Bilinear Matching Bandits

Wooseong Cho
Seoul National University
wooseong_cho@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

We study *bilinear matching bandits*, an online sequential decision problem in which at each round a learner assigns M users to N items via an injective matching $\pi : [M] \hookrightarrow [N]$ and observes semi-bandit feedback with bilinear mean reward $\mathbf{x}_i^\top \mathbf{\Theta}^\star \mathbf{y}_{\pi(i)}$, where the parameter $\mathbf{\Theta}^\star \in \mathbb{R}^{d_1 \times d_2}$ is unknown. We propose HYBRIDRRD, a hybrid algorithm that combines high-width elimination with a row-wise regularized design sampler over the fractional matching polytope. Without assuming $\mathbf{\Theta}^\star$ is low-rank, HYBRIDRRD achieves a high-probability regret bound of $\widetilde{\mathcal{O}}(d_2\sqrt{d_1 MT} + Md_1d_2)$, suppressing problem-scale constants. This improves the leading dimension dependence from d_1d_2 to $d_2\sqrt{d_1}$ over the standard Kronecker-linearized contextual combinatorial semi-bandit baseline. We further prove a minimax lower bound of $\Omega(d_2\sqrt{d_1 MT})$ in a large-item regime, matching the leading regret term up to logarithmic factors. Numerical experiments show that HYBRIDRRD performs favorably against Kronecker-linearized baselines.

1 Introduction

A wide range of online platforms repeatedly assign users to capacity-constrained items, providers, or content slots. In these settings, the value of a match depends on how user-side and item-side attributes interact, while the parameters of this interaction must be learned from sequential feedback. We study this problem through a model we call *bilinear matching bandits*. At each round, the learner observes user contexts in $\mathbb{R}^{d_1}$ and item contexts in $\mathbb{R}^{d_2}$, chooses an injective matching of M users to $N \geq M$ items, and observes one semi-bandit reward per matched pair. The mean reward of matching user i to item j is $\mathbf{x}_{t,i}^\top \mathbf{\Theta}^\star \mathbf{y}_{t,j}$ for an unknown shared matrix parameter $\mathbf{\Theta}^\star$.

This formulation is related to several existing lines of work, but none directly answers the dimension-dependence question studied here. Classical online bipartite matching and budgeted allocation [19, 2, 31] study online arrivals and competitive ratios against offline benchmarks, rather than regret for learning an unknown shared reward function. Bandit learning in matching markets [29, 36, 13] studies uncertain preferences or match rewards, but typically focuses on stable or equilibrium-oriented market outcomes rather than dimension-dependent minimax regret over feature spaces.

Contextual combinatorial semi-bandits with linear payoff functions [35, 39] study combinatorial arm selection under minimax regret and can be applied to our setting by treating each user–item edge as an arm with Kronecker feature $\mathbf{z}_{t,i,j} = \mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i} \in \mathbb{R}^{d_1 d_2}$. The resulting Kronecker-linearized baseline yields a regret bound with leading term $\widetilde{\mathcal{O}}(d_1 d_2\sqrt{MT})$.

Low-rank bilinear bandit methods exploit the rank structure of the unknown parameter [17], while action-space analysis can improve regret without a low-rank assumption [15]. Unlike our setting, these works select a single pair of arms rather than an injective matching with M semi-bandit observations. Whether this d_1d_2 dependence can be improved for arbitrary bounded $\mathbf{\Theta}^\star$, without low-rank or other structural assumptions on the parameter matrix, is our main question.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The challenge goes beyond the matrix-valued parameter. A single round selects an entire matching, so the M observed edge rewards are generated by a jointly constrained action rather than by independent arm pulls. To obtain a semi-bandit gain, the algorithm must turn these M observations into useful covariance growth, but the covariance must be useful for comparing whole matchings rather than isolated edges. At the same time, exploration cannot be completely unstructured in the full $d_1 d_2$-dimensional Kronecker feature space, as this would bring back the product-dimension cost. An effective design must provide information for comparing competing matchings while accounting for their predicted rewards.

Our approach is to separate exploration into two regimes. When some matching has large aggregate uncertainty, HYBRIDRRD plays the width-maximizing matching to reduce uncertainty directly. When all feasible matchings have small width, the algorithm uses a row-wise regularized design (RRD) sampler over the fractional matching polytope. The RRD objective promotes item-span coverage for each user while favoring higher predicted reward. The resulting fractional design is implemented by sampling feasible matchings through a Birkhoff distribution. Thus, although the estimator is maintained in the full Kronecker feature space, the exploration design and covariance certificate exploit the row-wise item-span structure of the bilinear reward.

Our contributions are summarized as follows.

- **Row-wise regularized design for matching bandits.** We introduce HYBRIDRRD, whose low-width component optimizes a reward-regularized design over the fractional matching polytope. Unlike a global design over Kronecker edge features, RRD promotes item-span coverage for each user while respecting the matching constraints. The RRD optimality condition and row-wise covariance decomposition yield a certificate that jointly controls aggregate matching uncertainty and predicted reward differences. The certificate depends on the item span dimension $q_t \le d_2$, rather than the full Kronecker dimension $d_1 d_2$.
- **Dimension-improved regret without low-rank structure.** We prove that HYBRIDRRD achieves the high-probability regret bound

$$\widetilde{\mathcal{O}}\left(\sigma_\eta d_2 \sqrt{d_1 M T} + B M d_1 d_2\right)$$

  without any low-rank restriction on $\mathbf{\Theta}^\star$ and for arbitrary context sequences. The leading regret term improves the Kronecker-linearized dimension dependence from $d_1 d_2$ to $d_2\sqrt{d_1}$.
- **Sharpness of the leading term.** We prove a minimax lower bound

$$\Omega(\sigma_\eta d_2 \sqrt{d_1 M T})$$

  in the large-item regime $N \ge M 2^{d_2}$, under the stated parameter-radius condition. The construction uses deterministic contexts and independent Gaussian noise, placing it within the same model class as the upper bound. Thus the leading regret term is sharp up to logarithmic factors under these conditions.
- **Empirical evaluation.** We complement the theory with synthetic experiments on bilinear matching instances. HYBRIDRRD performs favorably against Kronecker-linearized UCB and Thompson-sampling baselines in finite-horizon simulations.

2 Related Work

Online matching and learning in matching markets. Classical online matching [19], stochastic-reward matching [31], and budgeted allocation [2] study online arrivals and competitive ratios against offline benchmarks. Dynamic matching models study stochastic arrivals and service systems under type structure [12]. A separate line brings bandit learning into matching markets, focusing on learning preferences, stable outcomes, equilibria, or welfare-oriented objectives [29, 36, 13, 32]. Recent work further studies stable regret and contextual preferences in matching markets [23, 28]. Closer to our setting, Sentenac et al. [38] study matching bandits with semi-bandit feedback under a rank-one edge reward matrix, while Kasy and Teytelboym [20] analyze Thompson sampling for capacity-constrained matching with a prior-independent Bayesian regret bound. Other matching bandit models study queue stability under preference feedback [21]. In contrast, we study minimax regret for injective matchings under a shared bilinear reward model without a rank constraint on the unknown parameter.

Contextual combinatorial semi-bandits. The combinatorial bandit and semi-bandit literature studies how to select multiple arms per round and observe their individual or aggregated outcomes [4–6, 25, 7, 41]. Contextual extensions share reward information across arms through linear feature models [35, 42, 39] or more general supervised-learning oracle reductions [24]. Applied directly to our model, these methods treat user–item pairs as $d_1 d_2$-dimensional Kronecker arms. Our algorithm retains a Kronecker ridge estimator but uses the bilinear row-wise item-span structure in its exploration design and covariance certificate.

Bilinear and matrix bandits. Low-rank bilinear bandits exploit the rank structure of the unknown parameter [17]. Related work extends this approach to low-rank matrix actions and generalized reward models [30, 18]. In contrast, Jang et al. [15] improve regret bounds through bilinear action-space analysis without a low-rank assumption on the parameter. Recent work also develops matrix estimation and experimental design methods for low-rank bandits and generalized trace regression [16, 27]. These studies do not address our matching constraints and semi-bandit feedback. Our improvement comes from a row-wise covariance certificate without imposing a rank constraint on $\mathbf{\Theta}^\star$.

Linear partial monitoring. Our problem can also be viewed as contextual linear partial monitoring, where each matching has an aggregate reward feature and an observation operator formed by its selected edge features. General linear partial monitoring methods study the relation between information and regret [22]. In our setting, the improved dimension dependence relies on the bilinear row structure and the RRD covariance certificate rather than a direct application of generic partial monitoring bounds.

Design-based exploration and semi-bandit dependence. Our regret analysis uses self-normalized linear bandit tools [1]. The RRD sampler is related to experimental design for pure-exploration linear bandits [10], but incorporates predicted rewards and matching constraints into a design used for cumulative regret minimization. Dependence among simultaneously observed rewards is important in combinatorial semi-bandits [9, 33], while covariance-adaptive algorithms exploit estimates of the reward covariance [34]. Our RRD instead optimizes the expected design matrix of selected edge features, while Assumption 2 controls within-round reward noise.

3 Preliminaries

3.1 Notation

For an integer $n \geq 1$, write $[n] = \{1, \dots, n\}$. For an event $\mathcal{E}$, let $\mathbb{1}\{\mathcal{E}\}$ denote its indicator. All vectors are column vectors. For a vector $\mathbf{u}$, let $\|\mathbf{u}\|_2$ denote its Euclidean norm. For a matrix $\mathbf{A}$, let $\|\mathbf{A}\|_{\mathrm{op}}$ and $\|\mathbf{A}\|_{\mathrm{F}}$ denote its operator and Frobenius norms, respectively. We write $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ for the matrix inner product.

For symmetric matrices, $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. For a positive semidefinite matrix $\mathbf{A}$, let $\mathbf{A}^\dagger$ denote its Moore–Penrose pseudoinverse, and define $\|\mathbf{u}\|_{\mathbf{A}}^2 = \mathbf{u}^\top \mathbf{A} \mathbf{u}$ and $\|\mathbf{u}\|_{\mathbf{A}^\dagger}^2 = \mathbf{u}^\top \mathbf{A}^\dagger \mathbf{u}$. We write $\mathrm{range}(\mathbf{A})$ for the column space of $\mathbf{A}$, $\mathrm{rank}(\mathbf{A})$ for its rank, and $\mathrm{tr}(\mathbf{A})$ for its trace. The Kronecker product is denoted by $\otimes$, and $\mathrm{vec}(\cdot)$ stacks the columns of a matrix.

3.2 Problem Formulation

Model. At each round $t \in [T]$, the learner observes user contexts $\mathbf{x}_{t,1}, \dots, \mathbf{x}_{t,M} \in \mathbb{R}^{d_1}$ and item contexts $\mathbf{y}_{t,1}, \dots, \mathbf{y}_{t,N} \in \mathbb{R}^{d_2}$, where $N \geq M$. Let $C_t = (\mathbf{x}_{t,1:M}, \mathbf{y}_{t,1:N})$ denote the round-t context. After observing C_t, the learner chooses an injective matching $\pi_t : [M] \hookrightarrow [N]$. The reward of matching user i to item $\pi_t(i)$ is

$$r_{t,i} = \mathbf{x}_{t,i}^\top \mathbf{\Theta}^\star \mathbf{y}_{t,\pi_t(i)} + \eta_{t,i},$$

where $\mathbf{\Theta}^\star \in \mathbb{R}^{d_1 \times d_2}$ is unknown and $\eta_{t,i}$ denotes the reward noise. The learner observes all M selected rewards, corresponding to semi-bandit feedback. We impose no distributional assumption on the context sequence $(C_t)_{t=1}^T$. For any $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, define the matching-level mean reward

$$F_{\mathbf{\Theta}}(C_t, \pi) = \sum_{i=1}^{M} \mathbf{x}_{t,i}^\top \mathbf{\Theta} \mathbf{y}_{t,\pi(i)}.$$

Objective. Let $\pi_t^\star \in \operatorname{argmax}_{\pi:[M]\hookrightarrow[N]} F_{\boldsymbol{\Theta}^\star}(C_t, \pi)$ be an optimal matching under the true parameter. We measure performance by the cumulative pseudo-regret

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left[F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t^\star) - F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t) \right].$$

Linearization. Since the bilinear reward is linear in $\boldsymbol{\theta}^\star = \mathrm{vec}(\boldsymbol{\Theta}^\star)$, define the Kronecker feature

$$\mathbf{z}_{t,i,j} := \mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i} \in \mathbb{R}^p, \qquad p = d_1 d_2.$$

Then $\mathbf{x}_{t,i}^\top \boldsymbol{\Theta}^\star \mathbf{y}_{t,j} = \mathbf{z}_{t,i,j}^\top \boldsymbol{\theta}^\star$. For a vector $\boldsymbol{\theta} \in \mathbb{R}^p$, write $F_{\boldsymbol{\theta}}(C_t, \pi) = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi(i)}^\top \boldsymbol{\theta}$.

Matchings. Let $\Pi = \{\pi : [M] \hookrightarrow [N]\}$ be the set of feasible integral matchings. For a matching $\pi \in \Pi$, define its aggregate feature vector and semi-bandit design matrix by

$$\mathbf{a}_t(\pi) = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi(i)}, \qquad \mathbf{S}_t(\pi) = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi(i)} \mathbf{z}_{t,i,\pi(i)}^\top.$$

The aggregate feature determines the matching-level mean reward, $F_{\boldsymbol{\theta}}(C_t, \pi) = \mathbf{a}_t(\pi)^\top \boldsymbol{\theta}$, whereas $\mathbf{S}_t(\pi)$ represents the design contribution of the M individually observed edge features under semi-bandit feedback. For a score matrix $\mathbf{C} \in \mathbb{R}^{M \times N}$, let

$$\mathrm{Match}(\mathbf{C}) \in \operatorname*{argmax}_{\pi \in \Pi} \sum_{i=1}^{M} \mathbf{C}_{i,\pi(i)},$$

with ties broken by a fixed deterministic rule.

Fractional matching polytope. For each $\pi \in \Pi$, let $\mathbf{X}_\pi \in \{0,1\}^{M \times N}$ be its incidence matrix, $(\mathbf{X}_\pi)_{i,j} = \mathbb{1}\{\pi(i) = j\}$. Thus each row of $\mathbf{X}_\pi$ sums to one and each column sums to at most one. We define the fractional matching polytope of nonnegative matrices as

$$\mathcal{M} = \left\{ \mathbf{P} \in \mathbb{R}_+^{M \times N} : \sum_{j=1}^{N} \mathbf{P}_{i,j} = 1 \;\; \forall i \in [M], \quad \sum_{i=1}^{M} \mathbf{P}_{i,j} \le 1 \;\; \forall j \in [N] \right\}.$$

The rectangular Birkhoff property, proved for completeness in Lemma 8, identifies this polytope with the convex hull of injective matching incidence matrices:

$$\mathcal{M} = \mathrm{conv}\{\mathbf{X}_\pi : \pi \in \Pi\}.$$

Consequently, every $\mathbf{P} \in \mathcal{M}$ can be realized as the edge-marginal matrix of a randomized integral matching: there exists a distribution $\zeta_{\mathbf{P}}$ over Π such that

$$\mathbf{P} = \mathbb{E}_{\pi \sim \zeta_{\mathbf{P}}}[\mathbf{X}_\pi], \qquad \mathbf{P}_{i,j} = \mathbb{P}_{\pi \sim \zeta_{\mathbf{P}}}\{\pi(i) = j\} \quad \forall i \in [M],\ j \in [N].$$

Thus $\mathbf{P}_{i,j}$ is the probability that the randomized integral matching $\pi \sim \zeta_{\mathbf{P}}$ assigns user i to item j.

3.3 Assumptions

Filtration. Let $\mathcal{F}_t$ be the sigma-field generated by all observations before the learner chooses the round-t matching, together with the current context C_t. After the algorithmic randomization has been realized and the selected matching π_t is fixed, let

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(\pi_t).$$

For the analysis, enumerate the M selected rewards in the fixed order $i = 1, \dots, M$ and define

$$\mathcal{H}_{t,0} = \mathcal{G}_t, \qquad \mathcal{H}_{t,i} = \mathcal{G}_t \vee \sigma(r_{t,1}, \dots, r_{t,i}) = \mathcal{G}_t \vee \sigma(\eta_{t,1}, \dots, \eta_{t,i}).$$

This refined filtration is used to state within-round martingale conditions.

Assumption 1 (Bounded contexts and bounded parameter)**.** *There exist constants $L_x, L_y, S > 0$ such that, for all $t \in [T]$, $i \in [M]$, and $j \in [N]$,*

$$\|\mathbf{x}_{t,i}\|_2 \le L_x, \qquad \|\mathbf{y}_{t,j}\|_2 \le L_y, \qquad \|\boldsymbol{\Theta}^\star\|_F \le S.$$

Define

$$L = L_x L_y, \qquad B = SL_x L_y = SL.$$

It follows that $|\mathbf{z}_{t,i,j}^\top \boldsymbol{\theta}^\star| \le B$ for all t, $i \in [M]$, and $j \in [N]$.

Assumption 2 (Noise)**.** *There exists $\sigma_\eta > 0$ such that, for every t, every $i \in [M]$, and every $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\eta_{t,i} \mid \mathcal{H}_{t,i-1}] = 0, \qquad \mathbb{E}\left[\exp(\lambda \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] \le \exp\left(\frac{\lambda^2 \sigma_\eta^2}{2}\right).$$

Remark 3 (Within-round martingale condition)**.** The semi-bandit feedback in each round consists of the M selected rewards $r_{t,1}, \ldots, r_{t,M}$. The order $i = 1, \ldots, M$ is a fixed deterministic enumeration used only to state the within-round martingale condition; it is not part of the learner's observation protocol. The condition does not require conditional independence among the M same-round noises, but it does impose sequential martingale structure within a round.

By iterated conditioning, Assumption 2 implies that, for $\boldsymbol{\eta}_t = (\eta_{t,1}, \ldots, \eta_{t,M})$ and every $\mathcal{G}_t$-measurable $\mathbf{u} \in \mathbb{R}^M$,

$$\mathbb{E}\left[\exp\left(\mathbf{u}^\top \boldsymbol{\eta}_t\right) \mid \mathcal{G}_t\right] \le \exp\left(\frac{\sigma_\eta^2 \|\mathbf{u}\|_2^2}{2}\right). \tag{1}$$

This ℓ_2-scale vector bound is used for our self-normalized martingale analysis to retain the $\sqrt{M}$ semi-bandit gain after flattening the M per-round observations into regression samples.

A marginal coordinate-wise post-action sub-Gaussian condition alone does not imply (1) under arbitrary same-round dependence. One needs additional within-round structure, such as conditional independence, a sequential martingale condition as in Assumption 2, or a direct joint conditional vector sub-Gaussian assumption such as (1). Appendix D gives a quantitative comparison and a scalar common-shock example showing that the marginal-only condition is insufficient for a uniform $\sqrt{M}$ semi-bandit gain under arbitrary same-round dependence.

4 Algorithm

We now describe HYBRIDRRD (Algorithm 1). The algorithm maintains a ridge estimator in the Kronecker feature space, but uses different exploration rules depending on the current maximum matching-level width. High-width rounds play a width-maximizing matching, whereas low-width rounds invoke a row-wise regularized design sampler over the fractional matching polytope.

Hybrid exploration rule. The algorithm maintains the ridge design matrix $\mathbf{V}_t = \lambda \mathbf{I}_p + \sum_{s<t} \mathbf{S}_s(\pi_s)$ and response vector $\mathbf{b}_t = \sum_{s<t} \sum_{i=1}^M \mathbf{z}_{s,i,\pi_s(i)} r_{s,i}$. At each round t, it computes the ridge estimator $\boldsymbol{\theta}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$ and the width-score matrix $\mathbf{W}_t \in \mathbb{R}^{M \times N}$ with entries

$$(\mathbf{W}_t)_{i,j} = \mathbf{z}_{t,i,j}^\top \mathbf{V}_t^{-1} \mathbf{z}_{t,i,j},$$

measuring the statistical uncertainty of the predicted reward for user-item pair (i, j). The algorithm then runs in two branches determined by the *aggregate width*. Let $\pi_t^w = \mathrm{Match}(\mathbf{W}_t)$ be the width-maximizing matching, and define the aggregate width

$$\omega_t = \sum_{i=1}^M (\mathbf{W}_t)_{i,\pi_t^w(i)} = \max_{\pi \in \Pi} \sum_{i=1}^M (\mathbf{W}_t)_{i,\pi(i)}.$$

Equivalently, $\omega_t = \mathrm{tr}(\mathbf{V}_t^{-1} \mathbf{S}_t(\pi_t^w))$. Thus ω_t is the maximum total uncertainty over feasible matchings. When $\omega_t > \rho$, the algorithm is in the *high-width* regime and plays π_t^w. The selected edges add information in directions that remain poorly covered by the ridge design. Such rounds may also occur later when new context directions appear. When $\omega_t \le \rho$, every feasible matching has width at most ρ, and the algorithm enters the *low-width* regime. It forms predicted scores $(\widehat{\mathbf{R}}_t)_{i,j} = \mathbf{z}_{t,i,j}^\top \boldsymbol{\theta}_t$ and uses RRD to balance predicted reward and item coverage.

