Entropified Berk-Nash Equilibrium

Filippo Massari Jonathan Newton

Bocconi University Kyoto University

October 31, 2019

Abstract

Esponda and Pouzo (2016) propose Berk-Nash equilibrium as a solution concept for games that are misspecified in that it is impossible for players to learn the true probability distribution over outcomes. The beliefs that support Berk-Nash equilibrium are, for each player, the learning outcome of Bayes' rule. However, under misspecification, Bayes' rule might not converge to the model that leads to actions with the highest objective payoff among the models subjectively admitted by the player. From an evolutionary perspective, this renders the beliefs that support Berk-Nash vulnerable to invasion. Drawing on the machine learning literature, we propose entropified Berk-Nash equilibrium, which is immune to this critique.

Keywords: misspecified learning, evolutionary models, Berk-Nash Equilibrium.
JEL Classification: D8, C7,C4

1 Introduction

> *"No statistical model is "true" or "false", "right" or "wrong"; the models just have varying performance, which can be assessed."* – Rissanen (2007)

A desirable property of an equilibrium concept is that there be no profitable deviation (NPD). In the space of learning rules (or, a fortiori, beliefs), this means that an adopted learning rule should ensure that the learning outcome is not a model that, if believed true, leads a player to choose actions that have lower expected payoff (according to the true distribution) than the actions she would choose if she had learned another model.

When a learning problem is correctly specified, a Bayesian learner will learn the true model. In such an environment, no alternative learning rule strictly outperforms Bayesian learning. However, this is not the case when the learning problem is misspecified. To see this, consider a coin which can land either heads or tails and has a true probability of 0.7 of landing heads. Consider Alice, a Bayesian learner whose prior has full support over two models of the coin, one in which the probability of heads is 0.49 and one in which the probability of heads is 0.99. Every period, she earns a dollar if she correctly guesses the outcome of the coin toss. Bayesian updating will lead her, in the limit, to place probability one on the first model. Hence, she will predict tails and earn an average per period payoff of 0.3. However, she would achieve a higher payoff if she placed probability one on the second model, predicted heads and earned an average per period payoff of 0.7. NPD is not satisfied.

From an evolutionary perspective (see, e.g. Weibull, 1995; Sandholm, 2010), a population of Alice-like players who learn using Bayes' rule would thus be vulnerable to invasion by a mutant, say Bob, who follows a learning rule that eventually places probability one on the second model, leading him to predict heads. If payoff is positively related to replication, then over time the share of Bobs in the population will increase as they outperform the Alices in terms of realized payoff. Note that Bob in fact performs exactly as well as another player type, say Colm, who learns the correct belief that the probability of heads is 0.7. In pragmatic terms, Bob and Colm learn perfectly. Alice, in contrast, learns the model in her support that maximizes log-likelihood.

Here, we propose an equilibrium concept that satisfies NPD even if the learning environment is not well specified. The equilibrium we propose, *entropified Berk-Nash equilibrium* (eBNE), is similar to Berk-Nash equilibrium (BNE) of Esponda and Pouzo (2016) in requiring that players' beliefs attach probability one to the set of subjective distributions over consequences that are "closest" to the objective distribution. The main difference between BNE and eBNE is the way in which the respective concepts define the distance between distributions. Specifically, BNE uses weighted Kullback-

Leibler divergence (wKLD) whereas eBNE uses *entropified Kullback-Leibler divergence* (eKLD), the difference being that the latter uses *entropified probabilities* (Grünwald, 1998) that are constructed using payoffs. This approach weights the learning process in favour of beliefs supporting actions with higher objective expected payoffs. That is to say, the players learn what they care about: payoffs.

In well-specified learning settings with a proper information structure, BNE coincides with NE and every NE is a eBNE. The reverse inclusion does not hold because eBNE does not uniquely pin down beliefs: the same actions can be, and typically are, a best response to more than one belief. In misspecified learning settings, a BNE is observationally equivalent to some eBNE if and only if its beliefs are a *useful* model in that they lead to the highest payoff possible amongst the models subjectively believed possible by the decision maker.

Esponda and Pouzo (2016) justify the use of wKLD from a Bayesian learning perspective. A Bayesian player will eventually attach positive probability only to models that have minimal Kullback-Leibler divergence. This justification for BNE relies on the implicit assumption that Bayes' rule is the "rational" way to learn even if the learning problem is misspecified.

However, the use of Bayes' rule in misspecified problems is controversial in the (more pragmatic) statistical learning and computer science literature because they are mainly concerned with empirical validation and Bayes' rule is known to lack robustness to model misspecification.[1] When convergence of the posterior occurs, Bayes' rule converges to the maximum likelihood model (Berk, 1966; White, 1982), but there is no guarantee that the maximum likelihood model is also the model that maximizes payoffs (Grünwald et al., 2017; Csaba and Szoke, 2018; Massari, 2019). Hence, for a decision maker that wishes to maximize payoff, the adoption of Bayes' rule in (possibly) misspecified learning problems is irrational in the sense that

> *"...a mode of behavior is irrational for a given decision maker, if, when the decision maker behaves in this mode and is then exposed to the analysis of her behavior, she feels embarrassed"* – Gilboa (2009, pp.139)

[1] See Timmermann (2006); Grünwald (2007); Grünwald and Langford (2007). Note that in this literature, maximizing expected payoff is usually described as minimizing an expected loss.

There are many candidate solutions to "robustify" a learning problem. Most of them are obtained by incorporating the objective of the decision maker into his learning rule to give more weight to models that induce actions that lead to high expected payoffs (according to the objective distribution), rather than to the models with the highest likelihood. We rely on the *entropification* approach of Grünwald (1998) to transform the original learning problem of the player to fit the generalized Bayesian learning framework (a.k.a. aggregation algorithm, Vovk, 1990; Rissanen, 1989). This choice is to maintain a close parallel between our analysis and that of Esponda and Pouzo (2016) so that we can preserve the role played by subjective beliefs in the resulting equilibrium.

The paper is organized as follows. Section 2 gives the model. Section 3 defines and discusses eBNE, and compares eBNE and BNE. Section 4 presents examples. Section 5 discusses the learning foundation of eBNE.

2 Model

Our model is that of Esponda and Pouzo (2016). A **game** $\mathcal{G} = \langle \mathcal{O}, \mathcal{Q} \rangle$ is composed of a (simultaneous-move) objective game $\mathcal{O}$ and a subjective model $\mathcal{Q}$. The objective game represents the players' true environment. The subjective model represents the players' perception of their environment.

