Pooled Exploration Closes the Distributed Linear-Bandit Gap

Pahan Dewasurendra
Johns Hopkins University

Abstract

Distributed adversarial linear bandits have two sources of difficulty. A network needs time to mix information, while bandit feedback needs enough collective exploration to identify a d-dimensional loss. The best known finite-action guarantee multiplies both costs by d. Its lower bound only multiplies the bandit cost by d, leaving a factor $\sqrt{d}$ gap in the communication term.

We close this gap. The key is that exploration should stabilize the *pooled* block estimate, not every local estimate separately. An exponential-potential argument with conditional Bernstein control allows a spanner exploration rate of order $\eta d/N$ across N agents. The previous pointwise analysis requires order $\eta B d$, where B is the communication block length. A separate leverage lemma shows that the tiny uniform floor already present in the algorithm keeps every communicated spanner coordinate polynomially bounded, so reducing exploration does not weaken gossip.

For K actions spanning dimension $r \leq d$, spectral gap ρ, and horizon T, our algorithm communicates $O(r)$ scalars per agent and round and has per-agent regret

$$\widetilde{O}\left(\sqrt{(\rho^{-1/2} + r/N)T \log K}\right).$$

This matches the known lower bound up to logarithmic factors and resolves the finite-action linear-bandit gap. The proof also gives a general mini-batch EXP2 theorem showing that independent agents pool both variance and the exploration needed for exponential stability.

1 Introduction

In a distributed adversarial bandit, N agents face heterogeneous local losses and communicate only with neighbors. Each agent is judged against the best action for the global average loss. This model combines partial feedback with a network bottleneck and has become a basic test case for decentralized online learning (Cesa-Bianchi et al., 2019; Yi and Vojnović, 2023).

For K unrelated arms, Qiu et al. (2026) recently identified the minimax rate

$$\widetilde{\Theta}\left(\sqrt{(\rho^{-1/2} + K/N)T}\right),$$

where ρ is the spectral gap of the gossip matrix. Their reduction holds a policy fixed for a block of length $B = \Theta(\rho^{-1/2} \log(KNT))$. While agents play the block, accelerated gossip mixes the previous block's loss estimates. Blocking converts the network problem into online learning with one-block-delayed feedback.

The linear case is subtler. Suppose the K actions are vectors $a_1, \ldots, a_K \in \mathbb{R}^d$ and losses are linear. A volumetric spanner lets each agent communicate only $O(d)$ reconstructed losses rather than all K coordinates (Hazan and Karnin, 2016; Bhaskara et al., 2023). Qiu et al. obtain

$$\widetilde{O}\left(\sqrt{d(\rho^{-1/2} + 1/N)T \log K}\right). \tag{1}$$

Their lower bound is

$$\Omega\left(\sqrt{(\rho^{-1/2} + d/N)T \log K}\right). \tag{2}$$

Thus (1) has an extra factor $\sqrt{d}$ exactly in the communication term. Closing this gap is posed as an open problem.

The loss is not caused by the estimator's second moment. The published proof already obtains the sharp blockwise quantity

$$B^2 + O(Bd/N). \tag{3}$$

Instead, it controls every reconstructed loss pointwise before applying the standard EXP2 inequality. If β is the mass assigned to a size-$O(d)$ spanner and η is the learning rate, this requires $\beta = \Omega(\eta B d)$. Exploration then costs $\Omega(\eta BdT)$ and inserts d into the first term of (1).

Our starting observation is that the exponential update receives a global average of NB conditionally independent estimates. It does not receive one local estimate. Conditional Bernstein control therefore needs only

$$\beta = \Theta(\eta d/N). \tag{4}$$

This exploration costs $O(\eta dT/N)$, exactly the scale already present in (3). The factor B disappears because Bernstein's range condition is applied to each summand after the $1/N$ averaging factor. The factor N appears for the same reason.

There is a second obstacle. Gossip convergence is deterministic and still needs bounded local messages. We show that the $1/T$ uniform action mixture already used by Qiu et al. implies

$$\left|\langle a_k, \widehat{\theta}_t(i)\rangle\right| \le KT \quad \text{for all } i, t, k. \tag{5}$$

This bound does not use β. It follows from a leverage comparison with the unweighted action covariance. Hence (4) preserves the original inverse-polynomial consensus error and $O(d)$ communication.

Contributions. Our results are entirely theoretical.

- We prove a pooled EXP2 lemma for heterogeneous mini-batches. It replaces pointwise exponential stability by a conditional log-moment bound. With N independent agents, both estimator variance and required exploration improve by N.
- We prove the uniform-floor leverage bound (5). It decouples message boundedness from the spanner exploration used for statistical stability.
- Combining these facts with the blocking and accelerated-gossip reduction of Qiu et al. (2026) gives regret

  $$O\Big(\sqrt{\log K\,(TB + rT/N)} + (B + r/N)\log K\Big)$$

  for effective dimension r. This matches (2) up to logarithmic factors and uses $O(r)$ communicated scalars per agent and round.

Related work. Classical EXP2 mixes exponential weights with a geometric exploration design. Its analysis couples the exploration mass to a worst-case reconstructed loss (Bubeck et al., 2012). Delayed adversarial linear bandits can instead use distribution truncation or barrier stability (Ito et al., 2020; van der Hoeven et al., 2023). Those methods address temporal delay for one loss stream. They do not exploit averaging across heterogeneous agents and their linear-bandit delay terms retain dimension dependence. Earlier decentralized adversarial work treats unrelated arms (Yi and Vojnović, 2023). Accelerated-gossip blocking was developed for decentralized online convex optimization (Wan et al., 2024). The recent black-box reduction of Qiu et al. (2026) is the first result that combines finite-action adversarial linear bandits, heterogeneous local losses, gossip-only communication, and $O(d)$-dimensional messages. Our theorem changes its exploration and potential analysis while retaining its network protocol.

Table 1: Per-agent adversarial regret under gossip communication.

