Learning from ambiguous and misspecified models[*]

Massimo Marinacci

Department of Decision Sciences and IGIER, Bocconi University

Filippo Massari

School of Banking and Finance, UNSW

April 13, 2018

Abstract

We model inter-temporal ambiguity as the scenario in which a Bayesian learner holds more than one prior distribution over a set of models and provide necessary and sufficient conditions for ambiguity to fade away because of learning. Our condition applies to most learning environments: iid and non-idd model-classes, well-specified and misspecified model-classes/prior support pairs. It shows that a Bayesian agent does not suffer from long-run ambiguity if and only if the data support a unique model.

Keywords: Ambiguity, Learning.

JEL Classification: D81, D83, C11

1 Introduction

Let $\mathcal{M}$ be a family of models and $\mathcal{C}$ a set of prior distributions on it. If $\mathcal{C}$ contains more than one prior distribution, its multiplicity represents the *a priori* ambiguity

[*] We thank Werner Ploberger for his comments. Massimo Marinacci acknowledges the financial support of the European Research Council (advanced grant INDIMACRO).

perceived by a Bayesian decision maker (DM). This setting has been used to highlight the interaction between learning and ambiguity.[1]

Marinacci (2002) formalizes the intuition that if a DM observes repeated draws (with replacement) from the same ambiguous urn, ambiguity fades away over time because he eventually learns the true composition. If the learning problem is well-specified — in the sense that the true probability belongs to the model-class/prior support pair adopted by the DM — ambiguity fades away because all posterior distributions converge to a Dirac distribution on the true model.

Here, we generalize the result in Marinacci (2002) to the case in which the DM does not learn the true probability because his prior view of the world is incorrect — that is, when the learning problem is misspecified in the sense that the model-class/prior support pair does not contain the true model/parameter. We show that ambiguity fades away if and only if the data clearly designates a unique most accurate model (or a set of models with equivalent predictions), a condition that is always satisfied in well-specified learning problems and in most cases of misspecification. In a nutshell, ambiguity fades away in all cases in which the empirical evidence eventually dominates the effect of heterogeneity in the prior distributions. On the contrary, ambiguity persists in those sequences in which two or more models with different predictions have comparable likelihood infinitely often. When this happens, the posteriors are "split" between these models with weights that depend on the priors, and the DM perceives ambiguity.

Our key contribution is to formalize necessary and sufficient conditions for the posteriors obtained from all priors to concentrate on the same model. Our findings rely on and generalize standard results in statical learning theory. With a unique prior, a sufficient condition for the Bayesian posterior to concentrate on the true model (consistency) is that the prior μ attaches a positive mass to the true parameter(s) (Doob, 1949; Freedman, 1963). In a multiple priors setting, this result continues to hold: if all priors give positive mass to the true model, then all posteriors concentrate on it and ambiguity fades away (Marinacci, 2002). On the other hand, in an iid setting

[1] Epstein and Schneider (2003) provides an axiomatization of prior-by-prior updating which requires the process of conditional preferences to be dynamically consistent. Because we are focusing on one-step-ahead decisions, the consistency issue has no bite in our setting.

and if the true parameter set does not belong to the prior support, the posterior concentrates on the model that is the closest in terms of K-L divergence to the truth if it is unique (Berk, 1966; White, 1982). In a multiple priors setting, this result suggests that if the minimizer of the K-L divergence, P^*, is unique and all priors give it a positive weight, then ambiguity fades away because all posteriors concentrate on P^*. Theorem 2 proves this conjecture and generalizes it to the non-iid setting, while Theorem 1 provides a condition for the posteriors derived from all priors to concentrate on the same model (on a set of models with identical predictions) that is both necessary and sufficient.

2 Discussion

We prove that a Bayesian agent with multiple priors does not suffer from long-run ambiguity in all those cases in which the data support a unique model (or a set of models with identical predictions). How common are these situations? A precise answer to this question is hard to give because it depends on the true probability measure, which is typically unknown. If all measures in $\mathcal{M}$ and the true model are iid, ambiguity fades away on a set of parameters that has Lebesgue measure 1 (as an implication of Theorem 2), thus suggesting that ambiguity should be the exception, rather than the norm. However, we are cautious about concluding that ambiguity typically fades away in real world situations because models and parameters are hardly iid and chosen at random. For example, consider the standard problem of predicting stock market returns. Several models have been proposed and, to this date, it is not clear which model is the closest to the truth — there is no definite statistical test that favors a unique model over another. Because the empirical evidence does not support a unique model, an investor with a set of priors on available models of stock market returns suffers ambiguity despite the large amount of available financial data.