Objective game. A (simultaneous-move) **objective** game is a tuple

$$\mathcal{O} = \langle I, \Omega, \mathbb{S}, p, \mathbb{X}, \mathbb{Y}, f, \pi \rangle .$$

I is the set of players. Ω is the set of payoff-relevant states. $\mathbb{S} = \times_{i \in I} \mathbb{S}^i$ is the set of profiles of signals, where $\mathbb{S}^i$ is the set of signals for player i. p is a probability distribution over $\Omega \times \mathbb{S}$ and is assumed to have marginals with full support. Standard notation is used to denote marginal and conditional distributions, for example $p_{\Omega|S^i}(\cdot|s^i)$ denotes the conditional distribution over Ω given $S^i = s^i$. $\mathbb{X} = \times_{i \in I} \mathbb{X}^i$ is a set of profiles of actions, where $\mathbb{X}^i$ is the set of actions of player i. $\mathbb{Y} = \times_{i \in I} \mathbb{Y}^i$ is a set of profiles of (observable) consequences, where $\mathbb{Y}^i$ is the set of consequences for player i. $f = (f^i)_{i \in I}$

is a profile of feedback or consequence functions, where $f^i : \mathbb{X} \times \Omega \to \mathbb{Y}^i$ maps outcomes in $\Omega \times \mathbb{X}$ into consequences for player i. $\pi = (\pi^i)_{i \in I}$, where $\pi^i : \mathbb{X}^i \times \mathbb{Y}^i \to \mathbb{R}$ is the payoff function of player i. All of the above sets are finite.

A strategy for player i is a mapping $\sigma^i : \mathbb{S}^i \to \Delta(\mathbb{X}^i)$. The probability that player i chooses action x^i after observing signal s^i is denoted by $\sigma^i(x^i|s^i)$. A strategy profile is a vector of strategies $\sigma = (\sigma^i)_{i \in I}$. Let Σ denote the space of all strategy profiles.

Fix an objective game. For each strategy profile σ, there is an **objective distribution** over player i's consequences, $Q^i_\sigma : \mathbb{S}^i \times \mathbb{X}^i \to \Delta(\mathbb{Y}^i)$, where

$$Q^i_\sigma(y^i|s^i, x^i) = \sum_{\{(\omega, x^{-i}) : f^i(x^i, x^{-i}, \omega) = y^i\}} \sum_{s^{-i}} \prod_{j \neq i} \sigma^j(x^j|s^j) p_{\Omega \times S^{-i}|S^i}(\omega, s^{-i}|s^i). \tag{1}$$

That is, when the strategy profile is σ, player i observes signal s^i and takes action x^i, then the distribution over consequences for player i is given by $Q^i_\sigma(\cdot|s^i, x^i)$.

Subjective model. The subjective model is the set of distributions over consequences that players consider possible a priori. For a fixed objective game, a **subjective model** is a tuple

$$\mathcal{Q} = \langle \Theta, (Q_\theta)_{\theta \in \Theta} \rangle ,$$

$\Theta = \times_{i \in I} \Theta^i$ and Θ^i is player i's parameter set. $Q_\theta = (Q^i_{\theta^i})_{i \in I}$, where $Q^i_{\theta^i} : \mathbb{S}^i \times \mathbb{X}^i \to \Delta(\mathbb{Y}^i)$ is the conditional distribution over player i's consequences parameterized by $\theta^i \in \Theta^i$. Denote the conditional distribution by $Q^i_{\theta^i}(\cdot|s^i, x^i)$.

3 Equilibrium

Like Esponda and Pouzo (2016), we shall require players' actions to be best responses to subjective beliefs and subjective beliefs to be close to the true distribution. The main difference is that their concept (Berk-Nash equilibrium) and our concept (entropified Berk-Nash equilibrium) use different notions of distance between distributions. For ease of comparison, we first recall the definitions used by Esponda and Pouzo (2016).

Definition 1. Weighted Kullback-Leibler divergence (wKLD):

$$K^i(\sigma,\theta^i) = \sum_{(s^i,x^i)\in\mathbb{S}^i\times\mathbb{X}^i} E_{Q^i_\sigma(\cdot|s^i,x^i)}\left[\log\frac{Q^i_\sigma(Y^i|s^i,x^i)}{Q^i_{\theta^i}(Y^i|s^i,x^i)}\right]\sigma^i(x^i|s^i)\,p_{S^i}(s^i). \tag{2}$$

Definition 2. A strategy profile σ is a **Berk-Nash equilibrium** (BNE) of game $\mathcal{G}$ if, for all players $i \in I$, there exists $\mu^i \in \Delta(\Theta^i)$ such that

(i) σ^i is optimal given μ^i, and

(ii) If $\hat{\theta}^i$ is in the support of μ^i then $\hat{\theta}^i \in \operatorname{argmin}_{\theta^i\in\Theta^i} K^i(\sigma,\theta^i)$.

The definition of BNE depends on wKLD, which is independent of players' payoffs. In contrast, our concept of entropified Berk-Nash Equilibrium will rely on *entropified* Kullback-Leibler divergence, which depends on the objective expected payoff from the optimal action induced by each subjective belief. The following definitions are instrumental to our definition of entropified Kullback-Leibler divergence.

First, we define the set of *best responses* induced by every subjective belief and the set of *subjectively non-dominated* model-response pairs.

Definition 3. The set of **best responses** of player i to $Q^i_{\theta^i}$ is

$$X^*(Q^i_{\theta^i}) = \times_{s^i\in\mathbb{S}^i} X^*(Q^i_{\theta^i}, s^i),$$

where

$$X^*(Q^i_{\theta^i}, s^i) = \operatorname*{argmax}_{x^i\in\mathbb{X}^i} E_{Q^i_{\theta^i}(\cdot|s^i,x^i)}\pi^i(x^i,Y^i).$$

So $(\bar{x}^i_{s^i})_{s^i\in\mathbb{S}^i} \in X^*(Q^i_{\theta^i})$ is a vector, each element of which comprises a best response for some signal. For some $Q^i_{\theta^i}$, it may be that $X^*(Q^i_{\theta^i})$ has multiple elements. When this is the case, it suits to consider each pair $(Q^i_{\theta^i}, \bar{x}^i)$ as a distinct object that can be learned. We define the set of all such model-response pairs.

Definition 4. The set of **subjectively non-dominated** model-response pairs of player i is

$$\Lambda^i = \left\{(\theta^i,\bar{x}^i) : \theta^i\in\Theta^i,\, \bar{x}^i \in X^*(Q^i_{\theta^i})\right\}.$$

It must be that every $\theta^i \in \Theta^i$ appears in at least one element of Λ^i, but the same is not true for $\bar{x}^i \in (\mathbb{X}^i)^{\mathbb{S}^i}$. If $\bar{x}^i$ is not a best response for any subjective model considered by player i, then it will not be part of any element of Λ^i. Conversely, the same actions can occur in multiple elements of Λ^i. For example, considering the coin toss example from our introduction, if there are multiple models that give a probability of heads of at least half, then each of these models paired with the action "predict heads" will be an element of Λ^i.

There is more than one way to consider such pairs $(\theta^i, \bar{x}^i)$ bonded by a best response correspondence. Our preferred interpretation is that subjective beliefs are ancillary to actions in the sense that it is possible to omit beliefs from the decision model and still have a model, but the model without actions would be nonsensical. What the beliefs do is to restrict the set of possible actions to those that are justifiable by some model in the prior. Actions that are unjustifiable are never taken.

Second, given any belief and an associated subjective best response, we calculate the objective (expected) payoff for player i against a given strategy profile σ.

