Almost the Best of Three Worlds: Risk, Consistency and Optional Stopping for the Switch Criterion in Nested Model Selection

Stéphanie van der Pas and Peter Grünwald

Leiden University and CWI, Amsterdam

Supplementary Material

We start by listing some well-known properties of exponential families which we will repeatedly use in the proofs. Then, in Section S4, we provide a sequence of technical lemmata that lead up to the proof of our main result, Theorem 1. Finally, in Section S5, we compare the switch distribution and criterion as defined here to the original switch distribution and criterion of Van Erven et al. (2012).

Additional Notation Our results will often involve displays involving several constants. The following abbreviation proves useful: when we write 'for positive constants $\vec{c}$, we have ...', we mean that there exist some $(c_1, \ldots, c_N) \in \mathbb{R}^N$, with $c_1, \ldots, c_N > 0$, such that ... holds; here N is left unspecified but it will always be clear from the application what N is. Further, for positive constants $\vec{b} = (b_1, b_2, b_3)$, we define $\texttt{small}_{\vec{b}}(n)$ as

$$\texttt{small}_{\vec{b}}(n) = \begin{cases} 1 & \text{if } n < b_1 \\ b_2 e^{-b_3 n} & \text{if } n \geq b_1, \end{cases}$$

and we frequently use the following fact. Suppose that $\mathcal{E}_1, \mathcal{E}_2, \ldots$ is a sequence of events such that $\mathbb{P}(\mathcal{E}_n) \leq \texttt{small}_{\vec{b}}(n)$. Then we also have, for any event $\mathcal{A}$, and for all n,

$$\mathbb{P}(\mathcal{A}, \mathcal{E}_n^c) \geq \mathbb{P}(\mathcal{A}) - \texttt{small}_{\vec{b}}(n), \tag{1}$$

as is immediate from $\mathbb{P}(\mathcal{A}, \mathcal{E}_n^c) = \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A}, \mathcal{E}_n) \geq \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{E}_n)$.

The components of a vector $\mu \in \mathbb{R}^n$ are given by $(\mu_1, \mu_2, \ldots, \mu_n)$. If the vector already has an index, we add a comma, for example $\mu_1 = (\mu_{1,1}, \mu_{1,2}, \ldots, \mu_{1,n})$. A sequence of vectors is denoted by $\mu^{(1)}, \mu^{(2)}, \ldots$.

S1 Definitions Concerning and Properties of Exponential Families

The following definitions and properties can all be found in the standard reference (Barndorff-Nielsen, 1978) and, less formally, in (Grünwald, 2007, Chapters 18 and 19).

A k-dimensional exponential family is a set of distributions on $\mathcal{X}$, which we invariably represent by the corresponding set of densities $\{p_\theta \mid \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$, such that any member p_θ can be written as

$$p_\theta(x) = \frac{1}{z(\theta)} e^{\theta^T \phi(x)} r(x) = e^{\theta^T \phi(x) - \psi(\theta)} r(x), \tag{S1.1}$$

where $\phi(x) = (\phi_1(x), \ldots, \phi_k(x))$ is a *sufficient statistic*, r is a non-negative function called the *carrier*, z the *partition function* and $\psi(\theta) = \log z(\theta)$. We assume the representation (S1) to be *minimal*, meaning that the components of $\phi(x)$ are linearly independent.

The parameterization in (S1.1) is referred to as the *canonical* or *natural parameterization*; we only consider families for which the set Θ is open and connected. Every exponential family can alternatively be parameterized in terms of its *mean-value parameterization*, where the family is parameterized by the mean $\mu = \mathbb{E}_\theta[\phi(X)]$, with μ taking values in $M \subset \mathbb{R}$, where μ as a function of θ is smooth and strictly increasing; as a consequence, the set M of mean-value parameters corresponding to an open and connected set Θ is itself also open and connected. Whenever for data $x_1, \ldots, x_n$, we have $\frac{1}{n} \sum_{i=1}^n \phi(x_i) \in M$, then the maximum likelihood is

uniquely achieved by the μ that is itself equal to this value,

$$\widehat{\mu}(x^n) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i). \tag{S1.2}$$

We thus define the maximum likelihood estimator (MLE) to be equal to (S1.2) whenever

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \in M. \tag{S1.3}$$

Since the result below which directly involves the MLE (Lemma 3) does not depend on its value for x^n with $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \notin M$, we can leave $\widehat{\mu}(x^n)$ undefined for such values. However, if we want to use the MLE as a 'sufficiently efficient' estimator as used in the statement of Theorem 1, we need to define $\widehat{\mu}(x^n)$ for such values in such a way that the 'sufficiently efficient property' (4.1) is satisfied. The following examples show various ways of constructing such sufficiently efficient estimators.

Example 1. [Sufficient Efficiency for MLE's for squared (standardized) error and Hellinger] For many full families such as the full (multivariate) Gaussians, Gamma and many others, (S1.3) holds μ-almost surely for each n, for all $\mu \in M$. If we compare two families $\mathcal{M}_0$ and $\mathcal{M}_1$ given in their mean-value parameterization with $M_0 \subset M_1$ where $\mathcal{M}_1$ is any such family, then the MLE is almost surely well-defined for M_1 and thus we need not worry about the issue indicated above. We can then take $\breve{\mu}_1 := \widehat{\mu}_1$ to be the MLE for $\mathcal{M}_1$. To get a sufficiently efficient estimator for M_0, we take $\breve{\mu}_0$ to be the projection of $\widehat{\mu}_1$ on the first m_0 coordinates (usually (S1.3) will still hold for $\mathcal{M}_0$ and then this $\breve{\mu}_0$ will also be the MLE for $\mathcal{M}_0$). This pair of estimators will be sufficiently efficient for (standardized) squared error and squared Hellinger distance, i.e. (4.1) holds for these three losses. To show this, note that from Proposition 1, Eq. (S1.7), we see that it is sufficient to show that (4.1) holds for the squared error loss. Since the j-th component of $\widehat{\mu}_1$ is equal to $n^{-1}\sum_{i=1}^{n}\phi_j(X_i)$ and $\mathbb{E}_{\mu_1}[n^{-1}\sum_{i=1}^{n}\phi_j(X_i)] = \mu_{1,j}$ and

$\mathrm{VAR}_{\mu_1}\left[n^{-1}\sum_{i=1}^{n}\phi_j(X_i)\right] = n^{-1}\mathrm{VAR}_{\mu_1}\left[\phi_j(X_1)\right]$, it suffices to show that

$$\sup_{\mu_1 \in M_1'} \sup_{j=1,\ldots,m_1} \mathrm{VAR}_{\mu_1}\left[\phi_j(X_1)\right] = O\left(1\right),$$

which is indeed the case since M_1' is a CINECSI set, so that the variance of all ϕ_j's is uniformly bounded on M_1' (Barndorff-Nielsen, 1978).

Example 2. [Other sufficiently efficient estimators for squared (standardized) error and Hellinger] For models such as the Bernoulli or multinomial, (S1.3) may fail to hold with positive probability: the full Bernoulli exponential family does not contain the distributions with $P(X_1 = 1) = 1$ and $P(X_1 = 0) = 1$, so if after n examples, only zeros or only ones have been observed, the MLE is undefined. We can then go either of three ways. The first way, which we shall not pursue in detail here, is to work with so-called 'aggregate' exponential families, which are extensions of full families to their limit points. For models with finite support (such as the multinomial) these are well-defined (Barndorff-Nielsen, 1978, page 154–158) and then the MLE's for these extended families are almost surely well-defined again, and the MLE's are sufficiently efficient by the same reasoning as above. Another approach that works in some cases (e.g. multinomial) is to take $\breve{\mu}_1$ to be a truncated MLE, that, at sample size n, maps X^n to the MLE within some CINECSI subset $M_1^{(n)}$ of M_1, where $M_1^{(n)}$ converges to M_1 as n increases in the sense that $\sup_{\mu \in M_1^{(n)}, \mu' \in M_1 \setminus M_1^{(n)}} \|\mu - \mu'\|_2^2 = O(1/n)$. The resulting truncated MLE, and its projection on M_0 (usually itself a truncated MLE) will then again be sufficiently efficient. This approach also works if the models $\mathcal{M}_0$ and $\mathcal{M}_1$ are not full but restricted families to begin with. For full families though, a more elegant approach than truncating MLE's is to work with Bayesian posterior MAP estimates with conjugate priors. For steep exponential families (nearly all families one encounters in practice are steep), one can always find conjugate priors such that the Bayes MAP estimates based on these priors exist and take a value in M_1 almost surely

(Grünwald and de Rooij, 2005). They then take the form $\breve{\mu}_1 = \sum_{i=1}^{n}(\phi(X_i) + \lambda_0\mu_1^\circ)/(n+\lambda_0)$, where $\lambda_0 > 0$ and $\mu_1^\circ \in M_1$ are determined by the prior. $\breve{\mu}_0$ can then again be taken to be the projection of $\breve{\mu}_1$ onto M_0. Under the assumption that μ_1 is contained in a CINECSI set M_1', one can now again show, using the same arguments as in Example 1, that such estimators are sufficiently efficient for squared (standardized) error and Hellinger loss.

