SOME EXISTENCE RESULTS FOR MAXIMIN PRIORS IN STATISTICAL MINIMAX THEOREMS

PAUL DELATTE

University of Southern California, *delatte@usc.edu*

We extend statistical minimax theorems for the average risk by providing general conditions under which maximin priors exist and are saddle points. We show that these conditions apply not only when the parameter space is compact, but also under the weaker condition that the priors have bounded moments. We illustrate the practicality of these conditions in the normal mean problem and the sparse normal mean problem where we readily recover known existence results and derive new ones. We then cast doubt on the possibility of extending Huber's approach to derive new existence results in minimax games for the average risk without any boundedness conditions on the parameter or the priors. We illustrate this issue in the normal mean problem and the sparse normal mean problem when the parameter space is the whole real line. As a corollary of independent interest, we show that Brown's identity does not hold for subprobability measures on the reals. This corrects a number of results available in Johnstone (2019). We finally show that existence results obtained in Bickel (1983) and Bickel and Collins (1983) when the parameter space is the extended real line $[-\infty, +\infty]$ is not equivalent to Huber's approach and impose much stronger bounds on the parameter than may appear at first. As a consequence, we call for caution when working with maximin priors for the average risk in absence of explicit bounds on the parameter space or on the priors' moments.

1 Introduction

1.1 Roadmap

Statistical minimax theorems guarantee under general conditions that the max-min inequality for the average risk obtained under minimizing decision rules and maximizing priors is an equality. This is a fundamental first step for evaluating the minimax properties of statistical procedures. Statistical minimax theorems were obtained in the 1940s and 1950s, notably under the influence of Wald and then Le Cam. The contribution of Le Cam was to extend the results of Wald to infinite parameter spaces by finding a satisfactory topology on decision rules so as to apply Kneser's minimax theorem. For these general results to hold, no topological conditions need to be imposed on the set of priors. As a consequence, the existence of maximin priors is generally not guaranteed. However, both from a theoretical viewpoint and from a computational viewpoint, it is of great interest to ensure that maximin priors exist. From a theoretical viewpoint, the existence of maximin priors completes the bridge built

This version: October 18, 2024.

between frequentists and Bayesians by statistical minimax theorems. It ensures, moreover, that maximin priors can be used for further theoretical constructions. From a computational viewpoint, the existence of maximin priors allows one to recover minimax rules by first computing a maximin prior and evaluating the Bayes rule for this distribution. Since this is often the only way to compute minimax rules, the existence of maximin priors is a problem of high practical relevance.

If we want maximin priors to exist, it is natural to start by endowing the set of priors with the weak topology. In Section 2, we provide general conditions for maximin priors to exist in this setting. The general existence result we obtain unifies a number of classical results scattered in the literature: it provides the weakest conditions for applying Kneser's minimax theorem and generalizations of the extreme value theorem. Interestingly, we show that these conditions still yield a saddle point property in full generality. Moreover, we provide general conditions under which the Bayes rule in the saddle point can be taken to be deterministic. We then show in Section 2.2 that these conditions obtain in a number of important cases. Beyond the natural applications when the parameter lies in a compact metric space, we show that it is possible to relax compactness if the priors' moments are bounded. This formalizes a tightening procedure first considered in Feldman (1991) and expanded in Donoho and Johnstone (1994). Because the bounded moments assumption on the priors is weaker than the compactness assumption on the parameter space[1], we advocate for its use in applications. In Section 2.3, we illustrate the practicality of the conditions we obtained by applying them to the normal mean problem and the sparse normal mean problem where we recover known existence results and derive new ones.

It is tempting to enlarge these results by considering a coarser topology on the set of priors so as to get compactness (and hence the existence of maximin priors) without any boundedness conditions on the parameter space or on the moments of the priors. The vague topology is a natural candidate that has been considered in the literature – see, e.g., Johnstone (2019) – following the seminal contributions of Huber in Huber (1964) for the minimization of the Fisher information. Indeed, under separability and local compactness of the parameter set, compactness directly obtains by considering the closure of the set of priors in the set of subprobability measures endowed with the vague topology[2]. While it is possible to derive a general existence result under this topology (as done in Appendix B), we show in Section 3 that the conditions for existence are rarely satisfied in practice. This is the manifestation of an implicit tension between (proper) subprobability measures and semicontinuity of the Bayes risk function – a tension that is absent in the minimization of the Fisher information as considered in Huber (1964). We illustrate this issue in the normal mean problem and the sparse normal mean problem when the parameter space is the whole real line. For these problems, we show that the Bayes risk fails to be vaguely upper semicontinuous, which prevents the application of compactness arguments to prove the existence of maximin priors. In the course of these illustrations, we obtain as

[1]The parameter space is the sample space of the prior. Compactness of the parameter space then implies that all moments of the priors are finite.

[2]Provided the Bayes risk is vaguely upper semicontinuous, the existence of maximin subprobability measures is guaranteed (see Appendix B). Then to prove the existence of maximin priors, one has to show that the maximin subprobability measures are probability measures.

a corollary of independent interest that Brown's identity does not hold for subprobability measures on the whole real line. This corrects a number of results available in Johnstone (2019).

These negative outcomes invite us to revisit existence results that were positively derived in the literature for the sparse normal mean problem when the parameter space is the extended real line $[-\infty, +\infty]$ – see Bickel (1983) and Bickel and Collins (1983). While these results suggest that boundedness assumptions can be relaxed at no cost, we show, however, that the use of the two-point compactification of the real line impose much stronger bounds on the parameter space than may appear at first. We prove that this construction is, contrarily to what has been claimed, not equivalent to the one used by Huber in Huber (1964) based on the one-point compactification of the real line. To make this clear, we exhibit the topological imports of both compactifications, recover the construction of Huber (1964), and show that the existence results in Bickel (1983) and Bickel and Collins (1983) for the sparse normal mean problem are actually obtained as if the parameter space was $[-\pi/2, \pi/2]$. We actually show that the bounds $-\pi/2$ and $\pi/2$ are arbitrary and depend ultimately on the homeomorphism considered by the modeler when defining $[-\infty, \infty]$. As a consequence, we call for caution when working with maximin priors for the average risk whenever neither the parameter space nor the priors' moments are explicitly bounded by some known and interpretable constants $m > 0$ or $M \in \mathbb{R}$. We also suggest to reconsider as an open problem the existence of a maximin prior in the sparse normal mean problem when the parameter space is the whole real line[3].

It is important to note that our results do not exhaust the question of whether maximin priors exist. Their main raison d'être is to guide and discipline the use of general tools to non-constructively prove existence. In particular, our negative results for the use of the vague topology do not prove that maximin priors do not exist, but only that standard arguments based on semicontinuity and compactness using variations of the extreme value theorem generally fail to deliver positive results in this case. Moreover, our results should not be viewed as saying that the vague topology cannot be used to prove existence in minimax problems involving probability measures: the vague topology has been used profitably in other minimax problems (as by Huber when minimizing the Fisher information); if it cannot deliver existence in the statistical games we consider, it is only due to the inherent properties of the Bayes risk. It is finally of interest to note that there likely exist workarounds for the negative results we obtained. In particular, if one has no issue with pure subprobability measures as priors, then there are two promising directions to obtain encompassing existence results: either using the q-vague topology as developed in Bioche and Druilhet (2016) or redefining the Bayes risk for pure subprobability measures. The second approach is briefly explored in Section 3.4, but a thorough treatment of the problem is left for future research.

1.2 Related literature

The notion of minimax in decision theory has a long history, starting at least with the early contributions of Borel and von Neumann in the 1920s – see Fréchet (1953).

[3]Numerical simulations, especially when compared with the normal mean problem where non-existence is known, leads us to conjecture that such a prior does not exist.

The use of minimax considerations for the determination and evaluation of statistical rules was initiated by Wald in the 1940s – see his textbook treatment in Wald (1950). His approach based on the average risk was rapidly extended by Blyth, Ghosh, Hodges, Kiefer, Lehmann, Le Cam, Stein, Wolfowitz, among others – see, for instance, Hodges and Lehmann (1950) and Lehmann (1952). An important generalization for our purpose was obtained by Le Cam in Le Cam (1955) with the first derivation of statistical minimax theorems for the average risk holding under general conditions. The general treatment of Wald and Le Cam rapidly crystallized in stable forms that can be found in textbooks – see, e.g., Ferguson (1967), Brown (1974), Berger (1985), Strasser (1985), and Le Cam (1986).

As far as we know, the existence of maximin priors was never considered in full generality in the context of these theorems. It was, however, tackled in the many applications of these results to specific problems. Due to the inherent difficulty of double optimization in minimax problems, the literature has had a tendency to focus on simple models. Among them, the normal mean problem has received a lot of attention, owing in particular to approximation results connecting it asymptotically to much more complicated models. The version with restricted parameter space became a popular research topic in the 1980s as it was (re)discovered, long after Ghosh (1964), that bounds on the parameter space altered minimax results drastically, from non-existence of maximin priors to the existence of finitely supported ones: see Casella and Strawderman (1981), Bickel (1981), Levit (1981), Zinzius (1981); see Marchand and Strawderman (2004) for a review. A second wave of results were obtained for more complicated versions of normal mean estimation, either due to more elaborate parameter sets as in Donoho, Liu, and MacGibbon (1990), Donoho and Johnstone (1994), Donoho and Johnstone (1996), Donoho and Johnstone (1998), or due to the presence of performance constraints or prior information as in Bickel (1983), Bickel and Collins (1983), Feldman (1991), Donoho and Johnstone (1994), Johnstone (1994). A textbook treatment of these different results can be found in Johnstone (2019).

In these problems, the existence of maximin priors has been mostly motivated by computational considerations. Indeed, in problems for which Bayes rules and maximin priors exist, it is often possible to numerically compute minimax rules and their minimax risk by first computing maximin priors and then evaluating the associated Bayes rules. This procedure is of great importance, since it is often the only one available to compute minimax rules. For an illustration of this procedure, see Nelson (1966), Casella and Strawderman (1981), Kempthorne (1987), Eichenauer and Lehn (1989), Gourdin, Jaumard, and MacGibbon (1994), Johnstone (1994), Chamberlain (2000), and Noubiap and Seidel (2001).

Interest in the existence of maximin priors of the nature considered in this paper has been recently revived in a number of fields, not the least in econometrics and in optimization. Such maximin priors have appeared both in the construction of theoretical solutions for uniform inference as well as in the computation of minimax rules in applied decision-making settings. Examples in econometrics include Elliott, Müller, and Watson (2015), Müller and Wang (2019), or Kline and Walters (2021). In the optimization literature, these maximin priors have gained prominence due notably to an increased interest in distributionally robust solutions. This interest has been fueled by recent advances in

computational optimal transport, in particular for the computation of maximin priors in Wasserstein balls. See, for instance, Mohajerin Esfahani and Kuhn (2018), Shafieezadeh Abadeh, Nguyen, Kuhn, and Mohajerin Esfahani (2018), Blanchet and Murthy (2019), or Gao and Kleywegt (2023). The results in our paper are here to assist these recent developments by providing easily verifiable conditions for existence while calling for caution when trying to relax too abruptly boundedness conditions on the priors.

1.3 Definitions

We collect here a number of definitions that will be used repeatedly in the paper. Other notations, definitions, and results that appear in the paper can be found in Appendix A.

A statistical experiment is a pair $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\})$ where $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is a measurable space and $\{P_\theta : \theta \in \Theta\}$ is a set of probability measures on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ indexed by a set Θ.

A statistical decision problem is a quadruple $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ where $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\})$ is a statistical experiment, $(\mathcal{A}, \mathcal{B}_\mathcal{A})$ is a measurable space called action space, and $L\colon \mathcal{A} \times \Theta \to [0, \infty]$ is a loss function.

A decision rule for the statistical experiment is a Markov kernel from $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ to $(\mathcal{A}, \mathcal{B}_\mathcal{A})$. The set of all decision rules for a given decision problem is denoted $\mathcal{D}$. If for every $x \in \mathcal{X}$, the decision rule δ is the Dirac measure at $T(x)$ for some measurable function $T\colon \mathcal{X} \to \mathcal{A}$, then the decision rule is said to be non-randomized or deterministic. In this case, the decision rule is often directly considered to be the function T. When $\mathcal{A}$ is a convex subset of a Euclidean space, a non-randomized rule can be obtained from a randomized one by averaging over $\mathcal{A}$, that is, by taking $T_\delta(x) = \int_\mathcal{A} a d\delta(x, a)$.

