FEDERATED MULTI-ARMED BANDITS VIA DISTRIBUTED TSALLIS MIRROR DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a multi-agent multi-armed bandit problem in which the agents of a network learn personalized (agent-specific) policies. Our method is designed for networks whose agents form clusters with similar loss distributions, in particular a shared optimal arm, while the cluster assignments are unknown; each agent's regret is measured against the optimal arm of its own cluster. We propose GTV-TSALLIS, which couples local instances of Tsallis-INF, a best-of-both-worlds mirror-descent bandit algorithm, through generalized total variation minimization (GTVMin) on the agents' policies. Consensus-based methods force all agents onto one shared policy. GTVMin instead penalizes *disagreement between neighbors*, so policies stay piecewise constant across cluster boundaries and distinct clusters converge to different arms without first estimating the clusters. A policy-similarity gate restricts the coupling to neighbors whose policies agree. We evaluate GTV-TSALLIS in homogeneous and heterogeneous settings, under stochastic and adversarial losses, across four network topologies and two real-world graphs. Whenever agents have same-cluster neighbors it reduces regret relative to non-cooperative Tsallis-INF and to cooperative baselines that assume a shared optimal arm: final regret falls by more than half in the homogeneous setting and by 17–30% on the real-world graphs. When every edge crosses clusters, the policy-similarity gate matches the non-cooperative baseline.

1 INTRODUCTION

Multi-armed bandits (MABs) model sequential decision-making under uncertainty: in each round an agent plays one of K arms, observes its loss, and aims to minimize cumulative regret against the best arm (Lattimore & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012). Increasingly, many such agents run in parallel, face similar but not identical problems, and can communicate over a graph (Landgren et al., 2021; Cesa-Bianchi et al., 2019; Y. SarcheshmehPour & Jung, 2022). When and how should they share information, given that their optimal arms may differ?

For the homogeneous case, where all agents share an optimal arm, the answer is known: pooling observations with any neighbor reduces estimator variance, and per-agent regret shrinks by a factor that grows with the number of cooperating agents (Bar-On & Mansour, 2019; Shi & Shen, 2021) or with the connectivity of the graph (Cesa-Bianchi et al., 2019; Landgren et al., 2021). When agents in different clusters have different optimal arms, the same pooling corrupts local estimates and cooperation can increase regret: in Appendix D, ungated pooling reaches four times the regret of not cooperating at all.

We study federated MABs whose agents form latent clusters: agents within a cluster share a loss distribution and hence an optimal arm, while clusters may disagree. The assumption is reasonable for recommendation across regions, where users of one region share preferences over items but regions differ (our MovieLens and LastFM experiments), and for the configuration of edge devices, where devices of one hardware type share the best setting but types differ. In both, an agent gains from sharing with similar peers and must resist contamination from dissimilar neighbors.

Our approach. We run a Tsallis-INF instance (Zimmert & Seldin, 2021) at every node and couple them by adding a *generalized total variation* (GTV) penalty to the linearized loss that each agent's mirror-descent step minimizes (Section 3). The penalty sums a variation measure $d(\mathbf{w}^{(i)}, \mathbf{w}^{(j)})$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

over the edges of the communication graph: it discourages neighboring policies from disagreeing but leaves weakly connected groups free to converge to different policies. The algorithm is not told the cluster structure; it discovers the clusters by solving GTVMin. The variation measure determines the statistical and computational properties of the method, and we study the squared Euclidean norm (Section 4.3) and the ℓ_1 norm (Section 4.2).

Contributions. (i) We formalize the *clustered federated bandit* setting as a federated learning network whose nodes carry MABs with cluster-wise identical loss distributions, with per-cluster regret as the objective (Section 2). (ii) We propose GTV-TSALLIS (Algorithm 1), which couples local Tsallis-INF instances through GTVMin on the agents' policies (Section 4). (iii) We evaluate it in homogeneous and heterogeneous settings, under stochastic and adversarial losses, on four synthetic graph families and two real-world graphs, against COOP-UCB2, DFTRL, and local Tsallis-INF (Section 5).

