ONLINE GRAPH CLASSIFICATION WITH ONE-BIT FEEDBACK: A GRAPH CONVOLUTIONAL BANDIT APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online graph classification with one-bit feedback, where the learner observes only whether its predicted class is correct. We train a graph convolutional classifier using the selected-class Bernoulli likelihood and explore through upper confidence bound (UCB) or Thompson sampling (TS). As the graph encoder changes, the head's accumulated uncertainty can become inconsistent with the current features. Moment transport (MT) updates the joint head mean and covariance to account for this change. We establish prediction and decision invariance under invertible feature transformations, give a counterexample for stale uncertainty, and bound decision perturbations under approximate transport. Experiments verify exact invariance and show improved decision quality over stale covariance during learned GCN updates. On four real datasets, `UCB-MT` ranks among the top four of nineteen methods in mean online mistake rate. Comparisons with stronger variance fitting give mixed results, showing that the benefit depends on the uncertainty model and exploration rule.

1 INTRODUCTION

Graph classification predicts an outcome from both the attributes of individual objects and their relations. Molecular structures, protein networks and social interaction graphs are common examples (Morris et al., 2020). In an online setting, graphs arrive sequentially and a prediction must be made before its outcome is observed. Feedback can be more limited than a class label: a proposed classification may be accepted or rejected without revealing the correct class after an error. This setting connects graph classification to contextual bandits, where the graph is the context, the predicted class is the action, and correctness is the reward. The learner must use these partial observations both to improve its classifier and to decide which class to predict next.

Multiclass bandit methods learn from correctness feedback (Kakade et al., 2008; Foster & Krishnamurthy, 2021), while neural bandits learn representations for exploration (Zhou et al., 2020; Zhang et al., 2021). Combining these approaches introduces a difficulty when the representation changes during learning. The decision layer accumulates uncertainty in the coordinates of its previous features. Updating the encoder and classifier mean does not transform that uncertainty into the new coordinates. Likelihood matching and variance fitting address related problems in neural-linear bandits (Zahavy & Mannor, 2019; Nabati et al., 2021). A multinomial classifier must also preserve dependence among class predictions and learn from the likelihood of the selected correctness bit. Our aim is to update the graph representation while retaining the joint uncertainty needed for UCB and Thompson sampling.

We propose `OGC-GCN-UCB-MT` and `OGC-GCN-TS-MT`, which train a graph convolutional encoder and a joint multinomial head from one-bit correctness feedback. Both algorithms use moment transport to update the head mean and covariance after an encoder change, with UCB and TS providing alternative exploration rules. We establish exact and sequential decision invariance, exhibit a failure of stale uncertainty, and derive approximate-transport bounds and conditional regret guarantees. Experiments examine these effects during learned encoder updates and compare the complete algorithms on four real graph datasets. The results show improvements over stale covariance and competitive classification performance. Comparisons with replay and stronger variance fitting further characterize the accuracy and computational tradeoffs.

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.

2 Related work

Multiclass prediction with bandit feedback. `Banditron` and `Gaptron` learn from the correctness of a selected class (Kakade et al., 2008; van der Hoeven, 2020). Recent work develops first- and second-order methods for this setting (Feng et al., 2025). `FastCB` and `SquareCB` reduce contextual bandits to prediction oracles (Foster & Krishnamurthy, 2021; Foster & Rakhlin, 2020). In our problem, each input is an attributed graph and the classifier predicts a jointly normalized class distribution. We train this model using the likelihood of the selected correctness bit.

Neural representations and uncertainty. Neural bandits learn representations for exploration (Zhou et al., 2020; Zhang et al., 2021; Salgia, 2023; Deb et al., 2024), with guarantees that depend on assumptions about the model and its optimization (Tirinzoni et al., 2023). `LinReBoot` explores by resampling linear-model residuals (Wu et al., 2022). Likelihood matching and `LiM2` update uncertainty after feature changes (Zahavy & Mannor, 2019; Nabati et al., 2021). We study this problem for a joint multinomial head trained from correctness feedback. The joint predictive distribution matters for TS (Osband et al., 2023), and approximation errors can affect its regret (Phan et al., 2019). Guarantees for posterior samplers (Mazumdar et al., 2020; Clavier et al., 2024; Zheng et al., 2024) require assumptions that are not established for our Fisher-based Gaussian model.

Graphs as bandit inputs. Graphs can represent relations among users, actions or observations (Qi et al., 2023; Zhang et al., 2023). Laplacian kernelized bandits couple user rewards through graph smoothness (Wu & Amini, 2026). Graph-action bandits select a graph at each round (Kassraie et al., 2022), and `GNN-TS` uses graph neural tangent features for exploration (Wu & Amini, 2024). In our setting, the learner observes a graph and selects a class. We study how to update joint class uncertainty as its graph representation changes.

3 Problem formulation

Let $\mathcal{G}$ be a space of attributed graphs. Each graph $G = (\boldsymbol{A}, \boldsymbol{X}) \in \mathcal{G}$ has an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and a node-feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d_x}$, whose i-th row contains the features of node i. The number of nodes n may vary across graphs. Given G, the classification task is to predict a label $Y \in [K] = \{1, \dots, K\}$, where $K \geq 2$. Its conditional class distribution is

$$\boldsymbol{\mu}(G) = (\mu_1(G), \dots, \mu_K(G))^\top, \qquad \mu_k(G) = \mathbb{P}(Y = k \mid G), \qquad \boldsymbol{\mu}(G) \in \Delta^{K-1}, \tag{1}$$

where $\mathbf{1}$ and $\mathbf{0}$ are the all-one and all-zero vectors, and $\Delta^{K-1} = \{\boldsymbol{p} \in \mathbb{R}^K : \boldsymbol{p} \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{p} = 1\}$. If this distribution were known, a Bayes classifier would choose $f^\star(G) \in \arg\max_{k\in[K]} \mu_k(G)$, minimizing the conditional mistake probability $1 - \mu_{f(G)}(G)$.

We consider the online version of this problem. At round t, the learner observes G_t, predicts $A_t \in [K]$, and receives only

$$R_t = \mathbf{1}\{A_t = Y_t\}. \tag{2}$$

Thus a correct prediction identifies Y_t, whereas an incorrect prediction only excludes A_t. Let $\mathcal{F}_{t-1}$ denote the learner's completed history before round t. Under a stable conditional label law and action randomization independent of the hidden label given the graph and history,

$$R_t \mid (\mathcal{F}_{t-1}, G_t, A_t) \sim \text{Bernoulli}(\mu_{A_t}(G_t)), \qquad \mathbb{E}[R_t \mid \mathcal{F}_{t-1}, G_t, A_t] = \mu_{A_t}(G_t). \tag{3}$$

Online graph classification is therefore a K-armed contextual bandit with Bernoulli rewards. Learning the classifier amounts to learning which coordinates of $\boldsymbol{\mu}(G)$ are large. Unlike general Bernoulli bandits, the class reward means are constrained to sum to one.

The learner aims to minimize cumulative mistakes M_T. To separate the cost of its decisions from intrinsic label uncertainty, we also consider Bayes pseudo-regret $\mathcal{R}_T$:

$$M_T = \sum_{t=1}^{T}(1 - R_t), \qquad \mathcal{R}_T = \sum_{t=1}^{T}\left[\mu_t^\star - \mu_{A_t}(G_t)\right], \qquad \mu_t^\star = \max_k \mu_k(G_t). \tag{4}$$

Writing $N_T^\star = \sum_{t=1}^{T}(1-\mu_t^\star)$, we have $\mathbb{E}[M_T] = \mathbb{E}[N_T^\star] + \mathbb{E}[\mathcal{R}_T]$. Thus even zero pseudo-regret can coexist with positive mistake rates. We evaluate $\mathcal{R}_T$ when the generating probabilities are known and M_T on ordinary labeled datasets. Proposition 8 gives the pathwise decomposition and feedback-noise bound.

4 GCN MULTINOMIAL LOGISTIC BANDITS

We model the unknown class distribution with a GCN encoder and a multinomial logistic head. The encoder produces a graph representation, and the head uses correctness feedback to estimate class probabilities and uncertainty. Moment transport updates the head state after each accepted encoder change.

4.1 GCN CLASSIFIER

A graph convolutional network aggregates information from neighboring nodes and then pools the node representations into a graph feature (Kipf & Welling, 2017). Let $\boldsymbol{D}$ be the diagonal degree matrix of $\boldsymbol{A}$ and $\boldsymbol{I}$ an identity matrix of the required dimension. Add self-loops through $\widehat{\boldsymbol{A}} = (\boldsymbol{D}+\boldsymbol{I})^{-1/2}(\boldsymbol{A}+\boldsymbol{I})(\boldsymbol{D}+\boldsymbol{I})^{-1/2}$. For a network with L layers,

$$\boldsymbol{H}^{(0)} = \boldsymbol{X}, \qquad \boldsymbol{H}^{(\ell+1)} = \sigma_\ell(\widehat{\boldsymbol{A}}\boldsymbol{H}^{(\ell)}\boldsymbol{W}^{(\ell)}), \qquad \boldsymbol{z}_{\boldsymbol{\phi}}(G) = \begin{bmatrix} n^{-1}\sum_i (\boldsymbol{H}^{(L)}_{i,:})^\top \\ 1 \end{bmatrix} \in \mathbb{R}^m. \tag{5}$$

Here $\boldsymbol{W}^{(\ell)}$ maps layer ℓ to layer $\ell+1$, σ_ℓ is its activation, $\boldsymbol{H}^{(\ell)}$ is the matrix of node representations at layer ℓ, and $\boldsymbol{\phi}$ collects the encoder parameters. The final coordinate of $\boldsymbol{z}_{\boldsymbol{\phi}}(G)$ is the intercept term.

The multinomial head converts this representation into class probabilities. Fix $\boldsymbol{C} \in \mathbb{R}^{K\times(K-1)}$ with $\boldsymbol{C}^\top\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{C}^\top\boldsymbol{1} = \boldsymbol{0}$, and let $\boldsymbol{\Theta} \in \mathbb{R}^{(K-1)\times m}$. Define

$$\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\phi}}(G) = \boldsymbol{C}\boldsymbol{\Theta}\boldsymbol{z}_{\boldsymbol{\phi}}(G), \qquad \widehat{p}_{\boldsymbol{\theta},\boldsymbol{\phi},k}(G) = \frac{e^{g_k}}{\sum_{j=1}^K e^{g_j}}, \qquad \boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}) \in \mathbb{R}^d. \tag{6}$$

Here $d = m(K-1)$, vec stacks matrix columns, and $\otimes$ denotes the Kronecker product. The components $g_k := [\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\phi}}(G)]_k$ are the class logits. The contrast basis removes the redundant common shift of all logits while retaining their joint dependence. With $\boldsymbol{Z}_{\boldsymbol{\phi}}(G) = \boldsymbol{z}_{\boldsymbol{\phi}}(G)^\top \otimes \boldsymbol{C}$, the logits are linear in the head: $\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\phi}}(G) = \boldsymbol{Z}_{\boldsymbol{\phi}}(G)\boldsymbol{\theta}$.

4.2 LEARNING FROM CORRECTNESS FEEDBACK

The learner observes a Bernoulli outcome for its selected class, rather than a complete multinomial label. Its partial-feedback negative log likelihood (PLL) is

$$\ell_t^{\mathrm{PLL}}(\boldsymbol{\theta},\boldsymbol{\phi}) = -R_t \log \widehat{p}_{\boldsymbol{\theta},\boldsymbol{\phi},A_t}(G_t) - (1-R_t)\log\{1-\widehat{p}_{\boldsymbol{\theta},\boldsymbol{\phi},A_t}(G_t)\}. \tag{7}$$

When $R_t = 0$, this loss increases the total probability of the other classes without assigning an unobserved label. Both the encoder and head are trained from this same objective.

For a given graph, write the class-versus-rest log odds and its head derivative as

$$\nu_k = \log\frac{\widehat{p}_k}{1-\widehat{p}_k}, \qquad \boldsymbol{d}_k = \nabla_{\boldsymbol{\theta}}\nu_k = \frac{\boldsymbol{Z}^\top(\boldsymbol{e}_k - \widehat{\boldsymbol{p}})}{1-\widehat{p}_k}. \tag{8}$$

Here $\boldsymbol{e}_k \in \mathbb{R}^K$ is the k-th coordinate vector, and the graph and encoder arguments are suppressed only within a fixed state. A subscript t means that the quantity is evaluated on G_t before assimilating R_t. At the pre-update parameter state, $\nabla_{\boldsymbol{\theta}}\ell_t^{\mathrm{PLL}} = (\widehat{p}_{t,A_t} - R_t)\boldsymbol{d}_{t,A_t}$. The head takes a covariance-preconditioned step, while scheduled encoder steps use the PLL gradient. Appendix D.1.2 gives the clipping, update schedule and exact numerical equations. Each round first selects a class and observes its bit, then assimilates that bit in the current coordinates. Any encoder change and transport follow this assimilation.

4.3 UNCERTAINTY AND BANDIT DECISIONS

We write $(\boldsymbol{\phi}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma})$ for the current encoder, head estimate and covariance. Within a round, $-$ denotes the state after assimilating the bit but before changing the encoder, and $+$ denotes its transported counterpart. These superscripts indicate update stages, not positive and negative matrix parts. The selected Bernoulli model contributes the Fisher information matrix:

$$\boldsymbol{F}_t = \widehat{p}_{t,A_t}(1-\widehat{p}_{t,A_t})\boldsymbol{d}_{t,A_t}\boldsymbol{d}_{t,A_t}^\top, \qquad \boldsymbol{V}^- \leftarrow \boldsymbol{V} + \boldsymbol{F}_t, \qquad \boldsymbol{\Sigma}^- = (\boldsymbol{V}^-)^{-1}. \tag{9}$$

Initialize $\boldsymbol{V} = \lambda_0 \boldsymbol{I}_d$ and $\boldsymbol{\Sigma} = \lambda_0^{-1} \boldsymbol{I}_d$, with $\lambda_0 > 0$. The inverse precision $\boldsymbol{\Sigma}$ estimates the uncertainty of the joint head. Its statistical coverage requires a separate assumption, since positive definiteness alone does not ensure calibration. Before feedback, define $s_{t,k} = (\boldsymbol{d}_{t,k}^\top \boldsymbol{\Sigma} \boldsymbol{d}_{t,k})^{1/2}$. We use two alternative exploration rules:

$$U_{t,k} = \mathrm{sigmoid}(\nu_{t,k} + b_t s_{t,k}), \qquad a_t^{\mathrm{UCB}} \in \arg\max_{k \in [K]} U_{t,k}, \tag{10}$$

$$\widetilde{\boldsymbol{\theta}}_t \sim N(\widehat{\boldsymbol{\theta}}, \alpha_t^2 \boldsymbol{\Sigma}), \qquad a_t^{\mathrm{TS}} \in \arg\max_{k \in [K]} [\boldsymbol{Z}(G_t) \widetilde{\boldsymbol{\theta}}_t]_k. \tag{11}$$

Here $\mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$ and $\boldsymbol{Z}(G_t) = \boldsymbol{Z}_{\boldsymbol{\phi}}(G_t)$ at the current encoder. The positive scalars b_t and α_t control the UCB width and TS sampling scale, respectively. UCB favors classes with a large optimistic score. TS draws one joint head and chooses its most probable class. The latter retains dependence among class logits. We call the resulting algorithms `OGC-GCN-UCB-MT` and `OGC-GCN-TS-MT`.

4.4 Moment transport

Updating $\boldsymbol{\phi}$ changes the features on which the head was trained. Suppose the old and new features satisfy $\boldsymbol{z}^+(G) = \boldsymbol{M} \boldsymbol{z}^-(G)$ for an invertible $\boldsymbol{M} \in \mathbb{R}^{m \times m}$. Using the update-stage convention above, preserving the logits requires $\boldsymbol{\Theta}^+ = \boldsymbol{\Theta}^- \boldsymbol{M}^{-1}$. In vector form, with $\boldsymbol{P} = \boldsymbol{M}^{-\top} \otimes \boldsymbol{I}_{K-1}$, moment transport is

$$\widehat{\boldsymbol{\theta}}^+ = \boldsymbol{P} \widehat{\boldsymbol{\theta}}^-, \qquad \boldsymbol{\Sigma}^+ = \boldsymbol{P} \boldsymbol{\Sigma}^- \boldsymbol{P}^\top, \qquad \boldsymbol{V}^+ = \boldsymbol{P}^{-\top} \boldsymbol{V}^- \boldsymbol{P}^{-1}. \tag{12}$$

Both the learned information and its prior or regularizer are transformed. Moving only the head mean leaves predictive uncertainty inconsistent with the new features.

A nonlinear GCN update need not admit one exact map. We estimate $\boldsymbol{M}$ from paired old and new features of previously observed anchor graphs:

$$\widehat{\boldsymbol{M}} \in \arg\min_{\boldsymbol{M}} \sum_{G \in \mathcal{B}_{\mathrm{anc}}} \|\boldsymbol{z}_{\boldsymbol{\phi}'}(G) - \boldsymbol{M} \boldsymbol{z}_{\boldsymbol{\phi}}(G)\|_2^2 + \lambda_{\mathrm{map}} \|\boldsymbol{M} - \boldsymbol{I}\|_F^2. \tag{13}$$

Here $\mathcal{B}_{\mathrm{anc}}$ contains at most m_{anc} previously observed graphs, $\lambda_{\mathrm{map}} > 0$ penalizes deviation from the identity, and $\|\cdot\|_F$ is the Frobenius norm. The candidate encoder is $\boldsymbol{\phi}'$. We accept a finite invertible map when $\kappa_2(\widehat{\boldsymbol{M}}) = \sigma_{\max}(\widehat{\boldsymbol{M}}) / \sigma_{\min}(\widehat{\boldsymbol{M}}) \le \kappa_{\max}$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values and $\kappa_{\max} \ge 1$ is a prespecified threshold. Otherwise the previous encoder and assimilated head state are retained. Algorithm 1 gives both methods. The positive encoder learning rate λ_{enc}, integer update period $\tau \ge 1$, anchor capacity and acceptance parameters are fixed before each run. Appendix D.1.2 gives the experimental values. The normalization c_t in Algorithm 1 clips the head step in its local information norm, as derived in Appendix B.2.

