When does ambiguity fade away?

Filippo Massari* Jonathan Newton

Bocconi University Kyoto University

January 24, 2020

Abstract

Is long-run ambiguity a possible outcome of the multiple prior Bayesian learning model? If 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). Conversely, here we show that, under natural assumptions, ambiguity fades away on most paths if the prior support is naturally rich.

Keywords: Ambiguity, learning, robust statistical decisions, misspecified learning.

JEL Classification: D81, D83, C11

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, 2019), 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

*Corresponding author. E-mail: *massari3141@gmail.com*.

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

fades away. However, many applications, including all those mentioned above, feature decision-makers with sets of priors that are not finite, but are instead compact sets with positive Lebesgue measure on some parameter space. It is this latter setup that we study in the current paper, showing that for the exponential family of models, ambiguity fades away on most sequences.

The conditions for our result are relatively weak. It applies to any sequence of observations such that a unique maximum likelihood estimate exists at any given date sufficiently far along the sequence. This holds, for example, if the prior support is convex, in which case the concavity of the log-likelihood function implies a unique maximum likelihood estimate. 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.

From an applied perspective, this result suggests that ambiguity should not be a concern for a decision-maker who makes a large number of decisions and whose payoff gets averaged over time. This is because the strong law of large numbers allows us to average payoffs irrespective of the priors. Of course, 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, then the model eventually returns to the classical mean-variance model (Markowitz, 1952; Sharpe, 1970).[3]

When does our result not apply? Firstly, it does not apply to situations in which the decision-maker needs to make an important one-off decision such as buying a house or health insurance. In this case, there is no long term learning as there is no long term. Another possibility is that ambiguity may persist for some exogenous reason. For example, there may be aspects of the problem that cannot be learned or that the decision-maker refuses to learn for some reason. Clearly, ambiguity can then persist with respect to these aspects (see, e.g. Epstein and

[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.

[3]Garlappi et al. (2006); 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.

Schneider, 2007).

2 Probabilities

We consider a family of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ parameterized by a positive Lebesgue measure parameter set $\Theta \subset \mathbb{R}^n$, defined on a σ-algebra Σ^∞ of subsets of X^∞ with representative element $x^\infty = x_1, x_2, ...$, 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 $(i.e., Cyl(x^t) := \{x_1, ..., x_t, X_{t+1}, X_{t+2}, ...\})$, as well as the likelihood that model P_θ attaches to the partial sequence $(x_1, ..., x_t)$. In this paper we focus on the case in which $\mathcal{M}$ is an exponential family, which covers most standard learning settings, including those cited in the introduction.

Definition 1. *Let Θ be some subset of $\mathbb{R}^n$ and let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions on a sample space X. $\mathcal{M}$ is an* exponential family *(in the canonical form) if it is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form*

$$P_\theta(x) := exp\left(\theta^T \phi(x) - \psi(\theta)\right) r(x)$$

For some known (vector valued) functions $\phi(x), \psi(\theta)$ and $r(x)$; where $\theta^T \phi(x)$ is the standard inner product between θ and $\phi(x)$.

Most of the commonly used distributions form an exponential family or a subset of an exponential family (e.g. Gaussian, Multinomial, Poisson, Gamma, Beta). Furthermore, some non-i.i.d. models such as finite Markov models can be constructed as sequences of exponential families and maintain their essential properties so that our results are preserved (Grünwald, 2007).

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,[4] 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

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

definition, for all $A \subseteq \Theta$ we have

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

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 posteriors concentrate on the same parameter (or on a set of parameters with identical predictions) on the realized path. In well-specified learning problems, all posteriors concentrate on the true model almost surely, and Definition 2 is equivalent to the notion of *weak merging*, (Kalai and Lehrer, 1994).