Example 3. [Sufficient Efficiency for Rényi and KL divergence] As is well-known, for the multivariate Gaussian model with fixed covariance matrix, the squared error risk and KL divergence are identical up to constant factors, so the unrestricted MLE's will still be sufficiently efficient for KL divergence. For other models, though, the MLE will not always be sufficiently efficient. For example, with the Bernoulli model and other models with finite support, to make the unrestricted MLE's well-defined, we would have to extend the family to its boundary points as indicated in Example 1. Since, however, for any $0 < \mu < 1$ and $\mu' = 0$, the KL divergence $D(\mu\|\mu') = \infty$ and $\mathbb{P}_\mu(\hat{\mu}(X^n) = \mu') > 0$, the unrestricted MLE in the full Bernoulli model including the boundaries will have infinite risk and thus will not be sufficiently efficient. The MAP estimators tend to behave better though: Grünwald and de Rooij (2005) implicitly show that for 1-dimensional families, under weak conditions on the family (Condition 1 underneath Theorem 1 in their paper) — which were shown to hold for a number of families such as Bernoulli, Poisson, geometric — sufficient efficiency for the KL divergence still holds for MAP estimators of the form above. We conjecture that a similar result can be shown for multidimensional families, but will not attempt to do so here.

A standard property of exponential families says that, for any $\mu \in M$, any distribution $\mathbb{Q}$ on $\mathcal{X}$ with $\mathbb{E}_{X\sim\mathbb{Q}}[\phi(X)] = \mu$, any $\mu' \in M$, we have

$$\mathbb{E}_{X\sim\mathbb{Q}}\left[\log \frac{p_\mu(X)}{p_{\mu'}(X)}\right] = \mathbb{E}_{X\sim\mathbb{P}_\mu}\left[\log \frac{p_\mu(X)}{p_{\mu'}(X)}\right] = D(\mu\|\mu'), \tag{S1.4}$$

the final equality being just the definition of $D(\cdot\|\cdot)$. Now fix an arbitry sample x^n. By taking $\mathbb{Q}$ to be the empirical distribution on $\mathcal{X}$ corresponding to sample x^n, it follows from (S1.4) that if $\widehat{\mu}(x^n) \in M$ then also the following relationship holds for any $\mu' \in M$:

$$\frac{1}{n} \log \frac{p_{\widehat{\mu}(x^n)}(x^n)}{p_{\mu'}(x^n)} = D(\widehat{\mu}(x^n)\|\mu'). \tag{S1.5}$$

(S1.4) and (S1.5) are a direct consequence of the sufficiency of $\widehat{\mu}_1(X^n)$, and folklore among information theorists. For a proof of (S1.4) and more details on (S1.5), see e.g. (Grünwald, 2007, Chapter 19), who calls this the *robustness property* of the KL divergence for exponential families.

We are now in a position to prove Proposition 1, which we repeat for convenience.

Proposition 1 Let M, a product of open intervals, be the mean-value parameter space of an exponential family, and let M' be a CINECSI subset of M. Then there exist positive constants $\vec{c}$ such that for all $\mu, \mu' \in M'$,

$$c_1\|\mu' - \mu\|_2^2 \leq c_2 \cdot d_{ST}(\mu'\|\mu) \leq d_{H^2}(\mu', \mu) \leq d_R(\mu', \mu) \leq D(\mu'\|\mu) \leq c_3\|\mu' - \mu\|_2^2. \tag{S1.6}$$

and for all $\mu' \in M', \mu \in M$ (i.e. μ is now not restricted to lie in M'),

$$d_{H^2}(\mu', \mu) \leq c_4\|\mu' - \mu\|_2^2 \leq c_5 \cdot d_{ST}(\mu'\|\mu) \leq c_6\|\mu' - \mu\|_2^2. \tag{S1.7}$$

Proof. We start with (S1.6). The third and fourth inequality are immediate by using $-\log x \geq 1 - x$ and Jensen's inequality, respectively. From standard properties of Fisher information for exponential families (Barndorff-Nielsen, 1978) we have that, for any CINECSI (hence compact and bounded away from the boundaries of M) subset M' of M, there exists positive $\vec{C}$ with

$$0 < C_1 = \inf_{\mu \in M'} \det I(\mu) < \sup_{\mu \in M'} \det I(\mu) = C_2 < \infty, \tag{S1.8}$$

from which we infer that for all $\mu' \in M'$, $\mu, \mu'' \in \mathbb{R}^m$,

$$C_3\|\mu - \mu''\|_2^2 \leq (\mu - \mu'')^T I(\mu')(\mu - \mu'') \leq C_4\|\mu - \mu''\|_2^2, \tag{S1.9}$$

for some $0 < C_3 \leq C_4 < \infty$. Using (S1.9), the first inequality is immediate, and the final inequality follows straightforwardly from a second-order Taylor approximation of KL divergence as in (Grünwald, 2007, Chapter 4). It only remains to establish the second inequality. Now, since M' is CINECSI and hence compact the fifth (rightmost) inequality implies that there is a $C_5 < \infty$ such that $\sup_{\mu,\mu' \in M'} D(\mu'\|\mu) < C_5$ and hence, via the fourth inequality, that $\sup_{\mu,\mu' \in M'} d_R(\mu', \mu) < C_5$. Equality (3.2) now implies that there is a C_6 such that

$$\sup_{\mu,\mu' \in M'} d_R(\mu', \mu)/d_{H^2}(\mu', \mu) < C_6. \tag{S1.10}$$

Using again (S1.8), a second order Taylor approximation as in Van Erven and Harremoës (2014) now gives that for some constant $C_7 > 0$, $\|\mu - \mu'\|_2^2 \leq C_7 d_R(\mu', \mu)$ for all $\mu, \mu' \in M'$. The first result, (S1.6), now follows upon combining this with (S1.10).

As to (S1.7), the second and third inequality are immediate from (S1.9). For the first inequality, note that, since M' is CINECSI and we assume M to be a product of open intervals, there must exist another CINECSI subset M'' of M strictly containing M' such that $\inf_{\mu' \in M', \mu \in M \setminus M''} \|\mu' - \mu\|_2^2 = \delta$ for some $\delta > 0$. We now distinguish between μ in (S1.7) being an element of (a) M'' or (b) $M \setminus M''$. For case (a) (S1.6), with M'' in the role of M', gives that there is a constant C_8 such that for all $\mu \in M''$, $d_{H^2}(\mu', \mu) \leq C_8\|\mu' - \mu\|_2^2$. For case (b), $\mu \in M \setminus M''$, we have $\|\mu' - \mu\|_2^2 \geq \delta$ and, using that squared Hellinger distance for any pair of distributions is bounded by 2, we have $d_{H^2}(\mu', \mu) \leq (2/\delta)\|\mu' - \mu\|_2^2$. Thus, by taking $c_4 = \max\{C_8, 2/\delta\}$, case (a) and (b) together establish the first inequality in (S1.7). □

S2 Preparation for Proof of Main Result: Results on Large Deviations

Let $\mathcal{M}_1$ and M_1 be as in Theorem 1. For the following result, Lemma 1, we set $\widehat{\mu}_1'(X^n) := n^{-1}\sum \phi(X_i)$, so that $\widehat{\mu}_1'(X^n) = \widehat{\mu}_1(X^n)$ whenever $n^{-1}\sum \phi(X_i) \in M_1$. It is essentially a multidimensional extension of a standard information-theoretic result, with KL divergence replaced by squared error loss. This standard result states the following: whenever $\mathcal{M}_1$ is a single-parameter exponential family (that is, $m_1 = 1$), then for any $\mu \in M_1$, all $a, a' > 0$ with $\mu + a \in M_1$, $\mu - a' \in M_1$,

$$\mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \geq \mu + a) \leq e^{-nD(\mu+a\|\mu)}. \;\; ; \;\; \mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \leq \mu - a') \leq e^{-nD(\mu-a'\|\mu)}. \tag{S2.1}$$

For a simple proof, see (Grünwald, 2007, Section 19.4.2); for discussion see (Csiszár, 1984) — the latter reference gives a multidimensional extension of (S2.1) but of a very different kind than Lemma 1 below. To prepare for the lemma, let $\mathcal{M}_1$ and M_1 be as in Theorem 1 and, for any $\mu \in M_1$ and any $\vec{a}, \vec{b} \in \mathbb{R}^{m_1}_{>0}$, define the ℓ_∞-rectangle $R_\infty(\mu, \vec{a}, \vec{b}) = \{\mu' \in \mathbb{R}^{m_1} : \forall j = 1, \ldots, m_1, -b_j \leq \mu_j' - \mu_j \leq a_j\}$.