RRD sampler. The low-width branch calls $\mathrm{RRD}_t(\widehat{\mathbf{R}}_t, \xi, \varepsilon)$, a certified sampler that returns a distribution μ_t over feasible matchings. It is defined as follows. Given a predicted-score matrix $\widehat{\mathbf{R}}_t \in \mathbb{R}^{M\times N}$, let $q_t = \dim \mathrm{span}\{\mathbf{y}_{t,1}, \ldots, \mathbf{y}_{t,N}\}$. If $q_t = 0$, then all item contexts lie in the zero-dimensional span, so $\mathbf{z}_{t,i,j} = \mathbf{0}$ for every edge. In this degenerate case no fractional design is needed. Otherwise, proceed as follows. Choose $\mathbf{U}_t \in \mathbb{R}^{d_2\times q_t}$ with orthonormal columns spanning the item span and set $\widetilde{\mathbf{y}}_{t,j} = \mathbf{U}_t^\top \mathbf{y}_{t,j}$. For $\mathbf{P} \in \mathcal{M}$, define

$$\mathbf{V}_{t,i}(\mathbf{P}) = \sum_{j=1}^{N} (\mathbf{P})_{i,j} \widetilde{\mathbf{y}}_{t,j} \widetilde{\mathbf{y}}_{t,j}^\top.$$

Let $\mathcal{D}_t = \{\mathbf{P} \in \mathcal{M} : \mathbf{V}_{t,i}(\mathbf{P}) \succ 0 \text{ for all } i \in [M]\}$. For each user, the candidate edge features share the same user context and differ through the item contexts. Thus $\mathbf{V}_{t,i}(\mathbf{P})$ measures item coverage for user i. The matching constraints couple these designs across users. The historical ridge design $\mathbf{V}_t$ determines whether the round enters the low-width regime. RRD then constructs a design using the current item contexts and predicted scores. The RRD objective is

$$\Phi_t(\mathbf{P}) = \sum_{i=1}^{M} \log\det \mathbf{V}_{t,i}(\mathbf{P}) + \xi \sum_{i=1}^{M}\sum_{j=1}^{N} (\mathbf{P})_{i,j} (\widehat{\mathbf{R}}_t)_{i,j}.$$

The log determinant terms form a row-wise D-optimal design criterion, promoting coverage of the current item span, while the linear term favors higher predicted reward. For each user, the marginal-weighted item leverage scores sum to q_t. The Frank–Wolfe certificate combines this identity with predicted reward differences to control the leverage of every competing matching. The resulting row-wise bound is converted into a matching-level covariance certificate in Lemma 5. The sampler computes $\mathbf{P}_t \in \mathcal{D}_t$ satisfying

$$g_t(\mathbf{P}_t) := \max_{\mathbf{Q}\in\mathcal{M}} \langle \nabla\Phi_t(\mathbf{P}_t), \mathbf{Q} - \mathbf{P}_t\rangle \le \varepsilon.$$

The Frank–Wolfe gap measures the largest improvement in the linearized objective over feasible fractional matchings. An exact maximizer exists and has zero gap by Lemma 7. By the rectangular Birkhoff decomposition, let $\zeta_{\mathbf{P}_t}$ be any edge-marginal representation of $\mathbf{P}_t$, i.e., a distribution over injective matchings satisfying

$$\mathbb{P}_{\pi\sim\zeta_{\mathbf{P}_t}}\{\pi(i) = j\} = (\mathbf{P}_t)_{i,j}, \qquad \forall i \in [M], j \in [N].$$

Both the expected predicted reward and the expected design matrix depend only on these edge marginals. Thus, sampling from $\zeta_{\mathbf{P}_t}$ realizes the fractional design through feasible matchings. Finally, let $\pi_0 \in \Pi$ be any fixed feasible matching. The sampler $\mathrm{RRD}_t(\widehat{\mathbf{R}}_t, \xi, \varepsilon)$ returns

$$\mu_t = \begin{cases} \delta_{\pi_0}, & q_t = 0, \\ \zeta_{\mathbf{P}_t}, & q_t > 0. \end{cases}$$

Computational aspects. On low-width rounds with $q_t > 0$, the Frank–Wolfe linear oracle is a maximum weight matching problem with edge scores given by $\nabla\Phi_t(\mathbf{P})$. The matching returned by the linear oracle also gives the Frank–Wolfe gap $g_t(\mathbf{P})$, so checking the stopping criterion requires no additional matching solve. Forming and factorizing the row covariance matrices and evaluating all leverage scores takes $O(MNq_t^2 + Mq_t^3)$ time. The rectangular Hungarian algorithm takes $O(M^2N)$ time for the matching oracle. The gradient evaluation and linear oracle therefore cost $O(MNq_t^2 + Mq_t^3 + M^2N)$ per iteration.

With the leverage scores already available, the matrix determinant lemma allows each trial step in a line search to be evaluated in $O(M)$ time. The inverse ridge design can be maintained through sequential rank-one updates in $O(Mp^2)$ time per round. Sampling from the final fractional design requires an additional Birkhoff decomposition.

Standard Frank–Wolfe convergence bounds depend on the curvature of the objective [14]. This curvature need not be bounded over $\mathcal{D}_t$, since the row covariance matrices can approach singularity. We therefore use the computable Frank–Wolfe gap as a stopping certificate, without claiming a uniform bound on the required iterations.

Algorithm 1 HYBRIDRRD

Require: Horizon T, regularization $\lambda > 0$, threshold $\rho > 0$, RRD reward weight $\xi > 0$, and RRD tolerance $\varepsilon \geq 0$

Initialize $\mathbf{V}_1 = \lambda \mathbf{I}_p$ ($p = d_1 d_2$), $\mathbf{b}_1 = \mathbf{0}$, and $\boldsymbol{\theta}_1 = \mathbf{0}$
for $t = 1, 2, \ldots, T$ **do**
  Observe $C_t = (\mathbf{x}_{t,1:M}, \mathbf{y}_{t,1:N})$
  Form $\mathbf{z}_{t,i,j} = \mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i}$ for all $i \in [M]$, $j \in [N]$
  Compute $\boldsymbol{\theta}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$
  Compute $(\mathbf{W}_t)_{i,j} = \mathbf{z}_{t,i,j}^\top \mathbf{V}_t^{-1} \mathbf{z}_{t,i,j}$ for all $i \in [M]$, $j \in [N]$
  Set $\pi_t^w = \mathrm{Match}(\mathbf{W}_t)$ and $\omega_t = \sum_{i=1}^M (\mathbf{W}_t)_{i,\pi_t^w(i)}$
  if $\omega_t > \rho$ **then**
    Set $\pi_t = \pi_t^w$
  else
    Set $(\widehat{\mathbf{R}}_t)_{i,j} = \mathbf{z}_{t,i,j}^\top \boldsymbol{\theta}_t$ for all $i \in [M]$, $j \in [N]$
    Compute $\mu_t = \mathrm{RRD}_t(\widehat{\mathbf{R}}_t, \xi, \varepsilon)$
    Sample $\pi_t \sim \mu_t$
  Play π_t and observe $r_{t,1:M}$
  Update $\mathbf{V}_{t+1} = \mathbf{V}_t + \mathbf{S}_t(\pi_t)$
  Update $\mathbf{b}_{t+1} = \mathbf{b}_t + \sum_{i=1}^M \mathbf{z}_{t,i,\pi_t(i)} r_{t,i}$

5 Regret Analysis

5.1 Regret Upper Bound

Setup. Fix $\delta \in (0,1)$ and let $p = d_1 d_2$. Recall that $\omega_t = \max_{\pi \in \Pi} \mathrm{tr}(\mathbf{V}_t^{-1} \mathbf{S}_t(\pi))$, and define

$$\mathcal{A} = \{t \in [T] : \omega_t > \rho\}, \qquad \mathcal{B} = [T] \setminus \mathcal{A}.$$

Also let

$$K_T = p \log\left(1 + \frac{MTL^2}{\lambda p}\right), \quad \ell_\delta = \max\left\{1, \log\frac{cT}{\delta}\right\}, \quad A_T = \lambda S^2 + \sigma_\eta^2 (K_T + \ell_\delta),$$

where $c > 0$ is a sufficiently large numerical constant.

Theorem 4 (Regret upper bound)**.** *Suppose Assumption 1 and Assumption 2 hold. Run Algorithm 1 with*

$$\lambda = \frac{L^2}{T}, \quad \rho = \frac{c_\rho}{\ell_\delta}, \quad \xi = c_\xi \sqrt{\frac{(Md_2 + \varepsilon)T}{A_T}},$$

where $c_\rho, c_\xi > 0$ are numerical constants chosen as in the proof. Suppose that the RRD oracle returns an ε-certificate on every low-width round with $q_t > 0$, with $0 \leq \varepsilon \leq Md_2$. Then there exist numerical constants $C, C' > 0$ such that, with probability at least $1 - \delta$,

$$\mathrm{Reg}(T) \leq C\sqrt{(Md_2 + \varepsilon)TA_T} + C'BM(K_T + 1)\ell_\delta.$$

Consequently,

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sigma_\eta d_2 \sqrt{d_1 MT} + BMd_1 d_2\right).$$

Proof overview. The key challenge is to use information from individual edge observations to compare the mean rewards of competing matchings. We separate rounds according to their maximum matching width. For high-width rounds, a determinant-potential argument gives

$$|\mathcal{A}| \leq \frac{2K_T}{\rho}, \qquad \mathrm{Reg}_{\mathcal{A}}(T) \leq \frac{4BMK_T}{\rho}.$$

On low-width rounds, let $\mathbf{H}_t = \mathbb{E}_{\pi \sim \mu_t} \mathbf{S}_t(\pi)$ denote the expected design matrix. Then

$$\mathbf{S}_t(\pi) \preceq \rho \mathbf{V}_t \quad \text{for all } \pi \in \Pi, \qquad \mathbf{H}_t \preceq \rho \mathbf{V}_t.$$

The following RRD certificate uses row-wise item-span coverage to jointly control aggregate matching uncertainty and predicted reward differences.

Lemma 5 (RRD certificate)**.** *Fix a low-width round $t \in \mathcal{B}$ and let $\mathbf{H}_t = \mathbb{E}_{\pi\sim\mu_t}\mathbf{S}_t(\pi)$. If $q_t = 0$, then $\mathbf{a}_t(\pi) = \mathbf{0}$ and $\mathbf{H}_t = \mathbf{0}$ for every feasible matching $\pi \in \Pi$. If $q_t > 0$, suppose that $\mathbf{P}_t \in \mathcal{D}_t$ satisfies $g_t(\mathbf{P}_t) \le \varepsilon$, with $\mu_t = \zeta_{\mathbf{P}_t}$ as defined in Algorithm 1. Then, for every feasible matching $\pi \in \Pi$,*

$$\|\mathbf{a}_t(\pi)\|^2_{\mathbf{H}_t^\dagger} + \xi\left(\sum_{i=1}^{M}(\widehat{\mathbf{R}}_t)_{i,\pi(i)} - \langle \mathbf{P}_t, \widehat{\mathbf{R}}_t\rangle\right) \le Mq_t + \varepsilon,$$

and $\mathbf{a}_t(\pi) \in \mathrm{range}(\mathbf{H}_t)$.

The factor Mq_t reflects the item span rather than the full Kronecker dimension. For each nonzero user context, $\mathbf{x}_{t,i}^\top(\mathbf{x}_{t,i}\mathbf{x}_{t,i}^\top)^\dagger\mathbf{x}_{t,i} = 1$, while a zero context contributes no edge feature. The marginal-weighted item leverage scores sum to q_t for each user, yielding $Mq_t \le Md_2$ across all users. To convert this certificate into regret, define the estimation error $\mathbf{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star$ and the low-width prediction error $E_t = \mathbf{e}_t^\top\mathbf{H}_t\mathbf{e}_t$. For $t \in \mathcal{B}$, also define the conditional one-step regret

$$\overline{\mathrm{reg}}_t = F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t^\star) - \mathbb{E}\left[F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t) \mid \mathcal{F}_t\right].$$

Let $\mathbf{a}_t^\mu = \mathbb{E}[\mathbf{a}_t(\pi_t) \mid \mathcal{F}_t]$ be the mean aggregate feature. Since $\mathbf{a}_t(\pi)\mathbf{a}_t(\pi)^\top \preceq M\mathbf{S}_t(\pi)$ for every feasible matching π, we have $\mathbf{a}_t^\mu \in \mathrm{range}(\mathbf{H}_t)$ and $\|\mathbf{a}_t^\mu\|^2_{\mathbf{H}_t^\dagger} \le M$. Applying Lemma 5 and then generalized Cauchy–Schwarz and Young's inequality yields the one-step regret bound in Lemma 15,

$$\overline{\mathrm{reg}}_t \le \frac{\xi}{2}E_t + \frac{2(Md_2+\varepsilon)}{\xi}.$$

The ridge design is updated using the matching actually played, while E_t measures prediction error under its conditional expected design $\mathbf{H}_t$. The ridge error recursion and noise concentration then give

$$\sum_{t\in\mathcal{B}} E_t \le CA_T$$

with high probability. Therefore, with $\overline{\mathrm{Reg}}_{\mathcal{B}}(T) = \sum_{t\in\mathcal{B}}\overline{\mathrm{reg}}_t$,

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) \le C\left\{\xi A_T + \frac{(Md_2+\varepsilon)T}{\xi}\right\}.$$

Choosing $\xi \asymp \sqrt{(Md_2+\varepsilon)T/A_T}$ gives

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) \le C\sqrt{(Md_2+\varepsilon)TA_T}.$$

The asymmetric dimension dependence follows from the Md_2 contribution of the RRD certificate and estimation in the full Kronecker space of dimension $p = d_1d_2$. These factors yield the leading stochastic term $\sigma_\eta d_2\sqrt{d_1MT}$, up to logarithmic factors. The additive term comes from the high-width rounds and retains the full product dimension. Ignoring logarithmic factors, the leading stochastic term dominates the additive term when $T \gtrsim (B^2/\sigma_\eta^2)Md_1$. With $\rho = c_\rho/\ell_\delta$, the high-width and low-width bounds control cumulative conditional regret. Since low-width matchings are sampled randomly, a final martingale comparison converts this bound into the realized pseudo-regret bound in Theorem 4, as shown in Lemma 18. The complete proofs and probability bounds are given in Appendices A and B.

5.2 Regret Lower Bound

Theorem 6 (Lower bound)**.** *Suppose $d_1, d_2, M \ge 1$, $T \ge 2d_1$, and $N \ge M2^{d_2}$. Let $T_0 = \lfloor T/d_1 \rfloor$, $L = L_xL_y$, and suppose*

$$S \ge \frac{\sigma_\eta d_2}{4L}\sqrt{\frac{d_1}{MT_0}}.$$

Then there exists a numerical constant $c > 0$ such that, for every possibly randomized algorithm, there exists an instance satisfying Assumption 1 and Assumption 2 such that

$$\mathbb{E}_{\mathcal{I},\mathsf{Alg}}\,\mathrm{Reg}(T) \ge c\sigma_\eta d_2\sqrt{d_1MT}. \tag{2}$$

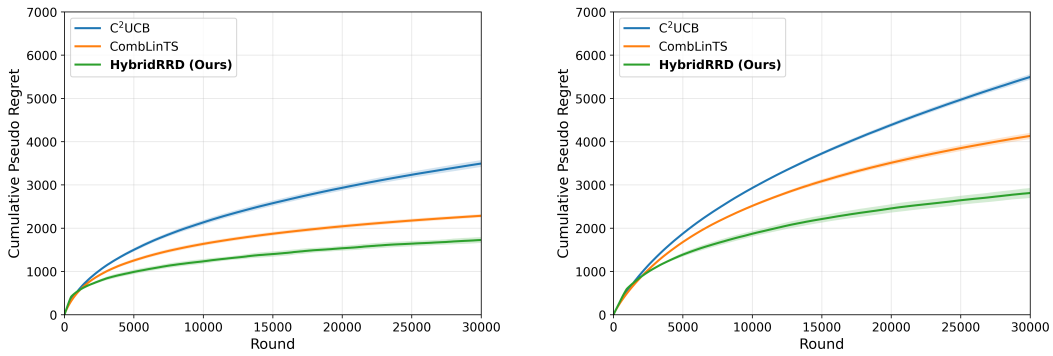
Consequently, $\inf_{\mathsf{Alg}}\sup_{\mathcal{I}}\mathbb{E}\,\mathrm{Reg}(T) = \Omega(\sigma_\eta d_2\sqrt{d_1MT})$ *over this large-item subclass.*

Proof sketch. The lower bound embeds $d_1 d_2$ independent binary testing problems into the bilinear parameter. The item set contains M copies of every sign vector $\mathbf{y}(\mathbf{s}) = L_y \mathbf{s}/\sqrt{d_2}$, so an optimal sign vector can be assigned to all M users simultaneously. During the first $d_1 T_0$ rounds, the user contexts cycle through the d_1 coordinate directions. In rounds assigned to row g, all users have context $L_x \mathbf{e}_g$, so those rounds reveal information only about the g-th row of the unknown matrix. The remaining user contexts are zero.

For each sign matrix $\mathbf{\Omega} \in \{-1, +1\}^{d_1 \times d_2}$, set $\mathbf{\Theta}_{\mathbf{\Omega}} = \Delta \mathbf{\Omega}$, where $\Delta = (\sigma_\eta/(4L))\sqrt{d_2/(MT_0)}$ and $T_0 = \lfloor T/d_1 \rfloor$. The radius condition ensures feasibility. Under $\mathbf{\Theta}_{\mathbf{\Omega}}$, the optimal item sign in row g is the g-th row of $\mathbf{\Omega}$, and each incorrectly chosen sign bit causes regret of order $L\Delta/\sqrt{d_2}$ per user.

It remains to show that each bit is hard to identify. For two parameters differing in one bit of row g, the KL chain rule involves only the T_0 rounds assigned to row g, with M observations per round. With the chosen Δ, the KL divergence between the full interaction histories is $1/8$. Pinsker's inequality and the standard testing lower bound then give a constant error probability for each bit. Summing the resulting per-bit regret over $d_1 d_2$ bits, M users, and T_0 repetitions yields $\mathbb{E}\,\mathrm{Reg}(T) \geq c\sigma_\eta d_2 \sqrt{d_1 MT}$. Full details are in Appendix C.

6 Numerical Experiments



(a) $(d_1, d_2) = (20, 20)$ (b) $(d_1, d_2) = (40, 20)$

Figure 1: Cumulative pseudo-regret on arbitrary-context bilinear matching instances

We evaluate HybridRRD on synthetic bilinear matching bandit instances with bounded contexts. We use independent Gaussian reward noise with standard deviation $\sigma_\eta = 0.2$. We set $M = 10$, $N = 30$, $L_x = L_y = 1$, $S = 1$, and $T = 30000$, and average over five random seeds. We consider two representative dimension settings, $(d_1, d_2) = (20, 20)$ and $(d_1, d_2) = (40, 20)$, with all other environment parameters fixed.

We compare HybridRRD with two Kronecker-linearized baselines. C^{2}UCB [35] uses the Kronecker edge features $\mathbf{z}_{t,i,j} = \mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i}$, with edge-wise optimistic scores and the same matching oracle. This serves as our UCB-type baseline; after the same Kronecker-linearized reduction, CombLinUCB [42] gives a closely related edge-wise linear UCB rule. We also include CombLinTS [42] as a randomized counterpart in the same Kronecker-linearized model. All methods receive the same semi-bandit feedback, and exact configurations are provided in the supplementary code.

Figure 1 reports cumulative pseudo-regret against the optimal matching under the true parameter, averaged over five random seeds; shaded regions denote one standard deviation across seeds. HybridRRD performs favorably in both settings, and the gap remains visible when d_1 is increased while d_2 is fixed. We further examine dimension dependence by varying $d_1 \in \{10, 20, 40, 80\}$ with $d_2 = 20$ and $d_2 \in \{20, 25, 30\}$ with $d_1 = 40$. Table 1 reports the final cumulative pseudo-regret at $T = 30000$. HybridRRD achieves the lowest mean regret in all six reported configurations. For $d_1 = 40$, its regret reduction relative to CombLinTS remains around 32% as d_2 increases from 20 to 30. All experiments were run on a dual-socket Intel Xeon Silver 4214R machine with 24 physical cores and 376 GB RAM.

Table 1: Final cumulative pseudo-regret at $T = 30000$ for the dimension sweeps, reported as mean $\pm$ standard deviation over five random seeds.

(d_1, d_2)	C^{2}UCB	CombLinTS	HYBRIDRRD
$(10, 20)$	1734.2 ± 66.1	1121.3 ± 23.6	$\mathbf{1073.6 \pm 47.9}$
$(20, 20)$	3476.6 ± 78.6	2288.3 ± 74.6	$\mathbf{1673.3 \pm 103.6}$
$(40, 20)$	5448.2 ± 48.8	4156.6 ± 64.3	$\mathbf{2816.0 \pm 110.1}$
$(80, 20)$	6789.5 ± 89.9	6061.6 ± 41.9	$\mathbf{4343.1 \pm 47.2}$
$(40, 25)$	6257.7 ± 95.3	4936.7 ± 67.5	$\mathbf{3341.5 \pm 100.5}$
$(40, 30)$	6666.1 ± 93.0	5496.3 ± 88.6	$\mathbf{3733.5 \pm 115.8}$

7 Limitations and Societal Impact

HYBRIDRRD is statistically efficient, but its low-width branch requires solving a row-wise regularized design problem over the fractional matching polytope. This design step is the main computational overhead beyond the Kronecker ridge estimator used by the baselines, and scaling it to large matching instances without losing the leading regret term remains open.

The regret bound also contains an additive $\widetilde{\mathcal{O}}(BMd_1d_2)$ term from high-width rounds, which may dominate at shorter horizons. Moreover, the lower bound establishes sharpness of the leading term only in the large-item regime $N \geq M \cdot 2^{d_2}$ under the stated parameter-radius condition.

Our model treats context vectors as given. In deployed matching platforms, these vectors may encode biased, sensitive, or privacy-relevant information, so regret minimization alone does not ensure fairness or privacy. Addressing such concerns is outside the scope of the present theory.

8 Conclusion

We studied a matching bandit problem in which the learner chooses an entire injective matching, rather than a single edge or a collection of independent edge pulls. This distinction is important. Feedback is observed at the edge level, but the benchmark and the relevant uncertainty are both defined at the level of whole matchings. Thus the learner has to use semi-bandit observations on individual matched edges to reduce uncertainty about whole matching decisions. A direct Kronecker linearization makes the reward model compatible with existing contextual combinatorial semi-bandit algorithms, but does not by itself exploit the row-wise structure of the bilinear reward model.

To exploit this structure, we propose HYBRIDRRD. In high-width rounds, the algorithm reduces the largest matching-level uncertainty directly. In low-width rounds, it constructs a fractional matching that balances item-span coverage for each user with predicted reward, while respecting the matching constraints. The algorithm then samples an injective matching whose edge marginals are given by this fractional solution. The key step in the analysis is a row-wise covariance certificate that controls the aggregate uncertainty of competing matchings. Although the ridge estimator operates in the full d_1d_2-dimensional space, the certificate depends on the item-span dimension $q_t \leq d_2$ through the leverage structure of each user row. This separation allows the algorithm to improve the leading dimension dependence without assuming that $\mathbf{\Theta}^\star$ is low-rank.

