Section 5.1 is primarily due to more iterations in SGDmax and the additional complexity of the JVHW estimator.

We compare the risk of the JVHW estimator and our trained Γ-minimax estimator under our set data-generating mechanism computed with 20000 Monte Carlo runs. We also compare their Bayes risk under the computed prior from Algorithm 6 using the last and finest grid in the computation with 20000 Monte Carlo runs. The results are summarized in Table 3. In this simulation experiment, our Γ-minimax estimator reduces the risk by a fair percentage compared with the JVHW estimator and achieves lower worst-case Bayes risk. According to these simulation results, we conclude that incorporating informative prior knowledge into the estimator may result in some improvement in estimating entropy. Thus, for well-posed statistical problems with satisfactory solutions, we expect mild or no substantial improvement and little deterioration from using a Gamma-minimax estimator.

Fig. 6 presents the unbiased estimator of Bayes risks over iterations when

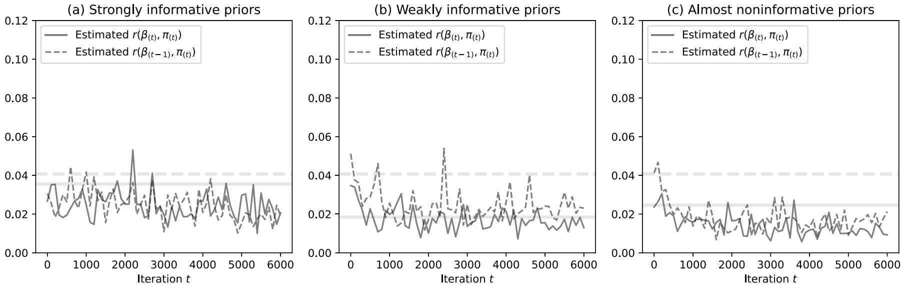

Fig 6. *Estimated Bayes risks of the estimator over iterations when computing a Γ_1-minimax estimator. The lines are unbiased estimates of the current Bayes risks (y-axis) with 30 Monte Carlo runs over iterations (x-axis). The horizontal lines are the risks of the JVHW (dashed) and the computed Γ-minimax (solid) estimators, respectively, under P_0 in the simulation. For clearness of visualization, the estimated Bayes risks are plotted every 100 iterations.*

Table 3
Risks and Bayes risks of estimators. $R(d, P_0)$: risk of the estimator under the true data-generating mechanism P_0. $r(d, \hat{\pi}^*)$: Bayes risk under prior $\hat{\pi}^*$, the computed prior from Algorithm 6 in the last and finest grid in the computation.

Strength of prior	Estimator	$R(d, P_0)$	$r(d, \hat{\pi}^*)$
strong	JVHW	0.041	0.035
	Γ-minimax	0.036	0.021
weak	JVHW	0.041	0.028
	Γ-minimax	0.018	0.024
almost none	JVHW	0.041	0.031
	Γ-minimax	0.025	0.016

computing a Γ_1-minimax estimator. With strongly informative prior information present, the Bayes risks appear to fluctuate without an increasing or decreasing trend at the beginning and decrease slowly after several thousand iterations. With weakly informative or almost no prior information, the Bayes risks also decrease slowly. A reason may be that the JVHW estimator is already minimax rate optimal (Jiao et al., 2015). The computed Γ-minimax estimators also appear to be somewhat similar to the JVHW estimator: in the output layer of the three settings with different prior information, the coefficients for the JVHW estimator are 0.97, 0.90 and 0.89, respectively; the coefficients for the previous hidden layer are 0.17, 0.17 and 0.20, respectively; the intercepts are 0.06, 0.30 and 0.30, respectively.

We further use the above methods to estimate entropy based on the dataset used as the true population in the simulation. The tuning parameters of the Γ-minimax estimators are exactly the same as those in the above simulation except that the starting point in Algorithm 4 has more categories. The estimates are presented in Table 4. All methods produce almost identical estimates. Because the sample size is more than ten times the sample size in the simulation and the JVHW estimator is minimax rate optimal (Jiao et al., 2015), we expect

TABLE 4

Estimated entropy based on the sample with size 1088 studied in Miller and Wiegert (1989). The strength of prior information in Γ-minimax estimators is shown in brackets.

Estimator	Estimated entropy
JVHW	4.709
Γ-minimax (strong)	4.709
Γ-minimax (weak)	4.708
Γ-minimax (almost none)	4.703

the JVHW estimator to have little room for improvement, which explains why the three Γ-minimax estimators perform similarly to the JVHW estimator. In other words, Gamma-minimax estimators appear to maintain, if not improve, the performance of the original JVHW estimator.

6. Discussion

We propose adversarial meta-learning algorithms to compute a Gamma-minimax estimator with theoretical guarantees under fairly general settings. These algorithms still leave room for improvement. As we discussed in Section 3.1, the stopping criterion we employ does not necessarily indicate that the maximal Bayes risk is close to the true minimax Bayes risk. In future work, it would be interesting to derive a better criterion that necessarily does indicate this near optimality. Our algorithms also require the user to choose increasingly fine approximating grids to the model space. Although we propose a heuristic algorithm for this procedure that performed well in our experiments, at this point, we have not provided optimality guarantees for this scheme. It may also be possible to improve our proposed algorithms to solve intermediate minimax problems in Section 3.1 by utilizing recent and ongoing advances from the machine learning literature that can be used to improve the training of generative adversarial networks.

We do not explicitly consider uncertainty quantification such as confidence intervals or credible intervals under a Gamma-minimax framework. Uncertainty quantification is important in practice since it provides more information than a point estimator and can be used for decision-making. In theory, our method may be directly applied if such a problem can be formulated into a Gamma-minimax problem. However, such a formulation remains unclear. The most challenging part is to identify a suitable risk function that correctly balances the level of uncertainty and the size of the output interval/region. Though the risk function used in Schafer and Stark (2009) appears to provide one possible starting point, it is not clear how to extend this approach to nonparametric settings.

It is possible to allow the space of estimators $\mathcal{D}$ to increase as the grid $\mathcal{M}_\ell$ increase. For example, we may specify an increasing sequence of estimator spaces $\{\mathcal{D}_\ell\}_{\ell=1}^{\infty}$ whose limit is dense in a general space $\mathcal{D}_0$; then, in Line 3 of Algorithm 1, we compute a Γ_ℓ-minimax estimator in $\mathcal{D}_\ell$, namely replace $\mathcal{D}$ with $\mathcal{D}_\ell$. This sequence of estimators might converge to a Γ-minimax estimator in $\mathcal{D}_0$. One possible choice of $\mathcal{D}_\ell$ ($\ell > 1$) in this approach is a space of statistical knowledge networks with the given estimator being the computed $\Gamma_{\ell-1}$-minimax estimator in $\mathcal{D}_{\ell-1}$. It is of future interest to investigate the properties of such an approach.

In conclusion, we propose adversarial meta-learning algorithms to compute a Gamma-minimax estimator under general models that can incorporate prior information in the form of generalized moment conditions. They can be useful when a parametric model is undesirable, semi-parametric efficiency theory does not apply, or we wish to utilize prior information to improve estimation.

Appendix A: Two counterexamples of Condition 3

We provide two counterexamples of Condition 3 to illustrate that this condition fails in extremely ill cases.

In the first counterexample, $P \mapsto R(d, P)$ is discontinuous: we set $R(d, P^*)$ to be zero for a fixed $P^* \in \mathcal{M}$ and $R(d, P)$ to be one for all other $P \in \mathcal{M}$. If we choose the grid $\mathcal{M}_\ell$ to be dense in $\mathcal{M}$ but to never contain P^*, then Condition 3 does not hold since $r_{\sup}(d, \tilde{\Gamma}_\ell) = 1$ for sufficiently large ℓ such that $P^* \in \Omega_\ell$ but $r_{\sup}(d, \tilde{\Gamma}_{i|\ell}) = 0$ for all i and ℓ. This issue can be resolved by choosing a continuous risk function.

In the second counterexample, $\mathcal{M}_\ell$ does not contain distributions that are consistent with prior information. Suppose that $\Gamma = \{\pi \in \Pi : \int \Phi(P)\pi(\mathrm{d}P) = 0\}$ where $\Phi(P) := \mathbb{E}_P[X^2]$. In other words, it is known that the true data-generating mechanism P_0 must be a distribution that is a point mass at zero, and thus Γ also only contains a point mass at P_0. If $\Phi(P) \neq 0$ for every $P \in \cup_{i=1}^\infty \mathcal{M}_i$, then, even if $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in $\mathcal{M}$, $\tilde{\Gamma}_{i|\ell} = \emptyset$ and thus Condition 3 does not hold. This issue can be resolved by rewriting the problem such that these hard constraints on $\mathcal{M}$ are incorporated into the specification of $\mathcal{M}$ rather than Γ.

Appendix B: Additional gradient-based algorithms

If we can evaluate $R(\beta, P)$ exactly for all $\beta \in \mathcal{H}$ and $P \in \mathcal{M}_\ell$, then the GDmax algorithm (Algorithm 5) may be used. Note that Line 3 can be formulated into a linear program, which can always be solved in polynomial time with an interior point method (e.g., Jiang et al., 2020) and often be solved in polynomial time with a simplex method (Spielman and Teng, 2004).

Algorithm 5 Gradient descent with max-oracle (GDmax) to compute a Γ_ℓ-minimax estimator

```
1: Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$ and max-oracle accuracy $\zeta > 0$.
2: for $t = 1, 2, \ldots$ do
3:     Maximization: find $\pi_{(t)} \in \Gamma_\ell$ such that $r(\beta_{(t-1)}, \pi_{(t)}) \geq \max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \zeta$
4:     Gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \eta \nabla_\beta r(\beta, \pi_{(t)})|_{\beta=\beta_{(t-1)}}$
```

We have the following result on the validity of GDmax.

Theorem 5 (Validity of GDmax (Algorithm 5))**.** *Under conditions in Theorem 2, in Algorithm 5, with $\eta = \epsilon^2/(L_1 L_2^2)$ and $\zeta = \epsilon^2/(24 L_1)$, $\beta_{(t)}$ is an*

ϵ-stationary point of $\beta \mapsto r_{\text{sup}}(\beta, \Gamma_\ell)$ for $t = O(L_1 L_2 \Delta/\epsilon^4)$, and is thus close to a local minimum of $\beta \mapsto r_{\text{sup}}(\beta, \Gamma_\ell)$.

Therefore, we propose a variant (Algorithm 6) by replacing this line with Lines 3–4 so that ordinary linear program solvers can be directly applied. The following theorem justifies this variant.

