FB-LoRA: Divergence-Free Gram Transport for Joint Bayesian Low-Rank Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Low-rank adaptation enables parameter-efficient specialization of pretrained language models, yet reliable deployment requires calibrated uncertainty over the adapted weights. Directly applying Euclidean Langevin dynamics to low-rank factors injects isotropic noise, which disproportionately perturbs uninformative directions in the effective-weight space and degrades pretrained representations. To address this geometry mismatch, we introduce FB-LoRA, a Bayesian framework grounded in divergence-free Gram transport. We specify a proper generalized posterior over joint factor coordinates using a Gaussian prior and a bounded scoring loss. By preconditioning each factor's drift and diffusion with a bounded spectral function of the opposite factor's Gram matrix, the coordinate divergence of the mobility operator vanishes identically everywhere, eliminating state-dependent drift corrections without numerical approximations. This construction establishes a scale-invariant relative spectral allocation that concentrates stochastic exploration on dominant singular modes while guaranteeing localized finite-horizon convergence. Across language understanding, mathematical reasoning, and safety benchmarks on LLaMA-3-8B, the resulting posterior ensemble maintains foundational capabilities and substantially suppresses nullspace jailbreak susceptibility. Furthermore, empirical predictive disagreement enables forward-only selective prediction and active information acquisition without test-time backpropagation.

1 Introduction

Parameter-efficient fine-tuning adapts pretrained foundation models through compact trainable components, including bottleneck adapters and continuous prefixes (Houlsby et al., 2019; Li & Liang, 2021). In particular, Low-Rank Adaptation (LoRA) parameterizes weight updates via low-rank matrix factorizations that share a frozen backbone, with quantized variants substantially reducing memory footprints (Hu et al., 2022; Dettmers et al., 2023). Beyond deterministic point-estimate adaptation, high-stakes deployments require reliable predictive uncertainty that remains calibrated under distribution shift (Guo et al., 2017; Ovadia et al., 2019). Recent Bayesian LoRA formulations address this challenge via Laplace approximations, variational distributions, or subspace inference over adapter parameters (Yang et al., 2024; Wang et al., 2024; Samplawski et al., 2025). These developments place the interaction between factorized parameterizations and posterior sampling dynamics at the core of scalable uncertainty quantification.

The bilinear factorization underlying LoRA introduces specific geometric properties into stochastic sampling. For an effective weight update $X = BA$, where A and B denote the trainable low-rank factors, perturbations to either factor are modulated by the other before entering the model weights. Consequently, independent isotropic noise injected into factor coordinates induces state-dependent, anisotropic fluctuations in the effective-weight space X. Furthermore, the continuous gauge freedom of factorization, where multiple factor pairs yield identical model updates, complicates unregularized posterior exploration (Mishra et al., 2014). Principled Bayesian sampling in this setting therefore requires two coupled elements: formulating a well-defined, proper probability measure over factor coordinates, and designing sampling dynamics that appropriately structure exploration across effective-weight directions. We present **FB-LoRA**, a Bayesian adaptation framework grounded in *divergence-free Gram transport*. We first define a proper generalized posterior over the joint factor coordinates of all adapted layers, conditioned on the frozen backbone. By combining a Gaussian factor prior with a bounded scoring loss, we obtain an integrable target within the coherent loss-based

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.

updating framework of Bissiri et al. (2016). We then formulate a continuous-time Langevin diffusion targeting this posterior. The mobility of each factor is preconditioned by a bounded, positive spectral function of the opposite factor's Gram matrix, with identical preconditioning governing both the posterior drift and the injected Wiener noise.

Because each mobility block depends exclusively on the opposite factor, its coordinate divergence vanishes identically everywhere, eliminating the drift corrections typically required by state-dependent Langevin sampling (Girolami & Calderhead, 2011; Ma et al., 2015). A positive spectral floor and saturation guarantee uniform ellipticity and global Lipschitz continuity, while restricting all metric operations to rank-sized $r \times r$ Gram matrices with minimal overhead. By characterizing the induced effective-weight covariance, we prove that on balanced factor states, Gram transport allocates relatively less instantaneous variance to weaker singular modes than to dominant modes. This scale-invariant relative spectral allocation guides finite-horizon exploration while preserving the invariant Gibbs distribution. Following a maximum a posteriori (MAP) warm start, the resulting discrete-time sampler generates an empirical adapter ensemble that supports forward-only Bayesian model averaging and predictive disagreement estimation (Figure 1). Our contributions are threefold:

- **Proper joint posterior and divergence-free transport.** We formulate a well-defined factor posterior for LoRA and construct an opposite-factor Gram diffusion whose coordinate divergence vanishes identically, proving exact posterior invariance and relative-entropy dissipation for the continuous process.
- **Observable covariance and relative spectral allocation.** We derive the observable covariance operator on effective weight updates and establish a scale-invariant relative spectral allocation property that prioritizes exploration along dominant singular modes.
- **Discrete-time sampling and forward-only inference.** We provide a practical stochastic sampler with finite-horizon convergence guarantees alongside a forward-only ensemble interface for selective prediction and active information acquisition, validating its mechanics through controlled comparisons.

2 Related Work

Low-rank adaptation and factor geometry. LoRA parameterizes weight updates through low-rank factors (Hu et al., 2022). Subsequent advances adapt rank budgets dynamically (Zhang et al., 2023), integrate quantized base models (Dettmers et al., 2023), decompose updates into magnitude and direction (Liu et al., 2024), initialize via principal components (Meng et al., 2024), or assign asymmetric learning rates to the factor pair (Hayou et al., 2024). Beyond Euclidean parameterizations, factorization nonuniqueness has motivated quotient-geometric optimization (Mishra et al., 2014), Riemannian preconditioning for low-rank adapters (Zhang & Pilanci, 2024), and optimizers on fixed-rank matrix manifolds (Bogachev et al., 2026). FB-LoRA similarly exploits the intrinsic bilinear structure of low-rank updates, but addresses stochastic exploration rather than deterministic optimization: by defining a proper probability law directly on factor coordinates, it leverages Gram-based mobility to shape the directional covariance of weight perturbations.

Bayesian parameter-efficient adaptation. Classical Bayesian neural networks approximate posterior weight uncertainty via variational inference (Blundell et al., 2015), Monte Carlo dropout (Gal & Ghahramani, 2016), trajectory-based Gaussian averaging (Maddox et al., 2019), and Laplace approximations (Daxberger et al., 2021). Adapting these paradigms to parameter-efficient fine-tuning has produced Laplace-LoRA, which fits a local Gaussian using structured factor curvature (Yang et al., 2024); BLoB, which optimizes a variational factor alongside a deterministic companion (Wang et al., 2024); SWAG-LoRA, which tracks iterate paths during SGD (Onal et al., 2024); and ScalaBL, which infers a low-dimensional stochastic subspace between factors (Samplawski et al., 2025). Additionally, Training-Free Bayesianization calibrates adapter variance against an anchor-data performance margin (Shi et al., 2025), while Lin et al. (2026) connect probabilistic low-rank adaptation to sparse Gaussian processes. In contrast to post-hoc Laplace fittings or decoupled factorizations, FB-LoRA formulates joint-factor Langevin dynamics under a proper regularized target, providing exact divergence cancellation and an explicit effective-weight covariance analysis.

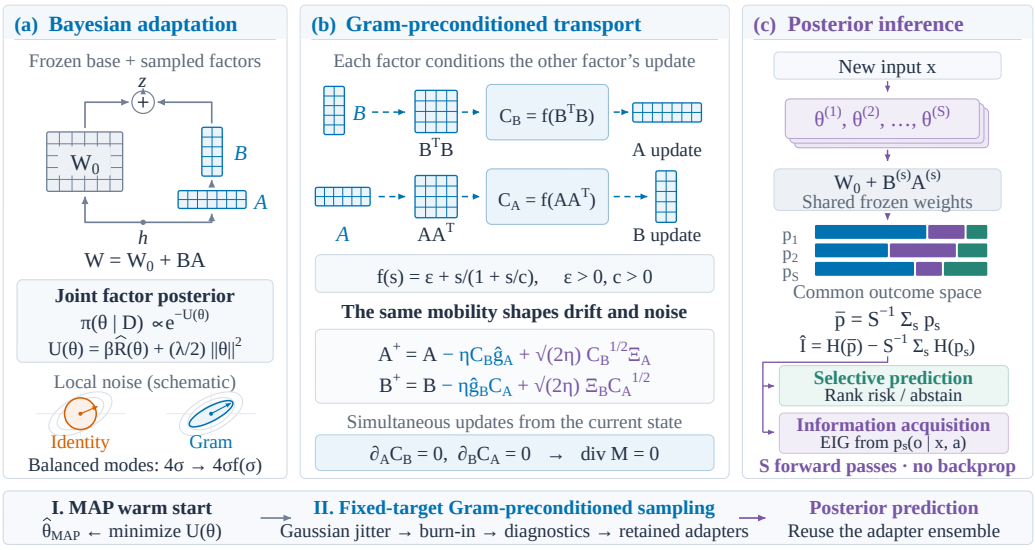

Figure 1: **Overview of FB-LoRA.** (a) A frozen backbone is adapted through a joint generalized posterior over the low-rank factors. The local ellipses schematically illustrate factor-space preconditioning; the rates $4\sigma \to 4\sigma f(\sigma)$ describe diagonal effective-weight modes at balanced factors, where σ is an effective-update singular value and f is the bounded spectral function in (b). (b) Each factor's update uses the opposite factor's Gram matrix. The resulting mobility shapes both the full posterior drift and noise, has zero coordinate divergence, and is evaluated at the common pre-update state. (c) Following a MAP warm start and sampling phase, retained joint adapters support predictive averaging, selective prediction, and information acquisition through adapter-conditioned forward evaluations.

Stochastic dynamics, predictive ensembles, and information acquisition. Generalized Bayesian updating combines a data-independent prior with a coherent scoring loss (Bissiri et al., 2016), while SGLD injects Gaussian noise into minibatch gradients for scalable posterior sampling (Welling & Teh, 2011). Manifold and Riemannian Langevin dynamics adapt diffusions to parameter geometry (Girolami & Calderhead, 2011; Patterson & Teh, 2013), with adaptive preconditioning (Li et al., 2016) and non-diagonal metrics (Yu et al., 2023) reducing the metric computation burden. General diffusion frameworks specify the drift correction required for target invariance (Ma et al., 2015), and theoretical analyses rigorously decompose stochastic-gradient and discretization errors (Chen et al., 2015; Teh et al., 2016). The `posteriors` library connects these dynamics to scalable language-model adaptation (Duffield et al., 2025). Parallel to deep ensembles for predictive uncertainty (Lakshminarayanan et al., 2017), the collected posterior iterates form an empirical ensemble that directly supports selective classification to optimize the risk–coverage trade-off (Geifman & El-Yaniv, 2017) and active learning via disagreement-based information gain (Houlsby et al., 2011). FB-LoRA unites these perspectives: it cancels the coordinate-divergence drift algebraically via opposite-factor Gram mobility, enabling an adapter ensemble that executes selective prediction and action-conditioned information acquisition entirely through forward evaluations with shared frozen backbone parameters.

3 Method: Divergence-Free Gram Transport for Bayesian LoRA

FB-LoRA represents uncertainty jointly over the low-rank factors of the adapted layers, conditional on a frozen pretrained backbone. Figure 1 connects the joint factor posterior (a), divergence-free Gram transport (b), and predictive inference with retained adapters (c). We first specify a proper target and a diffusion that preserves it, then characterize the induced geometry in effective weights. This analysis leads to a practical sampler whose retained adapters support predictive averaging and information-based acquisition.