Result	Regret up to logs	Message
Qiu et al., K arms	$\sqrt{(B+K/N)T}$	K
Qiu et al., linear	$\sqrt{r(B+1/N)T}$	r
This paper, linear	$\sqrt{(B+r/N)T}$	r
Known lower bound	$\sqrt{(B+r/N)T}$	any

Table 1 isolates the gap. We suppress logarithmic factors in K, N, and T, and write $B \asymp \rho^{-1/2}$ for the communication cost. The action-dependent term improves with the number of agents, while the network term cannot improve with N in the heterogeneous model.

2 Problem and Algorithm

There are N agents connected by an undirected graph. Communication uses a symmetric doubly stochastic matrix W supported on the graph. Write $\sigma_2(W)$ for its second singular value and $\rho = 1 - \sigma_2(W) > 0$.

The common finite action set is $\Omega = \{a_1, \ldots, a_K\} \subset \mathbb{R}^d$. Let

$$r = \dim \operatorname{span}(\Omega).$$

All inverses below are restricted to this span. This is equivalent to representing the actions in $\mathbb{R}^r$. An oblivious adversary chooses $\theta_t(i) \in \mathbb{R}^r$ such that

$$\ell_t(i,k) = \langle a_k, \theta_t(i)\rangle \in [-1, 1]$$

for every agent, round, and action. Define the global average loss

$$\bar{\ell}_t(k) = \frac{1}{N}\sum_{i=1}^{N} \ell_t(i,k).$$

Agent i samples $A_t(i)$ from $p_t(i) \in \Delta_K$ and observes only $\ell_t(i, A_t(i))$. Its pseudo-regret is

$$\operatorname{Reg}_T(i) = \max_{k\in[K]} \mathbb{E}\sum_{t=1}^{T} \langle p_t(i) - e_k, \bar{\ell}_t\rangle.$$

We bound $\mathrm{Reg}_T = \max_i \mathrm{Reg}_T(i)$.

Let $S = \{b_1, \ldots, b_s\} \subseteq \Omega$ be an ℓ_2 volumetric spanner. Thus $s \le 3r$ and every action has a representation

$$a_k = \sum_{j=1}^{s} \lambda_j^{(k)} b_j, \qquad \|\lambda^{(k)}\|_2 \le 1. \tag{6}$$

Such a spanner is efficiently computable (Bhaskara et al., 2023).

Our algorithm uses the reduction and notation of Qiu et al. (2026). Divide time into blocks $\mathcal{T}_\tau$ of length B, where

$$B = \Theta\Big(\rho^{-1/2} \log(KNT)\Big) \tag{7}$$

is large enough for accelerated gossip (Liu and Morse, 2011) to have contraction at most $(KT)^{-6} N^{-1/2}$. Two entropy-FTRL instances alternate between odd and even blocks, removing the one-block feedback delay. Let $q_\tau(i)$ be the base policy produced for agent i. The sampling policy is

$$p_\tau(i) = (1-\alpha-\beta) q_\tau(i) + \alpha u_K + \beta u_S, \qquad \alpha = T^{-1}, \tag{8}$$

where u_K is uniform on all actions and u_S is uniform on the spanner. The change from prior work is

$$\beta = 32\eta s/N, \tag{9}$$

instead of $\beta = \Theta(\eta B r)$.

Within a block, define

$$M_\tau(i) = \sum_{k=1}^{K} p_\tau(i,k) a_k a_k^\top, \tag{10}$$

$$\widehat{\theta}_t(i) = M_\tau(i)^{-1} a_{A_t(i)} \ell_t(i, A_t(i)). \tag{11}$$

Agent i accumulates only $(\langle b_j, \widehat{\theta}_t(i)\rangle)_{j=1}^{s}$ and performs one accelerated gossip step per round on the previous block's s-vector. At block end, the local FTRL instance reconstructs all K coordinates through (6). Thus communication is $s \le 3r$ scalars per agent and round.

Block protocol. For clarity, one block performs four operations. First, agent i queries the appropriate parity instance of entropy FTRL and forms (8). Second, it keeps this sampling policy fixed for B rounds and accumulates the s spanner coordinates of (11). Third, in parallel with play, it takes B accelerated-gossip steps on the preceding block's accumulated vector. Fourth, it reconstructs all action losses from the mixed vector and feeds them to the FTRL instance that next acts on the same parity. Each block's feedback is therefore available before that instance is queried again. This is the standard two-copy reduction from a one-block delay (Joulani et al., 2013; Qiu et al., 2026).

The protocol only needs the spectral gap and the common spanner as shared preprocessing information. It does not transmit action identities, local loss coefficients, or K-dimensional vectors. The estimator is unbiased under heterogeneous losses because each local covariance $M_\tau(i)$ matches the agent's own sampling policy.

3 Pooled Exponential Stability

We first isolate the online-learning fact that removes the dimension loss. Fix one block and condition on all prior randomness. Let $q \in \Delta_K$ be a ghost entropy-FTRL policy and let p_i be the sampling policy of agent i. For b conditionally independent samples per agent, define

$$X_{i,t,k} = \langle a_k, M_i^{-1} a_{A_t(i)} \rangle \ell_t(i, A_t(i)), \tag{12}$$

$$Z_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{b} X_{i,t,k}, \qquad L_k = \mathbb{E}[Z_k \mid \mathcal{F}]. \tag{13}$$

Here $M_i = \sum_k p_i(k) a_k a_k^\top$. Unbiasedness gives $L_k = N^{-1} \sum_{i,t} \ell_t(i,k)$, so $|L_k| \le b$.