Algorithm 6 Convenient variant of SGDmax (Algorithm 2) to compute a Γ_ℓ-minimax estimator

```
1: Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$ and batch sizes $J$, $J'$.
2: for $t = 1, 2, \ldots$ do
3:     Generate iid copies $\xi_1, \ldots, \xi_{J'}$ of $\xi$.
4:     Stochastic maximization: $\pi_{(t)} \leftarrow \text{argmax}_{\pi \in \Gamma_\ell} \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi, \xi_j)$.
5:     Generate iid copies of $\xi_{J'+1}, \ldots, \xi_{J'+J}$ of $\xi$.
6:     Stochastic gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \frac{\eta}{J} \sum_{j=J'+1}^{J'+J} \nabla_\beta \hat{r}(\beta, \pi_{(t)}, \xi_j)|_{\beta=\beta_{(t-1)}}$.
```

The validity of this variant of SGDmax is given in Theorem 6 below.

Theorem 6 (Validity of convenient variant of SGDmax (Algorithm 6))**.** *Suppose that $\{\xi \mapsto \hat{r}(\beta, \pi, \xi) : \beta \in \mathbb{R}^D, \pi \in \Gamma_\ell\}$ is a Ξ-Glivenko-Centelli class (van der Vaart and Wellner, 2000). Then, for any $\zeta > 0$, there exists a sufficiently large J' such that*

$$\mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq \max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \zeta$$

for all t, where the expectation is taken over $\pi_{(t)}$ and $\beta_{(t-1)}$ is fixed. Therefore, with the chosen parameters in Theorem 2, we may choose a sufficiently large J' so that $\beta_{(t)}$ is an ϵ-stationary point of $\beta \mapsto r_{\text{sup}}(\beta, \Gamma_\ell)$ in expectation for $t = O(L_1(L_2^2 + \sigma^2)\Delta/\epsilon^4)$ and is thus close to a local minimum of $\beta \mapsto r_{\text{sup}}(\beta, \Gamma_\ell)$ with high probability.

We prove Theorem 6 by showing that $\max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})]$ converges to 0 as $J' \to \infty$. The proof is essentially an application of empirical process theory to the study of an M-estimator.

Appendix C: Additional simulation: mean estimation

In this appendix, we illustrate our proposed methods via simulation in a special case of Example 1, namely for estimating the mean of a distribution. We assume that $\mathcal{M}$ consists of all probability distributions defined on the Borel σ-algebra on $[0, 1]$ and we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_0 \in \mathcal{M}$. Here we take $n = 10$. The estimand is $\Psi(P_0) = \int x \, P_0(\mathrm{d}x)$. We use the mean squared error risk introduced in Example 1. Suppose that we represent the prior information by $\Gamma = \{\pi \in \Pi : \int \Psi(P) \, \pi(\mathrm{d}P) = 0.3\}$, which corresponds to the set of prior distributions in Π that satisfy an equality constraint on the prior mean of $\Psi(P)$.

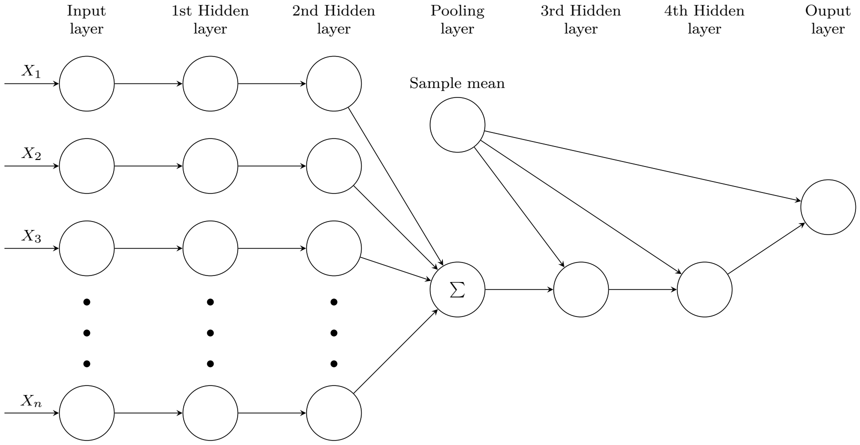

Fig 7. *Architecture of the permutation invariant neural network estimator of the mean in* $\mathcal{D}_{\text{skn}}$*.* X_i*: observation* i *in the sample;* $\sum$*: the node that sums up all ancestor nodes. In the first two hidden layers, all input nodes are transformed by the same function. The arrows from the input nodes to the sample mean estimator are omitted from this graph. Each node in the hidden layers represents a vector.*

We apply our method to three spaces of estimators separately. The first space, $\mathcal{D}_{\text{linear}}$, is the set of affine transformations of the sample mean, that is, $\mathcal{D}_{\text{linear}} = \{d : d(\mathbf{X}) = \beta_0 + \beta_1 \sum_{i=1}^{n} X_i/n, \beta_0, \beta_1 \in \mathbb{R}\}$. As shown in Proposition 1 in Appendix E.5, there is an estimator d^* in $\mathcal{D}_{\text{linear}}$ that is Γ-minimax in the space of all estimators that are square-integrable with respect to all $P \in \mathcal{M}$, so we consider this simple space to better compare our computed estimator with that theoretical Γ-minimax estimator. When computing a Γ-minimax estimator in $\mathcal{D}_{\text{linear}}$, we initialize the estimator to be the sample mean, that is, we let $\beta_0 = 0$ and $\beta_1 = 1$.

The second space, $\mathcal{D}_{\text{skn}}$ (statistical knowledge network), is a set of neural networks designed based on statistical knowledge that includes the sample mean as an input. We consider this space to illustrate our proposal in Section 4.2. More precisely, we use the architecture in Fig. 7, which is similar to the deep set architecture (Zaheer et al., 2017; Maron et al., 2019) and is a permutation invariant neural network. We use such an architecture to account for the fact that the sample is iid. In this architecture, the sample mean node is used as an augmenting node to an ordinary deep set network and is combined with the output of that ordinary network in the fourth hidden layer to obtain the final output. Note that $\mathcal{D}_{\text{skn}} \supset \mathcal{D}_{\text{linear}}$. When computing a Γ-minimax estimator for this class, we also initialize the network to be exactly the sample mean, which is a reasonable choice given that the sample mean is known to be a sensible estimator. In this simulation experiment, we choose the dimensionality of nodes in each hidden layer in Fig. 7 as follows: each node in the first, second, third and fourth hidden layer represents a vector in $\mathbb{R}^{10}$, $\mathbb{R}^5$, $\mathbb{R}^{10}$ and $\mathbb{R}$, respectively. We do not use larger architectures because usually the sample mean is already a good estimator, and we expect to obtain a useful estimator as a small perturbation of this estimator. We also use the ReLU as the activation function. We did not use ELMs in this and the following simulations because we found that neural networks perform well.

The third space, $\mathcal{D}_{\text{nn}}$, is a set of neural networks that do not utilize knowledge of the sample mean. We consider this space to illustrate our method without utilizing existing estimators. These estimators are also deep set networks with similar architecture as $\mathcal{D}_{\text{skn}}$ in Fig. 7. The main difference is that the explicit sample mean node and the fourth hidden layer are removed. When computing a Γ-minimax estimator in $\mathcal{D}_{\text{nn}}$, we also randomly initialize the network, unlike $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$, in order not to input statistical knowledge. Because the ReLU activation function is used, $\mathcal{D}_{\text{nn}} \supset \mathcal{D}_{\text{linear}}$, and we do not expect that optimizing over $\mathcal{D}_{\text{nn}}$ should not lead to a Γ-minimax estimator with worse performance than those in $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$.

To construct the grid $\mathcal{M}_\ell$ for this problem, we use a simpler method than Algorithm 4. As indicated by Lemma 6 in Appendix E.5, for estimators in $\mathcal{D}_{\text{linear}}$, Bernoulli distributions tend to have high risks since all probability weights lie on the boundary of $[0,1]$; in addition, a prior π^* for which d^* is Bayes is a Beta prior over Bernoulli distributions. Therefore, we randomly generate 2000 Bernoulli distributions as grid points in $\mathcal{M}_1$. We also include two degenerate distributions in this grid, namely the distribution that places all of its mass at 0 and that which places all of its mass at 1. When constructing $\mathcal{M}_\ell$ from $\mathcal{M}_{\ell-1}$, we still add in more complicated distributions to make the grid dense in the limit: we first randomly generate 500 discrete distributions with support being those in $\mathcal{M}_{\ell-1}$; then we randomly generate 10 new support points in $[0,1]$ and 1000 distributions with support points being the union of the new support points and the existing support points in $\mathcal{M}_{\ell-1}$.

When computing the Γ-minimax estimator, for each grid $\mathcal{M}_\ell$, we compute the Γ_ℓ-minimax estimator for all three estimator spaces with Algorithm 6. We set the learning rate $\eta = 0.005$, the batch size $J = 50$ and the number of iterations to be 200 for Γ_ℓ ($\ell > 1$). The number of iterations for Γ_1 is larger because, in our experiments, we saw that a Γ_1-minimax estimator is already close to a Γ-minimax estimator, and using a large number of iterations in this step can improve the initial estimator substantially. For $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$, the number of iterations for Γ_1 is 2000; the corresponding number for $\mathcal{D}_{\text{nn}}$ is 6000 to account for the lack of human knowledge input. We also use Algorithm 3 with 10000 iterations to compute a Γ_ℓ-minimax estimator for $\mathcal{D}_{\text{linear}}$ for illustration. In this setup, as described in Section 3.3, we take the average of the computed Γ-minimax stochastic estimator as the final output estimator in $\mathcal{D}_{\text{linear}}$. We do not apply Algorithm 3 to $\mathcal{D}_{\text{skn}}$ or $\mathcal{D}_{\text{nn}}$ because it is computationally intractable for these estimator spaces.

We set the stopping criterion in Algorithm 1 as follows. When Algorithm 6 is used to compute Γ_ℓ-minimax estimators, we estimate both $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_{\ell-1})$ with 2000 Monte Carlo runs as described in Section 3.1; when Algorithm 3 is used, $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_{\ell-1})$ are computed exactly because $R(d,P)$ has a closed-form expression for all $d \in \mathcal{D}_{\text{linear}}$ and $P \in \mathcal{M}_\ell$.

Table 5

Coefficients and Bayes risks of estimators of the mean. Unrestricted space: the space of all estimators that are square-integrable with respect to all $P \in \mathcal{M}$.