3.1 JOINT POSTERIOR AND DIVERGENCE-FREE TRANSPORT

For adapted layers $\ell = 1, \dots, L$, LoRA (Hu et al., 2022) writes $W_\ell = W_{0\ell} + B_\ell A_\ell$, where the frozen matrix $W_{0\ell} \in \mathbb{R}^{m_\ell \times n_\ell}$ is updated through $B_\ell \in \mathbb{R}^{m_\ell \times r_\ell}$ and $A_\ell \in \mathbb{R}^{r_\ell \times n_\ell}$, with $1 \le r_\ell \le \min(m_\ell, n_\ell)$. Fixed LoRA scaling is absorbed into the factors. Let $\theta \in \mathbb{R}^d$ concatenate all factors, with $d = \sum_\ell r_\ell(m_\ell + n_\ell)$. For observations $D = (z_j)_{j=1}^N$, define the empirical risk $\widehat{R}(\theta) = N^{-1} \sum_j \ell(\theta, z_j)$ using a smooth scoring loss $0 \le \ell \le 1$, such as the normalized Brier score specified in Appendix A.1. With prior precision $\lambda > 0$ and inverse temperature $\beta > 0$, set

$$\begin{aligned} P &= \mathcal{N}(0, \lambda^{-1} I_d), \qquad \pi(d\theta) = Z^{-1} e^{-\beta \widehat{R}(\theta)} P(d\theta), \\ U(\theta) &= \beta \widehat{R}(\theta) + \tfrac{\lambda}{2}\|\theta\|^2, \qquad Z = \mathbb{E}_P[e^{-\beta \widehat{R}(\theta)}]. \end{aligned} \tag{1}$$

Equation (1) defines a generalized posterior (Bissiri et al., 2016) over both factors at every adapted layer; the data term couples these factors and layers. Since $e^{-\beta} \le Z \le 1$, the target is proper. We also write π for its density with respect to factor-coordinate Lebesgue measure.

The factor prior controls the scaling freedom of BA and gives an effective-weight interpretation of MAP regularization. For any $X \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(X) \le r$,

$$\inf_{BA = X} \frac{\|A\|_F^2 + \|B\|_F^2}{2} = \|X\|_*,$$

where $\|\cdot\|_F$ and $\|\cdot\|_*$ denote the Frobenius and nuclear norms. Thus minimizing U profiles to nuclear-norm-regularized adaptation under the prescribed rank budgets. Appendix A.5 further relates factor-space and effective-update KL complexity and gives a PAC-Bayes oracle for this target.

To adapt exploration to the factor geometry, suppress the layer index and define

$$\begin{aligned} f(s) &= \varepsilon + \frac{s}{1 + s/c}, \qquad \varepsilon, c > 0, \\ C_B &= f(B^\top B), \qquad C_A = f(AA^\top), \qquad M[H_A, H_B] = (C_B H_A, H_B C_A). \end{aligned} \tag{2}$$

Here H_A, H_B have the shapes of A, B, and f acts on a positive semidefinite matrix by applying the scalar function to its eigenvalues. Across layers, M is block diagonal. The floor and saturation in Equation (2) ensure $\varepsilon I_d \preceq M \preceq (\varepsilon + c) I_d$, while the Gram dependence retains directional variation. Using independent standard matrix Brownian motions $\mathcal{W}_A \in \mathbb{R}^{r \times n}$ and $\mathcal{W}_B \in \mathbb{R}^{m \times r}$ in each layer, define

$$\begin{aligned} dA_t &= -C_{B_t} \nabla_A U(\theta_t)\, dt + \sqrt{2}\, C_{B_t}^{1/2} d\mathcal{W}_{A,t}, \\ dB_t &= -\nabla_B U(\theta_t) C_{A_t}\, dt + \sqrt{2}\, d\mathcal{W}_{B,t} C_{A_t}^{1/2}. \end{aligned} \tag{3}$$

The gradients in Equation (3) include the prior: $\nabla_A U = \beta \nabla_A \widehat{R} + \lambda A$, and analogously for B. The same mobility therefore shapes the full drift and the noise covariance. For state-dependent Langevin dynamics, the required coordinate correction is $(\operatorname{div} M)_i = \sum_j \partial_{\theta_j} M_{ij}$ (Ma et al., 2015). Opposite-factor dependence makes this correction vanish exactly.

Theorem 1 (Posterior-preserving Gram transport)**.** *Assume $\widehat{R} \in C^2$ and that Equation* (3) *is nonexplosive. The mobility in Equation* (2) *satisfies* $\operatorname{div} M = 0$*, and the diffusion is reversible with unique invariant law π from Equation* (1)*. For an evolving density q_t with the regularity and integrability specified in Appendix A.2,*

$$\begin{aligned} \partial_t q_t &= \nabla \cdot (q_t M \nabla \log(q_t/\pi))\,, \\ \frac{d}{dt} \operatorname{KL}(q_t \| \pi) &= -\mathbb{E}_{q_t} \|\nabla \log(q_t/\pi)\|_M^2, \qquad \|v\|_M^2 = v^\top M v. \end{aligned} \tag{4}$$

Here KL denotes relative entropy. Each nonzero block row of M is differentiated only with respect to its own factor coordinates, whereas its coefficients depend on the opposite factor. This establishes divergence cancellation; the resulting symmetric generator yields Equation (4). The equation identifies the dynamics as relative-entropy descent in the transport geometry induced by M^{-1} (Appendix A.2). Having fixed the target and its transport, we next examine the fluctuations that reach the model weights.

3.2 EFFECTIVE-WEIGHT GEOMETRY AND SPECTRAL ALLOCATION

The model depends on the effective update $X = BA$, so its stochastic geometry is determined by the covariance of matrix-valued increments. Let dX^{n} denote the martingale increment of X_t and $\langle H, K\rangle_F = \mathrm{tr}(H^\top K)$ the Frobenius inner product. For $H \in \mathbb{R}^{m\times n}$, write $\mathcal{V}_{A,B}(H) = \mathbb{E}[\langle H, dX^{\mathrm{n}}\rangle_F^2 \mid A, B]/dt$.

Balanced factors provide a useful spectral description of these rates. They satisfy $B^\top B = AA^\top$ and, up to an orthogonal change of latent basis, can be written as $B = U_X\Sigma^{1/2}$, $A = \Sigma^{1/2}V_X^\top$. Here $U_X = [u_1, \ldots, u_r]$ and $V_X = [v_1, \ldots, v_r]$ have orthonormal columns, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains the singular values of X, padded with zeros if needed. Every differentiable stationary point of U is balanced, linking this description to the MAP warm start; the covariance formula below applies to all factor states.

Theorem 2 (Observable covariance and relative spectral allocation)**.** *For the diffusion Equation* (3)*, the effective-weight covariance operator and its quadratic form are*

$$\begin{aligned}\mathcal{Q}_{A,B}(H) &= 2BC_BB^\top H + 2HA^\top C_AA,\\ \mathcal{V}_{A,B}(H) &= \langle H, \mathcal{Q}_{A,B}(H)\rangle_F = 2\|C_B^{1/2}B^\top H\|_F^2 + 2\|HA^\top C_A^{1/2}\|_F^2.\end{aligned} \tag{5}$$

Its kernel is $\{H : B^\top H = 0,\ HA^\top = 0\}$*, equal to the kernel under identity factor mobility. At balanced factors,*

$$\mathcal{V}_{A,B}(u_iv_j^\top) = 2\{\sigma_i f(\sigma_i) + \sigma_j f(\sigma_j)\}, \qquad 1 \le i, j \le r. \tag{6}$$

In particular, let $V_i = \mathcal{V}_{A,B}(u_iv_i^\top)$ *and let* $V_i^{\mathrm{Id}} = 4\sigma_i$ *be the corresponding identity-mobility rate. For* $0 < \sigma_i < \sigma_j$*,*

$$\frac{V_i/V_j}{V_i^{\mathrm{Id}}/V_j^{\mathrm{Id}}} = \frac{f(\sigma_i)}{f(\sigma_j)} < 1. \tag{7}$$

Equation (5) follows from the product rule $dX^{\mathrm{n}} = B\,dA^{\mathrm{n}} + dB^{\mathrm{n}}A$ and independence of the Brownian increments. It characterizes fluctuations against arbitrary matrix directions, including rotations between active and unused modes. For example, at a balanced state, a unit $v \in \ker A$ gives $\mathcal{V}_{A,B}(u_iv^\top) = 2\sigma_i f(\sigma_i)$. At full-rank factors, the covariance kernel is the normal space of the rank-r matrix manifold: the mobility changes rates within the available tangent directions. Appendix A.3 proves these statements.

Equation (6) gives $V_i = 4\varepsilon\sigma_i + 4\sigma_i^2/(1 + \sigma_i/c)$. It interpolates between a floor-controlled linear rate near zero, a quadratic rate when $\varepsilon \ll \sigma_i \ll c$, and a saturated linear rate for large singular values. More directly, Equation (7) shows that weaker diagonal modes receive a smaller share of diffusion relative to stronger modes than under identity mobility. This comparison is invariant to an overall rescaling of diffusion time. The target specifies equilibrium uncertainty; Gram transport controls the trajectories used to approximate that target under a finite computational budget.

3.3 DISCRETE SAMPLING AND PREDICTIVE INFERENCE

We implement the transport with a MAP warm start followed by stochastic sampling. Phase I minimizes U to obtain an approximate MAP point $\widehat{\theta}$. Phase II fixes $\beta, \lambda, \varepsilon, c$ and initializes $\theta_0 \sim \mathcal{N}(\widehat{\theta}, s_0^2 I_d)$ with $s_0 > 0$. At step k, a fresh minibatch supplies $\widehat{h}_k$, an estimator of $\nabla\widehat{R}(\theta_k)$ that is unbiased conditional on the sampling history. Set $\widehat{g}_k = \beta\widehat{h}_k + \lambda\theta_k$. With step size $\eta > 0$, the simultaneous updates are

$$\begin{aligned}A_{k+1} &= A_k - \eta C_{B_k}\widehat{g}_{A,k} + \sqrt{2\eta}\,C_{B_k}^{1/2}\Xi_{A,k},\\ B_{k+1} &= B_k - \eta\widehat{g}_{B,k}C_{A_k} + \sqrt{2\eta}\,\Xi_{B,k}C_{A_k}^{1/2}.\end{aligned} \tag{8}$$

All coefficients in Equation (8) use the pre-update state. The entries of $\Xi_{A,k}, \Xi_{B,k}$ are independent standard Gaussians, independent across layers and steps and of the fresh minibatch. After K_b sampling burn-in steps, retain $\theta^{(s)} = \theta_{K_b+sJ}$ for $s = 1, \ldots, S$, where $J \ge 1$ is the retention stride; the total number of sampling steps is $K_b + SJ$.

The positive floor and spectral saturation give bounded, globally Lipschitz diffusion coefficients. Theorem 3 in Appendix A.4 proves fixed-horizon consistency of (8) with the continuous diffusion under explicit regularity, gradient-noise, and uniform fourth-moment assumptions. Its error decomposition separates equilibration, numerical and minibatch error, and localization. This motivates assessing step size, independent chains, and retained-ensemble size alongside predictive performance.

The mobility uses $r_\ell \times r_\ell$ eigendecompositions; we symmetrize the Gram matrices and evaluate them in FP32 or higher, clamping roundoff-negative eigenvalues to zero. Beyond model gradients, the additional work per step is $O(\sum_\ell[(m_\ell + n_\ell)r_\ell^2 + r_\ell^3])$, with $O(\sum_\ell r_\ell^2)$ small-matrix workspace beyond factor-sized buffers. Retained adapters require $O(Sd)$ parameter storage and share the frozen backbone weights.