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 member of the exponential family.*

A2: *the set of priors, $\mathcal{C}$, is compact;*

A3: *there is a subset $\Theta_0 := \Theta_0(\mathcal{C})$ of Θ such that each prior in $\mathcal{C}$ is continuous and positive on the compact closure of Θ_0;*

A4: *The interior of Θ_0 is nonempty, and the closure of Θ_0 is a compact subset of the interior of Θ.*

Condition **A1** limits our attention to densities that are measurable jointly in x and θ and regular enough to admit an accurate Taylor expansion of their K-L divergence. This assumption is stronger than condition i) of Berk (1966), and allows to go beyond the i.i.d. setting. **A2** is needed to ensure uniform convergence in the set of priors (Marinacci, 2002). **A3** requires that priors are not orthogonal on at least some subsets of parameters. If there were no set of parameters on which all priors agree it would be impossible for all posterior to converge to a common model. **A4**, together with **A3**, is a technical requirement to avoid singular behavior of the approximating function used in the Laplace method when the maximum likelihood parameter is near the boundary of Θ. Intuitively, if no parameter is ruled out a priori, Θ_0 has the same dimensionality as Θ but does not come arbitrarily close to the "boundary" of Θ, where, if Θ is unbounded, "boundary" points may be vectors with infinite components. Of particular interest is the subset of sequences such that the maximum likelihood estimator $\hat{\theta}(x^t)$ is in Θ_0 for all large t.

Definition 4. *Let $\hat{\theta}(x^t)$ be a maximum likelihood estimator at x^t :*

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

$\hat{S}(\mathcal{M}, \Theta_0) \subset X^\infty$ is the subset of sequences such that $\hat{\theta}(x^t)$ is — all $\hat{\theta}(x^t)$ are, if not unique — in Θ_0 for all large t.

If the maximum likelihood parameter is in Θ_0 for t large, then it is unique and the Laplace method can be used to show that the posterior of each prior is concentrated in a neighborhood of the maximum likelihood parameter; so, ambiguity fades away.

Theorem 1.
If the learning problem is regular, ambiguity fades away on all sequences in $\hat{S}(\mathcal{M}, \Theta_0) \subset X^\infty$.

Proof. See Appendix. □

Theorem 1 makes no references 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); 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 belongs to a well-behaved subset of Θ on which all priors are positive. Being agnostic agnostic about the

true distribution renders our approach particularly suited to discuss convergence in possibly misspecified learning environments.

A crucial question is how *large* is the set $\hat{S}(\mathcal{M}, \Theta_0)$. In learning problems in which the prior support is naturally rich, in that priors do not rule out specific parameter choices, $\hat{S}(\mathcal{M}, \Theta_0)$ contains most sequences. Specifically, it contains all sequences on which the maximum likelihood parameter does not come arbitrarily close to the "boundary" of Θ, where, if Θ is unbounded, boundary points may be vectors with infinite components. Furthermore, for Θ bounded, the requirement that $\hat{\theta}(x^t)$ is in Θ_0 is instrumental in the proof of Theorem 1, not substantial. Convergence to parameters on the boundary of Θ can be obtained using ad hoc methods. For instance, if the decision-maker is interested in learning parameters of the multinomial model and priors are defined on the whole simplex, ambiguity fades away in all sequences.

Corollary 1. *If $\mathcal{M}$ is the multinomial family (with arbitrary but finite Markov correlation structure), $\mathcal{C}$ satisfies* **A2** *and priors in $\mathcal{C}$ are strictly positive with full support, then ambiguity fades away in every sequence.*

Proof. If the maximum likelihood parameter is in the interior of Θ for most periods, Theorem 1 holds. If it converges to the boundary of Θ, the technique of Xie and Barron (2000) guarantees convergence of the priors. □

To get an intuition of the size of $\hat{S}(\mathcal{M}, \Theta_0)$, Theorem 1 and Corollary 1 guarantee that

- ambiguity fades away on all sequences with finite average if the Bayesian decision-maker believes that realisations are Gaussian with known variance and $\mathcal{C}$ is a compact set of not-degenerate Gaussian priors for the mean;
- ambiguity fades away on all sequences 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 head (i.e., $\mathcal{C} = \{Beta(\alpha, \beta),\ \alpha \in [a, b], \beta \in [c, d]\}$, with $[a, b], [c, d]$ strictly positive, finite intervals).

5 Appendix

In this appendix $\hat{\theta}_t := \hat{\theta}(x^t)$, and we make use of the K-L divergence.