Lemma 1 (Pooled log-moment bound)**.** *Suppose $q(k)/p_i(k) \le C_0$ for all i,k, and $M_i \succeq (\beta/s) \sum_{j=1}^{s} b_j b_j^\top$. If*

$$0 < \beta \le 1, \qquad \eta b \le \frac{1}{8}, \qquad \beta \ge 32\eta s/N,$$

then

$$\mathbb{E}\left[\log \sum_{k=1}^{K} q(k) e^{-\eta Z_k} \,\middle|\, \mathcal{F}\right] \le -\eta \langle q, L\rangle + 3\eta^2 \left(b^2 + C_0 b r/N\right). \tag{14}$$

The distinction from standard EXP2 is central. A pointwise argument requires $\eta |Z_k| = O(1)$ and therefore $\beta = \Omega(\eta b r)$. Lemma 1 requires only that each centered summand in Z_k has small exponential scale. Its $1/N$ coefficient gives (9).

Proof. The spanner property gives $a_k^T M_i^{-1} a_k \le s/\beta$. Hence

$$|X_{i,t,k}| \le s/\beta,$$

$$\left| \frac{X_{i,t,k} - \mathbb{E}[X_{i,t,k} \mid \mathcal{F}]}{N} \right| \le \frac{2s}{\beta N} \le \frac{1}{16\eta}. \tag{15}$$

Let $v_k = \mathrm{Var}(Z_k \mid \mathcal{F})$. The standard covariance-estimator calculation gives

$$v_k \le \frac{1}{N^2} \sum_{i,t} a_k^T M_i^{-1} a_k. \tag{16}$$

In particular, $v_k \le bs/(\beta N) \le b/(32\eta)$. Also, with $M_q = \sum_k q(k) a_k a_k^\top$,

$$\sum_k q(k) v_k \le \frac{1}{N^2} \sum_{i,t} \mathrm{tr}(M_i^{-1} M_q) \\ \le C_0 br/N. \tag{17}$$

The last step uses $M_q \preceq C_0 M_i$.

Conditional Bernstein and (15) imply

$$\mathbb{E}[e^{-\eta Z_k} \mid \mathcal{F}] \le e^{-\eta L_k + \eta^2 v_k}.$$

Conditional Jensen moves the expectation outside the logarithm. Since $|\eta L_k| \le 1/8$ and $\eta^2 v_k \le 1/256$, the elementary bound $e^x \le 1 + x + x^2$ for $|x| \le 1/4$ yields

$$\mathbb{E}\left[\log \sum_k q(k) e^{-\eta Z_k} \mid \mathcal{F}\right] \\ \le \log \sum_k q(k) e^{-\eta L_k + \eta^2 v_k} \\ \le -\eta \langle q, L\rangle + 3\eta^2 \left(b^2 + \sum_k q(k) v_k\right).$$

Apply (17). Full scalar details are in the supplement. $\square$

Telescoping the exponential potential gives the following mini-batch result.

Theorem 2 (Pooled EXP2)**.** *Consider m meta-rounds, each containing b independent samples at each of N agents. Let entropy FTRL update q_j with the pooled estimates Z_j. Suppose the conditions of Lemma 1 hold at every meta-round. Then for every action k,*

$$\mathbb{E} \sum_{j=1}^{m} \langle q_j - e_k, Z_j\rangle \le \frac{\log K}{\eta} + 3\eta m \left(b^2 + C_0 br/N\right). \tag{18}$$

Proof. Initialize weights $w_1(k) = 1$ and update $w_{j+1}(k) = w_j(k) e^{-\eta Z_j(k)}$. Their normalized values are q_j. For every fixed comparator k,

$$-\eta \sum_{j=1}^{m} Z_j(k) - \log K \le \log \frac{\sum_u w_{m+1}(u)}{\sum_u w_1(u)} \\ = \sum_{j=1}^{m} \log \sum_u q_j(u) e^{-\eta Z_j(u)}.$$

Condition on the pre-round filtration and apply Lemma 1 to each summand. Since q_j is measurable before Z_j is generated, $\mathbb{E}\langle q_j, Z_j\rangle = \mathbb{E}\langle q_j, L_j\rangle$. Rearranging proves (18). $\square$

For $N = 1$ and $b = 1$, the exploration scale in this theorem is the classical $\Theta(\eta r)$. For general N, both its variance and its required exploration are divided by N. Increasing the batch length b increases the mean-square term but does not increase the exploration needed for exponential stability.

The same argument separates pooling into two norms for nonuniform objectives.

Corollary 3 (Weighted pooling)**.** *Let $w \in \Delta_N$ and replace (13) by $Z_k = \sum_i w_i \sum_{t=1}^{b} X_{i,t,k}$. If*

$$\beta \ge 32\eta s \|w\|_\infty, \tag{19}$$

then the conclusion of Theorem 2 holds with its variance term $C_0 br/N$ replaced by $C_0 br\|w\|_2^2$.

Thus the effective number of samples is $\|w\|_2^{-2}$, while exponential stability is controlled by the largest aggregation weight. These quantities coincide at N for uniform averaging but can differ under personalized or importance-weighted network objectives. The supplement gives the weighted calculation.

4 Uniform Exploration Is Enough for Gossip

Reducing β appears to make the local coefficient estimator large. The next lemma shows that the uniform floor α already provides the polynomial bound needed by accelerated gossip.

Lemma 4 (Uniform-floor leverage)**.** *Let $\Sigma = \sum_{k=1}^{K} a_k a_k^\top$ on the action span. If $p(k) \ge \alpha/K$ and $M = \sum_k p(k) a_k a_k^\top$, then*

$$a_k^T M^{-1} a_k \le K/\alpha \quad \text{for every } k. \tag{20}$$

Consequently, when $\alpha = 1/T$ and $|\ell(A)| \le 1$,

$$|\langle a_k, M^{-1} a_A \ell(A)\rangle| \le KT \quad \text{for all } k, A. \tag{21}$$

Proof. We have $M \succeq (\alpha/K)\Sigma$. Also $a_k a_k^\top \preceq \Sigma$, which implies $a_k^T \Sigma^{-1} a_k \le 1$. This proves (20). Equation (21) follows by Cauchy–Schwarz in the M^{-1} norm. $\square$

A block's s-dimensional spanner message therefore has norm at most $B\sqrt{s}KT$. Substitution into the accelerated-gossip contraction used in Qiu et al. (2026) gives, for the exact global spanner estimate $\bar{Z}_\tau^S$ and every local output $Z_\tau^B(i)$,

$$\|Z_\tau^B(i) - \bar{Z}_\tau^S\|_2 \le \delta_S := \frac{2B\sqrt{s}}{T^5 K^5}. \tag{22}$$

Reconstruction through (6) does not enlarge any coordinate error and enlarges the full K-vector norm by at most $\sqrt{K}$.

