Bernoulli **0**(00), 0, 1–30
https://doi.org/10.3150/21-BEJ1398

Posterior probabilities: Nonmonotonicity, asymptotic rates, log-concavity, and Turán's inequality

SERGIU HART[1] and YOSEF RINOTT[2]

[1]*The Hebrew University of Jerusalem, Federmann Center for the Study of Rationality, Einstein Institute of Mathematics, and Department of Economics. E-mail: hart@huji.ac.il*
[2]*The Hebrew University of Jerusalem, Federmann Center for the Study of Rationality, and Department of Statistics. E-mail: yosef.rinott@mail.huji.ac.il*

In the standard Bayesian framework data are assumed to be generated by a distribution parametrized by θ in a parameter space Θ, over which a prior distribution π is given. A Bayesian statistician quantifies the belief that the true parameter is θ_0 in Θ by its posterior probability given the observed data. We investigate the behavior of the posterior belief in θ_0 when the data are generated under some parameter θ_1, which may or may not be the same as θ_0. Starting from stochastic orders, specifically, likelihood ratio dominance, that obtain for resulting distributions of posteriors, we consider monotonicity properties of the posterior probabilities as a function of the sample size when data arrive sequentially. While the θ_0-posterior is monotonically increasing (i.e., it is a submartingale) when the data are generated under that same θ_0, it need not be monotonically decreasing in general, not even in terms of its overall expectation, when the data are generated under a different θ_1. In fact, it may keep going up and down many times, even in simple cases such as iid coin tosses. We obtain precise asymptotic rates when the data come from the wide class of exponential families of distributions; these rates imply in particular that the expectation of the θ_0-posterior under $\theta_1 \neq \theta_0$ is eventually strictly decreasing. Finally, we show that in a number of interesting cases this expectation is a log-concave function of the sample size, and thus unimodal. In the Bernoulli case we obtain this result by developing an inequality that is related to Turán's inequality for Legendre polynomials.

Keywords: Bayesian analysis; stochastic and likelihood ratio orders; sequential observations; expected posteriors; unimodality; Legendre polynomials; exponential families

1. Introduction

Consider a sequence of observations $x_1, x_2, \ldots$ whose distribution is governed by a parameter θ, in a standard Bayesian setup with a prior distribution π on the space of parameters Θ. For simplicity assume for now that Θ and the space of observations are finite (the results extend readily to general observations and parameter spaces, as we will see later). Let $\mathbb{P}_\theta$ denote the probability distribution under the parameter θ, and let $\mathbb{P} = \sum_{\theta\in\Theta} \pi(\theta)\mathbb{P}_\theta$ denote the marginal probability; the corresponding expectations are denoted by $\mathbb{E}_\theta$ and $\mathbb{E}$, respectively (thus $\mathbb{P}_\theta(\cdot) \equiv \mathbb{P}(\,\cdot\,|\theta)$ and $\mathbb{E}_\theta[\cdot] \equiv \mathbb{E}[\cdot|\theta]$).

Let θ_0 in Θ be a fixed value of the parameter. We are interested in the way the belief in θ_0 varies as one gets more and more observations, i.e., as n increases. We denote by $q_n^{\theta_0}$ the *posterior* probability of θ_0 at time n; i.e., for every sequence $s_n = (x_1, \ldots, x_n)$ of observations up to time n,

$$q_n^{\theta_0} \equiv q_n^{\theta_0}(s_n) := \mathbb{P}(\theta_0|s_n) = \frac{\mathbb{P}_{\theta_0}(s_n)\pi(\theta)}{\mathbb{P}(s_n)}.$$

1350-7265 © 0 ISI/BS

As is well known, the sequence $q_n^{\theta_0}$ of posteriors is a martingale with respect to the marginal probability $\mathbb{P}$, i.e.,

$$\mathbb{E}[q_{n+1}^{\theta_0}|s_n] = q_n^{\theta_0}(s_n) \tag{1}$$

for every n and s_n. Thus, given s_n, a new observation x_{n+1} distributed according to $\mathbb{P}$ may increase or decrease the posterior of θ_0, but on average this posterior does not change.

Posterior probabilities are used in Bayesian hypothesis testing, where the decision between hypotheses depends on their posterior probabilities (see [4], Section 4.3.3). In sequential Bayes decision rules, monotonicity properties of posteriors determine situations where more data lead to better decisions vs. ones where more data could at times be misleading. This question was raised and discussed in [11]. For a discussion of sequential Bayesian inference and references see, e.g., [9] and [4]. Setups where more data may be harmful according to certain criteria of the statistician are given, for example, in [3] and [12]. In addition to the Bayesian setup, the questions addressed here are pertinent to setups with agents that have different prior beliefs on the parameters. See [13] for details, including applications to models with informed agents vs. an uninformed market, and reputation-building models.

Suppose that the true parameter is θ_0 (but this is, of course, unknown to the observer). What can one say about the sequence of θ_0-posteriors under $\mathbb{P}_{\theta_0}$? When the observations x_n are iid and the distributions $\mathbb{P}_\theta$ are distinct (i.e., $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ for $\theta \neq \theta'$, referred to as "identifiable parameters"), the Doob consistency theorem (see, e.g., [25] and [18]) says that under $\mathbb{P}_{\theta_0}$ the sequence of θ_0-posteriors $q_n^{\theta_0}$ converges to 1 almost surely (i.e., except in a $\mathbb{P}_{\theta_0}$-null set). In fact, this happens monotonically (see [11] and [13], and the references therein), in the sense that under $\mathbb{P}_{\theta_0}$ the posterior of θ_0 always increases on average; that is,

$$\mathbb{E}_{\theta_0}[q_{n+1}^{\theta_0}|s_n] \geq q_n^{\theta_0}(s_n) \tag{2}$$

for every n and s_n (recall that $\mathbb{E}_{\theta_0}[\cdot] \equiv \mathbb{E}[\cdot|\theta_0]$ stands for the expectation with respect to $\mathbb{P}_{\theta_0}$). This submartingale inequality means that *each additional observation increases on average the posterior of the true parameter, under the probability law of the true parameter.*

Taking the posterior probability of θ_0 as one's belief in the model determined by θ_0, the expected belief in θ_0 increases with more data generated under θ_0. Thus, as stated in [11], Bayesian inference does not lead one astray on average. In [13] we show that this is in fact a consequence of an even stronger result, which holds for any observation (and thus, in particular, for x_{n+1} after s_n): the distribution of the θ_0-posterior under $\mathbb{P}_{\theta_0}$ dominates the distribution of that same θ_0-posterior under the marginal probability $\mathbb{P}$, where the domination is in the likelihood ratio order, which is a strengthening of the usual stochastic order; see Sections 2 and 3. We thus get the submartingale inequality (2), from which it follows that the overall expectation $\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \equiv \mathbb{E}_{\theta_0}[\mathbb{P}(\theta_0|s_n)]$ of the θ_0-posterior $q_n^{\theta_0}$ under the probability $\mathbb{P}_{\theta_0}$ is an increasing function of the number of observations n.

Now suppose that the true parameter, θ_1, is different from θ_0. In the above case of iid observations and identifiable parameters, the Doob consistency theorem now says that under $\mathbb{P}_{\theta_1}$ the sequence of θ_0-posteriors $q_n^{\theta_0}$ converges to 0 almost surely. When θ_0 and θ_1 are the only possible parameter values (i.e., $\Theta = \{\theta_0, \theta_1\}$), by (1), (2), and $\mathbb{P}$ being the average of $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_1}$, it immediately follows that $\mathbb{E}_{\theta_1}[q_{n+1}^{\theta_0}|s_n] \leq q_n^{\theta_0}(s_n)$; thus, each additional observation decreases on average the posterior of a "false" parameter, under the probability law of the true parameter. However, this seemingly natural property need *not* hold when there are more than two possible values of θ (see [13] for a simple example, and Section 4 below).

We next turn to consider monotonicity as a function of n of $\psi(n) := \mathbb{E}_{\theta_1}[q_n^{\theta_0}] \equiv \mathbb{E}_{\theta_1}[\mathbb{P}(\theta_0|s_n)]$, the expectation over s_n of the θ_0-posterior $q_n^{\theta_0}$ under the probability $\mathbb{P}_{\theta_1}$, where $\theta_1 \neq \theta_0$. Consider for concreteness the simple setup of iid coin tosses. By Doob's consistency result, $\psi(n) \to 0$ as $n \to \infty$.

This convergence to zero need not however be monotonic. Indeed, when the true parameter θ_1 is "close" to θ_0 in a suitable sense (which we will quantify in Proposition 6), it is natural for data under θ_1 to strengthen the belief in θ_0 at first (i.e., for small n), and so for $\psi(n)$ to increase for small n (as is indeed the case when $\theta_1 = \theta_0$; see (2)). Eventually, however, $\psi(n)$ must approach zero. Once $\psi(n)$ starts decreasing in n, suggesting that evidence against θ_0 is mounting, can $\psi(n)$ increase again? While there may well be particular realizations after which the θ_0-posterior goes up (i.e., $q_{n+1}^{\theta_0} > q_n^{\theta_0}$), and perhaps even particular data s_n after which the θ_0-posterior is expected to go up (i.e., $\mathbb{E}_{\theta_1}[q_{n+1}^{\theta_0}|s_n] > q_n^{\theta_0}(s_n)$), the *overall expectation* is expected to be well behaved and continue to go down (i.e., $\mathbb{E}_{\theta_1}[q_{n+1}^{\theta_0}] \leq \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$). Perhaps surprisingly, that is *not* the case: we provide simple examples where $\psi(n)$ is *not* unimodal in n and may have multiple local maxima; that is, it can go down and then up many times. Thus, after $\psi(n)$ starts decreasing and the statistician may begin to doubt that θ_0 is the true parameter, the increase in $\psi(n)$ with more observations strengthens the statistician's wrong belief that θ_0 is the true parameter; the new observations *do* "lead one astray on average." While this is not a knife-edge phenomenon and may indeed happen, we show that certain natural assumptions rule it out: the expected posterior is eventually decreasing, and even log-concave and thus unimodal.

We now summarize the results on the behavior of $\psi(n) \equiv \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$:

1. When $\theta_1 = \theta_0$, the sequence $\psi(n)$ is increasing in n (as stated above, this is a consequence of the likelihood ratio dominance, which in turn implies the inequality (2); see Sections 2 and 3).
2. When $\theta_1 \neq \theta_0$, the sequence $\psi(n)$ may increase in some range of values of n and decrease in others, and may, for example, decrease first, then increase, and then decrease again (which, as we will show in Section 6.3, can happen already in the case of iid normal observations with a normal prior). Moreover, a large number of modes (i.e., local maxima) can occur in the simple case of iid Bernoulli coin tosses (see Section 4).
3. The sequence $\psi(n)$ is asymptotically equivalent to $C\sqrt{n}\,w^n$ as $n \to \infty$ (for constants $C > 0$ and $0 < w \leq 1$) in the case of exponential families of distributions (discrete and continuous) with a continuous prior, where $w = 1$ when $\theta_1 = \theta_0$, and $w < 1$ when $\theta_1 \neq \theta_0$; in the latter case $\psi(n)$ is therefore strictly decreasing from some n on (see Section 5).
4. The sequence $\psi(n)$ is log-concave in n, and thus unimodal, in a number of scenarios: iid coin tosses with a uniform prior, iid normal observations with a normal prior (in a wide region of parameters), and iid exponential observations with an exponential prior (see Section 6).

The paper is organized as follows. In Section 2 we discuss various order relations between posteriors under different distributions. In Section 3 we extend these results to sequences of observations and see that $\psi(n)$ is increasing when $\theta_1 = \theta_0$. In Section 4 we exhibit situations in which $\psi(n)$ for $\theta_1 \neq \theta_0$ is not unimodal in n, and quantify, for Bernoulli observations, the notion of θ_0 and θ_1 being close together such that an observation under θ_1 increases the belief in θ_0 on average. Section 5 provides the asymptotic analysis of $\psi(n)$, and Section 6 deals with cases where the sequence $\psi(n)$ is unimodal and even log-concave. For Bernoulli observations this is obtained by proving in Section 6.2 a reversal of the reverse Turán inequality for orthogonal polynomials, which is of independent interest. The Appendix contains discussions on possible extensions of the results as well as some technical details.

2. Preliminaries: A single signal

In this section we summarize results on various order relations for posterior distributions that appeared with a somewhat different emphasis in [13]. We consider a signal s, which can be a single observation as well as multiple ones. For simplicity we now consider discrete random variables s and finite parameter spaces Θ. The extension to continuous random variables and prior distributions is straightforward; see Appendix A.1 and Section 5.

We use the following standard notions and notation. The random variables x and y satisfy $y \geq_{\mathrm{st}} x$ (*stochastic order*; also known as *first-order stochastic dominance*) or $y \geq_{\mathrm{icx}} x$ (*increasing convex order*) if $\mathbb{E}[f(y)] \geq \mathbb{E}[f(x)]$ for every increasing function or increasing convex function f, respectively. By "increasing" or "convex" we do not mean "strictly" unless otherwise stated. The probability law (distribution) of a random variable x is denoted by $\mathcal{L}(x)$ (formally, $\mathcal{L}(x) = \mathbb{P} \circ x^{-1}$). Instead of $y \geq_{\mathrm{st}} x$ we may write $\mathcal{L}(y) \geq_{\mathrm{st}} \mathcal{L}(x)$.

The *likelihood ratio order*, denoted by $y \geq_{\mathrm{lr}} x$ or $\mathcal{L}(y) \geq_{\mathrm{lr}} \mathcal{L}(x)$, is said to hold if $\mathbb{P}(y = t)/\mathbb{P}(x = t)$ is increasing in t, or equivalently, $\mathbb{P}(y = t')\mathbb{P}(x = t) \geq \mathbb{P}(y = t)\mathbb{P}(x = t')$ for all $t' > t$. This is stronger than the stochastic order: $\mathcal{L}(y) \geq_{\mathrm{lr}} \mathcal{L}(x)$ implies $\mathcal{L}(y) \geq_{\mathrm{st}} \mathcal{L}(x)$. Moreover, $\mathcal{L}(y) \geq_{\mathrm{lr}} \mathcal{L}(x)$ implies $\mathcal{L}(y|y \in A) \geq_{\mathrm{lr}} \mathcal{L}(x|x \in A)$, and hence $\mathcal{L}(y|y \in A) \geq_{\mathrm{st}} \mathcal{L}(x|x \in A)$, for any measurable subset A of the real line. In fact, the latter condition of stochastic dominance for every A is equivalent to $\mathcal{L}(y) \geq_{\mathrm{lr}} \mathcal{L}(x)$; see [22].

In the standard Bayesian setup we have a prior π with support Θ, and so $\pi(\theta) > 0$ for every $\theta \in \Theta$, and a random variable s whose distribution depends on θ. The conditional distribution $\mathbb{P}(s|\theta)$, namely, the distribution of s given θ, is denoted by $\mathbb{P}_\theta$, and its probability law by $\mathcal{L}_\theta$. We also consider the marginal probability (also called the "prior predictive probability") $\mathbb{P}(s) := \sum_{\theta \in \Theta} \mathbb{P}_\theta(s)\pi(\theta)$, and denote its law by $\mathcal{L}$. Expectations with respect to $\mathbb{P}$, $\mathbb{P}_\theta$, and $\mathbb{P}_\Gamma(s) := \mathbb{P}(s|\Gamma) = \sum_{\theta \in \Gamma} \pi(\theta)\mathbb{P}_\theta(s) / \sum_{\theta \in \Gamma} \pi(\theta)$, for a set of parameters $\Gamma \subset \Theta$, are denoted by $\mathbb{E}$, $\mathbb{E}_\theta$, and $\mathbb{E}_\Gamma$, respectively. We use the notation

$$q^\theta \equiv q^\theta(s) := \mathbb{P}(\theta|s)$$

for the *posterior* probability of θ (the "*θ-posterior*" for short) given s. We will compare random variables like q^θ under different distributions, such as $\mathbb{P}$ and $\mathbb{P}_\theta$.

