Exact Posterior Probability of Optimality for Factor-Gaussian Beliefs

the integral that Thompson-style methods approximate, computed in closed form at scale

Peter Cotton

September 1, 2026

Abstract

The posterior probability that alternative i is best, $p_i = \Pr(\theta_i = \max_j \theta_j \mid \mathcal{D})$, drives stopping rules, budget allocation, derandomized Thompson sampling, and safety constraints, and the reinforcement-learning and Bayesian optimization literatures describe it as intractable: recent work variationally approximates it from marginal means and variances alone. We observe that in two standard settings the posterior covariance is exactly of factor form $\Sigma = WW' + \mathrm{diag}(d)$ at every step: simulation replicates under common random numbers, and per-arm Gaussian observations against a factor prior. In that case the whole vector p is a one-dimensional integral per factor node, computable for all n alternatives at once in time linear in n by machinery previously developed for pricing correlated contests, and certified here against Monte Carlo. In a stopping problem under common random numbers, rules that stop on marginal-only estimates of p hold their nominal error by overspending, between 1.5 and 3.4 times the replicates of the exact rule, which stops earliest in every correlation regime at the nominal error level. In a correlated Gaussian bandit, sampling from the exact vector reproduces Thompson sampling's regret to within replication noise, as the occupancy identity requires, while the variational policy pays ten percent more regret and sits at measured total-variation distance 0.13 to 0.16 from the object it stands in for.

1 Introduction

Let $\theta \in R^n$ be unknown values of n alternatives (simulated systems, bandit arms, candidate molecules) and let $\mathcal{D}$ be data. The posterior probability of optimality is the vector

$$p_i \;=\; \Pr\left(\theta_i = \max_j \theta_j \,\middle|\, \mathcal{D}\right), \qquad i = 1, \ldots, n. \tag{1}$$

This vector is the natural currency of sequential decision problems. A stopping rule halts when $\max_i p_i$ clears a confidence target; a budget allocator spends where p is contested; Thompson sampling draws its action from p implicitly; a constrained or multi-agent policy needs p explicitly, because a single posterior sample no longer suffices.

The literature treats Equation (1) as out of reach. Tarbouriech et al. (2023), introducing the quantity as the key object of Bayesian reinforcement learning, write that computing it "involves computing several complicated integrals with respect to the posterior and is intractable in most cases," and open their solution section with "computing this probability is intractable in general." Their remedy, VAPOR, is a variational program over occupancy measures. Menet et al. (2026)

fine-tune a large language model toward the same object under a Gaussian-process posterior; the tractable policy they target consumes only the marginal means and marginal standard deviations of that posterior, and they note of the prior guarantee for this family that "the structure of the kernel is not taken into account, i.e., the worst-case bandit with independent arms is assumed." In ranking and selection, procedures bound the probability of correct selection with Bonferroni-type inequalities or estimate it by posterior sampling; the joint vector is not evaluated.

This paper makes a single observation and develops its consequence. In the settings these literatures study, the posterior covariance frequently has exact factor form

$$\Sigma \;=\; WW' + \mathrm{diag}(d), \qquad W \in R^{n\times\rho}, \quad \rho \ll n, \tag{2}$$

not as a modeling approximation but as a consequence of the observation process, and for factor-form covariance Equation (1) is computable exactly, for all i at once, in time linear in n. The computation is not new to this paper: it is the correlated-contest pricing pass of Cotton (2026), where it is developed together with its derivatives and its inverse and benchmarked to $n = 10^6$. This paper adds the derivation that two canonical posteriors stay inside the factor grammar at every update, the identification of the probability-of-optimality literature's central object with that pass, and measurements of what the exact vector buys where surrogates are currently used.