We assume that $\mathcal{X}$, $\mathcal{A}$, and Θ are topological spaces. We assume that $\mathcal{B}_\mathcal{A}$ is the Baire σ-algebra for $\mathcal{A}$. We endow Θ with its Borel σ-algebra which we denote $\mathcal{B}_\Theta$.

Given a statistical decision problem, we define the risk function $r\colon \mathcal{D} \times \Theta \to [0, +\infty]$ by

$$r(\delta, \theta) := \int_\mathcal{X} \int_\mathcal{A} L(a, \theta)\, d\delta(x, a)\, dP_\theta(x).$$

If the decision rule is a non-randomized rule δ_T inducing a measurable function T, then

$$r(T, \theta) := r(\delta_T, \theta) = \int_\mathcal{X} L(T(x), \theta)\, dP_\theta(x) = \mathbb{E}_{x \sim P_\theta}[L(T(x), \theta)].$$

Given a set $\mathcal{P}$ of probability measures on $(\Theta, \mathcal{B}_\Theta)$ called priors, we define the average risk or integrated risk $B\colon \mathcal{D} \times \mathcal{P} \to [0, +\infty]$ by

$$B(\delta, \pi) := \int_\Theta r(\delta, \theta)\, d\pi(\theta) = \mathbb{E}_{\theta \sim \pi}[r(\delta, \theta)].$$

The Bayes(ian) risk over $\mathcal{D}_0 \subseteq \mathcal{D}$ is then defined as the function $\underline{B}\colon \mathcal{P} \to [0, +\infty]$ given by

$$\underline{B}(\pi) := \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi) = \inf_{\delta \in \mathcal{D}_0} \int_\Theta r(\delta, \theta)\, d\pi(\theta).$$

We sometimes allow priors to be subprobability measures and so define the average risk and Bayesian risk over sets $\mathcal{P} \subseteq \mathcal{M}_{\leq 1}(\Theta)$ in a natural way by integrating in the Lebesgue sense over finite measures. The integral expressions and notations in this case remain unchanged.

A maximin prior or least favorable prior is a measure $\pi_0 \in \mathcal{P}$ such that

$$\underline{B}(\pi_0) = \sup_{\pi \in \mathcal{P}} \underline{B}(\pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi),$$

and $B(\mathcal{P}) := \underline{B}(\pi_0)$ is then said to be the maximum Bayesian risk.

A Bayes(ian) rule with respect to a prior $\pi \in \mathcal{P}$ is any rule $\delta_b \in \mathcal{D}_0$ such that

$$B(\delta_b, \pi) = \underline{B}(\pi) = \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

A minimax rule (for the risk) is any rule $\delta_m \in \mathcal{D}_0$ such that

$$\sup_{\theta \in \Theta} r(\delta_m, \theta) = \inf_{\delta \in \mathcal{D}_0} \sup_{\theta \in \Theta} r(\delta, \theta),$$

and $R_N(\Theta) := \inf_{\delta \in \mathcal{D}_0} \sup_{\theta \in \Theta} r(\delta, \theta)$ is then said to be the minimax risk (over $\mathcal{D}_0$). Note that minimaxity for the risk is a frequentist notion (as there is no mention of priors).

1.4 Regular statistical decision problems

We review here the notion of regular statistical decision problems as well as topological considerations on the set of decision rules that were initiated by Le Cam (1955) to derive statistical minimax theorems. These results will be used in later sections. We follow the construction of Brown (1974), which is partially reproduced in Johnstone (2019).

Definition 1 (Regular Statistical Decision Problem)**.** A statistical decision problem $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ is said to be regular if:

1. the action space $\mathcal{A}$ is a compact metric space (and hence also second-countable);
2. the family of probability measures $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure P_0 such that the space $L^1(\mathcal{X}, P_0)$ is a separable Banach space.

Lemma 1.1. *Let $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ be a regular statistical experiment. Then there is a vector topological space such that the set $\mathcal{D}$ of all decision rules for the statistical decision problem is a compact and convex subset of this space.*

Proof. The proof follows directly from Theorem 42.3 and Corollary 42.8 in Strasser (1985). The result Theorem 42.3 can be traced back to Le Cam (1955) and the result leading to Corollary 42.8 from Farrell (1966). For our purpose, we will use the construction of Theorem 3.9 in Brown (1974) (p.219) where a direct imbedding is worked out for the dominated case. The proof of Brown is partially reproduced p.405 in Johnstone (2019). □

Remark 1.1. A topology satisfying Lemma 1.1 can be constructed in two ways. Le Cam and Strasser identify $\mathcal{D}$ and a subset of $C_b(\mathcal{A}) \times L(\mathcal{X})$ where $L(\mathcal{X})$ is some space of finite measures called

L-space. In the dominated case, Brown identifies $\mathcal{D}$ and a subset of $C(\mathcal{A}) \times L_1(\mathcal{X}, P_0)$ through the imbedding $\delta \mapsto b_\delta$ where

$$b_\delta(c, g) := \int \int c(a)g(x)\, d\delta(x, a)\, dP_0(x).$$

The set $\mathcal{D}$ inherits both a topology and a linear structure from the bijection $\delta \mapsto b_\delta$. It is important to note that there is no assumption on the loss L for this result to hold, in particular L need not be in $C(\mathcal{A})$ nor $C_b(\mathcal{A})$ for the imbedding to be valid. In the rest of the paper, we will use the topology constructed above and call it, as typically done, the weak topology on $\mathcal{D}$.

Lemma 1.2. *Let $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ be a regular statistical experiment. Let $\mathcal{P}$ be a set of probability measures on $(\Theta, \mathcal{B}_\Theta)$. Let $\mathcal{D}_0 \subseteq \mathcal{D}$ be a set of decision rules where $\mathcal{D}$ is endowed with the weak topology. If the loss function L is lower semicontinuous in a for each $\theta \in \Theta$, then the integrated risk is lower semicontinuous in δ for each $\pi \in \mathcal{P}$.*

Proof. Corollary 11 in Brown (1974) (p.221) yields lower semicontinuity of the risk function. Then an application of Fatou's lemma concludes the proof (see Lemma A.3). See also p.405 in Johnstone (2019). The construction is identical to Theorem 43.2 in Strasser (1985). □

Remark 1.2. In the course of the proof of the previous result, a useful characterization of the risk is obtained. We reproduce it and the argument justifying it as it will be used in our existence results. Since $\mathcal{A}$ is second-countable and the loss function L is lower semicontinuous in a, it follows that, for a fixed θ, the loss L can be expressed as the limit of an increasing sequence of continuous functions of a. It thus follows that

$$r(\delta, \theta) = \sup_{c \in C(\mathcal{A})} \{b_\delta(c, f_\theta) : c \leq L(a, \theta)\}.$$

2 Existence of maximin priors under the weak topology

2.1 An existence result under the weak topology

We are now ready to state and prove a general existence result under the weak topology. The first part of the result, which yields a statistical minimax theorem for the average risk, is already known (modulo some small variations) – see, e.g., Theorem 1 in Le Cam (1986) (p.16) and the subsequent remarks; see also Theorem A.5 in Johnstone (2019) for a particular application. The second part, which proves the existence of a maximin prior, has, as far as we know, no counterpart in the existing literature, but proceeds by very standard arguments. The merit of the result is to gather assumptions as general as possible. The difficult part of verifying these assumptions in common applications is relegated to Section 2.2.

Proposition 2.1. *Let $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ be a regular statistical decision problem. Let $\mathcal{D}_0 \subseteq \mathcal{D}$ be a set of decision rules for the statistical decision problem. Let $\mathcal{P}$ be a set of probability measures on $(\Theta, \mathcal{B}_\Theta)$. Suppose that*

1. *for each θ, the loss L is lower semicontinuous in a;*
2. *the Bayes risk function $\pi \mapsto \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi)$ is weakly upper semicontinuous;*
3. *the set $\mathcal{P}$ is weakly compact and convex;*
4. *the set $\mathcal{D}_0$ is closed and convex as a subset of $\mathcal{D}$ endowed with the weak topology.*

Then there exists a pair $(\delta^*, \pi^*) \in \mathcal{D}_0 \times \mathcal{P}$ such that

$$B(\delta^*, \pi^*) = \inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \mathcal{P}} B(\delta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi). \tag{2.1.1}$$

and that (δ^*, π^*) is a saddle point in the sense that

$$B(\delta^*, \pi) \leq B(\delta^*, \pi^*) \leq B(\delta, \pi^*) \tag{2.1.2}$$

for all $\delta \in \mathcal{D}_0$ and all $\pi \in \mathcal{P}$.

Proof. By linearity of the space of finite measures, the integrated risk B is linear in π for each $\delta \in \mathcal{D}_0$. Moreover, from Remark 1.2, we have

$$r(\delta, \theta) = \sup_{c \in C(\mathcal{A})} \{b_\delta(c, f_\theta) : c \leq L(a, \theta)\},$$

hence r is convex in δ for each $\theta \in \Theta$, and so the integrated risk B is convex in δ for each $\theta \in \Theta$. By assumption, $\mathcal{P}$ is convex and $\mathcal{D}_0$ is convex and closed. By Lemma 1.1, $\mathcal{D}$ is compact and so $\mathcal{D}_0$ is compact. It thus follows from Kneser's minimax theorem (see Kuhn (1953)) that

$$\inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \mathcal{P}} B(\delta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

By assumption, $\mathcal{P}$ is weakly compact and $\pi \mapsto \inf_\delta B(\delta, \pi)$ is weakly upper semicontinuous, hence the supremum on the right-hand side is attained. Denote $\pi^* \in \mathcal{P}$ the distribution that attains this supremum, that is,

$$\sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi) = \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi^*)$$

By Lemma 1.2, $\delta \mapsto B(\delta, \pi)$ is lower semicontinuous for each $\pi \in \mathcal{P}$, and in particular for $\pi^* \in \mathcal{P}$. Since $\mathcal{D}_0$ is compact, the infimum on the right-hand side is also attained for some $\delta' \in \mathcal{D}_0$. Therefore, there exists a pair $(\delta', \pi^*) \in \mathcal{D}_0 \times \mathcal{P}$ such that

$$B(\delta', \pi^*) = \inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \mathcal{P}} B(\delta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

We now show that there exists a rule $\delta^* \in \mathcal{D}_0$ such that (δ^*, π^*) attains the minimax equality and is a saddle point. Since $\delta \mapsto B(\delta, \pi)$ is lower semicontinuous for each $\pi \in \mathcal{P}$, we have that $\delta \mapsto \sup_\pi B(\delta, \pi)$ is lower semicontinuous as the pointwise supremum of lower semicontinuous functions. Since $\mathcal{D}_0$ is compact, the infimum on the left-hand side of the minimax equality is attained. We thus have

$$\sup_{\pi \in \mathcal{P}} B(\delta^*, \pi) = B(\delta', \pi^*)$$

for some $\delta^* \in \mathcal{D}_0$. Then $B(\delta', \pi^*) \geq B(\delta^*, \pi^*) \geq \inf_\delta B(\delta, \pi^*) = B(\delta', \pi^*)$, and so the supremum on the left-hand side is also achieved for π^*. It follows that

$$B(\delta^*, \pi^*) = B(\delta', \pi^*) = \sup_{\pi \in \mathcal{P}} B(\delta^*, \pi) = \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi^*).$$

This concludes the proof. □

Remark 2.1. The saddle point characterization (2.1.2) rewrites as

$$B(\delta^*, \pi^*) = \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi^*) = \sup_{\pi \in \mathcal{P}} B(\delta^*, \pi). \tag{2.1.3}$$

Being a saddle point is sufficient for δ^* to be a Bayes rule for π^* (from the first equality, by definition) and for π^* to be a maximin prior (by combining the two equalities, since the first guarantees that $B(\delta^*, \pi^*) = \underline{B}(\pi^*)$ and the second that $\underline{B}(\pi^*) \geq B(\delta^*, \pi)$ for all $\pi \in \mathcal{P}$, and so that $\underline{B}(\pi^*) \geq \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi) = \underline{B}(\pi)$ for all $\pi \in \mathcal{P}$). Note that the fact that π^* is a maximin prior also follows from the minimax equality (2.1.1) and first equality of the saddle point formula (2.1.3).