1.1 RELATED WORK

The closest prior method is GTVMin-based clustered federated learning (SarcheshmehPour et al., 2023), whose graph-total-variation penalty implicitly clusters connected devices in supervised learning. We port this penalty from model space to the bandit *policy* space, where an agent sees only its own realized loss rather than a gradient, and pair it with a policy-similarity gate.

Cooperative bandits with a shared optimal arm. Most multi-agent bandit work assumes all agents share the same loss distribution and hence the same optimal arm; in this homogeneous regime sharing observations with any neighbor reduces estimator variance and hence regret. COOP-UCB2 (Landgren et al., 2021) runs a consensus protocol on empirical means and Shi & Shen (2021) obtain similar guarantees with a central server; such UCB-based methods commit to the stochastic assumption and lack the adversarial guarantees of mirror-descent approaches. EXP3-COOP (Cesa-Bianchi et al., 2019) achieves $\tilde{O}(\sqrt{\alpha(\mathcal{G})KT})$ adversarial regret, Bar-On & Mansour (2019) close the gap to the lower bound with a Tsallis regularizer, and Yi & Vojnović (2023) propose DFTRL, a hybrid Tsallis and log-barrier regularizer with a collaborative inverse-propensity estimator over closed neighborhoods, our closest adversarial baseline. All pool the observations of every neighbor, which is unbiased only when every neighbor shares the mean-loss vector; in our clustered setting the pooling contaminates the per-cluster estimates.

Heterogeneous cooperative bandits with a global objective. Xu & Klabjan (2023) and Xu et al. (2025) allow heterogeneous rewards but keep a single global arm, the one with the best network-averaged loss, as the target. This is the right target when one decision must serve the whole network, such as one default configuration for every device, and the wrong one in the settings above, where per-cluster performance is what matters and agents with different preferences are forced toward an arm that is suboptimal for most of them. Neither work provides a mechanism for suppressing coupling across cluster boundaries while preserving it within them.

Clustered bandits with explicit cluster recovery. CLUB (Gentile et al., 2014) and COFIBA (Li et al., 2016) interleave playing arms with estimating which users share a cluster, and use the estimate to decide what to share; Korda et al. (2016) do the same over a peer-to-peer network, recovering cluster identity through gossip on estimated preference vectors. All three operate in a linear contextual model. Our method never estimates cluster membership, needs no number of clusters, and makes no linear or parametric assumption on the losses: clustering is a side effect of the GTV penalty pushing neighboring policies toward agreement, not a computation the algorithm performs.

Clustered federated learning. Personalized models for clients are learned by partitioning clients by model similarity (Ghosh et al., 2020), by a graph-total-variation penalty whose minimizers cluster connected devices (SarcheshmehPour et al., 2023), or by attentive message passing (Huang et al., 2021). All assume labeled data and full per-round gradient access, which is incompatible with bandit feedback. Instead of these supervised methods, we use GTVMin to couple the policies of agents, recovering the same implicit clustering under bandit feedback.

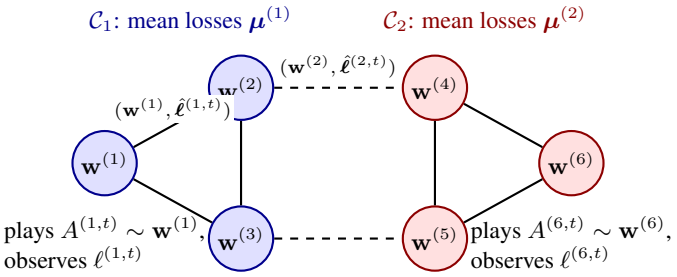

Figure 1: The clustered federated bandit setting. Each agent i (node) holds a policy $\mathbf{w}^{(i)}$ over the K arms, plays an arm drawn from it, and observes the loss of that arm only. Agents exchange their policy and loss estimate with their neighbors over the edges of the communication graph $\mathcal{G}$. Colors mark the latent clusters, whose members share a mean-loss vector and hence an optimal arm; dashed edges cross clusters. The algorithm is not told the clusters.

2 Problem Setting

We study interconnected agents, each learning a policy for a separate MAB. The agents interact across the edges of an undirected graph (Section 2.1), each is judged by its own regret (Section 2.2), and minimizing the sum of these regrets is the problem that our method and all baselines address through one algorithmic blueprint (Section 2.3); Figure 1 illustrates the setting.

