Communication-Corruption Coupling and Verification in Cooperative Multi-Objective Bandits

Ming Shi

Dept. of Electrical Engineering
University at Buffalo
Buffalo, NY, USA
mshi24@buffalo.edu

Abstract—We study cooperative stochastic multi-armed bandits with vector-valued rewards under adversarial corruption and limited verification. In each of T rounds, each of N agents selects an arm, the environment generates a clean reward vector, and an adversary perturbs the observed feedback subject to a global corruption budget Γ. Performance is measured by team regret under a coordinate-wise nondecreasing, L-Lipschitz scalarization ϕ, covering linear, Chebyshev, and smooth monotone utilities. Our main contribution is a communication-corruption coupling: we show that a fixed environment-side budget Γ can translate into an effective corruption level ranging from Γ to $N\Gamma$, depending on whether agents share raw samples, sufficient statistics, or only arm recommendations. We formalize this via a protocol-induced multiplicity functional and prove regret bounds parameterized by the resulting effective corruption. As corollaries, raw-sample sharing can suffer an N-fold larger additive corruption penalty, whereas summary sharing and recommendation-only sharing preserve an unamplified $O(\Gamma)$ term and achieve centralized-rate team regret. We further establish information-theoretic limits, including an unavoidable additive $\Omega(\Gamma)$ penalty and a high-corruption regime $\Gamma = \Theta(NT)$ where sublinear regret is impossible without clean information. Finally, we characterize how a global budget ν of verified observations restores learnability. That is, verification is necessary in the high-corruption regime, and sufficient once it crosses the identification threshold, with certified sharing enabling the team's regret to become independent of Γ.

Index Terms—Cooperative multi-agent bandits, multi-objective bandits, adversarial corruption, verified feedback, phase transitions, robust regret analysis, information sharing

I. INTRODUCTION

Stochastic multi-armed bandits (MAB) provide a canonical model for sequential decision-making with unknown reward distributions and a regret objective [1]–[3]. A growing class of applications, however, is inherently distributed and multi-metric: multiple cooperative agents (devices, robots, base stations, edge servers) repeatedly probe the same set of actions while optimizing a vector of objectives such as latency, reliability, energy, and fairness. These settings naturally lead to *multi-agent multi-objective bandits*, where learning speed is shaped not only by the stochasticity of the environment but also by *how agents communicate and aggregate information* [4], [5].

A central obstacle in such deployments is that feedback can be *strategically corrupted*, e.g., measurements may be perturbed by sensor faults, compromised telemetry, click fraud, fake reviews, or data poisoning. This has motivated the study of *stochastic bandits with adversarial corruptions*, where an adaptive adversary perturbs observed rewards subject to a global budget [6]–[8]. Existing theory largely characterizes a single learner's regret as a stochastic term plus corruption term tradeoff. In contrast, in cooperative systems the communication layer can *replicate* corrupted samples across agents (e.g., via raw-sample broadcasting), thereby altering the statistical effect of the same environment-side corruption budget. This communication-induced amplification phenomenon is orthogonal to (and not captured by) standard models of cooperative bandits under clean feedback and limited/imperfect communication [4], [5], and decentralized robustness to malicious/Byzantine agents [9], [10]. Moreover, corruption is especially consequential in *multi-objective* learning. Each pull returns a reward vector, and small perturbations can distort implied trade-offs, invalidate dominance relations, and break optimism based on scalar utilities. While multi-objective bandits have been studied under clean feedback (e.g., Pareto or scalarization objectives) [11]–[15], their intersection with budgeted adversarial corruption and cooperative communication is not well understood.

We study cooperative N-agent stochastic multi-objective bandits with vector rewards in $[0,1]^d$. Agents exchange messages according to one of three canonical protocols: raw-sample sharing (S1), sufficient-statistic sharing (S2), or recommendation-only sharing (S3). An adaptive adversary corrupts unverified observations under a *global* corruption budget Γ, and the team may additionally acquire at most ν *verified* (clean) observations system-wide. Performance is measured by *team regret* under a monotone L-Lipschitz scalarization ϕ (covering linear, Chebyshev/min, and smooth log-sum-exp utilities) [11]. This setting leads to three information-theoretic questions that are specific to cooperative systems:

1) *Communication–corruption coupling:* how does the regret depend on (Γ, N) under different sharing protocols?
2) *Protocol-induced phase transitions:* can a fixed environment-side budget Γ yield qualitatively different learnability depending on whether corrupted data are replicated through communication?
3) *Clean side information:* how much verification budget ν is necessary and sufficient to recover learnability when

corruption is severe?

Our main contributions provide a protocol-level characterization of corruption robustness in cooperative multi-objective bandits.

- **A protocol-level corruption functional (effective corruption).** We formalize how a cooperative algorithm uses each underlying agent-round observation across the network via a *multiplicity* (replication) factor, which induces an *effective corruption* level. This yields a sharp distinction. That is, (S1) can inflate the corruption impact by up to a factor N via replication, whereas (S2)-(S3) avoid replication and preserve the original $O(\Gamma)$ corruption penalty.
- **A meta regret theorem parameterized by effective corruption.** We prove a protocol-agnostic robust-UCB guarantee in which the corruption term scales with the induced effective corruption (rather than Γ itself), and we show how monotone scalarizations can be handled via a dominance-preserving vector-to-scalar optimism principle.
- **Tight separation between sharing modes (S1) versus (S2) and (S3).** For (S1), we provide a matching lower bound showing that naive "append-all" raw sharing can be *information-theoretically* N-times worse in the corruption term. For (S2) and (S3), we obtain centralized-rate team regret (up to logarithmic factors) with only an *unamplified* $O(\Gamma)$ corruption penalty.
- **Verification and certified sharing under high corruption.** To address regimes where corrupted feedback alone is insufficient, we introduce a verification channel and show that sharing *verified certificates* enables the team to filter corrupted recommendations and regain identifiability with ν scaling on the order of $K \log(\cdot)/\Delta_{\min}^2$ (up to constants and Lipschitz factors), even when Γ is linear in the total number of pulls. This mechanism is distinct from prior multi-agent robustness models (e.g., Byzantine agents) [9] and from existing corruption-robust multi-agent bandits that do not treat multi-objective scalarization and verification jointly [16].

Overall, our results show that in cooperative multi-objective bandits, *what agents share* can be as consequential as *how much corruption the environment injects*. Communication protocols reshape the effective statistical contamination, and limited verified information can be leveraged to restore learnability in the high-corruption regime.

II. PROBLEM FORMULATION

Notation: Let $[K] \triangleq \{1, \ldots, K\}$ denote the arm set and d the number of objectives. For $x, y \in \mathbb{R}^d$, write $x \preceq y$ for coordinate-wise inequality. Unless stated otherwise, $\|\cdot\| = \|\cdot\|_\infty$. Let $\mathbf{1} \in \mathbb{R}^d$ be the all-ones vector. There are N agents indexed by $n \in [N]$ and the horizon is T rounds.

A. Multi-Agent Stochastic Multi-Objective Bandits

Each arm k is associated with an unknown distribution $\mathcal{D}_k$ supported on $[0,1]^d$, with mean vector $\mu_k \triangleq \mathbb{E}_{R \sim \mathcal{D}_k}[R] \in$ $[0,1]^d$. At each round $t = 1, 2, \ldots, T$, agent n chooses an arm $k_{n,t} \in [K]$. Conditioned on joint action $\{k_{n,t}\}_{n=1}^N$, the environment draws clean rewards $R_{n,t} \sim \mathcal{D}_{k_{n,t}}$, independently across agents and time (conditioned on the chosen arms). Let $\mathcal{H}_{n,t-1}$ be the σ-field generated by agent n's past observations and all messages received up to time $t-1$ under the communication model in Section II-D. All decisions of agent n at round t are measurable with respect to (w.r.t.) $\mathcal{H}_{n,t-1}$.

B. Verification and Corrupted Observations

After choosing $k_{n,t}$ and before receiving reward feedback, each agent n selects a verification decision $V_{n,t} \in \{0, 1\}$ that is measurable w.r.t. $\mathcal{H}_{n,t-1}$. If $V_{n,t} = 1$, agent n observes the clean reward $R_{n,t}$ (clean side information). We restrict the verification to be constrained by a global budget, i.e.,

$$\sum_{t=1}^{T} \sum_{n=1}^{N} V_{n,t} \leq \nu. \tag{1}$$

If $V_{n,t} = 0$, an adversary selects a corruption vector $C_{n,t} \in \mathbb{R}^d$ and agent n will then observe a corrupted reward

$$\widetilde{R}_{n,t} = \Pi_{[0,1]^d}\big(R_{n,t} + C_{n,t}\big), \tag{2}$$

where $\Pi_{[0,1]^d}$ is coordinate-wise projection. The adversary may be adaptive to the past public transcript (joint actions, all messages, and all previously revealed feedback), and may also depend on the current $(k_{n,t}, V_{n,t})$ before choosing $C_{n,t}$. However, corruption is constrained by a global budget, i.e.,

$$\sum_{t=1}^{T} \sum_{n=1}^{N} \|C_{n,t}\| \leq \Gamma. \tag{3}$$

Thus, we define the actually used feedback

$$X_{n,t} \triangleq \begin{cases} R_{n,t}, & V_{n,t} = 1 \text{ (verified round)}, \\ \widetilde{R}_{n,t}, & V_{n,t} = 0 \text{ (unverified round)}. \end{cases} \tag{4}$$

C. Scalarization and Team Regret

We consider multi-objective scalarizations[1] $\phi : [0,1]^d \to \mathbb{R}$ that are coordinate-wise nondecreasing (i.e., $x \preceq y \Rightarrow \phi(x) \leq \phi(y)$) and L-Lipschitz under ℓ_∞ (i.e., $|\phi(x) - \phi(y)| \leq L\|x - y\|_\infty$ for all vectors $x, y \in [0,1]^d$). Define $\theta_k \triangleq \phi(\mu_k)$ and let $k^* \in \arg\max_{k \in [K]} \theta_k$ be the best arm with largest scalarized mean reward. For $k \neq k^*$, define the gap $\Delta_k \triangleq \theta_{k^*} - \theta_k > 0$.

We measure team scalarization regret as follows,

$$\text{Reg}_\phi^{\text{team}}(T) \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \big(\theta_{k^*} - \theta_{k_{n,t}}\big) = \sum_{k \neq k^*} \Delta_k N_k^{\text{team}}(T), \tag{5}$$

where $N_k^{\text{team}}(T) \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbf{1}\{k_{n,t} = k\}$.

---

[1] This class includes linear scalarizations $\phi(x) = w^\top x$ with $w \in \Delta_d$, Chebyshev scalarization $\phi(x) = \min_{i \in [d]} x_i$, and log-sum-exp smoothing $\phi(x) = \beta^{-1} \log \sum_{i=1}^{d} e^{\beta x_i}$.

D. Communication and Sharing Models

At the end of each round t, all agents broadcast messages over a reliable channel. We consider three canonical message types as follows (and the algorithm additionally specifies how received information is aggregated).