Lemma 1. *Let $\mathcal{M}_1$ and M_1 be as in Theorem 1 and fix an arbitrary CINECSI subset M_1' of M_1. Then there is a $c > 0$ (depending on M_1') such that, for all $\mu \in M_1$, all n, all $\vec{a}, \vec{b} \in \mathbb{R}^{m_1}_{>0}$ such that $R_\infty(\mu, \vec{a}, \vec{b}) \subset M_1'$,*

$$\mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \notin R_\infty(\mu, \vec{a}, \vec{b})) \leq 2m_1 e^{-nc\cdot(\min_j \min\{a_j, b_j\})^2}. \tag{S2.2}$$

Proof. For $j = 1, \ldots, m_1$, $d \in \mathbb{R}$, let $\vec{e_j}$ represent the jth standard basis vector, such that $\mu + d\vec{e_j} = (\mu_1, \ldots, \mu_{j-1}, \mu_j + d, \mu_{j+1}, \ldots, \mu_{m_1})$, and let $D_{\mu+d\vec{e_j}} := D(\mu + d\vec{e_j}\|\mu)$. We now have that there exist constants $c_{a,1}, \ldots, c_{a,m_1}, c_{b,1}, \ldots, c_{b,m_1} > 0$ such that for

$c := \min\{c_{a,1}, \ldots, c_{a,m_1}, c_{b,1}, \ldots, c_{b,m_1}\}$, all n,

$$\begin{aligned}\mathbb{P}_\mu(\widehat{\mu}_1(X^n) \notin R_\infty(\mu, \vec{a}, \vec{b})) &\leq \sum_{j=1}^{m_1} \mathbb{P}_\mu(\widehat{\mu}_{1,j}(X_n) \geq \mu_j + a_j) + \sum_{j=1}^{m_1} \mathbb{P}_\mu(\widehat{\mu}_{1,j}(X^n) \leq \mu_j - b_j) \\ &\leq \sum_{j=1}^{m_1} \left(e^{-nD_{\mu + a_j \vec{e_j}}} + e^{-nD_{\mu - b_j \vec{e_j}}}\right) \leq \sum_{j=1}^{m_1} \left(e^{-nc_{a,j}a_j^2} + e^{-nc_{b,j}b_j^2}\right) \\ &\leq 2m_1 e^{-nc \cdot (\min_j \min\{a_j, b_j\})^2},\end{aligned}$$

Here the first inequality follows from the union bound, and the second follows by applying, for each of the $2m_1$ terms, (S2.1) above to the one-dimensional exponential sub-family $\{p_\mu \mid \mu \in M_1 \cap \{\mu : \mu = \mu + d\vec{e_j} \text{ for some } d \in \mathbb{R}\}\}$. The third follows by Proposition 1 together with the equivalence of the ℓ_2 and sup norms on $\mathbb{R}^{m_1}$, and the final inequality is immediate. □

Lemma 2. *Under conditions and notations as in Theorem 1, let μ, μ' be elements of M_1 and suppose $X^N = (X_{n_1}, \ldots, X_{n_2})$ is a sequence of i.i.d. observations of length N from p_μ. Then, for any $A \in \mathbb{R}$:*

$$\mathbb{P}_\mu\left(\log \frac{p_\mu(X^N)}{p_{\mu'}(X^N)} < A\right) \leq e^{\frac{1}{2}A} e^{-\frac{N}{2} d_R(\mu', \mu)}. \tag{S2.3}$$

Proof. For any A, by Markov's inequality:

$$\begin{aligned}\mathbb{P}_\mu\left(\log \frac{p_\mu(X^N)}{p_{\mu'}(X^N)} < A\right) &= \mathbb{P}_\mu\left(\left(\frac{p_{\mu'}(X^N)}{p_\mu(X^N)}\right)^{\frac{1}{2}} > e^{-\frac{1}{2}A}\right) \leq e^{\frac{1}{2}A} \mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X^N)}{p_\mu(X^N)}\right)^{\frac{1}{2}}\right] \\ &= e^{\frac{1}{2}A}\left(\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)^N = e^{\frac{1}{2}A} e^{\log\left(\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)^N} \\ &= e^{\frac{1}{2}A} e^{-\frac{N}{2}\left(-\frac{1}{1-1/2}\log \mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)} = e^{\frac{1}{2}A} e^{-\frac{N}{2} d_R(\mu, \mu')}.\end{aligned} \tag{S2.4}$$

□

Proposition 2. *Let $\mathcal{M}_0, \mathcal{M}_1, M_0, M_1$ be as in Theorem 1 and let M_1' be a CINECSI subset of M_1. Then there exists another, larger, CINECSI subset M_1'' of M_1 and positive constants $\vec{b}$ such*

that M_1' is itself a CINECSI subset of M_1'' and for both $j \in \{0,1\}$, the ML estimator $\widehat{\mu}_j(x^n)$ satisfies

$$\sup_{\mu \in M_1'} P_\mu(\widehat{\mu}_j(X^n) \notin M_1'') \leq \texttt{small}_{\vec{b}}(n).$$

Proof. M_1 can be written as in (3.7), and hence we can define a set

$$M_1'' = [\zeta_{1,1}^*, \eta_{1,1}^*] \times \ldots \times [\zeta_{1,m_1}^*, \eta_{1,m_1}^*]$$

for values $\zeta_{1,j}^*, \eta_{1,j}^* \in \mathbb{R}$ such that M_1'' is a CINECSI subset of M_1. Since M_1' is connected with compact closure in interior of M_1 and M_1'' is a subset of M_1, we can choose the $\zeta_{1,j}^*, \eta_{1,j}^* \in \mathbb{R}$ such that M_1' is itself a CINECSI subset of M_1''. Since M_1' is connected and its closure is in the interior of M_1'' which is itself compact, it follows that there is some $\delta > 0$ such that, for all $\mu_1' \in M_1', \mu_1'' \notin M_1''$, all $j \in \{1, \ldots, m_1\}$, it holds $|\mu_{1,j}' - \mu_{1,j}''| > \delta$. It now follows from Lemma 1, applied with $\vec{a}$ chosen such that $R_\infty(\mu', \vec{a}) = M_1''$, that for every $\mu' \in M_1'$, all n,

$$\mathbb{P}_{\mu'}\left(\widehat{\mu}_1(X^n) \notin M_1''\right) \leq C_1 e^{-nC_2\delta^2}$$

for some constants C_1, C_2. Here we used that by construction, each entry of $\vec{a}$ must be at least as large as δ. Since $\widehat{\mu}_{1,j}(x^n)$ and $\widehat{\mu}_{0,j}(x^n)$ coincide for $0 < j \leq m_0$ and $\widehat{\mu}_{0,j}(x^n)$ is constant for $m_0 < j \leq m_1$, the result follows for $\widehat{\mu}_0(x^n)$ as well. □

S3 Preparation for Proof of Main Result: Results on Bayes Factor Model Selection

Lemma 3. *Let $\mathcal{M}_0, \mathcal{M}_1, M_0, M_1$ be as in Theorem 1 and let, for $j \in \{0,1\}$, M_j' be a CINECSI subset of M_j. For both $j \in \{0,1\}$, there exist positive constants $\vec{c}, \vec{b}$ such that for all $\mu_1 \in M_1'$,*

$$c_1 \leq n^{-m_j/2} \cdot \frac{p_{\widehat{\mu}_j(X^n)}(X^n)}{p_{B,j}(X^n)} \leq c_2, \tag{S3.1}$$

with $\mathbb{P}_{\mu_1}$-probability at least $1 - \mathtt{small}_{\vec{b}}(n)$.