Definition 5. The **objective payoff** of player i from $(\theta^i, \bar{x}^i) \in \Lambda^i$ is

$$\Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i) = \sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(\cdot|s^i, \bar{x}^i_{s^i})} \pi^i(\bar{x}^i_{s^i}, Y^i)\, p_{S^i}(s^i).$$

If θ^i is the true model σ, then any choice of $\bar{x}^i$ gives the same objective payoff, so we omit the second argument and write $\Pi^i_\sigma(Q^i_\sigma)$.

Last, we use objective payoffs to define our measure of "distance" between distributions, the *entropified K-L divergence.* In Section 5 we clarify the close relation between the "standard" K-L divergence and the *entropified K-L divergence.*

Definition 6. Entropified Kullback-Leibler divergence (eKLD):

$$eK^i(\sigma, \theta^i, \bar{x}^i) = \Pi^i_\sigma(Q^i_\sigma) - \Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i).$$

$\Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i)$ is player i's objective expected payoff when he plays the subjective best response $\bar{x}^i$ to beliefs $Q^i_{\theta^i}$ and the other players follow strategies $(\sigma^j)_{j \neq i}$. So the eKLD measures the distance between model-response pairs in terms of differences in true expected payoffs. In Section 5 we show that the *entropified weighted K-L divergence* plays a similar role in the generalised Bayesian framework to the role played by the regular Kullback-Leibler divergence in standard Bayes. Note that no model can give a higher objective payoff than the true model σ. That is, $\Pi^i_\sigma(Q^i_\sigma) \geq \Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i)$ for all $(\theta^i, \bar{x}^i) \in \Lambda^i$. Therefore, $eK^i(\sigma, \theta^i, \bar{x}^i) \geq 0$ and equals 0, by definition, if and only if $\bar{x}^i \in \operatorname{argmin} \Pi^i(Q^i_\sigma)$ for all s^i. In an iid learning environment in which the player observes all consequences and in which all other players play the equilibrium, these properties and standard Bayesian arguments guarantee that a player i that entropifies his beliefs and uses (generalized) Bayes' rule on the entropified belief set will eventually give zero weight to all beliefs outside the $\operatorname{argmin}_{(\theta^i,\bar{x}^i)} eK^i(\sigma, \theta^i, \bar{x}^i)$ (Lemma 5). He will learn a (set) of beliefs which induce the actions with the highest objective payoff.

Definition 7. A profile $(\theta^{i*}, \bar{x}^{i*})_{i\in I}$, is a (pure) **entropified Berk-Nash equilibrium** (eBNE) of game $\mathcal{G}$ if, for all players $i \in I$,

(i) $(\theta^{i*}, \bar{x}^{i*}) \in \Lambda^i$, and

(ii) $(\theta^{i*}, \bar{x}^{i*}) \in \operatorname{argmin}_{(\theta^i,\bar{x}^i)\in\Lambda^i} eK^i(\sigma, \theta^i, \bar{x}^i)$.

eBNE is a solution concept for players who (i) care about obtaining as high a payoff as possible for themselves, similarly to all Nash-style concepts, and (ii) learn about what they care about.[2] In equilibrium, there are no (subjective) beliefs that player i could learn that could lead him to act in a way that would increase his (objective) expected payoff. In other words, there does not exist an (objectively) profitable deviation to a different set of beliefs together with (subjectively) optimal actions.

Lemma 1. *$(\theta^{i*}, \bar{x}^{i*})_{i\in I}$ is a eBNE if and only if for all $i \in I$,*

$$(\theta^{i*}, \bar{x}^{i*}) \in \underset{(\theta^i,\bar{x}^i)\in\Lambda^i}{\operatorname{argmax}} \Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i). \tag{3}$$

[2] Point (ii) is the intuitive reason that eBNE differs from BNE. eBNE players learn about payoffs, whereas BNE players learn about log-likelihoods (compare the definitions of wKLD and eKLD). In some situations, learning about log-likelihoods is also learning about payoffs. This is the case when the correct model is included in the parameter set (see Section 5) or when a decision maker with log utility solves an investment problem (see Section 4.2). However, in general, the two things are not equivalent.

Proof. Follows immediately from Definitions 6 and 7. □

Our analysis has effectively reduced the problem to a game with player set I, strategy sets Λ^i for $i \in I$, and payoff functions given by the objective payoffs. Lemma 1 tells us that each eBNE corresponds to a pure Nash equilibrium of the reduced game. This illustrates that under an appropriate learning procedure, the role of model misspecification is to reduce the choice of strategies available to a decision maker. This reduces the choice of possible profitable deviations and consequently, if strategies that constitute a pure Nash equilibrium of the objective game $\mathcal{O}$ are still available to players in the game $\mathcal{G} = \langle \mathcal{O}, \mathcal{Q} \rangle$, then there exists a eBNE in these strategies.

Lemma 2. *If $(\bar{x}^{i*})_{i \in I}$ is a pure Nash Equilibrium of the objective game and, for all $i \in I$, there exists $\theta^{i*} \in \Theta^i$ such that $(\theta^{i*}, \bar{x}^{i*}) \in \Lambda^i$, then $(\theta^{i*}, \bar{x}^{i*})_{i \in I}$ is a eBNE.*

Proof. For given i, by definition of Nash equilibrium, $\bar{x}^{i*}$ is a best response under correct beliefs $(Q^i_\sigma)_{i \in I}$. This best response gives an expected payoff of $\Pi^i_\sigma(Q^i_\sigma)$. As $\Pi^i_\sigma(Q^i_\sigma) \geq \Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i)$ for all $(\theta^i, \bar{x}^i) \in \Lambda^i$, and $\Pi^i_\sigma(Q^i_\sigma) = \Pi^i_\sigma(Q^i_{\theta^{i*}}, \bar{x}^{i*})$, it must be that $(\theta^{i*}, \bar{x}^{i*})$ solves (3). □

A question that remains is whether model misspecification should reduce the choice of strategies even further. Specifically, should it be permissible to consider mixing over elements of Λ^i? We can think of two interpretations of such a mixture. The first is that a player mixing between $(\theta^{i1}, \bar{x}^{i1})$ and $(\theta^{i2}, \bar{x}^{i2})$ should act according to beliefs that are a convex combination of $Q^i_{\theta^{i1}}$ and $Q^i_{\theta^{i2}}$. However, it may be that neither $\bar{x}^{i1}$ nor $\bar{x}^{i2}$ is a best response to such beliefs. The second interpretation is the "mass action" interpretation of John Nash's PhD thesis (Nash, 1950a). Under this interpretation, a mixture between $(\theta^{i1}, \bar{x}^{i1})$ and $(\theta^{i2}, \bar{x}^{i2})$ would indicate that player i is drawn from some population and that such a draw renders some chance of player i being of type $i1$, for whom $\bar{x}^{i1}$ is a best response, and some chance of player i being of type $i2$, for whom $\bar{x}^{i2}$ is a best response. This latter interpretation motivates the following.

Let Ξ^i be the set of all probability measures over model-response pairs $(\theta^i, \bar{x}^i)$. Let

ς^i denote an element of Ξ^i. Note that $\varsigma^i \in \Xi^i$ induces a distribution σ^i on $\mathbb{X}^i$ given by

$$\sigma^i(x^i|s^i) = \sum_{(\theta^i,\bar{x}^i)\in\Lambda^i \colon \bar{x}^i_{s^i}=x^i} \varsigma^i\left((\theta^i,\bar{x}^i)\right).$$

It follows that if $(\varsigma^i)_{i\in I}$ is given, then probabilities Q^i_σ under the true model are well-defined and, consequently, so are Π^i_σ and $eK(\sigma,\cdot,\cdot)$.