- *(S1) Raw-sample sharing (append-all):* Agent n broadcasts $(k_{n,t}, X_{n,t}, V_{n,t})$. Each agent maintains a local multiset of received triples and updates its per-arm estimates by appending all received samples as unit-weight data.
- *(S2) Sufficient-statistic sharing (synchronized global aggregates):* Agent n broadcasts per-arm cumulative statistics $\left(H_{n,k}(t), S_{n,k}(t), H_{n,k}^{\mathrm{ver}}(t), S_{n,k}^{\mathrm{ver}}(t)\right)_{k \in [K]}$, where $H_{n,k}(t) \triangleq \sum_{\tau \leq t} \mathbf{1}\{k_{n,\tau} = k\}$ is the number of times agent n has pulled arm k up to round t, $S_{n,k}(t) \triangleq \sum_{\tau \leq t : k_{n,\tau} = k, V_{n,\tau} = 0} \widetilde{R}_{n,\tau}$ is the cumulative sum of the (possibly corrupted) observed reward vectors $\widetilde{R}_{n,\tau}$ obtained by agent n from arm k up to round t, $H_{n,k}^{\mathrm{ver}}(t) \triangleq \sum_{\tau \leq t} \mathbf{1}\{k_{n,\tau} = k, V_{n,\tau} = 1\}$ is the number of *verified* pulls of arm k by agent n up to round t, $S_{n,k}^{\mathrm{ver}}(t) \triangleq \sum_{\tau \leq t : k_{n,\tau} = k, V_{n,\tau} = 1} R_{n,\tau}$ is the cumulative sum of the corresponding clean rewards $R_{n,\tau}$ from those verified pulls. Upon receiving all broadcasts, corresponding synchronized global aggregates are $H_k(t) = \sum_{n=1}^{N} H_{n,k}(t)$, $S_k(t) = \sum_{n=1}^{N} S_{n,k}(t)$, $H_k^{\mathrm{ver}}(t) = \sum_{n=1}^{N} H_{n,k}^{\mathrm{ver}}(t)$, $S_k^{\mathrm{ver}}(t) = \sum_{n=1}^{N} S_{n,k}^{\mathrm{ver}}(t)$, and all agents can use the same global statistics to compute indices.
- *(S3) Recommendation-only sharing:* Agent n broadcasts only an arm index $M_{n,t} \in [K]$ (e.g., its local argmax index). The message could include a *verified certificate* computed only from verified samples (defined in Section III-F). No reward vectors are communicated.

III. MAIN RESULTS

We present our main theoretical results in this section. Please see our technical report for the complete proofs [17].

A. Multiplicity and Effective Corruption

A cooperative protocol specifies the message content (S1)-(S3) and an aggregation rule, i.e., which observed samples are incorporated (possibly multiple times) into the estimators that drive arm indices. We formalize this by tracking sample reuse.

Let $\widehat{\mu}_{j,k}(t)$ denote the empirical mean vector for arm k used by *estimator* j at the end of round t. For each estimator j and arm k, let $\mathcal{I}_{j,k}(t) \subseteq [N] \times [t]$ be the multiset of agent-round indices (n, τ) whose (unverified) observations $\widetilde{R}_{n,\tau}$ are included with unit weight in $\widehat{\mu}_{j,k}(t)$.

Definition 1 (Multiplicity and effective corruption). *For each agent-round (n, t), we define its multiplicity*

$$\rho_{n,t} \triangleq \#\{j : (n,t) \in \mathcal{I}_{j,k_{n,t}}(T)\}, \quad (6)$$

i.e., the number of distinct index-driving estimators that include $\widetilde{R}_{n,t}$ with unit weight. Define the effective corruption

$$\Gamma_{\mathrm{eff}} \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \rho_{n,t} \|C_{n,t}\|_\infty, \quad (7)$$

and the arm-wise effective corruption

$$\Gamma_{\mathrm{eff},k} \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \rho_{n,t} \|C_{n,t}\|_\infty \mathbf{1}\{k_{n,t} = k\}. \quad (8)$$

Lemma 1 (Effective corruption under (S1)-(S3)). *Under (S1) raw-sample sharing with append-all, each agent maintains its own local estimator, hence $\rho_{n,t} = N$ and $\Gamma_{\mathrm{eff}} = N\Gamma$. Under (S2) synchronized sufficient-statistic sharing, all agents compute indices from a single synchronized global estimator, hence $\rho_{n,t} = 1$ and $\Gamma_{\mathrm{eff}} = \Gamma$. Under (S3) recommendation-only sharing, no raw samples are transmitted and each unverified observation is used only locally, hence $\rho_{n,t} = 1$ and $\Gamma_{\mathrm{eff}} = \Gamma$.*

B. A Concentration Lemma with Arm-Wise Effective Corruption

The next lemma is the workhorse: any estimator whose arm-k empirical mean is formed from m unverified samples of arm k incurs a stochastic fluctuation term $O(1/\sqrt{m})$ plus a bias term proportional to the total effective corruption mass assigned to arm k divided by m.

Lemma 2 (Vector mean concentration under effective corruption). *Fix $\delta \in (0,1)$. With probability at least $1 - \delta$, simultaneously for all estimators j, all arms k, and all times t,*

$$\|\widehat{\mu}_{j,k}(t) - \mu_k\|_\infty \leq \sqrt{\frac{\log(2dKNT/\delta)}{2 \max\{1, |\mathcal{I}_{j,k}(t)|\}}} + \frac{\Gamma_{\mathrm{eff},k}}{\max\{1, |\mathcal{I}_{j,k}(t)|\}}. \quad (9)$$

Proof sketch. Decompose $\widehat{\mu}_{j,k}(t) - \mu_k = (\overline{\mu}_{j,k}(t) - \mu_k) + (\widehat{\mu}_{j,k}(t) - \overline{\mu}_{j,k}(t))$, where $\overline{\mu}_{j,k}(t)$ is the mean of the corresponding clean rewards. The first term is bounded by coordinate-wise Hoeffding and union bound. For the second term, projection is nonexpansive in ℓ_∞, so $\|\widetilde{R}_{n,t} - R_{n,t}\|_\infty \leq \|C_{n,t}\|_\infty$. Summing over the multiset $\mathcal{I}_{j,k}(t)$ yields a bias bounded by the total corruption mass assigned to arm k across all samples used by estimator j. Summing this worst-case usage over all estimators is precisely captured by $\Gamma_{\mathrm{eff},k}$. $\square$

C. Dominance Lifting for Monotone Scalarizations

Definition 2 (Upper-closed confidence sets). *A set $\mathcal{C} \subseteq [0,1]^d$ is upper-closed if $x \in \mathcal{C}$ and $x \preceq y \preceq \mathbf{1}$ implies $y \in \mathcal{C}$.*

Lemma 3 (Upper-corner reduction). *Let $\mathcal{C} \subseteq [0,1]^d$ be upper-closed and define $x^{\mathrm{max}} \in [0,1]^d$ by $x_j^{\mathrm{max}} \triangleq \sup\{x_j : x \in \mathcal{C}\}$. Then, $\sup_{x \in \mathcal{C}} \phi(x) = \phi(x^{\mathrm{max}})$. In particular, for ℓ_∞ rectangles $\mathcal{C} = \{x : \|x - \widehat{\mu}\|_\infty \leq \beta\} \cap [0,1]^d$, we have*

$$\sup_{x \in \mathcal{C}} \phi(x) = \phi\big(\Pi_{[0,1]^d}(\widehat{\mu} + \beta\mathbf{1})\big). \quad (10)$$

D. A Meta Regret Theorem Parameterized by Γ_{eff}

Denote the upper bound, i.e., the right-hand side, of (9) by $\beta_{j,k}(t)$. Consider any protocol that produces (possibly multiple) estimators $\widehat{\mu}_{j,k}(t)$ and uses the following index rule: each estimator j forms an optimistic index $U_{j,k}(t) \triangleq \phi\Big(\Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big)\Big)$ with radius $\beta_{j,k}(t)$. Agents choose arms based on their designated estimator's indices (local in S1, global in S2, local in S3).

Theorem 1 (Team regret bound via effective corruption). *Fix $\delta \in (0,1)$. On an event of probability at least $1 - \delta$, any cooperative UCB-type protocol using radii $\beta_{j,k}(t)$ satisfies*

$$Reg_\phi^{\text{team}}(T) \leq cL\Big(\sqrt{KNT\log(2dKNT/\delta)} \\ + \Gamma_{\text{eff}}\log(1 + NT)\Big), \qquad (11)$$

for a universal constant $c > 0$.

Proof sketch. On the concentration event from Lemma 2, the rectangle $\{x : \|x - \widehat{\mu}_{j,k}(t)\|_\infty \leq \beta_{j,k}(t)\}$ contains μ_k. Lemma 3 implies optimism for $\theta_k = \phi(\mu_k)$. Whenever a suboptimal arm $k \neq k^*$ is selected by an estimator j, standard UCB reasoning yields $\Delta_k \leq 2L\beta_{j,k}(t)$. Summing the resulting pull-count constraints across all estimators and arms produces a stochastic term $\tilde{O}(L\sqrt{KNT})$ and a corruption term proportional to $\sum_{n,t} \rho_{n,t}\|C_{n,t}\|_\infty = \Gamma_{\text{eff}}$, since each corrupted sample can only "pay for" a bounded number of additional optimistic selections before confidence shrinks. $\square$

Corollary 1 ((S1) Raw-sample sharing incurs N-fold amplification). *Under (S1), $\Gamma_{\text{eff}} = N\Gamma$ (Lemma 1), hence we have*

$$Reg_\phi^{\text{team}}(T) \leq \widetilde{O}\big(L\sqrt{KNT} + LN\Gamma\big). \qquad (12)$$

Corollary 2 ((S2)–(S3) Achieve centralized corruption penalty). *Under (S2) or (S3), $\Gamma_{\text{eff}} = \Gamma$, hence we have*

$$Reg_\phi^{\text{team}}(T) \leq \widetilde{O}\big(L\sqrt{KNT} + L\Gamma\big). \qquad (13)$$

E. Lower Bound for Naive Raw Sharing

Theorem 2 (Lower bound). *Under (S1) raw-sample sharing with append-all local estimators, there exist instances and adversaries with $\sum_{t,n}|C_{n,t}| \leq \Gamma$ such that*

$$\mathbb{E}\big[Reg_\phi^{\text{team}}(T)\big] \geq c_1\sqrt{KNT} + c_2 N\Gamma \qquad (14)$$

for universal constants $c_1, c_2 > 0$.

Proof sketch. Construct a two-point testing instance where identifying the best arm requires resolving a mean gap Δ. An adversary spends corruption mass $\Theta(\Gamma)$ on early samples of the informative arm. Under (S1), the same corrupted samples enter N local estimators, reducing N estimators' effective KL simultaneously. Le Cam's method yields an additional error probability bounded away from zero unless the protocol spends $\Omega(N\Gamma)$ regret mass to compensate. $\square$

F. Verification and Certified Recommendation Sharing

Verification provides a clean channel immune to corruption. The key multi-agent issue is how to convert scattered verified reward samples into *network-wide reliable decisions* under limited communication among all agnets. For each arm $k \in [K]$, we define the verified-only empirical mean ($\max\{1, H_k^{\text{ver}}(t)\}$ is used to avoid division by zero when no verification has occurred yet)

$$\widehat{\mu}_k^{\text{ver}}(t) \triangleq \frac{1}{\max\{1, H_k^{\text{ver}}(t)\}} \sum_{\tau \leq t} \sum_{n:k_{n,\tau}=k, V_{n,\tau}=1} R_{n,\tau}, \quad (15)$$

which averages only the clean reward vectors $R_{n,\tau}$ collected from verified pulls of arm k up to time t.

