Economics Letters 194 (2020) 109404

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

When does ambiguity fade away?

Filippo Massari [a,b,*], Jonathan Newton [c]

[a] University of East Anglia, United Kingdom of Great Britain and Northern Ireland
[b] Bocconi University, Italy
[c] Kyoto University, Japan

ARTICLE INFO

Article history:
Received 20 October 2019
Received in revised form 8 July 2020
Accepted 9 July 2020
Available online 11 July 2020

JEL classification:
D81
D83
C11

Keywords:
Ambiguity
Learning
Robust statistical decisions
Misspecified learning

ABSTRACT

the prior support is finite, long-run ambiguity is known to be a possible outcome only if the learning problem is misspecified (Marinacci and Massari, 2019). We show that if the prior support is naturally rich, long-run ambiguity cannot occur.

1. Introduction

Researchers have considered the implications of ambiguity for many economic phenomena. Examples include trade (Kajii and Ui, 2006), portfolio selection (Garlappi et al., 2006), risk pricing (Augustin and Izhakian, 2020), savings behavior (Hansen et al., 1999), job search (Nishimura and Ozaki, 2004) and the possibility of speculative bubbles (Werner, 2019).[1] Given the salience of ambiguity in economic and financial research, it is natural to wonder about how persistent it is. In the current paper, we focus on the multiple prior model of ambiguity and consider conditions under which ambiguity fades away in the long run as a consequence of learning.

When a Bayesian decision-maker's set of priors comprises a finite set of iid models that includes the true model, Marinacci (2002) shows that ambiguity fades away over time as the decision-maker learns the true model. Marinacci and Massari (2019) drop the iid assumption and allow the problem to be misspecified so that it is impossible for the decision-maker to learn the true model. Nevertheless, they can still provide tight conditions under which ambiguity fades away. However, many applications, including all those mentioned above, feature decision-makers with sets of priors on the whole parameter space, a set of positive Lebesgue measure. It is this latter setup that we study in the current paper. We demonstrate that, under natural assumptions, ambiguity fades away on all sequences with finite maximum likelihood. Over time, all the posteriors concentrate on a shrinking neighborhood of this estimate and ambiguity fades away. Notably, the result holds even if the maximum likelihood estimate does not converge to a limit: all priors eventually concentrate around the estimate, even if the estimate itself changes over time.

The impact of ambiguity fading away will differ across models. For example **(I)** (Kajii and Ui, 2006) give necessary and sufficient conditions under which trade can take place under ambiguity. Trade that does take place in these conditions will be unaffected by ambiguity fading away, but additional opportunities for trade may arise.[2] **(II)** (Werner, 2019) shows that speculative trading bubbles can arise when market participants have common but ambiguous beliefs. Consequently, if ambiguity fades away, then another explanation for long-run speculative trade is required. **(III)** (Garlappi et al., 2006) consider mean–variance portfolio selection with an ambiguous parameter. If ambiguity fades away,

* Corresponding author at: Bocconi University, Italy.
E-mail address: F.Massari@uea.ac.uk (F. Massari).

[1] The reader is referred to the survey article by Gilboa and Marinacci (2016) for more examples.

[2] In the model of Kajii and Ui (2006), trade between two players is possible if and only if their sets of priors do not overlap. It is easy to see that if their sets of priors do not overlap under ambiguity, then the players will differ in their beliefs after ambiguity has faded away. Conversely, even if their sets of priors overlap under ambiguity, it is possible that the players will differ in their beliefs after ambiguity has faded away.

https://doi.org/10.1016/j.econlet.2020.109404
0165-1765/

then the model eventually returns to the classical mean–variance model (Markowitz, 1952; Sharpe, 1970).[3]

There are other models that study the effect that learning has on ambiguity, and some of these models (see, e.g. Epstein and Schneider, 2007) allow for persistent ambiguity. The multiple prior model we describe relies on the strong law of large numbers. Because the strong law of large numbers holds for each prior, all priors concentrate on the same model and ambiguity fades away.

2. Probabilities

We consider a family of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ parametrized by a positive Lebesgue measure parameter set $\Theta \subset \mathbb{R}^k$, defined on a σ-algebra Σ^∞ of subsets of X^∞ with representative element $x^\infty = x_1, x_2, \ldots$, where $X^\infty := \times^\infty X$ is the infinite Cartesian product of a state space X with representative element x and σ-algebra Σ. With a slight abuse of notation, we use $P_\theta(x^t)$ to denote the probability that model P_θ attaches to the cylinder with base x^t (*i.e.*, $Cyl(x^t) := \{x_1, \ldots, x_t, X_{t+1}, X_{t+2}, \ldots\}$), as well as the likelihood that model P_θ attaches to the partial sequence $(x_1, \ldots, x_t)$.