Proof. For a Bayesian marginal distribution p_B defined relative to m-dimensional exponential family $\mathcal{M}$ given in its mean-value parameterization M, with a prior $\omega(\cdot)$ that is continuous and strictly positive on M, we have as a consequence of the familiar Laplace approximation of the Bayesian marginal distribution of exponential famlies as in e.g. (Kass and Raftery, 1995),

$$p_B(x^n) \sim \left(\frac{n}{2\pi}\right)^{-m/2} \cdot \frac{\omega(\widehat{\mu}(x^n))}{\sqrt{\det I(\widehat{\mu}(x^n))}} p_{\widehat{\mu}(x^n)}(x^n).$$

As shown in Theorem 8.1 in (Grünwald, 2007), this statement holds uniformly for all sequences x^n with ML estimators in any fixed *CINECSI* subset M' of M. By compactness of M', and by positive definiteness and continuity of Fisher information for exponential families, the quantity $\omega(\widehat{\mu})/\sqrt{\det I(\widehat{\mu})}$ will be bounded away from zero and infinity on such sequences, and, applying the result to both the families $\mathcal{M}_0$ and $\mathcal{M}_1$ it follows that there exist $c_1, c_2 > 0$ such that for all n larger than some n_0, uniformly for all sequences x^n with $\widehat{\mu}_j(x^n) \in M'_j$, we have:

$$c_1 \leq n^{-m_j/2} \cdot \frac{p_{\widehat{\mu}_j(x^n)}(x^n)}{p_{B,j}(x^n)} \leq c_2. \tag{S3.2}$$

The result now follows by combining this statement with Proposition 2. □

Lemma 4. *Let $\mathcal{M}_0, \mathcal{M}_1, M_0, M_1$ and the Bayesian marginal distribution $p_{B,0}$ be as in Theorem 1. Let M'_1 be a CINECSI subset of M_1. Then there exist positive constants $\vec{c}$ and $\vec{b}$ such that for all n, all $\mu_1 \in M'_1$, all $A \in \mathbb{R}$,*

$$\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A\right) \leq n^{m_1/2} \cdot c_1 \cdot e^{\frac{1}{2}c_2 A} e^{-\frac{n}{2}c_3 \|\mu_1 - \mu_0\|_2^2} + \mathtt{small}_{\vec{b}}(n),$$

where for each μ_1, $\mu_0 = \Pi_0(\mu_1)$ as in (3.8).

Proof. Fix constants C_1, C_2 such that they are smaller and larger respectively than the constants c_1, c_2 from Lemma 3 and define

$$\mathcal{E}_n = \left\{X^n : C_1 \leq n^{-m_1/2} \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{B,1}(X^n)} \leq C_2\right\}.$$

Using Lemma 3, we have that there exists positive $\vec{b}$ such that for all $A \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A\right) \\
=&\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n\right) + \mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n^c\right) \\
\leq&\mathbb{P}_{\mu_1}\left(\log \frac{C_2^{-1} n^{-m_1/2} p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n\right) + \texttt{small}_{\vec{b}}(n) \\
\leq&\mathbb{P}_{\mu_1}\left(\log \frac{C_2^{-1} n^{-m_1/2} p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A\right) + \texttt{small}_{\vec{b}}(n) \\
=&\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A + \log C_2 n^{m_1/2}\right) + \texttt{small}_{\vec{b}}(n). \qquad \text{(S3.3)}
\end{aligned}$$

To bound this probability further, we need to relate $p_{B,0}$ to $p_{B',0}$, the Bayesian marginal likelihood under model M_0 under a prior with support restricted to a compact set M_0'. To define M_0', note first that there must exist a CINECSI subset, say M_1'', of M_1 such that M_1' is itself a CINECSI subset of M_1''. Take any such M_1'' and let M_0' be the closure of $M_1'' \cap M_0$. Given ω, the prior density on $\Pi'(M_0)$ used in the definition of $p_{B,0}$, define $\omega'(\nu) = \omega(\nu)/\int_{\nu \in \Pi'(M_0')} \omega(\nu) d\nu$ as the prior density restricted to and normalized on $\Pi'(M_0')$ and let $p_{B',0}$ be the corresponding Bayesian marginal density on X^n.

To continue bounding (S3.3), define

$$\mathcal{E}_n' = \left\{X^n : C_3 \leq n^{-m_0/2} \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{B,0}(X^n)} \leq C_4 \text{ and } \; C_3 \leq n^{-m_0/2} \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{B',0}(X^n)} \leq C_4\right\},$$

with C_3 and C_4 smaller and larger respectively than the constants c_1 and c_2 resulting from Lemma 3 (note that Lemma 3 can be applied to $p_{B',0}$ as well, by taking M_0 in that lemma to be the interior of M_0' as defined here). Set $C_5 > C_4/C_3$, and note that for any $A_1 \in \mathbb{R}$,

abbreviating $\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{C_5 p_{B',0}(X^n)} < A_1\right)$ to p^*, we have

$$\begin{aligned}
&\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1\right)\\
=&\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1,\ \frac{p_{B_0}(X^n)}{p_{B',0}(X^n)} < C_5\right) + \mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1,\ \frac{p_{B_0}(X^n)}{p_{B',0}(X^n)} \geq C_5\right)\\
\leq&\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{C_5 p_{B',0}(X^n)} < A_1\right) + \mathbb{P}_{\mu_1}\left(p_{B,0}(X^n) \geq C_5 p_{B',0}(X^n)\right)\\
=&p^* + \mathbb{P}_{\mu_1}\left(p_{B,0}(X^n) \geq C_5 p_{B',0}(X^n)\right)\\
\leq&p^* + \mathbb{P}_{\mu_1}\left(\frac{p_{B,0}(X^n)}{p_{B',0}(X^n)} \geq C_5, \mathcal{E}'_n\right) + \mathbb{P}_{\mu_1}\left(\frac{p_{B,0}(X^n)}{p_{B',0}(X^n)} \geq C_5, (\mathcal{E}'_n)^c\right)\\
\leq&p^* + 0 + \mathtt{small}_{\vec{b}}(n). \qquad \text{(S3.4)}
\end{aligned}$$

Now it only remains to bound p^*. To this end, let

$$C_6 := \int_{\nu \in \Pi'(M'_0)} \sqrt{\omega(\nu)} d\nu. \qquad \text{(S3.5)}$$

Since M'_0 has compact closure in the interior of M_0 and we are assuming that ω has full support on M_0, we have that $C_6 < \infty$.

Now using Markov's inequality as in the proof of Lemma 2, that is, the first line of (S2.4) with $p_{B',0}$ in the role of $p_{\mu'}$, gives, for any $A_2 \in \mathbb{R}$,

$$\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B',0}(X^n)} < A_2\right) \leq e^{\frac{1}{2}A_2} \mathbb{E}_{\mu_1}\left[\left(\frac{p_{B',0}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right]. \qquad \text{(S3.6)}$$

The expectation on the right can be further bounded, defining $\omega'' = \sqrt{\omega}/C_6$ and noting that ω'' is a probability density, as

$$\begin{aligned}
\mathbb{E}_{\mu_1}\left[\left(\frac{p_{B',0}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] &\leq \mathbb{E}_{\mu_1}\left[\left(\frac{\int_{\nu \in \Pi'(M'_0)} \omega(\nu)^{1/2} p_\nu(X^n)^{1/2} d\nu}{p_{\mu_1}(X^n)^{1/2}}\right)\right]\\
&= C_6 \cdot \mathbb{E}_{\mu \sim \omega''} \mathbb{E}_{\mu_1}\left[\left(\frac{p_\mu(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] \leq C_6 \cdot \mathbb{E}_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right],
\end{aligned}$$

where $\mu^\circ \in M'_0$ achieves the supremum of $E_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right]$ within M'_0. By compactness of M'_0 and continuity, this supremum is achieved. The final term can be rewritten, following the

same steps as in the second and third line of (S2.4), as

$$\mathbb{E}_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] = e^{-\frac{n}{2}d_R(\mu_1,\mu^\circ)}. \tag{S3.7}$$