Definition 8. $(\varsigma^i)_{i\in I}$ is a (mixed) **entropified Berk-Nash equilibrium** (meBNE) of game $\mathcal{G}$ if, for all players $i \in I$, for all $(\theta^{i*}, \bar{x}^{i*})$ in the support of ς^i,

(i) $(\theta^{i*}, \bar{x}^{i*}) \in \Lambda^i$, and

(ii) $(\theta^{i*}, \bar{x}^{i*}) \in \operatorname{argmin}_{(\theta^i,\bar{x}^i)\in\Lambda^i} eK^i(\sigma, \theta^i, \bar{x}^i)$.

Lemma 3. *A mixed eBNE exists.*

Proof. For all $i \in I$, for all $\bar{x}^i$ such that $(\theta^i, \bar{x}^i) \in \Lambda^i$ for some θ^i, choose one such θ^i. Denote the finite set of $(\theta^i, \bar{x}^i) \in \Lambda^i$ chosen this way by $\tilde{\Lambda}^i \subseteq \Lambda^i$.

The game $\tilde{G}$ with player set I, pure strategies $(\tilde{\Lambda}^i)_{i\in I}$ and payoffs equal to objective payoffs is finite and thus has at least one, possibly mixed, Nash equilibrium by Nash's existence theorem (Nash, 1950b). Choose one such equilibrium and denote it by $(\tilde{\varsigma}^{i*})_i$.

Define G to be identical to $\tilde{G}$ except that the strategy sets are Λ^i instead of $\tilde{\Lambda}^i$. For all $i \in I$, let $\varsigma^{i*} = \tilde{\varsigma}^{i*}$ on $\tilde{\Lambda}^i$ and $\varsigma^{i*}(\Lambda^i \setminus \tilde{\Lambda}^i) = 0$.

If $(\varsigma^{i*})_i$ is not a Nash equilibrium of G, there exists a profitable deviation for some player i to some $(\theta^{i1}, \bar{x}^{i1}) \in \Lambda^i$. Note that by construction of $\tilde{\Lambda}^i$ there exists $\theta^i \in \Theta^i$ such that $(\theta^i, \bar{x}^{i1}) \in \tilde{\Lambda}^i \subseteq \Lambda^i$. Objective payoffs do not depend directly on beliefs, so if $(\theta^{i1}, \bar{x}^{i1})$ is a profitable deviation from $(\varsigma^{i*})_i$, then $(\theta^i, \bar{x}^{i1})$ is also a profitable deviation from $(\varsigma^{i*})_i$. However, as $(\tilde{\varsigma}^{i*})_i$ and $(\varsigma^{i*})_i$ induce the same distributions over consequences, it must be that $(\theta^i, \bar{x}^{i1})$ is also a profitable deviation from $(\tilde{\varsigma}^{i*})_i$. Contradiction. Therefore, $(\varsigma^{i*})_i$, is a Nash equilibrium of G.

By definition of Nash equilibrium, if $(\theta^{i*}, \bar{x}^{i*})$ is in the support of ς^{i*}, then $(\theta^{i*}, \bar{x}^{i*}) \in$

$\operatorname{argmax}_{(\theta^i,\bar{x}^i)\in\Lambda^i} \Pi^i_\sigma(\theta^i,\bar{x}^i)$. Definition 6 then implies that

$$(\theta^{i*},\bar{x}^{i*}) \in \underset{(\theta^i,\bar{x}^i)\in\Lambda^i}{\operatorname{argmin}}\, eK^i(\sigma,\theta^i,\bar{x}^i).$$

Therefore $(\varsigma^{i*})_i$ is a eBNE. □

4 Examples

4.1 Coin tosses

Here we discuss the illustrative example from the introduction. A decision maker guesses the outcome of a coin toss, $\mathbb{X}^i = \{H, T\}$. The outcome of the coin toss is independent of the decision maker's action and is given by $y = f(x,\omega) = \omega$, where $\omega = H$ with probability 0.7 and $\omega = T$ with probability 0.3. There are no signals. Hence we have $Q^i_\sigma(H|x^i) = Q^i_\sigma(H) = 0.7$ for all σ, x^i. Payoffs are given by $\pi^i(x,y) = 1$ if $x = y$ and $\pi^i(x,y) = 0$ if $x \neq y$. The parameter set is $\Theta^i = \{\theta^{i1},\theta^{i2}\}$ and we let $Q^i_{\theta^{i1}}(H|x^i) = Q^i_{\theta^{i1}}(H) = 0.49$ and $Q^i_{\theta^{i2}}(H|x^i) = Q^i_{\theta^{i2}}(H) = 0.99$ for all x^i. Note that T is the unique best response to beliefs $Q^i_{\theta^{i1}}$, whereas H is the unique best response to beliefs $Q^i_{\theta^{i2}}$ or to the true model Q^i_σ.

Berk-Nash equilibrium. Substituting into (1) we obtain, for $\theta^i \in \Theta^i$,

$$K^i(\sigma,\theta^i) = E_{Q^i_\sigma(\cdot)}\left[\log \frac{Q^i_\sigma(Y^i)}{Q^i_{\theta^i}(Y^i)}\right]. \tag{4}$$

Therefore, as Q^i_σ is independent of σ, we have that, for all σ,

$$K^i(\sigma,\theta^{i1}) = 0.7\left(\log\frac{0.7}{0.49}\right) + 0.3\left(\log\frac{0.3}{0.51}\right) \approx 0.09$$

and

$$K^i(\sigma,\theta^{i2}) = 0.7\left(\log\frac{0.7}{0.99}\right) + 0.3\left(\log\frac{0.3}{0.01}\right) \approx 0.78.$$

Therefore, at BNE, it must be that model θ^{i1} is believed with probability one. These are the beliefs that would be learned by applying Bayes' rule. The unique best response for θ^{i1} is T, so the unique BNE has $\sigma^i(H) = 0$, $\sigma^i(T) = 1$, supported by the belief $\mu^i(\theta^i) = 1$. At this equilibrium, the decision maker obtains an expected objective payoff of 0.3 and thus he would correctly guess the outcome of the coin toss only 0.3 of the time.

Entropified Berk-Nash equilibrium. The set of subjectively non-dominated model-response pairs is given by $\Lambda^i = \{(\theta^{i1}, T), (\theta^{i2}, H)\}$. We obtain objective expected payoffs

$$\Pi^i_\sigma(Q^i_{\theta^{i1}}, T) = 0.3, \qquad \Pi^i_\sigma(Q^i_{\theta^{i2}}, H) = 0.7, \qquad \Pi^i_\sigma(Q^i_\sigma) = 0.7.$$

Therefore, for all σ, we have

$$eK^i(\sigma, \theta^{i1}, T) = 0.7 - 0.3 = 0.4, \qquad eK^i(\sigma, \theta^{i2}, H) = 0.7 - 0.7 = 0.$$

It follows that the unique eBNE is (θ^{i2}, H). At this equilibrium, the decision maker obtains an expected objective payoff of 0.7 and thus he would correctly guess the outcome of the coin toss 0.7 of the time. As we shall see in Section 5, this is the pair that will be learned by applying generalized Bayes' rule. The player learns what he cares about: payoffs.