We now summarize several simple results with short proofs. They are given in [13] with more details, interpretations, and related references.

Proposition 1.

(i) *Let P_1 and P_2 be two probability measures on a measure space $\mathcal{S}$ such that $P_1 \ll P_2$ (i.e., $P_2(s) = 0$ implies $P_1(s) = 0$), and let $r(s) = P_1(s)/P_2(s)$ be the likelihood ratio.*[1] *Then*

$$\mathcal{L}_{P_1}(r) \geq_{\mathrm{lr}} \mathcal{L}_{P_2}(r),$$

where $\mathcal{L}_{P_i}$ denotes the probability law with respect to P_i, and for any increasing function f,

$$\mathcal{L}_{P_1}(f(r)) \geq_{\mathrm{lr}} \mathcal{L}_{P_2}(f(r)). \tag{3}$$

(ii) *In the Bayesian setup, for every θ the posterior $q^\theta(s) \equiv \mathbb{P}(\theta|s)$ of θ satisfies*

$$\mathcal{L}_\theta(q^\theta) \geq_{\mathrm{lr}} \mathcal{L}(q^\theta). \tag{4}$$

Proof. (i) For every value t of r let $B := \{s : \frac{P_1(s)}{P_2(s)} = t\}$; then $P_1(r = t) = \sum_{s \in B} P_1(s) = t \sum_{s \in B} P_2(s) = t P_2(r = t)$. It follows that $\frac{P_1(r=t)}{P_2(r=t)} = t$, which is an increasing function of t, and so

[1] When the ratio is $0/0$ define $r(s)$ arbitrarily; this will not matter since it occurs on a null event for both P_1 and P_2.

we have the result for r by definition. The result for increasing f then follows readily (see [22], Theorem 1.C.8).

(ii) Setting $P_1 = \mathbb{P}_\theta$ and $P_2 = \mathbb{P}$ we have $q^\theta(s) = \pi(\theta)\frac{P_1(s)}{P_2(s)} = \pi(\theta)r(s)$, and the result follows from (i) since $\mathbb{P}_\theta \ll \mathbb{P}$. □

Remark. In (i) the ratio $P_1(r = t)/P_2(r = t) = t$ is a *strictly* increasing function of t—unless r is constant, which happens only when $P_1 \equiv P_2$ (and then $r \equiv 1$)—and the domination is therefore *strict*. In (ii) this happens except when the signal s is completely uniformative, i.e., when the posterior is identical to the prior: $q^\theta(s) = \pi(\theta)$ for all s. The strict domination implies that the resulting inequalities, such as (6) below, are strict (this is pointed out in [13]).

Since likelihood ratio order implies stochastic order, (3) implies $\mathcal{L}_{P_1}(f(r)) \geq_{\mathrm{st}} \mathcal{L}_{P_2}(f(r))$, and so for any increasing function f we have

$$\sum_s P_1(s) f\left(\frac{P_1(s)}{P_2(s)}\right) \geq \sum_s P_2(s) f\left(\frac{P_1(s)}{P_2(s)}\right).$$

The quantity on the right-hand side is known (for convex f) as *f-divergence*. Similarly, the likelihood ratio order relation (4) implies stochastic order; that is, $\mathbb{E}_\theta[f(q^\theta(s))] \geq \mathbb{E}[f(q^\theta(s))]$ for increasing f. Moreover, this holds also when conditioning on a set of values of the posterior (such as being, say, more than 1/2):

$$\mathbb{E}_\theta[f(q^\theta(s))|q^\theta \in A] \geq \mathbb{E}[f(q^\theta(s))|q^\theta \in A] \tag{5}$$

for increasing f and $A \subseteq [0, 1]$. Thus the posterior of θ when the data s are generated according to $\mathbb{P}_\theta(s)$ is stochastically larger than the posterior when the data are generated under $\mathbb{P}(s)$. Taking $f(x) = x$ we obtain

$$\mathbb{E}_\theta[q^\theta] \geq \mathbb{E}[q^\theta] = \pi(\theta), \tag{6}$$

where the inequality is strict for any informative signal s (see the above remark), and the equality is the *martingale property* of posteriors under $\mathbb{P}$ (see (1)). Replacing the single parameter θ_0 with a set $\Gamma \subset \Theta$ of parameter values, the result of (4) readily implies that

$$\mathcal{L}_\Gamma(\mathbb{P}(\Gamma|s)) \geq_{\mathrm{lr}} \mathcal{L}(\mathbb{P}(\Gamma|s)) \geq_{\mathrm{lr}} \mathcal{L}_{\Gamma^c}(\mathbb{P}(\Gamma|s)) \tag{7}$$

(for the second $\geq_{\mathrm{lr}}$ use $\mathbb{P} = \pi(\Gamma)\mathbb{P}_\Gamma + \pi(\Gamma^c)\mathbb{P}_{\Gamma^c}$), and thus

$$\mathbb{E}_\Gamma[\mathbb{P}(\Gamma|s)] \geq \mathbb{E}[\mathbb{P}(\Gamma|s)] = \pi(\Gamma) \geq \mathbb{E}_{\Gamma^c}[\mathbb{P}(\Gamma|s)]. \tag{8}$$

Result (7) is given in [13], and (8) is given in [11]; see these papers for further results, references, and history.

When there are only two parameter values, say $\Theta = \{\theta_0, \theta_1\}$, (7) becomes $\mathcal{L}_{\theta_0}(q^{\theta_0}) \geq_{\mathrm{lr}} \mathcal{L}(q^{\theta_0}) \geq_{\mathrm{lr}} \mathcal{L}_{\theta_1}(q^{\theta_0})$. However, the latter dominance relation need *not* hold when there are additional parameter values in Θ; see, for instance, the example at the end of Section 1 in [13].[2] A case where it does hold is provided in the proposition below, which applies, for instance, to the Bernoulli and normal distributions, and many other exponential families discussed later; see, e.g., [16] or [14]. The parameter space is now an interval on the real line, θ_0 and θ_1 are the two interval ends, and the family of distributions $\mathbb{P}_\theta$ satisfies the *monotone likelihood ratio property* (*MLRP*), i.e., $\mathbb{P}_{\theta'}(s)/\mathbb{P}_\theta(s)$ is increasing in $s \in \mathbb{R}$

[2] Where $\Theta = \{\alpha, \beta, \gamma\}$ and the dominance is reversed: $\mathcal{L}_\gamma(q^\alpha) >_{\mathrm{lr}} \mathcal{L}(q^\alpha)$ (in the notation of the present paper, [13] shows that $\mathcal{L}_\gamma(1 - q^\alpha) <_{\mathrm{lr}} \mathcal{L}(1-q^\alpha)$).

for $\theta' > \theta$. The order $\geq_{\text{lr}}$ below can of course be replaced by the weaker $\geq_{\text{st}}$ or by comparisons of expectations.

Proposition 2. *Let $\Theta = [\theta_0, \theta_1] \subset \mathbb{R}$ and assume that $\mathbb{P}_\theta$ is a monotone likelihood ratio (MLRP) family. Then*

$$\mathcal{L}_{\theta_0}(q^{\theta_0}) \geq_{\text{lr}} \mathcal{L}(q^{\theta_0}) \geq_{\text{lr}} \mathcal{L}_{\theta_1}(q^{\theta_0}).$$

Proof. The only new part is $\mathcal{L}(q^{\theta_0}) \geq_{\text{lr}} \mathcal{L}_{\theta_1}(q^{\theta_0})$. First, $\frac{\mathbb{P}_{\theta_1}(s)}{\mathbb{P}(s)}$ is increasing in s by the MLRP assumption, because the denominator is a mixture of $\mathbb{P}_\theta$'s over $\theta \leq \theta_1$, and so $\mathcal{L}_{\theta_1}(s) \geq_{\text{lr}} \mathcal{L}(s)$. Similarly, $q^{\theta_0}(s) = \pi(\theta_0)\frac{\mathbb{P}_{\theta_0}(s)}{\mathbb{P}(s)}$ is decreasing in s, and the argument in the proof of Proposition 1 (i) leading to (3), now applied to a decreasing rather than increasing function and thus reversing the order, implies $\mathcal{L}(q^{\theta_0}) \geq_{\text{lr}} \mathcal{L}_{\theta_1}(q^{\theta_0})$. □

We conclude with a simple symmetry between the θ_0-posterior under θ_1 and the θ_1-posterior under θ_0.

Proposition 3. *Let θ_0 and θ_1 be in Θ. Then*

$$\frac{1}{\pi(\theta_0)}\mathbb{E}_{\theta_1}[q^{\theta_0}] = \frac{1}{\pi(\theta_1)}\mathbb{E}_{\theta_0}[q^{\theta_1}] = \frac{1}{\pi(\theta_0)\pi(\theta_1)}\mathbb{E}[q^{\theta_0} \cdot q^{\theta_1}].$$

Proof. We have

$$\begin{aligned}\frac{1}{\pi(\theta_0)}\mathbb{E}_{\theta_1}[q^{\theta_0}] &= \frac{1}{\pi(\theta_0)}\sum_s \frac{\mathbb{P}_{\theta_0}(s)\pi(\theta_0)}{\mathbb{P}(s)}\mathbb{P}_{\theta_1}(s)\\ &= \frac{1}{\pi(\theta_0)\pi(\theta_1)}\sum_s \frac{\mathbb{P}_{\theta_0}(s)\pi(\theta_0)}{\mathbb{P}(s)}\frac{\mathbb{P}_{\theta_1}(s)\pi(\theta_1)}{\mathbb{P}(s)}\mathbb{P}(s) = \frac{1}{\pi(\theta_0)\pi(\theta_1)}\mathbb{E}[q^{\theta_0} \cdot q^{\theta_1}].\end{aligned}$$

The last expression is symmetric in θ_0 and θ_1, and so it is equal to $\frac{1}{\pi(\theta_1)}\mathbb{E}_{\theta_0}[q^{\theta_1}]$ as well. □

The same symmetry applies of course when we consider sequences of observations (see for instance Corollary 7 below).

3. Increasing posterior of the true state

We now consider observations that arrive sequentially and apply the results of Section 2 to obtain monotonicity and order relations as a function of the sample size n.

The data consist of a process of observations $x_1, x_2, \ldots$ whose distribution is $\mathbb{P}_\theta$, where θ lies in the parameter space Θ. At this point we make no assumptions about the distributions of the observation process and the dependence structure (over n). We assume the standard Bayesian framework given in Section 1. Given the vector of observations up to stage n, which we denote by $s_n = (x_1, \ldots, x_n)$, the *posterior* of θ at time n is $q_n^\theta \equiv \mathbb{P}(\theta|s_n)$. Viewing q_n^θ as an s_n-measurable random variable, we obtain that the sequence q_n^θ is a martingale with respect to the probability $\mathbb{P}$, i.e., $\mathbb{E}[q_{n+1}^\theta|s_n] = q_n^\theta$.

Proposition 1 (ii) applied to $x_{n+1}|s_n$ yields

Proposition 4. *For every θ, the posterior q_{n+1}^{θ} of θ at time $n+1$ satisfies*

$$\mathcal{L}_\theta\left(q_{n+1}^{\theta}|s_n\right) \geq_{\mathrm{lr}} \mathcal{L}\left(q_{n+1}^{\theta}|s_n\right).$$

Thus, given s_n, the θ-posterior $q_{n+1}^{\theta} \equiv \mathbb{P}(\theta|x_{n+1}, s_n)$ under the probability $\mathbb{P}_\theta(\cdot|s_n)$ likelihood-ratio dominates that same θ-posterior under the probability $\mathbb{P}(\cdot|s_n)$. Since, again, $\geq_{\mathrm{lr}}$ implies $\geq_{\mathrm{st}}$, by Proposition 4 and the martingale property of $q_n^{\theta} \equiv \mathbb{P}(\theta|s_n)$ we get

$$\mathbb{E}_\theta[q_{n+1}^{\theta}|s_n] \geq \mathbb{E}[q_{n+1}^{\theta}|s_n] = q_n^{\theta} \tag{9}$$

(as in (5), one may also condition on q_{n+1}^{θ} lying in a certain set). Proposition 4 generalizes Proposition 1 (ii): given any past data s_n, the posterior belief in θ with an additional observation x_{n+1} distributed according to $\mathbb{P}_\theta$ likelihood-ratio dominates the same posterior when the additional observation is distributed according to $\mathbb{P}$; (9) is then the consequent expectation comparison.

Inequality (9) means that under $\mathbb{P}_\theta$ the process q_n^{θ} is a submartingale. Since every increasing convex (integrable) function of a submartingale is a submartingale, for any increasing convex f we have

$$\mathbb{E}_\theta[f(q_{n+1}^{\theta})|s_n] \geq f(q_n^{\theta}). \tag{10}$$

Taking expectations on the two sides of (9) and (10) with respect to s_n distributed under θ we obtain

Corollary 5. *The expectation $\mathbb{E}_\theta[q_n^{\theta}]$ of the posterior probability of θ with respect to $\mathbb{P}_\theta$ is increasing in n, and, more generally, so is $\mathbb{E}_\theta[f(q_n^{\theta})]$ for any convex increasing function f.*

Thus, with more data generated under θ, the θ-posterior increases in the increasing convex order, and in particular in expectation. The convergence of the expected posterior to 1 (by Doob's theorem) is thus monotone. We have obtained this result starting from a strong ordering: the likelihood ratio order. That q_n^{θ} is a submartingale under $\mathbb{P}_\theta$ is shown in [17] and [11] (see also the references therein), where other relevant results are given.

4. Nonmonotonic and multimodal expected posteriors

Assume now that the observations are generated under θ_1, which is different from θ_0; then the posterior probability of θ_0 given s_n converges to zero as $n \to \infty$ by Doob's theorem. This convergence may not be monotone, and in fact, if θ_1 and θ_0 are close together, the expected posterior $\psi(n) \equiv \psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ may increase as a function of n for small n before it starts decreasing. The question that we address in this section is whether it is possible for the expected θ_0-posterior, with data generated under θ_1, to increase again after it starts decreasing. If some observations distributed under θ_1 cause the expected θ_0-posterior to decrease, the average belief in θ_0 decreases, as it should under θ_1, but as we will show below it is possible for further observations generated under the same θ_1 to cause $\psi(n)$ to increase before it eventually decreases to zero. Such an increase leads to an erroneous upturn of the Bayesian statistician's degree of belief in θ_0. Furthermore, $\psi(n)$ need not be unimodal, and may fluctuate many times.

We present two examples of this behavior in the simplest case of iid coin tosses, i.e., Bernoulli(θ) observations. In Figure 1 the sequence $\psi(n)$ decreases, then increases, and then decreases again to 0. Figure 2 provides a further counterexample to the unimodality of the sequence $\psi(n)$, showing that it may have many modes, and thus may alternate several times between being increasing and decreasing.

