Bernoulli **22**(1), 2016, 615–651
DOI: 10.3150/14-BEJ670

Asymptotic optimality of myopic information-based strategies for Bayesian adaptive estimation

JANNE V. KUJALA

Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35, FI-40014 Jyväskylä, Finland. E-mail: jvk@iki.fi

This paper presents a general asymptotic theory of sequential Bayesian estimation giving results for the strongest, almost sure convergence. We show that under certain smoothness conditions on the probability model, the greedy information gain maximization algorithm for adaptive Bayesian estimation is asymptotically optimal in the sense that the determinant of the posterior covariance in a certain neighborhood of the true parameter value is asymptotically minimal. Using this result, we also obtain an asymptotic expression for the posterior entropy based on a novel definition of almost sure convergence on "most trials" (meaning that the convergence holds on a fraction of trials that converges to one). Then, we extend the results to a recently published framework, which generalizes the usual adaptive estimation setting by allowing different trial placements to be associated with different, random costs of observation. For this setting, the author has proposed the heuristic of maximizing the expected information gain divided by the expected cost of that placement. In this paper, we show that this myopic strategy satisfies an analogous asymptotic optimality result when the convergence of the posterior distribution is considered as a function of the total cost (as opposed to the number of observations).

Keywords: active data selection; active learning; asymptotic optimality; Bayesian adaptive estimation; cost of observation; D-optimality; decision theory; differential entropy; sequential estimation

1. Introduction

The theoretical framework of this paper is that of Bayesian adaptive estimation with an information based objective function (see, e.g., MacKay [9], Kujala and Lukka [7], Kujala [6]). Following the notation of Kujala [5,6], the basic problem we consider is the estimation of an unobservable random variable $\Theta : \Omega \mapsto \boldsymbol{\Theta}$ based on a sequence $y_{x_1}, \dots, y_{x_t}$ of independent (given θ) realizations from some conditional densities $p(y_{x_t} \mid \theta)$ indexed by trial *placements* x_t, each of which can be adaptively chosen from some set X based on the outcomes $(y_{x_1}, \dots, y_{x_{t-1}})$ of the earlier observations. A commonly used greedy strategy is to choose the next placement so as to maximize the expected immediate information gain, that is, the decrease of the (differential) entropy of the posterior distribution given the next observation.

Previous work on the asymptotics of Bayesian estimation (see, e.g., Schervish [11], van der Vaart [13]) has mostly concentrated on the i.i.d. case, and in the few cases where the independent (given θ) but not identical case is considered, it is customarily assumed that a certain fixed sequence of variables is given. Hence, these results do not apply to the present situation where the sequence X_t of placements is also random.

Paninski [10] has developed an asymptotic theory for this adaptive setting. He states consistency and asymptotic normality results for the greedy information maximization placement strategy and quantifies the asymptotic efficiency of the method. However, the proofs therein are not complete and hence do not provide a sufficient foundation for some generalizations and theorems we are interested in. In this paper, we develop a more general theory which allows us to generalize the main results of Paninski [10] to almost sure convergence (with novel proofs) and to show that the greedy method is in a certain sense asymptotically optimal among *all* placement methods. Furthermore, we provide a rigorous and general framework that lends itself to further extensions of the theory.

One particular extension we are interested in is analyzing the asymptotic properties of the novel framework proposed in Kujala [5]. In this framework, the observation of Y_x is associated with some random cost C_x (see Section 4.4 for details). To make measurement "cost-effective", a myopic placement rule is considered that on each trial t maximizes the expected value of the information gain (decrease of entropy)

$$G_t = \mathrm{H}(\Theta \mid Y_{X_1}, \dots, Y_{X_{t-1}}) - \mathrm{H}(\Theta \mid Y_{X_1}, \dots, Y_{X_t})$$

divided by the expected value of the cost $C_t = C_{X_t}$. This is called a myopic strategy as it looks only one step ahead. However, it is not a greedy strategy as it does not optimize the immediate gain.

In Kujala [5], the following fairly simple asymptotic optimality result is given for this myopic strategy.

Theorem 1.1. *Suppose that there exists a constant $\alpha > 0$ such that*

$$\max_{x \in \mathsf{X}} \frac{\mathrm{E}(G_t \mid \mathbf{y}, X_t = x)}{\mathrm{E}(C_t \mid \mathbf{y}, X_t = x)} = \alpha \tag{1.1}$$

for all possible sets $\mathbf{y}$ of past observations. If the next placement X_t is defined as the maximizer of (1.1) and if for some $\sigma^2 < \infty$ and $\varepsilon > 0$,

$$\begin{cases} \mathrm{Var}(G_t \mid Y_{X_1}, \dots, Y_{X_{t-1}}) \le \sigma^2, \\ \mathrm{Var}(C_t \mid Y_{X_1}, \dots, Y_{X_{t-1}}) \le \sigma^2, \\ \mathrm{E}(C_t \mid Y_{X_1}, \dots, Y_{X_{t-1}}) \ge \varepsilon \end{cases} \tag{1.2}$$

for all t, then the gain-to-cost ratio satisfies

$$\lim_{t\to\infty} \frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t} \overset{a.s.}{=} \alpha.$$

This is asymptotically optimal in the sense that for any other strategy that satisfies (1.2), we have

$$\limsup_{t\to\infty} \frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t} \overset{a.s.}{\le} \alpha.$$

However, this result requires the obtainable information gains to not decrease over time for the optimality condition to make sense and hence does not in general apply to smooth models. In this paper, we provide a counterpart of the above result using an optimality criterion (D-optimality) relevant to smooth models.

Our results are structured as follows. In Section 2, we derive strong consistency of the posterior distributions under extremely mild, purely topological conditions on the family of likelihood functions. In Section 3, we consider the local smoothness assumptions (to be assumed in a certain neighborhood of the true parameter value) required for asymptotic normality. In Section 4.1, we develop a theory of asymptotic proportions and use it for a novel type of convergence of random variables that is required in our analysis. Then, in Sections 4.2 and 4.3, we are able to quantify the asymptotic covariance and asymptotic entropy of the posterior distribution and to show a form of asymptotic optimality for the standard greedy information maximization strategy. In Section 4.4, these results are generalized to the situation with random costs of observation associated with each placement as discussed above. The heuristically justified, myopic placement strategy proposed in Kujala [5] turns out to be asymptotically optimal also in the sense of the present paper, supporting the view that this strategy is the most natural generalization of the greedy information maximization strategy to the situation where the costs of observation can vary. We give concrete examples of the optimality results in Section 5 and then end with general discussion in Section 6.

1.1. Preliminaries

We shall denote random variables by upper case letters and their specific values by lower case letters. The information theoretic definitions that we will use are the (differential) entropy $\mathrm{H}(A) = -\int p(a)\log p(a)\,\mathrm{d}a$, which does depend on the parameterization of a, the Kullback–Leibler divergence

$$D_{\mathrm{KL}}\big(p(a)\,\|\,p(b)\big) = \int p(a)\log\frac{p(a)}{p(b)}\,\mathrm{d}a,$$

which is independent of the parameterization, and the mutual information

$$\begin{aligned}\mathrm{I}(A;B) &= \int p(a,b)\log\frac{p(a,b)}{p(a)p(b)}\,\mathrm{d}(a,b)\\ &= \int p(a)D_{\mathrm{KL}}\big(p(b\mid a)\,\|\,p(b)\big)\,\mathrm{d}a\\ &= \int p(b)D_{\mathrm{KL}}\big(p(a\mid b)\,\|\,p(a)\big)\,\mathrm{d}b,\end{aligned}$$

which is also independent of the parameterization as well as symmetric. Also, the identities $\mathrm{I}(A;B) = \mathrm{H}(A) - \mathrm{E}(\mathrm{H}(A\mid B)) = \mathrm{H}(B) - \mathrm{E}(\mathrm{H}(B\mid A))$ hold whenever the differences are well defined. This is all standard notation (see, e.g., Cover and Thomas [3]) except that in our notation, there is no implicit expectation over the values of A in $\mathrm{H}(B\mid A)$, and so it is a random variable depending on the value of A. Similarly, a conditional density $p(b \mid a)$ as an argument to $D_{\mathrm{KL}}(\cdots)$ is treated the same way as any other density of b, with no implicit expectation over a.

The densities $p(a)$ and $p(b)$ above are assumed to be taken w.r.t. arbitrary dominating measures "da" and "db". Thus, following Lindley [8], we are in fact working in full measure theoretic generality even though we use the more familiar notation. The underlying probability space is $(\Omega, \mathcal{F}, \mathsf{P})$ and so, for example, $\mathsf{P}\{\Theta \in U\}$ means the probability that the value of $\Theta : \Omega \to \boldsymbol{\Theta}$ is within the measurable set $U \subset \boldsymbol{\Theta}$. In some places we may abbreviate this by $p(U)$, but it will be clear from the context what random variable is referred to. When we say "for a.e. θ", it is w.r.t. the prior distribution of Θ. The σ-algebra of $\boldsymbol{\Theta}$ is assumed to contain at least the Borel sets of the topology which $\boldsymbol{\Theta}$ is assumed to be endowed with.

For any fixed $x \in \mathsf{X}$, we assume that the conditional densities $p(y_x \mid \theta)$ are given w.r.t. the same dominating σ-finite measure "dy_x" for all $\theta \in \boldsymbol{\Theta}$ and when we say "for a.e. y_x", it is w.r.t. this measure. For brevity, we shall indicate conditioning on the data $\mathbf{Y}_t := (Y_{X_1}, \ldots, Y_{X_t})$ by the subscript t on any quantities that depend on them. For example, $p_t(\theta) = p(\theta \mid \mathbf{Y}_t)$ is the posterior density of Θ given $\mathbf{Y}_t$ and $\mathrm{E}_t(f(\Theta)) = \mathrm{E}_t(f(\Theta) \mid \mathbf{Y}_t)$ is the posterior expectation of $f(\Theta)$ given $\mathbf{Y}_t$.

It is often assumed that one can observe multiple independent (given θ) copies of the same random variable Y_x. However, instead of complicating the general notation with something like $Y_{x_t}^{(t)}$, we rely on the fact that the set X can explicitly include separate indices for any identically distributed copies, for example, one might have $[Y_{(x,t)} \mid \theta] \overset{\text{i.i.d.}}{\sim} [Y_{(x,t')} \mid \theta]$ for all $t, t' \in \mathbb{N}$, $t \neq t'$. Hence, we can use the simple notation with no loss of generality.

The greedy information gain maximization strategy can be formally defined as choosing the placement X_t to be the value x that maximizes the mutual information $\mathrm{I}_{t-1}(\Theta; Y_x) = \mathrm{H}_{t-1}(\Theta) - \mathrm{E}_{t-1}(\mathrm{H}_{t-1}(\Theta \mid Y_x))$, the expected decrease in the entropy of Θ after the next observation. In some models, there may be no maximum of the mutual information in which case the placement should be chosen sufficiently close to the supremum, which we formally define as the ratio of the mutual information and its supremum converging to one (condition O4 in Section 4).

2. Consistency

The general assumptions for consistency are:

C1. The parameter space $\boldsymbol{\Theta}$ is a compact topological space.

C2. The family of log-likelihoods is (essentially) equicontinuous, that is, for all $\theta \in \boldsymbol{\Theta}$ and $\varepsilon > 0$, there exists a neighborhood U of θ such that whenever $\theta' \in U$,

$$\left|\log p(y_x \mid \theta) - \log p\big(y_x \mid \theta'\big)\right| < \varepsilon$$

for a.e. y_x for all $x \in \mathsf{X}$.

C3. All points in $\boldsymbol{\Theta}$ are statistically distinguishable from each other. That is, for all distinct $\theta, \theta' \in \boldsymbol{\Theta}$,

$$d_x\big(\theta, \theta'\big) := \int \left|p(y_x \mid \theta) - p_t\big(y_x \mid \theta'\big)\right| \mathrm{d}y_x > 0$$

for some $x \in \mathsf{X}$.

C4. For some $\gamma > 0$, the placements X_t satisfy

$$\mathrm{I}_{t-1}(\Theta; Y_{X_t}) \geq \gamma \sup_{x \in \mathsf{X}} \mathrm{I}_{t-1}(\Theta; Y_x)$$

for all sufficiently large t.

Remark 2.1. These assumptions for consistency are considerably weaker than those formulated in Paninski [10]. In particular, the assumptions C1–C3 only pertain to the likelihood function $p(y_x \mid \theta)$, absolutely nothing is assumed about the prior distribution of Θ. Furthermore, these assumptions are purely topological in the sense that they are preserved by all homeomorphic transformations of $\boldsymbol{\Theta}$. Also, in C4, we do not require perfect maximization of information gain; this is useful as it allows us to apply the same result to the non-greedy strategy discussed in Section 4.4 as well.

Remark 2.2. Non-compact spaces can be handled if the log-likelihood has an (essentially) equicontinuous extension to a compactification of $\boldsymbol{\Theta}$. This happens precisely when the following conditions hold:

C1′. The parameter space $\boldsymbol{\Theta}$ is a topological space.

C2′. The function $f(\theta) = ((x, y_x) \mapsto \log p(y_x \mid \theta))$, with the topology of the target space induced by the ($[0, \infty]$-valued) norm

$$\|v\| = \sup_{x \in \mathsf{X}} \operatorname*{ess\,sup}_{y_x} \bigl|v(x, y_x)\bigr|,$$

is continuous (this is just restating C2) and the closure of the range $f(\boldsymbol{\Theta})$ is compact (this is the extra condition needed for non-compact spaces).

C3′. For all distinct $\theta, \theta' \in \boldsymbol{\Theta}$, the inequality $f(\theta) \neq f(\theta')$ holds true, where equality is interpreted w.r.t. a.e. y_x. (This is equivalent to C3.)

In that case, f lifts continuously to the Stone–Čech compactification $\beta\boldsymbol{\Theta}$ of $\boldsymbol{\Theta}$ (Theorem A.1). Condition C3 may not hold for the points added by the compactification, but this can be fixed by moving to the compact quotient space $\beta\boldsymbol{\Theta}/\ker(f)$. Thus, C1–C3 can always be replaced by the strictly weaker conditions C1′–C3′.

Lemma 2.1. *Suppose that* C1–C3 *hold. Then, there exists a metric* $d : \boldsymbol{\Theta} \times \boldsymbol{\Theta} \to \mathbb{R}$ *that is consistent with the topology of* $\boldsymbol{\Theta}$, *and an estimator* $\hat{\Theta}_t$ *such that for each* t *there exists* $x \in \mathsf{X}$ *such that*

$$\mathrm{I}_t(Y_x; \Theta) \geq \mathrm{E}_t\bigl(d(\Theta, \hat{\Theta}_t)^2\bigr).$$

Proof. First, we show that the pseudometric d_x defined in C3 is continuous in $\boldsymbol{\Theta} \times \boldsymbol{\Theta}$ for all $x \in \mathsf{X}$. It can be shown using C2 that for any $\theta \in \boldsymbol{\Theta}$ and $\varepsilon > 0$, there exists a neighborhood $U_{\theta,\varepsilon}$ such that $d_x(\theta, \theta') \leq \varepsilon$ for all $\theta' \in U_{\theta,\varepsilon}$. Thus, for any $\varepsilon > 0$ and $\theta_1, \theta_2 \in \boldsymbol{\Theta}$, the triangle inequality implies

$$\bigl|d_x\bigl(\theta_1', \theta_2'\bigr) - d_x(\theta_1, \theta_2)\bigr| \leq d_x\bigl(\theta_1, \theta_1'\bigr) + d_x\bigl(\theta_2, \theta_2'\bigr) \leq 2\varepsilon$$

whenever $(\theta_1', \theta_2') \in U_{\theta_1,\varepsilon} \times U_{\theta_2,\varepsilon}$, and so d_x is continuous.

As d_x is continuous, the set

$$S_x = \{(\theta, \theta') \in \boldsymbol{\Theta} \times \boldsymbol{\Theta}\colon\ d_x(\theta, \theta') > 0\}$$

is open for every $x \in \mathsf{X}$. Now C3 implies that $\bigcup_{x\in\mathsf{X}} S_x$ covers $\boldsymbol{\Theta} \times \boldsymbol{\Theta}$, and as $\boldsymbol{\Theta} \times \boldsymbol{\Theta}$ is compact, there exists a finite subcover $\bigcup_{x\in\mathsf{X}'} S_x$. It follows that

$$d(\theta, \theta') = \left[\frac{1}{8|\mathsf{X}'|} \sum_{x\in\mathsf{X}'} \left(\int \left|p(y_x \mid \theta) - p_t(y_x \mid \theta')\right| \mathrm{d}y_x\right)^2\right]^{1/2}$$

is positive definite and hence a metric. Since X' is finite, this metric inherits the continuity of d_x.