4.2 Arrow-Debreu securities

We extend the example of the preceding subsection so that the decision maker chooses a share $x^i \in \mathbb{X}^i = \{0, 0.01, \ldots, 0.99, 1\}$ of a unit of Arrow-Debreu security to invest in outcome H. The remainder is invested in outcome T. Similar to before, $y = f(x, \omega) = \omega$, where $\omega = H$ with probability p_H and $\omega = T$ with probability $1 - p_H$. There are no signals and the decision maker's action does not affect outcome probabilities. Hence we have $Q^i_\sigma(H|x^i) = Q^i_\sigma(H) = p_H$ for all σ, x^i. The decision maker is aware that his action does not affect outcome probabilities and has Bernoulli beliefs parametrized by $\Theta = \{0, 0.01, \ldots, 0.99, 1\}$, so that $\forall \theta^i \in \Theta^i, Q^i_{\theta^i}(H) = \theta^i$. Payoffs are given by $\pi^i(x^i, H) = u(x^i)$ and $\pi^i(x^i, T) = u(1 - x^i)$, where u is a utility function.

Berk-Nash equilibrium. To find a BNE, choose θ^i to minimize (4), then choose a strategy in which any action x^i played with positive probability maximizes $E_{Q^i_{\theta^i}(\cdot)}\pi^i(x^i, Y^i)$.

Entropified Berk-Nash equilibrium. eKLD is minimized by $(\theta^i, x^i) \in \Lambda^i$ that maximize $E_{Q^i_\sigma(\cdot)}\pi^i(x^i, Y^i)$. Hence, in general, actions played in eBNE differ from those played in BNE.

Correctly specified model. If there exists $\theta^{i*} \in \Theta^i$ such that $Q^i_{\theta^{i*}} = Q^i_\sigma$, then wKLD is minimized at θ^{i*}. Then BNE is a NE. Furthermore, $E_{Q^i_{\theta^{i*}}(\cdot)}u(x^i) = E_{Q^i_\sigma(\cdot)}u(x^i)$, therefore we also have that (θ^{i*}, x^i) is a eBNE. Note that there may also exist other eBNE that choose the same actions but are based on incorrect beliefs. However, if (θ^i, x^i) is a eBNE, then (θ^{i*}, x^i) is also a eBNE and σ^i, $\sigma^i(x^i) = 1$, is a BNE supported by the belief $\mu^i(\theta^{i*}) = 1$.

Log utility. Now, let $u(\cdot) = \log(\cdot)$. When this is the case, for any $(\theta^i, x^i) \in \Lambda^i$, we have that $x^i = Q^i_{\theta^i}(H)$ and $1 - x^i = Q^i_{\theta^i}(T)$. That is, the share of asset invested in H equals the subjective probability of H. Readers will recognize this as the celebrated Kelly criterion. Consequently, eKLD will be minimized by $(\theta^i, x^i) \in \Lambda^i$ that maximize $E_{Q^i_\sigma(\cdot)} \log Q^i_{\theta^i}(Y^i)$. This is equivalent to minimizing (4), therefore (θ^i, x^i) is a eBNE if and only if σ^i, $\sigma^i(x^i) = 1$, is a BNE supported by the belief $\mu^i(\theta^i) = 1$.

4.3 Monopoly with unknown demand

Here we consider Example 2.1 of Esponda and Pouzo (2016). A monopolist chooses a price $x^i \in \mathbb{X}^i = \{2, 10\}$ that generates demand $y^i = f(x^i, \omega) = \phi_0(x^i) + \omega$, where ω is a mean-zero shock with distribution $p \in \Delta(\Omega)$. It is assumed that $\phi_0(2) = 34$ and $\phi_0(10) = 2$. There are no signals. The payoff is $\pi^i(x^i, y^i) = x^i y^i$.

The monopolist's uncertainty about p and f is described by a parametric model f_θ, p_θ, where $y = f_\theta(x^i, \omega) = a - bx^i + \omega$ is the demand function, $\theta = (a, b) \in \Theta$ is a parameter vector, and $\omega \sim N(0, 1)$. The set of possible models is given by $\Theta = [33, 40] \times [3, 3.5]$. Let $\theta_0 \in \mathbb{R}^2$ provide a perfect fit for the demand so that $\phi_0(x^i) = \phi_{\theta_0}(x^i)$ for all $x^i \in \mathbb{X}^i$. This gives $\theta_0 = (a_0, b_0) = (42, 4) \notin \Theta$ and therefore the monopolist has a misspecified model. Note that, as there are no other players, the conditional objective distribution $Q^i_\sigma(\cdot|x^i)$ does not depend on σ and is normal with mean $\phi_0(x^i)$ and unit variance. Similarly, $Q_\theta(\cdot|x^i)$ is normal with mean $\phi_\theta(x^i) = a - bx$ and unit variance.

Berk-Nash equilibrium. Esponda and Pouzo (2016) show that if σ is a BNE, then σ^i must put strictly positive probability on each action. The reason for this is that if $\sigma^i(2) = 1$ so that the monopolist only plays action 2, then all $\theta^i \in \Theta^i$ that minimize $K^i(\sigma, \theta^i)$ are such that $X^*(Q^i_{\theta^i}) = \{10\}$. That is, playing $\sigma^i(2) = 1$ causes the monopolist to learn beliefs to which the unique best response is to play 10. Similarly, playing $\sigma^i(10) = 1$ causes the monopolist to learn beliefs to which the unique best response is to play 2. Consequently, BNE involves mixing between actions 2 and 10. The frequencies with which each action is played under σ^i are chosen so that the θ^i that minimizes $K^i(\sigma, \theta^i)$ makes the monopolist (subjectively) indifferent between the actions. The unique Berk-Nash equilibrium of this game is $\sigma^i = (35/36, 1/36)$ with supporting beliefs given by the parameters $\theta^i = (40, 10/3)$. These beliefs are those that a Bayesian learner would learn from observing the outcomes induced by the BNE strategy σ. That is, the equilibrium θ^i maximizes log-likelihood amongst all possible parameters in Θ^i. However, the monopolist does not learn about payoffs. As we shall see below, he obtains a higher payoff when he plays 2 than he obtains when he plays 10. The learning justification of BNE relies on his learning a mixed strategy whilst never noticing the difference in payoffs obtained from each of the actions in his support. Furthermore, there exist beliefs arbitrarily close to the BNE beliefs that induce a unique subjective best response 2 that is associated with the best possible objective payoff.