Since M_0' and M_1' are both CINECSI, it now follows from Proposition 1 that for some fixed $C_7 > 0$,

$$d_R(\mu_1,\mu^\circ) \geq C_7\|\mu_1 - \mu^\circ\|_2^2 \geq C_7\|\mu_1 - \mu_0\|_2^2, \tag{S3.8}$$

where the latter inequality follows by the definition of $\mu_0 = \Pi_0(\mu_1)$, see the explanation below (3.8). Combining (S3.6), (S3.7) and (S3.8), we have thus shown that for all n, all $\mu_1 \in M_1$, all $A_2 \in \mathbb{R}$,

$$\mathbb{P}_{\mu_1}\left(\log\frac{p_{\mu_1}(X^n)}{p_{B',0}(X^n)} < A_2\right) \leq C_6 e^{\frac{1}{2}A''} e^{-\frac{n}{2}C_7\|\mu_1-\mu_0\|_2^2}. \tag{S3.9}$$

The result now follows by combining (S3.3), (S3.4) and (S3.9). □

S4 Proof of Main Result, Theorem 1

Proof Idea The proof is based on analyzing what happens if $X_1, X_2, \ldots, X_n$ are sampled from $p_{\mu_1^{(n)}}$, where $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ are a sequence of parameters in M_1'. We consider three regimes, depending on how fast (if at all) $\mu_1^{(n)}$ converges to $\mu_0^{(n)}$ as $n \to \infty$. Here $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is the projection of $\mu^{(1)}$ onto M_0, i.e. the distribution in M_0 defined, for each n, as in (3.8), with μ_1 and μ_0 in the role of $\mu_1^{(n)}$ and $\mu_0^{(n)}$, respectively. Our regimes are defined in terms of the function f given by

$$f(n) := \frac{\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2}{\frac{\log\log n}{n}} = \frac{n\cdot\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2}{\log\log n}, \tag{S4.1}$$

which indicates how fast $d_{SQ}(\mu_1^{(n)}, \mu_0^{(n)})$ grows relative to the best possible rate $(\log\log n)/n$. We fix appropriate constants Γ_1 and Γ_2, and we distinguish, for all n with $\Gamma_2 \log n \geq \Gamma_1$, the

cases:

$$f(n) \in \begin{cases} [0, \Gamma_1] & \text{Case 1} \\ [\Gamma_1, \Gamma_2 \log n] & \text{Case 2 (Theorem 4)} \\ [\Gamma_2 \log n, \infty] & \text{Case 3 (Theorem 3).} \end{cases}$$

For Case 1, the rate is easily seen to be upper bounded by $O((\log\log n)/n)$, as shown inside the proof of Theorem 1. In Case 2, Theorem 4 establishes that the probability that model $\mathcal{M}_0$ is chosen is at most of order $1/(\log n)$, which, as shown inside the proof of Theorem 1, again implies an upper-bound on the rate-of convergence of $O((\log\log n)/n)$. Theorem 3 shows that in Case 3, which includes the case that $\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2$ does not converge at all, the probability that model $\mathcal{M}_0$ is chosen is at most of order $1/n$, which, as again shown inside the proof of Theorem 1, again implies an upper-bound on the rate-of convergence of $O((\log\log n)/n)$.

The two theorems take into account that $\mu_1^{(n)}$ is not just a fixed function of n, but may in reality be chosen by nature in a worst-case manner, and that $f(n)$ may actually fluctuate between regions for different n. Combining these two results, we finally prove the main theorem, Theorem 1.

Theorem 3. *Let M_0, M_1, M_1' and $p_{sw,1}(x^n)$ be as in Theorem 1. Then there exist positive constants $\vec{b}, \vec{c}$ such that for all $\mu_1 \in M_1'$, all n,*

$$\mathbb{P}_{\mu_1}\left(\delta_{sw}(X^n) = 0\right) \leq c_1 \cdot n^{m_1/2} \cdot e^{-c_2 n \|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2} + \texttt{small}_{\vec{b}}(n), \tag{S4.2}$$

where $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is as in (3.8). As a consequence, with $\Gamma_2 := c_2^{-1}(1 + m_1/2)$, we have the following: for every sequence $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ with $f(n)$ as in (S4.1) larger than $\Gamma_2 \log n$, we have

$$\mathbb{P}_{\mu_1^{(n)}}\left(\delta_{sw}(X^n) = 0\right) \leq \frac{c_1}{n} + \texttt{small}_{\vec{b}}(n).$$

Proof. We can bound the probability of selecting the simple model by:

$$\mathbb{P}_{\mu_1^{(n)}}\left(\delta_{\mathrm{sw}}(X^n)=0\right)=\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)}\leq 1\right)=\mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\infty}\pi(2^i)\bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)}\leq 1\right)$$
$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\pi(1)p_{B,1}(X^n)}{p_{B,0}(X^n)}\leq 1\right).$$

Now (S4.2) follows directly by applying Lemma 4 to the rightmost probability. For the second part, set $\Gamma_2 = c_2^{-1}(1+m_1/2)$. By assumption $f(n) > \Gamma_2 \log n$, we have $\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2 > \Gamma_2(\log n)(\log\log n)/n$. Applying (S4.2) now gives the desired result. □

Theorem 4. *Let f be as in (S4.1) and M_1' be as in Theorem 1. For any $\gamma > 0$, there exist constants $\Gamma_1, \Gamma_3 > 0$ such that, for every sequence $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ of elements of M_1' with for all n, $f(n) > \Gamma_1$, we have*

$$\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)}\leq \gamma\right)\leq \frac{\Gamma_3}{\log n}. \tag{S4.3}$$

In particular, by taking $\gamma = 1$, we have

$$\mathbb{P}_{\mu_1^{(n)}}(\delta_{sw}(X^n)=0)\leq \frac{\Gamma_3}{\log n}.$$

The probabilities thus converge uniformly at rate $O(1/(\log n))$ for all such sequences $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$.

Proof. We specify Γ_1 later. By assumption, we have $\pi(2^i) \gtrsim (\log n)^{-\kappa}$ for $i \in \{0, \ldots, \lfloor \log_2 n \rfloor\}$. We can restrict our attention to the strategy that switches to the complex model at the penultimate switching index, due to the following inequality: for any fixed γ, there exist positive constants $\vec{C}$ such that for all large n:

$$\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)}\leq \gamma\right)\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\lfloor \log_2 n\rfloor}\pi(2^i)\bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)}\leq \gamma\right)$$
$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\lfloor \log_2 n\rfloor}\bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)}\leq C_1(\log n)^{\kappa}\right)$$
$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\bar{p}_{2^{\lfloor \log_2 n\rfloor-1}}(X^n)}{p_{B,0}(X^n)}\leq C_1(\log n)^{\kappa}\right)$$
$$= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{\bar{p}_{2^{\lfloor \log_2 n\rfloor-1}}(X^n)}{p_{B,0}(X^n)}\leq \kappa\log\log n + C_2\right). \tag{S4.4}$$

For the remainder of this proof, we will denote the penultimate switching index by n^*, that is: $n^* = 2^{\lfloor \log_2 n \rfloor - 1}$. Now apply Lemma 3 twice, which gives that there exist C_3, C_4 such that, with probability at least $1 - \texttt{small}_{\vec{b}}(n)$,

$$\begin{aligned}
\log \bar{p}_{n^*}(X^n) &= \log p_{B,0}(X^{n^*}) + \log p_{B,1}(X^n | X^{n^*}) = \\
&= \log p_{B,0}(X^{n^*}) + \log p_{B,1}(X^n) - \log p_{B,1}(X^{n^*}) \\
&\geq \log p_{B,0}(X^{n^*}) + \log p_{\widehat{\mu}_1(X^n)}(X^n) - \log p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*}) + \frac{m_1}{2} \log \frac{n^*}{n} - C_3 \\
&\geq \log p_{B,0}(X^{n^*}) + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_4, \qquad \text{(S4.5)}
\end{aligned}$$

where we used that $\log \frac{n^*}{n}$ is of the order of a constant, because n^* is between $\frac{n}{4}$ and $\frac{n}{2}$. From this, applying again Lemma 3 twice, it follows that there exists $\vec{b}$ and C_5, C_6 such that for all n, with probability at least $1 - \texttt{small}_{\vec{b}}(n)$,

$$\begin{aligned}
\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} &\geq \log \frac{p_{B,0}(X^{n^*})}{p_{B,0}(X^n)} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_4 \\
&= -\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{m_0}{2} \log \frac{n^*}{n} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_5 \\
&\geq -\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_6 \qquad \text{(S4.6)}
\end{aligned}$$