Notation. Vectors are boldface ($\mathbf{w}$, $\hat{\boldsymbol{\ell}}$, $\boldsymbol{\mu}$) with subscripted entries (w_a); matrices are bold capitals ($\mathbf{L}$). Parenthesized superscripts index agents and rounds: $\mathbf{w}^{(i,t)}$ is agent i's policy at round t and $\hat{\ell}_a^{(i,t)}$ its loss estimate for arm a. The probability simplex over K arms is $\Delta^{K-1} = \{\mathbf{w} \in \mathbb{R}_+^K : \sum_{a=1}^K w_a = 1\}$, and $\mathcal{N}^{(i)}$ is the (open) neighborhood of agent i.

2.1 Federated Learning Network

A *federated learning network* (Y. SarcheshmehPour & Jung, 2022) is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose N nodes are agents and whose edges are communication links: agents i and j exchange information iff $(i,j) \in \mathcal{E}$, and $\mathcal{N}^{(i)} = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$. The agents are partitioned into C clusters $\mathcal{C}_1, \dots, \mathcal{C}_C$ and, following the *cluster assumption* of clustered federated learning (SarcheshmehPour et al., 2023), agents in the same cluster share their loss distribution. Nothing is assumed about how the graph relates to the clusters; the algorithms of Section 4 let the cluster structure emerge from observed losses and neighbor communication alone.

2.2 A Node-wise MAB

Each agent $i \in \mathcal{V}$ faces a K-armed stochastic bandit with mean-loss vector $\boldsymbol{\mu}^{(i)} \in [0,1]^K$. Under the cluster assumption $\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}^{(c)}$ for all $i \in \mathcal{C}_c$, so each cluster has one optimal arm $\tilde{a}^{(c)} = \arg\min_{a \in [K]} \mu_a^{(c)}$ with mean loss $\tilde{\mu}^{(c)} = \min_{a \in [K]} \mu_a^{(c)}$. Over T rounds, agent i plays mixed policies $\mathbf{w}^{(i,t)} \in \Delta^{K-1}$, drawing $A^{(i,t)} \sim \mathbf{w}^{(i,t)}$ and observing $\ell^{(i,t)}$ with $\mathbb{E}[\ell^{(i,t)}] = \mu^{(i)}_{A^{(i,t)}}$. Single-agent bandits measure regret against the best arm of the agent's own mean-loss vector; cooperative bandits with a global objective, against the arm with the best network-averaged loss. We measure each node against the optimal arm of its own cluster, which coincides with the former under the cluster assumption and differs from the latter whenever clusters disagree:

$$\mathrm{Reg}_T(i) := \sum_{t=1}^{T} \big(\mathbb{E}[\ell^{(i,t)}] - \tilde{\mu}^{(c(i))}\big) = \sum_{t=1}^{T} \big\langle \boldsymbol{\mu}^{(c(i))} - \tilde{\mu}^{(c(i))}\mathbf{1},\, \mathbf{w}^{(i,t)} \big\rangle. \tag{1}$$

If $\boldsymbol{\mu}^{(c(i))}$ were known, playing $\tilde{a}^{(c(i))}$ in every round would give $\mathrm{Reg}_T(i) = 0$. Since it is unknown and each round reveals a noisy realization of a single coordinate, the node must trade exploration against exploitation: every arm must be played often enough to estimate its mean loss, and arms that appear suboptimal rarely.

2.3 FEDERATED MAB ALGORITHM MODEL

A federated MAB algorithm chooses the policy sequences to minimize the network regret $\mathrm{Reg}_T = \sum_{i\in\mathcal{V}} \mathrm{Reg}_T(i)$,

$$\min_{\{\mathbf{w}^{(i,t)}\}\subset\Delta^{K-1}} \mathrm{Reg}_T, \tag{2}$$

where $\mathbf{w}^{(i,t+1)}$ may depend only on agent i's observations and on messages from $\mathcal{N}^{(i)}$ up to round t. In each round every agent (1) draws an arm from its policy and observes its loss, (2) forms a per-arm loss estimate $\hat{\boldsymbol{\ell}}^{(i,t)} \in \mathbb{R}^K$, (3) exchanges a message, typically $(\mathbf{w}^{(i,t)}, \hat{\boldsymbol{\ell}}^{(i,t)})$, with its neighbors, and (4) updates its policy from the estimate and the received messages. Algorithms differ in the estimator, the message, and the update; our contribution is the update. The blueprint covers TSALLIS-INF and DFTRL directly and COOP-UCB2 with running per-arm statistics in place of a loss vector and a deterministic UCB choice in step 1.