Lemma 4 (Verified scalar certificate). *Fix $\delta \in (0,1)$. With probability at least $1 - \delta$, for all k and all t, we have*

$$\big|\phi(\widehat{\mu}_k^{\text{ver}}(t)) - \theta_k\big| \leq L\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, H_k^{\text{ver}}(t)\}}}. \qquad (16)$$

Under the recommendation-only communication model (S3), agents do not transmit reward vectors (which could replicate corrupted samples). However, pure recommendations $M_{n,t}$ can still be misleading under corruption. To make recommendations *verifiable*, we allow an agent to optionally attach a certificate computed *only from verified (clean) data*.

Concretely, at the end of round t, agent n broadcasts a pair $(M_{n,t}, \text{cert}_{n,t})$ for $M_{n,t} \in [K]$, where $M_{n,t}$ is the arm it recommends (e.g., its current optimistic maximizer or most-played arm in a round). The certificate $\text{cert}_{n,t} = \big(\text{LCB}_{n,t}, \text{UCB}_{n,t}\big)$ is a scalar confidence interval for the *true* scalarized value $\theta_{M_{n,t}} = \phi(\mu_{M_{n,t}})$ of the recommended arm, based solely on verified-only statistics. Specifically, we set $\text{LCB}_{n,t} = \phi(\widehat{\mu}_{M_{n,t}}^{\text{ver}}(t)) - \varepsilon_{M_{n,t}}(t)$, $\text{UCB}_{n,t} = \phi(\widehat{\mu}_{M_{n,t}}^{\text{ver}}(t)) + \varepsilon_{M_{n,t}}(t)$, where the half-width $\varepsilon_k(t) \triangleq L\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, H_k^{\text{ver}}(t)\}}}$ follows from Hoeffding concentration applied coordinate-wise to verified rewards in $[0,1]^d$ (with a union bound over d objectives, K arms, and NT agent-rounds) together with L-Lipschitzness of ϕ under ℓ_∞. Because $\text{cert}_{n,t}$ depends only on verified samples, it is immune to arbitrary corruption on unverified rounds. Thus, even if an adversary can manipulate the unverified feedback that led agent n to propose $M_{n,t}$, the interval $[\text{LCB}_{n,t}, \text{UCB}_{n,t}]$ remains a valid high-probability bracket for $\theta_{M_{n,t}}$. Recipients can therefore *filter* incoming recommendations by keeping only those with sufficiently tight certificates (large $H_k^{\text{ver}}(t)$) and competitive lower bounds relative to others (formalized in the filtering rule below), ensuring that network-wide coordination is driven by clean evidence rather than corrupted impressions.

Theorem 3 (High-corruption learnability via certified sharing). *There exists a cooperative algorithm using (S3) recommendation-only sharing with optional verified certificates and at most ν verifications such that:*

1) *(Always-valid robust baseline) For all instances and adversaries with budget (3),*

$$Reg_\phi^{\text{team}}(T) \leq \widetilde{O}\big(L \cdot Reg_{\text{comm}}^{\text{clean}}(K, N, T) + L\Gamma\big), \quad (17)$$

   where $Reg_{\text{comm}}^{\text{clean}}(K, N, T)$ is the regret overhead of the same coordination protocol in the clean case.

2) *(Verification-driven override) If $\Delta_{\min} \triangleq \min_{k \neq k^*} \Delta_k > 0$ and*

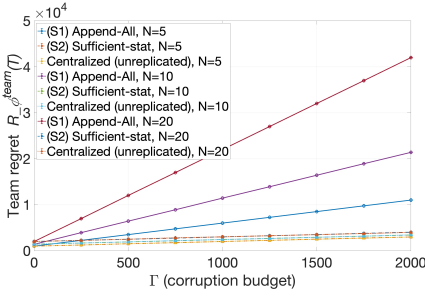$$\nu \geq cK\frac{L^2}{\Delta_{\min}^2}\log\Big(\frac{2dKNT}{\delta}\Big), \qquad (18)$$

Fig. 1. Team regret versus corruption budget Γ for (S1) and (S2): (S1) scales like $\widetilde{O}(\sqrt{KNT} + N\Gamma)$ while (S2) scales like $\widetilde{O}(\sqrt{KNT} + \Gamma)$.
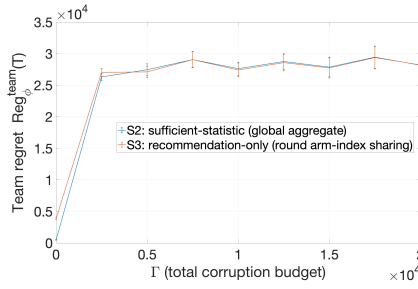

Fig. 2. Team regret versus Γ under (S2) and (S3): (S3) preserves the unamplified $O(\Gamma)$ term while trading off a clean-case coordination overhead.
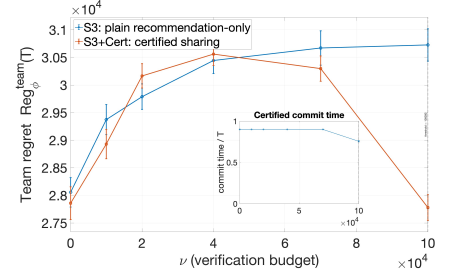

Fig. 3. Team regret versus verification budget ν under (S3) with/without certified sharing in high-corruption regime $\Gamma = \Theta(NT)$: sharp improvement once $\nu \gtrsim KL^2\Delta_{\min}^{-2} \log(dKNT)$.

then, with probability at least $1 - \delta$, there exists $t^* \leq T$ after which all agents play k^* and

$$Reg_\phi^{\text{team}}(T) \leq O(\nu\Delta_{\max})$$
$$+ \widetilde{O}\big(L \cdot Reg_{\text{comm}}^{\text{clean}}(K, N, T) \wedge NT\Delta_{\max}\big), \qquad (19)$$

independently of Γ (even when $\Gamma = \Theta(NT)$).

Theorem 3 gives a two-regime guarantee. In general, recommendation-only sharing avoids sample replication, so corruption contributes only an additive $O(\Gamma)$ term at the team level, plus $Reg_{\text{comm}}^{\text{clean}}(K, N, T)$. Moreover, once $\nu \gtrsim \widetilde{\Theta}(KL^2/\Delta_{\min}^2)$, verified certificates become tight enough to rule out suboptimal arms, since they rely only on clean samples, the adversary cannot forge them. Thus, the team can filter misleading recommendations and commit to k^*, making regret independent of Γ even when $\Gamma = \Theta(NT)$.

Proof sketch (certificate dominance mechanism). Under (18), the verified confidence width satisfies $\varepsilon_k(t) \leq \Delta_{\min}/4$ once $H_k^{\text{ver}}(t)$ crosses threshold. By Lemma 4, any certified suboptimal arm $k \neq k^*$ then satisfies $\text{UCB}_k^{\text{ver}}(t) \leq \theta_k + \Delta_{\min}/4 < \theta_{k^*} - \Delta_{\min}/4 \leq \text{LCB}_{k^*}^{\text{ver}}(t)$, so certified filtering retains only k^*. Because certificates depend solely on verified samples, corruption on unverified rounds cannot invalidate them. Broadcast ensures that once a valid k^* certificate appears, all agents receive it and commit. $\square$

IV. NUMERICAL RESULTS

We empirically validate the communication-corruption-verification theory in Section III. Our experiments target four qualitative predictions, corruption amplification under raw-sample sharing (S1), centralized-rate robustness under sufficient-statistic sharing (S2), unamplified $O(\Gamma)$ robustness under recommendation-only sharing (S3) with reduced communication, and high-corruption recovery under (S3) via certified sharing with verification budget ν.

We simulate by setting $K = 20$, $d = 5$, $N \in \{5, 10, 20\}$, and $T = 10^4$. Rewards are coordinate-wise Bernoulli: for each arm k, $R_{n,t} \sim \text{Bernoulli}(\mu_k)$ independently across n, t and coordinates, with $\mu_k \in [0, 1]^d$. We test three monotone L-Lipschitz scalarizations: linear $\phi(x) = w^\top x$ ($w \in \Delta_d$), Chebyshev $\phi(x) = \min_j x^{(j)}$, and log-sum-exp $\phi(x) = $ $\beta^{-1} \log \sum_j e^{\beta x^{(j)}}$. We compare all the three sharing modes (S1), (S2), and (S3). We report team regret $Reg_\phi^{\text{team}}(T)$, and average results over 50 i.i.d. instances.

Figure 1 plots $Reg_\phi^{\text{team}}(T)$ versus Γ for (S1) and (S2). Consistent with Lemma 1 and Theorem 1, (S1) exhibits an N-fold degradation in the corruption-dominated regime. The slope of regret versus Γ scales approximately linearly with N. In contrast, (S2) closely tracks the centralized (unreplicated) benchmark and remains stable as N increases.

Figure 2 compares (S3) recommendation-only sharing with (S2). As predicted by Corollary 2 and Theorem 3, (S3) avoids corruption amplification (the Γ-slope matches (S2)), but incurs an additional clean-case coordination overhead that is visible when Γ is small. Communication is reduced from transmitting reward vectors or per-arm d-dimensional summaries to transmitting $O(N \log T)$ arm indices.

We set $\Gamma = \Theta(NT)$ and vary the verification budget ν, and Figure 3 shows that plain (S3) recommendations can be misled by corrupted local estimates, yielding large regret even with moderate ν. In contrast, certified sharing produces a sharp improvement once ν crosses the identification threshold in (18). Regret drops rapidly and the team commits to k^* shortly thereafter, consistent with Theorem 3.

V. CONCLUSION

We studied cooperative N-agent stochastic multi-objective bandits under a global corruption budget Γ and a verification budget ν. Our main message is that robustness is jointly governed by *communication* and *corruption*: protocol-induced replication converts Γ into an effective budget $\Gamma_{\text{eff}} \in [\Gamma, N\Gamma]$, producing an N-fold gap between raw-sample sharing and summary/recommendation sharing. We proved a protocol-agnostic regret bound parameterized by Γ_{eff} for general monotone L-Lipschitz scalarizations, yielding tight corollaries for (S1)–(S3), and matching lower bounds showing the amplification under naive raw sharing is unavoidable. Finally, we identified a high-corruption regime $\Gamma = \Theta(NT)$ where sublinear team regret is impossible without clean information, and showed that verification restores learnability when shared as certified evidence: once ν exceeds the identification threshold, certificates enable reliable filtering and team-wide commitment with regret independent of Γ.

REFERENCES

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[4] M. Agarwal, S. Aggarwal, and K. Azizzadenesheli, "Multi-agent multi-armed bandits with limited communication," *Journal of Machine Learning Research*, vol. 23, no. 305, pp. 1–24, 2022. JMLR 23(305):9529–9552 (article numbering varies by index).

[5] U. Madhushani, A. Dubey, N. Leonard, and A. Pentland, "One more step towards reality: Cooperative bandits with imperfect communication," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 7813–7824, 2021.

[6] T. Lykouris, V. Mirrokni, and R. Paes Leme, "Stochastic bandits robust to adversarial corruptions." arXiv preprint arXiv:1803.09353, 2018.

[7] A. Gupta, T. Koren, and K. Talwar, "Better algorithms for stochastic bandits with adversarial corruptions," in *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, vol. 99 of *Proceedings of Machine Learning Research*, pp. 1562–1578, PMLR, 2019.

[8] S. Ito, "On optimal robustness to adversarial corruption in online decision problems," in *Advances in Neural Information Processing Systems*, 2021.

