Pathwise Concentration Bounds for Bayesian Beliefs∗

Drew Fudenberg† Giacomo Lanzani‡ Philipp Strack§

First posted version: March 10, 2021
This version: May 5, 2022

Abstract

We show that Bayesian posteriors concentrate on the outcome distributions that approximately minimize the Kullback-Leibler divergence from the empirical distribution, uniformly over sample paths, even when the prior does not have full support. This generalizes Diaconis and Freedman (1990)’s uniform convergence result to e.g. finite-support priors, or priors that satisfy parametric restrictions or independence assumptions. Our result lets us provide a rate of convergence for Berk (1996)’s result on the limiting behavior of posterior beliefs when the prior is misspecified. The main result also facilitates the derivation of a concentration bound for beliefs over the parameters of a parametric model when some probability distributions may not be representable by any parameter. We use the concentration result to provide a bound on approximation errors in “anticipated-utility” models, and extend our analysis to outcomes that are perceived to follow a Markov process.

*We thank Simone Cerreia-Vioglio, Xiaohong Chen, Jana Gieselmann, Paul Heidhues, Yuhta Ishii, Mats Köster, and Pooya Molavi for helpful conversations, and National Science Foundation grant SES 1951056, the Guido Cazzavillan Scholarship, and the Sloan Foundation for financial support.

†Department of Economics, MIT, drew.fudenberg@gmail.com
‡Department of Economics, MIT, lanzani@mit.edu
§Department of Economics, Yale University, philipp.strack@yale.edu
1 Introduction

Learning from repeated observations is a key feature of many economic settings, and almost all economic studies of learning model it as Bayesian inference. To understand the medium and long-run implications of Bayesian learning, it is very useful to know how quickly beliefs concentrate around the data generating processes that best explain the observations. Our main result, Theorem 1, shows that the probability the posterior assigns to distributions that do not approximately maximize the likelihood assigned to the data vanishes exponentially quickly. Importantly, we identify conditions for this to hold not only with high probability, but for every possible realization of the data. More specifically, Theorem 1 establishes that for every $\varepsilon > 0$ the posterior probability of the distributions that do not $\varepsilon$-minimize the Kullback-Leibler (KL) divergence vanishes at an exponential rate. Importantly, our bound holds even if the agent’s prior does not have full support, or satisfies parametric restrictions, and holds regardless of the nature of the true data generating process.

Our results generalize Diaconis and Freedman (1990), which showed that a $\phi$-positivity condition implies that Bayesian posteriors converge to the empirical distribution at a uniform exponential rate. This condition implies that the support of the agent’s prior includes every distribution over outcomes, and thus rules out many settings of economic interest in which the set of outcome distributions is naturally restricted. For example, it does not apply to agents who have a prior with finite support (which is natural in settings such as urn problems with a finite number of balls), agents who each period observe a set of Bernoulli trials that they think are i.i.d., or agents who are certain that money supply and output are positively correlated, as in Spiegler (2020) where the agent neglects the mediating role of expectations in the Phillips curve. In addition, $\phi$-positivity rules out any case where the support of the agent’s prior does not contain the true data generating process, so that the agent is misspecified.

Theorem 1 guarantees that beliefs concentrate on the approximate KL minimizers for the empirical frequency. We show that this is equivalent to concentration on a ball around the exact KL minimizers when priors have full support, but not in general. However, Corollary 1 shows that if the support of the prior is convex, which rules out priors with finite or countable support, then beliefs do concentrate with a uniform rate on a neighborhood of the exact KL minimizer, although the size of the neighborhood
depends on the distance between the realized empirical distribution and the boundary of the probability simplex. The corollary is a consequence of Proposition 1, which gives a more general sufficient condition for beliefs to concentrate around the exact minimizers.

We use Theorem 1 to prove Theorem 2, which provides a rate of convergence for Berk (1966)’s result that posterior beliefs concentrate around the Kullback-Leibler minimizers with respect to the true data generating process. Berk’s result, like our paper, is stated for an exogenous data generating process. It was extended to learning from endogenous data by Esponda and Pouzo (2016), which led to a renewed interest in misspecified learning in the economics literature. A key step in the analysis of such models is often to establish that Bayesian beliefs concentrate quickly around the KL minimizers, and as our Theorem 1 holds pathwise it immediately implies such a result. In Fudenberg, Lanzani, and Strack (2021b) we use the same concentration result to characterize the long-run beliefs of a correctly specified agent who has an imperfect and selective memory.

The Diaconis and Freedman (1990) result has been used in a number of economic applications, starting with the analysis of non-equilibrium learning in games in Fudenberg and Levine (1993). The result has also been used to analyze selective attention (Schwartzstein, 2014), the “fragility of agreement” (Acemoglu, Chernozhukov, and Yildiz, 2016), recursive utility functions (Al-Najjar and Shmaya, 2019), and persuasion (Schwartzstein and Sunderam, 2021). Our results enable extensions of these results to e.g. priors that impose parametric restrictions, or assume independence or finite support.

In Proposition 2 we leverage this result to show that the play of a Bayesian agent

1Frick, Iijima, and Ishii (2022) characterized the rates of convergence for finite-support priors and uses it to develop welfare comparisons for different departures from Bayesian updating.
2Subsequent papers include Fudenberg, Romanyuk, and Strack (2017), He (2020), Molavi (2019), Boluren and Hauser (2021), Frick, Iijima, and Ishii (2021), He and Libgober (2021), Esponda, Pouzo, and Yamamoto (2021), Heidhues, Kőszege, and Strack (2021), and Levy, Barreda, and Razin (2021). Before this, Arrow and Green (1973) gave the first general framework for this problem, and Nyarko (1991) pointed out that the combination of misspecification and endogenous observations can lead to cycles.
3For example, Theorem 1 provides a shorter way to prove Proposition 1 of Fudenberg, Lanzani, and Strack (2021a), which guarantees a uniform convergence rate on a probability one set of empirical frequencies.
converges to that predicted by the anticipated utility model commonly used in the macroeconomics literature (e.g. Kreps, 1998; Preston, 2005; Eusepi and Preston, 2018), and moreover we quantify the rate of convergence. This helps clarify when the anticipated utility model is a good approximation of rational play, and complements recent numerical studies on the same topic (Cogley, Colacito, and Sargent, 2007; Cogley and Sargent, 2008; Cogley, Colacito, Hansen, and Sargent, 2008).

Finally, Theorem 3 extends Theorem 1 (and thus Diaconis and Freedman, 1990) to beliefs that result from observing a Markov process whose transition probabilities are unknown. This provides a complement to recent work by Molavi (2019) and Esponda and Pouzo (2021) which studied analogs of Berk (1966) and Esponda and Pouzo (2016) for Markovian environments.

2 Setup

We study the time path of Bayesian beliefs when observing a sequence of subjectively i.i.d. data. Let $Y$ be a finite set of possible outcomes, and let $P = \Delta(Y)$ be the set of probability measures over $Y$ endowed with $\| \cdot \|$, the total variation norm on the space of signed measures.$^5$

Let $\mu_0 \in \Delta(P) = \Delta(\Delta(Y))$ denote a prior belief over distributions of outcomes, and $\Theta = \text{supp} \mu_0$ denote its support. A data set $y^t = (y_1, y_2, \ldots, y_t) \in Y^t$ is a vector of outcomes. For every data set $y^t$ we let $\mu_t$ be the posterior belief, which is required to satisfy Bayes rule whenever the denominator is different from 0:

$$\mu_t(C) = \frac{\int_{p \in C} \prod_{\tau=1}^{t} p(y_{\tau}) d\mu_0(p)}{\int_{p \in P} \prod_{\tau=1}^{t} p(y_{\tau}) d\mu_0(p)} .$$

(Bayes Rule)

The empirical distribution $f_t \in P$ is

$$f_t(z) = \frac{1}{t} \sum_{\tau=1}^{t} 1_{y_{\tau}=z} .$$

Our main result is that along any path of realized outcomes, the probability the posterior belief assigns to the outcome distributions that do not best approximate

---

$^5$For a signed measure $\chi$ on $Y$, $\| \chi \| = \frac{1}{2} \sum_{z \in Y} |\chi(z)|$. For every $X \subseteq \mathbb{R}_k$, we let $\Delta(X)$ denote the set of Borel probability distributions on $X$. 


the empirical distribution converges to zero at a uniform and exponential rate. To state this conclusion formally, we adopt the convention that \( \frac{0}{0} = 0 \) and \( 0 \log 0 = 0 \), and define \( H : P \times P \to \mathbb{R} \) to be the (possibly infinite) Kullback–Leibler divergence of \( q \) with respect to \( p \):

\[
H(q, p) = \sum_{z \in Y} q(z) \log \left( \frac{q(z)}{p(z)} \right).
\]

We denote by \( M : P \Rightarrow P \) the correspondence that maps a distribution \( q \) to the set of minimizers of the Kullback–Leibler divergence over the support of the prior:

\[
M(q) = \arg\min_{p \in \Theta} H(q, p).
\]

The log-likelihood assigned to the data set \( y^t \) under outcome distribution \( p \) is

\[
\log \left( \prod_{\tau=1}^{t} p(y_\tau) \right) = \sum_{z \in Y} t f_i(z) \log p(z) = -t H(f_i, p) + t \sum_{z \in Y} f_i(z) \log f_i(z). \tag{1}
\]

Minimizing the Kullback–Leibler divergence relative to the empirical distribution is hence the same as maximizing the log-likelihood assigned to the data set, so the KL minimizers \( M(f_i) \) at time \( t \) correspond to the outcome distributions that maximize the likelihood of \( y^t \). Throughout, we denote by \( B_\varepsilon(D) \) be the ball of radius \( \varepsilon \) around a set \( D \) in total variation norm.\(^6\) We denote by \( M_\varepsilon : P \Rightarrow P \) the correspondence that maps a distribution \( q \) to the set of distributions that come within \( \varepsilon \) of the minimum KL divergence:

\[
M_\varepsilon(q) = \left\{ p' \in \Theta : H(q, p') \leq \min_{p \in \Theta} H(q, p) + \varepsilon \right\}.
\]

### 3 The Rate of Convergence of Bayesian Beliefs

To show that Bayesian beliefs concentrate around the empirical distribution at a uniform rate, Diaconis and Freedman (1990) used the following condition:

**Definition 1** (\( \phi \) positivity). The prior \( \mu_0 \) is \( \phi \) positive if for \( \phi : \mathbb{R}_{++} \to \mathbb{R}_{++} \), we have \( \mu_0(B_\varepsilon(p)) \geq \phi(\varepsilon) \) for every \( p \in P \) and \( \varepsilon > 0 \).

\(^6\)\( B_\varepsilon(D) = \{ q \in \Delta(Y) : \exists p \in D, ||q - p|| \leq \varepsilon \} \).
Since \( \phi \) positivity requires the prior to assign strictly positive probability to every \( \varepsilon \) ball, it requires the prior to have full support.

**Theorem A** (Diaconis and Freedman 1990). For every \( \phi : \mathbb{R}^+ \to \mathbb{R}^+ \) and every \( \alpha \in (0, 1) \) there are \( \tilde{A} : \mathbb{R}^+ \to \mathbb{R}^+ \) and \( g : \mathbb{R}^+ \to \mathbb{R}^+ \) such that
\[
\frac{\mu_t(B_\varepsilon(f_t))}{1 - \mu_t(B_\varepsilon(f_t))} \geq \tilde{A}(\varepsilon) \exp(g(\varepsilon)t),
\]
for all \( \phi \) positive \( \mu_0, \varepsilon \in (0, 1), t \in \mathbb{N}, \) and \( f_t \in \Delta(Y) \).

Theorem A shows that for \( \phi \) positive priors, the probability that Bayesian beliefs assign to distributions that are more than \( \varepsilon \) away from the empirical distribution vanishes exponentially quickly, so it quantifies the speed at which a Bayesian with full support prior becomes more certain when observing i.i.d. data. The strength of this theorem lies in the fact that it holds not only in probability, but for every realization of outcomes.