Our condition for ambiguity to persist in the long run is harder to satisfy than conditions based on the multiple-likelihood setting (e.g., Epstein and Schneider, 2007; Epstein and Seo, 2015). Our, multiple-prior, model describes a DM who is uncertain about the a priori probability of each model in the support but updates each model in a unique way. On the other hand, the multiple-likelihoods model describes a DM who believes that signals have multiple, hence uncertain, interpretations. Such signals can generate ambiguity even where none is present a priori. Learning models, that accommodate such a possibility generate posterior sets different from those defined in this paper, and they lead to different results regarding if/when ambiguity fades away.

3 Probabilities

We consider a family of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ with a finite 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 finite observation space X with representative element x and σ-algebra Σ.[2] 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, and the likelihood that model P_θ attaches to the partial sequence $(x_1, ..., x_t)$. 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 $P_\mu \in \Delta(\Theta \times X^\infty)$. By definition, for all $A \subseteq \Theta$ we have that:

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

We denote by $\mu(.|x^t) \in \Delta\Theta$ the usual posterior given the observations x^t,[3] 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 $A \subseteq \Theta$:

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

[2] In the rest of the paper, we focus on the case of extractions from ambiguous urns. However, this setting can accommodate most prediction tasks with minor changes which do not affect our results. For example, x could be a vector of stock market returns, $\mathcal{M}$ a set of regression models with parameters to be estimated and $\mathcal{C}$ a (meta)prior over the set of regression.

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

4 Decisions

Let C be the space of consequences on which the DM has a bounded utility function $u : C \to \mathbb{R}$. An act $f : X \to C$ is a $\Sigma-$measurable map that associates a consequence to each observation in X. We are considering one step ahead acts. The decision criterion adopted by the DM depends on the quality of his prior information. For illustrative purposes, we briefly provide examples of the DM's decision criterion when facing *risk*, *unambiguous uncertainty*, and *ambiguity*.

Suppose there is an urn with 3 balls, each of which is either white, x_W, or red, x_R. Suppose the DM chooses a color and draws a ball from the urn. If this ball matches the DM's color, he wins \$100. Otherwise, he gets nothing. The consequence space is $C = \{\$0, \$100\}$, the observation space $X = \{x_R, x_W\}$, and the DM can choose between two acts: f_R, he bets on a red ball; and f_W, he bets on a white ball. The following table summarizes this decision problem:

$$\begin{array}{ccc} & x_R & x_W \\ f_R & \$100 & 0 \\ f_W & 0 & \$100 \end{array} \tag{1}$$

Finally, θ is the fraction of white balls in the urn, so that $\Theta = \{0, 1/3, 2/3, 1\}$. If draws are made with replacement from the same urn, $\mathcal{M}$ is the iid Bernoulli distribution family with parameter set Θ.

- **Scenario 1: Risky Urns.** The DM knows the true composition of the urn θ_0 (e.g., he knows that it contains exactly two white balls). In this case, the DM's choice criterion is, for every act f, given by:

$$\int_X u(f(x))dP_{\theta_0}.$$

- **Scenario 2: Bayesian Urns.** The DM does not know the composition of the urn but has enough prior information to uniquely pin down a prior distribution μ on the set of possible compositions Θ. That is, $\mathcal{C}$ is a singleton. For example, the DM might believe that all the compositions of the urn are equally likely. Unlike the previous case, the DM's choice criterion now changes over time because of learning. In the first period, the DM's choice criterion is, for every act f, given by:

$$\int_{\Theta}\left[\int_X u(f(x_1))dP_{\theta}\right]d\mu = \int_X u(f(x_1))dP_{\mu}(x_1|\emptyset).$$