To show that the topology induced by d coincides with that of $\boldsymbol{\Theta}$, let U be an arbitrary open neighborhood of θ_0. Then U^c is compact and so its continuous image $S := \{d(\theta_0, \theta)\colon\ \theta \in U^c\}$ is compact, too. It follows that S^c is open and as $0 \in S^c$, we obtain $[0, \delta_U) \subset S^c$ for some $\delta_U > 0$. Thus, we obtain $\{\theta \in \boldsymbol{\Theta}\colon\ d(\theta_0, \theta) < \delta_U\} \subset U$, and so the topology induced by d is finer than the default topology of $\boldsymbol{\Theta}$. As d is continuous, we obtain the converse, and so the topologies coincide.

Let then t be arbitrary. We extend $d(\theta, \theta')$ with a special point $\bar{\Theta}_t \notin \boldsymbol{\Theta}$ for which we define the distances

$$d(\theta, \bar{\Theta}_t) = \left[\frac{1}{8|\mathsf{X}'|} \sum_{x\in\mathsf{X}'} \left(\int \left|p(y_x \mid \theta) - p_t(y_x)\right| \mathrm{d}y_x\right)^2\right]^{1/2}.$$

The extended distance function may not be strictly positive definite, but it is still a pseudometric and satisfies the triangle inequality. Denoting

$$\hat{\Theta}_t = \arg\min_{\theta\in\boldsymbol{\Theta}} d(\theta, \bar{\Theta}_t),$$

we have $d(\theta, \bar{\Theta}_t) \geq d(\hat{\Theta}_t, \bar{\Theta}_t)$ for all $\theta \in \boldsymbol{\Theta}$, and the triangle inequality yields $d(\theta, \bar{\Theta}_t) \geq d(\theta, \hat{\Theta}_t) - d(\hat{\Theta}_t, \bar{\Theta}_t)$. Adding both inequalities, we obtain $2d(\theta, \bar{\Theta}_t) \geq d(\theta, \hat{\Theta}_t)$ for all $\theta \in \boldsymbol{\Theta}$. Now, the L^1-bound of Kullback–Leibler divergence [3], Lemma 11.6.1, yields

$$\begin{aligned}
\max_{x\in\mathsf{X}'} \mathrm{I}_t(Y_x; \Theta) &\geq \frac{1}{|\mathsf{X}'|} \sum_{x'\in\mathsf{X}'} \mathrm{I}_t(Y_x; \Theta) \\
&= \int \frac{1}{|\mathsf{X}'|} \sum_{x'\in\mathsf{X}'} D_{\mathrm{KL}}\big(p(y_x \mid \theta) \parallel p_t(y_x)\big) p_t(\theta)\,\mathrm{d}\theta \\
(L^1 \text{ bound}) \qquad &\geq \int \frac{1}{|\mathsf{X}'|} \sum_{x'\in\mathsf{X}'} \frac{1}{2}\left[\int \left|p(y_x \mid \theta) - p_t(y_x)\right| \mathrm{d}y_x\right]^2 p_t(\theta)\,\mathrm{d}\theta \\
&= 4\int d(\theta, \bar{\Theta}_t)^2 p_t(\theta)\,\mathrm{d}\theta \geq \int d(\theta, \hat{\Theta}_t)^2 p_t(\theta)\,\mathrm{d}\theta.
\end{aligned}$$

□

Lemma 2.2. *Suppose that* K *is a function of* Θ *and has a finite range* K. *Then, for arbitrarily chosen placements* X_t, *the inequality* $\sum_{t=1}^{\infty} \mathrm{I}_{t-1}(K; Y_{X_t}) < \infty$ *holds almost surely (which implies* $\mathrm{I}_{t-1}(K; Y_{X_t}) \xrightarrow{a.s.} 0$).

Proof. As $\mathrm{I}_{t-1}(K; Y_{X_t}) = \mathrm{H}_{t-1}(K) - \mathrm{E}_{t-1}(\mathrm{H}_t(K))$, where $0 \le \mathrm{H}_t(K) \le \log|\mathsf{K}|$ for all t, we obtain

$$\mathrm{E}\left(\sum_{k=1}^{t} \mathrm{I}_{k-1}(K; Y_{X_k})\right) = \mathrm{E}\big(\mathrm{H}_0(K) - \mathrm{E}_{t-1}\big(\mathrm{H}_t(K)\big)\big) \le \log|\mathsf{K}|$$

for all t. As $\mathrm{I}_{t-1}(K; Y_{X_t})$ is nonnegative, the sequence of partial sums is non-decreasing, and Lebesgue's monotone convergence theorem yields

$$\mathrm{E}\left(\sum_{k=1}^{\infty} \mathrm{I}_{k-1}(K; Y_{X_k})\right) = \lim_{t\to\infty} \mathrm{E}\left(\sum_{k=1}^{t} \mathrm{I}_{k-1}(K; Y_{X_k})\right) \le \log|\mathsf{K}| < \infty,$$

which implies the statement. □

Lemma 2.3. *Suppose that* C1 *and* C2 *hold. Then* $\mathrm{I}_{t-1}(\Theta; Y_{X_t}) \xrightarrow{a.s.} 0$ *for arbitrarily chosen placements* X_t.

Proof. Let $\varepsilon > 0$ be arbitrary. As $\boldsymbol{\Theta}$ is compact, a finite number of the sets $U_{\theta,\varepsilon}$ given by C2 cover it. Thus, we can partition the parameter space into a finite number of subsets $\boldsymbol{\Theta}_k$ each one contained in some $U_{\theta,\varepsilon}$. Letting the random variable K denote the index of the subset that Θ falls into, the chain rule of mutual information yields

$$\mathrm{I}_{t-1}(\Theta; Y_t) = \mathrm{I}_{t-1}(\Theta, K; Y_t) = \mathrm{I}_{t-1}(K; Y_t) + \sum_k p_{t-1}(k)\mathrm{I}_{t-1}(\Theta; Y_t \mid k), \tag{2.1}$$

where $Y_t := Y_{X_t}$ and Lemma 2.2 implies that $\mathrm{I}_{t-1}(K; Y_t) \xrightarrow{\text{a.s.}} 0$. Let us then look at the latter term. Convexity of the Kullback–Leibler divergence yields

$$\begin{aligned}
\mathrm{I}_{t-1}(\Theta; Y_t \mid k) &= \int p_{t-1}(\theta \mid k) D_{\mathrm{KL}}\big(p(y_t \mid \theta) \,\|\, p_{t-1}(y_t \mid k)\big)\, \mathrm{d}\theta \\
&\le \int p_{t-1}(\theta \mid k)\left[\int p_{t-1}\big(\theta' \mid k\big) D_{\mathrm{KL}}\big(p(y_t \mid \theta) \,\|\, p\big(y_t \mid \theta'\big)\big)\, \mathrm{d}\theta'\right] \mathrm{d}\theta \\
&= \iint p_{t-1}(\theta \mid k) p_{t-1}\big(\theta' \mid k\big)\left[\int p(y_t \mid \theta) \underbrace{\log \frac{p(y_t \mid \theta)}{p(y_t \mid \theta')}}_{\le 2\varepsilon \text{ for a.e. } y_t} \mathrm{d}y_t\right] \mathrm{d}\theta\, \mathrm{d}\theta' \\
&\le 2\varepsilon
\end{aligned}$$

for all t. Thus,

$$\limsup_{t\to\infty} \mathrm{I}_{t-1}(\Theta; Y_t) \le 2\varepsilon$$

almost surely. As $\varepsilon > 0$ was arbitrary, we obtain $\mathrm{I}_{t-1}(\Theta; Y_t) \xrightarrow{\text{a.s.}} 0$. □

Lemma 2.4. *For any measurable function* $f : \boldsymbol{\Theta} \to \mathbb{R}$, *if the prior expectation* $\mathrm{E} f(\Theta)$ *is well-defined and finite, then* $\lim_{t\to\infty} \mathrm{E}_t f(\Theta)$ *exists as a finite number almost surely.*

Proof. The finiteness of $\mathrm{E} f(\Theta)$ implies that $\mathrm{E}|f(\Theta)|$ must also be finite and so $Z_t := \mathrm{E}_t f(\Theta)$ satisfies $\mathrm{E}|Z_t| = \mathrm{E}|\mathrm{E}_t f(\Theta)| \le \mathrm{E}|f(\Theta)| < \infty$ for all t. Furthermore, since Z_{t+1} depends linearly on the posterior p_{t+1} whose expectation $\mathrm{E}_t(p_{t+1})$ equals the prior p_t, we obtain $\mathrm{E}_t(Z_{t+1}) = Z_t$ for all t and so Z_t is a martingale. As $\sup_t \mathrm{E}|Z_t| \le \mathrm{E}|f(\Theta)| < \infty$, Theorem A.2 implies that $\lim Z_t$ exists as a finite number almost surely. □

Theorem 2.1 (Strong consistency). *Suppose that* C1–C4 *hold. Then, conditioned on almost any* $\theta_0 \in \boldsymbol{\Theta}$ *as the true parameter value, the posteriors are strongly consistent, that is,* $\mathsf{P}_t\{\Theta \in U\} \xrightarrow{a.s.} 1$ *for any neighborhood* U *of* θ_0.

Proof. As the metric d given by Lemma 2.1 is bounded, Lemma 2.4 implies that $\lim_{t\to\infty} \mathrm{E}_t(d(\Theta, \theta))$ exists and is finite for all θ in a countable dense subset of $\boldsymbol{\Theta}$ almost surely, in which case continuity of d implies the same for all $\theta \in \boldsymbol{\Theta}$.

Lemmas 2.1 and 2.3 and C4 yield $\mathrm{E}_t(d(\Theta, \hat{\Theta}_t)) \xrightarrow{\text{a.s.}} 0$. As d is bounded, Lebesgue's dominated convergence theorem and Markov's inequality imply

$$\mathsf{P}\big\{d(\Theta, \hat{\Theta}_t) > \varepsilon\big\} \le \frac{\mathrm{E}(d(\Theta, \hat{\Theta}_t))}{\varepsilon} = \frac{\mathrm{E}(\mathrm{E}_t(d(\Theta, \hat{\Theta}_t)))}{\varepsilon} \to 0$$

for all $\varepsilon > 0$ and so $d(\Theta, \hat{\Theta}_t) \xrightarrow{\mathrm{P}} 0$. Convergence in probability implies that there exists a subsequence t_k such that $d(\Theta, \hat{\Theta}_{t_k}) \xrightarrow{\text{a.s.}} 0$. Thus, conditioned on almost any θ_0 as the true value, we obtain $d(\theta_0, \hat{\Theta}_{t_k}) \xrightarrow{\text{a.s.}} 0$, and the triangle inequality yields

$$\mathrm{E}_{t_k}\big(d(\Theta, \theta_0)\big) \le \mathrm{E}_{t_k}\big(d(\Theta, \hat{\Theta}_{t_k})\big) + d(\theta_0, \hat{\Theta}_{t_k}) \xrightarrow{\text{a.s.}} 0.$$

As we have already established that the full sequence $\mathrm{E}_t(d(\Theta, \theta_0))$ almost surely converges, it now follows that the limit must almost surely be zero. Thus, given any neighborhood $U \supset B_d(\theta_0, \varepsilon)$ of θ_0, Markov's inequality yields

$$\mathsf{P}_t\big\{\Theta \in U^c\big\} \le \mathsf{P}_t\big\{\Theta \in B_d(\theta_0, \varepsilon)^c\big\} \le \frac{\mathrm{E}_t(d(\Theta, \theta_0))}{\varepsilon} \xrightarrow{\text{a.s.}} 0.$$ □

Lemma 2.5. *Suppose that* C1–C3 *hold and assume that conditioned on* $\theta_0 \in \boldsymbol{\Theta}$ *as the true parameter value, the posteriors are strongly consistent. Then:*

1. *Given any metric d consistent with the topology of $\boldsymbol{\Theta}$,*

$$\Theta_t^* := \arg\min_{\theta\in\boldsymbol{\Theta}} \mathrm{E}_t\big(d(\Theta,\theta)^2\big) \xrightarrow{a.s.} \theta_0.$$

2. *For any neighborhood U of θ_0 there exists a constant $c > 0$ such that, almost surely, $\mathrm{I}_t(Y_x;\Theta) \geq c\mathsf{P}_t\{\Theta \in U^c\}$ for some $x \in \mathsf{X}$ for all sufficiently large t.*

Proof. Let D be the diameter of Θ. The triangle inequality $a \leq b+c$ implies $a^2 \leq (b+c)^2 \leq 2(b^2+c^2)$ and so consistency of the posteriors yields

$$\begin{aligned} d\big(\theta_0, \Theta_t^*\big)^2 &\leq 2\mathrm{E}_t\big(d(\Theta,\theta_0)^2 + d\big(\Theta,\Theta_t^*\big)^2\big) \leq 4\mathrm{E}_t\big(d(\Theta,\theta_0)^2\big) \\ &\leq 4\big(r^2 + D^2\mathsf{P}_t\big\{\Theta\in B_d(\theta_0,r)^c\big\}\big) \xrightarrow{\text{a.s.}} 4\big(r^2 + D^2\cdot 0\big) \end{aligned}$$

for all $r > 0$, which implies $\Theta_t^* \xrightarrow{\text{a.s.}} \theta_0$.

Let us then assume that the metric d is the one given by Lemma 2.1 and choose $\varepsilon > 0$ such that $B_d(\theta_0, 2\varepsilon) \subset U$. As $\Theta_t^* \xrightarrow{\text{a.s.}} \theta_0$, we have $B_d(\Theta_t^*, \varepsilon) \subset U$ for all sufficiently large t, and so Lemma 2.1 and Markov's inequality yield

$$\begin{aligned} \mathrm{I}_t(Y_x;\Theta) &\geq \mathrm{E}_t\big(d(\Theta,\hat{\Theta}_t)^2\big) \\ &\geq \mathrm{E}_t\big(d\big(\Theta,\Theta_t^*\big)^2\big) \geq \varepsilon^2\mathsf{P}_t\big\{\Theta\in B_d\big(\Theta_t^*,\varepsilon\big)^c\big\} \geq \varepsilon^2\mathsf{P}_t\big\{\Theta\in U^c\big\} \end{aligned}$$

for some $x \in \mathsf{X}$. □

2.1. Asymptotic entropy

The differential entropy is sensitive to the parameterization, but asymptotically, we can in most cases ignore this due to the following lemma.

Lemma 2.6. *Suppose that the prior entropy* $\mathrm{H}(\Theta)$ *is well-defined and finite. Then,*

$$\lim_{t\to\infty}\big[\mathrm{H}_t(\Theta) + D_{\mathrm{KL}}\big(p_t(\theta)\,\|\,p(\theta)\big)\big]$$

exists as a finite number almost surely.

Proof. As $\mathrm{H}_t(\Theta) + D_{\mathrm{KL}}(p_t(\theta)\,\|\,p(\theta)) = \mathrm{E}_t \log p(\Theta)$ and $\mathrm{E}\log p(\Theta) = -\mathrm{H}(\Theta)$ is well-defined and finite, the statement follows from Lemma 2.4. □

Lemma 2.7. *Suppose that* C1′ *holds and let f be defined as in* C2′. *Then, for any subset $S \subset \boldsymbol{\Theta}$,*

$$\big|\log p_{t+1}(\theta \mid S) - \log p_t(\theta\mid S)\big| \leq 2\,\mathrm{diam}\, f(S)$$

for all $\theta \in S$. If C2′ *holds, then this upper bound is finite.*

Proof. Let $\theta_1 \in S$ be fixed. If $p_t(\theta \mid S)$ is multiplied by $p(y_x \mid \theta)/p(y_x \mid \theta_1)$, it can change by at most a factor of $\exp(\operatorname{diam} f(S))$, and for the same reason, the normalization constant for this density is within a factor of $\exp(\operatorname{diam} f(S))$ from 1. The statement follows.

Suppose then that C2′ holds. As $f(\boldsymbol{\Theta})$ is compact, it follows that $f(S) \subset f(\boldsymbol{\Theta})$ must be bounded. □

Lemma 2.8. *Suppose that* C1 *and* C2 *hold. Then, for any* $\varepsilon > 0$, *the inequality* $D_{\mathrm{KL}}(p_t(\theta) \,\|\, p(\theta)) < \varepsilon t$ *holds true for all sufficiently large* t.