where we again used that $\log \frac{n^*}{n}$ can be bounded by constants. Let $\mathcal{B}_n$ be the event that (S4.6) holds. By (S4.4) and (S4.6), for all large n, all $\beta \geq 1$,

$$\begin{aligned}
&\mathbb{P}_{\mu_1^{(n)}} \left(\frac{p_{\text{sw},1}(X^n)}{p_{B,0}(X^n)} \leq \gamma \right) \leq \mathbb{P}_{\mu_1^{(n)}} \left(\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} \leq \kappa \log\log n + C_2 \right) \\
\leq &\mathbb{P}_{\mu_1^{(n)}} \left(\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} \leq \kappa \log\log n + C_2, \mathcal{B}_n \right) + \mathbb{P}_{\mu_1^{(n)}} \left(\mathcal{B}_n^c \right) \\
\leq &\mathbb{P}_{\mu_1^{(n)}} \left(-\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_6 \leq \kappa \log\log n + C_2 \right) + \texttt{small}_{\vec{b}}(n) \\
= &\mathbb{P}_{\mu_1^{(n)}} \left(\mathcal{E}_n^{(1)} \right) + \texttt{small}_{\vec{b}}(n) \leq \mathbb{P}_{\mu_1^{(n)}} \left(\mathcal{E}_n^{(\beta)} \right) + \texttt{small}_{\vec{b}}(n), \qquad \text{(S4.7)}
\end{aligned}$$

where we defined

$$\mathcal{E}_n^{(\beta)} = \left\{ \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} \cdot \frac{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})}{p_{\widehat{\mu}_0(X^n)}(X^n)} \leq A_n^{(\beta)} \right\} \qquad \text{(S4.8)}$$

and, for $\beta \geq 1$, we set $A_n^{(\beta)} = \beta\kappa \log\log n + C_2 - C_6$.

Below, if a sample is split up into two parts $x_1, \ldots, x_{n^*}$ and $x_{n^*+1}, \ldots, x_n$, these partial samples will be referred to as x^{n^*} and $x^{>n^*}$ respectively. We also suppress in our notation the dependency of A_n, $\mathcal{E}_n$ and $\mathcal{D}_{j,n}$ as defined below on β; all results below hold, with the same constants, for any $\beta \geq 1$.

We will now bound the right-hand side of (S4.7) further. Define the events

$$\mathcal{D}_{1,n} = \left\{ \log \frac{p_{\mu_1^{(n)}}(x^n)}{p_{\mu_1^{(n)}}(x^{n^*})} \leq \log \frac{p_{\widehat{\mu}_1(X^n)}(x^n)}{p_{\widehat{\mu}_1(X^{n^*})}(x^{n^*})} + A_n \right\}$$
$$\mathcal{D}_{0,n} = \left\{ \log \frac{p_{\mu_0^{(n)}}(x^n)}{p_{\mu_0^{(n)}}(x^{n^*})} \geq \log \frac{p_{\widehat{\mu}_0(X^n)}(x^n)}{p_{\widehat{\mu}_0(X^{n^*})}(x^{n^*})} - A_n \right\}.$$

The probability in (S4.7) can be bounded, for all $\beta \geq 1$, as

$$\begin{aligned}
\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n) &= \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n} \cap \mathcal{D}_{1,n}) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, (\mathcal{D}_{0,n} \cap \mathcal{D}_{1,n})^c) + \texttt{small}_{\vec{b}}(n) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n}, \mathcal{D}_{1,n}) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{1,n}^c) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{0,n}^c) + \texttt{small}_{\vec{b}}(n). \qquad \text{(S4.9)}
\end{aligned}$$

We first consider the first probability in (S4.9): there are constants $\vec{C}$ such that, for all large n,

$$\begin{aligned}
&\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n}, \mathcal{D}_{1,n}) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\mu_1^{(n)}}(X^n)}{p_{\mu_1^{(n^*)}}(X^{n^*})} - A_n + \log \frac{p_{\mu_0^{(n)}}(X^n)}{p_{\mu_0^{(n)}}(X^{n*})} - A_n \leq A_n \right) \\
&= \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\mu_1^{(n)}}(X^{>n^*})}{p_{\mu_0^{(n)}}(X^{>n^*})} \leq 3A_n \right) \\
&\leq e^{\frac{3}{2}A_n} e^{-\frac{n}{4} d_R(\mu_1^{(n)}, \mu_0^{(n)})} \leq e^{(3/2)\beta\kappa \log\log n + C_7} e^{-C_8 n \|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2} = e^{C_7} (\log n)^{(3/2)\beta\kappa - \Gamma_1 \cdot C_8},
\end{aligned} \qquad \text{(S4.10)}$$

where Γ_1 is as in the statement of the theorem, the second inequality follows by Lemma 2 and noting $n^* < \frac{n}{2}$, we used Proposition 1.

We now consider the second probability in (S4.9). Using $p_{\widehat{\mu}_1(X^n)}(x^n) \geq p_{\mu_1^{(n)}}(x^n)$ we have the following, where we define the event $\mathcal{F}_n = \{\widehat{\mu}_1(X^{n^*}) \in M_1'\}$ with M_1' the CINECSI subset

of M_1 mentioned in the theorem statement: there is $C_9, C_{10} > 0$ such that for al large n,

$$\begin{aligned}
\mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{1,n}^c) &= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\mu_1^{(n)}}(X^n)}{p_{\mu_1^{(n)}}(X^{n^*})} > \log\frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} + A_n\right) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\mu_1^{(n)}}(X^{n^*})} > \log\frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} + A_n\right) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})}{p_{\mu_1^{(n)}}(X^{n^*})} > A_n, \mathcal{F}_n\right) + \mathbb{P}_{\mu_1^{(n)}}\left(\mathcal{F}_n^c\right) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(D(\widehat{\mu}_1(X^{n^*})\|\mu_1^{(n)}) > A_n, \mathcal{F}_n\right) + \texttt{small}_{\vec{b}}(n) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\|\widehat{\mu}_1(X^{n^*}) - \mu_1^{(n)}\|_2^2 > C_9 A_n, \mathcal{F}_n\right) + \texttt{small}_{\vec{b}}(n) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\|\widehat{\mu}_1(X^{n^*}) - \mu_1^{(n)}\|_\infty > \sqrt{C_9 A_n/m_1}\right) + \texttt{small}_{\vec{b}}(n) \qquad \text{(S4.11)} \\
&\leq e^{-C_{10}A_n} = e^{-C_{10}(C_2-C_6)}\frac{1}{(\log n)^{C_{10}\beta\kappa}}, \qquad \text{(S4.12)}
\end{aligned}$$

where we used the KL robustness property (S1.5), Proposition 1 and Lemma 1.

The third probability in (S4.9) is considered in a similar way. Using $p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*}) \geq p_{\mu_0^{(n)}}(X^{n^*})$ we have $C_{11}, C_{12} > 0$ such that:

$$\begin{aligned}
\mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{0,n}^c) &= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\mu_0^{(n)}}(X^n)}{p_{\mu_0^{(n)}}(X^{n^*})} < \log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{1}{3}A_n\right) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\mu_0^{(n)}}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(x^{n^*})} < \log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{1}{3}A_n\right) \\
&= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\mu_0^{(n)}}(X^n)} > \frac{1}{3}A_n\right) \\
&\leq C_{11}\frac{1}{(\log n)^{C_{12}\beta\kappa}} \qquad \text{(S4.13)}
\end{aligned}$$

where we omitted the last few steps which are exactly as in (S4.11).

We now finish the proof by combining (S4.9), (S4.10), (S4.11) and (S4.13), which gives that, if we choose $\beta \geq \max\{1/(\kappa C_{10}), 1/(\kappa C_{12})\}$ and, for this choice β, we choose Γ_1 as in (S4.10) as $\Gamma_1 \geq (1 + (3/2)\beta\kappa)/C_8$, then we have $\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n) \leq \Gamma_4/(\log n)$ for some constant Γ_4 independent of n; the result now follows from (S4.7).

□

Proof of Theorem 1

Proof. We show the result in two stages. In Stage 1 we provide a tight upper bound on the risk, based on an extension of the decomposition of the risk (3.9) to general families and estimators $\breve{\mu}_0$ and $\breve{\mu}_1$ that are sufficiently efficient, i.e. that satisfy (4.1), and to losses $d_{\texttt{gen}}(\cdot\|\cdot)$ equal to squared error loss, standardized squared error loss and KL divergence (it is not sufficient to refer to Proposition 1 and prove the result only for squared error loss, because the equivalence result of Proposition 1 only holds on CINECSI sets and our estimators may take values outside of these; we do not need to consider Rényi and squared Hellinger divergences though, because these are uniformly upper bounded by KL divergence even for μ outside any CINECSI set). In Stage 2 we show how the bound implies the result.