We establish a high-probability regret bound of $\widetilde{\mathcal{O}}(\sigma_\eta d_2\sqrt{d_1MT} + BMd_1d_2)$ for HYBRIDRRD, improving the leading dimension dependence from d_1d_2 to $d_2\sqrt{d_1}$ over the Kronecker-linearized baseline. We also prove a minimax lower bound matching the leading term in the large-item regime. The synthetic experiments show lower cumulative regret than Kronecker-linearized baselines.

Our work also suggests several open questions. One is to develop a more computationally efficient method that preserves the sharp leading regret term. Another is to extend the row-wise design viewpoint developed here to generalized bilinear rewards, capacitated matching constraints, and dynamic matching models.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023-00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008). This research was also supported by the Basic Science Research Program through the NRF, funded by the Ministry of Education (No. RS-2025-25421486).

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Gagan Aggarwal, Gagan Goel, Chinmay Karande, and Aranyak Mehta. Online vertex-weighted bipartite matching and single-bid budgeted allocations. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1253–1264. SIAM, 2011.

[3] Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Revista Facultad de Ciencias Exactas, Puras y Aplicadas Universidad Nacional de Tucumán, Serie A (Matemáticas y Física Teórica)*, 5:147–151, 1946.

[4] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[5] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

[6] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17 (50):1–33, 2016.

[7] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X.

[9] Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. *Advances in Neural Information Processing Systems*, 29, 2016.

[10] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

[11] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

[12] Ming Hu and Yun Zhou. Dynamic type matching. *Manufacturing & Service Operations Management*, 24(1):125–142, 2022.

[13] Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

[14] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

[15] Kyoungseok Jang, Kwang-Sung Jun, Se-Young Yun, and Wanmo Kang. Improved regret bounds of bilinear bandits using action space analysis. In *International Conference on Machine Learning*, pages 4744–4754. PMLR, 2021.

[16] Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21329–21372. PMLR, 2024.

[17] Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pages 3163–3172. PMLR, 2019.

[18] Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35: 19971–19983, 2022.

[19] Richard M Karp, Umesh V Vazirani, and Vijay V Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358, 1990.

[20] Maximilian Kasy and Alexander Teytelboym. Matching with semi-bandits. *The Econometrics Journal*, 26(1):45–66, 2023.

[21] Jung-hun Kim and Min-hwan Oh. Queueing matching bandits with preference feedback. *Advances in Neural Information Processing Systems*, 37:64649–64702, 2024.

[22] Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications. *Journal of Machine Learning Research*, 24(346):1–45, 2023.

[23] Fang Kong, Zilong Wang, and Shuai Li. Improved analysis for bandit learning in matching markets. *Advances in Neural Information Processing Systems*, 37:91904–91929, 2024.

[24] Akshay Krishnamurthy, Alekh Agarwal, and Miro Dudik. Contextual semibandits via supervised learning oracles. *Advances In Neural Information Processing Systems*, 29, 2016.

[25] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

[26] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[27] Junghyun Lee, Kyoungseok Jang, Kwang-Sung Jun, Milan Vojnovic, and Se-Young Yun. GL-LowPopArt: A nearly instance-wise minimax-optimal estimator for generalized low-rank trace regression. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 2872–2880. PMLR, 2026.

[28] Shiyun Lin, Simon Mauras, Vianney Perchet, and Nadav Merlis. Adaptive bandit algorithms for contextual matching markets. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 74800–74843. PMLR, 2026.

[29] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

[30] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 460–468. PMLR, 2021.

[31] Aranyak Mehta and Debmalya Panigrahi. Online matching with stochastic rewards. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 728–737. IEEE, 2012.

[32] Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *Advances in Neural Information Processing Systems*, 35:19956–19970, 2022.

[33] Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 33:5429–5440, 2020.

[34] Pierre Perrault, Michal Valko, and Vianney Perchet. Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Conference on Learning Theory*, pages 3152–3184. PMLR, 2020.

[35] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

[36] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pages 1252–1260. PMLR, 2021.

[37] Alexander Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume 24 of *Algorithms and Combinatorics*. Springer, Berlin, Heidelberg, 2003. ISBN 978-3-540-44389-6.

[38] Flore Sentenac, Jialin Yi, Clement Calauzenes, Vianney Perchet, and Milan Vojnovic. Pure exploration and regret minimization in matching bandits. In *International Conference on Machine Learning*, pages 9434–9442. PMLR, 2021.

[39] Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9791–9798, 2021.

[40] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[41] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR, 2018.

[42] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR, 2015.

Appendix

A Row-wise Regularized Design

A.1 Well-definedness of the RRD oracle

Lemma 7 (Nonempty RRD domain and exact maximizer)**.** *Fix a round t, a context $C_t = (\mathbf{x}_{t,1:M}, \mathbf{y}_{t,1:N})$, a predicted-score matrix $\widehat{\mathbf{R}}_t \in \mathbb{R}^{M\times N}$, and $\xi > 0$. Let*

$$q_t := \dim \operatorname{span}\{\mathbf{y}_{t,1}, \dots, \mathbf{y}_{t,N}\}.$$

Suppose that $q_t > 0$. Define the RRD domain

$$\mathcal{D}_t := \{\mathbf{P} \in \mathcal{M} : \mathbf{V}_{t,i}(\mathbf{P}) \succ 0 \text{ for all } i \in [M]\}.$$

Then $\mathcal{D}_t$ is nonempty, and the exact RRD maximization problem

$$\sup_{\mathbf{P}\in\mathcal{D}_t} \Phi_t(\mathbf{P})$$

attains its supremum at some $\mathbf{P}_t^\star \in \mathcal{D}_t$. Moreover, every exact maximizer $\mathbf{P}_t^\star \in \operatorname{argmax}_{\mathbf{P}\in\mathcal{D}_t} \Phi_t(\mathbf{P})$ satisfies

$$g_t(\mathbf{P}_t^\star) := \max_{\mathbf{Q}\in\mathcal{M}} \langle \nabla\Phi_t(\mathbf{P}_t^\star), \mathbf{Q} - \mathbf{P}_t^\star\rangle = 0.$$

Proof. First we show that $\mathcal{D}_t \neq \emptyset$. Define

$$(\mathbf{P}^0)_{i,j} = \frac{1}{N}, \qquad i \in [M],\ j \in [N].$$

Since $N \geq M$, we have

$$\sum_{j=1}^{N} (\mathbf{P}^0)_{i,j} = 1 \qquad \text{for all } i \in [M],$$

and

$$\sum_{i=1}^{M} (\mathbf{P}^0)_{i,j} = \frac{M}{N} \leq 1 \qquad \text{for all } j \in [N].$$

Thus $\mathbf{P}^0 \in \mathcal{M}$.

Let $\mathbf{U}_t \in \mathbb{R}^{d_2\times q_t}$ have orthonormal columns spanning $\operatorname{span}\{\mathbf{y}_{t,1}, \dots, \mathbf{y}_{t,N}\}$, and define

$$\widetilde{\mathbf{y}}_{t,j} = \mathbf{U}_t^\top \mathbf{y}_{t,j} \in \mathbb{R}^{q_t}.$$

Then $\widetilde{\mathbf{y}}_{t,1}, \dots, \widetilde{\mathbf{y}}_{t,N}$ span $\mathbb{R}^{q_t}$. For every row i,

$$\mathbf{V}_{t,i}(\mathbf{P}^0) = \frac{1}{N} \sum_{j=1}^{N} \widetilde{\mathbf{y}}_{t,j} \widetilde{\mathbf{y}}_{t,j}^\top.$$

For any nonzero vector $\mathbf{v} \in \mathbb{R}^{q_t}$,

$$\begin{aligned} \mathbf{v}^\top \mathbf{V}_{t,i}(\mathbf{P}^0)\mathbf{v} &= \frac{1}{N} \sum_{j=1}^{N} \left(\mathbf{v}^\top \widetilde{\mathbf{y}}_{t,j}\right)^2 \\ &> 0, \end{aligned}$$

because $\widetilde{\mathbf{y}}_{t,1}, \dots, \widetilde{\mathbf{y}}_{t,N}$ span $\mathbb{R}^{q_t}$. Hence

$$\mathbf{V}_{t,i}(\mathbf{P}^0) \succ 0 \qquad \text{for all } i \in [M].$$

Therefore $\mathbf{P}^0 \in \mathcal{D}_t$, and so $\mathcal{D}_t \neq \emptyset$.

We next prove that the supremum is attained. Let

$$\phi_t^\star := \sup_{\mathbf{P}\in\mathcal{D}_t} \Phi_t(\mathbf{P}).$$

Since $\mathbf{P}^0 \in \mathcal{D}_t$, we have

$$\phi_t^\star \geq \Phi_t(\mathbf{P}^0) > -\infty.$$

It remains to show that $\phi_t^\star < \infty$. For every $\mathbf{P} \in \mathcal{M}$, every row i, and every unit vector $\mathbf{u} \in \mathbb{R}^{q_t}$,

$$\begin{aligned}
\mathbf{u}^\top \mathbf{V}_{t,i}(\mathbf{P})\mathbf{u} &= \sum_{j=1}^{N} \mathbf{P}_{i,j} \left(\mathbf{u}^\top \widetilde{\mathbf{y}}_{t,j}\right)^2 \\
&\leq \sum_{j=1}^{N} \mathbf{P}_{i,j} \|\mathbf{u}\|_2^2 \|\widetilde{\mathbf{y}}_{t,j}\|_2^2 \\
&= \sum_{j=1}^{N} \mathbf{P}_{i,j} \|\widetilde{\mathbf{y}}_{t,j}\|_2^2 \\
&\leq L_y^2.
\end{aligned}$$

Here the first inequality uses $\|\mathbf{u}\|_2 = 1$ and $(\mathbf{u}^\top \widetilde{\mathbf{y}}_{t,j})^2 \leq \|\widetilde{\mathbf{y}}_{t,j}\|_2^2$, and the last uses $\|\widetilde{\mathbf{y}}_{t,j}\|_2 \leq \|\mathbf{y}_{t,j}\|_2 \leq L_y$ and $\sum_{j=1}^{N} (\mathbf{P})_{i,j} = 1$. Therefore,

$$\lambda_{\max}(\mathbf{V}_{t,i}(\mathbf{P})) \leq L_y^2.$$

Whenever $\mathbf{V}_{t,i}(\mathbf{P}) \succ 0$, this gives

$$\log\det \mathbf{V}_{t,i}(\mathbf{P}) \leq q_t \log L_y^2.$$

The linear predicted-reward term is also bounded on $\mathcal{M}$. Let

$$R_t^{\max} := \max_{i \in [M],\, j \in [N]} |(\widehat{\mathbf{R}}_t)_{i,j}|.$$

Since every $\mathbf{P} \in \mathcal{M}$ satisfies $\sum_{j=1}^{N} (\mathbf{P})_{i,j} = 1$ for all $i \in [M]$, summing over rows gives $\sum_{i=1}^{M} \sum_{j=1}^{N} (\mathbf{P})_{i,j} = M$, and therefore

$$\begin{aligned}
\left| \sum_{i=1}^{M} \sum_{j=1}^{N} (\mathbf{P})_{i,j} (\widehat{\mathbf{R}}_t)_{i,j} \right| &\leq \sum_{i=1}^{M} \sum_{j=1}^{N} (\mathbf{P})_{i,j} |(\widehat{\mathbf{R}}_t)_{i,j}| \\
&\leq M R_t^{\max}.
\end{aligned}$$

Thus Φ_t is bounded above on $\mathcal{D}_t$, and consequently

$$-\infty < \phi_t^\star < \infty.$$

Now take a maximizing sequence $\mathbf{P}^n \in \mathcal{D}_t$, meaning

$$\Phi_t(\mathbf{P}^n) \to \phi_t^\star.$$

Since $\mathcal{M}$ is a closed and bounded polytope in a finite-dimensional Euclidean space, it is compact. Since $\mathbf{P}^n \in \mathcal{M}$, there exists a subsequence, still denoted by $\mathbf{P}^n$, and a matrix $\mathbf{P}^\infty \in \mathcal{M}$ such that

$$\mathbf{P}^n \to \mathbf{P}^\infty.$$

We show that $\mathbf{P}^\infty \in \mathcal{D}_t$. Suppose, toward a contradiction, that there exists a row $i_0 \in [M]$ such that $\mathbf{V}_{t,i_0}(\mathbf{P}^\infty)$ is singular. Since $\mathbf{V}_{t,i_0}(P)$ is linear in P,

$$\mathbf{V}_{t,i_0}(\mathbf{P}^n) \to \mathbf{V}_{t,i_0}(\mathbf{P}^\infty).$$

Therefore, by continuity of the determinant,

$$\det \mathbf{V}_{t,i_0}(\mathbf{P}^n) \to \det \mathbf{V}_{t,i_0}(\mathbf{P}^\infty) = 0.$$

Since $\mathbf{P}^n \in \mathcal{D}_t$, each $\mathbf{V}_{t,i_0}(\mathbf{P}^n)$ is positive definite, and hence $\det \mathbf{V}_{t,i_0}(\mathbf{P}^n) > 0$. It follows that

$$\log\det \mathbf{V}_{t,i_0}(\mathbf{P}^n) \to -\infty.$$

For every other row $i \neq i_0$, the uniform eigenvalue bound above gives

$$\log\det \mathbf{V}_{t,i}(\mathbf{P}^n) \leq q_t \log L_y^2.$$

Thus the other row log-determinant terms cannot diverge to $+\infty$. The linear predicted-reward term is also bounded along the sequence $\mathbf{P}^n$. Consequently,

$$\Phi_t(\mathbf{P}^n) \to -\infty.$$

This contradicts $\Phi_t(\mathbf{P}^n) \to \phi_t^\star \geq \Phi_t(\mathbf{P}^0) > -\infty$. Therefore no such row i_0 exists, and

$$\mathbf{V}_{t,i}(\mathbf{P}^\infty) \succ 0 \qquad \text{for all } i \in [M].$$

Hence $\mathbf{P}^\infty \in \mathcal{D}_t$. Since Φ_t is continuous on $\mathcal{D}_t$ and $\mathbf{P}^n \to \mathbf{P}^\infty \in \mathcal{D}_t$, we obtain

$$\Phi_t(\mathbf{P}^\infty) = \lim_{n\to\infty} \Phi_t(\mathbf{P}^n) = \phi_t^\star.$$

Thus $\mathbf{P}^\infty$ attains the supremum of Φ_t over $\mathcal{D}_t$.

It remains to prove the zero Frank-Wolfe gap assertion. Let $\mathbf{P}_t^\star \in \operatorname{argmax}_{\mathbf{P}\in\mathcal{D}_t} \Phi_t(\mathbf{P})$ be any exact maximizer. Fix any $\mathbf{Q} \in \mathcal{M}$ and any $\alpha \in (0,1)$. Define

$$\mathbf{P}_\alpha := (1-\alpha)\mathbf{P}_t^\star + \alpha\mathbf{Q}.$$

Since $\mathcal{M}$ is convex, $\mathbf{P}_\alpha \in \mathcal{M}$. Moreover, for every row $i \in [M]$,

$$\mathbf{V}_{t,i}(\mathbf{P}_\alpha) = (1-\alpha)\mathbf{V}_{t,i}(\mathbf{P}_t^\star) + \alpha\mathbf{V}_{t,i}(\mathbf{Q}).$$

Here $\mathbf{V}_{t,i}(\mathbf{P}_t^\star) \succ 0$ because $\mathbf{P}_t^\star \in \mathcal{D}_t$, and $\mathbf{V}_{t,i}(\mathbf{Q}) \succeq 0$ because $\mathbf{Q} \in \mathcal{M}$ has nonnegative entries. Therefore

$$\mathbf{V}_{t,i}(\mathbf{P}_\alpha) \succ 0 \qquad \text{for all } i \in [M],$$

and hence $\mathbf{P}_\alpha \in \mathcal{D}_t$.

By optimality of $\mathbf{P}_t^\star$ over $\mathcal{D}_t$,

$$\Phi_t(\mathbf{P}_\alpha) - \Phi_t(\mathbf{P}_t^\star) \leq 0.$$

Dividing by $\alpha > 0$ and sending $\alpha \downarrow 0$, using differentiability of Φ_t at $\mathbf{P}_t^\star$, gives

$$\langle \nabla\Phi_t(\mathbf{P}_t^\star), \mathbf{Q} - \mathbf{P}_t^\star \rangle \leq 0.$$

Since $\mathbf{Q} \in \mathcal{M}$ was arbitrary,

$$\max_{\mathbf{Q}\in\mathcal{M}} \langle \nabla\Phi_t(\mathbf{P}_t^\star), \mathbf{Q} - \mathbf{P}_t^\star \rangle \leq 0.$$

On the other hand, taking $\mathbf{Q} = \mathbf{P}_t^\star$ shows that the maximum is at least 0. Therefore

$$g_t(\mathbf{P}_t^\star) = 0.$$

This completes the proof. □

A.2 Rectangular Birkhoff decomposition

A matrix is *doubly stochastic* if all entries are nonnegative and every row sum and every column sum equals one. The Birkhoff–von Neumann theorem states that every doubly stochastic matrix is a convex combination of permutation matrices [3]. The following lemma extends this to the rectangular case $N \geq M$, showing that $\mathcal{M}$ is the convex hull of injective matching incidence matrices [37, Corollary 18.1b]. We include a self-contained proof via a padding argument for completeness.

Lemma 8 (Rectangular Birkhoff decomposition)**.** *For every $\mathbf{P} \in \mathcal{M}$, there exists a distribution $\zeta_{\mathbf{P}}$ over injective matchings $\pi : [M] \hookrightarrow [N]$ such that*

$$\mathbb{P}_{\pi\sim\zeta_{\mathbf{P}}}[\pi(i) = j] = \mathbf{P}_{i,j} \qquad \forall i \in [M],\ j \in [N].$$

Equivalently,

$$\mathcal{M} = \operatorname{conv}\left\{\mathbf{X}_\pi : \pi : [M] \hookrightarrow [N]\right\},$$

where $\mathbf{X}_\pi \in \{0,1\}^{M\times N}$ is the incidence matrix of the matching π.

Proof. First suppose $N = M$. Since $\mathbf{P} \in \mathcal{M}$, we have

$$\sum_{j=1}^{N} \mathbf{P}_{i,j} = 1 \qquad \forall i \in [M],$$

and

$$\sum_{i=1}^{M} \mathbf{P}_{i,j} \leq 1 \qquad \forall j \in [N].$$

Summing the row constraints gives

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{P}_{i,j} = M.$$

Since $N = M$, the N column sums are nonnegative, each is at most one, and their total sum is $M = N$. Therefore every column sum is exactly one:

$$\sum_{i=1}^{M} \mathbf{P}_{i,j} = 1 \qquad \forall j \in [N].$$

Thus $\mathbf{P}$ is doubly stochastic. By the Birkhoff-von Neumann theorem [3], $\mathbf{P}$ is a convex combination of permutation matrices. Each permutation matrix is the incidence matrix of a bijective matching from $[M]$ to $[N]$. Hence the desired distribution $\zeta_{\mathbf{P}}$ exists.

Now suppose $N > M$. Define the column slacks

$$s_j = 1 - \sum_{i=1}^{M} \mathbf{P}_{i,j}, \qquad j \in [N].$$

Since $\mathbf{P} \in \mathcal{M}$, we have $0 \leq s_j \leq 1$ for every j. Moreover,

$$\begin{aligned} \sum_{j=1}^{N} s_j &= \sum_{j=1}^{N} \left(1 - \sum_{i=1}^{M} \mathbf{P}_{i,j}\right) \\ &= N - \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{P}_{i,j} \\ &= N - M. \end{aligned}$$

Construct a nonnegative matrix $\mathbf{D} \in \mathbb{R}_{+}^{(N-M)\times N}$ by setting

$$\mathbf{D}_{a,j} = \frac{s_j}{N - M}, \qquad a \in [N - M],\ j \in [N].$$

Then each dummy row has sum one, because

$$\begin{aligned} \sum_{j=1}^{N} \mathbf{D}_{a,j} &= \sum_{j=1}^{N} \frac{s_j}{N - M} \\ &= \frac{1}{N - M} \sum_{j=1}^{N} s_j \\ &= 1. \end{aligned}$$

Also, each column of $\mathbf{D}$ has sum equal to the corresponding slack:

$$\begin{aligned} \sum_{a=1}^{N-M} \mathbf{D}_{a,j} &= \sum_{a=1}^{N-M} \frac{s_j}{N - M} \\ &= s_j. \end{aligned}$$

Define

$$\mathbf{Q} = \begin{pmatrix} \mathbf{P} \\ \mathbf{D} \end{pmatrix} \in \mathbb{R}_+^{N \times N}.$$

The first M rows of $\mathbf{Q}$ have row sum one because $\mathbf{P} \in \mathcal{M}$, and the last $N - M$ rows have row sum one by construction of $\mathbf{D}$. Hence every row of $\mathbf{Q}$ sums to one. For each column j,

$$\begin{aligned} \sum_{\ell=1}^{N} \mathbf{Q}_{\ell,j} &= \sum_{i=1}^{M} \mathbf{P}_{i,j} + \sum_{a=1}^{N-M} \mathbf{D}_{a,j} \\ &= \sum_{i=1}^{M} \mathbf{P}_{i,j} + s_j \\ &= 1. \end{aligned}$$

Thus $\mathbf{Q}$ is doubly stochastic.

By the Birkhoff-von Neumann theorem, there exist permutation matrices $\mathbf{\Pi}^1, \ldots, \mathbf{\Pi}^K \in \{0,1\}^{N \times N}$ and coefficients $\alpha_1, \ldots, \alpha_K$ such that

$$\mathbf{Q} = \sum_{k=1}^{K} \alpha_k \mathbf{\Pi}^k, \qquad \alpha_k \geq 0, \qquad \sum_{k=1}^{K} \alpha_k = 1.$$

For each k, let $\mathbf{X}_{\pi^k} \in \{0,1\}^{M \times N}$ be the matrix obtained by restricting $\mathbf{\Pi}^k$ to its first M rows. Since $\mathbf{\Pi}^k$ is a permutation matrix, each row of $\mathbf{X}_{\pi^k}$ contains exactly one 1, and each column of $\mathbf{X}_{\pi^k}$ contains at most one 1. Hence $\mathbf{X}_{\pi^k}$ is the incidence matrix of an injective matching $\pi^k : [M] \hookrightarrow [N]$.