Entropified Berk-Nash equilibrium. By substituting the true model parameters into the payoff function, we obtain expected objective payoffs from playing model-response pairs $(Q^i_{\theta^i}, x^i)$.

$$\begin{aligned}\Pi^i_\sigma(Q^i_{\theta^i}, x^i) &= E[\pi^i(x^i, y^i)] = E\left[x^i(42 - 4x^i + \omega))\right] \\ &= x^i(42 - 4x^i) + x^i E(\omega) = \begin{cases} 68 \; if \; x^i = 2 \\ 20 \; if \; x^i = 10 \end{cases}\end{aligned}$$

As there are no actions other than 2 and 10, and no uncertainty about other players, with true beliefs the monopolist should always choose action 2, so $\Pi^i_\sigma(Q^i_\sigma) = 68$. Applying Definition 6, we see that $eK^i(\sigma, \theta^i, 2) = 68 - 68 = 0$ for all pairs $(\theta^i, 2) \in \Lambda^i$

and $eK^i(\sigma, \theta^i, 2) = 68 - 20 = 48$ for all pairs $(\theta^i, 10) \in \Lambda^i$. Consequently, the set of eBNE is the set of all pairs $(\theta^i, 2) \in \Lambda^i$. As there is only one player, the set of meBNE is the set of all mixtures on these pairs. As we shall see in Section 5, these are exactly the pairs that can be learned by applying generalized Bayes' rule. The player learns what he cares about: payoffs.

4.4 Regression to the mean

Here we consider Example 2.3 of Esponda and Pouzo (2016). An instructor observes the initial performance $s^i = y_1^i$ of a student and decides to praise or criticize him, $x^i \in \{C, P\}$. The student then performs again and the instructor observes his final performance, y_2^i. The truth is that performances y_1^i and y_2^i are drawn independently from the same distribution with mean zero. The instructor's payoff is $\pi^i(x^i, y^i) = y_2^i - c(x^i, y_1^i)$, where $c(x^i, y_1^i) = \kappa |y_1^i| > 0$ if either $y_1^i > 0$, $x = C$ or $y_1^i < 0$, $x = P$, and, in all other cases, $c(x^i, y_1^i) = 0$. The function c represents a reputation cost from lying – criticizing above average performance or praising below average performance – that increases in the size of the lie.

The instructor believes that $y_2^i = y_1^i + \theta_x^i + \varepsilon$, where $\varepsilon \sim N(0, 1)$, and $\theta^i = (\theta_C^i, \theta_P^i) \in \Theta = \mathbb{R}^2$. Note that as the instructor is informed of y^{i1} by her signal s^i, it is acceptable for her to condition her action on the value of y^{i1}. In fact, considering her subjective model, we see that best responses are characterized by a cutoff $\bar{s}$, whereby she plays P if $s^i = y^{i1} > \bar{s}$ and plays C if $s^i = y^{i1} \leq \bar{s}$. Observe that as she cannot influence performance, objective payoff is maximized when $\bar{s} = 0$.

Berk-Nash equilibrium. Esponda and Pouzo (2016) show that at BNE the instructor believes that playing C improves the student's performance whereas playing P worsens it, and thus chooses an inefficiently high $\bar{s} > 0$.

Entropified Berk-Nash equilibrium. Consider any model-response pair $(\theta^{i*}, \bar{s}^*)$ such that $\theta_C^{i*} = \theta_P^{i*}$ and $\bar{s}^* = 0$. This pair is in Λ^i as $\bar{s} = 0$ is a best response for such θ^{i*}. It is optimal under the correct beliefs to choose $\bar{s} = 0$, therefore $\bar{s} = 0$ is a Nash equilibrium of the (one player) objective game. Consequently, by Lemma 2, $(\theta^{i*}, \bar{s}^*)$ is a eBNE. Further note that if θ^i is such that $\theta_C^i \neq \theta_P^i$, then $\bar{s} = 0$ will not be a best response. Therefore, the eBNE described above are the only eBNE. So an instructor who learns from payoffs should converge on a belief that praise and criticism are equally (in)effective. Given this, she will prefer to tell the truth —i.e., to set $\bar{s} = 0$.

5 A learning foundation for eBNE

BNE has been justified as a possible learning outcome when players learn according to Bayes rule from a prior with full support on Θ. This justification assumes that players learn according to a rule that is axiomatically sound but independent of the payoff of the players. In contrast, eBNE represents the learning outcome of players that learn directly about the consequences of their actions, rather than completely separate the belief space from the payoffs corresponding to the actions chosen. This approach is typical of reinforcement learning (see, e.g. Erev and Roth, 1998; Roth and Erev, 1995) and is arguably closer to the way learning occurs in real-world situations because it is both less abstract and more robust than Bayes' rule. Less abstract because players learn directly from and about rewards and punishments rather than learning from observations about a hypothetical parameter characterizing a true distribution. More robust because, unlike Bayes' rule, it guarantees that a player will learn a model that induces an action that leads to as high an objective expected payoff as possible.

To incorporate the subjective prior beliefs of our player into the learning problem we postulate that players utilize the *aggregation algorithm* of Vovk (1990) (generalized Bayes' algorithm, Rissanen, 1989, with our parameter choice). First, players transform their original beliefs to a new set of entropified probabilities (Grünwald, 1998) which incorporate payoffs that correspond to the best responses induced by each subjective belief. Second, players update their prior beliefs iteratively using (generalized) Bayes' rule on the set of entropified probabilities.

Here we briefly describe the entropification procedure, define generalized Bayes' rule and provide a simple proof of the fact that a player who follows this rule will learn to play actions that correspond to the highest objective payoff that can be justified by some model in his prior. That is, players learn to play as per the definition of eBNE.

We then illustrate the differing learning outcomes of Bayes' and generalized Bayes' rule by revisiting Example 4.1 (coin tosses).

Definition 9. For each $(\theta^i, \bar{x}^i) \in \Lambda^i$, the *entropified probability* of consequence y^i given s^i is

$$eQ^i_{(\theta^i,\bar{x}^i)}(\, y^i \mid s^i \,) = \frac{e^{\beta\,\pi^i(\bar{x}^i_{s^i},y^i)}}{\int_{\mathbb{Y}^i} e^{\beta\,\pi^i(\bar{x}^i_{s^i},\hat{y}^i)} d\hat{y}^i},$$

where we will fix $\beta = 1$ for the rest of the paper.[3] For given σ, we similarly define eQ^i_σ by replacing $\bar{x}^i$ with an arbitrary best response to Q^i_σ.

Entropified probabilities are defined with reference to the finite set of best responses rather than the possibly infinite set of model-response pairs. Thus, dealing with entropified probabilities effectively reduces the domain of the learning problem to a finite set of classes of model-response pairs indexed by the set of subjectively non-dominated best responses. This finiteness guarantees that assumptions **C1** – **C4** of Grünwald (1998) are satisfied (π is bounded) and makes Assumption 1 of Esponda and Pouzo (2016) redundant.

It is easy to verify that if we replace $\pi^i(\bar{x}_{s^i}, y^i)$ by $\log Q^i_{\theta^i}(y^i|s^i, \bar{x}_{s^i})$ in Definition 9, then we obtain the standard likelihood function. This analogy goes further. In fact, if probabilities over outcomes are independent of a player's action, then the entropified Kullback-Leibler divergence of Definition 6 is simply the definition of standard Kullback-Leibler divergence applied to the entropified probabilities.

