NONPARAMETRIC BINARY REGRESSION WITH RANDOM COVARIATES

BY

PERSI DIACONIS and DAVID FREEDMAN

TECHNICAL REPORT NO. 439

SEPTEMBER 1993

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS92–04864

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Nonparametric Binary Regression With Random Covarites

By

Persi Diaconis and David Freedman

Technical Report No. 439

September 1993

Prepared Under the Auspices

Of

National Science Foundation Grant DMS92-04864

Department of Statistics
Stanford University
Stanford, California

Nonparametric Binary Regression With Random Covariates

PERSI DIACONIS[1]
DAVID FREEDMAN[2]

Abstract

The performance of Bayes' estimates are studied under an assumption of conditional exchangeability. More exactly, for each subject in a data set, let ξ be a covariate and let η be a binary response variable, with $P\{\eta = 1|\xi\} = f(\xi)$. Here, f is an unknown function to be estimated from the data; the subjects are independent, and the ξ's are iid uniform in [0,1]. Define a prior distribution on f as $\sum_k w_k \pi_k / \sum_k w_k$, where π_k is uniform on the set of f which only depend on the first k bits of ξ. And $w_k > 0$ for infinitely many k. Bayes' estimates are consistent at all f if w_k decreases rapidly as k increase. Otherwise, the estimates are inconsistent at $f \equiv \frac{1}{2}$.

1. Introduction

This paper studies non-parametric binary regression in a Bayesian context. Let $\xi \in [0,1]$ be an observable covariate. Let η be a binary response variable with $P\{\eta = 1|\xi\} = f(\xi)$. The function f is assumed to be measurable from [0,1] to [0,1]. The data are $(\eta_1, \xi(1)), \cdots (\eta_n, \xi(n))$ with the ξ's independent and uniform in [0,1]. Furthermore,

(1) Given the covariates, the response variables are independent across subjects and $P\{\eta_i = 1|\xi\} = f(\xi(i))$.

The function f is an infinite-dimensional parameter, to be estimated from the data by Bayesian methods. We introduce non-parametric priors on the space of all f's. The posterior is computed in Section 2. The main issue to be studied is consistency: does the posterior concentrate near the true f as more data comes in? We show that the answer is usually "yes," but not always. For some classes of priors, the posterior does not converge when $f \equiv \frac{1}{2}$.

We next describe the class of priors. These were motivated by an example of de Finetti [1959]. Regard a point ξ in the unit interval as an infinite sequence of binary digits (bits) $\xi_1, \xi_2 \ldots$). The first bit might stand for treated or not, the second bit for fat or thin,

---

[1] Mathematics Dept., Harvard University, Cambridge, MA 02138; research partially supported by NSF Grant DMS 86-00235.

[2] Statistics Department, University of California, Berkeley, CA 94720; research partially supported by NSF Grant DMS 86-01634.

and so on. There is a fairly conventional prior which is "nested" or "hierarchical." Begin with a prior π_k supported on the class of functions f that depend only on the first k bits in ξ. Then treat k as a "hyper-parameter" putting prior weight w_k on k. This gives a prior

$$\pi = \sum_{k=0}^{\infty} w_k \pi_k / \sum_{k=0}^{\infty} w_k \tag{2a}$$

where

$$w_k > 0 \text{ for infinitely many } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty. \tag{2b}$$

Let C_k be the set of strings of 0's and 1's of length k. The prior π_k is defined by the joint distribution it assigns to the 2^k parameters $\theta_s: s \in C_k$. Here θ_s is the probability of success for subjects whose first k covariates are given by s. One simple choice is to take θ_s independent and uniform in [0,1] for all $s \in C_k$. This is the example to keep in mind. The calculations and results work for a generalization which we now introduce. Fix $0 < b \leq B < \infty$ and a finite subset F of (0,1). Consider the class Γ of all densities γ on [0,1] with $b \leq \gamma \leq B$ and $\int_0^1 \theta\gamma(\theta)d\theta \in \Gamma$. Consider the π_k which make the 2^k success probabilities θ_s independent, with density $\gamma_s \in \Gamma$. We will require that the various choices fit together for large k in the following sense. Let $g_s = \int \theta\gamma_s(\theta)\, d\theta$. We assume that the g_s all lie in a finite subset F of (0,1), given a priori. Furthermore, for all large k, for all $s \in C_k$, $g_s = g_\infty(x)$ for all x with $x_j = s_j$ for $1 \leq j \leq k$. And g_∞ depends only on finitely many covariates. This completes the definition of a Γ-uniform prior. If $b = B = 1$ and $F = \{\frac{1}{2}\}$ we get the uniform priors.

To define consistency, a topology must be specified. Let $C_\infty = \{0,1\}^\infty$, so $x \in C_\infty$ has coordinates $x_1, x_2, \cdots$ which are 0 or 1. Write λ^∞ for the uniform measure on C_∞, i.e., Lebesgue measure. With respect to λ^∞, the coordinates are independent and $\lambda^\infty\{x_j = 1\} = \frac{1}{2}$. By definition, the parameter space Θ is the set of measurable functions from C_∞ to [0,1]; functions which are equal a.e. are identified. Put the L_2 metric on the parameter space. (The same topology is given by convergence in the L_p norm for any $p \geq 1$, or by convergence in measure.) A typical neighborhood $N(f, \delta, \epsilon)$ of f is defined by

(3) If $f \in \Theta$ and $\epsilon, \delta > 0$, let $N(f, \delta, \epsilon)$ be the set of $h \epsilon \Theta$ with

$$\lambda^\infty\{x: x \in C_\infty \text{ and } |h(x) - f(x)| \leq \epsilon\} \geq 1 - \delta.$$

If π is a prior probability on Θ, the posterior $\tilde{\pi}_n$ on Θ is the conditional law of f given the data. This will be computed for our setup in Section 2. The prior π is "consistent at f" if $\tilde{\pi}_n\{N(f, \delta, \epsilon)\} \to 1$ almost surely as $n \to \infty$ for all positive ϵ, δ.

Returning to the data, at stage n, there are n subjects, indexed by $i, 1 \leq i \leq n$. Each subject has a covariate $\xi \in C_\infty$ and $\eta \in \{0,1\}$ distributed according to (1). We assume

(4) $\xi(i)$ are independent and have a common uniform distribution λ^∞.

With these preliminaries, the main theorems of this paper can now be stated.

(5) **Theorem.** *Suppose (1) and (4). Suppose too the π_k are Γ-uniform, the prior π is hierarchical in the sense of (2), and $f \not\equiv g_\infty$. Then π is consistent.*

(6) **Theorem.** *Suppose (1) and (4). Suppose too the* π_k *are*

a) Suppose $\sum_{k=n}^{\infty} w_k < \exp\left[-\frac{1}{4}(\log 2)n2^\ell - \delta_0 n2^\ell\right]$ *for all large* n, *for some* $\delta_0 > 0$. *Then* π *is consistent at* f.

b) Suppose $\sum_{k=n}^{\infty} w_k > \exp\left[-\frac{1}{4}(\log 2)n2^\ell - \delta_0 n2^\ell\right]$ *for infinitely many* n, *for some* $\delta_0 > 0$. *Then* π *is inconsistent at* f.

The critical rate is different here and in DF93; see Theorems (8) and (9) there. The δ_0 in the statement of Theorem 6 is a fixed quantity.

Discussion and Literature Review

Theorems 5 and 6 show that our Bayes rules are consistent for all f, provided that the weights w_k fall off suitably fast. For example, suppose $b = B = 1$ so that π_k makes the coordinates independent and uniform on [0,1]. If $w_k = 1/2^k$ for $k = 0, 1, 2, ...$, then (π, f) is consistent for all f. If $w_k = 1/(k+1)^2$ for $k = 0, 1, 2, ...$ then (π, f) is consistent for all f except $f \equiv \frac{1}{2}$.

Consistency of Bayes rules is a classical problem going back to Laplace [1774]. A survey of the literature with emphasis on the problems that can occur in infinite dimensions is given by Diaconis and Freedman [1988].

The present paper is a modification of Diaconis and Freedman [1993], referred to throughout as DF93. That paper studies the model (1) with a different sampling design for the $\xi(i)$. These were taken balanced, so that at stage n, all $s \in C_n$ occurred exactly once. This eliminated the annoying inhomogeneity we have to deal with in the present paper. The results are similar but the initial rate differs from Theorem 6; see Theorems (8) and (9) in DF93. In DF93, similar results were obtained for a class of priors more general than Γ-uniform. We presume such results hold in the present setup. DF93 reviews the relationship between consistency of Bayes rules and model selection, sieves, and orthogonal series estimation. In Diaconis and Freedman [1993A], we discuss variants such as real valued observables. There, the analogous priors are consistent for all f.

The remainder of this paper is organized as follows. Section 2 computes the posterior. Section 3 has some preliminary estimates, including large-derivation results for balls dropped at random into boxes. A proof of Theorem 5 will be found in Section 4, and a proof of Theorem 6 in Section 5. The arguments are modifications of those presented in DF93.

2. Computing the Posterior

Let Ω be an underlying probability space on which the response variables $\eta(i)$ and covariates $\xi_j(i)$ are defined. Recall that $f \in \Theta$ maps C_∞ to [0,1]. For $f \in \Theta$, let P_f be the probability on Ω which makes the response variables and covariates distributed so that (1) and (4) hold. The dependence between the data at stage n and stage $n+1$ is simple: there is one extra subject with covariate sequence $\xi(n+1)$. The joint distribution across n's will matter for some of the arguments here, as opposed to DF93.

Let

$$(2.1) \qquad f_k(x) = E\{f|x_1, \cdots, x_k\} = \int_{C_\infty} f(x_1, \cdots, x_k, y)\,\lambda^\infty(dy)$$

and write $f_k(s)$ for $f_k(x)$ when $s \in C_k$ and $x_1 = s_1, ..., x_k = s_k$.

For now, fix n and k. For $s \in C_k$, let N_s be the number of subjects $i = 1, ..., n$ such that $\xi_j(i) = s_j$ for $j = 1, ..., k$. In other words, N_s is the number of subjects $i = 1, ..., n$ whose

first k covariates are given by s. Of course, N_s is random; that is the new technical difficulty. Let X_s be the number of successes among subjects whose covariate sequence begins with s. More formally, $\eta(i)$ is the response for subject i, and

$$X_s = \sum_{i=1}^{n} \{\eta(i): \xi_j(i) = s_j \quad \text{for } i = 1, ..., n\}. \tag{2.2}$$

Write $\text{bin}(m, p)$ for the binomial distribution with m trials and success probability p.

(2.3) LEMMA. *Assume (4). With respect to* P_f :

a) $\{N_s: s \in C_k\}$ *is distributed like the result of dropping* n *balls at random into* 2^k *cells.*

b) *Given* $\{N_s: s \in C_k\}$, *the random variables* X_s *are independent as* s *ranges over* C_k, *each being* $\text{bin}[N_s, f_k(s)]$.

As usual, π_k can be extended to a probability on $\Theta \times \Omega$, by the formula

$$\pi_k(A \times B) = \int_A P_f\{B\}\, \pi_k\{df\}.$$

In this formula, A is a measurable subset of Θ and B is a measurable subset of Ω. The proofs of (2.3-4) are omitted as routine. In (2.4) and similar contexts, π_k is viewed as a probability on $\Theta \times \Omega$.

(2.4) LEMMA. *Suppose* π_k *is* Γ*-uniform. With respect to* π_k, *the* N_s *have the ball-dropping distribution given by (2.3). Given* $\{N_s: s \in C_k\}$, *the pairs* (θ_s, X_s) *are independent as* s *ranges over* C_k. *The parameter* θ_s *has density* $\gamma_s \in \Gamma$. *Given* N_s *and* θ_s, *the number of successes* X_s *is* $\text{bin}(N_s, \theta_s)$.