Taking the first M rows in the decomposition of $\mathbf{Q}$ gives

$$\mathbf{P} = \sum_{k=1}^{K} \alpha_k \mathbf{X}_{\pi^k}.$$

Define a distribution $\zeta_{\mathbf{P}}$ on injective matchings by

$$\zeta_{\mathbf{P}}(\pi^k) = \alpha_k, \qquad k = 1, \ldots, K.$$

If the same matching appears multiple times in the decomposition, combine the corresponding masses. Then, for every $i \in [M]$ and $j \in [N]$,

$$\begin{aligned} \mathbb{P}_{\pi \sim \zeta_{\mathbf{P}}}[\pi(i) = j] &= \sum_{k=1}^{K} \alpha_k \mathbb{1}\{\pi^k(i) = j\} \\ &= \sum_{k=1}^{K} \alpha_k (\mathbf{X}_{\pi^k})_{i,j} \\ &= \mathbf{P}_{i,j}. \end{aligned}$$

This completes the proof. □

A.3 A PSD decomposition inequality

Lemma 9 (PSD decomposition inequality)**.** *Let $\mathbf{A}_1, \ldots, \mathbf{A}_K \in \mathbb{R}^{d \times d}$ be symmetric positive semidefinite matrices, and let*

$$\mathbf{A} = \sum_{k=1}^{K} \mathbf{A}_k.$$

Suppose that $\mathbf{v}_k \in \mathrm{range}(\mathbf{A}_k)$ for every $k \in [K]$, and define

$$\mathbf{v} = \sum_{k=1}^{K} \mathbf{v}_k.$$

Then $\mathbf{v} \in \mathrm{range}(\mathbf{A})$ and

$$\|\mathbf{v}\|_{\mathbf{A}^\dagger}^2 \leq \sum_{k=1}^{K} \|\mathbf{v}_k\|_{\mathbf{A}_k^\dagger}^2.$$

Proof. For each $k \in [K]$, since $\mathbf{A}_k \succeq \mathbf{0}$, its positive semidefinite square root $\mathbf{A}_k^{1/2}$ is well-defined. Moreover,

$$\mathrm{range}(\mathbf{A}_k) = \mathrm{range}(\mathbf{A}_k^{1/2}).$$

Since $\mathbf{v}_k \in \mathrm{range}(\mathbf{A}_k)$, there exists $\mathbf{u}_k$ such that

$$\mathbf{v}_k = \mathbf{A}_k^{1/2}\mathbf{u}_k.$$

Choose $\mathbf{u}_k$ to be the minimum Euclidean norm solution of this equation. By the minimum-norm characterization of the Moore–Penrose pseudoinverse, this choice satisfies

$$\|\mathbf{u}_k\|_2^2 = \|\mathbf{v}_k\|_{\mathbf{A}_k^\dagger}^2.$$

Define the block operator $\mathbf{L} : \mathbb{R}^{Kd} \to \mathbb{R}^d$ by

$$\mathbf{L}(\mathbf{u}_1, \ldots, \mathbf{u}_K) = \sum_{k=1}^K \mathbf{A}_k^{1/2}\mathbf{u}_k.$$

Then

$$\begin{aligned}\mathbf{L}\mathbf{L}^\top &= \sum_{k=1}^K \mathbf{A}_k^{1/2}(\mathbf{A}_k^{1/2})^\top \\ &= \sum_{k=1}^K \mathbf{A}_k \\ &= \mathbf{A}.\end{aligned}$$

Also, the block vector $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$ is feasible for the linear system $\mathbf{L}\mathbf{w} = \mathbf{v}$, because

$$\begin{aligned}\mathbf{L}(\mathbf{u}_1, \ldots, \mathbf{u}_K) &= \sum_{k=1}^K \mathbf{A}_k^{1/2}\mathbf{u}_k \\ &= \sum_{k=1}^K \mathbf{v}_k \\ &= \mathbf{v}.\end{aligned}$$

Thus $\mathbf{v} \in \mathrm{range}(\mathbf{L})$. Since $\mathrm{null}(\mathbf{L}\mathbf{L}^\top) = \mathrm{null}(\mathbf{L}^\top)$, we have $\mathrm{range}(\mathbf{L}\mathbf{L}^\top) = \mathrm{range}(\mathbf{L})$, and hence

$$\mathbf{v} \in \mathrm{range}(\mathbf{A}).$$

It remains to relate $\|\mathbf{v}\|_{\mathbf{A}^\dagger}^2$ to the minimum norm of a preimage under $\mathbf{L}$. We claim that

$$\|\mathbf{v}\|_{\mathbf{A}^\dagger}^2 = \min_{\mathbf{w}:\, \mathbf{L}\mathbf{w}=\mathbf{v}} \|\mathbf{w}\|_2^2.$$

Indeed, since $\mathbf{v} \in \mathrm{range}(\mathbf{L})$, the minimum-Euclidean-norm solution of $\mathbf{L}\mathbf{w} = \mathbf{v}$ is $\mathbf{w}^\star = \mathbf{L}^\dagger\mathbf{v}$. Therefore,

$$\min_{\mathbf{w}:\, \mathbf{L}\mathbf{w}=\mathbf{v}} \|\mathbf{w}\|_2^2 = \|\mathbf{L}^\dagger\mathbf{v}\|_2^2.$$

Using the reduced singular value decomposition

$$\mathbf{L} = \boldsymbol{\Upsilon}\boldsymbol{\Sigma}\boldsymbol{\Omega}^\top,$$

we have

$$\mathbf{L}^\dagger = \boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Upsilon}^\top.$$

Therefore,

$$\begin{aligned}(\mathbf{L}^\dagger)^\top\mathbf{L}^\dagger &= \boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}^\top\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Upsilon}^\top \\ &= \boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-2}\boldsymbol{\Upsilon}^\top.\end{aligned}$$

On the other hand,

$$\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Upsilon}\boldsymbol{\Sigma}^2\boldsymbol{\Upsilon}^\top,$$

and hence

$$(\mathbf{L}\mathbf{L}^\top)^\dagger = \boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-2}\boldsymbol{\Upsilon}^\top.$$

Thus

$$(\mathbf{L}^\dagger)^\top\mathbf{L}^\dagger = (\mathbf{L}\mathbf{L}^\top)^\dagger.$$

Hence

$$\begin{aligned}\|\mathbf{L}^\dagger\mathbf{v}\|_2^2 &= \mathbf{v}^\top(\mathbf{L}^\dagger)^\top\mathbf{L}^\dagger\mathbf{v} \\ &= \mathbf{v}^\top(\mathbf{L}\mathbf{L}^\top)^\dagger\mathbf{v} \\ &= \mathbf{v}^\top\mathbf{A}^\dagger\mathbf{v} \\ &= \|\mathbf{v}\|_{\mathbf{A}^\dagger}^2.\end{aligned}$$

Since the particular block vector $(\mathbf{u}_1, \ldots, \mathbf{u}_K)$ constructed above is feasible for $\mathbf{L}\mathbf{w} = \mathbf{v}$, the minimum over all feasible preimages is no larger than its squared norm. Therefore,

$$\begin{aligned}\|\mathbf{v}\|_{\mathbf{A}^\dagger}^2 &= \min_{\mathbf{w}:\,\mathbf{L}\mathbf{w}=\mathbf{v}} \|\mathbf{w}\|_2^2 \\ &\le \sum_{k=1}^{K} \|\mathbf{u}_k\|_2^2 \\ &= \sum_{k=1}^{K} \|\mathbf{v}_k\|_{\mathbf{A}_k^\dagger}^2.\end{aligned}$$

This completes the proof. □

A.4 Proof of RRD certificate

Lemma 5 (RRD certificate, restated)**.** *Fix a round t, and suppose $\xi > 0$ and $\varepsilon \ge 0$. Let μ_t be as returned by $\mathrm{RRD}_t(\widehat{\mathbf{R}}_t, \xi, \varepsilon)$, and define*

$$\mathbf{H}_t = \mathbb{E}_{\pi\sim\mu_t}\mathbf{S}_t(\pi).$$

If $q_t = 0$, then $\mathbf{a}_t(\pi) = \mathbf{0}$ and $\mathbf{H}_t = \mathbf{0}$ for every feasible matching $\pi \in \Pi$.

If $q_t > 0$, let $\mathbf{P}_t \in \mathcal{D}_t$ be a certified approximate RRD solution satisfying $g_t(\mathbf{P}_t) \le \varepsilon$. Then, for every feasible matching $\pi \in \Pi$,

$$\|\mathbf{a}_t(\pi)\|_{\mathbf{H}_t^\dagger}^2 + \xi\left(\sum_{i=1}^{M}(\widehat{\mathbf{R}}_t)_{i,\pi(i)} - \langle\mathbf{P}_t, \widehat{\mathbf{R}}_t\rangle\right) \le Mq_t + \varepsilon, \tag{3}$$

and

$$\mathbf{a}_t(\pi) \in \mathrm{range}(\mathbf{H}_t). \tag{4}$$

Proof. If $q_t = 0$, then

$$\mathbf{y}_{t,j} = \mathbf{0} \qquad \forall j \in [N].$$

Hence

$$\mathbf{z}_{t,i,j} = \mathbf{0} \qquad \forall i \in [M],\ j \in [N].$$

It follows that $\mathbf{a}_t(\pi) = \mathbf{0}$, $\mathbf{S}_t(\pi) = \mathbf{0}$, and $\mathbf{H}_t = \mathbf{0}$ for every feasible matching π. Hence the claim holds trivially. Assume $q_t > 0$ henceforth.

Fix the round t and suppress the subscript t when no confusion can arise. Write

$$\mathbf{P} = \mathbf{P}_t, \qquad \mathbf{H} = \mathbf{H}_t, \qquad \widehat{\mathbf{R}} = \widehat{\mathbf{R}}_t.$$

Work first in item-span coordinates. Define the leverage scores

$$\ell_{i,j}(\mathbf{P}) = \widetilde{\mathbf{y}}_j^\top\mathbf{V}_i(\mathbf{P})^{-1}\widetilde{\mathbf{y}}_j.$$

Using $d \log\det \mathbf{X} = \operatorname{tr}(\mathbf{X}^{-1} d\mathbf{X})$, the gradient of Φ is

$$\nabla_{i,j}\Phi(\mathbf{P}) = \ell_{i,j}(\mathbf{P}) + \xi(\widehat{\mathbf{R}})_{i,j}. \tag{5}$$

This is a standard matrix calculus identity.

For every row i,

$$\begin{aligned}\sum_{j=1}^{N}(\mathbf{P})_{i,j}\ell_{i,j}(\mathbf{P}) &= \sum_{j=1}^{N}(\mathbf{P})_{i,j}\widetilde{\mathbf{y}}_j^\top \mathbf{V}_i(\mathbf{P})^{-1}\widetilde{\mathbf{y}}_j \\ &= \operatorname{tr}\left(\mathbf{V}_i(\mathbf{P})^{-1}\sum_{j=1}^{N}(\mathbf{P})_{i,j}\widetilde{\mathbf{y}}_j\widetilde{\mathbf{y}}_j^\top\right) \\ &= \operatorname{tr}(\mathbf{I}_{q_t}) \\ &= q_t.\end{aligned}$$

Thus

$$\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{P})_{i,j}\ell_{i,j}(\mathbf{P}) = Mq_t. \tag{6}$$

For a feasible matching π, let $\mathbf{X}_\pi$ be its incidence matrix. The Frank-Wolfe gap certificate gives

$$\langle \nabla\Phi(\mathbf{P}), \mathbf{X}_\pi - \mathbf{P}\rangle \le \varepsilon. \tag{7}$$

Frank-Wolfe gap certificates are standard computable optimality certificates for constrained smooth optimization [14]. Expanding (7) using (5) and (6) gives

$$\begin{aligned}\varepsilon &\ge \langle \nabla\Phi(\mathbf{P}), \mathbf{X}_\pi - \mathbf{P}\rangle \\ &= \sum_{i=1}^{M}\ell_{i,\pi(i)}(\mathbf{P}) - Mq_t + \xi\left(\sum_{i=1}^{M}(\widehat{\mathbf{R}})_{i,\pi(i)} - \langle \mathbf{P}, \widehat{\mathbf{R}}\rangle\right).\end{aligned} \tag{8}$$

Consequently, for every feasible matching π,

$$\sum_{i=1}^{M}\ell_{i,\pi(i)}(\mathbf{P}) + \xi\left(\sum_{i=1}^{M}(\widehat{\mathbf{R}})_{i,\pi(i)} - \langle \mathbf{P}, \widehat{\mathbf{R}}\rangle\right) \le Mq_t + \varepsilon. \tag{9}$$

It remains to prove the width certificate. The global feature convention is

$$\mathbf{z}_{i,j} = \mathbf{y}_j \otimes \mathbf{x}_i.$$

For the row-wise leverage argument, it is cleaner to use $\mathbf{x}$-first item-span coordinates. Let $\mathbf{K} \in \mathbb{R}^{d_1 q_t \times q_t d_1}$ be the commutation matrix satisfying

$$\mathbf{K}(\widetilde{\mathbf{y}}_j \otimes \mathbf{x}_i) = \mathbf{x}_i \otimes \widetilde{\mathbf{y}}_j.$$

Since $\mathbf{K}$ is a permutation matrix, it is orthogonal. Define

$$\mathbf{T} = (\mathbf{U} \otimes \mathbf{I}_{d_1})\mathbf{K}^\top \in \mathbb{R}^{d_1 d_2 \times d_1 q_t}.$$

Then

$$\mathbf{T}^\top\mathbf{T} = \mathbf{I}_{d_1 q_t}.$$

For

$$\overline{\mathbf{z}}_{i,j} = \mathbf{x}_i \otimes \widetilde{\mathbf{y}}_j \in \mathbb{R}^{d_1 q_t},$$

we have

$$\mathbf{z}_{i,j} = \mathbf{y}_j \otimes \mathbf{x}_i = \mathbf{T}\overline{\mathbf{z}}_{i,j}.$$

For a feasible matching π, define

$$\overline{\mathbf{a}}(\pi) = \sum_{i=1}^{M}\overline{\mathbf{z}}_{i,\pi(i)},$$

and

$$\overline{\mathbf{S}}(\pi) = \sum_{i=1}^{M} \overline{\mathbf{z}}_{i,\pi(i)} \overline{\mathbf{z}}_{i,\pi(i)}^{\top}.$$

Then

$$\mathbf{a}(\pi) = \mathbf{T}\overline{\mathbf{a}}(\pi), \qquad \mathbf{S}(\pi) = \mathbf{T}\overline{\mathbf{S}}(\pi)\mathbf{T}^{\top}. \tag{10}$$

Also define

$$\overline{\mathbf{H}}_{\mathbf{P}} = \sum_{i=1}^{M}\sum_{j=1}^{N} (\mathbf{P})_{i,j} \overline{\mathbf{z}}_{i,j} \overline{\mathbf{z}}_{i,j}^{\top}.$$

Since $\overline{\mathbf{z}}_{i,j} = \mathbf{x}_i \otimes \widetilde{\mathbf{y}}_j$,

$$\begin{aligned}\overline{\mathbf{H}}_{\mathbf{P}} &= \sum_{i=1}^{M}\sum_{j=1}^{N} (\mathbf{P})_{i,j} (\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes (\widetilde{\mathbf{y}}_j \widetilde{\mathbf{y}}_j^{\top}) \\ &= \sum_{i=1}^{M} (\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes \mathbf{V}_i(\mathbf{P}). \end{aligned} \tag{11}$$

By the marginal property of $\zeta_{\mathbf{P}}$ and by (10),

$$\mathbf{H} = \mathbf{T}\overline{\mathbf{H}}_{\mathbf{P}}\mathbf{T}^{\top}.$$

Since $\mathbf{T}^{\top}\mathbf{T} = \mathbf{I}_{d_1 q_t}$, we have

$$(\mathbf{T}\overline{\mathbf{H}}_{\mathbf{P}}\mathbf{T}^{\top})^{\dagger} = \mathbf{T}\overline{\mathbf{H}}_{\mathbf{P}}^{\dagger}\mathbf{T}^{\top}.$$

This follows from the eigenvalue decomposition of $\overline{\mathbf{H}}_{\mathbf{P}}$ and $\mathbf{T}^{\top}\mathbf{T} = \mathbf{I}_{d_1 q_t}$. Therefore, for every feasible matching π,

$$\|\mathbf{a}(\pi)\|_{\mathbf{H}^{\dagger}}^2 = \|\overline{\mathbf{a}}(\pi)\|_{\overline{\mathbf{H}}_{\mathbf{P}}^{\dagger}}^2. \tag{12}$$

We next bound $\|\overline{\mathbf{a}}(\pi)\|_{\overline{\mathbf{H}}_{\mathbf{P}}^{\dagger}}^2$. Before applying Lemma 9 to the decomposition (11), we verify its range condition. For each row i, let

$$\mathbf{A}_i = (\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes \mathbf{V}_i(\mathbf{P}), \qquad \mathbf{v}_i = \mathbf{x}_i \otimes \widetilde{\mathbf{y}}_{\pi(i)}.$$

If $\mathbf{x}_i = \mathbf{0}$, then $\mathbf{v}_i = \mathbf{0} \in \mathrm{range}(\mathbf{A}_i)$. Otherwise, $\mathbf{x}_i \in \mathrm{range}(\mathbf{x}_i \mathbf{x}_i^{\top})$. Moreover, since $\mathbf{P} \in \mathcal{D}_t$, we have $\mathbf{V}_i(\mathbf{P}) \succ \mathbf{0}$, and hence $\mathrm{range}(\mathbf{V}_i(\mathbf{P})) = \mathbb{R}^{q_t}$. Thus $\widetilde{\mathbf{y}}_{\pi(i)} \in \mathrm{range}(\mathbf{V}_i(\mathbf{P}))$, which implies

$$\mathbf{x}_i \otimes \widetilde{\mathbf{y}}_{\pi(i)} \in \mathrm{range}\left((\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes \mathbf{V}_i(\mathbf{P})\right).$$

Therefore Lemma 9 applies and gives

$$\|\overline{\mathbf{a}}(\pi)\|_{\overline{\mathbf{H}}_{\mathbf{P}}^{\dagger}}^2 \le \sum_{i=1}^{M} \left\|\mathbf{x}_i \otimes \widetilde{\mathbf{y}}_{\pi(i)}\right\|_{\left[(\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes \mathbf{V}_i(\mathbf{P})\right]^{\dagger}}^2. \tag{13}$$

For rows with $\mathbf{x}_i = \mathbf{0}$, the contribution in (13) is zero.

For rows with $\mathbf{x}_i \neq \mathbf{0}$, using the pseudoinverse identity for Kronecker products gives

$$\begin{aligned}&\left\|\mathbf{x}_i \otimes \widetilde{\mathbf{y}}_{\pi(i)}\right\|_{\left[(\mathbf{x}_i \mathbf{x}_i^{\top}) \otimes \mathbf{V}_i(\mathbf{P})\right]^{\dagger}}^2 \\ &\quad = \left(\mathbf{x}_i^{\top} (\mathbf{x}_i \mathbf{x}_i^{\top})^{\dagger} \mathbf{x}_i\right) \left(\widetilde{\mathbf{y}}_{\pi(i)}^{\top} \mathbf{V}_i(\mathbf{P})^{-1} \widetilde{\mathbf{y}}_{\pi(i)}\right) \\ &\quad = \widetilde{\mathbf{y}}_{\pi(i)}^{\top} \mathbf{V}_i(\mathbf{P})^{-1} \widetilde{\mathbf{y}}_{\pi(i)}.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\overline{\mathbf{a}}(\pi)\|_{\overline{\mathbf{H}}_{\mathbf{P}}^{\dagger}}^2 &\le \sum_{i=1}^{M} \widetilde{\mathbf{y}}_{\pi(i)}^{\top} \mathbf{V}_i(\mathbf{P})^{-1} \widetilde{\mathbf{y}}_{\pi(i)} \\ &\le M q_t + \varepsilon - \xi \left(\sum_{i=1}^{M} (\widehat{\mathbf{R}})_{i,\pi(i)} - \langle \mathbf{P}, \widehat{\mathbf{R}} \rangle\right),\end{aligned}$$

where the last step uses (9). In particular,

$$\overline{\mathbf{a}}(\pi) \in \mathrm{range}(\overline{\mathbf{H}}_{\mathbf{P}})$$

by Lemma 9.

Combining the preceding bound with (12), we obtain

$$\|\mathbf{a}(\pi)\|^2_{\mathbf{H}^\dagger} + \xi \left(\sum_{i=1}^{M} (\widehat{\mathbf{R}})_{i,\pi(i)} - \langle \mathbf{P}, \widehat{\mathbf{R}} \rangle \right) \le M q_t + \varepsilon,$$

which proves (3).

It remains to prove the range assertion. Since $\overline{\mathbf{a}}(\pi) \in \mathrm{range}(\overline{\mathbf{H}}_{\mathbf{P}})$, there exists $\mathbf{w}$ such that

$$\overline{\mathbf{a}}(\pi) = \overline{\mathbf{H}}_{\mathbf{P}} \mathbf{w}.$$

Using $\mathbf{T}^\top \mathbf{T} = \mathbf{I}_{d_1 q_t}$ and $\mathbf{H} = \mathbf{T} \overline{\mathbf{H}}_{\mathbf{P}} \mathbf{T}^\top$, we have

$$\begin{aligned} \mathbf{a}(\pi) &= \mathbf{T} \overline{\mathbf{a}}(\pi) \\ &= \mathbf{T} \overline{\mathbf{H}}_{\mathbf{P}} \mathbf{w} \\ &= \mathbf{H}(\mathbf{T}\mathbf{w}). \end{aligned}$$

Hence $\mathbf{a}(\pi) \in \mathrm{range}(\mathbf{H})$, which proves (4). □

B Regret Upper Bound

B.1 Setup

Throughout this section, $\mathcal{F}_t$ denotes the pre-action filtration defined in Section 3.3, and $\mathcal{G}_t$ denotes the post-action, pre-reward filtration. For any feasible matching $\pi \in \Pi$, define the matching-level squared width

$$\omega_t(\pi) = \mathrm{tr}\left(\mathbf{V}_t^{-1} \mathbf{S}_t(\pi) \right) = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi(i)}^\top \mathbf{V}_t^{-1} \mathbf{z}_{t,i,\pi(i)}.$$

Thus the algorithm computes

$$\pi_t^w \in \operatorname*{argmax}_{\pi \in \Pi} \omega_t(\pi), \qquad \omega_t = \omega_t(\pi_t^w).$$

We split the rounds into the high-width and low-width sets:

$$\mathcal{A} = \{t \in [T] : \omega_t > \rho\}, \qquad \mathcal{B} = [T] \setminus \mathcal{A}.$$

Let $p = d_1 d_2$ and define

$$K_T = p \log\left(1 + \frac{MTL^2}{\lambda p}\right), \quad \ell_\delta = \max\left\{1, \log\frac{cT}{\delta}\right\}, \quad A_T = \lambda S^2 + \sigma_\eta^2 (K_T + \ell_\delta), \tag{14}$$

where $c > 0$ is a sufficiently large numerical constant. In the main theorem, we take $\rho = c_\rho / \ell_\delta$ with $0 < c_\rho \le 1$, so that $0 < \rho \le 1$.