Proof. Let $\varepsilon > 0$ be arbitrary. As in the proof of Lemma 2.3, we partition $\boldsymbol{\Theta}$ into a finite number of subsets $\boldsymbol{\Theta}_k$ such that $|\log p(y_x \mid \theta) - \log p(y_x \mid \theta_k)| \le \varepsilon$ for all $\theta \in \boldsymbol{\Theta}_k$, y_x, and $x \in \mathsf{X}$, where θ_k is some fixed point of $\boldsymbol{\Theta}_k$. Let the random variable K denote the index of the subset that Θ falls into. Lemma 2.7 implies that

$$\left|\log p_{t+1}(\theta \mid k) - \log p_t(\theta \mid k)\right| \le 2\varepsilon$$

for all $\theta \in \boldsymbol{\Theta}_k$, which yields

$$D_{\mathrm{KL}}\big(p_t(\theta \mid k) \,\|\, p(\theta \mid k)\big) = \mathrm{E}_t\left(\log \frac{p_t(\Theta \mid k)}{p(\Theta \mid k)} \,\middle|\, k\right) \le 2\varepsilon t$$

for all t and k. The chain rule of Kullback–Leibler divergence now yields

$$\begin{aligned} D_{\mathrm{KL}}\big(p_t(\theta) \,\|\, p(\theta)\big) &= D_{\mathrm{KL}}\big(p_t(k) \,\|\, p(k)\big) + \sum_k p_t(k) D_{\mathrm{KL}}\big(p_t(\theta \mid k) \,\|\, p(\theta \mid k)\big) \\ &\le \log \max_k p(k)^{-1} + 2\varepsilon t, \end{aligned}$$

where we may assume that $p(k)$ is positive since we can drop any set $\boldsymbol{\Theta}_k$ with $p(k) = 0$ from the partition. □

Lemma 2.9. *Suppose that* $\boldsymbol{\Theta} \subset \mathbb{R}^n$ *is bounded and the family of log-likelihoods is uniformly Lipschitz*, *that is*,

$$\left|\log p(y_x \mid \theta) - \log p\big(y_x \mid \theta'\big)\right| \le M\left|\theta - \theta'\right|$$

for all $\theta, \theta' \in \boldsymbol{\Theta}$ *for all* y_x *and* $x \in \mathsf{X}$. *Then, for arbitrarily chosen placements* X_t, *the expected gain over* t *trials is bounded by* $\mathrm{I}(\Theta; \mathbf{Y}_t) \le n \log t + c$ *for some constant* $c < \infty$.

Proof. For each t, we can subdivide the bounded parameter space $\boldsymbol{\Theta}$ into $\le ct^n$ subsets $\boldsymbol{\Theta}_k$, each having diameter $\le t^{-1}$. Letting the random variable K_t denote the index of the subset that Θ falls into, the chain rule of mutual information yields

$$\mathrm{I}(\Theta; \mathbf{Y}_t) = \underbrace{\mathrm{I}(K_t; \mathbf{Y}_t)}_{\le \log(ct^n)} + \sum_{k_t} p(k_t) \underbrace{\mathrm{I}(\Theta; \mathbf{Y}_t \mid k_t)}_{\le M} \le n \log t + \log c + M \tag{2.2}$$

as in equation (2.1) in the proof Lemma 2.3. □

3. Asymptotic normality

In this section, we assume that:

N1. The parameter space Θ is a subset of $\mathbb{R}^n$.
N2. The true parameter value θ_0 is an interior point of Θ.
N3. The log-likelihood $\theta \mapsto \log p(y_x \mid \theta)$ is twice continuously differentiable with $|\nabla_\theta \log p(y_x \mid \theta)| \le M$ and $|\nabla_\theta^2 \log p(y_x \mid \theta)| \le M$ for all $x \in \mathsf{X}$ and y_x.
N4. The family of Hessians $\theta \mapsto \nabla_\theta^2 \log p(y_x \mid \theta)$ is equicontinuous at θ_0 over all $x \in \mathsf{X}$ and y_x.
N5. The prior density is absolutely continuous w.r.t. the Lebesgue measure with positive and continuous density at θ_0.

For simplicity of notation, all statements are implicitly conditioned on θ_0 being the true parameter value. Throughout this section, we will denote the posterior mean and covariance by $\hat{\Theta}_t := \mathrm{E}_t(\Theta)$ and $\Sigma_t = \mathrm{Cov}_t(\Theta)$. Note that the expected square error $\mathrm{E}_t(|\Theta - \theta|^2)$ is minimized by the mean $\theta = \mathrm{E}_t(\Theta)$. Thus, if the posteriors are strongly consistent, then Lemma 2.5 implies that $\hat{\Theta}_t \xrightarrow{\text{a.s.}} \theta_0$. Note also that the square error is related to the variance through the identity $\mathrm{E}_t(|\Theta - \hat{\Theta}_t|^2) = \mathrm{tr}(\Sigma_t)$.

Lemma 3.1. *Suppose that* N1 *and* N3 *hold and* Θ *is a bounded convex set with diameter* $\le D < \infty$. *Then, there exists a constant* $C_{M,D} < \infty$ *such that for all* t, *and* x,

$$\left|\mathrm{I}_t(Y_x; \Theta) - \left(\tfrac{1}{2}\Sigma_t\right) \odot I_x(\hat{\Theta}_t)\right| \le C_{M,D} \mathrm{E}_t\left(|\Theta - \hat{\Theta}_t|^3\right),$$

where $\odot$ *denotes the Frobenius product* $A \odot B = \sum_{i,j} A_{ij} B_{ij} = \mathrm{tr}(A^T B)$, *and* $I_x(\theta)$ *is the Fisher information matrix*

$$I_x(\theta) := \int \left[\frac{\nabla_\theta p(y_x \mid \theta)}{p(y_x \mid \theta)}\right]\left[\frac{\nabla_\theta p(y_x \mid \theta)}{p(y_x \mid \theta)}\right]^T p(y_x \mid \theta)\, \mathrm{d}y_x.$$

Proof. We can formally expand the mutual information as

$$\begin{aligned}
\mathrm{I}_t(Y_x; \Theta) &= \mathrm{H}_t(Y_x) - \mathrm{E}_t\big(\mathrm{H}(Y_x \mid \Theta)\big) \\
&= \int g\left(\int p(y_x \mid \theta) p_t(\theta)\, \mathrm{d}\theta\right) \mathrm{d}y_x - \int \left(\int g\big(p(y_x \mid \theta)\big)\, \mathrm{d}y_x\right) p_t(\theta)\, \mathrm{d}\theta \\
&= \int \left[g\left(\int p(y_x \mid \theta) p_t(\theta)\, \mathrm{d}\theta\right) - \int g\big(p(y_x \mid \theta)\big) p_t(\theta)\, \mathrm{d}\theta\right] \mathrm{d}y_x,
\end{aligned}$$

where $g(p) = -p \log p$. (Although $\mathrm{H}_t(Y_x) - \mathrm{E}_t(\mathrm{H}(Y_x \mid \Theta))$ may not be well defined here, the last line is always well-defined and equal to the mutual information.) Denoting $p_{y_x} := p(y_x \mid \hat{\Theta}_t)$, Taylor's theorem yields

$$g(p) = -p_{y_x} \log p_{y_x} - (1 + \log p_{y_x})(p - p_{y_x}) - \frac{(p - p_{y_x})^2}{2 p_{y_x}} + \frac{(p - p_{y_x})^3}{6 q_{p,y_x}^2},$$

where q_{p,y_x} is some number between p_{y_x} and p. The error term is bounded by

$$\left|\varepsilon_{y_x}(p)\right| := \left|\frac{(p-p_{y_x})^3}{6q_{p,y_x}^2}\right| \leq \frac{|p-p_{y_x}|^3}{6\min\{p,\, p_{y_x}\}^3} p_{y_x} = \frac{1}{6}\bigl(\exp\bigl(|\log p - \log p_{y_x}|\bigr) - 1\bigr)^3 p_{y_x},$$

and as $|\log p(y_x \mid \theta) - \log p(y_x \mid \hat{\Theta}_t)| \leq M|\theta - \hat{\Theta}_t| \leq MD$, we further obtain

$$\begin{aligned}
\left|\varepsilon_{y_x}\bigl(p(y_x \mid \theta)\bigr)\right| &\leq \frac{1}{6}\bigl(\exp\bigl(\left|\log p(y_x \mid \theta) - \log p(y_x \mid \hat{\Theta}_t)\right|\bigr) - 1\bigr)^3 p(y_x \mid \hat{\Theta}_t) \\
&\leq \frac{1}{6}\bigl(\exp\bigl(M|\theta - \hat{\Theta}_t|\bigr) - 1\bigr)^3 p(y_x \mid \hat{\Theta}_t) \\
&\leq \frac{1}{6}\left(\frac{\exp(MD) - 1}{MD} M|\theta - \hat{\Theta}_t|\right)^3 p(y_x \mid \hat{\Theta}_t) \\
&= C_1|\theta - \hat{\Theta}_t|^3 p(y_x \mid \hat{\Theta}_t).
\end{aligned}$$

Due to the linearity of the integral, the constant and first order terms of the expansion cancel out, leaving just

$$\begin{aligned}
\mathrm{I}_t(Y_x;\Theta) &\approx \int \frac{-[\int p(y_x \mid \theta) p_t(\theta)\,\mathrm{d}\theta - p_{y_x}]^2 + \int [p(y_x \mid \theta) - p_{y_x}]^2 p_t(\theta)\,\mathrm{d}\theta}{2p_{y_x}}\,\mathrm{d}y_x \\
&= \int \frac{1}{2}\operatorname{Var}_t\left(\frac{p(y_x \mid \Theta)}{p(y_x \mid \hat{\Theta}_t)}\right) p(y_x \mid \hat{\Theta}_t)\,\mathrm{d}y_x,
\end{aligned}$$

where the error is bounded by

$$\begin{aligned}
&\left|\int \varepsilon_{y_x}\left(\int p(y_x \mid \theta) p_t(\theta)\,\mathrm{d}\theta\right)\mathrm{d}_{y_x} - \iint \varepsilon_{y_x}\bigl(p(y_x \mid \theta)\bigr) p_t(\theta)\,\mathrm{d}\theta\,\mathrm{d}y_x\right| \\
&\quad\leq \int \left\{\left|\varepsilon_{y_x}\left(\int p(y_x \mid \theta) p_t(\theta)\,\mathrm{d}\theta\right)\right| + \int \left|\varepsilon_{y_x}\bigl(p(y_x \mid \theta)\bigr)\right| p_t(\theta)\,\mathrm{d}\theta\right\}\mathrm{d}y_x \\
&\quad\overset{\text{Jensen}}{\leq} \int \left\{\int \left|\varepsilon_{y_x}\bigl(p(y_x \mid \theta)\bigr)\right| p_t(\theta)\,\mathrm{d}\theta + \int \left|\varepsilon_{y_x}\bigl(p(y_x \mid \theta)\bigr)\right| p_t(\theta)\,\mathrm{d}\theta\right\}\mathrm{d}y_x \\
&\quad\leq \int 2\int C_1|\theta - \hat{\Theta}_t|^3 p(y_x \mid \hat{\Theta}_t) p_t(\theta)\,\mathrm{d}\theta\,\mathrm{d}y_x \leq 2C_1 \mathrm{E}_t\bigl(|\Theta - \hat{\Theta}_t|^3\bigr)
\end{aligned}$$

for all t, $\hat{\Theta}_t$, and x (Jensen's inequality applies as $|\varepsilon_{y_x}(p)|$ is convex).

Now Taylor's theorem yields

$$\frac{p(y_x \mid \theta)}{p(y_x \mid \hat{\Theta}_t)} = 1 + \frac{\nabla_\theta p(y_x \mid \hat{\Theta}_t)^T}{p(y_x \mid \hat{\Theta}_t)}(\theta - \hat{\Theta}_t) + \frac{1}{2}(\theta - \hat{\Theta}_t)^T \frac{\nabla_\theta^2 p(y_x \mid \theta')}{p(y_x \mid \hat{\Theta}_t)}(\theta - \hat{\Theta}_t)^T,$$

where θ' is a convex combination of $\hat{\Theta}_t$ and θ. The coefficients are uniformly bounded by

$$\left|\frac{\nabla_\theta p(y_x \mid \hat{\Theta}_t)}{p(y_x \mid \hat{\Theta}_t)}\right| = \left|\nabla_\theta \log p(y_x \mid \hat{\Theta}_t)\right| \le M$$

and

$$\begin{aligned}\left|\frac{\nabla_\theta^2 p(y_x \mid \theta')}{p(y_x \mid \hat{\Theta}_t)}\right| &= \underbrace{\frac{p(y_x \mid \theta')}{p(y_x \mid \hat{\Theta}_t)}}_{\le \exp(MD)} \Big|\underbrace{\nabla_\theta \log p\big(y_x \mid \theta'\big)}_{|\cdot| \le M} \underbrace{\nabla_\theta \log p\big(y_x \mid \theta'\big)^T}_{|\cdot| \le M} + \underbrace{\nabla_\theta^2 \log p\big(y_x \mid \theta'\big)}_{|\cdot| \le M}\Big| \\ &\le \exp(MD)\big(M^2 + M\big) =: C_2.\end{aligned}$$

Thus, denoting the linear term by A and the error term by B, we obtain

$$\mathrm{Var}_t\left(\frac{p(y_x \mid \Theta)}{p(y_x \mid \hat{\Theta}_t)}\right) = \mathrm{Var}_t(A) + \mathrm{Var}_t(B) + 2\,\mathrm{Cov}_t(A, B),$$

where

$$\begin{aligned}\mathrm{Var}_t(A) &= \Sigma_t \odot \left[\frac{\nabla_\theta p(y_x \mid \hat{\Theta}_t)}{p(y_x \mid \hat{\Theta}_t)}\right]\left[\frac{\nabla_\theta p(y_x \mid \hat{\Theta}_t)}{p(y_x \mid \hat{\Theta}_t)}\right]^T, \\ \mathrm{Var}_t(B) &\le \mathrm{E}_t\big(|B|^2\big) \le \big(\tfrac{1}{2}C_2\big)^2 \mathrm{E}_t\big(|\Theta - \hat{\Theta}_t|^4\big) \le \big(\tfrac{1}{2}C_2\big)^2 D\mathrm{E}_t\big(|\Theta - \hat{\Theta}_t|^3\big), \\ \big|\mathrm{Cov}_t(A, B)\big| &= \big|\mathrm{E}_t(AB) - \underbrace{\mathrm{E}_t(A)}_{=0}\mathrm{E}_t(B)\big| \le \mathrm{E}_t\big(|A||B|\big) \le M\tfrac{1}{2}C_2\mathrm{E}_t\big(|\Theta - \hat{\Theta}_t|^3\big).\end{aligned}$$

□

For the next theorems and lemmas, we define the following conditions that depend on a subset $U \subset \Theta$:

- L1. $|\nabla_\theta^2 \log|p(y_x \mid \theta) - \nabla_\theta^2 \log|p(y_x \mid \theta')| < \mu/2$ for all $\theta, \theta' \in U$, $x \in \mathsf{X}$, and y_x.
- L2. $|\log p(\theta) - \log p(\theta')| \le C$ for all $\theta, \theta' \in U$.
- L3. The maximum likelihood estimator $\Theta_t^* := \arg\max_{\theta \in U} p(\mathbf{Y}_t \mid \theta)$ is eventually well-defined and converges to θ_0 as t increases within indices satisfying $\lambda_t \ge t\mu$, where λ_t is the smallest eigenvalue of $-\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.

Lemma 3.2. *Suppose that* N4 *and* N5 *hold. Then, for any* $\mu, C > 0$, *there exists a constant* $\delta_{\mu,C} < \infty$ *such that* L1 *and* L2 *hold for any neighborhood* U *of* θ_0 *having diameter less than* $\delta_{\mu,C}$.

Lemma 3.3. *Suppose that* N1, N3, *and* L1 *hold. If* $p(\mathbf{Y}_t \mid \theta) \ge p(\mathbf{Y}_t \mid \theta_0)$ *for some* $\theta \in U$, *then*

$$|\theta - \theta_0| \le \frac{2|A_t|}{t^{1/2}\mu},$$

where $A_t = t^{-1/2}\nabla \log p(\mathbf{Y}_t \mid \theta_0)$. *Furthermore, conditioned on* θ_0 *as the true parameter value,*

$$\mathsf{P}\big\{|A_t| \geq a\big\} \leq 2n \exp\!\left(-\frac{a^2}{2nM^2}\right)$$

for all t *satisfying* $\lambda_t \geq t\mu$, *where* λ_t *is the smallest eigenvalue of* $-\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.