Clearly \( \phi \) positivity plays a crucial role in Theorem A, as if the prior is not \( \phi \) positive the empirical distribution need not to be in its support, so beliefs cannot concentrate around it. However, requiring the prior to satisfy \( \phi \) positivity rules out several practically relevant cases. For example, \( \phi \) positivity cannot be satisfied if the prior has finite support, reduces the dimensionality of the problem, or is supported only on unimodal distributions. Moreover, models of misspecified learning suppose that the true data generating process is not in the support of the prior, which rules out \( \phi \) positivity. We extend Theorem A to cases where \( \phi \) positivity fails. Loosely, we require that either the prior gives all neighborhoods of a distribution sufficient weight or the prior gives zero weight to a small neighborhood of the distribution.

In Diaconis and Freedman (1990), Bayes rule is well defined everywhere. This is not true in our setup where the prior need not have full support. We define \( \Delta^\Theta(Y) \) to be the (compact) set of empirical frequencies such that Bayesian updating is well defined for a prior with support \( \Theta \).\(^7\) We next generalize the definition of \( \phi \) positivity to priors that do not necessarily have full or convex support.

**Definition 2** (\( \phi \) positivity on \( \Theta \)). The prior \( \mu_0 \) is \( \phi \) positive on \( \Theta \) if for \( \phi : \mathbb{R}^+ \to \mathbb{R}^+ \) we have \( \mu_0(B_\varepsilon(p)) \geq \phi(\varepsilon) \) for every \( p \in \Theta \) and \( \varepsilon > 0 \).

\(^7\)That is \( \Delta^\Theta(Y) = \{ q \in \Delta(Y) : \exists p \in \Theta, \text{supp } q \subseteq \text{supp } p \} \).
We will show that if beliefs are $\phi$ positive on $\Theta$, the posterior concentrates on $M_\varepsilon(f_t)$. Note that $\phi$ positivity on $\Theta$ reduces to $\phi$ positivity when $\Theta = P$, i.e. the prior has full support.

**Theorem 1.** For every $\phi : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ and every $\alpha \in (0, 1)$ there is a function $A : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ such that

$$\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} \geq A(\varepsilon) \exp(\alpha \varepsilon t)$$

for all $\mu_0$ that are $\phi$ positive on $\Theta$, $\varepsilon \in (0, 1)$, $t \in \mathbb{N}$, and $f_t \in \Delta^\Theta(Y)$.

Moreover, if $q := \inf_{q \in \Theta} \min_{z \in \text{supp} q} q(z) > 0$, then we can set

$$A(\varepsilon) = \phi(\min\{q/2, (1 - \alpha)\varepsilon\}q/2).$$

Theorem 1 only requires $\phi$ positivity on $\Theta$, in contrast to Theorem A, which assumes $\phi$ positivity on the whole space of distributions.\(^8\) When the prior is not $\phi$ positive on all of $\Delta(Y)$, beliefs need not concentrate around the empirical frequency, because this frequency might not be in the prior’s support. This is why Theorem 1 bounds the probability assigned to the distributions outside $M_\varepsilon(f_t)$, which are the $\varepsilon$ minimizers of the KL divergence, while Theorem A bounds the probability assigned to $B_\varepsilon(f_t)$, the $\varepsilon$ ball around the empirical distribution. Moreover, as Example 1 below shows, the theorem does not apply to the $\varepsilon$ ball $B_\varepsilon(M(f_t))$ around the exact minimizers $M(f_t)$, because when $\Theta$ is not convex points far from the minimizers can attain almost the same divergence.

The theorem implies that the probability assigned to all distributions that do not $\varepsilon$ best explain the empirical frequency $f_t$ vanishes at the exponential rate $\alpha \varepsilon$:

$$\mu_t(\Theta \setminus M_\varepsilon(f_t)) \leq \frac{1}{A(\varepsilon)} \exp(-\alpha \varepsilon t).$$

Notice that the multiplicative constant $A(\varepsilon)$ depends on the prior $\mu_0$ only through $\Theta$ and the function $\phi$. The second part of the statement guarantees that in the widely-

\(^8\)As with the original Diaconis and Freedman (1990) result, our Theorem 1 extends to the case of a compact set of outcomes under some additional regularity assumptions on the support of the prior. With an unbounded set of outcomes, the convergence is uniform with respect to the compact-open topology. That is, the rate of convergence is uniform only over compact subsets of empirical frequencies, because extreme realizations can induce arbitrarily large belief revisions, so our Lemma 2 becomes vacuous. Here both our theorem and that of Diaconis and Freedman (1990) fail.
studied case of finite support priors, there is an explicit formula to compute the rate of convergence as a function of Θ and φ, with the intuitive comparative statics that the rate of convergence improves when φ is higher and when the support is smaller.\(^9\)

The proofs of all our results are in the appendix. The proof of Theorem 1 has three steps. Step 1 proves a local Lipschitz property of the KL divergence, step 2 gives an explicit rate of concentration for each realized empirical frequency, while step 3 concludes by turning this explicit local rate of convergence into an exponential (but with possibly implicit constant) global rate of convergence.

Specifically, we first shows that Kullback-Leibler divergence \(H(p, q)\) is locally Lipschitz continuous in its second argument:

\[
|H(q, p) - H(q, \tilde{p})| \leq 2 \max_{z \in Y} \max \left\{ \frac{q(z)}{p(z)}, \frac{q(z)}{\tilde{p}(z)} \right\} ||p - \tilde{p}||. \tag{2}
\]

With this we are able to prove the next lemma, which is at the heart of our results.

**Lemma 1.** If \(\mu_0\) is \(\phi\) positive on \(\Theta\) then for every \(\varepsilon, \varepsilon', \kappa \in \mathbb{R}_+, \ t \in \mathbb{N}, \ f_t \in \Delta^Y, \ \bar{q} \in M_{\varepsilon'}(f_t)\), with \(\varepsilon' + \kappa \leq \varepsilon\),

\[
\frac{\mu_t(M_{\varepsilon}(f_t))}{1 - \mu_t(M_{\varepsilon}(f_t))} \geq \phi(\kappa/2R(f_t, \kappa, \bar{q})) \exp((\varepsilon - \kappa - \varepsilon')t) \tag{3}
\]

where

\[
R(f_t, \kappa, \bar{q}) = \max_{q \in \Theta \cap B_{\kappa}(\bar{q})} \frac{f_t(z)}{q(z)}.
\]

To prove the lemma, we use the Lipschitz condition (2) to establish that the Kullback-Leibler divergence from \(f_t\) is at most \(\min_{p \in \Theta} H(f_t, p) + \varepsilon' + \kappa\) in a ball of radius \(\kappa/2R(f_t, \kappa, \bar{q})\) around the \(\varepsilon'\)-minimizer \(\bar{q}\). Therefore, as by assumption \(M_{\varepsilon' + \kappa}(f_t) \subseteq M_{\varepsilon}(f_t)\), \(M_{\varepsilon}(f_t)\) contains a ball of radius \(\kappa/2R(f_t, \kappa, \bar{q})\) around \(\bar{q} \in \Theta\). As \(\mu_0\) is \(\phi\) positive on \(\Theta\), that ball has prior probability at least \(\phi(\kappa/2R(f_t, \kappa, \bar{q}))\), so the odds ratio between \(M_{\varepsilon}(f_t) \supseteq B_{\kappa/2R(f_t, \kappa, \bar{q})}(\bar{q})\) and \(\Theta \setminus M_{\varepsilon}(f_t)\) under the prior is at least

\[
\frac{\mu_0(M_{\varepsilon}(f_t))}{1 - \mu_0(M_{\varepsilon}(f_t))} \geq \frac{\mu_0(B_{\kappa/2R(f_t, \kappa, \bar{q})}(\bar{q}))}{1} \geq \phi(\kappa/2R(f_t, \kappa, \bar{q}))
\]

and this delivers the multiplicative constant in the RHS of the lemma. The exponential term follows from the fact that the increase in the posterior odds ratio for

\(^9\)The explicit formula for \(A(\varepsilon)\) yields a vacuous bound when \(\Theta = \Delta^Y\), as then \(q = 0\).
an arbitrary set \( D \) is growing exponentially in the difference between their KL divergence from the empirical distribution \( f_t \) and the probability distributions outside \( D \) according to

\[
\frac{\mu_t(D)}{1 - \mu_t(D)} = \frac{\int_{p \in D} \exp(-H(f_t, p)t) d\mu_0(p)}{\int_{p \notin D} \exp(-H(f_t, p)t) d\mu_0(p)},
\]

and from the fact that by definition, distributions outside \( M_\varepsilon(f_t) \) have a KL-divergence from \( f_t \) of at least \( \min_{p \in \Theta} H(f_t, p) + \varepsilon \). Thus at time \( t \) the odd ratio is equal to the prior odd ratio \( \phi(\kappa/2R(f_t, \kappa, \tilde{q})) \) multiplied \( t \) times the exponential of the difference in divergence between the two sets \( \varepsilon - \kappa - \varepsilon' \).

To derive Theorem 1 from Lemma 1, we need to bound the multiplicative constant away from 0 on \( \Delta^\Theta(Y) \). \(^{10}\) We prove this by contradiction, using the compactness of \( \Delta^\Theta(Y) \) and the lower semicontinuity of \( H \) to show there are \( c, \kappa > 0 \) such that for every \( f_t \in \Delta^\Theta(Y) \) we can pick \( \hat{q}_{f_t} \in M_{(1-\alpha)\varepsilon/2} \) such that \( R(f_t, \kappa, \hat{q}_{f_t}) < c \). In general, this \( \hat{q}_{f_t} \) may not be an exact KL minimizer for \( f_t \), since a minimizer \( q' \) too close (or on) the boundary of the simplex may have an excessively high value of the ratio \( f_t(z)/q'(z) \). Loosely speaking, moving away from the boundary decreases this ratio, and since the minimizers assign very low probability only to outcomes with very low probability under \( f_t \), this does have much effect on the KL fit, i.e., there is a \( (1-\alpha)\varepsilon/2 \)-minimizer sufficiently far from the boundary. Finally, the exponential bound obtained in the Lemma 1 does not directly relate the size of the minimization error \( \varepsilon \) to the concentration speed, while in Theorem 1 it scales linearly in \( \varepsilon \) (i.e. \( \exp(\alpha \varepsilon t) \) for some \( \alpha \)). \(^{11}\) This follows from the fact that \( \hat{q}_{f_t} \) can be chosen in \( M_{(1-\alpha)\varepsilon/2} \).

Theorem 1 only requires \( \phi \) positivity on \( \Theta \), in contrast to Theorem A, which assumes \( \phi \) positivity on the whole space of distributions. As mentioned above, misspecified beliefs need not concentrate around the empirical frequency; this is why Theorem 1 bounds the probability assigned to the distributions outside \( M_\varepsilon(f_t) \), which are the \( \varepsilon \) minimizers of the KL divergence, while Theorem A bounds the probability assigned to \( B_\varepsilon(f_t) \), the \( \varepsilon \) ball around the empirical distribution. In general, beliefs need not concentrate around the empirical frequency because this frequency might

\(^{10}\)This is not a concern in the lemma because \( f_t \in \Delta^\Theta(Y) \), which implies that there is at least one \( p \) in \( \Theta \) with finite Kullback-Leibler divergence from \( f_t \). Thus the \( \varepsilon' \)-minimizer \( \hat{q} \) also has a finite divergence from \( f_t \), and so has \( \hat{q}(z) > 0 \) for all \( z \in \text{supp } f_t \). The same then holds for the elements of \( B_\varepsilon(\tilde{q}) \) for sufficiently small \( \kappa \), so \( R \) is finite, and the bound given by the lemma is not trivial.

\(^{11}\)Although \( \varepsilon \) enters linearly in the exponential term of Lemma 1, different values of \( \varepsilon \) may need different values of \( \varepsilon' \) and \( \kappa \), so the overall effect of \( \varepsilon \) is not linear.
not be in the support of the prior. The next example illustrates why Theorem 1 does not apply to the \( \varepsilon \) ball \( B_\varepsilon(M(f_t)) \) around the exact minimizers \( M(f_t) \).