For $\gamma \in \Gamma$, $m = 0, 1, 2, ...$, and $j = 0, 1, ..., m$, let

$$\gamma(m, j, \cdot): \theta \to \frac{\theta^j (1-\theta)^{m-j} \gamma(\theta)}{\phi(m, j, \gamma)} \tag{2.5a}$$

where the normalizing constant is

$$\phi(m, j, \gamma) = \int_0^1 \theta^j (1-\theta)^{m-j} \gamma(\theta)\, d\theta. \tag{2.5b}$$

In particular, $\phi(0, 0, \gamma) = 1$ and $\gamma(0, 0, \cdot) = \gamma(\cdot)$.

Let $\widetilde{\pi}_{k,n}$ be the posterior distribution of f, computed relative to π_k, given the data from a design of order n.

(2.6) LEMMA. *Suppose* π_k *is* Γ*-uniform. According to the posterior* $\widetilde{\pi}_{k,n}$, *the success probabilities* θ_s *are independent as* s *ranges over* C_k, *and* θ_s *has density* $\gamma_s(N_s, X_s, \cdot)$ *with respect to Lebesgue measure on [0,1].*

To compute the posterior relative to π, the π_k-predictive probability of the data is needed. To set up the notation, recall the normalizing constant ϕ from (2.5b). Let

(2.7) $$\rho_{k,n} = \prod_{s \in C_k} \phi(N_s, X_s, \gamma_s).$$

If $N_s = 0$, the corresponding factor in $\rho_{k,n}$ is taken as 1. By (2.4), $\rho_{k,n}$ is the π_k-predictive probability of the data, given $\{N_s\}$.

Turn now to the posterior $\tilde{\pi}_n$, computed relative to π. Informally, the "theory index" k in (2) is a parameter, which has a posterior distribution relative to π. Let

(2.8) $$\tilde{w}_{k,n} = w_k \rho_{k,n}.$$

Now, $\pi_k\{data\}/\pi\{data\} = \tilde{w}_{k,n} / \sum_{k=0}^{\infty} \tilde{w}_{k,n}$. So

(2.9) $$\tilde{\pi}_n(k) = \frac{\tilde{w}_{k,n}}{\sum_{k=0}^{\infty} \tilde{w}_{k,n}}.$$

(2.10) LEMMA. *Suppose π is hierarchical in the sense of (2), and the π_k are Γ-uniform. Given the data from a design of order n, the posterior is*

$$\tilde{\pi}_n = \sum_{k=0}^{\infty} \tilde{w}_{k,n} \frac{\tilde{\pi}_{k,n}}{\sum_{k=0}^{\infty} \tilde{w}_{k,n}}.$$

The proof is omitted as routine. Of course, $\tilde{\pi}_n$ can be written as

$$\sum_{k=0}^{\infty} \tilde{\pi}_n(k) \tilde{\pi}_{k,n}.$$

3. Some Estimates

(3.1) LEMMA. *Let $0 \le p \le 1$. Let X be* bin(m,p) *and $Y = (X - mp)^2/m$. If $m = 0$, or $m > 0$ but $p = 0$ or 1, let $Y = 0$. Then*

a) $P\{X \le mp - \sqrt{mx}\} < \exp\left(-\frac{1}{2}x\right)$ *for all* $x > 0$.

b) $P\{X \ge mp + \sqrt{mx}\} < \exp\left(-\frac{1}{2}x\right)$ *for all* $x > 0$.

c) Y *is stochastically smaller than* $\chi_2^2 + 2\log 2$.

Proof. Suppose $m > 0$ and $0 < p < 1$. Claim a) follows from Bernstein's inequality. For example, use (4) in Freedman [1973] to see that

$$P\{X \le mp - \sqrt{mx}\} < \exp\left(-\frac{1}{2p}x\right) < \exp\left(-\frac{1}{2}x\right)$$

To get claim b), write $q = 1 - p$, and observe that $X \ge mp + \sqrt{mx}$ iff $(m - X) \le mq - \sqrt{mx}$. Now use a). For c),

$$P\{Y \ge x\} < 2\exp\left(-\frac{1}{2}x\right). \quad \blacksquare$$

(3.2) LEMMA. *Suppose the random variable ξ has a Laplace transform for $h < h_0$, where h_0 is positive. Let $\mathcal{H}$ be the class of random variables Y for which $E\{[Y - EY]^j\} \le E\{\xi^j\}$ for $j \ge 2$. There are positive, finite σ^2 and h_1, depending only on ξ, such that*

$$P\left\{\sum_{i=1}^{m} Y_i \ge \sum_{i=1}^{m} E\{Y_i\} + y\right\} < \exp\left(\frac{-y^2}{2\sigma^2 m}\right)$$

provided the Y_i's are independent, $Y_i \in \mathcal{H}$ for all i, and $0 < y \le h_1 m$.

Note. This lemma is set up to give one-sided bounds. In some cases, of course, it can also be applied to $\{-Y\}$. Then a lower bound can be obtained the same way, with a slightly smaller σ. More detailed results can be obtained by matching variances or Esscher tilting, but these refinements will not be needed here. See [Feller, 1966, sec. XVI.6].

Proof. Assume without real loss of generality that $EY = 0$. Let $\phi_Y(h) = E\{e^{hY}\} = 1 + \sum_{j=2}^{\infty} E\{Y^j\}h^j/j!$. The sum is bounded above by

$$\sum_{j=2}^{\infty} h^j E\{\xi^j\}/j! < \frac{1}{2}\sigma^2 h^2 \quad \text{for } 0 \le h \le h'.$$

Here, σ^2 is a suitable positive, finite number, slightly larger than the second moment of ξ. For $0 \le h \le h'$,

$$\phi_Y(h) < 1 + \frac{1}{2}\sigma^2 h^2$$

and

$$\log \phi_Y(h) < \frac{1}{2}\sigma^2 h^2.$$

The constants σ^2 and h' depend on ξ, not Y or h.

We are assuming $E\{Y_i\} = 0$. Chebychev's inequality can be applied to bound $P\{e^{h(Y_1+\cdots+Y_m)} \ge e^{hy}\}$:

$$\begin{aligned}\log P\{Y_1 + \cdots + Y_m \ge y\} &\le -hy + m \log \phi_Y(h) \\ &< -hy + \frac{1}{2} m\sigma^2 h^2.\end{aligned}$$

Put $h = y/\sigma^2 m$. We require $h \le h'$, i.e., $y \le h'\sigma^2 m$: set $h_1 = h'\sigma^2$. ∎

(3.3) LEMMA. *Suppose $|U|$ is stochastically smaller than V. Then $|U - EU|$ is stochastically smaller than $V + EV$.*

(3.4) COROLLARY. *Let n_i be non-negative integers, and $0 \le p_i \le 1$. Let X_i be independent* $\text{bin}(n_i, p_i)$ *and $Y_i = (X_i - n_i p_i)^2/n_i$. There are universal positive constants σ^2 and h_1 such that*

$$P\left\{Y_1 + \cdots + Y_m \ge \sum_{i=1}^{m} p_i(1 - p_i) + y\right\} < \exp\left(\frac{-y^2}{2\sigma^2 m}\right)$$

provided $0 < y \leq h_1 m$.

Proof. Combine (3.1-3.) ∎

(3.5) LEMMA. *Let* N_λ *be Pois*(λ), *i.e., Poisson with parameter* λ. *If* $N_\lambda = 0$, *let* $\log(N_\lambda) = 0$. *Let* $z > 0$.

a) $P\left\{\sqrt{\lambda}\,(\log N_\lambda - \log \lambda) \geq z\right\} < \exp\left(-\frac{1}{2}z^2\right)$ for all $z > 0$.

b) $P\left\{\sqrt{\lambda}\,(\log N_\lambda - \log \lambda) \leq z\right\} < \exp\left(-\frac{1}{2}z^2\,[(1-e^{-\epsilon})/\epsilon]^2\right)$

provided $0 < z \leq \epsilon\sqrt{\lambda}$. *If* $\epsilon = \frac{1}{2}$, *an upper bound is* $\exp(-z^2/4)$.

Proof. This follows from Bernstein's inequality: see (4) in Freedman [1973]. Some auxiliary calculations are needed to estimate the function in (9) of that paper. We claim

(3.6) $u \to (e^u - 1)^2/u^2 e^u$ is strictly convex, with a minimum at $u = 0$.

Indeed, the function in (3.6) is $[(e^{u/2} - e^{-u/2})/u]^2$, which is readily expanded in even powers of u, with positive coefficients.

(3.7) $\lambda\left(e^{z/\sqrt{\lambda}} - 1\right)^2 / e^{z\sqrt{\lambda}} > z^2$, for $\lambda > 0$ and $z > 0$.

(3.8) $u \to (1 - e^{-u})/u$ is strictly decreasing for $u > 0$.

(3.9) $\lambda\left(1 - e^{-z/\sqrt{\lambda}}\right)^2 > z^2[(1-e^{-\epsilon})/\epsilon]^2$ for $0 < z < \epsilon\sqrt{\lambda}$. ∎

(3.10) COROLLARY. *Let* $N'_\lambda = N_\lambda$ *if* $N_\lambda > \lambda e^{-1/2}$, *else let* $N'_\lambda = \lambda e^{-1/2}$. *Let* $Z_\lambda = \sqrt{\lambda}(\log N'_\lambda - \log \lambda)$. *Then* Z_λ^2 *is stochastically smaller than* $4\log 2 + 2\chi_2^2$.

(3.11) COROLLARY. *There are finite positive constants* σ^2 *and* h_1 *such that*

$$P\left\{\left|\sum_{i=1}^{m}(Z_i - E\{Z_i\})\right| \geq y\right\} < 2\exp\left(\frac{-y^2}{2\sigma^2 m}\right)$$

provided the Z_i *are independent, each* Z_i *is distributed as* Z_{λ_i} *in (3.10), and* $0 \leq y \leq h_1 m$.

(3.12) LEMMA. $\lim_{\lambda\to\infty} E\{\log N'_\lambda\}/\log\lambda = \lim_{\lambda\to\infty} E\{\log N_\lambda\}/\log\lambda = 1$.

Note. As $\lambda \to \infty$, the law of Z_λ tends to the standard normal. The bound in (3.5b) can be improved, but there is mass $P\{N_\lambda = 0\} = e^{-\lambda}$ at $z = -\sqrt{\lambda}\log\lambda$; no upper bound of the form $\exp(-\delta z^2)$ can be valid for large λ.

Results (3.13-16) are familiar, but are included for ease of reference. The elementary proof of (3.13) is omitted.

(3.13) LEMMA. *Let* f *be a convex function. Let* $a, b > 0$ *and let the random variable* X_{ab} *take values* $-a$ *or* b *and* $E\{X_{ab}\} = 0$. *Then* $E\{f(X_{ab})\}$ *increases with* b *for fixed* a; *likewise,* $E\{f(X_{ab})\}$ *increases with* a *for fixed* b.

(3.14) LEMMA. *Let* f *be a convex function. Fix* A, B *and* μ *with* $\infty < A < \mu < B < \infty$. *Let* $\mathcal{H}$ *be the class of random variables* X *such that* $A \leq X \leq B$ *and* $E\{X\} = \mu$. *Let* $\xi \in \mathcal{H}$ *take only the values* A, B *and let* $E\{\xi\} = \mu$. *Then* $E\{f(X)\} \leq E\{f(\xi)\}$.

Proof. Assume without loss of generality that $\mu = 0$. The extreme X have two-point distributions and (3.13) applies. ∎

(3.15) COROLLARY. *Let f be convex and increasing. Fix L and ϵ positive and finite. Let $\mathcal{H}$ be the class of random variables X such that $|X| \leq L$ and $E\{X\} = -\epsilon$. Let $\xi \in \mathcal{H}$ take only the values $\pm L$ and let $E\{\xi\} = -\epsilon$. Then $E\{f(X)\} \leq E\{f(\xi)\}$.*

(3.16) LEMMA. *Define $\mathcal{H}$ as in (3.15). There is a ρ with $0 < \rho < 1$, depending only on L and ϵ, such that: for independent $X_i \in \mathcal{H}$ and $y > 0$,*

$$P\left\{\sum_{i=1}^{m} X_i \geq y \quad \text{for some } m\right\} < \rho^Y.$$

Proof. Define ξ as in (3.15). De Moivre solved the gambler's ruin problem by finding the unique $r > 1$ with $E\{r^\xi\} = 1$. Continuing his argument, let $S(m) = \sum_{i=1}^{m} X_i$. Then $r^{S(m)}$ is an expectation-decreasing martingale, which can be stopped at the crossing time; take $\rho = 1/r$. ∎

Remark. Lemma (3.16) is easily extended to partial sums of variables X_i such that the conditional law of X_i given the past falls in $\mathcal{H}$. See Dubins and Savage [1965, p. 164].