Proof. Taylor's theorem yields

$$\begin{aligned}\log p(\mathbf{Y_t} \mid \theta) = \log p(\mathbf{Y_t} \mid \theta_0) + \overbrace{\nabla_\theta \log p(\mathbf{Y_t} \mid \theta_0)}^{=:Z_t}{}^T(\theta-\theta_0) \\ + \underbrace{\tfrac{1}{2}(\theta-\theta_0)^T \nabla_\theta^2 \log p\big(\mathbf{Y_t} \mid \theta'\big)(\theta-\theta_0)}_{\leq -(1/2)\lambda_t|\theta-\theta_0|^2 \leq -(1/2)t\mu|\theta-\theta_0|^2},\end{aligned}$$

for some θ' between θ_0 and θ. Thus, $p(\mathbf{Y}_t \mid \theta) \geq p(\mathbf{Y}_t \mid \theta_0)$ implies $Z_t^T(\theta-\theta_0) \geq \frac{1}{2}t\mu|\theta-\theta_0|^2$, which in turn implies $|Z_t| \geq \frac{1}{2}t\mu|\theta-\theta_0|$. This is equivalent to the first statement.

Let us then prove the latter statement. Now $|Z_t|t^{-1/2} = |A|^t \geq a$ implies that $|Z_t^{(k)}| \geq t^{1/2}a/\sqrt{n}$ holds for at least one component $k \in \{1,\ldots,n\}$. But as each $Z_t^{(k)}$ is a martingale satisfying $Z_0^{(k)} = 0$ and $|Z_{k+1}^{(k)} - Z_k^{(k)}| \leq M$, Theorem A.4 yields

$$\mathsf{P}\big\{\big|Z_t^{(k)}\big| \geq t^{1/2}a/\sqrt{n}\big\} \leq 2\exp\!\left(-\frac{ta^2}{2ntM^2}\right)$$

for all $k \in \{1,\ldots,n\}$. Summing these probabilities over k so as to give an upper bound on the probability that at least one component is over the limit gives the statement. □

Lemma 3.4. *Suppose that* N1–N3 *and* L1 *hold. Then,* L3 *holds almost surely.*

Proof. For any sufficiently small $\varepsilon > 0$, N2 implies that the set $V = B(\theta_0, \varepsilon)$ is a subset of $\boldsymbol{\Theta}$. Lemma 3.3 applied to this set implies that Θ_t^* converges fast in probability to θ_0, that is, the probability $\mathsf{P}\{\Theta_t^* \notin B(\theta_0,\varepsilon)\}$ sums to a finite value over all t. This implies that $\Theta_t^* \xrightarrow{\text{a.s.}} \theta_0$. □

Theorem 3.1 (Asymptotic normality). *Suppose that* N1–N5 *hold and let* L1–L3 *hold for some* $\mu > 0$, $C > 0$, *and* $U \subset \boldsymbol{\Theta}$. *Then, the following conditions surely hold when* t *increases within indices satisfying* $\lambda_t \geq t\mu$:

1. *The posterior density of the scaled variable* $\Phi_t = t^{1/2}(\Theta - \Theta_t^*)$ *satisfies*

$$\int \big|p_t(\phi_t \mid \Theta \in U) - N\big(\phi_t; 0, B_t^{-1}\big)\big|\, \mathrm{d}\phi_t \to 0,$$

where $N(\cdots)$ *denotes a normal density with given mean and covariance and* $B_t = -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.

2. *All moments as well as the entropy of* $p_t(\phi_t \mid \Theta \in U)$ *are asymptotically equal to those of* $N(\phi_t; 0, B_t^{-1})$, *that is, the difference converges to zero.*
3. *Adjusting for the* $t^{1/2}$ *scaling factor, this implies in particular that* $t\,\mathrm{Cov}_t(\Theta \mid U) - B_t^{-1} \to 0$ *and* $t^{3/2}\mathrm{E}_t(|\Theta - \mathrm{E}_t(\Theta \mid U)|^3 \mid U) \le c_n \mu^{-3/2}$ *for sufficiently large* t *for some constant* c_n, *and so* (*assuming that* U *is bounded and convex*), *Lemma* 3.1 *yields*

$$\sup_{x \in \mathsf{X}} \left| t\mathrm{I}_t(\Theta; Y_x \mid U) - \frac{1}{2} B_t^{-1} \odot I_x(\theta_0) \right| \to 0.$$

Proof. The scaled variable Φ_t takes values in the set $V_t := \{\phi_t \in \mathbb{R}^n\colon\ \Theta_t^* + t^{-1/2}\phi_t \in U\}$. A Taylor expansion of $\log p(\mathbf{Y}_t \mid \phi_t)$ at $\phi_t = 0$ yields

$$\frac{p_t(\phi_t)}{p_t(\phi_t = 0)} = \exp\bigl(\pm\varepsilon(r)\bigr) \exp\left(-\frac{1}{2}\phi_t^T B_t \phi_t \pm \frac{1}{2}\varepsilon(r)|\phi_t|^2\right)$$

for all ϕ_t satisfying $\Theta_t^* + t^{-1/2}\phi_t \in B(\theta_0, r)$, where

$$\varepsilon(r) = \sup_{x, y_x, \theta \in B(\theta_0, r)} \max\left\{ \left| \log \frac{p(\theta)}{p(\theta')} \right|, \left| \nabla_\theta^2 \log p(y_x \mid \theta) - \nabla_\theta^2 \log p\bigl(y_x \mid \theta'\bigr) \right| \right\}.$$

Denoting $r_t = t^{1/4}$, we have $S_t := B(0, r_t) \subset V_t$ for sufficiently large t and $\varepsilon_t = \varepsilon(r_t t^{-1/2} + |\Theta_t^* - \theta_0|) \to 0$. It follows

$$p_t(\phi_t) \propto f_t(\phi_t) := \underbrace{\exp\bigl(-\tfrac{1}{2}\phi_t^T B_t \phi_t\bigr)}_{=: N_t(\phi_t)} g_t(\phi_t)$$

for all $\phi_t \in V_t$, where $g_t(\phi) = \exp(\pm\varepsilon_t \pm \frac{1}{2}\varepsilon_t^{1/2}) \to 1$ for $\phi \in S_t$. As $N_t(\phi)$ is uniformly bounded and $S_t \to \mathbb{R}^n$, it follows $[\phi \in V_t] f_t(\phi) - N_t(\phi) \to 0$ for all $\phi \in \mathbb{R}^n$. Furthermore, as $N_t(\phi_t) \le \exp(-\frac{1}{2}\mu|\phi|^2)$ and $g_t(\phi) = \exp(\pm C \pm \frac{1}{4}\mu|\phi|^2)$ for all $\phi \in V_t$, it follows

$$\int [\phi \in V_t] f_t(\phi)|\phi|^k \le \int \exp\left(C - \frac{1}{4}\mu|\phi|^2\right)|\phi|^k < \infty, \qquad \int N_t(\phi)|\phi|^k < \infty$$

for all $k \ge 0$, and so Lebesgue's dominated convergence theorem implies that

$$\int \bigl| [\phi \in V_t] f_t(\phi) u(\phi) - N_t(\phi) u(\phi) \bigr| \,\mathrm{d}\phi \to 0$$

for any function $|u(\phi)| \le |\phi|^k$. This implies that all moments of $[\phi \in V_t] f_t(\phi)$ are asymptotically equal to those of $N_t(\phi)$. As the eigenvalues of B_t are between μ and M, the normalization constant $Z := \int N_t(\phi)\,\mathrm{d}\phi$ is within the constant range $[(2\pi/M)^{n/2}, (2\pi/\mu)^{n/2}]$, and it follows that the moments of the normalized densities $p_t(\phi_t)$ and $N(\phi_t; 0, B_t^{-1})$ are also asymptotically equal. Similarly, as $f_t(\phi)\log f_t(\phi) - N_t(\phi)\log N_t(\phi) \to 0$, where the log-factors can be bounded by polynomials of $|\phi|$, it follows that the entropies of $p_t(\phi_t)$ and $N(\phi_t; 0, B_t^{-1})$ are asymptotically equal. (Note that the entropy of a density $p(x) = f(x)/Z$ can be calculated as $-(\int f \log f)/Z + \log(Z)$.) □

Lemma 3.5. *Suppose that* N1 *and* N3 *hold. Then*, *conditioned on* θ_0 *as the true parameter value*, $\mathrm{E}(-\nabla_\theta^2 \log p(Y_x \mid \theta_0)) = I_x(\theta_0)$ *for all* $x \in \mathsf{X}$, *and*

$$B_t - \frac{\sum_{k=1}^t I_{X_t}(\theta_0)}{t} \xrightarrow{a.s.} 0,$$

where $B_t = -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.

Proof.

$$\begin{aligned}
&\mathrm{E}\bigl(-\nabla_\theta^2 \log p(Y_x \mid \theta_0) \mid \Theta = \theta_0\bigr)\\
&= \int p(y_x \mid \theta_0)\left\{\left[\frac{\nabla_\theta p(y_x \mid \theta_0)}{p(y_x \mid \theta_0)}\right]\left[\frac{\nabla_\theta p(y_x \mid \theta_0)}{p(y_x \mid \theta_0)}\right]^T - \frac{\nabla_\theta^2 p(y_x \mid \theta_0)}{p(y_x \mid \theta_0)}\right\} \mathrm{d}y_x\\
&= I_x(\theta_0) - \int \nabla_\theta^2 p(y_x \mid \theta_0)\,\mathrm{d}y_x\\
&= I_x(\theta_0) - \nabla_\theta \int \nabla_\theta p(y_x \mid \theta_0)\,\mathrm{d}y_x\\
&= I_x(\theta_0) - \nabla_\theta^2 \int p(y_x \mid \theta_0)\,\mathrm{d}y_x = I_x(\theta_0),
\end{aligned}$$

where the interchange of the order of integration and differentiation is justified by Lebesgue's dominated convergence theorem for the $\mathrm{d}y_x$-integrable dominating functions $f_x(y_x)$ and $g_x(y_x)$ given by

$$\begin{aligned}
\bigl|\nabla_\theta^2 p(y_x \mid \theta)\bigr| &= p(y_x \mid \theta)\bigl|\nabla_\theta \log p(y_x \mid \theta)\nabla_\theta \log p(y_x \mid \theta)^T + \nabla_\theta^2 \log p(y_x \mid \theta)\bigr|\\
&\le p(y_x \mid \theta_0)\exp\bigl(M|\theta - \theta_0|\bigr)\cdot\bigl(M^2 + M\bigr)\\
&\le p(y_x \mid \theta_0)\exp(MD)\cdot\bigl(M^2 + M\bigr) =: f_x(y_x)
\end{aligned}$$

and

$$\begin{aligned}
\bigl|\nabla_\theta p(y_x \mid \theta)\bigr| &= p(y_x \mid \theta)\bigl|\nabla_\theta \log p(y_x \mid \theta)\bigr|\\
&\le p(y_x \mid \theta_0)\exp(MD)\cdot M =: g_x(y_x).
\end{aligned}$$

Thus, denoting $Z_k = -\nabla_\theta^2 \log p(Y_{x_k} \mid \theta_0) - I_{X_k}(\theta_0)$, given $\Theta = \theta_0$, the sequence $Z_1 + \cdots + Z_k$ of partial sums is a martingale and satisfies $\mathrm{E}(|Z_k|^2) \le (M + M)^2 < \infty$ for all k, and so Theorem A.3 implies that $(Z_1 + \cdots + Z_t)/t \xrightarrow{\text{a.s.}} 0$, which is the statement. □

Corollary 3.1. *Suppose that* N1–N5 *hold. Then*, *for all* $\mu > 0$, *almost surely* $t\Sigma_t > (B_t + \mu I)^{-1}$ (*meaning that the difference is positive definite*) *for all sufficiently large* t, *where* $B_t := -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$. *In particular*, $\mathrm{tr}(t\Sigma_t) \ge (2\mu)^{-1}$ *and* $\det(t\Sigma_t) \ge (2\mu)^{-1}(2M)^{-(n-1)}$ *for all sufficiently large* t *satisfying* $\min\lambda_{B_t} \le \mu \le M$, *where* $\min\lambda_{B_t}$ *denotes the smallest eigenvalue of* B_t.

Proof. Let $\mu > 0$ be arbitrary and define an augmented observation model $Y'_x := (Y_x, Z)$, where $Z \sim N(\Theta, \mu^{-1} I)$ is independent (given θ) from Y_x. Let U be a neighborhood of θ_0 satisfying L1 and L2 as well as L3 almost surely. If we choose the auxiliary component z_t so as to obtain $t^{-1} \sum_{k=1}^{t} z_k = \mathrm{E}(\Theta \mid \mathbf{y}_t)$ for each t, then L3 remains satisfied given the augmented data and we also obtain $\Sigma_t > \Sigma'_t$, because the augmented data will strictly decrease the square error from the original mean, and moving to the new mean can only further reduce this error. The normalized Hessian at θ_0 for the augmented data is $B'_t = B_t + \mu I$, and so, due to Lemma 3.5, $\min \lambda_{B'_t} \geq \mu/2$ for all sufficiently large t (although we have fiddled with the z_k values, Lemma 3.5 still applies as it does not depend on these values). Thus, Theorem 3.1(3) implies that $t\,\mathrm{Cov}(\Theta \mid \mathbf{y}'_t, U) - (B'_t)^{-1} \to 0$ (note that Theorem 3.1 is a sure result and hence applies even with our fiddled z_k values). Since $\mathsf{P}_t\{\Theta \in U^c\}$ decays exponentially in the augmented model, it follows that also $t\Sigma'_t - (B'_t)^{-1} \to 0$. As the eigenvalues of B'_t are within the range $[\mu/2, M + \mu/2]$, the matrix inverse behaves nicely and we obtain $(t\Sigma'_t)^{-1} - B'_t \to 0$, which implies $(t\Sigma'_t)^{-1} - B'_t < \varepsilon I$ for all sufficiently large t for any $\varepsilon > 0$. It follows $t\Sigma_t > t\Sigma'_t > (B_t + (\mu + \varepsilon) I)^{-1}$ for all sufficiently large t. □

4. Asymptotic optimality

In this section, we assume that:

O1. C1–C4 hold globally.
O2. Some neighborhood U_0 of $\theta_0 \in \boldsymbol{\Theta}$ is homeomorphic to a subset of $\mathbb{R}^n$ that satisfies N1–N5.
O3. There exists placements $x_1, \ldots, x_m \in \mathsf{X}$ and nonnegative weights $\alpha_1 + \cdots + \alpha_m = 1$ such that $\sum_{j=1}^{m} \alpha_j I_{x_j}(\theta_0)$ is positive definite.
O4. The placements X_t satisfy

$$R_t := \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}})}{\sup_{x \in \mathsf{X}} \mathrm{I}_t(\Theta; Y_x)} \rightsquigarrow 1.$$

(See Section 4.1 below for the definition of "$\rightsquigarrow$".)

First, let us say a few words about the main difficulty related to the adaptivity of the placements, namely the complications caused by any secondary modes in the posterior distribution. This issue is discussed by Paninski [10] in the context of consistency, but it seems that even after consistency has been established, the issue cannot be ignored.

The information maximization strategy decreases the relative weights of any secondary modes only at a rate approximately proportional to $1/t$ [10]. Therefore, any secondary mode may have a contribution proportional to $1/t$ to all moments of the posterior distribution. This means that only the first order moments of the approximating normal distribution remain asymptotically accurate, even though its total variation distance from the posterior does tend to zero. In particular, the inverse Hessian of the likelihood generally *does not* give an asymptotically accurate approximation of the global posterior covariance. (In fact, the global posterior covariance may be undefined as $\boldsymbol{\Theta}$ need not have a global Euclidean structure.)

For this reason, the asymptotic approximation to the expected information gain $\mathrm{I}_t(\Theta; Y_x \mid U)$ given by Theorem 3.1(3) only applies within a sufficiently small neighborhood U of the true parameter value, where the posterior can be shown to be asymptotically unimodal. Nonetheless, even though the local and global moments are not in good agreement asymptotically, it turns out that $\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)$ is in fact in good agreement with $\mathrm{I}_t(\Theta; Y_{X_{t+1}})$ on “most trials”. Indeed, as the relative weights of any secondary modes typically decay at an exponential rate with the number of trials whose placements can distinguish between them, it follows that the placements of only a decreasing fraction of trials can be significantly affected by the secondary modes.

To formalize this intuition, we will first develop a theory for measuring asymptotic proportions.