**Example 1.** Let \( Y = \{0, 1\} \), identify each \( p \in \Delta(Y) \) with the probability of \( y = 1 \), and let \( \mu_0(\{1/4\}) = \mu_0(\{3/4\}) = 1/2 \). Consider the sequence of outcomes \( (y_t)_{t=1}^\infty \) where \( y_t = 1 \) if \( t \) is odd and \( y_t = 0 \) if \( t \) is even. At every odd period \( 2t+1 \), \( M(f_{2t+1}) = \{3/4\} \), so for \( \varepsilon < 1/2 \), \( B_\varepsilon(M(f_{2t+1})) = \{3/4\} \) and

\[
\frac{\mu_{2t+1}(B_\varepsilon(M(f_{2t+1})))}{1 - \mu_{2t+1}(B_\varepsilon(M(f_{2t+1})))} = \frac{\mu_{2t+1}(\{3/4\})}{\mu_{2t+1}(\{1/4\})} = \frac{\mu_0(\{1/4\})(1/4)^t(3/4)^{t+1}}{\mu_0(\{3/4\})(1/4)^t+1(3/4)^t} = 3.
\]

Thus beliefs do not concentrate on the neighborhood of the KL minimizer.\(^{12}\)

However, when \( \Theta \) is convex and the empirical distribution has full support, Corollary 1 shows that the beliefs do concentrate around the (unique) exact minimizer, though the relation between the size of the ball and the rate depends on how close \( f_t \) is to the boundary of the probability simplex.

## 4 Implications of Theorem 1

We now discuss some implications of Theorem 1. We start with its relation to the classic results of Diaconis and Freedman (1990) and Berk (1966). We then discuss the case of beliefs about parameters, followed by how Theorem 1 relates to the anticipated utility model used in macroeconomics, then explain how the theorem can be used to analyze learning from endogenous data.

### 4.1 Classic Results

The most direct implication of Theorem 1 is Theorem A, which is the special case where \( \Theta = \Delta(Y) \). Here we use Pinsker’s inequality (which bounds the KL divergence of \( q \) from \( f_t \) as a function of \( \|q - f_t\| \)) and the fact that for a full support prior, \( M(f_t) = \{f_t\} \), i.e. the unconstrained minimizer of the Kullback-Leibler divergence is the distribution itself.

---

\(^{12}\)In this example, \( \Theta \) is not connected, but Example 4 in the Appendix shows that the same problem can arise when it is. The example there extends this one to a setting with a third outcome, and specifies a \( \Theta \) that connects 1/4 and 3/4 through a set of distributions that are not relevant under the empirical frequency described above.
Proof of Theorem A. Consider $\varepsilon \in (0, 1)$ and a $\phi$ positive prior $\mu_0$. As $H(f_t, f_t) = 0$, all $p \in M_\varepsilon(f_t)$ satisfy $H(p, f_t) \leq \varepsilon$. Pinsker’s inequality implies that $M_\varepsilon(f_t) \subseteq B_{\sqrt{\varepsilon/2}}(f_t)$. Defining $\bar{\varepsilon} = \sqrt{\varepsilon/2}$, by Theorem 1 there exists $A : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ such that

$$\frac{\mu_t(B_\varepsilon(f_t))}{1 - \mu_t(B_\varepsilon(f_t))} \geq \frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} \geq A(\varepsilon) \exp\left(\frac{\varepsilon t}{2}\right) = A(2\bar{\varepsilon}^2) \exp(\bar{\varepsilon}^2 t).$$

The result follows by letting $\tilde{A}(\varepsilon) = A(2\varepsilon^2)$ and $g(\varepsilon) = \bar{\varepsilon}^2$.

Our result is also closely related to the seminal work by Berk (1966) on long-run beliefs in a misspecified model. The paper showed that when the objective data generating process is i.i.d., beliefs almost surely concentrate on every $\varepsilon$ ball around the set of KL minimizers relative to the true outcome distribution $p^*$ provided that the prior is “sufficiently regular.” We say that $\Theta$ is regular if for every $p \in \Theta$ and $\varepsilon \in (0, 1)$ there is $p' \in \Theta$ with $\text{supp } p' = Y$ such that $||p - p'|| \leq \varepsilon$.

Theorem B (Berk 1966). If $\Theta$ is regular then for all $\varepsilon \in (0, 1)$

$$\lim_{t \rightarrow \infty} \mu_t(B_\varepsilon(M(p^*))) = 1 \quad p^*\text{-a.s.}$$

Theorem 1 lets us use the assumption of $\phi$ positivity to add a rate of convergence to Theorem B when the number of outcomes is finite, as we assume throughout. The rate of convergence has important implications when the beliefs of an agent are used to solve a decision problem, as it lets us bound the probability of choosing actions that are not optimal with respect to the KL-minimizers as a function of how many outcomes have been observed.

Theorem 2. Let $\mathbb{P}$ be the probability measure induced if the outcomes are i.i.d. draws from $p^*$. If $\Theta$ is regular and $\mu_0$ is $\phi$ positive on $\Theta$, then for every $\varepsilon \in (0, 1)$ there is a $K \in \mathbb{R}_{++}$ such that

$$\mathbb{P}\left[\mu_t\left(B_\varepsilon(M(p^*))\right) < 1 - K \exp(-Kt)\right] = O(\exp(-Kt)).$$

The idea is that if $\mu_0$ is $\phi$ positive on $\Theta$ then for all $\alpha \in (0, 1)$ there exists a
function $A : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ such that for all $\epsilon \in (0, 1), t \in \mathbb{N}$ and $f_t \in \Delta^\Theta(Y)$

$$
\mu_t(M_{\epsilon}(f_t)) \geq 1 - \frac{1}{A(\epsilon)} \exp(-\alpha \epsilon t).
$$

We then show there is $\hat{\epsilon}$ such that if the empirical frequency is in an $\hat{\epsilon}$ ball around the objective distribution, i.e. $f_t \in B_{\hat{\epsilon}}(p^*)$, then $M_{\hat{\epsilon}}(f_t) \subseteq B_{\epsilon}(M(p^*))$, so we can use Theorem 1 and Sanov’s theorem to obtain the stated conclusion.

### 4.2 Parametric Models

**Finite Support Priors** In some problems the reasonable priors have finite support, and are in that sense “parametric” whether or not they are correctly specified. For example, if outcomes correspond to the color of balls drawn with replacement from an urn with known size but with unknown composition, then the reasonable priors have finite support. Also, whenever the agent put a qualitative restriction on the correlation between two variables (e.g., postulates that they are positively correlated) the Diaconis and Freedman (1990) result does not apply, while ours does.

**Smooth Parametric Models** In our setting of a finite number of outcomes, we can view the probability distributions themselves as the parameters, and take $\Xi = \Delta(Y)$. However, when the size $\#Y$ of the outcome space is large, so is the dimension of $\Delta(Y)$, and people tend to use lower-dimensional parametric models to make the distribution of outcomes easier to think about and analyze. Here the Diaconis and Freedman (1990) result does not apply, but our extension guarantees that beliefs concentrate exponentially fast on the KL-minimizing parameters.

Consider a parametric model where $\Theta = \{p_\theta : \theta \in \Xi\}$ with $p$ differentiable in $\theta$, and $\Xi \subset \mathbb{R}^k$ closed and convex. Let

$$
\tilde{H}(f, \theta) = H(f, p_\theta),
$$

and define

$$
\theta^*(f) = \arg\min_{\theta \in \Xi} \tilde{H}(f, \theta)
$$

to be the parameters that minimize the KL divergence between the $p_\theta$ and $f$. 

11
**Definition 3.** $\tilde{H}$ is uniformly strongly $m$-convex on $\mathcal{F} \subseteq \Delta(Y)$ if there is $m > 0$ such that for all $f \in \mathcal{F}$,

$$(\nabla_{\theta} \tilde{H}(f, \theta) - \nabla_{\theta} \tilde{H}(f, \theta'))^T(\theta - \theta') \geq m||\theta - \theta'||^2$$

for all $\theta, \theta' \in \Xi$.

In the single dimensional case $\theta \in \mathbb{R}$, strong $m$-convexity requires that the second derivative of $\tilde{H}$ is bounded away from zero. In the multidimensional case strong $m$-convexity is equivalent to the smallest eigenvalue of the second derivative being greater than $m$. Uniform $m$-convexity extends this property to parametric models.

We define $B_{\varepsilon}(\theta) = \{p_0 : ||\eta - \theta||_2 \leq \varepsilon\}$ to be the set of all distributions whose parameter are at most $\varepsilon$ away from the $\theta$. We then have the following result establishing concentration of the beliefs in the parameter space around the KL minimizer:

**Proposition 1.** If $\tilde{H}$ is uniformly strongly $m$-convex on $\mathcal{F}$ then for every $\phi : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ and every $\alpha \in (0, 1)$

$$\frac{\mu_0 \left(B_{\sqrt{2\varepsilon/m}}(\theta^*(f_t))\right)}{1 - \mu_0 \left(B_{\sqrt{2\varepsilon/m}}(\theta^*(f_t))\right)} \geq A(\varepsilon) \exp(\alpha \varepsilon t)$$

for all $\mu_0$ that are $\phi$ positive on $\Theta$, $\varepsilon \in (0, 1)$, $t \in \mathbb{N}$, and $f_t \in \Delta^\Theta(Y) \cap \mathcal{F}$, where $A$ is the function whose existence is guaranteed by Theorem 1.

To see why the proposition is true, note that because $\tilde{H}$ is continuous and $\Xi$ is convex, when $\tilde{H}$ is strongly $m$-convex the KL minimizer $\theta^*(f_t)$ is a singleton. In addition, the convexity of $\Xi$ guarantees that small movements away from the minimizer to other parameters in $\Xi$ increase the KL divergence. Uniform strong convexity provides a lower bound on this increase, so we can conclude that parameters outside of $B_{\sqrt{2\varepsilon/m}}(\theta^*(f_t))$ are not $\varepsilon$ minimizers. The proposition then follows from Theorem 1.

The proposition does not assume that $\Xi$ is a convex subset of $\Delta(Y)$, but it has additional implications when it is.

Note that if $\Xi \subseteq \Delta(Y)$, then for all $m \in (0, 1)$, $\tilde{H}$ is uniformly strongly $m$-convex on $\mathcal{F} = \{f : \min_{z \in Y} f(z) \geq m > 0\}$ with constant $m = \min_{z \in Y} f(z)$. This observation gives the following corollary.
Corollary 1. Suppose that $\Theta \subseteq \Delta(Y)$ is convex and that $\mu_0$ is $\phi$ positive on $\Theta$. Then for every $\alpha, \varepsilon \in (0, 1)$

$$
\frac{\mu_t \left( B_{\sqrt{\varepsilon/\min_z f_t(z)}(M(f_t))} \right)}{1 - \mu_t \left( B_{\sqrt{\varepsilon/\min_z f_t(z)}(M(f_t))} \right)} \geq \exp(\alpha \varepsilon t)
$$

for all $t \in \mathbb{N}$ and $f_t \in \Delta^\Theta(Y)$ such that $\min_z f_t(z) > 0$, where $A$ is the function whose existence is guaranteed by Theorem 1.

The corollary delivers a uniform rate of convergence for all sequences of outcomes for which $\lim \inf_t \min_z f_t(z)$ is bounded away from 0. Some forms of prior restrictions do correspond to a convex $\Theta$, for example if the outcomes are linearly ordered and the prior assigns probability 1 to the distributions where the probability of each outcome is weakly increasing or to all distributions that first-order stochastically dominate a fixed alternative. Other sorts of prior restrictions can lead to sets $\Theta$ that are connected but not convex; for example the prior might only assign positive probability to unimodal distributions. Example 5 in the Appendix shows that convergence need not be uniform over frequencies that approach the boundary of the simplex. When $f_t$ assigns vanishingly small probability to some outcomes, distributions that are very different from the minimizers in terms of total variation norm but are similar on the outcomes that have non-trivial empirical frequency can have non-vanishing posterior probability.

The next example considers a parametric $\Xi$ that is not a subset of $\Delta(Y)$. Here the support restriction comes from the assumption that successive trials are i.i.d., which is a way of simplifying a complex environment.