Lemmas (3.17-18) are elementary, and proofs are omitted.

(3.17) LEMMA. *Let j be a non-negative integer, and x be a positive real number. Let $f_j(x) = \sum_{i=j}^{\infty} x^i/i!$. Then $f_j(x)/x^j$ is continuous, convex, and strictly increasing on $(0, \infty)$, with a limit of $1/j!$ as x decreases to 0.*

(3.18) LEMMA. *Let m be a positive integer and $0 < p < 1$. Let X be* bin(m, p), *and let j be a non-negative integer.*

a) $P\{X = j\} < (mp)^j/j!$.

b) $P\{X \geq j\} < f_j(mp)$.

(3.19) LEMMA. *Assume (4). Fix $c > 5/3$. Almost surely, for all sufficiently n, for all $k > c \log_2 n$, there are no $s \in C_k$ with $N_s \geq 4$.*

Proof. By (2.3a), N_s is bin$(n, 1/2^k)$. Write $\lambda = n/2^k = E\{N_s\}$. So $\lambda < 1/n^{c-1}$. By (3.17) and (3.18), $P\{N_s \geq 4\} < C\lambda^4$, where C is a suitable positive constant (a bit larger than 1/4!). The expected number of s with $N_s \geq 4$ is then smaller than $C2^k\lambda^4 = Cn\lambda^3 < C/n^{3c-4}$. The chance of having at least one box with $N_s \geq 4$ is also smaller than C/n^{3c-4}, by Chebychev's inequality. Since $3c > 5$, the Borel Cantelli lemma implies that for all sufficiently large n, for k the least integer exceeding $c \log_2 n$ there are no $s \in C_k$ with $N_s \geq 3$. Finally, for n fixed, $|\{s: s \in C_k \text{ and } N_s \geq 4\}|$ is decreasing as k increases. ∎

Note. We write $|S|$ for the cardinality of a set S.

(3.20) LEMMA. *Assume (4). Fix $c > 7/4$. Almost surely, for all sufficiently large n, for all $k > c \log_2 n$,*

(i) there are no $s \in C_k$ with $N_s \geq 4$, and

(ii) there is at most one $s \in C_k$ with $N_s = 3$.

Proof. (i) follows from (3.19), since $7/4 > 5/3$. For (ii), let Q_k be the event that $N_s = 3$ for two or more $s \in C_k$. Thus,

$$Q_k = \cup\{N_s = 3 \text{ and } N_t = 3 | s, t \in C_k \text{ and } s \neq t\}$$

and

$$P\{Q_k\} = \binom{2^k}{2} P_f\{N_s = 3 \text{ and } N_t = 3\}.$$

Now $P_f\{N_s = 3\} < \lambda^3/6$ by (3.18a). Given $\{N_s = 3\}$, N_t is bin$(n-3,\ 1/(2^k - 1))$, so

$$P\{N_t = 3 | N_s = 3\} < \frac{1}{6}\left(\frac{n-3}{2^k - 1}\right)^3 < \frac{\lambda^3}{6}.$$

Thus

$$P\{Q_k\} < \frac{1}{72} 2^{2k}\lambda^6 = \frac{1}{72} n^2\lambda^4 \leq \frac{1}{72n^{4c-6}}.$$

The balance of the argument is omitted, as similar to (3.19). ∎

(3.21) LEMMA. *Assume (4). Fix $c > 3$. Almost surely, for all sufficiently large n, for all $k > c\log_2 n$, there are no $s \in C_k$ with $N_s \geq 2$.*

Proof. Only the minimal k needs to be considered. Now $P_f\{N_s \geq 2\} < C\lambda^2$ by (3.17-18), so the expected number of $s \in C_k$ with $N_s \geq 2$ is at most $C2^k\lambda^2 = Cn^2/2^k < C/n^{c-2}$. Since $c > 3$, the Borel Cantelli lemma completes the proof. ∎

Lemmas (3.19-21) involve the dependence structure of the ball-dropping process, as k and n vary. The next result does not. Consider dropping n balls at random into b boxes, where n is much smaller than b: in the case of interest, b is of order $n^2/\log n$. Let $\lambda = n/b$, the expected number of balls in each box.

(3.22) Definition. Let $|M|$ be the number of multiply-occupied boxes, and T the total number of balls in the set of multiply-occupied boxes. Let $S_n = T - |M|$, with $S_0 = 0$. Let p_j be the conditional probability that ball j drops into a previously-occupied box, given the results of dropping the first $j-1$ balls.

Clearly, $S_n \leq n - 1$, where the bound is sharp; $n - S_n$ is the number of occupied cells; $S_1 = 0$; and for $n \geq 2$, $S_n = \sum_{j=2}^n X_j$, where X_j is 1 if ball j drops into a previously-occupied box; else, X_j is 0. Recall p_j from (3.21). Of course, p_j is itself a random variable, and

$$p_j = \frac{(j - 1 - S_{j-1})}{b}.$$

(3.24) LEMMA. *Let $\mu = n(n-1)/2b$.*

a) *If $0 < \delta < 1$, then* $\Pr\{S_n \geq (1+\delta)\mu\} < \exp(-\delta^2\mu/4)$.

b) Suppose $0 < \delta < 1$, and $n/b < \delta/2$. Then

$$\Pr\{S_n \leq (1-\delta)\mu\} < \exp\left(-\delta^2\mu/8\right).$$

Proof. Claim a) is Bernstein's inequality: see e.g. (4) in Freedman [1973], noting that $\sum_{j=1}^n p_j \leq \mu$.

Claim b) is similar. Indeed,

$$\Pr\left\{S_n \le -\frac{\delta}{2}\,\mu + \sum_{j=1}^{\infty} p_j\right\} < \exp\left(\frac{-\delta^2\mu}{8}\right).$$

Clearly, $\mu - (nS_n/b) \le \sum_{j=1}^{\infty} p_j$ because s_j increases with j. So

$$\left\{S_n \le \left(1-\frac{\delta}{2}\right)\mu - \frac{nS_n}{b}\right\} \subset \left\{S_n \le -\frac{\delta}{2}\,\mu + \sum_{j=1}^{\infty} p_j\right\}.$$

Furthermore, $S_n \le \left(1-\frac{\delta}{2}\right)\mu - \frac{nS_n}{b}$ iff $S_n \le \alpha\mu$, where

$$\alpha = \frac{1-\frac{\delta}{2}}{1+\frac{n}{b}} > (1-\delta).$$

Therefore, $\{S_n \le (1-\delta)\mu\} \subset \{S_n \le \alpha\,\mu\}$. ∎

Note. The argument shows S_n to be stochastically smaller than $\sum_{j=2}^{n} Y_j$, where the Y_j are independent 0-1 valued random variables, and $\Pr\{Y_j = 1\} = (j-1)/b$.

(3.25) LEMMA. *Fix $j \ge 0$. Let N_λ be Poisson, but conditioned to be j or more. Then N_λ is stochastically increasing with λ.*

Proof. Let $f_j(\lambda) = \sum_{k=j}^{\infty} \lambda^k/k!$. If $i > j$, we claim that $f_i(\lambda)/f_j(\lambda)$ increases with λ. This comes down to showing

$$\frac{f_i'(\lambda)}{f_i(\lambda)} > \frac{f_j'(\lambda)}{f_j(\lambda)}. \tag{3.26}$$

However,

$$f_i'(\lambda) = \frac{\lambda^{i-1}}{(i-1)!} + f_i(\lambda).$$

So (3.26) in turn reduces to

$$\sum_{k=j}^{\infty} \frac{\lambda^{k-j}}{k!} > \sum_{k=i}^{\infty} \frac{\lambda^{k-i}\,(i-1)!}{k!} \tag{3.27}$$

which holds term by term. ∎

4. The Proof of Theorem 5

This is proved like (8) in DF93; only the main points are given. Zones are defined in terms of positive integers K_i to be chosen later.

Early zone: $0 \le k \le K_1$.

Lower midzone: $K_1 \le k \le \log_2 \log n + K_2$.

Upper midzone: $\log_2 \log n + K_2 \le k \le \log_2 n - K_3$.
End zone: $\log_2 n - K_3 \le k \le \log n + K_4$.
High zone: $\log_2 n + K_4 \le k$.

The end zone and high zone have negligible posterior mass; the early zone is negligible too, unless $f = f_k$ for some k. Almost surely, for all large n, for all k in the midzone, for most $s \in C_k$, N_s is large and the MLE $\widehat{p}_s = X_s/N_s$ is close to $f_k(s)$. Of course, the latter tends to f : see (2.1). Finally, the posterior piles up around the MLE, by Diaconis and Freedman [1990]. We turn to details; lemmas (4.2-4) do most of the work for the midzone.

(4.1) $$\text{Let } \lambda = \frac{n}{2^k} \text{ so } E\{N_s\} = \lambda.$$

(4.2) LEMMA. *Assume (1) and (4). Fix any positive integer K. Almost surely $[P_f]$, for all sufficiently large n, for all k with $0 \le k \le \log_2 \log n + K$ and all $s \in C_k$:*

a) $N_s > n/2^{k+1} \ge n/(2^{K+1} \log n)$.

b) $|\widehat{p}_s - f_k(s)| < (2\sqrt{2^{K+1} \log n})/\sqrt{n}$.

Proof. Claim a). By (2.3a), N_s is $\text{bin}(n, 1/2^k)$. Abbreviate $C = 1/2^{2K+3}$. By Bernstein's inequality (3.1a),

$$P\left\{\frac{N_s \le \lambda}{2}\right\} < \exp\left(\frac{-\lambda^2}{8n}\right) \le \exp\left[\frac{-Cn}{(\log n)^2}\right].$$

The number of strings $s \in C_k$ with $0 \le k \le \log_2 \log n + K$ is

$$\sum_{k=0}^{\log_2 \log n + K} 2^k \le 2^{K+1} \log n$$

and

$$\sum_{n=1}^{\infty} (\log n) \exp\left[\frac{-Cn}{(\log n)^2}\right] < \infty.$$

The Borel-Cantelli lemma completes the proof of a).

