SHARP CONCENTRATION FOR GIBBS DISTRIBUTIONS UNDER GAUSSIAN SCORE NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gibbs distributions, or softmax distributions under a uniform reference measure, convert scores into probabilistic decisions. We study their stability when the scores are perturbed by correlated Gaussian estimation errors. The forward Kullback–Leibler (KL) divergence between the population and noisy Gibbs distributions is exactly a nonlinear log-partition remainder after its first-order response to the score error is removed. Our main result gives a sharp, dimension-independent concentration bound for this remainder: after normalization by the largest variance of a pairwise score difference, its centered positive moment-generating function is bounded by that of a Gamma random variable with shape $1/2$. The result holds for arbitrary finite action sets, reference probabilities, temperatures, Gaussian correlations, and noise magnitudes, and a balanced two-action problem approaches equality in the small-noise limit. The proof combines decreasing covariance envelopes under Gaussian smoothing with an explicit bound on the nonzero average gradient caused by asymmetric priors. We derive high-probability KL certificates, $O(1/\epsilon)$ sample-complexity bounds, a data-dependent guarantee for unknown covariance, stronger results for structured action spaces, and trajectory-level guarantees for deterministic entropy-regularized planning. Numerical experiments examine tightness and the practical size of these certificates.

1 INTRODUCTION

Gibbs and softmax distributions turn scores into probabilistic decisions, including in entropy-regularized reinforcement learning (Geist et al., 2019) and soft Bellman updates (Fox et al., 2016). In practice, these scores are estimated from data and therefore contain uncertainty. We ask a basic stability question: *How much can Gaussian score uncertainty change the resulting Gibbs distribution?*

Let p denote the Gibbs distribution formed from population scores and let $\widehat{p}$ denote the distribution formed after adding jointly Gaussian score errors, which may be correlated across alternatives. A central identity in our analysis is $B = \tau D(p\|\widehat{p})$, where B is the nonlinear part of the log-partition change left after subtracting its first-order response to the score error and τ is the temperature. Thus concentration of B directly controls the forward-KL change in the decision rule.

Our main result shows that the centered fluctuations of B obey a sharp Gamma-type moment-generating-function (MGF) bound at every noise magnitude. The natural scale is the largest variance of a pairwise score difference, so a random shift shared equally by all scores cancels automatically. After this problem-dependent scale is removed, the numerical constant does not grow with the number of alternatives or the dimension of the underlying Gaussian vector. A balanced two-action problem approaches equality in the small-noise limit, showing that the constant cannot be uniformly improved. Sections 3–4 explain why ordinary Gaussian concentration misses this sharper second-order behavior and how the finite-noise bound is proved.

The concentration result yields computable high-probability KL certificates and an $O(1/\epsilon)$ sample requirement when score variances scale as $1/n$. We also give a data-dependent certificate when the reward covariance is unknown, stronger guarantees for collinear and product-structured action spaces, and a trajectory-level extension to deterministic entropy-regularized planning that preserves correlations created by shared reward estimates. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We prove a sharp finite-noise Gamma-$1/2$ MGF bound for the Gibbs adaptation remainder under arbitrary finite jointly Gaussian score noise, with scale determined by the largest variance of a pairwise score difference.
- We characterize the small-noise quadratic limit, show why it determines the best uniform MGF constant even though it need not upper-bound a fixed model at nonzero noise, and prove the finite-noise theorem using contracting covariance bounds, an explicit average-gradient bound, and a martingale comparison.
- We derive forward-KL stability certificates, $O(1/\epsilon)$ sample-complexity guarantees, a same-data covariance-estimation correction, and sharper results under one-dimensional or product structure.
- We extend the analysis to deterministic entropy-regularized planning by treating complete trajectories as Gibbs alternatives, and we numerically examine tightness, certificate size, covariance estimation, and trajectory-level scaling.

2 Problem setup: Gibbs decisions under Gaussian score noise

We first define the noisy Gibbs decision problem and the scalar quantity whose concentration controls policy stability. The goal is to establish the object of study and its natural uncertainty scale. Let $r_i \in \mathbb{R}$ be the population score of alternative i, let $q_i > 0$ be a reference probability, and let $\tau > 0$ be the *temperature*. The associated Gibbs distribution is

$$p_i = \frac{q_i e^{r_i/\tau}}{\sum_j q_j e^{r_j/\tau}}. \tag{1}$$

When q is uniform, this is the usual softmax distribution. If the observed score is $r_i + \xi_i$, where $\xi = (\xi_1, \dots, \xi_k)$ is centered jointly Gaussian, the perturbed distribution is

$$\widehat{p}_i = \frac{q_i e^{(r_i+\xi_i)/\tau}}{\sum_j q_j e^{(r_j+\xi_j)/\tau}} = \frac{p_i e^{\xi_i/\tau}}{\sum_j p_j e^{\xi_j/\tau}}.$$

Define the scaled log-partition function $V_\tau(r; q) = \tau \log \sum_i q_i e^{r_i/\tau}$. A direct calculation gives

$$B := V_\tau(r+\xi; q) - V_\tau(r; q) - \sum\nolimits_i p_i \xi_i = \tau \log \sum\nolimits_i p_i e^{\xi_i/\tau} - \sum\nolimits_i p_i \xi_i = \tau D(p\|\widehat{p}). \tag{2}$$

We call B the *entropic adaptation remainder*. It is the part of the noisy log-partition value that remains after subtracting the direct linear score error $p^\top \xi$. Equivalently, it is exactly the forward KL change of the Gibbs distribution, multiplied by τ. Thus controlling B is the same as controlling the stability of the noisy Gibbs decision.

Any centered jointly Gaussian vector can be written as $\xi_i = \langle a_i, G \rangle,\ G \sim N(0, I_d)$, for vectors $a_i \in \mathbb{R}^d$. This representation is simply a factorization of the covariance matrix, since $\mathrm{Cov}(\xi_i, \xi_j) = a_i^\top a_j$. With $\mathrm{supp}(p) = \{i : p_i > 0\}$, we define

$$\bar{a} = \sum\nolimits_i p_i a_i, \qquad \Delta = \max\nolimits_{i,j \in \mathrm{supp}(p)} \|a_i - a_j\|, \qquad c = \frac{\Delta^2}{4\tau}. \tag{3}$$

Since $\|a_i - a_j\|^2 = \mathrm{Var}(\xi_i - \xi_j)$, Δ^2 is the largest variance of a pairwise score difference. Any Gaussian component shared equally by all scores thus cancels from both Δ and the Gibbs distribution. This is important because softmax decisions depend only on relative scores: uncertainty that shifts every score by the same amount should not affect the stability measure. We have therefore reduced policy stability to concentration of the deterministic function $G \mapsto B(G)$ at the natural scale Δ^2.

3 From Gaussian concentration to the sharp Gamma bound

We first present the natural Gaussian-concentration baseline, then explain the second-order structure that it misses, and state the Gamma MGF bound that remains valid at arbitrary noise magnitudes.

3.1 Standard Gaussian concentration as a baseline

Substituting $\xi_i = \langle a_i, G\rangle$ into (2) gives $B(G) = \tau \log \sum_i p_i e^{\langle a_i, G\rangle/\tau} - \bar{a}^\top G$. Differentiating yields

$$\nabla B(G) = \sum\nolimits_i \widehat{p}_i(G) a_i - \bar{a}, \qquad \nabla^2 B(G) = \tau^{-1} \operatorname{Cov}_{\widehat{p}(G)}(a). \tag{4}$$

Both feature means in the gradient lie in the convex hull of the a_i's, so $\|\nabla B(G)\| \le \Delta$. Hence Gaussian Lipschitz concentration gives

$$K_B(t) := \log \mathbb{E} e^{t(B - \mathbb{E}B)} \le \frac{\Delta^2 t^2}{2}, \qquad t \ge 0.$$

The Lipschitz bound is the most direct reference point: it is valid without additional structure, but it uses only the largest possible gradient. The Hessian in (4) gives more information. Standard covariance-diameter inequalities imply

$$\|\nabla^2 B\|_{\mathrm{op}} \le c, \quad \|\nabla^2 B\|_{\mathrm{HS}}^2 \le c^2, \quad \|\nabla B\|^2 \le 2cB, \quad \mathbb{E}B \le c. \tag{5}$$

Combining these inequalities with the Gaussian log-Sobolev inequality gives the stronger baseline

$$K_B(t) \le \frac{c\,\mathbb{E}B\, t^2}{1 - ct} \le \frac{c^2 t^2}{1 - ct}, \qquad 0 \le t < 1/c. \tag{6}$$

The derivation is given in Section B. Both baselines are dimension independent. The self-bounding log-Sobolev argument already captures the second-order scale; the sharp constant requires finer Gibbs covariance bounds and control of the average gradient.

3.2 Second-order behavior and the finite-noise challenge

At zero perturbation, $\widehat{p}(0) = p$, so (4) gives $\nabla B(0) = 0$. A Taylor expansion therefore begins with

$$B(G) \approx \frac{1}{2} G^\top \nabla^2 B(0) G = \frac{1}{2\tau} G^\top \operatorname{Cov}_p(a) G. \tag{7}$$

A Gaussian quadratic form is a weighted sum of squared Gaussian variables, suggesting chi-square/Gamma-type rather than ordinary Gaussian fluctuations. In the balanced two-alternative case, the normalized quadratic limit is exactly $Z^2/2$ for $Z \sim N(0,1)$, which already has the Gamma shape that appears in the main theorem. This approximation is only local, however. At nonzero noise, $\widehat{p}(G)$ changes with G, so the Hessian in (4) is not fixed at its value at the origin. Thus the small-noise quadratic form is not a pointwise upper bound on $B(G)$, and a fixed model can even become more variable as the noise first moves away from zero. The central question is therefore whether the Gamma-type behavior suggested by (7) can nevertheless yield one centered-MGF upper bound for all admissible Gibbs models and all noise magnitudes.

3.3 Main result: a sharp Gamma MGF bound

With $\Sigma = \operatorname{Cov}(\xi)$, the scale in (3) can be computed as $\Delta^2 = \max_{i,j \in \mathrm{supp}(p)} \{\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}\}$.

Theorem 3.1 (Gaussian entropic adaptation)**.** *For every model above with $\Delta > 0$,*

$$K_B(t) := \log \mathbb{E} e^{t(B - \mathbb{E}B)} \le \tfrac{1}{2}\psi(ct), \qquad 0 \le t < 1/c. \tag{8}$$

where $\psi(s) = -s - \log(1-s)$ for $0 \le s < 1$. If $\Delta = 0$, then $B = 0$ almost surely.

The right-hand side of (8) is the centered log-MGF of a $\mathrm{Gamma}(1/2, c)$ random variable. Thus Theorem 3.1 is an exponential-moment comparison, not a claim that B is Gamma distributed. After the problem-dependent scale $c = \Delta^2/(4\tau)$ is removed, the same bound applies independently of the Gaussian dimension, number of alternatives, prior probabilities, correlation structure, and noise magnitude. The theorem is therefore stronger than the local quadratic approximation: it retains the same Gamma benchmark even after the Gibbs weights have changed substantially under the noise.

Corollary 3.2 (Variance and upper deviations)**.** *Let $q_\alpha(x)$ be the unique nonnegative solution of*

$$q - \alpha \log(1 + q/\alpha) = x, \qquad \alpha > 0, \quad x \ge 0. \tag{9}$$

Then $\mathrm{Var}(B) \le c^2/2$, and for $c > 0$ and $x > 0$,

$$\mathbb{P}\{B - \mathbb{E}B \ge c\, q_{1/2}(x)\} \le e^{-x}. \tag{10}$$

In particular, $q_{1/2}(x) \le \sqrt{x} + x$. For $c = 0$, the centered remainder is identically zero.

The corollary converts the MGF comparison into two usable quantities. The variance bound gives the second-order fluctuation scale, while the tail bound turns a failure probability into a deviation radius. These are the ingredients used below to obtain high-probability policy-stability certificates.

Sharpness. Let B_u denote the remainder obtained after replacing every feature a_i by ua_i.

Proposition 3.3 (The extremal small-noise limit)**.** *Fix $\Delta > 0$. If $C = \mathrm{Cov}_p(a)$, then*

$$\frac{4\tau B_u}{u^2\Delta^2} \longrightarrow Q_C := \frac{2G^\top CG}{\Delta^2} \quad \text{as } u \downarrow 0. \tag{11}$$

The centered MGFs converge for $s < 1$, and $K_{Q_C}(s) \le \psi(s)/2$ for $0 \le s < 1$. For two equally weighted distinct feature locations, $Q_C \overset{d}{=} Z^2/2$, $Z \sim N(0,1)$, with equality at every $s \in [0,1)$.

The proposition explains both the Gamma shape and the optimal coefficient. After the first-order response has been removed, the normalized small-noise remainder converges to a Gaussian quadratic form. Diagonalizing C expresses this limit as a weighted sum of squared independent Gaussians, and the balanced two-action case has only one active direction, giving exactly $Z^2/2$. Hence the coefficient $1/2$ in (8) cannot be uniformly reduced over the full model class.

Uniform sharpness does *not* imply that a fixed model is always most variable at zero noise. For binary weights $(p, 1-p)$, with $q = 1-p$, independent score errors uG_1, uG_2, $\tau = 1$, and $W_u = 2B_u/u^2$,

$$\mathrm{Var}(W_u) = 8p^2q^2 + \frac{8}{3}p^2q^2(11 - 56pq)u^2 + O(u^4). \tag{12}$$

The correction is positive at $p = 0.1$, so a same-prior quadratic upper bound cannot hold at every noise level. This is not an effect of an imbalanced prior. For a uniform prior on the vertices of a regular simplex, Proposition B.1 shows that once the number of alternatives is at least nine, the normalized variance rises above its zero-noise limit and reaches its maximum at positive noise magnitude. These examples separate two statements that are easy to conflate: a model need not be dominated by its own quadratic limit, yet every model still satisfies the Gamma MGF bound in Theorem 3.1. The finite-noise theorem therefore requires an argument beyond the local quadratic approximation.

4 Why the Gamma bound holds at arbitrary noise levels

The proof tracks the Hessian of the Gibbs log-partition function as Gaussian noise is revealed rather than freezing it at its zero-noise value. Details are in Section A; here we summarize the three ingredients. For $\Delta > 0$, center and rescale the features as $x_i = 2(a_i - \bar{a})/\Delta$ and define

$$h(z) = \log \sum\nolimits_i p_i e^{\langle x_i, z\rangle}, \quad \beta = \frac{\Delta}{2\tau}, \quad v = \beta^2, \quad W = \frac{B}{c} = \frac{h(\beta G) - h(0) - \beta \nabla h(0)\cdot G}{\beta^2}. \tag{13}$$

The target becomes $\log \mathbb{E}e^{s(W - \mathbb{E}W)} \le \psi(s)/2$ for $0 \le s < 1$.

Covariance bounds under Gaussian smoothing. The Hessian of h is the covariance matrix of an exponentially tilted feature distribution. For a centered tilted feature vector X, write $C = \mathrm{Cov}(X)$ and $T = \mathrm{tr}\, C$. The diameter constraint implies the coupled inequalities

$$0 \preceq C \preceq I, \qquad \mathrm{tr}(C^2) + (T-1)^2 \le 1, \qquad \mathrm{Var}(\|X\|^2) \le 2[2T - T^2 - \mathrm{tr}(C^2)]. \tag{14}$$

These bounds control both directional and total covariance. Let $P_w f(z) = \mathbb{E}[f(z + \sqrt{w}Z)]$, $Z \sim N(0, I)$, and $\bar{C} = P_w \nabla^2 h(z)$. Draw the Gaussian parameter, then a feature from its tilted law, and let C_* be the covariance of this mixture. The law of total covariance and projection onto linear Gaussian functions give

$$\bar{C} + w\bar{C}^2 \preceq C_*. \tag{15}$$

For each w, C_* obeys the same diameter constraint. The matrix inequality therefore yields decreasing uniform upper bounds on the smoothed covariance:

$$\|\bar{C}\|_{\mathrm{op}} \le \alpha(w) = \frac{2}{1 + \sqrt{1+4w}}, \qquad \|\bar{C}\|_{\mathrm{HS}}^2 \le b(w) = \frac{2y-1}{y^4}, \quad y^2(y-1) = w, \tag{16}$$

with $y \ge 1$; see Lemma A.1. These envelopes decrease with smoothing variance. An individual model's smoothed covariance can increase because C_* also depends on w. The proof uses the shrinking admissible bounds rather than a Hessian frozen at zero noise.

The nonzero average gradient. A purely second-order argument would be simplest if $\mathbb{E}\nabla W = 0$. However, $\nabla B(0) = 0$ is only a local cancellation: asymmetry in the prior weights or feature configuration can make the Gaussian average of the gradient nonzero. At an exponentially reweighted feature distribution, let $\zeta = \nabla \operatorname{tr} C$. The same diameter-based covariance control yields

$$\|P_w\zeta\|^2 \le \gamma(w) := \max_{0\le\rho\le1} \frac{2\rho(1-\rho^2)}{1+2w\rho}, \qquad \|\mathbb{E}\nabla W\|^2 \le D(v) := \frac{1}{4v}\left(\int_0^v \sqrt{\gamma(w)}\,dw\right)^2. \tag{17}$$

The quantity $D(v)$ therefore controls the contribution that remains because of asymmetry. Tracking it explicitly is essential: replacing it by a generic first-order estimate would lose the constant needed for the sharp Gamma comparison; see Lemma A.2.