Algorithm 1 `OGC-GCN-UCB-MT` or `OGC-GCN-TS-MT`

Require: Initial $(\boldsymbol{\phi}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma} \succ 0)$; rule rule $\in \{\mathrm{UCB}, \mathrm{TS}\}$.
Require: Scales $\{b_t, \alpha_t\}$; $\lambda_{\mathrm{enc}}, \tau, m_{\mathrm{anc}}, \lambda_{\mathrm{map}}, \kappa_{\max}$.
$\mathcal{B}_{\mathrm{anc}} \leftarrow \varnothing$.
for $t = 1, \dots, T$ **do**
  Observe G_t; compute $(\boldsymbol{z}_t, \widehat{\boldsymbol{p}}_t, \nu_{t,k}, \boldsymbol{d}_{t,k})$ by (5)–(8).
  $A_t \leftarrow a_t^{\mathrm{rule}}$ using (10) or (11); observe $R_t = \mathbf{1}\{A_t = Y_t\}$.
  $\boldsymbol{u}_\phi \leftarrow \nabla_{\boldsymbol{\phi}} \ell_t^{\mathrm{PLL}}$ at the current parameters; $e_t \leftarrow R_t - \widehat{p}_{t,A_t}$.
  $\boldsymbol{\Sigma}^- \leftarrow (\boldsymbol{\Sigma}^{-1} + \boldsymbol{F}_t)^{-1}$; $c_t \leftarrow \max\{1, |e_t| \sqrt{\boldsymbol{d}_{t,A_t}^\top \boldsymbol{\Sigma}^- \boldsymbol{d}_{t,A_t}}\}$.
  $\widehat{\boldsymbol{\theta}}^- \leftarrow \widehat{\boldsymbol{\theta}} + \boldsymbol{\Sigma}^- \boldsymbol{d}_{t,A_t} e_t / c_t$.
  $(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}) \leftarrow (\widehat{\boldsymbol{\theta}}^-, \boldsymbol{\Sigma}^-)$; append G_t to $\mathcal{B}_{\mathrm{anc}}$, retaining the last m_{anc} graphs.
  if $t \bmod \tau = 0$ **then**
    $\boldsymbol{\phi}' \leftarrow \boldsymbol{\phi} - \lambda_{\mathrm{enc}} \boldsymbol{u}_\phi / \max\{1, \|\boldsymbol{u}_\phi\|_2\}$; fit $\widehat{\boldsymbol{M}}$ by (13).
    if $\widehat{\boldsymbol{M}}$ is finite, $\det \widehat{\boldsymbol{M}} \neq 0$, and $\kappa_2(\widehat{\boldsymbol{M}}) \le \kappa_{\max}$ **then**
      $\boldsymbol{P} \leftarrow \widehat{\boldsymbol{M}}^{-\top} \otimes \boldsymbol{I}_{K-1}$; $(\boldsymbol{\phi}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}) \leftarrow (\boldsymbol{\phi}', \boldsymbol{P} \widehat{\boldsymbol{\theta}}^-, \boldsymbol{P} \boldsymbol{\Sigma}^- \boldsymbol{P}^\top)$.
    end if
  end if
end for

We abbreviate `OGC-GCN-UCB-MT` and `OGC-GCN-TS-MT` as `UCB-MT` and `TS-MT` in experimental descriptions, figures and tables. The stale, replay and variance-fitting controls differ only in how they update uncertainty after an encoder change.

5 Theoretical Analysis

Our analysis first establishes exact decision invariance and then bounds the effect of approximate transport in terms of decision margins. We next express accumulated information in transported coordinates and use it to derive conditional regret bounds. For $\boldsymbol{V} \succ 0$, write $\|\boldsymbol{v}\|_{\boldsymbol{V}} = (\boldsymbol{v}^\top \boldsymbol{V} \boldsymbol{v})^{1/2}$. Proofs are given in Appendix C.

5.1 Exact transport and sequential equivalence

Condition on completed feedback, so the old and new feature maps are fixed.

Theorem 1 (Exact moment and decision invariance)**.** *Suppose $\boldsymbol{\Sigma}^- \succ 0$ and $\boldsymbol{z}^+(G) = \boldsymbol{M}\boldsymbol{z}^-(G)$ for every relevant G, with $\det \boldsymbol{M} \neq 0$. Under (12), common b_t, α_t and tie-breaking, for every graph and class,*

$$\boldsymbol{g}^+(G) = \boldsymbol{g}^-(G), \qquad \widehat{\boldsymbol{p}}^+(G) = \widehat{\boldsymbol{p}}^-(G),$$
$$(\boldsymbol{d}_k^+)^\top \boldsymbol{\Sigma}^+ \boldsymbol{d}_k^+ = (\boldsymbol{d}_k^-)^\top \boldsymbol{\Sigma}^- \boldsymbol{d}_k^-, \qquad U_k^+ = U_k^-.$$

Moreover, $\|\boldsymbol{P}\boldsymbol{e}\|_{\boldsymbol{V}^+} = \|\boldsymbol{e}\|_{\boldsymbol{V}^-}$ for every $\boldsymbol{e} \in \mathbb{R}^d$. Coupling $\widetilde{\boldsymbol{\theta}}^+ = \boldsymbol{P}\widetilde{\boldsymbol{\theta}}^-$ gives identical TS actions. UCB actions also agree. These identities hold pathwise for maps selected from completed feedback.

Corollary 2 (Sequential equivalence with equivariant updates)**.** *In addition to exact transport, transform the regularizer, optimizer and retained statistics consistently. Let $\mathcal{T}_{\boldsymbol{P}}$ transport this complete state S and let $\mathcal{U}^\pm$ be the subsequent learning updates. If*

$$\mathcal{U}^+(\mathcal{T}_{\boldsymbol{P}}(\mathsf{S}), \boldsymbol{z}^+, A, R) = \mathcal{T}_{\boldsymbol{P}}(\mathcal{U}^-(\mathsf{S}, \boldsymbol{z}^-, A, R)),$$

then coupled environment and action randomness give $(A_t^+, R_t^+) = (A_t^-, R_t^-)$ for every t and $\mathcal{R}_T^+ = \mathcal{R}_T^-$. Arbitrary Euclidean SGD after a nonorthogonal map need not satisfy this condition.

5.2 Why transporting only the mean can fail

The following linear-bandit example isolates the effect of stale uncertainty. Its arm means need not sum to one, so the result is not a lower bound for the multinomial GCN model.

Theorem 3 (Failure of stale precision)**.** *For every integer $n \geq 1$, there is a two-dimensional, two-arm linear bandit with $T = 2n$, deterministic rewards $\mu_{t,a} \in \{0, 1\}$ and one orthogonal feature change such that a UCB learner transporting only its center incurs*

$$\mathcal{R}_{2n}^{\mathrm{stale}} = \sum_{t=1}^{2n} [\max_a \mu_{t,a} - \mu_{t,A_t^{\mathrm{stale}}}] = n = T/2.$$

Exact precision transport instead selects an optimal arm at round $n+1$, immediately after the change.

5.3 Approximate transport and decision margins

For $\boldsymbol{v} \in \mathbb{R}^K$, let $\Gamma(\boldsymbol{v}) = v_{(1)} - v_{(2)}$ be the gap between its two largest entries. Approximate transport is useful when its perturbation is small relative to this gap.

Proposition 4 (Margin-sensitive decision stability). *Fix a graph and completed history. Couple TS logits as* $\boldsymbol{g}^{\pm} = \boldsymbol{m}^{\pm} + \boldsymbol{B}^{\pm}\boldsymbol{\xi}$, *with* $\boldsymbol{m}^{\pm} \in \mathbb{R}^K$, $\boldsymbol{B}^{\pm} \in \mathbb{R}^{K\times r}$, $r \geq 1$ *and* $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{I}_r)$. *The factors include the TS scale. For* $0 < \delta < 1$, *define*

$$e = \|\boldsymbol{m}^+ - \boldsymbol{m}^-\|_\infty, \quad d_B = \max_k \|(\boldsymbol{B}^+ - \boldsymbol{B}^-)_{k,:}\|_2, \quad \epsilon_\delta = e + d_B\sqrt{2\log(2K/\delta)}.$$

With common tie-breaking and $a^{\pm} \in \arg\max_k g_k^{\pm}$,

$$\mathbb{P}(a^+ \neq a^-) \leq \mathbb{P}\{\Gamma(\boldsymbol{g}^-) \leq 2\epsilon_\delta\} + \delta, \tag{14}$$

$$|\mathbb{E}[\mu_{a^-} - \mu_{a^+}]| \leq \mathbb{P}(a^+ \neq a^-), \qquad \boldsymbol{\mu} \in [0,1]^K. \tag{15}$$

For deterministic UCB score vectors $\boldsymbol{U}^{\pm}$, *setting* $e_U = \|\boldsymbol{U}^+ - \boldsymbol{U}^-\|_\infty$ *gives* $\Gamma(\boldsymbol{U}^-) > 2e_U \Rightarrow \arg\max_k U_k^+ = \arg\max_k U_k^-$.

For MT, choose $\boldsymbol{B}^- = \alpha_t \boldsymbol{Z}^- \boldsymbol{S}$, $\boldsymbol{B}^+ = \alpha_t \boldsymbol{Z}^+ \boldsymbol{P}\boldsymbol{S}$ and $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{\Sigma}^-$. With $\boldsymbol{E}(G) = \boldsymbol{Z}^+(G)\boldsymbol{P} - \boldsymbol{Z}^-(G)$, the transported mean gives $e = \|\boldsymbol{E}(G)\widehat{\boldsymbol{\theta}}^-\|_\infty$ and $d_B = \alpha_t \max_k \|[\boldsymbol{E}(G)\boldsymbol{S}]_{k,:}\|_2$. Thus feature-map residuals control decision perturbations. Preserving a decision does not imply its optimality or the calibration required below.

5.4 Transported information and regret

Let $\mathcal{H}_t = \sigma(\mathcal{F}_{t-1}, G_t)$ be the information before action randomization. Fitted means, features and widths are evaluated at this stage. Define

$$w_t = \widehat{p}_{t,A_t}(1 - \widehat{p}_{t,A_t}), \qquad \mathcal{J}_T = \sum_{t=1}^{T} \log(1 + w_t s_{t,A_t}^2). \tag{16}$$

Write $\boldsymbol{V}_t = \boldsymbol{\Sigma}_t^{-1}$ and let $\boldsymbol{P}_t$ be the invertible head map after round t's assimilation, or $\boldsymbol{I}_d$ if no map is accepted. Set $\boldsymbol{Q}_1 = \boldsymbol{I}_d$ and $\boldsymbol{Q}_{t+1} = \boldsymbol{P}_t \boldsymbol{Q}_t$.

Lemma 5 (Transported information identity). *Suppose* $\boldsymbol{V}_1 = \lambda_0 \boldsymbol{I}_d$ *and* $\boldsymbol{V}_{t+1} = \boldsymbol{P}_t^{-\top}(\boldsymbol{V}_t + w_t \boldsymbol{d}_{t,A_t}\boldsymbol{d}_{t,A_t}^\top)\boldsymbol{P}_t^{-1}$. *For* $\boldsymbol{x}_t = \sqrt{w_t}\boldsymbol{Q}_t^\top \boldsymbol{d}_{t,A_t}$ *and* $\overline{\boldsymbol{V}}_t = \boldsymbol{Q}_t^\top \boldsymbol{V}_t \boldsymbol{Q}_t$, *every realized trajectory satisfies*

$$\overline{\boldsymbol{V}}_{T+1} = \lambda_0 \boldsymbol{I}_d + \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top,$$

$$\mathcal{J}_T = \log \frac{\det \overline{\boldsymbol{V}}_{T+1}}{\det(\lambda_0 \boldsymbol{I}_d)}, \qquad \sum_{t=1}^{T} \min\{1, w_t s_{t,A_t}^2\} \leq 2\mathcal{J}_T. \tag{17}$$

If $\|\boldsymbol{x}_t\|_2 \leq L_x$ *almost surely, then* $\mathcal{J}_T \leq \overline{J}_T := d\log(1 + TL_x^2/(d\lambda_0))$.

Adaptive congruences cancel before applying the elliptical-potential argument (Abbasi-Yadkori et al., 2011). The identity holds for fitted maps without exact nonlinear feature matching. A uniform L_x requires an additional growth condition, since bounding each map's condition number does not bound their product. Appendix A.2 discusses this distinction and the role of the regularizer.

For both regret bounds, fix $0 < \delta < 1$ and assume $w_t \geq \omega > 0$ and $\mathcal{J}_T \leq \overline{J}_T$ almost surely for a deterministic $\overline{J}_T$.

Theorem 6 (UCB Bayes pseudo-regret under calibration). *Assume $0 < \mu_k(G_t) < 1$, deterministic $\eta_t \geq 0$ and $b_t \leq \bar{b}$. Suppose, with probability at least $1-\delta$, simultaneously for all $t \leq T$ and $k \in [K]$,*

$$\left| \nu_{t,k} - \log \frac{\mu_k(G_t)}{1-\mu_k(G_t)} \right| \leq b_t s_{t,k} + \eta_t. \tag{18}$$

On this event, Algorithm 1 with UCB satisfies

$$\mathcal{R}_T \leq C_{\mathrm{U}} \sqrt{2T\overline{J}_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t, \qquad C_{\mathrm{U}} = \max\{1, \bar{b}/(2\sqrt{\omega})\}. \tag{19}$$

The expectation bound adds δT.

For TS, write $\widetilde{\boldsymbol{p}}_t = \mathrm{softmax}(\boldsymbol{Z}(G_t)\widetilde{\boldsymbol{\theta}}_t)$ for the joint draw in (11), and let $k_t^\star$ be a Bayes-optimal class with fixed tie-breaking.

Theorem 7 (TS Bayes pseudo-regret under sampling conditions). *Let $\mathcal{C}_t \in \mathcal{H}_t$ satisfy $\mathbb{P}(\bigcap_{t\leq T} \mathcal{C}_t) \geq 1-\delta$. On every history in $\mathcal{C}_t$, assume*

$$\mathbb{P}\big(\forall k:\ |\widetilde{p}_{t,k} - \mu_k(G_t)| \leq c_t^{\mathrm{TS}} s_{t,k} + e_t^{\mathrm{TS}} \mid \mathcal{H}_t\big) \geq 1-\zeta_t, \tag{20}$$

$$\mathbb{P}(\widetilde{p}_{t,k_t^\star} \geq \mu_t^\star \mid \mathcal{H}_t) \geq p_0. \tag{21}$$

Here $c_t^{\mathrm{TS}} \leq \bar{c}^{\mathrm{TS}}$, $e_t^{\mathrm{TS}} \geq 0$ and $0 \leq \zeta_t \leq p_0/2$ are deterministic, and $p_0 > 0$. Algorithm 1 with TS satisfies

$$\mathbb{E}[\mathcal{R}_T] \leq \left(1 + \frac{6}{p_0}\right)\left[C_{\mathrm{TS}}\sqrt{2T\overline{J}_T} + \sum_{t=1}^{T} e_t^{\mathrm{TS}}\right] + \sum_{t=1}^{T}\zeta_t + \delta T, \tag{22}$$

where $C_{\mathrm{TS}} = \max\{1, \bar{c}^{\mathrm{TS}}/\sqrt{\omega}\}$.

The TS proof controls selected widths on the learner's actual history using optimism (Agrawal & Goyal, 2013). For our Gaussian head, Lemma 10 supplies $c_t^{\mathrm{TS}} = (b_t + a_t)/4$ and $e_t^{\mathrm{TS}} = (\eta_t + \rho_t)/4$, where a_t is the Gaussian width and ρ_t the nonlinear log-odds remainder defined in Appendix A.3. The theorem uses deterministic upper envelopes when these quantities depend on history. The lemma also gives an explicit sufficient lower bound for p_0. Positive covariance alone does not imply optimism.

Rates and classification mistakes. If $\overline{J}_T = \mathcal{O}(d\log(1+T))$, ω, p_0 are bounded away from zero and $\bar{b}, \bar{c}^{\mathrm{TS}}$ grow at most polylogarithmically, the information terms are $\widetilde{\mathcal{O}}(\sqrt{dT})$. Sublinear regret further requires $\sum_t \eta_t = o(T)$ for UCB, $\sum_t e_t^{\mathrm{TS}} + \sum_t \zeta_t = o(T)$ for TS, and $\delta = o(1)$ for expectations. The regret guarantees therefore require calibration, sampling and growth conditions beyond the Fisher update and map acceptance test.

Proposition 8 (Mistakes, Bayes loss and feedback noise). *Assume $\mathbb{E}[R_t \mid \mathcal{A}_t] = \mu_{A_t}(G_t)$, where $\mathcal{A}_t$ is the information immediately before feedback, including A_t. With $\Delta_T^{\mathrm{noise}} = \sum_t (\mu_{A_t}(G_t) - R_t)$ and $0 < \delta < 1$,*

$$M_T = N_T^\star + \mathcal{R}_T + \Delta_T^{\mathrm{noise}}, \qquad \mathbb{E}[\Delta_T^{\mathrm{noise}}] = 0, \qquad \mathbb{P}\left\{|\Delta_T^{\mathrm{noise}}| > \sqrt{\tfrac{T}{2}\log\tfrac{2}{\delta}}\right\} \leq \delta.$$

Thus the regret bounds control excess mistakes above the unavoidable Bayes loss, connecting the synthetic regret and real-data mistake metrics used next.

6 Experiments

We evaluate the complete methods on real graph datasets and synthetic graphs with known class probabilities. Online mistakes measure classification performance, while synthetic Bayes pseudo-regret

separates decision error from irreducible label noise. Every learner observes only the correctness of its selected class.

6.1 Online classification performance and computation

We use ENZYMES, IMDB-MULTI, COLLAB and REDDIT-MULTI-5K (Morris et al., 2020). Each of twenty seeds defines a stratified 70/15/15 training/validation/test split and a shared random training order. Each training graph is classified once. Learned methods use a two-layer GCN with hidden widths 16 and 12 and the same encoder-update schedule. Encoder learning rates are selected on three development seeds. We compare the proposed methods with graph adaptations of `FastCB`, `SquareCB`, `LiM2` and neural-linear UCB/TS. Appendix D.1 describes the selected-bit oracle adaptations. The guarantees of the original methods are not assumed to hold for the GCN optimizer.

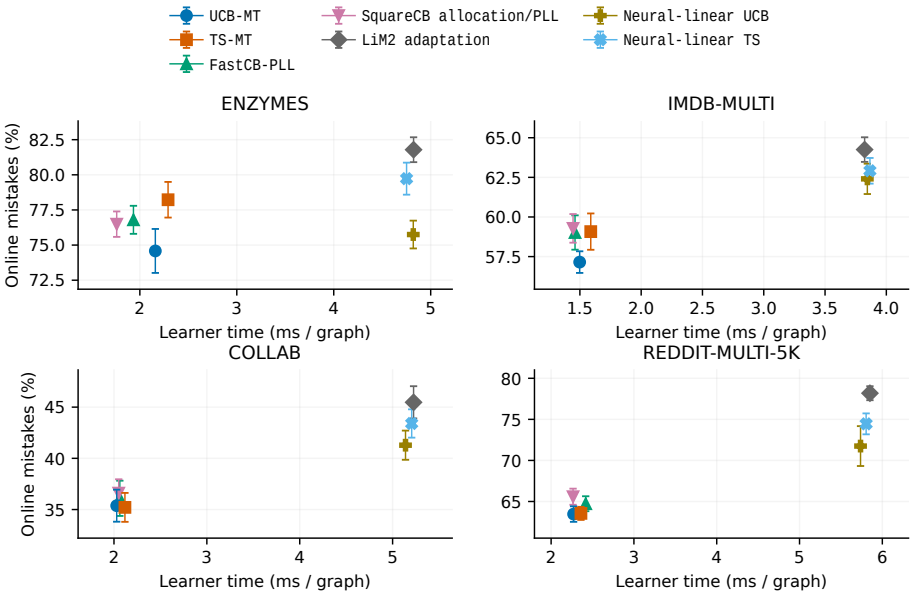

Figure 1: Classification performance and computational cost. Each panel shows one dataset. Lower values indicate fewer mistakes, and leftward positions indicate less learner time. Points are means over twenty seeds. Vertical bars are pointwise 95% t intervals for mean mistake rate. Horizontal coordinates give mean learner time per training round. The intervals are descriptive and do not represent paired tests. `LiM2` and neural-linear methods are the graph adaptations described in Appendix D.1.