Let q_τ be the ghost entropy-FTRL policy updated with exact global estimates, and let $q_\tau(i)$ be agent i's policy updated with its gossip approximations. Strong convexity of negative entropy and (22) give

$$\|q_\tau - q_\tau(i)\|_2 \le \eta(T/B)\sqrt{K}\,\delta_S \le \frac{1}{KT}. \tag{23}$$

For $\alpha+\beta \le 3/4$, equations (8) and (23) imply

$$\frac{q_\tau(k)}{p_\tau(i,k)} \le 4 \quad \text{for all } \tau, i, k. \tag{24}$$

Indeed, (23) gives $q_\tau(k) \le q_\tau(i,k) + \alpha/K$, while $p_\tau(i,k) \ge (1-\alpha-\beta)q_\tau(i,k)+\alpha/K$ and $1-\alpha-\beta \ge 1/4$. This proves (24) and verifies Lemma 1 with $C_0 = 4$.

The coupling also has constant regret cost. By Lemma 4, every coordinate of an exact global block estimate has magnitude at most BKT, so its Euclidean norm is at most $BK^{3/2}T$. Before the final simplification in (23), policy distance is at most $2\eta\sqrt{sK}/(T^4K^5)$. Summing the product of these two bounds over at most T/B blocks gives

$$\sum_\tau \left|\langle q_\tau(i) - q_\tau, \bar{Z}_\tau\rangle\right| \le \frac{2\eta\sqrt{s}}{T^2K^3} \le 1. \tag{25}$$

Thus local policies can be analyzed through the exact pooled estimates without paying a dimension or network factor. Full indexing for the two parity subsequences is in the supplement.

5 Minimax Regret

Set $D = TB + rT/N$ and choose

$$\eta = c\min\left\{\sqrt{\frac{\log K}{D}}, \frac{1}{B}, \frac{N}{r}\right\}, \qquad \beta = 32\eta s/N, \tag{26}$$

for a sufficiently small numerical c. The caps ensure $\eta B \le 1/8$, $\beta \le 1/4$, and $\alpha+\beta \le 3/4$.

Theorem 5 (Distributed linear-bandit upper bound)**.** *Assume $T \ge 3$, $K \ge 2$, and use a block length satisfying (7). The algorithm of Section 2, with parameters* (26)*, communicates at most* $3r$ *real numbers per agent and round and satisfies*

$$\mathrm{Reg}_T \le C\Big[\sqrt{\log K\,(TB + rT/N)} + (B + r/N)\log K + 1\Big] \tag{27}$$

for a numerical constant C.

Proof. For block τ, let

$$L_\tau(k) = \sum_{t\in\mathcal{T}_\tau} \bar{\ell}_t(k)$$

be the true cumulative global loss and let $\bar{Z}_\tau(k)$ be the exact global average of the reconstructed estimates. Conditional unbiasedness of (11) gives

$$\mathbb{E}[\bar{Z}_\tau(k) \mid \mathcal{F}_\tau] = L_\tau(k). \tag{28}$$

Fix one parity subsequence. The ghost policy is measurable before the current block's actions. Equations (24) and (10) verify the domination and spanner-covariance conditions of Theorem 2 with $b = B$ and $C_0 = 4$. If this parity has m blocks, then for every comparator k its ghost regret is at most

$$\frac{\log K}{\eta} + 3\eta m(B^2 + 4Br/N).$$

Equation (28) replaces exact estimates by true block losses in expectation. Equation (25) then transfers the bound to agent i at additive cost one. Summing both parities and using $m_0 + m_1 = T/B$ gives

$$\mathbb{E}\sum_\tau \langle q_\tau(i) - e_k, L_\tau\rangle \le \frac{2\log K}{\eta} + C_1\eta(TB + rT/N) + 2. \tag{29}$$

The actual sampling policy has two exploration mixtures. For every block,

$$\begin{aligned}\langle p_\tau(i) - e_k, L_\tau\rangle &= (1-\alpha-\beta)\langle q_\tau(i) - e_k, L_\tau\rangle \\ &\quad + \alpha\langle u_K - e_k, L_\tau\rangle + \beta\langle u_S - e_k, L_\tau\rangle.\end{aligned}$$

Every coordinate of L_τ is in $[-B, B]$. The last two terms sum to at most

$$2\alpha T + 2\beta T \le 2 + 64\eta sT/N = O(1 + \eta rT/N).$$

This is absorbed by the variance term in (29), proving

$$\mathrm{Reg}_T \le \frac{2\log K}{\eta} + C_2\eta D + C_3.$$

If the square-root choice in (26) is active, balancing gives $O(\sqrt{D\log K})$. If the $1/B$ cap is active, then $D < B^2\log K$ up to constants and both terms are $O(B\log K)$. The N/r cap similarly contributes $O((r/N)\log K)$. This gives (27). $\square$

Substituting (7) yields the advertised network rate.

Corollary 6 (Gap closure)**.** *Under the conditions of Theorem 5,*

$$\mathrm{Reg}_T = \widetilde{O}\left(\sqrt{(\rho^{-1/2} + r/N)T\log K}\right), \tag{30}$$

up to lower-order terms, with $O(r)$ communication per agent and round. Together with the lower bound of Qiu et al. (2026, Theorem 9), this is minimax optimal up to logarithmic factors.

Where pooling matters. If $N = 1$, (26) uses $\beta = \Theta(\eta r)$ and recovers the usual finite-action linear-bandit scaling. If N is large, the information term r/N and the exploration mass shrink together. The communication term does not vanish because a policy still waits one block for network-wide information. This is exactly the decomposition in the lower bound.