[9] J. Zhu, A. Koppel, Á. Velasquez, and J. Liu, "Byzantine-resilient decentralized multi-armed bandits under stochastic and adversarial models," *arXiv preprint arXiv:2310.07320*, 2023.

[10] M. Shi, X. Lin, and L. Jiao, "Power-of-2-arms for adversarial bandit learning with switching costs," *IEEE Transactions on Networking*, 2025.

[11] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[12] M. M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2013.

[13] P. Auer, C.-K. Chiang, R. Ortner, and M. M. Drugan, "Pareto front identification from stochastic bandit feedback," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 51 of *Proceedings of Machine Learning Research*, pp. 939–947, PMLR, 2016.

[14] É. Crépon, A. Garivier, and W. M. Koolen, "Sequential learning of the pareto front for multi-objective bandits," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 238 of *Proceedings of Machine Learning Research*, pp. 3583–3591, PMLR, 2024.

[15] L. Cao, M. Shi, and N. B. Shroff, "Provably efficient multi-objective bandit algorithms under preference-centric customization," *arXiv preprint arXiv:2502.13457*, 2025.

[16] F. Ghaffari, M. H. Movahedian, and O. Darwiche, "Multi-agent stochastic bandits robust to adversarial corruptions," in *Proceedings of The 7th Annual Learning for Dynamics and Control Conference*, Proceedings of Machine Learning Research, 2025.

[17] , "Communication–corruption coupling and verification in cooperative multi-objective bandits." https://mingshihomepage.com/CommunicationCorruption%20Coupling%20and%20Verification%20in%20Cooperative%20MOB_ISIT2026.pdf, 2026.

APPENDIX A
PROOF OF LEMMA 1

Proof. We first make the multiplicity $\rho_{n,t}$ precise and then specialize it to the three protocols.

A. Step 0: Formal Definition of Multiplicity

Fix a cooperative algorithm Alg and fix a time $t \in [T]$. For each agent-round sample (n,τ) with $\tau \leq t$, define its (protocol-induced) multiplicity

$$\rho_{n,\tau}(t) \triangleq \sum_{n'=1}^{N} w_{n',t}(n,\tau), \qquad (20)$$

where $w_{n',t}(n,\tau) \geq 0$ is the *weight* with which agent n''s estimator at time t uses the single observed vector $\widetilde{R}_{n,\tau}$ (or $R_{n,\tau}$ if verified) when forming its arm-wise sufficient statistics (e.g., sums and counts) that determine its indices at time $t+1$. In other words, for each arm k and agent n', the statistic used by n' at time t can be written as

$$S_k^{(n')}(t) = \sum_{\tau \leq t} \sum_{n=1}^{N} w_{n',t}(n,\tau)\, \widetilde{R}_{n,\tau}\, \mathbf{1}\{k_{n,\tau} = k, V_{n,\tau} = 0\}$$

$$+ \sum_{\tau \leq t} \sum_{n=1}^{N} w_{n',t}^{\text{ver}}(n,\tau)\, R_{n,\tau}\, \mathbf{1}\{k_{n,\tau} = k,\, V_{n,\tau} = 1\}, \qquad (21)$$

with the corresponding counts defined analogously by replacing $\widetilde{R}_{n,\tau}$ (resp. $R_{n,\tau}$) with 1. The weights $\{w_{n',t}(n,\tau)\}$ are completely determined by the communication/sharing rule and the deterministic updates of Alg; they do not depend on the reward distribution. For the present lemma, we only need the following facts: (i) under append-all updates each received raw sample is appended with unit weight; (ii) under synchronized sufficient-statistic sharing there is a single global sum/count in which each raw sample appears once; (iii) under recommendation-only sharing raw samples are never transmitted, hence a sample can only be used by its owner.

For conciseness, we will write $\rho_{n,\tau}$ instead of $\rho_{n,\tau}(T)$, i.e., the total multiplicity over the entire horizon, since the proof below is pointwise in t and hence also holds at $t = T$.

Given multiplicities $\{\rho_{n,t}\}$, the (protocol-dependent) effective corruption is

$$\Gamma_{\text{eff}} \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \rho_{n,t}\, \|C_{n,t}\|_{\infty}. \qquad (22)$$

The environment-side budget is

$$\Gamma \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} \|C_{n,t}\|_{\infty}. \qquad (23)$$

B. Step 1: (S1) Raw-Sample Sharing with Append-All Implies $\rho_{n,t} = N$

Under (S1), at the end of each round t, agent n broadcasts the tuple $(k_{n,t}, \widetilde{R}_{n,t}, V_{n,t})$ (or $(k_{n,t}, R_{n,t}, V_{n,t})$ if verified). Under the append-all update rule, every recipient agent $n' \in [N]$ *appends* each received raw sample into its local dataset

as if it were an additional fresh sample with unit weight. Therefore, for every agent-round pair (n, t) and every recipient $n' \in [N]$, $w_{n', T}(n, t) = 1$, (and similarly $w_{n', T}^{\text{ver}}(n, t) = 1$ on verified rounds). Plugging into (20) yields

$$\rho_{n,t} = \sum_{n'=1}^{N} w_{n',T}(n,t) = \sum_{n'=1}^{N} 1 = N.$$

Substituting into (22) and using (23) gives

$$\Gamma_{\text{eff}} = \sum_{t=1}^{T} \sum_{n=1}^{N} N \|C_{n,t}\|_\infty = N \sum_{t=1}^{T} \sum_{n=1}^{N} \|C_{n,t}\|_\infty = N\Gamma.$$

C. Step 2: (S2) Synchronized Sufficient-Statistic Sharing Implies $\rho_{n,t} = 1$

Under (S2), each agent n broadcasts the per-arm cumulative statistics $\left(H_{n,k}(t), S_{n,k}(t), H_{n,k}^{\text{ver}}(t), S_{n,k}^{\text{ver}}(t)\right)_{k \in [K]}$. The protocol specifies that all agents compute indices using the *same* synchronized global aggregates $H_k(t) = \sum_{n=1}^{N} H_{n,k}(t)$, $S_k(t) = \sum_{n=1}^{N} S_{n,k}(t)$, $H_k^{\text{ver}}(t) = \sum_{n=1}^{N} H_{n,k}^{\text{ver}}(t)$, $S_k^{\text{ver}}(t) = \sum_{n=1}^{N} S_{n,k}^{\text{ver}}(t)$. Crucially, by the definitions of $H_{n,k}(t)$ and $S_{n,k}(t)$ in the formulation, each raw observation $\widetilde{R}_{n,\tau}$ contributes *exactly once* to exactly one summand in $S_{k_{n,\tau}}(t)$, namely to $S_{n,k_{n,\tau}}(t)$, and hence exactly once to the global sum $S_{k_{n,\tau}}(t)$. There is no mechanism that duplicates $\widetilde{R}_{n,\tau}$ across multiple $S_{n',k}$'s, because agents transmit only already-aggregated sums/counts and the global aggregate is formed by addition, not by appending all raw samples into each local dataset.

Formally, for each (n, τ), the global sum $S_k(t)$ (and similarly $S_k^{\text{ver}}(t)$) can be written as

$$S_k(t) = \sum_{\tau' \le t} \sum_{n'=1}^{N} \widetilde{R}_{n',\tau'} \mathbf{1}\{k_{n',\tau'} = k, V_{n',\tau'} = 0\},$$

where the coefficient of $\widetilde{R}_{n,\tau}$ is exactly 1 if $(k_{n,\tau} = k, V_{n,\tau} = 0)$ and 0 otherwise. Thus, in the weight notation, for the estimator actually used to form indices, we have $w_{n',t}(n,\tau) = \mathbf{1}\{n' = n^*\}$, for some single (conceptual) synchronized estimator n^*, or equivalently (and more simply): the aggregate uses each sample with unit weight exactly once. Hence $\rho_{n,\tau} = 1$ for all (n, τ). Substituting into (22) yields

$$\Gamma_{\text{eff}} = \sum_{t=1}^{T} \sum_{n=1}^{N} 1 \cdot \|C_{n,t}\|_\infty = \Gamma.$$

D. Step 3: (S3) Recommendation-Only Sharing Implies $\rho_{n,t} = 1$

Under (S3), agents broadcast only an arm index $M_n(t) \in [K]$ (optionally with a certificate based on verified-only statistics), and *no reward vectors $\widetilde{R}_{n,t}$ (or $R_{n,t}$) are ever transmitted*. Therefore, the only agent that can possibly include $\widetilde{R}_{n,t}$ in any estimator is the originating agent n itself; other agents do not observe $\widetilde{R}_{n,t}$ and cannot reconstruct it from $M_n(t)$. Consequently, for every (n, t) and for every other agent $n' \ne n$, $w_{n',T}(n, t) = 0$, while $w_{n,T}(n, t) = 1$ because

agent n uses its own sample once in its own local sufficient statistics. Hence, (20) gives

$$\rho_{n,t} = \sum_{n'=1}^{N} w_{n',T}(n,t) = 1.$$

Substituting into (22) yields $\Gamma_{\text{eff}} = \Gamma$.

Combining Steps 1–3 proves the stated values of $\rho_{n,t}$ and Γ_{eff} for (S1)–(S3). $\square$

APPENDIX B
PROOF OF LEMMA 2

Proof. We prove a coordinate-wise bound and then take a union bound over coordinates, arms, times, and estimators.

A. Step 0: Estimator/Sample-Set Notation

Fix an estimator index j, an arm k, and a time t. Let $\mathcal{I}_{j,k}(t)$ denote the (multi-)set of agent-round indices whose observations are used with unit weight by estimator j to form the empirical mean of arm k up to time t.[2] Let $m \triangleq |\mathcal{I}_{j,k}(t)|$. If $m = 0$, then $\widehat{\mu}_{j,k}(t)$ can be defined arbitrarily (e.g., 0); since $\|\widehat{\mu}_{j,k}(t) - \mu_k\|_\infty \le 1$ and the right-hand side of (9) is at least $\sqrt{\log(2dKNT/\delta)/2}$, the bound holds trivially. Henceforth assume $m \ge 1$.

Index the elements of $\mathcal{I}_{j,k}(t)$ as $\{(n_r, \tau_r)\}_{r=1}^{m}$ (including repetitions if any). For each $r \in [m]$, the underlying clean reward is $R_{n_r,\tau_r} \in [0,1]^d$ and the (possibly corrupted) observed reward is

$$\widetilde{R}_{n_r,\tau_r} = \begin{cases} R_{n_r,\tau_r}, & \text{if } V_{n_r,\tau_r} = 1, \\ \Pi_{[0,1]^d}(R_{n_r,\tau_r} + C_{n_r,\tau_r}), & \text{if } V_{n_r,\tau_r} = 0. \end{cases}$$

Then, the empirical mean used by estimator j for arm k at time t is

$$\widehat{\mu}_{j,k}(t) = \frac{1}{m} \sum_{r=1}^{m} \widetilde{R}_{n_r,\tau_r}.$$

B. Step 1: Decomposition Into Stochastic Fluctuation and Corruption Bias

Using $\mu_k = \mathbb{E}[R \mid k]$ and adding/subtracting clean rewards,

$$\widehat{\mu}_{j,k}(t) - \mu_k = \underbrace{\frac{1}{m} \sum_{r=1}^{m} (R_{n_r,\tau_r} - \mu_k)}_{\text{stochastic term}}$$
$$+ \underbrace{\frac{1}{m} \sum_{r=1}^{m} (\widetilde{R}_{n_r,\tau_r} - R_{n_r,\tau_r})}_{\text{corruption term}}. \quad (24)$$

Taking $\|\cdot\|_\infty$ and using the triangle inequality gives

$$\|\widehat{\mu}_{j,k}(t) - \mu_k\|_\infty \le \left\| \frac{1}{m} \sum_{r=1}^{m} (R_{n_r,\tau_r} - \mu_k) \right\|_\infty$$
$$+ \left\| \frac{1}{m} \sum_{r=1}^{m} (\widetilde{R}_{n_r,\tau_r} - R_{n_r,\tau_r}) \right\|_\infty. \quad (25)$$

[2] If the protocol duplicates samples (e.g., append-all under (S1)), then $\mathcal{I}_{j,k}(t)$ is naturally viewed as a multiset that may contain multiple copies of the same agent-round observation.