In Figure 1, `UCB-MT` has a lower mean mistake rate than each displayed external adaptation on all four datasets. `TS-MT` has the lowest displayed mean on COLLAB and performs similarly to `UCB-MT` on REDDIT. Both proposed methods require less learner time than the graph `LiM2` and neural-linear adaptations. `FastCB-PLL` and `SquareCB allocation/PLL` have comparable computational costs. Timing includes prediction, encoder training, head assimilation and uncertainty updates. These rankings compare means rather than establish statistical significance. Table 4 and Appendix D.2 report exact values and results for all nineteen methods.

6.2 Bayes pseudo-regret of the complete methods

We compare the seven complete methods on dense and sparse role-SBM graphs with three classes and a fixed topology-based teacher. The evaluator knows $\boldsymbol{\mu}(G_t)$, while each learner observes only G_t and the correctness of its selected class. All methods use GCN widths $(8, 6)$, update period 64 and paired streams. Each is tuned over fifteen encoder-rate and exploration configurations on two development seeds. The selected settings are then evaluated on twenty independent seeds, each with $T = 4096$

rounds per family. We average both families within each seed before inference. Appendix D.3 gives the protocol, selected settings and ten prespecified paired comparisons.

Table 1: Cumulative Bayes pseudo-regret at $T = 4096$, reported as mean $\pm$ standard deviation over twenty independent evaluation seeds. The Average column combines the dense and sparse families within each seed. Lower values are better, and bold marks the lowest mean in each column. All seven methods receive equal tuning budgets.

Method	Dense	Sparse	Average
`UCB-MT`	**330.3** $\pm$ 201.1	**440.7** $\pm$ 280.2	**385.5** $\pm$ 233.4
`TS-MT`	723.2 $\pm$ 370.8	746.3 $\pm$ 311.8	734.7 $\pm$ 322.3
`FastCB-PLL`	742.4 $\pm$ 377.6	789.7 $\pm$ 330.2	766.0 $\pm$ 328.5
`SquareCB allocation/PLL`	731.1 $\pm$ 367.0	794.7 $\pm$ 331.5	762.9 $\pm$ 320.8
`Graph LiM2 adaptation`	1779.5 $\pm$ 526.7	1589.4 $\pm$ 440.4	1684.4 $\pm$ 453.8
`Graph neural-linear UCB`	1311.7 $\pm$ 354.2	1298.5 $\pm$ 344.4	1305.1 $\pm$ 313.9
`Graph neural-linear TS`	1554.8 $\pm$ 434.2	1413.5 $\pm$ 431.2	1484.1 $\pm$ 424.3

`UCB-MT` achieves the lowest mean pseudo-regret in both families: its pooled value is 385.5, compared with 766.0 for `FastCB-PLL` and 762.9 for `SquareCB allocation/PLL`, reductions of 49.7% and 49.5%. It also improves on `LiM2` and both neural-linear adaptations. All five UCB differences remain significant after the prespecified ten-test Holm correction ($p < 2 \times 10^{-5}$). `TS-MT` has pooled regret 734.7 and improves significantly on `LiM2` and both neural-linear adaptations, while its differences from `FastCB` and `SquareCB` are inconclusive (adjusted $p = 0.201$ for each). On this benchmark, `UCB-MT` improves on all five baselines, while the significant gains for `TS-MT` are limited to `LiM2` and the neural-linear methods.

The contribution of uncertainty transport. To isolate uncertainty transport, we compare MT, stale covariance, variance fitting and covariance replay with the same GCN–PLL architecture and update rules. At $T = 4096$, `UCB-MT` reduces mean Bayes pseudo-regret from 550.50 to 451.73 (17.9%) and `TS-MT` from 894.58 to 687.42 (23.2%) relative to stale covariance. Both paired differences remain significant after the six-test Holm correction ($p = 0.045$ and 0.00366). `UCB-MT` has the lowest mean among the four covariance methods, whereas refined variance fitting gives lower TS regret (548.72). Figure 4 and Appendix D.4 report the learning curves and all comparisons. These experiments evaluate the regret defined in Section 5.4 at a finite horizon. They do not establish the calibration, optimism or asymptotic conditions of the regret theorems.

AI USE STATEMENT

Generative AI tools assisted literature search and experiment design, and proof drafting, code implementation, synthetic-data generation, figure preparation, and manuscript revision. Verification included source checks, feedback-access audits, numerical checks and compilation. The authors are responsible for the final content and for reviewing all AI-assisted claims, code and results.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011. URL https://papers.nips.cc/paper/4417-improved-algorithms-for-linear-stochastic-bandits.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 127–135. PMLR, 2013. URL https://proceedings.mlr.press/v28/agrawal13.html.

Pierre Clavier, Tom Huix, and Alain Oliviero Durmus. VITS: Variational inference thompson sampling for contextual bandits. In *International Conference on Machine Learning*, volume 235,

pp. 9033–9075, 2024. URL https://proceedings.mlr.press/v235/clavier24a.html.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5ep85sakT3.

Wanjin Feng, Xingyu Gao, Peilin Zhao, and Steven C. H. Hoi. A unified framework for bandit online multiclass prediction. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2200–2211, 2025. doi: 10.1109/TKDE.2024.3458419. URL https://ieeexplore.ieee.org/document/10689269/.

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2107.02237.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 440–447, 2008. doi: 10.1145/1390156.1390212.

Parnian Kassraie, Andreas Krause, and Ilija Bogunovic. Graph neural network bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34519–34531, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate thompson sampling with langevin algorithms. In *International Conference on Machine Learning*, volume 119, pp. 6797–6807, 2020. URL https://proceedings.mlr.press/v119/mazumdar20a.html.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

Ofir Nabati, Tom Zahavy, and Shie Mannor. Online limited memory neural-linear bandits with likelihood matching. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7905–7915. PMLR, 2021. URL https://proceedings.mlr.press/v139/nabati21a.html.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.03271.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate thompson sampling via epistemic neural networks. In *Uncertainty in Artificial Intelligence*, volume 216, pp. 1586–1595, 2023. URL https://proceedings.mlr.press/v216/osband23a.html.

My Phan, Yasin Abbasi-Yadkori, and Justin Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, 2019. URL https://papers.neurips.cc/paper/9084-thompson-sampling-and-approximate-inference.pdf.

Yunzhe Qi, Yikun Ban, and Jingrui He. Graph neural bandits. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1920–1931. ACM, 2023. doi: 10.1145/3580305.3599371. URL https://doi.org/10.1145/3580305.3599371.

Sudeep Salgia. Provably and practically efficient neural contextual bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29800–29844. PMLR, 2023. URL https://proceedings.mlr.press/v202/salgia23a.html.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. On the complexity of representation learning in contextual linear bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7871–7896, 2023. URL https://proceedings.mlr.press/v206/tirinzoni23a.html.

Dirk van der Hoeven. Exploiting the surrogate gap in online multiclass classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9562–9572, 2020.

Shuang Wu and Arash A. Amini. Graph neural thompson sampling. *Reinforcement Learning Journal*, 1:29–63, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper12.html.

Shuang Wu and Arash A. Amini. Laplacian kernelized bandit. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5majEbTAZa.

Shuang Wu, Chi-Hua Wang, Yuantong Li, and Guang Cheng. Residual bootstrap exploration for stochastic linear bandit. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2117–2127. PMLR, 2022. URL https://proceedings.mlr.press/v180/wu22a.html.

Tom Zahavy and Shie Mannor. Deep neural linear bandits: Overcoming catastrophic forgetting through likelihood matching. *arXiv preprint arXiv:1901.08612*, 2019.

Mengxiao Zhang, Yuheng Zhang, Olga Vrousgou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/618c95f4557c15b253fb0e6f548ea0c0-Abstract-Conference.html.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tkAtoZkcUnm.

Haoyang Zheng, Wei Deng, Christian Moya, and Guang Lin. Accelerating approximate thompson sampling with underdamped langevin monte carlo. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 2611–2619, 2024. URL https://proceedings.mlr.press/v238/zheng24b.html.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11492–11502, 2020. URL https://proceedings.mlr.press/v119/zhou20a.html.

A Further Theoretical Results

This appendix gives the predictable-feature condition, discusses the assumptions needed to control information growth, and derives sufficient Gaussian sampling conditions for Section 5. Proofs are provided in Appendix C.

A.1 Feedback chronology and predictable features

The learner's completed history is $\mathcal{F}_{t-1}$. Define the information available before the graph-dependent decision and before feedback by

$$\mathcal{H}_t = \sigma(\mathcal{F}_{t-1}, G_t), \qquad \mathcal{A}_t = \sigma(\mathcal{H}_t, \xi_t^{\mathrm{alg}}, A_t).$$

where ξ_t^{alg} contains the current learner randomness. This randomness uses no additional information about the hidden label. In particular, $\mathbb{E}[R_t \mid \mathcal{A}_t] = \mu_{A_t}(G_t)$ under the stable-law assumptions.

Proposition 9 (Predictable-feature score condition)**.** *If $\boldsymbol{x}_t \in \mathbb{R}^d$ is $\mathcal{A}_t$-measurable with $\mathbb{E}[\|\boldsymbol{x}_t\|_2] < \infty$ and $\varepsilon_t = R_t - \mathbb{E}[R_t \mid \mathcal{A}_t]$, then*

$$\mathbb{E}[\boldsymbol{x}_t \varepsilon_t \mid \mathcal{A}_t] = \mathbf{0}.$$

The conclusion need not hold for a feature recomputed by an encoder trained on the same R_t.

Proof: Appendix C.8.

A.2 Information growth under transported precision

The canonical coordinates in Lemma 5 follow the actual sequence of fitted invertible maps, whether or not a nonlinear encoder admits exact feature matching.

The determinant argument is the usual elliptical-potential calculation (Abbasi-Yadkori et al., 2011), expressed in transported coordinates. It holds for adaptive, invertible fitted maps because the congruences cancel algebraically. However, bounding each map's condition number does not bound $\|\boldsymbol{Q}_t\|_{\mathrm{op}}$ over many updates. A constant L_x is an additional growth condition, not an implication of Algorithm 1's acceptance test. Resetting the regularizer or truncating the Fisher history changes the recursion and requires a separate bound. Under an exact feature change, $\boldsymbol{d}^+ = \boldsymbol{P}^{-\top}\boldsymbol{d}^-$ preserves both the width and its canonical feature.

A.3 Sufficient sampling conditions for the joint Gaussian head

The TS assumptions involve the nonlinear probabilities of one joint head draw, not independent classwise samples. The following calculation relates them to the fitted covariance without assuming that it is a true posterior. Condition on $\mathcal{H}_t$, factor $\boldsymbol{\Sigma}_t = \boldsymbol{S}_t\boldsymbol{S}_t^\top$, and let $\boldsymbol{Z}_{t,k,:}$ be the kth row of $\boldsymbol{Z}(G_t)$. Define

$$D_t^{\mathrm{logit}} = \max_{i,j} \|(\boldsymbol{Z}_{t,i,:} - \boldsymbol{Z}_{t,j,:})\boldsymbol{S}_t\|_2, \quad a_t = \alpha_t\sqrt{2\log(4K/\zeta_t)}, \quad \rho_t = \frac{\alpha_t^2 (D_t^{\mathrm{logit}})^2 d}{4}\log(4d/\zeta_t),$$

for $0 < \zeta_t < 1$. The scalar D_t^{logit} is a logit-difference scale, distinct from the graph degree matrix $\boldsymbol{D}$. Write $\widetilde{\nu}_{t,k} = \log[\widetilde{p}_{t,k}/(1-\widetilde{p}_{t,k})]$, and let Φ denote the standard normal distribution function.

Lemma 10 (Gaussian concentration and an optimism condition)**.** *For the joint draw in* (11)*,*

$$\mathbb{P}\left(\forall k:\ |\widetilde{\nu}_{t,k} - \nu_{t,k}| \le a_t s_{t,k} + \rho_t \mid \mathcal{H}_t\right) \ge 1 - \zeta_t.$$

On a history satisfying (18)*,* (20) *holds with*

$$c_t^{\mathrm{TS}} = (b_t + a_t)/4, \qquad e_t^{\mathrm{TS}} = (\eta_t + \rho_t)/4.$$

If $s_{t,k_t^\star} > 0$, its conditional optimism probability is at least

$$1 - \Phi\left(\frac{b_t s_{t,k_t^\star} + \eta_t + \rho_t}{\alpha_t s_{t,k_t^\star}}\right) - \frac{\zeta_t}{2}.$$

Any uniform positive lower bound on this expression supplies p_0 in (21)*. For Theorem 7, deterministic upper envelopes for $c_t^{\mathrm{TS}}, e_t^{\mathrm{TS}}$ may replace these history-dependent quantities.*

Proof: Appendix C.4.

This sufficient condition may be conservative. A nondegenerate Gaussian draw alone does not give a uniform p_0 relative to the unknown Bayes mean, nor does it make $\sum_t \rho_t$ sublinear. With ω, p_0 fixed, $\overline{b}, \overline{c}^{\mathrm{TS}}$ polylogarithmic, $\overline{J}_T = \mathcal{O}(d \log(1+T))$, $\delta = T^{-2}$, and summable ζ_t, the remaining rate requirements are $\sum_t \eta_t = o(T)$ and $\sum_t e_t^{\mathrm{TS}} = o(T)$. These conditions concern calibration and nonlinear sampling and require justification beyond the PLL and Fisher update formulas. The main experiments measure pseudo-regret independently of these sufficient conditions.

B Additional Algorithmic Details

B.1 Notation reference

Vectors and matrices are bold. The table lists dimensions for the main method. Additional experimental quantities are defined where they are used. Transpose is $\top$, vec stacks columns, and $\otimes$ is the Kronecker product. For a matrix $\boldsymbol{B}$, $\|\boldsymbol{B}\|_F$ and $\|\boldsymbol{B}\|_{\mathrm{op}}$ are its Frobenius and spectral norms. For symmetric matrices, $\boldsymbol{B} \preceq \boldsymbol{C}$ means $\boldsymbol{C} - \boldsymbol{B}$ is positive semidefinite. The functions diag and Diag extract a matrix diagonal and form a diagonal matrix from a vector, respectively. Scalar indicators are written $\mathbf{1}\{\cdot\}$. Expectations use square brackets, including conditional expectations.

Table 2: Main notation. The feature dimension includes the intercept.

Symbol	Meaning
$G = (\boldsymbol{A}, \boldsymbol{X}), n, d_x$	Graph, $n \times n$ adjacency, $n \times d_x$ node features
K, T, A_t, Y_t, R_t	Number of classes, horizon, selected class, hidden label, correctness bit
$\boldsymbol{\mu}(G), \widehat{\boldsymbol{p}}_t$	True and estimated class-probability vectors in $\mathbb{R}^K$
$M_T, N_T^\star, \mathcal{R}_T$	Cumulative mistakes, conditional Bayes loss, Bayes pseudo-regret
$\boldsymbol{\phi}, \boldsymbol{z}_{\boldsymbol{\phi}}(G), m$	Encoder parameters and m-dimensional graph representation
$\boldsymbol{C}, \boldsymbol{\Theta}, \boldsymbol{\theta}, d$	$K \times (K-1)$ contrast basis; $(K-1) \times m$ head; vectorized head; $d = (K-1)m$
$\boldsymbol{Z}_{\boldsymbol{\phi}}(G), \boldsymbol{g}$	$K \times d$ logit design and K-dimensional logits
$\nu_{t,k}, \boldsymbol{d}_{t,k}$	Class-versus-rest log odds and its d-dimensional head derivative
$\boldsymbol{F}_t, \boldsymbol{V}, \boldsymbol{\Sigma}$	$d \times d$ fitted Fisher contribution, precision, inverse precision
$s_{t,k}, b_t, \alpha_t$	Log-odds uncertainty, `UCB` multiplier, `TS` sampling scale
$\boldsymbol{M}, \boldsymbol{P}$	$m \times m$ feature map and $d \times d$ head transport map
$\mathcal{B}_{\mathrm{anc}}, m_{\mathrm{anc}}$	Observed anchor graphs and buffer capacity
$\lambda_0, \lambda_{\mathrm{map}}$	Initial precision scale and identity-map regularization
$\lambda_{\mathrm{enc}}, \tau, \kappa_{\max}$	Encoder rate, update period, map condition threshold
$\Gamma(\boldsymbol{v}), e, d_B, \epsilon_\delta$	Decision margin and perturbation quantities in Proposition 4

B.2 PLL DERIVATIVES AND ONLINE ASSIMILATION

Fix the encoder and a graph, and write $\widehat{\boldsymbol{p}} = \mathrm{softmax}(\boldsymbol{Z}\boldsymbol{\theta})$. For the selected class a, the softmax derivative and log-odds derivative are

$$\nabla_{\boldsymbol{\theta}} \widehat{p}_a = \widehat{p}_a \boldsymbol{Z}^\top (\boldsymbol{e}_a - \widehat{\boldsymbol{p}}),$$
$$\nabla_{\boldsymbol{\theta}} \nu_a = \left(\widehat{p}_a^{-1} + (1-\widehat{p}_a)^{-1}\right) \nabla_{\boldsymbol{\theta}} \widehat{p}_a = \frac{\boldsymbol{Z}^\top (\boldsymbol{e}_a - \widehat{\boldsymbol{p}})}{1-\widehat{p}_a} =: \boldsymbol{d}_a.$$

Writing $\widehat{p}_a = \mathrm{sigmoid}(\nu_a)$ gives the Bernoulli loss and its derivatives:

$$\ell^{\mathrm{PLL}} = \log(1+e^{\nu_a}) - R\nu_a,$$
$$\nabla_{\boldsymbol{\theta}} \ell^{\mathrm{PLL}} = (\widehat{p}_a - R)\boldsymbol{d}_a,$$
$$\nabla^2_{\boldsymbol{\theta}} \ell^{\mathrm{PLL}} = \widehat{p}_a(1-\widehat{p}_a)\boldsymbol{d}_a\boldsymbol{d}_a^\top + (\widehat{p}_a - R)\nabla^2_{\boldsymbol{\theta}} \nu_a.$$

Under the *fitted* Bernoulli model $R \sim \mathrm{Bernoulli}(\widehat{p}_a)$,

$$\mathbb{E}_{R\sim\mathrm{Bernoulli}(\widehat{p}_a)}[\nabla^2_{\boldsymbol{\theta}} \ell^{\mathrm{PLL}}] = \widehat{p}_a(1-\widehat{p}_a)\boldsymbol{d}_a\boldsymbol{d}_a^\top = \boldsymbol{F}.$$

Thus $\boldsymbol{F}$ is the expected Hessian under the fitted Bernoulli model, which gives the update in (9). The realized Hessian can differ, and the fitted label distribution need not equal the true one.