Subsequently, as the DM incorporates past realization, x^t, to his prior distribution using Bayes' rule, his choice criterion becomes:

$$\int_{\Theta}\left[\int_X u(f(x_{t+1}))dP_{\theta}\right]d\mu(.|x^t) = \int_X u(f(x_{t+1}))dP_{\mu}(x_{t+1}|x^t).$$

- **Scenario 3: Ambiguous Urns.** The DM does not know the composition of the urn and does not have enough prior information to uniquely pin down a distribution on the set of possible compositions of the urn. That is, $\mathcal{C}$ is not a singleton. For example, the DM might only know that every composition has at least a $\frac{1}{10}$ probability to be the correct one: $\mathcal{C} := \left\{\mu \in \Delta : \forall\theta \in \Theta, \mu(\theta) \geq \frac{1}{10}\right\}$. In evaluating an act in this scenario, the DM has to use a set criterion. In the first period, the DM's set criterion is, for every act f, given by:

$$\left\{\int_X u(f(x_1))dP_{\mu}(x_1|\emptyset) : \mu \in \mathcal{C}\right\}.$$

  Subsequently, as the DM incorporates past realizations using Bayes' rule, his choice criterion becomes:

$$\left\{\int_X u(f(x_{t+1}))dP_{\mu}(x_{t+1}|x^t) : \mu \in \mathcal{C}\right\}.$$

  Possible summaries of this set criteria are the infimum and supremum:

$$\sup_{\mu\in\mathcal{C}}\int_X u(f(x_{t+1}))dP_{\mu}(x_{t+1}|x^t) \quad ; \quad \inf_{\mu\in\mathcal{C}}\int_X u(f(x_{t+1}))dP_{\mu}(x_{t+1}|x^t).$$

5 Long-run ambiguity

As in Marinacci (2002), we consider the difference between the DM'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 a DM perceives in evaluating an act f. We are ultimately interested in verifying whether this quantity converges to 0 as the number of past observations goes to infinity and each prior gets independently updated using Bayes' rule. A tight sufficient condition for the most conservative and the least conservative expected utility to coincide is to require that the posteriors calculated from all priors in $\mathcal{C}$ eventually coincide (see Lemma 1 in Appendix).[4] We say that

Definition 1. *Ambiguity fades away at path $x^\infty \in X^\infty$ if,*

$$\lim_{t\to\infty}\left[\sup_{\mu',\mu''\in\mathcal{C}}\int_X \left|dP_{\mu''}(x_{t+1}|x^t) - dP_{\mu'}(x_{t+1}|x^t)\right|\right] = 0; \tag{2}$$

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

Definition 1 requires that all posteriors concentrate on the same model (or on a set of models with identical predictions) on the realized path. Unlike the definition proposed by Marinacci (2002) — which requires all the posteriors to converge to a Dirac measure on the true model on a set of sequences of true measure 1 — ours does not assume an iid structure, and it does not depend on the true model. Thus, it can be used to discuss long-run ambiguity when the model class support contains models with learning, a time series structure, or is misspecified. In those cases in which all posteriors concentrate on the true model, our definition is equivalent to the notion of *weak merging* (Lehrer and Smorodinsky, 1996).

[4] This condition fails to be necessary only in those knife-edge cases in which the posteriors do not concentrate on a unique model but the expected utilities of the preferred act calculated from all posteriors coincide.

6 Main result

In this section, we present our necessary and sufficient condition for ambiguity to fade away. The driving force of our result is the observation that the key component of Bayesian learning is the existence of a unique most accurate model, rather than the true model belonging to the prior support. For instance, Berk (1966) shows that if all models in the support and the truth are iid, then the posterior obtained from a unique prior eventually assigns probability 1 to the set of parameters that minimize the K-L divergence from the truth, if unique. Here, we generalize Berk (1966)'s result to the case of multi-prior, non-iid setting and provide a condition that is both necessary and sufficient for all posteriors to concentrate on models that deliver the same predictions. Let's start by formalizing an appropriate generalization of the notion of *unique most accurate model.*

Definition 2. *Given a path $x^\infty \in X^\infty$ and a family of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$. We say that $\hat{\theta} := \hat{\theta}(x^\infty, \Theta)$ is a* strong maximum likelihood (SML) *model if $\hat{\theta} \in \Theta$ and*

$$\forall \theta \in \Theta, \lim_{t \to \infty} \frac{P_\theta(x^t)}{P_{\hat{\theta}}(x^t)} \in [0, \infty) \text{ exists};$$

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

Theorem 1 shows that the existence of a SML is a necessary and sufficient condition for ambiguity to fade away.