Specifically, we focus on the case in which $\mathcal{M}$ is a regular exponential family in the natural parametrization — most of the commonly used distributions form a regular exponential family and can be re-parametrized into their natural parametrization form (e.g. Gaussian, Multinomial, Poisson, …), which covers most standard learning settings, including those cited in the introduction.[4]

Definition 1 (*Exponential Family*)**.** Let ν be a σ-finite measure on the Borel subsets of $\mathbb{R}^k$ and $\mathcal{H}$ be the support of ν. Define

$$\Theta := \left\{\theta \in \mathbb{R}^k : \int_{\mathcal{H}} \exp(\theta^T x)\nu(dx) < \infty\right\};$$

define a function ψ and a probability densities P_θ on X with respect to ν by $\psi(\theta) := \ln \int_X exp(\theta^T x)\nu(dx)$ and $P_\theta(x) := exp(\theta^T x - \psi(\theta))$. We refer to $\mathcal{M} := \{P_\theta(x)|\theta \in \Theta\}$ as an *exponential family* in the natural parametrization. An exponential family is *regular* if Θ is an open set.

The prior information about the parameters is summarized by prior distributions $\mu \in \Delta\Theta$. The set of prior distributions is $\mathcal{C}$. For any prior distribution $\mu \in \mathcal{C}$ the joint distribution of the parameters and the observations is $P^\mu \in \Delta(\Theta \times X^\infty)$, defined by, for all sets $A \subseteq \Theta$ and all cylinders x^t,

$$P^\mu(A \times x^t) := \int_A P_\theta(x^t)d\mu.$$

We denote by $\mu(.|x^t) \in \Delta\Theta$ the usual posterior given the observations x^t,[5] while $P^\mu(.|x^t) \in \Delta(\Theta \times X)$ is the one-step-ahead predictive distribution of x_{t+1}, given observations x^t. By definition, for all $A \subseteq \Theta$ we have

$$P^\mu(A \times x_{t+1}|x^t) := \int_A P_\theta(x_{t+1}|x^t)d\mu(.|x^t)$$

$$:= \int_A P_\theta(x_{t+1}|x^t)\frac{P_\theta(x^t)d\mu}{\int_\Theta P_\theta(x^t)d\mu}.$$

When $A = \Theta$ we use the lighter notation $P^\mu(x|x^t) := P^\mu(\Theta \times x_{t+1}|x^t)$.

3. Long-run ambiguity

As in Marinacci (2002), we consider the difference between a decision-maker's expected utility under the most advantageous prior and under the least advantageous prior in $\mathcal{C}$ to be a measure of the ambiguity that the decision-maker perceives in evaluating an act. If the set of priors $\mathcal{C}$ is compact, as we always assume, a tight sufficient condition for this difference to be zero is that the posteriors calculated from all priors in $\mathcal{C}$ eventually coincide (Marinacci and Massari, 2019).

Definition 2. Ambiguity fades away at path $x^\infty \in X^\infty$ if,

$$\lim_{t\to\infty}\left[\sup_{\mu',\mu''\in\mathcal{C}}\int_X \left|dP^{\mu''}(x|x^t) - dP^{\mu'}(x|x^t)\right|\right] = 0 \tag{1}$$

where, for each $t > 0$, x^t indicates the first t realizations of path x^∞.

Definition 2 does not depend on the true model, which in any practical learning situation is not known by the decision-maker. It requires that all priors concentrate eventually on the same parameter (or on a set of parameters with identical predictions) on the realized path. Its relation with the familiar notion of *weak merging* (Kalai and Lehrer, 1994) is as follows. In well-specified learning problems all priors *weakly merge* to the true and ambiguity fades away. However, ambiguity might and does fade away in many misspecified learning problems in which the priors do not *weakly merge* with the truth.

4. Main result

In this section, we identify conditions that guarantee that ambiguity fades away in the long-run when Θ has positive Lebesgue measure. These regularity conditions are borrowed from Grünwald (2007) conditions for the BIC approximation (Schwarz, 1978; Clarke and Barron, 1990), to which we add a compactness assumption on the set of priors $\mathcal{C}$ to ensure convergence.

Definition 3. The learning problem is **regular** if

- **A1:** $\mathcal{M}$ is a regular exponential family;
- **A2:** the set of priors, $\mathcal{C}$, is compact;
- **A3:** priors in $\mathcal{C}$ are continuous and strictly positive on every compact subset of Θ.