Martingale comparison. The final step must combine covariance fluctuation and average-gradient bias without maximizing them independently. Revealing the standard Gaussian input through Brownian motion produces a gradient martingale whose initial value is controlled by (17) and whose fluctuations are controlled by (16). Writing $q \in [0,1]$ for remaining Brownian time, an exponential-supermartingale argument gives

$$\log \mathbb{E}e^{s(W-\mathbb{E}W)} \le b_s(1) + \tfrac{1}{2}D(v)a_s(1), \qquad \begin{cases} a_s'(q) = [s+\alpha(vq)a_s(q)]^2, & a_s(0)=0, \\ b_s'(q) = \tfrac{1}{2}b(vq)a_s(q), & b_s(0)=0. \end{cases} \tag{18}$$

The important point is that the Hessian fluctuation and the mean-gradient term are propagated through the same comparison. The remaining high-dimensional problem reduces to the scalar condition

$$D(v)+I(v) \le \tfrac{1}{2}, \qquad I(v) := \int_0^1 qb(vq)\,dq \implies b_s(1) + \tfrac{1}{2}D(v)a_s(1) \le \tfrac{1}{2}\psi(s) \quad (0 \le s < 1). \tag{19}$$

For $0 \le v \le 8$, the premise is verified by explicit positive-coefficient polynomial identities; for $v \ge 8$, a simpler uniform-gradient bound is already sufficient. The same analysis gives

$$\operatorname{Var}(W) \le \min\{D(v)+I(v),\, 4/v\} < \tfrac{1}{2} \qquad (v>0), \tag{20}$$

showing that the Gamma variance constant is approached as noise tends to zero but is not attained by a fixed nondegenerate model at positive noise. See Lemmas A.3 to A.5 for the reduction. Thus the local quadratic behavior identifies the sharp benchmark, while contracting covariance bounds and the gradient correction explain why the same benchmark remains valid at arbitrary noise magnitudes.

5 Consequences and extensions for decision making

We next translate the concentration theorem into statistical and decision-level guarantees.

5.1 Policy stability and sample complexity

Let $x = \log(1/\delta)$ with $0 < \delta < 1$. Combining Corollary 3.2 with $\mathbb{E}B \le c$ and $B = \tau D(p\|\widehat{p})$ yields

$$\mathbb{P}\Big\{D(p\|\widehat{p}) > \frac{\Delta^2}{4\tau^2}[1+q_{1/2}(x)]\Big\} \le \delta. \tag{21}$$

The threshold depends only on pairwise score-difference variances and temperature; it does not require knowing policy p or unknown mean $\mathbb{E}B$. When $\Delta^2 = \Delta_0^2/n$, as for sample means,

$$n \ge \max\left\{1, \left\lceil \frac{\Delta_0^2}{4\tau^2\epsilon}[1+q_{1/2}(x)]\right\rceil\right\} \tag{22}$$

is sufficient for forward KL at most ϵ, giving $O(1/\epsilon)$ sample complexity. For independent reward observations of variance σ^2, this becomes

$$n \ge \max\left\{1, \left\lceil \frac{\sigma^2}{2\tau^2\epsilon}[1+q_{1/2}(\log(1/\delta))]\right\rceil\right\}. \tag{23}$$

The forward KL also has a decision-theoretic interpretation. For $V_\tau(r) = \tau\log\sum_i e^{r_i/\tau}$,

$$V_\tau(r+\xi) - V_\tau(r) = p^\top\xi + B. \tag{24}$$

Thus $B = \tau D(p\|\widehat{p})$ is the additional apparent gain from re-optimizing on noisy scores. This gives the forward KL a direct decision-theoretic interpretation rather than treating it only as a statistical distance. By contrast, evaluating the noisy policy under the population rewards produces the reverse divergence $D(\widehat{p}\|p)$; see Section B.8. The covariance need not be known in advance.

Corollary 5.1 (Unknown reward covariance)**.** *Let $Y_1, \dots, Y_n \sim N_k(r, \Sigma_0)$ be independent, with $n \geq 2$, $k \geq 2$, and let S be the unbiased sample covariance matrix. Define $\widehat{p} = \mathrm{softmax}(\overline{Y}/\tau)$, $p = \mathrm{softmax}(r/\tau)$, and $\widehat{v} = \max_{i<j}(S_{ii} + S_{jj} - 2S_{ij})$, $\nu = n - 1$, $u_\nu(\eta) = \frac{\nu}{\chi^2_{\nu,\eta}}$.*

Here $\chi^2_{\nu,\eta}$ is the lower η quantile of χ^2_ν. Then, for $\delta, \eta > 0$ with $\delta + \eta < 1$, we have:

$$\mathbb{P}\Big\{ D(p\|\widehat{p}) > \frac{u_\nu(\eta)\widehat{v}}{4n\tau^2}[1 + q_{1/2}(\log(1/\delta))] \Big\} \leq \delta + \eta. \tag{25}$$

The sample covariance estimates the largest pairwise score-difference variance, while the chi-square factor accounts for estimation uncertainty. The failure probability is split between the adaptation event controlled by δ and covariance-estimation event controlled by η. The correction has no explicit factor depending on action count: the argument protects the population pair attaining largest variance rather than constructing simultaneous confidence intervals for all pairs. The proof is in Section B.4.

5.2 Stronger guarantees from structured action spaces

The certificate above is deliberately worst-case: it uses only the largest pairwise feature distance and ignores how the alternatives are arranged. When the feature set has additional geometry, the same analysis can yield stronger comparisons. Two useful cases are collinear features and product-structured decisions. Additional structure can improve the generic diameter scale. If all feature vectors lie on one affine line, we obtain a stronger comparison.

Theorem 5.2 (Scalar convex order)**.** *Let $Z \sim N(0,1)$ and $f \in C^2(\mathbb{R})$ satisfy $f(0) = f'(0) = 0$ and $0 \leq f'' \leq c$. Then*

$$f(Z) - \mathbb{E}f(Z) \leq_{\mathrm{cx}} \tfrac{c}{2}(Z^2 - 1). \tag{26}$$

Here $X \leq_{\mathrm{cx}} Y$ denotes convex order*: $\mathbb{E}\phi(X) \leq \mathbb{E}\phi(Y)$ for every convex function ϕ for which both expectations are finite, with the two variables having the same mean. In particular, the theorem applies whenever all feature vectors a_i lie on one affine line.*

Thus collinear uncertainty is convex-order dominated by the universal benchmark $c(Z^2 - 1)/2$. A model's local quadratic uses $f''(0)$, which may be smaller than c and need not dominate finite-noise fluctuations. For product-structured decisions, the gain can be larger.

Theorem 5.3 (Product priors with dependent scores)**.** *Suppose $p_{\boldsymbol{i}} = \prod_{j=1}^m p_j(i_j)$ and $a_{\boldsymbol{i}} = a_0 + \sum_{j=1}^m b_{j,i_j} v_j$. Let ℓ_j be the range of b_{j,i_j} over the positive support, $c_j = \ell_j^2 \|v_j\|^2/(4\tau)$, and $C = \sum_j c_j$. The directions v_j need not be orthogonal. Then $C \leq c$ and*

$$K_B(t) \leq \tfrac{1}{2}\psi(Ct), \qquad \mathbb{E}B \leq C/2, \qquad \mathbb{E}e^{tB} \leq (1 - Ct)^{-1/2} \quad (0 \leq t < 1/C). \tag{27}$$

If $C = 0$, then $B = 0$.

Product structure replaces the diameter of the full combinatorial action set by the sum of component scales. This can be substantially smaller because the generic diameter first combines all component variations and then squares the total distance. The component directions v_j need not be orthogonal, and the score errors need not be independent because dependence is already represented through the shared Gaussian vector G. Section C.3 gives a binary-product example where the improvement grows linearly with the number of components.

5.3 Entropy-regularized planning via trajectory-level Gibbs distributions

Consider a finite-horizon MDP with finite state and action spaces, known deterministic transitions, a fixed initial state, and entropy coefficient τ. Writing $f_h(s,a)$ for the next state and $r_h(s,a)$ for the population reward, the population soft Bellman recursion is

$$\begin{aligned} Q_h(s,a) &= r_h(s,a) + V_{h+1}(f_h(s,a)), \\ V_h(s) &= \tau \log \sum\nolimits_a e^{Q_h(s,a)/\tau}, \qquad \pi_{*,h}(a \mid s) = e^{[Q_h(s,a) - V_h(s)]/\tau}. \end{aligned} \tag{28}$$

Suppose the reward error at time h for (s, a) is $\langle a_h(s, a), G\rangle$, and define the perturbed recursion by replacing r_h with its noisy estimate. Rather than applying concentration at each Bellman update, treat each feasible trajectory $\omega = (S_1, A_1, \ldots, S_H, A_H)$ as one Gibbs alternative and define $A(\omega) = \sum_{h=1}^{H} a_h(S_h, A_h)$, $\Delta_{\mathrm{traj}} = \mathrm{diam}\{A(\omega) : \omega \text{ feasible}\}$. The trajectory law is the distribution over feasible trajectories induced by the policy and deterministic dynamics. This is useful because the product of soft action probabilities in (28) telescopes: intermediate value terms cancel, leaving a Gibbs distribution over total trajectory reward. Thus a sequential planning problem can be analyzed by one application of the static Gibbs theorem rather than accumulating one-step concentration bounds. If Q_* and $\widehat{Q}$ are the population and perturbed trajectory laws, this gives

$$\widehat{V}_\tau - V_\tau - \langle \mathbb{E}_{Q_*} A, G\rangle = \tau D(Q_* \| \widehat{Q}). \tag{29}$$

Hence Theorem 3.1 applies directly with trajectories as alternatives. This viewpoint avoids applying a separate concentration bound at every Bellman update and automatically preserves correlations created by shared states, edges, or reward coordinates because all trajectories depend on the same Gaussian vector. For independent time-indexed reward means of variance σ^2/n, we have:

$$\Delta_{\mathrm{traj}}^2 \le 2H\sigma^2/n, \qquad c_{\mathrm{traj}} \le H\sigma^2/(2n\tau). \tag{30}$$

If one stationary noisy reward estimate is reused along a trajectory, the same error can accumulate coherently before the diameter is squared, and the corresponding scale can instead be H^2. Thus horizon dependence is determined by the covariance structure of accumulated trajectory errors, not by the number of Bellman updates alone. Let $d_{*,h}(s) = \Pr_{\pi_*}(S_h = s)$. The KL chain rule gives:

$$D(Q_* \| \widehat{Q}) = H\mathcal{D}_H, \qquad \mathcal{D}_H := \frac{1}{H} \sum\nolimits_{h=1}^{H} \sum\nolimits_{s} d_{*,h}(s) D(\pi_{*,h}(\cdot \mid s) \| \widehat{\pi}_h(\cdot \mid s)). \tag{31}$$

Under the independent time-indexed design in (30),

$$\mathbb{P}\Big\{\mathcal{D}_H > \frac{\sigma^2}{2n\tau^2}\Big[1 + q_{1/2}\Big(\log\frac{1}{\delta}\Big)\Big]\Big\} \le \delta. \tag{32}$$

The factor H in the trajectory diameter cancels after averaging the KL divergence per decision, so the one-step budget (23) also suffices for $\mathcal{D}_H \le \epsilon$. This cancellation holds for the independent time-indexed design; it need not hold when the same uncertain reward estimate is repeatedly reused. Full trajectory derivations and a two-path extremal example are in Section B.10.

6 Experiments

Our experiments ask: when is the Gamma MGF bound nearly attained, how much data does it save relative to standard certificates, and do the predicted covariance effects persist in planning and large softmax action spaces? Full protocols and results are in Section D. For the MGF experiments, let $W = B/c$ and define the *tightness ratio* $R(s) = K_W(s)/[\psi(s)/2]$. By Theorem 3.1, $R(s) \le 1$; thus a value near one means the Gamma bound is nearly attained, while a smaller value indicates slack.

Tightness and finite-noise behavior We test whether the bound is nearly tight in the balanced two-action, small-noise regime identified by Proposition 3.3, and how it behaves away from that regime. We evaluate 11,324 feature/prior configurations, including Gaussian feature clouds, clustered configurations, outliers, collinear grids, and nearly coincident pairs. Low-dimensional MGFs are evaluated by Gauss–Hermite quadrature and higher-dimensional cases by importance sampling; independent CPU integration checks representative binary cases (Table 2 and Section D.6).

For two equally weighted alternatives with separation $\Delta = 0.001$, $R(s) = 0.999999$ at $s = 0.1$ and 0.999896 at $s = 0.98$, confirming the bound can be essentially attained. With prior $(0.3, 0.7)$, the small-noise ratio falls to 70% at $s = 0.1$, showing the bound becomes more conservative away from the extremal balanced case. No configuration is numerically flagged as violating the Gamma bound.

The experiment also separates the finite-noise theorem from the local quadratic approximation. At $p = 0.2$, $u = 0.5$, and $s = 0.98$, the actual log-MGF is $K_W(s) = 0.3075$, compared with $K_{Q_C}(s) = 0.1798$ for the quadratic approximation, so the actual log-MGF is 71 % larger than its quadratic approximation, while Theorem 3.1 still holds. Thus the quadratic limit explains the sharp constant but is not itself a finite-noise certificate.

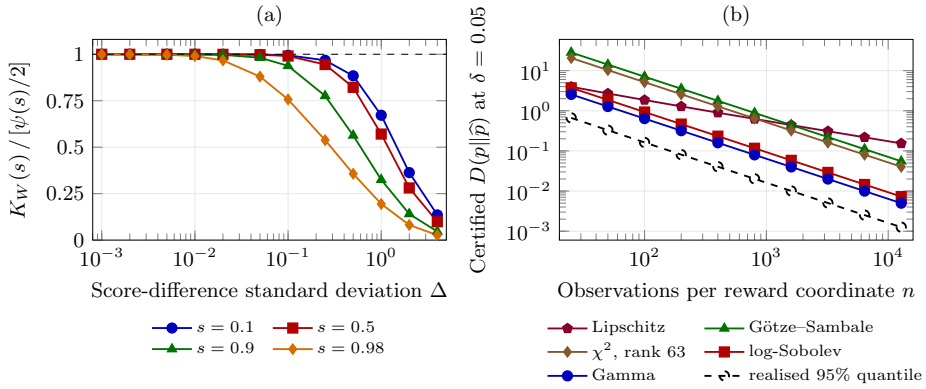

Figure 1: (a) Tightness ratio $R(s)$ for a balanced two-action prior, features $\pm\Delta/2$, and $\tau = 1$; one denotes equality in the Gamma MGF bound. (b) Forward-KL certificates for independent reward errors with $\sigma = 1$, $\delta = 0.05$, $k = 64$, and $\tau = 0.2$. The empirical 95% quantile uses a uniform population prior and 200,000 draws at each budget.

Practical cost of policy-stability certificates For target error ϵ and failure probability δ, we compare the sample size n sufficient to certify $D(p\|\widehat{p}) \le \epsilon$; smaller n means a tighter certificate. The main setting uses independent reward noise with $\sigma = 1$, $k = 64$, $\tau = 0.2$, $\epsilon = 0.05$, and $\delta = 0.05$. The Gamma certificate needs 1,277 observations per reward coordinate, versus 1,865 for log-Sobolev, a 32 % reduction (Tables 5 and 8). The first-order Lipschitz certificate is much costlier because its budget scales as ϵ^{-2} rather than ϵ^{-1} at fixed temperature. Structure can change the best certificate. The rank-dependent χ^2 bound is tighter below centered feature rank five but degrades quickly: at $\delta = 0.05$, its radius for B grows from $1.92c$ at rank one to $41.3c$ at rank 63, versus $5.11c$ for Gamma; KL radii divide these values by τ. Gamma remains 2.7–52 times the empirical 95% quantile across the Gaussian policy instances. With product structure, Theorem 5.3 reduces the radius by a factor of 35 at $m = 32$ (Table 6). Thus the generic bound improves concentration, while known structure should be exploited when available.

Planning under reward uncertainty We test whether the trajectory-level theory correctly predicts how policy error scales with horizon and reward correlation. For each horizon $H \in \{1, 8, 32, 128\}$, we generate eight deterministic MDPs with $S = 64$, $A = 4$, and $n = 200$, and draw 2,048 Gaussian reward perturbations per MDP. After each draw we recompute the soft policy and measure $\mathcal{D}_H$, the population-occupancy-weighted forward KL averaged over decisions.

Under independent time-indexed reward errors, the Gamma radius is independent of H, as predicted by (32), and the empirical means and 95% quantiles remain on a similar scale across horizons (Table 14). Only one of 65,536 reward draws exceeds the Gamma certificate; a simultaneous 95% upper bound on the configuration-specific exceedance probabilities is 0.0043, below the nominal 0.05. Thus the certificate is conservative here, but its horizon dependence matches the theory.

Why the reward representation matters. A two-path example isolates the role of correlation. Independent time-indexed errors give $\Delta^2_{\text{traj}} = 2H/n$, whereas reusing one reward along each path gives $2H^2/n$. At $H = 64$, measured per-decision KL differs by a factor of 63.90 (Table 15), matching the predicted factor-H change. Hence horizon dependence comes from how reward uncertainty is shared across time, not from the Bellman recursion itself.

Large-vocabulary softmax policies We use language-model outputs as a controlled large-action test: the goal is not to model realistic language-model uncertainty, but to check whether the certificate applies when the softmax has tens of thousands of alternatives. For fixed context, adding $G/\sqrt{n}$ to the final representation gives logit error $\xi_i = w_i^\top G/\sqrt{n}$, so unembedding rows w_i are Gaussian features. We evaluate six checkpoints across three families, with vocabularies up to 151,936 tokens.

The squared pairwise row diameter ranges from 1.5 to 57.5. At $\tau = 1$, $\delta = 0.05$, and $\epsilon = 0.01$, the sufficient budgets range from 195 to 7,340 representation observations per coordinate. With $n = 10{,}000$, 24 fixed contexts, and 1,024 Gaussian draws per context, the certificates are 15–27 times

the pooled empirical 95% quantiles, with no observed exceedance (Table 16). This shows that the theorem can be evaluated for vocabulary-scale action sets without an explicit action-count factor, although the generic diameter certificate can be conservative for a fixed model and context.

Estimating covariance from data Finally, we test whether the correction in Corollary 5.1 is needed when covariance is estimated from the same samples as the scores. Across 27 configurations, the equal-pair case with $n = 3$ is a small-sample stress test: plugging sample covariance into the known-covariance certificate yields an 8.51 % exceedance rate, versus the nominal 5 % and 8.55 % from independent integration. The calibrated certificate reduces this to 0.32 %. For correlated settings, median calibrated-to-known radius ratios are 1.43–1.45 at $n = 128$ and 1.21 at 512; for independent $k = 256$, they are 2.27 and 1.54. Thus covariance uncertainty matters at small n, while the correction shrinks with more data and adds no explicit action-count factor (Section D.12).