Estimator space	Method to obtain d^*	β_0	β_1	$r(d, \pi^*)$
Unrestricted space	Theoretical derivation	0.072	0.760	0.012
$\mathcal{D}_{\text{linear}}$	Algorithms 1 &6	0.072	0.763	0.012
$\mathcal{D}_{\text{skn}}$	Algorithms 1 &6	0.071	0.767	0.012
$\mathcal{D}_{\text{nn}}$	Algorithms 1 &6	—	—	0.012
$\mathcal{D}_{\text{linear}}$	Algorithms 1 &3	0.072	0.760	0.012

We set the tolerance ϵ to be equal to 0.0001 so that we stop Algorithm 1 if $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_\ell) - r_{\text{sup}}(d^*_{\ell-1}, \Gamma_{\ell-1}) \leq \epsilon$.

After computation, we report the Bayes risk of the computed and theoretical Γ-minimax estimators under π^*, the prior such that $r(d^*, \pi^*) = \inf_{d \in \mathcal{D}} r_{\text{sup}}(d, \Gamma)$. For the estimators in $\mathcal{D}_{\text{linear}}$, we further report their coefficients. We also report two coefficients of the computed estimator in $\mathcal{D}_{\text{skn}}$ as follows. Since $\mathcal{D}_{\text{linear}} \subseteq \mathcal{D}_{\text{skn}}$ and we initialize the estimator to be the sample mean for $\mathcal{D}_{\text{skn}}$, we would expect that the bias β_0 and the weight of the sample mean β_1 in the output layer for the computed Γ-minimax estimator in $\mathcal{D}_{\text{skn}}$ may correspond to those in $\mathcal{D}_{\text{linear}}$. Therefore, we also report these two coefficients β_0 and β_1 for $\mathcal{D}_{\text{skn}}$. This may not be the case for $\mathcal{D}_{\text{nn}}$ because the sample mean is not explicit in its parameterization and all coefficients are randomly initialized, so we do not report any coefficients for $\mathcal{D}_{\text{nn}}$.

Table 5 presents the computation results. By Theorem 7 in Appendix E.5, these computed estimators are all approximately Γ-minimax since their Bayes risks for π^* are all close to that of a theoretical Γ-minimax estimator. The coefficients β_0 and β_1 of the computed estimators in $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$ are also close to a theoretically derived estimator. For the computed estimator in $\mathcal{D}_{\text{skn}}$, the weight of the other ancestor node in the output layer (i.e., the node in the 4th hidden layer in Fig. 7) is 0.000. Therefore, our computed Γ-minimax estimator in $\mathcal{D}_{\text{skn}}$ is also close to a theoretically derived Γ-minimax estimator.

In our experiments, Algorithm 1 converged after computing a Γ_1-minimax estimator except when using Algorithm 6 for $\mathcal{D}_{\text{linear}}$. Even in this exceptional case, the computed Γ_1-minimax estimator is still approximately Γ-minimax. We think the algorithm does not stop then in these cases because of Monte Carlo errors when computing $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\text{sup}}(d^*_{\ell-1}, \Gamma_{\ell-1})$.

Fig. 3 presents the Bayes risks (or its unbiased estimates) over iterations when computing a Γ_1-minimax estimator. In all cases using Algorithm 6, the Bayes risks appear to decrease and converge. When using Algorithm 3, the upper and lower bounds both converge to the same limit. The limiting values of the Bayes risks in all cases are close to $r(d^*, \pi^*)$ because Γ_1 can approximate π^* well.

Appendix D: Sensitivity analysis for tuning parameter selection

For the simulation in Section 5.1 with strongly informative prior information, we conduct three simulations to investigate the sensitivity of our proposed method

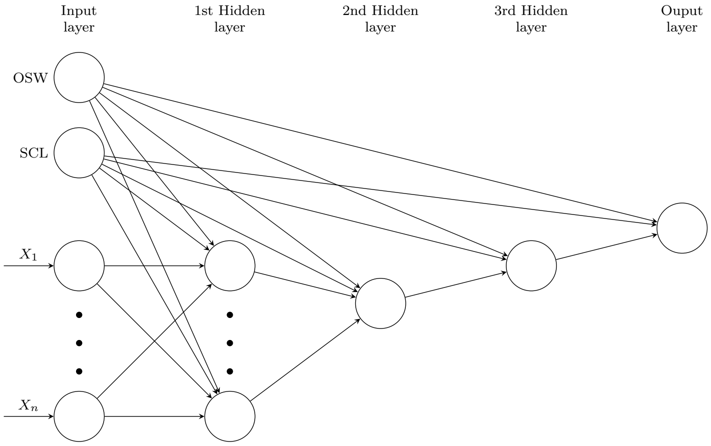

Fig 8. *Architecture of the deeper and wider neural network estimator of the expected number of new categories.*

Table 6

Table similar to Table 1 for sensitivity analysis with strongly informative prior information.

Varied tuning parameter	$R(d, P_0)$	$r(d, \hat{\pi}^*)$
Initial distribution in MCMC	19	44
Grid size	15	34
Statistical knowledge network structure	17	38

to the selection of tuning parameters. In each simulation below, we vary one set of tuning parameters and rerun the algorithm to obtain an estimator. In the first simulation, we vary the starting point of Algorithm 4 to construct the first grid $\mathcal{M}_1$. The new starting point is a distribution with 173 categories and $\Phi(P_{(0)}) = 61$, and so this starting point is qualitatively different from the one chosen in the original simulation. In the second simulation, we vary the grid sizes: There are 500 grid points in $\mathcal{M}_1$ and we add 500 grid points each time we enlarge the grid. In the third simulation, we chose a wider and deeper statistical knowledge network (see Fig. 8): Compared to the original simulation, we add one more hidden layer and increased the number of hidden nodes in the first two hidden layers to 100. As shown in Table 6, the results in these sensitivity simulations appear similar to that in Section 5.1 within the variation due to randomness in MCMC (Algorithm 4) and SGDmax (Algorithm 6).

Appendix E: Proofs

E.1. Proof of Theorem 1 and Corollary 1

Lemma 1. *If $\{\Omega_\ell\}_{\ell=1}^\infty$ is an increasing sequence of subsets of $\mathcal{M}$ such that $\bigcup_{\ell=1}^\infty \Omega_\ell = \mathcal{M}$, then, for any $d \in \mathcal{D}$, $r_{\sup}(d, \tilde{\Gamma}_\ell) \nearrow r_{\sup}(d, \Gamma)$ $(\ell \to \infty)$.*

Proof of Lemma 1. Since $\tilde{\Gamma}_\ell \subseteq \tilde{\Gamma}_{\ell+1} \subseteq \Gamma$, it holds that

$$r_{\sup}(d, \tilde{\Gamma}_\ell) \leq r_{\sup}(d, \tilde{\Gamma}_{\ell+1}) \leq r_{\sup}(d, \Gamma),$$

and so we only need to lower bound $r_{\sup}(d, \tilde{\Gamma}_\ell)$. Fix $\epsilon > 0$. By Corollary 5 of Pinelis (2016), $r_{\sup}(d, \Gamma)$ can be approximated by $r(d, \nu)$ arbitrarily well for priors $\nu \in \Gamma$ with a finite support; that is, there exists $\nu \in \Gamma$ with finite support such that $r(d, \nu) \geq r_{\sup}(d, \Gamma) - \epsilon$. For sufficiently large ℓ, Ω_ℓ contains all support points of ν and hence $r_{\sup}(d, \tilde{\Gamma}_\ell) \geq r(d, \nu) \geq r_{\sup}(d, \Gamma) - \epsilon$. The desired result follows. □

Lemma 2. *Under Condition 2, for any $\Gamma' \subseteq \Gamma$ and $\epsilon > 0$, there exists $\delta > 0$ such that $r_{\sup}(d^*, \Gamma') - r_{\sup}(d, \Gamma') \leq \epsilon$ for all $d \in \mathcal{D}$ such that $\varrho(d, d^*) \leq \delta$.*

Proof of Lemma 2. By Corollary 5 of Pinelis (2016), there exists $\nu \in \Gamma'$ with a finite support such that $r_{\sup}(d^*, \Gamma') \leq r(d^*, \nu) + \epsilon/2$. By Condition 2 and the fact that ν has a finite support, there exists $\delta > 0$ such that, for any $d \in \mathcal{D}$ such that $\varrho(d, d^*) \leq \delta$, $|r(d, \nu) - r(d^*, \nu)| \leq \epsilon/2$. Since $\nu \in \Gamma'$, we have that $r_{\sup}(d, \Gamma') \geq r(d, \nu)$ and thus $r_{\sup}(d^*, \Gamma') - r_{\sup}(d, \Gamma') \leq r(d^*, \nu) + \epsilon/2 - r(d, \nu) \leq \epsilon$ for any $d \in \mathcal{D}$ such that $\varrho(d, d^*) \leq \delta$. □

Lemma 3. *Under Condition 3, it holds that $\lim_{i \to \infty} r_{\sup}(d, \tilde{\Gamma}_{i|\ell}) = r_{\sup}(d, \tilde{\Gamma}_\ell)$.*

Proof of Lemma 3. Let $d \in \mathcal{D}$, ℓ and $\epsilon > 0$ be fixed. By Corollary 5 of Pinelis (2016), $r_{\sup}(d, \tilde{\Gamma}_\ell) \leq r(d, \pi) + \epsilon/2$ for some $\pi \in \tilde{\Gamma}_\ell$ with a finite support. Under Condition 2, there exists a sequence $\pi_i \in \tilde{\Gamma}_{i|\ell}$ such that, for all sufficiently large i, $r(d, \pi_i) \geq r(d, \pi) - \epsilon/2$. For such i, $r_{\sup}(d, \tilde{\Gamma}_\ell) \leq r(d, \pi_i) + \epsilon$. Since $r_{\sup}(d, \tilde{\Gamma}_\ell) \geq r_{\sup}(d, \tilde{\Gamma}_{i|\ell}) \geq r(d, \pi_i)$, we have that $r(d, \pi_i) \leq r_{\sup}(d, \tilde{\Gamma}_{i|\ell}) \leq r_{\sup}(d, \tilde{\Gamma}_\ell) \leq r(d, \pi_i) + \epsilon$ for all sufficiently large i, and thus we have proved Lemma 3. □

Proof of Theorem 1. Let $\epsilon > 0$. There exists $d' \in \mathcal{D}$ such that

$$r_{\sup}(d', \Gamma) \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) + \epsilon.$$

Moreover, there exists $\pi_\ell \in \Gamma_\ell$ such that

$$r_{\sup}(d', \Gamma_\ell) \leq r(d', \pi_\ell) + \epsilon.$$

Using the fact that d_ℓ^* is Γ_ℓ-minimax and the definition of $r_{\sup}$, it holds that

$$\begin{aligned} r_{\sup}(d_\ell^*, \Gamma_\ell) &\leq r_{\sup}(d', \Gamma_\ell) \leq r(d', \pi_\ell) + \epsilon \\ &\leq r_{\sup}(d', \Gamma) + \epsilon \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) + 2\epsilon. \end{aligned}$$