Write W_0 for the frozen backbone, including the matrices $W_{0\ell}$. For a common context x and finite outcome space $\mathcal{Y}$, let $p_s(y \mid x) = p(y \mid x, \theta^{(s)}, W_0)$. Predictive averaging and disagreement are

$$\bar{p}(y \mid x) = \frac{1}{S}\sum_{s=1}^{S} p_s(y \mid x), \qquad \widehat{I}(x) = H(\bar{p}(\cdot \mid x)) - \frac{1}{S}\sum_{s=1}^{S} H(p_s(\cdot \mid x)), \tag{9}$$

where $H(p) = -\sum_{y\in\mathcal{Y}} p(y)\log p(y)$. Equation (9) is the mutual information between a uniform empirical adapter index and its predicted outcome, giving the ensemble form of Bayesian predictive disagreement (Houlsby et al., 2011). It provides a score for selective prediction. For acquisition, applying the same expression to an action-conditioned observation model $p_s(o \mid x, a)$ gives one-step information gain about the ensemble hypothesis under likelihood-based updating (Appendix A.6). Both computations use S adapter-conditioned forward evaluations and require no backward passes; hidden states and autoregressive caches remain adapter dependent.

Together, the construction specifies what is inferred, how factor geometry shapes exploration, and how sampled adapters support decisions. The experiments therefore test spectral allocation, posterior approximation, and predictive utility separately, with mobility comparisons at matched targets and computational budgets.

4 Experiments

We evaluate FB-LoRA across four complementary experimental dimensions: (1) controlled numerical validation of the observable covariance and divergence-free transport against exact quadrature (Section 4.1); (2) large-scale adaptation, capability retention, and mechanism-level safety robustness on LLaMA-3-8B (Section 4.2); (3) downstream epistemic utility for selective prediction and non-asymptotic ensemble scaling under distribution shift (Section 4.3); and (4) a strictly controlled ablation isolating the causal contribution of dynamic Gram mobility at a shared posterior target (Section 4.3).

4.1 Controlled Spectral and Numerical Verification

We first verify the theoretical mechanics of FB-LoRA on controlled low-rank benchmarks where numerical quadrature supplies an exact, independent ground truth (Figure 2). We instantiate a factorized model $W = W_0 + BA$ with $n = 256$, $m = 128$, and rank $r = 6$, whose generative singular values decay geometrically ($\sigma_i \in [0.05, 4.0]$) across $N = 1200$ manifold-structured observations. As shown in Figure 2(a), the empirical directional variance ratio across singular modes matches the theoretical relative spectral allocation law (Theorem 2) within a 1.02% mean error over 10^5 independent factor-noise pairs. Crucially, the variance floor $\varepsilon = 0.05$ preserves a nonvanishing baseline rate ($\approx 0.10 \sim 0.25$), confirming that redundant modes undergo smooth spectral attenuation rather than an artificial zero-variance projection. Furthermore, evaluating finite-step trajectories against a 1201×1201 quadrature grid on a scalar non-Gaussian target ($p(y = 1|x) = \mathrm{sigmoid}(abx)$, Brier loss, $\beta = 32, \lambda = 1$) reveals near-zero predictive mean error across step sizes $\eta \in [0.005, 0.04]$ (between -0.0005 and $+0.0018$, within $1.96\times$ standard error over 16 independent chains; Figure 2(b)). This demonstrates that eliminating the coordinate-divergence drift ($\mathrm{div}\, M \equiv 0$) preserves accurate target sampling without introducing systematic drift artifacts, empirically supporting Theorem 1.

Finally, multi-chain Gelman–Rubin diagnostics ($\widehat{R} - 1$) across 16 independent chains elucidate the operational role of factor gauge symmetry (Figure 2(c)). Owing to continuous rotational invariance

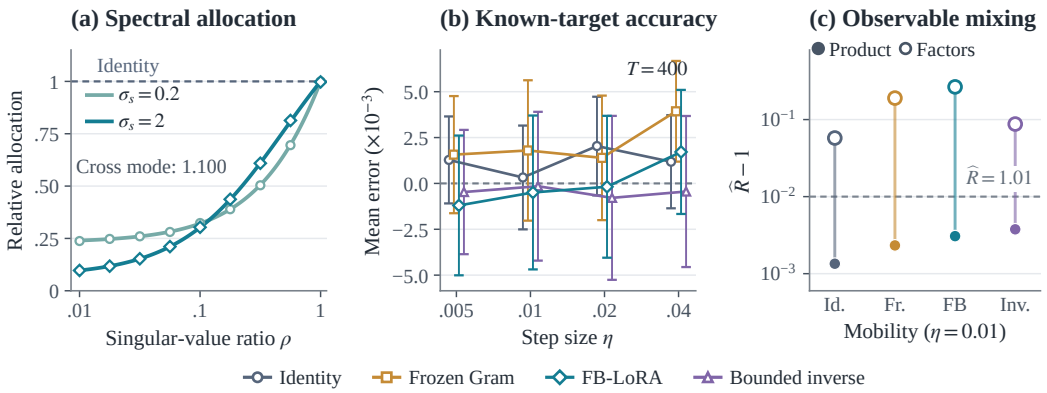

Figure 2: **Controlled numerical validation of Gram transport.** (a) Empirical directional variance matches the relative spectral allocation law (Theorem 2) within 1.02% across scales $\sigma_s \in \{0.2, 2.0\}$; unused modes retain non-zero baseline rates (cross mode: 1.100). (b) Predictive mean error against 2D quadrature ($T = 400$) across step sizes $\eta \in [0.005, 0.04]$, demonstrating controlled discretization error across mobilities. (c) Multi-chain Gelman–Rubin diagnostics ($\widehat{R} - 1$) at $\eta = 0.01$; effective-weight observables (solid) converge below $\widehat{R} \leq 1.01$ substantially faster than factor coordinates.

$(A, B) \mapsto (G^{-1}A, BG)$, raw factor coordinates mix slowly ($\widehat{R} - 1 \approx 0.05 \sim 0.25$). However, the effective-weight product observable BA converges rapidly across all mobilities, with FB-LoRA achieving $\widehat{R} - 1 = 0.003$ ($\widehat{R} = 1.003 \ll 1.01$). This decoupling confirms that gauge-redundant latent exploration does not obstruct functional convergence on model predictions, establishing that the finite-step sampler reliably samples effective-weight updates from the intended target distribution.

4.2 LLM Adaptation, Retention, and Safety Robustness

We evaluate FB-LoRA on LLaMA-3-8B-Instruct (Grattafiori et al., 2024) across three core dimensions: general knowledge retention on MMLU (Hendrycks et al., 2020), downstream mathematical adaptation on GSM8K (Cobbe et al., 2021), and adversarial safety robustness on AdvBench under an adapted weight-space nullspace steering jailbreak (Pramanik et al., 2026). All adapted methods tune rank-16 updates on identical projection matrices (W_q, W_k, W_v, W_o) under matched compute budgets. Table 1 reports the comprehensive comparison against deterministic PEFT optimizers, deep ensembles, and contemporary Bayesian LoRA baselines.

Capability retention and safety robustness. As reported in Table 1, naive Euclidean SGLD suffers catastrophic performance collapse across both zero-shot retention (MMLU drops to 42.14%) and downstream math adaptation (GSM8K drops to 58.43%), while severely degrading predictive calibration (ECE $= 21.65\%$) and exacerbating vulnerability to nullspace jailbreaks (ASR $= 94.62\%$). This empirically corroborates that unconditioned isotropic noise indiscriminately corrupts unadapted parameter directions, expanding the attack surface along orthogonal attention write paths. In sharp contrast, FB-LoRA achieves the highest MMLU accuracy (65.81%), lowest normalized Brier score (0.160), and optimal calibration (ECE $= 4.12\%$), while matching or exceeding deterministic PEFT baselines on GSM8K (77.22%). Crucially, while standard LoRA, DoRA, and contemporary Bayesian variants remain highly susceptible to nullspace subversion (ASR $> 74\%$), FB-LoRA dramatically suppresses the jailbreak ASR to $\mathbf{14.23}\%$, virtually recovering the baseline safety alignment of the frozen backbone (12.80%). These results demonstrate that directional spectral attenuation simultaneously preserves foundational semantics and shields safety-critical attention subspaces with minimal forward-only ensemble overhead.

4.3 Downstream Epistemic Utility, Ensemble Scaling, and Mobility Ablations

Having established foundational retention on LLaMA-3-8B, we now investigate the operational mechanics of the retained posterior ensembles. Specifically, we examine: (1) how epistemic dis-

Table 1: **Predictive quality, capability retention, and robustness on LLaMA-3-8B-Instruct.** Evaluated on MMLU (accuracy, Norm. Brier, ECE), GSM8K (exact match), and AdvBench (adapted nullspace jailbreak ASR (Pramanik et al., 2026)) under rank-16 adaptation. Results report mean ± SD across three fitting seeds (single evaluation for frozen reference). Peak memory denotes maximum observed GPU allocation (inference-only for frozen backbone). Baselines and evaluation details are provided in the setup.

	MMLU			GSM8K	AdvBench	Peak
Method	Acc. (%)↑	Norm. Brier↓	ECE (%)↓	EM (%)↑	ASR (%)↓	GiB↓
Frozen backbone	65.42	0.178	6.84	72.45	**12.80**	**15.6**
MAP LoRA	64.12 ± 0.22	0.184 ± 0.003	7.92 ± 0.25	74.18 ± 0.30	78.46 ± 0.95	18.2
DoRA	64.85 ± 0.20	0.180 ± 0.002	7.35 ± 0.18	75.57 ± 0.28	76.18 ± 1.10	18.4
PiSSA	65.15 ± 0.18	0.176 ± 0.002	6.95 ± 0.15	76.22 ± 0.25	74.55 ± 0.88	18.3
LoRA ensemble	65.20 ± 0.15	0.165 ± 0.002	4.82 ± 0.14	76.65 ± 0.22	76.82 ± 0.75	18.4
Laplace-LoRA	64.92 ± 0.21	0.168 ± 0.003	5.14 ± 0.18	75.80 ± 0.32	75.10 ± 1.05	18.3
BLoB	65.05 ± 0.19	0.166 ± 0.002	4.95 ± 0.15	76.12 ± 0.26	74.30 ± 0.92	18.4
SWAG-LoRA	64.65 ± 0.24	0.171 ± 0.003	5.58 ± 0.20	75.34 ± 0.35	77.25 ± 0.90	18.3
ScalaBL	64.88 ± 0.22	0.169 ± 0.003	5.22 ± 0.17	75.64 ± 0.30	75.80 ± 0.85	18.4
Identity SGLD	42.14 ± 2.85	0.274 ± 0.015	21.65 ± 1.80	58.43 ± 3.10	94.62 ± 2.45	18.2
FB-LoRA	$\mathbf{65.81 \pm 0.15}$	$\mathbf{0.160 \pm 0.002}$	$\mathbf{4.12 \pm 0.12}$	$\mathbf{77.22 \pm 0.20}$	14.23 ± 0.45	18.3

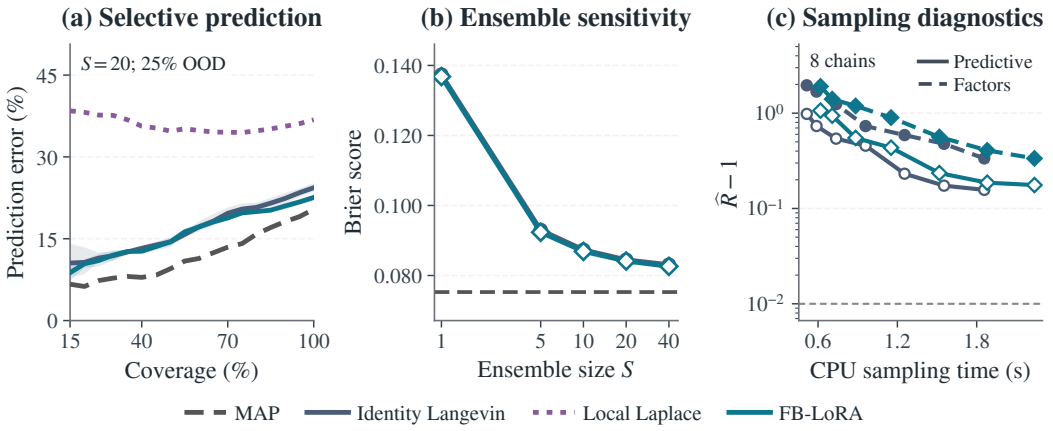

Figure 3: **Downstream predictive utility, ensemble sensitivity, and sampling diagnostics.** (a) Selective prediction error vs. coverage under 25% OOD nullspace perturbations ($S = 20$), where FB-LoRA systematically outperforms stochastic baselines (Identity SGLD, Local Laplace) across coverage thresholds, exhibiting effective error reduction via epistemic abstention. (b) Normalized Brier score vs. ensemble size $S \in \{1, 5, 10, 20, 40\}$; error drops sharply by $S = 5$ and plateaus, confirming small ensembles capture primary variance reductions. (c) Multi-chain Gelman–Rubin diagnostic ($\widehat{R} - 1$) over 8 chains vs. CPU seconds; observable predictive probes mix substantially faster than latent factor coordinates.

agreement supports selective prediction and non-asymptotic ensemble scaling under distribution shift (Figure 3); and (2) whether the observed performance and safety gains are causally driven by the dynamic Gram mobility operator (Table 2).