Condition **A1** is a high order assumption that limits our attention to densities that are measurable jointly in x and θ and regular enough for the empirical maximum likelihood to be unique in every period (in the canonical representation Θ is a convex set). Further, it allows writing the Fisher information matrix as the Hessian of the relative entropy. This assumption is stronger than condition (*i*) of Berk (1966) and, together with **A3**, allows us to drop all assumptions about the data generating process. Unlike (Berk, 1966), we do not require draws to be iid. **A2** is needed to ensure uniform convergence in the set of priors (Marinacci, 2002). **A3** requires that priors have full, and thus common, support. This assumption reflects the attitude of an agent that does not rule out a-priori any parameter choice. The restriction on priors to be strictly positive in every compact subset of Θ, rather than Θ itself, is due to the fact that $|\Theta| = \mathbb{R}^k$ for many members of the exponential family in the canonical representation.

[3] Garlappi et al. (2006) and Hansen et al. (1999) belong to a special class of ambiguous models known as ε-*contamination* models (see, e.g. Berger, 2013), in which the set of priors consists of all models within some distance ε of an estimated model. Such models satisfy our condition of a positive Lebesgue measure of models in the support of the decision-maker.

[4] We refer the reader to Nielsen and Garcia (2009) for a brief and schematic description of the main characteristic of the exponential family and a useful mapping between their canonical and natural parametrization.

[5] We rule out the possibility of observing an event which is impossible according to all models in $\mathcal{M}$.

Definition 4. $\hat{\theta}(x^t)$ denotes the (vector valued) maximum likelihood estimator at x^t :

$$\hat{\theta}(x^t) = \operatorname*{argmax}_{\theta\in\Theta} P_\theta(x^t);$$

The equality in the definition above is justified because Assumptions **A1** and **A3** guarantee that the support, Θ, is convex, so that a unique maximum likelihood exists at every finite history. We now present our main result: ambiguity fades away on all sequences such that the sequence of maximum likelihoods is bounded.

Theorem 1. *If the learning problem is regular, ambiguity fades away on all sequences such that* $\limsup \|\hat{\theta}_t\| < \infty$.

Proof. See Appendix. □

Theorem 1 makes no reference to the truth. The point of view we adopt in Theorem 1 is empirical and differs from that of standard convergence results (e.g., Blackwell and Dubins (1962), Doob (1949) and Berk (1966)). Instead of postulating the existence of a true distribution and deriving almost sure results, we show that convergence to the same predictive distributions occurs on all paths in which the sequence of maximum likelihood parameters is uniformly bounded. Being agnostic about the true distribution renders our approach particularly suited to discuss convergence in possibly misspecified learning environments.

Theorem 1 shows that ambiguity cannot persist in the long-run in the multiple-priors Bayesian learning model with naturally rich support. Its result naturally translates to exponential families when expressed in their canonical parametrization. For example, it tells us that:

- Ambiguity fades away on all sequences with frequency of heads uniformly bounded away from 0 and 1 if the Bayesian decision-maker believes draws are from an ambiguous coin and $\mathcal{C}$ is a compact set of not-degenerate Beta priors on the probability of heads p^h (i.e., $\mathcal{C} = \{Beta(\alpha,\beta), \alpha \in [a,b], \beta \in [c,d]\}$, with $[a,b]$, $[c,d]$ strictly positive, finite intervals). In order to apply Theorem 1, we need to verify that on all sequences with frequency of heads p^h uniformly bounded away from 0 and 1 the maximum likelihood estimator of the natural parameter of Bernoulli family $\hat{\theta}$ satisfies the condition $\limsup \|\hat{\theta}_t\| < \infty$. From the conversion table of Nielsen and Garcia (2009), we see that $\theta^{p_h} = \ln \frac{p_h}{1-p_h}$; so, $(\|\hat{\theta}_t^{p_h}\|)_{t=1}^\infty$ is bounded if and only if $(\hat{p}_t^h)_{t=1}^\infty$ is uniformly bounded away from 0 and 1.
- Ambiguity fades away on all bounded sequences, if the Bayesian decision-maker believes that realizations are Gaussian with known positive variance σ and unknown mean μ, and he has a compact set of non-degenerate Gaussian priors on the values of μ(i.e., $\mathcal{C}_\mu := \{\mathcal{N}(\mu^\mu, \sigma^{\mu^*}), \mu^\mu \in [a,b], \sigma^\mu \in [c,d]\}$ with $[a,b]$ finite and $[c,d]$ finite strictly positive intervals). In order to apply Theorem 1, we need to verify that if the sequence is bounded, the maximum likelihood estimator of the natural parameters of the Gaussian family $\hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2) \in \mathbb{R}\times(R^-)$ satisfies the condition $\limsup \|\hat{\theta}_t\| < \infty$. From the conversion table of Nielsen and Garcia (2009), we see that $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$, so that $\hat{\theta} = (\frac{\hat{\mu}_t}{\hat{\sigma}_t^2}, -\frac{1}{2\hat{\sigma}_t^2})$, where $\hat{\mu}_t$ and $\hat{\sigma}_t$ are the maximum likelihood estimators of mean, $\hat{\mu}_t := \frac{1}{t}\sum_{\tau=1}^t x_t$, and variance, $\hat{\sigma}_t^2 := \frac{1}{t}\sum_{\tau=1}^t (x_\tau - \hat{\mu}_t)$. So, $\limsup \|\hat{\theta}_t\| < \infty$ because for all t, $\hat{\mu}_t < \infty$ on all bounded sequences and $\sigma > 0 \Rightarrow \hat{\sigma}_t > 0$ for all large t.