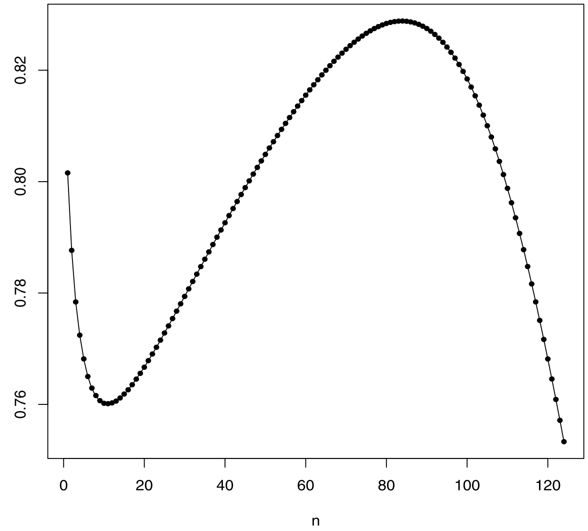

Figure 1. The sequence $\psi(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ for iid Bernoulli observations: $\Theta = \{\theta_0, \theta_1, \theta_2\}$ with $\theta_j = 0.5,\ 0.65,\ 0.85$, and prior probabilities $\alpha_j = \pi(\theta_j) = 4100/5001,\ 1/5001,\ 900/5001$ (for $j = 0, 1, 2$).

In both examples the priors concentrate on three points θ_j (for $j = 0, 1, 2$) with probabilities $\pi(\theta_j) = \alpha_j$, and so the distribution of the sufficient statistic $u_n := \sum_{i=1}^n x_i$ is a mixture of three Binomial(n, θ_i) distributions: $\mathbb{P}(u_n = k) = \sum_{j=0}^2 \alpha_j \mathbb{P}_{\theta_j}(u_n = k)$. By continuity, it is clear that such examples are robust to small changes in the parameters and their associated probabilities, and priors having a larger or even continuous support with a similar behavior can be constructed. The function $\psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$, which is given by the formula

$$\psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}] = \sum_{k=0}^{n} \frac{\binom{n}{k}^2 \theta_0^k (1-\theta_0)^{n-k} \theta_1^k (1-\theta_1)^{n-k} \pi(\theta_0)}{\mathbb{P}(u_n = k)} \tag{11}$$

(and thus equals $\frac{\pi(\theta_0)}{\pi(\theta_1)} \mathbb{E}_{\theta_0}[q_n^{\theta_1}]$, as in Proposition 3), is depicted in Figures 1 and 2. In Figure 3 we provide an example of iid observations with a normal prior, which will be analyzed in Section 6.3.

For some intuitive explanations, take Figure 1 first, where we consider the posterior of $\theta_0 = 0.5$, the data are generated under $\theta_1 = 0.65$, and there is another possible parameter, $\theta_2 = 0.85$. Initially the expected posterior belief in θ_0 decreases, as the observations under θ_1 make both θ_1 and θ_2 seem more likely (on average). After 11 observations the expected belief in θ_0 starts to increase, as θ_2, whose distance from θ_1 (under which the data are generated) is greater than that of θ_0, begins to seem less likely. Eventually, after a further 70 observations, the expected θ_0-posterior begins its final descent to 0.

Turning to Figure 2, we see that, in addition to long-term fluctuations similar to those of Figure 1 (see also Figure 3), there are many short-term up and down fluctuations[3] that are likely due to the discreteness of the data (cf. Figure 3) and of the time steps. In Proposition 6 and Corollary 7 below we try to shed some light on these fluctuations. Interestingly, the fluctuations of the expected posterior beliefs are rather sensitive to the values of the parameters; for example, changing θ_1 from 0.5 to 0.45, or θ_2 from 0.85 to 0.8, yields a strictly decreasing sequence $\psi(n)$, with no up and down fluctuations.

[3] We have generated other examples where the number of modes exceeds 8 by far.

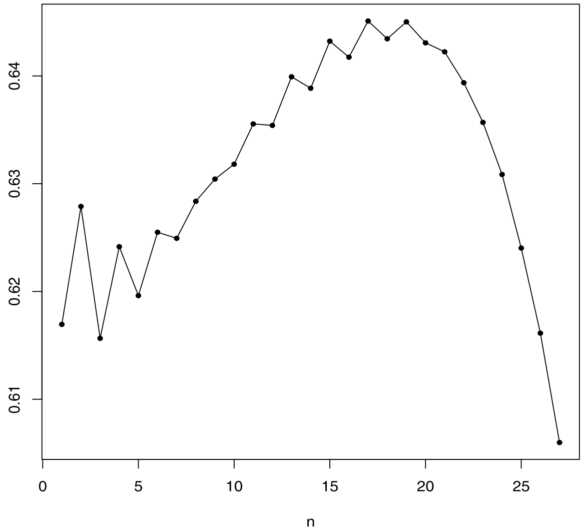

Figure 2. The sequence $\psi(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ for iid Bernoulli observations: $\Theta = \{\theta_0, \theta_1, \theta_2\}$ with $\theta_j = 0.2,\ 0.5,\ 0.85$, and prior probabilities $\alpha_j = \pi(\theta_j) = 2000/3001,\ 1/3001,\ 1000/3001$ (for $j = 0, 1, 2$); we see 8 modes.

In Figure 3 we have iid normal $\mathcal{N}(\theta, \sigma^2)$ observations with a large variance ($\sigma = 100$) and a standard normal prior (see Section 6.3). For data generated under $\theta_1 = 1/3$, the expected belief in $\theta_0 = -1/3$ starts by decreasing for nearly 2,000 periods, following which it goes in the "wrong direction" for a very long time, increasing for about 30,000 additional observations; only then it starts decreasing

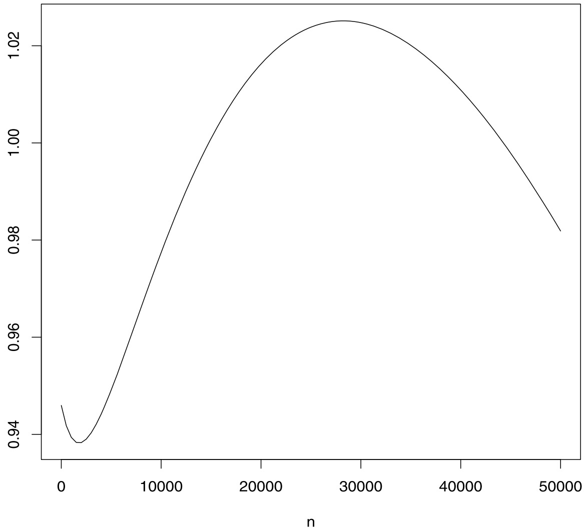

Figure 3. The sequence $\psi(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ for iid $\mathcal{N}(\theta, \sigma^2)$ normal observations with a standard normal prior on θ (see (31)): $\theta_0 = -1/3$, $\theta_1 = 1/3$ and $\sigma = 100$.

monotonically to its limit of 0. As we will show in Theorem 16 (iv), additional up and down turns are not possible for normal distributions.

Consider now the effect of a single observation x on the θ_0-belief when x is generated under a different θ_1. As seen above, while in many cases the belief in θ_0 is reduced, this need not be so when θ_1 is close to θ_0 and there are other points in Θ; indeed, the θ_0-posterior strictly increases on average when $\theta_1 = \theta_0$ (see the remark after Proposition 1), and thus, by continuity, also when θ_1 and θ_0 are close enough together. We quantify precisely this "closeness" of parameters in the simplest case of a Bernoulli observation, as follows.

Proposition 6. *Let x be a* Bernoulli(θ) *observation, and let π be the prior probability of θ for $\theta \in \Theta \subseteq [0, 1]$. Let $\overline{\theta} := \mathbb{E}[\theta] = \sum_{\theta\in\Theta} \theta\pi(\theta)$ be the (prior) average parameter. For any θ_0 and θ_1 in Θ, the inequality $\mathbb{E}_{\theta_1}[q^{\theta_0}] \le \pi(\theta_0)$ holds if and only if $\overline{\theta}$ lies between θ_0 and θ_1, with equality if and only if $\overline{\theta} = \theta_0$ or $\overline{\theta} = \theta_1$.*

The condition that $\overline{\theta}$ lies between θ_0 and θ_1 is symmetric in θ_0 and θ_1, and so it is also equivalent to $\mathbb{E}_{\theta_0}[q^{\theta_1}] \le \pi(\theta_1)$ (this equivalence follows from Proposition 3 as well). Assume therefore without loss of generality that $\theta_0 \le \theta_1$. Proposition 6 says that data under θ_1 make the expected posterior of θ_0 lower than the prior of θ_0 if and only if $\theta_0 \le \overline{\theta} \le \theta_1$, and make it higher than the prior if and only if either $\theta_0, \theta_1 \le \overline{\theta}$ or $\theta_0, \theta_1 \ge \overline{\theta}$. Thus θ_1 being "close" to θ_0 in the sense that an observation under θ_1 increases the expected belief in θ_0 (i.e., $\mathbb{E}_{\theta_1}[q^{\theta_0}] \ge \pi(\theta_0)$ holds) is equivalent to θ_1 being on the same side of $\overline{\theta}$ as θ_0. In fact, this notion of "belief-close" is related to "metric-close" as follows: the shorter the distance $|\theta_1 - \theta_0|$ between θ_1 and θ_0 is, the fewer the priors π yielding $\overline{\theta}$ in the interval $[\theta_0, \theta_1]$ there are, and so the fewer the priors π yielding $\mathbb{E}_{\theta_1}[q^{\theta_0}] \ge \pi(\theta_0)$ there are. Finally, when $\theta_1 = \theta_0$, Proposition 6 gives (6), with strict inequality unless $\overline{\theta} = \theta_0$.

Proof. Using $\mathbb{P}(x = 1) = \overline{\theta}$ we have

$$\begin{aligned}\mathbb{E}_{\theta_1}[q^{\theta_0}] &= \theta_1 q^{\theta_0}(1) + (1-\theta_1)q^{\theta_0}(0) = \theta_1 \frac{\theta_0\pi(\theta_0)}{\overline{\theta}} + (1-\theta_1)\frac{(1-\theta_0)\pi(\theta_0)}{1-\overline{\theta}}\\ &= \pi(\theta_0)V(\overline{\theta}),\end{aligned} \tag{12}$$

where

$$V(y) := \frac{\theta_0\theta_1}{y} + \frac{(1-\theta_0)(1-\theta_1)}{1-y}. \tag{13}$$

The function V is strictly convex and satisfies $V(\theta_0) = V(\theta_1) = 1$, and so $V(\overline{\theta}) \le 1$ if and only if $\theta_0 \le \overline{\theta} \le \theta_1$. □

This extends to a sequence $x_1, x_2, \ldots$ of iid Bernoulli(θ) observations, where we obtain the condition for the expected posterior to increase after the next observation, given the past data. Indeed, using (12) for the observation x_{n+1} that comes after data $s_n = (x_1, \ldots, x_n)$ yields

Corollary 7. *Let $x_1, x_2, \ldots$ be iid* Bernoulli(θ) *observations, let π be the prior probability of θ for $\theta \in \Theta \subseteq [0, 1]$, and denote by $\overline{\theta}_n := \mathbb{E}[\theta|s_n] = \sum_{\theta\in\Theta} \theta\mathbb{P}(\theta|s_n)$ the average parameter conditional on s_n. For any θ_0 and θ_1 in Θ we have*

$$\mathbb{E}_{\theta_1}[q^{\theta_0}_{n+1}|s_n] = q^{\theta_0}_n V(\overline{\theta}_n) \;\; and \;\; \mathbb{E}_{\theta_0}[q^{\theta_1}_{n+1}|s_n] = q^{\theta_1}_n V(\overline{\theta}_n), \tag{14}$$

where the function V is given by (13), *and so each one of the inequalities $\mathbb{E}_{\theta_1}[q_{n+1}^{\theta_0}|s_n] \le q_n^{\theta_0}$ and $\mathbb{E}_{\theta_0}[q_{n+1}^{\theta_1}|s_n] \le q_n^{\theta_1}$ holds if and only if $\overline{\theta}_n$ lies between θ_0 and θ_1.*

Thus θ_0 and θ_1 are "not close together" if they are separated by $\overline{\theta}_n$, in which case an additional observation under θ_1 decreases the expected belief in θ_0, and vice versa. The notion of closeness depends, of course, on n and s_n. Since for data generated under θ_1 (in the support of π) we have $\overline{\theta}_n \to \theta_1$, as n increases θ_0 and θ_1 may well fluctuate between being and not being close together given s_n. Despite these fluctuations, we will see a large class of natural setups (Theorem 9 and Corollary 10) where the overall expectation $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is strictly decreasing from some n on.

5. Asymptotic rates and eventual monotonicity

In this section we obtain the precise asymptotic behavior of $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ as $n \to \infty$ in the commonly used rich class of *exponential families* of distributions; for clarity we focus on one-dimensional families (see Appendix A.2 for extensions). The prior is now assumed to be continuous (which allows the use of analysis tools), while the data may be discrete or continuous. All the results of the previous sections are clearly seen to extend here, with densities replacing probabilities as needed.

The general setup is as follows. Let $\mathcal{X} \subseteq \mathbb{R}$ be the space of observations, and let ν be a σ-finite measure on the Borel sets of $\mathcal{X}$; for instance, take ν to be the counting measure in the discrete case where $\mathcal{X}$ is a finite or countable set, and the Lebesgue measure in the continuous case where $\mathcal{X}$ is a bounded or unbounded interval. Let $\Theta \subseteq \mathbb{R}$ be the space of parameters, and let Π, the prior, be a probability measure on Θ; we assume that Θ is a convex set and that Π has a density $\pi(\theta)$ that is continuous and strictly positive on Θ. Finally, the conditional-on-θ probability $\mathbb{P}_\theta(\cdot) \equiv \mathbb{P}(\cdot|\theta)$ on $\mathcal{X}$ has a density p_θ with respect to ν, given by[4]

$$p_\theta(x) = \exp(\eta(\theta)T(x) - A(\eta(\theta)) - B(x)) \tag{15}$$

for some functions η, T, A, and B, where η is differentiable and $\eta'(\theta) > 0$ for all $\theta \in \Theta$. The following are well known (see, e.g., [5]): the function $A(\eta)$ is determined by the functions B and T (use the condition $\int_{\mathcal{X}} p_\theta(x)\,\mathrm{d}\nu(x) = 1$ for every $\theta \in \Theta$), it is infinitely differentiable, $A'(\eta(\theta)) = \mathbb{E}_\theta[T(x)]$, and $A''(\eta(\theta)) = \mathbb{V}ar_\theta(T(x))$. We assume that $A''(\eta(\theta)) > 0$ (that is, $T(x)$ is not $\mathbb{P}_\theta$-a.s. constant) for all $\theta \in \Theta$. Let $I(\theta) := \mathbb{E}_\theta[(\partial/\partial\theta \log p_\theta(x))^2] = -\mathbb{E}_\theta[\partial^2/\partial\theta^2 \log p_\theta(x)]$ be the *Fisher information* at θ; for exponential families (15) we have $I(\theta) = A''(\eta(\theta)) \cdot (\eta'(\theta))^2 = \mathbb{V}ar_\theta(T(x)) \cdot (\eta'(\theta))^2 > 0$.