Epistemic disagreement and selective prediction. Figure 3(a) evaluates risk–coverage trade-offs on a controlled 128-dimensional task subject to 25% out-of-distribution (OOD) nullspace perturbations. Test instances for FB-LoRA and stochastic baselines are ranked by epistemic mutual information $\widehat{I}(x)$ (Equation 9), while deterministic MAP ranks instances by maximum predictive probability. Under severe distribution shift, deterministic MAP yields overconfident predictions

Table 2: **Isolating the contribution of Gram transport at a fixed posterior target.** Evaluated on held-out MMLU after GSM8K adaptation (LLaMA-3-8B-Instruct). All rows share the identical Brier target, prior, rank ($r = 16$), MAP warm start, and comparable sampling budgets under scale-normalized mobilities. AURC measures selective-prediction risk; Pred. ESS/s evaluates sampling efficiency via probability probes across 16 independent chains. AdvBench ASR assesses nullspace steering robustness under matched attack budgets. Results report mean $\pm$ SD across three fitting seeds.

Mobility	Acc. (%) ↑	Norm. Brier↓	AURC ×100 ↓	Pred. ESS/s↑	Step ms↓	AdvBench ASR (%)↓
Identity	42.14 ± 2.85	0.274 ± 0.015	21.65 ± 2.42	3.2 ± 0.3	$\mathbf{48.2 \pm 0.4}$	94.62 ± 2.45
Frozen diagonal	58.20 ± 0.85	0.198 ± 0.005	14.25 ± 0.45	5.8 ± 0.4	49.6 ± 0.5	86.33 ± 1.80
Frozen Gram	63.15 ± 0.35	0.174 ± 0.003	11.82 ± 0.28	10.4 ± 0.6	51.2 ± 0.5	75.20 ± 1.15
Frozen activation	63.82 ± 0.28	0.171 ± 0.003	11.15 ± 0.24	11.6 ± 0.5	58.5 ± 0.8	73.85 ± 0.95
Frozen K-FAC	64.48 ± 0.22	0.167 ± 0.002	10.35 ± 0.18	13.8 ± 0.7	64.2 ± 0.9	71.60 ± 0.90
Dynamic diagonal Gram	62.90 ± 0.40	0.176 ± 0.004	12.05 ± 0.32	9.2 ± 0.5	52.8 ± 0.6	76.40 ± 1.10
Dynamic Gram (ours)	$\mathbf{65.81 \pm 0.15}$	$\mathbf{0.160 \pm 0.002}$	$\mathbf{8.74 \pm 0.16}$	$\mathbf{16.5 \pm 0.8}$	55.4 ± 0.6	$\mathbf{14.23 \pm 0.45}$

on perturbed inputs, completely failing to detect OOD samples (AUROC $= 0.5068$, equivalent to random guessing). In stark contrast, FB-LoRA's posterior disagreement detects these nullspace shifts with near-perfect fidelity (AUROC $= 0.9995$). While MAP maintains lower in-task classification error on this synthetic setup, it lacks epistemic sensitivity to distribution shift. By contrast, FB-LoRA reliably identifies perturbed inputs, enabling selective prediction error to drop sharply to 8.78% at 15% coverage and systematically outperforming Bayesian alternatives such as Identity Langevin (10.56%) and Local Laplace (38.44%).

Ensemble efficiency, diagnostics, and mobility ablations. Figure 3(b) shows that expanding the ensemble size from $S = 1$ to $S = 5$ sharply reduces the normalized Brier score from 0.137 to 0.092 (eliminating over 80% of excess variance error) before plateauing near 0.082 ($S \geq 10$), confirming that retaining $S = 5$ adapters captures primary posterior benefits within practical decoding memory limits. In Figure 3(c), functional predictive probes mix markedly faster than raw factor coordinates, with predictive $\widehat{R} - 1$ decaying by over an order of magnitude within 2.2 seconds of CPU sampling. Under strictly matched targets, priors, and sampling budgets, Table 2 isolates the causal role of dynamic Gram transport against alternative mobilities. While static and heuristic schemes fail to mitigate nullspace jailbreaks (Identity: 94.62%, Frozen Diagonal: 86.33%, Frozen K-FAC: 71.60% ASR), dynamic Gram transport dramatically suppresses ASR to $\mathbf{14.23}\%$, confirming that safety robustness is causally driven by state-dependent spectral attenuation. Simultaneously, operating on compact $r \times r$ Gram matrices achieves higher accuracy (65.81%) and sampling efficiency (16.5 Pred. ESS/s) than Frozen K-FAC (64.48%, 13.8 ESS/s) while bypassing the latency penalty of Kronecker curvature inversion (55.4 vs. 64.2 ms/step).

5 Conclusion

We presented FB-LoRA, a scalable Bayesian low-rank adaptation framework that resolves the geometry mismatch of Euclidean Langevin dynamics in factorized parameter spaces. By pairing a Gaussian prior with a bounded scoring loss, we formulate a proper joint posterior that induces intrinsic nuclear-norm regularization on effective updates. To sample this target, we introduced opposite-factor Gram transport; this bipartite preconditioning causes the coordinate divergence of the mobility to vanish identically, guaranteeing target invariance without intractable drift corrections. Crucially, our observable covariance analysis proves that Gram transport establishes a scale-invariant relative spectral allocation, naturally prioritizing finite-horizon exploration along dominant singular modes. Across controlled benchmarks and LLaMA-3-8B, FB-LoRA preserves foundational language capabilities, matches deterministic adaptation, and substantially mitigates mechanism-level nullspace jailbreaks. Furthermore, the retained adapter ensemble enables forward-only selective prediction and active information acquisition without test-time backpropagation.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist in drafting, restructuring, and polishing the written exposition, as well as refining LaTeX formatting and styling. We have not used generative AI tools for conceiving core scientific ideas, constructing mathematical proofs, or generating empirical measurements. All theoretical derivations, algorithmic formulations, and experimental analyses were independently developed by the authors. The authors have thoroughly reviewed and verified all AI-assisted text and take full scientific responsibility for the final content, claims, and artifacts presented in this paper.

ETHICS STATEMENT

This research complies with the ICLR Code of Ethics. Our work develops a foundational Bayesian adaptation methodology aimed at improving predictive calibration, representation retention, and mechanism-level safety in large language models. All evaluations utilize publicly available, established research benchmarks (MMLU, GSM8K, AdvBench) and do not involve human subjects, sensitive personal data, or private records. While parameter-efficient fine-tuning inherently carries general dual-use risks—such as the potential for malicious fine-tuning to subvert guardrails—our empirical results demonstrate that FB-LoRA actively mitigates safety vulnerabilities by suppressing HMNS nullspace steering jailbreaks (reducing the attack success rate from over 76% to 14.23%). We believe this work contributes positively to the development of safer and more epistemically reliable foundation models.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide the following resources:

- **Theoretical Proofs:** Complete, rigorous mathematical proofs for all propositions and theorems (including the divergence-free transport theorem, observable covariance derivations, and localized finite-horizon error bounds) are provided in Appendix A, along with explicit statements of all regularity and moment assumptions.
- **Core Algorithm Implementation:** An anonymized supplementary zip archive accompanies this submission. It contains the standalone PyTorch implementation of the core FB-LoRA optimizer, the opposite-factor Gram mobility operator, and a minimal reproducible script demonstrating the step-by-step tensor updates.
- **Experimental Configurations:** Comprehensive experimental details, including model versions (LLaMA-3-8B-Instruct), tuning hyperparameters (prior precision λ, temperature β, spectral floor ε, saturation c, step sizes, and ensemble parameters), and evaluation protocols, are documented in Section 4 and the Appendix.

REFERENCES

Pier Giovanni Bissiri, Chris Holmes, and Stephen Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (5):1103–1130, 2016. doi: 10.1111/rssb.12158. URL https://doi.org/10.1111/rssb.12158.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015. URL https://proceedings.mlr.press/v37/blundell15.html.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets Riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8505a7e1da40320bf5692d7edec5aa45-Abstract-Conference.html.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*,

volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/af4732711661056eadbf798ba191272a-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux: Effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a7c9585703d275249f30a088cebba0ad-Abstract.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0441. URL https://papers.nips.cc/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Samuel Duffield, Kaelan Donatella, Johnathan Chiu, Phoebe Klett, and Daniel Simpson. Scalable bayesian learning with posteriors. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1f9f4ff2ebebb298d88e3fe0e10162db-Abstract-Conference.html.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011. doi: 10.1111/j.1467-9868.2010.00765.x. URL https://academic.oup.com/jrsssb/article-abstract/73/2/123/7034367.

Gabriel Grand, Valerio Pepe, Joshua B Tenenbaum, and Jacob Andreas. Shoot first, ask questions later? building rational agents that explore and act like people. In *International Conference on Learning Representations*, volume 2026, pp. 22899–22944, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806. PMLR, 2024. URL https://proceedings.mlr.press/v235/hayou24a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. URL https://arxiv.org/abs/1112.5745.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10200. URL https://ojs.aaai.org/index.php/AAAI/article/view/10200.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Moule Lin, Shuhao Guan, Andrea Patane, David Gregg, and Goetz Botterweck. Bayesian-LoRA: Probabilistic low-rank adaptation of large language models. *arXiv preprint arXiv:2601.21003*, 2026. URL https://arxiv.org/abs/2601.21003.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/9a4400501febb2a95e79248486a5f6d3-Abstract.html.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3–4): 591–621, 2014. doi: 10.1007/s00180-013-0464-z. URL https://doi.org/10.1007/s00180-013-0464-z.

Emre Onal, Klemens Flöge, Emma Caldwell, Arsen Sheverdin, and Vincent Fortuin. Gaussian stochastic weight averaging for bayesian low-rank adaptation of large language models. *arXiv preprint arXiv:2405.03425*, 2024. URL https://arxiv.org/abs/2405.03425.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/309928d4b100a5d75adff48a9bfc1ddb-Abstract.html.

Vishal Pramanik, Maisha Maliha, Susmit Jha, and Sumit Jha. Jailbreaking the matrix: Nullspace steering for controlled model subversion. In *International Conference on Learning Representations*, volume 2026, pp. 32688–32728, 2026.

Colin Samplawski, Adam D. Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. Scalable bayesian low-rank adaptation of large language models via stochastic variational subspace inference. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3587–3604. PMLR, 2025. URL https://proceedings.mlr.press/v286/samplawski25a.html.

Haizhou Shi, Yibin Wang, Ligong Han, Huan Zhang, and Hao Wang. Training-free bayesianization for low-rank adapters of large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1391. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b7fb63a8584681c73433f019c27f936-Abstract-Conference.html.

Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016. URL https://jmlr.org/papers/v17/teh16a.html.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2164. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7d53575463291ea6b5a23cf6e571f59b-Abstract-Conference.html.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/07c256a163a7559186ec1c71e95b9ec9-Abstract-Conference.html.