The proof of b) is similar. Indeed, by (3.1),

$$P_f\left\{|X_s - N_s f_k(s)| \ge \sqrt{N_s} \cdot 2\sqrt{\log n}\right\} < 2\exp(-2 \log n) = \frac{2}{n^2}. \quad \blacksquare$$

(4.3) LEMMA. *Assume (1) and (4). Fix any large, finite M and small, positive δ. There are positive integers K_2, K_3 (depending on M, δ) such that: almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \le k \le \log_2 n - K_3$, for all but $\delta 2^k$ strings $s \in C_k$, $N_s > M$.*

Proof. The argument is by Poissonization. For now, fix k. Let N_s^* be iid $\text{Pois}(\lambda)$ as s varies over C_k. Thus, $\{N_s\}$ is distributed as $\{N_s^*\}$, given that $\{\sum_s N_s^* = n\}$. The conditioning event has probability asymptotic to $1/\sqrt{2\pi n}$. Choose K_3 so large that $\Pr\{\text{Pois}(2^{K_3}) \le M\} < \delta/2$. The chance that $\delta 2^k$ or more of the $s \in C_k$ have $N_s^* \le M$ is bounded above

by $\exp(-\delta^2 2^k/8)$. This follows from Bernstein's inequality (3.1b); also see (3.25). Now $2^k \geq 2^{K_2} \log n$. Choose K_2 so large that $C = 2^{K_2}\delta^2/8 > 1.5$. There are fewer than $\log_2 n$ theories k to consider, and

$$\sum (\log_2 n)\, \sqrt{n}/n^C < \infty.$$

The Borel-Cantelli lemma completes the proof. ∎

(4.4) LEMMA. *Assume (1) and (4). Fix δ, ϵ positive but small. There are positive integers K_2, K_3 such that: almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \leq k \leq \log_2 n - K_3$, for all but $\delta 2^{k+1}$ strings $s \in C_k$, $|\widehat{p}_s - f_k(s)| < \epsilon$.*

Proof. By Chebychev's inequality, if X is $\text{bin}(m,p)$, then $P\{|X - mp| \geq \epsilon m\} \leq 1/(4\epsilon^2 m)$. Choose M finite but so large that $1/(4\epsilon^2 M) < \delta/2$. By (4.3), apart from $\delta 2^k$ strings $s \in C_k$, $N_s > M$. Given $\{N_s\}$, the X_s are independent $\text{bin}[N_s, f_k(s)]$ random variables. Bernstein's inequality—with no Poissonization needed—completes the argument, as in (4.3): There are another $\delta 2^k$ exceptional strings, and setting them aside, $|\widehat{p}_s - f_k(s)| < \epsilon$. ∎

The early zone: $k \leq K_1$

Let

$$L_{k,n} = \frac{1}{n} \log \rho_{k,n} = \frac{1}{n} \sum_{s \in C_k} \log \phi\,(N_s, X_s, \gamma_s). \tag{4.5}$$

We also need the entropy function:

$$\begin{aligned} H(p) &= p \log p + (1-p)\log(1-p) && \text{for } 0 < p < 1 \\ &= 0 && \text{for } p = 0 \text{ or } 1. \end{aligned} \tag{4.6}$$

(4.7) LEMMA. *Suppose (1) and (4). Suppose the π_k are Γ-uniform. Fix k. Then*

$$\lim_{n\to\infty} L_{k,n} = \int H(f_k)\, d\lambda^\infty \quad \textit{almost surely } [P_f].$$

Proof. This like (4.12) in DF93. Since k is fixed, C_k is finite. We have $N_s \approx n/2^k$ almost surely by the ordinary strong law: see (2.3a). And $\widehat{p}_s \to f_k(s)$ by the strong law or (4.2b). By (3.2-3c) in DF93,

$$\frac{1}{n} \log \phi\,(N_s, X_s, \gamma_s) \to \frac{1}{2^k} H\,[f_k(s)] \quad \text{a.s.} \qquad ∎$$

The end zone: $\log_2 n - K_3 \leq k \leq \log_2 n + K_4$.

(4.8) LEMMA. *Suppose (1) and (4). Suppose the π_k are Γ-uniform. Fix any positive integers K_3, K_4. There is a positive $\rho < 1$, a finite positive constant A, and a small positive δ (all depending on K_3, K_4) such that, for all n, for all k with $\log_2 n - K_3 \leq \log_2 n + K_4$,*

$$P_f\left\{L_{k,n} \geq \int H(f_k) d\lambda^\infty - \delta\right\} < A\sqrt{n}\,\rho^n.$$

Proof. The argument proceeds by Poissonization, as in (4.3). For the moment, fix k. Recall that $\lambda = n/2^k$.

(4.9a) Let N_s^* be iid Pois(λ) for $s \in C_k$.

(4.9b) Given $\{N_s^*\}$, let the X_s^* be independent bin $[N_s^*, f_k(s)]$.

Let

$$Y_s^* = \log \int_0^1 \theta^{X_s^*} (1-\theta)^{N_s^* - X_s^*} \gamma_s(\theta)\, d\theta. \tag{4.9c}$$

It suffices to prove

$$\Pr\left\{ \frac{1}{n} \sum_{s \in C_k} Y_s^* \geq \int H(f_k) - \delta \right\} < \rho^n. \tag{4.10}$$

Choose L^* with $2 \leq L^* < \infty$. We claim:

$$E\{Y_s^* \mid N_s^*\} \leq N_s^* H(f_k(s)). \tag{4.11a}$$

(4.11b) There is a positive ϵ (which depends on L^* but not k or n) such that

$$E\{Y_s^* \mid N_s^*\} \leq N_s^* H(f_k(s)) - \epsilon N_s^* \text{ on } \{2 \leq N_s^* \leq L^*\}.$$

These results follow from (3.8) in DF93. Thus,

$$E\{Y_s^*\} \leq \lambda H(f_k(s)) - \epsilon \Pr\{2 \leq N_s^* \leq L^*\}.$$

Because $2^{-K_4} \leq \lambda \leq 2^{K_3}$, $\Pr\{2 \leq N_s^* \leq L^*\}/\lambda$ is bounded above and below. There is a small positive ϵ', which does not depend on k or n, such that

$$E\{Y_s^*\} \leq \lambda [H(f_k(s)) - \epsilon']. \tag{4.12}$$

We can now use Bernstein's inequality (3.2). Indeed, by the definition of Γ-uniformity, $\gamma_s \geq b > 0$; see (7) in DF93. And

$$-N_s^* + \log b < Y_s^* < 0,$$

by (3.3d) in DF93. Furthermore, the Y_s^* are independent. We take $m = 2^k$, $\xi = \text{Pois}(2^{K_3}) + 2\log b + 2^{K_3}$, $y = \epsilon'' n$, where ϵ'' is fixed with $0 < \epsilon'' < \min\{\epsilon', h_1/2^{K_3}\}$, so $\epsilon'' < \epsilon'$ and $y < h_1 m$. See (3.3) to motivate the definition of ξ. Then

$$\Pr\left\{ \sum_{s \in C_k} Y_s^* \geq \sum_{s \in C_k} E\{Y_s^*\} + \epsilon'' n \right\} < r^n \tag{4.13}$$

where $r = \exp(-Cn/m)$ and $C = \epsilon''^2/2\sigma^2$. But $n/m = n/2^k \geq 2^{-K_4}$. We take $\rho = \exp(-C2^{-K_4})$. Combine (4.12-13):

$$\Pr\left\{\frac{1}{n}\sum_{s\in C_k} Y_s^* \geq \int H(f_k) - \epsilon' + \epsilon''\right\} < \rho^n. \tag{4.14}$$

We take $\delta = \epsilon' - \epsilon'' > 0$. This proves (4.10). ∎

(4.15) COROLLARY. *Suppose (1) and (4). Suppose the π_k are Γ-uniform. Fix positive integers K_3, K_4. There is a small positive δ (depending on K_3, K_4) such that, almost surely, for all sufficiently large n, for all k with $\log_2 n - K_3 \leq k \leq \log_2 n + K_4$,*

$$L_{k,n} < \int H(f_k)\, d\lambda^\infty - \delta.$$

This completes our discussion of the end zone.

The high zone: $\log_2 n + K_4 \leq k$

Let

$$H(p,\theta) = p\log\theta + (1-p)\log(1-\theta). \tag{4.16}$$

The relative entropy function H is left undefined at the corners $p = \theta = 0$ or 1, where it has singularities.

Let $s \in S_k$ iff $s \in C_k$ and $N_s = 1$: the "S" is for "singly-occupied." Let

$$S_{k,n} = \sum_{s\in S_k} \log\phi(N_s, X_s, \gamma_s). \tag{4.17}$$

In other words, $S_{k,n}$ represents the sum defining $L_{k,n}$, extended only over the singly-occupied s. Since $0 < \phi < 1$, $L_{k,n} \leq S_{k,n}/n$.

From the definition of Γ-uniformity, given as (7) in DF93, g_s is the mean of γ_s; if $k > k_1$ and $s \in C_k$, $g_s = g_\infty(s)$, the function g_∞ being constant on each s in C_k.

(4.18) LEMMA. *Suppose (1) and (4). Suppose the π_k are Γ-uniform. For any positive δ, there is a positive $\rho < 1$ and a positive integer K_4 (both depending on δ) and a finite positive constant A such that for all n and all $k \geq \log_2 n + K_4$,*

$$P_f\left\{S_{k,n} \geq n\left[e^{-\lambda}\int H(f_k, g_\infty)\, d\lambda^\infty + \delta\right]\right\} < A\sqrt{n}\,\rho^{n^2/2^k}.$$

Proof. The argument is by Poissonization. Define N_s^*, X_s^*, and Y_s^* as in (4.9). It suffices to prove

$$\Pr\left\{\sum_{s\in C_k}\widetilde{Y}_s \geq n\left[e^{-\lambda}\int H(f_k, g_\infty)\, d\lambda^\infty + \delta\right]\right\} < \rho^{n^2/2^k}, \tag{4.19}$$

where $\widetilde{Y}_s = Y_s^*$ when $N_s^* = 1$, and $\widetilde{Y}_s = 0$ elsewhere. Now, for $N_s^* = 1$,

$$\begin{aligned}\widetilde{Y}_s &= X_s^* \log g_s + (1 - X_s^*) \log (1 - g_s) \\ &= X_s^* \log g_\infty(s) + (1 - X_s^*) \log (1 - g_\infty(s)) .\end{aligned}$$

In particular,

$$E\left\{\widetilde{Y}_s | N_s^* = 1\right\} = H\left[f_k(s),\ g_\infty(s)\right].$$

Since $\Pr\{N_s^* = 1\} = \lambda e^{-\lambda}$, and $\lambda = n/2^k$,

$$\begin{aligned} E\left\{\sum_{s \in C_k} \widetilde{Y}_s\right\} &= \lambda e^{-\lambda} \sum_{s \in C_k} H\left[f_k(s),\ g_\infty(s)\right] \\ &= n e^{-\lambda} \int H\left[f_k(s),\ g_\infty(s)\right] \end{aligned} \tag{4.20}$$

The function g_∞ is bounded between b, B, with $0 < b < B < 1$, again by definition. So the random variables $\widetilde{Y}_s$ are uniformly bounded, say by C. We use Bernstein's inequality (3.2) with $\xi = 2C$, $y = n\delta$, $m = 2^k$:

$$\Pr\left\{\sum_{s \in C_k} \widetilde{Y}_s \geq n\left[e^{-\lambda} \int H\left(f_k, g_\infty\right) d\lambda^\infty + \delta\right]\right\} < \exp\left[-\frac{n^2\delta^2}{2\sigma^2 2^k}\right]. \tag{4.21}$$

The condition $y \leq h_1 m$ is satisfied if K_4 is large enough. This proves (4.19), with $\rho = \exp(-\delta^2/2\sigma^2)$. ∎

(4.22) LEMMA. *Suppose (1) and (4). Suppose the π_k are Γ-uniform, and $f \not\equiv g_\infty$. There is a small positive δ and a large positive integer K_4 such that, almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 n + K_4 \leq k$,*

$$L_{k,n} < \int H(f)\, d\lambda^\infty - \delta.$$

Proof. $L_{k,n} \leq S_{k,n}/n$, and $S_{k,n}$ decreases with increasing k. (Eventually, $S_{k,n}$ stabilizes.) The reason is that S_k, the set of singly-occupied cells, increases with k. Thus, it suffices to consider the least $k \geq \log_2 n + K_4$, We must show that almost surely, for all sufficiently large n, for the least $k \geq \log_2 n + K_4$,

$$\frac{S_{k,n}}{n} < \int H(f)\, d\lambda^\infty - \delta. \tag{4.23}$$

We choose $\delta > 0$ so small that $\int H(f, g_\infty) d\lambda^\infty < \int H(f) d\lambda^\infty - 4\delta$. Now $f_k \to f$, so for K_4 large and $k \geq \log_2 n + K_4$,

$$\int H(f_k, g_\infty)\, d\lambda^\infty \leq \int H(f, g_\infty) d\lambda^\infty + \delta.$$

But $\lambda = n/2^k \leq 1/2^{K_4}$; H is negative; for K_4 large,

$$e^{-\lambda} \int H(f_k, g_\infty)\, d\lambda^\infty + \delta < e^{-\lambda} \left[\int H(f, g_\infty)\, d\lambda^\infty + \delta \right] + \delta$$

$$< \int H(f, g_\infty)\, d\lambda^\infty + 3\delta < \int H(f) d\lambda^\infty - \delta.$$

Now (4.18) proves (4.23), because

$$\sum_{n=1}^{\infty} \sqrt{n}\, \rho^{n^2/2^k} < \sum_{n=1}^{\infty} \sqrt{n}\, \rho^{n/2^{K_4}} < \infty. \quad \blacksquare$$

Discussion. For this part of the argument, we do not need that F, the set of prior means, is finite; we do need $b \leq \gamma \leq B$. We also do not need that $g_k \equiv g_\infty$ for all large k; uniform convergence would be enough, or even convergence in measure. Finally, we do not need that g_∞ is finitary, continuous, etc.