2 The computation

Condition on the factor variables. Write $\theta = m + Wz + \varepsilon$ with $z \sim N(0, I_\rho)$ and $\varepsilon_i \sim N(0, d_i)$ independent. Given z, the alternatives are independent with means $m_i + w_i'z$ and variances d_i, so with $F_{i|z}$ and $f_{i|z}$ the conditional distribution and density functions,

$$p_i \;=\; \int_{R^\rho} \phi_\rho(z) \int_R f_{i|z}(x) \prod_{j\neq i} \big(1 - F_{j|z}(x)\big)\, dx\, dz. \tag{3}$$

The inner integrand for every i shares the survival field $G_z(x) = \prod_j (1 - F_{j|z}(x))$: dividing out the ith factor prices all n alternatives from one field,

$$p_i \;=\; E_z \int f_{i|z}(x)\, \frac{G_z(x)}{1 - F_{i|z}(x)}\, dx. \tag{4}$$

On a lattice of L points with Q quadrature nodes over z, the cost is $O(nLQ)$ for the whole vector, with the numerical care (window placement, log-domain products, tail handling) described in Cotton (2026), whose implementation is used unchanged here. Every exact vector reported below is certified against Monte Carlo argmax frequencies at 2×10^5 draws; the worst certificate total-variation distance in this paper is 0.0042, at the Monte Carlo noise floor for that sample size.

3 Two posteriors with exact factor form

3.1 Common random numbers

Ranking and selection compares n simulated systems. Under common random numbers, replicate r scores every system on one shared scenario:

$$Y_{ir} \;=\; \theta_i + v_i'F_r + \varepsilon_{ir}, \qquad F_r \sim N(0, I_\rho), \quad \varepsilon_{ir} \sim N(0, \sigma_i^2), \tag{5}$$

with F_r shared across i. The scenario effects v_i and noise scales σ_i are the declared simulation structure; the unknown is θ, with prior $\theta \sim N(0, s_0^2 I)$. The per-replicate covariance is

$$\Lambda \;=\; VV' + D, \qquad D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2). \tag{6}$$

After n_r replicates with mean vector $\bar{Y}$, the posterior of θ is Gaussian with precision

$$A \;=\; s_0^{-2} I + n_r \Lambda^{-1}. \tag{7}$$

The Woodbury identity applied to Λ gives

$$\Lambda^{-1} = D^{-1} - D^{-1} V M^{-1} V' D^{-1}, \qquad M = I_\rho + V' D^{-1} V, \tag{8}$$

so that, writing $U = D^{-1}V$ and $S = n_r M^{-1}$, Equation (7) becomes

$$A = \mathrm{diag}(a) - USU', \qquad a_i = s_0^{-2} + n_r \sigma_i^{-2}. \tag{9}$$

The precision is a positive diagonal minus a rank-ρ positive semidefinite term. Applying Woodbury a second time, now to Equation (9),

$$A^{-1} = \mathrm{diag}(a)^{-1} + \mathrm{diag}(a)^{-1} U\, C\, U' \mathrm{diag}(a)^{-1}, \qquad C = \left(S^{-1} - U' \mathrm{diag}(a)^{-1} U\right)^{-1}. \tag{10}$$

The middle matrix C is positive definite: consider the block matrix $K = \left(\begin{smallmatrix} \mathrm{diag}(a) & U \\ U' & S^{-1} \end{smallmatrix}\right)$. The Schur complement of S^{-1} in K is $\mathrm{diag}(a) - USU' = A \succ 0$, and $S^{-1} \succ 0$, so $K \succ 0$; but then the Schur complement of $\mathrm{diag}(a)$ in K, which is C^{-1}, is also positive definite. Therefore

$$\Sigma_{\text{post}} \;=\; \mathrm{diag}(d) + WW', \qquad d_i = a_i^{-1}, \quad W = \mathrm{diag}(a)^{-1} U\, C^{1/2}, \tag{11}$$

exactly, at every n_r, with W of rank ρ: the number of shared scenario streams, not the number of systems. The posterior mean is $m = A^{-1}h$ with $h = n_r \Lambda^{-1} \bar{Y}$, both factors already expanded above. Nothing here is approximated; the factor form of the posterior is inherited from the factor form of the scenario noise.