Since this inequality holds for any $\epsilon > 0$, we have that

$$r_{\sup}(d_\ell^*, \Gamma_\ell) \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma).$$

An almost identical argument shows that the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ is nondecreasing. Therefore, this sequence converges to some limit

$$\mathcal{R} \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) \leq r_{\sup}(d^*, \Gamma).$$

We next prove that $r_{\sup}(d^*, \Gamma) \leq \mathcal{R}$. Let $\epsilon > 0$. Without loss of generality, we may assume that $\mathcal{M}_\ell \subseteq \Omega_\ell$ for all $\ell = 1, 2, \ldots$ in Condition 3. (Otherwise, we may instead consider the sequence $\{\Omega_{\tilde{\ell}}\}_{\tilde{\ell}=1}^\infty$ where $\Omega_{\tilde{\ell}} = \bigcap_{\ell' : \Omega_{\ell'} \supseteq \mathcal{M}_\ell} \Omega_{\ell'}$. Note that Condition 3 also holds for $\{\Omega_{\tilde{\ell}}\}_{\tilde{\ell}=1}^\infty$.) By Lemma 1, there exists ℓ_0 such that $r_{\sup}(d^*, \tilde{\Gamma}_{\ell_0}) \geq r_{\sup}(d^*, \Gamma) - \epsilon/3$. By Condition 3, there exists i_1 such that $r_{\sup}(d^*, \tilde{\Gamma}_{i_1|\ell_0}) \geq r_{\sup}(d^*, \tilde{\Gamma}_{\ell_0}) - \epsilon/3$. Without loss of generality, suppose that $d^*_\ell \to d^*$ (otherwise, take a convergent subsequence to this limit point). This then implies that there exists $i_2 > i_1$ such that $\varrho(d^*_{i_2}, d^*)$ is sufficiently small, such that, by Lemma 2, $r_{\sup}(d^*_{i_2}, \tilde{\Gamma}_{i_1|\ell_0}) \geq r_{\sup}(d^*, \tilde{\Gamma}_{i_1|\ell_0}) - \epsilon/3$. Moreover, since $\tilde{\Gamma}_{i_1|\ell_0} \subseteq \tilde{\Gamma}_{i_1} \subseteq \tilde{\Gamma}_{i_2}$, it holds that $r_{\sup}(d^*_{i_2}, \tilde{\Gamma}_{i_2}) \geq r_{\sup}(d^*_{i_2}, \tilde{\Gamma}_{i_1|\ell_0})$. Therefore, $r_{\sup}(d^*_{i_2}, \tilde{\Gamma}_{i_2}) \geq r_{\sup}(d^*, \Gamma) - \epsilon$. Since the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ is nondecreasing, it holds that $r_{\sup}(d^*_\ell, \Gamma_\ell) \geq r_{\sup}(d^*, \Gamma) - \epsilon$ for all $\ell \geq i_2$. Since ϵ is arbitrary, we have that $\liminf_{\ell \to \infty} r_{\sup}(d^*_\ell, \Gamma_\ell) \geq r_{\sup}(d^*, \Gamma)$, and hence $\mathcal{R} \geq r_{\sup}(d^*, \Gamma)$.

Combining the results from the preceding two paragraphs,

$$\mathcal{R} = \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) = r_{\sup}(d^*, \Gamma).$$

Consequently, d^* is Γ-minimax. Moreover, as $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ increases to $\mathcal{R}$, this sequence also increases to $r_{\sup}(d^*, \Gamma)$. This concludes the proof. □

Proof of Corollary 1. We first establish the strict convexity of $d \mapsto r(d, \pi)$ for any $\pi \in \Gamma$. We then establish the strict convexity of $d \mapsto r_{\sup}(d, \Gamma)$. We then establish that there is a unique minimizer of $d \mapsto r_{\sup}(d, \Gamma)$ and show that the desired result follows from Theorem 1.

Let $d_1, d_2 \in \mathcal{D}$ and $c \in (0, 1)$ be arbitrary, then by the convexity of $\mathcal{D}$ and the strict convexity of $d \mapsto R(d, P)$ for each $P \in \mathcal{M}$,

$$\begin{aligned} r(cd_1 + (1-c)d_2, \pi) &= \int R(cd_1 + (1-c)d_2, P)\pi(\mathrm{d}P) \\ &< \int \{cR(d_1, P) + (1-c)R(d_2, P)\}\pi(\mathrm{d}P) \\ &= cr(d_1, \pi) + (1-c)r(d_2, \pi). \end{aligned}$$

Therefore, $d \mapsto r(d, \pi)$ is strictly convex for any $\pi \in \Gamma$.

Let $d_1, d_2 \in \mathcal{D}$ be distinct and $c \in (0, 1)$ be arbitrary. Since $r_{\sup}(d, \Gamma)$ is attainable for any $d \in \mathcal{D}$, there exists $\tilde{\pi} \in \Gamma$ such that

$$\begin{aligned} r_{\sup}(cd_1 + (1-c)d_2, \Gamma) &= r(cd_1 + (1-c)d_2, \tilde{\pi}) \\ &< cr(d_1, \tilde{\pi}) + (1-c)r(d_2, \tilde{\pi}) \\ &\leq cr_{\sup}(d_1, \Gamma) + (1-c)r_{\sup}(d_2, \Gamma). \end{aligned}$$

Thus, $d \mapsto r_{\mathrm{sup}}(d, \Gamma)$ is strictly convex.

As $d \mapsto r_{\mathrm{sup}}(d, \Gamma)$ is strictly convex and $\mathcal{D}$ is convex, this function achieves exactly one minimum on $\mathcal{D}$. By Theorem 1, any limit point d^* of $\{d_\ell^*\}_{\ell=1}^\infty$ is a minimizer of $d \mapsto r_{\mathrm{sup}}(d, \Gamma)$, and so the sequence has a limit point, which is also the unique Γ-minimax estimator. □

E.2. Proof of Theorems 2 & 5

We prove Theorems 2 and 5 by checking that Assumptions 3.1 and 3.6 in Lin, Jin and Jordan (2020) are satisfied and using Theorem E.3 and E.4 in Lin, Jin and Jordan (2020), respectively. Since Assumption 3.1 is satisfied by our construction of $\hat{R}$, we focus on Assumption 3.6 for the rest of this section.

Let $\mathcal{M}_\ell = \{P_1, P_2, \ldots, P_\Lambda\} \subseteq \mathcal{M}$. For any $\pi \in \Gamma_\ell$, let π_λ denote the probability weight of π on P_λ ($\lambda = 1, \ldots, \Lambda$). For the rest of this section, we also use π to denote the vector $(\pi_1, \ldots, \pi_\Lambda)$. We also use $\lesssim$ to denote less than equal to up to a universal positive constant that may depend on ℓ. Then, straightforward calculations imply that $\nabla_\beta r(\beta, \pi) = \sum_{\lambda=1}^{\Lambda} \pi_\lambda \nabla_\beta R(\beta, P_\lambda)$ and $\nabla_\pi r(\beta, \pi) = (R(\beta, P_1), \ldots, R(\beta, P_\Lambda))^\top$.

For each $\ell = 1, 2, \ldots$, for any $\beta^1, \beta^2 \in \mathcal{H}$ and $\pi^1, \pi^2 \in \Gamma_\ell$, by Conditions 4 and 5,

$$\begin{aligned}
&\left\| \nabla_\beta r(\beta, \pi)|_{\beta=\beta^1, \pi=\pi^1} - \nabla_\beta r(\beta, \pi)|_{\beta=\beta^2, \pi=\pi^2} \right\| \\
&= \left\| \sum_{\lambda=1}^{\Lambda} \left\{ \pi_\lambda^1 \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_1} - \pi_\lambda^2 \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\} \right\| \\
&\leq \sum_{\lambda=1}^{\Lambda} \pi_\lambda^1 \left\| \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_1} - \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\| \\
&\quad + \left\| \sum_{\lambda=1}^{\Lambda} (\pi_\lambda^1 - \pi_\lambda^2) \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\| \\
&\lesssim \|\beta^1 - \beta^2\| + \|\pi^1 - \pi^2\| \\
&\lesssim \|(\beta^1, \pi^1) - (\beta^2, \pi^2)\|,
\end{aligned}$$

and similarly for $\nabla_\pi r(\beta, \pi)$,

$$\begin{aligned}
&\left\| \nabla_\pi r(\beta, \pi)|_{\beta=\beta^1, \pi=\pi^1} - \nabla_\pi r(\beta, \pi)|_{\beta=\beta^2, \pi=\pi^2} \right\| \\
&= \Big\| \big(R(\beta^1, P_1) - R(\beta^2, P_1), R(\beta^1, P_2) \\
&\quad - R(\beta^2, P_2), \ldots, R(\beta^1, P_\Lambda) - R(\beta^2, P_\Lambda) \big)^\top \Big\| \\
&\lesssim \|\beta^1 - \beta^2\| \leq \|(\beta^1, \pi^1) - (\beta^2, \pi^2)\|.
\end{aligned}$$

This implies that for each ℓ, the gradient of $r(\beta, \pi)$ ($\beta \in \mathcal{H}$, $\pi \in \Gamma_\ell$) is Lipschitz continuous.

For each $\ell = 1, 2, \ldots$, for any $\beta^1, \beta^2 \in \mathcal{H}$ and $\pi \in \Gamma_\ell$, Condition 4 implies that

$$\begin{aligned}\left|r(\beta^1, \pi) - r(\beta^2, \pi)\right| &= \left|\sum_{\lambda=1}^{\Lambda} \pi_\lambda \left[R(\beta^1, P_\lambda) - R(\beta^2, P_\lambda)\right]\right| \\ &\leq \sum_{\lambda=1}^{\Lambda} \pi_\lambda \left|R(\beta^1, P_\lambda) - R(\beta^2, P_\lambda)\right| \lesssim \|\beta^1 - \beta^2\|.\end{aligned}$$

Therefore, $\beta \mapsto r(\beta, \pi)$ is Lipschitz continuous with a universal Lipschitz constant independent of $\pi \in \Gamma_\ell$.

Finally, it is straightforward to check that (i) $\pi \mapsto r(\beta, \pi)$ is concave for any $\beta \in \mathcal{H}$, and (ii) Γ_ℓ is parameterized by a convex subset of a simplex in a Euclidean space, which is a convex and bounded set. These results show that Assumption 3.6 in Lin, Jin and Jordan (2020) is satisfied for Algorithm 5 and 2.