Appendix

In this appendix $\hat{\theta}_t := \hat{\theta}(x^t)$, and we make use of the K–L divergence. Let $S(\mathcal{M},\Theta)$ be the set of sequences such that $\limsup \|\hat{\theta}_t\| < \infty$.

Definition 5. The *K–L divergence* from $P_{\hat{\theta}_t}$ to P_θ is

$$D\left(P_{\hat{\theta}_t} \,\|\, P_\theta\right) := E_{P_{\hat{\theta}_t}}\left[\ln \frac{P_{\hat{\theta}_t}(x)}{P_\theta(x)}\right].$$

The proof is a standard application of the Laplace method. The strategy is to show that for t large, for all priors in $\mathcal{C}$, the value of the integral of the unconditional probabilities is well approximated by the value it assumes on a shrinking interval around the minimizer of the K–L divergence (i.e., by the maximum likelihood model). Because Θ is convex and $-D(P_{\hat{\theta}_t} \,\|\, P_\theta)$ is strictly concave, this minimizer is unique. Because the exponential family is regular $x^\infty \in S(\mathcal{M},\Theta) \Rightarrow (\hat{\theta}_t)_{t=1}^\infty$ belongs to a compact subset of Θ and the approximation below is never on the boundary of the support.[6]

Proof of Theorem 1

Proof. $\mathcal{C}$ compact $\Rightarrow$ $\forall x^t$, $\operatorname{argmax}_{\mu\in\mathcal{C}} \lim_{t\to\infty} \int_X |dP^g(x|x^t) - dP^h(x|x^t)|$ exists. Thus, it suffices to show that if the learning problem is regular, then $\forall x^\infty \in \hat{S}(\mathcal{M},\Theta)$ and $\forall g, h \in \mathcal{C}$, $\lim_{t\to\infty}\int_X |dP^g(x|x^t) - dP^h(x|x^t)| = 0$.

$$\begin{aligned}
0 &\le \lim_{t\to\infty}\int_X |dP^g(x|x^t) - dP^h(x|x^t)| \\
&:= \lim_{t\to\infty}\int_X \left|\int_\Theta P_\theta(x)\left(\frac{P_\theta(x^t)g(\theta)}{P^g(x^t)} - \frac{P_\theta(x^t)h(\theta)}{P^h(x^t)}\right)d\theta\right| dx \\
&=^a \lim_{t\to\infty}\int_X \left|\int_\Theta P_\theta(x)\left(\frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)}{\int_\Theta e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)\,d\theta}\frac{P_{\hat{\theta}_t}(x^t)}{P_{\hat{\theta}_t}(x^t)}\right.\right. \\
&\quad \left.\left. - \frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)}{\int_\Theta e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)\,d\theta}\frac{P_{\hat{\theta}_t}(x^t)}{P_{\hat{\theta}_t}(x^t)}\right)d\theta\right| dx \\
&=^b \int_X \lim_{t\to\infty}\left|\int_\Theta P_\theta(x)\left(\frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)}{\int_\Theta e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)\,d\theta}\right.\right. \\
&\quad \left.\left. - \frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)}{\int_\Theta e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)\,d\theta}\right)d\theta\right| dx \\
&=^{c,d} \int_X \lim_{t\to\infty}\left|\int_{B_t} P_\theta(x)\left(\frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)}{\int_{B_t} e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}g(\theta)\,d\theta}\right.\right. \\
&\quad \left.\left. - \frac{e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)}{\int_{B_t} e^{-tD\left(P_{\hat{\theta}_t}\|P_\theta\right)}h(\theta)\,d\theta}\right)d\theta\right| dx \\
&\le^e \int_X \lim_{t\to\infty}\left|\int_{B_t} P_\theta(x)\max\left\{\left|\frac{\frac{\sqrt{(2\pi)^k}g_t^+}{\sqrt{t^k\det(I_t^-)}}}{\frac{\sqrt{(2\pi)^k}g_t^-}{\sqrt{t^k\det(I_t^+)}}} - \frac{\frac{\sqrt{(2\pi)^k}h_t^-}{\sqrt{t^k\det(I_t^+)}}}{\frac{\sqrt{(2\pi)^k}h_t^+}{\sqrt{t^k\det(I_t^-)}}}\right|\right.\right.
\end{aligned}$$