3 TSALLIS MIRROR DESCENT: BACKGROUND

TSALLIS-INF (Zimmert & Seldin, 2021) solves (2) for a single agent ($N = 1$) by online mirror descent on the simplex with the Bregman divergence of the negative $\frac{1}{2}$-Tsallis entropy $H_{1/2}(\mathbf{w}) = 2(1 - \sum_a \sqrt{w_a})$:

$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\Delta^{K-1}} \{\eta_t \langle \hat{\boldsymbol{\ell}}^{(t)}, \mathbf{w}\rangle + D_{H_{1/2}}\Big(\mathbf{w}, \mathbf{w}^{(t)}\Big)\}, \quad D_{H_{1/2}}(\mathbf{u}, \mathbf{w}) = 2\sum_a (\sqrt{u_a} - \sqrt{w_a})^2. \tag{3}$$

Among the α-Tsallis regularizers, $\alpha = \frac{1}{2}$ is the one whose optimal tuning does not require the suboptimality gaps (Zimmert & Seldin, 2021). By the FTRL equivalence the update accumulates $\hat{\boldsymbol{L}}^{(t)} := \sum_{s=1}^t \hat{\boldsymbol{\ell}}^{(s)}$ and solves

$$w_a^{(t+1)} = \frac{4}{(\eta_t(\hat{L}_a^{(t)} - x))^2}, \tag{4}$$

with x found by a one-dimensional Newton iteration that enforces $\sum_a w_a^{(t+1)} = 1$. Since only the played arm's loss is observed, the agent uses the importance-weighted (IW) or the reduced-variance (RV) estimator. The RV estimator subtracts a baseline of $\frac{1}{2}$ on arms with non-negligible probability, which lowers the variance without changing the bias and improves both the stochastic constants and the empirical regret in Zimmert & Seldin (2021); we use it throughout with $\eta_t = 4/\sqrt{t}$:

$$\hat{\ell}_a^{(t)} = \frac{(\ell^{(t)} - \mathbb{B}_a^{(t)})\mathbb{1}\{A^{(t)}=a\}}{w_a^{(t)}} + \mathbb{B}_a^{(t)}, \quad \mathbb{B}_a^{(t)} = \tfrac{1}{2}\mathbb{1}\{w_a^{(t)} \geq \eta_t^2\}. \tag{5}$$

Theorem 1 (Tsallis-INF (Zimmert & Seldin, 2021)). *With $\eta_t = 4/\sqrt{t}$ (RV), the algorithm simultaneously achieves $\mathbb{E}[\mathrm{Reg}_T] = O(\sqrt{KT})$ adversarially and $\mathbb{E}[\mathrm{Reg}_T] = O\big(\sum_{a:\Delta_a>0} \log T/\Delta_a\big)$ stochastically, without knowing T or which regime applies.*

Tsallis-INF therefore needs no knowledge of whether the losses are stochastic or adversarial, which is why we build on it.

4 FEDERATED TSALLIS MIRROR DESCENT

GTVMin penalizes the disagreement between the policies of neighboring agents, so within-cluster pairs converge to a common policy while inter-cluster pairs settle on different ones, unlike consensus methods, which enforce one policy for all. Section 4.1 augments the update (4) with the GTVMin coupling, which requires a disagreement measure; Sections 4.2 and 4.3 instantiate it with the ℓ_1 and the squared ℓ_2 measure; Section 4.4 adds a gate that reweights each neighbor by the distance between the two policies.