We fix $\delta > 0$ and choose K_4 large to control the high zone by an entropy rate argument. For any choice of K_3, K_4, the end zone goes away. We choose K_3, K_2 large to control the upper midzone, in the sense of showing that $\widetilde{\pi}_{k,n}$ will be close to f_k and hence f: see (4.3-4). This may be inefficient because the upper midzone is probably irrelevant. For any K_2, we get consistency in the lower midzone by (4.2); and likewise for the early zone, if $f = f_k$. Details are omitted because they parallel DF93. This concludes our discussion of Theorem 5.

5. The Proof of Theorem 6

The argument is more delicate; the rate of convergence of g_k to g_∞ matters, and so does the behavior of g_∞. We assume that

(5.1) $$f_k = f = p \text{ for all } k; \text{ and } g_k = g_\infty = p \text{ for } k > k_1.$$

The high zone splits:

Early high zone:

$$\log_2 n + K_4 \leq k \leq 2 \log_2 n - \log_2 \log n - K_5.$$

Middle high zone:

$$2 \log_2 n - \log_2 \log n - K_5 \leq k \leq 2 \log_2 n - \log_2 \log n + K_6.$$

Late high zone:

$$2 \log_2 n - \log_2 \log n + K_6 \leq k \leq 3.1 \log_2 n.$$

Very late high zone:

$$3.1 \log_2 n \leq k.$$

We now give some heuristics for the early zone, lower midzone, and upper midzone; that is, for $k \leq \log_2 n - K_3$:

(5.2a) $$\log \rho_{k,n} \doteq \sum_{s \in C_k} \left[N_s H(\widehat{p}_s) - \frac{1}{2} \log N_s \right]$$

and

(5.2b) $$\sum_{s \in C_k} N_s H(\widehat{p}_s) \doteq nH(p) + T_n + H'(p) Q_{k,n}$$

where

(5.3a) $$T_n = H'(p) \sum_{i=1}^{n} (\eta_i - p)$$

and

(5.3b) $$Q_{k,n} = \sum_{s \in C_k} N_s (\widehat{p}_s - p)^2 .$$

Furthermore,

$$\sum_{s \in C_k} \log N_s \doteq 2^k \log (n/2^k).$$

(The expression $n/2^k$ represents the number of observations per parameter.) To sum up,

$$\log \rho_{k,n} \doteq nH(p) + T_n - \frac{1}{2} 2^k \log \left(\frac{n}{2^k} \right).$$

The class of theories k with $k \leq \log_2 n - K_3$ is dominated by the smallest k with positive prior weight, namely, theory ℓ. (In the upper midzone, another nuisance term appears in the expansion; but the argument goes through anyway.) The end zone goes away by previous arguments. The early and middle high zones can also be eliminated.

The late and very late high zones remain, and the term in $2^k \log (n/2^k)$ drops out:

$$\log \rho_{k,n} \doteq nH(p) + T_n.$$

Therefore, late theories compete—on entropy grounds—with theory ℓ. It is the rate of decay of the theory weights w_k which decides the issue. The competitive late zone starts more or less at $k = 2 \log_2 n$, when there are $1/n$ data points per parameter. In DF93, the cutoff was 1 data point per parameter; the extra randomness in N_s helps the Bayesian statistician, and changes the critical rate for w_k from $[1/\sqrt{2}]^k$ to $[1/\sqrt[4]{2}]^k$.

Now for the details. We begin by showing that $Q_{k,n}$ is small relative to $2^k \log (n/2^k)$, provided $k \leq \log_2 n - K_3$.

(5.4) LEMMA. *Define $Q_{k,n}$ by (5.3b). For each n, $Q_{k,n}$ increases with k.*

Proof. Use Jensen's inequality. ∎

(5.5) LEMMA. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Let K_1 be an arbitrary positive integer. Almost surely $[P_f]$, for all sufficiently large n, for all $k \leq K_1$, $Q_{k,n} < 2^k \cdot 2 \cdot \log \log n$.*

Proof. Use the law of the iterated logarithm. ∎

(5.6) LEMMA. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Define σ^2 and h_1 as in (3.4). Fix $B > 2$. There is a large positive integer K_2 (depending on B) such that: almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \le k \le \log_2 n$,*

$$Q_{k,n} < p(1-p)2^k + \sigma\sqrt{B2^k \log n}.$$

Proof. This is immediate from (3.4), with $m = 2^k$ and $y = \sigma\sqrt{B2^k \log n}$. The test sum for the Borel Cantelli lemma is at most

$$\sum_n \frac{(\log_2 n)}{n^{B/2}} < \infty.$$

And the condition $y \le h_1 m$ is satisfied if K_2 is large enough. ∎

(5.7) LEMMA. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Fix $\delta > 0$. Choose K_2 as in (5.6). There is a large positive integer K_3 (depending on δ) such that: almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \le k \le \log_2 n - K_3$,*

$$Q_{k,n} \;<\; \delta 2^k \log_2\left(\frac{n}{2^k}\right).$$

Proof. Suppose first that $\log_2 \log n + K_2 \le k \le \frac{1}{2}\log_2 n$. Write $a_n << b_n$ iff $a_n/b_n \to 0$. Then

$$p(1-p)2^k + \sigma\sqrt{B2^k \log n} \;<< \frac{1}{2}\, 2^k \log_2 n \;\le\; 2^k \log_2\left(\frac{n}{2^k}\right).$$

Suppose next that $\frac{1}{2}\log_2 n \le k \le \log_2 n - K_3$. Then

$$p(1-p)2^k + \sigma\sqrt{B2^k \log n} \;\le\; \delta 2^k \log_2\left(\frac{n}{2^k}\right)$$

provided K_3 is large. Indeed, $p(1-p) \le 1/4$ and $\log_2(n/2^k) \ge K_3$ which is large, taking care of the term $p(1-p)2^k$. Finally, $\sigma\sqrt{B2^k \log n} << 2^k$. ∎

(5.8) LEMMA. *Assume (1), (4) and (5.1). Suppose the π_k are Γ-uniform. Fix $\delta > 0$. Choose K_2 as in (5.6). There is a large positive integer K_1 (depending on δ) such that: almost surely $[P_f]$, for all sufficiently large n, for all k with $K_1 \le k \le \log_2 \log n + K_2$,*

$$Q_{k,n} < \delta 2^k \;\log_2\left(\frac{n}{2^k}\right).$$

Proof. Let k_n be the least positive integer which is $\log_2 \log n + K_2$ or more. Now

$$\begin{aligned}
Q_{k,n} &\le Q_{k_n,n} && \text{by (5.4)}\\
&< p(1-p)2^{k_n} \;+\; \sigma\sqrt{B2^{k_n} \log n} && \text{by (5.6)}\\
&\le \left[p(1-p)2^{K_2+1} \;+\; \sigma\sqrt{B2^{K_2+1}}\,\right] \log n\\
&< \delta 2^k \log_2\left(\frac{n}{2^k}\right)
\end{aligned}$$

for k with $K_1 \le k \le \log_2 \log n + K_2$, provided K_1 is large. (The 1st and 3rd inequalities hold for all n; the 2nd and 4th for n large.) ∎

(5.9) COROLLARY. *Assume (1), (4) and (5.1). Suppose the π_k are Γ-uniform. Fix $\delta > 0$. Choose K_3 as in (5.7). Almost surely $[P_f]$, for all sufficiently large n, for all k with $k \le \log_2 n - K_3$,*

$$Q_{k,n} < \delta 2^k \log_2 \left(\frac{n}{2^k}\right).$$

Note: From here on, K_3 is forced large; but K_1, K_2 are free again.

Proof. Combine (5.5), (5.7), and (5.8). ∎

This completes the discussion of $Q_{k,n}$, and we turn to the term $\sum_{s \in C_k} \log N_s$ in the expansion (5.2a) of $\log \rho_{k,n}$. The sum is $[1 + o(1)]2^k \log \lambda$, where $\lambda = n/2^k$ as in (4.1). The main technique is Poissonization to approximate the ball-dropping distribution (2.3a). Unfortunately, there are zones which do not quite match those previously defined. We begin with $k \le (\log_2 n)/4$.

(5.10) LEMMA. *Assume (4). Fix $\delta > 0$. For all n, all $k \le (\log_2 n)/2$, and all $s \in C_k$,*

a) $P_f\{N_s/\lambda \ge 1 + \delta\} < \exp(-\delta^2 \sqrt{n}/2)$

b) $P_f\{N_s/\lambda \le 1 - \delta\} < \exp(-\delta^2 \sqrt{n}/2)$.

Proof. As (2.3a) shows, N_s is $\text{bin}(n, 1/2^k)$. Now use (3.1). Of course, $\lambda^2/2^k = n^2/2^{3k} \ge \sqrt{n}$ since $k \le (\log_2 n)/2$. ∎

(5.11) LEMMA. *Assume (4). Fix $\delta > 0$. Almost surely $[P_f]$, for all sufficiently large n, and all $k \le (\log_2 n)/2$,*

$$\left| \sum_{s \in C_k} (\log N_s - \log \lambda) \right| < \delta 2^k \log \lambda.$$

Proof. By (5.10) and the Borel Cantelli lemma, $1 - \delta \le N_s/\lambda \le 1 + \delta$ for all $s \in C_k$ and all $k \le (\log_2 n)/2$, for all sufficiently large n, almost surely: the test sum is bounded by

$$2 \sum_n \sum_{k=0}^{(\log_2 n)/2} 2^k \exp\left(-\delta^2 \sqrt{n}/2\right) < 4 \sum_n \sqrt{n} \exp\left(-\delta^2 \sqrt{n}/2\right) < \infty.$$

Finally, $k \le (\log_2 n)/2$ entails

$$2^k |\log(1 \pm \delta)| << 2^k \log\left(\frac{n}{2^k}\right). \quad ∎$$

We turn now to larger k; the lower end point of range is not material, but $\log_2 \log n$ is a convenient cut-point.

(5.12) LEMMA. *Assume (4). For $s \in C_k$, let N_s^* be independent Pois(λ) variables. Let $\tilde{N}_s = N_s^*$ when $N_s^* > \lambda e^{-1/2}$, else let $\tilde{N}_s = \lambda e^{-1/2}$.*

a) *Fix $B > 2$. There is a positive integer K_2 so large (depending on B) that for all n and all $k \geq \log_2 \log n + K_2$,*

$$\Pr\left\{ \left| \sum_{s \in C_k} \left[\log \tilde{N}_s - E\{\log \tilde{N}_s\} \right] \right| \geq B\sigma 2^k \sqrt{\frac{\log n}{n}} \right\} < \frac{2}{n^{B^2/2}}.$$

b) *Fix $\delta > 0$ and $C > 2$. There are positive integers K_2, K_3 so large (depending on δ and C) that for all n and all k with $\log_2 \log n + K_2 \leq k \leq \log_2 n - K_3$, the chance that $N_s^* \leq \lambda e^{-1/2}$ for $\delta 2^k$ or more indices $s \in C_k$ is bounded above by $1/n^C$.*

Proof. Claim a). This follows from (3.11), with $m = 2^k$, all $\lambda_i = \lambda = n/2^k$, and $y = B\sigma\sqrt{2^k \log n}$. The condition $y \leq h_1 m$ is satisfied if K_2 is large.