Theorem 1. *Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ be a family of models and $\mathcal{C}$ a compact set of non-degenerate prior distributions on Θ, ambiguity fades away at path x^∞ if and only if $\hat{\theta}(x^\infty, \Theta)$ exists.*

Suppose $\hat{\theta}$ is unique, by Definition 2 this implies that the SML model is the model whose likelihood converges to zero at the slowest rate — i.e., $\forall \theta \neq \hat{\theta}, \lim_{t \to \infty} \frac{P_\theta(x^t)}{P_{\hat{\theta}}(x^t)} = 0$. Therefore, ambiguity vanishes because eventually the posteriors calculated from all priors attach unitary weight to model SML. Otherwise, if Θ contains more than one SML model, all SML models must eventually deliver identical predictions on x^∞ because the limit exists. Ambiguity vanishes because eventually the posteriors calculated from all priors attach positive weights only to SML models, and all SML models deliver identical posterior distributions.

An alternative condition for ambiguity to fade away can be obtained by noticing that a sufficient condition for the existence of a unique SML model is the presence of a unique model with the lowest average K-L divergence.

Definition 3. *The average K-L divergence from P_θ to the true probability P_{θ_0} is*

$$\bar{D}(P_{\theta_0}||P_\theta) := \lim_{t\to\infty} \frac{1}{t} E_{P_{\theta_0}}\left[\ln \frac{P_{\theta_0}(x^t)}{P_\theta(x^t)}\right].$$

This approach delivers a sufficient condition for ambiguity to fade away P_{θ_0}-a.s.[5] which generalizes Berk (1966)'s results to the non-iid setting and includes Marinacci (2002)'s condition as a special case. When $\mathcal{C}$ is a singleton, if all models in $\mathcal{M}$ and the true measure are iid, Berk (1966)'s result follows from Theorem 2 because the average K-L divergence coincides with the K-L divergence P_{θ_0}-a.s. as an implication of the Strong Law of Large Numbers. Whereas, if all models in $\mathcal{M}$ are iid and the truth belongs to $\mathcal{M}$, Marinacci (2002)'s condition follows because the true model is the unique maximizer of the K-L divergence.

Theorem 2. *Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ be a family of models and $\mathcal{C}$ a compact set of non-degenerate priors on Θ, ambiguity fades away P_{θ_0}-a.s. if $\operatorname*{argmin}_{\theta\in\Theta} \bar{D}(P_{\theta_0}||P_\theta)$ exists and is unique.*

Unlike the condition of Theorem 1, which depends only on the properties of the sequence of realizations, Theorem 2's condition requires an apriori knowledge of the true probability distribution —to calculate the K-L divergence. In a nutshell, the difference between the two conditions is as follows: Theorem 1 tells us that ambiguity persists if and only if the data is inconclusive, while Theorem 2 tells us that ambiguity persists if the true probability generates inconclusive data P_{θ_0}-a.s.. We prefer the former because it directly links ambiguity to properties of observables, rather than the true model, which is hardly known in practice and unobservable. For example, while a sequence of

[5] The necessary part of Theorem 2's condition is lost because the existence of two models such that $\bar{D}(P_{\theta_0}||P_{\theta_i}) = \bar{D}(P_{\theta_0}||P_{\theta_j})$ does not rule out the existence of a unique SML model (Massari, 2017), nor does it imply that P_{θ_i}'s predictions eventually coincide with P_{θ_j}'s.

stock market returns is observable, their distribution is not.

We conclude by presenting scenarios illustrating the use of Theorem 1. Scenarios 4, 5 and 7 show cases in which ambiguity fades away because the truth generates sequences that support only one model among those believed possible by the DM. Conversely, scenarios 6 and 8 shows that ambiguity persists in the long run when the truth generates sequences which equally endorse at least two models with different predictions in the prior support infinitely often. That is, if and only if prior distributions affect next period predictions even after an arbitrarily large sample. The iid assumption for models in $\mathcal{M}$, the repeated urn setting, and the simplicity of the deterministic sequences are chosen for illustrative purposes. More complicated examples can be easily constructed.