Example 2. (Bernoulli trials) Suppose the outcome $y \in Y = \{1, \ldots, \#Y\}$ corresponds to the number of Bernoulli trials needed to get one success, with $y = \#Y - 1$ denoting the maximum number of allowed trials. If the agent believes the trials are independent, their subjective distribution for outcome $y$ is a truncated geometric

---

13Beliefs that are product measures need not have convex support. For example, let $Y = \{0, 1\} \times \{0, 1\}$, let $p$ be the uniform probability on $Y$ and $q = \delta_{(1,1)}$. Then $p$ and $q$ are product measures while $p/2 + q/2$ is not. However, the conclusion of the proposition obtains when $Y$ is the Cartesian product of sets $Y_i$, and $\Theta$ is composed of product measures of convex subsets of the $\Delta(Y_i)$.

14So when $y = \#Y$, no success occurred in the allowed $\#Y - 1$ trials.
distribution

\[ p_\theta(y) = \begin{cases} 
\theta(1 - \theta)^{y-1} & \text{for } y < \#Y \\
(1 - \theta)^y & \text{for } y = \#Y 
\end{cases} \]

and the support of the agent’s prior \( \Theta \) only includes these distributions, i.e. \( \Theta \subseteq \{p_\theta : \theta \in [0, 1] \} \). No prior with this \( \Theta \) can satisfy the \( \phi \)-positivity condition of Diaconis and Freedman (1990), but our results apply if the prior \( \mu_0 \) has a density that is bounded away from zero on \([0, 1]\) or is Beta.\(^{15}\) We have

\[ H(f, \theta) = f(\#Y)\#Y \log(1 - \theta) + \sum_{z=1}^{\#Y-1} f(z)[(z - 1) \log(1 - \theta) + \log(\theta)] - \sum_{z=1}^{\#Y} z \log(f(z)) \]

\[ = \log(1 - \theta) \left[ \sum_{z=1}^{\#Y} zf(z) \right] + \log \left( \frac{\theta}{1 - \theta} \right) (1 - f(\#Y)) - \sum_{z=1}^{\#Y} z \log(f(z)). \]

This function is uniformly strongly 1-convex, and the first-order condition shows that the unique KL minimizing parameter is given by

\[ \theta^*(f) = \frac{1 - f(\#Y)}{\sum_{z \in Y} zf(z)}. \]

Thus from Proposition 1 beliefs about \( \theta \) concentrate on any \( B_{\sqrt{\varepsilon}}(\theta^*(f)) \) exponentially fast. Moreover, if the data has no realizations of \( \#Y \), the KL minimizer is the reciprocal of the average outcome \( \sum_{z \in Y} zf(z) \). This is intuitive, as the expectation of a geometric distribution is the reciprocal of the parameter \( \theta \), i.e. \( \lim_{\#Y \to \infty} \sum_{z \in Y} p_\theta(z)z = 1/\theta. \)

The next support restriction arises from a different way of simplifying a complex environment.

**Example 3.** (Oversimplified Predictors) Suppose that the outcome \( y = (y_0, \ldots, y_n) \in Y = \prod_{i=0}^n Y_i \) has \( n + 1 \) dimensions, and each dimension \( i \) has \( D_i \) possible outcomes. The agent is trying to learn the relationship between dimensions \((y_1, \ldots, y_n)\) and \( y_0\), and postulates that the distribution of \( y_0 \) is a polynomial function of \((y_1, \ldots, y_n)\)

\[ p_{y_0}(y_{0,i}) = \alpha_{0,i} + \alpha_{1,i} y_1 + \ldots + \alpha_{l,n} y_n + \beta_{1,2,i} y_1 y_2 + \beta_{1,3,i} y_1 y_3 + \ldots \]

\(^{15}\)Of course, the particular function \( \phi \) will change. If the density is bounded, \( \phi \) can be chosen linear in \( \varepsilon \), while for Beta priors \( \phi \) can be chosen to be a power function of \( \varepsilon \).
so each parameter vector maps to a conditional distribution for \( y_0 \). Any initial belief \( \nu \) about the parameter values together with a distribution over \( \times_{i=1}^{n} Y_i \) corresponds to some \( \mu \in \Delta(P) \), but if the agent uses a model that only gives positive probability to polynomials up to a order \( K \), the resulting \( \mu \) will not have full support. By Theorem 1 the beliefs will concentrate around the \( \varepsilon \) minimizing predictors uniformly over all the empirical frequencies. Moreover, as \( p \) is a linear function of the parameters, and \( H \) is uniformly strongly convex on the subsets of frequencies that are bounded away from the boundaries, by Proposition 1 the belief will concentrate on the ball around the unique minimizer.

4.3 Anticipated Utility

The macroeconomics literature on “anticipated utility” (Kreps, 1998) assumes that agents in the economy choose actions that maximize their payoff under a point estimate that maximizes the likelihood of their sample, ignoring uncertainty about the state. This is a simpler problem than the maximization of expected utility, and the reduction in complexity and dimension makes anticipated utility models more tractable and easier to analyze. However, it has not been clear how much error the approximation induces. For example, Cogley and Sargent (2008) wrote

Macroeconomists might justify anticipated-utility models as an approximation to a correctly formulated Bayesian decision problem... (the models) would be more compelling if one could also show that anticipated-utility decisions well approximate Bayesian decisions. As far as we know, no one has assessed the quality of the approximation...

There is also a small literature that addresses this question using numerical simulations (Cogley, Colacito, and Sargent, 2007; Cogley and Sargent, 2008; Cogley, Colacito, Hansen, and Sargent, 2008). Our result on Bayesian updating can be used to derive analytical results that complement these numerical studies. Corollary 1 implies that the long-run behavior under anticipated utility models converges to that of an expected utility maximizer when the empirical distribution has full support.\(^{16}\)

\(^{16}\)This does not follow from Theorem 1, because the best response to a parameter that \( \varepsilon \) minimizes the KL divergence can be very different than the best response to parameters in the neighborhood of an exact KL minimizer. For instance, if we augment Example 1 with a two-action decision problem in which the first action is a strict best reply to a belief that assigns probability 3/4 to the
This provides a formal justification for the use of anticipated can be used as an approximation of expected utility models in studies of long-run behavior.

Our results can be used to study how quickly an optimizing Bayesian agent approximates this behavior. To develop this link, suppose that in each period \( t \in \{1, 2, 3, \ldots \} \) the agent chooses an action from \( A \). We assume that \( A \) is a convex set, endowed with a metric \( d \) that makes it a compact set. The action does not affect the outcome distribution but, paired with the realized outcome, determines the flow payoff of the agent via a utility function \( u : A \times Y \to \mathbb{R} \) that is strictly concave in \( a \). Let \( A^*(\nu) \) denote the (unique) optimal action given belief \( \nu \), i.e.,

\[
A^*(\nu) = \arg\max_{a \in A} \int_P \mathbb{E}_p[u(a, y)] d\nu(p)
\]

and suppose \( A^* \) is Lipschitz continuous when \( \Delta(P) \) is endowed with the topology of weak convergence of measures.\(^\text{17}\)

Let \( A^*(M(f_t)) \) denote the action that is optimal for a point belief at the likelihood maximizer \( M(f_t) \).

**Proposition 2.** Suppose that \( \Theta \) is convex and that \( \mu_0 \) is \( \phi \) positive on \( \Theta \). Then for all \( \varepsilon, c > 0 \) there is a time period \( T \) such that \( d(A^*(\mu_t), A^*(M(f_t))) \leq \varepsilon \) for every \( t > T \) and every \( f_t \in \Delta^\Theta(Y) \) with \( \min_{z \in Y} f_t(z) > c \).

### 4.4 Learning from Endogenous Data

Theorem 1 can be used to study the limit points of misspecified learning when the distribution of outcomes depends on the action played by an agent, and that action depends on the agent’s beliefs. For example, the agent might be a customer who wants to learn which of two products she prefers, decides every period which one to buy, and receives a signal about the product they bought. If a fixed action \( a \) is chosen in every period after some random time \( \tau \), the situation is much like the one we considered in Theorem 2, with the posterior belief at time \( \tau \) taking the role of the prior, and the objective outcome distribution under action \( a \) taking the role of the distribution \( 3/4 \), while the second action is the strict best reply to the Dirac belief over \( 3/4 \), then along the outcome path considered in the example the agent’s best response converges to the first action, while the adaptive learning heuristic would prescribe the second action in all odd periods.

\(^\text{17}\)A sufficient condition for this is that \( A \subseteq \mathbb{R} \) is a compact interval, \( Y \) can be ordered to make \( u \) supermodular, and for all \( y \in Y \) \( u(\cdot, y) \) is differentiable with a bounded derivative, see, e.g. Frankel, Morris, and Pauzner (2003).
fixed distribution over outcomes \( p^* \). To understand if the action \( a \) can be played in the long-run we need to understand whether the resulting process of beliefs makes it optimal to play \( a \).

Suppose that there are multiple KL minimizers given the action \( a \). From the central limit theorem, when \( a \) is played every period, the empirical outcome distribution generated by \( a \) converges to \( p^* \) at rate \( \sqrt{t} \). Thus there will be infinitely many oscillations of the empirical frequency of size \( \frac{1}{\sqrt{t}} \) from \( p^* \) in the direction of every KL minimizer. If the beliefs resulting from these empirical frequencies concentrate around the minimizer sufficiently quickly, any optimal action must be optimal against a point belief at this KL minimizer. As a consequence, a limit action must be a best reply to all of the KL minimizers if the beliefs concentrate around the minimizers more rapidly than the empirical distribution oscillates. Fudenberg, Lanzani, and Strack (2021a) showed that this is the case when the prior has subexponential decay, i.e. if \( \lim \inf_{t \to \infty} \phi(t) \exp(t) = \infty \). That result follows immediately from the next proposition, which in turn is an easy consequence of Theorem 1. This makes the logic of the result easier to see, and shows how it relates to Theorem A.

For every outcome distribution \( q \) and every \( \alpha \in (0, 1) \), denote the convex combination of \( q \) with the outcome distribution \( p^* \) by \( p_{q,\alpha}^* = \alpha q + (1 - \alpha) p^* \).

**Proposition 3.** If \( \text{supp } p = \text{supp } p^* \) for every \( p \in \Theta \) and \( \mu_0 \) is \( \phi \) positive on \( \Theta \), for every \( q \in M(p^*) \), \( \varepsilon \in (0, 1/4) \)

\[
\frac{\mu_t(B_{\varepsilon}(q))}{1 - \mu_t(B_{\varepsilon}(q))} > \phi \left( \frac{q\varepsilon^2}{2\sqrt{t}} \right) \exp \left( \varepsilon^2 \sqrt{t}/2 \right),
\]

for all \( f_t \in B_{\varepsilon^2/(2\log q\sqrt{t})} \left( p_{q,1/\sqrt{t}}^* \right) \cap \Delta^\Theta(Y) \).

The proof has several steps. We first show a form of Lipschitz continuity of the KL divergence when the distributions involved are absolutely continuous. Then we generalize Pinsker’s inequality to obtain a lower bound on how much the KL divergence changes as the distribution moves away from the KL minimizers.\(^\text{18}\) Then we apply Theorem 1 to get an exponential rate for the concentration around the unique minimizer \( q \) when the empirical frequency is exactly \( p_{q,1/\sqrt{t}}^* \), and use the Lipschitz continuity and minimal increase conditions to extend the result to a ball of size \( \varepsilon^2/(4 \log q(y) \sqrt{t}) \) around \( p_{q,1/\sqrt{t}}^* \).

\(^\text{18}\)Pinsker’s inequality is the special case in which the unique KL minimizer for \( p \) is \( p \) itself.
5 Subjectively Markovian Environments

We now show how to generalize Theorem 1 to the case of beliefs that the signals $y$ are generated by a Markov process, which is a key environment in macroeconomics. For example, this is the setting where Cogley and Sargent (2008) numerically analyzed the approximation properties of the anticipated utility model.