C. Step 2: Bounding the Corruption Term by $\Gamma_{\mathrm{eff},k}/m$

We claim that for every (n,τ),

$$\|\widetilde{R}_{n,\tau} - R_{n,\tau}\|_{\infty} \leq \|C_{n,\tau}\|_{\infty}, \qquad (26)$$

where we adopt the convention $C_{n,\tau} \equiv 0$ on verified rounds $(V_{n,\tau} = 1)$, so that (26) holds uniformly. Indeed, coordinate-wise projection onto $[0,1]$ is 1-Lipschitz and fixes $R_{n,\tau}^{(\ell)} \in [0,1]$, hence

$$\big|\widetilde{R}_{n,\tau}^{(\ell)} - R_{n,\tau}^{(\ell)}\big| = \big|\Pi_{[0,1]}\big(R_{n,\tau}^{(\ell)} + C_{n,\tau}^{(\ell)}\big) - \Pi_{[0,1]}\big(R_{n,\tau}^{(\ell)}\big)\big|$$
$$\leq |C_{n,\tau}^{(\ell)}| \leq \|C_{n,\tau}\|_{\infty}, \quad (27)$$

and taking the maximum over $\ell \in [d]$ yields (26).

Therefore,

$$\left\|\frac{1}{m}\sum_{r=1}^{m}(\widetilde{R}_{n_r,\tau_r} - R_{n_r,\tau_r})\right\|_{\infty} \leq \frac{1}{m}\sum_{r=1}^{m}\|\widetilde{R}_{n_r,\tau_r} - R_{n_r,\tau_r}\|_{\infty}$$
$$\leq \frac{1}{m}\sum_{r=1}^{m}\|C_{n_r,\tau_r}\|_{\infty}. \qquad (28)$$

We now relate the right-hand side to $\Gamma_{\mathrm{eff},k}$. Let $w_j(n,\tau;t) \in \{0,1,2,\dots\}$ denote the number of times the agent-round observation (n,τ) appears in the multiset $\mathcal{I}_{j,k}(t)$. Then $m = \sum_{\tau' \leq t}\sum_{n'=1}^{N} w_j(n',\tau';t)\,\mathbf{1}\{k_{n',\tau'} = k\}$, and

$$\sum_{r=1}^{m}\|C_{n_r,\tau_r}\|_{\infty}$$
$$= \sum_{\tau' \leq t}\sum_{n'=1}^{N} w_j(n',\tau';t)\,\|C_{n',\tau'}\|_{\infty}\,\mathbf{1}\{k_{n',\tau'} = k\}. \quad (29)$$

By definition of the protocol-induced multiplicity $\rho_{n,\tau}$ (total weight across all estimators/statistics), we have $w_j(n,\tau;t) \leq \rho_{n,\tau}$ for all (n,τ). Hence,

$$\sum_{r=1}^{m}\|C_{n_r,\tau_r}\|_{\infty} \leq \sum_{\tau' \leq t}\sum_{n'=1}^{N} \rho_{n',\tau'}\,\|C_{n',\tau'}\|_{\infty}\,\mathbf{1}\{k_{n',\tau'} = k\}$$
$$\leq \Gamma_{\mathrm{eff},k}, \qquad (30)$$

where $\Gamma_{\mathrm{eff},k}$ is the (arm-wise) effective corruption defined by $\Gamma_{\mathrm{eff},k} \triangleq \sum_{\tau=1}^{T}\sum_{n=1}^{N} \rho_{n,\tau}\|C_{n,\tau}\|_{\infty}\mathbf{1}\{k_{n,\tau} = k\}$. Combining (28)–(30) yields

$$\left\|\frac{1}{m}\sum_{r=1}^{m}(\widetilde{R}_{n_r,\tau_r} - R_{n_r,\tau_r})\right\|_{\infty} \leq \frac{\Gamma_{\mathrm{eff},k}}{m}. \qquad (31)$$

D. Step 3: Concentration of the Stochastic Term (Coordinate-Wise Hoeffding)

Fix a coordinate $\ell \in [d]$ and consider the scalar random variables

$$X_r \triangleq R_{n_r,\tau_r}^{(\ell)} \in [0,1], \qquad r = 1,\dots,m.$$

Conditioned on the event $\{k_{n_r,\tau_r} = k\}$ for all r (which holds by construction of $\mathcal{I}_{j,k}(t)$), each X_r is distributed according to the ℓ-th coordinate marginal of $\mathcal{D}_k$ and has mean $\mu_k^{(\ell)}$. Moreover, since clean rewards are independent across agents

and times conditioned on the chosen arms, the collection $\{X_r\}_{r=1}^{m}$ is independent conditioned on $\mathcal{I}_{j,k}(t)$ (even though $\mathcal{I}_{j,k}(t)$ is random and adaptively chosen). Therefore, Hoeffding's inequality applies conditionally on $\mathcal{I}_{j,k}(t)$: for any $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{m}\sum_{r=1}^{m}(X_r - \mu_k^{(\ell)})\right| \geq \varepsilon \;\Big|\; \mathcal{I}_{j,k}(t)\right) \leq 2\exp(-2m\varepsilon^2).$$

Let $\delta' \triangleq \delta/(dKNT)$ and set

$$\varepsilon(m) \triangleq \sqrt{\frac{\log(2/\delta')}{2m}} = \sqrt{\frac{\log(2dKNT/\delta)}{2m}}.$$

Then the conditional probability above is at most δ'. Removing the conditioning gives

$$\Pr\left(\left|\frac{1}{m}\sum_{r=1}^{m}(R_{n_r,\tau_r}^{(\ell)} - \mu_k^{(\ell)})\right| \geq \sqrt{\frac{\log(2dKNT/\delta)}{2m}}\right) \leq \delta'. \qquad (32)$$

E. Step 4: Union Bound Over (j,k,t,ℓ)

There are at most N estimators j, K arms, T times, and d coordinates. Applying (32) and a union bound over all (j,k,t,ℓ) shows that with probability at least $1 - dKNT \cdot \delta' = 1 - \delta$, simultaneously for all j,k,t and all $\ell \in [d]$,

$$\left|\frac{1}{m}\sum_{r=1}^{m}(R_{n_r,\tau_r}^{(\ell)} - \mu_k^{(\ell)})\right| \leq \sqrt{\frac{\log(2dKNT/\delta)}{2m}}.$$

Equivalently,

$$\left\|\frac{1}{m}\sum_{r=1}^{m}(R_{n_r,\tau_r} - \mu_k)\right\|_{\infty} \leq \sqrt{\frac{\log(2dKNT/\delta)}{2m}}. \qquad (33)$$

F. Step 5: Combine the Bounds

Plugging (33) and (31) into (25) yields, on the same high-probability event,

$$\|\widehat{\mu}_{j,k}(t) - \mu_k\|_{\infty} \leq \sqrt{\frac{\log(2dKNT/\delta)}{2m}} + \frac{\Gamma_{\mathrm{eff},k}}{m}.$$

Finally, replacing m by $\max\{1,m\}$ accommodates the case $m = 0$ discussed in Step 0, giving (9). $\qquad\square$

APPENDIX C
PROOF OF LEMMA 3

Proof. We first check that x^{max} is well-defined and belongs to $[0,1]^d$, then show it is (coordinate-wise) an upper bound for $\mathcal{C}$, and finally use monotonicity of ϕ to identify the supremum.

A. Step 1: Well-Definedness of x^{max} and Basic Properties

Fix any coordinate $j \in [d]$. Since $\mathcal{C} \subseteq [0,1]^d$, the set

$$\mathcal{C}_j \triangleq \{x_j : x \in \mathcal{C}\}$$

is nonempty (unless $\mathcal{C} = \emptyset$; see the remark at the end) and is bounded above by 1. Hence the supremum

$$x_j^{\mathrm{max}} \triangleq \sup \mathcal{C}_j$$

satisfies $0 \leq x_j^{\mathrm{max}} \leq 1$. Therefore $x^{\mathrm{max}} \in [0,1]^d$.

B. Step 2: $x^{\max}$ Dominates Every Point in $\mathcal{C}$

By definition of supremum, for every $x \in \mathcal{C}$ and every coordinate j,

$$x_j \leq \sup\{z_j : z \in \mathcal{C}\} = x_j^{\max}.$$

Thus, $x \preceq x^{\max}$ holds for all $x \in \mathcal{C}$.

C. Step 3: Upper Bound on the Objective: $\sup_{x \in \mathcal{C}} \phi(x) \leq \phi(x^{\max})$

Since ϕ is coordinate-wise nondecreasing, $x \preceq x^{\max}$ implies $\phi(x) \leq \phi(x^{\max})$. Applying this to every $x \in \mathcal{C}$ yields

$$\sup_{x \in \mathcal{C}} \phi(x) \leq \phi(x^{\max}).$$

D. Step 4: Achievability of $\phi(x^{\max})$ as the Supremum

We show that values of ϕ on $\mathcal{C}$ can get arbitrarily close to $\phi(x^{\max})$.

Fix $\varepsilon > 0$. For each coordinate $j \in [d]$, by the definition of $x_j^{\max}$ as a supremum, there exists $x^{(j)} \in \mathcal{C}$ such that

$$x_j^{(j)} \geq x_j^{\max} - \varepsilon.$$

Define the coordinate-wise maximum (join) $y \triangleq \bigvee_{j=1}^d x^{(j)}$ where $y_\ell \triangleq \max_{j \in [d]} x_\ell^{(j)} \ \forall \ell \in [d]$. Then, for each coordinate ℓ,

$$y_\ell \geq x_\ell^{(\ell)} \geq x_\ell^{\max} - \varepsilon,$$

so

$$x^{\max} - \varepsilon \mathbf{1} \preceq y \preceq \mathbf{1}.$$

It remains to justify $y \in \mathcal{C}$. This is where *upper-closedness* is used. Pick any $x^{(1)} \in \mathcal{C}$ (one of the selected points). By construction, $x^{(1)} \preceq y \preceq \mathbf{1}$. Since $\mathcal{C}$ is upper-closed, $x^{(1)} \in \mathcal{C}$ and $x^{(1)} \preceq y \preceq \mathbf{1}$ imply $y \in \mathcal{C}$.

