Rare-event estimation for Bayesian predictive rollouts: rollout Monte Carlo vs. posterior sampling (Rao–Blackwellization)

Vinayak Pathak

Abstract

We study the probability of a rare event A defined on the future outputs of a Bayesian predictive model after conditioning on observed data $y_{1:n}$. Two Monte Carlo strategies are compared: (i) direct sampling of future rollouts from the posterior predictive and counting hits of A (an indicator estimator), and (ii) sampling parameters θ from the posterior and computing $f(\theta) = \mathbb{P}(A \mid \theta, y_{1:n})$ analytically (a Rao–Blackwellized estimator). We give a self-contained multiplicative Chernoff bound for bounded random variables, derive high-probability relative-error sample complexity for both methods, and characterize when an exponential gap is possible. We also record extensions beyond iid Bernoulli to multinomials and HMM/state-space models.

1 Bayesian predictive distributions and autoregressive rollouts

1.1 General Bayesian setup

Let $\theta \in \Theta$ be a parameter with prior $\pi_0(d\theta)$. Let $Y_{1:N}$ be observations with likelihood

$$p_\theta(y_{1:N}) = p(y_{1:N} \mid \theta).$$

After observing $D := y_{1:n}$, the posterior is

$$\pi(d\theta \mid D) \propto p_\theta(D)\, \pi_0(d\theta).$$

Fix a future horizon $m := N - n$ and consider the future block $Y_{n+1:n+m}$.

Theorem 1 (Posterior predictive mixture / marginalization identity)**.** *For any measurable set B in the space of length-m sequences,*

$$\mathbb{P}\big(Y_{n+1:n+m} \in B \mid D\big) = \int \mathbb{P}_\theta\big(Y_{n+1:n+m} \in B \mid D\big)\, \pi(d\theta \mid D).$$

Proof. By the law of total probability w.r.t. θ,

$$\mathbb{P}(Y_{n+1:n+m} \in B \mid D) = \int \mathbb{P}(Y_{n+1:n+m} \in B \mid \theta, D)\, \pi(d\theta \mid D),$$

and $\mathbb{P}(Y_{n+1:n+m} \in B \mid \theta, D) = \mathbb{P}_\theta(Y_{n+1:n+m} \in B \mid D)$ by definition of the model under θ. □

Remark 1 (Autoregressive sampling vs. block sampling). If one sequentially samples

$$Y_{n+t} \sim \mathbb{P}(\cdot \mid D, Y_{n+1:n+t-1}), \qquad t = 1, 2, \ldots, m,$$

then by the chain rule the joint distribution of $Y_{n+1:n+m}$ equals the posterior predictive $\mathbb{P}(\cdot \mid D)$. Thus "autoregressive rollout" (sampling from one-step conditionals and feeding back) and "block sampling" from $\mathbb{P}(Y_{n+1:n+m} \mid D)$ are equivalent ways to generate future sequences.

2 Rare events and two estimators

Definition 1 (Rare-event probability)**.** Fix an event A measurable w.r.t. the future block $Y_{n+1:n+m}$. Define

$$q := \mathbb{P}(A \mid D).$$

For each θ, define the conditional event probability

$$f(\theta) := \mathbb{P}_\theta(A \mid D).$$

Corollary 1 (Rare-event probability as a posterior expectation)**.**

$$q = \mathbb{E}_{\theta \sim \pi(\cdot|D)}\big[f(\theta)\big].$$

Proof. Apply Theorem 1 with $B = A$. □

2.1 Estimator 1: rollout hit-rate

Sample R independent rollouts $Y^{(r)}_{n+1:n+m} \sim \mathbb{P}(\cdot \mid D)$ and define

$$I_r := \mathbf{1}\{Y^{(r)}_{n+1:n+m} \in A\}, \qquad \widehat{q}_{\text{roll}} := \frac{1}{R}\sum_{r=1}^{R} I_r.$$

Then $I_r \sim \text{Bernoulli}(q)$ and $\mathbb{E}[\widehat{q}_{\text{roll}}] = q$.

2.2 Estimator 2: posterior sampling + analytic $f(\theta)$ (Rao–Blackwell)

Sample $\theta_1, \ldots, \theta_M \overset{\text{iid}}{\sim} \pi(\cdot \mid D)$ and compute

$$X_i := f(\theta_i), \qquad \widehat{q}_{\text{post}} := \frac{1}{M}\sum_{i=1}^{M} X_i.$$

Then $\mathbb{E}[\widehat{q}_{\text{post}}] = q$. In what follows we assume $f(\theta)$ can be computed exactly (or to negligible error) given θ.