3.2 Per-arm observations against a factor prior

In a finite Gaussian bandit, the prior on arm means carries the correlation: $\theta \sim N(0, \Sigma_0)$ with $\Sigma_0 = VV' + \mathrm{diag}(d_0)$ — the form produced by linear models $\theta_i = v_i'\beta + e_i$, by inducing-point Gaussian-process posteriors, and by the heteroscedastic output layers of large classifiers. Pulling arm a returns $\theta_a + N(0, \sigma_n^2)$. With per-arm counts c the posterior precision is

$$A \;=\; \Sigma_0^{-1} + \sigma_n^{-2}\, \mathrm{diag}(c), \tag{12}$$

and Equation (8) applied to Σ_0 puts A in the form of Equation (9) with $a_i = d_{0i}^{-1} + c_i \sigma_n^{-2}$ and $S = M_0^{-1}$, $M_0 = I_\rho + V' \mathrm{diag}(d_0)^{-1} V$. The argument from Equation (10) onward is unchanged: the posterior is exactly factor-plus-diagonal after every pull, with the same rank ρ as the prior.

4 The marginal-only surrogates and the Thompson identity

Two computable policies stand in for Equation (1) in current practice, and both consume only the marginal means m_i and marginal standard deviations $s_i = \sqrt{d_i + \|w_i\|^2}$.

The first is the independence approximation: replace the posterior by $\prod_i N(m_i, s_i^2)$ and compute the resulting maximality probabilities, which Menet et al. (2025) do with a shared threshold κ^* solving $\sum_i \Phi((m_i - \kappa^*)/s_i) = 1$. The second is the variational optimistic policy of O'Donoghue and Lattimore (2021), extended by Tarbouriech et al. (2023) and given a near-closed form by Menet et al. (2026):

$$\tilde{\pi}_i \;=\; v\left(\frac{m_i - \kappa^*}{s_i}\right), \qquad v(c) = \exp\!\Big(-\tfrac{1}{8}\big(\sqrt{c^2+4} - c\big)^2\Big), \tag{13}$$

with κ^* again normalizing the sum to one. The tilt away from the true maximality probabilities is deliberate — the regret analysis uses it — so Equation (13) should be read as a policy, not as an estimate of p; the measurements below report its distance from p because the fine-tuning literature uses it as the trainable stand-in for p.

The identity that connects Equation (1) to Thompson sampling is Lemma 8 of Tarbouriech et al. (2023): the expected occupancy of the Thompson-sampling policy equals the posterior probability of optimality. In the bandit case this is immediate — Thompson sampling draws $\tilde{\theta}$ from the posterior and plays its argmax, so its action distribution at any history is exactly p at that history. Sampling from the exact vector is therefore not a new policy but a derandomization-ready form of the old one, and the regret curves of the two must coincide in expectation. That coincidence is measured, not assumed, in Section 6. Under simple-regret loss the one-step value of a measurement is the knowledge-gradient quantity (Frazier et al., 2008). Tuisov (2026) shows this is the normative myopic value of computation, noting that the exact dynamic version "is itself generally intractable"; for factor-Gaussian beliefs the myopic quantities are exact, at scale.

5 Stopping under common random numbers

The experiment races three stopping rules on the posterior of Section 3.1, with $n = 100$ systems, $\rho = 2$ scenario streams, prior scale $s_0 = 1$, unit total replicate variance for every system, and true means drawn from the prior. Each rule computes its own estimate of p after every replicate and stops the first time its own maximum clears $1 - \delta$ with $\delta = 0.05$, selecting its argmax: the exact vector by the pass of Section 2; the independence construction; and Equation (13). Correlation geometry is the treatment. In the *aligned* regime all systems load on one stream with factor share 0.6, so the leaders' differences are less uncertain than their marginals suggest; in the *opposed* regime two blocks load oppositely, so the leaders' differences are more uncertain; the *independent* regime is the control with factor share zero.

The mechanism is the size of the substitution error: along its own pre-stopping trajectory in the aligned regime, the independence construction sits at median total-variation distance 0.123 from the exact vector it replaces (ninetieth percentile 0.174), and the variational policy at 0.141 (0.206). In the control regime the gap does not vanish — median 0.063 and 0.096 respectively — because the independence construction is itself a threshold approximation with its own bias even when the posterior really is independent, and the variational tilt is away from p by design. The exact rule's distance is zero by construction, and its certificate against Monte Carlo is total-variation 0.0015 at worst.

regime	rule	mean replicates	realized PCS	censored of 200
independent	exact	141.2	0.970	35
	independence	207.0	0.985	65
	variational	271.5	1.000	102
aligned	exact	69.7	0.958	9
	independence	175.6	1.000	60
	variational	240.1	1.000	91
opposed	exact	106.3	0.944	21
	independence	171.7	0.979	57
	variational	239.0	0.991	88