Computational cost. The algorithm retains the finite-action entropy update of Qiu et al. (2026). It can require $O(K)$ local work and storage per block. Our result concerns communication and statistical regret, not oracle-efficient optimization over exponentially represented action sets. The spanner is computable in $O(Kr^3 \log r)$ time (Bhaskara et al., 2023).

6 Discussion

The proof separates two roles that explicit exploration often conflates. Spanner exploration controls the exponential moment of the pooled estimator. A much smaller uniform floor controls the deterministic size of messages. Once these roles are separated, the exploration mass tracks collective information, not block length.

This distinction is specific enough to close the open distributed linear-bandit gap, but the pooled log-moment lemma is more general. Whenever a bandit update aggregates independent structured estimators before applying an exponential potential, exploration should scale with the largest aggregation weight. For uniform averaging that weight is $1/N$. A natural extension is to nonuniform network objectives, where the variance involves squared weights while exponential stability involves their maximum. High-probability and adaptive first-order versions are also open.

A A Conditional Exponential-Potential Lemma

We give complete details for the pooled log-moment argument. The following version of conditional Bernstein follows by applying the usual scalar proof inside each regular conditional law.

Lemma 7 (Conditional Bernstein moment bound)**.** *Let $Y_1, \ldots, Y_m$ be conditionally independent given $\mathcal{F}$, with $\mathbb{E}[Y_j \mid \mathcal{F}] = 0$ and $|Y_j| \le R$ almost surely. Write $V = \sum_j \mathbb{E}[Y_j^2 \mid \mathcal{F}]$. If $|\lambda|R < 3$, then*

$$\mathbb{E}\left[\exp\left(\lambda \sum_j Y_j\right) \middle| \mathcal{F}\right] \le \exp\left(\frac{\lambda^2 V}{2(1 - |\lambda|R/3)}\right). \quad (31)$$

Proof. For any centered Y with $|Y| \le R$, expansion of the exponential and $|Y|^k \le R^{k-2}Y^2$ for $k \ge 2$ give

$$\mathbb{E}[e^{\lambda Y} \mid \mathcal{F}] \le 1 + \frac{\lambda^2 \mathbb{E}[Y^2 \mid \mathcal{F}]}{2(1 - |\lambda|R/3)}.$$

Here we used $k! \ge 2\,3^{k-2}$ for $k \ge 2$. Apply $1 + x \le e^x$ and multiply the conditional moment bounds using conditional independence. □

We next record the two elementary scalar estimates used below.

Lemma 8 (Small exponential mixture)**.** *If $|x_k| \le 1/4$ and $q \in \Delta_K$, then*

$$\log \sum_{k=1}^{K} q(k) e^{x_k} \le \sum_{k=1}^{K} q(k)(x_k + x_k^2). \quad (32)$$

Proof. The inequality $e^x \le 1 + x + x^2$ holds on $[-1/4, 1/4]$. Therefore

$$\sum_k q(k) e^{x_k} \le 1 + \sum_k q(k)(x_k + x_k^2).$$

The right side is positive. Apply $\log(1 + u) \le u$. □

We now prove the main pooled estimate with explicit constants. Let $\Omega = \{a_1, \ldots, a_K\}$ span $\mathbb{R}^r$. Let $S = \{b_1, \ldots, b_s\}$ satisfy

$$a_k = S\lambda^{(k)}, \qquad \|\lambda^{(k)}\|_2 \le 1, \qquad s \le 3r, \quad (33)$$

where S also denotes the matrix with columns b_j. Set $\Sigma_S = SS^\top$.

Fix a sigma-field $\mathcal{F}$. For each agent $i \in [N]$, let $p_i \in \Delta_K$ be $\mathcal{F}$-measurable and define

$$M_i = \sum_{k=1}^{K} p_i(k) a_k a_k^\top.$$

For times $t \in [b]$, conditionally independent actions $A_t(i) \sim p_i$ produce losses $\ell_t(i, A_t(i)) = \langle a_{A_t(i)}, \theta_t(i)\rangle$, where $|\langle a_k, \theta_t(i)\rangle| \le 1$. Define

$$X_{i,t,k} = \langle a_k, M_i^{-1} a_{A_t(i)}\rangle \ell_t(i, A_t(i)), \quad (34)$$

$$Z_k = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{b} X_{i,t,k}, \qquad L_k = \mathbb{E}[Z_k \mid \mathcal{F}]. \quad (35)$$

Lemma 9 (Pooled log-moment bound)**.** *Suppose a ghost policy $q \in \Delta_K$ is $\mathcal{F}$-measurable and satisfies*

$$q(k) \le C_0 p_i(k) \quad \textit{for all } i, k. \quad (36)$$

Suppose also

$$M_i \succeq \frac{\beta}{s}\Sigma_S, \qquad 0 < \beta \le 1,$$
$$\eta b \le \frac{1}{8}, \qquad \beta \ge \frac{32\eta s}{N}. \quad (37)$$

Then

$$\mathbb{E}\left[\log \sum_{k=1}^{K} q(k) e^{-\eta Z_k} \middle| \mathcal{F}\right] \le -\eta\langle q, L\rangle + 3\eta^2\left(b^2 + C_0 br/N\right). \quad (38)$$

Proof. The estimator is conditionally unbiased because

$$\mathbb{E}[X_{i,t,k} \mid \mathcal{F}] = \sum_{j=1}^{K} p_i(j)\langle a_k, M_i^{-1} a_j\rangle\langle a_j, \theta_t(i)\rangle = \langle a_k, \theta_t(i)\rangle.$$

Thus $|L_k| \le b$.