B.2 High-width elimination

Lemma 10 (High-width elimination)**.** *On the high-width set $\mathcal{A} = \{t \in [T] : \omega_t > \rho\}$,*

$$|\mathcal{A}| \le \frac{2K_T}{\rho}.$$

Moreover,

$$\mathrm{Reg}_{\mathcal{A}}(T) \le \frac{4BMK_T}{\rho}.$$

Proof. For $t \in \mathcal{A}$, the algorithm plays π_t^w and updates

$$\mathbf{V}_{t+1} = \mathbf{V}_t + \mathbf{S}_t(\pi_t^w).$$

By definition of $\mathcal{A}$, for every $t \in \mathcal{A}$,

$$\operatorname{tr}\left(\mathbf{V}_t^{-1}\mathbf{S}_t(\pi_t^w)\right) = \omega_t > \rho.$$

Let

$$\mathbf{Y}_t = \mathbf{V}_t^{-1/2}\mathbf{S}_t(\pi_t^w)\mathbf{V}_t^{-1/2}.$$

Then $\mathbf{Y}_t \succeq 0$ and $\operatorname{tr}(\mathbf{Y}_t) > \rho$. If $\alpha_{t,1}, \dots, \alpha_{t,p}$ are the eigenvalues of $\mathbf{Y}_t$, then $\alpha_{t,k} \geq 0$ and

$$\begin{aligned}
\log \frac{\det \mathbf{V}_{t+1}}{\det \mathbf{V}_t} &= \log \det(\mathbf{I}_p + \mathbf{Y}_t) \\
&= \sum_{k=1}^{p} \log(1 + \alpha_{t,k}) \\
&\geq \log\left(1 + \sum_{k=1}^{p} \alpha_{t,k}\right) \\
&= \log(1 + \operatorname{tr}(\mathbf{Y}_t)) \\
&> \log(1 + \rho) \\
&\geq \frac{\rho}{2},
\end{aligned}$$

where the first inequality uses $\prod_{k=1}^{p}(1 + \alpha_{t,k}) \geq 1 + \sum_{k=1}^{p} \alpha_{t,k}$, and the last inequality uses $0 < \rho \leq 1$. Summing over $t \in \mathcal{A}$ and using the telescoping determinant potential gives

$$\begin{aligned}
\frac{\rho}{2}|\mathcal{A}| &\leq \sum_{t \in \mathcal{A}} \log \frac{\det \mathbf{V}_{t+1}}{\det \mathbf{V}_t} \\
&\leq \sum_{t=1}^{T} \log \frac{\det \mathbf{V}_{t+1}}{\det \mathbf{V}_t} \\
&= \log \frac{\det \mathbf{V}_{T+1}}{\det \mathbf{V}_1}.
\end{aligned}$$

Therefore,

$$|\mathcal{A}| \leq \frac{2}{\rho} \log \frac{\det \mathbf{V}_{T+1}}{\det \mathbf{V}_1}.$$

It remains to upper bound the determinant ratio. Define the realized update matrix

$$\mathbf{\Delta}_t = \mathbf{V}_{t+1} - \mathbf{V}_t = \mathbf{S}_t(\pi_t),$$

which is defined for all rounds $t \in [T]$. Then $\mathbf{\Delta}_t \succeq 0$. Moreover, for every matching $\pi \in \Pi$,

$$\begin{aligned}
\operatorname{tr}(\mathbf{S}_t(\pi)) &= \sum_{i=1}^{M} \|\mathbf{z}_{t,i,\pi(i)}\|_2^2 \\
&\leq ML^2,
\end{aligned}$$

where $\|\mathbf{z}_{t,i,j}\|_2 = \|\mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i}\|_2 = \|\mathbf{y}_{t,j}\|_2 \|\mathbf{x}_{t,i}\|_2 \leq L_y L_x = L$. Hence

$$\operatorname{tr}(\mathbf{\Delta}_t) \leq ML^2 \qquad \forall t \in [T].$$

Let

$$\mathbf{\Gamma}_T = \sum_{t=1}^{T} \mathbf{\Delta}_t.$$

Then

$$\mathbf{V}_{T+1} = \mathbf{V}_1 + \mathbf{\Gamma}_T = \lambda \mathbf{I}_p + \mathbf{\Gamma}_T,$$

and

$$\operatorname{tr}(\mathbf{\Gamma}_T) \le MTL^2.$$

By the arithmetic-geometric mean inequality applied to the eigenvalues of $\mathbf{I}_p + \lambda^{-1}\mathbf{\Gamma}_T$,

$$\begin{aligned}\log \frac{\det \mathbf{V}_{T+1}}{\det \mathbf{V}_1} &= \log\det(\mathbf{I}_p + \lambda^{-1}\mathbf{\Gamma}_T) \\ &\le p \log\left(1 + \frac{\operatorname{tr}(\mathbf{\Gamma}_T)}{\lambda p}\right) \\ &\le p \log\left(1 + \frac{MTL^2}{\lambda p}\right) \\ &= K_T.\end{aligned}$$

Thus

$$|\mathcal{A}| \le \frac{2K_T}{\rho}.$$

Since each matched-pair mean reward is bounded in absolute value by $B = SL$, the matching-level pseudo-regret in any round is at most $2BM$. Let

$$\operatorname{Reg}_{\mathcal{A}}(T) = \sum_{t \in \mathcal{A}} \left[F_{\mathbf{\Theta}^\star}(C_t, \pi_t^\star) - F_{\mathbf{\Theta}^\star}(C_t, \pi_t)\right].$$

Then

$$\operatorname{Reg}_{\mathcal{A}}(T) \le 2BM|\mathcal{A}| \le \frac{4BMK_T}{\rho}.$$

□

B.3 Low-width covariance domination

Lemma 11 (Low-width covariance domination)**.** *For every $t \in \mathcal{B}$ and every feasible matching $\pi \in \Pi$,*

$$\mathbf{S}_t(\pi) \preceq \rho \mathbf{V}_t.$$

Consequently, on every low-width round,

$$\mathbf{H}_t = \mathbb{E}_{\pi \sim \mu_t} \mathbf{S}_t(\pi) \preceq \rho \mathbf{V}_t.$$

Proof. For $t \in \mathcal{B}$, we have $\omega_t \le \rho$. Since π_t^w maximizes $\omega_t(\pi)$ over Π, every feasible matching $\pi \in \Pi$ satisfies

$$\begin{aligned}\operatorname{tr}\left(\mathbf{V}_t^{-1}\mathbf{S}_t(\pi)\right) &= \omega_t(\pi) \\ &\le \omega_t(\pi_t^w) \\ &= \omega_t \\ &\le \rho.\end{aligned}$$

For any feasible matching π, define

$$\mathbf{Y}_t(\pi) = \mathbf{V}_t^{-1/2}\mathbf{S}_t(\pi)\mathbf{V}_t^{-1/2}.$$

Then $\mathbf{Y}_t(\pi) \succeq \mathbf{0}$ and

$$\operatorname{tr}(\mathbf{Y}_t(\pi)) = \operatorname{tr}\left(\mathbf{V}_t^{-1}\mathbf{S}_t(\pi)\right) \le \rho.$$

Hence

$$\lambda_{\max}(\mathbf{Y}_t(\pi)) \le \operatorname{tr}(\mathbf{Y}_t(\pi)) \le \rho,$$

which implies

$$\mathbf{Y}_t(\pi) \preceq \rho \mathbf{I}_p.$$

Multiplying on the left and right by $\mathbf{V}_t^{1/2}$ gives

$$\mathbf{S}_t(\pi) \preceq \rho \mathbf{V}_t.$$

On a low-width round, the algorithm samples $\pi_t \sim \mu_t$ and we define

$$\mathbf{H}_t = \mathbb{E}_{\pi \sim \mu_t} \mathbf{S}_t(\pi) = \mathbb{E}[\mathbf{S}_t(\pi_t) \mid \mathcal{F}_t].$$

Since $\mathbf{S}_t(\pi) \preceq \rho \mathbf{V}_t$ for every $\pi \in \Pi$, taking expectation with respect to μ_t gives

$$\mathbf{H}_t \preceq \rho \mathbf{V}_t.$$

□

B.4 Noise concentration

The next lemmas are used only in the proof of the low-width cumulative prediction error bound.

Lemma 12 (Within-round vector sub-Gaussianity)**.** *Fix a round t and let $\boldsymbol{\eta}_t = (\eta_{t,1}, \dots, \eta_{t,M})$. Then, for every $\mathcal{G}_t$-measurable vector $\mathbf{u}_t = (u_{t,1}, \dots, u_{t,M}) \in \mathbb{R}^M$ with finite Euclidean norm,*

$$\mathbb{E}\left[\exp(\mathbf{u}_t^\top \boldsymbol{\eta}_t) \mid \mathcal{G}_t\right] \le \exp\left(\frac{\sigma_\eta^2 \|\mathbf{u}_t\|_2^2}{2}\right).$$

In particular, the same inequality holds for every deterministic $\mathbf{u} \in \mathbb{R}^M$.

Proof. We first note that Assumption 2 extends from deterministic scalars to predictable random scalars. More precisely, if λ is $\mathcal{H}_{t,i-1}$-measurable and finite almost surely, then

$$\mathbb{E}\left[\exp(\lambda \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] \le \exp\left(\frac{\sigma_\eta^2 \lambda^2}{2}\right).$$

Indeed, first suppose that λ is simple:

$$\lambda = \sum_{m=1}^{R} a_m \mathbb{1}_{A_m}, \qquad A_m \in \mathcal{H}_{t,i-1}.$$

Then, using the pull-out property of conditional expectation,

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] &= \sum_{m=1}^{R} \mathbb{1}_{A_m} \mathbb{E}\left[\exp(a_m \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] \\
&\le \sum_{m=1}^{R} \mathbb{1}_{A_m} \exp\left(\frac{\sigma_\eta^2 a_m^2}{2}\right) \\
&= \exp\left(\frac{\sigma_\eta^2 \lambda^2}{2}\right).
\end{aligned}$$

For a general finite $\mathcal{H}_{t,i-1}$-measurable λ, choose simple $\mathcal{H}_{t,i-1}$-measurable random variables λ_n such that $\lambda_n \to \lambda$ almost surely. Since

$$\exp(\lambda_n \eta_{t,i}) \to \exp(\lambda \eta_{t,i}) \qquad \text{almost surely,}$$

and all these random variables are nonnegative, conditional Fatou's lemma gives

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] &\le \liminf_{n\to\infty} \mathbb{E}\left[\exp(\lambda_n \eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] \\
&\le \liminf_{n\to\infty} \exp\left(\frac{\sigma_\eta^2 \lambda_n^2}{2}\right) \\
&= \exp\left(\frac{\sigma_\eta^2 \lambda^2}{2}\right),
\end{aligned}$$

where the last equality follows from the almost sure convergence $\lambda_n \to \lambda$ and continuity of $x \mapsto \exp(\sigma_\eta^2 x^2/2)$.

The rewards are revealed in the forward order $i = 1, \dots, M$, with

$$\mathcal{H}_{t,0} = \mathcal{G}_t, \qquad \mathcal{H}_{t,i} = \mathcal{G}_t \vee \sigma(\eta_{t,1}, \dots, \eta_{t,i}).$$

Since $\mathbf{u}_t$ is $\mathcal{G}_t$-measurable, it is also $\mathcal{H}_{t,i-1}$-measurable for every i. Using the tower property and peeling the last coordinate first, we get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sum_{i=1}^{M} u_{t,i}\eta_{t,i}\right) \mid \mathcal{G}_t\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{i=1}^{M-1} u_{t,i}\eta_{t,i}\right)\mathbb{E}\left[\exp(u_{t,M}\eta_{t,M}) \mid \mathcal{H}_{t,M-1}\right] \mid \mathcal{G}_t\right] \\
&\le \mathbb{E}\left[\exp\left(\sum_{i=1}^{M-1} u_{t,i}\eta_{t,i}\right)\exp\left(\frac{\sigma_\eta^2 u_{t,M}^2}{2}\right) \mid \mathcal{G}_t\right] \\
&= \exp\left(\frac{\sigma_\eta^2 u_{t,M}^2}{2}\right)\mathbb{E}\left[\exp\left(\sum_{i=1}^{M-1} u_{t,i}\eta_{t,i}\right) \mid \mathcal{G}_t\right].
\end{aligned}$$

The last equality uses the fact that $u_{t,M}$ is $\mathcal{G}_t$-measurable. Repeating the same argument for $M-1, M-2, \ldots, 1$ gives

$$\begin{aligned}
\mathbb{E}\left[\exp(\mathbf{u}_t^\top \boldsymbol{\eta}_t) \mid \mathcal{G}_t\right] &\le \prod_{i=1}^{M}\exp\left(\frac{\sigma_\eta^2 u_{t,i}^2}{2}\right) \\
&= \exp\left(\frac{\sigma_\eta^2}{2}\sum_{i=1}^{M} u_{t,i}^2\right) \\
&= \exp\left(\frac{\sigma_\eta^2\|\mathbf{u}_t\|_2^2}{2}\right).
\end{aligned}$$

This proves the claim.

The same bound will also be used after adjoining external auxiliary randomness. Let $\mathcal{K}_t$ be a sigma-field generated by auxiliary randomness such that, conditionally on $\mathcal{G}_t$, it is independent of the whole same-round noise vector $\boldsymbol{\eta}_t = (\eta_{t,1}, \ldots, \eta_{t,M})$:

$$\mathcal{K}_t \perp\!\!\!\perp \sigma(\boldsymbol{\eta}_t) \mid \mathcal{G}_t.$$

We claim that the one-coordinate conditional sub-Gaussian bound remains valid after this enlargement. Fix $i \in [M]$. Since $\mathcal{H}_{t,i-1} = \mathcal{G}_t \vee \sigma(\eta_{t,1}, \ldots, \eta_{t,i-1})$, the conditional independence of $\mathcal{K}_t$ from the full vector $\boldsymbol{\eta}_t$ given $\mathcal{G}_t$ implies, by testing against bounded $\mathcal{H}_{t,i-1} \vee \mathcal{K}_t$-measurable functions and a monotone-class argument, that for every bounded measurable function h,

$$\mathbb{E}\left[h(\eta_{t,i}) \mid \mathcal{H}_{t,i-1} \vee \mathcal{K}_t\right] = \mathbb{E}\left[h(\eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right].$$

Applying this identity to the bounded functions $h_R(x) = \exp(ax) \wedge R$ and then letting $R \uparrow \infty$ by conditional monotone convergence gives, for every deterministic $a \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(a\eta_{t,i}) \mid \mathcal{H}_{t,i-1} \vee \mathcal{K}_t\right] = \mathbb{E}\left[\exp(a\eta_{t,i}) \mid \mathcal{H}_{t,i-1}\right] \le \exp\left(\frac{\sigma_\eta^2 a^2}{2}\right).$$

As before, this deterministic-a statement extends to every finite $\mathcal{H}_{t,i-1} \vee \mathcal{K}_t$-measurable scalar λ by simple approximation and conditional Fatou:

$$\mathbb{E}\left[\exp(\lambda\eta_{t,i}) \mid \mathcal{H}_{t,i-1} \vee \mathcal{K}_t\right] \le \exp\left(\frac{\sigma_\eta^2 \lambda^2}{2}\right).$$

Consequently, if $\mathbf{u}_t$ is $\mathcal{G}_t \vee \mathcal{K}_t$-measurable with finite Euclidean norm, the same backward peeling argument applied to the enlarged within-round filtration $(\mathcal{H}_{t,i} \vee \mathcal{K}_t)_{i=0}^{M}$ yields

$$\mathbb{E}\left[\exp(\mathbf{u}_t^\top \boldsymbol{\eta}_t) \mid \mathcal{G}_t \vee \mathcal{K}_t\right] \le \exp\left(\frac{\sigma_\eta^2\|\mathbf{u}_t\|_2^2}{2}\right).$$

In Lemma 13 we use this extension with $\mathcal{K}_t = \sigma(\boldsymbol{\omega})$, where $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$ is sampled on a product extension and is conditionally independent of $\boldsymbol{\eta}_t$ given $\mathcal{G}_t$. □

Lemma 13 (Self-normalized quadratic noise bound)**.** *For each round, let*

$$\boldsymbol{\Delta}_t = \mathbf{V}_{t+1} - \mathbf{V}_t = \mathbf{S}_t(\pi_t).$$

Define

$$\tau_t = \operatorname{tr}(\mathbf{V}_{t+1}^{-1}\boldsymbol{\Delta}_t).$$

Let

$$\boldsymbol{\nu}_t = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi_t(i)}\eta_{t,i}.$$

Then there exists a numerical constant $C > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} \le C\sigma_\eta^2 \left(K_T + \log\frac{1}{\delta}\right).$$

Proof. For a selected matching π_t, let

$$\mathbf{Z}_t = \begin{bmatrix} \mathbf{z}_{t,1,\pi_t(1)} & \cdots & \mathbf{z}_{t,M,\pi_t(M)} \end{bmatrix} \in \mathbb{R}^{p\times M}.$$

Also let

$$\boldsymbol{\eta}_t = (\eta_{t,1},\dots,\eta_{t,M})^\top.$$

Then

$$\boldsymbol{\nu}_t = \mathbf{Z}_t\boldsymbol{\eta}_t, \qquad \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} = \boldsymbol{\eta}_t^\top \mathbf{A}_t\boldsymbol{\eta}_t,$$

where

$$\mathbf{A}_t = \mathbf{Z}_t^\top \mathbf{V}_{t+1}^{-1}\mathbf{Z}_t.$$

We first condition on $\mathcal{G}_t$. Then π_t, $\mathbf{Z}_t$, $\mathbf{V}_{t+1}$, and $\mathbf{A}_t$ are fixed. In every round,

$$\mathbf{Z}_t\mathbf{Z}_t^\top = \mathbf{S}_t(\pi_t) = \boldsymbol{\Delta}_t \preceq \mathbf{V}_{t+1}.$$

Therefore,

$$\mathbf{V}_{t+1}^{-1/2}\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{V}_{t+1}^{-1/2} \preceq \mathbf{I}_p.$$

Since $\mathbf{A}_t = \mathbf{Z}_t^\top\mathbf{V}_{t+1}^{-1}\mathbf{Z}_t$ has the same nonzero eigenvalues as $\mathbf{V}_{t+1}^{-1/2}\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{V}_{t+1}^{-1/2}$, we obtain

$$\|\mathbf{A}_t\|_{\mathrm{op}} \le 1.$$

Let $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$ be an auxiliary Gaussian vector sampled on a product extension such that

$$\sigma(\boldsymbol{\omega}) \perp\!\!\!\perp \sigma(\boldsymbol{\eta}_t) \mid \mathcal{G}_t.$$

For every $\alpha \ge 0$, the Gaussian-mixture identity gives

$$\exp(\alpha\boldsymbol{\eta}_t^\top\mathbf{A}_t\boldsymbol{\eta}_t) = \mathbb{E}_{\boldsymbol{\omega}}\left[\exp\left(\sqrt{2\alpha}\,\boldsymbol{\omega}^\top\mathbf{A}_t^{1/2}\boldsymbol{\eta}_t\right)\right].$$

Conditioning further on $\boldsymbol{\omega}$, the vector $\sqrt{2\alpha}\mathbf{A}_t^{1/2}\boldsymbol{\omega}$ is $\mathcal{G}_t \vee \sigma(\boldsymbol{\omega})$-measurable. Since the integrand is nonnegative, conditional Tonelli's theorem on the product extension allows us to exchange the auxiliary Gaussian expectation and the conditional expectation over $\boldsymbol{\eta}_t$. Hence

$$\begin{aligned}
\mathbb{E}\left[\exp(\alpha\boldsymbol{\eta}_t^\top\mathbf{A}_t\boldsymbol{\eta}_t) \mid \mathcal{G}_t\right] &= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\omega}}\exp\left(\sqrt{2\alpha}\,\boldsymbol{\omega}^\top\mathbf{A}_t^{1/2}\boldsymbol{\eta}_t\right) \mid \mathcal{G}_t\right] \\
&= \mathbb{E}_{\boldsymbol{\omega}}\mathbb{E}\left[\exp\left(\sqrt{2\alpha}\,\boldsymbol{\omega}^\top\mathbf{A}_t^{1/2}\boldsymbol{\eta}_t\right) \mid \mathcal{G}_t \vee \sigma(\boldsymbol{\omega})\right] \\
&\le \mathbb{E}_{\boldsymbol{\omega}}\exp\left(\sigma_\eta^2\alpha\,\boldsymbol{\omega}^\top\mathbf{A}_t\boldsymbol{\omega}\right),
\end{aligned}$$

where the last inequality uses the enlarged-filtration version of Lemma 12. Choose a numerical constant $c_1 \le 1/4$. For $0 \le \alpha \le c_1/\sigma_\eta^2$, the bound $\|\mathbf{A}_t\|_{\mathrm{op}} \le 1$ implies

$$2\sigma_\eta^2\alpha\|\mathbf{A}_t\|_{\mathrm{op}} \le 2c_1 \le \frac{1}{2}.$$

Therefore, the Gaussian quadratic-form moment formula yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\omega}} \exp\left(\sigma_\eta^2 \alpha \boldsymbol{\omega}^\top \mathbf{A}_t \boldsymbol{\omega}\right) &= \det\left(\mathbf{I}_M - 2\sigma_\eta^2 \alpha \mathbf{A}_t\right)^{-1/2} \\ &\le \exp\left(C\sigma_\eta^2 \alpha \operatorname{tr}(\mathbf{A}_t)\right).\end{aligned}$$