Let $e_t = R_t - \widehat{p}_{t,A_t}$, $w_t = \widehat{p}_{t,A_t}(1-\widehat{p}_{t,A_t})$ and $\boldsymbol{v}_t = \boldsymbol{\Sigma}\boldsymbol{d}_{t,A_t}$. Sherman–Morrison yields

$$\boldsymbol{\Sigma}^- = (\boldsymbol{\Sigma}^{-1} + w_t \boldsymbol{d}_{t,A_t}\boldsymbol{d}_{t,A_t}^\top)^{-1} = \boldsymbol{\Sigma} - \frac{w_t \boldsymbol{v}_t \boldsymbol{v}_t^\top}{1 + w_t \boldsymbol{d}_{t,A_t}^\top \boldsymbol{v}_t}. \tag{23}$$

The implemented mean update is

$$c_t = \max\{1, |e_t|\sqrt{\boldsymbol{d}_{t,A_t}^\top \boldsymbol{\Sigma}^- \boldsymbol{d}_{t,A_t}}\}, \qquad \widehat{\boldsymbol{\theta}}^- = \widehat{\boldsymbol{\theta}} + \boldsymbol{\Sigma}^- \boldsymbol{d}_{t,A_t} e_t / c_t. \tag{24}$$

The encoder gradient is evaluated at the same pre-update parameters using the observed bit. For a scheduled encoder step,

$$\boldsymbol{u}_{\phi,t} = \nabla_{\boldsymbol{\phi}} \ell_t^{\mathrm{PLL}}, \qquad \boldsymbol{\phi}' = \boldsymbol{\phi} - \lambda_{\mathrm{enc}} \boldsymbol{u}_{\phi,t} / \max\{1, \|\boldsymbol{u}_{\phi,t}\|_2\}.$$

The bit is assimilated before the head is transported to $\boldsymbol{\phi}'$. Implementation parameters and baseline-specific updates are in Appendix D.1.2.

C PROOFS

Table 3: Proof locations for the theoretical results.

Statement	Proof
Theorem 6	Appendix C.2
Theorem 7	Appendix C.3
Lemma 5	Appendix C.1
Lemma 10	Appendix C.4
Theorem 1	Appendix C.5
Theorem 3	Appendix C.6
Proposition 4	Appendix C.7
Proposition 9	Appendix C.8
Proposition 8	Appendix C.9
Corollary 2	Appendix C.10

C.1 PROOF OF LEMMA 5

Proof. **Step 1: express precision in the initial coordinates.** Substituting $\boldsymbol{Q}_{t+1} = \boldsymbol{P}_t \boldsymbol{Q}_t$ into the precision recursion gives

$$\overline{\boldsymbol{V}}_{t+1} = \boldsymbol{Q}_t^\top \boldsymbol{P}_t^\top \boldsymbol{P}_t^{-\top} (\boldsymbol{V}_t + w_t \boldsymbol{d}_{t,A_t}\boldsymbol{d}_{t,A_t}^\top) \boldsymbol{P}_t^{-1} \boldsymbol{P}_t \boldsymbol{Q}_t$$
$$= \overline{\boldsymbol{V}}_t + \boldsymbol{x}_t \boldsymbol{x}_t^\top.$$

Induction from $\overline{\boldsymbol{V}}_1 = \lambda_0 \boldsymbol{I}_d$ proves the first identity. This cancellation is pathwise even when $\boldsymbol{P}_t$ is chosen using the completed feedback.

Step 2: preserve the information increment. Since $\overline{\boldsymbol{V}}_t^{-1} = \boldsymbol{Q}_t^{-1}\boldsymbol{V}_t^{-1}\boldsymbol{Q}_t^{-\top}$,

$$\boldsymbol{x}_t^\top \overline{\boldsymbol{V}}_t^{-1}\boldsymbol{x}_t = w_t \boldsymbol{d}_{t,A_t}^\top \boldsymbol{V}_t^{-1}\boldsymbol{d}_{t,A_t} = w_t s_{t,A_t}^2.$$

The matrix determinant lemma therefore yields

$$\frac{\det \overline{\boldsymbol{V}}_{t+1}}{\det \overline{\boldsymbol{V}}_t} = 1 + w_t s_{t,A_t}^2.$$

Take logarithms and sum to obtain the determinant identity for $\mathcal{J}_T$.

Step 3: bound the clipped widths. For $x \geq 0$, $\min\{1, x\} \leq 2\log(1+x)$. Applying this with $x = w_t s_{t,A_t}^2$ proves (17). Consequently, for every $c \geq 0$ and $w_t \geq \omega$,

$$\min\{1, c s_{t,A_t}\} \leq \max\{1, c/\sqrt{\omega}\}\sqrt{\min\{1, w_t s_{t,A_t}^2\}},$$

$$\sum_{t=1}^{T} \min\{1, c s_{t,A_t}\} \leq \max\{1, c/\sqrt{\omega}\}\sqrt{2T\mathcal{J}_T}. \tag{25}$$

The second inequality is Cauchy–Schwarz.

Step 4: bound the determinant by the trace. Under $\|\boldsymbol{x}_t\|_2 \leq L_x$,

$$\operatorname{tr}(\overline{\boldsymbol{V}}_{T+1}) \leq d\lambda_0 + TL_x^2.$$

The arithmetic–geometric mean inequality for its positive eigenvalues gives

$$\det \overline{\boldsymbol{V}}_{T+1} \leq \left(\lambda_0 + \frac{TL_x^2}{d}\right)^d.$$

Divide by λ_0^d and take logarithms. □

C.2 Proof of Theorem 6

Proof. **Step 1: simultaneous calibration.** Set $\nu_{t,k}^\star = \log[\mu_k(G_t)/(1-\mu_k(G_t))]$ and $f(x) = \operatorname{sigmoid}(x)$. The derivative satisfies $0 < f'(x) \leq 1/4$. Equation (18) implies

$$\nu_{t,k}^\star \leq \nu_{t,k} + b_t s_{t,k} + \eta_t, \qquad \nu_{t,k} + b_t s_{t,k} \leq \nu_{t,k}^\star + 2b_t s_{t,k} + \eta_t.$$

Monotonicity and the Lipschitz bound yield

$$\mu_k(G_t) \leq U_{t,k} + \eta_t/4, \qquad U_{t,k} \leq \mu_k(G_t) + b_t s_{t,k}/2 + \eta_t/4.$$

Step 2: regret of the selected UCB action. With $k_t^\star$ Bayes optimal and $A_t \in \arg\max_k U_{t,k}$,

$$\begin{aligned} r_t := \mu_t^\star - \mu_{A_t}(G_t) &\leq U_{t,k_t^\star} - \mu_{A_t}(G_t) + \eta_t/4 \\ &\leq U_{t,A_t} - \mu_{A_t}(G_t) + \eta_t/4 \\ &\leq b_t s_{t,A_t}/2 + \eta_t/2. \end{aligned}$$

Because $0 \leq r_t \leq 1$,

$$r_t \leq \min\{1, \bar{b}\, s_{t,A_t}/2\} + \eta_t/2.$$

Step 3: cumulative regret. Lemma 5, specifically (25) with $c = \bar{b}/2$, gives

$$\mathcal{R}_T \leq C_{\mathrm{U}}\sqrt{2T\mathcal{J}_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t \leq C_{\mathrm{U}}\sqrt{2T\overline{J}_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t.$$

The bound is evaluated along the trajectory generated by the learner.

Step 4: expectation and mistakes. Outside the calibration event use $\mathcal{R}_T \leq T$. If $B_{\mathrm{U}}(T)$ denotes the right-hand side of (19), then

$$\mathbb{E}[\mathcal{R}_T] \leq B_{\mathrm{U}}(T) + \delta T, \qquad \mathbb{E}[M_T] - \mathbb{E}[N_T^\star] \leq B_{\mathrm{U}}(T) + \delta T.$$

The second relation uses Proposition 8. For any $\delta_{\mathrm{noise}} \in (0,1)$, a union bound also gives, with probability at least $1 - \delta - \delta_{\mathrm{noise}}$,

$$M_T \leq N_T^\star + B_{\mathrm{U}}(T) + \sqrt{\frac{T}{2}\log\frac{2}{\delta_{\mathrm{noise}}}}.$$

The guarantee controls excess mistakes above the Bayes loss, which may itself grow linearly. □

C.3 Proof of Theorem 7

Proof. **Step 1: fix a calibrated pre-action history.** Condition on $\mathcal{H}_t$ in $\mathcal{C}_t$. Write $\mu_k = \mu_k(G_t)$, $\widetilde{p}_k = \widetilde{p}_{t,k}$, $\Delta_k = \mu_t^\star - \mu_k$, and

$$h_k = \min\{1, c_t^{\mathrm{TS}} s_{t,k} + e_t^{\mathrm{TS}}\}, \qquad \mathcal{U}_t = \{k : \Delta_k \le 2h_k\}.$$

The set $\mathcal{U}_t$ is nonempty since it contains $k_t^\star$. All these quantities are fixed under the conditioning. Let $\mathcal{E}_t$ denote the simultaneous sampling event in (20). Since both $\widetilde{p}_k$ and μ_k are probabilities, on $\mathcal{E}_t$ their difference is bounded by h_k.

Step 2: lower bound the probability of an unsaturated action. On $\mathcal{E}_t \cap \{\widetilde{p}_{k_t^\star} \ge \mu_t^\star\}$, the TS argmax satisfies

$$\mu_t^\star \le \widetilde{p}_{k_t^\star} \le \widetilde{p}_{A_t} \le \mu_{A_t} + h_{A_t}, \qquad \Delta_{A_t} \le h_{A_t}.$$

Therefore $A_t \in \mathcal{U}_t$. No independence between the two events is needed:

$$\mathbb{P}(A_t \in \mathcal{U}_t \mid \mathcal{H}_t) \ge p_0 - \zeta_t \ge p_0/2.$$

Choose $j_t \in \arg\min_{k \in \mathcal{U}_t} h_k$ with fixed tie-breaking. Nonnegativity gives

$$\mathbb{E}[h_{A_t} \mid \mathcal{H}_t] \ge h_{j_t} \mathbb{P}(A_t \in \mathcal{U}_t \mid \mathcal{H}_t), \qquad h_{j_t} \le \frac{2}{p_0} \mathbb{E}[h_{A_t} \mid \mathcal{H}_t].$$

Step 3: bound regret by the selected width. On $\mathcal{E}_t$, using the TS argmax and $j_t \in \mathcal{U}_t$,

$$\begin{aligned} \Delta_{A_t} &= \Delta_{j_t} + \mu_{j_t} - \mu_{A_t} \\ &\le 2h_{j_t} + (\mu_{j_t} - \widetilde{p}_{j_t}) + (\widetilde{p}_{j_t} - \widetilde{p}_{A_t}) + (\widetilde{p}_{A_t} - \mu_{A_t}) \\ &\le 3h_{j_t} + h_{A_t}. \end{aligned}$$

On its complement use $\Delta_{A_t} \le 1$, giving

$$\mathbb{E}[\Delta_{A_t} \mid \mathcal{H}_t] \le \left(1 + \frac{6}{p_0}\right) \mathbb{E}[h_{A_t} \mid \mathcal{H}_t] + \zeta_t.$$

Step 4: histories outside the calibration event. The event $\mathcal{C}_t$ is $\mathcal{H}_t$-measurable. Thus the conditional inequality, multiplied by $\mathbf{1}_{\mathcal{C}_t}$ and integrated, implies

$$\mathbb{E}[\Delta_{A_t}] \le \left(1 + \frac{6}{p_0}\right) \mathbb{E}[h_{A_t}] + \zeta_t + \mathbb{P}(\mathcal{C}_t^c).$$

The action law remains conditioned on the current history rather than the future event $\bigcap_s \mathcal{C}_s$. Its assumed probability implies $\mathbb{P}(\mathcal{C}_t^c) \le \delta$ for each t.

Step 5: sum widths and errors. Since $h_{A_t} \le \min\{1, \bar{c}^{\mathrm{TS}}\, s_{t,A_t}\} + e_t^{\mathrm{TS}}$, the pathwise information bound yields

$$\sum_{t=1}^{T} h_{A_t} \le C_{\mathrm{TS}} \sqrt{2T\overline{J}_T} + \sum_{t=1}^{T} e_t^{\mathrm{TS}}.$$

Sum Step 4 to obtain (22). Proposition 8 gives the identical bound for $\mathbb{E}[M_T] - \mathbb{E}[N_T^\star]$. □

C.4 Proof of Lemma 10

Proof. **Step 1: write the nonlinear log odds along the draw.** Suppress t after conditioning. Draw $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + \alpha \boldsymbol{S}\boldsymbol{\xi}$, with $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{I}_d)$. Let $\boldsymbol{h} = \alpha \boldsymbol{Z}\boldsymbol{S}\boldsymbol{\xi}$ be the logit perturbation, and define

$$f_k(u) = g_k + u h_k - \log \sum_{j \ne k} \exp(g_j + u h_j), \qquad 0 \le u \le 1.$$

Then $f_k(0) = \nu_k$, $f_k(1) = \widetilde{\nu}_k$, and

$$f_k'(0) = \alpha \boldsymbol{d}_k^\top \boldsymbol{S}\boldsymbol{\xi}.$$

The second derivative is minus the variance of the numbers $\{h_j : j \neq k\}$ under their normalized exponential weights. For any distribution supported on an interval of length R, the variance is at most $R^2/4$. Hence

$$-\frac{\alpha^2 (D^{\mathrm{logit}})^2 \|\boldsymbol{\xi}\|_2^2}{4} \leq f_k''(u) \leq 0.$$

Taylor's formula with integral remainder gives

$$\left|\widetilde{\nu}_k - \nu_k - \alpha \boldsymbol{d}_k^\top \boldsymbol{S}\boldsymbol{\xi}\right| \leq \frac{\alpha^2 (D^{\mathrm{logit}})^2 \|\boldsymbol{\xi}\|_2^2}{8}.$$

Step 2: control the Gaussian linear terms and the remainder. Each $\boldsymbol{d}_k^\top \boldsymbol{S}\boldsymbol{\xi}$ is centered Gaussian with variance s_k^2, including the degenerate case $s_k = 0$. A Gaussian tail bound and a union bound imply

$$\mathbb{P}\left(\exists k : \ |\alpha \boldsymbol{d}_k^\top \boldsymbol{S}\boldsymbol{\xi}| > \alpha s_k \sqrt{2\log(4K/\zeta)} \mid \mathcal{H}_t\right) \leq \zeta/2.$$

Separately,

$$\mathbb{P}\left(\max_{i \leq d} |\xi_i| > \sqrt{2\log(4d/\zeta)}\right) \leq \zeta/2.$$

On the second event's complement, $\|\boldsymbol{\xi}\|_2^2 \leq 2d\log(4d/\zeta)$, so the Taylor remainder is at most ρ_t. Combining the two events proves the simultaneous log-odds bound with failure at most ζ_t.

Step 3: convert log odds to probabilities. On a calibration history, the triangle inequality gives

$$|\widetilde{\nu}_{t,k} - \nu_{t,k}^\star| \leq (b_t + a_t) s_{t,k} + \eta_t + \rho_t.$$

Apply the $1/4$-Lipschitz property of the sigmoid to get (20) with the claimed $c_t^{\mathrm{TS}}, e_t^{\mathrm{TS}}$.

Step 4: sufficient optimism. On the norm event from Step 2,

$$\widetilde{\nu}_{t,k_t^\star} \geq \nu_{t,k_t^\star} + \alpha_t \boldsymbol{d}_{t,k_t^\star}^\top \boldsymbol{S}_t \boldsymbol{\xi} - \rho_t.$$

Calibration gives $\nu_{t,k_t^\star}^\star - \nu_{t,k_t^\star} \leq b_t s_{t,k_t^\star} + \eta_t$. Thus a sufficient event for optimism is

$$\alpha_t \boldsymbol{d}_{t,k_t^\star}^\top \boldsymbol{S}_t \boldsymbol{\xi} \geq b_t s_{t,k_t^\star} + \eta_t + \rho_t.$$

When $s_{t,k_t^\star} > 0$, standardize this Gaussian variable and subtract the norm-event failure probability $\zeta_t/2$. This yields the stated lower bound. No independence between the linear Gaussian and the norm event is assumed. □

C.5 Proof of Theorem 1

Proof.

Step 1: preserve the head's represented function. Let $\boldsymbol{\Theta}$ be the $(K-1) \times m$ identified head matrix and let the fixed contrast basis be $\boldsymbol{C} \in \mathbb{R}^{K \times (K-1)}$. Under $\boldsymbol{z}^+ = \boldsymbol{M}\boldsymbol{z}^-$ set $\boldsymbol{\Theta}^+ = \boldsymbol{\Theta}^- \boldsymbol{M}^{-1}$. Then for every relevant graph,

$$\boldsymbol{C}\boldsymbol{\Theta}^+ \boldsymbol{z}^+ = \boldsymbol{C}\boldsymbol{\Theta}^- \boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{z}^- = \boldsymbol{C}\boldsymbol{\Theta}^- \boldsymbol{z}^-.$$

Column-major vectorization gives $\boldsymbol{\theta}^+ = (\boldsymbol{M}^{-\top} \otimes \boldsymbol{I}_{K-1})\boldsymbol{\theta}^- = \boldsymbol{P}\boldsymbol{\theta}^-$. Equal centered logits imply equal softmax probabilities, class-versus-rest log odds, and selected-bit PLL losses for both $R = 0$ and $R = 1$.

Step 2: transform derivatives by the chain rule. For an invariant scalar prediction function $f^+(\boldsymbol{\theta}^+) = f^-(\boldsymbol{P}^{-1}\boldsymbol{\theta}^+)$,

$$\boldsymbol{d}^+ = \nabla_{\boldsymbol{\theta}^+} f^+(\boldsymbol{\theta}^+) = \boldsymbol{P}^{-\top} \boldsymbol{d}^-.$$

This applies to the log-odds gradient defining the predictive width.