Lemma 1 (Variance decomposition / Rao–Blackwell)**.** *Let $I := \mathbf{1}\{A\}$ be the indicator of A under a single posterior predictive draw. Then*

$$(I) = \mathbb{E}\big[(I \mid \theta)\big] + \big(\mathbb{E}[I \mid \theta]\big),$$

and since $\mathbb{E}[I \mid \theta] = f(\theta)$ we have

$$\big(f(\theta)\big) \le (I) = q(1-q).$$

Proof. This is the law of total variance applied to (I, θ), using $f(\theta) = \mathbb{P}(A \mid \theta, D) = \mathbb{E}[I \mid \theta]$. □

3 A Chernoff bound for bounded random variables (with proof)

We will use a multiplicative Chernoff bound that holds for *any* independent bounded random variables in $[0,1]$, not just Bernoullis.

Lemma 2 (MGF domination by a Bernoulli)**.** *Let $Z \in [0,1]$ be a random variable with $\mathbb{E}[Z] = \mu$. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\lambda Z}\big] \le \exp\big(\mu(e^{\lambda} - 1)\big).$$

Proof. The function $x \mapsto e^{\lambda x}$ is convex, so for $x \in [0,1]$,

$$e^{\lambda x} \le (1-x)e^{0} + xe^{\lambda} = 1 + x(e^{\lambda} - 1).$$

Taking expectations and using $\mathbb{E}[Z] = \mu$ gives

$$\mathbb{E}[e^{\lambda Z}] \le 1 + \mu(e^{\lambda} - 1) \le \exp\big(\mu(e^{\lambda} - 1)\big),$$

where the last inequality uses $1 + u \le e^{u}$. □

Theorem 2 (Multiplicative Chernoff bound for $[0,1)$. *$variables]Let Z_1, \dots, Z_M$ be independent random variables with $Z_i \in [0,1]$ and $\mathbb{E}[Z_i] = \mu$. Let $\bar{Z} := \frac{1}{M}\sum_{i=1}^{M} Z_i$. Then for any $\rho \in (0,1)$,*

$$\mathbb{P}\big(\bar{Z} \ge (1+\rho)\mu\big) \le \exp\Big(-\mu M\big((1+\rho)\ln(1+\rho) - \rho\big)\Big) \le \exp\Big(-\frac{\rho^2 \mu M}{3}\Big),$$

and

$$\mathbb{P}\big(\bar{Z} \le (1-\rho)\mu\big) \le \exp\Big(-\mu M\big(\rho + (1-\rho)\ln(1-\rho)\big)\Big) \le \exp\Big(-\frac{\rho^2 \mu M}{2}\Big).$$

Consequently,

$$\mathbb{P}\big(|\bar{Z} - \mu| \ge \rho\mu\big) \le 2\exp\Big(-\frac{\rho^2 \mu M}{3}\Big).$$

Proof. Let $S := \sum_{i=1}^{M} Z_i$, so $\mathbb{E}[S] = \mu M$.

Upper tail. For $\lambda > 0$, Markov's inequality gives

$$\mathbb{P}(S \ge (1+\rho)\mu M) = \mathbb{P}\big(e^{\lambda S} \ge e^{\lambda(1+\rho)\mu M}\big) \le \frac{\mathbb{E}[e^{\lambda S}]}{e^{\lambda(1+\rho)\mu M}}.$$

By independence and Lemma 2,

$$\mathbb{E}[e^{\lambda S}] = \prod_{i=1}^{M} \mathbb{E}[e^{\lambda Z_i}] \le \prod_{i=1}^{M} \exp\big(\mu(e^{\lambda} - 1)\big) = \exp\big(\mu M(e^{\lambda} - 1)\big).$$

Thus

$$\mathbb{P}(S \ge (1+\rho)\mu M) \le \exp\big(\mu M(e^{\lambda} - 1) - \lambda(1+\rho)\mu M\big).$$

Optimize over $\lambda > 0$; the minimizer is $\lambda^\star = \ln(1+\rho)$, yielding

$$\mathbb{P}(S \ge (1+\rho)\mu M) \le \exp\Big(-\mu M\big((1+\rho)\ln(1+\rho) - \rho\big)\Big).$$

To get the simpler $\rho^2/3$ bound for $\rho \in (0,1)$, note that for $\rho \in [0,1]$,

$$(1+\rho)\ln(1+\rho) - \rho \ge \frac{\rho^2}{3}.$$

A short proof: define $g(\rho) = (1+\rho)\ln(1+\rho) - \rho - \rho^2/3$. Then $g(0) = 0$ and

$$g'(\rho) = \ln(1+\rho) - \frac{2\rho}{3}.$$

Since $\ln(1+\rho)$ is concave, it lies above the chord from $(0,0)$ to $(1, \ln 2)$: $\ln(1+\rho) \geq \rho \ln 2$ for $\rho \in [0,1]$. Because $\ln 2 > 2/3$, we get $g'(\rho) \geq \rho(\ln 2 - 2/3) \geq 0$, hence $g(\rho) \geq 0$.