7 Related work

Our result sits between two lines of work: entropy-regularized decision making, where Gibbs distributions arise naturally, and second-order Gaussian concentration, which provides tools for nonlinear functions of Gaussian noise. Entropy regularization determines the exact nonlinear function whose stability we need to study, while Gaussian concentration determines how estimation noise propagates through that function. Our contribution is not a new decision rule, but a sharper probabilistic description of an existing and widely used one.

Regularization and noisy soft updates. Geist et al. (2019) develop the convex-duality framework for regularized MDPs, G-learning uses log-partition Bellman updates (Fox et al., 2016), and Fox (2019) corrects the bias of a binary Gaussian softmax-weighted reward average. Our target is the centered log-partition remainder at fixed temperature: the Gibbs identities are established, contribution is their sharp nonlinear concentration under correlated Gaussian perturbations. Chan et al. (2022) study forward and reverse KL in policy improvement toward a Boltzmann distribution, a question about improvement operator; here soft policy is fixed and reward errors move the policy selected from data.

Second-order Gaussian concentration. Second-order Gaussian concentration controls nonlinear Gaussian functionals using Hessian information rather than a global Lipschitz constant. Götze & Sambale (2020) give exponential concentration under Hessian bounds and a zero expected-gradient condition, not required here; their nonzero-gradient extension allows nonzero expected gradient and specialises to $\mathbb{E}\exp|B - \mathbb{E}B|/[(8 + 5\sqrt{2})c] \le 2$, applicable but not (8). Adamczak & Wolff (2013) control Gaussian functionals through norms of higher derivatives up to constants; our proof uses the exact covariance under the perturbed Gibbs probabilities to obtain the sharp MGF bound. The Gaussian log-Sobolev, interpolation and diffusion identities are standard (Bakry et al., 2014).

Gamma comparisons and Gaussian dilation. Chasapis et al. (2023) and Pang (2026) study Gamma sums and Gamma MGF bounds for Gaussian quadratic forms; both concern quadratic forms, whereas B at fixed nonzero noise is generally not itself a quadratic form. Dafnis & Paouris (2016, Theorem 1.1) compare moments of log-convex functions under Gaussian dilation, which does not bound the centered MGF at a fixed noise level. The distinction is that our centered MGF comparison holds for every noise magnitude and uses the optimal constant based on the pairwise feature diameter.

8 Discussion and conclusion

We studied how correlated Gaussian score errors affect Gibbs or softmax decisions. The identity $B = \tau D(p\|\widehat{p})$ reduces forward-KL policy stability to concentration of a log-partition remainder. Our main result shows that, after scaling by the largest variance of a pairwise score difference, the centered MGF of this remainder is bounded by a sharp Gamma-$1/2$ benchmark. The bound is valid at arbitrary noise levels: although the small-noise quadratic approximation identifies the sharp benchmark, extending it to finite noise requires decreasing covariance envelopes under Gaussian smoothing together with control of the average gradient and a martingale comparison. The resulting theory yields KL stability certificates, $O(1/\epsilon)$ sample complexity, covariance-estimation corrections, sharper guarantees under structural assumptions, and trajectory-level bounds for deterministic entropy-regularized planning. A natural direction is to replace the worst-case pairwise diameter by more instance-adaptive quantities while retaining explicit finite-sample guarantees.

Statement of Use of AI

The research questions, the main ideas, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure: refining the conceptual framing, stress-testing mathematical claims and proof arguments, and implementing and running the numerical experiments. In the theoretical component, generative AI was used in part as a "theory attacker": it assisted in implementing computational searches and large-scale numerical tests over extensive collections of generated examples in order to identify potential counterexamples, edge cases, or weaknesses in candidate theorem statements and proof arguments. These searches were used to help the authors examine and refine the theoretical claims. The failure of such searches to identify a counterexample does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, numerical testing, or search.

The authors independently verified all theorem statements and proofs, citations, experimental specifications, and AI-assisted text against their research artifacts. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature discovery and for drafting and editing the manuscript, including checks of notation and of the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

Reproducibility statement

All formal results have analytic proofs in the appendix. The two polynomial identities closing the main proof are printed with exact integer coefficients and independently reconstructed in rational arithmetic. Section D specifies the instance distributions, integration and sampling estimators, replication counts, and error budgets. CUDA and Triton calculations are checked against direct implementations and independent integration. The numerical studies evaluate proved statements and certificate size; simulations are not premises of the theorems. The experiments use synthetic Gaussian errors and fixed public-domain text contexts for frozen language models; no human-subject data or annotation is involved. The supplementary material provides the source code for all experiments and figures, together with the recorded outputs behind every reported result.

References

Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. https://arxiv.org/abs/1304.1826, 2013.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. doi: 10.1007/978-3-319-00227-9.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A. Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse KL divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022. URL https://www.jmlr.org/papers/v23/21-054.html.

Giorgos Chasapis, Salil Singh, and Tomasz Tkocz. Entropies of sums of independent gamma random variables. *Journal of Theoretical Probability*, 36(2):1227–1242, 2023. doi: 10.1007/s10959-022-01192-y. URL https://arxiv.org/abs/2201.08284.

William G. Cochran. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30 (2):178–191, 1934. doi: 10.1017/S0305004100016595. URL https://doi.org/10.1017/S0305004100016595.

Nikos Dafnis and Grigoris Paouris. An inequality for moments of log-concave functions on gaussian random vectors. https://arxiv.org/abs/1601.02492, 2016.

Roy Fox. Toward provably unbiased temporal-difference value estimation. In *Optimization Foundations for Reinforcement Learning Workshop at NeurIPS*, 2019. URL https://royf.org/pub/Fox2019EQL/.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016. URL https://www.auai.org/uai2016/proceedings/papers/219.pdf.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Friedrich Götze and Holger Sambale. Second order concentration via logarithmic Sobolev inequalities. *Bernoulli*, 26(1):93–126, 2020. doi: 10.3150/19-BEJ1118. URL https://arxiv.org/abs/1605.08635.

Zhekai Pang. A Gamma envelope and sharp moment inequalities for Gaussian quadratic forms. https://arxiv.org/abs/2609.05914, 2026. Preprint, submitted 5 September 2026.

A Proof of Theorem 3.1

The appendix gives the complete proof of the main theorem. The argument follows the four-step roadmap in Section 4: first establish covariance inequalities implied by a bound on the feature diameter, then derive contraction under Gaussian smoothing, next control the mean-gradient term and the gradient martingale, and finally reduce the MGF comparison to an explicit scalar inequality. We keep the proof self-contained so that every constant in the main theorem can be checked directly.

We prove the bound uniformly over finite feature laws, without a symmetry or independence assumption on their scores. Throughout this section, s denotes a normalized MGF argument, $\beta > 0$ denotes noise amplitude, and $v = \beta^2$ denotes noise variance. Brownian time is denoted by r, and $q = 1 - r$ is remaining time.

Proof notation. A symmetric matrix is *positive semidefinite* if $x^T M x \geq 0$ for every vector x, and $A \preceq B$ denotes the Loewner order, meaning that $B - A$ is positive semidefinite. For a symmetric matrix M, $\|M\|_{\mathrm{op}}$ is its operator norm, $\|M\|_{\mathrm{HS}} = \sqrt{\mathrm{tr}(M^T M)}$ its Hilbert–Schmidt/Frobenius norm, and $\mathrm{tr}(M)$ its trace. Gaussian convolution is $P_w f(x) = \mathbb{E}[f(x + \sqrt{w}Z)]$ with $Z \sim N(0, I)$; it averages f over an independent Gaussian perturbation. Brownian motion, used later to reveal the Gaussian input gradually, is a continuous-time process with Gaussian independent increments.

For clarity, write

$$B = \tau \log \sum_i p_i e^{a_i \cdot G/\tau} - \Big(\sum_i p_i a_i \Big) \cdot G, \qquad \Delta = \max_{i,j \in \mathrm{supp}(p)} \|a_i - a_j\|, \qquad c = \frac{\Delta^2}{4\tau}.$$

We prove the bound for every finite feature law and every $\tau > 0$; zero masses can be discarded, and zero diameter makes B identically zero.

A.1 Covariance bounds from the feature diameter and Gaussian smoothing

Lemma A.1 (Covariance and smoothing bounds)**.** *Let X be centered with support diameter at most two, and write $C = \mathbb{E}XX^T$, $T = \mathrm{tr}\, C$, $S = \mathrm{tr}(C^2)$, and $R = \|X\|^2$. Then*

$$0 \preceq C \preceq I, \qquad S + (T-1)^2 \leq 1, \qquad \mathrm{Var}(R) \leq 2(2T - T^2 - S). \tag{33}$$

The region $\mathcal{C} = \{H \succeq 0 : \mathrm{tr}(H^2) + (\mathrm{tr}\, H - 1)^2 \leq 1\}$ is convex and downward closed in Loewner order.

Let $h(x) = \log \sum_i p_i e^{x_i \cdot x}$ for a feature law of diameter at most two, and let $C(x) = \nabla^2 h(x)$ be its Gibbs-weighted covariance. If P_w is Gaussian convolution with covariance wI, then $\bar{C} = P_w C(x)$ satisfies

$$\|\bar{C}\|_{\mathrm{op}} \leq \alpha(w) := \frac{2}{1 + \sqrt{1 + 4w}}, \qquad \|\bar{C}\|_{\mathrm{HS}}^2 \leq b(w) := \frac{2y - 1}{y^4}, \quad y^2(y-1) = w, \quad y \geq 1. \tag{34}$$

Both profiles equal one at zero, and $b(w) \geq \alpha(w)^2$.

Interpretation. This lemma supplies the geometric control used throughout the finite-noise proof. A bounded feature diameter restricts both the size and the spectrum of every Gibbs-weighted covariance matrix. Gaussian smoothing then improves these restrictions according to the explicit profiles $\alpha(w)$ and $b(w)$. In the main argument, these profiles bound admissible curvature as unrevealed Gaussian noise is averaged out.

Proof. For an independent copy X', direct expansion gives

$$\mathbb{E}\|X - X'\|^4 = 4T^2 + 2\,\mathrm{Var}(R) + 4S \leq 4\mathbb{E}\|X - X'\|^2 = 8T.$$

This proves the last two assertions in (33); the operator bound follows from the scalar variance–range inequality. Convexity of $\mathcal{C}$ is immediate. For downward closure, ordered eigenvalues decrease under Loewner order. If their sum u is at most one, their squared sum plus $(u-1)^2$ is at most $u^2 + (u-1)^2 \leq 1$. If $u > 1$, the expression $\sum_j \lambda_j^2 + (\sum_j \lambda_j - 1)^2$ is coordinatewise nondecreasing along any increase of the nonnegative eigenvalues.

Sample the Gaussian parameter and then a feature from its posterior. The averaged feature law still has diameter at most two. Its covariance $C_* = \bar{C} + \mathrm{Cov}(m_\pi)$ obeys

$$\bar{C} + w\bar{C}^2 \preceq C_*.$$

Here Gaussian first-chaos projection gives $\mathrm{Cov}(m_\pi) \succeq w\bar{C}^2$: for the smooth bounded posterior mean function $F(G)$, subtracting its orthogonal projection $(\mathbb{E}DF)G$ proves $\mathrm{Cov}(F) \succeq (\mathbb{E}DF)(\mathbb{E}DF)^T$. Downward closure therefore puts $\bar{C} + w\bar{C}^2$ in $\mathcal{C}$. Its operator bound proves the first inequality in (34).

For the second, let $T_0 = \sum_j \lambda_j$, $S_0 = \sum_j \lambda_j^2$ for the eigenvalues of $\bar{C}$, and assume $S_0 > 0$. Moment Cauchy–Schwarz and Jensen give

$$\sum_j (\lambda_j + w\lambda_j^2)^2 \geq S_0(1 + wS_0/T_0)^2.$$

The coupled constraint then gives, with $\ell = S_0/T_0$, $(T_0 + \ell)(1 + w\ell) \leq 2$, and hence

$$S_0 \leq \frac{2\ell}{1 + w\ell} - \ell^2.$$

The right side is maximized when $\ell(1 + w\ell)^2 = 1$. Writing $\ell = y^{-2}$ gives the claimed $b(w)$. Evaluating the same expression at $\ell = \alpha(w)$ gives $\alpha(w)^2$, proving $b \geq \alpha^2$. The zero case and limits at $w = 0$ are immediate. □

Normalize the remainder to

$$W_\beta(G) = \beta^{-2}[h(\beta G) - h(0) - \beta\nabla h(0) \cdot G], \qquad v = \beta^2, \tag{35}$$

where the feature support of h has diameter at most two. Then $0 \leq W_\beta(G) \leq \|G\|^2/2$ and $\|\nabla W_\beta\| \leq 2/\beta$; both feature means in the gradient belong to the convex hull of a set of diameter at most two.

Lemma A.2 (Initial gradient bias)**.** *For* (35)*,*

$$\|\mathbb{E}\nabla W_\beta\|^2 \leq D(v) := \frac{(1-z)(1+z+z^2-z^3)^2}{(1+z)(1+z^2)}, \qquad v = \frac{3\sqrt{3}}{4}\frac{1-z^4}{z^6}, \quad 0 < z \leq 1. \tag{36}$$

The continuous value is $D(0) = 0$*.*

Interpretation. The covariance bounds alone are not sufficient because the smoothed gradient can have a nonzero mean when the Gibbs weights are asymmetric. This lemma bounds that deterministic bias by the explicit function $D(v)$. The bound vanishes at zero noise and complements the covariance profiles from the previous lemma, so the proof can track both the initial gradient offset and its subsequent random fluctuation using the same feature geometry.

Proof. At a natural parameter, center the posterior feature and put $\zeta = \mathbb{E}[X\|X\|^2] = \nabla \operatorname{tr} C$. Fix a smoothing variance w, and set $a(w) = P_w\zeta$, $\bar{C} = P_wC$, and $\rho = \|\bar{C}\|_{\mathrm{op}}$. Under the joint law of the Gaussian parameter and its centered posterior feature, the covariance of X and $R - T$ is $a(w)$. Thus

$$\|a(w)\|^2 \leq \rho P_w \operatorname{Var}_\pi(R).$$

Gaussian first-chaos projection of the scalar field T gives $\mathrm{Var}_G(T) \geq w\|P_w\nabla T\|^2 = w\|a(w)\|^2$. Together with $P_wS \geq \mathrm{tr}(\bar{C}^2)$ and (33), this yields

$$(1 + 2w\rho)\|a(w)\|^2 \leq 2\rho[2\operatorname{tr}\bar{C} - (\operatorname{tr}\bar{C})^2 - \mathrm{tr}(\bar{C}^2)] \leq 2\rho(1 - \rho^2).$$

Consequently $\|a(w)\|^2 \leq \gamma(w)$, where the maximum of $2\rho(1-\rho^2)/(1+2w\rho)$ on $[0,1]$ is attained at the unique $\rho(w) \in (0, 1/\sqrt{3}]$ satisfying

$$3\rho(w)^2 + 4w\rho(w)^3 = 1, \qquad \gamma(w) = 4\rho(w)^3.$$

Since $\Delta\nabla h = \nabla \operatorname{tr} C$, the heat identity gives

$$u := P_v\nabla h(0) - \nabla h(0) = \frac{1}{2}\int_0^v a(w)\,dw, \qquad \mathbb{E}\nabla W_\beta = u/\sqrt{v}.$$

Hence the squared bias is at most $(4v)^{-1}(\int_0^v \sqrt{\gamma(w)}\,dw)^2$. Substitution of $w = (1-3\rho^2)/(4\rho^3)$ evaluates the integral as

$$\rho(v)^{-3/2} + 3\sqrt{\rho(v)} - 2\,3^{3/4}.$$

Putting $z = 3^{1/4}\sqrt{\rho(v)}$ gives (36). All cumulants of the exponentially reweighted feature distribution used here are bounded for a fixed bounded feature support, so the heat differentiations are justified. The bounds have no dimension or support-size factor. □

A.2 Martingale comparison and elimination of the MGF argument

The martingale can be read as revealing the Gaussian reward error over time. At time r, M_r is the gradient averaged over the part of the noise still to be revealed. Its initial value is $\mathbb{E}\nabla W$, and its random changes are determined by the smoothed Hessian. The functions a_s and b_s in (18) track the resulting exponential bound. The first accounts for the size of the gradient; the second accumulates its fluctuations. Equation (36) bounds the initial gradient, and (34) bounds its subsequent changes. Both depend on the same feature vectors and their pairwise distances.

Lemma A.3 (Gradient-martingale comparison)**.** *Fix $v > 0$ and $0 \le s < 1$. Let $A(q), B(q)$ solve*

$$A' = (s + \alpha(vq)A)^2, \quad A(0) = 0, \qquad B' = \tfrac{1}{2}b(vq)A, \quad B(0) = 0. \tag{37}$$

Then

$$\log \mathbb{E}e^{s(W_\beta - \mathbb{E}W_\beta)} \le B(1) + \tfrac{1}{2}D(v)A(1).$$

Interpretation. This lemma is the step that combines the two quantities controlled earlier. The Brownian representation reveals the Gaussian perturbation gradually: the initial value of the gradient is controlled by $D(v)$, while its changes are controlled by the smoothed Hessian profiles α and b. The two scalar ODEs summarize these effects and convert them into an upper bound on the centered log-MGF. Thus the high-dimensional Gibbs problem is reduced to controlling the scalar expression $B(1) + \frac{1}{2}D(v)A(1)$.