Suppose a DM confronted with decision problem 1 subjectively believes that he is facing iid realizations from an ambiguous urn with three balls, two of which have the same color. In our notation, he believes that $\mathcal{M}$ is the class of iid Bernoulli distributions with possible parameters $\Theta = \{\frac{1}{3}, \frac{2}{3}\}$. His prior information is accurate enough to reduce ambiguity to only two possible priors on the composition of the urn: $\mathcal{C} = \{\mu'(\theta), \mu''(\theta)\}$. These are, $\mu'(\theta) = \{\mu'(\frac{1}{3}) = \frac{1}{2}, \mu'(\frac{2}{3}){=}\frac{1}{2}\}$ and $\mu''(\theta){=}\{\mu''(\frac{1}{3}){=}\frac{1}{4}, \mu''(\frac{2}{3}){=}\frac{3}{4}\}$.

- **Scenario 4: Well-specified model and ambiguity fades away.** Draws are indeed iid from an urn whose composition is $\theta_0 = \frac{2}{3}$. Because the learning problem is well-specified ($\theta_0 \in \Theta$), by the Strong Law of Large Numbers θ_0 is the SML model P_{θ_0}-a.s.. Thus, Theorem 1 (and Marinacci (2002)) implies that ambiguity fades away P_{θ_0}-a.s..
- **Scenario 5: Incorrect $\mathcal{M}$ and ambiguity fades away.** Draws are not iid, as the DM incorrectly believes. Instead, the urn is secretly changed before every draw to deliver the deterministic sequence $x^\infty := \{W, W, R, W, W, R, ...\}$. Because the frequency of W converges to $\frac{2}{3}$, then $\hat{\theta} = \frac{2}{3}$ is the SML parameter in Θ. By Theorem 1, all posteriors concentrate on $\hat{\theta}$ and ambiguity fades away. Although the DM fails to realize that draws are not iid, he successfully learns the best

parameter in Θ and ambiguity fades away.

- **Scenario 6: Incorrect $\mathcal{M}$ and ambiguity persists.** Draws are not iid, as the DM incorrectly believes. Instead, he is facing the deterministic sequence $x^{\infty} := \{W, R, W, R, ...\}$. Because of symmetry around $\frac{1}{2}$, there is not a strong maximum likelihood in Θ and, by Theorem 1, ambiguity does not fade away. Intuitively, in every even period $P_{\theta=\frac{1}{3}}$ and $P_{\theta=\frac{2}{3}}$ have identical likelihood. Therefore, each odd period prediction obtained from priors μ' and μ'' coincides with their first period prediction. Because μ''s and μ'''s first period predictions differ, their predictions differ for every odd period, and ambiguity does not fade away.[6]

- **Scenario 7: Correct $\mathcal{M}$, $\theta_0 \notin \Theta$ and ambiguity fades away.** Draws are iid from an urn whose composition is $\theta_0 = \frac{3}{5}$. In this case, $\mathcal{M}$ is correctly specified because draws are indeed iid, but the learning problem is misspecified because Θ does not contain the true parameter: $\theta_0 \notin \Theta$. It is easy to verify that $\hat{\theta} = \frac{2}{3}$ is the strong maximum likelihood P_{θ_0}-a.s..[7] Thus, Theorem 1 implies that both posteriors concentrate on $\hat{\theta}$. Although the DM cannot learn the true model, ambiguity fades away because the data clearly indicates which model is the most accurate.

- **Scenario 8: Correct $\mathcal{M}$, $\theta_0 \notin \Theta$ and ambiguity persists.** Draws are iid from an urn whose composition is $\theta_0 = \frac{1}{2}$. In this case, $\mathcal{M}$ is correctly specified because draws are iid, but Θ is not, since $\theta_0 \notin \Theta$. Because of symmetry around $\frac{1}{2}$, there is not a strong maximum likelihood in Θ P-a.s. and, by Theorem 1, ambiguity persists P-a.s.. Intuitively, $P_{\theta=\frac{2}{3}}$ and $P_{\theta=\frac{1}{3}}$ can be shown to have identical likelihood infinitely often P_{θ_0}-a.s. (Massari, 2013). When this happens μ''s and