Since $\boldsymbol{\Delta}_t = \mathbf{S}_t(\pi_t) = \mathbf{Z}_t \mathbf{Z}_t^\top$,

$$\begin{aligned}\operatorname{tr}(\mathbf{A}_t) &= \operatorname{tr}(\mathbf{Z}_t^\top \mathbf{V}_{t+1}^{-1} \mathbf{Z}_t) \\ &= \operatorname{tr}(\mathbf{V}_{t+1}^{-1} \mathbf{Z}_t \mathbf{Z}_t^\top) \\ &= \operatorname{tr}(\mathbf{V}_{t+1}^{-1} \boldsymbol{\Delta}_t) \\ &= \tau_t.\end{aligned}$$

Thus, for all rounds and every $0 \le \alpha \le c_1/\sigma_\eta^2$,

$$\mathbb{E}\left[\exp\left(\alpha \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} - C\sigma_\eta^2 \alpha \tau_t\right) \mid \mathcal{G}_t\right] \le 1.$$

Fix $\alpha = c_1/\sigma_\eta^2$. Define the post-reward filtration

$$\mathcal{J}_t = \mathcal{H}_{t,M}, \qquad t \ge 1,$$

and let $\mathcal{J}_0$ be the initial sigma-field before the first round. Then

$$\mathcal{J}_{t-1} \subseteq \mathcal{F}_t \subseteq \mathcal{G}_t,$$

where $\mathcal{F}_t$ is the pre-action filtration. Define

$$Y_t = \exp\left(\alpha \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} - C\sigma_\eta^2 \alpha \tau_t\right).$$

The preceding conditional moment bound gives

$$\begin{aligned}\mathbb{E}[Y_t \mid \mathcal{J}_{t-1}] &= \mathbb{E}\left[\mathbb{E}[Y_t \mid \mathcal{G}_t] \mid \mathcal{J}_{t-1}\right] \\ &\le 1.\end{aligned}$$

Therefore, the process

$$M_n = \prod_{t=1}^n Y_t = \exp\left(\alpha \sum_{t=1}^n \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} - C\sigma_\eta^2 \alpha \sum_{t=1}^n \tau_t\right)$$

is a nonnegative supermartingale with respect to $(\mathcal{J}_n)_{n \ge 0}$. Indeed,

$$\begin{aligned}\mathbb{E}[M_n \mid \mathcal{J}_{n-1}] &= M_{n-1} \mathbb{E}[Y_n \mid \mathcal{J}_{n-1}] \\ &\le M_{n-1}.\end{aligned}$$

Therefore, by Markov's inequality, with probability at least $1 - \delta$,

$$\sum_{t=1}^T \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}} \le C\sigma_\eta^2 \left(\sum_{t=1}^T \tau_t + \log \frac{1}{\delta}\right).$$

It remains to control $\sum_{t=1}^T \tau_t$. Let $\alpha_{t,1}, \dots, \alpha_{t,p}$ be the eigenvalues of $\mathbf{V}_t^{-1/2} \boldsymbol{\Delta}_t \mathbf{V}_t^{-1/2}$. Then

$$\begin{aligned}\tau_t &= \operatorname{tr}(\mathbf{V}_{t+1}^{-1} \boldsymbol{\Delta}_t) \\ &= \sum_{k=1}^p \frac{\alpha_{t,k}}{1 + \alpha_{t,k}} \\ &\le \sum_{k=1}^p \log(1 + \alpha_{t,k}) \\ &= \log \frac{\det \mathbf{V}_{t+1}}{\det \mathbf{V}_t}.\end{aligned}$$

Summing over t and using the determinant-potential bound gives

$$\sum_{t=1}^T \tau_t \le \log \frac{\det \mathbf{V}_{T+1}}{\det \mathbf{V}_1} \le K_T.$$

This proves the claim. □

B.5 Prediction-error recursion and cumulative error

Let

$$\mathbf{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star, \qquad \Psi_t = \mathbf{e}_t^\top \mathbf{V}_t \mathbf{e}_t.$$

For low-width rounds $t \in \mathcal{B}$, define

$$E_t = \mathbf{e}_t^\top \mathbf{H}_t \mathbf{e}_t, \qquad \widetilde{E}_t = \mathbf{e}_t^\top \mathbf{S}_t(\pi_t)\mathbf{e}_t, \qquad \Delta_t^E = \widetilde{E}_t - E_t.$$

Since $\pi_t \sim \mu_t$ conditionally on $\mathcal{F}_t$ and $\mathbf{H}_t = \mathbb{E}_{\pi\sim\mu_t}\mathbf{S}_t(\pi)$, we have

$$\mathbb{E}[\Delta_t^E \mid \mathcal{F}_t] = 0.$$

For all rounds, define

$$\boldsymbol{\nu}_t = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi_t(i)}\eta_{t,i}.$$

Under the direct realized-design update, in every round we have

$$\mathbf{V}_{t+1}\mathbf{e}_{t+1} = \mathbf{V}_t\mathbf{e}_t + \boldsymbol{\nu}_t.$$

Indeed,

$$\begin{aligned}
\mathbf{V}_{t+1}\mathbf{e}_{t+1} &= \mathbf{b}_{t+1} - \mathbf{V}_{t+1}\boldsymbol{\theta}^\star \\
&= \mathbf{b}_t + \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi_t(i)} r_{t,i} - \left(\mathbf{V}_t + \mathbf{S}_t(\pi_t)\right)\boldsymbol{\theta}^\star \\
&= \mathbf{b}_t - \mathbf{V}_t\boldsymbol{\theta}^\star + \boldsymbol{\nu}_t \\
&= \mathbf{V}_t\mathbf{e}_t + \boldsymbol{\nu}_t.
\end{aligned}$$

Therefore,

$$\Psi_{t+1} = \Psi_t - D_t + n_t + \chi_t,$$

where

$$D_t = \Psi_t - \|\mathbf{V}_t\mathbf{e}_t\|^2_{\mathbf{V}_{t+1}^{-1}} \geq 0,$$

and

$$n_t = 2\boldsymbol{\nu}_t^\top \mathbf{V}_{t+1}^{-1}\mathbf{V}_t\mathbf{e}_t, \qquad \chi_t = \|\boldsymbol{\nu}_t\|^2_{\mathbf{V}_{t+1}^{-1}}.$$

$D_t \geq 0$ follows from $\mathbf{V}_{t+1} \succeq \mathbf{V}_t$, which implies $\mathbf{V}_{t+1}^{-1} \preceq \mathbf{V}_t^{-1}$.

For low-width rounds, the dissipation term controls the realized quadratic prediction error. For any feasible matching $\pi \in \Pi$, write

$$\mathbf{Y}_t(\pi) = \mathbf{V}_t^{-1/2}\mathbf{S}_t(\pi)\mathbf{V}_t^{-1/2}.$$

Let $\mathbf{v}_t = \mathbf{V}_t^{1/2}\mathbf{e}_t$. By Lemma 11, $\mathbf{Y}_t(\pi_t) \preceq \rho\mathbf{I}_p$, and

$$D_t = \mathbf{v}_t^\top \mathbf{Y}_t(\pi_t)(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}\mathbf{v}_t.$$

This follows from $\mathbf{V}_{t+1}^{-1} = (\mathbf{V}_t + \mathbf{S}_t(\pi_t))^{-1}$ and the Sherman–Morrison–Woodbury identity. Hence

$$\begin{aligned}
D_t &\geq \frac{1}{1+\rho}\mathbf{v}_t^\top \mathbf{Y}_t(\pi_t)\mathbf{v}_t \\
&= \frac{1}{1+\rho}\mathbf{e}_t^\top \mathbf{S}_t(\pi_t)\mathbf{e}_t \\
&= \frac{1}{1+\rho}\widetilde{E}_t.
\end{aligned}$$

Equivalently,

$$\widetilde{E}_t \leq (1+\rho)D_t.$$

Finally, we record two elementary consequences of the low-width condition that will be used below. Since $\mathbf{H}_t \preceq \rho\mathbf{V}_t$ and $\mathbf{S}_t(\pi_t) \preceq \rho\mathbf{V}_t$ on low-width rounds,

$$0 \leq E_t \leq \rho\Psi_t, \qquad 0 \leq \widetilde{E}_t \leq \rho\Psi_t.$$

Therefore,

$$|\Delta_t^E| \le \rho \Psi_t.$$

Moreover,

$$\begin{aligned}
\mathrm{Var}(\Delta_t^E \mid \mathcal{F}_t) &\le \mathbb{E}[(\Delta_t^E)^2 \mid \mathcal{F}_t] \\
&= \mathbb{E}[(\widetilde{E}_t - E_t)^2 \mid \mathcal{F}_t] \\
&\le \mathbb{E}[\widetilde{E}_t^2 \mid \mathcal{F}_t] \\
&\le \rho \Psi_t \mathbb{E}[\widetilde{E}_t \mid \mathcal{F}_t] \\
&= \rho \Psi_t E_t,
\end{aligned}$$

where the second inequality uses $\mathbb{E}[\widetilde{E}_t \mid \mathcal{F}_t] = E_t$, so that $\mathbb{E}[(\widetilde{E}_t - E_t)^2 \mid \mathcal{F}_t] = \mathbb{E}[\widetilde{E}_t^2 \mid \mathcal{F}_t] - E_t^2 \le \mathbb{E}[\widetilde{E}_t^2 \mid \mathcal{F}_t]$, and the third uses $\widetilde{E}_t \le \rho \Psi_t$.

Auxiliary Freedman comparison. We will use the following one-sided comparison form of Freedman's inequality [11]. Let $\mathcal{J}_{t-1} \subseteq \mathcal{F}_t \subseteq \mathcal{J}_t$ be a two-stage filtration, and let X_t be $\mathcal{J}_t$-measurable with $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. Suppose that $a_t \ge 0$ is $\mathcal{F}_t$-measurable and that, for some deterministic $b > 0$,

$$|X_t| \le b, \qquad \mathrm{Var}(X_t \mid \mathcal{F}_t) \le b a_t \qquad \text{a.s. for all } t.$$

Then there exists a numerical constant $C_{\mathrm{F}} > 0$ such that, for every $\varepsilon_{\mathrm{F}} \in (0,1)$ and every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} X_t \le \varepsilon_{\mathrm{F}} \sum_{t=1}^{T} a_t + \frac{C_{\mathrm{F}} b}{\varepsilon_{\mathrm{F}}} \log \frac{1}{\delta}. \tag{15}$$

Indeed, the boundedness and conditional-variance assumptions imply the conditional Bernstein bound

$$\mathbb{E}\left[\exp(\alpha X_t) \mid \mathcal{F}_t\right] \le \exp(C \alpha^2 b a_t) \qquad \text{for } 0 \le \alpha \le c/b,$$

with numerical constants $c, C > 0$. Hence

$$\exp\left(\alpha \sum_{t=1}^{n} X_t - C\alpha^2 b \sum_{t=1}^{n} a_t\right)$$

is a nonnegative supermartingale with respect to $(\mathcal{J}_n)_{n \ge 0}$. Markov's inequality and the choice $\alpha \asymp \varepsilon_{\mathrm{F}}/b$ give (15).

Lemma 14 (Low-width cumulative prediction-error bound)**.** *There exist numerical constants* $0 < c_\rho \le 1$ *and* $C_1 > 0$ *such that if*

$$\rho \le \frac{c_\rho}{\ell_\delta},$$

then, with probability at least $1 - \delta$*,*

$$\sum_{t \in \mathcal{B}} E_t \le C_1 A_T,$$

where A_T *is defined in* (14)*.*

Proof. By increasing the numerical constant in the definition of ℓ_δ, we ensure that logarithmic factors such as $\log(5/\delta)$ are bounded by ℓ_δ. Fix a deterministic threshold $\Gamma > 0$ and define the stopping time

$$\tau_\Gamma = \inf\{t \ge 1 : \Psi_t > \Gamma\},$$

with the convention $\inf \emptyset = \infty$. Throughout the proof, sums of the form $\sum_{t < \tau_\Gamma}$ mean $\sum_{t=1}^{T} \mathbb{1}\{t < \tau_\Gamma\}(\cdot)$. On this stopped process, $\Psi_t \le \Gamma$ whenever $t < \tau_\Gamma$.

For $t \in \mathcal{B}$, the low-width covariance domination gives

$$\mathbf{S}_t(\pi_t) \preceq \rho \mathbf{V}_t, \qquad \mathbf{H}_t \preceq \rho \mathbf{V}_t.$$

Hence, on the stopped process,

$$0 \le \widetilde{E}_t \le \rho\Gamma, \qquad 0 \le E_t \le \rho\Gamma, \qquad |\Delta_t^E| \le \rho\Gamma.$$

Moreover,

$$\begin{aligned}\mathrm{Var}(\Delta_t^E \mid \mathcal{F}_t) &\le \mathbb{E}[(\Delta_t^E)^2 \mid \mathcal{F}_t] \\ &= \mathbb{E}[(\widetilde{E}_t - E_t)^2 \mid \mathcal{F}_t] \\ &\le \mathbb{E}[\widetilde{E}_t^2 \mid \mathcal{F}_t] \\ &\le \rho\Gamma\mathbb{E}[\widetilde{E}_t \mid \mathcal{F}_t] \\ &= \rho\Gamma E_t.\end{aligned}$$

Apply the auxiliary Freedman comparison to

$$X_t = -\mathbb{1}\{t \in \mathcal{B},\ t < \tau_\Gamma\}\Delta_t^E, \qquad a_t = \mathbb{1}\{t \in \mathcal{B},\ t < \tau_\Gamma\}E_t, \qquad b = \rho\Gamma.$$

The indicators are $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$, and the bounds above give

$$|X_t| \le b, \qquad \mathrm{Var}(X_t \mid \mathcal{F}_t) \le b a_t.$$

Therefore, with probability at least $1 - \delta/5$,

$$-\sum_{t\in\mathcal{B},t<\tau_\Gamma} \Delta_t^E \le \epsilon_0 \sum_{t\in\mathcal{B},t<\tau_\Gamma} E_t + \frac{C\rho\Gamma}{\epsilon_0}\ell_\delta. \tag{16}$$

Next consider the martingale term n_t. The following control of n_t is over all stopped rounds, not only low-width rounds, because the ridge recursion is summed over the entire realized design. Conditional on $\mathcal{G}_t$, the term n_t is a mean-zero linear form in the same-round noise vector $\boldsymbol{\eta}_t$. Write

$$\mathbf{S}_t(\pi_t) = \mathbf{Z}_t\mathbf{Z}_t^\top.$$

Under the direct update,

$$n_t = 2\boldsymbol{\eta}_t^\top \mathbf{Z}_t^\top \mathbf{V}_{t+1}^{-1}\mathbf{V}_t\mathbf{e}_t.$$

By Lemma 12, its conditional sub-Gaussian variance proxy is at most

$$4\sigma_\eta^2 \left\|\mathbf{Z}_t^\top \mathbf{V}_{t+1}^{-1}\mathbf{V}_t\mathbf{e}_t\right\|_2^2.$$

Then

$$D_t = \mathbf{v}_t^\top \mathbf{Y}_t(\pi_t)(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}\mathbf{v}_t,$$

and

$$\begin{aligned}\left\|\mathbf{Z}_t^\top \mathbf{V}_{t+1}^{-1}\mathbf{V}_t\mathbf{e}_t\right\|_2^2 &= \mathbf{v}_t^\top(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}\mathbf{Y}_t(\pi_t)(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}\mathbf{v}_t \\ &\le \mathbf{v}_t^\top \mathbf{Y}_t(\pi_t)(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}\mathbf{v}_t \\ &= D_t.\end{aligned}$$

Here the inequality uses $(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1} \preceq \mathbf{I}_p$ and the fact that $\mathbf{Y}_t(\pi_t)$ commutes with $(\mathbf{I}_p + \mathbf{Y}_t(\pi_t))^{-1}$. Let $\mathbb{1}_t^\Gamma = \mathbb{1}\{t < \tau_\Gamma\}$. Since $\mathbb{1}_t^\Gamma$ is $\mathcal{F}_t$-measurable and hence $\mathcal{G}_t$-measurable, the preceding conditional sub-Gaussian bound implies

$$\mathbb{E}\left[\exp\left(\alpha\mathbb{1}_t^\Gamma n_t - C\sigma_\eta^2\alpha^2\mathbb{1}_t^\Gamma D_t\right) \mid \mathcal{G}_t\right] \le 1 \qquad \forall\alpha \in \mathbb{R}.$$

Define the post-reward filtration

$$\mathcal{J}_t = \mathcal{H}_{t,M}, \qquad t \ge 1,$$

and let $\mathcal{J}_0$ be the initial sigma-field before the first round. Since $\mathcal{J}_{t-1} \subseteq \mathcal{F}_t \subseteq \mathcal{G}_t$, the process

$$M_n = \exp\left(\alpha\sum_{t=1}^n \mathbb{1}_t^\Gamma n_t - C\sigma_\eta^2\alpha^2\sum_{t=1}^n \mathbb{1}_t^\Gamma D_t\right)$$

is a nonnegative supermartingale with respect to $(\mathcal{J}_n)_{n\ge0}$. Applying Markov's inequality with $\alpha = c\epsilon_0/\sigma_\eta^2$ for a sufficiently small numerical constant $c > 0$ gives, with probability at least $1 - \delta/5$,

$$\sum_{t<\tau_\Gamma} n_t \le \epsilon_0 \sum_{t<\tau_\Gamma} D_t + \frac{C\sigma_\eta^2}{\epsilon_0}\ell_\delta. \tag{17}$$

By Lemma 13, with probability at least $1 - \delta/5$,

$$\sum_{t<\tau_\Gamma} \chi_t \le \sum_{t=1}^{T} \chi_t \le C\sigma_\eta^2(K_T + \ell_\delta). \tag{18}$$

Now sum the realized-design recursion over all stopped rounds $t < \tau_\Gamma$:

$$\Psi_{t+1} = \Psi_t - D_t + n_t + \chi_t.$$

The left-hand side telescopes and gives

$$\Psi_{\tau_\Gamma \wedge (T+1)} + \sum_{t<\tau_\Gamma} D_t \le \Psi_1 + \sum_{t<\tau_\Gamma} n_t + \sum_{t<\tau_\Gamma} \chi_t.$$

We work on the intersection of the three high-probability events (16), (17), and (18). This intersection has probability at least $1 - 3\delta/5$. No additional union bound is needed below, since the same three events are reused throughout the stopping argument. Since $\Psi_1 = \lambda\|\boldsymbol{\theta}^\star\|_2^2 \le \lambda S^2$, choosing $\epsilon_0 > 0$ sufficiently small yields numerical constants $c_2, C_2 > 0$ such that

$$\Psi_{\tau_\Gamma \wedge (T+1)} + c_2 \sum_{t<\tau_\Gamma} D_t \le C_2 \left(\lambda S^2 + \sigma_\eta^2(K_T + \ell_\delta)\right). \tag{19}$$

Here

$$\Psi_1 = \lambda\|\boldsymbol{\theta}^\star\|_2^2 \le \lambda S^2.$$

Set

$$\Gamma_\star = 2C_2 A_T.$$

Apply (19) with $\Gamma = \Gamma_\star$. Then

$$\Psi_{\tau_{\Gamma_\star} \wedge (T+1)} \le \frac{\Gamma_\star}{2}.$$

If $\tau_{\Gamma_\star} \le T + 1$, then by the definition of the stopping time,

$$\Psi_{\tau_{\Gamma_\star}} > \Gamma_\star,$$

which contradicts the preceding display. Hence

$$\tau_{\Gamma_\star} > T + 1.$$

Therefore $\Psi_t \le \Gamma_\star$ for all $t \le T + 1$. In particular, for $\Gamma = \Gamma_\star$, the stopped process coincides with the unstopped process through horizon T: for every $t \le T$, we have $\mathbb{1}\{t < \tau_{\Gamma_\star}\} = 1$. Thus, on the same high-probability event, (19) gives

$$\Psi_{T+1} + c_2 \sum_{t=1}^{T} D_t \le C_2 A_T.$$

Consequently,

$$\sum_{t=1}^{T} D_t \le C A_T. \tag{20}$$

It remains to convert this realized-design dissipation bound into the desired bound on $\sum_{t\in\mathcal{B}} E_t$. Using the martingale comparison event (16) with $\Gamma = \Gamma_\star$ and the fact that $\tau_{\Gamma_\star} > T + 1$, we obtain

$$-\sum_{t\in\mathcal{B}} \Delta_t^E \le \epsilon_0 \sum_{t\in\mathcal{B}} E_t + \frac{C\rho\Gamma_\star}{\epsilon_0}\ell_\delta.$$

Since

$$\sum_{t\in\mathcal{B}} E_t = \sum_{t\in\mathcal{B}} \widetilde{E}_t - \sum_{t\in\mathcal{B}} \Delta_t^E,$$

we get

$$\sum_{t\in\mathcal{B}} E_t \le \sum_{t\in\mathcal{B}} \widetilde{E}_t + \epsilon_0 \sum_{t\in\mathcal{B}} E_t + \frac{C\rho\Gamma_\star}{\epsilon_0}\ell_\delta.$$

Using $\widetilde{E}_t \le (1+\rho)D_t$ on low-width rounds,

$$\sum_{t\in\mathcal{B}} \widetilde{E}_t \le (1+\rho)\sum_{t\in\mathcal{B}} D_t \le (1+\rho)\sum_{t=1}^{T} D_t. \tag{21}$$

Combining (20) and (21), and choosing $\epsilon_0 > 0$ small enough, gives

$$\sum_{t\in\mathcal{B}} E_t \le CA_T + C\rho\Gamma_\star\ell_\delta.$$

Since $\rho \le c_\rho/\ell_\delta$ and $\Gamma_\star \le CA_T$, the second term is at most $C'c_\rho A_T$ for another numerical constant $C' > 0$. Taking $c_\rho > 0$ sufficiently small and renaming the final numerical constant gives

$$\sum_{t\in\mathcal{B}} E_t \le C_1 A_T.$$

This proves the lemma. □

B.6 Low-width regret and conditional regret

For $t \in \mathcal{B}$, define the conditional one-step regret

$$\overline{\mathrm{reg}}_t = F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t^\star) - \mathbb{E}\left[F_{\boldsymbol{\Theta}^\star}(C_t, \pi_t) \mid \mathcal{F}_t\right]. \tag{22}$$