Proof. Let B_r be standard Brownian motion. A stochastic process (M_r) is a martingale if its current value is the conditional expectation of its future value; a supermartingale allows the conditional expectation to decrease. For Brownian motion B_r, put

$$M_r = P_{1-r}\nabla W_\beta(B_r), \quad H_r = P_{1-r}\nabla^2 W_\beta(B_r), \quad Y_r = \int_0^r M_u \cdot dB_u.$$

The backward heat equation and Itô's formula (the stochastic analogue of the chain rule) give $dM_r = H_r\,dB_r$, $M_0 = \mathbb{E}\nabla W_\beta$, and $Y_1 = W_\beta(B_1) - \mathbb{E}W_\beta$. Remaining time $q = 1 - r$ corresponds to natural smoothing variance vq, so Lemma A.1 bounds H_r by $\alpha(vq)$ in operator norm and $b(vq)$ in squared HS norm. The drift, divided by the current value, of

$$\exp\{sY_r + \tfrac{1}{2}A(1-r)\|M_r\|^2 + B(1-r)\}$$

is

$$-\tfrac{1}{2}A'\|M_r\|^2 - B' + \tfrac{1}{2}A\,\mathrm{tr}(H_r^2) + \tfrac{1}{2}\|sM_r + AH_rM_r\|^2 \le 0.$$

The ODE is finite through $q = 1$, since $0 \le A(q) \le s^2q/(1-sq)$. Stop this positive stopped supermartingale (a stochastic process whose conditional expected future value cannot exceed its current value) on bounded state sets and apply Fatou's lemma at the terminal time. This proves the asserted bound after Lemma A.2. Bounded Hessians justify the heat calculation; localization does not assume the exponential bound being proved. □

Lemma A.4 (One variance condition controls all positive MGF parameters)**.** *Define*

$$I(v) = \int_0^1 q\,b(vq)\,dq.$$

If $D(v) + I(v) \le 1/2$, then the right side of Lemma A.3 is at most $\psi(s)/2$ for every $0 \le s < 1$.

Interpretation. The preceding martingale bound still depends on the MGF parameter s through the ODE solution. This lemma removes that complication: it shows that the single scalar inequality $D(v) + I(v) \le 1/2$ is sufficient to guarantee the desired Gamma bound simultaneously for every positive MGF parameter $0 \le s < 1$. In other words, an infinite family of exponential-moment inequalities is reduced to one variance-like condition depending only on the normalized noise level v.

Proof. On $L^2[0,1]$, consider the kernel

$$\mathsf{K}(u,w) = F(\min(u,w)), \qquad F(r) = \alpha(v(1-r)).$$

The function F is increasing and lies in $[0,1]$. Therefore

$$\mathsf{K} = F(0)\, 1 \otimes 1 + \int_0^1 1_{[r,1]} \otimes 1_{[r,1]}\, dF(r)$$

is positive and trace class, with trace $\int_0^1 F(r)\, dr$. The Schur bound gives $R := \|\mathsf{K}\|_{\mathrm{op}} \le 1$. Let K_q be its compression to $[1-q, 1]$.

The scalar deterministic-control process $dM_r = F(r) dB_r$, $dY_r = M_r dB_r$ gives, when started from zero, the L^2 Gaussian-quadratic expansion $Y_1 = \frac{1}{2}\sum_j \lambda_j (Z_j^2 - 1)$, where the Z_j are independent standard normals and the λ_j are eigenvalues of K. For a deterministic initial value at time $1-q$, the extra term is that value times $B_1 - B_{1-q}$. The Gaussian quadratic MGF formula therefore gives

$$B_0(s) = \frac{1}{2}\sum_j \psi(s\lambda_j), \qquad A(s,q) = s^2 \langle 1, (I - s\mathsf{K}_q)^{-1} 1\rangle,$$

where λ_j are the eigenvalues of K. The conditional Gaussian expectation, as a function of the initial value of M, has logarithm $A(s,q)M^2/2$ plus a constant. Its backward equation gives (37) with $b(vq)$ replaced by $\alpha(vq)^2$. Uniqueness of the finite Riccati solution therefore identifies the same A, and $B_0 = \frac{1}{2}\int_0^1 \alpha(vq)^2 A(s,q)\, dq$. These formulas follow first in finite Gaussian coordinates and then by positive trace-class approximation. For any $s < 1$, choose $s' \in (s,1)$; the finite-rank Gaussian formulas bound the s' exponential moments uniformly, justifying convergence of the MGFs as well as the L^2 limits.

Since $b - \alpha^2 \ge 0$, the Bellman upper bound has the representation

$$B(1) + \tfrac{1}{2} D(v) A(s,1) = \frac{1}{2}\sum_j \psi(s\lambda_j) + \frac{s^2}{2}\int_{[0,R]} \frac{\nu(d\lambda)}{1 - s\lambda}, \tag{38}$$

where ν is positive and has total mass $D(v) + \int_0^1 q[b(vq) - \alpha(vq)^2]\, dq$. Indeed, the spectral measure of the constant function for each K_q has mass q and is supported in $[0,R]$.

Put $S = \mathrm{tr}(\mathsf{K}^2) = 2\int_0^1 q\alpha(vq)^2\, dq$. Then $S \ge R^2$ and the assumed variance condition is $S/2 + \nu([0,R]) \le 1/2$. For every integer $n \ge 2$, the coefficient of s^n in (38) is at most

$$R^{n-2}\left[\frac{S}{2n} + \frac{1-S}{4}\right] \le R^{n-2}\left[\frac{R^2}{2n} + \frac{1-R^2}{4}\right] \le \frac{1}{2n}.$$

The last inequality is $x^a[(a+1) - ax] \le 1$ on $[0,1]$, with $x = R^2$ and $a = (n-2)/2$; its left side is nondecreasing. If $R = 0$, only an s^2 term remains and the conclusion is immediate. The nonnegative series converge for $s < 1$, proving the coefficientwise comparison with $\psi(s)/2$. □

A.3 An exact algebraic certificate for the scalar inequality

Lemma A.5 (Scalar closure)**.** *For every* $0 \le v \le 8$, $D(v) + I(v) \le 1/2$.

Interpretation. This lemma closes the moderate-noise part of the proof by verifying the scalar condition required by the previous lemma for every $0 \le v \le 8$. Its role is algebraic but decisive: once $D(v) + I(v) \le 1/2$ is established, the martingale comparison yields the target Gamma MGF bound throughout this range. The remaining large-noise regime $v \ge 8$ is easier and is handled directly by the standard Gaussian log-Sobolev bound in the proof of Theorem 3.1.

Proof. Let $y \geq 1$ solve $y^2(y-1) = v$. Direct integration gives

$$I(v) = \frac{2y^3 - \frac{13}{2}y^2 + 9y - 2\log y - \frac{9}{2}}{v^2}, \qquad I(0) = \frac{1}{2}.$$

Since $\log y \geq 2(y-1)/(y+1)$ for $y \geq 1$,

$$I(v) \leq I_{\mathrm{rat}}(y) := \frac{4y^2 - y - 1}{2y^4(y+1)}. \tag{39}$$

The logarithmic inequality follows by differentiation, with derivative $(y-1)^2/[y(y+1)^2] \geq 0$ and equality at one.

For $0 \leq v \leq 8$, the parameter z in (36) lies in $[7/10, 1]$, and $y \in [1, 12/5]$. Indeed, $v(z)$ is decreasing, $3\sqrt{3} > 51/10$ makes $v(7/10) > 8$, and $(12/5)^2(12/5 - 1) = 1008/125 > 8$.

Put $P(z) = 1 + z + z^2 - z^3$ and define, in ascending power order,

$$Q_D(z) = -4 - 12z - 24z^2 - 4z^3 + 29z^4 + 75z^5 + 78z^6 + 38z^7 - 31z^8 - 49z^9 - 20z^{10} + 20z^{11}.$$

The exact identity

$$5(1+z)^2(1+z^2)^2 - [20z^6 + 9(1-z^4)]P(z)^2 = (1-z)Q_D(z)$$

and positivity of Q_D on $[7/10, 1]$ imply

$$D(v) \leq \frac{v}{3\sqrt{3} + 9v/5} \leq \frac{10v}{51 + 18v}. \tag{40}$$

Next put

$$Q_V(y) = -51 - 102y + 120y^2 + 120y^3 - 21y^4 - 2y^6 - 2y^7.$$

Using $v = y^2(y-1)$, a second exact identity is

$$\frac{1}{2} - I_{\mathrm{rat}}(y) - \frac{10v}{51 + 18v} = \frac{(y-1)Q_V(y)}{2y^4(y+1)(51 + 18v)}.$$

Thus positivity of Q_V on $[1, 12/5]$ completes the proof.

For completeness, both polynomial signs are certified by the finite positive-coefficient identities

$$(10+u)^{11} Q_D\left(\frac{7+u}{10+u}\right) = \sum_{j=0}^{11} d_j u^j, \qquad (1+5w)^7 Q_V\left(\frac{1+12w}{1+5w}\right) = \sum_{j=0}^{7} e_j w^j,$$

whose exact integer coefficients are displayed below. The substitutions map $u, w \geq 0$ onto the required intervals apart from their upper endpoints; continuity handles those endpoints. Every coefficient is positive, so no numerical approximation or external interval certificate is required.

j	d_j	e_j
0	240833519760	62
1	686698839540	4886
2	596563830807	127848
3	269022859128	1609818
4	74896412238	10802589
5	13901887620	37731057
6	1780602867	55449261
7	158643360	4535169
8	9681840	—
9	386616	—
10	9108	—
11	96	—

At $v = 0$, the continuous values are $D(0) = 0$ and $I(0) = 1/2$. □

Proof of Theorem 3.1. Discard zero masses. If $\Delta = 0$, all scores coincide and $B = 0$. Otherwise put

$$x_i = \frac{2(a_i - \bar{a})}{\Delta}, \qquad \bar{a} = \sum_i p_i a_i, \qquad \beta = \frac{\Delta}{2\tau}, \qquad v = \beta^2.$$

The x_i have diameter two and mean zero, and $B/c = W_\beta$. For $0 < v \le 8$, Lemmas A.3, A.4, and A.5 give $\log \mathbb{E}e^{s(W_\beta - \mathbb{E}W_\beta)} \le \psi(s)/2$ for every $0 \le s < 1$. For $v \ge 8$, Gaussian log-Sobolev and $\|\nabla W_\beta\| \le 2/\sqrt{v}$ give directly

$$\log \mathbb{E}e^{s(W_\beta - \mathbb{E}W_\beta)} \le \frac{2s^2}{v} \le \frac{s^2}{4} \le \frac{1}{2}\psi(s).$$

Substitute $s = ct$ to obtain the asserted physical-unit bound. All arguments take place in the finite span of the centered features, so singular score covariance causes no difficulty. □

Variance as a function of the normalized noise magnitude. The same proof gives a strict, explicit variance refinement. With $v = \beta^2 > 0$ and the scalar functions above,

$$\mathrm{Var}(W_\beta) \le \min\{D(v) + I(v),\, 4/v\} < \frac{1}{2}. \tag{41}$$

Indeed, the coefficient of s^2 in the Bellman comparison is $[D(v) + I(v)]/2$. The positive polynomial identities make $D(v) + I(v) < 1/2$ for $0 < v \le 8$, while Gaussian Poincare gives $\mathrm{Var}(W_\beta) \le 4/v < 1/2$ for $v > 8$. Multiplication by c^2 yields the corresponding profile for B.

B Sharpness, baseline bounds, and decision identities

B.1 Covariance bounds and the small-noise quadratic limit

For centered X supported on a set of diameter Δ, let $C = \mathrm{Cov}(X)$, let Y be an independent copy, and put $q = \operatorname{tr} C$. Scalar projection gives $\|C\|_{\mathrm{op}} \le \Delta^2/4$. Direct expansion and Jensen give

$$\mathbb{E}\|X - Y\|^4 = 2\mathbb{E}\|X\|^4 + 2q^2 + 4\operatorname{tr}(C^2) \ge 4q^2 + 4\operatorname{tr}(C^2).$$

The diameter bound also gives $\mathbb{E}\|X - Y\|^4 \le \Delta^2 \mathbb{E}\|X - Y\|^2 = 2\Delta^2 q$. Consequently

$$\operatorname{tr}(C^2) \le \Delta^2 q/2 - q^2 \le \Delta^4/16.$$

Both the operator and Hilbert–Schmidt bounds are attained by two equiprobable points at distance Δ.

Proof of Proposition 3.3. Take $X = a_I - \bar{a}$ with $I \sim p$. Taylor expansion of the finite log-partition function gives (11) pointwise in G. The Hessian bound gives $0 \le 4\tau B_u/(u^2\Delta^2) \le \|G\|^2/2$. Dominated convergence therefore gives convergence of the means and the centered MGFs for positive parameters for each fixed $s < 1$. For eigenvalues $\lambda_j = 4\ \mathrm{eig}_j(C)/\Delta^2$, the covariance bounds imply $0 \le \lambda_j \le 1$ and $\sum_j \lambda_j^m \le 1$ for every integer $m \ge 2$. Hence

$$K_{Q_C}(s) = \frac{1}{2}\sum_j \psi(s\lambda_j) = \frac{1}{2}\sum_{m \ge 2} \frac{s^m}{m} \sum_j \lambda_j^m \le \frac{1}{2}\psi(s), \qquad 0 \le s < 1.$$

For two equally weighted feature locations, the only nonzero eigenvalue is one, proving the equality and sharpness assertions. For negative s, convergence of MGFs follows from nonnegativity of the normalized remainder and convergence of its mean. □

B.2 Standard concentration and the nonzero mean gradient

The covariance-diameter inequalities imply

$$\|\nabla^2 B\|_{\mathrm{op}} \le c, \quad \|\nabla^2 B\|_{\mathrm{HS}}^2 \le c^2, \quad \|\nabla B\|^2 \le 2cB, \quad \mathbb{E}B \le c. \tag{42}$$

Equation (4) and the covariance bounds prove the first two inequalities in (42). For a nonnegative convex function with c-Lipschitz gradient, the descent inequality at $x - \nabla B(x)/c$ gives $\|\nabla B(x)\|^2 \le$

$2cB(x)$. To bound the mean, define $f(x) = \tau \log \sum_i p_i e^{a_i \cdot x/\tau} - \bar{a} \cdot x$. Gaussian heat differentiation yields

$$\frac{d}{du}\mathbb{E}f(\sqrt{u}G) = \tfrac{1}{2}\mathbb{E}\,\mathrm{tr}\,\nabla^2 f(\sqrt{u}G) = \frac{1}{2\tau}\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}_{\widehat{p}(\sqrt{u}G)}(a).$$

For every probability law π on the feature set, $\mathrm{tr}\,\mathrm{Cov}_\pi(a) = \frac{1}{2}\mathbb{E}_{\pi\otimes\pi}\|a_I - a_J\|^2 \le \Delta^2/2$. Integrating from zero to one gives $\mathbb{E}B \le \Delta^2/(4\tau) = c$.

Let $H(t) = \log \mathbb{E}e^{tB}$. Gaussian log-Sobolev applied to $e^{tB/2}$ and the self-bounding gradient imply

$$tH'(t) - H(t) \le \tfrac{1}{2}t^2\mathbb{E}_t\|\nabla B\|^2 \le ct^2H'(t).$$

For $0 < t < 1/c$, integration with $H'(0) = \mathbb{E}B$ gives $H(t) \le t\mathbb{E}B/(1-ct)$ and hence the first bound in (6). The gradient is the difference of two feature means, so $\|\nabla B\| \le \Delta$. Gaussian Lipschitz concentration gives the second bound. Finite support makes all derivatives used here integrable and B globally Lipschitz.

For the stated specialization of Götze & Sambale (2020), take $f = (B - \mathbb{E}B)/c$ and unit Gaussian log-Sobolev constant. Then $\|\nabla^2 f\|_{\mathrm{op}} \le 1$, $\mathbb{E}\|\nabla^2 f\|_{\mathrm{HS}}^2 \le 1$, and $\|\mathbb{E}\nabla f\|^2 \le \mathbb{E}\|\nabla f\|^2 \le 2$. Their choices $b^2 = 1$ and $b_0 = \sqrt{2}$ give

$$\mathbb{E}\exp\left\{\frac{|B - \mathbb{E}B|}{(8+5\sqrt{2})c}\right\} \le 2.$$

The zero expected-gradient hypothesis of their zero-gradient theorem is not implied by $\nabla B(0) = 0$.

B.3 Chernoff optimization and computable certificates

Proof of Corollary 3.2. Expansion of (8) at zero gives $\mathrm{Var}(B)/2 \le c^2/4$. For normalized deviation $q > 0$, maximize $sq - \alpha\psi(s)$ on $0 \le s < 1$. The maximizer is $s = q/(q+\alpha)$ and its value is $q - \alpha\log(1 + q/\alpha)$. This function is strictly increasing for $q > 0$, starts at zero, and diverges, proving the quantile definition and (10) with $\alpha = 1/2$. Finally $\psi(s) \le s^2/[2(1-s)]$. The resulting sub-gamma bound with variance factor $1/2$ and scale one gives deviation $\sqrt{x} + x$; equivalently its Chernoff rate is no larger than the exact rate above, so $q_{1/2}(x) \le \sqrt{x} + x$. □

Equation (21) follows from $B = \tau D(p\|\widehat{p})$ and $\mathbb{E}B \le c$. Substitution of $\Delta^2 = \Delta_0^2/n$ and rounding upward gives (22).

For completeness, let P project onto the span of all $a_i - \bar{a}$. Since the linear common component cancels, $B(G) = B(PG)$, $B(0) = 0$, $\nabla B(0) = 0$, and $\nabla^2 B \preceq cP$ on this span. Taylor's integral formula yields $B \le (c/2)\|PG\|^2$. Therefore $\mathbb{P}\{B > (c/2)\chi^2_{\rho,1-\delta}\} \le \delta$, where $\rho = \mathrm{rank}\,P$. For any fixed score vector ξ, Hoeffding's lemma for the categorical variable ξ_I/τ, $I \sim p$, gives $B \le (\max_i \xi_i - \min_i \xi_i)^2/(8\tau)$.