E.3. Proof of Theorem 6

Proof of Theorem 6. Let $\pi_{(t),0}$ denote a maximizer of $\pi \mapsto r(\beta_{(t-1)}, \pi)$. It holds that

$$\begin{aligned} 0 &\leq r(\beta_{(t-1)}, \pi_{(t),0}) - r(\beta_{(t-1)}, \pi_{(t)}) \\ &\leq \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi_j) - \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi_j) \\ &\quad + r(\beta_{(t-1)}, \pi_{(t),0}) - r(\beta_{(t-1)}, \pi_{(t)}) \\ &= \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi_j) - \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi_j)\right] \\ &\quad - \mathbb{E}\left[\hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi) - \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi)\right] \Bigg\} \\ &\leq \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \Bigg| \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)\right] \\ &\quad - \mathbb{E}\left[\hat{r}(\beta, \pi_1, \xi) - \hat{r}(\beta, \pi_2, \xi)\right] \Bigg\} \Bigg|. \end{aligned}$$

Note that the right hand side does not depend on t. Therefore,

$$\begin{aligned} 0 &\leq \sup_t \left\{ r(\beta_{(t-1)}, \pi_{(t),0}) - \mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})] \right\} \\ &\leq \mathbb{E}^* \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \Bigg| \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)\right] \end{aligned}$$

$$- \mathbb{E}\left[\hat{r}(\beta,\pi_1,\xi) - \hat{r}(\beta,\pi_2,\xi)\right] \Bigg\}\Bigg|,$$

where $\mathbb{E}^*$ stands for outer expectation. We may apply Corollary 9.27 in Kosorok (2008) to $\mathcal{F} := \{\xi \mapsto \hat{r}(\beta,\pi,\xi) : \beta \in \mathbb{R}^D, \pi \in \Gamma_\ell\}$ and show that $\mathcal{F} - \mathcal{F} := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\} \supseteq \{\xi \mapsto \hat{r}(\beta,\pi_1,\xi) - \hat{r}(\beta,\pi_2,\xi) : \beta \in \mathbb{R}^D, \pi_1,\pi_2 \in \Gamma_\ell\}$ is a Ξ-Glivenko-Cantelli class. Therefore,

$$\begin{aligned}&\left\{ \sup_{\beta\in\mathbb{R}^D,\pi_1,\pi_2\in\Gamma_\ell} \left| \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta,\pi_1,\xi_j) - \hat{r}(\beta,\pi_2,\xi_j)\right] \right.\right. \\ &\qquad \left.\left. - \mathbb{E}\left[\hat{r}(\beta,\pi_1,\xi) - \hat{r}(\beta,\pi_2,\xi)\right] \Bigg\} \right| \right\}^* \\ &\leq \left\{ \sup_{f\in\mathcal{F}-\mathcal{F}} \left| \frac{1}{J'} \sum_{j=1}^{J'} \{f(\xi_j) - \mathbb{E}[f(\xi)]\} \right| \right\}^* \overset{a.s.}{\to} 0,\end{aligned}$$

as $J' \to \infty$. Here, X^* stands for the minimal measurable majorant with respect to Ξ of a (possibly non-measurable) mapping X (van der Vaart and Wellner, 2000).

By Problem 1 of Section 2.4 in van der Vaart and Wellner (2000), there exists a random variable F such that $F \geq \sup_{f\in\mathcal{F}-\mathcal{F}} |f(\xi) - \mathbb{E}[f(\xi')]|$ Ξ-almost surely and $\mathbb{E}[F] < \infty$. Then,

$$\sup_{f\in\mathcal{F}-\mathcal{F}} \left| \frac{1}{J'} \sum_{j=1}^{J'} \{f(\xi_j) - \mathbb{E}[f(\xi_j)]\} \right| \leq F$$

Ξ-almost surely. By dominated convergence theorem,

$$\begin{aligned}&\mathbb{E}^* \sup_{\beta\in\mathbb{R}^D,\pi_1,\pi_2\in\Gamma_\ell} \left| \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta,\pi_1,\xi_j) - \hat{r}(\beta,\pi_2,\xi_j)\right] \right. \\ &\qquad \left. - \mathbb{E}\left[\hat{r}(\beta,\pi_1,\xi_j) - \hat{r}(\beta,\pi_2,\xi_j)\right] \Bigg\} \right| \to 0\end{aligned}$$

as $J' \to \infty$, and so does $\sup_t \left\{ r(\beta_{(t-1)}, \pi_{(t),0}) - \mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})] \right\}$. Thus, for any $\zeta > 0$, there exists a sufficiently large J' such that $\mathbb{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq r(\beta_{(t-1)}, \pi_{(t),0}) - \zeta$ for all t. □

E.4. Proof of Theorem 3

Our proof of Theorem 3 builds on that of Robinson (1951). Major modifications are needed to allow for more general definitions that can accommodate for potentially infinite spaces of pure strategies and a more careful control on a bound on $r(\overline{d}(\varpi_{(t-1)}), \pi^\dagger_{(t)}) - r(d^\dagger_{(t)}, \pi_{(t-1)})$ towards the end of the proof.

In this appendix, we slightly abuse the notation and use $\mathcal{D}$ to denote the compact set $\overline{\mathcal{D}}$ that contains all $d^{\dagger}_{(t)}$ $(t = 1, 2, \ldots)$. We first introduce the notion of cumulative Bayes risk functions. Under Algorithm 3, we let $U_0 : \mathcal{D} \to \mathbb{R}$ and $V_0 : \Gamma_\ell \to \mathbb{R}$ be any two continuous functions such that

$$\min_{d \in \mathcal{D}} U_0(d) = \max_{\pi \in \Gamma_\ell} V_0(\pi) \tag{3}$$

and recursively define

$$U_{t+1}(d) := U_t(d) + r(d, \pi^{\dagger}_{(t)}), \quad V_{t+1}(\pi) := V_t(\pi) + r(d^{\dagger}_{(t)}, \pi) \tag{4}$$

for $d \in \mathcal{D}$ and $\pi \in \Gamma_\ell$. Here, we let $\pi^{\dagger}_{(t)} \in \operatorname{argmax}_{\pi \in \Gamma_\ell} V_{t-1}(\pi)$ and $d^{\dagger}_{(t)} \in \operatorname{argmin}_{d \in \mathcal{D}} U_{t-1}(d)$. Note that the choices of $\pi^{\dagger}_{(t)}$ and $d_{(t)}$ in Algorithm 3 corresponds to setting $U_0 \equiv 0$ and $V_0 \equiv 0$, in which case $U_t(d) = t \cdot r(d, \pi_{(t)})$ and $V_t(\pi) = t \cdot r(\overline{d}(\varpi_{(t)}), \pi)$. In general,

$$U_t(d) = U_0(d) + t \cdot r(d, \pi_{(t)}), \quad V_t(\pi) = V_0(\pi) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi) \tag{5}$$

for some $\pi_{(t)} \in \Gamma$ and $\overline{d}(\varpi_{(t)}) \in \overline{\mathcal{D}}$. Later in this section, we will also make use of U_t and V_t with other initializations U_0 and V_0.

To make notations concise, we define $\min_{d \in \mathcal{D}'} U_t := \min_{d \in \mathcal{D}'} U_t(d)$ for any $\mathcal{D}' \subseteq \mathcal{D}$, and define $\max_{\mathcal{D}'} U_t$, $\min_{\Pi'} V_t$ and $\max_{\Pi'} V_t$ $(\Pi' \subseteq \Gamma_\ell)$ similarly. We also drop the subscript denoting the set when the set is the whole space we consider, that is, $\mathcal{D}$ or Γ_ℓ. Note that for any $t_1, t_2 = 1, 2, \ldots$, under the setting of Algorithm 3 and (2), it holds that

$$\begin{aligned} &\min U_{t_1}/t_1 = \min_{\overline{d} \in \overline{\mathcal{D}}} r(\overline{d}, \pi_{(t_1)}) \\ &\leq \max_{\pi \in \Gamma_\ell} \min_{\overline{d} \in \overline{\mathcal{D}}} r(\overline{d}, \pi) = r(\overline{d}(\varpi^*_\ell), \pi^*_\ell) = \min_{\overline{d} \in \overline{\mathcal{D}}} \max_{\pi \in \Gamma_\ell} r(\overline{d}, \pi) \\ &\leq \max_{\pi \in \Gamma_\ell} r(\overline{d}(\varpi_{(t_2)}), \pi) = \max V_{t_2}/t_2 \end{aligned}$$

Therefore, to prove the first result in Theorem 3, it suffices to show that $\limsup_{t \to \infty} (\max V_t - \min U_t)/t \leq 0$.

We next introduce additional definitions related to iterations. We say that $\pi \in \Gamma_\ell$ is eligible in the interval $[t_1, t_2]$ if there exists $t \in [t_1, t_2]$ such that $V_t(\pi) = \max V_t$; we say that $d \in \mathcal{D}$ is eligible in the interval $[t_1, t_2]$ if there exists $t \in [t_1, t_2]$ such that $U_t(d) = \min U_t$. We also define eligibility for sets. We say that $\Pi' \subseteq \Gamma_\ell$ is eligible in the interval $[t_1, t_2]$ if there exists $\pi \in \Pi'$ that is eligible in that interval; we say that $\mathcal{D}' \subseteq \mathcal{D}$ is eligible in the interval $[t_1, t_2]$ if there exists $d \in \mathcal{D}'$ that is eligible in the interval $[t_1, t_2]$. In addition, for any $\mathcal{D}' \subseteq \mathcal{D}$, we define maximum variation $\mathrm{MV}_t(\mathcal{D}') := \sup_{d \in \mathcal{D}'} U_t(d) - \inf_{d \in \mathcal{D}'} U_t(d)$ and $\mathrm{MV}_t(\Pi')$ similarly for any $\Pi' \subset \Gamma_\ell$. By Condition 2, there exists $B \in (0, \infty)$ such that $R \in [-B, B]$. Note that by Condition 1 and Lemma 2, given an arbitrary desired approximation accuracy $\epsilon > 0$, $\mathcal{D}$ can be covered by finitely many compact subsets with the maximum variation of each subset bounded by

ϵt for all t; by Condition 2, since Γ_ℓ is parameterized by a compact subset of a simplex in a Euclidean space, Γ_ℓ can also be covered by finitely many compact subsets with the maximum variation of each subset bounded by ϵt for all t. These covers can be viewed as discrete finite approximations to $\mathcal{D}$ and Γ_ℓ, respectively.