Lower tail. Similarly, for $\lambda < 0$,

$$\mathbb{P}(S \leq (1-\rho)\mu M) = \mathbb{P}\big(e^{\lambda S} \geq e^{\lambda(1-\rho)\mu M}\big) \leq \frac{\mathbb{E}[e^{\lambda S}]}{e^{\lambda(1-\rho)\mu M}} \leq \exp\big(\mu M(e^{\lambda} - 1) - \lambda(1-\rho)\mu M\big).$$

Optimize over $\lambda < 0$; the minimizer is $\lambda^\star = \ln(1-\rho)$, giving

$$\mathbb{P}(S \leq (1-\rho)\mu M) \leq \exp\Big(-\mu M\big(\rho + (1-\rho)\ln(1-\rho)\big)\Big).$$

To get the simpler $\rho^2/2$ bound, define $h(\rho) = \rho + (1-\rho)\ln(1-\rho) - \rho^2/2$. Then $h(0) = 0$ and

$$h'(\rho) = -\ln(1-\rho) - \rho \geq 0$$

because $-\ln(1-\rho) \geq \rho$ for $\rho \in [0,1)$ (equivalently $\ln(1-\rho) \leq -\rho$). Hence $h(\rho) \geq 0$.

Finally, combine the two tails and use the weaker constant 3 to obtain the stated two-sided bound. □

Corollary 2 (Relative-error sample size for bounded variables)**.** *Under the conditions of Theorem 2, for $\rho \in (0,1)$ and $\delta \in (0,1)$,*

$$M \;\geq\; \frac{3}{\rho^2 \mu} \ln \frac{2}{\delta} \quad \Longrightarrow \quad \mathbb{P}\big(|\bar{Z} - \mu| \leq \rho\mu\big) \geq 1 - \delta.$$

4 Sample complexity: rollouts vs posterior sampling

Throughout this section, fix $\rho \in (0,1)$ and $\delta \in (0,1)$.

4.1 Estimating q via rollouts

Theorem 3 (Rollout estimator sample complexity)**.** *Let $I_r = \mathbf{1}\{A\}$ from an iid posterior-predictive rollout. Then for*

$$R \;\geq\; \frac{3}{\rho^2 q} \ln \frac{2}{\delta},$$

we have $\mathbb{P}(|\widehat{q}_{\mathrm{roll}} - q| \leq \rho q) \geq 1 - \delta$.

Proof. Apply Corollary 2 to $Z_r = I_r \in [0,1]$ with mean $\mu = q$. □

4.2 Estimating q via posterior sampling and analytic $f(\theta)$

Define

$$b := \sup_{\theta \in \Theta} f(\theta) \in (0,1].$$

Since $f(\theta)$ is a probability, $b \leq 1$, but for many "thin" events $b \ll 1$.

Theorem 4 (Posterior-sampling estimator sample complexity)**.** *Assume* $0 \le f(\theta) \le b$ *for all* θ*, and define* $X_i = f(\theta_i)$ *with* $\theta_i \overset{iid}{\sim} \pi(\cdot \mid D)$*. Then for*

$$M \;\ge\; \frac{3b}{\rho^2 q} \ln\frac{2}{\delta},$$

we have $\mathbb{P}(|\widehat{q}_{\text{post}} - q| \le \rho q) \ge 1 - \delta$*.*

Proof. Let $Z_i := X_i/b \in [0,1]$. Then $\mathbb{E}[Z_i] = \mathbb{E}[X_i]/b = q/b$ and $\widehat{q}_{\text{post}} = b\,\bar{Z}$. Moreover,

$$\frac{|\widehat{q}_{\text{post}} - q|}{q} = \frac{|b\bar{Z} - b\mathbb{E}[Z]|}{b\mathbb{E}[Z]} = \frac{|\bar{Z} - \mathbb{E}[Z]|}{\mathbb{E}[Z]}.$$

Apply Corollary 2 with $\mu = q/b$. □

Corollary 3 (Improvement factor in sample count)**.** *Comparing Theorems 3 and 4, the posterior method reduces the required number of Monte Carlo samples by a factor*

$$\frac{R}{M} \approx \frac{1}{b}.$$

Thus a large gap is possible exactly when $b = \sup_\theta f(\theta)$ *is very small.*