Remark 2.2. The assumptions of Proposition 2.1 guarantee that the minimax equality (2.1.1) holds and is always attained for some pair of decision rule and prior. Given an arbitrary solution (δ', π') for (2.1.1), the proof of Proposition 2.1 shows that π' is always a maximin distribution and that δ' is always a Bayes rule for π'. On the other hand, (δ', π') need not be a saddle point, since the left-hand side inequality in (2.1.2) (or, equivalently, the second equality in (2.1.3)) need not hold. However, if the Bayes rule δ' for π' is unique, then the solution (δ', π') is necessarily a saddle point.

We now provide general conditions for the Bayes rule in the saddle point to be deterministic. The deterministic nature of the rule has important consequences in practice.

Corollary 2.2. *Suppose that:*

1. *$\mathcal{A}$ is a convex subset of a (possibly compactified) Euclidean space;*
2. *the loss function is convex in a for each $\theta \in \Theta$;*

then for any rule $\delta \in \mathcal{D}$, there exist a deterministic rule $T_\delta \in \mathcal{D}$ such that

$$B(T_\delta, \pi) \leq B(\delta, \pi) \tag{2.1.4}$$

for all $\pi \in \mathcal{P}$. If, moreover, the assumptions of Proposition 2.1 hold and $\mathcal{D}_0 = \mathcal{D}$, then the rule δ^* in Proposition 2.1 can be taken to be deterministic, and (2.1.1) and (2.1.2) hold both for $\mathcal{D}$ and for the restriction of $\mathcal{D}$ to the set $\mathcal{D}_d$ of deterministic rules.

Proof. Define $T_\delta(x) = \int_{\mathcal{A}} a d\delta(x, a)$, which is a measurable function from X to $\mathcal{A}$ since $\mathcal{A}$ is assumed convex. By convexity of L in a, we have by Jensen's inequality that

$$L(T_\delta(x), \theta) = L\left(\int a d\delta(x, a), \theta \right) \leq \int L(a, \theta)\, d\delta(x, a).$$

Integrating over P_θ yields by monotonicity of the integral that for all $\theta \in \Theta$

$$r(T_\delta, \theta) \leq r(\delta, \theta).$$

Integrating over arbitrary $\pi \in \mathcal{P}$ yields

$$B(T_\delta, \pi) \leq B(\delta, \pi) \tag{2.1.5}$$

and taking the supremum over $\mathcal{P}$ yield

$$\sup_{\pi \in \mathcal{P}} B(T_\delta, \pi) \leq \sup_{\pi \in \mathcal{P}} B(\delta, \pi). \tag{2.1.6}$$

Let us denote (δ^*, π^*) the solution in Proposition 2.1. Then

$$\begin{aligned} B(T_{\delta^*}, \pi^*) &\leq B(\delta^*, \pi^*) \\ &= \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi^*) \\ &\leq B(T_{\delta^*}, \pi^*). \end{aligned}$$

It follows that

$$B(T_{\delta^*}, \pi^*) = \inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \mathcal{P}} B(\delta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

We now prove that the minimax equality holds when restricting $\mathcal{D}$ to $\mathcal{D}_d$ and that it is also equal to $B(T_{\delta^*}, \pi^*)$. We have from (2.1.6) that

$$\sup_{\pi \in \mathcal{P}} B(T_{\delta'}, \pi) \leq \inf_{\delta \in \mathcal{D}} \sup_{\pi \in \mathcal{P}} B(\delta, \pi)$$

for any rule δ', and so

$$\inf_{T \in \mathcal{D}_d} \sup_{\pi \in \mathcal{P}} B(T, \pi) \leq \inf_{\delta \in \mathcal{D}} \sup_{\pi \in \mathcal{P}} B(\delta, \pi)$$

From the inclusion $\mathcal{D}_d \subseteq \mathcal{D}$, we have that

$$\sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}} B(\delta, \pi) \leq \sup_{\pi \in \mathcal{P}} \inf_{T \in \mathcal{D}_d} B(T, \pi).$$

From the minimax inequality, we have that

$$\sup_{\pi \in \mathcal{P}} \inf_{T \in \mathcal{D}_d} B(T, \pi) \leq \inf_{T \in \mathcal{D}_d} \sup_{\pi \in \mathcal{P}} B(T, \pi).$$

Combining everything, we can conclude that

$$B(T_{\delta^*}, \pi^*) = \inf_{T \in \mathcal{D}_d} \sup_{\pi \in \mathcal{P}} B(T, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{T \in \mathcal{D}_d} B(T, \pi).$$

We finally prove that (2.1.2) holds for (T_{δ^*}, π^*) for $\mathcal{D}$ and $\mathcal{D}_d$. We have

$$B(T_{\delta^*}, \pi^*) = \inf_{\delta \in \mathcal{D}} B(\delta, \pi^*) \leq \inf_{T \in \mathcal{D}_d} B(T, \pi^*) \leq \inf_{\delta \in \mathcal{D}} B(\delta, \pi^*) = B(T_{\delta^*}, \pi^*),$$

where the first inequality follows from the inclusion $\mathcal{D}_d \subseteq \mathcal{D}$ and the second from (2.1.5). Similarly, we have

$$B(T_{\delta^*}, \pi^*) \leq \sup_{\pi \in \mathcal{P}} B(T_{\delta^*}, \pi) \leq \sup_{\pi \in \mathcal{P}} B(\delta^*, \pi) = B(T_{\delta^*}, \pi^*),$$

where the second inequality follows again from (2.1.5). This concludes the proof. □

Before closing this section, we provide another existence result under slightly different conditions. The merit of this result is mostly in its proof, which differs from Proposition 2.1 and provides an early illustration of the tension between compactness and semicontinuity that will be explored in greater details later.

Corollary 2.3. *Suppose that Θ is a separable metric space. Suppose, moreover, that*

1. *the Bayesian risk is upper semicontinuous on the weak closure $\bar{\mathcal{P}}$;*
2. *for each $\delta \in \mathcal{D}_0$, the risk function r is lower semicontinuous in θ*

Then Proposition 2.1 holds under the weaker condition that $\mathcal{P}$ is weakly relatively compact instead of weakly compact. In particular, a maximin prior that belongs to $\mathcal{P}$ exists.

Proof. We first prove a preliminary result. Suppose π_n weakly converges to π. Then $\liminf \pi_n(O) \geq \pi(O)$ for all open sets $O \subseteq \Theta$. Since r is assumed lower semicontinuous, it has open superlevel sets. Since it is bounded from below by 0, we have that $\int r(\delta, \theta)\, d\pi_n(\theta) = \int_0^\infty \pi_n(\{\theta : r(\delta, \theta) > y\})\, dy$. Therefore, $\liminf_{n\to\infty} \int r(\delta, \theta)\, d\pi_n(\theta) \geq \int r(\delta, \theta)\, d\pi$ for every $\delta \in \mathcal{D}_0$.

Since the Bayesian risk is weakly continuous on $\bar{\mathcal{P}}$ and $\bar{\mathcal{P}}$ is weakly compact, we can apply Proposition 2.1 to the closure $\bar{\mathcal{P}}$ instead of $\mathcal{P}$. In particular, there exists a maximin prior π^* that belongs to $\bar{\mathcal{P}}$. We now prove that π^* can be taken to belong $\mathcal{P}$. If $\pi^* \in \mathcal{P}$, there is nothing to prove. Now take $\pi^* \in \bar{\mathcal{P}} \setminus \mathcal{P}$. Since the weak topology on $\mathcal{M}_1$ is metrizable under separability (see Theorem 11.3.3 in Dudley (2004)), there exists a sequence (π_n) in $\mathcal{P}$ such that μ_n converge weakly to π^*. Since r is lower semi-continuous in θ by assumption, we have $\liminf_{n\to\infty} B(\delta^*, \pi_n) \geq B(\delta^*, \pi^*)$ from the preliminary result. In particular, there is some $n \in \mathbb{N}$ such that $B(\delta^*, \pi_n) \geq B(\delta^*, \pi^*)$. Since (δ^*, π^*) is a saddle point, we also have $B(\delta^*, \pi_n) \leq B(\delta^*, \pi^*)$. Thus $B(\delta^*, \pi_n) = B(\delta^*, \pi^*)$ and $\pi_n \in \mathcal{P}$, which proves the claim. □

2.2 Comments on the assumptions of Proposition 2.1

Assumption 1. The lower semicontinuity in a of the loss function is typically verified in estimation problems. This includes, for instance, all problems with $\Theta \subseteq \mathbb{R}^n$, $\mathcal{A} \subseteq \bar{\mathbb{R}}^n$, and weighted l_p losses, $p \geq 1$, of the form

$$L(a, \theta) = \begin{cases} w(\|a - \theta\|_p) & \text{if } (a, \theta) \in \mathbb{R}^n \times \mathbb{R}^n \\ +\infty & \text{otherwise} \end{cases},$$

where $\|\cdot\|_p$ is the p-norm on $\mathbb{R}^n$ and $w\colon \mathbb{R} \to [0, +\infty)$ is a continuous convex increasing function.

Assumption 2. The weak upper semicontinuity of the Bayes risk holds under fairly general conditions. We consider two cases: 1. conditions under which the integrated risk is weakly continuous; 2. more general conditions under which the integrated risk may not be weakly continuous.

1. For the assumption to hold, it suffices that the integrated risk B is weakly continuous in π. Under separability of the parameter space Θ (implying metrizability of weak convergence), this holds in two common cases that we detail below.

U.1. The risk function r is continuous and bounded in θ, which directly implies continuity of B by definition of weak convergence. This happens regularly when the parameter space Θ and the action space $\mathcal{A}$ are compact. For instance, consider models with

(i) $\Theta = \mathcal{A} = [-m, m]$ for some $m > 0$,

(ii) $\{P_\theta : \theta \in \Theta\}$ a family of distributions with density functions continuous in θ,

(iii) any loss function L that is continuous in θ and a.

Then the risk r is continuous in θ as the double integral of a continuous integrable function (see Lemma 16.1 in Bauer (2011)), and so r is bounded on the compact set Ω.

U.2. The risk function r is continuous in θ but potentially unbounded and for each $\delta \in \mathcal{D}_0$, the random variables $r(\delta, \theta)$ are uniformly integrable over $\mathcal{P}$. Indeed, weak convergence and uniform integrability imply convergence of the first moment (see Theorem 3.2.8 in Durrett (2019)), that is,

$$\pi_n \xrightarrow[n\to\infty]{weak} \pi \quad \text{and} \quad \lim_{K\to\infty} \sup_{\pi\in\mathcal{P}} \int r(\delta,\theta)\mathbb{1}_{r(\delta,\theta)\geq K}\, d\pi(\theta) = 0$$

imply

$$\int r(\delta,\theta)\, d\pi_n(\theta) \xrightarrow[n\to\infty]{} \int r(\delta,\theta)\, d\pi(\theta),$$

which corresponds to weak continuity of the integrated risk. If $\Theta \subseteq \mathbb{R}$, this typically happens when θ is uniformly integrable over $\mathcal{P}$ and r satisfies for each $\delta \in \mathcal{D}_0$ a growth condition of the form $r(\theta, \delta) \leq C|\theta|$ for θ large enough.

2. The last conditions apply in a number of problems, but since weak semicontinuity of the Bayes risk is (much) weaker than weak continuity of the integrated risk, we suggest weaker conditions that cover a number of important cases.

U.3. If the integral sign and the infimum can be interchanged in the definition of the Bayes risk, then the conditions in (U.2.) need only be verified for the function $\theta \mapsto \inf_{\delta\in\mathcal{D}_0} r(\delta, \theta)$. In particular, for $\Theta \subseteq \mathbb{R}$, the growth conditions $\inf_{\delta\in\mathcal{D}_0} r(\delta, \theta) \leq C|\theta|$ is more easily verified. For instance, if $\mathcal{D}_0$ includes deterministic rules, then $\inf_{\delta\in\mathcal{D}_0} r(\delta, \theta) \leq \int L(x, \theta)\, dP_\theta(x)$, and for

some standard parametric families, $L(x,\theta)\,dP_\theta(x) \le C|\theta|$ for θ large enough. The interchange of integral and infimum is justified in a number of cases. In particular, if the infimum on both sides can be approximated by a sequence of measurable functions, then standard convergence theorems for the Lebesgue integral can be invoked. We leave it for future research to apply such approximation procedure for existing statistical problems.