[6] For non-regular member of the exponential family, the proof below cannot be adopted for those sequences on which the maximum likelihood estimator are within an order $1/\sqrt{t}$ to the boundary of θ because Laplace approximation is truncated. For those sequences the shape of $P_\theta(x^t)$ becomes a truncated Gaussian with a reduced value of the integral in Lemma 1. For those sequences, however, the discrepancy in the approximation is only a constant (Xie and Barron, 2000), and it would not affect our result.

$$\left|\frac{\frac{\sqrt{(2\pi)^k}g_t^-}{\sqrt{t^k\det(I_t^+)}}}{\frac{\sqrt{(2\pi)^k}g_t^+}{\sqrt{t^k\det(I_t^-)}}} - \frac{\frac{\sqrt{(2\pi)^k}h_t^+}{\sqrt{t^k\det(I_t^-)}}}{\frac{\sqrt{(2\pi)^k}h_t^-}{\sqrt{t^k\det(I_t^+)}}}\right|\Bigg\}\Bigg| dx$$

$$\leq^f \int_X \lim_{t\to\infty} P^+(x) \max\left\{\left|\frac{\frac{\sqrt{(2\pi)^k}g_t^+}{\sqrt{t^k\det(I_t^-)}}}{\frac{\sqrt{(2\pi)^k}g_t^-}{\sqrt{t^k\det(I_t^+)}}} - \frac{\frac{\sqrt{(2\pi)^k}h_t^-}{\sqrt{t^k\det(I_t^+)}}}{\frac{\sqrt{(2\pi)^k}h_t^+}{\sqrt{t^k\det(I_t^-)}}}\right|;\right.$$

$$\left.\left|\frac{\frac{\sqrt{(2\pi)^k}g_t^-}{\sqrt{t^k\det(I_t^+)}}}{\frac{\sqrt{(2\pi)^k}g_t^+}{\sqrt{t^k\det(I_t^-)}}} - \frac{\frac{\sqrt{(2\pi)^k}h_t^+}{\sqrt{t^k\det(I_t^-)}}}{\frac{\sqrt{(2\pi)^k}h_t^-}{\sqrt{t^k\det(I_t^+)}}}\right|\right\} dx$$

$$= \int_X \lim_{t\to\infty} P^+(x) \max\left\{\left|\frac{g_t^+}{g_t^-}\frac{\sqrt{\det(I_t^+)}}{\sqrt{\det(I_t^-)}} - \frac{h_t^-}{h_t^+}\frac{\sqrt{\det(I_t^-)}}{\sqrt{\det(I_t^+)}}\right|;\right.$$

$$\left.\left|\frac{g_t^-}{g_t^+}\frac{\sqrt{\det(I_t^-)}}{\sqrt{\det(I_t^+)}} - \frac{h_t^+}{h_t^-}\frac{\sqrt{\det(I_t^+)}}{\sqrt{\det(I_t^-)}}\right|\right\} dx$$

$$=^g 0.$$

(a) A known result for members of the exponential family (e.g., Grünwald, 2007, Chapter 8) is that

$$P^g(x^t) = \int_\Theta P_\theta(x^t) g(\theta)\, d\theta = \frac{\int_\Theta e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)} g(\theta)\, d\theta}{P_{\hat\theta_t}(x^t)}.$$

(b) We can exchange the order of limit and integration by the Lebesgue dominated convergence theorem.