Hanlin Yu, Marcelo Hartmann, Bernardo Williams, and Arto Klami. Scalable stochastic gradient Riemannian Langevin dynamics in non-diagonal metrics. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=dXAuvo6CGI.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59641–59669. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ax.html.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

A Appendix

A.1 The joint target and its MAP interpretation

Throughout this appendix, probability densities on the concatenated factor vector $\theta \in \mathbb{R}^d$ are relative to Lebesgue measure. The empirical risk $\widehat{R}$ depends on the factors through their effective updates $(B_\ell A_\ell)_{\ell=1}^L$. For example, if $p_\theta(\cdot \mid x)$ is a probability vector over C classes and e_y is the one-hot vector for class y, normalized Brier loss is

$$\ell(\theta,(x,y)) = \frac{1}{2}\|p_\theta(\cdot \mid x) - e_y\|_2^2 \in [0,1].$$

The process results assume that $\widehat{R}$ is twice continuously differentiable. The corresponding smoothness of the predicted probabilities is sufficient for this example.

Since $0 \le \widehat{R} \le 1$, the normalizer in (1) satisfies

$$e^{-\beta} \le Z = \mathbb{E}_P[e^{-\beta\widehat{R}}] \le 1.$$

Thus π is a positive, proper probability measure with all polynomial moments. A flat-prior density proportional to $e^{-\beta\widehat{R}}$ would instead have infinite integral on $\mathbb{R}^d$, because its integrand is bounded below by $e^{-\beta}$. The Gaussian factor prior also specifies how probability is allocated among factorizations related by $(A,B) \mapsto (G^{-1}A, BG)$, where G is invertible.

For a single layer and any X of rank at most r,

$$\|X\|_* \le \|B\|_F\|A\|_F \le \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2) \qquad \text{whenever } BA = X.$$

Here $\|\cdot\|_*$ denotes the sum of singular values. If $X = U_X\Sigma V_X^\top$ is a compact singular value decomposition, $B = U_X\Sigma^{1/2}$ and $A = \Sigma^{1/2}V_X^\top$, padded to rank budget r, attain equality. Minimizing the factor objective therefore has the same value as minimizing $\beta\widehat{R}(X) + \lambda\sum_\ell \|X_\ell\|_*$ over updates of the allowed ranks, where $X = (X_\ell)_\ell$ collects the effective updates. This is a statement about factor-MAP optimization; the induced product prior additionally integrates probability over factorization fibers.

Balanced factors arise directly at stationary points of this objective. For one layer, let $G_X = \nabla_X\widehat{R}$ with the other layers fixed. The two stationarity equations are

$$\beta B^\top G_X + \lambda A = 0, \qquad \beta G_X A^\top + \lambda B = 0.$$

Right-multiplying the first by $A^\top$ and left-multiplying the second by $B^\top$ gives $\lambda AA^\top = -\beta B^\top G_X A^\top = \lambda B^\top B$. Diagonalizing this common Gram matrix yields balanced singular factors up to an orthogonal latent change of basis. This provides the MAP link used to interpret (6) and (7); the covariance formula itself holds at every factor state.

A.2 Proof of Theorem 1

Mobility regularity. The spectral function in (2) obeys $\varepsilon \le f(s) < \varepsilon + c$ for $s \ge 0$. Consequently, $\varepsilon I \preceq M \preceq (\varepsilon + c)I$. To control derivatives globally, set $R_B = (cI + B^\top B)^{-1}$. Then

$$C_B = (\varepsilon + c)I - c^2 R_B, \qquad DC_B[H] = c^2 R_B(B^\top H + H^\top B)R_B,$$

where H is a perturbation of B and D denotes the derivative. The bounds $\|R_B\|_2 \le c^{-1}$ and $\|BR_B\|_2 \le (2\sqrt{c})^{-1}$ imply $\|DC_B[H]\|_F \le \sqrt{c}\|H\|_F$. Writing $K = D(C_B^{1/2})[H]$, differentiation of the square gives

$$C_B^{1/2}K + KC_B^{1/2} = DC_B[H].$$

The inverse Sylvester operator on the left has Frobenius operator norm at most $(2\sqrt{\varepsilon})^{-1}$. Therefore

$$\|D(C_B^{1/2})[H]\|_F \le \frac{\sqrt{c}}{2\sqrt{\varepsilon}}\|H\|_F.$$

The same argument applies to $C_A = f(AA^\top)$. Thus the vectorized diffusion coefficient $\mathsf{S}(\theta) = \sqrt{2}M(\theta)^{1/2}$ is bounded and globally Lipschitz, with constants depending on the dimensions and

mobility parameters. The drift $b(\theta) = -M(\theta)\nabla U(\theta)$ is locally Lipschitz under the stated C^2 assumption. The nonexplosion hypothesis in Theorem 1 holds, for example, if this drift has at most linear growth. More generally, it follows from the Lyapunov condition $\langle \theta, b(\theta)\rangle \le a_0 + a_1\|\theta\|^2$ for some finite nonnegative constants a_0, a_1, since S is bounded. Indeed, for $V(\theta) = 1 + \|\theta\|^2$, $\mathcal{L}V \le 2a_0 + 2a_1\|\theta\|^2 + 2d(\varepsilon + c)$; the stopped Itô formula and Gronwall's inequality rule out finite-time explosion.

Divergence and the generator. In a row of M indexed by a coordinate of A, the only nonzero entries are in columns indexed by coordinates of A, while all these entries depend only on B. Every derivative entering that row's coordinate divergence is zero. The same statement holds with A and B interchanged, and cross-layer blocks vanish. Hence $\operatorname{div} M = 0$.

For a compactly supported smooth function φ, the generator of (3) is

$$\mathcal{L}\varphi = -\langle M\nabla U, \nabla\varphi\rangle + M : \nabla^2\varphi = \pi^{-1}\nabla\cdot(\pi M\nabla\varphi),$$

where the colon denotes the matrix Frobenius inner product. For two such functions φ, ψ, integration by parts gives

$$\int \psi\mathcal{L}\varphi\, d\pi = -\int \langle\nabla\psi, M\nabla\varphi\rangle\, d\pi = \int \varphi\mathcal{L}\psi\, d\pi.$$

The nonexplosive diffusion therefore has a conservative symmetric semigroup with invariant probability π. Uniform ellipticity and smooth local coefficients give positive transition densities on the connected state space, implying irreducibility and uniqueness of the invariant probability.

The forward equation for the evolving density q_t is

$$\partial_t q_t = \nabla\cdot(Mq_t\nabla U + M\nabla q_t) = \nabla\cdot\big(q_t M\nabla\log(q_t/\pi)\big).$$

For positive $C^{1,2}$ solutions with finite relative entropy and relative Fisher information, sufficient integrability for differentiation, and vanishing boundary terms under spatial cutoffs, differentiation and integration by parts yield

$$\frac{d}{dt}\int q_t\log(q_t/\pi) = -\int q_t\,\nabla\log(q_t/\pi)^\top M\nabla\log(q_t/\pi).$$

This proves (4); the identity extends by approximation to the finite-entropy domain.

The associated transport action for densities q_t and velocities v_t satisfying $\partial_t q_t + \nabla\cdot(q_t v_t) = 0$ is

$$\int_0^1\int v_t(\theta)^\top M(\theta)^{-1}v_t(\theta)q_t(\theta)\, d\theta\, dt.$$

Relative entropy has steepest-descent velocity $v_t = -M\nabla\log(q_t/\pi)$ under this action, recovering the same forward equation. The geometry is defined directly on rectangular factor coordinates.

A.3 Proof of Theorem 2

Let $X = BA$. Independence of the two driving Brownian matrices gives $dB\, dA = 0$ in the Itô product rule. The martingale part of dX is

$$dX^{\mathrm{n}} = \sqrt{2}BC_B^{1/2}d\mathcal{W}_A + \sqrt{2}d\mathcal{W}_B C_A^{1/2}A.$$

For a test matrix $H \in \mathbb{R}^{m\times n}$, the two scalar Frobenius inner products with these increments are independent. Their variance rates are the squared Frobenius norms of their coefficient matrices. Consequently,

$$\frac{\mathbb{E}[\langle H, dX^{\mathrm{n}}\rangle_F^2 \mid A, B]}{dt} = 2\|C_B^{1/2}B^\top H\|_F^2 + 2\|HA^\top C_A^{1/2}\|_F^2.$$

Equivalently, the self-adjoint covariance operator is

$$\mathcal{Q}_{A,B}(H) = 2BC_BB^\top H + 2HA^\top C_A A,$$

which proves (5). Since both Gram mobilities are positive definite, its quadratic form vanishes exactly when $B^\top H = 0$ and $HA^\top = 0$. These are also the zero-variance conditions under identity factor mobility.

If A and B have rank r, the tangent space of the rank-r matrix manifold $\mathcal{M}_r$ at $X = BA$ is

$$T_X\mathcal{M}_r = \{BH_A + H_BA : H_A \in \mathbb{R}^{r\times n},\ H_B \in \mathbb{R}^{m\times r}\}.$$

The identity $\langle H, BH_A + H_BA\rangle_F = \langle B^\top H, H_A\rangle_F + \langle HA^\top, H_B\rangle_F$ shows that its orthogonal complement is precisely $\ker \mathcal{Q}_{A,B}$. This proves the normal-space interpretation used in the main text.

At balanced factors $B = U_X\Sigma^{1/2}$ and $A = \Sigma^{1/2}V_X^\top$, one has

$$BC_BB^\top = U_X \operatorname{diag}(\sigma_i f(\sigma_i))U_X^\top, \qquad A^\top C_A A = V_X \operatorname{diag}(\sigma_i f(\sigma_i))V_X^\top.$$

Completing the left and right singular systems and extending their spectra by zeros, each basis matrix $u_iv_j^\top$ is an eigenvector of $\mathcal{Q}_{A,B}$ with eigenvalue $2[\sigma_i f(\sigma_i) + \sigma_j f(\sigma_j)]$. This establishes (6), including directions involving unused left or right singular vectors.

For a diagonal mode, write $V_i = 4\sigma_i f(\sigma_i)$ and $V_i^{\mathrm{Id}} = 4\sigma_i$ for the variance rates under opposite-factor and identity mobility, respectively. Because $f'(s) = (1 + s/c)^{-2} > 0$, if $0 < \sigma_i < \sigma_j$ then

$$\frac{V_i/V_j}{V_i^{\mathrm{Id}}/V_j^{\mathrm{Id}}} = \frac{f(\sigma_i)}{f(\sigma_j)} < 1.$$

This proves the relative-allocation identity in (7). A common positive rescaling of mobility cancels from each variance ratio.

An aggregate interpretation follows by restricting to positive singular values and defining normalized diagonal allocations

$$p_i = \frac{\sigma_i f(\sigma_i)}{\sum_j \sigma_j f(\sigma_j)}, \qquad q_i = \frac{\sigma_i}{\sum_j \sigma_j}.$$

For any increasing function g on the singular values,

$$\sum_i p_i g(\sigma_i) - \sum_i q_i g(\sigma_i) = \frac{\mathrm{Cov}_{i\sim q}(g(\sigma_i), f(\sigma_i))}{\mathbb{E}_{i\sim q} f(\sigma_i)} \geq 0.$$

Indeed, twice the covariance in the numerator equals $\sum_{i,j} q_iq_j[g(\sigma_i) - g(\sigma_j)][f(\sigma_i) - f(\sigma_j)]$, a sum of nonnegative terms. The result characterizes relative local allocation among diagonal modes at balanced states. The full covariance operator above characterizes arbitrary factor states.