4.1. Asymptotic proportions

Definition 4.1. *To measure subsets $K \subset \mathbb{N}$, we use the* proportion measures

$$\rho(K) = \lim_{n\to\infty} \rho_{1,n}(K), \qquad \rho_{a,b}(K) = \frac{|K \cap [a, b[|}{b-a},$$

where $|\cdot|$ indicates the cardinality of a set. (Note that although $\rho_{a,b}$ is a measure in the measure-theoretic sense for any $a, b \in \mathbb{N}$, the limit ρ is only a finitely additive *measure.) When we say “for almost every $n \in \mathbb{N}$”, we mean that the set where the statement does not hold is a null set w.r.t. ρ. We use the notation $x_k \rightsquigarrow x$ to mean that there exists a subset $K \subset \mathbb{N}$ with $\rho(K) = 1$ such that $[k \in K](x_k - x) \to 0$. We also define*

$$\limsup_{k\rightsquigarrow\infty} x_k := \inf\{x \in \mathbb{R}\colon\ x_k \le x \text{ for a.e. } k \in \mathbb{N}\},$$
$$\liminf_{k\rightsquigarrow\infty} x_k := \sup\{x \in \mathbb{R}\colon\ x_k \ge x \text{ for a.e. } k \in \mathbb{N}\},$$

and when both equal x, we write $\lim_{k\rightsquigarrow\infty} x_k = x$.

Lemma 4.1. *Suppose that for all $j \in \mathbb{N}$, the proposition P_k^j holds for a.e. $k \in \mathbb{N}$. Then there exists an increasing sequence $j(k) \to \infty$ such that $P_k^1 \wedge \cdots \wedge P_k^{j(k)}$ holds for a.e. $k \in \mathbb{N}$.*

Proof. For all $j \in \mathbb{N}$, $Q_k^j := P_k^1 \wedge \cdots \wedge P_k^j$ holds for a.e. $k \in \mathbb{N}$. Thus, for all $j \in \mathbb{N}$,

$$f_j(k) := \inf_{k'\ge k} \frac{\sum_{i=1}^{k'} Q_i^j}{k'}$$

is increasing in k and tends to one as $k \to \infty$. Choosing

$$j(k) = \max\bigl\{j' \in \mathbb{N}\colon\ f_{j'}(k) \ge 1 - 1/j'\bigr\}$$

yields the statement. □

Lemma 4.2. *If x_k is a bounded sequence, then the following are equivalent:*

1. $x_k \rightsquigarrow x$,
2. $|x_k - x| < \varepsilon$ *for a.e.* $k \in \mathbb{N}$ *for all* $\varepsilon > 0$,
3. $\lim_{k \rightsquigarrow \infty} x_k = x$,
4. $\frac{1}{t} \sum_{k=1}^{t} |x_k - x| \to 0$.

If x_k is not bounded, then 1–3 *are equivalent and implied by* 4.

Proof. All implications are fairly obvious. As an example, "2 ⇒ 1" follows from Lemma 4.1 applied to $P_k^j = [|x_k - x| < 1/j]$. □

Lemma 4.3. *Let x_k be a nonnegative sequence. If $\sum_{k=1}^{\infty} x_k < \infty$, then for any $\varepsilon > 0$, the inequality $x_k < \varepsilon/k$ holds true for almost every $k \in \mathbb{N}$ (which implies $k \cdot x_k \rightsquigarrow 0$).*

Proof. Assume the contrary: for some $\varepsilon > 0$ there exists a set $K \subset \mathbb{N}$ such that $x_k \geq \varepsilon/k$ for all $k \in K$ and for some $c > 0$, $\rho_{1,k}(K) > c$ for arbitrarily large k. As $\rho_{1,n+1}(K) - \rho_{k,n+k}(K) \leq 2k/n \to 0$ as $n \to \infty$ for all k, we can recursively find an increasing sequence of indices $k_1 = 1$, $k_{i+1} \geq 2k_i$, such that $\rho_{k_i,k_{i+1}}(K) \geq c$ for all i. This yields

$$\sum_{k=1}^{\infty} x_k \geq \sum_{i=1}^{\infty} c(k_{i+1} - k_i)\frac{\varepsilon}{k_i} \geq \sum_{i=1}^{\infty} c(2k_i - k_i)\frac{\varepsilon}{k_i} = \infty,$$

which contradicts the assumption. □

Lemma 4.4. *Suppose that a sequence of random variables $X_k : \Omega \to [-M, M]$ satisfies $X_k \rightsquigarrow X$ almost surely. Then,* $\mathrm{E}(|X_k - X|) \rightsquigarrow 0$.

Proof. By Lemma 4.2(4) and the dominated convergence theorem,

$$\frac{1}{t}\sum_{k=1}^{t} \mathrm{E}\big(|X_k - X|\big) = \mathrm{E}\left(\frac{1}{t}\sum_{k=1}^{t} |X_k - X|\right) \to \mathrm{E}\left(\lim_{t \to \infty} \frac{1}{t}\sum_{k=1}^{t} |X_k - X|\right) = 0.$$ □

Corollary 4.1. *Suppose that the event A_k happens for a.e. $k \in \mathbb{N}$ a.s. Then,* $\mathsf{P}\{A_k\} \rightsquigarrow 1$.

Definition 4.2. *We use the notation $X_k \overset{\mathrm{P}}{\rightsquigarrow} X$ to mean that there exists a subset $K \subset \mathbb{N}$ with $\rho(K) = 1$ such that $[k \in K](X_k - X) \overset{\mathrm{P}}{\to} 0$.*

Lemma 4.5. *$X_k \overset{\mathrm{P}}{\rightsquigarrow} X$ if and only if $\mathsf{P}\{|X_k - X| \geq \varepsilon\} \rightsquigarrow 0$ for all $\varepsilon > 0$.*

Proof. The "only if" direction is obvious. We will prove the "if" direction.

By definition, we have $\mathsf{P}\{|X_k - X| \geq 1/j\} \leq 1/j$ for a.e. $k \in \mathbb{N}$ for all $j \in \mathbb{N}$. Lemma 4.1 then implies that there exists an increasing sequence $j(k) \to \infty$ such that

$$\mathsf{P}\big\{|X_k - X| \geq 1/j(k)\big\} \leq 1/j(k) \to 0$$

for a.e. $k \in \mathbb{N}$. □

Lemma 4.6. *Suppose that a sequence of random variables X_k satisfies $X_k \rightsquigarrow X$ almost surely. Then, $X_k \overset{\mathrm{P}}{\rightsquigarrow} X$.*

Proof. Let $\varepsilon > 0$ be arbitrary. Denoting

$$Y_t = \frac{1}{t}\sum_{k=1}^{t}\big[|X_k - X| \geq \varepsilon\big],$$

$X_k \rightsquigarrow X$ implies that $Y_t \to 0$. As Y_t is bounded, the dominated convergence theorem implies

$$0 = \mathrm{E}\Big(\lim_{t\to\infty} Y_t\Big) = \lim_{t\to\infty}\mathrm{E}(Y_t) = \lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}\mathsf{P}\big\{|X_k - X| \geq \varepsilon\big\}$$

and so Lemma 4.2(4) yields $\mathsf{P}\{|X_k - X| \geq \varepsilon\} \rightsquigarrow 0$. Now Lemma 4.5 implies the statement. □

4.2. Asymptotic D-optimality

In this section, we show that the greedy information maximization strategy satisfies asymptotically a condition known as D-optimality. This condition is defined as maximality of the determinant of the Fisher information matrix of the experiment at the true parameter value θ_0. The D-optimality criterion is special among all functionals of the information matrix (such as the trace, minimum eigenvalue, etc.) in that it is insensitive to linear or affine transformations of the parameter space $\boldsymbol{\Theta}$. Furthermore, in the asymptotically normal models that we are interested in, it yields a (local) approximation of the posterior entropy, which is the utility function commonly used in adaptive estimation settings. We will make use of this fact in the next section to derive an asymptotic expression of the posterior entropy.

Lemma 4.7. *For almost any $\theta_0 \in \boldsymbol{\Theta}$ satisfying* O1–O3, *there exists a constant c such that for all $\mu > 0$, given θ_0 as the true parameter value, almost surely $\mathrm{I}_t(\Theta; Y_{X_{t+1}}) \geq c(t\mu)^{-1}$ for all sufficiently large t satisfying $\lambda_t \leq t\mu$, where λ_t denotes the smallest eigenvalue of $-\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.*

Proof. Denoting $I := \sum_{j=1}^{m}\alpha_j I_j$, where α_j and $I_j := I_{x_j}(\theta_0)$ are given by O3, the smallest eigenvalue $\min \lambda_I$ is positive.

Suppose that U_0 has diameter D and let $C_{M,D}$ be the constant of Lemma 3.1 applied to U_0 as the parameter space. The same constant also applies to any subset $U = B(\theta_0, \delta/2) \subset U_0$ with diameter $\delta \leq D$ and as the posteriors are strongly consistent in U, too, Lemma 2.5 implies that $\mathrm{E}_t(\Theta \mid U) \xrightarrow{\text{a.s.}} \theta_0$. Thus, N3 and N4 imply that $|I_x(\mathrm{E}_t(\Theta \mid U)) - I_x(\theta_0)| < \delta$ for all x for all sufficiently large t. We obtain

$$\begin{aligned}
\mathrm{I}_t(Y_x;\Theta\mid U) &\geq \frac{1}{2}\operatorname{Cov}_t(\Theta\mid U)\odot I_x\big(\mathrm{E}_t(\Theta\mid U)\big) - C_{M,D}\mathrm{E}_t\big(\big|\Theta-\mathrm{E}_t(\Theta\mid U)\big|^3\mid U\big)\\
&\geq \frac{1}{2}\operatorname{Cov}_t(\Theta\mid U)\odot I_x\big(\mathrm{E}_t(\Theta\mid U)\big) - C_{M,D}\mathrm{E}_t\big(\delta\big|\Theta-\mathrm{E}_t(\Theta\mid U)\big|^2\mid U\big)\\
&= \frac{1}{2}\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)I_x\big(\mathrm{E}_t(\Theta\mid U)\big)\big) - C_{M,D}\delta\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\\
&\geq \frac{1}{2}\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)I_x(\theta_0)\big) - \bigg(C_{M,D}+\frac{1}{2}\bigg)\delta\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\\
&\geq \frac{1}{2}\max_{j=1,\dots,m}\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)I_j\big) - \bigg(C_{M,D}+\frac{1}{2}\bigg)\delta\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\\
&\geq \frac{1}{2}\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)I\big) - \bigg(C_{M,D}+\frac{1}{2}\bigg)\delta\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\\
&\geq \frac{1}{2}\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\min\lambda_I - \bigg(C_{M,D}+\frac{1}{2}\bigg)\delta\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\\
&= \bigg(\frac{\min\lambda_I}{2} - \bigg(C_{M,D}+\frac{1}{2}\bigg)\delta\bigg)\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big) =: c\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)
\end{aligned}$$

for some $x\in\mathsf{X}$ (fourth inequality) for all sufficiently large t (third inequality), where we have used the fact that $\operatorname{tr}(A)\min\lambda_B \leq \operatorname{tr}(AB)\leq\operatorname{tr}(A)\max\lambda_B$ (sixth and third inequalities). Let us then choose $\delta<\min\lambda_I/(2C_{M,D}+1)$ so that c as defined above is positive. Now, the inequality $\mathrm{I}_t(\Theta;Y_x)\geq p_t(U)\mathrm{I}_t(\Theta;Y_x\mid U)$, which follows from the chain rule of mutual information (cf. the proof of the next lemma), and C4 + Corollary 3.1 imply

$$\begin{aligned}
\mathrm{I}_t(\Theta;Y_{t+1}) &\geq \gamma\sup_{x\in\mathsf{X}}\mathrm{I}_t(\Theta;Y_x)\geq\gamma\sup_{x\in\mathsf{X}}p_t(U)\mathrm{I}_t(\Theta;Y_x\mid U)\\
&\geq \gamma p_t(U)c\operatorname{tr}\big(\operatorname{Cov}_t(\Theta\mid U)\big)\geq\gamma p_t(U)c(2t\mu)^{-1}.
\end{aligned}$$

As Lemma 2.5 yields $p_t(U)\xrightarrow{\text{a.s.}}1$, the statement follows. □

Lemma 4.8. *For almost any $\theta_0\in\boldsymbol{\Theta}$ satisfying* O1–O3, *there exists a neighborhood $U\subset U_0$ of θ_0 such that conditioned on θ_0 as the true parameter value*, *almost surely*,

$$Q_t := \frac{\mathrm{I}_t(\Theta;Y_{X_{t+1}}\mid U)}{\mathrm{I}_t(\Theta;Y_{X_{t+1}})}\rightsquigarrow 1.$$

Proof. By Lemmas 2.2, 2.3 and 4.3, almost surely, the convergences

$$\begin{aligned}
\mathrm{I}_t(\Theta;Y_{X_{t+1}}\mid U)&\to 0,\\
t\mathrm{I}_t\big([\Theta\in U];Y_{X_{t+1}}\big)&\rightsquigarrow 0
\end{aligned}$$

hold for all neighborhoods U in a countable basis of the compact metrizable space $\boldsymbol{\Theta}$. It follows that the same is true conditioned on almost any $\theta_0 \in \boldsymbol{\Theta}$ as the true parameter value. Thus, given almost any $\theta_0 \in \boldsymbol{\Theta}$, we can pick a neighborhood $U \subset U_0$ of θ_0 from the countable basis such that the above convergences almost surely hold.

Lemma 4.7 (applied to $\mu = M$) almost surely yields

$$\mathrm{I}_t(\Theta; Y_{t+1}) \geq c(Mt)^{-1} =: c_1 t^{-1}$$

for all sufficiently large t, where we denote $Y_{t+1} = Y_{X_{t+1}}$. Condition C4 + Lemma 2.5 yields

$$\mathrm{I}_t(\Theta; Y_{t+1}) \geq \gamma \sup_{x \in \mathsf{X}} \mathrm{I}_t(\Theta; Y_x) \geq \gamma c p_t\big(U^c\big) =: c_2 p_t\big(U^c\big)$$

for all sufficiently large t, and the chain rule of mutual information yields

$$\mathrm{I}_t(\Theta; Y_{t+1}) = \mathrm{I}_t\big([\Theta \in U]; Y_{t+1}\big) + p_t(U)\mathrm{I}_t(\Theta; Y_{t+1} \mid U) + p_t\big(U^c\big)\mathrm{I}_t\big(\Theta; Y_{t+1} \mid U^c\big).$$

Thus, almost surely,

$$\frac{\mathrm{I}_t(\Theta; Y_{t+1} \mid U)}{\mathrm{I}_t(\Theta; Y_{t+1})} = \underbrace{\frac{1}{p_t(U)}}_{\to 1}\left[1 - \frac{\overbrace{p_t(U^c)}^{\leq \mathrm{I}_t(\Theta; Y_{t+1})/c_2} \quad \overbrace{\mathrm{I}_t(\Theta; Y_{t+1} \mid U^c)}^{\to 0} + \overbrace{t\mathrm{I}_t([\Theta \in U]; Y_{t+1})}^{\rightsquigarrow 0} t^{-1}}{\underbrace{\mathrm{I}_t(\Theta; Y_{t+1})}_{\geq c_1 t^{-1}}}\right] \rightsquigarrow 1.$$

□

Corollary 4.2. *Conditioned on almost any* θ_0 *satisfying* O1–O4, *the sequence*

$$D_t := \sup_{x \in \mathsf{X}} B_t^{-1} \odot I_x(\theta_0) - B_t^{-1} \odot I_{X_{t+1}}(\theta_0)$$

satisfies $[\min \lambda_{B_t} \geq \mu] D_t \rightsquigarrow 0$ *a.s. for any given* $\mu > 0$, *where* $\min \lambda_{B_t}$ *denotes the smallest eigenvalue of* $B_t := -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$.