Claim b). From (3.5b), $\Pr\{N_s^* \leq \lambda e^{-1/2}\} < e^{-\lambda/16} < \delta/2$ provided K_3 is large; indeed, $\lambda \geq 2^{K_3}$. The chance that $\delta 2^k$ or more of these unlikely events occur can be bounded above by (3.1b). The bound is

$$\exp\left(-\frac{\delta^2 2^k}{8}\right) \leq \exp\left(-\delta^2 2^{K_2 - 3} \log n\right)$$

because $2^k/8 \geq 2^{K_2-3} \log n$. ∎

(5.13) LEMMA. *Assume (4). Let $N_s' = N_s$ when $N_s > \lambda e^{-1/2}$, else let $N_s' = \lambda e^{-1/2}$. Fix $\delta > 0$; choose K_2, K_3 as in (5.12).*

a) *Almost surely $[P_f]$, for all sufficiently large n, and all k with $\log_2 \log n + K_2 \leq k \leq \log_2 n - 1$,*

$$\left| \sum_{s \in C_k} (\log N_s' - \log \lambda) \right| < \delta 2^k \log \lambda.$$

b) *Almost surely $[P_f]$, for all sufficiently large n, and all k with $\log_2 \log n + K_2 \leq k \leq \log_2 n - K_3$,*

$$0 \leq \sum_{s \in C_k} (\log N_s' - \log N_s) < \delta 2^k \log \lambda.$$

Proof. Claim a). We de-Poissonize (5.12a):

$$P_f\left\{ \left| \sum_{s \in C_k} \left[\log N_s' - E\{\log \tilde{N}_s\} \right] \right| \geq B\sigma 2^k \sqrt{\frac{\log n}{n}} \right\} < \frac{A}{n^{(B^2-1)/2}}. \tag{5.14}$$

By (5.14) and the Borel Cantelli lemma, almost surely, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \leq k \leq \log_2 n - 1$,

$$\left| \sum_{s \in C_k} \left[\log N_s' - E\{\log \tilde{N}_s\} \right] \right| < B\sigma 2^k \sqrt{\frac{\log n}{n}};$$

indeed, the test sum is bounded above by

$$A\sum_n \frac{(\log_2 n)}{n^{(B^2-1)/2}} \;<\; \infty.$$

Since $k \le \log_2 n - 1$,

$$B\sigma 2^k \sqrt{(\log n)/n} \;<< 2^k \log_2\left(\frac{n}{2^k}\right).$$

Now use (3.12).

Claim b). We de-Poissonize (5.12b). Let $s \in S_k$ iff $s \in C_k$ and $N_s \le \lambda e^{-1/2}$: the S is for "small". Write $|S_k|$ for the cardinality of S_k. Now $P_f\{|S_k| \ge \delta 2^k\} < A/n^{C-.5}$. There are at most $\log_2 n$ indices k to consider, and $\sum(\log_2 n)n^{C-.5} < \infty$ because $C > 2$. Thus, almost surely, for all sufficiently large n, for all k with $\log_2\log n + K_2 \le k \le \log_2 n - K_3$, $|S_k| < \delta 2^k$.

If $s \notin C_k$, then $N'_s = N_s$. Now suppose $s \in S_k$. If $N_s = 0$, then $\log N_s = 0$ by definition. Thus,

$$0 \;\le\; \log N'_s - \log N_s \;\le\; \log\lambda - \frac{1}{2} < \log\lambda.$$

Consequently,

$$0 \;\le\; \sum_{s\in C_k} (\log N'_s - \log N_s) < |Q_k| \log\lambda < \delta 2^k \log\lambda. \qquad \blacksquare$$

(5.15) Remark. Assume (4). Fix $L > 6$. Almost surely, for all sufficiently large n, for all k with $k \le \log_2 n - \log_2\log n - L$, and all $s \in C_k$, $N'_s = N_s$.

Proof. This follows from (3.5b) and Poissonization:

$$P\{N_s \;\le\; \lambda e^{-1/2}\} \;\le\; A\sqrt{n}\, e^{-\lambda/16} \;\le\; \frac{A}{n^C}$$

The test sum for the Borel Cantelli lemma is bounded above by

$$A\sum_n \sum_{k=0}^{\log_2 n} \frac{2^k}{n^C} < 2A\sum_n \frac{1}{n^{C-1}} < \infty. \qquad \blacksquare$$

(5.16) COROLLARY. *Assume (4). Fix $\delta > 0$. Almost surely $[P_f]$, for all sufficiently large n, for all k with $k \le \log_2 n - K_3$,*

$$\Big|\sum_{s\in C_k} (\log N_s - \log\lambda)\Big| \;<\; \delta 2^k \log\lambda.$$

Proof. For k with $\log_2\log n + K_2 \le k \le \log_2 n - K_3$, use (5.13). For $k \le \log_2\log n + K_2$, use (5.11). $\blacksquare$

In the early zone and lower midzone, $k \le \log_2\log n + K$; then $\widehat{p}_s$ is nearly p: see (4.2). In these zones, we can estimate $\log \rho_{k,n}$, as follows.

(5.17) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Define T_n by (5.3a). Fix $\delta > 0$ and $K < \infty$. Almost surely, for all sufficiently large n, for all k with $0 \le k \le \log_2 \log n + K$,*

$$\left| \log \rho_{k,n} - nH(p) - T_n + \frac{1}{2} 2^k \log\left(\frac{n}{2^k}\right) \right| < \delta 2^k \log\left(\frac{n}{2^k}\right).$$

Proof. We estimate $\log \rho_{k,n}$ by (3.3) in DF93, making (5.2a) rigorous by adding $O(2^k) = o(2^k \log(n/2^k))$. Now

$$\sum_{s \in C_k} N_s H(\widehat{p}_s)$$

can be expanded around p by (3.14) in DF93. The lead term is $nH(p)$. The linear term gives T_n, after a bit of algebra. The quadratic remainder is negligible by (5.9). Finally,

$$\frac{1}{2} \sum_{s \in C_k} \log N_s$$

can be estimated by (5.16). ∎

In the upper midzone, N_s may be 0 for some s. The corresponding terms contribute 0 to the sum defining $\log \rho_{k,n}$. Even if $N_s > 0$, $\widehat{p}_s$ may be 0 or 1. This necessitates some additional nuisance terms in the expansion of $\log \rho_{k,n}$, because the approximation to $\phi(m, j, \gamma)$ changes form when $j = 0$ or m. See (3.2) and (5.4) in DF93.

(5.18a) Let N_k be the number of $s \in C_k$ with $N_s > 0$ and $X_s = 0$ or N_s.

(5.18b) Let $s \in G_k$ iff $0 < X_s < N_s$.

(5.18c) Let $\Xi_{k,n} = -\frac{1}{2} \log(n/2^k) N_k + \sum_{s \in G_k} \log \sqrt{\widehat{p}_s(1 - \widehat{p}_s)}$.

All terms in $\Xi_{k,n}$ are negative, because $0 < \widehat{p}_s < 1$.

(5.19) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Fix $\delta > 0$ and $K < \infty$. Define K_3 as in (5.16). Almost surely $[P_f]$, for all sufficiently large n, for all k with $\log_2 \log n + K_2 \le k \le \log_2 n - K_3$,*

$$\left| \log \rho_{k,n} - nH(p) - T_n + \frac{1}{2} 2^k \log\left(\frac{n}{2^k}\right) - \Xi_{k,n} \right| < \delta 2^k \log\left(\frac{n}{2^k}\right).$$

Proof. This is argued like (5.17). ∎

This completes the discussion of the early zone and midzone. The end zone goes away by (4.15), and we turn to the high zone.

The early high zone

The early high zone is defined by the condition

(5.20) $$\log_2 n + K_4 \le k \le 2 \log_2 n - \log_2 \log n - K_5.$$

K_4 defines the right edge of the end zone, but from our perspective it is a free parameter: (4.15) imposed no condition on K_4. For present purposes too, K_4 is not really material; we can set $K_4 = 3$. We will prove:

(5.21) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Fix a large positive number L. There is a large positive integer K_5 such that: almost surely $[P_f]$, for all sufficiently large n, for all k satisfying (5.20),*

$$\log \rho_{k,n} \ < \ nH(p) + T_n - L \log n.$$

Suppose $s \in C_k$. As in (4.17), let $s \in S_{k,n}$ iff $N_s = 1$; likewise, $s \in M_{k,n}$ iff $N_s > 1$. The S is for "single occupancy" and M for "multiple occupancy;" the dependence on n will matter later. Write $i \in s$ iff $\xi_j(i) = s_j$ for $1 \le j \le k$; in other words, the first k covariates for subject i agree with s. Suppose k is so large that $g_k \equiv p$: see (5.1). A bit of algebra shows

(5.22) If $s \in S_{k,n}$ then $\log \phi(N_s, X_s, \gamma_s) = H(p) + (\eta_i - p)H'(p)$ for the unique $i \in s$.

For $0 \le j \le m$ and $m \ge 2$, let

$$\phi_0(m, j, \gamma) \ = \ \log \phi(m, j, \gamma) - mH(p) - (j - mp)H'(p). \tag{5.23}$$

For $s \in M_{k,n}$, let $\Delta_s = \phi_0(N_s, X_s, \gamma_s)$. By (5.22) and a bit more algebra,

$$\log \rho_{k,n} \ = \ nH(p) + T_n + \sum_{s \in M_{k,n}} \Delta_s. \tag{5.24}$$

To prove (5.21), we must estimate $\sum_{s \in M_{k,n}} \Delta_s$. The main technique is Poissonization, and here are some preliminaries. Recall from the definition (7) in DF93 of Γ-uniformity that $\gamma \in \Gamma$ entails $\gamma \ge b > 0$. The next result is immediate from (3.3d) in DF93.

(5.25) LEMMA. $|\phi_0(m, j, \gamma)| \le [1 + |H'(p)|]m + |\log b|$ *for $\gamma \in \Gamma$ with lower bound b.*

(5.26) Definition. Fix k. For $s \in C_k$, let N_s^* be Pois(λ), where $\lambda = n/2^k$. Given $\{N_s^*\}$, let $\{X_s^*\}$ be independent bin(N_s^*, p). *Let $\Delta_s^* = \phi_0(N_s^*, X_s^*, \gamma_s)$. Let M^* be the number of $s \in C_k$ with $N_s^* \ge 2$.*

Relationships (5.27-5.30) are obvious:

(5.27) $M^* = \sum_{s \in C_k} I_s^*$, where I_s^* is 0 if $N_s^* < 2$ and I_s^* is 1 if $N_s^* \ge 2$. The I_s^* are iid.

(5.28) $\frac{1}{2}\lambda^2(1 - \lambda) \le [1 - (1 + \lambda)e^{-\lambda}] \le \frac{1}{2}\lambda^2$ for all λ.

(5.29) $E\{M^*\} = 2^k[1 - (1 + \lambda)e^{-\lambda}]$.

(5.30) $\frac{1}{2}n\lambda(1 - \lambda) \le E\{M^*\} \le \frac{1}{2}n\lambda$ for all λ.

(5.31) LEMMA. *Fix δ with $0 < \delta < 1$. Suppose $0 < \lambda < \delta/2$.*

a) $P\{M^* \geq (1+\delta)n\lambda/2\} < \exp(-\delta^2 n\lambda/8)$.

b) $P\{M^* \leq (1-\delta)n\lambda/2\} < \exp(-\delta^2 n\lambda/16)$.

Proof. Claim a). This is Bernstein's inequality. Theorem (4) in Freedman [1973], coupled with the estimate (5.30) for $E\{M^*\}$, gives the bound

$$\exp\left[-\frac{1}{2}\frac{(\delta n\lambda/2)^2}{(1+\delta)n\lambda/2}\right] < \exp\left(-\frac{\delta^2 n\lambda}{8}\right)$$

because $0 < \delta < 1$.