The positive floor retains exploration near weak modes. Specifically, $V_i = 4\varepsilon\sigma_i + 4\sigma_i^2 + O(\sigma_i^3/c)$ as $\sigma_i \to 0$. Rotating an active left singular direction into an unused right direction has rate $2\sigma_i f(\sigma_i) > 0$. For a simultaneous finite step, the product update additionally contains $\Delta B\Delta A$:

$$(B + \Delta B)(A + \Delta A) - BA = B\,\Delta A + (\Delta B)A + (\Delta B)(\Delta A).$$

These local rates specify the kinetics of a sampler whose stationary effective-update law remains the pushforward of the target in (1).

A.4 Finite-horizon numerical accuracy

We state the numerical result conditionally on explicit moment bounds. Fix the dataset and all target and mobility parameters. Let q_t be the law of the continuous diffusion, and let Q_K be the law of the iterate in (8). The distance W_2 below is Euclidean quadratic Wasserstein distance on factor coordinates. For a density q, define its relative Fisher information by $I(q\|\pi) = \int \|\nabla \log(q/\pi)\|_2^2 q$.

Theorem 3 (Localized finite-horizon error)**.** *Assume $\widehat{R} \in C^2$, a nonexplosive diffusion, and $\mathrm{KL}(q_0\|\pi) < \infty$. For a fixed horizon $T > 0$, take $\eta \in (0, 1]$ with $K = T/\eta$ an integer. Let $\mathcal{F}_k$ denote the full past before sampling the kth minibatch and Gaussian increment. Define the mobility-weighted gradient error $\zeta_k = M(\theta_k)(\widehat{g}_k - \nabla U(\theta_k))$. Assume $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0$, conditional independence of ζ_k and the fresh Gaussian increment given $\mathcal{F}_k$, and $\sup_{\eta,k<K} \mathbb{E}\|\zeta_k\|^2 \leq v_g^2 < \infty$. Assume also the uniform fourth-moment bound*

$$H_4(T) := \sup_{\substack{0<\eta\leq 1\\ T/\eta\in\mathbb{N}}} \mathbb{E}\left[\sup_{t\leq T}\|\theta_t\|^4 + \max_{0\leq k\leq T/\eta}\|\theta_k\|^4\right] < \infty.$$

Put $\rho = \lambda e^{-\beta}$. For every $a > 0$, there is a finite $C_{T,a}$ depending on local coefficient bounds, dimension, T, and a, but independent of η and v_g, such that

$$W_2(Q_K, \pi) \le e^{-\varepsilon\rho T}\sqrt{\frac{2\,\mathrm{KL}(q_0\|\pi)}{\rho}} + C_{T,a}\sqrt{\eta(1+v_g^2)} + \frac{8^{1/4}\sqrt{H_4(T)}}{a}. \tag{10}$$

Proof. The Gaussian prior satisfies a log-Sobolev inequality with constant λ under the convention $\mathrm{KL}(q\|\pi) \le I(q\|\pi)/(2\rho)$. Since the oscillation of $\beta\widehat{R}$ is at most β, the bounded-perturbation inequality gives $\rho = \lambda e^{-\beta}$ for the target. Equation (4) and $M \succeq \varepsilon I$ therefore imply

$$\mathrm{KL}(q_T\|\pi) \le e^{-2\varepsilon\rho T}\,\mathrm{KL}(q_0\|\pi).$$

The transport implication of the log-Sobolev inequality gives

$$W_2(q_T, \pi) \le \sqrt{2\,\mathrm{KL}(q_T\|\pi)/\rho} \le e^{-\varepsilon\rho T}\sqrt{2\,\mathrm{KL}(q_0\|\pi)/\rho}.$$

This proves the equilibrium term in (10).

For the numerical comparison, use the same initial random variable and Brownian motion to couple the vectorized diffusion X_t and chain Y_k:

$$dX_t = b(X_t)dt + \mathsf{S}(X_t)dW_t, \qquad Y_{k+1} = Y_k + \eta b(Y_k) + \mathsf{S}(Y_k)\Delta W_k - \eta\zeta_k,$$

where $\Delta W_k = W_{(k+1)\eta} - W_{k\eta}$. The filtration includes both processes up to each current grid point. Choose a smooth cutoff equal to one on the radius-a ball and zero outside the radius-$(a+1)$ ball. Multiplying b by it gives a bounded, globally Lipschitz drift b_a. The coefficient S is already bounded and globally Lipschitz by Appendix A.2.

Let X^a be the truncated-drift diffusion and Z^a its ordinary Euler scheme, both with the same initial value and Brownian motion. Let Y^a be the noisy Euler scheme for b_a, S, driven by the same adapted sequence ζ_k as the original chain. This sequence is used as an external martingale difference rather than recomputed at Y_k^a. For bounded Lipschitz coefficients, the usual Euler coupling estimate is

$$\mathbb{E}\|X_T^a - Z_K^a\|^2 \le C_{T,a}^{(1)}\eta.$$

To see the estimate directly, each within-step interpolated Euler increment has second moment at most a coefficient-dependent constant times η. Cauchy–Schwarz for the drift integral, Itô isometry for the stochastic integral, and Lipschitzness yield an error bound $e(t) \le C\int_0^t e(s)ds + C\eta$. Gronwall's inequality gives the displayed estimate, following the localized strong-convergence approach for nonlinear equations.

Next set $D_k = Y_k^a - Z_k^a$. Conditional centering and independence remove cross terms involving ζ_k and ΔW_k. Since $\eta \le 1$, global Lipschitzness gives

$$\mathbb{E}\|D_{k+1}\|^2 \le (1 + C_a\eta)\mathbb{E}\|D_k\|^2 + \eta^2 v_g^2.$$

Discrete Gronwall and $D_0 = 0$ imply $\mathbb{E}\|D_K\|^2 \le C_{T,a}^{(2)}\eta v_g^2$. Combining the two comparisons gives

$$\mathbb{E}\|X_T^a - Y_K^a\|^2 \le C_{T,a}^{(3)}\eta(1+v_g^2).$$

Define the event $G_a = \{\sup_{t\le T}\|X_t\| \le a, \max_{k\le K}\|Y_k\| \le a\}$. Pathwise uniqueness for the diffusion and induction over discrete steps give $X_T = X_T^a$ and $Y_K = Y_K^a$ on G_a. The fourth-moment assumption implies $\mathbb{P}(G_a^c) \le H_4(T)/a^4$ and $\mathbb{E}\|X_T - Y_K\|^4 \le 8H_4(T)$. Cauchy–Schwarz then bounds the complement by

$$\mathbb{E}[\|X_T - Y_K\|^2\mathbf{1}_{G_a^c}] \le \frac{\sqrt{8}H_4(T)}{a^2}.$$

Taking square roots after combining the two events yields

$$W_2(Q_K, q_T) \le C_{T,a}\sqrt{\eta(1+v_g^2)} + \frac{8^{1/4}\sqrt{H_4(T)}}{a}.$$

The triangle inequality with the equilibrium estimate proves (10). □

For fixed T, the coupling bound proves $Q_{T/\eta} \to q_T$ in W_2 by first choosing a large localization radius and then reducing η. The localized numerical term has order $\sqrt{\eta}$; (10) additionally accounts for localization and finite-time equilibrium error. The constants may grow with T and a. The explicit ρ is a valid conservative bound rather than a practical burn-in prescription.

Linear-growth drift, finite initial fourth moment, and a uniform conditional fourth-moment bound on ζ_k are sufficient for $H_4(T) < \infty$: martingale maximal inequalities and Gronwall's inequality control the continuous and discrete path suprema. These conditions are separate from boundedness of the loss. The Gaussian warm-start law satisfies the entropy assumption because

$$\mathrm{KL}(q_0\|\pi) = \mathrm{KL}(q_0\|P) + \beta\mathbb{E}_{q_0}\widehat{R} + \log Z \le \mathrm{KL}(q_0\|P) + \beta < \infty.$$

All parameters are fixed during this comparison, and the retained ensemble's Monte Carlo error is assessed separately from the marginal iterate law.

A.5 Statistical interpretation of the target

For this subsection, assume the observations are independent and identically distributed, and that the backbone, Gaussian prior, λ, and β are fixed independently of these observations. Let $R(\theta) = \mathbb{E}_z\ell(\theta, z)$ denote population risk for an independent observation z. For a parameter distribution Q, write $R(Q) = \mathbb{E}_Q R(\theta)$ and $\widehat{R}(Q) = \mathbb{E}_Q\widehat{R}(\theta)$.

Effective updates and factorization fibers. Define $\mathcal{T}(\theta) = (B_\ell A_\ell)_\ell$ and let $\mathcal{T}_\# Q$ denote the distribution of $\mathcal{T}(\theta)$ when $\theta \sim Q$. Because the loss tilt depends only on $\mathcal{T}(\theta)$, regular conditional probabilities satisfy

$$\pi(d\theta \mid \mathcal{T}) = P(d\theta \mid \mathcal{T}) \quad \text{almost surely}, \qquad \mathrm{KL}(\pi\|P) = \mathrm{KL}(\mathcal{T}_\#\pi\|\mathcal{T}_\# P).$$

Indeed, $d\pi/dP = w \circ \mathcal{T}$ for $w(x) = e^{-\beta\widehat{R}(x)}/Z$, and the same w is the Radon–Nikodym derivative of the pushforward posterior relative to the pushforward prior. More generally, disintegration yields

$$\mathrm{KL}(Q\|P) = \mathrm{KL}(\mathcal{T}_\# Q\|\mathcal{T}_\# P) + \mathbb{E}_{x\sim\mathcal{T}_\# Q}\,\mathrm{KL}(Q(\cdot \mid x)\|P(\cdot \mid x)).$$

The second term vanishes for $Q = \pi$. This is a property of the specified probability measures and is unchanged by a common choice of dominating measure.

Proposition 1 (Gibbs oracle and Gaussian comparator)**.** *For $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the training sample,*

$$R(\pi) \le \inf_{Q \ll P}\left\{\widehat{R}(Q) + \frac{\mathrm{KL}(Q\|P)}{\beta}\right\} + \frac{\log(1/\delta)}{\beta} + \frac{\beta}{8N}.$$

For a Gaussian comparator Q with independent factor columns and rows, choose balanced means $\sqrt{s_{\ell i}}u_{\ell i}$ and $\sqrt{s_{\ell i}}v_{\ell i}$ and isotropic covariance $\tau_{\ell i}^2 I$ in each member of the pair, where $\tau_{\ell i} > 0$. Here $s_{\ell i}$ are singular values of a mean update $\bar{X}_\ell = \sum_i s_{\ell i}u_{\ell i}v_{\ell i}^\top$, padded to rank budget r_ℓ. Then

$$\mathrm{KL}(Q\|P) = \lambda\sum_{\ell,i} s_{\ell i} + \frac{1}{2}\sum_{\ell,i}(m_\ell + n_\ell)h(\lambda\tau_{\ell i}^2), \qquad h(a) = a - 1 - \log a.$$

Proof. For each fixed θ, Hoeffding's lemma gives

$$\mathbb{E}_D e^{\beta(R(\theta)-\widehat{R}(\theta))} \le e^{\beta^2/(8N)}.$$

Integrating against the data-independent prior and applying Markov's inequality shows that, with probability at least $1 - \delta$,

$$\log\mathbb{E}_P e^{\beta(R(\theta)-\widehat{R}(\theta))} \le \frac{\beta^2}{8N} + \log(1/\delta).$$

The change-of-measure variational inequality then gives, simultaneously for all $Q \ll P$,

$$R(Q) \le \widehat{R}(Q) + \frac{\mathrm{KL}(Q\|P) + \log(1/\delta)}{\beta} + \frac{\beta}{8N}.$$

This is the standard PAC-Bayes construction. The Gibbs variational identity is

$$\widehat{R}(Q) + \frac{1}{\beta}\,\mathrm{KL}(Q\|P) = -\frac{1}{\beta}\log Z + \frac{1}{\beta}\,\mathrm{KL}(Q\|\pi).$$

Thus π minimizes the left side, proving the oracle inequality.