All of the above definitions are associated with the space of estimators $\mathcal{D}$ and the set of priors Γ_ℓ. We call $\{(U_t, V_t)\}_t$ a pair of cumulative Bayes risk functions constructed from the pair $(\mathcal{D}, \Gamma_\ell)$ of the space of estimators and the set of priors, and will consider pairs of cumulative Bayes risk functions constructed from other pairs $(\mathcal{D}', \Pi')$ of the space of estimators and the set of priors in the subsequent proof. We can define the above quantities similarly for such cases.

The following lemma gives an upper bound on the maximum variation of U_{s+t} and V_{s+t} over the corresponding entire space from which they are constructed after t iterations from s when essentially all parts of these spaces are eligible in $[s, s+t]$.

Lemma 4. *Suppose that $\{(U_t, V_t)\}_t$ is a pair of cumulative Bayes risk functions constructed from $(\mathcal{D}', \Pi')$. Suppose that $\mathcal{D}' = \bigcup_{i=1}^I \mathcal{D}_i$ and $\Pi' = \bigcup_{j=1}^J \Pi_j$ where*

$$\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \leq A, \quad \sup_{j,t} \mathrm{MV}_t(\Pi_j)/t \leq A$$

for $A < \infty$. If all $\mathcal{D}_i$ and Π_j are eligible in $[s, s+t]$, then $\max_{\mathcal{D}'} U_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (2B + A)t$ and $\max_{\Pi'} V_{s+t} - \min_{\Pi'} V_{s+t} \leq (2B + A)t$.

Proof of Lemma 4. Without loss of generality, assume that

$$\tilde{d} \in (\operatorname*{argmax}_{d \in \mathcal{D}'} U_{s+t}) \bigcap \mathcal{D}_1.$$

Since $\mathcal{D}_1$ is eligible in $[s, t]$, there exists $\tilde{t} \in [s, s+t]$ such that $(\operatorname{argmin}_{d \in \mathcal{D}'} U_{\tilde{t}}) \bigcap \mathcal{D}_1 \neq \emptyset$. By the recursive definition of the sequence $\{U_t\}_t$ in (4), the bound on the risk, and the assumption that $\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \leq A$, we have that $\max_{\mathcal{D}'} U_{s+t} = U_{s+t}(\tilde{d}) \leq U_{\tilde{t}}(\tilde{d}) + B(s+t-\tilde{t}) \leq \min_{\mathcal{D}'} U_{\tilde{t}} + At + B(s+t-\tilde{t}) \leq \min_{\mathcal{D}'} U_{\tilde{t}} + (A+B)t$. Letting $\tilde{d}' \in \operatorname{argmin}_{d \in \mathcal{D}'} U_{s+t}$, by the bound on the risk, we can derive that $\min_{\mathcal{D}'} U_{s+t} = U_{s+t}(\tilde{d}') \geq U_{\tilde{t}}(\tilde{d}') - B(s+t-\tilde{t}) \geq \min_{\mathcal{D}'} U_{\tilde{t}} - Bt$. Combine these two inequalities and we have that $\max_{\mathcal{D}'} U_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (2B + A)t$. An identical argument applied to the sequence $\{V_t\}_t$ shows that $\max_{\Pi'} V_{s+t} - \min_{\Pi'} V_{s+t} \leq (2B + A)t$. □

The next lemma builds on the previous lemma and provides an upper bound on $\max V_{s+t} - \min U_{s+t}$ under the same conditions.

Lemma 5. *Under the same setup and conditions as in Lemma 4, $\max_{\Pi'} V_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (4B + 2A)t$.*

Proof of Lemma 5. Summing the two inequalities in Lemma 4 and rearranging the terms, we have that $\max_{\Pi'} V_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (4B + 2A)t + \min_{\Pi'} V_{s+t} - \max_{\mathcal{D}'} U_{s+t}$. It therefore suffices to show that $\min_{\Pi'} V_{s+t} \leq \max_{\mathcal{D}'} U_{s+t}$.

Let $\tau := s + t$. There exists $\pi' \in \Pi'$ and a stochastic strategy $\overline{d}' \in \mathcal{D}'$ such that $U_\tau(d) = U_0(d) + \tau \cdot r(d, \pi')$ and $V_\tau(\pi) = V_0(\pi) + \tau \cdot r(\overline{d}', \pi)$ for all $d \in \mathcal{D}'$

and all $\pi \in \Pi'$. Therefore, for this choice of π' and $\overline{d}'$, using (3), $\min_{\Pi'} V_\tau \leq V_\tau(\pi') = V_0(\pi') + \tau \cdot r(\overline{d}', \pi') \leq \max_{\Pi'} V_0 + \tau \cdot r(\overline{d}', \pi') = \min_{\mathcal{D}'} U_0 + \tau \cdot r(\overline{d}', \pi') \leq U_0(\overline{d}') + \tau \cdot r(\overline{d}', \pi') = U_\tau(\overline{d}') \leq \max_{\mathcal{D}'} U_\tau$. □

Proof of Theorem 3. It suffices to show that $\limsup_{t\to\infty}(\max V_t - \min U_t)/t \leq 0$ by letting $U_0 \equiv 0$ and $V_0 \equiv 0$, which corresponds to Algorithm 3. Let $\epsilon > 0$. Note that r is Lipschitz continuous by Lemma 2 and the fact that $r(d, \pi)$ is an average of bounded risks with weights π. Furthermore, $\mathcal{D}$ and Γ_ℓ are both compact. In addition, U_0 and V_0 are both continuous. Therefore, there exist covers $\mathcal{D} = \bigcup_{i=1}^{I} \mathcal{D}_i$ and $\Gamma_\ell = \bigcup_{j=1}^{J} \Pi_j$ such that (i) $\mathcal{D}_i$ and Π_j are all compact, and (ii) $\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \leq \epsilon$, $\sup_{j,t} \mathrm{MV}_t(\Pi_j)/t \leq \epsilon$. (Note that I and J may depend on ϵ.) For index sets $\mathcal{I} \subseteq \{1, 2, \ldots, I\}$ and $\mathcal{J} \subseteq \{1, 2, \ldots, J\}$, define $\mathcal{D}_\mathcal{I} := \bigcup_{i\in\mathcal{I}} \mathcal{D}_i$ and $\Pi_\mathcal{J} := \bigcup_{j\in\mathcal{J}} \Pi_j$. We show that $\max V_t - \min U_t \leq C\epsilon t$ for an absolute constant C and all sufficiently large t via induction on the sizes of $\mathcal{I}$ and $\mathcal{J}$.

Let $\{(U_t, V_t)\}_t$ be a pair of cumulative Bayes risk functions constructed from $(\mathcal{D}_\mathcal{I}, \Pi_\mathcal{J})$ where $|\mathcal{I}| = |\mathcal{J}| = 1$. By (5) and the fact that $\mathrm{MV}_t(\mathcal{D}_\mathcal{I}) \leq \epsilon t$ and $\mathrm{MV}_t(\Pi_\mathcal{J}) \leq \epsilon t$, we have that

$$\begin{aligned}
\min_{\mathcal{D}_\mathcal{I}} U_t = \min_{d\in\mathcal{D}_\mathcal{I}} [U_0(d) + t \cdot r(d, \pi_{(t)})] &\geq \mathbb{E}_{d\sim\varpi_{(t)}}[U_0(d)] + t \cdot r(\overline{d}(\varpi_{(t)}), \pi_{(t)}) - \epsilon t \\
&\geq \min_{d\in\mathcal{D}_\mathcal{I}} U_0(d) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi_{(t)}) - \epsilon t \\
&= \max_{\pi\in\Pi_\mathcal{J}} V_0(\pi) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi_{(t)}) - \epsilon t \\
&\geq V_0(\pi_{(t)}) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi_{(t)}) - \epsilon t \\
&\geq \max_{\pi\in\Pi_\mathcal{J}} [V_0(\pi) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi)] - 2\epsilon t = \max_{\Pi_\mathcal{J}} V_t - 2\epsilon t.
\end{aligned}$$

Therefore, $\max_{\Pi_\mathcal{J}} V_t - \min_{\mathcal{D}_\mathcal{I}} U_t \leq 2\epsilon t$.

Let $\epsilon' > 0$ be arbitrary. Suppose that there exists t_0 such that, for any $\mathcal{I}' \subseteq \mathcal{I}$ and $\mathcal{J}' \subseteq \mathcal{J}$ such that $\mathcal{I}' \neq \mathcal{I}$ or $\mathcal{J}' \neq \mathcal{J}$, for any pair of cumulative Bayes risk functions $\{(U_t, V_t)\}_t$ constructed from $(\mathcal{D}_{\mathcal{I}'}, \Pi_{\mathcal{J}'})$, it holds that $\max_{\Pi_{\mathcal{J}'}} V_t - \min_{\mathcal{D}_{\mathcal{I}'}} U_t \leq \epsilon' t$ for all $t \geq t_0$. We next obtain a slightly greater bound on $\max_{\Pi_\mathcal{J}} V_t - \min_{\mathcal{D}_\mathcal{I}} U_t$ for all sufficiently large t.

We first prove that if, for a given pair of cumulative Bayes risk functions $\{(U_t, V_t)\}_t$ constructed from $(\mathcal{D}_\mathcal{I}, \Pi_\mathcal{J})$, there exists $i' \in \mathcal{I}$ or $j' \in \mathcal{J}$ such that $\mathcal{D}_{i'}$ or $\Pi_{j'}$ is not eligible in an interval $[s, s + t_0]$, then

$$\max_{\Pi_\mathcal{J}} V_{s+t_0} - \min_{\mathcal{D}_\mathcal{I}} U_{s+t_0} \leq \max_{\Pi_\mathcal{J}} V_s - \min_{\mathcal{D}_\mathcal{I}} U_s + \epsilon' t_0. \tag{6}$$

Suppose that $\mathcal{D}_{i'}$ is not eligible in $[s, s + t_0]$, then define $U'_t := U_{s+t}$ and $V'_t := V_{s+t} - \max_{\Pi_\mathcal{J}} V_s + \min_{\mathcal{D}_\mathcal{I}} U_s$ for all $t \geq 0$. It is straightforward to check that $\{(U'_t, V'_t)\}_{t=0}^{t_0}$ satisfies the recursive definition of a pair of cumulative Bayes risk functions constructed from $(\mathcal{D}_{\mathcal{I}\setminus\{i'\}}, \Pi_\mathcal{J})$. By the induction hypothesis, $\max_{\Pi_\mathcal{J}} V'_{t_0} - \min_{\mathcal{D}_{\mathcal{I}\setminus\{i'\}}} U'_{t_0} \leq \epsilon' t_0$. Therefore, $\max_{\Pi_\mathcal{J}} V_{s+t_0} - \min_{\mathcal{D}_\mathcal{I}} U_{s+t_0} = \max_{\Pi_\mathcal{J}} V'_{t_0} - \min_{\mathcal{D}_{\mathcal{I}\setminus\{i'\}}} U'_{t_0} + \max_{\Pi_\mathcal{J}} V_s - \min_{\mathcal{D}_\mathcal{I}} U_s \leq \max_{\Pi_\mathcal{J}} V_s - \min_{\mathcal{D}_\mathcal{I}} U_s + \epsilon' t_0$. Similar argument can be applied if $\Pi_{j'}$ is not eligible in $[s, s + t_0]$.