(c) B_t is a neighborhood of the maximum likelihood that, in all dimensions, converges to zero at a rate slightly slower than $\sqrt{\frac{1}{t}}$. That is $B_t := \{\theta \in \Theta \subset \mathbb{R}^k : \forall i = 1, \ldots, k, |\theta^i - \hat\theta^i| \leq t^{-\frac{1}{2}-\alpha}\}$ for some $\alpha \in (0, .5)$.

(d) By Lemma 1**(i)**, $\int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)} h(\theta)\, d\theta \to 0$ exponentially fast and it can be ignored in the calculation of the limit.

(e) By Lemma 1**(ii)**, with $I := E_{P_{\hat\theta_t}}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln P_\theta\right]_{\theta=\hat\theta_t}$, $\det(I_t^-) = \inf_{\theta'\in B_t}\det(I(\theta'))$, $\det(I_t^+) = \sup_{\theta'\in B_t}\det(I(\theta'))$, $g_t^- = \inf_{\theta'\in B_t} g(\theta')$, $g_t^+ = \sup_{\theta'\in B_t} g(\theta')$, $h_t^- = \inf_{\theta'\in B_t} h(\theta')$, $h_t^+ = \sup_{\theta'\in B_t} h(\theta')$.

(f) With $P^+(x) = \sup_{\theta\in B_t} P_\theta(x) < 1$.

(g) Continuity and strict positivity of $g(\cdot), h(\cdot)$ in $\det(I(\cdot))$ in B_t guarantee that for all $x^\infty \in \hat S(\mathcal{M}, \Theta)$ the following limit holds[7]:

$$\max\left\{\left|\frac{g_t^+}{g_t^-}\frac{\sqrt{\det(I_t^+)}}{\sqrt{\det(I_t^-)}} - \frac{h_t^-}{h_t^+}\frac{\sqrt{\det(I_t^-)}}{\sqrt{\det(I_t^+)}}\right|;\right.$$

$$\left.\left|\frac{g_t^-}{g_t^+}\frac{\sqrt{\det(I_t^-)}}{\sqrt{\det(I_t^+)}} - \frac{h_t^+}{h_t^-}\frac{\sqrt{\det(I_t^+)}}{\sqrt{\det(I_t^-)}}\right|\right\} \to 0. \quad \square$$

Lemma 1. *Let $\mathcal{M}$ be a regular member of the exponential family parametrized by Θ and μ a prior that satisfies **A3**, then, $\forall x^\infty \in \hat S(\mathcal{M}, \Theta)$,*

$$\int_\Theta e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)}\mu(\theta)d\theta = \int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)}\mu(\theta)\, d\theta + \int_{B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)}\mu(\theta)\, d\theta,$$

and, for t large, the following bounds holds uniformly when B_t is a neighborhood of the maximum likelihood such that $diam(B_t) \to^{t\to\infty} 0$ at a rate slightly slower than $\sqrt{\frac{1}{t}}$.

(i) First integral: $\exists k > 0 : \mathcal{I}_1 = \int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)}\mu(\theta)d\theta < e^{-rt^{2\alpha}}$.

(ii) Second integral: *Let $\mathcal{I}_2 = \int_{B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)}\mu(\theta)d\theta$; $I(\theta_t) := E_{P_{\hat\theta_t}}\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln P_\theta\right\}_{\theta=\hat\theta_t}$ be the Fisher information evaluated at the maximum likelihood parameter,[8] k be the dimensionality of Θ, $\det(I_t^-) = \inf_{\theta'\in B_t}\det(I(\theta'))$, $\det(I_t^+) = \sup_{\theta'\in B_t}\det(I(\theta'))$; and $\mu_t^- = \inf_{\theta'\in B_t}\mu(\theta')$, $\mu_t^+ = \sup_{\theta'\in B_t}\mu(\theta')$, then*

$$\frac{\mu_t^-(2\pi)^{k/2}}{\sqrt{t^k det(I_t^+)i}} \leq \mathcal{I}_2 \leq \frac{\mu_t^+(2\pi)^{k/2}}{\sqrt{t^k\det(I_t^-)}}.$$

Proof. Let $B_t := \{\theta \in \Theta \subset \mathbb{R}^k : \forall i = 1, \ldots, k, |\theta^i - \hat\theta^i| \leq t^{-\frac{1}{2}-\alpha}\}$ for some $\alpha \in (0, .5)$. To gain intuition, take α very small, so that B_t is a neighborhood of the maximum likelihood that shrinks to 0 at a rate slightly slower than $1/\sqrt{t}$ in all dimensions. Because $x^\infty \in \hat S(\mathcal{M}, \Theta)$ and μ is continuous and positive on all compact subsets of Θ (by **A3**), there is a $T : \forall t > T, B_t \subset \hat\Theta$ where $\hat\Theta$ is a compact subset of Θ in which $\mu > \epsilon > 0$ for some positive ϵ. We always assume $t > T$.