Lemma 15 (One-step low-width regret)**.** *Suppose that on every low-width round, the RRD certificate in Lemma 5 holds with tolerance ε. Then, for every $t \in \mathcal{B}$,*

$$\overline{\mathrm{reg}}_t \le \frac{\xi}{2}E_t + \frac{2(Md_2+\varepsilon)}{\xi}. \tag{23}$$

Proof. Write $\Lambda = Md_2 + \varepsilon$. Also write

$$\mathbf{a}_t(\pi) = \sum_{i=1}^{M} \mathbf{z}_{t,i,\pi(i)}, \qquad \mathbf{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star,$$

where $\boldsymbol{\theta}^\star = \mathrm{vec}(\boldsymbol{\Theta}^\star)$. If $q_t = 0$, then all edge features are zero, so $\overline{\mathrm{reg}}_t = E_t = 0$. Hence assume $q_t > 0$. Define the mean aggregate feature and the predicted reward difference by

$$\mathbf{a}_t^\mu := \mathbb{E}[\mathbf{a}_t(\pi_t) \mid \mathcal{F}_t], \qquad \Delta_t := F_{\boldsymbol{\theta}_t}(C_t, \pi_t^\star) - \langle \mathbf{P}_t, \widehat{\mathbf{R}}_t\rangle.$$

Since $\pi_t \sim \mu_t = \zeta_{\mathbf{P}_t}$ and $(\widehat{\mathbf{R}}_t)_{i,j} = \mathbf{z}_{t,i,j}^\top \boldsymbol{\theta}_t$, the marginal property gives

$$\mathbb{E}[F_{\boldsymbol{\theta}_t}(C_t, \pi_t) \mid \mathcal{F}_t] = \langle \mathbf{P}_t, \widehat{\mathbf{R}}_t\rangle.$$

We first bound the covariance norm of $\mathbf{a}_t^\mu$. For every feasible matching π and every $\mathbf{v} \in \mathbb{R}^p$, Cauchy–Schwarz gives

$$\begin{aligned}\left(\mathbf{v}^\top \mathbf{a}_t(\pi)\right)^2 &= \left(\sum_{i=1}^{M} \mathbf{v}^\top \mathbf{z}_{t,i,\pi(i)}\right)^2 \\ &\le M\sum_{i=1}^{M}\left(\mathbf{v}^\top \mathbf{z}_{t,i,\pi(i)}\right)^2 \\ &= M\mathbf{v}^\top \mathbf{S}_t(\pi)\mathbf{v}.\end{aligned}$$

Thus

$$\mathbf{a}_t(\pi)\mathbf{a}_t(\pi)^\top \preceq M\mathbf{S}_t(\pi).$$

Taking conditional expectations and using $\mathbf{a}_t^\mu = \mathbb{E}[\mathbf{a}_t(\pi_t) \mid \mathcal{F}_t]$ yields

$$\mathbf{a}_t^\mu(\mathbf{a}_t^\mu)^\top \preceq \mathbb{E}[\mathbf{a}_t(\pi_t)\mathbf{a}_t(\pi_t)^\top \mid \mathcal{F}_t] \preceq M\mathbf{H}_t.$$

Consequently,

$$\mathbf{a}_t^\mu \in \mathrm{range}(\mathbf{H}_t), \qquad \|\mathbf{a}_t^\mu\|_{\mathbf{H}_t^\dagger}^2 \le M. \tag{24}$$

Indeed, the range assertion follows by considering vectors in $\ker(\mathbf{H}_t)$. Taking $\mathbf{v} = \mathbf{H}_t^\dagger \mathbf{a}_t^\mu$ in the preceding quadratic inequality and using $\mathbf{H}_t^\dagger \mathbf{H}_t \mathbf{H}_t^\dagger = \mathbf{H}_t^\dagger$ then gives the norm bound.

By the RRD certificate in Lemma 5,

$$\mathbf{a}_t(\pi_t^\star) \in \mathrm{range}(\mathbf{H}_t)$$

and

$$\|\mathbf{a}_t(\pi_t^\star)\|_{\mathbf{H}_t^\dagger}^2 \le M q_t + \varepsilon - \xi \Delta_t.$$

Combining this with (24) gives

$$\begin{aligned} &\|\mathbf{a}_t(\pi_t^\star) - \mathbf{a}_t^\mu\|_{\mathbf{H}_t^\dagger}^2 \\ &\quad \le 2\|\mathbf{a}_t(\pi_t^\star)\|_{\mathbf{H}_t^\dagger}^2 + 2\|\mathbf{a}_t^\mu\|_{\mathbf{H}_t^\dagger}^2 \\ &\quad \le 2\left(M(q_t+1) + \varepsilon - \xi\Delta_t\right). \end{aligned} \tag{25}$$

Using $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_t - \mathbf{e}_t$, the conditional one-step regret can be written as

$$\begin{aligned} \overline{\mathrm{reg}}_t &= \langle \boldsymbol{\theta}^\star, \mathbf{a}_t(\pi_t^\star) - \mathbf{a}_t^\mu \rangle \\ &= \Delta_t - \langle \mathbf{e}_t, \mathbf{a}_t(\pi_t^\star) - \mathbf{a}_t^\mu \rangle . \end{aligned}$$

By the range assertions above, generalized Cauchy–Schwarz and (25) give

$$\overline{\mathrm{reg}}_t \le \Delta_t + \sqrt{2E_t\left(M(q_t+1) + \varepsilon - \xi\Delta_t\right)}.$$

The expression inside the square root is nonnegative by the preceding covariance-norm bound. Applying Young's inequality yields

$$\begin{aligned} \overline{\mathrm{reg}}_t &\le \Delta_t + \frac{\xi}{2}E_t + \frac{M(q_t+1) + \varepsilon - \xi\Delta_t}{\xi} \\ &= \frac{\xi}{2}E_t + \frac{M(q_t+1) + \varepsilon}{\xi} \\ &\le \frac{\xi}{2}E_t + \frac{2\Lambda}{\xi}, \end{aligned}$$

where the last inequality uses $1 \le q_t \le d_2$ and $\Lambda = M d_2 + \varepsilon$. This proves (23). □

Define the conditional low-width regret

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) = \sum_{t \in \mathcal{B}} \overline{\mathrm{reg}}_t.$$

Corollary 16 (Low-width conditional regret)**.** *Under the conditions of Lemmas 14 and 15, with the choice*

$$\xi = \frac{3}{8\sqrt{C_1}} \sqrt{\frac{(M d_2 + \varepsilon)T}{A_T}}, \tag{26}$$

we have with probability at least $1 - \delta$*,*

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) \le C_3 \sqrt{(M d_2 + \varepsilon) T A_T}, \tag{27}$$

where $C_3 = 6\sqrt{C_1}$ *and* A_T *is as defined in* (14)*.*

Proof. Write $\Lambda = M d_2 + \varepsilon$. Summing (23) over $t \in \mathcal{B}$ and applying Lemma 14 gives

$$\begin{aligned} \overline{\mathrm{Reg}}_{\mathcal{B}}(T) &\le \frac{\xi}{2} \sum_{t \in \mathcal{B}} E_t + \frac{2\Lambda T}{\xi} \\ &\le \frac{\xi C_1 A_T}{2} + \frac{2\Lambda T}{\xi}. \end{aligned}$$

Substituting the choice (26) yields

$$\begin{aligned}\overline{\mathrm{Reg}}_{\mathcal{B}}(T) &\le \left(\frac{3}{16}+\frac{16}{3}\right)\sqrt{C_1\Lambda T A_T} \\ &= \frac{265}{48}\sqrt{C_1\Lambda T A_T} \\ &\le 6\sqrt{C_1\Lambda T A_T}.\end{aligned}$$

This proves (27) with $C_3 = 6\sqrt{C_1}$. □

Conditional regret. Define the cumulative conditional regret by

$$\overline{\mathrm{Reg}}(T) = \mathrm{Reg}_{\mathcal{A}}(T) + \overline{\mathrm{Reg}}_{\mathcal{B}}(T),$$

where

$$\mathrm{Reg}_{\mathcal{A}}(T) = \sum_{t\in\mathcal{A}} \left[F_{\mathbf{\Theta}^\star}(C_t,\pi_t^\star) - F_{\mathbf{\Theta}^\star}(C_t,\pi_t)\right],$$

and

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) = \sum_{t\in\mathcal{B}} \left[F_{\mathbf{\Theta}^\star}(C_t,\pi_t^\star) - \mathbb{E}\left[F_{\mathbf{\Theta}^\star}(C_t,\pi_t)\mid\mathcal{F}_t\right]\right].$$

On high-width rounds, $\pi_t = \pi_t^w$ is $\mathcal{F}_t$-measurable under deterministic tie-breaking, so the high-width contribution is already equal to its conditional counterpart.

Corollary 17 (Conditional regret bound)**.** *Under the conditions of Corollary 16 and with $\rho = c_\rho/\ell_\delta$, with probability at least $1-\delta$,*

$$\overline{\mathrm{Reg}}(T) \le C_3\sqrt{(Md_2+\varepsilon)TA_T} + C_4 BMK_T\ell_\delta. \tag{28}$$

Proof. By Lemma 10,

$$\mathrm{Reg}_{\mathcal{A}}(T) \le \frac{4BMK_T}{\rho}.$$

By Corollary 16,

$$\overline{\mathrm{Reg}}_{\mathcal{B}}(T) \le C_3\sqrt{(Md_2+\varepsilon)TA_T}.$$

Using $\rho = c_\rho/\ell_\delta$ and absorbing $1/c_\rho$ into the numerical constant gives the claim. □

B.7 Conditional-to-actual regret conversion

Lemma 18 (Conditional-to-actual regret conversion)**.** *For every round, define the played-action one-step regret*

$$\mathrm{reg}_t = F_{\mathbf{\Theta}^\star}(C_t,\pi_t^\star) - F_{\mathbf{\Theta}^\star}(C_t,\pi_t),$$

and let $\overline{\mathrm{reg}}_t = \mathbb{E}[\mathrm{reg}_t \mid \mathcal{F}_t]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\mathrm{Reg}(T) \le \frac{3}{2}\overline{\mathrm{Reg}}(T) + \frac{3C_5}{2}BM\ell_\delta. \tag{29}$$

Proof. Note that

$$\mathrm{Reg}(T) = \sum_{t=1}^{T}\mathrm{reg}_t, \qquad \overline{\mathrm{Reg}}(T) = \sum_{t=1}^{T}\overline{\mathrm{reg}}_t,$$

where the second identity holds because high-width actions are $\mathcal{F}_t$-measurable, whereas low-width actions are sampled from μ_t. Set $X_t = \mathrm{reg}_t - \overline{\mathrm{reg}}_t$. Then $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. Since $\pi_t^\star$ maximizes the true mean reward and each matching mean reward is bounded in absolute value by BM,

$$0 \le \mathrm{reg}_t \le 2BM, \qquad |X_t| \le 2BM.$$

Moreover,

$$\mathrm{Var}(X_t\mid\mathcal{F}_t) \le \mathbb{E}\left[\mathrm{reg}_t^2 \mid \mathcal{F}_t\right] \le 2BM\,\mathbb{E}\left[\mathrm{reg}_t\mid\mathcal{F}_t\right] = 2BM\,\overline{\mathrm{reg}}_t.$$

Apply the auxiliary Freedman comparison with

$$a_t = \overline{\mathrm{reg}}_t, \qquad b = 2BM, \qquad \varepsilon_{\mathrm{F}} = \frac{1}{2}.$$

The assumptions are satisfied because $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$, $|X_t| \le 2BM$, and

$$\mathrm{Var}(X_t \mid \mathcal{F}_t) \le 2BM\,\overline{\mathrm{reg}}_t = b a_t.$$

Thus, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} X_t \le \frac{1}{2}\sum_{t=1}^{T} \overline{\mathrm{reg}}_t + C_5 BM\ell_\delta \le \frac{1}{2}\overline{\mathrm{Reg}}(T) + \frac{3C_5}{2} BM\ell_\delta,$$

where the last step only enlarges the numerical constant. Since $\mathrm{Reg}(T) = \overline{\mathrm{Reg}}(T) + \sum_{t=1}^{T} X_t$, this gives (29). □

Combining (28) and (29) on their common event gives

$$\mathrm{Reg}(T) \le \frac{3}{2} C_3 \sqrt{(Md_2+\varepsilon)TA_T} + \frac{3}{2} C_4 BMK_T\ell_\delta + \frac{3C_5}{2} BM\ell_\delta. \tag{30}$$

Equivalently,

$$\mathrm{Reg}(T) \le C_6 \sqrt{(Md_2+\varepsilon)TA_T} + C_7 BM(K_T+1)\ell_\delta. \tag{31}$$

B.8 Proof of Theorem 4

High-probability bookkeeping. The proof of Theorem 4 uses a fixed numerical failure split to accommodate a union bound over multiple high-probability events. Let $c_0 \ge 10$ be a numerical constant and set $\delta_0 = \delta/c_0$. Intermediate high-probability events are invoked with failure probability δ_0. Since

$$\max\left\{1, \log\frac{cT}{\delta_0}\right\} \le \max\left\{1, \log\frac{c'T}{\delta}\right\}$$

for another numerical constant $c' > 0$, increasing the numerical constant in the definition of ℓ_δ absorbs the replacement of δ by δ_0. We therefore keep writing ℓ_δ throughout this proof.

Theorem 4 (Regret upper bound, restated)**.** *Suppose Assumption 1 and Assumption 2 hold. Run Algorithm 1 with*

$$\lambda = \frac{L^2}{T}, \quad \rho = \frac{c_\rho}{\ell_\delta}, \quad \xi = \frac{3}{8\sqrt{C_1}}\sqrt{\frac{(Md_2+\varepsilon)T}{A_T}},$$

where c_ρ and C_1 are the numerical constants from Lemma 14, and A_T is as defined in (14)*. Suppose that on every low-width round with $q_t > 0$, the RRD subroutine returns $\mathbf{P}_t \in \mathcal{D}_t$ satisfying $g_t(\mathbf{P}_t) \le \varepsilon$, and that $\varepsilon \le Md_2$. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg}(T) \le C_6 \sqrt{(Md_2+\varepsilon)TA_T} + C_7 BM(K_T+1)\ell_\delta. \tag{32}$$

and consequently,

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sigma_\eta d_2\sqrt{d_1 MT} + BMd_1 d_2\right).$$

Proof. Invoke Lemma 14 with failure probability δ_0. The downstream conditional-regret bound (28) then holds on the same event. Invoke Lemma 18 with its failure parameter set to δ_0. The self-normalized quadratic noise bound is invoked inside the proof of Lemma 14 and is not union-bounded again as an independent final event. By the bookkeeping convention above, replacing δ by δ_0 only changes the numerical constant inside ℓ_δ. A union bound over the two external high-probability events gives an event of probability at least $1-\delta$.

On every low-width round with $q_t > 0$, the condition $g_t(\mathbf{P}_t) \le \varepsilon$ gives the RRD certificate (Lemma 5) with tolerance ε; the rectangular Birkhoff decomposition (Lemma 8) ensures μ_t is well defined. In the exact-RRD case, Lemma 7 guarantees an exact maximizer in $\mathcal{D}_t$ and shows that it satisfies $g_t(\mathbf{P}_t) = 0$, so the RRD certificate applies with $\varepsilon = 0$. On rounds with $q_t = 0$, the RRD certificate holds trivially.

On the high-probability event above, (31) gives

$$\mathrm{Reg}(T) \le C_6\sqrt{(Md_2+\varepsilon)TA_T} + C_7 BM(K_T+1)\ell_\delta.$$

This proves (32).

Moreover, since $B = SL$,

$$\lambda S^2 = \frac{B^2}{T}.$$

With $\lambda = L^2/T$ and $p = d_1 d_2$,

$$K_T = p\log\left(1+\frac{MTL^2}{\lambda p}\right) = d_1 d_2 \log\left(1+\frac{MT^2}{d_1 d_2}\right).$$

Therefore,

$$\begin{aligned} A_T &= \lambda S^2 + \sigma_\eta^2(K_T+\ell_\delta) \\ &= \frac{B^2}{T} + \sigma_\eta^2\left[d_1 d_2 \log\left(1+\frac{MT^2}{d_1 d_2}\right) + \ell_\delta\right]. \end{aligned}$$

Using $\varepsilon \le Md_2$, we have $(Md_2+\varepsilon) \le 2Md_2$, so $\sqrt{(Md_2+\varepsilon)TA_T} \le \sqrt{2Md_2TA_T}$. Applying $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}$ to $\sqrt{Md_2TA_T}$,

$$\sqrt{(Md_2+\varepsilon)TA_T} \le \sqrt{2}\left(B\sqrt{Md_2} + \sigma_\eta\sqrt{Md_2T\left[d_1 d_2\log\left(1+\frac{MT^2}{d_1 d_2}\right)+\ell_\delta\right]}\right).$$

The second term is

$$\widetilde{\mathcal{O}}\left(\sigma_\eta d_2\sqrt{d_1 MT}\right).$$

Also, since $M, d_1, d_2 \ge 1$,

$$B\sqrt{Md_2} \le BMd_1 d_2.$$

Finally,

$$BM(K_T+1)\ell_\delta = \widetilde{\mathcal{O}}\left(BMd_1 d_2\right).$$

This gives

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sigma_\eta d_2\sqrt{d_1 MT} + BMd_1 d_2\right).$$

□

C Regret Lower Bound

We now show that the stochastic term in the upper bound is minimax sharp up to logarithmic factors, over a rich item-context subclass. The construction below uses only deterministic contexts and independent Gaussian noise, and is therefore a subclass of the model satisfying Assumption 1 and Assumption 2.

Theorem 6 (Lower bound, restated)**.** *Suppose $d_1, d_2, M \ge 1$, $T \ge 2d_1$, and $N \ge M\,2^{d_2}$. Let*

$$T_0 = \left\lfloor \frac{T}{d_1} \right\rfloor, \qquad L = L_x L_y.$$

There exists a numerical constant $c > 0$ such that the following holds. If

$$S \ge \frac{\sigma_\eta d_2}{4L}\sqrt{\frac{d_1}{MT_0}}, \tag{33}$$

then, for every possibly randomized algorithm,

$$\sup_{\mathcal{I}} \mathbb{E}_{\mathcal{I},\mathsf{Alg}}\,\mathrm{Reg}(T) \ge c\sigma_\eta d_2\sqrt{d_1 MT}, \tag{34}$$

where the supremum is over bilinear matching semi-bandit instances satisfying Assumption 1 and Assumption 2, the context bounds, and the parameter bound. Consequently, under the same radius condition,

$$\inf_{\mathsf{Alg}}\sup_{\mathcal{I}} \mathbb{E}_{\mathcal{I},\mathsf{Alg}}\,\mathrm{Reg}(T) = \Omega\left(\sigma_\eta d_2\sqrt{d_1 MT}\right)$$

over this rich item-context subclass.

Proof. Write

$$d = d_2, \qquad \mathcal{S}_{\pm} = \{-1, +1\}^d.$$

For every sign vector $\mathbf{s} \in \mathcal{S}_{\pm}$, define an item context

$$\mathbf{y}(\mathbf{s}) = \frac{L_y}{\sqrt{d}}\mathbf{s}.$$

Then

$$\|\mathbf{y}(\mathbf{s})\|_2 = L_y.$$

The instance contains M copies of each item context $\mathbf{y}(\mathbf{s})$. Hence the first $M\,2^d$ items are indexed by pairs $(\mathbf{s}, m) \in \mathcal{S}_{\pm} \times [M]$. If $N > M\,2^d$, the remaining items are arbitrary duplicates of already constructed items. This does not change the effective set of sign vectors that can be assigned to the users, because M copies of every sign vector are already available.

For the first $d_1 T_0$ rounds, define the active row

$$g(t) = 1 + ((t-1) \bmod d_1).$$

Let $\mathbf{e}_g \in \mathbb{R}^{d_1}$ denote the g-th standard basis vector. For all users $i \in [M]$, set

$$\mathbf{x}_{t,i} = L_x \mathbf{e}_{g(t)}.$$

For the remaining rounds $t > d_1 T_0$, set $\mathbf{x}_{t,i} = \mathbf{0}$ for all i. These remaining rounds carry no information about the parameter and contribute zero regret. The item contexts are fixed as above in every round.

We now define a finite parameter family. For each sign matrix

$$\mathbf{\Omega} = (\Omega_{g,\ell})_{g \in [d_1],\, \ell \in [d]} \in \{-1, +1\}^{d_1 \times d},$$

let

$$\mathbf{\Theta}_{\mathbf{\Omega}} = \Delta \mathbf{\Omega}, \qquad \Delta = \frac{\sigma_\eta}{4L}\sqrt{\frac{d}{MT_0}}. \tag{35}$$

Then

$$\begin{aligned} \|\mathbf{\Theta}_{\mathbf{\Omega}}\|_F &= \Delta\sqrt{d_1 d} \\ &= \frac{\sigma_\eta d}{4L}\sqrt{\frac{d_1}{MT_0}} \\ &\le S \end{aligned}$$

by (33). Thus every parameter in the family satisfies the Frobenius-norm constraint.

The reward noise is independent Gaussian:

$$\eta_{t,i} \sim \mathcal{N}(0, \sigma_\eta^2),$$

independently over t and i, and independently of the algorithmic randomness. Hence Assumption 2 holds.

