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*—Adversarial corruption, cooperative multi-agent bandits, multi-objective bandits, information sharing, robust regret analysis, phase transitions, verified feedback**

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 $[K] \triangleq \{1,\dots,K\}$ 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 $t = 1,2,\dots,T$, 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.,

$$\sum\nolimits_{t=1}^{T}\sum\nolimits_{n=1}^{N} V_{n,t} \le \nu. \tag{1}$$

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.,

$$\sum\nolimits_{t=1}^{T}\sum\nolimits_{n=1}^{N} \|C_{n,t}\| \le \Gamma. \tag{3}$$

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., $x\preceq y \Rightarrow \phi(x)\le\phi(y)$) and L-Lipschitz under ℓ_∞ (i.e., $|\phi(x)-\phi(y)|\le L\|x-y\|_\infty$ 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 $k\ne k^*$, define the gap $\Delta_k\triangleq\theta_{k^*}-\theta_k>0$.

We measure team scalarization regret as follows,

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

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

[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 $\big(H_{n,k}(t), S_{n,k}(t), H^{\mathrm{ver}}_{n,k}(t), S^{\mathrm{ver}}_{n,k}(t)\big)_{k\in[K]}$, where $H_{n,k}(t) \triangleq \sum_{\tau\le t} \mathbf{1}\{k_{n,\tau} = k\}$ is the number of times agent n has pulled arm k up to round t, $S_{n,k}(t) \triangleq \sum_{\tau\le t: k_{n,\tau}=k, V_{n,\tau}=0} \widetilde{R}_{n,\tau}$ 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, $H^{\mathrm{ver}}_{n,k}(t) \triangleq \sum_{\tau\le t} \mathbf{1}\{k_{n,\tau} = k, V_{n,\tau} = 1\}$ is the number of *verified* pulls of arm k by agent n up to round t, $S^{\mathrm{ver}}_{n,k}(t) \triangleq \sum_{\tau\le t: k_{n,\tau}=k, V_{n,\tau}=1} R_{n,\tau}$ 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 $H_k(t) = \sum_{n=1}^N H_{n,k}(t), S_k(t) = \sum_{n=1}^N S_{n,k}(t), H^{\mathrm{ver}}_k(t) = \sum_{n=1}^N H^{\mathrm{ver}}_{n,k}(t), S^{\mathrm{ver}}_k(t) = \sum_{n=1}^N S^{\mathrm{ver}}_{n,k}(t)$, 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.

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*

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

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

$$\Gamma_{\mathrm{eff}} \triangleq \sum\nolimits_{t=1}^{T} \sum\nolimits_{n=1}^{N} \rho_{n,t} \|C_{n,t}\|_\infty, \tag{7}$$

and the arm-wise effective corruption

$$\Gamma_{\mathrm{eff},k} \triangleq \sum\nolimits_{t=1}^{T} \sum\nolimits_{n=1}^{N} \rho_{n,t}\, \|C_{n,t}\|_\infty\, \mathbf{1}\{k_{n,t} = k\}. \tag{8}$$

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,*

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

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 $\|\widetilde{R}_{n,t} - R_{n,t}\|_\infty \le \|C_{n,t}\|_\infty$. 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 $\Gamma_{\mathrm{eff},k}$. □

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 $x^{\max} \in [0, 1]^d$ by $x^{\max}_j \triangleq \sup\{x_j : x \in \mathcal{C}\}$. Then, $\sup_{x\in\mathcal{C}} \phi(x) = \phi(x^{\max})$. In particular, for ℓ_∞ rectangles $\mathcal{C} = \{x : \|x - \widehat{\mu}\|_\infty \le \beta\} \cap [0, 1]^d$, we have*

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

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*

$$\mathit{Reg}_{\phi}^{\mathrm{team}}(T) \le cL\Big(\sqrt{KNT\log(2dKNT/\delta)} + \Gamma_{\mathrm{eff}}\log(1+NT)\Big), \tag{11}$$

for a universal constant $c > 0$.

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

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

$$\mathit{Reg}_{\phi}^{\mathrm{team}}(T) \le \widetilde{O}\big(L\sqrt{KNT} + LN\Gamma\big). \tag{12}$$

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

$$\mathit{Reg}_{\phi}^{\mathrm{team}}(T) \le \widetilde{O}\big(L\sqrt{KNT} + L\Gamma\big). \tag{13}$$

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 $\sum_{t,n}|C_{n,t}| \le \Gamma$ such that*

$$\mathbb{E}\big[\mathit{Reg}_{\phi}^{\mathrm{team}}(T)\big] \ge c_1\sqrt{KNT} + c_2 N\Gamma \tag{14}$$

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. □

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

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

which averages only the clean reward vectors $R_{n,\tau}$ collected from verified pulls of arm k up to time t (and uses $\max\{1, H_k^{\mathrm{ver}}(t)\}$ to avoid division by zero when no verification has occurred yet).