Equation (33) gives

$$a_k^T \Sigma_S^{-1} a_k = \lambda^{(k)T} S^T (SS^T)^{-1} S \lambda^{(k)} \le \|\lambda^{(k)}\|_2^2 \le 1. \quad (39)$$

Together with (37), this implies

$$a_k^T M_i^{-1} a_k \le s/\beta. \quad (40)$$

Cauchy–Schwarz in the M_i^{-1} norm and bounded losses now give $|X_{i,t,k}| \le s/\beta$. Since the conditional mean has magnitude at most one,

$$\left|\frac{X_{i,t,k} - \mathbb{E}[X_{i,t,k} \mid \mathcal{F}]}{N}\right| \le \frac{s/\beta + 1}{N} \le \frac{2s}{\beta N} \le \frac{1}{16\eta}. \quad (41)$$

Let $v_k = \mathrm{Var}(Z_k \mid \mathcal{F})$. Conditional independence and the usual linear-bandit second-moment calculation yield

$$v_k \le \frac{1}{N^2}\sum_{i,t} \mathbb{E}[X_{i,t,k}^2 \mid \mathcal{F}] \le \frac{1}{N^2}\sum_{i,t} a_k^T M_i^{-1} a_k. \quad (42)$$

Indeed,

$$\mathbb{E}[X_{i,t,k}^2 \mid \mathcal{F}] \le \sum_j p_i(j)\langle a_k, M_i^{-1}a_j\rangle^2 = a_k^T M_i^{-1} a_k.$$

Equations (40) and (42) give

$$v_k \le \frac{bs}{\beta N} \le \frac{b}{32\eta}. \tag{43}$$

For the weighted variance, define $M_q = \sum_k q(k)a_k a_k^T$. The domination condition gives $M_q \preceq C_0 M_i$. Therefore

$$\sum_k q(k)v_k \le \frac{1}{N^2}\sum_{i,t}\sum_k q(k)a_k^T M_i^{-1}a_k = \frac{1}{N^2}\sum_{i,t}\operatorname{tr}(M_i^{-1}M_q) \le \frac{C_0 br}{N}. \tag{44}$$

Apply Lemma 7 to the centered summands in Z_k. By (41), its denominator is at least $2(1-1/48) > 1$, so

$$\mathbb{E}[e^{-\eta Z_k} \mid \mathcal{F}] \le \exp(-\eta L_k + \eta^2 v_k). \tag{45}$$

The logarithm is concave. Conditional Jensen and (45) imply

$$\mathbb{E}\left[\log \sum_k q(k)e^{-\eta Z_k}\middle|\mathcal{F}\right] \le \log \sum_k q(k)e^{x_k}, \quad x_k = -\eta L_k + \eta^2 v_k. \tag{46}$$

We have $|\eta L_k| \le 1/8$ and, by (43), $\eta^2 v_k \le \eta b/32 \le 1/256$. Thus $|x_k| < 1/4$. Lemma 8 gives

$$\log \sum_k q(k)e^{x_k} \le -\eta\langle q, L\rangle + \eta^2 \sum_k q(k)v_k + \sum_k q(k)x_k^2.$$

Finally,

$$x_k^2 \le 2\eta^2 L_k^2 + 2\eta^4 v_k^2 \le 2\eta^2 b^2 + \frac{1}{128}\eta^2 v_k.$$

Use (44) and enlarge constants to obtain (38). □

Theorem 10 (Pooled EXP2)**.** *For meta-rounds $j \in [m]$, let Z_j be generated as above conditional on the pre-round filtration $\mathcal{F}_j$. Let*

$$q_j(k) = \frac{\exp(-\eta \sum_{h<j} Z_h(k))}{\sum_{u=1}^K \exp(-\eta\sum_{h<j} Z_h(u))}. \tag{47}$$

If Lemma 9's conditions hold at every round, then for all $k \in [K]$,

$$\mathbb{E}\sum_{j=1}^m \langle q_j - e_k, Z_j\rangle \le \frac{\log K}{\eta} + 3\eta m(b^2 + C_0 br/N). \tag{48}$$

Proof. Set $w_1(k) = 1$ and $w_{j+1}(k) = w_j(k)e^{-\eta Z_j(k)}$. Then

$$\log \frac{\sum_k w_{m+1}(k)}{\sum_k w_1(k)} = \sum_{j=1}^m \log \sum_k q_j(k)e^{-\eta Z_j(k)}.$$

For any fixed action k, the left side is at least $-\eta\sum_j Z_j(k) - \log K$. Take expectations and apply Lemma 9 conditionally at each round. Since q_j is $\mathcal{F}_j$-measurable, $\mathbb{E}\langle q_j, Z_j\rangle = \mathbb{E}\langle q_j, L_j\rangle$. Rearrangement gives (48). □

Corollary 11 (Weighted pooling)**.** *Let $w \in \Delta_N$ and define*

$$Z_k = \sum_{i=1}^N w_i \sum_{t=1}^b X_{i,t,k}. \tag{49}$$

If the assumptions of Theorem 10 hold with

$$\beta \ge 32\eta s\|w\|_\infty, \tag{50}$$

then

$$\mathbb{E}\sum_{j=1}^m \langle q_j - e_k, Z_j\rangle \le \frac{\log K}{\eta} + 3\eta m\left(b^2 + C_0 br\|w\|_2^2\right). \tag{51}$$

Proof. Each centered summand in (49) has magnitude at most $2s\|w\|_\infty/\beta \le 1/(16\eta)$. Its conditional variance satisfies

$$v_k \le \sum_i w_i^2 \sum_{t=1}^b a_k^T M_i^{-1} a_k.$$

Therefore

$$\sum_k q(k)v_k \le C_0 br \sum_i w_i^2 = C_0 br\|w\|_2^2.$$

Also $\sum_i w_i^2 \le \|w\|_\infty$, so $v_k \le bs\|w\|_\infty/\beta \le b/(32\eta)$. The proof of Lemma 9 and the potential telescoping are otherwise unchanged. □

B Deterministic Message Control

The next argument makes precise why reducing spanner exploration does not weaken gossip. Let $\Sigma = \sum_{k=1}^K a_k a_k^T$ on the action span.