Step 3: preserve covariance and confidence geometry. Since $\boldsymbol{V}^+ = \boldsymbol{P}^{-\top}\boldsymbol{V}^-\boldsymbol{P}^{-1}$,

$$(\boldsymbol{V}^+)^{-1} = \boldsymbol{P}(\boldsymbol{V}^-)^{-1}\boldsymbol{P}^\top = \boldsymbol{\Sigma}^+.$$

Let $e^- = \widehat{\theta}^- - \theta^{\star,-}$ and $e^+ = Pe^-$. Direct substitution cancels the maps:

$$(e^+)^\top V^+ e^+ = (e^-)^\top P^\top P^{-\top} V^- P^{-1} P e^- = (e^-)^\top V^- e^-.$$

The same cancellation gives

$$(d^+)^\top \Sigma^+ d^+ = (d^-)^\top \Sigma^- d^-.$$

The entire matrix, including its prior or regularizer contribution, must transform. Resetting λI after a nonorthogonal map violates the first equality.

Step 4: couple decisions. Equal means and widths give equal UCB scores with the same width multiplier. Equal scores and the same tie convention give equal proposed actions. For TS, condition on the completed update history so P is fixed, and couple $\widetilde{\theta}^+ = P\widetilde{\theta}^-$ with $\widetilde{\theta}^- \sim N(\widehat{\theta}^-, \alpha_t^2 \Sigma^-)$. The Gaussian pushforward has mean $P\widehat{\theta}^-$ and covariance $\alpha_t^2 P \Sigma^- P^\top$. Step 1 implies identical sampled logits and hence identical argmax actions. Independent draws need only have the same action law.

Step 5: compose adaptive maps. All identities hold for each realized invertible P, so they remain true when P was chosen from completed feedback. Applying them successively to $P_j \cdots P_1$ proves the finite sequence of instantaneous invariances. Fresh feedback is still assimilated using predictable gradients. Equivalence across intervening optimization steps requires the additional condition in Corollary 2. □

C.6 Proof of Theorem 3

Proof.

We construct a linear bandit whose arm rewards are not constrained to sum to one.

Step 1: accumulate information in an irrelevant direction. For the linear-bandit construction, x_s is the selected arm feature and R_s its reward. Fix $\lambda > 0$, a strictly positive UCB multiplier b_t, and a ridge center $\widehat{\theta} = V^{-1} \sum_s x_s R_s$. For rounds $1, \ldots, n$, both actions have old-coordinate feature e_1 and the parameter is $\theta^{\star,-} = e_2$. Both actions have zero reward, so the learner incurs no regret during this phase. At its end,

$$V_n^- = \lambda I + n e_1 e_1^\top = \mathrm{Diag}(\lambda + n, \lambda), \qquad \widehat{\theta}_n^- = \mathbf{0}.$$

Step 2: swap coordinates and specify the next action sets. Use $M = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, so $\theta^{\star,+} = e_1$. For all remaining n rounds, the two new-coordinate features are e_1 (reward one) and e_2 (reward zero). The transported center is still zero. A stale precision matrix remains V_n^-.

Step 3: induct over the post-change decisions. Suppose $s < n$ post-change decisions have all selected e_2. Their rewards are zero, so the ridge center remains zero and stale assimilation produces $V = \mathrm{Diag}(\lambda + n, \lambda + s)$. The scores are

$$U(e_1) = \frac{b_{n+s+1}}{\sqrt{\lambda + n}}, \qquad U(e_2) = \frac{b_{n+s+1}}{\sqrt{\lambda + s}}.$$

Since $s < n$, $U(e_2) > U(e_1)$. This proves the induction, including $s = 0$. Each of these n decisions loses exactly one unit, hence total regret is $n = T/2$.

Step 4: contrast the correctly transported state. Congruence transport instead gives $V_n^+ = \mathrm{Diag}(\lambda, \lambda + n)$, so its first post-change scores strictly prefer e_1. This establishes optimality of the first action after the change. Later actions may still incur exploration loss. □

C.7 Proof of Proposition 4

Proof.

Step 1: bound coupled logit errors. Let $\boldsymbol{r} = \boldsymbol{m}^+ - \boldsymbol{m}^-$ and $\boldsymbol{\Delta} = \boldsymbol{B}^+ - \boldsymbol{B}^-$. For each k, $g_k^+ - g_k^- = r_k + \boldsymbol{\Delta}_{k,:}\boldsymbol{\xi}$, with scalar Gaussian noise of variance $\|\boldsymbol{\Delta}_{k,:}\|_2^2$. If $d_B > 0$, the Gaussian tail bound and a union bound give

$$\Pr\left\{\max_k |\boldsymbol{\Delta}_{k,:}\boldsymbol{\xi}| > d_B\sqrt{2\log(2K/\delta)}\right\} \le \delta.$$

If $d_B = 0$, the random residual is identically zero. In either case, $\|\boldsymbol{g}^+ - \boldsymbol{g}^-\|_\infty \le \epsilon_\delta$ outside an event of probability at most δ. Correlations among class logits do not invalidate this union bound.

Step 2: use the decision margin. On that event's complement, for the old unique maximizer a^- and any $b \neq a^-$,

$$g_{a^-}^+ - g_b^+ \ge g_{a^-}^- - g_b^- - 2\epsilon_\delta \ge \Gamma(\boldsymbol{g}^-) - 2\epsilon_\delta.$$

A margin greater than $2\epsilon_\delta$ therefore preserves the maximizing action. The event inclusion $\{a^+ \neq a^-\} \subseteq \{\Gamma(\boldsymbol{g}^-) \le 2\epsilon_\delta\} \cup \{\|\boldsymbol{g}^+ - \boldsymbol{g}^-\|_\infty > \epsilon_\delta\}$ proves (14). Ties belong to the small-margin event. The bound does not require an anti-concentration assumption. Applying the same inequality to deterministic scores proves the UCB statement.

Step 3: translate disagreement to one-step loss. For $\boldsymbol{\mu} \in [0,1]^K$, $|\mu_{a^-} - \mu_{a^+}| \le \mathbf{1}\{a^- \neq a^+\}$ pointwise. Jensen's inequality and expectation prove (15). Conditional versions hold at each common history, so a sum follows by iterated expectation and linearity. Histories generated by independently deployed policies differ. This argument does not couple their future training states.

Step 4: relate the residual to transport. Let $\bar{\boldsymbol{\theta}}^+ = \boldsymbol{P}\widehat{\boldsymbol{\theta}}^-$ be the transported mean. The Gaussian head coupling $\widetilde{\boldsymbol{\theta}}^- = \widehat{\boldsymbol{\theta}}^- + \alpha_t \boldsymbol{S}\boldsymbol{\xi}$ and $\widetilde{\boldsymbol{\theta}}^+ = \bar{\boldsymbol{\theta}}^+ + \alpha_t \boldsymbol{P}\boldsymbol{S}\boldsymbol{\xi}$ has the desired means and transported covariance. Multiplication by the logit designs gives $\boldsymbol{r} = \boldsymbol{Z}^+\bar{\boldsymbol{\theta}}^+ - \boldsymbol{Z}^-\widehat{\boldsymbol{\theta}}^-$ and $\boldsymbol{\Delta} = \alpha_t(\boldsymbol{Z}^+\boldsymbol{P} - \boldsymbol{Z}^-)\boldsymbol{S}$. Both vanish under the exact identity. The argument applies standard argmax perturbation and Gaussian-tail bounds to the residual induced by feature transport. □

C.8 Proof of Proposition 9

Proof.

Step 1: specify what is conditioned on. Let $\mathcal{A}_t$ contain the past, current graph, any sampled head, and the chosen action, immediately before feedback. The gradient $\boldsymbol{x}_t$ used for assimilation is $\mathcal{A}_t$-measurable. With $\varepsilon_t = R_t - \mu_{A_t}(G_t)$, the feedback model gives $\mathbb{E}[\varepsilon_t \mid \mathcal{A}_t] = 0$.

Step 2: take the predictable factor outside the expectation. For each coordinate i, conditional integrability and measurability imply

$$\mathbb{E}[x_{t,i}\varepsilon_t \mid \mathcal{A}_t] = x_{t,i}\mathbb{E}[\varepsilon_t \mid \mathcal{A}_t] = 0.$$

Stacking the coordinates proves the vector score identity. Summing these increments produces a martingale, the object required by a self-normalized concentration argument. A confidence bound for the estimator requires additional assumptions beyond this score identity.

Step 3: identify the failure of same-feedback re-embedding. For $R_t \sim \mathrm{Bernoulli}(1/2)$ and $x_t' = 2R_t - 1$, direct calculation gives $x_t'(R_t - 1/2) = 1/2$ for either possible outcome. Thus its conditional expectation is $1/2$, not zero. Assimilating the score in old predictable coordinates and then applying an exact algebraic change of coordinates avoids this failure. Re-fitting old rewards on a newly trained nonlinear feature map requires a separate statistical argument. □

C.9 Proof of Proposition 8

Proof.

Step 1: decompose one round. Write $\mu_t^\star = \max_k \mu_k(G_t)$ and $\Delta_t = \mu_t^\star - \mu_{A_t}(G_t)$. Pointwise,

$$1 - R_t = (1 - \mu_t^\star) + \Delta_t + \{\mu_{A_t}(G_t) - R_t\}.$$

The first term is the Bayes error of this context, the second is the decision's pseudo-regret, and the last has conditional mean zero given $\mathcal{A}_t$.

Step 2: sum and take expectations. With $N_T^\star = \sum_t (1 - \mu_t^\star)$,

$$M_T = N_T^\star + \mathcal{R}_T + \Delta_T^{\text{noise}}, \qquad \mathbb{E}[M_T] = \mathbb{E}[N_T^\star] + \mathbb{E}[\mathcal{R}_T], \tag{26}$$

where $D_t = \mu_{A_t}(G_t) - R_t$ has conditional mean zero and lies in an interval of length one. Conditional Hoeffding's lemma gives, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda D_t} \mid \mathcal{A}_t] \le e^{\lambda^2/8}, \qquad \mathbb{E}[e^{\lambda \sum_{t=1}^T D_t}] \le e^{T\lambda^2/8}.$$

For $u > 0$, exponential Markov and minimization over $\lambda > 0$ yield

$$\mathbb{P}(\Delta_T^{\text{noise}} \ge u) \le \inf_{\lambda > 0} e^{-\lambda u + T\lambda^2/8} = e^{-2u^2/T}.$$

Applying the same inequality to $-D_t$ gives

$$\mathbb{P}(|\Delta_T^{\text{noise}}| \ge u) \le 2e^{-2u^2/T}, \qquad u = \sqrt{\tfrac{T}{2}\log(2/\delta)}.$$

The conditional means are used only for this evaluation identity.

Step 3: interpret the rate correctly. If $\mathbb{E}[N_T^\star] \ge cT$ for $c > 0$, even the oracle $A_t \in \arg\max_k \mu_k(G_t)$ has at least cT expected mistakes while its pseudo-regret is zero. For example, constant class probabilities $(0.8, 0.1, 0.1)$ yield $\mathbb{E}[M_T] = 0.2T$ for that oracle. Therefore sublinear $\mathcal{R}_T$ does not imply sublinear M_T. For deterministic labels, the conditional Bayes loss is zero. □

C.10 Proof of Corollary 2

Proof.

Step 1: commuting updates imply an induction. Let S_t denote all learner state and $\mathcal{T}_{\boldsymbol{P}}$ its complete coordinate transform. The required property is

$$\mathcal{U}^+(\mathcal{T}_{\boldsymbol{P}}\mathsf{S}_t, \boldsymbol{z}^+, A_t, R_t) = \mathcal{T}_{\boldsymbol{P}}\mathcal{U}^-(\mathsf{S}_t, \boldsymbol{z}^-, A_t, R_t).$$

Assume transformed states agree before a round. The exact theorem and coupled randomness yield the same action. The same categorical-label draw yields the same bit. Commutation then gives corresponding next states. Initialization and induction establish the entire common trajectory.

Step 2: give a compatible example. Writing $\boldsymbol{u}^\pm$ for the PLL head gradients, the chain rule gives $\boldsymbol{u}^+ = \boldsymbol{P}^{-\top}\boldsymbol{u}^-$. For a step size $\lambda_{\text{step}} > 0$, a covariance-preconditioned step satisfies

$$\widehat{\boldsymbol{\theta}}^+ - \lambda_{\text{step}}\boldsymbol{\Sigma}^+\boldsymbol{u}^+ = \boldsymbol{P}\widehat{\boldsymbol{\theta}}^- - \lambda_{\text{step}}\boldsymbol{P}\boldsymbol{\Sigma}^-\boldsymbol{P}^\top\boldsymbol{P}^{-\top}\boldsymbol{u}^- = \boldsymbol{P}(\widehat{\boldsymbol{\theta}}^- - \lambda_{\text{step}}\boldsymbol{\Sigma}^-\boldsymbol{u}^-).$$

Clipping by the invariant norm $\sqrt{\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u}}$ preserves this property. An assimilation $\boldsymbol{V} \leftarrow \boldsymbol{V} + w\boldsymbol{d}\boldsymbol{d}^\top$ also commutes because w is an invariant probability function and $\boldsymbol{d}$ transforms covariantly. Optimizer momentum and regularization gradients, if present, must transform consistently too.

Step 3: exhibit the Euclidean-SGD obstruction. A Euclidean step in new coordinates is $\boldsymbol{P}\widehat{\boldsymbol{\theta}}^- - \lambda_{\text{step}}\boldsymbol{P}^{-\top}\boldsymbol{u}^-$, whereas the transformed old step is $\boldsymbol{P}\widehat{\boldsymbol{\theta}}^- - \lambda_{\text{step}}\boldsymbol{P}\boldsymbol{u}^-$. These differ for a general nonorthogonal $\boldsymbol{P}$. Consequently, equality of predictions immediately after transport does not ensure that later Euclidean-SGD trajectories remain equivalent. □

D Experimental Details and Supplementary Results

The experiments examine the performance of the complete algorithms, the effect of MT during representation learning, and the value of additional covariance fitting or policy aggregation. Appendix D.1 describes the common protocol. Appendices D.2 and D.3 compare classification performance and Bayes pseudo-regret. Appendices D.4 and D.5.1 examine learned encoder updates and exact transport invariance. The aggregation and covariance-correction studies in Appendices D.6 and D.7 report all prespecified comparisons, including negative and inconclusive results.

D.1 Data, learning protocol, and baseline implementations

D.1.1 Data and model selection

Each method receives a graph and then only the correctness bit for its selected class. The evaluator retains the label, conditional probabilities in simulation, and the losses of comparison policies. The learner cannot access these quantities during prediction, encoder training, head assimilation, replay selection or aggregation updates. In aggregation experiments, the GCN adviser updates using the action selected and executed by the aggregate. Held-out evaluation is deterministic and does not advance the TS random state.

The four benchmark datasets are ENZYMES, IMDB-MULTI, COLLAB, and REDDIT-MULTI-5K (Morris et al., 2020). Each seed defines a stratified 70/15/15 train/validation/test split and a paired random one-pass training order. The training horizons are 420, 1,050, 3,500, and 3,499 rounds, respectively. Node-feature preparation is fitted on training graphs. Degree features use eight bins where required. The confirmation split and stream are shared by all methods within a seed. Seed randomizations are the inferential units conditional on these four fixed datasets. The analysis does not treat the datasets as an independently sampled population.

Every learned variant uses a two-layer tanh GCN with widths 16 and 12 and a 32-round encoder update period. Seventeen learned variants receive five encoder rates $\{0.001, 0.003, 0.01, 0.03, 0.1\}$ on three development seeds. Two controls without a GCN are evaluated once per seed. Across four datasets this is 1,044 development runs. The rate minimizes development mean online mistake rate, with time and then rate as tie-breakers. Test labels are absent in development. This gives each learned method the same number of candidate settings, although method-specific parameters are not exhaustively tuned. The analysis includes all candidates. All learned methods use the same algebraically stable PLL arithmetic throughout development to avoid cancellation in $1 - \widehat{p}_a$.

D.1.2 Head and encoder updates

As in the main method, $\widehat{p}_a$ denotes the native GCN probability. For the PLL variants, the identified head starts at zero, its covariance at the identity, and its GCN weights are independently initialized with Gaussian scales 0.7 and 0.5. For the selected pre-feedback derivative $\boldsymbol{d}_a$, set $\boldsymbol{v} = \boldsymbol{\Sigma}\boldsymbol{d}_a$ and $w = \widehat{p}_a(1 - \widehat{p}_a)$. Assimilation uses

$$\boldsymbol{\Sigma}^- = \boldsymbol{\Sigma} - \frac{w\boldsymbol{v}\boldsymbol{v}^\top}{1 + w\boldsymbol{d}_a^\top \boldsymbol{v}}, \qquad \widehat{\boldsymbol{\theta}}^- = \widehat{\boldsymbol{\theta}} + \frac{\boldsymbol{\Sigma}^- \boldsymbol{d}_a (R - \widehat{p}_a)}{\max\{1, |R - \widehat{p}_a| \sqrt{\boldsymbol{d}_a^\top \boldsymbol{\Sigma}^- \boldsymbol{d}_a}\}}.$$

The covariance is symmetrized after the rank-one update. At a scheduled encoder step, the selected PLL encoder gradient is computed from the pre-feedback forward cache and head. Its norm is capped at one before applying the selected encoder rate. Thus the head assimilates the bit first, while the saved forward quantities identify the loss being differentiated.

The feature map is fitted on the last 32 observed graphs, with identity ridge 10^{-5}. A condition number exceeding 20 defers the encoder update. MT uses the congruence-transformed covariance. Stale keeps the old covariance after moving the head. The replay variant rebuilds an identity-regularized Fisher precision from the last 128 selected-bit rows in current features. For controlled memory accounting, the implementation retains 128 selected-bit rows for MT and stale as well. Only replay uses them to rebuild the Fisher state. The three rules have the same encoder-update opportunities, and replay work is included in runtime. In the broad confirmation study, the UCB width multiplier is 0.5. TS uses a single joint Gaussian head draw with scale one. The IX step size and prior are fixed by the protocol in Appendix D.6.1, without confidence-envelope fitting or a supplied true predictor.

D.1.3 BASELINE IMPLEMENTATIONS

`FastCB-PLL` uses the `FastCB` allocation on the same predicted loss $\widehat{\ell}_a = 1 - \widehat{p}_a$ that is passed to its selected-bit log-loss update (Foster & Krishnamurthy, 2021). Here q_a is the probability of selecting class a and $\gamma > 0$ controls the allocation. If b minimizes $\widehat{\ell}_a$, then for $a \neq b$,

$$q_a = \frac{\widehat{\ell}_b}{K\widehat{\ell}_b + \gamma(\widehat{\ell}_a - \widehat{\ell}_b)}, \qquad q_b = 1 - \sum_{a \neq b} q_a.$$

The zero-loss case uses the continuous tied-action limit. The `SquareCB-allocation/PLL` control uses $q_a = \{K + \gamma(\widehat{\ell}_a - \widehat{\ell}_b)\}^{-1}$ for $a \neq b$ (Foster & Rakhlin, 2020), with the same PLL prediction oracle. The PLL update replaces the original squared-loss oracle, so the original guarantee is not assumed to hold for this optimizer. Both allocations use $\gamma = \max\{10K, \sqrt{KT}\}$ and consistent pre-feedback predicted losses.