U.4. If the l_p loss is used and the priors are assumed to have bounded p moments, an approximation argument can be directly worked out without invoking the interchange of integral and infimum. The idea can be traced back to Donoho and Johnstone (1994) and combines nicely with tightness results ensuring compactness. Suppose $\Theta \subseteq \mathbb{R}$, $\mathcal{A} \subseteq \bar{\mathbb{R}}$ is convex, L is the l_p loss for $p \ge 1$, and $\mathcal{D}_0 = \mathcal{D}$. Suppose the priors in $\mathcal{P}$ have bounded first p moments. Suppose that the measures P_θ, $\theta \in \Theta$, have bounded first p moments and admit densities continuous in θ. Denote $\mathcal{D}_f$ the non-randomized rules T that take values in $\mathbb{R}$ (excluding $\pm\infty$) and $\mathcal{D}_m = \{T \in \mathcal{D}_f : T(x) = x \text{ if } |x| \ge m\}$. Then each rule in $\mathcal{D}_m$ has bounded risk, and so $\pi \mapsto \inf_{T\in\mathcal{D}_m} B(T,\pi)$ is weakly upper semicontinuous as the infimum of a family of weakly continuous functions. The sets $\mathcal{D}_m$ are increasing as $m \to \infty$, and so $\inf_{T\in\mathcal{D}_m} B(T,\pi)$ decreases as $m \to \infty$, hence $\pi \mapsto \inf_{T\in\mathcal{D}_m} B(T,\pi)$ has a limit which we denote $B^*(\pi)$. If we can prove that

$$B^*(\pi) = \inf_{\delta\in\mathcal{D}} B(\delta,\pi),$$

then $\pi \mapsto \inf_{\delta\in\mathcal{D}} B(\delta,\pi)$ is the decreasing (pointwise) limit of a family of weakly upper semicontinuous functions and so is weakly upper semicontinuous. We show that the equality holds under the conditions specified above. Since $\inf_{\delta\in\mathcal{D}} B(\delta,\pi) \le \inf_{T\in\mathcal{D}_m} B(T,\pi)$, then $\inf_{\delta\in\mathcal{D}} B(T,\pi) \le B^*(\pi)$ by taking limits as $m \to \infty$. To show the reverse inequality, we restrict attention to the subset $\mathcal{D}_F$ of deterministic rules in $\mathcal{D}_f$ with finite integrated risk (since otherwise the inequality is trivially verified). For any such rule $T \in \mathcal{D}_F$, define $T_m \in \mathcal{D}_m$ by $T_m(x) = T(x)\mathbb{1}\{|x| \le m\} + x\mathbb{1}\{|x| > m\}$. Then $r(T_m,\theta) \to r(T,\theta)$ uniformly on compact sets as $m \to \infty$. To see this, note that for any compact subset $K \subseteq \Theta$,

$$\begin{aligned}
\sup_{\theta\in K} \left|r(T_m,\theta) - r(T,\theta)\right| &= \sup_{\theta\in K} \left| - \int_{|x|>m} |T(x)-\theta)|^p\, dP_\theta(x) \right. \\
&\qquad \left. + \int_{|x|>m} |x-\theta|^p\, dP_\theta(x) \right| \\
&\le \sup_{\theta\in K} \int_{|x|>m} |T(x)-\theta)|^p\, dP_\theta(x) \\
&\qquad + \sup_{\theta\in K} \int_{|x|>m} |x-\theta|^p\, dP_\theta(x).
\end{aligned}$$

Then, by applying Dini's theorem to $\int_{|x|>m} |T(x)-\theta)|^p\, dP_\theta(x)$ and to $\int_{|x|>m} |x-\theta|^p\, dP_\theta(x)$, which are monotonically decreasing sequences of continuous functions in θ (see Lemma 16.1 in Bauer (2011)) that converge pointwise to 0 (since T has finite risk and P_θ has bounded p

moments), we obtain that

$$\sup_{\theta \in K} \left| r(T_m, \theta) - r(T, \theta) \right| \xrightarrow[m \to \infty]{} 0.$$

It follows that $\int r(T_m, \theta)\, d\pi(\theta) \to \int r(T, \theta)\, d\pi(\theta)$ as $m \to \infty$. Then for any rule $T \in \mathcal{D}_F$, $\inf_m B(T_m, \pi) \leq B(T, \pi)$. Thus $\inf_{T \in \mathcal{D}_m} B(T, \pi) \leq \inf_{T \in \mathcal{D}_F} B(T, \pi) \leq \inf_{\delta \in \mathcal{D}} B(\delta, \pi)$, where the last inequality follows from (2.1.4) (since L is convex in a). Therefore, $B^*(\pi) = \inf_{\delta \in \mathcal{D}} B(\delta, \pi)$, and the conclusion follows.

Assumption 3. The weak compactness of $\mathcal{P}$ is typically verified under two conditions which we detail below. For solutions involving the compactification of the parameter space, see Section 3.3.

C.1. If the parameter space Θ is a compact metric space, then the Banach– Alaoglu theorem guarantees that $\mathcal{M}_1(\Theta)$ is weakly compact (since it is weakly closed in the unit ball of $\mathcal{C}(\Theta)^*$), and so any weakly closed set of probability measures is weakly compact.

C.2. If the space Θ is not compact, then we can still hope to invoke Prokhorov's theorem (which guarantees weak relative compactness for tight families of probability measures). Due to the ubiquity of applications calling for tightness, many results for standard families are available. A fundamental case for our purpose is when the measures in $\mathcal{P}$ have uniformly bounded moments. In this case, Markov's inequality implies tightness (see Lemma A.4), and so we obtain a weakly compact family of priors without having to assume compactness of the parameter space. Under the l_p loss, this tightening procedure also ensures that (U.4.) is satisfied, and so we obtain upper semicontinuity of the Bayesian risk at no extra cost. This tightening procedure, which directly yields existence results for maximin priors without compactness of the parameter space, was first considered in Feldman (1991) and extended in Donoho and Johnstone (1994). Additional comments on the applicability of this procedure are provided in Remark 2.3.

Assumption 4. The condition holds for $\mathcal{D}_0 = \mathcal{D}$ by construction of the vector space topology on $\mathcal{D}$ (see Lemma 1.1). The possibility to restrict $\mathcal{D}_0$ to strict subsets of $\mathcal{D}$ is of interest, but its applicability is limited in practice as we now illustrate. Indeed, it is natural to consider for $\mathcal{D}_0$ a subset of deterministic decision rules (e.g., linear rules, thresholding rules, etc.). However, such a set will not be generally convex since the convex combination of Dirac measures are not Dirac measures. If it is still possible to consider the closure of the convex hull of a set $\mathcal{D}_d$ of deterministic rules, we general face an intricate problem: either this set is hard to characterize or it is directly equal to $\mathcal{D}$. For instance, if $\mathcal{D}_d$ includes Dirac measures at $T_a(x) = a$ for all $a \in \mathcal{A}$, then the closure of the convex hull of $\mathcal{D}_d$ is equal to $\mathcal{M}_1(\mathcal{A})$ by Choquet's theorem. (This should not come as a surprise since randomized rules have been (artificially) introduced to "fill holes".) We leave it for future research to characterize the convex closure of smaller sets of deterministic rules and to derive related complete class theorems so as to obtain existence results for strict subsets $\mathcal{D}_0 \subset \mathcal{D}$.

Remark 2.3. From the discussion above, it appears that there are at least two practical ways to ensure the existence of maximin priors under the weak topology when the l_p loss is used: either to work

with a compact parameter space or to work with priors with bounded moments. The latter approach has been much less explored than the former in spite of being more general (since the compactness of the parameter space ensures that the priors have finite moments of any order). While the choice of bounds on the priors' moments may be less intuitive than the choice of a range for the parameter, we still advocate for bounds on the priors' moments due to the increased robustness at no extra cost. It is an interesting avenue for research to investigate and compare the effects of these different constraints on minimax estimation. Some explorations in this direction can be found in Feldman (1991) and in Section 13.3 in Johnstone (2019). An important aspect of the comparison is computational: we expect some non-trivial gains in some problems; e.g., Feldman (1991) exhibits cases in the normal mean problem where the moment constraints make the maximin prior directly normal so that they can be computed much more easily than the ones under compact parameter spaces.

2.3 Applications

2.3.1 Bounded normal mean estimation under l_p loss

Consider estimation of θ in the model $P_\theta = N(\theta, 1)$ under the l_p loss, $p \geq 1$, where it is assumed that:

(i) the parameter space is $\Theta = [-m, m]$;

(ii) the action space is $\mathcal{A} = [-m, m]$;

(iii) the set of decision rules $\mathcal{D}_0$ is the whole set $\mathcal{D}$;

(iv) the set of priors is $\mathcal{P} = \mathcal{M}_1([-m, m])$;

for some $m > 0$.

Proposition 2.4. *In the above model, a maximin prior that generates a saddle point exists.*

Proof. This is a standard application of (U.1.) and (C.1.). The l_p loss function is continuous in θ and a and the normal distribution $N(\theta, 1)$ has continuous density as a function of θ. It follows that the risk function r is continuous and bounded (see Lemma 16.1 in Bauer (2011)), and so the Bayesian risk is upper semicontinuous. Since the parameter space is compact, the set of priors is weakly compact. Moreover, $\mathcal{P}$ is trivially convex since $\mathcal{P} = \mathcal{M}_1(\Theta)$. Since $\mathcal{D}_0 = \mathcal{D}$, the set of decision rules is weakly closed and convex. Then Proposition 2.1 applies, and there exists a maximin prior that generates a saddle point. □

Reference. This is a classical result in the literature, available at least as early as Ghosh (1964). See, for instance, Section 4.6 in Johnstone (2019).

2.3.2 Unbounded normal mean estimation under l_p loss with moment condition

Consider estimation of θ in the model $P_\theta = N(\theta, 1)$ under the l_p loss, $p \geq 1$, where it is assumed that:

(i) the parameter space is $\Theta = \mathbb{R}$;

(ii) the action space is $\mathcal{A} = \mathbb{R} \cup \{\infty\}$;

(iii) the set of decision rules $\mathcal{D}_0$ is the whole set $\mathcal{D}$;

(iv) the set of priors is

$$\mathcal{P}^p = \{\pi \in \mathcal{M}_1(\mathbb{R}) : \int_{\mathbb{R}} |\theta|^p \, d\pi(\theta) \leq M\}$$

for some $M \in \mathbb{R}$.

Proposition 2.5. *In the above model, a maximin prior that generates a saddle point exists.*

Proof. This is a standard example of the tightening procedure through the moment constraint on the set of priors. We thus rely on (U.4.) and (C.2.). The l_p loss function is continuous in θ and a and the normal distribution $N(\theta, 1)$ has finite moments of all orders and a continuous density as a function of θ. The argument in (U.4.) based on the moment condition guarantees that the Bayesian risk is upper semicontinuous. The moment condition also ensures the tightness of $\mathcal{P}^p$ using Markov's inequality (see Lemma A.4). Moreover, the (weak) inequality for the moment condition ensures that $\mathcal{P}^p$ is weakly closed. Indeed, by a version of Fatou's lemma (see Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014)), if $\pi_n \to \pi$ weakly, then

$$\int_{\mathbb{R}} |\theta|^p \, d\pi(\theta) \leq \liminf_{n\to\infty} |\theta|^p \, d\pi_n(\theta) \leq M.$$

Hence, $\mathcal{P}^p$ is weakly compact. Moreover, it is easily seen by linearity of the space of measures that $\mathcal{P}^p$ is convex. Indeed, for any $a \in (0, 1)$, we have that

$$\int |\theta|^p \, d(a\pi_1 + (1-a)\pi_2) = a \int |\theta|^p \, d\pi_1 + (1-a) \int |\theta|^p \, d\pi_2 \leq aM + (1-a)M = M$$

for any $\pi_1, \pi_2 \in \mathcal{P}^p$. Since $\mathcal{D}_0 = \mathcal{D}$, the set of decision rules is weakly closed and convex. Then Proposition 2.1 applies, and there exists a maximin prior that generates a saddle point. □

Reference. This result was first stated without proof in Feldman (1991). The main elements of the proof can be found in Donoho and Johnstone (1994) p.285 and p.297.