Proof. Let us first shrink the neighborhood U_0 of θ_0 as necessary to make its diameter smaller than the constant $\delta_{\mu,C}$ given by Lemma 3.2. Then, let $U \subset U_0$ be the neighborhood of θ_0 given by Lemma 4.8. By Theorem 3.1(3), there now exist random sequences $E_t \to 0$ and $E_t' \to 0$ such that conditioned on θ_0 as the true value,

$$\frac{1}{2} \sup_{x \in \mathsf{X}} B_t^{-1} \odot I_x(\theta_0) = \sup_{x \in \mathsf{X}} t\mathrm{I}_t(\Theta; Y_x \mid U) + E_t,$$

$$\frac{1}{2} B_t^{-1} \odot I_{X_{t+1}}(\theta_0) = t\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U) + E_t'$$

whenever $\min \lambda_{B_t} \geq \mu$. For these t, it follows

$$\frac{1}{2} D_t = \left(\frac{1}{2} \sup_{x \in \mathsf{X}} \underbrace{B_t^{-1} \odot I_x(\theta_0)}_{= \mathrm{tr}(B_t^{-1} I_x(\theta_0)) \leq n\mu^{-1} M} - E_t\right)\left(1 - \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)}{\sup_{x \in \mathsf{X}} \mathrm{I}_t(\Theta; Y_x \mid U)}\right) + E_t - E_t',$$

where Lemma 4.8 and the inequality $\mathrm{I}_t(\Theta; Y_x) \geq p_t(U)\mathrm{I}_t(\Theta; Y_x \mid U)$ yield

$$\frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)}{\sup_{x\in\mathsf{X}} \mathrm{I}_t(\Theta; Y_x \mid U)} \geq p_t(U)\frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)}{\sup_{x\in\mathsf{X}} \mathrm{I}_t(\Theta; Y_x)} = p_t(U)Q_t R_t \rightsquigarrow 1,$$

and so $[\min\lambda_{B_t} \geq \mu]D_t \rightsquigarrow 0$. □

Lemma 4.9. *Conditioned on almost any* θ_0 *satisfying* O1–O3, *there exists* μ *such that* $\min\lambda_{B_t} \geq \mu$ *for infinitely many* $t \in \mathbb{N}$, *where* $\min\lambda_{B_t}$ *denotes the smallest eigenvalue of* $B_t = -t^{-1}\nabla_\theta^2 p(\mathbf{Y}_t \mid \theta_0)$.

Proof. Let $\mu > 0$ be arbitrary. Lemma 4.7 almost surely yields $\mathrm{I}_{t-1}(\Theta; Y_{X_t}) \geq c(t\mu)^{-1}$ for all sufficiently large t satisfying $\min\lambda_{B_t} < \mu$ and Lemma 4.8 implies that $\mathrm{I}_{t-1}(\Theta; Y_{X_t} \mid U_0) \geq c(t\mu)^{-1}$ for a.e. t satisfying $\min\lambda_{B_t} \leq \mu$. Let then $K_\mu := \{t \in \mathbb{N}\colon \min\lambda_{B_t} \geq \mu\}$ and suppose that $\rho(K_\mu) = 0$. Then, $\rho_j := \rho_{2^j,2^{j+1}}(K_\mu) \to 0$, and then exists j_0 such that $\rho_j \leq 1/2$ for all $j \geq j_0$. It follows

$$\sum_{t=1}^{2^{j_1}-1} \mathrm{I}_{t-1}(\Theta; Y_{X_t} \mid U_0) \geq \frac{c}{\mu}\sum_{t=1}^{2^{j_1}-1}[t \notin K_\mu]\frac{1}{t} \geq \frac{c}{\mu}\sum_{j=j_0}^{j_1-1}\sum_{t=2^j(1+\rho_j)}^{2^{j+1}-1}\frac{1}{t} \geq \frac{c}{\mu}(j_1 - j_0)\log\frac{2}{3/2},$$

and so

$$\sum_{k=1}^{t}\mathrm{I}_{k-1}(\Theta; Y_{X_k} \mid U_0) \geq \left(\frac{c}{\mu}\log\frac{4}{3}\right)\log_2(t-1) - c_{c,\mu}$$

for all $t = 2^j$, $j \geq j_0$. Since μ was arbitrary, this implies that the sum grows asymptotically superlogarithmically if $\rho(K_\mu) = 0$ holds for all $\mu > 0$. If this event has positive probability among all $\theta_0 \in U_0$, then also

$$\mathrm{I}(\Theta; \mathbf{Y}_t \mid U_0) = \mathrm{E}\left(\sum_{k=1}^{t}\mathrm{I}_{k-1}(\Theta; Y_{X_k} \mid U_0)\,\middle|\, U_0\right)$$

grows superlogarithmically, contradicting Lemma 2.9. Thus, for almost all $\theta_0 \in U_0$ satisfying O1–O3, either K_μ is not ρ-measurable or $\rho(K_\mu) > 0$. In either case K_μ is infinite. □

Theorem 4.1 (Asymptotic D-optimality, part 1). *Conditioned on almost any* $\theta_0 \in \boldsymbol{\Theta}$ *satisfying* O1–O4, *almost surely,*

$$B_t := -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0) \to B^* := \operatorname*{arg\,max}_{B\in\mathcal{I}} \det(B),$$

where $\mathcal{I}$ *is the convex hull of the closure of* $\{I_x(\theta_0)\}_{x\in\mathsf{X}}$. *The maximizer* B^* *is unique, because the determinant is log-concave on the compact convex set* $\mathcal{I}$. *This result is optimal in the sense that for any strategy of choosing the placements* X_t (*instead of* O4 *and* C4), *almost surely* $\limsup_{t\to\infty}\det(B_t) \leq \det(B^*)$.

Proof. The objective function is

$$f(B) = \begin{cases} \log\det(B), & \min\lambda_B > 0, \\ -\infty, & \text{otherwise,} \end{cases}$$

where λ_B denotes the set of eigenvalues of B. Lemma 3.5 implies that B_t is asymptotically a convex combination of matrices in the closure of $\{I_x(\theta_0)\}_{x\in\mathsf{X}}$ and so $\limsup_{t\to\infty} f(B_t) \le f(B^*)$. Let us then show that this upper bound is tight.

First, we choose some representation $B^* = \sum_{k=1}^m \alpha_k I_k$ of the optimum point, where I_k are matrices in the closure of $\{I_x(\theta_0)\}_{x\in\mathsf{X}}$ and $\sum_{k=1}^m \alpha_k = 1$.

For any symmetric real matrix B_t, we have (with slight abuse of notation)

$$\begin{aligned}
\nabla f(B_t) &= B_t^{-1}, \\
\nabla^2 f(B_t) &= -\big[\big(B_t^{-1}\big)_i \big(B_t^{-1}\big)_j^T\big]_{i,j}^n, \\
\big[\nabla^2 f(B_t)\big]B &= -\big[\big(B_t^{-1}\big)_i \big(B_t^{-1}\big)_j^T \odot B\big]_{i,j}^n = -B_t^{-1} B B_t^{-1}, \\
B \odot \big[\nabla^2 f(B_t)\big]B &= -\operatorname{tr}\big(B_t^{-1} B B_t^{-1} B\big),
\end{aligned}$$

and Taylor's theorem yields

$$f(B_{t+1}) = f(B_t) + B_t^{-1} \odot (B_{t+1} - B_t) - \tfrac{1}{2}\operatorname{tr}\big(B_t^{-1} B' B_t^{-1} B'\big),$$

where B' is between 0 and $B_{t+1} - B_t$. Denoting $B := -\nabla_\theta^2 \log(p(Y_{X_{t+1}} \mid \theta_0))$, we obtain

$$\begin{aligned}
f(B_{t+1}) - f(B_t) &= f\left(\frac{tB_t + B}{t+1}\right) - f(B_t) \\
&= B_t^{-1} \odot \frac{B - B_t}{t+1} - \frac{1}{2} \underbrace{\operatorname{tr}\big(B_t^{-1} B' B_t^{-1} B'\big)}_{|\cdot| \le n4M^2\mu^{-2}(t+1)^{-2}} \\
&\ge \frac{1}{t+1}\left(B_t^{-1} \odot B - n - \frac{2nM^2\mu^{-2}}{t+1}\right),
\end{aligned}$$

for all indices t satisfying $\min\lambda_{B_t} \ge \mu$ for any $\mu > 0$. Denoting by λ_i the eigenvalues of $B_t^{-1}B^*$, Corollary 4.2 now implies that

$$\begin{aligned}
B_t^{-1} \odot I_{X_{t+1}}(\theta_0) + D_t &= \sup_{x\in\mathsf{X}} B_t^{-1} \odot I_x(\theta_0) \\
&\ge \max_k B_t^{-1} \odot I_k \ge \sum_k \alpha_k \big(B_t^{-1} \odot I_k\big) = B_t^{-1} \odot B^* \\
&= \operatorname{tr}\big(B_t^{-1} B^*\big) = \sum_{i=1}^n \lambda_i = n + \sum_{i=1}^n (\lambda_i - 1) \ge n + \sum_{i=1}^n \log(\lambda_i) \\
&= n + \log\det\big(B_t^{-1} B^*\big) = n + f\big(B^*\big) - f(B_t),
\end{aligned}$$

where $[\min \lambda_{B_t} \geq \mu] D_t \rightsquigarrow 0$ for any $\mu > 0$. Noting that $I_{X_{t+1}}(\theta_0) = \mathrm{E}_t(B \mid \theta_0)$, we obtain

$$\mathrm{E}_t\big(f(B_{t+1}) \mid \theta_0\big) - f(B_t) \geq \frac{1}{t+1}\big(f\big(B^*\big) - f(B_t) - D_{\mu,t}\big),$$

where $D_{\mu,t} = D_t + (2nM^2\mu^{-2})/(t+1)$.

From now on, in order to keep the notation clean, we will implicitly condition all probability statements on $\Theta = \theta_0$.

Let the constants $f_0 < f_1 < f(B^*)$ be arbitrary and define $\mu := \exp(f_0)M^{1-n}/2 > 0$. Suppose that some t_0 satisfies $f(B_{t_0}) \geq f_0$. Then, the definition of μ guarantees that $\min \lambda_{B_{t_0}} \geq 2\mu$. Let then $\alpha \in]1, \exp(\mu/M)]$ be arbitrary. Since $\min \lambda_{B_t}$ can decrease by at most M/t per each step, we obtain

$$\min \lambda_{B_t} \geq 2\mu - \sum_{t=t_0+1}^{t_1} \frac{M}{t} \geq 2\mu - M \log \frac{t_1}{t_0} \geq \mu$$

for all t between t_0 and $t_1 := \lfloor \alpha t_0 \rfloor$. Thus, the following inequalities hold true for all $t \in [t_0, t_1[$:

$$\begin{aligned}
\mathrm{E}_{t-1} f(B_t) - f(B_{t-1}) &\geq \frac{1}{t}\big(f\big(B^*\big) - f(B_{t-1}) - D_{\mu,t-1}\big),\\
\mathrm{E}_{t-1}\big(t f(B_t) - (t-1) f(B_{t-1})\big) &\geq f\big(B^*\big) - D_{\mu,t-1},\\
\mathrm{E}_{t_0}\big(t f(B_t) - (t-1) f(B_{t-1})\big) &\geq f\big(B^*\big) - \mathrm{E}_{t_0} D_{\mu,t-1},\\
\sum_{t=t_0+1}^{t_1} \mathrm{E}_{t_0}\big(t f(B_t) - (t-1) f(B_{t-1})\big) &\geq \sum_{t=t_0+1}^{t_1} \big(f\big(B^*\big) - \mathrm{E}_{t_0} D_{\mu,t-1}\big),\\
\mathrm{E}_{t_0}\big(t_1 f(B_{t_1})\big) - t_0 f(B_{t_0}) &\geq (t_1 - t_0) f\big(B^*\big) - \sum_{t=t_0}^{t_1-1} \mathrm{E}_{t_0} D_{\mu,t},
\end{aligned}$$

and dividing by t_1, we obtain the inequality

$$\begin{aligned}
\mathrm{E}_{t_0} f(B_{t_1}) - \alpha^{-1} f(B_{t_0}) \;&\geq\; \left(1 - \frac{t_0}{t_1}\right) f\big(B^*\big) - \mathrm{E}_{t_0}\left(\frac{1}{t_1}\sum_{t=t_0}^{t_1-1} D_{\mu,t}\right)\\
&\to \big(1 - \alpha^{-1}\big) f\big(B^*\big),
\end{aligned}$$

where we have used the fact that $t_1 \leq \alpha t_0$, and where the convergence holds for any increasing sequence of indices t_0 satisfying $f(B_{t_0}) \geq f_0$ (which implies $\min \lambda_{B_t} \geq \mu$ for all $t \in [t_0, t_1[$). This convergence is obtained by applying Lemma 4.2(3) to the bounded sequence $[\min \lambda_{B_t} \geq \mu] D_{\mu,t} \rightsquigarrow 0$, which yields

$$\left|\frac{1}{t_1}\sum_{t=t_0}^{t_1-1} D_{\mu,t}\right| \leq \frac{1}{t_1}\sum_{t=0}^{t_1-1}\big|[\min \lambda_{B_t} \geq \mu] D_{\mu,t}\big| \to 0$$

(and since $|D_{\mu,t}| \leq 2nM\mu^{-1} + 2nM^2\mu^{-2}$ for all t, Lebesgue's dominated convergence theorem allows us to take this limit inside the expectation). Thus, there exists a positive constant s such that

$$\mathrm{E}_{t_0} f(B_{t_1}) \geq f(B_{t_0}) + 2s$$

for all sufficiently large t_0 satisfying $f_0 \leq f(B_{t_0}) \leq f_1$. Also, since the maximum change in the value of f over one step is bounded by v/t for some constant $v > 0$ (depending on μ), we obtain

$$\mathrm{Var}_{t_0} f(B_{t_1}) \leq \sum_{t=t_0+1}^{t_1} \left(\frac{v}{t}\right)^2 \leq \int_{t_0}^{t_1} \left(\frac{v}{t}\right)^2 \mathrm{d}t = v^2\left(\frac{1}{t_0} - \frac{1}{t_1}\right) \leq \frac{v^2}{t_0}.$$

Now Markov's inequality yields

$$\begin{aligned}
\mathsf{P}_{t_0}\big\{f(B_{t_1}) < f(B_{t_0}) + s\big\} &\leq \mathsf{P}_{t_0}\big\{f(B_{t_1}) < \mathrm{E}_{t_0} f(B_{t_1}) - s\big\} \\
&\leq \mathsf{P}_{t_0}\big\{\big|\mathrm{E}_{t_0} f(B_{t_1}) - f(B_{t_1})\big|^2 > s^2\big\} \\
&\leq \frac{\mathrm{Var}_{t_0} f(B_{t_1})}{s^2} \leq \frac{v^2}{t_0 s^2}.
\end{aligned}$$

As this upper bound on the probability sums to a finite number over the sequence $t_0(k)$ determined by $t_0(k+1) = t_1(k) = \lfloor \alpha t_0(k) \rfloor$, the Borel–Cantelli lemma implies that almost surely $f(B_{t_0(k+1)}) < f(B_{t_0(k)}) + s$ holds for only finitely many indices $k \in \mathbb{N}$ satisfying $f_0 \leq f(B_{t_0(k)}) \leq f_1$. Thus, there exists k_0 such that for all $k \geq k_0$, whenever $f_0 \leq f(B_{t_0(k)}) \leq f_1$, the value $f(B_{t_0}(k))$ will increase by at least s on each step as k increases. Furthermore, since

$$\big|f(B_t) - f(B_{t_0(k)})\big| \leq \sum_{t=t_0(k)+1}^{t_1(k)} \frac{v}{t} \leq v \log \frac{t_1(k)}{t_0(k)} \leq v \log \alpha$$

for all $t \in [t_0(k), t_1(k)[$, it follows that if $f(B_{t_0(k)}) \geq f_0$ for any $k \geq k_0$, then $f(B_t) \geq f_1 - v\log\alpha$ for all sufficiently large t (provided that $f_1 - v\log\alpha \geq f_0$). Since $f_1 - v\log\alpha$ can be made arbitrarily close to $f(B^*)$ by appropriate choices of rational $\alpha > 1$ and rational $f_1 < f(B^*)$ for arbitrarily small rational f_0, we almost surely obtain $\liminf_{t\to\infty} f(B_t) \geq f(B^*)$ unless $f(B_t)$ eventually stays below any number. But this would imply that $\limsup_{t\to\infty} \min \lambda_{B_t} \leq 0$, which is almost surely contradicted by Lemma 4.9. □

Corollary 4.3 (Asymptotic D-optimality, part 2). *Conditioned on almost any $\theta_0 \in \mathbf{\Theta}$ satisfying* O1–O4, *there exists a neighborhood U of θ_0 such that $t\,\mathrm{Cov}_t(\Theta \mid U) \xrightarrow{a.s.} (B^*)^{-1}$. This is optimal in the sense that for any other strategy in place of* O4 *and* C4, *almost surely $\liminf_{t\to\infty} \det(t\,\mathrm{Cov}_t(\Theta \mid U)) \geq \det(B^*)^{-1}$.*

Proof. Given O4, Theorems 4.1 and 3.1(2) imply that $t\,\mathrm{Cov}_t(\Theta \mid U) \xrightarrow{\text{a.s.}} (B^*)^{-1}$. For any other strategy, we have $\limsup_{t\to\infty} \det(B_t) \leq \det(B^*)$ a.s., and so Theorem 3.1(2) yields $\liminf_{t\to\infty} \det(t\,\mathrm{Cov}_t(\Theta \mid U)) \geq \det(B^*)^{-1}$ a.s. as t increases within indices satisfying

$\min \lambda_{B_t} > \mu$ for some given $\mu > 0$. But Corollary 3.1 implies that if we choose a sufficiently small $\mu > 0$, then $\det(t \operatorname{Cov}_t(\Theta \mid U)) \geq \det(B^*)^{-1}$ also for $\min \lambda_{B_t} \leq \mu$, and the statement follows. □

Remark 4.1. As discussed in the beginning of this section, secondary modes with weights proportional to $1/t$ may remain outside U, and they do contribute to the asymptotic variance. Thus, the D-optimality result (part 2) shown here is only a local form of optimality.