Table 1: Stopping at nominal 95% confidence, 200 replications per regime, budget 400 replicates (censored runs count 400 toward the mean). Realized PCS is the fraction of stopped runs selecting the true best; the exact rule's 0.944 in the opposed regime is within binomial noise of nominal (179 stopped runs, standard error 0.016). Every rule held its nominal error in these geometries; the marginal-only rules did so by conservatism, spending 1.5 to 3.4 times the replicates of the exact rule and censoring at the budget far more often. The largest gap is where correlation helps most: with aligned loadings the exact rule exploits the leaders' tightly coupled differences and stops at 69.7 replicates on average.

6 A correlated Gaussian bandit

The bandit uses the posterior of Section 3.2: $n = 50$ arms, rank-2 prior, two regimes (independent, and two clusters with factor share 0.75), horizon 200, and 100 replications. Four policies act on identical posterior laws: Thompson sampling; probability matching on the exact vector; and sampling from Equation (13) and from the independence construction.

7 Scope

The derivations of Section 3 are exact because the observation processes there are themselves factor-structured. When scenario effects enter a simulation nonlinearly, the low-rank model of Λ is an approximation and the posterior inherits its error; when a Gaussian-process kernel is dense and short-scaled, a factor fit must precede the pass and its residual is the price. Nothing here computes the dynamic value of computation, which Tuisov (2026) is right to call intractable in general: this paper supplies the exact myopic quantities his Bellman characterization prices, and the stopping certificates. The optimistic tilt of Equation (13) carries the regret analysis of O'Donoghue and Lattimore (2021); measuring its distance from p quantifies the approximation, and does not by itself argue the tilt away. Finally, Thompson sampling needs one posterior sample and no vector; the vector pays where sampling does not suffice (stopping, allocation, constraints, derandomization), which is where the measurements above are taken.

regime	policy	regret	identification	shortfall	TV to exact p
independent	Thompson	172.1	0.72	0.073	0
	exact matching	173.9	0.75	0.055	0
	variational	190.3	0.76	0.061	0.159
	independence	179.2	0.79	0.052	0.057
clusters	Thompson	122.7	0.61	0.097	0
	exact matching	124.8	0.51	0.113	0
	variational	135.9	0.59	0.105	0.129
	independence	128.7	0.51	0.121	0.061

Table 2: Bayes regret over the horizon (mean of 100 replications), terminal identification rate, expected shortfall of the recommended arm, and the mid-horizon total-variation distance between each policy's action distribution and the exact vector along its own trajectory. Thompson sampling's distance is zero by the occupancy identity and exact matching's by construction. The Lemma-8 check holds: Thompson and exact matching differ by about one percent of cumulative regret in both regimes, within replication noise. The variational tilt costs a measured ten to eleven percent of cumulative regret relative to Thompson sampling while sitting at total-variation distance 0.13 to 0.16 from the vector it stands in for. Identification rates at this horizon are statistically flat across policies: differences are within two binomial standard errors at 100 replications, so the exhibit is the regret column and the gap column, not identification.

References

P. Cotton. Scalable inversion of contests with correlated performances, including softmax and multinomial probit. arXiv:2609.01133, 2026. Also SSRN 7307363.

P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

N. Menet, J. Huebotter, P. Kassraie, and A. Krause. Efficiently estimating Gaussian probability of maximality. In *AISTATS*, 2025. arXiv:2501.13535.

N. Menet, A. Terzić, M. Hersche, A. Krause, and A. Rahimi. Thompson sampling via fine-tuning of LLMs. In *ICLR*, 2026. arXiv:2510.13328.

B. O'Donoghue and T. Lattimore. Variational Bayesian optimistic sampling. In *Advances in Neural Information Processing Systems 34*, 2021. arXiv:2110.15688.

J. Tarbouriech, T. Lattimore, and B. O'Donoghue. Probabilistic inference in reinforcement learning done right. In *Advances in Neural Information Processing Systems 36*, 2023. arXiv:2311.13294.

A. Tuisov. Search as computation allocation. arXiv:2607.27871, 2026.