2.3.3 Sparse unbounded normal mean estimation under l_p loss with moment condition

Consider estimation of θ in the model $P_\theta = N(\theta, 1)$ under the l_p loss, $p \geq 1$, where it is assumed that:

(i) the parameter space is $\Theta = \mathbb{R}$;

(ii) the action space is $\mathcal{A} = \mathbb{R} \cup \{\infty\}$;

(iii) the set of decision rules $\mathcal{D}_0$ is the whole set $\mathcal{D}$;

(iv) the set of priors is

$$\mathfrak{m}_0^p(\varepsilon) = \{\varepsilon\mathfrak{d}_0 + (1-\varepsilon)\pi : \pi \in \mathcal{P}^p\}$$

where $\mathfrak{d}_0$ is the Dirac measure at 0 and $\mathcal{P}^p = \{\pi \in \mathcal{M}_1(\mathbb{R}) : \int_{\mathbb{R}} |\theta|^p \, d\pi(\theta) \leq M\}$ for some $M \in \mathbb{R}$ and some $\varepsilon \in (0,1)$.

Proposition 2.6. *In the above model, a maximin prior that generates a saddle point exists.*

Proof. This is another example of the tightening argument. We can easily show that the priors in $\mathfrak{m}_0^p(\varepsilon)$ have uniformly bounded p moments by the linearity of the integral with respect to measures. Indeed, the Dirac measure at 0 has moments of any order equal to zero and $\mathcal{P}^p$ has uniformly bounded moments by definition. It then follows from (U.4.) that the Bayesian risk is weakly upper continuous. It remains to prove that $\mathfrak{m}_0^p(\varepsilon)$ is weakly compact. For this, it suffices to note that $\mathfrak{m}_0^p(\varepsilon)$ is the weighted Minkowski sum of two weakly compact convex sets, namely the singleton $\{\mathfrak{d}_0\}$ and the set $\mathcal{P}^p$ (which was proved to be weakly closed and convex in Section 2.3.2). Then Proposition 2.1 applies, and there exists a maximin prior that generates a saddle point. □

Reference. This result, which leverages the tightening argument previously worked out, is new. The problem emerges from minimax estimation under performance constraint at 0 (e.g., for sparsity reasons – see Johnstone (2019)). The result relaxes the standard compactness assumption on the parameter space (see Remark 2.3). We show in the next sections that other results in the literature that try to completely relax the boudedness assumptions either are incorrect (Johnstone (2019)) or reimpose strong boundedness conditions on the parameter space (Bickel (1983) and Bickel and Collins (1983)).

3 Existence of maximin priors under the vague topology

3.1 An impossibility result under the vague topology

The previous results show that the existence of maximin priors under the weak topology generally requires some form of tightening. To completely relax boundedness assumptions of the parameter or the priors and restore full robustness, a natural solution is to work with a coarser topology on the priors. The most natural candidate for this purpose is the vague topology, as first considered by Huber in Huber (1964) for the minimization of the Fisher information. Indeed, by embedding the priors (as a subset of $\mathcal{M}_1(\Theta)$) in the set of subprobability measures $\mathcal{M}_{\leq 1}(\Theta)$ endowed with the vague topology, relative compactness of the initial set of priors easily obtains (see Lemma A.1). While this approach has proved successful in the context considered by Huber, we show that it may not be the case in minimax games for the average risk of the form previously considered. To see this, note that to obtain general existence results for maximin priors using compactness arguments through the extreme value theorem applied to any given topology on the priors, we need at least that:

1. the set of priors is compact for this topology;
2. the Bayesian risk is upper semicontinuous with respect to this topology.

The main issue is that the properties (1.) and (2.) work in opposite direction: the coarser the topology on a set Θ, the more compact sets on Θ, but the fewer (upper semi)continuous functions (with initial space Θ). While it is always possible to gather high-level conditions for a general existence result under the vague topology (see Proposition B.1 where Θ is assumed locally compact so that the vague topology is well-defined), it often happens in practice that the two conditions cannot be satisfied jointly in cases where the weak topology does not already deliver a valid solution. Indeed, we commonly face the two following disjoint cases:

- either the set of priors $\mathcal{P} \subseteq \mathcal{M}_1(\Theta)$ is closed in the vague topology (in particular, there is no "escape of mass at infinity"), and hence weakly compact (see Lemma A.2), and so Proposition 2.1 can be applied with only weak upper semicontinuity of the Bayesian risk to be verified;
- or the set of priors $\mathcal{P}$ is not closed in the vaguely topology and its vague closure include elements in $\mathcal{M}_{<1}(\Theta)$ (that is, there is "escape of mass at infinity"), and then the Bayesian risk is often not vaguely upper semicontinuous on the vague closure $\bar{\mathcal{P}}$.

This makes Proposition B.1 either superfluous or inapplicable, despite its implicit use in the literature – see, e.g., Johnstone (2019). Due to the large degree of freedom of statistical decision problems, it is impractical to derive general impossibility results linking "escape of mass at infinity" and the failure of upper semicontinuity of the Bayesian risk. However, we are still able to illustrate the tension between compactness and semicontinuity in specific problems. In the next subsections, we focus on the normal mean problem and the sparse normal mean problem where "escape of mass at infinity" prevents the Bayesian risk from being vaguely upper semicontinuous. In the course of these examples, we show that Brown's equality does not hold for subprobability measures on the whole real line. This corrects a number of results available in Johnstone (2019).

3.2 Examples

3.2.1 Unbounded normal mean estimation under l_2 loss

Consider estimation of θ in the model $P_\theta = N(\theta, 1)$ under the l_2 loss where it is assumed that:

(i) the parameter space is $\Theta = \mathbb{R}$;

(ii) the action space is $\mathcal{A} = \mathbb{R} \cup \{\infty\}$;

(iii) the set of decision rules $\mathcal{D}_0$ is the whole set $\mathcal{D}$;

(iv) the set of priors is $\mathcal{M}_1(\mathbb{R})$.

The set $\mathcal{M}_1(\mathbb{R})$ is not weakly compact, so Proposition 2.1 cannot be applied directly. The set $\mathcal{M}_{\leq 1}(\mathbb{R})$, which is the vague closure of $\mathcal{M}_1(\mathbb{R})$, is, however, vaguely compact, but we show that the Bayesian risk is typically not vaguely upper semicontinuous on it. For this, we review a number of standard results for unbounded normal mean estimation and vague convergence.

Lemma 3.1. *Let $\mu_n = N(0,n)$ be the Gaussian distribution on $\mathbb{R}$ with mean 0 and variance $n \in \mathbb{N}$, μ_0 the zero measure on $\mathbb{R}$ (that is, $\nu_0(B) = 0$ for all $B \in \mathcal{B}(\mathbb{R})$), and λ the Lebesgue measure on $\mathbb{R}$. Then*

1. *$N(0,n)$ does not converge weakly as $n \to \infty$;*
2.

$$N(0,n) \xrightarrow[n\to\infty]{vague} \nu_0.$$

Proof. We start by proving (2.). Let $f \in C_c(\mathbb{R})$ with support A compact. Then $\|f\|_\infty < \infty$ and for all $x \in \mathbb{R}$, $|f(x)| \leq \|f\|_\infty \mathbb{1}_A(x)$. Thus

$$\begin{aligned}\left|\int_{\mathbb{R}} f\, d\mu_n\right| \leq \|f\|_\infty \int_A d\mu_n &= \frac{1}{\sqrt{2\pi n}}\|f\|_\infty \int_A \exp\left(-\frac{x^2}{2n}\right) dx \\ &\leq \frac{1}{\sqrt{2\pi n}}\|f\|_\infty \int_A dx \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

We now prove (1.). Suppose by contradiction that $\mu_n \xrightarrow{weakly} \mu$. Since $1 \in C_b(\mathbb{R})$, then $\mu(\mathbb{R}) = 1$. Since $\mathbb{1}_A \in C_c(\mathbb{R})$ for any compact subset A, we must have from (2.) and the Portmanteau theorem that $\mu(A) = 0$. Since $\mathbb{R}$ can be written as the countable union of compact intervals, it follows from the union bound that $\mu(\mathbb{R}) = 0$: a contradiction. □

Lemma 3.2. *In the model of this subsection, we have that:*

1. *the Bayes rule with respect to $N(0,k)$ is given by the deterministic rule*

$$d_{N(0,k)}(x) = \mathbb{E}\left[\theta | X = x\right] = \frac{1}{1/k^2 + 1} x,$$

with integrated risk

$$B_k = \mathbb{E}_X[\mathrm{Var}(\theta|X)] = \frac{1}{1/k^2+1} \xrightarrow[k\to\infty]{} 1;$$

2. *the deterministic rule $T(x) = x$ is minimax with minimax risk $R_N(\Theta) = 1$;*
3. *$(N(0,k))_{k\in\mathbb{N}}$ is a sequence of maximin priors in the sense that*

$$\lim_{k\to\infty} \inf_{\delta\in\mathcal{D}} B(\delta, N(0,k)) = 1 \geq \inf_{\delta\in\mathcal{D}} B(\delta,\pi)$$

for all $\pi \in \mathcal{P}$, but there is no maximin prior in the weak topology.

Proof. This is a common result in the literature. See, for instance, Section 1.8 p.48 and Section 2.11 p.94 in Ferguson (1967), or Example 4.2.2, Example 4.2.4, and Example 5.1.14 in Lehmann and Casella (1998) pp.233-235 and p.317. □

Proposition 3.3. *In the model of this subsection, we have that:*

1. *the sequence of maximin priors $(N(0,k))_{k\in\mathbb{N}}$ does not converge vaguely to a maximin prior in the vague topology (if it exists);*
2. *the Bayes risk function $\pi \mapsto \inf_{\delta\in\mathcal{D}} B(\delta,\pi)$ is not vaguely upper semicontinuous on $\mathcal{M}_{\leq 1}(\mathbb{R})$.*

Proof. By Lemma 3.1, the sequence $(N(0,k))_{k\in\mathbb{N}}$ converges vaguely to the zero measure ν_0 whose Bayes risk is given by

$$\inf_{\delta\in\mathcal{D}} B(\delta,\nu_0) = 0 < 1 = \lim_{k\to\infty} \inf_{\delta\in\mathcal{D}} B(\delta, N(0,k)).$$

This concludes the proof for (1.) and (2.). □

This result does not prove that there does not exist a maximin prior in the vague topology, only that standard arguments to prove existence based on compactness and semicontinuity do not apply in this case. Nonetheless, this still implies, as shown below, that the following equality

$$\underline{B}(\pi) = 1 - I(\pi * \phi),$$

where I is the Fisher information, $*$ denotes convolution, and ϕ is the standard normal density, does not hold for subprobability measures on the whole real line. This equality has been referred to as Brown's identity in the literature (see Proposition 4.5 in Johnstone (2019)).

Corollary 3.4. *Brown's identity does not hold on* $\mathcal{M}_{\leq 1}(\mathbb{R})$.

Proof. Brown's identity implies that the Bayes risk function is vaguely upper semicontinuous as (a constant plus) the Fisher information, which is vaguely lower semicontinuous by definition: a contradiction with Proposition 3.3. □

3.2.2 Sparse unbounded normal mean estimation under l_2 loss

Consider estimation of θ in the model $P_\theta = N(\theta, 1)$ under the l_2 loss where it is assumed that:

(i) the parameter space is $\Theta = \mathbb{R}$;

(ii) the action space is $\mathcal{A} = \mathbb{R} \cup \{\infty\}$;

(iii) the set of decision rules $\mathcal{D}_0$ is the whole set $\mathcal{D}$;

(iv) the set of priors is

$$\mathfrak{m}_0(\varepsilon) = \{\varepsilon\mathfrak{d}_0 + (1-\varepsilon)\mu : \mu \in \mathcal{M}_1(\mathbb{R})\}$$

where $\varepsilon \in (0,1)$ and $\mathfrak{d}_0$ is the Dirac measure at 0.

The set of priors $\mathfrak{m}_0(\varepsilon)$ is not weakly compact, so Proposition 2.1 cannot be applied. However, we consider $\mathfrak{m}_0(\varepsilon)$ as a subset of the set of subprobability measures endowed with the vague topology and take its vague closure, that is, the set $\overline{\mathfrak{m}}_0(\varepsilon) = \{\varepsilon\mathfrak{d}_0 + (1-\varepsilon)\mu : \mu \in \mathcal{M}_{\leq 1}(\mathbb{R})\}$, which is then a compact subset of $\mathcal{M}_{\leq 1}(\mathbb{R})$. We show that the Bayesian risk function for this problem is not vaguely upper semicontinuous on $\overline{\mathfrak{m}}_0(\varepsilon)$, so that the standard arguments based on compactness and semicontinuity cannot be used to prove existence. The implications of this result for sparse unbounded normal mean estimation are further explored in next section.