In the Markov setting the agent is learning about $\#Y$ different outcome distributions; let $\mathcal{P} = \Delta(Y)^Y$ be the set of transition matrices over $Y$ endowed with the total variation norm.\(^{19}\) Let $\mu_0 \in \Delta(\mathcal{P}) = \Delta(\Delta(Y)^Y)$ denote a prior distribution over transition matrices and $\Theta = \text{supp} \mu_0$ its support.\(^{20}\)

To initialize the process, we fix an observed period 0 outcome $y_0$. For every data set $y^t$ we let $\mu_t$ be the posterior belief, which is required to satisfy Bayes rule whenever the data set has positive prior probability:

$$\mu_t(C) = \frac{\int_{\pi \in \Theta} \prod_{r=1}^t \pi(y_r | y_{r-1}) d\mu_0(\pi)}{\int_{\pi \in \Theta} \prod_{r=1}^t \pi(y_r | y_{r-1}) d\mu_0(\pi)}. \quad \text{(Bayes Rule)}$$

The empirical transition distribution $f_t \in \Delta(Y \times Y)$ is

$$f_t(z, z') = \frac{1}{t} \sum_{\tau=1}^t 1_{y_\tau = z', y_{\tau-1} = z}.$$

We define $\mathcal{H} : \Delta(Y \times Y) \times \mathcal{P} \rightarrow \mathbb{R}$ as

$$\mathcal{H}(f, \pi) = - \sum_{z \in Y} f(z, z') \log \left( \pi(z' | z) \right).$$

The function $\mathcal{H}$ generalizes $H$ to the non-i.i.d. case, as $\mathcal{H}(f, \pi)$ measures the log-likelihood assigned to the empirical transitions distribution $f$ given the transition probability $\pi$:

$$\log \left( \prod_{\tau=1}^t \pi(y_{\tau} | y_{\tau-1}) \right) = t \sum_{(z, z') \in Y \times Y} f_t(z, z') \log \pi(z' | z) = -t \mathcal{H}(f_t, \pi).$$

\(^{19}\)That is, for all $\chi \in (\mathbb{R}^Y)^Y$, $||\chi|| = \frac{1}{2} \sum_{z \in Y} |\chi(z'|z)|$.

\(^{20}\)Note that the $\mu_0$ need not be a product measure, and that this reduces to the subjectively i.i.d. environment of the previous sections if for every $\pi \in \Theta$ and every $z, z' \in Y$, $\pi(\cdot | z) = \pi(\cdot | z')$. 18
We denote by $\mathcal{M} : \Delta(Y \times Y) \rightarrow \mathcal{P}$ the correspondence that maps an empirical transition distribution $f$ to the minimizers of $\mathcal{H}$ over the support of the prior: $\mathcal{M}(f) = \text{argmin}_{\pi \in \Theta} \mathcal{H}(f, \pi)$, and let $\mathcal{M}_\varepsilon(f)$ be the set of distributions that come within $\varepsilon$ of the minimum of $\mathcal{H}$:

$$\mathcal{M}_\varepsilon(f) = \left\{ \pi' \in \Theta : \mathcal{H}(f, \pi') \leq \min_{\pi \in \Theta} \mathcal{H}(f, \pi) + \varepsilon \right\}.$$

Let $\Delta^\Theta(Y \times Y)$ denote the set of empirical transition distributions for which Bayes rule is well defined. That is, $f \in \Delta^\Theta(Y \times Y)$ if there is a $\pi \in \Theta$ such that for all $(z, z') \in \text{supp} f$, $\pi(z'|z) > 0$.

**Theorem 3.** Suppose that for all $\pi, \pi' \in \Theta$, $z, z' \in Y$, $\pi(z'|z) > 0$ if and only if $\pi'(z'|z) > 0$ and that $\mu_0$ is $\phi$ positive on $\Theta$. Then for all $\alpha \in (0, 1)$ there is a function $A : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ such that

$$\frac{\mu_t(\mathcal{M}_\varepsilon(f_t))}{1 - \mu_t(\mathcal{M}_\varepsilon(f_t))} \geq A(\varepsilon) \exp(\alpha \varepsilon t),$$

for all $\varepsilon > 0, t \in \mathbb{N}$ and $f_t \in \Delta^\Theta(Y \times Y)$.

The proof of this result is similar in spirit to the proof of Theorem 1, because we can consider the data set to be a sequence of pairs $(y_t, y_{t+1})$ in place of a sequence of $y_t$.

### 6 Conclusion

We have shown that under weak conditions on the prior, for every realization of the data, Bayesian beliefs concentrate exponentially quickly on the models that best explain the empirical frequency of outcomes. One implication of this concentration result is that optimal actions can be determined directly from the empirical frequency without computing beliefs. More precisely, once the sample is sufficiently large, neither the exact sample size nor calendar time is needed to compute the optimal action; the empirical frequency is sufficient. As the dynamics and distribution of the empirical frequency are well understood, this insight can greatly simplify the analysis of the long-run behavior of Bayesian agents.

In addition to the applications in Section 4, Theorem 1 may allow generalizations of other results from the literature on misspecified Bayesian agents who learn from
endogenous data. One recurrent theme in this literature is the possibility that when actions are endogenous, misspecified beliefs can lead to cycles in setting that would not occur with correctly specified beliefs, because repeated play of an action generates evidence in favor of another action.\textsuperscript{21} In such situations our concentration result may be used to bound the number of periods spent in each phase of the cycles. This would complement Esponda, Pouzo, and Yamamoto (2021), which characterized the asymptotic frequencies of these cycles when the space of beliefs can be partitioned into a finite number of attracting sets and the support of the prior is one dimensional. Our uniform speed of convergence result might be useful in extending this to more general settings. In addition, as we provide a concentration bound for every finite time, our result can be used to characterize behavior in the “medium-run” before the asymptotic results apply.

Mazumdar, Pacchiano, Ma, Bartlett, and Jordan (2020) proves that with high probability, the posteriors of a correctly-specified Bayesian concentrate around the true parameter at rate $\sqrt{n}$, and uses this result to study the long-run properties of Thompson sampling. The paper allows for infinitely many outcomes, but imposes additional strong conditions such as log-concavity of the true data generating process, and the a prior density that is bounded away from 0. Our results immediately enable extensions to Thompson sampling with less restricted priors in the finite outcome case.

In settings where multiple agents choose their actions based on the same observables, our concentration results can be used to quantify the minimal extent of the differences in their prior beliefs needed to rationalize different choices. For example, Montiel Olea, Ortoleva, Pai, and Prat (2021) showed that when observing signals of an object’s value, misspecified agents with lower dimensional models have a higher willingness to pay after the first few observations, while correctly specified agents have a higher willingness to pay in the long-run; our result on the speed of convergence may help to better identify the switching time.

The learning in games literature has assumed correctly specified beliefs in order to appeal to Diaconis and Freedman (1990). Our generalization will facilitate the extension of the results from this literature to cases where the agents in the learning model have misspecified beliefs about the extensive form of the game. It will also

\textsuperscript{21}See e.g. Nyarko (1991), Fudenberg, Romanyuk, and Strack (2017), and Levy, Razin, and Young (2020).
enable extensions to incorrect beliefs about a complex network structure in Bowen, Dmitriev, and Galperti (2021), and to overconfident agents as in Heidhues, Köszegi, and Strack (2018).

7 Appendix

7.1 Properties of the KL divergence

Lemma 2. For all $p, \tilde{p}, q \in P$

$$|H(q, p) - H(q, \tilde{p})| \leq 2 \max_{z \in Y} \max \left\{ \frac{q(z)}{p(z)}, \frac{q(z)}{\tilde{p}(z)} \right\} \|p - \tilde{p}\|.$$ 

Proof. Let $R := \max_{z \in Y} \max \left\{ \frac{q(z)}{p(z)}, \frac{q(z)}{\tilde{p}(z)} \right\}$, $\hat{Y} = \{z : p(z) > \tilde{p}(z)\}$, and suppose without loss of generality that $p(\text{supp } q) \geq \tilde{p}(\text{supp } q)$. Then

$$|H(q, p) - H(q, \tilde{p})| = \sum_{z \in \text{supp } q} \left( \log \left( \frac{p(z)}{q(z)} \right) - \log \left( \frac{\tilde{p}(z)}{q(z)} \right) \right) q(z) = \sum_{z \in \text{supp } q} \left( \frac{p(z)}{q(z)} - \frac{\tilde{p}(z)}{q(z)} \right) q(z)$$

$$\leq R \sum_{z \in \text{supp } q} \left( \frac{p(z)}{q(z)} - \frac{\tilde{p}(z)}{q(z)} \right) q(z) = R \sum_{z \in \text{supp } q} \left( 2\hat{Y} - 1 \right) \left( \frac{p(z)}{q(z)} - \frac{\tilde{p}(z)}{q(z)} \right) q(z)$$

$$= R \sum_{z \in \text{supp } q} \left( 2\hat{Y} - 1 \right) p(z) - R \sum_{z \in \text{supp } q} \left( 2\hat{Y} - 1 \right) \tilde{p}(z)$$

$$= 2R \left[ \sum_{z \in \text{supp } q} \hat{Y} p(z) - \sum_{z \in \text{supp } q} \hat{Y} \tilde{p}(z) \right] + R(\tilde{p}(\text{supp } q) - p(\text{supp } q))$$

$$\leq 2R \left[ \sum_{z \in \text{supp } q} \hat{Y} p(z) - \sum_{z \in \text{supp } q} \hat{Y} \tilde{p}(z) \right] \leq 2R \left[ \sum_{z \in Y} p(z) - \sum_{z \in Y} \tilde{p}(z) \right]$$

$$= 2R \|p - \tilde{p}\|.$$

The last equality follows directly from the definition of the total variation norm. ■
Lemma 3. For every $q, q', p, p' \in P$

$$\left| (H(p, q) - H(p, q')) - (H(p', q) - H(p', q')) \right| \leq 2\|p - p'\| \max_{y \in \text{supp} p \cup \text{supp} p'} \left| \log \frac{q'(y)}{q(y)} \right|.$$ 

Proof. We have

$$\left| (H(p, q) - H(p, q')) - (H(p', q) - H(p', q')) \right|$$

$$= \left| \sum_{z \in Y} p(z) \log \frac{p(z)}{q(z)} - \sum_{z \in Y} p(z) \log \frac{p(z)}{q'(z)} - \sum_{z \in Y} p'(z) \log \frac{p'(z)}{q(z)} + \sum_{z \in Y} p'(z) \log \frac{p'(z)}{q'(z)} \right|$$

$$= \left| \sum_{z \in Y} p(z) \log \frac{q'(z)}{q(z)} - \sum_{z \in Y} p'(z) \log \frac{q'(z)}{q(z)} \right| = \left| \sum_{z \in Y} (p(z) - p'(z)) \log \frac{q'(z)}{q(z)} \right|$$

$$\leq \sum_{z \in Y} \left| p(z) - p'(z) \right| \cdot \left| \log \frac{q'(z)}{q(z)} \right| \leq 2\|p - p'\| \max_{y \in \text{supp} p \cup \text{supp} p'} \left| \log \frac{q'(y)}{q(y)} \right|. \quad \blacksquare$$

Lemma 4. Let $p \in P$, and let $D \subseteq P$ be nonempty. Then $\Gamma_D(p) = \arg\min_{q \in D} H(p, q)$ is nonempty and compact-valued, and if $D$ is regular $\Gamma_D(p)$ is upper hemicontinuous.

Proof. If $H(p, q) = \infty$ for all $q \in D$, $\Gamma_D(p) = D$ is nonempty. If there is $\hat{q}$ such that $H(p, \hat{q}) = K < \infty$, the set $D' = \{q \in D : H(p, q) \leq K\}$ is compact by the continuity of $H(p, \cdot)$, so the continuous and real-valued restriction of $H(p, \cdot)$ to $D'$ attains a minimum, and the set of minimizers is compact. That $\Gamma_D$ is upper hemicontinuous if $D$ is regular follows from Lemma 1 in Esponda and Pouzo (2016). \hfill \blacksquare

7.2 Proof of Lemma 1

Lemma 1. If $\mu_0$ is $\phi$ positive on $\Theta$ then for every $\varepsilon, \varepsilon', \kappa \in \mathbb{R}_+$, $t \in \mathbb{N}$, $f_t \in \Delta^\Theta(Y)$, $\bar{q} \in M_\varepsilon(f_t)$, with $\varepsilon' + \kappa \leq \varepsilon$,

$$\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} \geq \phi(\kappa/2R(f_t, \kappa, \bar{q})) \exp((\varepsilon - \kappa - \varepsilon')t)$$

where

$$R(f_t, \kappa, \bar{q}) = \max_{z \in \Theta \cap B_\kappa(q)} \frac{f_t(z)}{q(z)}.$$
The lemma is a consequence of two facts. First, the continuity of $H$ established in Lemma 2 implies that distributions sufficiently close (relative to the inverse of the Lipschitz constant) to an $\varepsilon'$-minimizer have a KL divergence that is at most $\kappa + \varepsilon'$ larger than the one of the exact minimizers. Since by definition the distributions outside the set $M_\varepsilon$ have a KL divergence that is at least $\varepsilon$ larger than that of the exact minimizers, the difference between the value of the KL divergence in these two sets is at least $\varepsilon - \kappa - \varepsilon'$. Second, the posterior probability ratio between the two sets grows exponentially in the difference between the KL-divergences from the empirical distributions of the elements of the sets.