Now we obtain a bound on $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t$. Let $t > t_0$, $\mathscr{Q} := \lfloor t/t_0 \rfloor \geq 1$ and $\mathscr{R} := t/t_0 - \mathscr{Q} \in [0,1)$. There are two cases.

Case 1: There exists $s_0 \leq \mathscr{Q}$ such that $\mathcal{D}_i$ and Π_j are eligible in $[(s_0 - 1 + \mathscr{R})t_0, (s_0 + \mathscr{R})t_0]$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Take s_0 to be the largest such integer. Then, repeatedly apply (6) to intervals $[(s_0 + \mathscr{R})t_0, (s_0 + 1 + \mathscr{R})t_0], [(s_0 + 1 + \mathscr{R})t_0, (s_0 + 2 + \mathscr{R})t_0], \ldots, [(\mathscr{Q} - 1 + \mathscr{R})t_0, (\mathscr{Q} + \mathscr{R})t_0] = [t - t_0, t]$ and we derive that

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq \max_{\Pi_{\mathcal{J}}} V_{(s_0+\mathscr{R})t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{(s_0+\mathscr{R})t_0} + \epsilon'(\mathscr{Q} - s_0)t_0.$$

By Lemma 5, $\max_{\Pi_{\mathcal{J}}} V_{(s_0+\mathscr{R})t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{(s_0+\mathscr{R})t_0} \leq (4B + \epsilon)t_0$. Therefore,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B + \epsilon)t_0 + \epsilon'(\mathscr{Q} - s_0)t_0 \leq (4B + \epsilon)t_0 + \epsilon' t.$$

Case 2: There is no integer s_0 satisfying the condition in Case 1. Then, repeatedly apply (6) to intervals $[\mathscr{R}t_0, (1 + \mathscr{R})t_0], [(1 + \mathscr{R})t_0, (2 + \mathscr{R})t_0], \ldots, [(\mathscr{Q} - 1 + \mathscr{R})t_0, (\mathscr{Q} + \mathscr{R})t_0] = [t - t_0, t]$, we derive that

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq \max_{\Pi_{\mathcal{J}}} V_{\mathscr{R}t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{\mathscr{R}t_0} + \epsilon' \mathscr{Q} t_0.$$

By the bound on the risk, $\max_{\Pi_{\mathcal{J}}} V_{\mathscr{R}t_0} \leq B\mathscr{R}t_0$ and $\min_{\mathcal{D}_{\mathcal{I}}} U_{\mathscr{R}t_0} \geq -B\mathscr{R}t_0$. Hence,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq 2B\mathscr{R}t_0 + \epsilon' \mathscr{Q} t_0 \leq (4B + \epsilon)t_0 + \epsilon' t.$$

Thus, in both cases, it holds that $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B + \epsilon)t_0 + \epsilon' t$ for $t > t_0$. Let $C > 0$ be any constant (which may depend on ϵ, the approximation error of the covers, that is, the bound on MV_t/t). The following holds for any sufficiently large t,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B + \epsilon)t_0 + \epsilon' t \leq (1 + C)\epsilon' t. \tag{7}$$

In other words, we show that after increasing the size of either index set by 1, for all sufficiently large t, we obtain a bound on $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t$ that grows by a multiplicative factor of $(1 + C)$ relative to the original bound.

It takes finitely many, say N, steps to induct from the initial case where the sizes of both index sets are one to the case of interest with index sets $\{1, \ldots, I\}$ and $\{1, \ldots, J\}$. (Note that N may also depend on ϵ through its dependence on I and J.) Take $C = 1/N$ in (7) and we derive that, for all sufficiently large t,

$$\max V_t - \min U_t = \max_{\Pi_{\{1,\ldots,J\}}} V_t - \min_{\mathcal{D}_{\{1,\ldots,I\}}} U_t \leq (1 + 1/N)^N \cdot 2\epsilon t \leq 2e\epsilon t$$

where e is the base of natural logarithm. Since ϵ is arbitrary, we show that $\limsup_{t\to\infty}(\max V_t - \min U_t)/t \leq 0$. □

E.5. Derivation of Γ-minimax estimator of the mean in Section C

In this section, we show that, for the problem of estimating the mean in Section C, one Γ-minimax estimator lies in $\mathcal{D}_{\text{linear}}$. This is formally presented below.

Proposition 1. *Let $\mathcal{M}$ consist of all probability distributions defined on the Borel σ-algebra on $[0,1]$. Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} P_0 \in \mathcal{M}$ and $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be the observed data. Let $\Psi : P \mapsto \int xP(\mathrm{d}x)$ denote the mean parameter and $\Gamma = \{\pi \in \Pi : \int \Psi(P)\pi(\mathrm{d}P) = \mu\}$ be the set of priors that represent prior information. Let $\mathcal{D}$ denote the space of estimators that are square-integrable with respect to all $P \in \mathcal{M}$. Consider the risk in Example 1, $R : (d,P) \mapsto \mathbb{E}_P[(d(\mathbf{X}) - \Psi(P))^2]$. Define $\bar{X} = \sum_{i=1}^n X_i/n$ and $d_0 : \mathbf{X} \mapsto (\mu + \sqrt{n}\bar{X})/(1+\sqrt{n})$. Then $d_0 \in \mathcal{D}_{\text{linear}}$ is Γ-minimax over $\mathcal{D}$.*

We first present a theorem on a criterion of Γ-minimaxity.

Theorem 7. *Suppose that $d_0 \in \mathcal{D}$ is a Bayes estimator for $\pi_0 \in \Gamma$ and $r(d_0,\pi_0) = r_{\text{sup}}(d_0,\Gamma)$. Then d_0 is a Γ-minimax estimator in $\mathcal{D}$.*

Proof of Theorem 7. Clearly $r_{\text{sup}}(d_0,\Gamma) \geq \inf_{d\in\mathcal{D}} r_{\text{sup}}(d,\Gamma)$. Fix $d' \in \mathcal{D}$. Then, $r_{\text{sup}}(d',\Gamma) \geq r(d',\pi_0) \geq r(d_0,\pi_0) = r_{\text{sup}}(d_0,\Gamma)$. Since d' is arbitrary, this shows that $\inf_{d\in\mathcal{D}} r_{\text{sup}}(d,\Gamma) \geq r_{\text{sup}}(d_0,\Gamma)$. Thus, $r_{\text{sup}}(d_0,\Gamma) = \inf_{d\in\mathcal{D}} r_{\text{sup}}(d,\Gamma)$ and d_0 is Γ-minimax. □

We now present a lemma that is used to prove Proposition 1.

Lemma 6. *Let $a < b$ and suppose that $\mathcal{M}$ denotes the model space that consists of all probability distributions defined on the Borel σ-algebra on $[a,b] \subseteq \mathbb{R}$ with mean $\mu \in [a,b]$. Let X denote a generic random variable generated from some $P \in \mathcal{M}$. Then $\max_{P\in\mathcal{M}} \mathrm{Var}_P(X) = \mathrm{Var}_{P^*}(X) = (b-\mu)(\mu-a)$, where P^* is defined by $P^*(X=a) = (b-\mu)/(b-a)$ and $P^*(X=b) = (\mu-a)/(b-a)$.*

Proof of Lemma 6. Without loss of generality, we may assume that $a = -1$ and $b = 1$. Note that for any $P \in \mathcal{M}$, it holds that $\mathrm{Var}_P(X) = \mathbb{E}_P[X^2] - \mathbb{E}_P[X]^2 = \mathbb{E}_P[X^2] - \mu^2 \leq 1 - \mu^2$, where the equality is attained if $P(X \in \{-1,1\}) = 1$. Therefore, the maximum variance is achieved at the distribution with the specified mean μ and support being $\{a,b\}$, that is, at the distribution P^* defined in the lemma statement. Straightforward calculations show that $\mathrm{Var}_{P^*}(X) = (b-\mu)(\mu-a)$. □

Proof of Proposition 1. Let $\mathcal{M}' := \{\text{Bernoulli}(\theta) : \theta \in (0,1)\} \subseteq \mathcal{M}$ and let π_0 be a prior distribution over $\mathcal{M}'$ such that the prior distribution on the success probability θ is $\text{Beta}(\mu\sqrt{n}, (1-\mu)\sqrt{n})$. By Theorem 1.1 in Chapter 4 of Lehmann and Casella (1998), a Bayes estimator for π_0 minimizes the risk under the posterior distribution, whose minimizer over $\mathcal{D}$ is the posterior mean d_0 for our choice of risk. That is, d_0 is a Bayes estimator in $\mathcal{D}$ for π_0.

We next show that $r(d_0,\pi_0) = \sup_{\pi\in\Gamma} r(d_0,\pi)$. Let $\pi \in \Gamma$ be arbitrary. Since

Table 7
Summary of frequently used symbols.