Consider a sequence of iid observations $x_1, x_2, \ldots$ and set $s_n = (x_1, \ldots, x_n) \in \mathcal{X}^n$. The density of the θ_0-posterior (for $\theta_0 \in \Theta$) is[5]

$$q_n^{\theta_0} \equiv q_n^{\theta_0}(s_n) = \frac{p_{\theta_0}(s_n)\pi(\theta_0)}{p(s_n)},$$

where p denotes the marginal density, i.e., $p(s_n) = \int_\Theta p_\theta(s_n)\pi(\theta)\,\mathrm{d}\theta$. The θ_1-expectation (for $\theta_1 \in \Theta$) of the θ_0-posterior is[6]

$$\psi_{\theta_0,\theta_1}(n) \equiv \mathbb{E}_{\theta_1}[q_n^{\theta_0}] = \int_{\mathcal{X}^n} \frac{p_{\theta_0}(s_n)\pi(\theta_0)}{p(s_n)} p_{\theta_1}(s_n)\,\mathrm{d}\nu^n(s_n). \tag{16}$$

[4] Formally, p_θ is the Radon–Nikodym derivative $\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\nu$. In the discrete case where ν is the counting measure, $p_\theta(x) = \mathbb{P}_\theta(x)$ for all x, and $\int_Y p_\theta(x)\,\mathrm{d}\nu(x) = \sum_{x\in Y} p_\theta(x) = \sum_{x\in Y} \mathbb{P}_\theta(x)$ for every $Y \subseteq \mathcal{X}$.

[5] We use the notation $p_\theta(\cdot)$ for the conditional-on-θ density of any variable; thus, $p_\theta(s_n) = \prod_{i=1}^n p_\theta(x_i)$.

[6] The measure ν^n on $\mathcal{X}^n$ is the n-fold product of the measure ν on $\mathcal{X}$.

In exponential families there is a simple relation between ψ_{θ_0,θ_1} and $\psi_{\theta,\theta}$ for an appropriate $\theta \in \Theta$.

Proposition 8. *Let $(p_\theta)_{\theta\in\Theta}$ be a family of densities (15), and let π be a positive density on the convex set $\Theta \subseteq \mathbb{R}$. Let $\theta_0, \theta_1 \in \Theta$. Then*

$$\psi_{\theta_0,\theta_1}(n) = \frac{\pi(\theta_0)}{\pi(\theta_2)}\psi_{\theta_2,\theta_2}(n)\, w^n$$

for every $n \geq 1$, where

$$\theta_2 = \eta^{-1}\left(\frac{\eta(\theta_0)+\eta(\theta_1)}{2}\right) \in \Theta \tag{17}$$

and[7]

$$w = \left(\int_{\mathcal{X}} \sqrt{p_{\theta_0}(x)\, p_{\theta_1}(x)}\, \mathrm{d}\nu(x)\right)^2 \leq 1, \tag{18}$$

with equality (i.e., $w = 1$) if and only if $\theta_0 = \theta_1$.

Proof. The function η is strictly increasing and continuous, and so it attains the value $(\eta(\theta_0)+\eta(\theta_1))/2$ at a (unique) point between θ_0 and θ_1, and thus in Θ (which is a convex set); this is the point θ_2 given by (17). From (15) we have

$$p_{\theta_0}(x)\, p_{\theta_1}(x) = w\, p_{\theta_2}(x)^2, \tag{19}$$

for every x, where $w := \exp(2A(\eta_2) - A(\eta_0) - A(\eta_1))$ and $\eta_i := \eta(\theta_i)$ (and so $\eta_2 = (\eta_0 + \eta_1)/2$). Therefore $p_{\theta_0}(s_n)\, p_{\theta_1}(s_n) = w^n p_{\theta_2}(s_n)^2$, yielding the result by (16). Formula (18) follows from (19): $\int_{\mathcal{X}} \sqrt{p_{\theta_0}(x)\, p_{\theta_1}(x)} = \int_{\mathcal{X}} \sqrt{w}\, p_{\theta_2}(x) = \sqrt{w}$. Since $\int_{\mathcal{X}} \sqrt{p_{\theta_0}(x)\, p_{\theta_1}(x)} \leq \int_{\mathcal{X}}(p_{\theta_0}(x) + p_{\theta_1}(x))/2 = 1$ we get $w \leq 1$, with equality if and only if $p_{\theta_0}(x) = p_{\theta_1}(x)$ for all $x \in \mathcal{X}$, which occurs if and only if $\theta_0 = \theta_1$ (because η is one-to-one and T is not constant). □

Proposition 8 thus reduces the analysis of the $\theta_1 \neq \theta_0$ case to that of the $\theta_1 = \theta_0$ case. The relation (19), or, equivalently, $p_{\theta_2} = c\sqrt{p_{\theta_0}\, p_{\theta_1}}$ (for the constant $c = 1/\sqrt{w}$), says that the density p_{θ_2} is proportional to the geometric average of the densities p_{θ_1} and p_{θ_2}; it is their "normalized geometric average." The exponential family (15) is closed under this averaging operation by the convexity of Θ. For further discussions and extensions, see Appendix A.2.

When $\theta_1 = \theta_0$ the sequence $\mathbb{E}_{\theta_0}[q_n^{\theta_0}]$ is monotonically increasing (by Corollary 5), whereas when $\theta_1 \neq \theta_0$ the sequence $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ converges to 0 (by Doob's theorem). Theorem 9 strengthens these results by providing the precise asymptotic rates.

The standard notation "$f(n) \sim g(n)$ as $n \to \infty$" means "asymptotic equivalence," i.e., $\lim_{n\to\infty} f(n)/g(n) = 1$.

Theorem 9. *Let $x_1, x_2, \ldots$ be iid observations distributed according to an exponential family with density p_θ given by (15), and let θ be distributed according to a prior having a continuous strictly positive density π on a convex set $\Theta \subseteq \mathbb{R}$.*

[7] The constant $\sqrt{w}$ is known as the Bhattacharrya coefficient, and also as the Chernoff 1/2-coefficient (see [6] and [8], or [19] for a convenient reference).

(i) *Let θ_0 be an interior point of Θ; then*

$$\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \sim \frac{\sqrt{I(\theta_0)}}{2\sqrt{\pi}}\sqrt{n} \quad \textit{as } n \to \infty. \tag{20}$$

(ii) *Let $\theta_1 \neq \theta_0$ be in Θ; then*

$$\mathbb{E}_{\theta_1}[q_n^{\theta_0}] = \frac{\pi(\theta_0)}{\pi(\theta_2)}\mathbb{E}_{\theta_2}[q_n^{\theta_2}]\, w^n \sim \frac{\pi(\theta_0)}{\pi(\theta_2)}\frac{\sqrt{I(\theta_2)}}{2\sqrt{\pi}}\sqrt{n}\, w^n \quad \textit{as } n \to \infty, \tag{21}$$

where θ_2 and w are given by (17)–(18).

Corollary 10. *When $\theta_1 \neq \theta_0$ the sequence $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is eventually strictly decreasing.*

Proof. As $n \to \infty$ we have $\mathbb{E}_{\theta_1}[q_{n+1}^{\theta_0}]/\mathbb{E}_{\theta_1}[q_n^{\theta_0}] \sim (\sqrt{n+1}/\sqrt{n})w \to w < 1$. □

Note that the prior π does not appear in (20); as in the Bernstein–von Mises theorem (which is used in the proof below), the prior is "overwhelmed" by the data as the number of observations increases; this is however *not* the case in (21) when $\theta_1 \neq \theta_0$.

Before proving the theorem, we apply it to a number of classical examples:

- **Bernoulli**(θ) for $\theta \in \Theta = (0,1)$. Here $\mathcal{X} = \{0,1\}$, $\eta(\theta) = \log(\theta/(1-\theta))$, and $T(x) = x$, and so $\mathbb{V}ar_\theta(T(x)) = \theta(1-\theta)$ and $\eta'(\theta) = 1/(\theta(1-\theta))$, yielding $I(\theta) = 1/(\theta(1-\theta))$ and

$$\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \sim \frac{\sqrt{n}}{2\sqrt{\pi\theta_0(1-\theta_0)}} \quad (\text{for } 0 < \theta_0 < 1),$$
$$\mathbb{E}_{\theta_1}[q_n^{\theta_0}] \sim \frac{\pi(\theta_0)\sqrt{n}\, w^n}{2\pi(\theta_2)\sqrt{\pi\theta_2(1-\theta_2)}} \quad (\text{for } \theta_0 \neq \theta_1),$$

where

$$\theta_2 = \frac{\sqrt{\theta_0\theta_1}}{\sqrt{\theta_0\theta_1} + \sqrt{(1-\theta_0)(1-\theta_1)}} \text{ and } w = \left(\sqrt{\theta_0\theta_1} + \sqrt{(1-\theta_0)(1-\theta_1)}\right)^2.$$

- **Normal**(θ,σ^2) for $\theta \in \Theta = \mathbb{R}$ and fixed $\sigma > 0$. Here $\mathcal{X} = \mathbb{R}$, $\eta(\theta) = \theta$, and $T(x) = x/\sigma^2$, and so $I(\theta) = \mathbb{V}ar_\theta(T(x)) = 1/\sigma^2$, yielding

$$\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \sim \frac{\sqrt{n}}{2\sigma\sqrt{\pi}},$$
$$\mathbb{E}_{\theta_1}[q_n^{\theta_0}] \sim \frac{\pi(\theta_0)\sqrt{n}}{2\pi(\theta_2)\sigma\sqrt{\pi}}\exp\left(-\frac{n(\theta_0-\theta_1)^2}{4\sigma^2}\right),$$

where $\theta_2 = (\theta_0+\theta_1)/2$.

- **Exponential**(θ) for $\theta \in \Theta = (0,\infty)$ (i.e., $p_\theta(x) = \theta e^{-\theta x}$ for $x > 0$). Here $\mathcal{X} = (0,\infty)$, $\eta(\theta) = \theta$, and $T(x) = -x$, and so $I(\theta) = \mathbb{V}ar_\theta(T(x)) = 1/\theta^2$, yielding

$$\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \sim \frac{\sqrt{n}}{2\theta_0\sqrt{\pi}},$$

$$\mathbb{E}_{\theta_1}[q_n^{\theta_0}] \sim \frac{\pi(\theta_0)\sqrt{n}}{2\pi(\theta_2)\theta_2\sqrt{\pi}} \left(\frac{\theta_0\theta_1}{\theta_2^2}\right)^n, \tag{22}$$

where $\theta_2 = (\theta_0 + \theta_1)/2$.

- **Poisson**(θ) for $\theta \in \Theta = (0, \infty)$. Here $\mathcal{X} = \mathbb{N}$, $\eta(\theta) = \log\theta$, and $T(x) = x$, and so $\mathbb{V}ar_\theta(T(x)) = \theta$ and $\eta'(\theta) = 1/\theta$, yielding $I(\theta) = 1/\theta$ and

$$\mathbb{E}_{\theta_0}[q_n^{\theta_0}] \sim \frac{\sqrt{n}}{2\sqrt{\pi\theta_0}},$$

$$\mathbb{E}_{\theta_1}[q_n^{\theta_0}] \sim \frac{\pi(\theta_0)\sqrt{n}}{2\pi(\theta_2)\sqrt{\pi\theta_2}} \exp\left(-n(\theta_0 + \theta_1 - 2\theta_2)\right),$$

where $\theta_2 = \sqrt{\theta_0\theta_1}$.

Proof of Theorem 9. Part (ii) follows from part (i) by Proposition 8 (when $\theta_1 \neq \theta_0$ the point θ_2 lies strictly between θ_0 and θ_1, and so is an interior point of Θ). We will thus prove (i). It is convenient to assume without loss of generality that $\eta(\theta) \equiv \theta$ (this is called the “canonical” representation); indeed, since $\eta' > 0$, the transformation $\tilde{\theta} := \eta(\theta)$ (which preserves the convexity of the parameter space and maps interior points to interior points) yields: $\tilde{p}_{\tilde{\theta}}(x) = \exp(\tilde{\theta} \cdot T(x) - A(\tilde{\theta}) - B(x))$; $\tilde{\pi}(\tilde{\theta}) = \pi(\theta)/\eta'(\theta)$; $\tilde{q}_n^{\tilde{\theta}} = q_n^\theta/\eta'(\theta)$; and $\tilde{I}(\tilde{\theta}) = I(\theta)/(\eta'(\theta))^2$; hence $\tilde{q}_n^{\tilde{\theta}}/\sqrt{\tilde{I}(\tilde{\theta})} = q_n^\theta/\sqrt{I(\theta)}$, and so (20) for θ is equivalent to (20) for $\tilde{\theta}$. From now on we thus have

$$p_\theta(x) = \exp(\theta\, T(x) - A(\theta) - B(x)) \tag{23}$$

for all $x \in \mathcal{X}$ and $\theta \in \Theta$, and so $I(\theta) = A''(\theta)$. For $s_n = (x_1, \dots, x_n) \in \mathcal{X}^n$, let $\widehat{\theta}_n \equiv \widehat{\theta}_n(s_n) := \arg\max_{\theta\in\Theta} \sum_{i=1}^n \log p_\theta(x_i)$ denote the maximum likelihood estimator (MLE); thus $\widehat{\theta}_n$ minimizes the strictly convex function $h_n(\theta) := A(\theta) - \theta \cdot \bar{t}_n$, where $\bar{t}_n \equiv \bar{t}_n(s_n) := (1/n)\sum_{i=1}^n T(x_i)$; if $\widehat{\theta}_n$ is an interior point of Θ then $h_n'(\widehat{\theta}_n) = 0$, i.e., $A'(\widehat{\theta}_n) = \bar{t}_n$. Put $\mathbb{P}_0 \equiv \mathbb{P}_{\theta_0}$, $p_0 \equiv p_{\theta_0}$, and $\mathbb{E}_0 \equiv \mathbb{E}_{\theta_0}$, respectively, for the probability, density, and expectation under θ_0. Given iid observations $x_1, x_2, \dots$, under $\mathbb{P}_0$, we have (see, e.g., [20], Theorem 7.57 or [10], Theorem 18)[8] $\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, which implies $\widehat{\theta}_n \xrightarrow{\mathbb{P}_0} \theta_0$ (the “consistency” of the MLE).

For convenience we divide the proof of (20) (for $\theta_0 \in \mathrm{int}\Theta$) into a number of steps, as follows. First, we show that with high $\mathbb{P}_0$-probability the posterior $q_n^{\widehat{\theta}_n}$ at $\widehat{\theta}_n$ converges to the appropriate limit by the Bernstein–von Mises theorem (Step 1). Second, since $\widehat{\theta}_n - \theta_0$ is approximately normal (as seen above), we show that replacing $\widehat{\theta}_n$ with θ_0 requires a factor of $1/\sqrt{2}$ on expectation (Steps 2 and 3). We then prove that the θ_0-posterior $q_n^{\theta_0}$ is $\mathrm{O}(\sqrt{n})$ (Step 4), and so sets of small $\mathbb{P}_0$-probability can be ignored (Step 5); this completes the proof.