The proof is done by performing a second-order Taylor expansion of $D(P_{\hat\theta_t} \| P_\theta)$ to bound the two integrals. $\mathcal{M}$ is an exponential family; thus, $D(P_{\hat\theta_t} \| P_\theta)$ can be exactly approximated in B_t as follows (see Grünwald, 2007, chapter 19):

$$D\left(P_{\hat\theta_t} \| P_\theta\right) = \frac{1}{2}\left(\hat\theta_t - \theta\right)^T I(\theta^*)\left(\hat\theta_t - \theta\right), \tag{2}$$

for some $\theta^* \in B_t$ such that θ^* lies between θ and $\hat\theta_t$ – here, $I := E_{P_{\hat\theta_t}}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln P_\theta\right]_{\theta=\hat\theta_t}$; because $\mathcal{M}$ is an exponential family, this is the Fisher information matrix evaluated at the maximum likelihood estimator.

(i) First integral: Because $D\left(P_{\hat\theta_t} \| P_\theta\right)$, as a function of θ, is strictly convex, has a minimum at $\theta = \hat\theta_t$, and is increasing in $\|\theta - \hat\theta_t\|$, the following holds:

$$0 < \int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)} g(\theta)\, d\theta < \int_{\Theta\setminus B_t} e^{-t\min_{\theta\in\Theta\setminus B_t} D\left(P_{\hat\theta_t}\|P_\theta\right)} g(\theta)\, d\theta$$

where

$$\min_{\theta\in\Theta\setminus B_t} D\left(P_{\hat\theta_t} \| P_\theta\right) =^{(a)} \min_{\theta\in\partial B_t} D\left(P_{\hat\theta_t} \| P_\theta\right) \geq^{(b)} \frac{1}{2}t^{-1+2\alpha}\min_{\theta\in int(\Theta)} \mathbf{1}^T I(\theta)\mathbf{1},$$

where (a) holds because strict convexity of $D(\cdot\|\cdot)$ implies that the $D(\cdot\|\cdot)$ is minimal at the boundary of B_t; and (b) holds, with $\mathbf{1}$ being a k-dimensional vector of 1s, because of the definition of B_t and Eq. (2). So, since $I(\theta)$ is continuous and > 0 for all $\theta \in \Theta$, and also $\int_{\Theta\setminus B_t}\mu(\theta)\, d\theta \leq 1$,

$$0 < \int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)} g(\theta)\, d\theta < \int_{\Theta\setminus B_t} e^{-t\left(\frac{1}{2}t^{-1+2\alpha}\min_{\theta\in int(\Theta)} I(\theta)\right)} g(\theta)\, d\theta < e^{-rt^{2\alpha}},$$

for $r = \frac{1}{2}\min_{\theta\in int(\Theta)} I(\theta) > 0$.

(ii) Second integral: by Eq. (2),

$$\mathcal{I}_2 = \int_{B_t} e^{-tD\left(P_{\hat\theta_t}\|P_\theta\right)} g(\theta)\, d\theta = \int_{B_t} e^{-\frac{t}{2}(\hat\theta_t-\theta)^T I(\theta')(\hat\theta_t-\theta)} g(\theta)\, d\theta$$

[7] By construction, for all $\forall x^\infty \in \hat S(\mathcal{M}, \Theta)$, B_t is a subset of a compact of Θ; thus, by (**A3**), $g(\cdot)$ and $h(\cdot)$ are continuous and bounded away from zero. $\det(I(\cdot))$ is continuous bounded away from zero because $\mathcal{M}$ is a member of the exponential family (**A1**).

[8] Which is positive definite because $\mathcal{M}$ is an exponential family.

where θ' depends on θ. Let $I_t^- = \text{argmin}_{\theta' \in B_t} \det(I(\theta'))$ and $I_t^+ = \text{argmax}_{\theta' \in B_t} \det(I(\theta'))$; it follows that

$$g_t^- \int_{B_t} e^{-\frac{t}{2}(\hat{\theta}_t - \theta)^T I_t^+ (\hat{\theta}_t - \theta)} d\theta \leq \mathcal{I}_2 \leq g_t^+ \int_{B_t} e^{-\frac{t}{2}(\hat{\theta}_t - \theta)^T I_t^- (\hat{\theta}_t - \theta)} d\theta.$$