The graph `LiM2` adaptation implements selected-arm squared-error training, normal–inverse-gamma linear posteriors, and PSD predictive-variance matching from Nabati et al. (2021). It uses capacity 128, minibatches of at most 16, a projected-gradient variance-matching pass with step 0.01, and an SPD inversion floor 10^{-5}. The head and encoder gradient norm is capped at five. The new neural head supplies the prior mean. Each arm's retained rows appear once in the rebuilt posterior. The buffer removes the first tuple of the incoming arm when full. The inverse-gamma prior has shape two and scale one. This graph adaptation differs from the released TensorFlow system in its GCN architecture, epochic updates, buffer reset and variance-matching solver.

The matched neural-linear TS replay control uses the same architecture, MSE updates, posterior formulas and buffer, resetting the prior covariance to the identity instead of fitting old predictive variances. Its UCB counterpart uses posterior mean plus $0.5\sqrt{\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z}}$, without inverse-gamma noise scaling, where $\boldsymbol{z}$ is its current graph feature and $\boldsymbol{\Sigma}$ is the selected arm's linear posterior covariance. This implementation differs from variance-aware `Neural-LinUCB` algorithms in its treatment of posterior uncertainty. These Gaussian reward-surrogate methods do not natively produce a multinomial probability simplex. We therefore compare their mistakes and mean-head accuracy, not uncalibrated probability metrics.

D.1.4 STATISTICAL ANALYSIS AND REPRODUCIBILITY

For the broad synthetic and real-data aggregation studies, the declared families use paired two-sided t tests, Holm-adjusted p values, marginal 95% intervals and Bonferroni simultaneous intervals. Multiple conditions sharing one base seed are averaged before inference. The real-data primary outcome is the mean of four dataset mistake-rate differences, giving each benchmark equal weight. The synthetic primary outcome is cumulative pseudo-regret over the two recurring-pool conditions. Pointwise curve intervals are descriptive and are not simultaneous over methods or horizons. The prespecified confirmation budget uses 20 independent real-data seeds. A sensitivity calculation based only on development data estimates the detectable effect at this fixed sample size. Three development seeds provide an uncertain variance estimate, particularly after selecting hyperparameters. Consequently, small effects may remain statistically inconclusive or have wide confidence intervals.

Methods, hyperparameters, random seeds, and analysis contrasts are fixed before confirmation. Evaluation is separated from learning: labels reconstruct the selected correctness bits, while comparison-policy losses and held-out outcomes are computed without updating the learner. Development observations are excluded from confirmation, and confirmation outcomes are not used for model selection. The learned-update and correction studies specify their separate sign-randomization tests below.

Total learner time includes encoder, head, reconciliation and master computation. A separate 19-method IMDB-MULTI benchmark at rate 0.01 and development-only seed 62,004 measures isolated-process peak RSS and retained NumPy arrays, including replay graphs. RSS includes imports and the prepared dataset. The retained-array measure deduplicates by object identity and excludes Python-object overhead. It therefore approximates retained array memory rather than the exact memory allocated to the model.

D.2 Real-data ablations and additional performance measures

D.2.1 Online classification and general comparison

Table 4: Online mistake rate (%) of the complete methods: mean ± standard deviation over twenty seeds. Lower is better. Bold marks the lowest unrounded mean among the methods displayed. It does not denote a significance test. Uncertainty ablations and aggregation extensions are reported in Appendix D.2. IMDB and REDDIT denote IMDB-MULTI and REDDIT-MULTI-5K.

Method	ENZYMES	IMDB	COLLAB	REDDIT
UCB-MT	**74.58** ± 3.34	**57.15** ± 1.46	35.38 ± 3.34	**63.47** ± 2.02
TS-MT	78.23 ± 2.71	59.08 ± 2.45	**35.21** ± 3.02	63.55 ± 1.70
FastCB-PLL	76.80 ± 2.13	59.02 ± 2.31	36.10 ± 3.68	64.73 ± 1.95
SquareCB allocation/PLL	76.49 ± 1.93	59.28 ± 1.93	36.60 ± 2.88	65.58 ± 2.10
Graph LiM2 adaptation	81.79 ± 1.89	64.25 ± 1.66	45.47 ± 3.36	78.18 ± 1.82
Graph neural-linear UCB	75.75 ± 2.11	62.40 ± 2.04	41.28 ± 3.05	71.75 ± 5.19
Graph neural-linear TS	79.73 ± 2.43	62.91 ± 1.73	43.40 ± 2.95	74.45 ± 2.72

Table 5 includes the complete methods, the stale and replay ablations, and the aggregation extensions. UCB-replay improves over UCB-MT on ENZYMES, COLLAB, and REDDIT by 0.30, 0.48, and 1.44 percentage points, respectively. UCB-MT improves by 0.99 points on IMDB. TS-MT improves over TS-replay on each dataset. MT–stale differences are small except for UCB on REDDIT. These mean differences are descriptive. The declared aggregation comparisons are reported in Appendix D.6.

Table 5: Online mistake rate over the independent real-data seeds. Lower is better. IMDB and REDDIT abbreviate IMDB-MULTI and REDDIT-MULTI-5K.

Bold indicates the best unrounded column mean and does not denote statistical significance.

Method	ENZYMES	IMDB	COLLAB	REDDIT
UCB-MT	0.746	**0.572**	0.354	0.635
UCB-stale	0.744	0.573	0.355	0.692
UCB-replay	**0.743**	0.581	**0.349**	**0.620**
TS-MT	0.782	0.591	0.352	0.636
TS-stale	0.782	0.587	0.352	0.637
TS-replay	0.794	0.612	0.365	0.659
FastCB-PLL	0.768	0.590	0.361	0.647
SquareCB allocation/PLL	0.765	0.593	0.366	0.656
Graph LiM2 adapt.	0.818	0.643	0.455	0.782
Graph neural-linear TS	0.797	0.629	0.434	0.745
Graph neural-linear UCB	0.757	0.624	0.413	0.717
UCB-MT-IX	0.819	0.626	0.392	0.704
UCB-stale-IX	0.818	0.626	0.393	0.705
UCB-replay-IX	0.818	0.630	0.393	0.704
TS-MT-IX	0.824	0.637	0.394	0.709
TS-stale-IX	0.824	0.641	0.395	0.707
TS-replay-IX	0.824	0.647	0.408	0.730
IX-uniform	0.833	0.671	0.527	0.799
EXP3-IX	0.834	0.673	0.520	0.800

D.2.2 Computation and memory

Learner time includes representation learning, head updates, and uncertainty reconciliation. Replay rebuilds Fisher information from the most recent 128 selected-bit rows, while MT estimates its map on 32 anchors. The same buffer capacity is retained for all three uncertainty variants. Figure 2 shows the resulting mistake–time tradeoff: replay requires 1.9–2.5 times the direct MT runtime. In the isolated-process development benchmark, MT, stale and replay each retain 366,272 array bytes because buffer capacities are matched. Peak RSS ranges from 51.6 to 52.7 MiB across the

nineteen methods and includes the prepared dataset. These measurements provide no evidence of a model-memory advantage for MT.

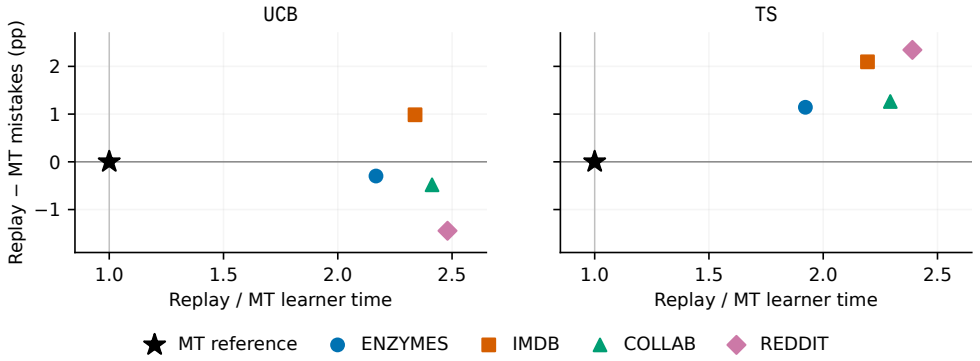

Figure 2: Replay relative to MT on each real dataset. Horizontal axis: ratio of mean learner runtimes. Vertical axis: replay minus MT mean mistake rate in percentage points. MT is the common reference $(1, 0)$. Points above zero favor MT in mistakes, while points below zero favor replay. These are descriptive means, not confidence regions.

D.3 Bayes pseudo-regret comparison of the complete methods

Graph generation and feedback. This study compares `UCB-MT` and `TS-MT` with the five complete external baseline adaptations from Figure 1. The study was designed after the covariance and real-data experiments. It uses the dense and sparse role-SBM families and topology teacher from the learned-update benchmark, fixed before hyperparameter tuning. For a graph G, let $\boldsymbol{\rho}(G) \in \mathbb{R}^3$ contain the realized edge densities between role pairs $(0,1), (1,2), (2,0)$ and let $\boldsymbol{B} \in \mathbb{R}^{3\times 2}$ be the fixed orthonormal contrast basis used by the generator. The teacher applies $\mathrm{softmax}(7.5\boldsymbol{H}\boldsymbol{B}^\top[\boldsymbol{\rho}(G) - \bar{\rho}(G)\mathbf{1}])$, where $\bar{\rho}(G)$ is the mean of the three density coordinates, $\mathbf{1} \in \mathbb{R}^3$ is the all-ones vector, and the three rows of $\boldsymbol{H} \in \mathbb{R}^{3\times 2}$ are unit vectors separated by $2\pi/3$ and their common orientation is randomized once per seed. This teacher remains fixed throughout each stream and is independent of the student GCN. Each graph has 18–24 nodes, one-hot role features and one independent Gaussian node feature. Within-role edge probabilities are 0.15 (dense) and 0.10 (sparse). One uniformly selected cross-role channel has probability 0.78 or 0.42, and the remaining channels have probability 0.08 or 0.05. The sampled graphs, teacher probabilities and label draws are shared across methods within each seed and family.

The evaluator alone draws the class label and computes $\mu_t^\star - \mu_{A_t}(G_t)$. Each learner receives only the graph before prediction and the selected action, correctness bit and round afterward. Trajectories record selected actions, correctness bits, chosen means, optimal means and cumulative regret at 256-round intervals. The implementation and protocol are fixed before development, and selected settings are fixed before confirmation. For the 280 confirmation runs and 1,146,880 rounds, we check consistency of the 40 shared streams, final regret sums, learning curves, observed correctness bits and method implementations. Confirmation outcomes are excluded from tuning, and all runs and families are included in the analysis.

Equal tuning budgets and adaptation details. All methods use the same initial GCN weights, two tanh layers of widths $(8, 6)$, encoder-update period 64 and retained-feedback capacity 128. MT uses 32 recent anchor graphs. The smaller synthetic backbone is shared by all seven methods. The separate real-data study uses widths $(16, 12)$. Each method receives the Cartesian grid of encoder rates $\{0.01, 0.03, 0.1\}$ and exploration multipliers $\{0.0625, 0.25, 1, 4, 16\}$: fifteen configurations, two development seeds and two graph families at horizon 1024, totaling 420 development runs. The smallest mean development pseudo-regret selects one configuration per method, with rate and multiplier breaking ties. Confirmation uses twenty independent seeds at horizon 4096, totaling 280 runs. The independent seed is the inferential unit, with both families averaged before testing.

For UCB-MT, the multiplier is b_t. For TS-MT, it is α_t. FastCB-PLL and SquareCB allocation/PLL set their allocation parameter to $\gamma = c \max\{30, \sqrt{3T}\}$ for multiplier c. Their encoder and joint head use the selected-bit PLL oracle and MT. The neural-linear UCB multiplier is its posterior-width coefficient. LiM2 and neural-linear TS multiply deviations of each normal–inverse-gamma posterior draw from its mean by c. The underlying selected-arm regression, buffer rebuilding, variance matching and encoder optimization are unchanged from the graph adaptations in Appendix D.1. These graph adaptations use practical learners whose compliance with the original oracle assumptions has not been established. Equal configuration counts and shared architecture do not imply identical losses, optimization work or computational costs.

Table 6: Settings selected using only the two development seeds and both families. Development pseudo-regret is measured at 1024 rounds.

Method	Encoder rate	Multiplier	Development regret
UCB-MT	0.01	16	212.28
TS-MT	0.01	0.25	258.15
FastCB-PLL	0.01	16	257.07
SquareCB allocation/PLL	0.01	16	258.87
Graph LiM2 adaptation	0.03	0.25	352.56
Graph neural-linear UCB	0.01	1	332.55
Graph neural-linear TS	0.1	1	368.05

Several selected multipliers lie at a grid boundary, so the comparison is limited to the stated tuning budget. A wider grid would require a new development stage and independent evaluation. Hyperparameters remain fixed throughout confirmation. The UCB results describe the complete tuned method on this benchmark. The separate covariance controls in Appendix D.4 assess how much of the improvement can be attributed to MT.

Statistical analysis. For each of the ten prespecified differences between a proposed method and an external baseline, we average the dense and sparse regrets within each seed, then enumerate all 2^{20} sign assignments for a two-sided paired mean-difference test. This test uses the paired sign-symmetry null. Holm adjustment covers the single family of ten tests. Table 7 reports every contrast. Its paired t intervals are descriptive, unadjusted 95% intervals, not simultaneous intervals. Eight differences remain significant after correction: all five UCB-MT comparisons and the three TS-MT comparisons with LiM2 and neural-linear methods. Neither TS-MT versus FastCB nor TS-MT versus SquareCB is significant. Table 1 gives all endpoint means and standard deviations. UCB and TS are not tested against one another because that comparison was outside the prespecified family.

Table 7: All ten predeclared pooled-seed contrasts. Negative differences favor the proposal. Intervals are unadjusted paired 95% t intervals. p values use exact sign-flip tests with ten-test Holm correction.

Proposal	Baseline	Difference	95% interval	Adjusted p
UCB-MT	FastCB-PLL	-380.5	$[-476.0, -285.1]$	$< 2 \times 10^{-5}$
UCB-MT	SquareCB allocation/PLL	-377.4	$[-469.3, -285.4]$	$< 2 \times 10^{-5}$
UCB-MT	Graph LiM2 adaptation	-1298.9	$[-1461.0, -1136.8]$	$< 2 \times 10^{-5}$
UCB-MT	Graph neural-linear UCB	-919.6	$[-1024.9, -814.2]$	$< 2 \times 10^{-5}$
UCB-MT	Graph neural-linear TS	-1098.6	$[-1252.4, -944.9]$	$< 2 \times 10^{-5}$
TS-MT	FastCB-PLL	-31.3	$[-69.2, 6.5]$	0.201
TS-MT	SquareCB allocation/PLL	-28.1	$[-66.1, 9.9]$	0.201
TS-MT	Graph LiM2 adaptation	-949.7	$[-1066.0, -833.4]$	$< 2 \times 10^{-5}$
TS-MT	Graph neural-linear UCB	-570.4	$[-645.8, -494.9]$	$< 2 \times 10^{-5}$
TS-MT	Graph neural-linear TS	-749.4	$[-862.6, -636.2]$	$< 2 \times 10^{-5}$

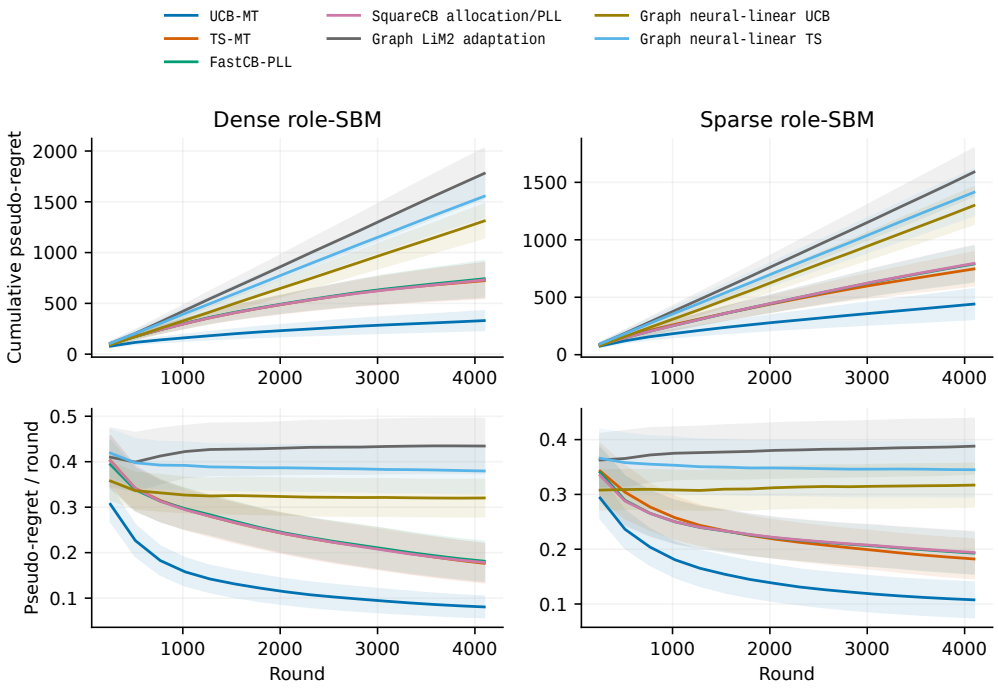

Figure 3: Bayes pseudo-regret for the seven complete methods. The top row shows cumulative pseudo-regret, and the bottom row shows pseudo-regret per round. Columns correspond to the two teacher families. Shading indicates pointwise 95% t intervals over twenty seeds. These are descriptive intervals, not simultaneous bands or paired tests. The final-round analysis averages both families within each seed.

These curves and endpoint contrasts provide finite-horizon evidence for the same Bayes pseudo-regret objective as Theorems 6–7. They do not identify an asymptotic exponent or establish calibration, Gaussian optimism, a uniform Fisher lower bound, or the growth bound on transported information. The separate learned-update mechanism study uses different tuning and seeds. Its final values cannot be compared with this study using paired tests.