- *Step 1:* Let $J(\theta) := \sqrt{I(\theta)/(2\pi)}$; then

$$\frac{1}{\sqrt{n}} q_n^{\widehat{\theta}_n}(s_n) \xrightarrow{\mathbb{P}_0} J(\theta_0) \text{ as } n \to \infty.$$

Proof. Let $\vartheta := \sqrt{nI(\widehat{\theta}_n)}(\theta - \widehat{\theta}_n)$; the Bernstein–von Mises theorem says that under θ_0 the posterior density of ϑ converges as $n \to \infty$ to the standard normal density φ; specifically, Theorem 7.89 in [20]

[8] Notation: $\xrightarrow{\mathcal{L}}$ means convergence in law (or distribution), and $\xrightarrow{\mathbb{P}_0}$ means convergence in probability with respect to the probability $\mathbb{P}_0$.

(in Appendix A.2.3 we show that all the "general regularity conditions" of the theorem hold) applied at $\vartheta = 0$, i.e., $\theta = \widehat{\theta}_n$, yields $q_n^{\widehat{\theta}_n}(s_n)/\sqrt{nI(\widehat{\theta}_n)} \xrightarrow{\mathbb{P}_0} \varphi(0) = 1/\sqrt{2\pi}$. Now use $\widehat{\theta}_n \xrightarrow{\mathbb{P}_0} \theta_0$ and the continuity of I. □

Given $0 < \varepsilon < 1$, let $\Omega_n^1 \equiv \Omega_n^1(\varepsilon)$ be the event that $\left|(1/\sqrt{n})q_n^{\widehat{\theta}_n}(s_n)/J(\theta_0) - 1\right| \le \varepsilon$; thus $\mathbb{P}_0(\Omega_n^1) \to 1$ (as $n \to \infty$) by Step 1. Let $\Omega_n^2 \equiv \Omega_n^2(\varepsilon)$ be the event that $\widehat{\theta}_n$ is an interior point of Θ and $|\pi(\theta_0)/\pi(\widehat{\theta}_n) - 1| \le \varepsilon$; since $\widehat{\theta}_n \xrightarrow{\mathbb{P}_0} \theta_0 \in \operatorname{int}\Theta$ and π is continuous and positive, $\mathbb{P}_0(\Omega_n^2) \to 1$. Put $\Omega_n := \Omega_n^1 \cap \Omega_n^2$; then $\mathbb{P}_0(\Omega_n) \to 1$.

- *Step 2:*

$$\mathbb{E}_{\theta_0}\left[\frac{p_0(s_n)}{p_{\widehat{\theta}_n}(s_n)}\mathbf{1}_{\Omega_n}\right] \to \frac{1}{\sqrt{2}} \text{ as } n \to \infty.$$

Proof. Recall that $h_n(\theta) := A(\theta) - \theta \cdot \bar{t}_n$; then

$$Y_n := \frac{p_0(s_n)}{p_{\widehat{\theta}_n}(s_n)}\mathbf{1}_{\Omega_n} = \exp\left(-n[h(\theta_0) - h(\widehat{\theta}_n)]\right)\mathbf{1}_{\Omega_n}.$$

In $\Omega_n \subseteq \Omega_n^2$ the point $\widehat{\theta}_n$ is an interior point of Θ and so the Taylor series of h around $\widehat{\theta}_n$, where $h'(\widehat{\theta}_n) = 0$ and $h''(\widehat{\theta}_n) = A''(\widehat{\theta}_n)$, yields an intermediate point θ_n between θ_0 and $\widehat{\theta}_n$ such that

$$Y_n = \exp\left(-\frac{1}{2}nA''(\theta_n)(\theta_0 - \widehat{\theta}_n)^2\right)\mathbf{1}_{\Omega_n}.$$

Now $\widehat{\theta}_n \xrightarrow{\mathbb{P}_0} \theta_0$ implies $\theta_n \xrightarrow{\mathbb{P}_0} \theta_0$ and thus $A''(\theta_n) \xrightarrow{\mathbb{P}_0} A''(\theta_0) = I(\theta_0)$; also, $\mathbf{1}_{\Omega_n} \xrightarrow{\mathbb{P}_0} 1$ (because $\mathbb{P}_0(\Omega_n) \to 1$). Under p_0 we have $\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} Z \equiv \mathcal{N}(0,1)$, as mentioned before Step 1; altogether, $Y_n \xrightarrow{\mathcal{L}} \exp(-Z^2/2)$. The Y_n are uniformly bounded ($0 \le Y_n \le 1$, because $A'' > 0$), and so $\mathbb{E}_0[Y_n] \to \mathbb{E}[\exp(-Z^2/2)] = 1/\sqrt{2}$. □

- *Step 3:*

$$\limsup_{n\to\infty}\left|\frac{1}{\sqrt{n}}\mathbb{E}_{\theta_0}[q_n^{\theta_0}\mathbf{1}_{\Omega_n}] - \frac{J(\theta_0)}{\sqrt{2}}\right| \le \varepsilon',$$

where $\varepsilon' := (3J(\theta_0)/\sqrt{2})\varepsilon$.

Proof. For every s_n we have

$$\frac{1}{\sqrt{n}}q_n^{\theta_0}(s_n) = \frac{q_n^{\widehat{\theta}_n}(s_n)}{\sqrt{n}} \cdot \frac{\pi(\theta_0)}{\pi(\widehat{\theta}_n)} \cdot \frac{p_0(s_n)}{p_{\widehat{\theta}_n}(s_n)}.$$

In Ω_n^1 the first factor is at most $(1+\varepsilon)J(\theta_0)$, and in Ω_n^2 the second factor is at most $1+\varepsilon$; hence

$$\frac{1}{\sqrt{n}}q_n^{\theta_0}(s_n)\mathbf{1}_{\Omega_n} \le (1+\varepsilon)^2 J(\theta_0)\frac{p_0(s_n)}{p_{\widehat{\theta}_n}(s_n)}\mathbf{1}_{\Omega_n}.$$

Taking expectation under $\mathbb{P}_0$ yields by Step 2

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}} \mathbb{E}_{\theta_0}[q_n^{\theta_0} \mathbf{1}_{\Omega_n}] \leq (1+\varepsilon)^2 \frac{J(\theta_0)}{\sqrt{2}} \leq \frac{J(\theta_0)}{\sqrt{2}} + \varepsilon'.$$

Similarly,

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}} \mathbb{E}_{\theta_0}[q_n^{\theta_0} \mathbf{1}_{\Omega_n}] \geq (1-\varepsilon)^2 \frac{J(\theta_0)}{\sqrt{2}} \geq \frac{J(\theta_0)}{\sqrt{2}} - \varepsilon',$$

completing the proof. □

- *Step 4:* There is a constant $C < \infty$ such that

$$q_n^{\theta_0}(s_n) \leq C\sqrt{n}$$

for all s_n and all $n \geq 1$.

Proof. We will show that there is a constant $c > 0$ such that

$$\frac{\pi(\theta_0)}{q_n^{\theta_0}(s_n)} = \frac{p(s_n)}{p_0(s_n)} \geq \frac{c}{\sqrt{n}}$$

for all $n \geq 1$. Take $\delta > 0$ such that $[\theta_0 - \delta, \theta_0 + \delta] \subset \operatorname{int}\Theta$; then

$$\frac{p(s_n)}{p_0(s_n)} = \int_\Theta \exp(-n[h_n(\theta) - h_n(\theta_0)])\pi(\theta)\,\mathrm{d}\theta \geq \rho \int_{\theta_0-\delta}^{\theta_0+\delta} H_n(\theta)\,\mathrm{d}\theta,$$

where $H_n(\theta) := \exp(-n[h_n(\theta) - h_n(\theta_0)])$ and $\rho := \min_{\theta\in[\theta_0-\delta,\theta_0+\delta]} \pi(\theta) > 0$ (recall that the density π is continuous and strictly positive on Θ). The strictly convex function $h_n(\theta) = A(\theta) - \theta \cdot \bar{t}_n$ has a unique minimizer in $[\theta_0 - \delta, \theta_0 + \delta]$; call it ξ. Without loss of generality assume that $\xi \geq \theta_0$. When $\xi \geq \theta_0 + \delta$ the function h_n is decreasing for $\theta \leq \xi$, and thus for all $\theta \in [\theta_0, \theta_0 + \delta]$ we have $h_n(\theta) \leq h_n(\theta_0)$, and hence $H_n(\theta) \geq 1$, which gives $p(s_n)/p_0(s_n) \geq \rho\delta \geq \rho\delta/\sqrt{n}$ for all $n \geq 1$. When $\theta_0 \leq \xi < \theta_0 + \delta$ we have[9] $h_n'(\xi) = 0$ and so $h_n(\theta) - h_n(\theta_0) \leq h_n(\theta) - h_n(\xi) = h_n''(\zeta)(\theta - \xi)^2/2$ for every θ in the interval $[\theta_0 - \delta, \theta_0 + \delta]$ (by the second-order Taylor expansion), where $\zeta \equiv \zeta_\theta$ is an intermediate point between θ and ξ, and so $\zeta \in [\theta_0 - \delta, \theta_0 + \delta]$. Since $h_n'' = A''$ we get $0 < h_n''(\zeta) \leq \alpha := \max_{\theta\in[\theta_0-\delta,\theta_0+\delta]} A''(\theta)$, and so $h_n(\theta) - h_n(\theta_0) \leq \alpha(\theta - \xi)^2/2$ and

$$\rho \int_{\theta_0-\delta}^{\theta_0+\delta} H(\theta)\,\mathrm{d}\theta \geq \rho \int_{\xi-\delta}^{\xi} \exp\Big(-\frac{n\alpha}{2}(\theta - \xi)^2\Big)\,\mathrm{d}\theta = \frac{\rho\sqrt{2\pi}}{\sqrt{n\alpha}} \left(\frac{1}{2} - \Phi(-\delta\sqrt{n\alpha})\right)$$

(because $[\xi - \delta, \xi] \subset [\theta_0 - \delta, \theta_0 + \delta]$; here Φ denotes the cumulative standard normal distribution). Since $\Phi(-\delta\sqrt{n\alpha}) \to 0$ as $n \to \infty$, the final expression is $\sim c/\sqrt{n}$ for some $c > 0$, which completes the proof. □

- *Step 5:* Let Ω_n^c denote the complement of Ω_n; we have

$$\frac{1}{\sqrt{n}} \mathbb{E}_{\theta_0}[q_n^{\theta_0} \mathbf{1}_{\Omega_n^c}] \to 0 \text{ as } n \to \infty.$$

[9] In this case, where ξ is an interior point, ξ is the minimum over all Θ, and thus $\xi = \widehat{\theta}_n$. The argument here is an instance of the Laplace method.

Proof. Use the uniform boundedness of $(1/\sqrt{n})q_n^{\theta_0}$ by Step 4 and $\mathbb{P}_0(\Omega_n) \to 1$. □

Adding the results of Steps 3 and 5 and noting that $\varepsilon > 0$ is arbitrary yields (20), and thus completes the proof of Theorem 9. □

See Appendix A.2 for additional comments, extensions, and technical details.

6. Log-concavity

In this section we discuss setups in which the sequence $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is log-concave in n. A sequence of positive numbers $\psi(n)$ for $n \geq 1$ is (*strictly*) *log-concave* if $\log \psi(n)$ is a (strictly) concave function of n; this is equivalent to $\psi(n)^2 \geq \psi(n-1)\psi(n+1)$ for every $n \geq 2$, with $>$ for the strict version. The sequence $\psi(n)$ is *unimodal* if there exists $0 \leq n_0 \leq \infty$ (possibly equal to 0 or ∞) such that $\psi(n)$ is increasing for $n \leq n_0$ and decreasing for $n \geq n_0$. Log-concavity clearly implies unimodality (with n_0 that maximizes $\psi(n)$).

In Corollary 10 we saw that for $\theta_1 \neq \theta_0$ the sequence $\psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is eventually strictly decreasing. We will now show that in certain natural setups, with conjugate priors, this can be strengthened to unimodality, and, in fact, to log-concavity, with respect to the time period n. We do this in three setups. The first one consists of iid Bernoulli observations with a uniform prior; the second, of iid normal observations with a normal prior; and the third, of iid exponential observations with an exponential prior. In the normal case log-concavity is obtained only from some n_0 on, where n_0 may be 1 or arbitrarily large, depending on the parameters; Figure 3 in Section 4 is typical of the latter case. The analysis suggests that general log-concavity results may be hard to obtain, as indicated by the different proofs in the three cases, as well as by the dependence of the result on the specific prior (see the normal case, or take a Beta prior in the Bernoulli case).

In Proposition 8 of Section 5 we get a log-linear relation between $\psi_{\theta_0,\theta_1}(n)$ and $\psi_{\theta_2,\theta_2}(n)$, and so it suffices to consider only the case where $\theta_1 = \theta_0$.

Corollary 11. *Under the assumptions of Proposition* 8, *the sequence* $\psi_{\theta_0,\theta_1}(n)$ *is log-concave/convex in* n *if and only if the sequence* $\psi_{\theta_2,\theta_2}(n)$ *is log-concave/convex in* n.

6.1. Bernoulli observations with uniform prior

This section deals with sequences of iid Bernoulli observations with a parameter θ that is uniformly distributed in $(0, 1)$. We show that in this case $\psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is a log-concave function of the time period n, and so unimodal in n. We obtain this result by proving in Section 6.2 a "reversal" of the reverse Turán inequality for Legendre polynomials, which may be of independent interest.

Theorem 12. *Let* $x_1, x_2, \ldots$ *be iid* Bernoulli(θ) *observations, and let the prior distribution of* θ *be the uniform distribution on* $\Theta = (0, 1)$. *For every* θ_0 *and* θ_1 *in* Θ *the sequence* $\psi_{\theta_0,\theta_1}(n) = \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ *is log-concave for* $n \geq 1$ (*and strictly log-concave for* $n \geq 2$), *and hence unimodal.*

Proof. As in Section 4, we work with the sufficient statistic $u_n := \sum_{i=1}^n x_i$, whose distribution given θ is Binomial(n, θ). The marginal distribution of u_n when the prior is uniform on $(0, 1)$ is then the uniform distribution on the set $\{0, 1, \ldots, n\}$; i.e., $\mathbb{P}(u_n = k) = 1/(n+1)$ for $0 \le k \le n$ (this is a well-known result; see, e.g., [15], Section 6.10.1, and in particular page 285). Therefore

$$\psi_{\theta,\theta}(n) = \sum_{k=0}^{n} \frac{(\mathbb{P}_\theta(u_n = k))^2}{\mathbb{P}(u_n = k)} = (n+1) \sum_{k=0}^{n} \binom{n}{k}^2 \theta^{2k}(1-\theta)^{2(n-k)}.$$

The log-concavity of the last expression follows from Corollary 15 in the next section, with $y = \theta^2$ and $z = (1-\theta)^2$. Finally, for $\theta_1 \neq \theta_0$ apply Corollary 11. □

It is natural to try to generalize from the uniform prior to other priors, such as Beta distributions (the class of conjugate priors here). However, numerical calculations show that $\mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ need *not* be log-concave: take, for example, the Beta(7, 1) prior, $\theta_0 = 3/4$, $\theta_1 = 9/10$, and $n = 2, 3, 4$ (this is clearly robust to small changes in the parameters).

6.2. Reversing the reverse Turán inequality for Legendre polynomials

This section proves an interesting reversal of the reverse Turán inequality for Legendre polynomials for $|x| > 1$; it yields in particular the log-concavity of the previous section.