B.4 Covariance estimated from the same reward observations

Write $v_{ij} = \mathrm{Var}(Y_{1i} - Y_{1j})$ and $v_* = \max_{i<j} v_{ij}$. If $v_* = 0$, all reward errors are common to the actions, and both $D(p\|\widehat{p})$ and $\widehat{v}$ vanish. Otherwise fix a pair (i_*, j_*) attaining $v_* > 0$; this pair depends only on the population covariance. For its sample contrast variance $S_* = S_{i_*i_*} + S_{j_*j_*} - 2S_{i_*j_*}$,

$$\frac{\nu S_*}{v_*} \sim \chi^2_\nu, \qquad S_* \le \widehat{v}. \tag{43}$$

This is the Gaussian residual-sum-of-squares pivot (Cochran, 1934): project the n independent standardized contrast observations onto the subspace orthogonal to the constant vector. The projection has rank $n-1$, so its squared norm is χ^2_{n-1}. No invertibility of Σ_0 is required. It follows that

$$\mathbb{P}\{v_* > u_\nu(\eta)\widehat{v}\} \le \mathbb{P}\{\nu S_*/v_* < \chi^2_{\nu,\eta}\} = \eta. \tag{44}$$

Only one population-maximizing pair enters this proof. Requiring an interval for every contrast would solve a stronger problem than estimating the single maximum used by the Gibbs certificate.

Since $\overline{Y} - r \sim N(0, \Sigma_0/n)$, the known-covariance certificate (21) gives

$$\mathbb{P}\left\{ D(p\|\widehat{p}) > \frac{v_*}{4n\tau^2}[1 + q_{1/2}(\log(1/\delta))] \right\} \le \delta.$$

On the intersection of this event's complement and that of (44), substituting the covariance upper bound gives (25). A union bound proves the result. This argument allows the sample mean and sample covariance to use the same observations; it does not condition on the estimated covariance. The sampling budget, temperature, action set and error levels are fixed before observing the data.

Sharpness of the covariance multiplier. Take a scalar standard Gaussian Z_ℓ and features $a_i \in \{-1, 1\}$, with both signs present, and set $Y_{\ell i} = r_i + a_i Z_\ell$. All nonzero empirical contrast variances are $4s_Z^2$, so $\widehat{v}/v_* = s_Z^2$ and $\nu\widehat{v}/v_* \sim \chi^2_\nu$ exactly. Therefore any constant $u < u_\nu(\eta)$ makes $\mathbb{P}\{v_* > u\widehat{v}\} > \eta$. This sharpness concerns covariance calibration; the final KL radius also uses the Gamma tail and its mean allowance.

Selecting the error split. For a target failure probability α, fix $\delta = \alpha - \eta$ and choose a split by numerical search on

$$u_{n-1}(\eta)[1 + q_{1/2}(\log(1/(\alpha - \eta)))] \quad \text{over } 0 < \eta < \alpha.$$

This scalar objective depends only on n and α. Thus numerical search can precede the observations and does not select a confidence level from the observed outcomes. Any fixed split gives the stated bound. For comparison, Bonferroni calibration replaces η by $\eta/\binom{k}{2}$ in u_{n-1} and uses the same data. It is valid but unnecessary for the maximum-variance target.

B.5 Binary and simplex variance expansions

The binary expansion (12) follows by expanding $W_u = 2B_u/u^2$ to fourth order. For the uniform-prior analogue, let X be uniform on the $d+1$ centered vertices of a regular simplex of diameter one, set $b_u = u^{-2} \log \mathbb{E}_X e^{u\langle X, G\rangle}$, and let $\lambda = 1/[2(d+1)]$.

Proposition B.1 (A finite-noise variance maximum)**.** *For each fixed $d \ge 1$,*

$$\mathbb{E}b_u = \frac{d\lambda}{2} - \frac{d\lambda^2}{4}u^2 + O_d(u^4), \tag{45}$$

$$\mathrm{Var}(b_u) = \frac{d\lambda^2}{2} + \frac{d(d-7)\lambda^3}{6}u^2 + O_d(u^4). \tag{46}$$

For $k = d + 1 \ge 9$, the normalized variance exceeds its zero-noise limit for sufficiently small positive u and attains a global maximum at some finite positive u.

This result shows that uniform sharpness of Theorem 3.1 is different from an instancewise claim that every model is most variable at zero noise. A fixed model can move above its own local quadratic behavior while remaining below the common Gamma MGF bound.

For binary probabilities (p, q), $q = 1 - p$, independent unit Gaussians, and noise amplitude u, let $Z = (G_1 - G_2)/\sqrt{2}$. Expanding $W_u = 2B_u/u^2$ gives

$$W_u = 2pqZ^2 + \frac{2\sqrt{2}}{3}pq(1-2p)uZ^3 + \frac{1}{3}pq(1-6pq)u^2Z^4 + O_{L^2}(u^3).$$

Using $\mathbb{E}Z^6 = 15$ and $\mathrm{Cov}(Z^2, Z^4) = 12$ proves (12). Expanding two further orders and using Gaussian symmetry gives the stated $O(u^4)$ remainder. Cumulants of the finite bounded feature law are bounded at every fixed order, which justifies these L^2 Taylor expansions.

Proof of Proposition B.1. The centered simplex vertices satisfy $\|X\|^2 = d\lambda$, $\mathrm{Cov}(X) = \lambda I_d$, and $\langle x_i, x_j\rangle = -\lambda$ for distinct vertices. Set

$$T_3(G) = \mathbb{E}_X\langle X, G\rangle^3, \qquad T_4(G) = \mathbb{E}_X\langle X, G\rangle^4 - 3(\lambda\|G\|^2)^2.$$

Taylor expansion yields $b_u = Q + uT_3/6 + u^2T_4/24 + \cdots$, where $Q = \lambda\|G\|^2/2$. Gaussian contractions give

$$\mathbb{E}T_4 = -6d\lambda^2, \qquad \mathbb{E}T_3^2 = 6d(d-1)\lambda^3, \qquad \mathrm{Cov}(Q, T_4) = -12d\lambda^3.$$

For the cubic identity, $\mathbb{E}_{X,Y}\langle X, Y\rangle^3 = d(d-1)\lambda^3$ and the remaining Wick term vanishes by constant radius and $\mathbb{E}X = 0$. For the last identity use $\mathrm{Cov}(\|G\|^2, T_4) = 4\mathbb{E}T_4$ for a homogeneous quartic. These equalities prove (45)–(46). Bounded feature support, Taylor expansion to sufficiently high order, Gaussian moments, and cancellation of odd terms justify the $O_d(u^4)$ remainders.

For $d \geq 8$ the variance increases strictly above its zero-noise limit for sufficiently small $u > 0$. The bound $0 \leq b_u \leq \max_i \|x_i\|\|G\|/u$ gives $b_u \to 0$ in L^2 as $u \to \infty$. Continuity, including at zero, then implies attainment of a global maximum at a finite positive u. □

B.6 Reading the scale from the reward covariance

Reading the scale from the reward covariance. The feature representation is a convenient way to write correlated Gaussian errors. A reader with the covariance matrix $\Sigma = \mathrm{Cov}(\xi)$ can compute the relevant scale directly:

$$\Delta^2 = \max_{i,j \in \mathrm{supp}(p)} \{\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}\}. \tag{47}$$

Thus the bound depends on uncertainty in reward differences. For example, suppose $\xi_i = \sigma_0 Z_0 + \sigma Z_i/\sqrt{n}$, where $Z_0, Z_1, \ldots, Z_k$ are independent standard Gaussians and the support has at least two alternatives. The shared error $\sigma_0 Z_0$ changes every score by the same amount and cancels from the softmax probabilities. Equation (47) gives $\Delta^2 = 2\sigma^2/n$, regardless of σ_0. This is useful when several action values are estimated from shared data: the covariance between their errors is part of the stability calculation.

Linear reward error and policy adaptation have different scales. For a fixed reward-sampling design, write $\xi_i = \langle a_i^{(0)}, G\rangle/\sqrt{n}$ and let Δ_0 be the diameter of the unscaled features. At fixed τ, the mean and variance bounds become

$$\mathbb{E}B \leq \frac{\Delta_0^2}{4n\tau}, \qquad \mathrm{Var}(B) \leq \frac{\Delta_0^4}{32n^2\tau^2}. \tag{48}$$

In comparison, the linear term $p^\top \xi$ has variance $p^\top \Sigma_0 p/n$, where Σ_0 is the unscaled noise covariance. Hence the soft value error contains a first-order Gaussian term and a second-order adaptation term. Theorem 3.1 controls the latter at finite sample size, while Proposition 3.3 identifies the limiting case that determines its constant. This distinction explains why a bound on the log-partition function itself can miss the scale of the change in policy.

B.7 Protocol for the simplex illustration

For a uniform prior on the k centered vertices of a regular simplex of diameter u, score errors can be represented as $uZ_i/\sqrt{2}$ after removing their common shift. On each Gaussian draw define the two quadratic variables at their respective scales:

$$Q_W = \sum_i p_i Z_i^2 - \Big(\sum_i p_i Z_i\Big)^2 \sim \frac{\chi^2_{k-1}}{k}, \qquad Q_b = \frac{Q_W}{4}.$$

Since $b_u = W/4$, $\mathbb{E}Q_b = (k-1)/(4k) = d\lambda/2$ and $\mathrm{Var}(Q_b) = (k-1)/(8k^2) = d\lambda^2/2$. The paired estimators use

$$\mathbb{E}b_u - \mathbb{E}Q_b = \tfrac{1}{4}\mathbb{E}[W - Q_W], \qquad \mathrm{Var}(b_u) - \mathrm{Var}(Q_b) = \tfrac{1}{16}[\mathrm{Var}(W) - \mathrm{Var}(Q_W)]. \tag{49}$$

The same draw evaluates W and Q_W, reducing the sampling error in these $O(u^2)$ differences. The implementation divides the mean difference by four and the variance difference by sixteen.

Each noise level uses 256 independent replicates of 2^{12} antithetic pairs for $k = 64$ and 2^{14} pairs for $k \leq 16$. Different noise levels use distinct random streams. Points are replicate means and bars are 95% Student half-widths over replicates. The plug-in variance estimator has a finite-sample bias of order the inverse number of pairs. The exact Triton Gaussian draws are replayed through a separate CPU moment calculation; all five accumulated moments agree within $3 \cdot 10^{-13}$. A CPU checker separately verifies (49).

B.8 THE TWO DIVERGENCES OF A SOFT DECISION

The two KL identities in a soft decision. For a distribution q over actions, define its regularized value by

$$J_r(q) = \sum_i q_i r_i - \tau \sum_i q_i \log q_i. \tag{50}$$

The optimizer is $p = \mathrm{softmax}(r/\tau)$ and its value is $V_\tau(r)$. Substituting $\log p_i = r_i/\tau - V_\tau(r)/\tau$ gives

$$\begin{aligned} V_\tau(r) - J_r(\widehat{p}) &= \tau D(\widehat{p}\|p), \\ V_\tau(r+\xi) - J_{r+\xi}(p) &= \tau D(p\|\widehat{p}) = B. \end{aligned} \tag{51}$$

The first line evaluates the learned policy under population rewards. The second evaluates the gain from changing the population policy under the observed rewards. Theorem 3.1 controls the fluctuation of the second quantity. This form is particularly convenient for a stability statement: its certificate in (21) can be evaluated from the noise covariance even when the population policy is unknown.

B.9 REGULARIZED PERFORMANCE AND TRAJECTORY KL

Let $J_r(q) = \sum_i q_i r_i - \tau \sum_i q_i \log q_i$. The softmax identity gives $V_\tau(r) - J_r(q) = \tau D(q\|p)$, so plugging in $q = \widehat{p}$ establishes the performance-loss orientation. Applying the same identity with perturbed rewards and $q = p$ proves (24) and the forward-KL representation of B.

For a finite-horizon MDP with fixed transition kernels, a fixed initial state and $V_{H+1} = 0$, let $P_h V_{h+1}$ denote the next-state value averaged under the transition kernel. The deterministic recursion is (28).

The soft Bellman identity is $r_h + P_h V_{h+1} - V_h = \tau \log \pi_{*,h}$. Summing under any history-dependent policy law Q gives

$$V_\tau(r) - J_r(Q) = \tau \mathbb{E}_Q \sum_h \log \frac{\pi_h(A_h \mid \text{history})}{\pi_{*,h}(A_h \mid S_h)} = \tau D(Q\|Q_*).$$

The trajectory KL chain rule cancels the common transition kernels. Applying the identity to perturbed rewards with comparator Q_* proves (29), also for stochastic transitions. The additional step needed for concentration is the Gibbs representation. With deterministic dynamics, the trajectory law of any history-dependent policy has entropy equal to the sum of its conditional action entropies. Every distribution on feasible trajectories can be implemented by its conditional action probabilities. The entropy variational principle therefore gives

$$\widehat{Q}(\omega) = \frac{Q_*(\omega) e^{\langle A(\omega), G\rangle/\tau}}{\sum_{\omega'} Q_*(\omega') e^{\langle A(\omega'), G\rangle/\tau}}.$$

This is exactly the setting of Theorem 3.1. The soft-optimal policy is Markov, although history-dependent policies can be used in the variational argument.

For independent time-indexed reward errors, a trajectory uses at most one coordinate at each time. Two feature sums differ in at most two coordinates per time, each of squared magnitude σ^2/n, proving (30). Repeated use of the same noisy coordinate can instead produce a coefficient of order H, whose squared contribution is of order H^2.

B.10 SHARED EDGES AND A TWO-PATH EXTREMIZER

Here the values are evaluated at the fixed initial state. To see the Gibbs structure directly, multiply the action probabilities in (28) along a feasible trajectory. The value terms telescope, giving

$$\begin{aligned} Q_*(\omega) &= \exp\left\{ \frac{\sum_{h=1}^H r_h(S_h, A_h) - V_1(S_1)}{\tau} \right\}, \\ \widehat{Q}(\omega) &= \frac{Q_*(\omega) e^{\langle A(\omega), G\rangle/\tau}}{\mathbb{E}_{\omega' \sim Q_*} e^{\langle A(\omega'), G\rangle/\tau}}. \end{aligned} \tag{52}$$

The denominator is a finite sum over feasible trajectories. This identity proves that the one-step theorem applies to the entire planning problem, with trajectories as alternatives.

Why shared edges are handled jointly. Two trajectories can use the same reward coordinate for several decisions, and different trajectories can share many edges. The feature sum $A(\omega)$ records these overlaps. Consequently the concentration argument acts on one Gaussian vector and the full trajectory remainder. It does not sum independent concentration bounds for Bellman updates. For time-indexed errors, shared edges cancel from $A(\omega) - A(\omega')$, leaving at most two distinct coordinates at each time, which yields (30). If a stationary reward estimate is reused, repeated visits add their coefficients before the squared diameter is computed. The same certificate then uses this resulting diameter. This keeps the sampling assumptions and the dependence on the horizon tied to the actual reward representation.

A two-path example attains the horizon variance scale. Consider an initial binary decision followed by two disjoint, deterministic paths of length H. All later actions are forced, and all population rewards are zero. The population policy chooses each path with probability $1/2$. Under the independent time-indexed sampling model, the total estimated rewards of the two paths are independent variables $X_+, X_- \sim N(0, H\sigma^2/n)$, where $\sigma > 0$. The fitted policy chooses the first path with probability $e^{X_+/\tau}/(e^{X_+/\tau} + e^{X_-/\tau})$. Subtracting the linear error $(X_+ + X_-)/2$ from the soft value gives the exact adaptation remainder

$$B_{\mathrm{path}} = \tau \log\cosh\left(\frac{X_+ - X_-}{2\tau}\right) \overset{d}{=} \tau \log\cosh\left(\frac{\sigma\sqrt{H}}{\tau\sqrt{2n}} Z\right), \qquad Z \sim N(0,1). \tag{53}$$

Here $\Delta_{\mathrm{traj}}^2 = 2H\sigma^2/n$ and $c_{\mathrm{traj}} = H\sigma^2/(2n\tau)$. For fixed H and τ, Taylor expansion of $\log\cosh$ gives

$$\frac{B_{\mathrm{path}}}{c_{\mathrm{traj}}} \longrightarrow \frac{Z^2}{2}, \qquad n^2 \operatorname{Var}(B_{\mathrm{path}}) \longrightarrow \frac{H^2\sigma^4}{8\tau^2} \quad \text{as } n \to \infty. \tag{54}$$

The bound $\log\cosh z \le z^2/2$ supplies an integrable dominating function for the required second moments. Thus the trajectory variance bound has the exact H^2/n^2 leading scale in this planning example. The average policy change also satisfies $n\mathbb{E}\mathcal{D}_H \to \sigma^2/(4\tau^2)$. This connects the binary extremal law in Proposition 3.3 to a finite-horizon decision problem: the sharp fluctuation can arise from one choice between two long, independently estimated paths.

C Proofs for scalar and product structure

C.1 Mixtures of Gamma contractions

Let $Q \sim \mathrm{Gamma}(1/2, 1)$, with density $\rho(q) = e^{-q}/\sqrt{\pi q}$ and survival function S. Suppose finitely many maps $g_j : [0,\infty) \to [0,\infty)$ are nondecreasing, one-Lipschitz and C^1 on $(0,\infty)$, with $g_j(0) = 0$, and J is independent of Q. We first prove that $g_J(Q)$ has law $T(Q)$ for an anchored nondecreasing one-Lipschitz map T. The maps needed below are continuously differentiable on $(0,\infty)$.