Proposition 3.5. *Consider the problem defined above for some $\varepsilon \in (0,1)$. Then the Bayesian risk function $\pi \mapsto \inf_{\delta\in\mathcal{D}} B(\delta,\pi)$ is not vaguely upper semicontinuous on $\overline{\mathfrak{m}}_0(\varepsilon)$.*

Proof. Take $\varepsilon \in (0,1)$ and $\mu_k = N(0,k)$. Then $\varepsilon\mathfrak{d}_0 + (1-\varepsilon)\mu_k \in \overline{\mathfrak{m}}_0(\varepsilon)$ for all $k \in \mathbb{N}$. By Lemma 3.1, we have

$$\varepsilon\mathfrak{d}_0 + (1-\varepsilon)\mu_k \xrightarrow[n\to\infty]{vague} \varepsilon\mathfrak{d}_0 + (1-\varepsilon)\nu_0$$

where ν_0 is the measure zero. We have that

$$\inf_{\delta\in\mathcal{D}} B(\delta, \varepsilon\mathfrak{d}_0 + (1-\varepsilon)\nu_0) = \varepsilon \inf_{\delta\in\mathcal{D}} B(\delta,\mathfrak{d}_0) = 0,$$

and that

$$\begin{aligned}\inf_{\delta\in\mathcal{D}} B(\delta, \varepsilon\mathfrak{d}_0 + (1-\varepsilon)N(0,k)) &\geq \varepsilon \inf_{\delta\in\mathcal{D}} B(\delta,\mathfrak{d}_0) + (1-\varepsilon)\inf_{\delta\in\mathcal{D}} B(\delta, N(0,k)) \\ &= 1-\varepsilon.\end{aligned}$$

Therefore,

$$\inf_{\delta\in\mathcal{D}} B(\delta, \varepsilon\mathfrak{d}_0 + (1-\varepsilon)\nu_0) = 0 < 1-\varepsilon = \lim_{k\to\infty} \inf_{\delta\in\mathcal{D}} B(\delta, \varepsilon\mathfrak{d}_0 + (1-\varepsilon)N(0,k)),$$

which proves that the Bayesian risk is not vaguely upper semicontinuous. □

3.3 A remark on the compactification of the parameter space

The approach considered above is exactly the one developed by Huber in Huber (1964) for the minimization of the Fisher information and redeveloped in Johnstone (2019) for the average risk (with a few important errors). The approach has a nice interpretation since $\mathcal{M}_{\leq 1}(\mathbb{R})$ endowed with the vague topology exactly corresponds to $\mathcal{M}_1(\mathbb{R}\cup\{\infty\})$ endowed with the weak topology where $\mathbb{R}\cup\{\infty\}$ is the one-point compactification of $\mathbb{R}$ (see Proposition 3.6). This suggests an alternative approach to relax boundedness assumptions based on using another compactification of the real line while sticking to the weak topology. This is exactly what is done in Bickel (1983) and Bickel and Collins (1983) for the sparse normal mean problem where the two-point compactification of the real line $\mathbb{R}\cup\{-\infty,\infty\} = [-\infty,+\infty]$ is used.

We first show that this construction is not equivalent to the one considered by Huber in Huber (1964) (contrarily to what has been implicitly suggested in Bickel (1983) and Bickel and Collins (1983) and explictly claimed in Johnstone (2019)). More importantly, we show that this construction cannot be considered as a solution to the problem of relaxing boundedness assumptions in existence results since $\mathcal{M}_1([-\infty,+\infty])$ is seen to correspond exactly to $\mathcal{M}_1([-\pi/2,\pi/2])$, both endowed with the weak topology. In other words, working with probability measures on the two-point compactification of the real line $[-\infty,+\infty]$ as done in Bickel (1983) and Bickel and Collins (1983) is equivalent to working with probability measures on $[-\pi/2,\pi/2]$ (and not subprobability measures on $\mathbb{R}$), hence imposing strong boundedness conditions on the parameter. Moreover, the bounds $-\pi/2$ and $\pi/2$

are arbitrary and depend only on the initial choice of the homeomorphism to define $[-\infty, +\infty]$: we have chosen a standard normalization, but any rescaling $-z\pi/2$ and $z\pi/2$ for $z > 0$ could have been considered. In other words, existence results when the parameter space is $[-\infty, \infty]$ are exactly existence results for the parameter space $[-z\pi/2, z\pi/2]$ for some $z > 0$ which depends on the choice of the homeomorphism by the modeler (and is often left unstated). For this reason, we strongly advocate against using this approach in practice. A more reasonable solution is to explicitly work with a parameter space $[-m, m]$ for some $m > 0$. A more general approach (if the l_p norm is used) is to simply assume that the moments of the priors are bounded by some explicit constant $M \in \mathbb{R}$ as recommended in Section 2.2. Given that the results of the last section suggest that Huber's approach does not transpose well to minimax games for the average risk, the option to bound the priors' moments is likely the most general to deliver existence results.

Proposition 3.6. *The set $\mathcal{M}_1(\mathbb{R} \cup \{\infty\})$ of probability measures on $\mathbb{R} \cup \{\infty\}$ endowed with the weak topology is homeomorphic to the set $\mathcal{M}_{\leq 1}(\mathbb{R})$ of subprobability measures on $\mathbb{R}$ endowed with the vague topology.*

Proof. We first prove that the function $\psi : \mathcal{M}_1(\mathbb{R} \cup \{\infty\}) \to \mathcal{M}_{\leq 1}(\mathbb{R})$ given by $\psi(\mu^\infty) = \mu$ where $\mu(A) := \mu^\infty(A)$ for all $A \in \mathcal{B}(\mathbb{R})$ is a continuous bijection. Consider the function ψ^{-1} given by $\psi^{-1}(\mu) = \mu_\infty$ where $\mu_\infty(A) := \mu(A)$ for all $A \in \mathcal{B}(\mathbb{R})$ and $\mu_\infty(\{\infty\}) = 1 - \mu(E)$. It is directly seen that ψ^{-1} is an inverse function for ψ. Suppose now that $\mu_n^\infty \to \mu^\infty$ weakly. By Proposition 4.36 in Folland (1999) (since $C_c(\mathbb{R}) \subseteq C_0(\mathbb{R}) \subseteq C(\mathbb{R})$), any compactly supported continuous function f on $\mathbb{R}$ extends continuously to a continuous function g on $\mathbb{R} \cup \{\infty\}$ such that $f = g$ on $\mathbb{R}$ and $g(\infty) = 0$. Hence for any such f,

$$\begin{aligned}\int_{\mathbb{R}} f \, d\psi(\mu_n^\infty) = \int_{\mathbb{R}} g \, d\mu_n + g(\infty) = \int_{\mathbb{R}\cup\{\infty\}} g \, d\mu_n^\infty \\ \xrightarrow[n\to\infty]{} \int_{\mathbb{R}\cup\{\infty\}} g \, d\mu^\infty = \int_{\mathbb{R}} g \, d\mu = \int_{\mathbb{R}} f \, d\psi(\mu^\infty).\end{aligned}$$

This proves that ψ is continuous. Since $\mathcal{M}_1(\mathbb{R} \cup \{\infty\})$ is compact and $\mathcal{M}_{\leq 1}(\mathbb{R})$ is Hausdorff (see p.192 in Bauer (2011)), it follows that ψ is a homeomorphism (see Proposition 4.28 in Folland (1999)). □

Proposition 3.7. *The set $\mathcal{M}_1(\mathbb{R} \cup \{-\infty, +\infty\})$ of probability measures on $\mathbb{R} \cup \{-\infty, +\infty\}$ endowed with the weak topology is homeomorphic to the set $\mathcal{M}_1([-\pi/2, \pi/2])$ of probability measures on $[-\pi/2, \pi/2]$ endowed with the weak topology.*

Proof. We first prove that the function $\psi : \mathcal{M}_1(\mathbb{R} \cup \{-\infty, +\infty\}) \to \mathcal{M}_1([-\pi/2, \pi/2])$ given by $\psi(\mu^\infty) = \mu := \arctan_*(\mu^\infty)$, where $\arctan_*(\mu^\infty)$ is the pushforward measure of μ for $\arctan : \mathbb{R} \cup \{-\infty, +\infty\} \to [-\pi/2, \pi/2]$, is a continuous bijection. Consider the function ψ^{-1} given by $\psi^{-1}(\mu) = \mu_\infty := \tan_*(\mu)$. It is directly seen that ψ^{-1} is an inverse function for ψ. To see this, note that $\psi^{-1}(\psi(\mu^\infty))(A) = \mu^\infty(\tan(\arctan(A))) = A$ for all $A \in \mathcal{B}(\mathbb{R} \cup \{-\infty, +\infty\})$ and $\psi(\psi^{-1}(\mu))(B) = \mu(\arctan(\tan(B))) = \mu(B)$ for all $B \in \mathcal{B}([-\pi/2, \pi/2])$. Suppose now that $\mu_n^\infty \to \mu^\infty$ weakly. We

have for any continuous function f that

$$\int_{[-\pi/2,\pi/2]} f \, d\psi(\mu_n^\infty) = \int_{\mathbb{R}\cup\{-\infty,\infty\}} f \circ \arctan \, d\mu_n^\infty$$
$$\xrightarrow[n\to\infty]{} \int_{\mathbb{R}\cup\{-\infty,\infty\}} f \circ \arctan \, d\mu^\infty = \int_{[-\pi/2,\pi/2]} f \, d\psi(\mu^\infty),$$

where the first and last equalities follow from the definition of the pushforward measure (under integrability) and the limit follows from the weak convergence of μ_n^∞ and the continuity of $f \circ \arctan$. Since $\mathcal{M}_1(\mathbb{R} \cup \{-\infty, \infty\})$ is compact and $\mathcal{M}_1([-\pi/2, \pi/2])$ is Hausdorff, it follows that ψ is a homeomorphism (see Proposition 4.28 in Folland (1999)). □

Remark 3.1. We believe that the compactification used in Bickel (1983) and Bickel and Collins (1983) can be traced back to an heuristic comment in Huber (1964) and Huber (1981). While Huber does not make use of this compactification in his proof (but directly works with subprobability measures on $\mathbb{R}$), this is not the case for Bickel (1983) and Bickel and Collins (1983) where the additional topological properties of the two-point compactification are put to work in the use of Brown's identity. Moreover, it follows from Proposition 3.7 that, contrarily to what has been implictly or explicitly claimed without proof in Bickel and Collins (1983) and Johnstone (2019) (see, for instance, p.440 in Johnstone (2019)), the two-point compactification of $\mathbb{R}$ does not make $\mathcal{M}_1(\mathbb{R} \cup \{-\infty, +\infty\})$ endowed with the weak topology homeomorphic to $\mathcal{M}_{\leq 1}(\mathbb{R})$ endowed with the vague topology.

3.4 A remark on the definition of the Bayes risk on $\mathcal{M}_{\leq 1}(\Theta)$

The average risk and the Bayes risk are traditionally defined on $\mathcal{M}_1(\Theta)$ but not on $\mathcal{M}_{\leq 1}(\Theta)$. There is some arbitrariness in the extension of these notions to $\mathcal{M}_{\leq 1}(\Theta)$. However, it is difficult (from both a mathematical perspective and a statistical perspective) not to take the natural extension of these notions by simply integrating over a finite measure. In our developments, we have always taken these natural extensions as their definitions.

Nonetheless, one may naturally wonder if it is possible to redefine the Bayes risk for strict subprobability measures so as to get upper semicontinuity. The simplest solution is to (upper semi)continuously extend the Bayes risk on the space $\mathcal{M}_{<1}(\Theta)$. For instance, define for any $\pi \in \mathcal{P} \cap \mathcal{M}_{<1}(\Theta)$,

$$\underline{B}(\pi) := \max\left\{ \lim_{n\to\infty} \underline{B}(\mu_n) : \mu_n \in V_\pi \right\}$$

where V_π is the set of sequences of probability measures that vaguely converge to π. This naturally guarantees upper semicontinuity of the Bayes risk. While this fix ensures the existence of a maximin prior, this prior will typically be a pure subprobability prior (i.e., an improper prior in the terminology of Johnstone (2019)). We exemplify this for the case of unbounded normal mean estimation.