Now apply monotonicity of ϕ:

$$\phi(y) \geq \phi(x^{\max} - \varepsilon \mathbf{1}).$$

Taking supremum over $x \in \mathcal{C}$ gives

$$\sup_{x \in \mathcal{C}} \phi(x) \geq \phi(y) \geq \phi(x^{\max} - \varepsilon \mathbf{1}).$$

Finally, since ϕ is L-Lipschitz under $\|\cdot\|_\infty$ on $[0,1]^d$,

$$\left| \phi(x^{\max}) - \phi(x^{\max} - \varepsilon \mathbf{1}) \right| \leq L \|\varepsilon \mathbf{1}\|_\infty = L\varepsilon,$$

so

$$\phi(x^{\max} - \varepsilon \mathbf{1}) \geq \phi(x^{\max}) - L\varepsilon.$$

Therefore,

$$\sup_{x \in \mathcal{C}} \phi(x) \geq \phi(x^{\max}) - L\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, letting $\varepsilon \downarrow 0$ yields

$$\sup_{x \in \mathcal{C}} \phi(x) \geq \phi(x^{\max}).$$

E. Step 5: Combine

Together with Step 3, we conclude $\sup_{x \in \mathcal{C}} \phi(x) = \phi(x^{\max})$.

a) Specialization to ℓ_∞ rectangles: Let $\mathcal{C} = \{x : \|x - \widehat{\mu}\|_\infty \leq \beta\} \cap [0,1]^d$. Then for each coordinate j,

$$\sup\{x_j : x \in \mathcal{C}\} = \min\{1, \widehat{\mu}_j + \beta\},$$

because the ℓ_∞ constraint allows increasing coordinate j up to $\widehat{\mu}_j + \beta$, and the intersection with $[0,1]^d$ caps it at 1. Hence

$$x^{\max} = \Pi_{[0,1]^d}(\widehat{\mu} + \beta \mathbf{1}),$$

and the first part implies

$$\sup_{x \in \mathcal{C}} \phi(x) = \phi\big(\Pi_{[0,1]^d}(\widehat{\mu} + \beta \mathbf{1})\big).$$

b) Remark (empty set): If $\mathcal{C} = \emptyset$, then $\sup_{x \in \mathcal{C}} \phi(x) = -\infty$ by convention, while $x^{\max}$ is undefined. In our application, confidence sets are nonempty (they at least intersect $[0,1]^d$), so this case does not arise. $\qquad \square$

APPENDIX D
PROOF OF THEOREM 1

Proof. Fix $\delta \in (0,1)$. Let $\mathcal{E}$ denote the "good" event on which the vector mean concentration bound in Lemma 2 holds simultaneously for all estimators j, all arms k, and all times t; namely,

$$\left\| \widehat{\mu}_{j,k}(t) - \mu_k \right\|_\infty \leq \beta_{j,k}(t) \qquad \forall (j,k,t), \quad (34)$$

where $\beta_{j,k}(t)$ is the radius used by the protocol (in particular, it upper bounds the stochastic fluctuation term and the bias term induced by the arm-wise effective corruption $\Gamma_{\text{eff},k}$). By Lemma 2, $\Pr(\mathcal{E}) \geq 1 - \delta$.

Throughout the proof, we condition on $\mathcal{E}$.

A. Step 1: UCB Indices are Valid Upper Bounds on $\theta_k = \phi(\mu_k)$

A UCB-type protocol maintains, for each estimator j and arm k at time t, an empirical mean $\widehat{\mu}_{j,k}(t)$ and uses the optimistic index

$$U_{j,k}(t) \triangleq \phi\Big(\Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big)\Big). \quad (35)$$

On $\mathcal{E}$, (34) implies coordinate-wise

$$\mu_k \preceq \widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}.$$

Since $\mu_k \in [0,1]^d$, applying coordinate-wise projection preserves dominance:

$$\mu_k \preceq \Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big).$$

Because ϕ is coordinate-wise nondecreasing, for all j, k, t,

$$\theta_k = \phi(\mu_k) \leq \phi\Big(\Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big)\Big) = U_{j,k}(t). \quad (36)$$

B. Step 2: Instantaneous Team Regret is Controlled by the Chosen Radius

Fix an agent-round decision (n, t), and let $j = j(n, t)$ be the estimator used to form the arm index at that decision. Let $k_{n,t}$ be the arm selected by the protocol. Since the protocol is UCB-type,

$$k_{n,t} \in \arg\max_{k \in [K]} U_{j,k}(t).$$

Let $k^* \in \arg\max_k \theta_k$ be an optimal arm. Then

$$U_{j,k_{n,t}}(t) \geq U_{j,k^*}(t) \geq \theta_{k^*},$$

where the last inequality uses (36). Therefore the instantaneous scalarized regret satisfies

$$\theta_{k^*} - \theta_{k_{n,t}} \leq U_{j,k_{n,t}}(t) - \theta_{k_{n,t}}. \tag{37}$$

We now bound $U_{j,k}(t) - \theta_k$ in terms of $\beta_{j,k}(t)$. Using (35), Lipschitzness of ϕ under $\|\cdot\|_\infty$, and nonexpansiveness of projection under $\|\cdot\|_\infty$,

$$
\begin{aligned}
U_{j,k}(t) - \theta_k &= \phi\Big(\Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big)\Big) - \phi(\mu_k) \\
&\leq L\Big\|\Pi_{[0,1]^d}\big(\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1}\big) - \mu_k\Big\|_\infty \\
&\leq L\Big\|\widehat{\mu}_{j,k}(t) + \beta_{j,k}(t)\mathbf{1} - \mu_k\Big\|_\infty \\
&\leq L\big\|\widehat{\mu}_{j,k}(t) - \mu_k\big\|_\infty + L\,\beta_{j,k}(t) \leq 2L\,\beta_{j,k}(t),
\end{aligned}
\tag{38}
$$

where the last step uses (34). Combining (37) and (38) yields, for every agent-round (n, t),

$$\theta_{k^*} - \theta_{k_{n,t}} \leq 2L\,\beta_{j(n,t),\,k_{n,t}}(t). \tag{39}$$

C. Step 3: Summing Over all NT Pulls and Bounding the Stochastic Part

Summing (39) over all (n, t) gives

$$\text{Reg}_\phi^{\text{team}}(T) \leq 2L \sum_{t=1}^{T} \sum_{n=1}^{N} \beta_{j(n,t),\,k_{n,t}}(t). \tag{40}$$

We upper bound the RHS using the explicit form of the radii (as in Lemma 2):

$$\beta_{j,k}(t) \leq \sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, |\mathcal{I}_{j,k}(t)|\}}} + \frac{\Gamma_{\text{eff},k}}{\max\{1, |\mathcal{I}_{j,k}(t)|\}}.$$

For the stochastic term, define $m \mapsto a(m) \triangleq \sqrt{\log(2dKNT/\delta)/(2m)}$. Every time an estimator uses an arm-k observation as a unit-weight sample, the corresponding counter $|\mathcal{I}_{j,k}(\cdot)|$ increases by 1. Hence, when we sum $a(|\mathcal{I}_{j,k}(t)|)$ over all agent-rounds in which arm k is played (and the chosen estimator uses its own observation), the

denominators can be pessimistically upper bounded by the harmonic sequence $1, 2, \ldots, N_k^{\text{team}}(T)$. Therefore,

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{n=1}^{N} &\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, |\mathcal{I}_{j(n,t),k_{n,t}}(t)|\}}} \\
&\leq \sqrt{\frac{\log(2dKNT/\delta)}{2}} \sum_{k=1}^{K} \sum_{r=1}^{N_k^{\text{team}}(T)} \frac{1}{\sqrt{r}} \\
&\leq \sqrt{\frac{\log(2dKNT/\delta)}{2}} \sum_{k=1}^{K} 2\sqrt{N_k^{\text{team}}(T)} \\
&\leq 2\sqrt{\frac{\log(2dKNT/\delta)}{2}} \sqrt{K \sum_{k=1}^{K} N_k^{\text{team}}(T)} \\
&= 2\sqrt{\frac{\log(2dKNT/\delta)}{2}} \sqrt{KNT}, \tag{41}
\end{aligned}
$$

where we used $\sum_{r=1}^{m} r^{-1/2} \leq 2\sqrt{m}$ and Cauchy–Schwarz.

D. Step 4: Bounding the Corruption Part via a Harmonic Sum

Similarly, for the corruption term,

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{n=1}^{N} \frac{\Gamma_{\text{eff},k_{n,t}}}{\max\{1, |\mathcal{I}_{j(n,t),k_{n,t}}(t)|\}} &\leq \sum_{k=1}^{K} \Gamma_{\text{eff},k} \sum_{r=1}^{N_k^{\text{team}}(T)} \frac{1}{r} \\
&\leq \sum_{k=1}^{K} \Gamma_{\text{eff},k} \,\log\big(1 + N_k^{\text{team}}(T)\big) \\
&\leq \log(1 + NT) \sum_{k=1}^{K} \Gamma_{\text{eff},k} = \Gamma_{\text{eff}} \log(1 + NT), \tag{42}
\end{aligned}
$$

where we used $\sum_{r=1}^{m} \frac{1}{r} \leq \log(1 + m)$ and $\sum_k \Gamma_{\text{eff},k} = \Gamma_{\text{eff}}$.

E. Step 5: Combine

Plugging (41) and (42) into (40) yields

$$
\begin{aligned}
&\text{Reg}_\phi^{\text{team}}(T) \\
&\leq 2L\left(2\sqrt{\frac{\log(2dKNT/\delta)}{2}}\sqrt{KNT} + \Gamma_{\text{eff}} \log(1 + NT)\right).
\end{aligned}
\tag{43}
$$

Absorbing numerical factors into a universal constant $c > 0$ gives (11). $\qquad\square$

APPENDIX E
PROOF OF THEOREM 2

Proof. We prove the claim for the scalar case $d = 1$ with $\phi(x) = x$ (so $\text{Reg}_\phi^{\text{team}}(T)$ is the standard team regret). This immediately implies the stated lower bound for general coordinate-wise nondecreasing L-Lipschitz scalarizations by scaling constants (e.g., by taking instances where only one coordinate varies and using L-Lipschitzness).

Let $H \triangleq NT$ denote the total number of *agent-round pulls*. Throughout, we work under the (S1) model: at the end of each round t, every agent broadcasts its raw observation, and under *append-all* each agent appends every received raw sample as a fresh sample into its local dataset. Thus, for any agent-round

observation (n, t), the realized (possibly corrupted) feedback is present in the local history of *all* N agents from the end of round t onward.

The proof has two components: (i) the classical stochastic minimax term $\Omega(\sqrt{KH})$ (already present when $\Gamma = 0$), and (ii) an additive corruption term $\Omega(N\Gamma)$ induced by (S1)-replication. Combining them via $\max\{a, b\} \geq \frac{1}{2}(a + b)$ yields the stated form with universal constants.

A. Step 1: The Stochastic Term $\Omega(\sqrt{KH})$

Consider the clean case $\Gamma = 0$. It is standard (see, e.g., [2], [3]) that for any (possibly adaptive) team policy making H total pulls there exists a K-armed Bernoulli instance such that

$$\mathbb{E}\big[\mathrm{Reg}^{\mathrm{team}}(T)\big] \geq c_{\mathrm{sto}}\sqrt{KH} \tag{44}$$

for a universal constant $c_{\mathrm{sto}} > 0$. (One convenient construction is: pick a uniformly random optimal arm $k^\star \sim \mathrm{Unif}([K])$, set $\mu_{k^\star} = \frac{1}{2} + \Delta$ and $\mu_k = \frac{1}{2}$ for $k \neq k^\star$, then choose $\Delta \asymp \sqrt{K/H}$ and apply a change-of-measure/Fano argument.)