Claim 1. For every $p' \in \Theta$, $f \in \Delta^\Theta(Y)$, $\varepsilon', \kappa \in \mathbb{R}_+$, and $\bar{q} \in M_{\varepsilon'}(f)$,

$$p' \in B_{\kappa/2R(f,\kappa,\bar{q})}(\bar{q}) \implies H(f, p') \leq \min_{p \in \Theta} H(f, p) + \varepsilon' + \kappa.$$

Proof. For every two distributions $f, q \in \Delta(Y)$, there is at least one outcome that is weakly more likely under $f$ than under $q$ so $R$ is bounded below by 1. Thus $p' \in B_{\kappa/2R(f,\kappa,\bar{q})}(\bar{q})$ implies $p' \in B_{\kappa}(\bar{q})$. Therefore, both $p', \bar{q}$ are in $\Theta \cap B_\kappa(\bar{q})$, so from the definition of $R$, $\max_{z \in Y} \max \left\{ \frac{f(z)}{p'(z)}, \frac{f(z)}{\bar{q}(z)} \right\} \leq R(f, \kappa, \bar{q})$. Moreover, $p' \in B_{\kappa/2R(f,\kappa,\bar{q})}(\bar{q})$ implies that $\bar{q} \in B_{\kappa/2R(f,\kappa,\bar{q})}(p') \cap M_{\varepsilon'}(f)$, so by Lemma 2

$$H(f, p') - H(f, \bar{q}) \leq \frac{\kappa}{R(f, \kappa, \bar{q})} \max_{z \in Y} \max \left\{ \frac{f(z)}{p'(z)}, \frac{f(z)}{\bar{q}(z)} \right\} \leq \frac{\kappa}{R(f, \kappa, \bar{q})} R(f, \kappa, \bar{q}) = \kappa,$$

and hence

$$H(f, p') \leq H(f, \bar{q}) + \kappa \leq \min_{p \in \Theta} H(f, p) + \varepsilon' + \kappa.$$  

23
Observe that
\[
\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} = \frac{\int_{M_\varepsilon(f_t)} \exp(-H(p, f_t)t) d\mu_0(dp)}{\int_{\Theta \setminus M_\varepsilon(f_t)} \exp(-H(p, f_t)t) d\mu_0(dp)} \\
\geq \frac{\int_{M_\varepsilon+\varepsilon'(f_t)} \exp(-H(p, f_t)t) d\mu_0(dp)}{\int_{\Theta \setminus M_\varepsilon(f_t)} \exp(-H(p, f_t)t) d\mu_0(dp)} \\
\geq \frac{\exp(- (\min_{p \in \Theta} H(f_t, p) + \kappa + \varepsilon') t) \mu_0(M_{\kappa+\varepsilon'}(f_t))}{\exp(- (\min_{p \in \Theta} H(f_t, p) + \varepsilon) t) 1 - \mu_0(\Theta \setminus M_\varepsilon(f_t))} \\
= \exp((\varepsilon - \kappa - \varepsilon') t) \frac{\mu_0(M_{\kappa+\varepsilon'}(f_t))}{\mu_0(\Theta \setminus M_\varepsilon(f_t))} \\
\geq \exp((\varepsilon - \kappa - \varepsilon') t) \mu_0(B_{\kappa/2R(f_t, \kappa, \bar{q})}(\bar{q})) \\
\geq \exp((\varepsilon - \kappa - \varepsilon') t) \phi(\kappa/2R(f_t, \kappa, \bar{q})).
\]

The first equality follows from (1), which shows how to rewrite the likelihood in terms of the Kullback-Leibler divergence. The first inequality follows from \(\varepsilon' + \kappa \leq \varepsilon\), the second inequality from pointwise bounding the integrands and the definition of \(M_\varepsilon\), the third inequality from Claim 1, and the fourth from the \(\phi\) positivity on \(\Theta\) of \(\mu_0\).

\[\Box\]

### 7.3 Proof of Theorem 1

**Theorem 1.** For every \(\phi : \mathbb{R}_{++} \to \mathbb{R}_{++}\) and every \(\alpha \in (0, 1)\) there is a function \(A : \mathbb{R}_{++} \to \mathbb{R}_{++}\) such that

\[
\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} \geq A(\varepsilon) \exp(\alpha \varepsilon t)
\]

for all \(\mu_0\) that are \(\phi\) positive on \(\Theta\), \(\varepsilon \in (0, 1)\), \(t \in \mathbb{N}\), and \(f_t \in \Delta^\Theta(Y)\).

Moreover, if \(q := \inf_{\bar{q} \in \Theta} \min_{z \in \text{supp } q} q(z) > 0\), then we can set

\[
A(\varepsilon) = \phi(\min\{q/2, (1 - \alpha)\varepsilon\}|q/2) .
\]

To see that if \(q := \inf_{\bar{q} \in \Theta} \min_{z \in \text{supp } q} q(z) > 0\), Lemma 1 yields the desired uniform rate of convergence, suppose first that \((1 - \alpha)\varepsilon < q\). Then for all \(f_t \in \Delta^\Theta(Y)\) and \(\bar{q} \in M(f_t)\) if \(p \in \Theta \cap B_{(1-\alpha)\varepsilon}(\bar{q})\), \(\text{supp } p = \text{supp } \bar{q} \subseteq \text{supp } f_t\) and so \(1/R(f_t, (1-\alpha)\varepsilon, \bar{q}) > q\). If instead \((1 - \alpha)\varepsilon \geq q\) it is enough to observe that \(1/R(f_t, q/2, \bar{q}) > q\) for all
$f_t \in \Delta^\Theta(Y), \bar{q} \in M(f_t)$.

Now we move to the proof of the general case.

**Claim 2.** For every $\varepsilon \in (0, 1)$, there exist $\bar{\kappa}(\varepsilon) \in (0, (1 - \alpha)\varepsilon/2]$ and $c \geq 1$ such that for all $\kappa \leq \bar{\kappa}$, and $f \in \Delta^\Theta(Y)$, there is $\bar{q} \in M_{(1-\alpha)c}^\Theta(f)$ such that $\max_{y \in Y, q \in B_{\kappa}(\bar{q})} \frac{f(y)}{q(y)} \leq c$.

**Proof.** If not, then since $\Delta^\Theta(Y)$ and $\Theta$ are compact, there is a sequence $(f_n, q_n) \in \Delta^\Theta(Y) \times \Theta$ with $q_n \in M(f_n)$ that converges to $(\hat{f}, \hat{q})$, and such that

$$\max_{y \in Y, q \in B_{\kappa}(\bar{q})} \frac{f_n(y)}{q(y)} \geq n \quad \forall \bar{q} \in M_{(1-\alpha)c}^\Theta(f_n). \tag{4}$$

Since $Y$ is finite, so is the set of possible supports, and thus there is a subsequence $f_n$ that has common support, with $f_n(y)$ weakly decreasing for all $y \in Y \setminus \text{supp } \hat{f}$.

Let $G = \inf_{n \in \mathbb{N}, y \in \text{supp } f} \log q_n(y)$. Since all the $f_n$ are in $\Delta^\Theta(Y)$ and have common support, and $q_n \in M(f_n)$, we have

$$H(f_n, q_n) \leq H(f_n, q_1) = \sum_{z \in \text{supp } f_1} f_n(z) \log q_1(z) \leq \min_{z \in \text{supp } f_1} \log q_1(z) < \infty,$$

Since $H(f_n, q_n)$ is bounded, $\lim_{n \to \infty} q_n(y) = 0$ implies $\lim_{n \to \infty} f_n(y) = 0$ and therefore $G > -\infty$.

Since $f_n$ is converging, there exists $N \in \mathbb{N}$ such that for all $n$ and $m$ larger than $N$

$$||f_n - f_m|| \leq \frac{1}{|G|} \frac{(1-\alpha)\varepsilon}{8}. \tag{5}$$

Because $\hat{H} = \lim \inf \ H(f_n, q_n) < \infty$, there exists $N' \geq N$ such that

$$H(f_{N'}, q_{N'}) \leq \hat{H} + \frac{(1-\alpha)\varepsilon}{8}. \tag{6}$$

Moreover, there exists $N'' > N'$ such that for all $n \geq N''$

$$H(f_n, q_n) \geq \hat{H} - \frac{(1-\alpha)\varepsilon}{8}. \tag{7}$$
Thus for every \( n \geq N'' \), we have

\[
H(f_n, q_{N'}) - H(f_n, q_n) \leq H(f_n, q_{N'}) - \hat{H} + \frac{(1 - \alpha)\varepsilon}{8}
\leq H(f_{N'}, q_{N'}) + \frac{1}{|G|} \frac{(1 - \alpha)\varepsilon}{8} \cdot 2|G| - \hat{H} + \frac{(1 - \alpha)\varepsilon}{8}
\leq \hat{H} + \frac{(1 - \alpha)\varepsilon}{8} + \frac{(1 - \alpha)\varepsilon}{4} - \hat{H} + \frac{(1 - \alpha)\varepsilon}{8} = \frac{(1 - \alpha)\varepsilon}{2}
\]

where the first inequality follows by equation (7), the second by equation (5) and the fact that \( f_m(z) \) is decreasing for the outcomes outside \( \text{supp} \hat{f} \) and the third by (6).

But then \( q_{N'} \in M_{\left(1 - \alpha\right)\varepsilon} (f_n) \) for all \( n \geq N'' \), a contradiction with equation (4).

Now for every \( f_t \in \Delta^\Theta(Y) \) let \( c, \bar{\kappa}_\alpha(\varepsilon) \), and \( \bar{q} \in M_{\left(1 - \alpha\right)\varepsilon} (f_t) \) be the values whose existence is established by Claim 2. Then

\[
\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} \geq \phi(\bar{\kappa}_\alpha(\varepsilon)/2R(f_t, \bar{\kappa}_\alpha(\varepsilon), \bar{q})) \exp \left( \left( \varepsilon - \bar{\kappa}_\alpha(\varepsilon) - \frac{(1 - \alpha)\varepsilon}{2} \right) t \right)
\geq \phi(\bar{\kappa}_\alpha(\varepsilon)/2c) \exp \left( \left( \varepsilon - \frac{(1 - \alpha)\varepsilon}{2} - \frac{(1 - \alpha)\varepsilon}{2} \right) t \right)
\geq \phi(\bar{\kappa}_\alpha(\varepsilon)/2c) \exp(\alpha \varepsilon t)
\]

where the first inequality follows from applying Lemma 1 with \( \varepsilon' = \frac{(1 - \alpha)\varepsilon}{2} \) and \( \kappa = \bar{\kappa}_\alpha(\varepsilon) \), the second inequality follows from the fact that by Claim 2 \( c \geq \max_{y \in Y, q \in B_{\bar{\kappa}_\alpha(\varepsilon)}(\bar{q})} R(f_t, \bar{\kappa}_\alpha(\varepsilon), \bar{q}) \) and \( \bar{\kappa}_\alpha(\varepsilon) \leq (1 - \alpha)\varepsilon/2 \), and the third inequality is algebra. Theorem 1 then follows by letting

\[
A(\varepsilon) = \phi \left( \frac{\bar{\kappa}_\alpha(\varepsilon)}{2c} \right).
\]