Lemma 4. *If* $\pi^i_\theta(\bar{x}_{s^i}, y^i) = \log Q^i_{\theta^i}(y^i|s^i, \bar{x}_{s^i})$ *and* $Q^i_\sigma(\cdot|s^i, x^i) = Q^i_\sigma(\cdot|s^i)$ *for all* $x^i \in \mathbb{X}^i$, *then*

$$eK^i(\sigma, \theta^i, \bar{x}^i) = \sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(Y^i|s^i)} \left[\log \frac{eQ^i_\sigma(Y^i|s^i)}{eQ^i_{(\theta^i,\bar{x}^i)}(Y^i|s^i)} \right] p_{S^i}(s^i)$$

Proof. Let $\bar{z}^i$ be a best response to Q^i_σ. Then,

$$\sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(Y^i|s^i)} \left[\log \frac{eQ^i_\sigma(Y^i|s^i)}{eQ^i_{(\theta^i,\bar{x}^i)}(Y^i|s^i)} \right] p_{S^i}(s^i)$$

[3] The appropriate value of β (a.k.a. the learning rate) is an active topic in the machine learning literature (e.g., Grünwald, 1998). WLOG, we set $\beta = 1$ because any time independent value of the learning rate converges to the same model when convergence occurs if the prior support is finite.

$$= \sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(Y^i|s^i)} \left[\log e^{\pi^i(\bar{z}^i_{s^i}, Y^i)} - \log e^{\pi^i(\bar{x}^i_{s^i}, Y^i)}\right] p_{S^i}(s^i)$$

$$= \sum_{s^i \in \mathbb{S}^i} \underbrace{E_{Q^i_\sigma(Y^i|s^i)} \left[\pi^i(\bar{z}^i_{s^i}, Y^i) - \pi^i(\bar{x}^i_{s^i}, Y^i)\right]}_{\substack{=E_{Q^i_\sigma(Y^i|s^i,\bar{z}^i_{s^i})}[\cdot]=E_{Q^i_\sigma(Y^i|s^i,\bar{x}^i_{s^i})}[\cdot] \\ \text{by assumption in lemma statement}}} p_{S^i}(s^i)$$

$$= \sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(Y^i|s^i,\bar{z}^i_{s^i})} \left[\pi^i(\bar{z}^i_{s^i}, Y^i)\right] p_{S^i}(s^i) - \sum_{s^i \in \mathbb{S}^i} E_{Q^i_\sigma(Y^i|s^i,\bar{x}^i_{s^i})} \left[\pi^i(\bar{x}^i_{s^i}, Y^i)\right] p_{S^i}(s^i)$$

$$\underbrace{=}_{\substack{\text{by} \\ \text{Definition 5}}} \Pi^i_\sigma(Q^i_\sigma) - \Pi^i_\sigma(Q^i_{\theta^i}, \bar{x}^i) \underbrace{=}_{\substack{\text{by} \\ \text{Definition 6}}} eK^i(\sigma, \theta^i, \bar{x}^i).$$

□

Given entropified probabilities, we can define the generalized likelihood of any model response pair $(\theta^i, \bar{x}^i)$ and the generalized Bayesian prior after t observations. These are simply the standard concepts applied to the entropified probabilities.

Definition 10. For each $(\theta^i, \bar{x}^i) \in \Lambda^i$, the *generalized likelihood* after t periods on $(s^i_\tau, y^i_\tau)^t_{\tau=0}$ is

$$gQ^i_{(\theta^i,\bar{x}^i)}\left((y^i_1, s^i_1), ..., (y^i_t, s^i_t)\right) = \prod_{\tau=1}^{t} \left(\frac{e^{\pi^i(\bar{x}^i_{s^i_\tau}, y^i_\tau)}}{\int_{\mathbb{Y}^i} e^{\pi^i(\bar{x}^i_{s^i_\tau}, \hat{y}^i)} d\hat{y}^i} \right).$$

Definition 11. Given prior distribution μ^i_0 on Λ^i, the *generalized Bayesian prior* distribution $g\mu^i_t$ given observations $(s_\tau, y_\tau)^t_{\tau=0}$ is given by

$$g\mu^i_t(A) = \frac{\int_A gQ_{(\theta^i,\bar{x}^i)}\left((y^i_1, s^i_1), ..., (y^i_t, s^i_t)\right)\, d\mu^i_0(\theta^i, \bar{x}^i)}{\int_{\Lambda^i} gQ^i_{(\theta^i,\bar{x}^i)}\left((y^i_1, s^i_1), ..., (y^i_t, s^i_t)\right)\, d\mu^i_0(\theta^i, \bar{x}^i)},$$

for $A \subseteq \Lambda^i$.

The following lemma applies the strong law of large numbers to show that generalized Bayes' rule gives more weight to entropified beliefs with lower eKLD, so that generalized Bayes' rule eventually gives positive weight only to beliefs that support actions that minimize eKLD. That is, if learning under generalized Bayes' rule converges, the outcome must be a eBNE.

Lemma 5. *Let Q_σ be generated by σ. Write $A := \operatorname{argmin}_{(\theta^i,\bar{x}^i)\in\Lambda^i} eK^i(\sigma,\theta^i,\bar{x}^i)$. If $\mu^i_0(A) > 0$, then $g\mu^i_t(A) \to 1$ Q_σ-a.s. as $t \to \infty$.*

Proof. Write

$$a := \max_{(\theta^i,\bar{x}^i)\in\Lambda^i} \Pi^i_\sigma(Q^i_{\theta^i},\bar{x}^i) \qquad \text{and} \qquad b := \max_{(\theta^i,\bar{x}^i)\in\Lambda^i\setminus A} \Pi^i_\sigma(Q^i_{\theta^i},\bar{x}^i). \tag{5}$$

The result follows from the strong law of large numbers (SLLN):

$$\begin{aligned}
g\mu^i_t(A) \qquad & = 1 - g\mu^i_t(\Lambda^i \setminus A) \\
& \underbrace{=}_{\substack{\text{by}\\ \text{Definition 11}}} \quad 1 - \frac{\int_{\Lambda^i\setminus A} gQ_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)\, d\mu^i_0(\theta^i,\bar{x}^i)}{\int_{\Lambda^i} gQ^i_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)\, d\mu^i_0(\theta^i,\bar{x}^i)} \\
& \geq 1 - \frac{\int_{\Lambda^i\setminus A} gQ_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)\, d\mu^i_0(\theta^i,\bar{x}^i)}{\int_{A} gQ^i_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)\, d\mu^i_0(\theta^i,\bar{x}^i)} \\
& = 1 - \frac{\int_{\Lambda^i\setminus A} e^{\log gQ_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)}\, d\mu^i_0(\theta^i,\bar{x}^i)}{\int_{A} e^{\log gQ^i_{(\theta^i,\bar{x}^i)}\left((y^i_1,s^i_1),...,(y^i_t,s^i_t)\right)}\, d\mu^i_0(\theta^i,\bar{x}^i)} \\
& \underbrace{=}_{\substack{\text{by}\\ \text{Definition 10}}} \quad 1 - \frac{\int_{\Lambda^i\setminus A} e^{t\sum_{\tau=1}^t \frac{1}{t}\pi(\bar{x}^i_{s^i_\tau},y_\tau)}\, d\mu^i_0(\theta^i,\bar{x}^i)}{\int_{A} e^{t\sum_{\tau=1}^t \frac{1}{t}\pi(\bar{x}^i_{s^i_\tau},y_\tau)}\, d\mu^i_0(\theta^i,\bar{x}^i)} \\
& \underbrace{\approx}_{\substack{Q^i_\sigma\text{-a.s. for } t \text{ large}\\ \text{by SSLN}}} \quad 1 - \frac{\int_{\Lambda^i\setminus A} e^{t\,\Pi_\sigma(Q^i_{\theta^i},\bar{x}^i)}\, d\mu^i_0(\theta^i,\bar{x}^i)}{\int_{A} e^{t\,\Pi_\sigma(Q^i_{\theta^i},\bar{x}^i)}\, d\mu^i_0(\theta^i,\bar{x}^i)} \\
& \underbrace{\geq}_{\text{by (5)}} 1 - \frac{e^{t\,b}\,\mu^i_0(\Lambda^i\setminus A)}{e^{t\,a}\,\mu^i_0(A)} \quad \underbrace{\xrightarrow{t\to\infty}}_{\text{by } a>b} \; 1.
\end{aligned}$$