The situation would be different if the placements were chosen so as to minimize the determinant of the posterior covariance $\operatorname{Cov}_t(\Theta)$ directly (which, of course, presupposes that the parameter space has global Euclidean structure). Then, slightly more trials would be spent to decrease the weights of the secondary modes, but they should remain insignificant in proportion. Thus, we can conjecture that $B_t \xrightarrow{\text{a.s.}} B^*$ would still obtain in Theorem 4.1 with $t \operatorname{Cov}_t(\Theta)$ asymptotically equal to $(B_t)^{-1}$, making the result globally optimal.

4.3. Asymptotic entropy

Here we use the D-optimality result to derive an expression for the asymptotic entropy.

Corollary 4.4. *Conditioned on almost any $\theta_0 \in \boldsymbol{\Theta}$ satisfying* O1–O4, *for any neighborhood U of θ_0, there exists a constant c_U such that almost surely, $p_t(U^c) \leq c_U/t$ for a.e. $t \in \mathbb{N}$.*

Proof. Theorem 4.1 implies that $\min \lambda_{B_t} \geq \mu$ for all sufficiently large t for some $\mu > 0$. Hence, given any $\varepsilon > 0$, Theorem 3.1(3) yields

$$t \mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U) \leq \sup_{x \in \mathsf{X}} B_t^{-1} \odot I_x(\theta_0) + \varepsilon \leq n\mu^{-1} M + \varepsilon =: c$$

for all sufficiently large t, where U is any sufficiently small neighborhood of θ_0. Combined with Lemma 4.8, this implies that $\mathrm{I}_t(\Theta; Y_{X_{t+1}}) \leq 2c/t$ for a.e. $t \in \mathbb{N}$, and so Lemma 2.5(2) yields the statement. □

Remark 4.2. Note that the statement of Corollary 4.4 holds only for a.e. $t \in \mathbb{N}$. What happens in a sufficiently long run is that most trials are spent on increasing the accuracy around the global mode and an approximately logarithmically growing number of trials is spent on placements that decrease the weights of secondary modes. However, on any such trial there is a small probability that the weight of the secondary mode actually increases, and given a sufficiently long run, this will eventually happen arbitrarily many times in a row, making the weight of the secondary mode temporarily arbitrarily much larger than the c/t bound that holds on most trials.

Theorem 4.2. *Conditioned on almost any $\theta_0 \in \boldsymbol{\Theta}$ satisfying* O1–O4, *if the prior entropy* $\mathrm{H}(\Theta)$ *w.r.t. a parameterization that is consistent with the local Euclidean structure* (*i.e.*, *the prior density $p(\theta)$ is given w.r.t. a measure that coincides with the Lebesgue measure on subsets of U_0*) *is well-defined and finite*, *then*, *almost surely*

$$\mathrm{H}_t(\Theta) + \frac{n}{2} \log t \rightsquigarrow H^* := -\frac{1}{2} \log \det\bigl(B^*\bigr) + \frac{n}{2} \log(2\pi \mathrm{e}).$$

Proof. Let us condition everything on θ_0 being the true value. Theorem 3.1(2) implies that for some sufficiently small neighborhood U of θ_0,

$$\mathrm{H}_t(\Theta \mid U) + \frac{n}{2}\log t \xrightarrow{\text{a.s.}} H^*.$$

Lemmas 2.6 and 2.8 imply that for any $\varepsilon > 0$, $|\mathrm{H}_t(\Theta \mid U^c)| < \varepsilon t$ for all sufficiently large t, and as Corollary 4.4 yields $p_t(U^c) \le c/t$ for a.e. t, Lemma 4.2(2) implies $p_t(U^c)\mathrm{H}_t(\Theta \mid U^c) \rightsquigarrow 0$. The statement now follows from the chain rule of entropy

$$\mathrm{H}_t(\Theta) = p_t(U)\mathrm{H}_t(\Theta \mid U) + \underbrace{p_t\big(U^c\big)\mathrm{H}_t\big(\Theta \mid U^c\big)}_{\rightsquigarrow 0} + \underbrace{\mathrm{H}_t\big([\Theta \in U]\big)}_{\to 0 \text{ a.s.}},$$

where the first term satisfies

$$p_t(U)\mathrm{H}_t(\Theta \mid U) + \frac{n}{2}\log t = p_t(U)\left[\mathrm{H}_t(\Theta \mid U) + \frac{n}{2}\log t\right] + \underbrace{p_t\big(U^c\big)}_{\le c/t}\frac{n}{2}\log t \rightsquigarrow H^*.$$

□

Corollary 4.5. *Suppose that* O1–O4 *hold for almost all* $\theta_0 \in \boldsymbol{\Theta}$ *and that the prior entropy* $\mathrm{H}(\Theta)$ *w.r.t. a parameterization that is consistent with the local Euclidean structures* U_0 *in* O2 *is well-defined and finite. Then,*

$$\mathrm{H}_t(\Theta) + \frac{n}{2}\log t \overset{\mathrm{P}}{\rightsquigarrow} H^*.$$

In other words, there exists a set $K \subset \mathbb{N}$ *of indices with* $\rho(K) = 1$ *such that*

$$\mathrm{H}_t(\Theta) + \frac{n}{2}\log t \xrightarrow{\mathrm{P}} H^*,$$

as t *increases within* K.

Proof. Apply Lemma 4.6 to the statement of Theorem 4.2. □

4.4. Varying cost of observation

In Kujala [5] the adaptive sequential estimation framework is generalized to the situation where the observation of Y_x is associated with some random cost C_x of observation, which given the value of Y_x, is independent of Θ and the results and costs of any other observations:

$$\begin{array}{ccc} & \Theta & \\ \swarrow & \downarrow & \searrow \\ Y_x & Y_{x'} & \cdots \\ \downarrow & \downarrow & \\ C_x & C_{x'} & \cdots \end{array}$$

The technical requirement that C_x depends on Θ only through Y_x is satisfied in particular if C_x is a component of Y_x. Thus, it leads to no loss of generality if the incurred costs are observable.

The goal considered in Kujala [5] is maximization of the expected information gain of a sequential experiment that terminates when the total cost overruns a given budget. To achieve this goal, the heuristic of maximizing the expected information gain $\mathrm{I}_t(\Theta; Y_x)$ divided by the expected cost $\mathrm{E}_t(C_x)$ on each trial is proposed. In this section, we are able to show that this heuristic is in fact asymptotically optimal (as the budget tends to infinity) under essentially the same conditions that the plain information gain maximization is.

Thus, condition O4 is now replaced by the following:

O4′. The placements satisfy

$$R'_t := \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}})/\mathrm{E}_t(C_{X_{t+1}})}{\sup_{x\in\mathsf{X}}(\mathrm{I}_t(\Theta; Y_x)/\mathrm{E}_t(C_x))} \rightsquigarrow 1,$$

where $|C_x| \le M$, $\mathrm{E}(C_x \mid \theta_0) \ge \gamma' > 0$, and the family of expected cost functions $\{\theta \mapsto \mathrm{E}(C_x \mid \theta)\colon x \in \mathsf{X}\}$ is equicontinuous at θ_0.

Due to the assumed bounds on the expected cost $\mathrm{E}(C_x \mid \theta_0)$, condition C4 is still satisfied and so all the previous lemmas depending on it apply. Together with the following lemma, these bounds also imply that the total cost grows asymptotically within linear bounds.

Lemma 4.10. *Suppose that* O4′ *holds. Then, conditioned on θ_0 as the true parameter value,*

$$\frac{C_t - \sum_{k=1}^t \mathrm{E}(C_{X_k} \mid \theta_0)}{t} \xrightarrow{a.s.} 0,$$

where $C_t := \sum_{k=1}^t C_{X_k}$. In particular, for any $\gamma < \gamma'$, almost surely $C_t \ge t\gamma$ for all sufficiently large t (as well as $C_t \le tM$ for all t).

Proof. Denoting $Z_k = C_{X_k} - \mathrm{E}(C_{X_k} \mid \theta_0)$, given $\Theta = \theta_0$, the sequence $Z_1 + \cdots + Z_k$ of partial sums is a martingale and satisfies $\mathrm{E}(|Z_k|^2) \le M^2 < \infty$ for all k, and so Theorem A.3 implies that $(Z_1 + \cdots + Z_t)/t \xrightarrow{\text{a.s.}} 0$, which is the statement. □

Next, we will generalize Corollary 4.2 for the cost-aware placements.

Corollary 4.6. *Conditioned on almost any θ_0 satisfying* O1–O3 *and* O4′, *the sequence*

$$D_t := \sup_{x\in\mathsf{X}} B_t^{-1} \odot \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)} - B_t^{-1} \odot \frac{I_{X_{t+1}}(\theta_0)}{\mathrm{E}_t(C_{X_{t+1}} \mid \theta_0)}$$

satisfies $[\min \lambda_{(C_t/t)B_t} \ge \mu]D_t \rightsquigarrow 0$ *a.s. for any given* $\mu > 0$, *where* $\min \lambda_{(C_t/t)B_t}$ *denotes the smallest eigenvalue of* $B_t := -C_t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$ *and* $C_t := \sum_{k=1}^t C_{X_k}$.

Proof. Let us first shrink the neighborhood U_0 of θ_0 as necessary to make its diameter smaller than the constant $\delta_{\mu,C}$ given by Lemma 3.2. Then, let $U \subset U_0$ be the neighborhood of θ_0 given by Lemma 4.8. The boundedness and equicontinuity at θ_0 of $\theta \mapsto \mathrm{E}(C_x \mid \theta) \in [\gamma', M]$ imply that conditioned on $\Theta = \theta_0$, almost surely, $\mathrm{E}_t(C_x) \to \mathrm{E}(C_x \mid \theta_0)$, uniformly over all $x \in \mathsf{X}$. Combined with Theorem 3.1(3), this implies that there exist random sequences $E_t \to 0$ and $E_t' \to 0$ such that conditioned on θ_0 as the true value,

$$\frac{1}{2}\sup_{x\in\mathsf{X}} B_t^{-1} \odot \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)} = \sup_{x\in\mathsf{X}} C_t \frac{\mathrm{I}_t(\Theta; Y_x \mid U)}{\mathrm{E}_t(C_x)} + E_t,$$
$$\frac{1}{2} B_t^{-1} \odot \frac{I_{X_{t+1}}(\theta_0)}{\mathrm{E}(C_{X_{t+1}} \mid \theta_0)} = C_t \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)}{\mathrm{E}_t(C_{X_{t+1}})} + E_t'$$

whenever $\min \lambda_{(C_t/t)B_t} \geq \mu$. For these t, it follows

$$\frac{1}{2} D_t = \Bigg(\frac{1}{2}\sup_{x\in\mathsf{X}} \underbrace{B_t^{-1} \odot \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)}}_{\leq \operatorname{tr}(B_t^{-1} I_x(\theta_0))/\gamma \leq n(\gamma\mu)^{-1} M} - E_t\Bigg)\Bigg(1 - \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)/\mathrm{E}_t(C_{X_{t+1}})}{\sup_{x\in\mathsf{X}}(\mathrm{I}_t(\Theta; Y_x \mid U)/\mathrm{E}_t(C_x))}\Bigg)$$
$$+ E_t - E_t',$$

where Lemma 4.8 and the inequality $\mathrm{I}_t(\Theta; Y_x) \geq p_t(U)\mathrm{I}_t(\Theta; Y_x \mid U)$ yield

$$\frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)/\mathrm{E}_t(C_{X_{t+1}})}{\sup_{x\in\mathsf{X}}(\mathrm{I}_t(\Theta; Y_x \mid U)/\mathrm{E}_t(C_x))} \geq p_t(U) \frac{\mathrm{I}_t(\Theta; Y_{X_{t+1}} \mid U)/\mathrm{E}_t(C_{X_{t+1}})}{\sup_{x\in\mathsf{X}}(\mathrm{I}_t(\Theta; Y_x)/\mathrm{E}_t(C_x))} = p_t(U) Q_t R_t' \rightsquigarrow 1,$$

and so $[\min \lambda_{(C_t/t)B_t} \geq \mu] D_t \rightsquigarrow 0$. □

Lemma 4.11. *The range of the expression*

$$r_t = \frac{\sum_{k=1}^t I_{x_k}(\theta_0)}{\sum_{k=1}^t \mathrm{E}(C_{x_k} \mid \theta_0)}$$

over all sequences x_k in X and all finite t is a dense subset of the set $\mathcal{I}$ defined as the closure of the convex hull of

$$S = \left\{ \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)} \right\}_{x\in\mathsf{X}}.$$

Furthermore, the range of the limits of all converging r_t equals $\mathcal{I}$.

Proof. For any sequence $\{x_k\}$, we have

$$r_t = \frac{\sum_{k=1}^t I_{x_k}(\theta_0)}{\sum_{k=1}^t \mathrm{E}(C_{x_k} \mid \theta_0)} = \sum_{k=1}^t \underbrace{\left(\frac{\mathrm{E}(C_{x_k} \mid \theta_0)}{\sum_{k=1}^t \mathrm{E}(C_{x_k} \mid \theta_0)}\right)}_{=:\alpha_{k,t}} \frac{I_{x_k}(\theta_0)}{\mathrm{E}(C_{x_k} \mid \theta_0)},$$

and so r_t is always a convex combination of elements in S. The convex combination is not exactly linear w.r.t. the number of different x in the sequence because of the different $\mathrm{E}(C_{x_k} \mid \theta_0)$

weights, but nonetheless, by varying the proportions of different x in a sufficiently long sequence, any convex combination can be approximated arbitrarily well. □

Theorem 4.3 (Asymptotic D-optimality, part 1). *Conditioned on almost any* $\theta_0 \in \Theta$ *satisfying* O1–O3, O4′, *almost surely,*

$$B_t := \frac{-\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)}{C_t} \to B^* := \operatorname*{arg\,max}_{B \in \mathcal{I}} \det(B),$$

where $C_t := \sum_{k=1}^{t} C_{X_k}$ *and* $\mathcal{I}$ *is the convex hull of the closure of*

$$S = \left\{ \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)} \colon\ x \in \mathsf{X} \right\}.$$

This is optimal in the sense that for any strategy of choosing the placements X_t (*instead of* O4′ *and* C4), *almost surely* $\limsup_{t\to\infty} \det(B_t) \le \det(B^*)$.

Proof. Since S is bounded, $\mathcal{I}$ is a compact convex set and B^* is well defined. Lemmas 3.5, 4.10, and 4.11 imply that $\limsup_{t\to\infty} \det(B_t) \le \det(B^*)$ a.s. Let us then show that this upper bound is tight.

Lemma 4.11 implies that there exists a representation

$$B^* = \lim_{m\to\infty} \frac{\sum_{k=1}^{m} I_k}{\sum_{k=1}^{m} c_k}$$

of the optimum point B^* where (I_k, c_k) are elements of $\{(I_x(\theta_0), \mathrm{E}(C_x \mid \theta_0)) \colon\ x \in \mathsf{X}\}$.