Initially assume $\eta \le g_j' \le 1$ for some $\eta > 0$. Write $r_j = g_j^{-1}$ and $I(v) = \rho(S^{-1}(v))$. For $x = S^{-1}(v)$,

$$I''(v) = \frac{(\log\rho)''(x)}{\rho(x)} = \frac{1}{2x^2\rho(x)} > 0.$$

For $Y = g_J(Q)$, with $\mathbb{P}(J = j) = w_j$, its survival and density obey

$$S_Y(y) = \sum_j w_j S(r_j(y)), \qquad \rho_Y(y) = \sum_j w_j \rho(r_j(y)) r_j'(y) \ge \sum_j w_j I(S(r_j(y))) \ge I(S_Y(y)).$$

The quantile transport $T(q) = S_Y^{-1}(S(q))$ therefore satisfies

$$0 < T'(q) = \frac{\rho(q)}{\rho_Y(T(q))} \le 1, \qquad T(0) = 0.$$

For $0 \le g_j' \le 1$, use $g_{j,\eta} = (1-\eta)g_j + \eta\,\mathrm{id}$. The corresponding transports have a locally uniformly convergent subsequence by their common Lipschitz bound and value at zero. Its limit is anchored, nondecreasing and one-Lipschitz. Almost sure convergence of both $T_\eta(Q)$ and $g_{J,\eta}(Q)$ identifies

their limiting laws, including atoms. All first moments converge since these variables are bounded by Q.

For any integrable nonnegative X and anchored nondecreasing one-Lipschitz T, let $H = \mathrm{id} - T$. Both T and H are nondecreasing. For a smooth convex Lipschitz test function ϕ, put $X_\theta = T(X) - \mathbb{E}T(X) + \theta(H(X) - \mathbb{E}H(X))$. Then

$$\frac{d}{d\theta}\mathbb{E}\phi(X_\theta) = \mathrm{Cov}(\phi'(X_\theta), H(X)) \geq 0.$$

The covariance sign follows by an independent-copy identity. Integrating and approximating stop-loss functions proves $T(X) - \mathbb{E}T(X) \leq_{\mathrm{cx}} X - \mathbb{E}X$. For exponential tests with $X = Q$ and $s < 1$, differentiation is dominated by a multiple of $(1+Q)e^{sQ}$.

C.2 Proof of the scalar and collinear theorem

If $c = 0$, the assumptions give $f = 0$ and the assertion follows. For $c > 0$, scale to $c = 1$. The curvature assumptions give $0 \leq f'(z)/z \leq 1$ for $z \neq 0$ and $0 \leq f(z) \leq z^2/2$. The branch maps $g_\pm(q) = f(\pm\sqrt{2q})$ are anchored, nondecreasing and one-Lipschitz. The sign of a Gaussian is independent of $Q = Z^2/2$, so $f(Z)$ is the equally weighted mixture of these maps. The preceding transport and centered-contraction argument proves Theorem 5.2. It also gives the raw bound $\mathbb{E}e^{tf(Z)} \leq (1-ct)^{-1/2}$.

For collinear features $a_i = a_0 + b_i v$ with $\|v\| = 1$, the common feature cancels and $B = f(\langle v, G\rangle)$, where

$$f(z) = \tau \log \sum_i p_i e^{b_i z/\tau} - z\sum_i p_i b_i, \qquad f''(z) = \tau^{-1}\,\mathrm{Var}_{\widehat{p}(z)}(b_i) \leq \frac{(\max b_i - \min b_i)^2}{4\tau}.$$

The last variance bound follows from $(b - \min b)(\max b - b) \geq 0$. This proves the claimed application, including arbitrary binary covariance.

C.3 Product structure over binary components

Example: actions with several binary components. Suppose an action is a vector $\boldsymbol{i} \in \{-1,1\}^m$, the population distribution factors over its components with positive probabilities for both choices, and its reward-error feature is $a_{\boldsymbol{i}} = a_0 + \sum_j i_j v_j$. Each component has range two, so Theorem 5.3 uses

$$C = \frac{1}{\tau}\sum_{j=1}^m \|v_j\|^2, \qquad c = \frac{1}{\tau}\max_{\varepsilon\in\{-1,1\}^m}\left\|\sum_{j=1}^m \varepsilon_j v_j\right\|^2. \tag{55}$$

If all directions are equal to a nonzero vector v, the same Gaussian error drives every component. Nevertheless, factorization gives $C = m\|v\|^2/\tau$, compared with the general diameter scale $c = m^2\|v\|^2/\tau$. If the directions are orthogonal, then $C = c$ and the component remainders are independent; their separate MGF bounds can be added. The product theorem therefore captures structure of the action distribution as well as dependence in the reward errors. The product certificate uses C, a sum of m component scales, even though the action set contains 2^m combinations.

C.4 Proof of product closure

Product probabilities and additive features factor the partition function, so $B = \sum_j B_j$, with each B_j a scalar remainder of curvature at most c_j. For positive c_j, put $w_j = c_j/C$. Hölder and the scalar theorem give

$$K_B(t) \leq \sum_j w_j K_{B_j}(t/w_j) \leq \frac{1}{2}\sum_j w_j \psi(c_j t/w_j) = \frac{1}{2}\psi(Ct).$$

All factors are finite for $Ct < 1$. The same calculation with the raw scalar MGF gives $\mathbb{E}e^{tB} \leq (1-Ct)^{-1/2}$; also $\mathbb{E}B = \sum_j \mathbb{E}B_j \leq C/2$.

For each factor choose its positive-support endpoints, and put $d_j = \ell_j v_j$. Every $\sum_j \epsilon_j d_j$, $\epsilon_j \in \{-1, 1\}$, is a difference between two points of the positive Cartesian support. Averaging its squared norm over independent symmetric signs gives

$$\sum_j \|d_j\|^2 = \mathbb{E}_\epsilon \left\| \sum_j \epsilon_j d_j \right\|^2 \le \Delta^2.$$

Thus $C \le \Delta^2/(4\tau)$ and Theorem 5.3 follows. For orthogonal factors, independence also yields the sharper noise-dependent bound $K_B(t) \le \frac{1}{2}\sum_j \psi(c_j t)$. Its diameter equality uses the positive-prior support.

D Numerical protocol and extended results

This appendix records how the reported quantities are computed. Nothing here is a premise of a theorem; the experiments test statements that are proved analytically and measure quantities that the theory does not predict, such as the conservatism of the resulting certificates.

D.1 What is evaluated

Temperature is absorbed before any computation. Writing $u_i = a_i/\tau$,

$$W := \frac{B}{c} = \frac{4b(u)}{\Delta_u^2}, \qquad b(u) = \log \sum_i p_i e^{\langle u_i, G\rangle} - \sum_i p_i \langle u_i, G\rangle, \tag{56}$$

with $\Delta_u = \max_{i,j \in \operatorname{supp} p} \|u_i - u_j\|$. Theorem 3.1 is the statement $K_W(s) \le \psi(s)/2$ for $0 \le s < 1$, and the tightness ratio $R(s) = K_W(s)/[\psi(s)/2]$ is the quantity reported. The reduction in (56), the invariance of b under a common translation of the features, the identity checks against a naive reference implementation, and the inequalities $\mathbb{E}W \le 1$, $\operatorname{Var} W \le 1/2$ and $\operatorname{tr}(C^2) \le \Delta^4/16$ are all verified numerically before any sweep; the recorded discrepancies are at the level of double-precision rounding.

Feature configurations and a second MGF-bound check. The input scale is the supported diameter of the normalized features u, not a multiplier applied to an unnormalized random cloud. After generating a cloud, we divide by its measured supported diameter and multiply by the desired Δ_u. The saved features independently recover this diameter. All prior draws use the stage's explicit random generator; positive Gamma draws used to generate Dirichlet priors are floored at 10^{-30} before normalization. For each sweep instance, the supported action count is uniform on $\{2, \dots, k_{\max}\}$, and the Dirichlet concentration is log-uniform on $[0.02, 50]$. The normalized diameter is independently log-uniform on $[0.001, 30]$. Five feature families are chosen with equal probability: standard Gaussian clouds; two random-sign clusters about a random unit direction with Gaussian jitter of scale 0.05; one unit-direction outlier and jittered remaining points; a regular grid on that line; and two opposite endpoints with all remaining points at zero. Only supported coordinates enter the diameter. The geometry ablation fixes dimension one, eight supported actions, and a Dirichlet(1) prior, sharing each draw across its four diameters.

Since $\|\nabla b\| \le \Delta_u$, Gaussian Lipschitz concentration gives

$$K_W(s) \le \frac{8s^2}{\Delta_u^2}, \qquad \frac{K_W(s)}{\psi(s)/2} \le \frac{32}{\Delta_u^2} \quad (0 < s < 1). \tag{57}$$

Every saved MGF is checked against both this Gaussian Lipschitz MGF upper bound and the Gamma MGF upper bound, using its own quadrature discrepancy or replicate interval plus a 10^{-8} rounding allowance. At $\Delta_u = 10$, a ratio above 0.32 would therefore fail a basic consistency check even if below one. Both checks use each instance's actual supported diameter and retain its numerical diagnostics.

D.2 Deterministic quadrature

For effective dimension $m \le 3$, we evaluate the Gaussian integral by Gauss–Hermite quadrature. The one-dimensional families use orders $(256, 512)$ or $(512, 1024)$, the two-dimensional family

$(96, 160)$, and the three-dimensional family $(64, 96)$ in each coordinate. For standard-normal nodes t_q, weights ω_q and design scale σ,

$$\mathbb{E}f(Z) \approx \sum_q \omega_q \, \sigma e^{-(\sigma^2-1)t_q^2/2} f(\sigma t_q). \tag{58}$$

The plain design, $\sigma = 1$, supplies both the reported mean and MGF. Its node-doubling discrepancy in $K_W(s)$, plus s times the mean discrepancy, defines the flag threshold. Three additional designs use $\sigma = (1-s)^{-1/2}$, $\min\{1, 3/\Delta\}$, and 4 as cross-checks. Their disagreement is recorded separately. The threshold measures numerical convergence; it is not a rigorous bound on quadrature error. The design ablation in Table 11 compares these choices against higher-order evaluation without changing the reported estimator.

D.3 Importance-sampled Monte Carlo

Each integration pass uses 16 independent replicates, each containing 2^{15} antithetic pairs, or 2^{20} individual Gaussian vectors across replicates. A single plain pass is shared across all 12 MGF parameters; each of the six parameters above $1/2$ has a separate importance-sampling pass. The tilts are $0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.98$. Beyond three dimensions the tensor grid is replaced by sampling. For $s \le 1/2$ the mean and the moment generating function are accumulated from the same draws, so that the two noise terms of $K_W(s) = \log \mathbb{E}e^{sW} - s\mathbb{E}W$ cancel to leading order. Each term is $O(s)$ while their difference is $O(s^2)$, so estimating them from independent draws would swamp the signal at small MGF parameters. Above $s = 1/2$ the estimator is importance sampled from $N(0, (I - 2sM)^{-1})$, where $g^\top Mg$ with $M = 2C/\Delta^2$ is the zero-noise quadratic limit of W. The covariance-diameter inequality gives $\|M\|_{\mathrm{op}} \le 1/2$, so this proposal is defined for every $s < 1$ and broadens the proposal in directions with large quadratic curvature. For $s > 1/2$, the proposal covariance is $\Lambda_s = (I - 2sM)^{-1}$, with density ratio $w_s(g) = \det(\Lambda_s)^{1/2} \exp\{-g^\top (I - \Lambda_s^{-1})g/2\}$. In each replicate, we average $w_s(G)e^{sW(G)}$ under that proposal and subtract s times the mean of W from the independent plain pass after taking the logarithm. The reported estimate averages these replicate log-MGFs; approximate two-sided 99.9% Student intervals use their sample standard deviation divided by four, with 15 degrees of freedom. For each fixed finite feature model, the remainder is Lipschitz, so the plain exponential estimator also has finite variance; importance sampling addresses its numerical size, not existence. Confidence intervals are formed from independent replicates of the whole estimator, so correlations between its parts are accounted for; a flag is counted when the lower replicate-based interval exceeds the Gamma MGF upper bound. These intervals are approximate and pointwise across the sweep.

D.4 Implementation

The remainder, its moments and all MGF parameters are accumulated inside one fused CUDA kernel, one block per instance, in double precision; no per-sample or per-node tensor is materialised, so memory is flat in the number of draws or nodes. Independent score errors, where the effective dimension equals the number of alternatives, are handled by a fused Triton kernel that keeps a whole instance in registers and draws its Gaussian coordinates in place. The planning and language-model pipelines use full-graph compiled PyTorch. All runs are on one consumer GPU with 16 GB of memory; the six-model sweep reaches 8.1 GB. The 11,324-configuration MGF sweep has a recorded elapsed time of 123.2 seconds; this is the time of that numerical stage, not a training-speed comparison.

D.5 Implementation validation

Table 1 records the checks run before any sweep. The fused kernels are compared against a naive PyTorch evaluation of the same integral, the temperature reduction and the translation invariance of the remainder are verified numerically, and the structural inequalities the proof uses are recomputed on random instances.

D.6 Independent integration and flag audit

The active sweep contains the freshly seeded configurations reported in Table 2. Features, priors, actual diameters, MGFs and numerical diagnostics are saved for every entry. Neither MGF upper

Table 1: Implementation checks. The first block reports maximum absolute discrepancies, the second the extreme value attained by a quantity the theory bounds.

quantity	value	expected
fused kernel against a naive reference, $\mathbb{E}W$	$2.5{\cdot}10^{-16}$	0
fused kernel against a naive reference, $K_W(s)$	$4.6{\cdot}10^{-16}$	0
temperature reduction (56)	$2.4{\cdot}10^{-11}$	0
invariance under a common feature translation	$4.5{\cdot}10^{-14}$	0
importance-sampled Monte Carlo against quadrature	$2.1{\cdot}10^{-4}$	0
$\mathbb{E}W$ over 20,000 random instances	0.6055	≤ 1
Var W over the same instances	0.499986	$\le 1/2$
$16\,\mathrm{tr}(C^2)/\Delta^4$ over the same instances	$1 - 3.6{\cdot}10^{-6}$	≤ 1

bound is clipped into the estimates. CPU checks of the saved arrays recover their pairwise distances and covariance quantities and test (57) as well as the Gamma MGF upper bound.

For each of the twenty feature-configuration–diameter cells in Table 9, the record stores the point attaining the largest ratio, including its feature vectors, probabilities, temperature, and MGF parameter. A separate CPU adaptive integral recomputes these points, with maximum absolute MGF difference below 10^{-12}. At $\Delta_u = 10$, the two-cluster and Gaussian-cloud maxima are 0.057643 and 0.046733, both below the Lipschitz ceiling 0.32.

A separate CPU implementation uses adaptive integration with extended-precision binary log-partition evaluation. It checks priors $p \in \{0.1, 0.2, 0.5\}$, separations $\Delta \in \{0.001, 0.05, 1/\sqrt{2}, 3\}$ and $s \in \{0.1, 0.5, 0.98\}$. The exponentially reweighted integral uses a rescaled Gaussian variable to resolve its tail. At the matched equal-pair points and the reported failure of quadratic domination, the maximum absolute difference from the GPU records is $5.9{\cdot}10^{-8}$. Adaptive integration tolerances are numerical diagnostics, separate from the analytic proof.

The certificate audit also checks temperature units directly against $\mathbb{E}B + \Delta\sqrt{2x}$ and checks the Gaussian-range integrator at $k = 2$ against $\sqrt{2}\Phi^{-1}(1 - \delta/2)$.

D.7 Extended numerical results

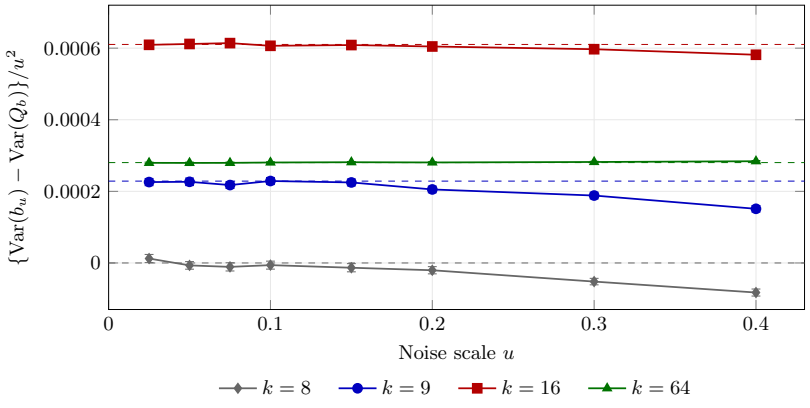

Figure 2: Normalized variance can initially increase with noise for uniform simplex priors; $Q_b = \lambda\|G\|^2/2$ is the quadratic limit of b_u. Dashed lines are the coefficients in (46); points use 2^{22} Gaussian antithetic pairs for $k \le 16$ and 2^{20} for $k = 64$, with the quadratic limit on the same draws, and bars are 95% half-widths over 256 replicates. The case $k = 8$, where the displayed correction vanishes, is included as a null control. The protocol is in Section B.7.

Table 2 reports the numerical sweep of Section 6 family by family.

Table 2: Numerical screening of (8) over 11,324 random instances and 135,888 instance–MGF-parameter comparisons, at twelve MGF parameters from $s = 0.05$ to $s = 0.98$. "Max ratio" is $\max K_W(s)/[\psi(s)/2]$; "max excess" is $\max\{K_W(s) - \psi(s)/2\}$, negative when the Gamma MGF upper bound is never reached. "Diagnostic" is the node-doubling discrepancy, or the Monte-Carlo standard error relative to the Gamma MGF upper bound. For the sampled families the ratio is a point estimate with that standard error, and values slightly above one require the uncertainty to be considered. A flag uses the entry's own convergence threshold or lower replicate interval; zero flags are a numerical finding, not an integration certificate.

m	k	method	instances	max ratio	max excess	diagnostic	flags
1	2–8	quadrature	5,000	0.999984	-10^{-8}	$6.4{\cdot}10^{-6}$	0
1	2–64	quadrature	5,000	0.998961	$-6.7{\cdot}10^{-7}$	$3.3{\cdot}10^{-5}$	0
2	2–16	quadrature	300	0.999732	$-1.7{\cdot}10^{-7}$	$4.3{\cdot}10^{-3}$	0
3	2–8	quadrature	256	0.997291	$-1.8{\cdot}10^{-6}$	$3.7{\cdot}10^{-4}$	0
4	2–64	sampling	256	0.969446	$-2.0{\cdot}10^{-5}$	$9.7{\cdot}10^{-3}$	0
6	2–32	sampling	256	1.008543	$8.3{\cdot}10^{-4}$	$1.4{\cdot}10^{-2}$	0
8	2–128	sampling	256	0.807502	$-1.2{\cdot}10^{-4}$	$1.5{\cdot}10^{-2}$	0

Table 3: Deterministic planning with $S = 256$ states, $A = 8$ actions, $\sigma = 1$, $\tau = 0.2$, $n = 200$ observations per reward coordinate and 4096 replications per horizon. The certificate of (32) does not depend on H; the log-Sobolev comparator uses the same trajectory feature vectors and therefore the same trajectory diameter. Realized values are measurements on one MDP per horizon.