Symbol	
P_0	True data-generating mechanism
$\mathcal{M}$	Space of data-generating mechanisms containing P_0
$\mathbf{X}^*$ and $\mathbf{X} = \mathcal{C}(\mathbf{X}^*)$	Full generated data and coarsened data
$\mathcal{D}$	Space of candidate estimators or decision functions (e.g., neural networks)
R	Risk function
r	Bayes risk function $r : (d, \pi) \mapsto \int R(d, P)\pi(\mathrm{d}P)$
$\Gamma(\subseteq \Pi)$	Set of prior distributions consistent with prior knowledge
Ψ	Functional defining the estimand $\Psi(P_0)$ in Examples 1–3
$\mathcal{M}_\ell$	An increasing sequence of finite subsets of $\mathcal{M}$
Γ_ℓ	Set of priors in Γ with support in $\mathcal{M}_\ell$
r_{sup}	Worst-case Bayes risk function $r_{\mathrm{sup}} : (d, \Gamma') \mapsto \sup_{\pi \in \Gamma'} r(d, \pi)$
d_ℓ^*	Γ_ℓ-minimax estimator in $\mathcal{D}$
d^*	A limit point of sequence $\{d_\ell^*\}_{\ell=1}^\infty$, which is Γ-minimax in $\mathcal{D}$ by Theorem 1
$\beta(\in \mathbb{R}^D)$	Coefficient of a finite-dimensional estimator (e.g., neural network)
$\xi \sim \Xi$	Exogenous randomness
$\hat{R}(\beta, P, \xi)$	Unbiased approximation of $R(\beta, P)$
$\overline{d}(\varpi)$	Stochastic estimator following distribution ϖ over $\mathcal{D}$
$\overline{\mathcal{D}}$	Space of stochastic estimators $\overline{d}(\varpi)$
$\overline{d}^* = \overline{d}(\varpi^*)$	Γ-minimax estimator in $\overline{\mathcal{D}}$

$\mathbb{E}_P[\bar{X}] = \Psi(P)$ and $\mathrm{Var}_P(\bar{X}) = \mathrm{Var}_P(X_1)/n$, we can derive that

$$\begin{aligned} r(d_0, \pi) &= \int \mathbb{E}_P\left[\left\{\frac{\mu + \sqrt{n}\bar{X}}{1+\sqrt{n}} - \Psi(P)\right\}^2\right]\pi(\mathrm{d}P) \\ &= \int \mathbb{E}_P\left[\left\{\frac{\sqrt{n}}{1+\sqrt{n}}\left(\bar{X} - \Psi(P)\right) + \frac{\mu - \Psi(P)}{1+\sqrt{n}}\right\}^2\right]\pi(\mathrm{d}P) \\ &= \int \left\{\frac{1}{(1+\sqrt{n})^2}\mathrm{Var}_P(X_1) + \frac{(\mu - \Psi(P))^2}{(1+\sqrt{n})^2}\right\}\pi(\mathrm{d}P) \end{aligned}$$

Apply Lemma 6 to $\mathrm{Var}_P(X_1)$ and the display continues as

$$\begin{aligned} &\leq \int \left\{\frac{1}{(1+\sqrt{n})^2}\Psi(P)(1-\Psi(P)) + \frac{(\mu - \Psi(P))^2}{(1+\sqrt{n})^2}\right\}\pi(\mathrm{d}P) \\ &= \int \frac{1}{(1+\sqrt{n})^2}\left\{\mu^2 + (1-2\mu)\Psi(P)\right\}\pi(\mathrm{d}P) = \frac{\mu(1-\mu)}{(1+\sqrt{n})^2}. \end{aligned}$$

This upper bound can be attained by any π with support contained in $\mathcal{M}'$, for example, π_0. Therefore, $r_{\mathrm{sup}}(d_0, \Gamma) = r(d_0, \pi_0)$. By Theorem 7, d_0 is Γ-minimax over $\mathcal{D}$. □

Funding

Generous support was provided by Amazon through an AWS Machine Learning Research Award, the NIH under award number DP2-LM013340, and the NSF under award number DMS-2210216. The content is solely the responsibility of the authors and does not necessarily represent the official views of Amazon, the NIH, or the NSF.

References

Amazon (2019). Amazon EC2 Instance Types – Amazon Web Services.

Bartlett, P. L. (1997). For valid generalization, the size of the weights is more important than the size of the network. *Advances in Neural Information Processing Systems* 134–140. MR1607706

Bartlett, P. L., Foster, D. J. and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems* **2017-December** 6241–6250.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics*. Springer New York, New York, NY. MR0804611

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models* **4**. Johns Hopkins University Press Baltimore. MR1245941

Birmingham, J., Rotnitzky, A. and Fitzmaurice, G. M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **65** 275–297. MR1959827

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity analysis of production and allocation* **13** 374–376. MR0056265

Bryan, B., McMahan, H. B., Schafer, C. M. and Schneider, J. (2007). Efficiently computing minimax expected-size confidence regions. *ACM International Conference Proceeding Series* **227** 97–104.

Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the Number of Species in Microbial Diversity Studies. *Annual Review of Statistics and Its Application* **1** 427–445.

Chen, X. (2007). Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. *Handbook of Econometrics* **6** 5549–5632.

Chen, L., Eichenauer-Herrmann, J. and Lehn, J. (1988). Gamma-minimax estimators for the parameters of a multinomial distribution. *Applicationes Mathematicae* **20** 561–564. MR1088727

Chen, L., Eichenauer-Herrmann, J., Hofmann, H. and Kindler, J. (1991). *Gamma-minimax estimators in the exponential family*. Polska Akademia Nauk, Instytut Matematyczny. MR1097077

Csáji, B. (2001). Approximation with artificial neural networks Technical Report.

Eckle, K. and Schmidt-Hieber, J. (2019). A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Networks* **110** 232–242.

Eichenauer-Herrmann, J. (1990). A gamma-minimax result for the class of symmetric and unimodal priors. *Statistical Papers* **31** 301–304. MR1124370

Eichenauer-Herrmann, J., Ickstadt, K. and Weiss, E. (1994). Gamma-minimax results for the class of unimodal priors. *Statistical Papers* **35** 43–56. MR1278642

Erhan, D., Courville, A., Bengio, Y. and Vincent, P. (2010). Why does unsupervised pre-training help deep learning? Technical Report.

Fan, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America* **39** 42. MR0055678

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures, Counter-Examples. 255–294. Springer, New York, NY.

Glorot, X., Bordes, A. and Bengio, Y. (2011). Deep sparse rectifier neural networks Technical Report.

Goel, S., Kanade, V., Klivans, A. and Thaler, J. (2016). Reliably Learning the ReLU in Polynomial Time.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets Technical Report No. January.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82** 711. MR1380810

Hanin, B. and Sellke, M. (2017). Approximating Continuous Functions by ReLU Nets of Minimal Width. *arXiv preprint arXiv:1710.11278v2*.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57** 97. MR3363437

Heitjan, D. F. (1993). Ignorability and Coarse Data: Some Biomedical Examples. *Biometrics* **49** 1099.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708. MR1326420

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and Coarse Data Technical Report No. 4. MR1135174

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4** 251–257.

Huang, G. B., Chen, L. and Siew, C. K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks* **17** 879–892.

Huang, F., Wu, X. and Huang, H. (2021). Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems* **34**.

Huang, G. B., Zhu, Q. Y. and Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing* **70** 489–501.

Jiang, S., Song, Z., Weinstein, O. and Zhang, H. (2020). Faster Dynamic Matrix Inverse for Faster LPs. *arXiv preprint arXiv:2004.07470v1*.

Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax Estimation of Functionals of Discrete Distributions. *IEEE Transactions on Information Theory* **61** 2835–2885. MR3342309

Kempthorne, P. J. (1987). Numerical Specification of Discrete Least Favorable Prior Distributions. *SIAM Journal on Scientific and Statistical Comput-*

ing **8** 171–184. MR0879409

Kidger, P. and Lyons, T. (2020). Universal Approximation with Deep Narrow Networks Technical Report.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics* **77**. Springer New York. MR2724368

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer. MR1639875

Lin, T., Jin, C. and Jordan, M. I. (2020). On gradient descent ascent for nonconvex-concave minimax problems. *37th International Conference on Machine Learning, ICML 2020* **PartF168147-8** 6039–6049.

Luedtke, A., Chung, I. and Sofrygin, O. (2020). Adversarial Monte Carlo Meta-Learning of Optimal Prediction Procedures. *arXiv preprint arXiv:2002.11275v1*. MR4353034

Luedtke, A., Carone, M., Simon, N. and Sofrygin, O. (2020). Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Science Advances* **6** eaaw2140.

Maron, H., Fetaya, E., Segol, N. and Lipman, Y. (2019). On the Universality of Invariant Networks.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21** 1087–1092.

Miller, R. I. and Wiegert, R. G. (1989). Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology* **70** 16–22. MR4497271

Nelson, W. (1966). Minimax Solution of Statistical Decision Problems by Iteration. *The Annals of Mathematical Statistics* **37** 1643–1657. MR0198635

Neyshabur, B., Bhojanapalli, S., McAllester, D. and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in Neural Information Processing Systems* **2017-December** 5948–5957.

Noubiap, R. F. and Seidel, W. (2001). An algorithm for calculating Γ-minimax decision rules under generalized moment conditions. *Annals of Statistics* **29** 1094–1116. MR1869242

Olman, V. and Shmundak, A. (1985). Minimax Bayes estimation of mean of normal law for the class of unimodal a priori distributions. *Proc. Acad. Sci. Estonian Physics Math* **34** 148–153. MR0796744

Orlitsky, A., Suresh, A. T. and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences of the United States of America* **113** 13283–13288. MR3582444

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox and R. Garnett, eds.)

8024–8035. Curran Associates, Inc.

Pfanzagl, J. (1990). Estimation in semiparametric models. 17–22. Springer, New York, NY. MR1048589

Pinelis, I. (2016). On the extreme points of moments sets. *Mathematical Methods of Operations Research* **83** 325–349. MR3513195

Qiu, H. (2022). QIU-Hongxiang-David/Gamma-minimax-learning: Simulation code for "Constructing Gamma-minimax estimators to leverage vague prior information". `https://github.com/QIU-Hongxiang-David/Gamma-minimax-learning/`. [Online; accessed 2022-03-14].

Robbins, H. (1951). Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability* 131–149. MR0044803

Robinson, J. (1951). An Iterative Method of Solving a Game. *The Annals of Mathematics* **54** 296. MR0043430

Sarma, A. and Kay, M. (2020). Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis. In *Conference on Human Factors in Computing Systems – Proceedings*. Association for Computing Machinery.

Schafer, C. M. and Stark, P. B. (2009). Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association* **104** 1080–1089. MR2562006

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27** 379–423. MR0026286

Shen, T. J., Chao, A. and Lin, C. F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84** 798–804.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics* **8** 171–176. MR0097026

Spielman, D. A. and Teng, S. H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM* **51** 385–463. MR2145860

Torrey, L. and Shavlik, J. (2009). Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* **11** 242–264.

v. Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* **100** 295–320. MR1512486

van der Vaart, A. and Wellner, J. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics.* Springer. MR4628026

Vidakovic, B. (2000). Γ-Minimax: A Paradigm for Conservative Robust Bayesians. 241–259. Springer, New York, NY. MR1795219

Wald, A. (1945). Statistical Decision Functions Which Minimize the Maximum Risk. *The Annals of Mathematics* **46** 265. MR0012402

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R. and Smola, A. J. (2017). Deep sets. *Advances in Neural Information Processing Systems* **2017-Decem** 3392–3402.

Zhang, Y., Lee, J. D. and Jordan, M. I. (2016). L1-regularized neural networks are improperly learnable in polynomial time. *33rd International Conference on Machine Learning, ICML 2016* **3** 1555–1563.