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 of strict concavity of $-D(P_{\hat{\theta}_t}||P_\theta)$ this minimizer is unique when $\hat{\theta}(x^t)$ is in Θ_0. So, if $\hat{\theta}(x^t)$ is in Θ_0 for all large t, all priors concentrate on the same parameter and ambiguity fades away.

Proof of Theorem 1

Proof. $\mathcal{C}$ compact $\Rightarrow$ $arg\max_{\mu\in\mathcal{C}} \int_X dP^\mu(.|x^t)$ and $arg\min_{\mu\in\mathcal{C}} \int_X dP^\mu(.|x^t)$ exist. Thus, it suffices to show that if the learning problem is regular, then $\forall x^\infty \in \hat{S}(\mathcal{M},\Theta_0)$ 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)} - \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} - \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-1}} e^{-tD\left(P_{\hat{\theta}_{t-1}}||P_\theta\right)}g(\theta)\ d\theta} - \frac{e^{-tD\left(P_{\hat{\theta}_t}||P_\theta\right)}h(\theta)}{\int_{B_{t-1}} e^{-tD\left(P_{\hat{\theta}_{t-1}}||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}g_t^+}{\sqrt{tI_t^-}}}{\frac{\sqrt{2\pi}g_{t-1}^-}{\sqrt{(t-1)I_{t-1}^+}}} - \frac{\frac{\sqrt{2\pi}h_t^-}{\sqrt{tI_t^+}}}{\frac{\sqrt{2\pi}h_{t-1}^+)}{\sqrt{(t-1)I_{t-1}^-}}}\right|; \left|\frac{\frac{\sqrt{2\pi}g_t^-}{\sqrt{tI_t^+}}}{\frac{\sqrt{2\pi}g_{t-1}^+}{\sqrt{(t-1)I_{t-1}^-}}} - \frac{\frac{\sqrt{2\pi}h_t^+}{\sqrt{tI_t^-}}}{\frac{\sqrt{2\pi}h_{t-1}^-}{\sqrt{(t-1)I_{t-1}^+}}}\right|\right\}\right| dx \\
&\le^f \int_X \lim_{t\to\infty} P^+(x)\max\left\{\left|\frac{\frac{\sqrt{2\pi}g_t^+}{\sqrt{tI_t^-}}}{\frac{\sqrt{2\pi}g_{t-1}^-}{\sqrt{(t-1)I_{t-1}^+}}} - \frac{\frac{\sqrt{2\pi}h_t^-}{\sqrt{tI_t^+}}}{\frac{\sqrt{2\pi}h_{t-1}^+)}{\sqrt{(t-1)I_{t-1}^-}}}\right|; \left|\frac{\frac{\sqrt{2\pi}g_t^-}{\sqrt{tI_t^+}}}{\frac{\sqrt{2\pi}g_{t-1}^+}{\sqrt{(t-1)I_{t-1}^-}}} - \frac{\frac{\sqrt{2\pi}h_t^+}{\sqrt{tI_t^-}}}{\frac{\sqrt{2\pi}h_{t-1}^-}{\sqrt{(t-1)I_{t-1}^+}}}\right|\right\} dx \\
&=^g 0.
\end{aligned}
$$

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) $B_t \in \Theta_0$ (for t large for all $x^\infty \in \hat{S}(\mathcal{M},\Theta_0)$) is a neighbourhood of the maximum likelihood such that $diam(B_t) \to^{t\to\infty} 0$ at a rate slightly slower than $\sqrt{\frac{1}{t}}$.

d) By Lemma 1 **(i)**, $\int_{\Theta_0\setminus B_t} e^{-tD\left(P_{\hat{\theta}_t}||P_\theta\right)}h(\theta)\ d\theta \to 0$ exponentially fast and it can be ignored in the calculation of the limit.

e) By Lemma 1 **(ii)**, with $I_t^- = \inf_{\theta'\in B_t} I(\theta')$, $I_t^+ = \sup_{\theta'\in B_t} I(\theta')$, $g_t^- = \inf_{\theta'\in B_t} g(\theta')$, $g_t^+ = \sup_{\theta'\in B_t} g(\theta')$, $h_t^- = \inf_{\theta'\in B_t} h(\theta')$, $h_t^+ = \sup_{\theta'\in B_t} h(\theta')$.