[6] It is straightforward to verify that $P_{\mu'}(x_W|x^t) \neq P_{\mu''}(x_W|x^t)$ for every t even:

$$P_{\mu'}(x_W|x^t) = \frac{1}{2}\frac{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{3}}{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{2}+\left(\frac{1}{2}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{2}} + \frac{1}{2}\frac{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{2}{3}}{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{2}+\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{2}} = \frac{1}{2}$$

$$P_{\mu''}(x_W|x^t) = \frac{1}{4}\frac{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{3}}{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{4}+\left(\frac{3}{4}\frac{2}{3}\right)^{\frac{t}{2}}\frac{3}{4}} + \frac{3}{4}\frac{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{2}{3}}{\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{1}{4}+\left(\frac{1}{3}\frac{2}{3}\right)^{\frac{t}{2}}\frac{3}{4}} = \frac{7}{12}$$

[7] By the Strong Law of Large Numbers, $\lim_{t\to\infty}\frac{P_{\theta=\frac{1}{3}}(x^t)}{P_{\theta=\frac{2}{3}}(x^t)} =^{P\text{-a.s.}} \lim_{t\to\infty}\left(\frac{\left(\frac{1}{3}\right)^{\frac{3}{5}}\left(\frac{2}{3}\right)^{\frac{2}{5}}}{\left(\frac{2}{3}\right)^{\frac{3}{5}}\left(\frac{1}{3}\right)^{\frac{2}{5}}}\right)^{t} = 0.$

μ'''s predictions differ and the DM suffers ambiguity, by the same argument used in Scenario 5.

7 Conclusion

In a multiple prior setting, ambiguity fades away if and only if the empirical evidence supports a unique model. Learning the true model is not a necessary condition for ambiguity to fade away.

8 Appendix

In this appendix,

- Given two functions, $f(.)$ and $g(.)$, $f(x) = o(g(x))$, abbreviates $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$;
- $\hat{\theta}_t := \hat{\theta}(x^t)$ is the maximum likelihood model on the partial history x^t;
- $D\left(P_{\hat{\theta}_t}||P_\theta\right):=E_{P_{\hat{\theta}_t}} \ln \frac{P_{\hat{\theta}_t}(x)}{P_\theta(x)}$ is the K-L divergence from P_θ to $P_{\hat{\theta}_t}$.

Proof of Theorem 1

Proof. Because $\mathcal{C}$ is compact, then $\operatorname*{argmax}_{\mu',\mu''\in\mathcal{C}} \lim_{t\to\infty} \int_X \left|dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)\right|$ is non empty. Thus, it suffices to prove that

$$\exists SML \Leftrightarrow \forall \mu', \mu'' \in \mathcal{C}, \lim_{t\to\infty} \int_X \left|dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)\right| = 0.$$

- Let start by analyzing the case in which the SML, $\hat{\theta}$, is unique.

$$\lim_{t\to\infty}\int_X \left|dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)\right|$$

$$:= \lim_{t\to\infty}\int_X \left|\sum_\theta P_\theta(x)\frac{P_\theta(x^t))\mu''(\theta)}{\sum_\theta P_\theta(x^t))\mu''(\theta)} - \sum_\theta P_\theta(x)\frac{P_\theta(x^t))\mu'(\theta)}{\sum_\theta P_\theta(x^t))\mu'(\theta)}\right| dx$$

$$=^a \int_X \lim_{t\to\infty}\left|\frac{P_{\hat{\theta}}(x)}{1+\sum\limits_{\theta\neq\hat{\theta}}\frac{P_\theta(x^t))\mu''(\theta)}{P_{\hat{\theta}}(x^t)\mu''(\hat{\theta})}} + \frac{\sum\limits_{\theta\neq\hat{\theta}} P_\theta(x)\frac{P_\theta(x^t))\mu''(\theta)}{P_{\hat{\theta}}(x^t)\mu''(\hat{\theta})}}{1+\sum\limits_{\theta\neq\hat{\theta}}\frac{P_\theta(x^t))\mu''(\theta)}{P_{\hat{\theta}}(x^t)\mu''(\hat{\theta})}} + \right.$$