Consider a round $t \le d_1 T_0$, and write $g = g(t)$. If user i is matched to an item with sign vector $\mathbf{s}_{t,i} \in \mathcal{S}_{\pm}$, then the conditional mean reward under $\mathbf{\Theta}_{\mathbf{\Omega}}$ is

$$\begin{aligned} \mathbf{x}_{t,i}^\top \mathbf{\Theta}_{\mathbf{\Omega}} \mathbf{y}(\mathbf{s}_{t,i}) &= L_x \mathbf{e}_g^\top (\Delta \mathbf{\Omega}) \frac{L_y}{\sqrt{d}} \mathbf{s}_{t,i} \\ &= \frac{L\Delta}{\sqrt{d}} \left\langle \boldsymbol{\omega}_g, \mathbf{s}_{t,i} \right\rangle, \end{aligned} \tag{36}$$

where $\boldsymbol{\omega}_g \in \{-1, +1\}^d$ denotes the g-th row of $\mathbf{\Omega}$. The optimal sign vector in this round is

$$\mathbf{s}^\star = \boldsymbol{\omega}_g,$$

and M copies of this sign vector are available. Hence the per-user regret from choosing $\mathbf{s}_{t,i}$ is

$$\begin{aligned} L\Delta\sqrt{d} - \frac{L\Delta}{\sqrt{d}} \left\langle \boldsymbol{\omega}_g, \mathbf{s}_{t,i} \right\rangle &= \frac{L\Delta}{\sqrt{d}} \left(d - \sum_{\ell=1}^{d} \Omega_{g,\ell} s_{t,i,\ell} \right) \\ &= \frac{2L\Delta}{\sqrt{d}} \sum_{\ell=1}^{d} \mathbb{1}\{s_{t,i,\ell} \neq \Omega_{g,\ell}\}. \end{aligned}$$

Let $\mathrm{Reg}_{\mathbf{\Omega}}(T)$ denote the regret on the instance with parameter $\mathbf{\Theta}_{\mathbf{\Omega}}$. For every $\mathbf{\Omega}$,

$$\mathrm{Reg}_{\mathbf{\Omega}}(T) \geq \frac{2L\Delta}{\sqrt{d}} \sum_{t=1}^{d_1 T_0} \sum_{i=1}^{M} \sum_{\ell=1}^{d} \mathbb{1}\{s_{t,i,\ell} \neq \Omega_{g(t),\ell}\}. \tag{37}$$

We next lower bound the expected number of bit errors. Fix a coordinate $(g, \ell) \in [d_1] \times [d]$, and fix all other bits $\mathbf{\Omega}_{-(g,\ell)}$. Let $\mathbf{\Omega}^+$ and $\mathbf{\Omega}^-$ be the two parameters that agree on all other coordinates and satisfy

$$\Omega^+_{g,\ell} = +1, \qquad \Omega^-_{g,\ell} = -1.$$

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the laws of the full interaction history under $\mathbf{\Theta}_{\mathbf{\Omega}^+}$ and $\mathbf{\Theta}_{\mathbf{\Omega}^-}$, respectively. These laws include the reward noise and the algorithm's internal randomization.

The two instances differ only on rounds with active row g. In such a round, if user i is assigned sign vector $\mathbf{s}_{t,i}$, then by (36),

$$\begin{aligned} \mu_{+,t,i} - \mu_{-,t,i} &= \frac{L\Delta}{\sqrt{d}} \left(\left\langle \boldsymbol{\omega}_g^+, \mathbf{s}_{t,i} \right\rangle - \left\langle \boldsymbol{\omega}_g^-, \mathbf{s}_{t,i} \right\rangle \right) \\ &= \frac{2L\Delta}{\sqrt{d}} s_{t,i,\ell}. \end{aligned}$$

Thus

$$(\mu_{+,t,i} - \mu_{-,t,i})^2 = \frac{4L^2\Delta^2}{d}.$$

By the chain rule for KL divergence, the algorithmic action kernels contribute zero KL because they are the same under the two instances; only the conditional Gaussian reward kernels contribute. Using the Gaussian KL formula,

$$\mathrm{KL}(\mathbb{P}_+ \,\|\, \mathbb{P}_-) = \mathbb{E}_+ \sum_{\substack{t \leq d_1 T_0 \\ g(t)=g}} \sum_{i=1}^{M} \frac{(\mu_{+,t,i} - \mu_{-,t,i})^2}{2\sigma_\eta^2}.$$

Since $(\mu_{+,t,i} - \mu_{-,t,i})^2 = 4L^2\Delta^2/d$ is deterministic and independent of the algorithm's decisions, the expectation $\mathbb{E}_+$ evaluates to a simple count, and we obtain

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_+ \,\|\, \mathbb{P}_-) &= MT_0 \cdot \frac{1}{2\sigma_\eta^2} \cdot \frac{4L^2\Delta^2}{d} \\ &= \frac{2MT_0L^2\Delta^2}{d\sigma_\eta^2}. \end{aligned} \tag{38}$$

This standard adaptive KL decomposition follows from the KL chain rule for sequential experiments [8, 26]. Using (35), (38) becomes

$$\mathrm{KL}(\mathbb{P}_+ \,\|\, \mathbb{P}_-) = \frac{1}{8}.$$

Pinsker's inequality gives

$$\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \leq \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_+ \,\|\, \mathbb{P}_-)} \leq \frac{1}{4}. \tag{39}$$

Now fix a round $t \leq d_1 T_0$ with $g(t) = g$, and fix a user $i \in [M]$. The decision $s_{t,i,\ell} \in \{-1, +1\}$ is determined after the matching π_t is selected and before the reward $r_{t,i}$ is observed; in particular, it is $\mathcal{G}_t$-measurable. Hence, by data processing for total variation, the total variation distance between the induced laws of $s_{t,i,\ell}$ under $\mathbb{P}_+$ and $\mathbb{P}_-$ is at most $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)$. Therefore, using (39),

$$\begin{aligned} \frac{1}{2}\mathbb{P}_+(s_{t,i,\ell} = -1) + \frac{1}{2}\mathbb{P}_-(s_{t,i,\ell} = +1) &\geq \frac{1}{2}\left(1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)\right) \\ &\geq \frac{3}{8}. \end{aligned} \tag{40}$$

Here the data-processing and testing step is a standard consequence of total variation contraction [8, 40]. Averaging (40) over the fixed values of $\mathbf{\Omega}_{-(g,\ell)}$ yields

$$\frac{1}{2^{d_1 d}} \sum_{\mathbf{\Omega} \in \{-1,+1\}^{d_1 \times d}} \mathbb{E}_{\mathbf{\Omega}} \mathbb{1}\{s_{t,i,\ell} \neq \Omega_{g,\ell}\} \geq \frac{3}{8}, \tag{41}$$

for every g, ℓ, every round $t \leq d_1 T_0$ with $g(t) = g$, and every user $i \in [M]$.

Combining (37) and (41), the Bayes average regret under the uniform prior on $\mathbf{\Omega}$ satisfies

$$\begin{aligned}
\frac{1}{2^{d_1 d}} \sum_{\mathbf{\Omega}} \mathbb{E}_{\mathbf{\Omega}} \operatorname{Reg}_{\mathbf{\Omega}}(T) &\geq \frac{2L\Delta}{\sqrt{d}} \sum_{g=1}^{d_1} \sum_{\ell=1}^{d} \sum_{\substack{t \leq d_1 T_0 \\ g(t)=g}} \sum_{i=1}^{M} \frac{3}{8} \\
&= \frac{2L\Delta}{\sqrt{d}} \cdot d_1 d \cdot M T_0 \cdot \frac{3}{8} \\
&= \frac{3}{4} L \Delta d_1 M T_0 \sqrt{d}.
\end{aligned}$$

Substituting (35),

$$\begin{aligned}
\frac{1}{2^{d_1 d}} \sum_{\mathbf{\Omega}} \mathbb{E}_{\mathbf{\Omega}} \operatorname{Reg}_{\mathbf{\Omega}}(T) &\geq \frac{3}{4} L \left(\frac{\sigma_\eta}{4L} \sqrt{\frac{d}{M T_0}} \right) d_1 M T_0 \sqrt{d} \\
&= \frac{3}{16} \sigma_\eta d_1 d \sqrt{M T_0}.
\end{aligned} \tag{42}$$

Since $T \geq 2d_1$,

$$T_0 = \left\lfloor \frac{T}{d_1} \right\rfloor \geq \frac{T}{2d_1}.$$

Therefore (42) implies

$$\frac{1}{2^{d_1 d}} \sum_{\mathbf{\Omega}} \mathbb{E}_{\mathbf{\Omega}} \operatorname{Reg}_{\mathbf{\Omega}}(T) \geq \frac{3}{16\sqrt{2}} \sigma_\eta d \sqrt{d_1 M T}. \tag{43}$$

Since the supremum over instances is at least the average over this finite family, every algorithm has an instance in the constructed subclass such that

$$\mathbb{E}_{\mathcal{I}, \mathsf{Alg}} \operatorname{Reg}(T) \geq c \sigma_\eta d \sqrt{d_1 M T} = c \sigma_\eta d_2 \sqrt{d_1 M T}$$

for a numerical constant $c > 0$. This proves (34). □

D On the Within-Round Noise Condition

We record a quantitative comparison illustrating why a marginal post-action sub-Gaussian condition alone is insufficient for the ℓ_2-scale vector bound (1). Suppose, in place of Assumption 2, only the marginal post-action condition holds: for every $\lambda \in \mathbb{R}$ and every $i \in [M]$,

$$\mathbb{E}\left[\exp(\lambda \eta_{t,i}) \mid \mathcal{G}_t\right] \leq \exp\left(\frac{\lambda^2 \sigma_\eta^2}{2}\right). \tag{44}$$

Let $\boldsymbol{\eta}_t = (\eta_{t,1}, \ldots, \eta_{t,M})$. As in the proof of Lemma 12, the deterministic-λ statement in (44) extends to every finite $\mathcal{G}_t$-measurable scalar λ by simple approximation and conditional Fatou. We use this extension below. The marginal-only condition (44) allows arbitrary within-round dependence among $\eta_{t,1}, \ldots, \eta_{t,M}$. It does not imply (1) in general. The implication does hold under conditional independence given $\mathcal{G}_t$, and more generally if one directly assumes the joint conditional vector sub-Gaussian bound (1). From (44) alone, for every finite $\mathcal{G}_t$-measurable vector $\mathbf{u} \in \mathbb{R}^M$, conditional Hölder's inequality gives the uniform implication

$$\begin{aligned}
\mathbb{E}\left[\exp(\mathbf{u}^\top \boldsymbol{\eta}_t) \mid \mathcal{G}_t\right] &\leq \prod_{i=1}^{M} \mathbb{E}\left[\exp(M u_i \eta_{t,i}) \mid \mathcal{G}_t\right]^{1/M} \\
&\leq \exp\left(\frac{M \sigma_\eta^2 \|\mathbf{u}\|_2^2}{2}\right).
\end{aligned} \tag{45}$$

The factor M in (45) is not merely an artifact of Hölder's inequality. Let $\vartheta_t \mid \mathcal{G}_t \sim \mathcal{N}(0, \sigma_\eta^2)$ be a shared noise component, or common shock, and set

$$\eta_{t,1} = \cdots = \eta_{t,M} = \vartheta_t.$$

Then every coordinate satisfies the marginal-only condition (44). However, for $\mathbf{u} = \mathbf{1}/\sqrt{M}$,

$$\begin{aligned}\mathbb{E}\left[\exp(\mathbf{u}^\top \boldsymbol{\eta}_t) \mid \mathcal{G}_t\right] &= \mathbb{E}\left[\exp(\sqrt{M}\vartheta_t) \mid \mathcal{G}_t\right] \\ &= \exp\left(\frac{M\sigma_\eta^2}{2}\right).\end{aligned}$$

By contrast, (1) would require the upper bound $\exp(\sigma_\eta^2/2)$ for the same vector $\mathbf{u}$. Thus, whenever $M \geq 2$, (1) fails. When $M \geq 2$, this common-shock example also violates Assumption 2: after observing $\eta_{t,1}$, the conditional mean of $\eta_{t,2}$ is $\eta_{t,1}$ rather than zero. Thus the variance proxy can be inflated by a factor of M relative to (1). If only (45) is available, then in Lemma 13 one must replace σ_η^2 by $M\sigma_\eta^2$. Consequently, the noise term in A_T is replaced by

$$M\sigma_\eta^2(K_T + \ell_\delta)$$

instead of

$$\sigma_\eta^2(K_T + \ell_\delta).$$

Propagating this change through the low-width prediction-error bound and the regret decomposition gives a stochastic term of order

$$\widetilde{\mathcal{O}}\left(\sigma_\eta M d_2 \sqrt{d_1 T}\right)$$

rather than

$$\widetilde{\mathcal{O}}\left(\sigma_\eta d_2 \sqrt{d_1 M T}\right).$$

Thus any analysis relying only on the flattened vector self-normalized bound available under the marginal-only condition loses the $\sqrt{M}$ semi-bandit gain, consistently with the role of within-action dependence structure in combinatorial semi-bandits [9].

The preceding common-shock example only shows that (1) can fail. The next proposition shows that this failure is not merely an artifact of our concentration argument: under the marginal-only condition and arbitrary same-round dependence, a uniform $\sqrt{M}$ semi-bandit gain is impossible even in the scalar bilinear matching subclass, under the simple richness and radius conditions stated below.

Proposition 19 (Algorithm-independent marginal-only lower bound)**.** *Suppose $M \geq 2$, $T \geq 1$, and $N \geq 2M$. Let $d_1 = d_2 = 1$ and $L = L_x L_y$. Define*

$$\Delta := \frac{\sigma_\eta}{4L\sqrt{T}}. \tag{46}$$

Suppose the parameter radius satisfies

$$S \geq \Delta. \tag{47}$$

There exists a deterministic-context bilinear matching instance satisfying the bounded-context and bounded-parameter conditions, and a same-round noise model satisfying the marginal-only condition (44)*, such that for every possibly randomized algorithm,*

$$\sup_{\Theta^\star \in \{+\Delta, -\Delta\}} \mathbb{E}_{\Theta^\star, \mathsf{Alg}} \operatorname{Reg}(T) \geq c\, \sigma_\eta M \sqrt{T},$$

for a numerical constant $c > 0$. Consequently, under the same richness and radius conditions, a uniform $\widetilde{\mathcal{O}}(\sigma_\eta \sqrt{MT})$ regret guarantee is impossible under the marginal-only condition with arbitrary same-round dependence.

Proof. By (47), both scalar parameters $\Theta^\star = +\Delta$ and $\Theta^\star = -\Delta$ are feasible. Set

$$x_{t,i} = L_x \qquad \text{for all } t \in [T],\ i \in [M].$$

The item set contains M copies of the positive item context $+L_y$ and M copies of the negative item context $-L_y$. If $N > 2M$, the remaining items are arbitrary duplicates of the two constructed

contexts, say additional copies of $+L_y$. This does not affect the argument, because M copies of each sign are already available. For a selected item, write

$$s_{t,i} := \frac{y_{t,\pi_t(i)}}{L_y} \in \{-1, +1\}.$$

This construction is used only as a counterexample under the marginal-only post-action condition (44); it is not intended to satisfy the within-round martingale condition in Assumption 2.

For each round t, draw a common shock

$$\vartheta_t \sim \mathcal{N}(0, \sigma_\eta^2),$$

independently across rounds and independently of the learner's history and randomization. For every item with sign $s \in \{-1, +1\}$, define its potential noise as $s\vartheta_t$. Hence, after the matching π_t is selected, the observed noise coordinates satisfy

$$\eta_{t,i} = s_{t,i}\vartheta_t, \qquad i \in [M].$$

Conditionally on $\mathcal{G}_t$, each selected sign $s_{t,i}$ is fixed and ϑ_t remains independent Gaussian; hence $\eta_{t,i} \mid \mathcal{G}_t \sim \mathcal{N}(0, \sigma_\eta^2)$. Therefore the constructed noise model satisfies the marginal post-action condition: for every $i \in [M]$,

$$\mathbb{E}[\eta_{t,i} \mid \mathcal{G}_t] = 0,$$

and for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\lambda\eta_{t,i}) \mid \mathcal{G}_t\right] = \exp\left(\frac{\lambda^2\sigma_\eta^2}{2}\right).$$

Thus the marginal-only condition (44) holds. However, when $M \geq 2$, this noise model does not satisfy Assumption 2. Indeed, since

$$\mathcal{H}_{t,1} = \mathcal{G}_t \vee \sigma(\eta_{t,1}),$$

and

$$\eta_{t,2} = s_{t,2}\vartheta_t = s_{t,2}s_{t,1}\eta_{t,1},$$

we have

$$\mathbb{E}[\eta_{t,2} \mid \mathcal{H}_{t,1}] = s_{t,2}s_{t,1}\eta_{t,1},$$

which is nonzero with positive probability. Thus the sequential martingale-difference condition in Assumption 2 fails. The observed reward is

$$\begin{aligned} r_{t,i} &= x_{t,i}\Theta^\star y_{t,\pi_t(i)} + \eta_{t,i} \\ &= s_{t,i}\left(L\Theta^\star + \vartheta_t\right). \end{aligned}$$

Thus, after multiplying by the known sign $s_{t,i}$, every selected reward in the same round reveals the same scalar observation

$$Y_t := s_{t,i}r_{t,i} = L\Theta^\star + \vartheta_t.$$

Consequently, the full semi-bandit feedback in a round carries no more information than one scalar Gaussian observation Y_t.

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the laws of the full interaction history under $\Theta^\star = +\Delta$ and $\Theta^\star = -\Delta$, respectively. Conditionally on the selected sign vector $(s_{t,1}, \ldots, s_{t,M})$, the full reward vector is the deterministic injective image of the scalar Gaussian observation

$$Y_t = L\Theta^\star + \vartheta_t$$

under the map

$$y \mapsto (s_{t,1}y, \ldots, s_{t,M}y).$$

Hence the conditional KL contribution of one round is exactly the scalar Gaussian KL between $\mathcal{N}(+L\Delta, \sigma_\eta^2)$ and $\mathcal{N}(-L\Delta, \sigma_\eta^2)$. The adaptive action kernels contribute zero KL because the algorithm uses the same decision rule under the two hypotheses. Therefore the adaptive KL chain rule gives

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-) &= T \cdot \frac{(2L\Delta)^2}{2\sigma_\eta^2} \\ &= \frac{2TL^2\Delta^2}{\sigma_\eta^2} \\ &= \frac{1}{8}, \end{aligned}$$

where the last equality uses (46). Pinsker's inequality implies

$$\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \le \frac{1}{4}. \tag{48}$$

Let J_t be the number of positive items selected in round t. Under $\Theta^\star = +\Delta$, the optimal matching selects all M positive items, and the one-step regret is

$$r_t^+ = 2L\Delta(M - J_t).$$

Under $\Theta^\star = -\Delta$, the optimal matching selects all M negative items, and the one-step regret is

$$r_t^- = 2L\Delta J_t.$$

Let $\mathbb{P}^0_{+,t}$ and $\mathbb{P}^0_{-,t}$ be the laws of the information available after the round-t action is chosen and before the reward is revealed. Since $J_t \in [0, M]$ is measurable with respect to this post-action, pre-reward information,

$$|\mathbb{E}_+ J_t - \mathbb{E}_- J_t| \le M\,\mathrm{TV}(\mathbb{P}^0_{+,t}, \mathbb{P}^0_{-,t}) \le M\,\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-).$$

Therefore, by (48),

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_+ r_t^+ + \frac{1}{2}\mathbb{E}_- r_t^- &= L\Delta\left(M - \mathbb{E}_+ J_t + \mathbb{E}_- J_t\right) \\
&\ge L\Delta M\left(1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)\right) \\
&\ge \frac{3}{4} L\Delta M.
\end{aligned} \tag{49}$$

Summing (49) over $t = 1, \ldots, T$ yields

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_+ \mathrm{Reg}(T) + \frac{1}{2}\mathbb{E}_- \mathrm{Reg}(T) &\ge \frac{3}{4} L\Delta MT \\
&= \frac{3}{16}\sigma_\eta M\sqrt{T}.
\end{aligned}$$

Therefore at least one of the two instances satisfies

$$\mathbb{E}_{\Theta^\star, \mathsf{Alg}}\,\mathrm{Reg}(T) \ge c\,\sigma_\eta M\sqrt{T}$$

for a numerical constant $c > 0$. This proves the proposition. □

Remark 20 (Scope of the marginal-only obstruction)**.** Proposition 19 is intentionally stated for the scalar subclass $d_1 = d_2 = 1$. This already rules out any uniform $\sqrt{M}$ semi-bandit gain under the marginal-only condition with arbitrary same-round dependence, because such a uniform guarantee would have to hold on every subclass. We do not claim a dimension-sharp marginal-only lower bound of order $\Omega(\sigma_\eta d_2 M\sqrt{d_1 T})$; establishing such a statement would require a separate high-dimensional construction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the problem setting, the proposed HybridRRD algorithm, the high-probability regret upper bound, the matching lower bound up to logarithmic factors, and the empirical scope. The assumptions and scope of the results are stated in Sections 3 and 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a Limitations section discussing the computational cost of the proposed algorithm, the scope of the theoretical guarantees, and the controlled nature of the empirical evaluation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The model and assumptions are stated in Section 3, and the main upper and lower bounds are stated in Section 5. Complete proofs of the RRD certificate, the upper-bound analysis, the lower bound, and the within-round noise comparison are provided in Appendices A–D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 6 describes the experimental setup needed to reproduce the main results, including the data generation process, compared baselines, and evaluation metric.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized source code in the supplemental material, including scripts, configuration files, and instructions for reproducing the reported experiments. The experiments use generated synthetic data, so no external dataset access is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 gives the main experimental protocol needed to understand the results, and the supplemental material includes the configuration files used for the reported runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 6 reports cumulative regret averaged over five independent runs. The shaded regions in the regret plots denote one standard deviation across runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 6 reports the compute resources used for the experiments, including the CPU model, number of cores, and memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on online learning algorithms, and the experiments use synthetic data. We do not use private data, human-subject data, scraped data, or deployed decision systems.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper notes that the work is theoretical and briefly discusses possible risks when matching-bandit methods are used with biased or sensitive real-world contexts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pretrained models, generative models, scraped datasets, or other assets with a high risk of misuse. The released materials, if any, consist only of algorithmic implementation code and synthetic experiment scripts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing datasets, pretrained models, or third-party code assets as part of the reported results. Baseline algorithms are reimplemented from the cited literature, and the experiments use synthetic data generated by our scripts.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The supplemental code is documented with a README, configuration files, and scripts for reproducing the experimental results.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects, crowdsourcing, or personally identifiable data, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development does not involve LLMs as an important, original, or non-standard algorithmic, experimental, or methodological component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.