D.4 Covariance ablations during learned representation updates

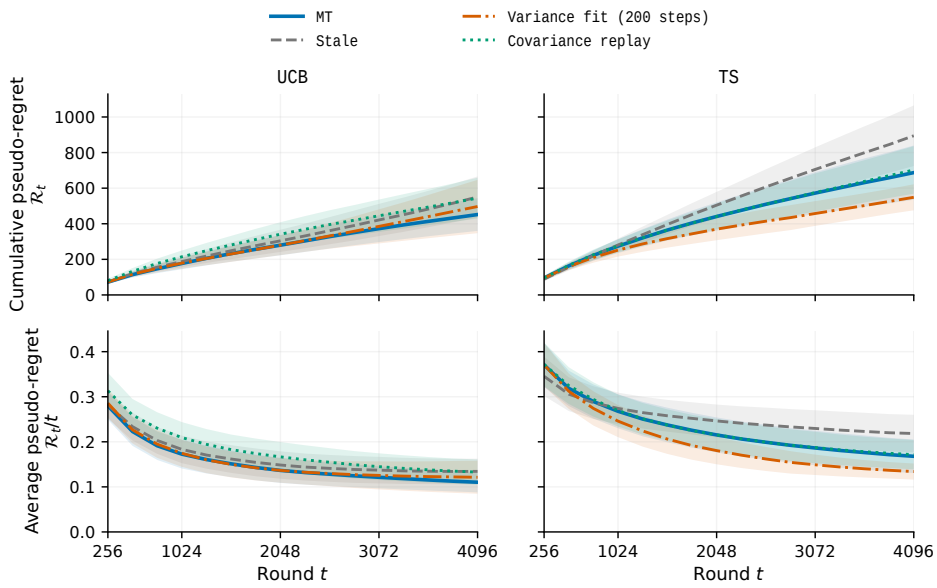

Figure 4: Bayes pseudo-regret of UCB and TS during learned encoder updates. The top row shows cumulative pseudo-regret, and the bottom row shows pseudo-regret per round. All methods use the same GCN–PLL learner with separately tuned exploration scales and no aggregation. Shading indicates pointwise 95% t intervals over twenty independent seeds after averaging the two teacher families within each seed. These are descriptive intervals, not simultaneous bands or paired tests.

D.4.1 Learning model and covariance controls

We test whether covariance transport improves decisions during PLL-driven encoder updates, with exploration scales tuned separately for each method. This study uses the same two-layer GCN, selected-bit mean update, mean transport, and encoder schedule in every method. The covariance choices are MT, stale state, *variance fit*, and *covariance replay*. Variance fit uses 20 projected full-batch steps to match old marginal class-logit variances on 32 fit graphs. The PSD projection and smoothness-based step size ensure that the objective does not increase, although 20 iterations need not reach the optimum. Covariance replay rebuilds an identity-regularized Fisher precision from 128 selected-action rows under the current mean and representation. These controls isolate the covariance update for a shared head. They differ from the complete `LiM2` and full-replay algorithms.

D.4.2 Initial comparison: design

Independent topology teachers generate three-class laws on dense and sparse role-SBM graphs. Teacher probabilities and sampled full labels are evaluator-only. The learner receives its selected correctness bit. Development comprises 96 runs: two seeds, two families, two policies, four covariance choices, and scales $\{0.25, 1, 4\}$, each for 1,024 rounds. Average development pseudo-regret selects the scale for each policy–method pair. Every UCB method selects 4. TS selects 1 except stale, which selects 0.25. These boundary selections limit interpretation to the declared tuning grid. Confirmation uses 20 independent seeds and all $20 \times 2 \times 2 \times 4 = 320$ runs for 4,096 rounds, totaling 1,310,720 confirmation rounds. The encoder rate is 0.1 and the update period is 64. Sample size is fixed by budget, without a claimed power guarantee.

D.4.3 Decision quality on a common history

After feedback assimilation and before each accepted encoder update on an MT trajectory, we retain the old mean and covariance. All four covariance alternatives share the same transported post-update

mean. They are evaluated on eight independent graphs excluded from fitting and feedback. UCB uses deterministic scores. TS uses 256 coupled Gaussian draws. MT uses its natural transported square root. Alternative square roots are aligned to it by orthogonal Procrustes, preserving their marginal covariances. The local comparison uses the MT trajectory and its exploration scale for every covariance choice. In the online experiments, each method uses its own selected scale. The local effect at update u is

$$D_{u,c} = \frac{1}{8}\sum_{j=1}^{8} \boldsymbol{\mu}(G_{u,j})^\top \big(\boldsymbol{q}_{u,\mathrm{MT}}(G_{u,j}) - \boldsymbol{q}_{u,c}(G_{u,j})\big),$$

where $G_{u,j}$ is a diagnostic graph, $\boldsymbol{\mu}(G)$ its true class-probability vector, and $\boldsymbol{q}_{u,c}(G)$ the action distribution under covariance choice c on the common post-update state. For UCB this distribution is a point mass. For TS we estimate the expectation with coupled draws. Positive $D_{u,c}$ favors MT. This quantity measures the immediate effect of replacing the covariance at a shared learner state. It differs from the regret comparison between policies with separate learning histories.

D.4.4 Initial comparison: inference and results

We average the two families within seed and use the 20 seed clusters for paired inference. Six primary MT-minus-control cumulative pseudo-regret contrasts use two-sided sign-flip tests with 99,999 random sign vectors and Holm correction, under paired sign-exchangeability. Six secondary local reward contrasts use a separate Holm family. Unadjusted 95% cluster-bootstrap intervals use 20,000 resamples. Results are reported for all prespecified settings and comparisons.

D.4.5 Refined variance fitting: design and results

After analyzing the first study, a development-only solver assessment found that 20 unconditioned variance-fitting steps left relative fitting residuals between 2.1% and 5.6% on eight development update states. The control was therefore refined before an independent confirmation study: a whitened covariance parameterization and 200 accelerated projected steps reduced those residuals to 0.09–0.22%. Whitening is an invertible congruence change of optimization variable. It does not replace the marginal-variance objective. We retain the incoming covariance if the final objective increases. Achieved residuals are recorded for every fit. The fixed iteration budget does not ensure convergence to the global optimum.

The development grid expands to $\{0.0625, 0.25, 1, 4, 16\}$ (160 runs, including 72 unchanged-method runs reused from the initial development). UCB selects 16 for MT, stale and variance fit, and 4 for covariance replay. TS selects 1 except stale, which selects 0.25. Three UCB selections still reach the upper boundary, so the study remains a finite-grid comparison. The refined confirmation study comprises 320 runs on 20 independent seeds, disjoint from both development and the first confirmation, with the same horizons, paired analysis and separate six-test primary/secondary Holm families. The design is informed by the first study. The second confirmation outcomes are never used for tuning. The tables below report all comparisons from this second study.

Table 8: Covariance controls in both independent studies. Effects are MT minus control: negative favors MT for regret, positive for local expected reward. Parentheses give adjusted p-values. Bold effects survive their study's separate six-test primary (regret) or secondary (local reward) Holm family. Replay denotes covariance replay.

		Online regret difference		Local reward gain (pp)	
Study	Control	UCB	TS	UCB	TS
Initial	Stale	**-106.35** (0.001)	**-159.65** (<0.001)	**0.439** (0.040)	**0.721** (<0.001)
Initial	Variance fit	**-101.81** (0.001)	**-81.97** (<0.001)	**0.434** (0.040)	**0.660** (0.002)
Initial	Replay	**169.80** (<0.001)	8.21 (0.887)	**-1.546** (0.002)	**1.186** (0.002)
Refined	Stale	**-98.77** (0.045)	**-207.16** (0.004)	**4.805** (<0.001)	**3.665** (<0.001)
Refined	Variance fit	-45.09 (1.000)	138.70 (0.05004)	0.088 (0.427)	0.020 (0.427)
Refined	Replay	**-92.29** (0.007)	-14.85 (1.000)	**16.314** (<0.001)	**2.662** (<0.001)

Table 9: Mean cumulative pseudo-regret at $T = 4096$ and local action disagreement in both covariance studies. Bold marks the lowest mean regret within each study and policy, not significance. Disagreement uses the declared same-history coupling, averaged over held-out graphs and updates.

		Cumulative pseudo-regret		Local disagreement (%)	
Study	Covariance	UCB	TS	UCB	TS
Initial	MT	610.13	646.81	1.079	1.417
Initial	Stale	716.48	806.46	2.568	6.198
Initial	Variance fit	711.94	728.78	2.451	6.332
Initial	Replay	**440.33**	**638.60**	5.913	6.655
Refined	MT	**451.73**	687.42	2.314	2.585
Refined	Stale	550.50	894.58	16.914	14.275
Refined	Variance fit	496.82	**548.72**	4.053	7.053
Refined	Replay	544.02	702.27	47.554	12.797

D.4.6 Interpretation and scope

The encoder updates in this study are learned nonlinear transformations. Across the refined covariance fits, the median relative fitting residual is 0.83%, the 90th percentile is 2.23%, and the maximum is 55.05%. Thus the finite anchor set and optimization budget can leave substantial fitting error. These updates differ from the scalar coordinate transformations used in the invariance experiment.

Figure 4 summarizes the confirmation runs from the refined study. For policy $\pi \in \{\mathrm{UCB}, \mathrm{TS}\}$, mechanism c, seed $s \in \{1, \dots, 20\}$ and family $h \in \{1, 2\}$, let $\mathcal{R}^{\pi,c}_{s,h}(t)$ be the saved cumulative pseudo-regret at checkpoint $t \in \{256, 512, \dots, 4096\}$. The plotted seed-level quantity and mean are

$$\overline{\mathcal{R}}^{\pi,c}_{s}(t) = \tfrac{1}{2}\sum_{h=1}^{2} \mathcal{R}^{\pi,c}_{s,h}(t), \qquad \widehat{m}^{\pi,c}(t) = \tfrac{1}{20}\sum_{s=1}^{20} \overline{\mathcal{R}}^{\pi,c}_{s}(t).$$

With $\widehat{s}^{\pi,c}(t)$ the sample standard deviation of these twenty seed-level quantities, the interval is

$$\widehat{m}^{\pi,c}(t) \;\pm\; t_{19,0.975}\widehat{s}^{\pi,c}(t)/\sqrt{20},$$

where $t_{19,0.975}$ is the 97.5th percentile of Student's t distribution with 19 degrees of freedom. Dividing the mean and interval endpoints by t gives the lower panels. The curves are unsmoothed, and no tests are conducted at intermediate checkpoints. Endpoint means are given in Table 9.

The refined UCB and TS regret reductions against stale covariance are 98.77 and 207.16, with corrected $p = 0.045$ and 0.00366, respectively. The corresponding local reward gains are 4.805 and 3.665 percentage points. The comparison with covariance replay depends on the study and policy. Under UCB, replay has lower regret in the first study, whereas MT has lower regret in the separately tuned second study. Local replay substitutions also differ from independently deployed replay.

With the refined variance fitter, neither refined local reward contrast against variance fit is significant (both corrected $p = 0.427$). For online TS, variance fit has lower mean regret (548.72 versus 687.42 for MT). The MT-minus-variance difference is 138.70, with unadjusted 95% interval $[40.29, 249.71]$ and corrected Monte Carlo $p = 0.05004$. Exact enumeration of all 2^{20} sign assignments, reported as a numerical sensitivity check, gives corrected $p = 0.05014$. Neither value establishes equivalence between the methods. UCB versus variance fit is also inconclusive. These results support the benefit of updating uncertainty relative to leaving covariance stale. They do not establish superiority over the refined variance fitter. Action disagreement alone also does not measure decision quality.

The conclusions are limited to three-class simulations, one shared encoder rate, the tested exploration grid and a 4,096-round horizon. They do not isolate joint covariance as the unique cause of an improvement over marginal-variance fitting, establish real-data superiority, or prove asymptotic regret for an unrestricted GCN. The results above include both solver configurations and their independent evaluations.

D.5 Transport invariance and local stability

D.5.1 Exact feature transformations

A separate study fixes an independently initialized GCN with hidden widths $(8, 6)$ and trains its PLL head from selected bits. After round 128, all feature coordinates are multiplied by $c \in \{1/4, 1, 4\}$. MT divides the head by c and the covariance by c^2. Stale divides only the head. Fixed coordinates provide the coupled reference. This is an exact coordinate intervention, not a learned nonlinear encoder update. The teacher is independently generated from graph topology. We evaluate both UCB and TS at all three scales under each of the three covariance controls. Three development seeds precede 20 independent confirmation seeds at 16,384 rounds. The two primary MT-minus-stale effects average all three scales within seed. Because these are working covariances, expansion can sometimes improve stale learning. The fixed-coordinate reference is not assumed to be statistically optimal.

The fixed-coordinate and transported learners produce identical action sequences in every paired comparison, as predicted by Theorem 1 under the compatible updates of Corollary 2. The study uses twenty seeds for each policy and feature scale. Figure 5 compares regret with stale covariance. Although MT preserves the reference policy exactly, that policy need not be statistically optimal.

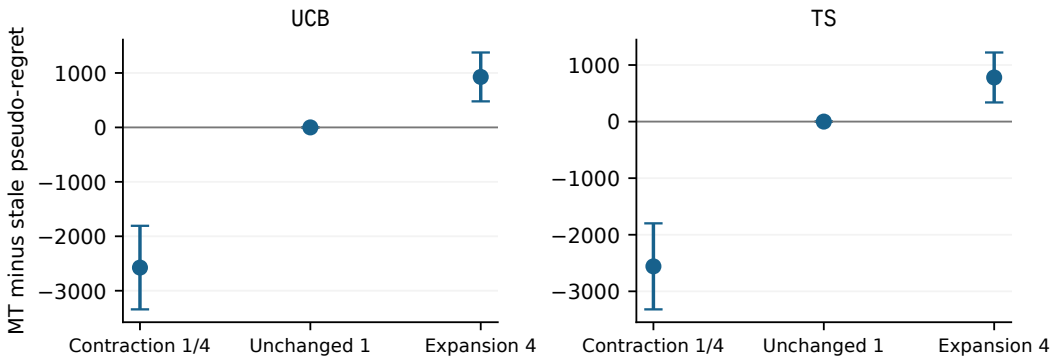

Figure 5: Effect of exact feature rescaling at each of the three scales. Points are paired mean differences over 20 seeds. Bars are marginal 95% intervals. Negative favors MT. These intervals are descriptive by scale. The two primary tests average all three scales within seed.

D.5.2 Predictive-variance continuity

An additional development diagnostic replays three 2,048-round fresh-topology streams for each of the six learned IX variants. At each scheduled encoder update, it clones the incoming adviser, assimilates the same selected bit in old coordinates, and compares its predictive variances with the actual post-update state on the last 32 anchors. Thus the comparison isolates the representation/reconciliation step from ordinary assimilation. The diagnostic reproduces every original development action and introduces no confirmation observations. The mean relative variance change is 0.070% for `UCB-MT` and 0.069% for `TS-MT`, compared with 2.471% and 2.463% for stale state. Replay gives 17.527% and 17.605%, respectively. Replay rebuilds the information matrix, so its larger change is not itself evidence of worse calibration. These measurements assess variance continuity on the anchor graphs. They do not establish a uniform approximation bound or confidence coverage.

D.5.3 Numerical verification of the perturbation inequalities

A numerical experiment evaluates Proposition 4 on 48 fixed Gaussian states: $K \in \{3, 6\}$, residual scales $\{0, 0.01, 0.05, 0.2\}$, mean margins $\{0.05, 1\}$, and three seeds. Each state uses 20,000 coupled Gaussian samples and 20,000 deterministic score pairs. All 960,000 TS draws satisfy the stated event inclusion, and all 960,000 UCB pairs satisfy the deterministic margin implication. Zero-residual cases preserve every action. This experiment checks the implementation of the local inequalities. It does not evaluate GCN performance or confidence coverage.

D.6 Graph-context generalization and the scope of aggregation

This study separates graph-dependent advice from learning a fixed recurring context and tests whether aggregation improves the direct decision rule. Recurring pools, fresh topology and unrelated random labels distinguish transfer from memorization. We evaluate aggregation as an empirical extension of the direct methods, separately from the covariance ablations.

D.6.1 Aggregation rule

The labels `UCB-MT-IX` and `TS-MT-IX` refer to an `EXP4-IX` aggregation adaptation (Neu, 2015). The underlying adviser is the same PLL-trained GCN with MT. The aggregate also uses a fixed graph partition and the GCN probability vector. The regret analysis in this paper covers the direct methods rather than this aggregation extension.

Fix a partition $h : \mathcal{G} \to [S]$, where S is the number of partition cells, before play. Let $\boldsymbol{Q}_s \in \mathbb{R}^K$ store estimated class losses in cell s and define three advice vectors

$$\boldsymbol{q}_{t,\mathrm{part}} = \mathrm{softmax}(-\eta \boldsymbol{Q}_{h(G_t)}), \qquad \boldsymbol{q}_{t,\mathrm{soft}} = \widehat{\boldsymbol{p}}_t, \qquad \boldsymbol{q}_{t,\mathrm{explore}} = \boldsymbol{e}_{a_t^{\mathrm{rule}}}.$$

The positive rate η and implicit-exploration offset β are fixed by

$$N_0 = K + 2, \qquad C_{\mathrm{IX}} = \max\{S \log K + \log(N_0/K), \log N_0\},$$
$$\eta = \sqrt{C_{\mathrm{IX}}/(KT)}, \qquad \beta = \eta/2.$$

Initialize $\boldsymbol{Q}_s = \boldsymbol{0}$ and group log weights $v_{\mathrm{part}} = \log(K/N_0)$, $v_{\mathrm{soft}} = v_{\mathrm{explore}} = -\log N_0$. At each round, normalize these three log weights into $w_{t,j}$ by softmax, and draw the actual class from

$$\boldsymbol{p}_t = \sum_{j \in \{\mathrm{part},\mathrm{soft},\mathrm{explore}\}} w_{t,j} \boldsymbol{q}_{t,j}, \qquad A_t \sim \boldsymbol{p}_t. \tag{27}$$

Here $p_{t,a}$ is an action probability. $\widehat{p}_{t,a}$ remains the GCN's class probability. Observe only R_t and form

$$\widetilde{\boldsymbol{\ell}}_t = \frac{(1 - R_t)\boldsymbol{e}_{A_t}}{p_{t,A_t} + \beta}, \qquad \boldsymbol{Q}'_{h(G_t)} = \boldsymbol{Q}_{h(G_t)} + \widetilde{\boldsymbol{\ell}}_t. \tag{28}$$

For $\mathrm{LSE}(\boldsymbol{v}) = \log \sum_a e^{v_a}$, update

$$v'_{\mathrm{part}} = v_{\mathrm{part}} + \mathrm{LSE}(-\eta \boldsymbol{Q}'_{h(G_t)}) - \mathrm{LSE}(-\eta \boldsymbol{Q}_{h(G_t)}),$$
$$v'_j = v_j - \eta \boldsymbol{q}_{t,j}^\top \widetilde{\boldsymbol{\ell}}_t, \qquad j \in \{\mathrm{soft}, \mathrm{explore}\}.$$

Replace the current-cell loss vector and group log weights by these updated values. The GCN assimilates the actual selected bit using Algorithm 1. The other cells are unchanged. The partition mass implements the tested prior with equal weight on the K^S fixed class maps within that group. Its initial total mass is K/N_0. The comparisons include MT, stale and replay advisers, uniform-advice aggregation, and the partition-only `EXP3-IX` control.