H	$\mathbb{E}\mathcal{D}_H$	95% quantile	Gamma	log-Sobolev	ratio
1	$1.4{\cdot}10^{-2}$	$4.8{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
2	$2.0{\cdot}10^{-3}$	$5.7{\cdot}10^{-3}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
4	$2.2{\cdot}10^{-2}$	$6.0{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
8	$1.4{\cdot}10^{-2}$	$5.0{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
16	$1.7{\cdot}10^{-2}$	$4.2{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
32	$1.3{\cdot}10^{-2}$	$4.4{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
64	$2.2{\cdot}10^{-2}$	$4.4{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46
128	$2.9{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$3.2{\cdot}10^{-1}$	$4.7{\cdot}10^{-1}$	1.46

Table 4 compares the simplex expansions (45) and (46) with the measured coefficients at the smallest noise scale, including the case $k = 8$ where the displayed variance correction vanishes. The quadratic limit is evaluated on the same Gaussian draws as the remainder, which is what makes the difference of two variances estimable at this precision.

Table 4: Simplex coefficients at $u = 0.025$, with 95% half-widths from 2^{22} antithetic pairs per point for $k \le 16$ and 2^{20} for $k = 64$.

k	mean, theory	mean, measured	variance, theory	variance, measured
8	-0.006836	$-0.00682 \pm 1.4{\cdot}10^{-5}$	0	$1.239{\cdot}10^{-5} \pm 1.1{\cdot}10^{-5}$
9	-0.006173	$-0.006172 \pm 1.2{\cdot}10^{-5}$	$2.286{\cdot}10^{-4}$	$2.260{\cdot}10^{-4} \pm 1.0{\cdot}10^{-5}$
16	-0.003662	$-0.00366 \pm 1.0{\cdot}10^{-5}$	$6.104{\cdot}10^{-4}$	$6.093{\cdot}10^{-4} \pm 5.6{\cdot}10^{-6}$
64	$-9.613{\cdot}10^{-4}$	$-9.629{\cdot}10^{-4} \pm 1.2{\cdot}10^{-5}$	$2.804{\cdot}10^{-4}$	$2.796{\cdot}10^{-4} \pm 2.0{\cdot}10^{-6}$

Table 5 gives the sample budget across temperatures and targets at $\delta = 0.05$; the relative saving against the log-Sobolev bound does not depend on the target, because both certificates are affine in c and c is proportional to $1/n$.

Table 5: Observations per reward coordinate sufficient for $D(p\|\widehat{p}) \le \epsilon$ with probability 0.95, for independent reward means with $\sigma = 1$.

τ	ϵ	Gamma	log-Sobolev	Götze–Sambale	Lipschitz	saving
0.05	0.01	102,120	149,148	1,131,908	47,971,709	32 %
0.05	0.05	20,424	29,830	226,382	1,925,261	32 %
0.05	0.1	10,212	14,915	113,191	483,309	32 %
0.1	0.01	25,530	37,287	282,977	11,992,928	32 %
0.1	0.05	5,106	7,458	56,596	481,316	32 %
0.1	0.1	2,553	3,729	28,298	120,828	32 %
0.2	0.01	6,383	9,322	70,745	2,998,232	32 %
0.2	0.05	1,277	1,865	14,149	120,329	32 %
0.2	0.1	639	933	7,075	30,207	32 %
0.5	0.01	1,022	1,492	11,320	479,718	32 %
0.5	0.05	205	299	2,264	19,253	31 %
0.5	0.1	103	150	1,132	4,834	31 %
1.0	0.01	256	373	2,830	119,930	31 %
1.0	0.05	52	75	566	4,814	31 %
1.0	0.1	26	38	283	1,209	32 %

Table 6 prices the product certificate of Theorem 5.3 on 2^m actions driven by one shared direction, where the diameter scale is $c = m^2\|v\|^2/\tau$ and the product scale is $C = m\|v\|^2/\tau$.

Table 6: Shared-direction product model at $\tau = 1$, $\|v\| = 1$, $\delta = 0.05$ and 2^{21} draws. The last column is the fraction of draws exceeding the product certificate.

m	actions	c	C	95% quantile	diameter cert.	product cert.	ratio	exceedance
2	2^2	4	2	2.57	20.4	9.2	2.2	$0/2{,}097{,}152$
4	2^4	16	4	5.15	81.7	18.4	4.4	$0/2{,}097{,}152$
8	2^8	64	8	10.28	326.8	36.8	8.9	$0/2{,}097{,}152$
16	2^{16}	256	16	20.59	1307.1	73.7	17.7	$4.8{\cdot}10^{-7}$
32	2^{32}	1024	32	41.15	5228.5	147.4	35.5	$0/2{,}097{,}152$

Table 7 checks the two-path limit (54): the variance of the trajectory remainder scaled by n^2 approaches $H^2\sigma^4/(8\tau^2)$ and the mean per-decision change scaled by n approaches $\sigma^2/(4\tau^2)$.

The certificate battery contains uniform priors at $k = 2, 8, 64, 256$; for each $k = 64, 256$, two priors $\mathrm{softmax}(r/t)$ from one fixed standard-normal reward vector with $t = 0.2, 1$; a separate balanced-pair run; and a 64-action prior with mass 0.98 on one action and $0.02/63$ on each other action. Every case uses $n = 1277$, noise temperature $\tau = 0.2$, and 400,000 draws. Across all ten cases, using each

Table 7: Two disjoint paths of length H with $\sigma = 1$, $\tau = 0.2$ and 2^{22} draws. The comparison columns are limits as $n \to \infty$ at fixed H and τ.

H	n	$n^2 \operatorname{Var}(B_{\mathrm{path}})$	limit	$n\,\mathbb{E}\mathcal{D}_H$	limit
1	50	2.154	3.125	5.648	6.25
1	200	2.792	3.125	6.071	6.25
1	800	3.026	3.125	6.195	6.25
4	50	19	50	4.688	6.25
4	200	34.43	50	5.646	6.25
4	800	44.68	50	6.073	6.25
16	50	114.8	800	3.3	6.25
16	200	304.1	800	4.684	6.25
16	800	550.5	800	5.643	6.25
64	50	540.2	12800	2.015	6.25
64	200	1838	12800	3.304	6.25
64	800	4856	12800	4.681	6.25
256	50	2285	$2.048{\cdot}10^5$	1.12	6.25
256	200	8646	$2.048{\cdot}10^5$	2.014	6.25
256	800	29400	$2.048{\cdot}10^5$	3.302	6.25

Table 8: Cost of each certificate on the primary configuration: independent reward means with $\sigma = 1$, $k = 64$ actions, $\tau = 0.2$, target $\epsilon = 0.05$ and $\delta = 0.05$. Columns are the observations per reward coordinate, the total across actions, the ratio to Theorem 3.1, the certified radius at the common budget $n = 1277$, and the largest observed exceedance frequency across the four $k = 64$ policies at that common budget. A zero is displayed as $0/N$ with $N = 400{,}000$ draws per instance; it is an observed count. Gaussian-range radii use numerical inversion of the exact one-dimensional range CDF. The mixed-action-count battery is reported separately below; its rank-dependent thresholds differ from this column.

certificate	n	total kn	ratio	radius	exceedance
Theorem 3.1	1,277	81,728	1.0	0.0500	$0/400{,}000$
log-Sobolev	1,865	119,360	1.5	0.0730	$0/400{,}000$
χ^2, rank 63	10,317	660,288	8.1	0.4039	$0/400{,}000$
Gaussian range	2,107	134,848	1.6	0.0825	$0/400{,}000$
Götze–Sambale	14,149	905,536	11.1	0.5540	$0/400{,}000$
Gaussian Lipschitz	120,329	7,701,056	94.2	0.4941	$0/400{,}000$

case's own rank and action-count threshold, the maximum observed rates are 0.001315 for Gamma, 0.0001175 for log-Sobolev, and 0.050075 for both the chi-square and range certificates. The latter maxima occur at $k = 2$ with radius 0.0188012, not at the 64-action radii in Table 8. Götze–Sambale and Lipschitz have zero observed exceedances throughout.

D.8 What the perturbation does

Figure 3 draws the statistic and its curvature for three equally weighted alternatives whose features are the corners of a triangle, so the Gaussian coordinate is two-dimensional and can be shown directly. Near the origin the remainder is the quadratic bowl of the zero-noise limit; away from it the surface flattens into a cone. The curvature, which is the covariance of the perturbed policy, is concentrated on the thin set where alternatives are still competing and decays to $2.5 \cdot 10^{-5}$ in the corners of the plotted square, while never exceeding the value one that the covariance-diameter inequality (5) allows. That decay is what Gaussian smoothing converts into the contraction of (16), and it is why a pointwise Hessian bound alone loses the constant.

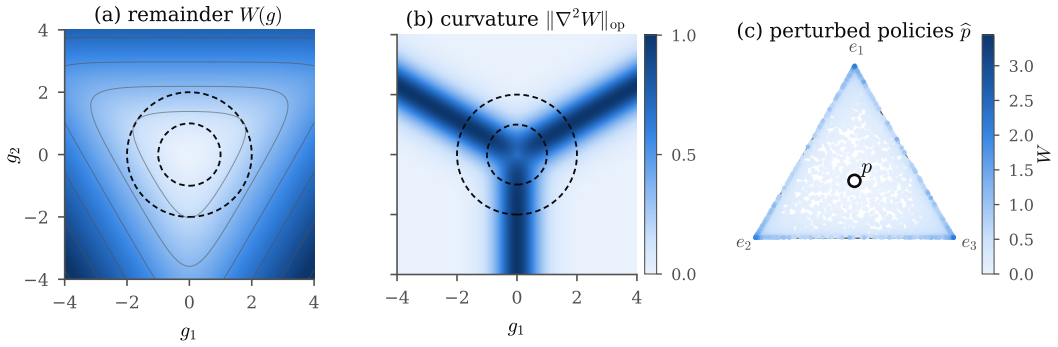

Figure 3: Three equally weighted alternatives at diameter $\Delta = 3$, with $\tau = 1$. (a) The normalised remainder $W(g) = 4B(g)/\Delta^2$ over the Gaussian coordinate; dashed circles are one and two standard deviations of the law that generates g. (b) The operator norm of its Hessian, equal to the covariance of the perturbed policy in these units and bounded by one. (c) Four thousand perturbed policies on the simplex, coloured by the remainder they produce, with the population policy p marked. Panels (a) and (c) share a colour scale.

Figure 4 maps how far the centred log-MGF falls short of equality in the Gamma MGF bound as a function of noise magnitude and the MGF parameter s. Along the extremal family the shortfall decreases as the separation decreases; taking the best of 4096 random instances per noise level reaches the same near-equality boundary without being told where to look.

D.9 Ablations

A random layered MDP draws each successor index independently and uniformly from its S next-layer states, then fixes that deterministic map; population rewards are independent $N(0, 1)$ draws and the initial state is zero. The eight-MDP horizon experiment uses 2,048 independent Gaussian reward perturbations per MDP. The supplementary size and temperature sweeps use one fixed MDP per setting and 1,024 perturbations, so their quantiles describe those fixed instances. The size sweep shares its generator at a given shape; the temperature sweep holds one MDP fixed. Self-loop controls set all rewards to zero and keep the state fixed under every action. Tables 9 and 10 report which instances approach the Gamma MGF upper bound, Table 11 what each quadrature design contributes, and Tables 12 and 13 the planning ablations.

D.10 Reward reuse on two forced paths

The population chooses between two paths with probability $1/2$ and all later actions are forced. Each reward has noise variance $1/n$, with $n = 2^{24}$ and $\tau = 0.2$. Let Z be standard Gaussian. The exact

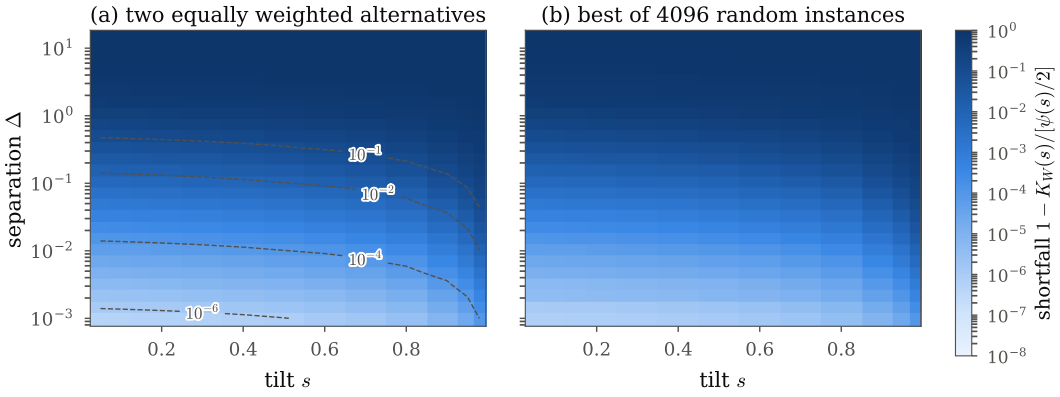

Figure 4: Shortfall $1 - K_W(s)/[\psi(s)/2]$ on a logarithmic colour scale, so lighter is tighter. (a) Two equally weighted alternatives, by numerical quadrature, with contours at 10^{-1}, 10^{-2}, 10^{-4} and 10^{-6}. (b) The largest ratio attained by any of 4096 random instances drawn independently at each actual normalized diameter; the banding across rows reflects that resampling.

Table 9: Tightness ratio by feature configuration and noise magnitude, at $k = 8$ with a Dirichlet(1) prior: maximum over 2048 instances and four tilts, with the median in parentheses. Columns are the actual supported normalized diameter Δ_u; temperature is one. All eight alternatives have positive mass, and a feature-configuration/prior draw is shared across the four diameter columns.

configuration	$\Delta_u = 0.01$	$\Delta_u = 0.1$	$\Delta_u = 1$	$\Delta_u = 10$
two clusters	0.958902 (0.41)	0.953867 (0.40)	0.651228 (0.27)	0.057643 (0.02)
two points, rest tied	0.538291 (0.02)	0.536627 (0.02)	0.417823 (0.02)	0.034611 (0.01)
Gaussian cloud	0.769196 (0.06)	0.765705 (0.06)	0.543423 (0.06)	0.046733 (0.01)
single outlier	0.949229 (0.05)	0.944483 (0.05)	0.650514 (0.10)	0.057079 (0.02)
collinear grid	0.686253 (0.07)	0.683434 (0.07)	0.497938 (0.06)	0.041995 (0.01)

Table 10: Tightness against the Dirichlet concentration that generates the prior, at actual normalized diameter $\Delta_u = 0.05$ with the same two-cluster feature draws and eight positive-mass alternatives.

α	max ratio	median ratio	$\max \mathbb{E}W$
0.02	0.997922	0.0000	0.4997
0.1	0.992622	0.0261	0.4984
0.5	0.976235	0.3143	0.4944
2	0.958690	0.4564	0.4901
10	0.951008	0.5014	0.4882
50	0.949814	0.5043	0.4879

Table 11: Quadrature design ablation on 4096 instances at four MGF parameters, all at 512 and 1024 nodes, against the plain design at 4096 and 8192 nodes whose own node-doubling discrepancy is $7.0 \cdot 10^{-10}$.

variant	max $\lvert K - K_{\mathrm{ref}} \rvert$	median	fraction above 10^{-6}
plain design alone	$2.4 \cdot 10^{-8}$	$7.6 \cdot 10^{-16}$	0.0000
tilted design alone	$1.5 \cdot 10^{-3}$	$9.5 \cdot 10^{-14}$	0.0802
contracted design alone	$1.2 \cdot 10^{-4}$	$2.8 \cdot 10^{-15}$	0.0283
wide design alone	$2.2 \cdot 10^{-5}$	$5.6 \cdot 10^{-14}$	0.0766
select by discrepancy and median	$1.7 \cdot 10^{-4}$	$3.0 \cdot 10^{-15}$	0.0106
plain estimator, others as cross-check	$2.4 \cdot 10^{-8}$	$7.6 \cdot 10^{-16}$	0.0000

Table 12: Reward representation, on self-looping MDPs with $S = 8$, $A = 4$, $\sigma = 1$, $\tau = 0.2$ and $n = 400$. Time-indexed coordinates leave two distinct entries per step; a reused stationary coordinate adds its coefficients first. The mean per-decision changes agree in this special self-loop model. The two forced paths in Table 15 test a case where reuse also changes the realized horizon dependence.

H	$\mathbb{E}\mathcal{D}_H$ time-indexed	$\mathbb{E}\mathcal{D}_H$ stationary	Gamma, time	Gamma, shared	ratio
1	$2.4{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$1.6{\cdot}10^{-1}$	1
2	$2.4{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$3.2{\cdot}10^{-1}$	2
4	$2.4{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$6.4{\cdot}10^{-1}$	4
8	$2.4{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$1.3{\cdot}10^{0}$	8
16	$2.3{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$2.6{\cdot}10^{0}$	16
32	$2.3{\cdot}10^{-2}$	$2.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	$5.1{\cdot}10^{0}$	32

Table 13: Problem size and temperature, on random layered MDPs at $H = 32$. The Gamma and log-Sobolev certificates use the same trajectory-diameter upper bound, independent of the state and action counts. The ratio is log-Sobolev/Gamma in the first block and Gamma/empirical quantile in the temperature block at fixed sampling.