Proposition 3.8. *Consider the extension of the Bayes risk as defined above for normal mean estimation with mean zero normal priors. Then a maximin prior exists, but no proper maximin prior exists.*

Proof. The conditions of Proposition B.1 are satisfied, so a maximin prior exists. By Lemma 3.2, we know that for any proper prior $N(0,k)$, $\underline{B}(N(0,k)) < 1$, while $\underline{B}(\nu_0) = \lim_{k\to\infty} \underline{B}(N(0,k)) = 1$. It follows that ν_0 is the unique maximin prior for this problem. □

There obviously exist different ways to upper semicontinuously extend the Bayes risk on $\mathcal{M}_{<1}(\Theta)$, but all such extensions rely on a similar idea which stands at the basis of the pathology. We conjecture that in most cases redefining the Bayes risk by upper semicontinuous extension does not yield existence of maximin priors that are proper probability measures. This is not an issue if one is more generally interested in the existence of improper maximin priors (that is, pure subprobability measures that achieve the minimax equality). If so, we conjecture that working with the q-vague topology of Bioche and Druilhet (2016) is another solution to this problem. We do not tackle this question here and leave the existence of improper maximin priors to future research.

4 Conclusion

In this paper, we contributed to statistical theory in two ways. We first extended the statistical minimax theorems of Wald and Le Cam by providing general applicable conditions under which maximin priors exist and are saddle points. We showed that these conditions not only obtained under compactness of the parameter space but also under the weaker condition that the priors have bounded moments, following a procedure first considered in Feldman (1991) and extended in Donoho and Johnstone (1994). We then exhibited the inherent difficulty of relaxing these boundedness conditions on the parameter space or the priors's moments to increase the robustness of the resulting minimax procedures. We first showed that Huber's approach, based on embedding the priors in the set of subprobability measures with the vague topology, does not transpose well in minimax games for the average risk. We illustrated this issue in the normal mean problem and the sparse normal mean problem when the parameter space is the whole real line where we found that the vague upper semicontinuity of the Bayes risk could not be saved. In the course of illustrating this issue, we corrected a number of results available in Johnstone (2019) and obtained as a result of independent interest that Brown's identity does not hold for subprobability measures on the whole real line. We then showed that an alternative approach, considered in Bickel (1983) and Bickel and Collins (1983) and based on taking the extended real line $[-\infty, +\infty]$ as parameter space, imposed much stronger boundedness conditions on the parameter than what the construction could suggest. In particular, we showed that this compactification was not equivalent to the one considered by Huber (hence correcting a number of claims in the literature) but amounted to exactly working with priors on $[-z\pi/2, z\pi/2]$ where $z > 0$ only depends on the arbitrary choice by the modeler of the homeomorphism to define $[-\infty, +\infty]$. While our results do not exhaust the problem of existence of maximin priors in minimax games for the average risk, they are still indicative of fundamental limits one faces when trying to gain robustness against any potential move of Nature. We believe that our results and the discussions that follow should be helpful for practitioners when deciding between different modeling assumptions for

the associated minimax games. In particular, we call for caution when working with maximin priors for the average risk without explicit bounds on the parameter or on the priors' moments.

References

Bauer, H. (2011): *Measure and integration theory*, vol. 26. Walter de Gruyter.

Berger, J. O. (1985): "Statistical decision theory and Bayesian analysis," *Springer Series in Statistics*.

Bickel, P. J. (1981): "Minimax estimation of the mean of a normal distribution when the parameter space is restricted," *The Annals of Statistics*, 9(6), 1301–1309.

——— (1983): "Minimax estimation of the mean of a normal distribution subject to doing well at a point," in *Recent Advances in Statistics*, pp. 511–528. Elsevier.

Bickel, P. J., and J. R. Collins (1983): "Minimizing Fisher information over mixtures of distributions," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 1–19.

Bioche, C., and P. Druilhet (2016): "Approximation of improper priors," *Bernoulli*, 22(3), 1709–1728.

Blanchet, J., and K. Murthy (2019): "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, 44(2), 565–600.

Brown, L. D. (1974): "Notes on Decision Theory," *Unpublished lecture notes*.

Casella, G., and W. E. Strawderman (1981): "Estimating a bounded normal mean," *The Annals of Statistics*, 9(4), 870–878.

Chamberlain, G. (2000): "Econometric applications of maxmin expected utility," *Journal of Applied Econometrics*, 15(6), 625–644.

Donoho, D. L., and I. M. Johnstone (1994): "Minimax risk over lp-balls for lp-error," *Probability theory and related fields*, 99, 277–303.

——— (1996): "Neo-classical minimax problems, thresholding and adaptive function estimation," *Bernoulli*, 2(1), 39–62.

——— (1998): "Minimax estimation via wavelet shrinkage," *The annals of Statistics*, 26(3), 879–921.

Donoho, D. L., R. C. Liu, and B. MacGibbon (1990): "Minimax risk over hyperrectangles, and implications," *The Annals of Statistics*, pp. 1416–1437.

Dudley, R. M. (2004): *Real analysis and probability*. Cambridge University Press.

Durrett, R. (2019): *Probability: theory and examples*, vol. 49. Cambridge university press.

Eichenauer, J., and J. Lehn (1989): "Computation of gamma-minimax estimators for a bounded normal mean under squared error loss," *Statistics & Risk Modeling*, 7(1-2), 37–62.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 83(2), 771–811.

Farrell, R. (1966): "Weak limits of sequences of Bayes procedures in estimation theory," in *Proc. Fifth Berkeley Symp. Math. Statist. Prob*, vol. 1, pp. 83–111.

Feinberg, E. A., P. O. Kasyanov, and N. V. Zadoianchuk (2014): "Fatou's lemma for weakly converging probabilities," *Theory of Probability & Its Applications*, 58(4), 683–689.

Feldman, I. (1991): "Constrained minimax estimation of the mean of the normal distribution with known variance," *The Annals of Statistics*, 19(4), 2259–2265.

Ferguson, T. S. (1967): *Mathematical statistics: a decision theoretic approach*. Academic Press, New York.

Folland, G. B. (1999): *Real analysis: modern techniques and their applications*, vol. 40. John Wiley & Sons.

Fréchet, M. (1953): "Commentary on the three notes of Emile Borel," *Econometrica*, pp. 118–124.

Gao, R., and A. Kleywegt (2023): "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, 48(2), 603–655.

Ghosh, M. (1964): "Uniform approximation of minimax point estimates," *The Annals of Mathematical Statistics*, pp. 1031–1047.

Gourdin, E., B. Jaumard, and B. MacGibbon (1994): "Global optimization decomposition methods for bounded parameter minimax risk evaluation," *SIAM Journal on Scientific Computing*, 15(1), 16–35.

Hodges, J., and E. Lehmann (1950): "Some Problems in Minimax Point Estimation," *The Annals of Mathematical Statistics*, 21(2), 182–197.

Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, pp. 73–101.

——— (1981): *Robust Statistics*. John Wiley & Sons.

Johnstone, I. M. (1994): "On minimax estimation of a sparse normal mean vector," *The Annals of Statistics*, pp. 271–289.

——— (2019): "Gaussian estimation: Sequence and wavelet models," *Unpublished lecture notes*.

Kempthorne, P. J. (1987): "Numerical specification of discrete least favorable prior distributions," *SIAM Journal on Scientific and Statistical Computing*, 8(2), 171–184.

Kline, P., and C. Walters (2021): "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 89(2), 765–792.

Kuhn, H. (1953): "Review of Kneser (1952)'," *Mathematical Reviews*, 14, 301–392.

Le Cam, L. (1955): "An extension of Wald's theory of statistical decision functions," *The Annals of Mathematical Statistics*, 26(1), 69–81.

——— (1986): *Asymptotic methods in statistical decision theory*. Springer.

Lehmann, E. (1952): "On the existence of least favorable distributions," *The Annals of Mathematical Statistics*, pp. 408–416.

Lehmann, E. L., and G. Casella (1998): *Theory of point estimation*. Springer.

Levit, B. Y. (1981): "On asymptotic minimax estimates of the second order," *Theory of Probability & Its Applications*, 25(3), 552–568.

Marchand, E., and W. E. Strawderman (2004): "Estimation in restricted parameter space: A review," *A Festshrift for Herman Rubin*, pp. 21–44.

Mohajerin Esfahani, P., and D. Kuhn (2018): "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, 171(1), 115–166.

Müller, U. K., and Y. Wang (2019): "Nearly weighted risk minimal unbiased estimation," *Journal of Econometrics*, 209(1), 18–34.

Nelson, W. (1966): "Minimax solution of statistical decision problems by iteration," *The Annals of Mathematical Statistics*, 37(6), 1643–1657.

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

Shafieezadeh Abadeh, S., V. A. Nguyen, D. Kuhn, and P. M. Mohajerin Esfahani (2018): "Wasserstein distributionally robust Kalman filtering," *Advances in Neural Information Processing Systems*, 31.

Strasser, H. (1985): *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*, vol. 7. Walter de Gruyter.

Wald, A. (1950): *Statistical Decision Functions*. Wiley.

Zinzius, E. (1981): "Minimaxschätzer für den mittelwert ϑ einer normalverteilten zufallsgröβe mit bekannter varianz bei vorgegebener oberer und unterer schranke fiir ϑ," *Statistics: A Journal of Theoretical and Applied Statistics*, 12(4), 551–557.

Appendices

Appendix A Additional notations, definitions, and useful theorems

We endow $\mathbb{R}$ with its standard metric $d(x, y) = |x - y|$, under which it is a locally compact separable complete metric space. The space $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ is the one-point compactification of $\mathbb{R}$, which is a compact metrizable space. The space $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\} = [-\infty, +\infty]$ is the two-point compactification of $\mathbb{R}$, which is endowed with the metric $d(x, y) = |A(x) - A(y)|$ where $A(x) = \arctan x$, under which it a compact separable complete metric space with Borel σ-algebra $\mathcal{B}(\overline{\mathbb{R}}) = \{E \subseteq \overline{\mathbb{R}} : E \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})\}$. We endow $\mathbb{R}$, $\mathbb{R}^*$, and $\overline{\mathbb{R}}$ with their respective Borel σ-algebra for the mentioned topology. See p.45 and p.132 in Folland (1999) for additional properties.

Given a metric space E, we define $C(E) = \{f : E \to \mathbb{R} : f \text{ is continuous}\}$, $C_b(E) = \{f \in C(E) : f \text{ is bounded}\}$, $C_c(E) = \{f \in C(E) : f \text{ has compact support}\}$, $C_0(E) = \{f \in C(E) : f \text{ vanishes at infinity}\}$ where the support of a function f is defined as $\overline{f^{-1}(\mathbb{R} \setminus \{0\})} = \overline{\{x : f(x) \neq 0\}}$ and f is said to vanish at infinity if for all $\varepsilon > 0$, the set $\{x : |f(x)| \geq \varepsilon\}$ is compact. We naturally have $C_c(E) \subseteq C_0(E) \subseteq C_b(E) \subseteq C(E)$. The vector spaces $C_b(E)$ and $C_c(E)$ are endowed with the supremum norm $\|f\|_\infty = \sup\{|f(x)| : x \in E\}$. See Chapter 4 in Folland (1999) for reference.

There exist some divergence among authors when it comes to defining Radon measures and related notions. For clarity, we briefly review the definitions and notations used in the paper. Given a metric space E endowed with its Borel σ-algebra $\mathcal{E}$, we define $\mathcal{M}(E)$ to be the set of all non-negative σ-finite measures on $(E, \mathcal{E})$ and $\mathcal{M}_r(E)$ the set of all non-negative σ-finite measures on $(E, \mathcal{E})$ that are inner regular and locally finite. The set $\mathcal{M}_r(E)$ is known as the set of Radon (non-negative) measures on E. We then respectively define the sets of all finite measures on $(E, \mathcal{E})$, the set of all subprobability measures on $(E, \mathcal{E})$, the set of all pure subprobability measures on $(E, \mathcal{E})$, and the set of all probability measures on $(E, \mathcal{E})$ by $\mathcal{M}_f(E) = \{\mu \in \mathcal{M}(E) : \mu(E) < \infty\}$, $\mathcal{M}_{\leq 1}(E) = \{\mu \in \mathcal{M}(E) : \mu(E) \leq 1\}$, $\mathcal{M}_{<1}(E) = \{\mu \in \mathcal{M}(E) : \mu(E) < 1\}$, and $\mathcal{M}_1(E) = \{\mu \in \mathcal{M}(E) : \mu(E) = 1\}$. It can be proved that: if E is Polish, then $\mathcal{M}_f(E) \subseteq \mathcal{M}_r(E)$; if E is locally compact and separable, then $\mathcal{M}_f(E) \subseteq \mathcal{M}_r(E)$.