$$\left. - \frac{P_{\hat{\theta}}(x)}{1+\sum\limits_{\theta\neq\hat{\theta}}\frac{P_\theta(x^t))\mu'(\theta)}{P_{\hat{\theta}}(x^t)\mu'(\hat{\theta})}} + \frac{\sum\limits_{\theta\neq\hat{\theta}} P_\theta(x)\frac{P_\theta(x^t))\mu'(\theta)}{P_{\hat{\theta}}(x^t)\mu'(\hat{\theta})}}{1+\sum\limits_{\theta\neq\hat{\theta}}\frac{P_\theta(x^t))\mu'(\theta)}{P_{\hat{\theta}}(x^t)\mu'(\hat{\theta})}}\right| dx$$

$$=^b \int_X \left|\frac{P_{\hat{\theta}}(x)}{1+o(1)} + o(1) - \frac{P_{\hat{\theta}}(x)}{1+o(1)} - o(1)\right| dx \quad \text{, if and only if } \hat{\theta} \text{ is SML, by definition;}$$

$$= 0$$

a : The Lebesgue's Dominated Convergence Theorem allows exchanging integral and limit signs (Williams, 1991).[8]

- Multiple SML.
Let $\hat{\theta}$ be a SML, note that all models, $\bar{\theta} \in \Theta$, that satisfy the condition $\lim_{t\to\infty}\frac{P_{\bar{\theta}}(x^t)}{P_{\hat{\theta}}(x^t)} > 0$ are also SML and must eventually deliver the same prediction $\bar{P}(x)$ — because the limit exists. The result follows substituting $\bar{P}$ and $\bar{\mu} = \sum_{\mu(\bar{\theta})}\mu(\theta)$ for $P_{\hat{\theta}}$ and $\hat{\mu}$ in (b), respectively.

□

Proof of Theorem 2

[8] Let $\{r_t(x)\}_{t=1}^{\infty} := \{|P_{\mu''}(x|x^t) - P_{\mu'}(x|x^t)|\}_{t=1}^{\infty}$ and note that $|r_1|, |r_2|$... are bounded above.

Proof. Let $\hat{\theta}$ be the unique $\underset{\theta\in\Theta}{\operatorname{argmin}}\, \bar{D}(P_{\theta_0}||P_\theta)$. Thus $\exists\epsilon > 0$:

$$
\begin{aligned}
&\forall\theta\in\Theta\setminus\hat{\theta},\ \bar{D}(P_{\theta_0}||P_{\hat{\theta}}) < \bar{D}(P_{\theta_0}||P_\theta) - \epsilon \\
\Leftrightarrow\ &\forall\theta\in\Theta\setminus\hat{\theta}, \lim_{t\to\infty}\frac{1}{t}E_{P_{\theta_0}}\left[\ln\frac{P_{\theta_0}(x^t)}{P_{\hat{\theta}}(x^t)}\right] - \lim_{t\to\infty}\frac{1}{t}E_{P_{\theta_0}}\left[\ln\frac{P_{\theta_0}(x^t)}{P_\theta(x^t)}\right] < -\epsilon \\
\Leftrightarrow^a\ &\forall\theta\in\Theta\setminus\hat{\theta}, \lim_{t\to\infty}E_{P_{\theta_0}}\left[\frac{1}{t}\sum_{\tau=1}^{t}E_{P_{\theta_0}(.|x^{\tau-1})}\ln\frac{P_{\theta_0}(.|x^{\tau-1})}{P_{\hat{\theta}}(.|x^{\tau-1})} - \frac{1}{t}\sum_{\tau=1}^{t}E_{P_{\theta_0}(.|x^{\tau-1})}\ln\frac{P_{\theta_0}(.|x^{\tau-1})}{P_\theta(.|x^{\tau-1})}\right] < -\epsilon \\
\Leftrightarrow^b\ &\forall\theta\in\Theta\setminus\hat{\theta}, \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\ln\frac{P_{\theta_0}(x_\tau|x^{\tau-1})}{P_{\hat{\theta}}(x_\tau|x^{\tau-1})} - \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\ln\frac{P_{\theta_0}(x_\tau|x^{\tau-1})}{P_\theta(x_\tau|x^{\tau-1})} < -\epsilon \quad P_{\theta_0}\text{-a.s.} \\
\Rightarrow\ &\forall\theta\in\Theta\setminus\hat{\theta},\ \lim_{t\to\infty}\sum_{\tau=1}^{t}\ln\frac{P_\theta(x_\tau|x^{\tau-1})}{P_{\hat{\theta}}(x_\tau|x^{\tau-1})} = -\infty \quad P_{\theta_0}\text{-a.s.} \\
\Leftrightarrow\ &\forall\theta\in\Theta\setminus\hat{\theta}, \lim_{t\to\infty}\frac{P_\theta(x^t)}{P_{\hat{\theta}}(x^t)} = -\infty \quad P_{\theta_0}\text{-a.s.} \\
\Leftrightarrow\ &\ \hat{\theta}\text{ is SML } P_{\theta_0}\text{-a.s.}
\end{aligned}
$$