□

It follows that the generalized Bayesian prior will identify responses $\bar{x}^i$ which give the highest objective expected payoff, but that each such response may correspond to a multiplicity of beliefs. By using generalized Bayes, players pragmatically learn how to act to maximize their average payoff according to the true distribution, rather than which of their probabilistic models is the most accurate in some abstract sense.

5.1 Coin tosses revisited

Again consider Example 4.1 in which a decision maker learns a probabilistic model of coin tosses from amongst the models θ^{i1} (probability of heads is 0.49) and θ^{i2} (probability of heads is 0.99).

Bayes' rule. By standard arguments, the prior probability of θ^{i1} after t periods, calculated via Bayes' rule is

$$\begin{aligned}
\mu_t^i(\theta^{i1}) &= \left(\frac{\mu_0^i(\theta^{i1}) \prod_{\tau=1}^t \prod_{\omega=H,T} Q_{\theta^{i1}}^i(\omega)^{I_{y_\tau=\omega}}}{\mu_0^i(\theta^{i1}) \prod_{\tau=1}^t \prod_{\omega=H,T} Q_{\theta^{i1}}^i(\omega)^{I_{y_\tau=\omega}} + \mu_0^i(\theta^{i2}) \prod_{\tau=1}^t \prod_{\omega=H,T} Q_{\theta^{i2}}^i(\omega)^{I_{y_\tau=\omega}}} \right) \\
&= \frac{1}{1 + \frac{\mu_0^i(\theta^{i2})}{\mu_0^i(\theta^{i1})} e^{\sum_{\tau=1}^t \sum_{\omega=H,T} I_{y_\tau=\omega} \log \frac{Q_{\theta i2}^i(\omega)}{Q_{\theta i1}^i(\omega)}}} \\
&= \frac{1}{1 + \frac{\mu_0^i(\theta^{i2})}{\mu_0^i(\theta^{i1})} e^{t\left(\frac{1}{t} \sum_{\tau=1}^t \sum_{\omega=H,T} I_{y_\tau=\omega} \log \frac{Q_\sigma^i(\omega)}{Q_{\theta i1}^i(\omega)} - \frac{1}{t} \sum_{\tau=1}^t \sum_{\omega=H,T} I_{y_\tau=\omega} \log \frac{Q_\sigma^i(\omega)}{Q_{\theta i2}^i(\omega)} \right)}} \\
&\approx_{\text{for t large}}^{Q_\sigma^i\text{-a.s.}} \left(\frac{1}{1 + \frac{\mu_0^i(\theta^{i2})}{\mu_0^i(\theta^{i1})} e^{t(K(\sigma,\theta^{i1}) - K(\sigma,\theta^{i2}))}} \right) \\
&\to^{Q_\sigma^i\text{-a.s.}} \begin{cases} 1 & \text{iff} \quad K(\sigma,\theta^{i1}) - K(\sigma,\theta^{i2}) < 0 \\ 0 & \text{iff} \quad K(\sigma,\theta^{i1}) - K(\sigma,\theta^{i2}) > 0 \end{cases}
\end{aligned}$$

For $Q_\sigma^i(H) = 0.7$, a quick calculation shows that $K(\sigma,\theta^{i1}) - K(\sigma,\theta^{i2}) < 0$, so that $\mu_t^i(\theta^{i1}) \to 1$ as $t \to \infty$. Accordingly, the Bayesian player becomes certain that tails is more likely than heads and bets on tail for all large t. These learned beliefs ensure him an objective expected payoff of 0.3, which is lower than the objective expected utility of 0.7 that he would have obtained had he learned the θ^{i2} model.

Generalized Bayes' rule. Adapting the previous argument (see also the proof of Lemma 5) the generalized Bayesian prior probability of θ^{i1} after t periods is

$$\begin{aligned}
\mu_t^e(\theta_1) \approx_{\text{for t large}}^{Q_\sigma^i\text{-a.s.}} &\left(\frac{1}{1 + \frac{\mu_0(\theta_2)}{\mu_0(\theta_2)} e^{t(eK(\sigma,\theta^{i1},T) - eK(\sigma,\theta^{i2},H))}} \right) \\
\to^{Q_\sigma^i\text{-a.s.}} &\begin{cases} 1 & \text{iff} \quad eK(\sigma,\theta^{i1},T) - eK(\sigma,\theta^{i2},H) < 0 \\ 0 & \text{iff} \quad eK(\sigma,\theta^{i1},T) - eK(\sigma,\theta^{i2},H) > 0 \end{cases}
\end{aligned}$$

which implies that the generalized Bayesian prior converges to a Dirac distribution on the parameter with the lowest eKL divergence from the truth. For $Q^i_\sigma(H) = 0.7$, our decision maker correctly learns that betting on heads is more profitable than betting on tails and that he is better off acting under the beliefs $Q_{\theta^{i2}}$ than he is acting under the beliefs $Q_{\theta^{i1}}$.

References

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

Csaba, D. and Szoke, B. (2018). Learning with misspecified models. *mimeo.*

Erev, I. and Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 88(4):848–881.

Esponda, I. and Pouzo, D. (2016). Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130.

Gilboa, I. (2009). *Theory of decision under uncertainty*, volume 1. Cambridge university press.

Grünwald, P. and Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149.

Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Grünwald, P. D. (1998). *The minimum description length principle and reasoning under uncertainty.* PhD thesis, Quantum Computing and Advanced System Research.

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

Massari, F. (2019). Ambiguity, robust statistics, and Raiffa's critique. SSRN Working Paper Series 3388410.

Nash, J. (1950a). *Non-cooperative games.* PhD thesis, Princeton University, USA.

Nash, J. F. (1950b). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* World Scientific.

Rissanen, J. (2007). *Information and complexity in statistical modeling.* Springer Science & Business Media.

Roth, A. E. and Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, 8(1):164–212.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics.* Economic learning and social evolution. Cambridge, Mass. MIT Press.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Vovk, V. G. (1990). Aggregating strategies. *Proc. of Computational Learning Theory, 1990.*

Weibull, J. (1995). *Evolutionary game theory.* MIT Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.