B. Step 2: A Corruption Term $\Omega(N\Gamma)$ Under (S1) Append-All

We now construct a family of two-arm instances and an adversary which, under (S1) append-all, makes the team essentially unable to identify the better arm unless the team pays regret linear in $N\Gamma$.

Two-point instance. Let $K = 2$ and consider two bandit instances $\mathcal{I}_+$ and $\mathcal{I}_-$:

$$\mathcal{I}_+ : \quad \mu_1 = \tfrac{1}{2} + \Delta, \ \mu_2 = \tfrac{1}{2}, \qquad \mathcal{I}_- : \quad \mu_1 = \tfrac{1}{2}, \ \mu_2 = \tfrac{1}{2} + \Delta,$$

with $\Delta \in (0, 1/8]$ to be chosen later. Rewards are Bernoulli with these means.

A coupling-style corruption strategy. We describe an adversary that, in either world $\mathcal{I}_+$ or $\mathcal{I}_-$, corrupts *only the samples from the currently-better arm* so that the *observed* reward distribution of that arm is pushed down to $\mathrm{Bernoulli}(\frac{1}{2})$. Concretely, suppose the clean reward is $R \in \{0, 1\}$ from a better arm with mean $\frac{1}{2} + \Delta$. Let the adversary produce the corrupted observation $\widetilde{R}$ by

$$\widetilde{R} = R \cdot Z, \qquad Z \sim \mathrm{Bernoulli}\left(\frac{\frac{1}{2}}{\frac{1}{2} + \Delta}\right) \text{ independent of } R.$$

Then $\mathbb{P}(\widetilde{R} = 1) = (\frac{1}{2} + \Delta) \cdot \frac{\frac{1}{2}}{\frac{1}{2} + \Delta} = \frac{1}{2}$, i.e., $\widetilde{R} \sim \mathrm{Bernoulli}(\frac{1}{2})$. This transformation flips a 1 to 0 with probability

$$\mathbb{P}(\widetilde{R} \neq R) = \mathbb{P}(R = 1) \cdot \mathbb{P}(Z = 0 \mid R = 1)$$
$$= (\tfrac{1}{2} + \Delta)\left(1 - \frac{\frac{1}{2}}{\frac{1}{2} + \Delta}\right) = \Delta. \tag{45}$$

Hence the expected *per-sample* corruption magnitude is $\mathbb{E}[|\widetilde{R} - R|] = \Delta$. (For samples from the worse arm with mean $\frac{1}{2}$, the adversary does nothing.)

Budget accounting under (S1) append-all. Let M be the total number of pulls of the better arm across all agents and

all rounds. Under the above strategy, the *physical* corruption budget used satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=1}^{N}|C_{n,t}|\right] = \mathbb{E}\left[\sum_{\text{better-arm pulls}}|\widetilde{R} - R|\right] = \Delta \cdot \mathbb{E}[M].$$

Now use the (S1) append-all property: each corrupted sample (n, t) is broadcast and appended by *all* N agents, so it appears N times in the *joint* team transcript (the collection of all agents' local histories that drive their future actions). Equivalently, one unit of physical corruption on a single agent-round observation induces N units of "statistical damage" across the team's decision rules. This is exactly the replication phenomenon quantified by $\rho_{n,t} = N$ in Lemma 1.

Formally, define the *replicated* (or effective) corruption cost of this adversary by

$$B \triangleq N\sum_{t=1}^{T}\sum_{n=1}^{N}|C_{n,t}|.$$

Then $\mathbb{E}[B] = N\Delta\,\mathbb{E}[M]$.

Choose

$$\Delta \triangleq \min\left\{\frac{1}{8}, \ \frac{N\Gamma}{8H}\right\}. \tag{46}$$

Since $M \leq H$, we have $\mathbb{E}[B] \leq N\Delta H \leq N \cdot \frac{N\Gamma}{8H} \cdot H = \frac{N^2\Gamma}{8}$, and in particular $\mathbb{E}\big[\sum_{t,n}|C_{n,t}|\big] \leq \Gamma/8$ whenever $\Delta = \frac{N\Gamma}{8H}$. By Markov's inequality,

$$\mathbb{P}\left(\sum_{t,n}|C_{n,t}| \leq \Gamma\right) \geq 1 - \frac{\mathbb{E}[\sum_{t,n}|C_{n,t}|]}{\Gamma} \geq \frac{7}{8}.$$

So with constant probability the strategy is feasible under the budget Γ.

Indistinguishability. Under $\mathcal{I}_+$, the adversary pushes the observed rewards of arm 1 down to $\mathrm{Bernoulli}(\frac{1}{2})$ and leaves arm 2 at $\mathrm{Bernoulli}(\frac{1}{2})$. Under $\mathcal{I}_-$, it pushes the observed rewards of arm 2 down to $\mathrm{Bernoulli}(\frac{1}{2})$ and leaves arm 1 at $\mathrm{Bernoulli}(\frac{1}{2})$. Therefore, *conditioned on the event that the budget constraint is satisfied*, the full observed transcript (actions, broadcasts, and received observations) has the same distribution under $\mathcal{I}_+$ and $\mathcal{I}_-$. Hence no (possibly randomized) cooperative policy can identify the better arm with probability exceeding $1/2$ in both instances.

Regret consequence. Let π be any cooperative protocol under (S1) append-all. By the above indistinguishability, for the uniform prior over $\{\mathcal{I}_+, \mathcal{I}_-\}$ we have

$$\mathbb{E}\big[\mathrm{Reg}^{\mathrm{team}}(T)\big]$$
$$\geq \frac{1}{2} \cdot \Delta \cdot \mathbb{E}\big[\#\{\text{suboptimal pulls across all } H \text{ pulls}\}\big]$$
$$\geq c'\Delta H \tag{47}$$

for a universal constant $c' > 0$ (e.g., $c' = 1/8$ suffices after conditioning on budget-feasibility with probability $\geq 7/8$). Using $\Delta = \frac{N\Gamma}{8H}$ (the non-saturated case in (46)), we get

$$\mathbb{E}\big[\mathrm{Reg}^{\mathrm{team}}(T)\big] \geq c_{\mathrm{cor}}\,N\Gamma \tag{48}$$

for a universal constant $c_{\text{cor}} > 0$, as long as $N\Gamma \leq H$ (otherwise the trivial bound $\text{Reg}^{\text{team}}(T) \leq H$ dominates and one can replace $N\Gamma$ by $\min\{N\Gamma, H\}$).

C. Step 3: Combine

From (44) and (48), the worst-case expected team regret under (S1) append-all satisfies

$$
\begin{aligned}
\sup_{\text{instances}} \ \inf_{\text{adv: } \sum_{t,n} |C_{n,t}| \leq \Gamma} & \ \mathbb{E}[\text{Reg}^{\text{team}}(T)] \\
& \geq \ \max\left\{ c_{\text{sto}}\sqrt{KH}, \ c_{\text{cor}}N\Gamma \right\} \\
& \geq \ \tfrac{1}{2}c_{\text{sto}}\sqrt{KH} + \tfrac{1}{2}c_{\text{cor}}N\Gamma. \quad (49)
\end{aligned}
$$

Absorbing the factor $1/2$ into constants yields the claim with $H = NT$. $\qquad\square$

APPENDIX F
PROOF OF LEMMA 4

Proof. Fix an arm $k \in [K]$. Let $\mathcal{T}_k^{\text{ver}}(t)$ be the set of verified agent-round indices up to time t in which arm k was pulled $\mathcal{T}_k^{\text{ver}}(t) \triangleq \{(n, \tau) : 1 \leq \tau \leq t, \ k_{n,\tau} = k, \ V_{n,\tau} = 1\}$, $H_k^{\text{ver}}(t) = |\mathcal{T}_k^{\text{ver}}(t)|$. Define the verified-only empirical mean

$$
\widehat{\mu}_k^{\text{ver}}(t) \triangleq \frac{1}{\max\{1, H_k^{\text{ver}}(t)\}} \sum_{(n,\tau) \in \mathcal{T}_k^{\text{ver}}(t)} R_{n,\tau}.
$$

A. Step 1: Verified Samples are i.i.d. From $\mathcal{D}_k$

For each agent-round (n, t), the verification decision $V_{n,t}$ is $\sigma(\mathcal{H}_{n,t-1})$-measurable and is made *before observing* the reward feedback at time t. Hence $V_{n,t}$ is independent of the realized reward $R_{n,t}$ conditional on the chosen arm $k_{n,t}$. Therefore, conditional on the event $\{k_{n,t} = k, \ V_{n,t} = 1\}$, the random vector $R_{n,t}$ is distributed as $\mathcal{D}_k$. Moreover, across distinct agent-rounds, the reward draws are independent (conditioned on actions). Consequently, for any fixed t, the multiset $\{R_{n,\tau} : (n, \tau) \in \mathcal{T}_k^{\text{ver}}(t)\}$ consists of $H_k^{\text{ver}}(t)$ independent samples from $\mathcal{D}_k$.

B. Step 2: Coordinate-Wise Hoeffding and Union Bound Over (k, j, s)

Fix a coordinate $j \in [d]$. For any integer $s \geq 1$, let

$$
\bar{R}_{k,j}(s) \triangleq \frac{1}{s} \sum_{\ell=1}^{s} X_{k,j,\ell},
$$

where $X_{k,j,\ell} \in [0, 1]$ are i.i.d. with mean $\mu_k^{(j)}$ (a generic copy of the verified samples of arm k in coordinate j). Hoeffding's inequality gives, for any $\varepsilon > 0$,

$$
\Pr\left(|\bar{R}_{k,j}(s) - \mu_k^{(j)}| \geq \varepsilon \right) \leq 2e^{-2s\varepsilon^2}.
$$

Set

$$
\varepsilon(s) \triangleq \sqrt{\frac{\log(2dKNT/\delta)}{2s}}.
$$

Then

$$
\begin{aligned}
& \Pr\left(|\bar{R}_{k,j}(s) - \mu_k^{(j)}| \geq \varepsilon(s) \right) \\
& \leq 2\exp\left(-2s \cdot \frac{\log(2dKNT/\delta)}{2s} \right) = \frac{\delta}{dKNT}. \quad (50)
\end{aligned}
$$

Applying a union bound over all arms $k \in [K]$, coordinates $j \in [d]$, and sample sizes $s \in \{1, 2, \ldots, NT\}$ yields an event $\mathcal{E}$ with $\Pr(\mathcal{E}) \geq 1 - \delta$ such that on $\mathcal{E}$, for all $k \in [K]$, $j \in [d]$, $s \in [NT]$,

$$
|\bar{R}_{k,j}(s) - \mu_k^{(j)}| \leq \sqrt{\frac{\log(2dKNT/\delta)}{2s}}. \quad (51)
$$

C. Step 3: Translate (51) to the Random-Time Estimator $\widehat{\mu}_k^{\text{ver}}(t)$

Fix any k and t. If $H_k^{\text{ver}}(t) = s \geq 1$, then by Step 1, $\widehat{\mu}_k^{\text{ver}}(t)$ is exactly the empirical mean of s i.i.d. samples from $\mathcal{D}_k$; thus (51) implies simultaneously for all $j \in [d]$,

$$
|\widehat{\mu}_k^{\text{ver}}(t)^{(j)} - \mu_k^{(j)}| \leq \sqrt{\frac{\log(2dKNT/\delta)}{2s}}.
$$

Equivalently,

$$
\|\widehat{\mu}_k^{\text{ver}}(t) - \mu_k\|_\infty \leq \sqrt{\frac{\log(2dKNT/\delta)}{2\, H_k^{\text{ver}}(t)}}.
$$

(When $H_k^{\text{ver}}(t) = 0$, the verified-only estimate is undefined/unused for certification; in implementations one can set the certificate radius to $+\infty$ and treat the interval as vacuous. If you keep the $\max\{1, H_k^{\text{ver}}(t)\}$ convention, the displayed bound remains valid for all t after defining $\widehat{\mu}_k^{\text{ver}}(t)$ arbitrarily when $H_k^{\text{ver}}(t) = 0$ and taking the confidence interval as vacuous.)