Claim b) is similar. By (5.30), $n\lambda/2 \geq E\{M^*\}$ and $E\{M^*\}-(1-\delta)n\lambda/2 \geq n\lambda(\delta-\lambda)/2 > \delta n\lambda/4$, so the bound is

$$\exp\left[-\frac{1}{2}\frac{(\delta n\lambda/4)^2}{n\lambda/2}\right] < \exp\left(-\frac{\delta^2 n\lambda}{16}\right). \quad \blacksquare$$

Note. Lemma 3.1 is quite inefficient for small p, when $\sqrt{mx}$ would—ideally—be replaced by $\sqrt{mpx}$. Hence the resort to other estimates.

(5.32) LEMMA. *For i = 1,2,... let $\widetilde{\Delta}_i$ be independent, and distributed as Δ_s^* given $N_s^* \geq 2$. Define $\epsilon_2 > 0$ as in (3.8) of* DF93. Suppose $0 < \lambda < 1/2$.

a) $E\{\widetilde{\Delta}_i\} < -\epsilon_2$.

b) There is an $\epsilon > 0$ and σ^2 with $0 < \sigma^2 < \infty$ such that:
(i) For all λ with $0 < \lambda < 1/2$ and all $m = 1, 2, ...$

$$\Pr\left\{\sum_{i=1}^{m} \widetilde{\Delta}_i \geq -\epsilon m\right\} < \exp\left(\frac{-\epsilon^2 m}{2\sigma^2}\right);$$

(ii)

$$\Pr\left\{\sum_{s\in C_k} \Delta_s^* I_s^* \geq -\epsilon M^*\right\} < \exp\left(-\frac{\epsilon^2 n\lambda}{8\sigma^2}\right).$$

Proof. Claim a). By (3.8) in DF93, $E\{\Delta_s^* | N_s^* = m\} < -m\epsilon_m$ for $m \geq 2$. So $E\{\widetilde{\Delta}_i\} = E\{\Delta_s^* | N_s^* \geq 2\} < -2\epsilon_2 P\{N_s^* = 2\}/P\{N_s^* \geq 2\} < -2\epsilon_2(1-\lambda) < -\epsilon_2$, with the help of (5.28).

Claim b). Let $\xi = |\log b| + \{1 + |H(p)| + |H'(p)|\}N$, where N is Pois(1/2) conditioned to be 2 or more. By (5.25) and (3.25), $|\widetilde{\Delta}_i|$ is stochastically bounded by ξ, so (3.2) applies. Compute σ^2 and h_1 according to that lemma. Let $\epsilon = \min\{h_1, \epsilon_2/2\}$ and $\rho = \exp(-\epsilon^2/2\sigma^2)$. Now

$$\Pr\left\{\sum_{i=1}^{m} \widetilde{\Delta}_i \geq -\epsilon m\right\} < \Pr\left\{\sum_{i=1}^{m} \widetilde{\Delta}_i \geq \sum_{i=1}^{m} E\{\widetilde{\Delta}_i\} + \epsilon m\right\} < \exp\left(\frac{-\epsilon^2 m}{2\sigma^2}\right).$$

The first inequality holds because $E\{\widetilde{\Delta}_i\} < -2\epsilon$; the second, by (3.2): the condition $y_1 \leq h_1 m$ holds because $\epsilon \leq h_1$. This proves (i), and we turn to (ii).

Conditional on $M^* = m$, the sum is distributed as $\sum_{i=1}^{m} \widetilde{\Delta}_i$, giving the bound

$$E\left\{\exp\left(\frac{-\epsilon^2 M^*}{2\sigma^2}\right)\right\} < \exp\left(\frac{-\epsilon^2 E\{M^*\}}{2\sigma^2}\right)$$

by Jensen's inequality. But $\lambda < 1/2$ so $E\{M^*\} > n\lambda/4$ by (5.30). ∎

Note. In the proof of c), if you just think of $\sum_{s\in C_k} \Delta_s^* I_s^*$ as the sum of 2^k terms, (3.2) gives the disappointing bound $\exp[-\frac{\epsilon^2(n\lambda)^2}{8\sigma^2 2^k}] = \exp(-\epsilon^2 n\lambda^3/8\sigma^2)$.

Recall that $M_{k,n} = \{s : s \in C_k \text{ and } N_s \geq 2\}$.

(5.33) LEMMA. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Fix δ with $0 < \delta < 1/2$ and suppose $0 < \lambda < \delta/2$. Choose $\epsilon > 0$ as in (5.32b).*

a) $P_f\{|M_{k,n}| \geq (1+\delta)n\lambda/2\} < A\sqrt{n}\exp(-\delta^2 n\lambda/8)$.

b) $P_f\{|M_{k,n}| \leq (1-\delta)n\lambda/2\} < A\sqrt{n}\exp(-\delta^2 n\lambda/16)$.

c) $P_f\left\{\sum_{s\in M_{k,n}} \Delta_s \geq -\epsilon|M_{k,n}|\right\} < A\sqrt{n}\exp(-\epsilon^2 n\lambda/8\sigma^2)$.

Proof. Claims a) and b) follow from (5.31) by de-Poissonization. Claim c) is similar, starting from (5.32). ∎

(5.34) COROLLARY. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Suppose k and n satisfy (5.20).*

a) $P_f\{|M_{k,n}| \leq (1-\delta)n\lambda/2\} < A/n^{C-.5}$, *with* $C = \delta^2 2^{K_5-4}$.

b) $P_f\left\{\sum_{s\in M_{k,n}} \Delta_s \geq -\epsilon|M_{k,n}|\right\} < A/n^{D-.5}$, *with* $D = \epsilon^2 2^{K_5-3}/\sigma^2$.

Proof of Proposition 5.21. Fix $\delta < 1/2$ in (5.34a); we require $2^{-K_4} < \delta/2$ so $\lambda < \delta/2$ in (5.31). Choose ϵ as in (5.32b). Choose K_5 so large that $C > 2$ and $D > 2$ in (5.34). There are at most $\log_2 n$ theories in the zone. So, almost surely, for all sufficiently large n, for all k satisfying (5.20),

$$|M_{k,n}| > (1-\delta)n\lambda/2 > 2^{K_5-2}\log n$$

$$\sum_{s\in M_{k,n}} \Delta_s < -\epsilon|M_{k,n}| < -\epsilon 2^{K_5-2}\log n. \quad ∎$$

Remark. We have assumed in (5.1) that the mean of γ_s equals p, for $s \in C_k$ and $k > n_1$. Suppose that γ_s is constant, say at $\gamma \in \Gamma$ with $\int \theta\gamma(\theta)\,d\theta = p$. Then $\{\Delta_s^* : s \in M_{k,n}\}$ are iid for each k, with $E\{\Delta_s^*\} = E\{\phi_0(N_\lambda, X, \gamma)\}$; N_λ is Pois(λ) conditioned to be 2 or more; given $N_\lambda = m$, X is bin(m,p). See (5.23) and (5.26). The argument for (5.33) shows that $\sum\{\Delta_s : s \in M_{k,n}\} \approx \frac{1}{2}n\lambda E\{\phi_0(N_\lambda, X, \gamma)\}$.

This completes our discussion of the early high zone.

Middle high zone

The middle high zone is the most delicate of all the zones. It is defined by the condition

$$2\log_2 n - \log_2\log n - K_5 \;\le\; k \;\le\; 2\log_2 n - \log_2\log n + K_6. \tag{5.35}$$

Here, K_5 and K_6 are large positive integers; K_5 is needed to control the early high zone; K_6 will control the late high zone.

(5.36) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. There is a small positive ϵ_0 (depending on K_5, K_6) such that: almost surely, for all sufficiently large n, for all k satisfying (5.35),*

$$\log \rho_{k,n} < nH(p) + T_n - \epsilon_0 \log n.$$

At stage n of the trial, we have data on n subjects; let $D_{k,n}$ be the set of $s \in C_k$ with $N_s = 2$; the D is for "doubly occupied." The main difficulty is showing that $|D_{k,n}| \approx n\lambda/2$. The dependence on n matters, and is displayed in the notation. Since $n\lambda$ is of order $\log n$, exponential bounds must be supplemented by passing to geometric subsequences, and the $\sqrt{n}$ for de-Poissonization cannot be afforded. We solve the latter problem first.

(5.37) LEMMA. *At stage n, let $D_{k,n}$ be the set of $s \in C_k$ with $N_s = 2$, and let $M_{k,n}$ be the set of $s \in C_k$ with $N_s \ge 2$. Fix δ with $0 < \delta < 1$. Fix positive integers K_5, K_6. Almost surely, for all sufficiently large n, for all k satisfying (5.35),*

a) $(1-\delta)n\lambda/2 < |M_{k,n}| < (1+\delta)n\lambda/2$

b) $(1-\delta)n\lambda/2 < |D_{k,n}| < (1+\delta)n\lambda/2$.

Note. λ depends on k and n, and $n\lambda = n^2/2^k$: see (4.1).

Proof. Claim a). Fix r slightly bigger than 1 and consider the sequence r^j. For each k, $|M_{k,n}|$ increases with n. As n increases from r^j to r^{j+1}, $n\lambda$ increases from $r^{2j}/2^k$ to $r^{2j+2}/2^k$, i.e., only by a factor of r^2. Thus, it suffices to prove claim a) for n of the form r^j. Recall S_n from (3.22). By (3.20), $|M_{k,n}| = S_n$ or $S_n - 1$; and it is enough to prove the claim for S_n and n of the form r^j. That is immediate from the Borel Cantelli lemma and (3.24) with $n = r^j$ and $b = 2^k$.

Claim b) follows from a), because $|D_{k,n}| = |M_{k,n}|$ or $|M_{k,n}| - 1$, by (3.20). ∎

Recall the function $\phi_0(m, j, \gamma)$ from (5.23). Let $\mathcal{H}$ be the class of random variables distributed as $\phi_0(2, X, \gamma)$ where X is bin$(2, p)$ and $\gamma \in \Gamma$. If $Y \in \mathcal{H}$, then Y is uniformly bounded by (5.25), and $E\{Y\} < -2\epsilon_2 < 0$ by (3.8) in DF93. As (3.2) shows,

(5.38) LEMMA. *There are positive constants h_1 and σ^2, depending only on $\mathcal{H}$, such that: if $Y_i \in \mathcal{H}$ are independent for $i = 1, ..., m$, and $0 < y \le h_1 m$, then*

$$\Pr\left\{\sum_{i=1}^{m} (Y_i - E\{Y_i\}) \ge y\right\} < \exp\left(-\frac{y^2}{2\sigma^2}\right).$$

Recall Δ_s, as defined for (5.24). Given $D_{k,n}$, $\{\Delta_s : s \in D_{k,n}\}$ are independent; and $\Delta_s \in \mathcal{H}$.

(5.39) LEMMA. *Define ϵ_2 as in (3.8) of* DF93; *h_1 and σ^2 as in (5.38). Let $0 < \epsilon < \min(h_1, \epsilon_2)$. Almost surely, for all sufficiently large n, for all n and k satisfying (5.35), $\sum\{\Delta_s : s \in D_{k,n}\} < -\epsilon(\log_2 n)/2^{K_6+4}$.*

Proof. First, consider only n of the form 2^j. By (5.38),

$$P_f\left\{\sum \{\Delta_s : s \in D_{k,n}\} \ge -\epsilon |D_{k,n}| \,|\, D_{k,n}\right\} < \exp\left(-\frac{\epsilon^2 |D_{k,n}|}{2\sigma^2}\right). \tag{5.40}$$

Then $E\{\exp(-\epsilon^2|D_{k,n}|/2\sigma^2)\} < \exp(-\epsilon^2 E\{|D_{k,n}|\}/2\sigma^2)$ and $E\{|D_{k,n}|\} \approx 2^k\lambda^2/2 = n^2/2^{k+1} > (\log n)/2^{K_6+1} > j/2^{K_6+2}$. The Borel Cantelli lemma shows that almost surely, for all sufficiently large n of the form 2^j, for all k satisfying (5.35), $\sum\{\Delta_s : s \in D_{k,n}\} < -\epsilon|D_{k,n}|$. Indeed, the test sum is bounded by

$$(1 + K_5 + K_6) \sum_j e^{-C_0 j} < \infty,$$

where $C_0 = \epsilon^2/2^{K_6+3}\sigma^2$. Finally, use (5.37) to bound $|D_{k,n}|$, noting that $n\lambda \ge (\log n)2^{K_6}$.