Performing the substitutions $z^T = \left(\sqrt{t}(\hat{\theta}_t - \theta)\right)^T A_t^+$ on the left integral and $z^T = \left(\sqrt{t}(\hat{\theta}_t - \theta)\right)^T A_t^-$ on the right integral — where A_t^+ and A_t^- are matrixes such that $A_t^+ (A_t^+)^T = I_t^+$ and $A_t^- (A_t^-)^T = I_t^-$, respectively —, we get

$$\frac{g_t^-}{\sqrt{t^k \det(I_t^+)}} \int_{|z| < |t^\alpha \mathbf{1}^T A_t^+|} e^{-\frac{1}{2} z^T z} dz \leq \mathcal{I}_2 \leq \frac{g_t^+}{\sqrt{t^k \det(I_t^-)}} \int_{|z| < |t^\alpha \mathbf{1}^T A_t^-|} e^{-\frac{1}{2} z^T z} dz,$$

— where for a vector x, the vector $|x|$ indicates the vector whose entries are the absolute values of x and k is the dimensionality of Θ^k — and recognize these integrals as proportional to standard multivariate Gaussian. Because, as $t \to \infty$, $I_t^- \to I(\hat{\theta}_t)$ and $I_t^+ \to I(\hat{\theta}_t)$, the domain of integration tends to infinity for both integrals, they both converge to $\sqrt{(2\pi)^k}$.

This approximation holds uniformly for all $x^\infty \in \hat{S}(\mathcal{M}, \Theta)$ because (i) the bound on $\mathcal{I}_1$ does not depend on x^t, and (ii) convergence of $\mathcal{I}_2$ is uniform because **A1** and **A3** guarantee that $g(\theta)$ and $I(\theta)$ are continuous, positive functions of θ over every compact subset of Θ. □

References

Augustin, P., Izhakian, Y., 2020. Ambiguity, volatility, and credit risk. The Review of Financial Studies 33 (4), 1618–1672.

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

Berk, R.H., 1966. Limiting behavior of posterior distributions when the model is incorrect. Ann. Math. Stat. 37 (1), 51–58.

Blackwell, D., Dubins, L., 1962. Merging of opinions with increasing information. Ann. Math. Stat. 33 (3), 882–886.

Clarke, B.S., Barron, A.R., 1990. Information-theoretic asymptotics of Bayes methods. IEEE Trans. Inform. Theory 36 (3), 453–471.

Doob, J.L., 1949. Application of the Theory of Martingales. Colloques Internationaux du Centre National de la Recherche Scientifique Paris, pp. 23–27.

Epstein, L.G., Schneider, M., 2007. Learning under ambiguity. Rev. Econom. Stud. 74 (4), 1275–1303.

Garlappi, L., Uppal, R., Wang, T., 2006. Portfolio selection with parameter and model uncertainty: A multi-prior approach. Rev. Financ. Stud. 20 (1), 41–81.

Gilboa, I., Marinacci, M., 2016. Ambiguity and the Bayesian paradigm. In: Readings in Formal Epistemology. Springer, pp. 385–439.

Grünwald, P.D., 2007. The Minimum Description Length Principle. MIT press.

Hansen, L.P., Sargent, T.J., Tallarini Jr, T.D., 1999. Robust permanent income and pricing. Rev. Econ. Stud. 873–907.

Kajii, A., Ui, T., 2006. Agreeable bets with multiple priors. J. Econom. Theory 128 (1), 299–305.

Kalai, E., Lehrer, E., 1994. Weak and strong merging of opinions. J. Math. Econom. 23 (1), 73–86.

Marinacci, M., 2002. Learning from ambiguous urns. Statist. Papers 43 (1), 143–151.

Marinacci, M., Massari, F., 2019. Learning from ambiguous and misspecified models. J. Math. Econom. 84, 144–149.

Markowitz, H., 1952. Portfolio selection. J. Finance 7 (1), 77–91.

Nielsen, F., Garcia, V., 2009. Statistical exponential families: A digest with flash cards. arXiv preprint arXiv:0911.4863.

Nishimura, K.G., Ozaki, H., 2004. Search and Knightian uncertainty. J. Econom. Theory 119 (2), 299–333.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Sharpe, W.F., 1970. Portfolio Theory and Capital Markets, Vol. 217. McGraw-Hill New York.

Werner, J., 2019. Speculative Trade under Ambiguity. Technical report, mimeo.

Xie, Q., Barron, A.R., 2000. Asymptotic minimax regret for data compression, gambling, and prediction. IEEE Trans. Inform. Theory 46 (2), 431–445.