D.6.2 Teacher families and evaluation protocol

Synthetic graphs have role-dependent block structure. The independent teacher uses two centered cross-role edge-density coordinates, with logits $4.5\boldsymbol{H}\boldsymbol{x}(G)$, where the three rows of $\boldsymbol{H}$ are directions separated by $2\pi/3$ and rotated by an independently seeded phase. This varies the optimal class with graph topology. Graphs and label uniforms are paired across methods. The learner receives neither the teacher parameters nor a predictor family known to contain the true conditional distribution. The graph learner starts from its own independent GCN parameters.

Two conditions repeatedly sample an unlabeled pool of 16 or 64 graphs. The observable graph hash identifies a partition cell. The pool size is public and all graph keys are distinct in the evaluated streams. The conditional law is fixed for each graph. A third condition uses fresh topology graphs and a single master cell. Its Bayes approximation term need not vanish. It evaluates advice transfer to new graphs. A fourth condition assigns independent uniformly random deterministic labels to distinct new graphs before play. Its Bayes loss is zero for the fixed assigned function, but no learner can predict its independent labels systematically. This is the negative control for transfer to unrelated unseen labels.

The synthetic methods are six learned IX variants, IX with two uniform advice rows, partition-only EXP3-IX, and the two direct MT policies without aggregation. The two non-GCN controls distinguish useful graph advice from generic master learning. The uniform-advice control matches the learned master's group priors exactly. Development uses three separate seeds and 2,048 rounds. Confirmation uses 20 independent seeds, four conditions, ten methods, and 16,384 rounds: 800 runs. Primary effects average the two recurring-pool conditions within each seed. The fresh-graph and random-label conditions separately assess generalization to unseen graphs.

D.6.3 PSEUDO-REGRET AND AGGREGATION EFFECTS

Table 10 reports final cumulative pseudo-regret across recurring, fresh-topology and random-label conditions. The direct TS-MT learner has the lowest mean in the three structured conditions. This experiment does not include direct replay or the stronger variance fitter. Tables 11 and 12 report the declared comparison families. On real graphs, aggregation increases mistakes relative to direct MT by 5.90 and 5.09 percentage points for UCB and TS.

Table 10: All synthetic endpoint pseudo-regrets. The random-label condition is an impossibility control and is not ranked.

Bold indicates the best unrounded column mean and does not denote statistical significance.

Method	Pool16	Pool64	Fresh topology	Random labels
UCB-MT-IX	360.58	457.34	579.94	10919.80
UCB-stale-IX	388.44	447.73	583.16	10927.15
UCB-replay-IX	388.64	484.81	629.34	10911.25
TS-MT-IX	366.04	478.72	617.75	10910.50
TS-stale-IX	380.09	477.99	627.68	10903.15
TS-replay-IX	490.41	648.05	810.70	10918.30
IX-uniform	1123.46	2253.37	8215.30	10917.30
EXP3-IX	1112.49	2250.24	8212.20	10912.00
UCB-MT	268.32	306.36	323.24	10919.20
TS-MT	**213.48**	**243.27**	**288.61**	10937.10

Table 11: Declared synthetic comparison family. Effects are cumulative pseudo-regret differences. Negative favors the first method.

Bold indicates Holm-adjusted $p < 0.05$, for either sign of the effect.

Contrast	Δ	95% CI	Simultaneous CI	Holm p
UCB-MT-IX – IX-uniform	**-1279.46**	[-1459.39, -1099.53]	[-1543.70, -1015.22]	$< 10^{-4}$
TS-MT-IX – IX-uniform	**-1266.04**	[-1447.60, -1084.47]	[-1532.68, -999.39]	$< 10^{-4}$
UCB-MT-IX – UCB-stale-IX	-9.13	[-28.06, 9.80]	[-36.92, 18.66]	0.6507
TS-MT-IX – TS-stale-IX	-6.65	[-27.87, 14.56]	[-37.81, 24.50]	0.6507
UCB-MT-IX – UCB-replay-IX	-27.77	[-60.31, 4.77]	[-75.56, 20.02]	0.2700
TS-MT-IX – TS-replay-IX	**-146.85**	[-212.07, -81.62]	[-242.63, -51.06]	0.0008
UCB-MT-IX – UCB-MT	**121.62**	[50.40, 192.84]	[17.03, 226.21]	0.0081
TS-MT-IX – TS-MT	**194.00**	[123.85, 264.16]	[90.97, 297.03]	$< 10^{-4}$

Table 12: Declared real-data comparison family. Effects are mistake-rate differences in percentage points. Negative favors the first method.

Bold indicates Holm-adjusted $p < 0.05$, for either sign of the effect.

Contrast	Δ	95% CI	Simultaneous CI	Holm p
UCB-MT-IX – IX-uniform	**-7.20**	[-7.69, -6.72]	[-7.91, -6.49]	$< 10^{-4}$
TS-MT-IX – IX-uniform	**-6.64**	[-7.15, -6.13]	[-7.40, -5.88]	$< 10^{-4}$
UCB-MT-IX – UCB-stale-IX	0.00	[-0.11, 0.12]	[-0.17, 0.17]	1.0000
TS-MT-IX – TS-stale-IX	-0.07	[-0.24, 0.11]	[-0.32, 0.19]	1.0000
UCB-MT-IX – UCB-replay-IX	-0.07	[-0.36, 0.21]	[-0.49, 0.34]	1.0000
TS-MT-IX – TS-replay-IX	**-1.12**	[-1.45, -0.79]	[-1.61, -0.64]	$< 10^{-4}$
UCB-MT-IX – UCB-MT	**5.90**	[5.36, 6.43]	[5.11, 6.68]	$< 10^{-4}$
TS-MT-IX – TS-MT	**5.09**	[4.55, 5.63]	[4.30, 5.88]	$< 10^{-4}$

D.7 Covariance fitting after transport

We test whether fitting the covariance more accurately improves decisions. This empirical extension adds a correction step to MT. We evaluate its fitting objective, online regret, immediate reward and computational cost separately.

D.7.1 Correction objective and implementation

For nonlinear encoder changes, let $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{\Sigma}_{\mathrm{MT}}$ denote the transported covariance and keep the transported mean fixed. We initialize $\boldsymbol{R}_0 = \boldsymbol{I}$ and take n_{corr} projected-gradient steps for

$$\min_{(1-\rho)\boldsymbol{I} \preceq \boldsymbol{R} \preceq (1+\rho)\boldsymbol{I}} F(\boldsymbol{R}) := \mathcal{L}_{\mathrm{anc}}(\boldsymbol{R}) + \lambda_{\mathrm{cor}} \|\boldsymbol{R} - \boldsymbol{I}\|_F^2, \qquad \boldsymbol{\Sigma}_{\mathrm{corr}} = \boldsymbol{S}\boldsymbol{R}\boldsymbol{S}^\top. \tag{29}$$

Here $0 < \rho < 1$, $\lambda_{\mathrm{cor}} > 0$, and $\mathcal{L}_{\mathrm{anc}}$ matches joint class-logit covariances and selected-bit log-odds variances on paired anchor graphs as specified below. We accept only steps that do not increase F. Both UCB and TS use the resulting covariance. $n_{\mathrm{corr}} = 0$ recovers MT. The labels `MT+5` and `MT+20` denote five and twenty correction steps within the same GCN method.

Anchor objective and projected iterations. All quantities here are evaluated after assimilating the current selected bit. The old encoder/head state has covariance $\boldsymbol{\Sigma}_{\mathrm{old}}$ and mean $\widehat{\boldsymbol{\theta}}_{\mathrm{old}}$. The candidate encoder and the transported mean $\widehat{\boldsymbol{\theta}}_{\mathrm{MT}} = \boldsymbol{P}\widehat{\boldsymbol{\theta}}_{\mathrm{old}}$ are then fixed during covariance correction. Let $\boldsymbol{S} = \boldsymbol{P}\boldsymbol{S}_{\mathrm{old}}$, where $\boldsymbol{S}_{\mathrm{old}}\boldsymbol{S}_{\mathrm{old}}^\top = \boldsymbol{\Sigma}_{\mathrm{old}}$. For paired anchor graph G_i, let $\boldsymbol{Z}_i^-$ and $\boldsymbol{Z}_i^+$ be its old/new logit design matrices from (6). Let $\boldsymbol{Q}_i^-$ and $\boldsymbol{Q}_i^+$ stack the old/new selected-bit log-odds derivatives $\boldsymbol{d}_{i,k}^\top$ over all classes, evaluated at the corresponding fixed means. The learner computes these derivatives without access to labels or oracle rewards. Define

$$\begin{aligned} \boldsymbol{A}_i &= \boldsymbol{Z}_i^+ \boldsymbol{S}, & \boldsymbol{B}_i &= \boldsymbol{Q}_i^+ \boldsymbol{S}, \\ \boldsymbol{C}_i &= \boldsymbol{Z}_i^- \boldsymbol{\Sigma}_{\mathrm{old}} (\boldsymbol{Z}_i^-)^\top, & \boldsymbol{v}_i &= \operatorname{diag}\big(\boldsymbol{Q}_i^- \boldsymbol{\Sigma}_{\mathrm{old}} (\boldsymbol{Q}_i^-)^\top\big). \end{aligned}$$

For m_{anc} anchors and a positive numerical floor ε, use fixed weights

$$w_i = [m_{\mathrm{anc}}(\|\boldsymbol{C}_i\|_F^2 + \varepsilon)]^{-1}, \qquad u_i = [m_{\mathrm{anc}}(\|\boldsymbol{v}_i\|_2^2 + \varepsilon)]^{-1}.$$

The common UCB/TS fitting loss is the convex quadratic

$$\mathcal{L}_{\mathrm{anc}}(\boldsymbol{R}) = \sum_{i=1}^{m_{\mathrm{anc}}} w_i \|\boldsymbol{A}_i \boldsymbol{R} \boldsymbol{A}_i^\top - \boldsymbol{C}_i\|_F^2 + \sum_{i=1}^{m_{\mathrm{anc}}} u_i \|\operatorname{diag}(\boldsymbol{B}_i \boldsymbol{R} \boldsymbol{B}_i^\top) - \boldsymbol{v}_i\|_2^2. \tag{30}$$

The first term matches joint logit covariances. The second targets the uncertainty used by our selected-bit UCB rule. Their positive weights are fixed before optimization. Let $\boldsymbol{E}_i = \boldsymbol{A}_i \boldsymbol{R} \boldsymbol{A}_i^\top - \boldsymbol{C}_i$ and $\boldsymbol{r}_i = \operatorname{diag}(\boldsymbol{B}_i \boldsymbol{R} \boldsymbol{B}_i^\top) - \boldsymbol{v}_i$. The gradient of F in (29) is

$$\nabla F(\boldsymbol{R}) = 2 \sum_i w_i \boldsymbol{A}_i^\top \boldsymbol{E}_i \boldsymbol{A}_i + 2 \sum_i u_i \boldsymbol{B}_i^\top \operatorname{Diag}(\boldsymbol{r}_i) \boldsymbol{B}_i + 2\lambda_{\mathrm{cor}}(\boldsymbol{R} - \boldsymbol{I}).$$

The implementation uses the step-size denominator

$$L_F = 2\sum_i w_i \|\boldsymbol{A}_i\|_{\mathrm{op}}^4 + 2\sum_i u_i \|\boldsymbol{B}_i\|_{\mathrm{op}}^4 + 2\lambda_{\mathrm{cor}}.$$

Starting from $\boldsymbol{R}_0 = \boldsymbol{I}$, compute

$$\boldsymbol{R}_{j+1} = \Pi_{[1-\rho,1+\rho]}\left(\boldsymbol{R}_j - L_F^{-1}\nabla F(\boldsymbol{R}_j)\right),$$

where the projection symmetrizes the matrix and clips its eigenvalues to the stated interval. Stop if the computed objective would increase. The implementation returns the last accepted iterate. Five or twenty steps need not reach the global optimum. The experiment fixes $m_{\mathrm{anc}} = 32$, $\varepsilon = 10^{-8}$, $\lambda_{\mathrm{cor}} = 0.01$, and $\rho = 0.5$ before tuning exploration scales. Only the step budget distinguishes `MT+5` from `MT+20`.

D.7.2 Evaluation protocol and results

We compare MT, the 200-step variance fitter from the refined study, and `MT+5`/`MT+20`. All four share the same GCN, selected-bit PLL update, mean transport, encoder rate 0.1 and update period 64. Only covariance reconciliation changes. Both policies use the same hybrid objective and fixed $(\rho, \lambda_{\mathrm{cor}}) = (0.5, 0.01)$. We evaluate the direct methods without IX aggregation. The variance-fitting control uses the same head as MT and therefore differs from the complete published `LiM2` algorithm.

Development uses 384 runs: two independent development seeds, dense/sparse topology families, three/five classes, UCB/TS, four variants, and six scales $\{0.0625, 0.25, 1, 4, 16, 64\}$ at horizon 1,024. Five-class teachers use five equally spaced directional class heads over the same realized edge statistics. We do not assume equally frequent optimal classes. Each policy/variant/class count selects its scale by mean development pseudo-regret across both seeds and graph families. UCB selects 16 for three classes and 4 for five classes in every variant. TS selects 0.0625 except the five-class variance fitter, which selects 0.25. Most TS selections therefore reach the lower grid boundary, leaving open whether smaller scales would perform better.

The prespecified confirmation contains $20 \times 2 \times 2 \times 2 \times 4 = 640$ runs and 2,621,440 rounds, with 20 independent seeds and $T = 4,096$. We average the four graph/class conditions within seed. All eight policy/budget contrasts against MT and variance fitting are primary, using exact paired sign-flip tests over 2^{20} assignments and a single eight-test Holm family. This inference assumes paired sign-exchangeability. Unadjusted 95% paired bootstrap intervals use 20,000 resamples. The sample size is determined by the computational budget rather than a specified power target.

A separate, predeclared four-test noninferiority family compares each hybrid with variance fitting using margin 0.01 in average pseudo-regret per round, one-sided paired t tests and Holm correction. This requires the corresponding seed-level sampling approximation. A nonsignificant difference is not treated as equivalence. We report results for both correction budgets and all conditions.

Same-history diagnostics separately replay the 160 MT trajectories, requiring agreement with their confirmed online regret. At every 512th round with an accepted encoder update, all alternatives use the same transported mean and MT exploration scale on eight diagnostic graphs excluded from fitting and feedback. TS uses 256 coupled draws with covariance-preserving Procrustes factor alignment. Four hybrid-minus-MT local reward contrasts form a separate Holm family. These are substitutions on MT histories, not additional independent online comparisons. Learner runtime excludes this diagnostic replay and includes prediction, PLL assimilation, encoder updates and all reconciliation work. An isolated single-worker timing check follows the parallel studies.

Results. Across 20,479 hybrid correction updates, there are no recorded objective-descent or spectral constraint violations. `MT+5` and `MT+20` reduce the regularized fitting objective by 7.21% and 13.20% on average, respectively. All scheduled correction steps are accepted. The measurements confirm that the optimization steps decrease the objective and satisfy the spectral constraints. Confidence calibration requires a separate statistical assessment.

Mean cumulative pseudo-regret, averaged over the four conditions, is 503.26/501.44/537.76/546.38 for UCB with MT/variance fit/`MT+5`/`MT+20`, and 763.70/730.79/780.94/781.19 for TS in the same order. Thus neither correction budget improves the pooled mean over MT or variance fitting. Under TS, `MT+5` increases regret relative to MT by 17.24, with unadjusted 95% interval $[2.52, 35.78]$ and

eight-test Holm $p = 0.04895$. The remaining seven primary differences are not significant after correction. This does not establish equivalence. None of the four declared noninferiority tests against variance fitting passes: unadjusted one-sided upper bounds range from 2.19 to 2.91 percentage points, above the 1-point margin.

The implemented covariance also preconditions subsequent PLL mean updates. Consequently, an online regret difference does not isolate exploration from learning dynamics. Same-history substitutions test immediate action effects separately. Lower anchor loss, smaller covariance mismatch and lower regret are distinct outcomes. Neither `MT+5` nor `MT+20` improves the pooled mean regret over MT. The finite scale grid, frequent lower-boundary TS selections, two development seeds, fixed correction hyperparameters and finite horizon delimit this result.

Local effects and computational cost. All 160 diagnostic MT trajectories reproduce the original online regret, yielding 1,280 accepted-update snapshots. No hybrid-minus-MT local reward contrast survives the separate four-test Holm family (all adjusted $p = 1$). UCB action disagreement from MT is 0.156% for both budgets. Coupled TS disagreement is 0.00351%/0.00549%. Under the selected exploration scales, additional moment fitting rarely changes the action on these MT histories. These findings do not identify the complete mechanism behind the online regret differences.

The isolated timing experiment uses one worker and 48 runs of 1,024 rounds. Across policy and budget, full learner time is 1.11–1.63 times MT and 0.51–0.57 times the 200-step variance fitter. Parallel confirmation gives ratios 1.32–1.66 and 0.42–0.57, respectively. The correction methods are faster than the variance fitter in this implementation, but their regret does not satisfy the prespecified noninferiority criterion.

Table 13: Independent correction study: mean cumulative pseudo-regret at 4,096 rounds. Bold marks the lowest descriptive mean in each column, not significance. Each entry averages two graph families and 20 seeds.

	UCB		TS	
Method	$K = 3$	$K = 5$	$K = 3$	$K = 5$
`MT`	451.24	**555.29**	755.20	772.20
`Variance fit`	**432.63**	570.26	**690.70**	**770.88**
`MT+5`	475.26	600.27	789.64	772.25
`MT+20`	489.40	603.37	786.23	776.15

Table 14: All eight primary correction contrasts. Effects are hybrid minus control, so negative favors the hybrid. Intervals are unadjusted 95% paired bootstrap intervals. Bold effects survive the single eight-test Holm family.

Policy	Contrast	Effect	95% interval	Holm p
`UCB`	`MT+5` vs `MT`	34.50	$[8.56, 67.24]$	0.0604
`UCB`	`MT+5` vs `Variance fit`	36.32	$[-20.66, 96.62]$	0.6895
`UCB`	`MT+20` vs `MT`	43.12	$[13.84, 78.69]$	0.0604
`UCB`	`MT+20` vs `Variance fit`	44.94	$[-12.11, 107.14]$	0.6895
`TS`	`MT+5` vs `MT`	**17.24**	$[2.52, 35.78]$	0.0490
`TS`	`MT+5` vs `Variance fit`	50.15	$[-22.10, 123.57]$	0.6895
`TS`	`MT+20` vs `MT`	17.49	$[1.05, 44.91]$	0.2032
`TS`	`MT+20` vs `Variance fit`	50.40	$[-24.46, 125.61]$	0.6895