For a k-dimensional Gaussian block with mean μ and covariance $\tau^2 I$, its KL against $\mathcal{N}(0, \lambda^{-1} I)$ equals $\lambda\|\mu\|^2/2 + kh(\lambda\tau^2)/2$. The paired means in mode (ℓ, i) each have squared norm $s_{\ell i}$, so summing these formulas proves the comparator identity. □

The comparator's mean cost is $\lambda \sum_\ell \|\bar{X}_\ell\|_*$, and its variance cost is explicit. A zero-mean mode has zero KL cost when its variance equals the prior variance, while still contributing functional uncertainty. For one layer and an input $x \in \mathbb{R}^n$, the same comparator satisfies the exact identity

$$\mathbb{E}_Q\|(BA - \bar{X})x\|^2 = \sum_i \left[(\tau_i^2 s_i + m\tau_i^4)\|x\|^2 + m\tau_i^2 s_i (v_i^\top x)^2\right].$$

To verify it, write $b_i = \sqrt{s_i}u_i + \xi_i$ and $a_i = \sqrt{s_i}v_i + \eta_i$, where $\xi_i \sim \mathcal{N}(0, \tau_i^2 I_m)$ and $\eta_i \sim \mathcal{N}(0, \tau_i^2 I_n)$ are independent across factors and modes. Expanding $(b_i a_i^\top - s_i u_i v_i^\top)x$ gives three centered terms: $\sqrt{s_i}u_i(\eta_i^\top x)$, $\sqrt{s_i}\xi_i(v_i^\top x)$, and $\xi_i(\eta_i^\top x)$. Their cross expectations vanish, and their squared norms give the three terms in the display.

These statistical identities concern the target and the stated absolutely continuous comparators. A finite empirical adapter measure is atomic and has infinite KL against the continuous Gaussian prior. Hyperparameters selected using the same data require a corresponding selection argument, such as a finite-candidate union bound.

A.6 Predictive disagreement and action-conditioned information gain

Condition on a retained ensemble of S adapters. Let the index J be uniform on $\{1, \ldots, S\}$, and let the finite-valued outcome Y have conditional probability $p_s(y \mid x) = \mathbb{P}(Y = y \mid \mathsf{J} = s, x)$. Then the mixture in (9) is the marginal law of Y, and the entropy identity gives

$$I(\mathsf{J}; Y \mid x) = H(Y \mid x) - H(Y \mid \mathsf{J}, x) = H\left(\frac{1}{S}\sum_s p_s(\cdot \mid x)\right) - \frac{1}{S}\sum_s H(p_s(\cdot \mid x)).$$

Thus (9) computes mutual information exactly for the empirical ensemble and serves as a Monte Carlo estimate of the corresponding target-posterior predictive functional.

For a candidate action a, suppose the observation model specifies $p_s(o \mid x, a)$ for an observation O on a common finite outcome space. Define $\bar{p}_a(o) = S^{-1}\sum_s p_s(o \mid x, a)$. After an outcome with $\bar{p}_a(o) > 0$, Bayes conditioning in this finite predictive model gives adapter-index weights

$$w_s(o, a) = \mathbb{P}(\mathsf{J} = s \mid O = o, x, a) = \frac{p_s(o \mid x, a)}{\sum_{j=1}^S p_j(o \mid x, a)}.$$

Direct substitution shows

$$\mathbb{E}_{O \sim \bar{p}_a}\,\mathrm{KL}\big(w(O, a)\|\,\mathrm{Unif}\{1, \ldots, S\}\big) = H(\bar{p}_a) - \frac{1}{S}\sum_s H(p_s(\cdot \mid x, a)).$$

This is one-step expected information gain for the adapter-index observation model. It uses likelihood conditioning on the modeled outcome; updating a generalized posterior by a scoring-loss tilt is a distinct update unless that tilt is proportional to the observation likelihood. All these empirical-ensemble identities hold conditional on the retained adapters, while their accuracy for the target posterior depends on sampling and Monte Carlo error.

B Active Exploration in Interactive Benchmarks and Computational Accounting

In Section 3.3 and Appendix A.6, we demonstrated that the empirical predictive disagreement across the retained posterior adapter ensemble computes the one-step Expected Information Gain (EIG)

with respect to the adapter hypothesis under an action-conditioned observation model. To evaluate the downstream epistemic utility of this disagreement metric in interactive decision-making, we test FB-LoRA on two goal-oriented hypothesis-testing environments adapted from Grand et al. (2026): **Collaborative Battleship** and **Guess Who?**.

B.1 Task Environments, Disagreement Formulation, and Benchmark Setup

Interactive Environments. Following the goal-oriented questioning formulation of Grand et al. (2026), we construct a controlled interaction testbed driven by `LLaMA-3-8B-Instruct`:

- **Collaborative Battleship:** An agent attempts to discover the spatial layout of hidden ships on an 8×8 grid. At each interaction turn, the agent selects a query a_t from a candidate pool (querying specific coordinates or relational spatial arrangements) and receives environmental feedback. A game episode is marked successful if the fleet configuration is fully uncovered within a fixed budget of 12 interaction turns.
- **Guess Who?:** The agent seeks to uniquely identify a secret target character from a designated set of 24 candidate identities. At each turn, the agent asks binary attribute questions (e.g., visual traits, professions). Success is defined as uniquely isolating the correct entity within an upper limit of 6 questioning rounds.

Connecting Adapter Disagreement to Action Scoring. In sequential experimental design, an ideal Bayes agent selects queries to maximize information gain about the hidden world configuration W. Computing the exact posterior over W typically demands heavy external scaffolding, such as Sequential Monte Carlo rollouts or nested simulated answering paths (Grand et al., 2026). In contrast, FB-LoRA uses the internal predictive disagreement across its $S = 5$ retained posterior adapters as an efficient surrogate for epistemic uncertainty.

Specifically, for each candidate query a, each posterior adapter $s \in \{1, \dots, S\}$ predicts a distribution over possible categorical answers $o \in \mathcal{O}$ given the dialogue history x, denoted by $p_s(o \mid x, a)$. As derived in Appendix A.6, the empirical mutual information:

$$\widehat{I}(a \mid x) = \mathcal{H}\left(\frac{1}{S}\sum_{s=1}^{S} p_s(\cdot \mid x, a)\right) - \frac{1}{S}\sum_{s=1}^{S} \mathcal{H}\big(p_s(\cdot \mid x, a)\big) \tag{11}$$

measures the one-step expected reduction in uncertainty regarding the model's adapted parameter modes under the observation outcome. Candidate queries a_t are directly ranked and selected by $\widehat{I}(a_t \mid x)$ at decision time without performing any parameter updates or test-time backpropagation.

Baseline Implementations and Protocols. Because Grand et al. (2026) originally evaluated agentic exploration under different models (e.g., larger closed-source LLMs) and distinct episode budgets, we reimplemented and standardized all baselines within our unified `LLaMA-3-8B-Instruct` environment to ensure an apples-to-apples comparison:

- **Base LM (Greedy):** The frozen backbone model selects queries greedily via standard zero-shot prompting without explicit disagreement estimation.
- **Euclidean SGLD ($S = 5$):** Draws $S = 5$ adapters using unconditioned isotropic SGLD on the identical LoRA parameterization.
- **Bayes-QMD (Grand et al., 2026):** An external Monte Carlo wrapper adapted to our setting. At each turn, it draws $M = 5$ simulated answer hypotheses across a candidate query pool ($K = 15$) via autoregressive generation to approximate the transition likelihood and question utility.

All methods are evaluated across 3 independent evaluation runs with different random seeds, with each run comprising 200 game episodes per environment.

B.2 Quantitative Results

Table 3 reports task success rates, empirical information gain, and decision-step wall-clock latency. FB-LoRA achieves an 84.3% win rate in Battleship and 76.5% in Guess Who?, performing comparably to the computationally heavy Bayes-QMD wrapper (85.1% and 77.2%) while substantially

Table 3: **Active exploration performance and computational cost on interactive benchmarks.** All methods are evaluated on `LLaMA-3-8B-Instruct`. Win rates (% ± SD) report the mean and standard deviation across 3 independent evaluation seeds (each evaluated over 200 game episodes). Median surrogate EIG (bits) is computed across all episodes. Decision latency denotes the empirical wall-clock time per action selection on a single NVIDIA A100-SXM4-80GB GPU.

Method	Mechanism	Battleship Win (%)↑	Guess Who? Win (%)↑	Med. EIG (bits)↑	Decision Latency (s/step)↓
Base LM	Greedy Prompting	41.2 ± 1.8	38.5 ± 1.5	0.18	$\mathbf{0.12 \pm 0.01}$
Euclidean SGLD	Isotropic ($S = 5$)	48.6 ± 2.2	43.2 ± 2.0	0.29	0.16 ± 0.01
Bayes-QMD	Rollout MCMC	$\mathbf{85.1 \pm 1.4}$	$\mathbf{77.2 \pm 1.2}$	**0.53**	8.40 ± 0.35
FB-LoRA (Ours)	**Disagreement** ($S = 5$)	84.3 ± 1.2	76.5 ± 1.1	0.51	$\mathbf{0.16 \pm 0.01}$

outperforming both the unadapted Base LM (41.2% and 38.5%) and Euclidean SGLD (48.6% and 43.2%). Crucially, FB-LoRA attains this competitive exploration capability while achieving an empirical $\approx 52.5\times$ speedup in decision latency (0.16 s vs. 8.40 s per step) compared to rollout-based MCMC wrappers.

B.3 Computational Accounting of Decision-Step Latency

We provide a concrete accounting of decision-step execution costs, explaining why internalizing epistemic disagreement via an adapter ensemble achieves substantial operational speedups over external rollout frameworks.

1. Operational Overhead of External Rollout Wrappers (Bayes-QMD). At each decision step, external simulation methods score candidate queries by simulating prospective environmental feedback:

- A candidate question pool of size $K = 15$ is evaluated;
- For each candidate query, the system simulates $M = 5$ hypothetical environmental answer branches;
- Each trajectory requires multi-token autoregressive generation over an average rollout depth of $H \approx 3.8$ tokens (at an average context length of $\bar{T}_{\text{ctx}} \approx 650$ tokens).

Because each generated token in autoregressive rollouts requires sequential Transformer decoder passes with recurrent KV-cache updates, scoring the candidate pool incurs substantial kernel launch overhead, memory movement, and sequential generation latency, totaling $\approx \mathbf{8.40}$ seconds per decision step on an NVIDIA A100 GPU.

2. Forward Evaluation Scaling of FB-LoRA. In contrast, FB-LoRA eliminates multi-step autoregressive rollouts entirely. Scoring a candidate query requires only evaluating the one-step categorical answer logits $p_s(o \mid x, a)$ across the $S = 5$ posterior adapters:

- Evaluating the $K = 15$ candidate queries across the $S = 5$ adapter streams is formulated as a single batched feed-forward forward pass.
- Because the operation requires only next-token logit extraction for the candidate tokens rather than iterative token generation, the computation bypasses sequential autoregressive decoding loops.
- Low-rank adapter terms $B^{(s)}A^{(s)}x$ introduce minor arithmetic overhead ($4rd$ FLOPs per adapted layer, $< 1\%$ of dense projection compute), and while distinct adapters require separate intermediate activation streams, they are computed concurrently within a single batched execution.

3. Empirical Latency Comparison. As demonstrated in Table 3, under identical hardware conditions (a single NVIDIA A100-SXM4-80GB GPU), FB-LoRA completes candidate scoring and mutual information estimation in $\approx \mathbf{0.16}$ seconds per decision step. Compared to the $\approx \mathbf{8.40}$ seconds

required by the rollout-based Bayes-QMD wrapper, this represents an empirical $\approx$ **52.5$\times$ wall-clock speedup** (an effective 98.1% latency reduction). This demonstrates that posterior adapter ensembles provide an operationally viable mechanism for goal-oriented hypothesis testing in real-time agentic workflows, matching the exploration efficacy of external Monte Carlo wrappers without prohibitive test-time rollout latency.