Lemma 12 (Uniform-floor leverage)**.** *If $p(k) \ge \alpha/K$ for every action and $M = \sum_k p(k)a_k a_k^T$, then*

$$a_k^T M^{-1} a_k \le K/\alpha \quad \textit{for every } k. \tag{52}$$

For any sampled action A and any loss $|\ell(A)| \le 1$,

$$|\langle a_k, M^{-1}a_A \ell(A)\rangle| \le K/\alpha. \tag{53}$$

Proof. We have $M \succeq (\alpha/K)\Sigma$. Since $a_k a_k^T \preceq \Sigma$, conjugation by $\Sigma^{-1/2}$ gives $a_k^T \Sigma^{-1} a_k \le 1$. Hence

$$a_k^T M^{-1} a_k \le (K/\alpha) a_k^T \Sigma^{-1} a_k \le K/\alpha.$$

Cauchy–Schwarz in the M^{-1} norm gives

$$|a_k^T M^{-1} a_A| \le \sqrt{a_k^T M^{-1} a_k}\sqrt{a_A^T M^{-1} a_A} \le K/\alpha.$$

Multiply by $|\ell(A)|$. □

We use the accelerated averaging guarantee in Ye et al. (2023, Proposition 1), in the normalization of Qiu et al. (2026, Lemma 1). If every agent initializes gossip with an s-vector x_i, their choice of B and momentum gives

$$\max_i \|x_i^B - \bar{x}\|_2 \le \frac{\|X^0 - \bar{X}\|_F}{T^6 K^6 \sqrt{N}} \le \frac{2(\sum_i \|x_i\|_2^2)^{1/2}}{T^6 K^6 \sqrt{N}}. \tag{54}$$

Lemma 13 (Spanner consensus)**.** *Use $\alpha = 1/T$ in the sampling mixture. For a block of at most B rounds, let $\bar{Z}_\tau^S$ be the exact global average of the accumulated spanner estimates and let $Z_\tau^B(i)$ be agent i's accelerated-gossip output. Then*

$$\max_i \|Z_\tau^B(i) - \bar{Z}_\tau^S\|_2 \le \delta_S := \frac{2B\sqrt{s}}{T^5 K^5}. \tag{55}$$

If $\bar{Z}_\tau \in \mathbb{R}^K$ and $\widetilde{Z}_\tau(i) \in \mathbb{R}^K$ are reconstructed using (33)*, then*

$$\|\widetilde{Z}_\tau(i) - \bar{Z}_\tau\|_2 \le \sqrt{K}\,\delta_S. \tag{56}$$

Proof. Lemma 12 with $\alpha = 1/T$ bounds each reconstructed action estimate, hence each spanner coordinate, by KT. The initial spanner vector of a block has norm at most $B\sqrt{s}KT$. Substitute this bound for every agent into (54) to obtain (55).

For each action k, equation (33) and Cauchy–Schwarz give

$$|\widetilde{Z}_\tau(i,k) - \bar{Z}_\tau(k)| \le \|\lambda^{(k)}\|_2 \|Z_\tau^B(i) - \bar{Z}_\tau^S\|_2 \le \delta_S.$$

Summing the squared coordinate bounds proves (56). □

C Local and Ghost Policies

The one-block delay is handled by alternating two FTRL instances. We analyze either parity subsequence. Let $\pi(1) < \cdots < \pi(m)$ be its block indices, where $m \le T/B$. The ghost policy is

$$q_j = \arg\min_{q \in \Delta_K} \left\langle q, \sum_{h<j} \bar{Z}_{\pi(h)} \right\rangle + \frac{1}{\eta} \sum_k q(k) \log q(k). \tag{57}$$

Agent i uses the same update with reconstructed gossip outputs $\widetilde{Z}_{\pi(h)}(i)$, producing $q_j(i)$.

Lemma 14 (Policy coupling and domination)**.** *Suppose $\eta \le 1$, $T \ge 3$, $K \ge 2$, $s \le K$, and $\alpha + \beta \le 3/4$. Then*

$$\|q_j - q_j(i)\|_2 \le \eta m \sqrt{K}\,\delta_S \le \frac{1}{KT}, \tag{58}$$

$$\frac{q_j(k)}{p_j(i,k)} \le 4 \quad \textit{for every } i, j, k. \tag{59}$$

Proof. Negative entropy is 1-strongly convex in Euclidean norm on the simplex. The standard sensitivity bound for a strongly convex argmin map and Lemma 13 give

$$\begin{aligned}\|q_j - q_j(i)\|_2 &\le \eta \left\| \sum_{h<j} (\bar{Z}_{\pi(h)} - \widetilde{Z}_{\pi(h)}(i)) \right\|_2 \\ &\le \eta m \sqrt{K}\,\delta_S \le \frac{2\eta\sqrt{sK}}{T^4 K^5}.\end{aligned}$$

The last expression is at most $1/(KT)$ under the stated elementary parameter bounds.

In particular, $q_j(k) \le q_j(i,k) + 1/(KT)$. Since $p_j(i,k) \ge (1-\alpha-\beta) q_j(i,k) + \alpha/K$ and $\alpha/K = 1/(KT)$,

$$\frac{q_j(k)}{p_j(i,k)} \le \frac{q_j(i,k) + \alpha/K}{(1-\alpha-\beta) q_j(i,k) + \alpha/K} \le 4.$$

□

The same calculation makes direct comparison with the ghost policy negligible.