D. Step 4: Apply Lipschitzness of ϕ

Since ϕ is L-Lipschitz under $\|\cdot\|_\infty$,

$$
\begin{aligned}
& |\phi(\widehat{\mu}_k^{\text{ver}}(t)) - \phi(\mu_k)| \\
& \leq L\|\widehat{\mu}_k^{\text{ver}}(t) - \mu_k\|_\infty \\
& \leq L\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, H_k^{\text{ver}}(t)\}}}. \quad (52)
\end{aligned}
$$

Recalling $\theta_k = \phi(\mu_k)$ completes the proof. $\qquad\square$

APPENDIX G
PROOF OF THEOREM 3

Proof. We describe an explicit cooperative protocol (using (S3) recommendation-only messages, with optional verified certificates) and then prove the two claims.

A. Algorithm: S3 with Certified Override (S3-Cert)

Fix confidence level $\delta \in (0, 1)$. The protocol maintains two tracks:

(A) Always-valid baseline track. All agents run a fixed recommendation-only coordination protocol Comm (e.g., an epoch-based coordination rule) whose arm-selection rule is UCB-type and uses only local (possibly corrupted) observations. Each agent broadcasts only an arm recommendation each round (as in (S3)). Let $\text{Reg}_{\text{comm}}^{\text{clean}}(K, N, T)$ denote the (high-probability) regret overhead of the same protocol in the clean case. (Concretely, in our paper this is the "clean-case coordination" term of Comm.)

(B) Certified-override track. The team spends up to ν verifications to create *verified-only* statistics and a *certificate-based stopping rule.* Let $H_k^{\mathrm{ver}}(t)$ and $\widehat{\mu}_k^{\mathrm{ver}}(t)$ be the verified-only count and empirical mean (defined in the main text) aggregated over all agents up to time t. At each verified agent-round (n,t) with $V_{n,t} = 1$, agent n broadcasts the tuple

$$(k_{n,t}, R_{n,t}, V_{n,t} = 1),$$

so that all agents can update the common verified-only statistics[3]. For each arm k and time t, define the certified radius $\varepsilon_k(t) \triangleq L\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, H_k^{\mathrm{ver}}(t)\}}}$, $\mathrm{LCB}_k(t) \triangleq \phi(\widehat{\mu}_k^{\mathrm{ver}}(t)) - \varepsilon_k(t)$, $\mathrm{UCB}_k(t) \triangleq \phi(\widehat{\mu}_k^{\mathrm{ver}}(t)) + \varepsilon_k(t)$. The protocol declares a *certified winner* at the first time

$$t_{\mathrm{cert}} \triangleq \inf\left\{t : \exists k \in [K] \text{ s.t. } \mathrm{LCB}_k(t) \geq \max_{k' \neq k} \mathrm{UCB}_{k'}(t)\right\}.$$

If such $t_{\mathrm{cert}} \leq T$ occurs, all agents commit from time $t_{\mathrm{cert}} + 1$ onward to the certified winner

$$\widehat{k} \in \arg\max_{k \in [K]} \mathrm{LCB}_k(t_{\mathrm{cert}}).$$

If no certified winner appears by time T, agents simply keep following the baseline track.

By construction, (S3-Cert) uses recommendation-only sharing on unverified rounds, and uses at most ν verified broadcasts. We now prove the two parts.

B. Part (1): Always-Valid Robust Baseline

Consider any instance and any adaptive adversary satisfying the global budget (3). On rounds before commitment, the arm-selection decisions are exactly those of the baseline protocol Comm (possibly with *more accurate* observations on verified rounds, since verified feedback is clean). Thus, any regret upper bound proven for Comm under corruption continues to hold for (S3-Cert).

In particular, instantiate Comm as the recommendation-only robust-UCB-type protocol analyzed in our paper. Under (S3), no raw rewards are replicated across agents, so by Lemma 1 we have $\Gamma_{\mathrm{eff}} = \Gamma$. Applying Theorem 1 (team regret bound via effective corruption) to the baseline track yields that on an event of probability at least $1 - \delta$,

$$\mathrm{Reg}_\phi^{\mathrm{team}}(T) \leq \widetilde{O}\Big(L \cdot \mathrm{Reg}_{\mathrm{comm}}^{\mathrm{clean}}(K, N, T) + L\Gamma\Big),$$

where $\mathrm{Reg}_{\mathrm{comm}}^{\mathrm{clean}}(K, N, T)$ denotes the clean-case coordination overhead of the same protocol. Since the certified-override can only (i) switch to a single arm and (ii) uses only clean information to decide the switch, it cannot increase regret on the high-probability event on which the baseline bound holds. This proves Part (1).

---

[3]This broadcast occurs only on verified rounds, at most ν times, and is therefore the "optional verified certificate" channel.

C. Part (2): Verification-Driven Override and Γ-Independence

Assume $\Delta_{\min} \triangleq \min_{k \neq k^*} \Delta_k > 0$ and the verification budget satisfies (18). Let $\mathcal{E}_{\mathrm{ver}}$ be the high-probability event from Lemma 4 applied to the verified-only statistics: with probability at least $1 - \delta$, simultaneously for all arms k and all times t,

$$\left|\phi(\widehat{\mu}_k^{\mathrm{ver}}(t)) - \theta_k\right| \leq \varepsilon_k(t). \tag{53}$$

We condition on $\mathcal{E}_{\mathrm{ver}}$ for the remainder of the argument.

Step 1: Once verified counts are large enough, the certified winner is k^*. Let $s \triangleq \min_{k \in [K]} H_k^{\mathrm{ver}}(t)$ be the minimum verified count across arms at some time t. If s is large enough so that

$$\max_{k \in [K]} \varepsilon_k(t) \leq \frac{\Delta_{\min}}{4}, \tag{54}$$

then the certification rule must accept k^* at (or before) time t. Indeed, on $\mathcal{E}_{\mathrm{ver}}$ we have for the optimal arm k^*,

$$\begin{aligned}\mathrm{LCB}_{k^*}(t) &= \phi(\widehat{\mu}_{k^*}^{\mathrm{ver}}(t)) - \varepsilon_{k^*}(t) \\ &\geq \theta_{k^*} - 2\varepsilon_{k^*}(t) \geq \theta_{k^*} - 2\max_k \varepsilon_k(t),\end{aligned} \tag{55}$$

and for any suboptimal arm $k \neq k^*$,

$$\begin{aligned}\mathrm{UCB}_k(t) &= \phi(\widehat{\mu}_k^{\mathrm{ver}}(t)) + \varepsilon_k(t) \\ &\leq \theta_k + 2\varepsilon_k(t) \leq \theta_{k^*} - \Delta_{\min} + 2\max_k \varepsilon_k(t).\end{aligned} \tag{56}$$

If (54) holds, then for every $k \neq k^*$,

$$\begin{aligned}\mathrm{LCB}_{k^*}(t) &\geq \theta_{k^*} - \frac{\Delta_{\min}}{2} \\ &> \theta_{k^*} - \Delta_{\min} + \frac{\Delta_{\min}}{2} \geq \mathrm{UCB}_k(t),\end{aligned} \tag{57}$$

so $\mathrm{LCB}_{k^*}(t) \geq \max_{k \neq k^*} \mathrm{UCB}_k(t)$ and the rule certifies k^*.

Step 2: The verification budget threshold implies (54). Under the condition (18), it is possible to allocate verifications so that each arm receives

$$H_k^{\mathrm{ver}}(t) \geq \left\lfloor \frac{\nu}{K} \right\rfloor$$

e.g., round-robin verification across arms and agents, by some time $t \leq T$ (assuming $\nu \leq NT$; otherwise verification can cover all pulls and the claim is trivial). For such a time t, we have for all k,

$$\varepsilon_k(t) \leq L\sqrt{\frac{\log(2dKNT/\delta)}{2\lfloor\nu/K\rfloor}} \leq \frac{\Delta_{\min}}{4}$$

by choosing the universal constant c in (18) large enough (e.g., $c \geq 8$ suffices up to floor/ceiling effects). Thus (54) holds, so by Step 1 the certified winner is k^* and commitment occurs at some $t^* \leq T$.

Step 3: Regret decomposition and independence from Γ. Let t^* be the (random) commitment time. After time t^*, all agents play k^*, hence incur zero regret. Therefore,

$$\begin{aligned}\mathrm{Reg}_\phi^{\mathrm{team}}(T) &= \sum_{t=1}^{t^*}\sum_{n=1}^N \left(\theta_{k^*} - \theta_{k_{n,t}}\right) \\ &\leq \Delta_{\max} \cdot \sum_{t=1}^{t^*}\sum_{n=1}^N \mathbf{1}\{k_{n,t} \neq k^*\}.\end{aligned} \tag{58}$$

Split the pre-commitment pulls into (i) verified pulls and (ii) all other pulls:

$$\text{Reg}_\phi^{\text{team}}(T) \leq \Delta_{\max} \cdot \underbrace{\sum_{t \leq t^*} \sum_{n=1}^{N} V_{n,t}}_{\leq \nu}$$

$$+ \underbrace{\sum_{t \leq t^*} \sum_{n=1}^{N} (\theta_{k^*} - \theta_{k_{n,t}}) \mathbf{1}\{V_{n,t} = 0\}}_{(\star)}. \qquad (59)$$

The first term is at most $\nu\Delta_{\max}$.

For the second term $(\star)$, we use a protocol-agnostic upper bound: trivially, each summand is at most $\Delta_{\max}$, hence

$$(\star) \leq (NT)\Delta_{\max}.$$

On the other hand, before commitment the actions are generated by the underlying coordination protocol Comm, so the regret accumulated on unverified rounds is also upper bounded (up to logs) by the clean-case coordination overhead of Comm, i.e.,

$$(\star) \leq \widetilde{O}\Big(L \cdot \text{Reg}_{\text{comm}}^{\text{clean}}(K, N, T)\Big),$$

because the certification/commitment rule depends only on verified samples and does not invalidate the clean-case coordination accounting up to time t^*. Combining the two bounds yields

$$(\star) \leq \widetilde{O}\Big(L \cdot \text{Reg}_{\text{comm}}^{\text{clean}}(K, N, T) \wedge NT\Delta_{\max}\Big).$$

Putting everything together gives, on $\mathcal{E}_{\text{ver}}$,

$$\text{Reg}_\phi^{\text{team}}(T) \leq O(\nu\Delta_{\max})$$
$$+ \widetilde{O}\Big(L \cdot \text{Reg}_{\text{comm}}^{\text{clean}}(K, N, T) \wedge NT\Delta_{\max}\Big), \qquad (60)$$

and crucially, this bound contains *no* dependence on Γ because the commitment decision is based solely on verified (clean) samples. Finally, since $\Pr(\mathcal{E}_{\text{ver}}) \geq 1 - \delta$, the same statement holds with probability at least $1 - \delta$. This completes Part (2) and the proof. $\qquad\square$