Stage 1: Decomposition of Upper Bound on the Risk Let A_n be the event that $\mathcal{M}_1$ is selected, as in Section 3.3. We will now show that, under the assumptions of Theorem 1, we have for the constant C appearing in (4.1), for all $\mu_1 \in M_1'$,

$$R(\mu_1, \delta, n) \leq \frac{3C}{n} + 2\mathbb{P}(A_n^c) d_{\texttt{gen}}(\mu_1 \| \mu_0), \tag{S4.14}$$

where the left inequality holds for all divergence measures mentioned in the theorem, and the right inequality holds for $d_{\texttt{gen}}(\cdot\|\cdot)$ set to any of the squared error, the standardized squared error or the KL divergence.

To prove (S4.14), we use that for the three divergences of interest, for any $\mu_1 \in M_1, \mu \in M_0$, with $\mu_0 \in M_0$ as in (3.8), we have

$$d_{\texttt{gen}}(\mu_1 \| \mu) \leq 2(d_{\texttt{gen}}(\mu_1 \| \mu_0) + d_{\texttt{gen}}(\mu_0 \| \mu)), \tag{S4.15}$$

For $d_{\text{gen}}(\cdot\|\cdot)$ the KL divergence, this follows because

$$\begin{aligned} D(\mu_1\|\mu) &= \mathbb{E}_{\mu_1}\left[-\log\frac{p_\mu(X)}{p_{\mu_1}(X)}\right] = \mathbb{E}_{\mu_1}\left[-\log\frac{p_\mu(X)}{p_{\mu_0}(X)}\right] + \mathbb{E}_{\mu_1}\left[-\log\frac{p_{\mu_0}(X)}{p_{\mu_1}(X)}\right] \\ &= \mathbb{E}_{\mu_0}\left[-\log\frac{p_\mu(X)}{p_{\mu_0}(X)}\right] + \mathbb{E}_{\mu_1}\left[-\log\frac{p_{\mu_0}(X)}{p_{\mu_1}(X)}\right], \end{aligned} \tag{S4.16}$$

where the last line follows by the robustness property of exponential families (S1.4), since μ and μ_0 are both in M_0.

For $d_{\text{gen}}(\cdot\|\cdot)$ the squared and standardized squared error case we show (S4.15) as follows: Fix a matrix-valued function $J : M_1 \to \mathbb{R}^{m_1^2}$ that maps each $\mu \in M_1$ to a positive definite matrix J_μ. We can write

$$d_{\text{gen}}(\mu\|\mu') = (\mu-\mu')^T J_\mu(\mu-\mu'). \tag{S4.17}$$

where J_μ is the identity matrix for the squared error case, and J_μ is the Fisher information matrix for the standardized squared error case. (S4.15) follows since we can write, for any function J_μ of the above type including these two:

$$\begin{aligned} (\mu_1-\mu)^T J_{\mu_1}(\mu_1-\mu) &= (\mu_1-\mu_0+\mu_0-\mu)^T J_{\mu_1}(\mu_1-\mu_0+\mu_0-\mu) \\ &= (\mu_1-\mu_0)^T J_{\mu_1}(\mu_1-\mu_0) + (\mu_0-\mu)^T J_{\mu_1}(\mu_0-\mu) + 2(\mu_1-\mu_0)J_{\mu_1}(\mu_0-\mu) \\ &\le 2\left((\mu_1-\mu_0)^T J_{\mu_1}(\mu_1-\mu_0) + (\mu_0-\mu)^T J_{\mu_1}(\mu_0-\mu)\right), \end{aligned}$$

where the last line follows because for general positive definite $m \times m$ matrices J and m-component column vectors a and b, $(b-a)^T J(b-a) \ge 0$ so that $b^T J(b-a) \ge a^T J(b-a)$ and, after rearranging, $b^T Jb + a^T Ja \ge 2a^T Jb$.

We have thus shown (S4.15). It now follows that

$$\begin{aligned} R(\mu_1,\delta,n) &= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n} d_{\text{gen}}(\mu_1\|\breve{\mu}_1(X^n)) + \mathbf{1}_{A_n^c} d_{\text{gen}}(\mu_1\|\breve{\mu}_0(X^n))\right] \\ &\le \mathbb{E}_{\mu_1}\left[d_{\text{gen}}(\mu_1\|\breve{\mu}_1(X^n)) + 2\cdot\mathbf{1}_{A_n^c}\left(d_{\text{gen}}(\mu_0\|\breve{\mu}_0(X^n)) + d_{\text{gen}}(\mu_1\|\mu_0)\right)\right] \\ &\le \frac{3C}{n} + 2\mathbb{P}(A_n^c) d_{\text{gen}}(\mu_1\|\mu_0), \end{aligned} \tag{S4.18}$$

where we used (S4.15) and our condition (4.1) on $\breve{\mu}_0$ and $\breve{\mu}_1$. We have thus shown (S4.14).

Stage 2 We proceed to prove our risk upper bound for the squared error loss, standardized squared error loss and KL divergence, for which the right inequality in (S4.14) holds; the result then follows for squared Hellinger and Rényi divergence because these are upper bounded by KL divergence. From (S4.14) we see that it is sufficient to show that for all n larger than some n_0,

$$\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(A_n^c) d_{\texttt{gen}}(\mu_1 \| \mu_0)\} = O\left(\frac{\log\log n}{n}\right), \tag{S4.19}$$

for our three choices of $d_{\texttt{gen}}(\cdot\|\cdot)$. We first note that, since M_1' is CINECSI, $\sup_{\mu_1 \in M_1'} d_{\texttt{gen}}(\mu_1\|\mu_0)$ is bounded by some constant C_1. It thus follows by Proposition 2 that there exists some CINECSI subset M_1'' of M_1 such that, with $B_n^c \subset A_n^c$ defined as $B_n^c = \{x^n : \delta(x^n) = 0; \widehat{\mu}_1(X^n) \in M_1''\}$, we have

$$\begin{aligned}
\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(A_n^c) d_{\texttt{gen}}(\mu_1\|\mu_0)\} &= \sup_{\mu_1 \in M_1'} \{(\mathbb{P}_{\mu_1}(B_n^c) + \mathbb{P}_{\mu_1}(A_n^c \setminus B_n^c)) d_{\texttt{gen}}(\mu_1\|\mu_0)\} \\
&= \sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) d_{\texttt{gen}}(\mu_1\|\mu_0)\} + C_1 \cdot \mathbb{P}_{\mu_1}(\widehat{\mu}^{(1)} \notin M_1'') \\
&= \sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) d_{\texttt{gen}}(\mu_1\|\mu_0)\} + \texttt{small}_{\vec{b}}(n),
\end{aligned}$$

so that it is sufficient if we can show (S4.19) with B_n^c instead of A_n^c. But on the set B_n^c, all three divergence measures considered are within constant factors of each other, so that it is sufficient if we can show that there is a constant C_2 such that for all n larger than some n_0,

$$\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) \cdot \|\mu_1 - \mu_0\|_2^2\} \le C_2 \cdot \frac{\log\log n}{n}. \tag{S4.20}$$

Now, fix some $\mu_1 \equiv \mu_1^{(n)}$ and consider $f(n)$ as in (S4.1). By Theorem 3, $\mathbb{P}_{\mu_1}(B_n^c) \le C_3/n$ for some constant C_3 that can be chosen uniformly for all $\mu_1 \in M_1'$ whenever $f(n) > \Gamma_2 \log n$ with Γ_2 as in that theorem. Using also that $\|\mu_1 - \mu_0\|_2^2$ is bounded by C_1 as above, it follows

that (S4.20) holds whenever $f(n) > \Gamma_2 \log n$ and $(C_1 C_3)/n \leq C_2(\log\log n)/n$, i.e. whenever $f(n) > \Gamma_2 \log n$ and $C_2 \geq C_1 C_3/(\log\log n)$.