### 7.4 Proof of Theorem 2

**Theorem 2.** Let \( \mathbb{P} \) be the probability measure induced if the outcomes are i.i.d. draws from \( p^* \). If \( \Theta \) is regular and \( \mu_0 \) is \( \phi \) positive on \( \Theta \), then for every \( \varepsilon \in (0, 1) \) there is a \( K \in \mathbb{R}_{++} \) such that

\[
\mathbb{P} \left[ \mu_t \left( B_{\varepsilon}(M(p^*)) \right) < 1 - K \exp(-Kt) \right] = O(\exp(-Kt)).
\]
Claim 3. For every $\varepsilon > 0$ and $p^* \in P$, there exists $\varepsilon' > 0$ such that

$$M_{\varepsilon'}(p^*) \subseteq B_\varepsilon(M(p^*)) .$$

Proof. Assume the claim is false, so for every $n \in \mathbb{N}$, there exists $q_n \in \Theta \setminus B_\varepsilon(M(p^*))$ such that $H(p^*, q_n) - \min_{p \in \Theta} H(p^*, p) \leq \frac{1}{n}$. Since $\Theta$ is compact, $(q_n)_{n \in \mathbb{N}}$ admits a convergent subsequence with limit $q^* \in \Theta$. Since $H(p^*, \cdot)$ is continuous in its second argument, $q^* \in M(p^*)$. But this would imply that the subsequence is eventually in $B_\varepsilon(M(p^*))$, a contradiction. \hfill \blacksquare

By Claim 3, $M_{\varepsilon'}(p^*) \subseteq B_\varepsilon(M(p^*))$ for some $\varepsilon' > 0$. Since $M_{\varepsilon'}(\cdot)$ is upper semicontinuous by Lemma 4, there exists $\hat{\varepsilon}$ such that if $q \in B_{\hat{\varepsilon}}(p^*)$, then $M_{\varepsilon' / 2}(q) \subseteq M_{\varepsilon'}(p^*)$. By Sanov’s theorem (Dupuis and Ellis, 2011) and the Pinsker inequality, $\mathbb{P} [f_t \notin B_{\varepsilon}(p^*)] \leq \mathbb{P} [H(p^*, f_t) \geq 2\varepsilon^2] \leq 2^{-2\varepsilon^2 t}$, and so

$$\mathbb{P} \left[ \mu_t \left( B_\varepsilon(M(p^*)) \right) < 1 - K \exp(-\hat{K}t) \right] = O(\exp(-K't))$$

follows from Theorem 1 by letting $K = \alpha\varepsilon'/2$ for $\alpha \in (0, 1)$ and $\hat{K} = \varepsilon'/2$. Since $\lim_{t \to \infty} \frac{K \exp(-\hat{K}t)}{\exp(-Ct)} = 0$ for all $C < \hat{K}$, the result follows by letting $K = \min\{\hat{K}/2, K'\}$. \hfill \blacksquare

7.5 Proof of Proposition 1

Proposition 1. If $\tilde{H}$ is uniformly strongly $m$-convex on $\mathcal{F}$ then for every $\phi : \mathbb{R}_{++} \to \mathbb{R}_{++}$ and every $\alpha \in (0, 1)$

$$\frac{\mu_t(B_{\sqrt{2\varepsilon/m} (\theta^*(f_i))})}{1 - \mu_t(B_{\sqrt{2\varepsilon/m} (\theta^*(f_i))})} \geq A(\varepsilon) \exp(\alpha\varepsilon t)$$

for all $\mu_0$ that are $\phi$ positive on $\Theta$, $\varepsilon \in (0, 1)$, $t \in \mathbb{N}$, and $f_i \in \Delta^\Theta(Y) \cap \mathcal{F}$, where $A$ is the function whose existence is guaranteed by Theorem 1.

Proof. We claim first that for every $\theta \in \Theta$ and $f \in \Delta^\Theta(Y) \cap \mathcal{F}$, $\nabla_\theta \tilde{H}(f, \theta^*(f))T(\theta -$
\( \theta^*(f) \geq 0 \). If not,

\[
0 > \nabla_{\theta} \tilde{H}(f, \theta^*(f))^T (\theta - \theta^*(f)) = \lim_{k \to 0} \frac{\tilde{H}(f, \theta^*(f)) - \tilde{H}(f, \theta^*(f) + k(\theta - \theta^*(f)))}{k}.
\]

But this means that there is \( \hat{k} \in (0, 1) \) such that \( \tilde{H}(f, \theta^*(f)) - \tilde{H}(f, \theta^*(f) + \hat{k}(\theta - \theta^*(f))) > 0 \) or \( \tilde{H}(f, (1 - \hat{k})\theta^*(f) + \hat{k}\theta) < \tilde{H}(f, \theta^*(f)) \). As \( \Theta \) is convex, \((1 - \hat{k})\theta^*(f) + \hat{k}\theta \) belongs to \( \Theta \), but this contradicts the fact that \( \theta^*(f) \) is a KL-minimizer, which proves the claim.

Next, as \( \tilde{H} \) is uniformly strongly \( m \)-convex,

\[
\tilde{H}(f, \theta) - \tilde{H}(f, \theta^*(f)) \geq \nabla_{\theta} \tilde{H}(f, \theta^*(f))^T (\theta - \theta^*(f)) + \frac{m}{2} ||\theta - \theta^*(f)||_2^2 \geq \frac{m}{2} ||\theta - \theta^*(f)||_2^2.
\]

As a consequence \( M_{\varepsilon}(f) \subseteq B_{\sqrt{2\varepsilon/m}(\theta^*(f))} \) and the result follows from Theorem 1.

### 7.6 Proof of Corollary 1

**Corollary 1.** Suppose that \( \Theta \subseteq \Delta(Y) \) is convex and that \( \mu_0 = \phi \) positive on \( \Theta \). Then for every \( \alpha, \varepsilon \in (0, 1) \)

\[
\frac{\mu_t(B_{\sqrt{\varepsilon/\min_{z\in Y} f_t(z)}(M(f_t))})}{1 - \mu_t(B_{\sqrt{\varepsilon/\min_{z\in Y} f_t(z)}(M(f_t))}) \geq A(\varepsilon) \exp(\alpha \varepsilon t)
\]

for all \( t \in \mathbb{N} \) and \( f_t \in \Delta^\Theta(Y) \) such that \( \min_{z \in Y} f_t(z) > 0 \), where \( A \) is the function whose existence is guaranteed by Theorem 1.

**Proof.** The result follows from Proposition 1 and observing that

\[
(\nabla_p H(f, p) - \nabla_p H(f, \tilde{p}))^T (p - \tilde{p}) = \sum_{z \in Y} f(z) \left( \frac{1}{\tilde{p}(z)} - \frac{1}{p(z)} \right) (p(z) - \tilde{p}(z)) = \sum_{z \in Y} \frac{f(z)}{p(z)\tilde{p}(z)} (p(z) - \tilde{p}(z))^2 \geq (\min_{z \in Y} f(z)) ||p - \tilde{p}||_2^2.
\]

\[28\]
7.7 Proof of Proposition 2

Proposition 2. Suppose that $\Theta$ is convex and that $\mu_0$ is $\phi$ positive on $\Theta$. Then for all $\varepsilon, c > 0$ there is a time period $T$ such that $d(A^*(\mu_t), A^*(M(f_t))) \leq \varepsilon$ for every $t > T$ and every $f_t \in \Delta^\Theta(Y)$ with $\min_{z \in Y} f_t(z) > c$.

Proof. Since $A^*$ is Lipschitz continuous, there exists $\varepsilon' \in \mathbb{R}_{++}$ such that $\mu_t \in B_{\varepsilon'}(\nu)$ implies $d(A^*(\mu_t), A^*(\nu)) \leq \varepsilon$. Since $Y$ is finite, $P$ is separable, so the topology of weak convergence on $\Delta(P)$ is metrized by the Lévy–Prokhorov metric. By the definition of this metric, $||\nu - \delta_p||_{LP} \leq \varepsilon'$ whenever

$$\frac{\mu((B_{\varepsilon'/2}(p))}{1 - \mu((B_{\varepsilon'/2}(p))} \geq 1 - \varepsilon'/2.$$ 

The statement follows from applying Corollary 1 with $\alpha = 1/2$ and choosing

$$T \geq \frac{2\log \left( \frac{1 - \varepsilon'/2}{A((\varepsilon'/2)^2)c^\varepsilon'/2} \right)}{A((\varepsilon'/2)^2)c^\varepsilon'/2}.$$ 

7.8 Proof of Proposition 3

Proposition 3. If $\text{supp } p = \text{supp } p^*$ for every $p \in \Theta$ and $\mu_0$ is $\phi$ positive on $\Theta$, for every $q \in M(p^*)$, $\varepsilon \in (0, 1/4)$

$$\frac{\mu_t(B_{\varepsilon}(q))}{1 - \mu_t(B_{\varepsilon}(q))} > \phi \left( \frac{q\varepsilon^2}{2\sqrt{i}} \right) \exp \left( \varepsilon^2 \sqrt{i}/2 \right).$$

for all $f_t \in B_{i^2/(-2\log_2 \varepsilon)} \left( p^*_{q,1/\sqrt{i}} \right) \cap \Delta^\Theta(Y)$.

First, we obtain a generalized version of the Pinsker inequality that bounds the increase in the KL divergence from below as we move away from the unique KL minimizer for $p^*_{q,\alpha}$.

Claim 4. For every $q \in M(p^*)$, $q' \in \Theta$ and $\alpha \in (0, 1)$, $H(p^*_{q,\alpha}, q') - H(p^*_{q,\alpha}, q) \geq 2\alpha ||q - q'||^2$. 

29
Proof. We have

\[
H(p^*_q, q') - H(p^*_q, q) = \sum_{y \in \mathcal{Y}} p^*_q(y) \log \frac{q(y)}{q'(y)}
\]

\[
= \sum_{y \in \mathcal{Y}} \alpha q(y) \log \frac{q(y)}{q'(y)} + \sum_{y \in \mathcal{Y}} (1 - \alpha) p^*(y) \log \frac{q(y)}{q'(y)}
\]

\[
\geq \alpha \sum_{y \in \mathcal{Y}} q(y) \log \frac{q(y)}{q'(y)} \geq 2\alpha \|q - q'\|^2,
\]

where the first inequality follows from the fact that \( q \) is a KL minimizer for \( p^* \) and the second is the Pinsker inequality.

\[\square\]

Claim 5. For every \( q \in M(p^*) \), for every \( \varepsilon \in (0, 1) \) if \( f \in B_{\varepsilon^2/(2\log q(y)\sqrt{t})} \left(p^*_{q,1/\sqrt{t}}\right) \cap \Delta^\Theta(Y) \), \( ||q - q'|| > \varepsilon \), and \( q' \in \Theta \), then \( H(f, q') - H(f, q) \geq \frac{\varepsilon^2}{\sqrt{t}} \).

Proof. Let \( \hat{c} = -\log \frac{q}{q'} \). Next, pick any \( q' \in \Theta \) such that \( ||q - q'|| > \varepsilon \). By Claim 4,

\[
H\left(p_{q,1/\sqrt{t}}, q'\right) - H\left(p^*_{q,1/\sqrt{t}}, q\right) \geq 2|q - q'|^2/\sqrt{t}.
\]

Let \( c = 2\hat{c}/\varepsilon^2 \) and take \( f \in B_{1/c\sqrt{t}} \left(p^*_{q,1/\sqrt{t}}\right) \cap \Delta^\Theta(Y) \). By Lemma 3

\[
\left(H\left(p^*_{q,1/\sqrt{t}}, q'\right) - H\left(p^*_{q,1/\sqrt{t}}, q\right)\right) - (H(f, q') - H(f, q))
\]

\[
\leq 2||f - p^*_{q,1/\sqrt{t}}|| \max_{y \in \text{supp}p^*} |\log \frac{q'(y)}{q(y)}| \leq 2||f - p^*_{q,1/\sqrt{t}}|| \hat{c} \leq \frac{1}{c\sqrt{t}} 2\hat{c} = \frac{\varepsilon^2}{2\sqrt{t}}
\]

so

\[
(H(f, q') - H(f, q)) \geq \left(H\left(p^*_{q,1/\sqrt{t}}, q'\right) - H\left(p^*_{q,1/\sqrt{t}}, q\right)\right) - \frac{\varepsilon^2}{2\sqrt{t}} \geq \frac{2||q - q'||^2}{\sqrt{t}} - \frac{\varepsilon^2}{2\sqrt{t}}
\]

\[
\geq \frac{\varepsilon^2}{\sqrt{t}}
\]

as desired. \[\square\]
Claim 6. For every \( \varepsilon \in (0,1) \), \( t \in \mathbb{N} \) and \( f \in B_{\varepsilon^2/(-2 \log_2 \varepsilon)} \left( p_{q,1/\sqrt{t}}^* \right) \cap \Delta^\Theta(Y) \), we have \( B_{\varepsilon^2/\sqrt{t}} \left( M_{\varepsilon^2} (f) \right) \subseteq B_{2\varepsilon} (q) \).