Lemma 15 (Coupling regret)**.** *For either parity subsequence,*

$$\left| \mathbb{E} \sum_{j=1}^m \langle q_j(i) - q_j, \bar{Z}_{\pi(j)} \rangle \right| \le 1. \tag{60}$$

Proof. Lemma 12 bounds every coordinate of each exact block estimate by BKT. Therefore $\|\bar{Z}_{\pi(j)}\|_2 \le BK^{3/2}T$. Using the pre-simplified bound in the proof of Lemma 14, the left side of (60) is at most

$$\frac{T}{B} \cdot \frac{2\eta\sqrt{sK}}{T^4 K^5} \cdot BK^{3/2} T = \frac{2\eta\sqrt{s}}{T^2 K^3} \le 1.$$

□

D Proof of the Distributed Regret Bound

We now assemble the pieces. If B does not divide T, append at most $B - 1$ zero-loss rounds. This changes neither regret nor any estimate. Let

$$D = TB + rT/N,$$
$$\eta = c\min\left\{\sqrt{\frac{\log K}{D}}, \frac{1}{B}, \frac{N}{r}\right\}, \quad \beta = 32\eta s/N, \quad (61)$$

where $c > 0$ is small enough that $\eta B \leq 1/8$, $\eta \leq 1$, $\beta \leq 1/4$, and $\alpha + \beta \leq 3/4$.

Theorem 16 (Distributed regret)**.** *For every agent i and every comparator action k,*

$$\mathbb{E}\sum_{t=1}^{T}\langle p_t(i) - e_k, \bar{\ell}_t\rangle \leq C\left[\sqrt{D\log K} + (B + r/N)\log K + 1\right] \quad (62)$$

for a numerical constant C.

Proof. For a block τ, define the true cumulative global loss

$$L_\tau(k) = \sum_{t\in\mathcal{T}_\tau}\bar{\ell}_t(k) = \frac{1}{N}\sum_{i'=1}^{N}\sum_{t\in\mathcal{T}_\tau}\ell_t(i', k).$$

Let $\bar{Z}_\tau(k)$ be the corresponding exact global average of the reconstructed unbiased estimates. Conditional on all pre-block randomness,

$$\mathbb{E}[\bar{Z}_\tau(k) \mid \mathcal{F}_\tau] = L_\tau(k). \quad (63)$$

Fix a parity subsequence. Lemma 14 verifies (36) with $C_0 = 4$. Spanner exploration in the sampling mixture gives $M_\tau(i) \succeq (\beta/s)\Sigma_S$. Therefore Theorem 10, with $b = B$, yields

$$\mathbb{E}\sum_{j=1}^{m}\langle q_j - e_k, \bar{Z}_{\pi(j)}\rangle \leq \frac{\log K}{\eta} + 3\eta m(B^2 + 4Br/N). \quad (64)$$

By (63), the same expected expression is obtained by replacing $\bar{Z}_{\pi(j)}$ with $L_{\pi(j)}$. Lemma 15 transfers (64) from q_j to the local policy $q_j(i)$ at additive cost one.

Sum over the two parity subsequences. Their block counts sum to T/B, so

$$\mathbb{E}\sum_{\tau}\langle q_\tau(i) - e_k, L_\tau\rangle \leq \frac{2\log K}{\eta} + C_1\eta(TB + rT/N) + 2. \quad (65)$$

It remains to account for the sampling mixtures. From $p_\tau(i) = (1-\alpha-\beta)q_\tau(i) + \alpha u_K + \beta u_S$,

$$\langle p_\tau(i) - e_k, L_\tau\rangle = (1-\alpha-\beta)\langle q_\tau(i) - e_k, L_\tau\rangle + \alpha\langle u_K - e_k, L_\tau\rangle + \beta\langle u_S - e_k, L_\tau\rangle.$$

Every coordinate of L_τ lies in $[-B, B]$. The last two terms therefore sum over blocks to at most

$$2\alpha T + 2\beta T \leq 2 + 64\eta sT/N \leq 2 + 192\eta rT/N. \quad (66)$$

Combining (65) and (66) gives

$$\mathrm{Reg}_T(i) \leq \frac{2\log K}{\eta} + C_2\eta D + C_3. \quad (67)$$

If the square-root term in (61) is the minimum, then (67) is $O(\sqrt{D\log K} + 1)$. If $1/B$ is the minimum, then $D = O(B^2\log K)$ and both nonconstant terms are $O(B\log K)$. If N/r is the minimum, then $D = O(r^2\log K/N^2)$ and both terms are $O((r/N)\log K)$. Enlarging the numerical constant proves (62). □

Finally, the block length of Qiu et al. (2026) satisfies $B = O(\rho^{-1/2}\log(KNT))$. Substitution in (62) gives

$$\widetilde{O}\left(\sqrt{(\rho^{-1/2} + r/N)T\log K}\right)$$

up to the displayed lower-order term. Their Theorem 9 gives the matching lower bound for the same finite-action, heterogeneous-loss, gossip-only model.

References

Bhaskara, A., Mahabadi, S., and Vakilian, A. (2023). Tight bounds for volumetric spanners and applications. In *Advances in Neural Information Processing Systems*, volume 36, pages 916–930.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. (2012). Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the Twenty-Fifth Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 41.1–41.14. PMLR.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2019). Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38.

Hazan, E. and Karnin, Z. (2016). Volumetric spanners: An efficient exploration basis for learning. *Journal of Machine Learning Research*, 17(119):1–34.

Ito, S., Hatano, D., Sumita, H., Takemura, K., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. (2020). Delay and cooperation in nonstochastic linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 4872–4883.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *Proceedings of the Thirtieth International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461. PMLR.

Liu, J. and Morse, A. S. (2011). Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165.

Qiu, H., Zhang, M., and Cesa-Bianchi, N. (2026). Near-optimal regret for distributed adversarial bandits: A black-box approach. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 5465–5517. PMLR.

van der Hoeven, D., Zierahn, L., Lancewicki, T., Rosenberg, A., and Cesa-Bianchi, N. (2023). A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and MDPs. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1285–1321. PMLR.

Wan, Y., Wei, T., Song, M., and Zhang, L. (2024). Nearly optimal regret for decentralized online convex optimization. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4862–4888. PMLR.

Ye, H., Luo, L., Zhou, Z., and Zhang, T. (2023). Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50.

Yi, J. and Vojnović, M. (2023). Doubly adversarial federated bandits. In *Proceedings of the Fortieth International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39951–39967. PMLR.