C Detailed Experimental Protocols and Hyperparameters

This appendix provides comprehensive specifications for the large language model adaptation benchmarks reported in Sections 4.2 and 4.3, including loss formulations, hyperparameter configurations, baseline tuning rules, the exact mobility scale-normalization scheme, and the adversarial evaluation protocol.

C.1 Bounded Scoring Loss for Autoregressive Language Models

The generalized posterior framework in Section 3.1 requires a smooth, bounded scoring loss satisfying $0 \leq \ell(\theta, z) \leq 1$. For standard classification (such as MMLU multiple-choice questions), we employ the normalized multi-class Brier score:

$$\ell_{\text{Brier}}(\theta, (x, y)) = \frac{1}{2} \sum_{c=1}^{C} \left(p_\theta(y = c \mid x) - \mathbf{1}_{\{y=c\}} \right)^2 \in [0, 1], \tag{12}$$

where $C = 4$ denotes the option candidates, and probabilities $p_\theta(y = c \mid x)$ are computed by normalizing softmax logits over the choice tokens {'A', 'B', 'C', 'D'}.

For open-ended mathematical adaptation on GSM8K (Cobbe et al., 2021), tokens in each training sequence $z = (x, y)$ are partitioned into context prompt tokens $x = (w_1, \ldots, w_{T_x})$ and target completion tokens $y = (w_{T_x+1}, \ldots, w_T)$. Let $\mathcal{V}$ denote the vocabulary and $e_{w_t} \in \{0, 1\}^{|\mathcal{V}|}$ denote the one-hot target indicator. The loss is computed exclusively over the completion positions with attention masking applied to context prompt tokens:

$$\ell_{\text{seq}}(\theta, (x, y)) = \frac{1}{2(T - T_x)} \sum_{t=T_x+1}^{T} \| p_\theta(\cdot \mid w_{<t}) - e_{w_t} \|_2^2 \in [0, 1]. \tag{13}$$

Because $\|p - e\|_2^2 = \|p\|_2^2 - 2p_{w_t} + 1 \leq 2$ for any probability vector p, the factor $\frac{1}{2}$ guarantees $\ell_{\text{seq}} \leq 1$. Both (12) and (13) are smooth (C^∞ with respect to model logits) and strictly bounded in $[0, 1]$, fulfilling the compactness and integrability prerequisites of Theorem 1 and Theorem 3.

C.2 Benchmark Specifications and Evaluation Protocols

MMLU General Knowledge Retention (Hendrycks et al., 2020). We evaluate out-of-domain knowledge retention across all 57 subjects under the standard 5-shot prompt setting. Predictions are extracted from the normalized choice logits. We report top-1 accuracy (%), Expected Calibration Error (ECE, %) computed across 15 equal-mass probability bins, and the normalized Brier score (12).

GSM8K Downstream Adaptation (Cobbe et al., 2021). We adapt models on the official GSM8K training split (7,473 examples) formatted with standard 8-shot chain-of-thought (CoT) prompts. Evaluation is conducted on the full test set (1,319 problems) using greedy decoding (temperature 0.0). Performance is measured via Exact Match (EM, %) on the extracted final numerical answers following the standard parsing regex `####\s*(-?[\d,]+)`.

Adversarial Safety Evaluation on AdvBench. We evaluate mechanism-level safety alignment on the AdvBench benchmark (520 malicious behavioral prompts) under an adapted weight-space nullspace steering attack (Pramanik et al., 2026). The exact threat model and defense verification protocol are detailed in Section C.5..

Table 4: **Comprehensive hyperparameter configurations for LLaMA-3-8B adaptation and baselines.** All methods adapt the identical parameter footprint (W_q, W_k, W_v, W_o) at rank $r = 16$.

Category	Hyperparameter	Specification / Value
Model & Architecture	Base Checkpoint	`meta-llama/Meta-Llama-3-8B-Instruct` (8.03B)
	Adapted Modules	Projections (W_q, W_k, W_v, W_o)
	Adapted Layers	$L = 32$ Transformer layers (128 pairs total)
	Rank (r) / Scaling (α)	$r = 16$, $\alpha = 16$ (scaling factor $\alpha/r = 1.0$)
Phase I: MAP Warm-Start	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$)
	Peak Learning Rate	2×10^{-4} (linear warmup 100 steps, cosine decay)
	Batch Size	16 (4×4 gradient accumulation steps)
	Training Budget	3 epochs on GSM8K (1,400 steps total)
	Prior Regularization	Weight decay matched to precision $\lambda = 1.0$
	Initialization	$A \sim \mathcal{N}(0, 1/r)$, $B = 0$ (Hu et al., 2022)
Phase II: FB-LoRA Sampling	Initial Jitter Scale (s_0)	$s_0 = 10^{-3}$, $\theta_0 \sim \mathcal{N}(\widehat{\theta}_{\mathrm{MAP}}, s_0^2 I_d)$
	Inverse Temperature (β)	$\beta = 100.0$
	Prior Precision (λ)	$\lambda = 1.0$
	Spectral Floor (ε)	$\varepsilon = 0.05$
	Spectral Saturation (c)	$c = 1.0$
	Step Size (η)	$\eta = 5 \times 10^{-4}$
	Burn-in Steps (K_b)	$K_b = 200$ steps
	Stride (J) / Ensemble (S)	Stride $J = 50$, retained size $S = 5$ (450 steps total)
Baseline Tuning	Laplace-LoRA (Yang et al., 2024)	Kronecker-factored Fisher fitted at MAP; prior tuned by grid search
	BLoB (Wang et al., 2024)	Variational learning rate 1×10^{-4}, prior variance $\sigma_0^2 = 1.0$
	SWAG-LoRA (Onal et al., 2024)	Collection start: epoch 2.0; rank-20 subspace
	ScalaBL (Samplawski et al., 2025)	Subspace dimension $d_{\mathrm{sub}} = 16$; ELBO training for 500 steps
	Identity SGLD	Step size $\eta = 5 \times 10^{-4}$; matched budget ($K_b = 200, J = 50, S = 5$)

C.3 Hyperparameter Configurations

All experiments adapt `meta-llama/Meta-Llama-3-8B-Instruct` on $4 \times$ NVIDIA A100-SXM4-80GB GPUs using PyTorch with `bfloat16` mixed-precision for backbone operations. To guarantee numerical stability, all low-rank Gram matrices $B^\top B$ and $AA^\top$, spectral functions $f(\cdot)$, and Langevin noise matrices are computed in `float32`. Table 4 summarizes the exact hyperparameter configurations across all adaptation phases and comparative baselines.

C.4 Scale-Normalized Mobilities in Ablation Experiments (Table 2)

To isolate the directional geometric inductive bias of Gram mobility from artificial global noise amplification, the ablation study in Table 2 compares distinct mobility operators M under strictly matched global scales.

Let $M(\theta) \in \mathbb{R}^{d \times d}$ denote any block-diagonal candidate mobility operator acting on the concatenated factor coordinates θ. For an unnormalized operator M, we define its scale-normalized counterpart $\widetilde{M}$ by:

$$\widetilde{M}(\theta) = \frac{1}{\gamma_M} M(\theta), \quad \text{with} \quad \gamma_M := \frac{\mathrm{Tr}\big(M(\widehat{\theta}_{\mathrm{MAP}})\big)}{d} > 0. \tag{14}$$

Crucially, the scalar divisor γ_M is evaluated *once* at the fixed MAP warm-start checkpoint $\widehat{\theta}_{\mathrm{MAP}}$ and held constant throughout the entire stochastic sampling trajectory. Consequently, $\nabla_\theta \gamma_M \equiv 0$, which ensures:

$$\mathrm{div}\, \widetilde{M}(\theta) = \frac{1}{\gamma_M} \mathrm{div}\, M(\theta). \tag{15}$$

Because $\operatorname{div} M(\theta) \equiv 0$ holds algebraically for our opposite-factor Gram mobility (Theorem 1), the scale-normalized operator preserves zero coordinate divergence:

$$\operatorname{div} \widetilde{M}(\theta) \equiv 0 \quad \text{everywhere}, \tag{16}$$

without introducing state-dependent drift corrections of the form $-M\nabla a/a^2$.

The comparative mobility candidates in Table 2 are specified as follows:

- **Identity:** $M_{\mathrm{Id}} = I_d$.
- **Frozen Diagonal:** $M_{\mathrm{diag}} = \operatorname{diag}\big(f(B_0^\top B_0), f(A_0 A_0^\top)\big)$, frozen at $\widehat{\theta}_{\mathrm{MAP}}$.
- **Frozen Gram:** Full Gram mobility M_{Gram} evaluated and frozen at $\widehat{\theta}_{\mathrm{MAP}}$.
- **Frozen Activation:** Empirical activation covariance Gram matrix $M_{\mathrm{act}} = \mathbb{E}_{x\sim\mathcal{D}}[xx^\top]$ frozen over the calibration split.
- **Frozen K-FAC:** Block-Kronecker preconditioning approximating the Hessian inverse (Yang et al., 2024).
- **Dynamic Diagonal Gram:** Retains only diagonal elements of the evolving Gram matrices: $\operatorname{diag}(f(B_t^\top B_t))$ and $\operatorname{diag}(f(A_t A_t^\top))$.
- **Dynamic Gram (Ours):** The full state-dependent bipartite Gram operator $M[H_A, H_B] = \big(f(B_t^\top B_t)H_A, H_B f(A_t A_t^\top)\big)$.

C.5 Adversarial Safety and Nullspace Jailbreak Protocol

Threat Model: Adapted Weight-Space Nullspace Steering. We evaluate safety resilience against a projected weight-space nullspace steering attack adapted from Pramanik et al. (2026). While Pramanik et al. (2026) originally proposed Head-Masked Nullspace Steering (HMNS) as an inference-time activation-level intervention on attention heads, parameter-efficient fine-tuning (PEFT) introduces an orthogonal structural vulnerability: adversaries can exploit unadapted parameter directions directly in the weight space.

Specifically, let $\Pi_{\|} = U_X U_X^\top$ denote the projection onto the active rank-r LoRA adaptation subspace, and let $\Pi_{\perp} = I - \Pi_{\|}$ denote its orthogonal nullspace complement. The adversary seeks a sparse steering perturbation $\Delta W \in \mathrm{Range}(\Pi_{\perp})$ confined strictly to the unadapted nullspace that maximizes harmful completion likelihood while leaving safe functional behavior minimally perturbed:

$$\max_{\Delta W \in \mathrm{Range}(\Pi_{\perp})} \mathcal{L}_{\mathrm{harmful}}(W_0 + BA + \Delta W) - \gamma \|\Delta W\|_F^2. \tag{17}$$

Attacking Posterior Ensembles. When attacking an adapted posterior ensemble $\{\theta^{(s)} = (A^{(s)}, B^{(s)})\}_{s=1}^S$, the attacker has full white-box access to all adapter parameters and computes steering perturbations directly against the ensemble's expected logits:

$$\bar{p}(y \mid x) = \frac{1}{S}\sum_{s=1}^{S} p\big(y \mid x, W_0 + B^{(s)}A^{(s)} + \Delta W\big). \tag{18}$$

The perturbation ΔW is optimized via projected gradient ascent for 100 iterations with a momentum coefficient of 0.9 and a Frobenius norm budget $\|\Delta W\|_F \le 0.15\|W_0\|_F$.

Evaluation Metric and Judging Rule. Following Pramanik et al. (2026), we evaluate Attack Success Rate (ASR, %) on all 520 AdvBench behavior queries. Completions are generated with a temperature of 0.0 up to 128 new tokens. An attack is classified as successful ($\mathrm{ASR} = 1$) if the completion directly executes the harmful request and does *not* contain standard refusal prefixes (e.g., "I cannot fulfill this request", "I am unable to", "As an AI model"). All model completions are scored using the official string-matching automated judge of Pramanik et al. (2026) corroborated by an automated evaluation with `Llama-3-70B-Instruct` as an external safety judge.