The Legendre polynomial of degree $n \ge 0$ is defined as

$$P_n(x) := \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k}^2 (x-1)^{n-k}(x+1)^k;$$

see, e.g., [24] and the references therein. The well-known Turán inequality for Legendre polynomials, first published in [23], states that

$$P_n^2(x) \ge P_{n-1}(x)P_{n+1}(x) \ \text{ for all } |x| \le 1 \tag{24}$$

holds for every $n \ge 1$, with equality if and only if[10] $|x| = 1$. Its reverse version,

$$P_n^2(x) < P_{n-1}(x)P_{n+1}(x) \ \text{ for all } |x| > 1, \tag{25}$$

holds for every $n \ge 1$; see, e.g., [24], Theorem 1.

We next show that multiplying P_n by $n+1$ reverses (25) for all[11] $|x| > 1$.

Theorem 13. *The inequality*

$$\frac{P_{n-1}(x)P_{n+1}(x)}{P_n^2(x)} \le \frac{(n+1)^2}{n(n+2)} \ \textit{for all } |x| > 1 \tag{26}$$

holds for every $n \ge 2$, with equality if and only if $n = 2$ and $|x| = \sqrt{3}$.

Proof. Putting

$$R_n := \frac{P_{n-1}P_{n+1}}{P_n^2} \ \text{ and } \ a_n := \frac{(n+1)^2}{n(n+2)}$$

[10] Put $0^0 = 1$; then $P_n(1) = 1$ and $P_n(-1) = (-1)^n$.

[11] Starting with P_1 rather than P_0 (which is what is needed for our Theorem 12).

we will prove that

$$1 < R_n(x) \le a_n$$

for all $|x| > 1$ and $n \ge 2$; the first inequality (which also holds for $n = 1$) is the reverse Turán inequality (25), for which we provide a simple proof as well. Because $P_n(-x) = (-1)^n P_n(x)$ it suffices to consider the range $x > 1$, where $P_n(x) > 0$ for all $n \ge 1$. Let b_n be the leading coefficient of P_n, i.e., the coefficient of its highest power, x^n; then

$$b_n = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k}^2 = \frac{1}{2^n} \binom{2n}{n},$$

and thus

$$R_n(\infty) := \lim_{x\to\infty} R_n(x) = \frac{b_{n-1}b_{n+1}}{b_n^2} = \frac{n(2n+1)}{(n+1)(2n-1)}. \tag{27}$$

It is straightforward to verify that

$$1 < R_n(\infty) < a_n \tag{28}$$

for every $n \ge 2$ (the first inequality is equivalent to $(2n^2+n)/(2n^2+n-1) > 1$, and the second to $n^2 - n - 1 > 0$). Next, differentiating $\log R_n(x)$ with respect to x yields

$$(\log R_n)' = \frac{R_n'}{R_n} = \frac{P_{n-1}'}{P_{n-1}} + \frac{P_{n+1}'}{P_{n+1}} - \frac{2P_n'}{P_n}.$$

Using the following well-known formula (see, e.g., [2], Equation (12.26) for a convenient reference)

$$(x^2-1)P_n'(x) = nxP_n(x) - nP_{n-1}(x)$$

for every $n \ge 1$ and every x we then obtain

$$\begin{aligned}(x^2-1)\frac{R_n'}{R_n} &= (n-1)x - (n-1)\frac{P_{n-2}}{P_{n-1}} + (n+1)x - (n+1)\frac{P_n}{P_{n+1}} - 2nx + 2n\frac{P_{n-1}}{P_n}\\ &= 2n\frac{P_{n-1}}{P_n} - (n-1)\frac{P_{n-2}}{P_{n-1}} - (n+1)\frac{P_n}{P_{n+1}} = (n+1)\frac{P_{n-1}}{P_n}\left(\frac{2n}{n+1} - \frac{n-1}{n+1}R_{n-1} - \frac{1}{R_n}\right).\end{aligned} \tag{29}$$

The proof of (26) is by induction on n, separately for each one of the two inequalities. For the first inequality, assume by induction that $R_{n-1}(x) > 1$ for every $x > 1$. If $R_n(x) \le 1$ for some $x > 1$, then, since $R_n(1) = 1$ and $R_n(\infty) > 1$ (see (27)), there is $x_* > 1$ where R_n attains its minimum, and so $R_n(x_*) \le 1$ and $R_n'(x_*) = 0$. Using (29) and then $R_{n-1}(x_*) > 1$ (by the induction hypothesis) yields

$$\frac{1}{R_n(x_*)} = \frac{2n}{n+1} - \frac{n-1}{n+1}R_{n-1}(x_*) < \frac{2n}{n+1} - \frac{n-1}{n+1} = 1,$$

in contradiction to $R_n(x_*) \le 1$. The induction starts with $n = 1$, where we have

$$P_0P_2 - P_1^2 = 1 \cdot \frac{1}{2}(3x^2-1) - (x)^2 = \frac{1}{2}(x^2-1) > 0 \text{ for } x > 1,$$

and so $R_1(x) > 1$ for every $x > 1$. For the second inequality, assume by induction that $R_{n-1}(x) < a_{n-1}$ for every $x > 1$. If $R_n(x) \ge a_n$ for some $x > 1$, then, because $R_n(\infty) < a_n$ (see (27) and (28)), there is $x^* > 1$ where R_n attains its maximum, and so $R_n(x^*) \geq a_n$ and $R'_n(x^*) = 0$. Using (29) whose left-hand side vanishes at x^*, and then $R_{n-1}(x^*) < a_{n-1}$ (by the induction hypothesis) yields

$$\begin{aligned}\frac{1}{R_n(x^*)} &= \frac{2n}{n+1} - \frac{n-1}{n+1} R_{n-1}(x^*) > \frac{2n}{n+1} - \frac{n-1}{n+1} a_{n-1} \\ &= \frac{2n}{n+1} - \frac{n-1}{n+1} \frac{n^2}{(n-1)(n+1)} = \frac{n(n+2)}{(n+1)^2} = \frac{1}{a_n},\end{aligned}$$

in contradiction to $R_n(x^*) \geq a_n$. The induction now starts with $n = 2$, where we have

$$P_1 P_3 - \frac{9}{8} P_2^2 = x \cdot \frac{1}{2}(5x^3 - 3x) - \frac{9}{8}\left(\frac{1}{2}(3x^2 - 1)\right)^2 = -\frac{(x^2-3)^2}{32} \leq 0 \text{ for } x > 1,$$

and so $R_2(x) \leq a_2$ for every $x > 1$, with equality only for $x = \sqrt{3}$; for $n = 3$ we have $R_3(\sqrt{3}) = 19/18 < 16/15 = a_3$, and so x^* cannot be $\sqrt{3}$, and thus $R_2(x^*) < a_2$ and the induction argument above gives $R_3(x) < a_3$ for every $x > 1$, and then $R_n(x) < a_n$ for every $x > 1$ and $n \geq 4$. □

Corollary 14. *For fixed* $x > 1$, *the sequence* $Q_n(x) := (n+1)P_n(x)$ *is log-concave in* n *for* $n \geq 1$, *and strictly log-concave for* $n \geq 2$.

Thus, $Q_n^2(x) \geq Q_{n-1}(x) Q_{n+1}(x)$ holds for every $n \geq 2$, with strict inequality for $n \geq 3$. The result holds for $x = 1$ as well, where $P_n(1) = 1$ for all n.

The sequence $P_n(x)$, for fixed $x \geq 1$, is log-convex in n by the reverse Turán inequality (25); Theorem 13 says that multiplying $P_n(x)$ by $n+1$ makes the sequence log-concave instead. Moreover, $(n+1)P_n(x)$ is the "right" multiple for which this reversal occurs: taking any smaller multiple, such as $nP_n(x)$, also reverses the inequality (by (26) and $\frac{(n+1)^2}{n(n+2)} \leq \frac{n^2}{(n-1)(n+1)}$), whereas any larger multiple, such as $(n+2)P_n(x)$, does not (consider $n = 2$ and $x = \sqrt{3}$).

Corollary 15. *For fixed* $y, z \geq 0$ *and not both* 0, *the sequence*

$$S_n(y,z) := (n+1) \sum_{k=0}^{n} \binom{n}{k}^2 y^k z^{n-k}$$

is log-concave in n *for* $n \geq 1$, *and strictly log-concave for* $n \geq 2$.

Proof. When $y \neq z$, say $y > z$, put $x = (y+z)/(y-z)$, and then $x \geq 1$ and

$$S_n(y,z) = (y-z)^n (n+1) P_n(x) = (y-z)^n Q_n(x),$$

and we use Corollary 14 for $x > 1$, and $P_n(1) = 1$ for $x = 1$. When $y = z > 0$ we have $S_n(y,y) = y^n(n+1)\binom{2n}{n}$, and then $S_n^2 > S_{n-1} S_{n+1}$ is obtained from (27)–(28) or by direct calculation. □

6.3. Normal observations with normal prior

We now consider normal observations whose mean is normally distributed; specifically, $x_1, x_2, \ldots$ are iid $\mathcal{N}(\theta, \sigma^2)$ observations (with $\sigma > 0$ fixed), and the prior on θ is the standard normal distribution $\mathcal{N}(0,1)$. We show that the expected posterior $\psi_{\theta_0,\theta_1}(n) \equiv \mathbb{E}_{\theta_1}[q_n^{\theta_0}]$ is either a log-concave function of

n, or a log-convex function up to some point and a log-concave function thereafter (as in Figure 3); which case it is depends on θ_0, θ_1, and σ.

Theorem 16. *Let $x_1, x_2, \ldots$ be iid $\mathcal{N}(\theta, \sigma^2)$ observations (with $\sigma > 0$ fixed), and let the prior distribution of θ be the normal standard distribution $\mathcal{N}(0, 1)$ on $\Theta = \mathbb{R}$.*

(i) *For every real θ and $n \geq 1$ we have*

$$\psi_{\theta,\theta}(n) = \frac{(n+\sigma^2)}{\sigma\sqrt{2\pi\,(2n+\sigma^2)}} \exp\left(-\frac{\theta^2\sigma^2}{4n+2\sigma^2}\right), \tag{30}$$

and thus

$$\psi_{\theta_0,\theta_1}(n) = \exp\left(\frac{\theta^2-\theta_0^2}{2}\right)\psi_{\theta,\theta}(n)\exp\left(-\frac{n(\theta_0-\theta_1)^2}{4\sigma^2}\right) \tag{31}$$

for every real θ_0 and θ_1 with $(\theta_0+\theta_1)/2 = \theta$.

(ii) *There is $n_0 \geq 0$ that depends on σ and θ such that the sequence $\psi_{\theta,\theta}(n)$ is strictly log-convex for $n < n_0$ and strictly log-concave for $n > n_0$, and thus so are the sequences $\psi_{\theta_0,\theta_1}(n)$ for every θ_0, θ_1 with $(\theta_0+\theta_1)/2 = \theta$.*

(iii) *When $\sigma^2 \leq \sqrt{2}$ or $|\theta| \geq 1/2$ the sequence $\psi_{\theta,\theta}(n)$ is strictly log-concave for $n \geq 1$.*

(iv) *For every $\theta_1 \neq \theta_0$ the sequence $\psi_{\theta_0,\theta_1}(n)$ has at most two critical points, and so it is of one of three types: always decreasing; increasing and then decreasing; decreasing, increasing, and then decreasing.*

Proof. (i) Take the sufficient statistic $u_n := \sum_{i=1}^n x_i$. The distribution of u_n given θ is $\mathcal{N}(n\theta, n\sigma^2)$, and the marginal distribution of u_n is $\mathcal{N}(0, n^2+n\sigma^2)$ (express u_n as the sum of $n\theta$ and $u_n - n\theta$). The result (30) is then obtained by a standard computation, which we relegate to Appendix A.3; as for (31), it then follows from Proposition 8.

(ii) Let $\xi(n) := \log\psi_{\theta,\theta}(n)$. Taking derivatives with respect to n (which we view as a continuous variable in (30)) yields $\xi''(n) = \gamma(n)/(2n+\sigma^2)^3$, where

$$\gamma(n) = \frac{(\sigma^4-2n^2)\left(2n+\sigma^2\right)}{\left(n+\sigma^2\right)^2} - 4\theta^2\sigma^2.$$

The function $\gamma(n)$ is strictly decreasing for $n > 0$ (because $\gamma'(n) = -2n(2n^2+6n\sigma^2+3\sigma^4)/(n+\sigma^2)^3 < 0$), and is negative for large enough n (for sure when $2n^2 \geq \sigma^4$). Thus either ξ'' is always negative, or it changes sign once from positive to negative, which means that either ξ is always concave, or it is first convex and then concave.

(iii) If $\sigma^2 \leq \sqrt{2}$ then $\sigma^4 \leq 2n^2$ for all $n \geq 1$, and so $\gamma(n) < 0$ for all $n \geq 1$. If $|\theta| \geq 1/2$ then $\gamma(0) \leq \sigma^2 - 4\theta^2\sigma^2 \leq 0$, and so $\gamma(n) < 0$ for all $n > 0$.

(iv) Let $\tilde{\xi}(n) := \log\psi_{\theta_0,\theta_1}(n)$ and $\xi(n) := \log\psi_{\theta,\theta}(n)$ for $\theta = (\theta_0+\theta_1)/2$; then $\tilde{\xi}' = \xi' + \log w$ (for the appropriate w) and $\tilde{\xi}'' = \xi''$. Since $\tilde{\xi}'' = \xi''$ changes sign at most once by (ii), $\tilde{\xi}'$ can vanish at most twice, and so $\tilde{\xi}$ has at most two critical points; the last one must be a maximum since $\tilde{\xi}(n) \to -\infty$ as $n \to \infty$. The same then holds for $\psi_{\theta_0,\theta_1} = \exp\tilde{\xi}$. □

Figure 3 provides an example with two critical points, a minimum followed by a maximum, which is thus the most that one may get in this normal setup.

General normal prior: When the prior on θ is a general $\mathcal{N}(\mu, \sigma_\Theta^2)$ distribution, and the observations x_i given θ are iid $\mathcal{N}(\theta, \sigma_X^2)$ (where $\mu, \sigma_\Theta^2, \sigma_X^2$ are fixed), the linear transformation $x \to (x-\mu)/\sigma_\Theta$ reduces it to the above case with $\sigma^2 = \sigma_X^2/\sigma_\Theta^2$. Thus, σ^2 is now the ratio of the variance of the observations to the variance of the prior. The above result (iii), for instance, says that when the variance of the observations is not too large relative to the variance of the prior (specifically, when $\sigma_X^2 \le \sqrt{2}\sigma_\Theta^2$), the sequence of expected posteriors is log-concave, and thus unimodal, for $n \ge 1$.