a) Telescoping the log and using the tower property of expectation.
b) The Strong Law of Large Numbers for Martingale Differences (Williams, 1991) allows substituting the limit average sum of conditional expected values with the limit average sum of realized values P-a.s.:

$$\forall\theta\in\Theta\ \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\left(E_{P_{\theta_0}(.|x^{\tau-1})}\left[\ln\frac{P_{\theta_0}(.|x^{\tau-1})}{P_\theta(.|x^{\tau-1})}\right] - \ln\frac{P_{\theta_0}(x_\tau|x^{\tau-1})}{P_\theta(x_\tau|x^{\tau-1})}\right) = 0 \ \ P_{\theta_0}\text{-a.s.}$$

□

Lemma 1. *Let μ' and μ'' be two prior on Θ, if* u *is bounded, then L1 convergence of the posteriors derived from μ' and μ'' implies convergence in expected utilities.*

$$\lim_{t\to\infty}\int_X |dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)| = 0 \Rightarrow \lim_{t\to\infty}\left[\int_X u(f(x))dP_{\mu''}(.|x^t) - \int_X u(f(x))dP_{\mu'}(.|x^t)\right] = 0$$

Proof.

$$
\begin{aligned}
&\lim_{t\to\infty}\int_X |dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)| = 0 \\
\Rightarrow &\lim_{t\to\infty}A_t^1 = \lim_{t\to\infty}\max_{x\in X}|u(f(x))|\int_X \left|dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)\right| = 0 \qquad \text{because } |u(f(x))| < \infty; \\
\Rightarrow &\lim_{t\to\infty}A_t^2 = \lim_{t\to\infty}\int_X |u(f(x))|\left|dP_{\mu''}(x|x^t) - dP_{\mu'}(x|x^t)\right| = 0 \qquad \text{because } \forall t, A_t^1 \geq A_t^2 \geq 0; \\
\Rightarrow &\lim_{t\to\infty}A_t^3 = \lim_{t\to\infty}\left|\int_X u(f(x))dP_{\mu''}(.|x^t) - \int_X u(f(x))dP_{\mu'}(.|x^t)\right| = 0 \quad \text{because } \forall t, A_t^2 \geq A_t^3 \geq 0; \\
\Rightarrow &\lim_{t\to\infty}\left[\int_X u(f(x))dP_{\mu''}(.|x^t) - \int_X u(f(x))dP_{\mu'}(.|x^t)\right] = 0.
\end{aligned}
$$

□

References

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

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. (2003). Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31.

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

Epstein, L. G. and Seo, K. (2015). Exchangeable capacities, parameters and incomplete theories. *Journal of Economic Theory*, 157:879–917.

Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403.

Lehrer, E. and Smorodinsky, R. (1996). Compatible measures and merging. *Mathematics of Operations Research*, 21(3):697–706.

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

Massari, F. (2013). Comment on 'if you're so smart, why aren't you rich? belief selection in complete and incomplete markets'. *Econometrica*, 81(2):849–851.

Massari, F. (2017). Markets with heterogeneous beliefs: A necessary and sufficient condition for a trader to vanish. *Journal of Economic Dynamics and Control*, 78:190–205.

Siniscalchi, M. (2011). Dynamic choice under ambiguity. *Theoretical Economics*, 6(3):379–421.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities.* Taylor & Francis.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Williams, D. (1991). *Probability with martingales.* Cambridge University Press.