Denoting $B := -\nabla_\theta^2 \log(p(Y_{X_{t+1}} \mid \theta_0))$ and $C := C_{X_{t+1}}$, and assuming $\min \lambda_{(C_t/t)B_t} \ge \mu$, we obtain

$$|B|, |C| \le M, \qquad \left|B_t^{-1}\right| \le (\mu/M)^{-1}, \qquad |B - CB_t| \le M + M^2/\mu, \qquad C_t + C \ge \gamma(t+1)$$

and so, for some B' between 0 and $B_{t+1} - B_t$, we obtain

$$\begin{aligned}
f(B_{t+1}) - f(B_t) &= f\left(\frac{C_t B_t + B}{C_t + C}\right) - f(B_t) \\
&= B_t^{-1} \odot \frac{B - CB_t}{C_t + C} - \frac{1}{2}\operatorname{tr}\bigl(B_t^{-1} B' B_t^{-1} B'\bigr) \\
&\ge \frac{1}{C_t + C}\left(B_t^{-1} \odot B - nC - \frac{[(\mu/M)^{-1}(M + M^2/\mu)]^2}{C_t + C}\right) \\
&\ge \underbrace{\frac{\mathrm{E}_t(C \mid \theta_0)}{C_t + C}}_{\ge (\gamma/M)/(t+1)} \left(B_t^{-1} \odot \frac{B}{\mathrm{E}_t(C \mid \theta_0)} - \frac{nC}{\mathrm{E}_t(C \mid \theta_0)} - \frac{c_{M,\mu,\gamma}}{t+1}\right).
\end{aligned}$$

Denoting by λ_i the eigenvalues of $B_t^{-1}B^*$, we obtain

$$\begin{aligned}
B_t^{-1} \odot \frac{\mathrm{I}_{X_{t+1}}(\theta_0)}{\mathrm{E}(C_{X_{t+1}} \mid \theta_0)} + D_t &= \sup_{x\in \mathsf{X}} B_t^{-1} \odot \frac{I_x(\theta_0)}{\mathrm{E}(C_x \mid \theta_0)} \\
&\geq \sup_k \left(B_t^{-1} \odot \frac{I_k}{c_k} \right) \geq \lim_{m\to\infty} \frac{\sum_{k=1}^m (B_t^{-1} \odot I_k)}{\sum_{k=1}^m c_k} = B_t^{-1} \odot B^* \\
&= \operatorname{tr}\bigl(B_t^{-1}B^*\bigr) = \sum_{i=1}^n \lambda_i = n + \sum_{i=1}^n (\lambda_i - 1) \geq n + \sum_{i=1}^n \log(\lambda_i) \\
&= n + \log\det\bigl(B_t^{-1}B^*\bigr) = n + f\bigl(B^*\bigr) - f(B_t),
\end{aligned}$$

where Corollary 4.6 implies that $[\min \lambda_{(C_t/t)B_t} \geq \mu]D_t \rightsquigarrow 0$. Noting that $\mathrm{E}_t(B/\mathrm{E}_t(C \mid \theta_0) \mid \theta_0) = \mathrm{I}_{X_{t+1}}(\theta_0)/\mathrm{E}(C_{X_{t+1}} \mid \theta_0)$, it follows

$$\mathrm{E}_t\bigl(f(B_{t+1}) \mid \theta_0\bigr) - f(B_t) \geq \frac{\gamma/M}{t+1}\bigl(f\bigl(B^*\bigr) - f(B_t) - D_{\mu,t}\bigr),$$

where $D_{\mu,t} = D_t + C_{M,\mu,\gamma}/(t+1)$.

From here on, the proof is essentially the same as in the maximum information case. We just use $\mu := \exp(f_0)M^{-n}/2$ to guarantee that $\min \lambda_{(C_t/t)B_t} \geq 2\mu$ for $f(B_t) \geq f_0$. □

The part 2 of the D-optimality result as well as analogs of the asymptotic entropy results follow with essentially the same proofs (just replacing t with C_t at appropriate places):

Corollary 4.7 (Asymptotic D-optimality, part 2). *Conditioned on almost any $\theta_0 \in \boldsymbol{\Theta}$ satisfying* O1–O3, O4′, *there exists a neighborhood U of θ_0 such that $C_t \operatorname{Cov}_t(\Theta \mid U) \xrightarrow{a.s.} (B^*)^{-1}$, where $C_t := \sum_{k=1}^t C_{X_k}$. This is optimal in the sense that for any other strategy in place of* O4′ *and* C4, *almost surely* $\liminf_{t\to\infty} \det(C_t \operatorname{Cov}_t(\Theta \mid U)) \geq \det(B^*)^{-1}$.

Theorem 4.4. *Conditioned on almost any $\theta_0 \in \boldsymbol{\Theta}$ satisfying* O1–O3, O4′, *if the prior entropy* $\mathrm{H}(\Theta)$ *w.r.t. a parameterization that is consistent with the local Euclidean structure* (*i.e.*, *the prior density $p(\theta)$ is given w.r.t. a measure that coincides with the Lebesgue measure on subsets of U_0*) *is well-defined and finite*, *then*, *almost surely*

$$\mathrm{H}_t(\Theta) + \frac{n}{2}\log C_t \rightsquigarrow H^* := -\frac{1}{2}\log\det\bigl(B^*\bigr) + \frac{n}{2}\log(2\pi \mathrm{e}),$$

where $C_t := \sum_{k=1}^t C_{X_k}$.

Corollary 4.8. *Suppose that* O1–O4 *hold for almost all $\theta_0 \in \boldsymbol{\Theta}$ and that the prior entropy* $\mathrm{H}(\Theta)$ *w.r.t. a parameterization that is consistent with the local Euclidean structures U_0 in* O2 *is well-defined and finite*. *Then*,

$$\mathrm{H}_t(\Theta) + \frac{n}{2}\log C_t \overset{\mathrm{P}}{\rightsquigarrow} H^*,$$

where $C_t := \sum_{k=1}^{t} C_{X_k}$. *In other words, there exists a set* $K \subset \mathbb{N}$ *of indices with* $\rho(K) = 1$ *such that*

$$\mathrm{H}_t(\Theta) + \frac{n}{2} \log C_t \xrightarrow{\mathrm{P}} H^*,$$

as t *increases within* K.

5. Examples

In this section, we give specific examples illustrating the optimality results.

Example 5.1 (Psychometric model). Consider the psychometric model, where an observer's unknown intensity threshold Θ for detecting a stimulus of intensity x is distributed uniformly on $[0, 100]$ and the trial result $Y_x \in \{0, 1\}$ for a test intensity $x \in [0, 100]$ is distributed as

$$p(y_x \mid \theta) = \begin{cases} \psi(\theta - x), & y_x = 1 \text{ (detected)}, \\ 1 - \psi(\theta - x), & y_x = 0 \text{ (not detected)}, \end{cases}$$

where $\psi(x)$ is the psychometric function, here assumed to be the sigmoid

$$\psi(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

for simplicity (for more general psychometric models, see Kujala and Lukka [7], and the references therein).

In this model, the Fisher information of a given placement x is calculated as

$$I_x(\theta) = \sum_{y_x=0}^{1} p(y_x \mid \theta)\left[\frac{\partial}{\partial\theta} \log p(y_x \mid \theta)\right]^2 = \frac{\psi'(\theta - x)^2}{\psi(\theta - x)[1 - \psi(\theta - x)]} = \frac{\mathrm{e}^{\theta - x}}{[1 + \mathrm{e}^{\theta - x}]^2}.$$

Thus, for any given θ_0, the D-optimal value of the averaged Fisher information in Theorem 4.1 is $B^* = \frac{1}{4}$ given by the placement $x = \theta_0$ to which the greedy algorithm eventually converges. Now Corollary 4.5 yields

$$\mathrm{H}_t(\Theta) + \frac{n}{2} \log t \overset{\mathrm{P}}{\rightsquigarrow} H^* = -\frac{1}{2} \log \underbrace{\det(B^*)}_{=0.25} + \frac{n}{2} \log(2\pi\mathrm{e}) \tag{5.1}$$

and this is the asymptotically optimal posterior entropy. In this example, the same expression also gives the asymptotically optimal expected utility $\mathrm{E}(\mathrm{H}_t(\Theta)) + \frac{n}{2} \log t$, which we will next compare to that of the offline design.

Example 5.2 (Offline design). A rigorous study of the optimal offline design is beyond the scope of the present article, so we will not go into detailed proofs here but only sketch the general ideas. Suffice it to say that for an offline design for optimizing the expected utility $\mathrm{E}(\mathrm{H}_t(\Theta))$, one cannot do much better than to use the usual strategy of placing the trials evenly on the interval [0, 100]. (Due to boundary effects, an exactly uniform distribution of placements is not really the global optimum, but for simplicity, we avoid a more complicated discussion here.)

For uniform placement of trials on [0, 100], Lemma 3.5 implies

$$B_t \xrightarrow{\text{a.s.}} \frac{1}{100}\int_0^{100} I_x(\theta_0)\,\mathrm{d}x = \frac{1}{100}\left(\frac{1}{1+\mathrm{e}^{-\theta_0}} - \frac{1}{1+\mathrm{e}^{100-\theta_0}}\right) \in [0.005, 0.01],$$

where $B_t = -t^{-1}\nabla_\theta^2 \log p(\mathbf{Y}_t \mid \theta_0)$, and it can be shown that the asymptotic posterior entropy satisfies

$$\mathrm{H}_t(\Theta) + \frac{n}{2}\log t - \left[-\frac{1}{2}\log \underbrace{\det(B_t)}_{\lim \le 0.01}\right] + \frac{n}{2}\log(2\pi \mathrm{e}) \xrightarrow{\text{a.s.}} 0,$$

which implies the asymptotic lower bound

$$\liminf_{t\to\infty}\left[\mathrm{H}_t(\Theta) + \frac{n}{2}\log t\right] \ge -\frac{1}{2}\log 0.01 + \frac{n}{2}\log(2\pi\mathrm{e})$$

on the posterior entropy. Comparing to the asymptotically optimal posterior entropy (5.1), it follows that the offline design needs asymptotically at least $(\frac{0.25}{0.01})^{1/n} = 25$ times as many trials as the optimal adaptive design for the same accuracy. If the range [0, 100] is doubled, then this number approximately doubles as well, so the gap to the asymptotically optimal adaptive design can be arbitrarily large.

Example 5.3 (Varying cost of observation). Let us then return to the adaptive case and suppose that instead of a unit cost, each trial costs

$$C_x = 1 + 3[Y_x = 0]$$

units. Such a formulation could be based on the assumption that the observer takes four times as long to respond when the stimulus is not detected. Then, the asymptotic efficiency of a placement x in Theorem 4.3 is characterized by the expression

$$\frac{I_x(\theta_0)}{\mathrm{E}(C_x)} = \frac{I_x(\theta_0)}{1+3[1-\psi(\theta_0 - x)]} = \frac{1}{5+5\cosh(\theta_0 - x) - 3\sinh(\theta_0 - x)}. \tag{5.2}$$

This expression is maximized by the placement $x = \theta_0 + \log 2$ to which the myopic algorithm eventually converges to (provided it is within the range [0, 100]). Thus, assuming that $\theta_0 \le 100 - \log 2 \approx 99.3069$ and substituting the maximizer in (5.2), we obtain in Theorem 4.3 the D-optimal asymptotic efficiency $B^* = \frac{1}{9}$. Comparing to the asymptotically optimal placement $x = \theta_0$ for unit cost (yielding $B^* = \frac{1}{10}$ in (5.2)), we see that the cost-aware strategy reaches the same accuracy in 10% less cost (time) in this example.

6. Discussion

We have derived an expression for the asymptotic efficiency of any sequential experiment design for both the standard framework with unit cost of observation as well as for the generalized framework with random costs of observation as proposed in Kujala [5]. We have shown an asymptotic D-optimality result for the greedy information optimization strategy in the standard framework and we have extended this result for the novel myopic strategy proposed in Kujala [5] for the situation with random costs of observations. These results indicate that for (almost) all true parameter values θ_0, the greedy or myopic adaptive design is asymptotically optimal among all placement strategies in a well-defined sense.

Assuming the standard sequential estimation framework with unit cost of observation, Lemma 3.5 together with the asymptotic normality result imply that the asymptotic efficiency of any given design is characterized by the average

$$\frac{\sum_{k=1}^{t} I_{X_k}(\theta_0)}{t}$$

of the Fisher information matrices $I_x(\theta_0)$ over the sequence of placements X_t and the D-optimality criterion of a design refers to maximality of the determinant of this averaged information matrix at the limit. For any given θ_0, there is a distribution (or sequence) of placements $x \in \mathsf{X}$ yielding the D-optimal average information matrix. For (almost) all θ_0, the placements of the greedy adaptive design converge to such an optimum, whereas the offline design cannot adjust the distribution of the placements $x \in \mathsf{X}$ depending on the true value θ_0. Thus, the offline design can be equally efficient for a given true value of Θ, but generally not for all values $\theta_0 \in \mathbf{\Theta}$ and depending on the model, the gap in efficiency can be arbitrarily large as seen in Example 5.2.

The situation is essentially the same in the framework with random costs of observation, the only difference being that the convergence of the estimate of Θ is not measured in relation to t but in relation to the total cost $C_t = C_{X_1} + \cdots + C_{X_t}$ of placements. In this situation, the asymptotic efficiency is characterized by the ratio

$$\frac{\sum_{k=1}^{t} I_{X_k}(\theta_0)}{\sum_{k=1}^{t} \mathrm{E}(C_{X_k} \mid \theta_0)}$$

and the limit is again determined by the distribution (or sequence) of the placements $x \in \mathsf{X}$. Theorem 4.3 shows that the myopic strategy of maximizing

$$\frac{\mathrm{I}_t(\Theta; Y_x)}{\mathrm{E}_t(C_x)}$$

yields the asymptotically D-optimal efficiency in this situation.

However, the actual utility function assumed in both of the frameworks considered is the differential entropy, and so the most relevant asymptotic optimality criterion should be based on the asymptotic properties of the differential entropy as shown in, for example, Corollaries 4.5 and 4.8. Thus, a topic for future work is finding conditions under which the results of Corollaries 4.5 and 4.8 can be said to be optimal among all placement strategies.

Appendix: Auxiliary theorems

Theorem A.1 (Stone–Čech compactification). *Suppose that X is a Tychonoff space. Then there exists a compact space βX that embeds X as a dense subspace. Any continuous map $f : X \to K$, where K is a compact Hausdorff space, lifts uniquely to a continuous map $\beta X \to K$.*

Theorem A.2 (Martingale convergence). *Let X_k be a submartingale (i.e., $\mathrm{E}(X_{k+1} \mid X_1, \dots, X_k) \geq X_k$) and suppose that $\sup_k \mathrm{E}|X_k| < \infty$. Then, $X = \lim_{k\to\infty}$ exists almost surely and $\mathrm{E}|X| < \infty$.*

Proof. For example, [11], Theorem B.117, page 648, or [12], Theorem 1, page 508. □

Theorem A.3 (A strong law of large numbers for martingales). *Let $X_k = Z_1 + \cdots + Z_k$ be a martingale and let $\delta > 0$. If*

$$\sum_{k=1}^{\infty} \frac{\mathrm{E}(|Z_k|^2)}{k^{2\delta}} < \infty,$$

then $X_k / k^\delta \xrightarrow{a.s.} 0$.

Proof. For example, [2] or [12], Theorem 4, page 519. □

Theorem A.4 (Hoeffding–Azuma inequality). *Let X_k be a martingale and suppose that $|X_k - X_{k-1}| \leq c_k$ for all k. Then, for all $t > 0$ and $k \in \mathbb{N}$,*

$$\mathsf{P}\{X_n - X_0 \geq t\} \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right),$$

and

$$\mathsf{P}\bigl\{|X_n - X_0| \geq t\bigr\} \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right).$$

Proof. See [4], Theorem 2 and note around (2.18) on page 18, or [1]. □

Acknowledgements

The author is grateful to Matti Vihola for many stimulating discussions. This research was supported by the Academy of Finland (grant number 121855).

References

[1] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.* (2) **19** 357–367. MR0221571

[2] Chow, Y.S. (1967). On a strong law of large numbers for martingales. *Ann. Math. Statist.* **38** 610. MR0208648

[3] Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley. MR2239987

[4] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[5] Kujala, J.V. (2010). Obtaining the best value for money in adaptive sequential estimation. *J. Math. Psych.* **54** 475–480. MR2736122

[6] Kujala, J.V. (2012). Bayesian adaptive estimation: A theoretical review. In *Descriptive and Normative Approaches to Human Behavior* (E.N. Dzhafarov and L. Perry, eds.). *Adv. Ser. Math. Psychol.* **3** 123–159. Hackensack, NJ: World Sci. Publ. MR2905468

[7] Kujala, J.V. and Lukka, T.J. (2006). Bayesian adaptive estimation: The next dimension. *J. Math. Psych.* **50** 369–389. MR2239290

[8] Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005. MR0083936

[9] MacKay, D.J.C. (1992). Information-based objective functions for active data selection. *Neural Comput.* **4** 590–604.

[10] Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Comput.* **17** 1480–1507.

[11] Schervish, M.J. (1995). *Theory of Statistics*. *Springer Series in Statistics*. New York: Springer. MR1354146

[12] Shiryaev, A.N. (1996). *Probability*, 2nd ed. *Graduate Texts in Mathematics* **95**. New York: Springer. MR1368405

[13] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

Received March 2012 and revised May 2014