Definition 2 (Weak Convergence)**.** Let E be a metric space. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of measures in $\mathcal{M}_f(E)$ and $\mu \in \mathcal{M}_f(E)$. The sequence $(\mu_n)_{n\in\mathbb{N}}$ is said to converge weakly to μ if

$$\int f \, d\mu_n \xrightarrow[n\to\infty]{} \int f \, d\mu \quad \text{for all } f \in C_b(E).$$

Definition 3 (Vague Convergence)**.** Let E be a locally compact metric space. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of measures in $\mathcal{M}_r(E)$ and $\mu \in \mathcal{M}_r(E)$. The sequence $(\mu_n)_{n\in\mathbb{N}}$ is said to converge vaguely to μ if

$$\int f \, d\mu_n \xrightarrow[n\to\infty]{} \int f \, d\mu \quad \text{for all } f \in C_c(E).$$

Lemma A.1. *If E is a locally compact separable metric space, then $\mathcal{M}_{\leq 1}(E)$ is vaguely compact.*

Proof. Under separablity, finite measures are Radon. Then apply Corollary 31.3. in Bauer (2011) (p.206) to guarantee vague sequential compactness and Theorem 31.5 in Bauer (2011) (p.208) to ensure metrizability of vague convergence. □

Lemma A.2. *Let Θ be a locally compact separable metric space. If $\mathcal{P} \subseteq \mathcal{M}_1(\Theta)$ is closed in the vague topology, then $\mathcal{P}$ is weakly compact.*

Proof. Since $\mathcal{P}$ is vaguely closed, it is vaguely compact by Lemma A.1 and metrizability of vague convergence. Let (μ_n) be any sequence in $\mathcal{P}$. Then, by sequential compactness, (μ_n) has a vaguely convergent subsequence in $\mathcal{P}$. Then the conclusion follows from Theorem 3.8 in Bauer (2011) (p.196) under separability of Θ (which also guarantees the metrizability of weak convergence). □

Lemma A.3. *Let Θ be a metric space and $(X, \mathcal{A}, \mu)$ an arbitrary measure space. If $f : X \times \Theta \to \overline{\mathbb{R}}$ is a nonnegative function such that:*

1. *for all $\theta \in \Theta$, $x \mapsto f(x, \theta)$ is $\mathcal{A}$-measurable;*
2. *for all $x \in X$, $\theta \mapsto f(x, \theta)$ is lower semicontinuous;*

then the function defined on Θ by $\theta \mapsto \int f(x, \theta)\, d\mu(x)$ is lower semicontinuous.

Proof. This follows directly from Fatou's lemma. Indeed, if $\theta_n \to \theta_0$, then for all $x \in X$,

$$\int f(\theta_0, x)\, d\mu(x) \leq \int \liminf_{n\to\infty} f(\theta_n, x) \leq \liminf_{n\to\infty} \int f(\theta_n, x),$$

where the first inequality follows from lower semicontinuity of $f(x, \cdot)$ and monotonicity of the integral, and the second from Fatou's lemma. □

Lemma A.4. *Let $p \geq 1$ ans $\{\mu_i\} \subseteq \mathcal{M}_1(\mathbb{R})$ a family of probability measures for which there is some $M \in \mathbb{R}$ such that $\int_{\mathbb{R}} |x|^p\, d\mu_i(x) \leq M$ for all i. Then $\{\mu_i\}$ is tight.*

Proof. Since $|x| \leq 1 + |x|^p$, we have $\int_{\mathbb{R}} |x|\, d\mu_i(x) \leq 1 + M$ by monotonicity of the integral. By Markov's inequality, we have for any $K > 0$,

$$\mu_i(\mathbb{R} \setminus [-K, K]) \leq \frac{\int_{\mathbb{R}} |x|\, d\mu_i(x)}{K} \leq \frac{1 + M}{K}.$$

This concludes the proof by taking K large enough. □

Appendix B A (weak) existence result in the vague topology

Proposition B.1. *Let $((\mathcal{X}, \mathcal{B}_\mathcal{X}), \{P_\theta : \theta \in \Theta\}, (\mathcal{A}, \mathcal{B}_\mathcal{A}), L)$ be a regular statistical decision problem. Suppose that the parameter space Θ is a locally compact separable metric space. Let $\mathcal{D}_0 \subseteq \mathcal{D}$ be a set of decision rules for the statistical decision problem. Let $\mathcal{P}$ be a set of probability measures on $(\Theta, \mathcal{B}_\Theta)$. Suppose that*

1. *for each θ, the loss L is lower semicontinuous in a;*
2. *the Bayes risk function $\pi \mapsto \inf_\delta B(\delta, \pi)$ is vaguely upper semicontinuous on $\bar{\mathcal{P}}$;*

3. the closure $\bar{\mathcal{P}}$ of the set $\mathcal{P}$ in the space of subprobability measures endowed with the vague topology is convex;

4. the set $\mathcal{D}_0$ is closed and convex as a subset of $\mathcal{D}$ endowed with the weak topology.

Then there exists a pair $(\delta^*, \pi^*) \in \mathcal{D}_0 \times \bar{\mathcal{P}}$ such that

$$B(\delta^*, \pi^*) = \inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \bar{\mathcal{P}}} B(\delta, \pi) = \sup_{\pi \in \bar{\mathcal{P}}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

and that (δ^*, π^*) is a saddle point in the sense that

$$B(\delta^*, \pi) \leq B(\delta^*, \pi^*) \leq B(\delta, \pi^*)$$

for all $\delta \in \mathcal{D}_0$ and all $\pi \in \bar{\mathcal{P}}$.

Proof. By linearity of the space of finite measures, the integrated risk B is linear in π for each $\delta \in \mathcal{D}_0$. Moreover,

$$r(\delta, \theta) = \sup_{c \in \mathcal{C}(\mathcal{A})} \{b_\delta(f_\theta, c) : c \leq L_\theta\},$$

hence r is convex in δ for each $\theta \in \Theta$, and so the integrated risk B is convex in δ for each $\theta \in \Theta$. By assumption, $\bar{\mathcal{P}}$ is convex and $\mathcal{D}_0$ is convex and closed. By Lemma 1.1, $\mathcal{D}$ is compact and so $\mathcal{D}_0$ is compact. It thus follows from Kneser's minimax theorem that

$$\inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \bar{\mathcal{P}}} B(\delta, \pi) = \sup_{\pi \in \bar{\mathcal{P}}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

The set $\mathcal{M}_{\leq 1}(\Theta)$ of subprobability measures on Θ endowed with the vague topology is compact under separability of Θ (see Lemma A.1), hence $\bar{\mathcal{P}}$ is vaguely compact due to metrizability of vague convergence. By assumption, $\pi \mapsto \inf_\delta B(\delta, \pi)$ is vaguely upper semicontinuous, hence the supremum on the right-hand side is attained. Denote $\pi^* \in \bar{\mathcal{P}}$ the distribution that attains this supremum, that is,

$$\sup_{\pi \in \bar{\mathcal{P}}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi) = \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi^*)$$

The proof of 1.2 can be extended to any finite measure using Lemma A.3, and in particular for $\bar{\mathcal{P}} \subseteq \mathcal{M}_{\leq 1}(\Theta)$. It follows that $\delta \mapsto B(\delta, \pi)$ is lower semicontinuous for each $\pi \in \bar{\mathcal{P}}$, and in particular for $\pi^* \in \bar{\mathcal{P}}$. Since $\mathcal{D}_0$ is compact, the infimum on the right-hand side is also attained for some $\delta' \in \mathcal{D}_0$. Therefore, there exists a pair $(\delta', \pi^*) \in \mathcal{D}_0 \times \bar{\mathcal{P}}$ such that

$$B(\delta', \pi^*) = \inf_{\delta \in \mathcal{D}_0} \sup_{\pi \in \bar{\mathcal{P}}} B(\delta, \pi) = \sup_{\pi \in \bar{\mathcal{P}}} \inf_{\delta \in \mathcal{D}_0} B(\delta, \pi).$$

We now show that there exists $\delta^* \in \mathcal{D}_0$ such that (δ^*, π^*) attains the minimax equality and is also a saddle point. Since $\delta \mapsto B(\delta, \pi)$ is lower semicontinuous for each $\pi \in \bar{\mathcal{P}}$, we have that $\delta \mapsto \sup_\pi B(\delta, \pi)$ is lower semicontinuous as the pointwise supremum of lower semicontinuous functions. Since $\mathcal{D}_0$ is compact, the infimum on the left-hand side of the minimax equality is attained.

We thus have

$$\sup_{\pi\in\mathcal{P}} B(\delta^*,\pi) = B(\delta',\pi^*)$$

for some $\delta^* \in \mathcal{D}_0$. Then $B(\delta',\pi^*) \geq B(\delta^*,\pi^*) \geq \inf_\delta B(\delta,\pi^*) = B(\delta',\pi^*)$, and so the supremum on the left-hand side is also achieved for π^*. It follows that $B(\delta',\pi^*) = B(\delta^*,\pi^*)$ and

$$B(\delta^*,\pi^*) = B(\delta',\pi^*) = \sup_{\pi\in\bar{\mathcal{P}}} B(\delta^*,\pi) = \inf_{\delta\in\mathcal{D}_0} B(\delta,\pi^*),$$

which proves that (δ^*,π^*) is a saddle point that attains the minimax equality. □

Lemma B.2. *Let Θ be a locally compact metric space. If π_n converges vaguely to π in $\mathcal{M}_{\leq 1}(\Theta)$, then*

$$\liminf_{n\to\infty} \int f\, d\pi_n \geq \int f\, d\pi$$

for every bounded from below, lower semicontinuous function $f\colon E \to \bar{\mathbb{R}}$.

Proof. Suppose π_n converges vaguely to π. Then by the portmanteau theorem for vague convergence, we have

$$\liminf \pi_n(O) \geq \pi(O)$$

for all open sets $O \subseteq \Theta$. Since f is assumed lower semicontinuous, it has open superlevel sets. Since it is bounded from below by 0, we have that $\int f(\theta)\, d\pi_n(\theta) = \int_0^\infty \pi_n(\{\theta : f(\theta) > y\})\, dy$. Therefore,

$$\liminf_{n\to\infty} \int f(\theta)\, d\pi_n(\theta) \geq \int f(\theta)\, d\pi.$$

□

Corollary B.3. *Suppose the assumptions of Proposition B.1 hold. Suppose, moreover, that:*

5. *for each $\delta \in \mathcal{D}_0$, the risk r is lower semicontinuous in θ;*
6. *the parameter space Θ is Polish.*

Then there exists $\pi' \in \mathcal{P} \subseteq \mathcal{M}_1(\Theta)$ such that $B(\delta^*,\pi') = B(\delta^*,\pi^*)$ where $(\delta^*,\pi^*) \in \mathcal{D}_0 \times \bar{\mathcal{P}}$ is a saddle point solution of the minimax equality whose existence is guaranteed by Proposition B.1.

Proof. If $\pi^* \in \mathcal{P}$, there is nothing to prove. Now take $\pi^* \in \bar{\mathcal{P}} \setminus \mathcal{P}$. Since the vague topology on $\mathcal{M}_{\leq 1}(\Theta)$ is metrizable under the assumptions on Θ (see Theorem 31.5 in Bauer (2011) and the separability argument in Lemma A.1), there exists a sequence (π_n) in $\mathcal{P}$ such that π_n converge vaguely to π^*. Since r is lower semi-continuous in θ by assumption, we have $\liminf_{n\to\infty} B(\delta,\pi_n) \geq B(\delta,\pi^*)$ for each $\delta \in \mathcal{D}_0$ by Lemma B.2. In particular, there is $n \in \mathbb{N}$ such that $B(\delta^*,\pi_n) \geq B(\delta^*,\pi^*)$. Since (δ^*,π^*) is a saddle point, we also have $B(\delta^*,\pi_n) \leq B(\delta^*,\pi^*)$. Thus $B(\delta^*,\pi_n) = B(\delta^*,\pi^*)$ and $\pi_n \in \mathcal{P}$, which proves the claim by taking $\pi' = \pi_n$. □