Second, suppose that $\Gamma_1 < f(n) \leq \Gamma_2 \log n$ with Γ_1 as in Theorem 4. Then by that theorem, uniformly for all $\mu_1^{(n)}$ with such $f(n)$, we have, with Γ_3 as in that theorem,

$$\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2 \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) = f(n) \cdot \frac{\log\log n}{n} \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \leq$$
$$\Gamma_2 \cdot (\log n) \cdot \frac{\log\log n}{n} \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \leq \Gamma_2\Gamma_3 \cdot \frac{\log\log n}{n},$$

where $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is defined as in (3.8), so that (S4.20) holds again whenever $C_2 \geq \Gamma_2\Gamma_3$.

Finally, suppose that $f(n) \leq \Gamma_1$ with Γ_1 as in Theorem 4. Then (S4.20) holds whenever $C_2 \geq \Gamma_1$. Combining the three cases we find that (S4.20) holds whenever $C_3 \geq \max\{\Gamma_1, \Gamma_2\Gamma_3, C_1C_3/(\log\log n)\}$; the result is proved.

□

S5 Switching as in Van Erven et al. (2012)

The basic building block of the switch distribution and criterion as formulated by Van Erven et al. (2012) is a countable set of *sequential prediction strategies* (also known as 'prequential forecasting systems' (Dawid, 1984)) $\{p_k \mid k \in \mathcal{K}\}$, where $\mathcal{K}$ is a finite or countable set indexing the basic models under consideration. Thus, each model is associated with a corresponding prediction strategy, where a prediction strategy p is a function from $\bigcup_{i\geq 0} \mathcal{X}^i$ to the set of densities on $\mathcal{X}$, where $p(\cdot \mid x^{n-1})$ denotes the density on $\mathcal{X}$ that x^{n-1} maps to, and $p(x_n \mid x^{n-1})$ is to be interpreted as the probabilistic prediction that strategy p makes for outcome X_n upon observation of the first $n-1$ outcomes, $X^{n-1} = x^{n-1}$. For example, for a parametric model $\{p_\theta \mid \theta \in \Theta\}$ one can base p_k on a Bayesian marginal likelihood, $p_B(x^n) := \int_\Theta \omega(\theta) p_\theta(x^n) d\theta$, where ω is a prior density on Θ. The corresponding prediction strategy could then be defined

by setting $p_k(x_n \mid x^{n-1}) := p_B(x^n)/p_B(x^{n-1})$, the standard Bayesian predictive distribution. In this paper, the basic strategies p_k were always Bayesian predictive distributions, but, in the spirit of Dawid (1984), one may consider other choices as well.

After constructing the set of basic prediction strategies, a new family of prediction strategies that switch between the strategies in the set $\{p_k \mid k \in \mathcal{K}\}$ is defined. Formally, let $\mathbb{S}$ be the set

$$\mathbb{S} = \{((t_1, k_1), \ldots, (t_m, k_m)) \in (\mathbb{N} \times \mathcal{K})^m \mid m \in \mathbb{N}, 1 = t_1 < t_2 < \ldots < t_m\}. \tag{S5.1}$$

Each $s \in \mathbb{S}$ specifies the times $t_1, \ldots, t_m$ at which a switch is made between the prediction strategies from the original set, identified by the indices $k_1, \ldots, k_m$. The new family $Q = \{q_s \mid s \in \mathbb{S}\}$ is then defined by setting, for all $n, x^n \in \mathcal{X}^n$:

$$q_s(x_n \mid x^{n-1}) = p_{k_j}(x_n \mid x^{n-1}), \quad t_j \leq n < t_{j+1}, \tag{S5.2}$$

with $t_{m+1} = \infty$ by convention. We now define $q_s(x^n) = \prod_{i=1}^n q_s(x_i \mid x^{i-1})$; one easily verifies that this defines a joint probability density on $\mathcal{X}^n$.

We now place a prior mass function π' on $\mathbb{S}$ and define, for each n, the *switch distribution* in terms of its joint density for $\mathcal{X}^n$ and $\mathbb{S}$:

$$p_{\mathrm{sw}}(x^n, s) = q_s(x^n)\pi'(s), \quad p_{\mathrm{sw}}(x^n) = \sum_{s \in \mathbb{S}} p_{\mathrm{sw}}(x^n, s) = \sum_{s \in \mathbb{S}} q_s(x^n)\pi'(s).$$

If the p_k are defined as Bayesian predictive distributions as above, then, as explained by Van Erven et al. (2012), the density $p_{\mathrm{sw}}(x^n)$ can be interpreted as a Bayesian marginal density of x^n under the prior π' on meta-models (model sequences) in $\mathbb{S}$.

The switch distribution can be used to define a model selection criterion δ'_{sw} by selecting the model with highest posterior probability under the switch distribution. This is done by

defining the random variable $K_{n+1}(s)$ on $\mathbb{S}$ to be the index of the prediction strategy that is used by q_s to predict the $(n+1)$th outcome. The model selection criterion is then:

$$\begin{aligned}\delta'_{\mathrm{sw}}(x^n) &= \arg\max_k p_{\mathrm{sw}}(K_{n+1}=k \mid x^n) = \arg\max_k \frac{\sum_{s:K_{n+1}(s)=k} p_{\mathrm{sw}}(x^n,s)}{p_{\mathrm{sw}}(x^n)} \\ &= \arg\max_k \frac{\sum_{s:K_{n+1}(s)=k} q_s(x^n)\pi'(s)}{\sum_{s\in\mathbb{S}} q_s(x^n)\pi'(s)}, \end{aligned} \tag{S5.3}$$

with ties resolved in any way desired.

In our nested two-model case, one might use, for example, a prior π' with support on

$$\mathbb{S}' = \{(1,0),(1,1),(\,(1,0),(2,1)\,),(\,(1,0),(4,1)\,),(\,(1,0),(8,1)\,),(\,(1,0),(16,1),\dots\,)\}.$$

Such a prior expresses that at time 1, for the first prediction, one can either switch to (i.e., start with), model 0, and keep predicting according to its Bayes predictive distribution — this strategy gets weight $\pi((1,0))$. Or one can start with model 1, and keep predicting according to its Bayes predictive distribution — this strategy gets weight $\pi((1,1))$. Or one can start with model 0 and switch to model 1 after 2^i observations and then stick with 1 forever — this strategy gets weight $\pi((\,(1,0),(2^i,1)\,))$. If we now start with a prior π on $\{1,2,\dots\}$ as in the main text and define $\pi'((1,0)) = 1/2$, $\pi'((1,1)) = (1/2)\cdot\pi(1)$, and for $i \geq 1$, $\pi'((1,0),(2^i,1)) = (1/2)\cdot\pi(2^i)$, then $\sum_{s\in\mathbb{S}'}\pi'(s) = 1$, so π' is a probability mass function. A simple calculation gives that (S5.3) based on switch prior π' now chooses model 1 if

$$\sum_{1\leq t<n} \bar{p}_t(x^n)\pi(t) > (1+g(n))\cdot p_{B,0}(x^n), \tag{S5.4}$$

where $g(n) = \sum_{t\geq n}\pi(t)$; note that $g(n)$ is decreasing and converges to 0 with increasing n. (S5.4) is thus an instance of the switch criterion of Van Erven et al. (2012). Comparing this to (2.2), the criterion used in this paper, after rearranging we see that it chooses model 1 if

$$\sum_{1\leq t<n} \bar{p}_t(x^n)\pi(t) > (1-g(n))\cdot p_{B,0}(x^n),$$

which is more likely by constant factor to select model $\mathcal{M}_0$, the factor however tending to 1 with increasing n. It is completely straightforward to check that Theorem 1 and all other results in this paper still hold if δ_{sw} with prior π as in the main text is replaced by δ'_{sw} with corresponding prior π' as defined here; thus our results carry over to the original definitions of Van Erven et al. (2012). Similarly, the proof for the strong consistency of δ'_{sw} given by Van Erven et al. (2012) carries through for δ_{sw}, needing only trivial modifications. From (S5.4) we see that modifying the prior π in either our or Van Erven et al.'s original criterion has a similar effect as keeping the same π but switching between the two versions of the switch criterion.

References

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley.

Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Prob.* **12**, 768-793.

Dawid, A. P. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. R. Stat. Soc. Ser. A Stat. Soc.* **147** 278-292.

Grünwald, P. D. and de Rooij, S. (2005). Asymptotic log-loss of prequential maximum likelihood codes. In *Proceedings of the Eighteenth Annual Conference on Computational Learning Theory*, 652-667. Springer-Verlag.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. The MIT Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* **90**, 773-795.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Ing. Theory* **60**, 3797-3820.

van Erven, T., Grünwald, P. D. and de Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.* **74**, 361-417.