6.4. Exponential observations with exponential prior

In this section we consider exponential observations whose parameter is also exponentially distributed; specifically, $x_1, x_2, \ldots$ are iid $\mathrm{Exp}(\theta)$ observations, and the prior distribution of θ is $\mathrm{Exp}(1)$. Thus $p_\theta(x) = \theta \exp(-\theta x)$ and $\pi(\theta) = \exp(-\theta)$ for $x \ge 0$ and $\theta > 0$ (with $\Theta = (0, \infty)$). Working again with the sufficient statistic $u_n := \sum_{i=1}^n x_i$, whose distribution conditional on θ is the $\Gamma(n, \theta)$ distribution (the n-fold convolution of $\mathrm{Exp}(\theta) \equiv \Gamma(1, \theta)$), we have $p_\theta(u_n) = \theta^n u_n^{n-1} \exp(-\theta u_n)/(n-1)!$, and $p(u_n) = \int_0^\infty p_\theta(u_n)\pi(\theta)\,\mathrm{d}\theta = n u_n^{n-1}(u_n+1)^{-(n+1)}$, and thus[12]

$$\psi_{\theta,\theta}(n) = \pi(\theta)\int_0^\infty \frac{p_\theta(u_n)^2}{p(u_n)}\,\mathrm{d}u_n = \frac{e^{-\theta}\theta^{2n}}{(n-1)!n!}\int_0^\infty u^{n-1}(u+1)^{n+1}\exp(-2\theta u)\,\mathrm{d}u. \tag{32}$$

Our result is

Theorem 17. *Let $x_1, x_2, \ldots$ be iid* $\mathrm{Exp}(\theta)$ *observations, and let the prior distribution of θ be the* $\mathrm{Exp}(1)$ *distribution on $\Theta = (0, \infty)$.*

(i) *For every $\theta > 0$ and $n \ge 1$ we have*

$$\psi_{\theta,\theta}(n) = \frac{\theta^{n-1/2}}{2^{n+1/2} n! \sqrt{\pi}}[(n+\theta)K_{n+1/2}(\theta) + \theta K_{n-1/2}(\theta)],$$

where K_ν denotes the modified Bessel function of the second kind,[13] *and by* (22) *we have for every $\theta_0, \theta_1 > 0$ with $(\theta_0+\theta_1)/2 = \theta$*

$$\psi_{\theta_0,\theta_1}(n) = \exp(\theta - \theta_0)\psi_{\theta,\theta}(n)\left(\frac{\theta_0\theta_1}{\theta^2}\right)^n.$$

(ii) *For every $\theta_0, \theta_1 > 0$ the sequence $\psi_{\theta_0,\theta_1}(n)$ is strictly log-concave for $n \ge 1$.*

We start with two preliminary results. Put

$$k_n := K_{n+1/2}(\theta)$$

and

$$I_{n,m} := \int_0^\infty u^n (u+1)^m e^{-2\theta u}\,\mathrm{d}u.$$

[12] This can be obtained also from the known fact that in this setup the posterior distribution on Θ is a gamma distribution, specifically, $\Gamma(n+1, u_n+1)$, and so $q_n^{\theta_0} = (u_n+1)^{n+1}\theta_0^n \exp(-(u_n+1)\theta_0)/n!$.

[13] See, e.g., [1], Chapter 9.6.

Lemma 18. *For every $n \geq 0$ we have*

$$I_{n,n} = \frac{e^{\theta}}{\sqrt{\pi}} \frac{n!}{(2\theta)^{n+1/2}} k_n.$$

Proof. The change of variable $v = 2u + 1$ and Equation (9.6.23) of [1] give

$$\begin{aligned} I_{n,n} &= \int_0^{\infty} u^n (u+1)^n \exp(-2\theta u)\,\mathrm{d}u = \frac{e^{\theta}}{2^{2n+1}} \int_0^{\infty} ((2u+1)^2 - 1)^n \exp(-\theta(2u+1))\,\mathrm{d}(2u+1) \\ &= \frac{e^{\theta}}{2^{2n+1}} \int_1^{\infty} (v^2-1)^n \exp(-\theta v)\,\mathrm{d}v = \frac{e^{\theta}}{2^{2n+1}} \frac{n!}{\sqrt{\pi}} \left(\frac{2}{\theta}\right)^{n+1/2} K_{n+1/2}(\theta). \end{aligned}$$

□

Lemma 19. *For every $n \geq 1$ we have*

$$I_{n-1,n+1} = \frac{n+\theta}{n} I_{n,n} + \frac{1}{2} I_{n-1,n-1}.$$

Proof. First, the identity $u^{n-1}(u+1)^n = u^n(u+1)^{n-1} + u^{n-1}(u+1)^{n-1}$ gives

$$I_{n-1,n} = I_{n,n-1} + I_{n-1,n-1}. \tag{33}$$

Second, integration by parts yields

$$\begin{aligned} 2\theta I_{n,n} &= \int_0^{\infty} u^n (u+1)^n 2\theta \exp(-2\theta u)\,\mathrm{d}u = \left[u^n (u+1)^n \left(-\exp(-2\theta u)\right)\right]_0^{\infty} \\ &- \int_0^{\infty} n u^{n-1} (u+1)^n \exp(-2\theta u)\,\mathrm{d}u - \int_0^{\infty} n u^n (u+1)^{n-1} \exp(-2\theta u)\,\mathrm{d}u \\ &= [0-0] + n I_{n-1,n} + n I_{n,n-1} \end{aligned}$$

(we used $n > 0$ for the value of the integrand at $u = 0$). By (33) we get

$$2\theta I_{n,n} = (n+n) I_{n,n-1} + n I_{n-1,n-1},$$

and thus

$$I_{n,n-1} = \frac{\theta}{n} I_{n,n} - \frac{1}{2} I_{n-1,n-1}. \tag{34}$$

Finally, the identity $u^{n-1}(u+1)^{n+1} = u^n(u+1)^n + u^{n-1}(u+1)^{n-1} + u^n(u+1)^{n-1}$ and (34) yield

$$I_{n-1,n+1} = I_{n,n} + I_{n-1,n-1} + I_{n,n-1} = \left(1 + \frac{\theta}{n}\right) I_{n,n} + \left(1 - \frac{1}{2}\right) I_{n-1,n-1}.$$

□

Proof of Theorem 17. (i) By the previous two lemmas and (32), and setting $\psi(n) := \psi_{\theta,\theta}(n)$, we have

$$\psi(n) = \frac{e^{-\theta} \theta^{2n} I_{n-1,n+1}}{(n-1)!n!} = \frac{1}{\sqrt{\pi}} \frac{\theta^{2n}}{(n-1)!n!} \left(\frac{n+\theta}{n} \frac{n!}{(2\theta)^{n+1/2}} k_n + \frac{1}{2} \frac{(n-1)!}{(2\theta)^{n-1/2}} k_{n-1} \right),$$

which simplifies to the claimed formula.

(ii) Let $\phi_n := (n+\theta)k_n + \theta k_{n-1}$ and $\rho_n := k_{n-1}/k_n$. Using the recursion

$$k_{n+1} = \frac{2n+1}{\theta} k_n + k_{n-1} \tag{35}$$

(by $K_{\nu+1}(u) = (2\nu/u)K_\nu(u) + K_{\nu-1}(u)$, which is (9.6.26) in [1]) we have

$$\begin{aligned}
\phi_{n+1} &= (n+1+\theta)k_{n+1} + \theta k_n = (n+1+\theta)\left(\frac{2n+1}{\theta}k_n + k_{n-1}\right) + \theta k_n \\
&= \left(\frac{2n^2+3n+1}{\theta} + 2n+1+\theta\right) k_n + (n+1+\theta)k_{n-1} \\
&= \left[\left(\frac{2n^2+3n+1}{\theta} + 2n+1+\theta\right) + (n+1+\theta)\rho_n\right] k_n, \\
\phi_n &= (n+\theta)k_n + \theta k_{n-1} = [n+\theta+\theta\rho_n]\, k_n, \\
\phi_{n-1} &= (n-1-\theta)k_{n-1} + \theta k_{n-2} = (n-1-\theta)k_{n-1} + \theta\left(k_n - \frac{2n-1}{\theta}k_{n-1}\right) \\
&= \theta k_n - (n-\theta)k_{n-1} = [\theta - (n-\theta)\rho_n]\, k_n.
\end{aligned}$$

Therefore

$$R_n := \frac{\psi(n-1)\psi(n+1)}{\psi(n)^2} = Q_n(\rho_n),$$

where

$$Q_n(\rho) := \frac{n\left[\left(\frac{2n^2+3n+1}{\theta} + 2n+1+\theta\right) + (n+1+\theta)\rho\right][\theta - (n-\theta)\rho]}{(n+1)(n+\theta+\theta\rho)^2}. \tag{36}$$

The proof that $Q_n(\rho_n) < 1$ for all $n \geq 2$ and $\theta \geq 0$ is quite technical and is relegated to Appendix A.4. □

General exponential prior: If the prior on θ is a general Exp(λ) distribution for some $\lambda > 0$ not necessarily equal to 1, then the linear transformation $x \to x/\lambda$ reduces it to the case of Theorem 17, and so the sequence of expected posteriors is log-concave, and thus unimodal, for $n \geq 1$.

Appendix

The appendix is devoted to extensions of the results and to the relegated technical proofs.

A.1. General probability models

The results in Sections 1–3 are stated for discrete models. We now discuss the changes needed for general models, as in Sections 5 and 6.

The general setup (see, e.g., [20]) consists of a space of observations $\mathcal{X}$ endowed with a measure ν and a space of parameters Θ endowed with a measure μ. The measures ν and μ are σ-finite; for example, the counting measure in the discrete case, and the Lebesgue measure in the continuous case. The prior probability Π on Θ has a density function $\pi(\theta)$ with respect to μ, and for every θ in Θ

the conditional-on-θ probability $\mathbb{P}_\theta$ on $\mathcal{X}$ has a density function $p_\theta(x)$ with respect to[14] ν. Let $\mathbb{P} = \int_\Theta \mathbb{P}_\theta\, \pi(\theta)\, \mathrm{d}\mu(\theta)$ be the marginal probability on $\mathcal{X}$, with density $p(x) = \int_\Theta p_\theta(x)\pi(\theta)\, \mathrm{d}\mu(\theta)$ (for $x \in \mathcal{X}$). Conditional on a sequence of observations $s_n = (x_1, \dots, x_n)$ in $\mathcal{X}^n$ (endowed with the measure ν^n), the *posterior density* (with respect to μ) on Θ is given at a point θ in Θ by

$$q_n^\theta(s_n) = \frac{p_\theta(s_n)\pi(\theta)}{p(s_n)};$$

the denominator is positive and finite for $\mathbb{P}$-a.e. s_n; see [20], Theorem 1.31. All the formulas and results in the discrete case carry over to densities in a straightforward manner; for example, the left-hand side of (1) is now

$$\mathbb{E}[q_{n+1}^{\theta_0} | s_n] = \int_{\mathcal{X}} q_{n+1}^{\theta_0}(s_{n+1}) \frac{p(s_{n+1})}{p(s_n)}\, \mathrm{d}\nu(x_{n+1}),$$

where $s_{n+1} = (s_n, x_{n+1})$.

The *likelihood ratio order* given in Section 2 is defined for general random variables x and y as follows: $y \geq_{\mathrm{lr}} x$ if

$$\mathbb{P}(x \in A)\,\mathbb{P}(y \in B) \geq \mathbb{P}(x \in B)\,\mathbb{P}(y \in A) \tag{35}$$

for any two sets A and B in $\mathbb{R}$ such that $A \leq B$, which means that $a \leq b$ for all $a \in A$ and $b \in B$; see [22], (1.C.3). Everything that is stated about this order relation in Section 2, including its relation to the stochastic order (whose definition in terms of expectations of increasing functions f remains the same) continues to hold.

Part (i) of Proposition 1 is now stated as follows. Let $P_1 \ll P_2$ be two probability measures on a space $\mathcal{S}$ with corresponding densities p_1 and p_2 with respect to the underlying measure ν on $\mathcal{S}$. Put $r(s) := p_1(s)/p_2(s)$ for every s (again, when the ratio is $0/0$, which has probability 0 under both P_1 and P_2, define $r(s)$ arbitrarily); then

$$\mathcal{L}_{P_1}(r) \geq_{\mathrm{lr}} \mathcal{L}_{P_2}(r),$$

i.e., $P_1 \circ r^{-1} \geq_{\mathrm{lr}} P_2 \circ r^{-1}$.

Proof. Let $A, B \subset [0, \infty)$ be such that $A \leq B$; i.e., there is c such that $\sup A \leq c \leq \inf B$. We need to show that

$$P_1(r^{-1}(B))P_2(r^{-1}(A)) \geq P_1(r^{-1}(A))P_2(r^{-1}(B)). \tag{36}$$

If $c = 0$ then $A = \{0\}$, and so for every $s \in r^{-1}(A)$ we have $r(s) = 0$, and thus $p_1(s) = 0$; therefore $P_1(r^{-1}(A)) = \int_{r^{-1}(A)} p_1(s)\, \mathrm{d}\nu(s) = 0$, and (36) holds since its right-hand side is 0.

Let thus $c > 0$. For every $s \in r^{-1}(A)$ we have $r(s) \leq c$, and thus $p_1(s) \leq c p_2(s)$; integrating over $r^{-1}(A)$ gives

$$P_1(r^{-1}(A)) \leq c P_2(r^{-1}(A)).$$

Similarly, for every $s \in r^{-1}(B)$ we have $r(s) \geq c$, and thus $p_2(s) \leq (1/c)p_1(s)$, which, integrating over $r^{-1}(B)$, gives

$$P_2(r^{-1}(B)) \leq \frac{1}{c} P_1(r^{-1}(B)).$$

[14] These densities (which may be discrete or continuous) are the corresponding Radon–Nikodym derivatives.

Multiplying the two inequalities yields (36). □

Part (ii) of Proposition 1, i.e.,

$$\mathcal{L}_\theta(q^\theta) \geq_{\mathrm{lr}} \mathcal{L}(q^\theta),$$

follows now from part (i) when $\mathbb{P}_\theta \ll \mathbb{P}$ (see [7], (34.15)), which clearly holds in our setup where $\Theta \subseteq \mathbb{R}$ is an interval, the prior π is positive on Θ, and the densities p_θ are continuous in θ.

Finally, we note that while posterior probabilites are always bounded from above by 1 when the space of parameters Θ is discrete, posterior densities need *not* be bounded in general (as seen in Sections 5–6 and Appendix A.2 below).

A.2. Asymptotic analysis

Section 5 deals with one-dimensional exponential families of distributions. We expect the analysis to extend to more general setups, in particular, to multidimensional exponential families. Such a family is given by densities $p_\theta(x) = \exp(\eta(\theta) \cdot T(x) - A(\eta(\theta)) - B(x))$ (for $\theta \in \Theta$ and $x \in \mathcal{X}$), where $d \geq 1$, $\eta : \Theta \to \mathbb{R}^d$, $T : \mathcal{X} \to \mathbb{R}^d$, $A : \mathbb{R}^d \to \mathbb{R}$, and $B : \mathcal{X} \to \mathbb{R}$ (in the "canonical" representation, η is the identity and $\Theta \subseteq \mathbb{R}^d$).

A.2.1. *Generalizing Proposition 8*

Consider first the useful reduction to the $\theta_1 = \theta_0$ case. Given two densities p_0 and p_1 on the space $\mathcal{X}$ (with respect to a positive σ-finite measure ν), define their *normalized geometric average* (*NGA*) to be the density r on $\mathcal{X}$ given by

$$r(x) := \frac{p_0(x)\, p_1(x)}{\int_{\mathcal{X}} \sqrt{p_0(x)\, p_1(x)}\, \mathrm{d}\nu(x)}$$

for every $x \in \mathcal{X}$. Using (16), Proposition 8 readily generalizes to

Proposition 20. *Let $(p_\theta)_{\theta\in\Theta}$ be a family of densities and let π be a prior density on Θ with $\pi(\theta) > 0$ for all $\theta \in \Theta$. Let $\theta_0, \theta_1 \in \Theta$ and let r be the normalized geometric average of p_{θ_0} and p_{θ_1}. If $r = p_{\theta_2}$ for some $\theta_2 \in \Theta$ then*

$$\psi_{\theta_0,\theta_1}(n) = \frac{\pi(\theta_0)}{\pi(\theta_2)} \psi_{\theta_2,\theta_2}(n)\, w^n, \tag{37}$$

where w is given by (18).