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

g) By definition of $\hat{\theta}$, $\|\hat{\theta}_t - \hat{\theta}_{t-1}\| \to 0$, which ensures that $diam(B_t) \to^{t\to\infty} 0 \Rightarrow \sup_{\theta' \in B_t, \theta'' \in B_{t-1}} ||\theta' - \theta''|| \to 0$. So, continuity and strict positivity of $g(\cdot), h(\cdot)$, and $I(\cdot)$ in $\hat{\Theta} \supset B_t$ implies that,[5] $\frac{g_t^+}{g_t^-} \to 1$, $\frac{h_t^+}{h_t^-} \to 1$, $\frac{I_t^+}{I_t^-} \to 1$, $\frac{t}{t-1} \to 1$ uniformly and the limit follows $\forall x^\infty \in \hat{S}(\mathcal{M}, \Theta_0)$.

□

Lemma 1. *Let $\mathcal{M}$ be a member of the exponential family parameterized by Θ and μ a prior that satisfies **A3**, then, $\forall x^\infty \in \hat{S}(\mathcal{M}, \Theta_0)$,*

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

and, for t large, the following bounds holds uniformly when B_t is a neighbourhood 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: *$\exists k > 0 : \mathcal{I}_1 = \int_{\Theta_0\setminus B_t} e^{-tD\left(P_{\hat{\theta}_t}||P_\theta\right)}\mu(\theta)d\theta < e^{-kt^{2\alpha}}$.*

(ii) Second integral: *Let $\mathcal{I}_2 = \int_{B_t} e^{-tD\left(P_{\hat{\theta}_t}||P_\theta\right)}\mu(\theta)d\theta$; $I(\theta_t)$ be the Fisher information evaluated at the maximum likelihood parameter,[6] $I_t^- = \inf_{\theta' \in B_t} I(\theta')$, $I_t^+ = \sup_{\theta' \in B_t} I(\theta')$; and $\mu_t^- = \inf_{\theta' \in B_t} \mu(\theta')$, $\mu_t^+ = \sup_{\theta' \in B_t} \mu(\theta')$, then*

$$\frac{\mu_t^-}{\sqrt{tI_t^+}2\pi} \leq \mathcal{I}_2 \leq \frac{\mu_t^+}{\sqrt{tI_t^-}2\pi}.$$

Proof. For ease of exposition,[7] we focus on the case in which $\mathcal{M}$ is the Bernoulli family, so that $P_\theta = \theta$ and B_t can be chosen as follow: $B_t = \{\theta \in [\hat{\theta}_t - t^{-\frac{1}{2}+\alpha}, \hat{\theta}_t + t^{-\frac{1}{2}+\alpha}]\}$ with $0 < \alpha < \frac{1}{2}$. 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 $1/\sqrt{t}$. Because $x^\infty \in \hat{S}(\mathcal{M}, \Theta_0)$, B_t concentrates around $\hat{\theta}_t$. Because μ is continuous and positive on the compact closure of Θ_0, 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 a member of the exponential family; thus , $D(P_{\hat{\theta}_t}||P_\theta)$ can be exactly approximated in B_t as follows (see Grünwald, 2007, chapter 19):

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

for some $\theta^* \in B_t$ such that θ^* lies between θ and $\hat{\theta}_t$.

(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:

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

[5] $g(\cdot)$ and $h(\cdot)$, by (**A3**) on Θ_0; and $I(\cdot)$ because $\mathcal{M}$ is a member of the exponential family (**A1**).

[6] Which is positive definite because $\mathcal{M}$ is a member of the exponential family.