4.3 Tightness: a worst-case posterior can saturate the bound

Proposition 1 (Worst-case tightness via a two-point posterior)**.** *Suppose there exist* θ_0, θ_1 *such that* $f(\theta_0) = 0$ *and* $f(\theta_1) = b$*. For any* $q \in (0, b)$*, define a posterior supported on* $\{\theta_0, \theta_1\}$ *by*

$$\pi(\theta = \theta_1 \mid D) = \frac{q}{b}, \qquad \pi(\theta = \theta_0 \mid D) = 1 - \frac{q}{b}.$$

Then $Z = f(\theta)/b$ *is exactly* Bernoulli(q/b)*. In particular, estimating* q *by posterior sampling is information-theoretically as hard as estimating the mean of a Bernoulli*(q/b)*, so* $M = \Omega\left(\frac{b}{\rho^2 q} \log\frac{1}{\delta}\right)$ *samples are necessary (up to constants).*

Proof. Under this posterior, $f(\theta)$ equals b with probability q/b and 0 otherwise, hence $Z = f(\theta)/b$ is Bernoulli(q/b). Standard lower bounds for Bernoulli mean estimation (e.g. Le Cam's method on μ vs. $(1+\rho)\mu$) imply the stated necessity; the upper bound in Theorem 4 matches this scaling up to constants. □

5 "Seeing one hit" vs. "estimating q"

If the goal is to *observe* at least one realization in A (rather than estimate q), then nothing beats the $1/q$ barrier.

Proposition 2 (Samples needed to see at least one event)**.** *If* $Y^{(1)}, \ldots, Y^{(R)}$ *are iid draws from the posterior predictive and* $\mathbb{P}(A \mid D) = q$*, then*

$$\mathbb{P}(\exists r : Y^{(r)} \in A) = 1 - (1-q)^R.$$

Thus to see at least one hit with probability $\ge 1 - \delta$*, it suffices (and is necessary up to constants for small* q*) that*

$$R \gtrsim \frac{1}{q} \ln\frac{1}{\delta}.$$

Moreover, sampling $\theta \sim \pi(\cdot \mid D)$ *and then sampling* $Y \sim p_\theta(\cdot \mid D)$ *produces exactly one posterior-predictive draw, so posterior sampling does not change this hit complexity unless one uses analytic* $f(\theta)$ *(i.e. one is no longer waiting for a literal hit).*

6 Bernoulli strings: two instructive extremes

Let $Y_t \in \{0,1\}$ and parameter $\theta = p \in [0,1]$. Let $m = N - n$ be the horizon.

6.1 Event $A =$ all ones (no sample-count gain)

If $A = \{Y_{n+1} = \cdots = Y_{n+m} = 1\}$, then $f(p) = p^m$ and

$$b = \sup_{p \in [0,1]} p^m = 1,$$

so Corollary 3 gives no sample-count improvement in the worst case: M and R both scale like $\Theta\big(\frac{1}{q}\big)$ for fixed (ρ, δ).

6.2 Event $A =$ one fixed mixed string (exponential gain possible)

Let $A = \{Y_{n+1:n+m} = s\}$ for a prespecified string $s \in \{0,1\}^m$ with k ones and $m-k$ zeros $(1 \le k \le m-1)$. Then

$$f(p) = p^k (1-p)^{m-k}, \qquad b = \max_{p \in [0,1]} f(p) = \left(\frac{k}{m}\right)^k \left(\frac{m-k}{m}\right)^{m-k}.$$

Hence the improvement factor is

$$\frac{1}{b} = \left(\frac{m}{k}\right)^k \left(\frac{m}{m-k}\right)^{m-k} = \exp\Big(m\, H\big(k/m\big)\Big),$$

where $H(x) = -x \ln x - (1-x)\ln(1-x)$ is the binary entropy (natural logs). In particular, for $k = m/2$ one gets $b = 2^{-m}$ and the improvement factor 2^m.

7 Beyond Bernoulli: multinomials, HMMs, and state-space models

7.1 Multinomial (Dirichlet–Categorical) generalization

Let $Y_t \in \{1, \dots, K\}$, $\theta = \pi \in \Delta^{K-1}$, and $p_\pi(y_{1:N}) = \prod_{t=1}^N \pi_{y_t}$ (iid categorical). For a prespecified length-m string s with counts $c_1, \dots, c_K$ (so $\sum_j c_j = m$),

$$f(\pi) = \mathbb{P}_\pi(Y_{n+1:n+m} = s \mid D) = \prod_{j=1}^K \pi_j^{c_j},$$

and

$$b = \sup_{\pi \in \Delta^{K-1}} \prod_{j=1}^K \pi_j^{c_j} = \prod_{j=1}^K \left(\frac{c_j}{m}\right)^{c_j}.$$

If $c_j = m/K$ (balanced), then $b = K^{-m}$ and the improvement factor is K^m.