For multidimensional exponential families, closure under normalized geometric averages (*NGA-closure* for short) is easily seen to amount to the convexity of the set of "natural parameters" $\eta(\Theta)$ (because $\eta(\theta_2)$ must equal $(\eta(\theta_0) + \eta(\theta_1))/2$). In the one-dimensional case with $\Theta \subseteq \mathbb{R}$ and $\eta' > 0$ the convexity of $\eta(\Theta)$ follows from the convexity of Θ.

For families of distributions that are not NGA-closed, one may consider for each p_{θ_0} and p_{θ_1} that member of the family, p_{θ_2} with $\theta_2 \in \Theta$, that is closest to their normalized geometric average r in terms of the Kullback–Leibler distance. Indeed, this p_{θ_2} yields the "leading exponential term" in (16) as $n \to \infty$ (cf. the Laplace method, which is used in the Bernstein–von Mises result as well; this explains why p_{θ_2} must belong to the family and $\pi(\theta_2) > 0$). The simplicity of the relation (37) is however lost when $p_{\theta_2} \neq r$.

Interestingly, for single-parameter families of densities $\mathcal{F} = \{p_\theta : \theta \in [a, b]\}$ (for $a < b$) where the dependence on θ is continuous and there is identifiability ($p_a \neq p_b$ suffices), $\mathcal{F}$ is NGA-closed if and only if $\mathcal{F}$ is a one-dimensional exponential family (15) on $[a, b]$. Indeed, NGA-closure implies by continuity closure with respect to all normalized geometric weighted averages; i.e., for every $p_{\theta_0}, p_{\theta_1} \in \mathcal{F}$ and every $\lambda \in [0, 1]$ there is $\tau(\lambda) \in [a, b]$ such that $c(\lambda)(p_{\theta_0})^{1-\lambda}(p_{\theta_1})^{\lambda} = p_{\tau(\lambda)}$, for the appropriate normalization constant $c(\lambda)$. Take $\theta_0 = a$ and $\theta_1 = b$; then $\tau(0) = a$ and $\tau(1) = b$; the function τ is easily seen to be continuous and one-to-one (because $p_a \neq p_b$ and so all the $p_{\tau(\lambda)}$ are distinct), and thus τ is strictly increasing and its range is the whole interval $[a, b]$. From $\log p_{\tau(\lambda)} = \log c(\lambda) + (1 - \lambda) \log p_a(x) + \lambda \log p_b(x)$ we thus get

$$\log p_\theta(x) = \tau^{-1}(\theta)(\log p_b(x) - \log p_a(x)) + \log p_a(x) + \log c(\tau^{-1}(\theta))$$

for every $\theta \in [a, b]$ and $x \in \mathcal{X}$, which yields the one-dimensional family (15) with $\eta = \tau^{-1}$, $T(x) = \log p_b(x) - \log p_a(x)$, $A(\eta) = -\log(c(\eta))$, and $B(x) = -\log p_a(x)$.

A.2.2. *Generalizing Theorem* 9

Consider the asymptotic result of Theorem 9 (i) when $\theta_1 = \theta_0$, but now for a multivariate d-dimensional exponential family as above. We expect the rate $\sqrt{n}$ to become $n^{d/2}$ (with appropriate constants). Indeed, in Step 1 the change of variable $\vartheta = (nI(\widehat{\theta}_n))^{1/2}(\theta - \widehat{\theta}_n)$ in the Bernstein–von Mises theorem (where nI is now a $d \times d$ matrix) involves a Jacobian, and so yields a factor of $n^{d/2}$; in Step 4, the bound on $q_n^{\theta_0}$ becomes $Cn^{d/2}$, as it involves a d-dimensional Gaussian integral (cf. the d-dimensional Laplace method).

We conjecture that the analysis extends beyond exponential families (for instance, in setups where the Bernstein–von Mises result applies).

A.2.3. *Proof of Theorem* 9: *Details for Step* 1

We provide here technical details for Step 1 of the proof of Theorem 9**.**

First, we show that all 7 regularity conditions of Theorem 7.89 in [20] (the Bernstein–von Mises result that we use) hold for an exponential family (23). Indeed, this is immediate for conditions 1–4 (see the discussion in Section 5). Condition 5: in our setup $\lambda_n = 1/(nA''(\hat{\theta}_n))$, and so $\lambda_n \xrightarrow{\mathbb{P}_0} 0$ follows from $A''(\hat{\theta}_n) \xrightarrow{\mathbb{P}_0} I(\theta_0) > 0$. Condition 6 is equivalent to the existence, for each $\delta > 0$, of a positive $K(\delta)$ such that

$$\lim_{n\to\infty} \mathbb{P}_0 \left(\sup_{\theta \notin [\theta_0 - \delta,\, \theta_0 + \delta]} \frac{1}{nA''(\hat{\theta}_n)} \sum_{i=1}^{n} [\log p_\theta(x_i) - \log p_{\theta_0}(x_i)] < -K(\delta) \right) = 1.$$

To see this note that $(1/n)\sum_{i=1}^{n}[\log p_\theta(x_i) - \log p_{\theta_0}(x_i)] = A(\theta_0) - A(\theta) + (\theta - \theta_0)\bar{t}_n$, and since $\bar{t}_n \xrightarrow{\mathbb{P}_0} A'(\theta_0)$, it suffices to show that $A(\theta_0) - A(\theta) + (\theta - \theta_0)A'(\theta_0) < 0$. The latter inequality follows readily from the strict convexity of A. Finally, condition 7 requires that for each $\varepsilon > 0$ there exist $\delta > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}_0 \left(\sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} \left| 1 + \lambda_n \sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log p_\theta(x_i) \right| < \varepsilon \right)$$
$$= \lim_{n\to\infty} \mathbb{P}_0 \left(\sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} \left| 1 - n\lambda_n A''(\theta) \right| < \varepsilon \right) = 1.$$

We have $n\lambda_n \xrightarrow{\mathbb{P}_0} 1/I(\theta_0) = 1/A''(\theta_0)$ implying, together with the continuity of A'', that there exists δ such that $n\lambda_n A''(\theta) = A''(\theta)/A''(\hat{\theta}_n) \in [1-\varepsilon, 1+\varepsilon]$ for $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$ holds with probability converging to 1 as $n \to \infty$.

Second, to translate the density $f_{\vartheta|s_n}$ of $\vartheta = \sqrt{nI(\hat{\theta}_n)}(\theta - \hat{\theta}_n)$ given s_n to that of θ given s_n—which is the posterior $q_n^\theta(s_n)$—we need to divide the latter by $\sqrt{nI(\hat{\theta}_n)}$.

A.3. The normal case

We provide here technical details for the proof of Theorem 16 (i) in Section 6.3.

Proof of Theorem 16 (i). Recall that the distribution $p_\theta(u_n)$ of the sufficient statistic $u_n = \sum_{i=1}^n x_i$ given θ is $\mathcal{N}(n\theta, n\sigma^2)$, and its marginal distribution $p(u_n)$ is $\mathcal{N}(0, n^2 + n\sigma^2)$. Therefore $\psi(n) \equiv \psi_{\theta,\theta}(n) = \int p_\theta^2(u_n)/p(u_n)\, \mathrm{d}u_n = a_n I_n$, where $I_n = \int_{-\infty}^{\infty} \exp(-h_n(u))\, \mathrm{d}u$ for

$$h_n(u) = 2\frac{(u-n\theta)^2}{2n\sigma^2} - \frac{u^2}{2(n^2+n\sigma^2)},$$

and a_n is the constant

$$a_n = \left(\frac{1}{\sqrt{2\pi}\sqrt{n}\,\sigma}\right)^2 \left(\frac{1}{\sqrt{2\pi}\sqrt{n^2+n\sigma^2}}\right)^{-1} = \frac{\sqrt{n^2+n\sigma^2}}{\sqrt{2\pi}\,n\sigma^2}.$$

The function $h_n(u)$ is a quadratic in u, namely, $h_n(u) = (b_n u - c_n)^2 - d_n$ for

$$b_n = \sqrt{\frac{2n+\sigma^2}{2n\sigma^2(n+\sigma^2)}} \quad \text{and} \quad d_n = \frac{\theta^2 n}{2n+\sigma^2}$$

(the value of c_n will not matter), which yields $I_n = (\sqrt{\pi}/b_n)\exp(d_n)$. Substituting in $\psi(n) = a_n I_n$ gives (30). □

A.4. The exponential case

We prove here the final inequality in the proof of Theorem 17 (ii) in Section 6.4, namely $Q_n(\rho_n) < 1$ for all $n \geq 2$, and $\theta \geq 0$; the function $Q_n(\rho)$ is defined in (36) and $\rho_n = k_{n-1}/k_n \equiv K_{n-1/2}(\theta)/K_{n+1/2}(\theta)$.

Proof. We will use the following bounds on ρ_n:

$$\eta_n := \frac{\theta}{n + \frac{1}{2} + \sqrt{\left(n - \frac{3}{2}\right)^2 + \theta^2}} < \rho_n \leq \theta; \tag{38}$$

see [21]: the lower bound is from (34), and the upper bound from (33). Taking the derivative of Q_n yields

$$Q_n'(\rho) = -\frac{2n^5 + (2\theta+3)n^4 + \left(2\theta^2 + \theta + 1 + 2\theta^2\rho - \theta\rho\right)n^3 + \left(2\theta^2 + 2\theta^2\rho - \theta\rho\right)n^2 + \left(\theta^2 + \theta^2\rho\right)n}{(n+1)(n+\theta+\theta\rho)^3\,\theta};$$

in the range $0 \le \rho \le \theta$ the coefficients of all powers of n in the numerator are positive (use $\theta\rho \le 2\theta^2$), and so Q_n is strictly decreasing in ρ there. Since $0 \le \eta_n < \rho_n \le \theta$ by (38), it follows that $R_n = Q_n(\rho_n) < Q_n(\eta_n)$, and so it suffices to show that $Q_n(\eta_n) < 1$ for all $n \ge 2$. Computing $Q_n(\eta_n)$ by substituting (38) in (36) yields

$$1 - Q_n(\eta_n) = \frac{A + B\sqrt{C}}{D},$$

where we put $n = m + 2$ (and so $m \ge 0$ when $n \ge 2$),

$$\begin{aligned} A &= 8m^4 + 72m^3 + \left(4\theta^2 - 16\theta + 230\right)m^2 + \left(8\theta^3 + 4\theta^2 - 56\theta + 302\right)m \\ &\quad + (8\theta^4 + 20\theta^3 + 2\theta^2 - 48\theta + 132), \\ B &= 4m^3 + 30m^2 + \left(-4\theta^2 + 74\right)m + (-4\theta^3 - 10\theta^2 + 60),\ C = 4m^2 + 4m + (4\theta^2 + 1), \end{aligned}$$

and the denominator D is positive. We claim that in the range $m \ge 0$ and $\theta \ge 0$ we have $A > 0$ and $E := A^2 - B^2C > 0$, and thus $A + B\sqrt{C} > 0$ (immediate when $B \ge 0$; when $B < 0$, use $A + B\sqrt{C} = E/(A - B\sqrt{C})$). Indeed, the coefficients of the powers of m in A and E are positive for all $\theta \ge 0$, as shown by direct calculations that we omit. Thus $1 - Q_n(\eta_n) > 0$ for all $n \ge 2$ and $\theta \ge 0$. □

Acknowledgments

Previous version: July 2020 (Hebrew University of Jerusalem, Center for Rationality DP-736).

We thank Marco Scarsini for many useful discussions. We are grateful to the referees, the associate editor, and the editor for many constructive comments, and in particular for encouraging us to consider exponential families and the Bernstein–von Mises Theorem.

References

[1] Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series, No.* 55. 10th Printing. Washington, D. C.: U. S. Government Printing Office.

[2] Arfken, G. (1985). *Mathematical Methods for Physicists*, 3rd ed. New York: Academic Press.

[3] Barron, A.R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics* **6**. (J.M. Bernardo, J.O. Berger, and A. Smith (Eds.)) 27–52. New York: Oxford Univ. Press. MR1723492

[4] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

[5] Berk, R.H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Stat.* **43** 193–204. MR0298810

[6] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35** 99–109. MR0010358

[7] Billingsley, P. (2013). *Convergence of Probability Measures*. New York: Wiley.

[8] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23** 493–507. MR0057518 https://doi.org/10.1214/aoms/1177729330

[9] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach. Probability and Mathematical Statistics, Vol.* 1. New York: Academic Press. MR0215390

[10] Ferguson, T.S. (1996). *A Course in Large Sample Theory. Texts in Statistical Science Series*. London: CRC Press. MR1699953 https://doi.org/10.1007/978-1-4899-4549-5

[11] Francetich, A. and Kreps, D. (2014). Bayesian inference does not lead you astray…on average. *Econom. Lett.* **125** 444–446. MR3281735 https://doi.org/10.1016/j.econlet.2014.10.022

[12] Grünwald, P.D. and Halpern, J.Y. (2014). When ignorance is bliss. Preprint. Available at arXiv:1407.7188.

[13] Hart, S. and Rinott, Y. (2020). Posterior probabilities: Dominance and optimism. *Econom. Lett.* **194** 109352, 3. MR4121083 https://doi.org/10.1016/j.econlet.2020.109352

[14] Jewitt, I. (1988). Justifying the first-order approach to principal-agent problems. *Econometrica* **56** 1177–1190. MR0964151 https://doi.org/10.2307/1911363

[15] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd ed. *Wiley Series in Probability and Statistics* **444**. Hoboken, NJ: Wiley. MR2163227 https://doi.org/10.1002/0471715816

[16] Karlin, S. (1968). *Total Positivity*. Stanford: Stanford Univ. Press.

[17] Mailath, G.J. and Samuelson, L. (2006). *Repeated Games and Reputations*. London: Oxford Univ. Press.

[18] Miller, J.W. (2018). A detailed treatment of Doob's theorem. Available at arXiv:1801.03122v1 [math.ST].

[19] Nielsen, F. (2011). Chernoff information of exponential families. Preprint. Available at arXiv:1102.2684.

[20] Schervish, M.J. (2012). *Theory of Statistics*. *Springer Series in Statistics*. New York: Springer.

[21] Segura, J. (2011). Bounds for ratios of modified Bessel functions and associated Turán-type inequalities. *J. Math. Anal. Appl.* **374** 516–528. MR2729238 https://doi.org/10.1016/j.jmaa.2010.09.030

[22] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. New York: Springer. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

[23] Szegö, G. (1948). On an inequality of P. Turán concerning Legendre polynomials. *Bull. Amer. Math. Soc.* **54** 401–405. MR0023954 https://doi.org/10.1090/S0002-9904-1948-09017-6

[24] Szwarc, R. (1998). Positivity of Turán determinants for orthogonal polynomials. In *Harmonic Analysis and Hypergroups* (*Delhi*, 1995). *Trends Math.* 165–182. Boston, MA: Birkhäuser. MR1616253

[25] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Received April 2021