[7] The result generalizes to other members of the exponential family in the canonical form straightforwardly.

where

$$\min_{\theta\in\Theta\setminus B_t} D\left(P_{\hat{\theta}_t}||P_\theta\right) = \min_{\theta\in\{\hat{\theta}_t - t^{-1/2-\alpha}, \hat{\theta}_t - t^{-1/2+\alpha}\}} D\left(P_{\hat{\theta}_t}||P_\theta\right) \geq^{\text{By eq.2}} \frac{1}{2} t^{-1+2\alpha} \min_{\theta\in int(\Theta_0)} I(\theta),$$

so that, since $I(\theta)$ is continuous and > 0 for all $\theta \in \Theta_0$, and also $\int_{\Theta\setminus B_t} \mu(\theta)\ d\theta \leq 1$,

$$0 < \int_{\Theta\setminus B_t} e^{-tD\left(P_{\hat{\theta}_t}||P_\theta\right)} g(\theta)\ d\theta < \int_{\Theta\setminus B_t} e^{-t\left(\frac{1}{2}t^{-1+2\alpha} \min_{\theta\in int(\Theta_0)} I(\theta)\right)} g(\theta)\ d\theta < e^{-kt^{2\alpha}},$$

for $k = \frac{1}{2} \min_{\theta\in int(\Theta_0)} I(\theta) > 0$.

(ii) Second integral: by Equation 2,

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

where θ' depends on θ. Using the notation defined above, we get

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

Performing the substitutions $z = (\hat{\theta}_t - \theta)\sqrt{tI_t^+}$ on the left integral and $z = (\hat{\theta}_t - \theta)\sqrt{tI_t^-}$ on the right integral, we get

$$\frac{g_t^-}{\sqrt{tI_t^+}} \int_{|z|<t^\alpha\sqrt{I_t^-}} e^{-\frac{1}{2}z^2} dz \leq \mathcal{I}_2 \leq \frac{g_t^+}{\sqrt{tI_t^-}} \int_{|z|<t^\alpha\sqrt{I_t^+}} e^{-\frac{1}{2}z^2} dz,$$

and recognize these integrals as standard 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, so that they both converge to $\sqrt{2\pi}$.

This approximation holds uniformly for all $x^\infty \in \hat{S}(\mathcal{M}, \Theta_0)$ because *i*) the bound on $\mathcal{I}_1$ does not depend on x^t, and *ii*) convergence of $\mathcal{I}_2$ is uniform because **A3** and **A4**. guarantee that $g(\theta)$ and $I(\theta)$ are continuous, positive functions of θ over the compact closure of Θ_0. □

References

Augustin, P. and Izhakian, Y. Y. (2019). Ambiguity, volatility, and credit risk. *Review of Financial Studies (forthcoming)*.

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. *The Annals of Mathematical Statistics*, 37(1):51–58.

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of bayes methods. *Information Theory, IEEE Transactions on*, 36(3):453–471.

Doob, J. L. (1949). *Application of the theory of martingales*. Colloques Internationaux du Centre National de la Recherche Scientifique Paris.

Epstein, L. G. and Schneider, M. (2007). Learning under ambiguity. *The Review of Economic Studies*, 74(4):1275–1303.

Garlappi, L., Uppal, R., and Wang, T. (2006). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *The Review of Financial Studies*, 20(1):41–81.

Gilboa, I. and Marinacci, M. (2016). Ambiguity and the bayesian paradigm. In *Readings in formal epistemology*, pages 385–439. Springer.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Hansen, L. P., Sargent, T. J., and Tallarini Jr, T. D. (1999). Robust permanent income and pricing. *Review of Economic studies*, pages 873–907.

Kajii, A. and Ui, T. (2006). Agreeable bets with multiple priors. *Journal of Economic Theory*, 128(1):299–305.

Kalai, E. and Lehrer, E. (1994). Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23(1):73–86.

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

Marinacci, M. and Massari, F. (2019). Learning from ambiguous and misspecified models. *Journal of Mathematical Economics*, 84:144–149.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

Nishimura, K. G. and Ozaki, H. (2004). Search and knightian uncertainty. *Journal of Economic Theory*, 119(2):299–333.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Sharpe, W. F. (1970). *Portfolio theory and capital markets*, volume 217. McGraw-Hill New York.

Werner, J. (2019). Speculative trade under ambiguity. Technical report, mimeo.

Xie, Q. and Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445.