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

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

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 $(M_{n,t}, \mathrm{cert}_{n,t})$ 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 $\mathrm{cert}_{n,t} = \big(\mathrm{LCB}_{n,t}, \mathrm{UCB}_{n,t}\big)$ 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 $\mathrm{LCB}_{n,t} = \phi\big(\widehat{\mu}_{M_{n,t}}^{\mathrm{ver}}(t)\big) - \varepsilon_{M_{n,t}}(t)$, $\mathrm{UCB}_{n,t} = \phi\big(\widehat{\mu}_{M_{n,t}}^{\mathrm{ver}}(t)\big) + \varepsilon_{M_{n,t}}(t)$, where the half-width $\varepsilon_k(t) \triangleq L\sqrt{\frac{\log(2dKNT/\delta)}{2\max\{1, H_k^{\mathrm{ver}}(t)\}}}$ 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 $\mathrm{cert}_{n,t}$ 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 $[\mathrm{LCB}_{n,t}, \mathrm{UCB}_{n,t}]$ 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 $H_k^{\mathrm{ver}}(t)$) 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)*,*

$$\mathit{Reg}_{\phi}^{\mathrm{team}}(T) \le \widetilde{O}\big(L \cdot \mathit{Reg}_{\mathrm{comm}}^{\mathrm{clean}}(K,N,T) + L\Gamma\big), \tag{17}$$

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

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

$$\nu \ge c\,K\frac{L^2}{\Delta_{\min}^2}\log\Big(\frac{2dKNT}{\delta}\Big), \tag{18}$$

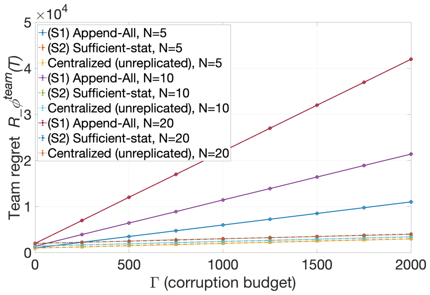

Fig. 1. Team regret versus corruption budget Γ for (S1) and (S2): (S1) scales like $\widetilde{O}(\sqrt{KNT}+N\Gamma)$ while (S2) scales like $\widetilde{O}(\sqrt{KNT}+\Gamma)$.

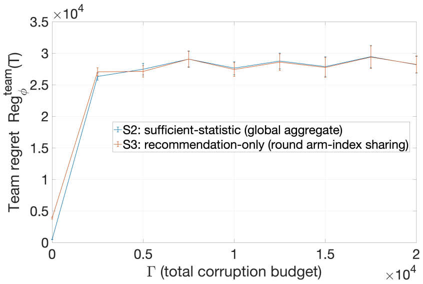

Fig. 2. Team regret versus Γ under (S2) and (S3): (S3) preserves the unamplified $O(\Gamma)$ term while trading off a clean-case coordination overhead.

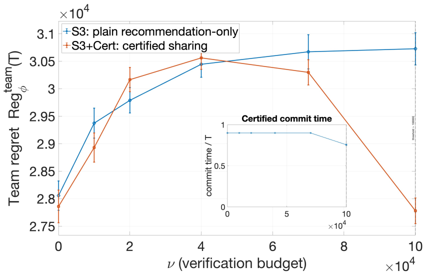

Fig. 3. Team regret versus verification budget ν under (S3) with/without certified sharing in high-corruption regime $\Gamma = \Theta(NT)$: sharp improvement once $\nu \gtrsim KL^2\Delta_{\min}^{-2}\log(dKNT)$.

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

$$\begin{aligned} \mathit{Reg}_\phi^{\mathrm{team}}(T) &\le O(\nu\Delta_{\max}) \\ &+ \widetilde{O}\big(L\cdot \mathit{Reg}_{\mathrm{comm}}^{\mathrm{clean}}(K,N,T)\wedge NT\Delta_{\max}\big), \end{aligned} \tag{19}$$

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 $\mathrm{Reg}_{\mathrm{comm}}^{\mathrm{clean}}(K,N,T)$. Moreover, once $\nu \gtrsim \tilde{\Theta}(KL^2/\Delta_{\min}^2)$, 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 $\varepsilon_k(t) \le \Delta_{\min}/4$ once $H_k^{\mathrm{ver}}(t)$ crosses threshold. By Lemma 4, any certified suboptimal arm $k \ne k^*$ then satisfies $\mathrm{UCB}_k^{\mathrm{ver}}(t) \le \theta_k + \Delta_{\min}/4 < \theta_{k^*} - \Delta_{\min}/4 \le \mathrm{LCB}_{k^*}^{\mathrm{ver}}(t)$, 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. □

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, $R_{n,t} \sim \mathrm{Bernoulli}(\mu_k)$ 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 $\mathrm{Reg}_\phi^{\mathrm{team}}(T)$, and average results over 50 i.i.d. instances.

Figure 1 plots $\mathrm{Reg}_\phi^{\mathrm{team}}(T)$ 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 $\Gamma_{\mathrm{eff}} \in [\Gamma, N\Gamma]$, 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.