Proof. By Claim 5 \( f \in B_{\varepsilon^2/(-2 \log_2 \varepsilon)} \left( p_{q,1/\sqrt{t}}^* \right) \cap \Delta^\Theta(Y) \) implies \( M_{\varepsilon^2} (f) \subseteq B_{\varepsilon} (q) \). Then \( B_{\varepsilon^2/\sqrt{t}} \left( M_{\varepsilon^2} (f) \right) \subseteq B_{\varepsilon} \left( B_{\varepsilon} (q) \right) \subseteq B_{2\varepsilon} (q) \), proving the claim.

7.9 Proof of Theorem 3

Theorem 3. Suppose that for all \( \pi, \pi' \in \Theta \), \( z, z' \in Y \), \( \pi(z'|z) > 0 \) if and only if \( \pi'(z'|z) > 0 \) and that \( \mu_0 \) is \( \phi \) positive on \( \Theta \). Then for all \( \alpha \in (0,1) \) there is a function \( A : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++} \) such that

\[
\frac{\mu_t(M_\varepsilon(f))}{1 - \mu_t(M_\varepsilon(f))} \geq A(\varepsilon) \exp(\alpha \varepsilon t),
\]

for all \( \varepsilon > 0, t \in \mathbb{N} \) and \( f_t \in \Delta^\Theta(Y \times Y) \).

We begin with a continuity result that extends Lemma 2 to the Markov setting.

Claim 7. Let \( R := \max_{\pi \in \Theta, z \in Y, z' \in \text{supp } \pi(z'|z)} (1/\pi(z'|z)) \). For all \( \pi, \tilde{\pi} \in \Theta \) and \( f \in \Delta^\Theta(Y \times Y) \)

\[
|\mathcal{H}(f, \pi) - \mathcal{H}(f, \tilde{\pi})| \leq 2R||\pi - \tilde{\pi}||.
\]

Proof. Suppose without loss of generality that \( \sum_{(z,z') \in \text{supp } f} \pi(z'|z) \geq \sum_{(z,z') \in \text{supp } f} \tilde{\pi}(z'|z) \). Then
\begin{align*}
|\mathcal{H}(f, \pi) - \mathcal{H}(f, \tilde{\pi})| &= \left| \sum_{(z, z') \in \text{supp} f} \left( \log \left( \frac{\pi(z'|z)}{f(z, z')} \right) - \log \left( \frac{\tilde{\pi}(z'|z)}{f(z, z')} \right) \right) f(z, z') \right| \\
&= \left| \sum_{(z, z') \in \text{supp} f} \int_{\tilde{\pi}(z'|z)/f(z, z')}^{\pi(z'|z)/f(z, z')} \frac{1}{r} dr f(z, z') \right| \\
&\leq \sum_{(z, z') \in \text{supp} f} \max \left\{ \frac{f(z, z')}{\tilde{\pi}(z'|z)/\pi(z'|z)}, \frac{\pi(z'|z)}{f(z, z')} \right\} \left| \frac{\pi(z'|z)}{f(z, z')} - \frac{\tilde{\pi}(z'|z)}{f(z, z')} \right| f(z, z') \\
&\leq R \sum_{(z, z') \in \text{supp} f} \left( 2 \mathbb{I}_{\frac{\pi(z'|z)}{f(z, z')} > \frac{\pi(z'|z)}{f(z, z')} > \frac{\tilde{\pi}(z'|z)}{f(z, z')} - 1 \right) \left( \frac{\pi(z'|z)}{f(z, z')} - \frac{\tilde{\pi}(z'|z)}{f(z, z')} \right) f(z, z') \\
&= R \sum_{(z, z') \in \text{supp} f} \left( 2 \mathbb{I}_{\frac{\pi(z'|z)}{f(z, z')} > \frac{\pi(z'|z)}{f(z, z')} > \frac{\tilde{\pi}(z'|z)}{f(z, z')} - 1 \right) \pi(z'|z) - R \sum_{(z, z') \in \text{supp} f} \left( 2 \mathbb{I}_{\frac{\pi(z'|z)}{f(z, z')} > \frac{\pi(z'|z)}{f(z, z')} > \frac{\tilde{\pi}(z'|z)}{f(z, z')} - 1 \right) \tilde{\pi}(z'|z) \\
&\leq 2R \left[ \sum_{(z, z') \in \text{supp} f} \mathbb{I}_{\frac{\pi(z'|z)}{f(z, z')} > \frac{\pi(z'|z)}{f(z, z')} > \frac{\tilde{\pi}(z'|z)}{f(z, z')} \pi(z'|z) - \sum_{(z, z') \in \text{supp} f} \mathbb{I}_{\frac{\pi(z'|z)}{f(z, z')} > \frac{\pi(z'|z)}{f(z, z')} \tilde{\pi}(z'|z)} \right] \\
&\leq 2R \|\pi - \tilde{\pi}\|.
\end{align*}

Here the first inequality follows from pointwise bounding the integrand, and the second inequality follows from the fact that for all \( \pi, \pi' \in \Theta, z, z' \in Y, \pi(z'|z) > 0 \) if and only if \( \pi'(z'|z) > 0 \) and \( f \in \Delta^\Theta(Y \times Y) \). The last inequality follows from the definition of the total variation distance.

We now use Claim 7 to establish the theorem. Fix \( \varepsilon \in \mathbb{R}_{++} \). Rewrite the likelihood ratio for distributions inside and outside of \( \mathcal{M}_c(f) \) as follows:
\[
\frac{\mu_t(M_\varepsilon(f_t))}{1 - \mu_t(M_\varepsilon(f_t))} = \frac{\int_{\pi \in M_\varepsilon(f_t)} \prod_{t=1}^t \pi(y_t | y_{t-1}) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \prod_{t=1}^t \pi(y_t | y_{t-1}) d\mu_0(\pi)} \\
= \frac{\int_{\pi \in M_\varepsilon(f_t)} \exp(\sum_{z, z' \in Y} f_t(z, z') \log(\pi(z | z')) t) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \exp(\sum_{z, z' \in Y} f_t(z, z') \log(\pi(z | z')) t) d\mu_0(\pi)} \\
= \frac{\int_{\pi \in M_\varepsilon(f_t)} \exp(-H(f_t, \pi) t) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \exp(-H(f_t, \pi) t) d\mu_0(\pi)}.
\]

Next we provide a lower bound on this likelihood ratio:

\[
\frac{\int_{\pi \in M_\varepsilon(f_t)} \exp(-H(f_t, \pi) t) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \exp(-H(f_t, \pi) t) d\mu_0(\pi)} \\
\geq \frac{\int_{\pi \in M_\varepsilon(f_t)} \exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + \varepsilon\right] t) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + \varepsilon\right] t) d\mu_0(\pi)} \\
\geq \frac{\int_{\pi \in M_\varepsilon(f_t)} \exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + (1 - \alpha)\varepsilon\right] t) d\mu_0(\pi)}{\int_{\pi \notin M_\varepsilon(f_t)} \exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + (1 - \alpha)\varepsilon\right] t) d\mu_0(\pi)} \\
\geq \phi((1 - \alpha)\varepsilon/2R) \frac{\exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + (1 - \alpha)\varepsilon\right] t)}{\exp(-\left[\min_{\pi \in \Theta} H(f_t, \pi) + \varepsilon\right] t)} \\
= \phi((1 - \alpha)\varepsilon/2R) \exp(\alpha \varepsilon t).
\]

Here the first and second inequalities follows from the definitions of $M_\varepsilon$, the third inequality from Claim 7, and the fourth inequality from $\phi$ positivity on $\Theta$. The result follows by setting $A(\varepsilon) = \phi((1 - \alpha)\varepsilon/2R)$. \]

### 7.10 Counterexamples to Corollary 1

**Example 4.** (Connected $\Theta$ is not sufficient)

Here we show that the conclusion of Corollary 1 can fail when $\Theta$ is connected but not convex.

Let $Y = \{A, B, C\}$, $\Theta = \{p : p(A)p(B)p(C) = 0\} \setminus \{p : p(A), p(B) \subseteq (1/4, 3/4), p(A) + p(B) = 1\}$, and $\hat{\mu}_0$ be the uniform measure on $\Theta$. Let $\mu'_0 = \delta_{(1/4,3/4,0)/2} + \delta_{(3/4,1/4,0)/2}$
and \( \mu_0 = \hat{\mu}_0/2 + \hat{\mu}_0'/2 \). Suppose \( y_t = B \) if \( t \) is odd and \( y_t = A \) if \( t \) is even. At every odd period \( t = 2n + 1 \), \( M(f_{2t+1}) = \{(1/4,3/4,0)\} \), and for \( \varepsilon < 1/12 \),

\[
\lim_{t \to \infty} \frac{\mu_{2t+1}(B_\varepsilon(M(f_{2t+1})))}{1 - \mu_{2t+1}(B_\varepsilon(M(f_{2t+1})))} = \frac{\mu_{2t+1}(B_\varepsilon(1/4,3/4,0))}{\mu_0(B_\varepsilon(1/4,3/4,0))} \leq \frac{\mu_0(B_\varepsilon(1/4,3/4,0))}{\mu_0(\{(1/4,3/4,0)\})} (1/4)^{t+1} (3/4)^t = \frac{3\mu_0(B_\varepsilon(1/4,3/4,0))}{\mu_0(\{(3/4,1/4,0)\})}
\]

so beliefs do not concentrate on the KL minimizer.

**Example 5.** (Convergence is not uniform over all paths)

Here we show that even if \( \Theta \) is convex, beliefs need not to converge to the KL minimizers along paths where the empirical distribution converges to the boundary.

Let \( Y = \{A,B,C\}, \Theta = \{p : p(A) = 1/3\} \), and \( \mu_0 \) be the uniform measure on \( \Theta \). Suppose \( f_{2n} = (1 - 1/n, 1/2n, 1/2n) \). We have \( M(f_{2n}) = \{(1/3,1/3,1/3)\} \) for all \( n \in \mathbb{N} \). However, fix an \( \varepsilon \in (0,1/12) \). Then we have

\[
\frac{\mu_{2n}(B_\varepsilon(M(f_{2n})))}{1 - \mu_{2n}(B_\varepsilon(M(f_{2n})))} \leq \frac{\int_{p_B \in B_\varepsilon(M(f_{2n}))} \exp(-H(f_{2n},p)2n) d\mu_0(p)}{\int_{p \notin B_\varepsilon(M(f_{2n}))} \exp(-H(f_{2n},p)2n) d\mu_0(p)}
\]

\[
\leq \frac{\mu_0(B_\varepsilon(M(f_{2n}))) \exp(-H(f_{2n},(1/3,1/3,1/3)2n))}{\mu_0(B_\varepsilon(M(f_{2n}))) \exp(-H(f_{2n},(1/3,1/3,1/3+2\varepsilon,1/3-2\varepsilon)2n))}
\]

\[
= \frac{\mu_0(B_\varepsilon(M(f_{2n})))}{\mu_0(B_\varepsilon(M(f_{2n})))} \frac{(1/3 + 2\varepsilon)(1/3 - 2\varepsilon)}{(1/3)(1/3)}
\]

\[
\to n \frac{\mu_0(B_\varepsilon(\{(1/4,3/4,0)\}))}{\mu_0(B_\varepsilon(\{(1/4,3/4,0)\}))} \frac{(1/3 + 2\varepsilon)(1/3 - 2\varepsilon)}{(1/3)(1/3)}
\]

so beliefs do not concentrate.

**References**


— (2021b). “Selective Memory Equilibrium”.