We must now interpolate between 2^j and 2^{j+1}; the argument is only sketched. Fix k. Then $\sum\{\Delta_s : s \in D_{k,n}\}$ is below $-\epsilon(\log n)/2^{K_6+2}$, say, when $n = 2^j$. As n increases from 2^j to 2^{j+1}, 2^j additional balls are dropped at random into the 2^k boxes, perturbing the sum. We claim that almost surely, for all sufficiently large n, the perturbations will amount at most to $\epsilon(\log n)/2^{K_2+3}$, so

$$\sum \{\Delta_s : s \in D_{k,n}\} < -\epsilon(\log n)/2^{K_6+3}$$

for all n and k satisfying (5.35), with $2^j \le n \le 2^{j+1}$, provided j is sufficiently large.

There are four kinds of perturbations: (i) an additional doubly-occupied box is created, adding an independent term $\Delta_i \in \mathcal{H}$; (ii) a triply-occupied box may be created; (iii) more than one triply-occupied boxes may be created; (iv) a box may become more than triply occupied. Perturbations (iii) and (iv) do not occur for large n, by (3.20), and need not be considered further. Perturbation (ii) changes the sum by a uniformly bounded amount; see (5.25).

We must now bound the effect of perturbations of type (i), showing they amount to less than $C_0 \log_2 n = C_0 j$, where $C_0 = \epsilon/2^{K_6+5}$; this leaves more than enough to absorb perturbations of type (ii). Now, dropping in 2^j balls increases $D_{k,n}$ from (essentially) $2^{2j}/2^{k+1}$ to $2^{2j+2}/2^{k+1}$, by (5.37); i.e., from cj to $4cj$. But, adding this number of Δ's— or any other— crosses the $C_0 j$ boundary with probability at most ρ^j, by (3.16). ∎

Proof of Proposition 5.36. Use (5.24), (3.20), and (5.39). ∎

Remark. We have assumed in (5.1) that the mean of γ_s equals p, for $s \in C_k$ and $k > n_1$. Suppose that γ_s is constant, say at $\gamma \in \Gamma$ with $\int \theta\gamma(\theta)\, d\theta = p$. Then $\{\Delta_s : s \in D_{k,n}\}$ are iid, with $E\{\Delta_s\} = E\{\phi_0(2, X, \gamma)\}$, X being bin(2,p): see (5.23). The argument for (5.39), pushed a little harder, shows that $\sum\{\Delta_s : s \in D_{k,n}\} \approx \frac{1}{2} n\lambda E\{\phi_0(2, X, \gamma)\}$; the idea is to split along the geometric sequence r^n with r just bigger than 1.

This completes our discussion of the middle high zone.

Late high zone

The late high zone is defined by the condition

$$2\log_2 n - \log_2 \log n + K_6 \;\le\; k \;\le\; 3.1\, \log_2 n. \tag{5.41}$$

(5.42) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Fix $\epsilon > 0$. There is a large positive integer K_6 (depending on ϵ) such that: almost surely, for all sufficiently large n, for all k satisfying (5.41),*

$$|\log \rho_{k,n} - nH(p) - T_n| < \epsilon \log n.$$

Proof. By (5.24), it is enough to bound $\sum\{\Delta_s : s \in M_{k,n}\}$. But (3.20) shows that $M_{k,n}$ (the set of multiply-occupied cells) differs from $D_{k,n}$ (the set of doubly-occupied cells) by at most one triply-occupied cell. So it is enough to bound $\sum\{\Delta_s : s \in D_{k,n}\}$ and hence $|D_{k,n}|$ by (5.25). For each n, $|D_{k,n}|$ decreases as k increases, so it is enough to consider k just larger than $2\log_2 n - \log_2 \log n + K_6$. Now (5.37) shows that almost surely, for all sufficiently large n,

$$|D_{k,n}| < n\lambda \;=\; \frac{n^2}{2^k} \;\le\; 2^{-K_6} \log n. \qquad ∎$$

Very late high zone

The very late high zone is defined by the condition

$$3.1\, \log_2 n \;\le\; k. \tag{5.43}$$

(5.44) PROPOSITION. *Assume (1), (4), and (5.1). Suppose the π_k are Γ-uniform. Almost surely, for all sufficiently large n, for all $k \ge 3.1 \log_2 n$,*

$$\log \rho_{k,n} \;=\; nH(p) + T_n$$

Proof. This is immediate from (3.21) and (5.24). ∎

Proof of Theorem 6

Claim a). By (5.17), with δ_1 any small positive number of our choice,

$$\widetilde{w}_{\ell,n} > w_\ell \exp\left[nH(p) + T_n - \frac{1}{2}\, 2^\ell \log\left(\frac{n}{2^\ell}\right) \;-\; \delta_1 2^\ell \log\left(\frac{n}{2^\ell}\right)\right]. \tag{5.45}$$

The random term T_n was defined by (5.3a), and is of order $\sqrt{n \log\log n}$ or less. We must now eliminate theories in the end zone and high zone.

Theories in the end zone ($\log_2 n - K_3$ to $\log_2 n + K_4$) are negligible relative to theory ℓ, by (4.15). The δ there depends on K_3, K_4; but no matter what that δ is, the end zone has entropy rate $H(p) - \delta < H(p)$.

The early high zone is defined by (5.20). For such theories, by (5.21).

$$\begin{aligned}\sum_k \widetilde{w}_{k,n} \;&<\; \left[\sum_{k=\log_2 n}^{\infty} w_k\right] \exp\left[nH(p) + T_n - L\log n\right] \\ &<\; \exp\left[nH(p) + T_n - L\log n - \frac{1}{4}\, 2^\ell \log n \left(-\frac{\delta_0}{\log 2}\right) 2^\ell \log n\right]\end{aligned}$$

using the condition of the theorem—with $\log_2 n = (\log n)/(\log 2)$ in place of n. (The sum of the high-zone weights starts at $\log_2 n$.) In total, the early high zone has negligible posterior weight, relative to theory ℓ, provided

$$L + \left(\frac{\delta_0}{\log 2}\right) 2^\ell \;>\; \frac{1}{4}\, 2^\ell + \delta_1 2^\ell.$$

But L can be made large by choosing K_5 large.

We combine the middle high zone, late high zone and very late high zone, i.e., we consider all theories

$$k \;\geq\; L(n) \;=\; 2\log_2 n - \log_2\log n - K_5.$$

The posterior weight in this combined zone is by (5.36), (5.42), and (5.44) at most

$$\begin{aligned}\sum_k \widetilde{w}_{k,n} \;&<\; \left[\sum_{k=L(n)}^{\infty} w_k\right] \exp\left[nH(p) + T_n + \epsilon \log n\right] \\ &<\; \exp\left[nH(p) + T_n + \epsilon\log n - \frac{1}{4}(\log 2)\; 2^\ell L(n) - \delta_0 2^\ell L(n)\right].\end{aligned} \tag{5.46}$$

Now

$$\frac{1}{4}(\log 2)\,2^\ell\,L(n) \;=\; \frac{1}{2}\,2^\ell \log n - C_0 \log\log n - C_1. \tag{5.47}$$

Again, this zone is negligible relative to theory ℓ, if we choose $\epsilon + \delta_1 2^\ell < (2\delta_0/\log 2)2^\ell$. The δ_1 in (5.45) is the δ of (5.17), and is at our disposition. The ϵ in (5.46) comes from (5.42). To make it small, we have to choose K_6 large. Choosing K_5 and K_6 large makes the ϵ_0 in (5.36) small. However, the value of ϵ_0 does not matter.

The balance of the argument for claim a) is omitted, being very similar to the argument for Theorem 5 in this paper, or Theorems 8 and 9 in DF93. Basically, posterior mass shifts into the early zone or midzone, where there are lots of observations per parameter.

We turn to claim b). Consider only n with

$$\sum_{k=n}^{\infty} w_k > \exp\left[-\frac{1}{4}(\log 2)n2^\ell + \delta_0 n 2^\ell\right].$$

We combine theories in the late and very late high zones, so

$$k \;\geq\; L(n) = 2\log_2 n - \log_2\log\, n + K_6.$$

By (5.42 and 5.44), the total posterior weight in these two zones is at least

$$\exp\left[nH(p) + T_n - \epsilon\log n - \frac{1}{4}(\log 2)2^\ell L(n) + \delta_0 2^\ell L(n)\right] \tag{5.48}$$

where ϵ is a small positive number at our disposition; δ_0 is fixed, by the conditions of the theorem. Of course,

$$\frac{1}{4}(\log 2)\;2^\ell\,L(n) \;=\; \frac{1}{2}\,2^\ell\log n - C_0\,\log\log n + C_1. \tag{5.49}$$

By comparison, the total posterior weight in the early zone and midzone ($k \leq \log_2 n - K_3$) is by (5.17 and 5.19) at most

$$\sum_{k=\ell}^{\log_2 n - K_3}\left[w_k \cdot \exp\left(nH(p) + T_n - \frac{1}{2}2^k\log\left(\frac{n}{2^k}\right) + \delta_1 2^k\,\log\left(\frac{n}{2^k}\right)\right)\right]$$

$$< \left[\sum_{k=\ell}^{\infty} w_k\right]\exp\left[nH(p) + T_n - \frac{1}{2}\,2^\ell\log\left(\frac{n}{2^\ell}\right) + \delta_1 2^\ell\,\log\left(\frac{n}{2^\ell}\right)\right]$$

where δ_1 is a small positive number, at our disposition. The term $\Xi_{k,n}$ in (5.19) was dropped, being negative: see (5.18c). The displayed inequality holds by (5.17) in DF93.

Compare (5.48 and 5.50): The early zone and midzone are negligible relative to the late and very late high zones, provided $\delta_1 2^\ell + \epsilon < 2\delta_0 2^\ell/\log 2$. It is the minor bit of algebra in (5.47 and 5.49) that seems to determine the critical rate of decay for the w's in Theorem 6.

The end zone goes away, as usual; the early high zone can also be eliminated. We do not know (or need to know) how much posterior mass is in the middle high zone. Informally, posterior weight shifts so far out that there are only $O(\log n/n)$ observations per parameter. The argument can be completed as in (5.18) in DF93. ∎

References

de Finetti, B. [1959]. La probabitità, la statistica, nei rapporti con l' induzione, secondo diversi punti di vista, *Centro Internazionale Matematica Estivo tremonese,* Rome (English translation in B. de Finetti [1972], *Probability, Induction Statistics,* Wiley, New York.

Diaconis, P. and Freedman, D. [1986]. On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1-67.

Diaconis, P. and Freedman, D. [1990]. On the uniform consistency of Bayes estimates for multinomial probabilities. *Ann. Statist.* **18** 317-327.

Diaconis, P. and Freedman, D. [1993]. Nonparametric binary regression: A Bayesian approach. *Ann. Statist.* **21**.

Dubins, L.E. and Savage, LJ. [1985] *How to Gamble if You Must: Inequalities For Stochastic Processes,* Dover, New York.

Feller, W. [1971]. *An Introduction to Probability Theory and its Applications.* Vol. II, 2nd ed., Wiley, New York.

Freedman, D. [1973]. Another note on the Borel-Cantelli lemma and the strong law with the Poisson approximation as a by-product. *Ann. Prob.* **6** 910-925.

Laplace, P.S. [1774]. Memoire sur la probabilité des causes par les évènements. *Memoires de mathématique et de physique presentés a l'Académe Royale des Sciences, par divers savants, et lûs dans ses assemblées.* **6** (English translation by S. Stigler [1986]. *Statist. Sci* **1** 359-378).