7.2 Hidden Markov models: analytic $f(\theta)$ via forward DP and exponential b

Consider an HMM with finite latent states $X_t \in \{1, \dots, S\}$ and observations $Y_t \in \mathcal{Y}$. A parameter θ specifies an initial distribution $\pi_\theta(x_1)$, transition matrix $T_\theta(x' \mid x)$, and emission probabilities $E_\theta(y \mid x)$. Given θ,

$$\mathbb{P}_\theta(y_{1:m}) = \sum_{x_{1:m}} \pi_\theta(x_1) \prod_{t=2}^{m} T_\theta(x_t \mid x_{t-1}) \prod_{t=1}^{m} E_\theta(y_t \mid x_t).$$

For a *fixed future string* $s \in \mathcal{Y}^m$ conditional on observed prefix $D = y_{1:n}$,

$$f(\theta) = \mathbb{P}_\theta\big(Y_{n+1:n+m} = s \mid D\big)$$

is computable by the standard forward algorithm: compute the filtered distribution over X_n from D, then propagate m steps multiplying by emissions for the fixed s. This costs $O(mS^2)$ time for discrete HMMs.

The key quantity for sample complexity is $b = \sup_\theta f(\theta)$. A simple sufficient condition for exponential smallness of b is a uniform emission peak bound.

Lemma 3 (Uniform per-step peak bound implies $b \le \eta^m$)**.** *Assume there exists $\eta \in (0, 1)$ such that for all θ, all states x, and all symbols $y \in \mathcal{Y}$,*

$$E_\theta(y \mid x) \le \eta.$$

Then for any fixed length-m observation string $s \in \mathcal{Y}^m$ and any prefix D,

$$\sup_\theta \mathbb{P}_\theta(Y_{n+1:n+m} = s \mid D) \le \eta^m.$$

Hence $b \le \eta^m$ and the rollout-vs-posterior improvement factor is at least η^{-m}.

Proof. Fix θ and condition on the prefix D. For each $t = 1, \dots, m$,

$$\mathbb{P}_\theta(Y_{n+t} = s_t \mid D, Y_{n+1:n+t-1} = s_{1:t-1}) = \sum_x \mathbb{P}_\theta(X_{n+t} = x \mid D, s_{1:t-1})\, E_\theta(s_t \mid x) \le \max_x E_\theta(s_t \mid x) \le \eta.$$

Multiplying the conditional probabilities via the chain rule gives

$$\mathbb{P}_\theta(Y_{n+1:n+m} = s \mid D) \le \eta^m.$$

Taking $\sup_\theta$ yields the claim. □

Remark 2 (State-space models)*.* For continuous-observation state-space models, the event "$Y_{n+1:n+m}$ equals an exact real-valued trajectory" typically has probability 0. A direct analogue is to take A to be a small neighborhood (e.g. an ε-tube) around a target trajectory, or to discretize/quantize observations. When the one-step observation likelihood/density is uniformly bounded above, an analogue of Lemma 3 typically yields $b \le (\text{const} \cdot \varepsilon)^m$ and therefore an exponential separation in m.

8 Executive summary of the main message

Let $q = \mathbb{P}(A \mid D) = \mathbb{E}_{\theta\sim\pi(\cdot|D)}[f(\theta)]$ with $f(\theta) = \mathbb{P}_\theta(A \mid D)$.

- **If you must literally wait for a hit of A in simulation:** you need $\Theta\big(\frac{1}{q}\log\frac{1}{\delta}\big)$ posterior-predictive draws to see one hit with high probability.
- **If you only need to estimate q and can compute $f(\theta)$:** rollout MC needs $R = \Theta\big(\frac{1}{\rho^2 q}\log\frac{1}{\delta}\big)$ samples, while posterior sampling needs

$$M = \Theta\left(\frac{b}{\rho^2 q}\log\frac{1}{\delta}\right), \qquad b = \sup_\theta f(\theta).$$

  Thus the **sample-count improvement factor** is $\Theta(1/b)$, which can be exponential in m for "thin" events where b decays exponentially (e.g. fixed strings in multinomials, noisy HMMs).