S	A	n	$\mathbb{E}\mathcal{D}_H$	95% quantile	Gamma	LSI/Gamma
64	4	100	$5.3{\cdot}10^{-2}$	$1.1{\cdot}10^{-1}$	$6.4{\cdot}10^{-1}$	1.46
64	4	400	$1.5{\cdot}10^{-2}$	$3.4{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	1.46
64	4	1600	$4.1{\cdot}10^{-3}$	10^{-2}	$4.0{\cdot}10^{-2}$	1.46
256	8	100	$6.5{\cdot}10^{-2}$	$1.1{\cdot}10^{-1}$	$6.4{\cdot}10^{-1}$	1.46
256	8	400	$1.6{\cdot}10^{-2}$	$2.8{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	1.46
256	8	1600	$4.1{\cdot}10^{-3}$	$7.1{\cdot}10^{-3}$	$4.0{\cdot}10^{-2}$	1.46
1024	16	100	$4.0{\cdot}10^{-2}$	$7.7{\cdot}10^{-2}$	$6.4{\cdot}10^{-1}$	1.46
1024	16	400	$1.1{\cdot}10^{-2}$	$2.3{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	1.46
1024	16	1600	$2.8{\cdot}10^{-3}$	$6.1{\cdot}10^{-3}$	$4.0{\cdot}10^{-2}$	1.46
Temperature sweep						Gamma/quantile
$\tau = 0.05$, $S = 256$, $A = 8$, $n = 400$			$1.4{\cdot}10^{-3}$	$3.5{\cdot}10^{-3}$	$2.6{\cdot}10^{0}$	737.8
$\tau = 0.1$, $S = 256$, $A = 8$, $n = 400$			$2.4{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$	$6.4{\cdot}10^{-1}$	57.7
$\tau = 0.2$, $S = 256$, $A = 8$, $n = 400$			$5.0{\cdot}10^{-3}$	$1.3{\cdot}10^{-2}$	$1.6{\cdot}10^{-1}$	12.0
$\tau = 0.5$, $S = 256$, $A = 8$, $n = 400$			$4.3{\cdot}10^{-3}$	$5.5{\cdot}10^{-3}$	$2.6{\cdot}10^{-2}$	4.6
$\tau = 1$, $S = 256$, $A = 8$, $n = 400$			$1.2{\cdot}10^{-3}$	$1.3{\cdot}10^{-3}$	$6.4{\cdot}10^{-3}$	4.9

Table 14: Planning replication at $\tau = 0.2$. Entries are means and 95% Student half-widths over eight MDPs. The quantile column averages the within-MDP 95% quantiles. Both certificates use $\delta = 0.05$.

H	mean $\mathcal{D}_H$	mean 95% quantile	Gamma	log-Sobolev
1	0.0119 ± 0.0115	0.0378 ± 0.0315	0.3191	0.4661
8	0.0196 ± 0.0113	0.0524 ± 0.0279	0.3191	0.4661
32	0.0178 ± 0.0092	0.0385 ± 0.0148	0.3191	0.4661
128	0.0193 ± 0.0032	0.0321 ± 0.0041	0.3191	0.4661

per-decision KL is

$$\mathcal{D}_H = H^{-1} \log\cosh\left(\frac{b_H Z}{\tau\sqrt{2n}}\right), \qquad b_H = \sqrt{H} \text{ (time-indexed)}, \quad b_H = H \text{ (stationary)}.$$

Both designs use the same 2^{20} Gaussian draws. The displayed half-widths are 95% Student intervals over 64 independent batches.

Table 15: Actual horizon dependence under reward reuse. Values are $n\mathbb{E}\mathcal{D}_H$; the small-noise limits are 6.25 and $6.25H$.

H	time-indexed	stationary	stationary/time-indexed
1	6.257 ± 0.016	6.257 ± 0.016	1.00
4	6.257 ± 0.016	25.027 ± 0.066	4.00
16	6.257 ± 0.016	100.097 ± 0.264	16.00
64	6.257 ± 0.016	399.817 ± 1.052	63.90

D.11 Language models

Contexts come from the public-domain text of *Pride and Prejudice*, starting after character 20,000. Each checkpoint tokenizes that same text and uses the first 24 consecutive 128-token windows; only the final position of each window is evaluated. Model-specific tokenization can produce different token windows from the same characters. Each context is fixed before drawing Gaussian representation noise. The scale uses the unembedding rows and the prescribed covariance. For the six-checkpoint sweep the covariance is I/n, with 24 contexts and 1024 draws per context. Quantiles pool these fixed contexts as a descriptive summary; confidence for a fixed context comes from the theorem. No probability interval is obtained by treating all context–draw pairs as independent when they share Gaussian draws. The coordinate scale is part of the noise model, so diameter alone is not a scale-invariant comparison of checkpoint quality. The unscaled diameter Δ_0^2 is computed by an exhaustive blocked row-pair scan of the unembedding matrix W: each row block is multiplied against all rows, and the maximum of $\|w_i\|^2 + \|w_j\|^2 - 2w_i^\top w_j$ is retained. Blocks have 512 rows in the six-checkpoint sweep and 1,024 in the GPT-2 covariance study; no vocabulary subsample is used for the diameter. Matrix operations use float32 with TF32 disabled, so this is exhaustive floating-point evaluation, not outward-rounded verification. The score-noise diameter is $\Delta^2 = \Delta_0^2/n$. For the covariance control, center the 24 fixed final representations, form their sample covariance and normalize its mean diagonal to one. Its positive-semidefinite square root, with negative numerical eigenvalues clipped to zero, transforms the rows to $W\Sigma_h^{1/2}$ before the same scan. This covariance is frozen before fresh errors are drawn. The GPT-2 temperature/covariance study uses 2,048 draws shared across contexts, temperatures and sample sizes; the six-checkpoint study uses 1,024 per context.

Table 16: Feature-diameter scale and certificate cost by checkpoint, at $\tau = 1$, $n = 10{,}000$ and $\delta = 0.05$, over 24 contexts. $H(p)$ is the mean entropy of the next-token distribution in nats; conservatism is the certified radius divided by the realised 95% quantile; the last column is the budget for $\epsilon = 0.01$.

model	params	V	d	Δ_0^2	$H(p)$	95% quantile	Theorem 3.1	cons.	n
GPT-2	124M	50,257	768	57.50	2.64	$4.8{\cdot}10^{-4}$	$7.3{\cdot}10^{-3}$	15	7,340
GPT-2 medium	354M	50,257	1024	44.21	2.59	$3.8{\cdot}10^{-4}$	$5.6{\cdot}10^{-3}$	15	5,644
GPT-2 large	774M	50,257	1280	28.84	1.97	$1.4{\cdot}10^{-4}$	$3.7{\cdot}10^{-3}$	27	3,682
GPT-2 xl	1,557M	50,257	1600	17.37	1.84	$1.2{\cdot}10^{-4}$	$2.2{\cdot}10^{-3}$	19	2,218
Pythia 410m	405M	50,304	1024	4.19	2.41	$3.6{\cdot}10^{-5}$	$5.4{\cdot}10^{-4}$	15	536
Qwen2.5 0.5B	494M	151,936	896	1.53	2.66	$1.2{\cdot}10^{-5}$	$1.9{\cdot}10^{-4}$	16	195

Table 17 reports the language-model measurements in full.

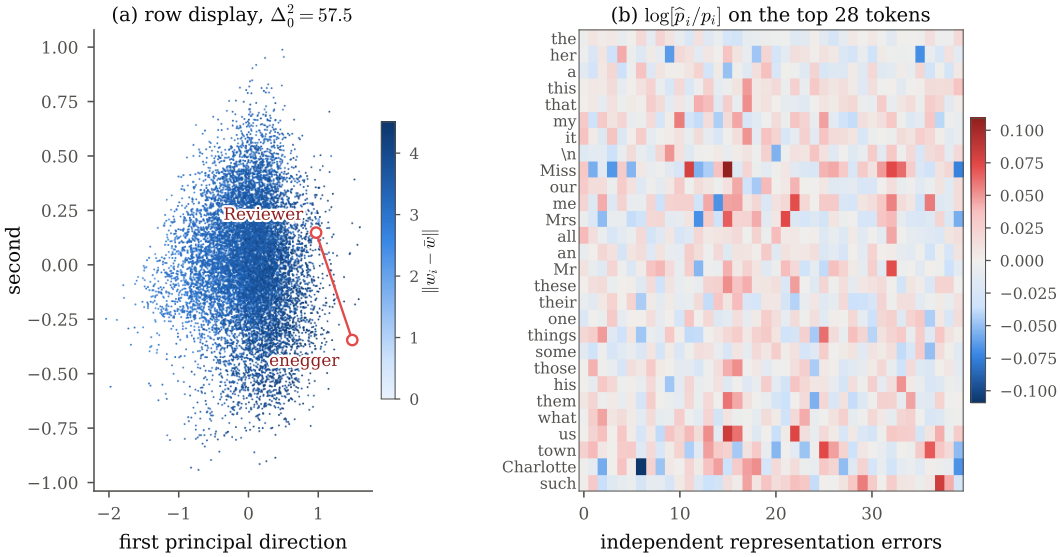

Figure 5: GPT-2 under representation error. (a) A display subset of 12,000 randomly selected unembedding rows plus the diameter pair, projected onto two principal directions computed from all 50,257 rows, and colored by distance from the mean row. The pair attaining the exhaustive floating-point maximum is joined in red. The diameter records the most separated token pair, irrespective of its probability under the fixed context. (b) For one context at $\tau = 1$ and $n = 10^4$, the log-ratio of the perturbed to the population probability of the top tokens across independent representation errors. Blue and red are the two directions of change about a neutral zero.

Table 17: GPT-2 next-token distribution under representation error. "iso" uses isotropic error, "cov" uses a covariance computed from the fixed representations before fresh Gaussian errors are drawn. Conservatism is the certified radius divided by the realised 95% quantile.

τ	n	geom.	mean KL	$\mathbb{E}W$	95% quantile	Gamma	log-Sobolev	ratio
0.5	10,000	iso	$2.9{\cdot}10^{-4}$	$5.0{\cdot}10^{-2}$	$1.3{\cdot}10^{-3}$	$2.9{\cdot}10^{-2}$	$4.3{\cdot}10^{-2}$	$2.2{\cdot}10^{1}$
1.0	10,000	iso	$1.8{\cdot}10^{-4}$	$1.3{\cdot}10^{-1}$	$4.8{\cdot}10^{-4}$	$7.3{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$	$1.5{\cdot}10^{1}$
2.0	10,000	iso	$1.2{\cdot}10^{-4}$	$3.2{\cdot}10^{-1}$	$1.5{\cdot}10^{-4}$	$1.8{\cdot}10^{-3}$	$2.7{\cdot}10^{-3}$	$1.2{\cdot}10^{1}$
0.5	100,000	iso	$2.9{\cdot}10^{-5}$	$5.0{\cdot}10^{-2}$	$1.3{\cdot}10^{-4}$	$2.9{\cdot}10^{-3}$	$4.3{\cdot}10^{-3}$	$2.2{\cdot}10^{1}$
1.0	100,000	iso	$1.8{\cdot}10^{-5}$	$1.3{\cdot}10^{-1}$	$4.8{\cdot}10^{-5}$	$7.3{\cdot}10^{-4}$	$1.1{\cdot}10^{-3}$	$1.5{\cdot}10^{1}$
2.0	100,000	iso	$1.2{\cdot}10^{-5}$	$3.2{\cdot}10^{-1}$	$1.5{\cdot}10^{-5}$	$1.8{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$	$1.2{\cdot}10^{1}$
0.5	10,000	cov	$5.4{\cdot}10^{-6}$	$6.1{\cdot}10^{-3}$	$2.5{\cdot}10^{-5}$	$4.6{\cdot}10^{-3}$	$6.6{\cdot}10^{-3}$	$1.8{\cdot}10^{2}$
1.0	10,000	cov	$2.8{\cdot}10^{-6}$	$1.2{\cdot}10^{-2}$	10^{-5}	$1.1{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$1.1{\cdot}10^{2}$
2.0	10,000	cov	$1.7{\cdot}10^{-6}$	$3.1{\cdot}10^{-2}$	$5.3{\cdot}10^{-6}$	$2.8{\cdot}10^{-4}$	$4.2{\cdot}10^{-4}$	$5.4{\cdot}10^{1}$
0.5	100,000	cov	$5.4{\cdot}10^{-7}$	$6.1{\cdot}10^{-3}$	$2.5{\cdot}10^{-6}$	$4.6{\cdot}10^{-4}$	$6.6{\cdot}10^{-4}$	$1.8{\cdot}10^{2}$
1.0	100,000	cov	$2.8{\cdot}10^{-7}$	$1.2{\cdot}10^{-2}$	10^{-6}	$1.1{\cdot}10^{-4}$	$1.7{\cdot}10^{-4}$	$1.1{\cdot}10^{2}$
2.0	100,000	cov	$1.7{\cdot}10^{-7}$	$3.1{\cdot}10^{-2}$	$5.3{\cdot}10^{-7}$	$2.8{\cdot}10^{-5}$	$4.2{\cdot}10^{-5}$	$5.4{\cdot}10^{1}$

D.12 COVARIANCE CALIBRATION FROM THE SAME DATA

We generate n independent Gaussian reward vectors for each repeated dataset and compute both the policy from their mean and S from their centered residuals. The simulation uses feature factors to generate the scores, but the certificate uses only empirical score-contrast variances. An observed-reward-only implementation returns the fitted policy and radius. A direct score-space implementation agrees with the compiled latent-space calculation to $2.2 \cdot 10^{-14}$ for the maximum variance and $4.5 \cdot 10^{-16}$ for the KL. The zero-diameter case and invariance to a large common feature shift are also checked. All calculations use CUDA float64 and full-graph PyTorch compilation.

The target is $\alpha = 0.05$, $\tau = 1$. Before drawing data, we select $\eta \in (0, \alpha)$ by numerical search on the scalar objective in Section B.4, and set $\delta = \alpha - \eta$. Bonferroni calibration uses a separate numerical search on its corresponding objective with $\eta/\binom{k}{2}$ in the lower chi-square quantile. Both use the same observations and target failure probability. The unadjusted diagnostic replaces the population variance maximum by $\widehat{v}$ without inflation and uses the full error budget α for the Gamma tail.

A check where covariance uncertainty matters. For two equal-prior actions, take opposite features at separation 0.05. We use $262{,}144$ datasets per $n \in \{3, 8, 32, 128\}$; repeating the two endpoints to obtain $k = 64$ uses $65{,}536$ datasets at $n = 3, 32$. Table 18 reports every configuration. The naive failure rate at $n = 3$ is above α, while the fixed-sample certificate accounts for covariance uncertainty. For two actions, Hoeffding's range inequality and the exact Student pivot also give the valid radius

$$R_t = \frac{\widehat{v}}{8n\tau^2} t^2_{n-1,1-\alpha/2}.$$

This structural comparator is included; the general covariance corollary is not asserted to improve it.

Table 18: Equal-pair covariance experiment. Rates are percentages; intervals are pointwise 95% Clopper–Pearson intervals for naive substitution. "Radius" is the median calibrated radius divided by the known-covariance radius. Rows with $k = 64$ duplicate the two feature locations.

k	n	naive rate [interval]	calibrated	Student	radius
2	3	8.51 [8.41, 8.62]	0.322	5.02	22.35
2	8	1.51 [1.47, 1.56]	0.020	4.92	4.21
2	32	0.32 [0.30, 0.35]	0.006	5.01	1.98
2	128	0.18 [0.16, 0.19]	0.020	4.94	1.43
64	3	8.52 [8.31, 8.74]	0.320	—	22.31
64	32	0.32 [0.28, 0.36]	0.003	—	1.98

An independent one-dimensional calculation integrates out the sample variance. With $\nu = n - 1$, $h = 1 + q_{1/2}(\log(1/\alpha))$ and $Z \sim N(0, 1)$,

$$\mathbb{P}\{D > R_{\text{naive}}\} = \mathbb{E}_Z F_{\chi^2_\nu}\left(\frac{4n\nu}{\Delta_0^2 h} \log\cosh\frac{\Delta_0 Z}{2\sqrt{n}}\right).$$

As $\Delta_0 \downarrow 0$, this tends to $2\mathbb{P}\{T_\nu > \sqrt{2h}\}$, where T_ν is Student-t. This calculation checks the observed failure without using the GPU implementation. The same scalar variance pivot makes the covariance coverage bound exact on the duplicated-endpoint family.

More actions and correlated rewards. The correlated cases have $(k, \rho) = (8, 4), (64, 8), (256, 8)$, where ρ is the centered feature rank; each uses a uniform and a fixed skewed prior. Features are Gaussian clouds normalized to population diameter one. The independent cases use $k = 256$, $\rho = 255$ and a uniform prior. Each setting is evaluated at $n = 32, 128, 512$ with 4096 datasets. Thus several settings have fewer observations than the feature rank, without requiring a covariance inverse. Priors in a matched pair share the same reward errors. All results are in Table 19.

Table 19: General covariance experiment. "Radius" is the median calibrated radius relative to known covariance; "Bonf." is the Bonferroni radius relative to the calibrated one on the same data. Counts show calibrated exceedances per 4096 datasets, separately for uniform and skewed priors where both are tested. Zero is an observed count.

configuration	k	ρ	n	radius	Bonf.	exceedances
correlated	8	4	32	2.04	1.44	0, 0 /4096
correlated	8	4	128	1.43	1.16	0, 0 /4096
correlated	8	4	512	1.21	1.07	0, 0 /4096
correlated	64	8	32	2.25	2.12	0, 0 /4096
correlated	64	8	128	1.44	1.35	0, 0 /4096
correlated	64	8	512	1.21	1.14	0, 0 /4096
correlated	256	8	32	2.26	2.66	0, 0 /4096
correlated	256	8	128	1.45	1.47	0, 0 /4096
correlated	256	8	512	1.21	1.19	0, 0 /4096
independent	256	255	32	4.76	2.66	0/4096
independent	256	255	128	2.27	1.47	0/4096
independent	256	255	512	1.54	1.19	0/4096