4.1 GTVMIN-AUGMENTED MIRROR DESCENT

In round t, agent i averages its loss estimate with those of its neighbors,

$$\bar{\boldsymbol{\ell}}^{(i,t)} := \frac{1}{1 + |\mathcal{N}^{(i)}|}\Big(\hat{\boldsymbol{\ell}}^{(i,t)} + \textstyle\sum_{j\in\mathcal{N}^{(i)}} \hat{\boldsymbol{\ell}}^{(j,t)}\Big), \tag{6}$$

which within a cluster reduces the variance of the estimate by a factor of order $1/(1+|\mathcal{N}^{(i)}|)$, the source of the collaborative gain. It then minimizes the augmented objective

$$\Phi_i^{(t)}(\mathbf{w}^{(i)}) := \langle \bar{\boldsymbol{\ell}}^{(i,t)}, \mathbf{w}^{(i)} \rangle + \frac{\lambda}{2} \sum_{j \in \mathcal{N}^{(i)}} d\Big(\mathbf{w}^{(i)}, \mathbf{w}^{(j,t)}\Big), \tag{7}$$

with coupling strength $\lambda > 0$ and a disagreement measure $d : \Delta^{K-1} \times \Delta^{K-1} \to \mathbb{R}_{\geq 0}$. The (sub)gradient of the coupling term gives the effective gradient

$$\mathbf{g}^{(i,t)} := \bar{\boldsymbol{\ell}}^{(i,t)} + \frac{\lambda}{2}\, \partial_{\mathbf{w}^{(i)}} \sum_{j \in \mathcal{N}^{(i)}} d\Big(\mathbf{w}^{(i)}, \mathbf{w}^{(j,t)}\Big), \tag{8}$$

which replaces $\hat{\boldsymbol{\ell}}^{(t)}$ in the update (4).

4.2 GTV-Tsallis

Following the GTVMin principle (SarcheshmehPour et al., 2023), our default measure is $d(\mathbf{w}^{(i)}, \mathbf{w}^{(j)}) = \big\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big\|_1$. Its subdifferential is the elementwise sign, so

$$\mathbf{g}^{(i,t),\mathrm{GTV}} = \bar{\boldsymbol{\ell}}^{(i,t)} + \frac{\lambda}{2} \sum_{j \in \mathcal{N}^{(i)}} \operatorname{sign}\Big(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\Big), \tag{9}$$

and each agent accumulates this gradient and solves the Tsallis Newton problem:

$$\boxed{\hat{\boldsymbol{L}}^{(i,t)} \leftarrow \hat{\boldsymbol{L}}^{(i,t-1)} + \mathbf{g}^{(i,t),\mathrm{GTV}}, \qquad w_a^{(i,t+1)} = \frac{4}{\big(\eta_t(\hat{L}_a^{(i,t)} - x^{(i,t)})\big)^2}, \quad \textstyle\sum_{a=1}^K w_a^{(i,t+1)} = 1.} \tag{10}$$

The ℓ_1 subgradient vanishes as soon as neighboring policies coincide, so within-cluster agreement is reached exactly rather than asymptotically. This is the piecewise-constant behavior of GTVMin on graphs (see also Network Lasso, Hallac et al., 2015): agents are hard-clustered into homogeneous groups without any estimate of the cluster assignment.

4.3 GTV-Tsallis (ℓ_2)

The ℓ_2 variant uses the smooth measure $d(\mathbf{w}^{(i)}, \mathbf{w}^{(j)}) = \big\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big\|_2^2$, the ℓ_2 instance of GTVMin, whose gradient, with the $\frac{1}{2}$ of (7), is graph-Laplacian feedback,

$$\mathbf{g}^{(i,t),\mathrm{L2}} = \bar{\boldsymbol{\ell}}^{(i,t)} + \lambda \sum_{j \in \mathcal{N}^{(i)}} \big(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\big) = \bar{\boldsymbol{\ell}}^{(i,t)} + \lambda(\mathbf{L}\mathbf{W})_i, \tag{11}$$

with $\mathbf{L}$ the graph Laplacian; the policy update is (10) with $\mathbf{g}^{\mathrm{L2}}$ in place of $\mathbf{g}^{\mathrm{GTV}}$. Since this gradient is non-zero whenever neighbors disagree, agents are pulled toward each other without ever agreeing exactly: consensus is soft, and clustering only implicit.

4.4 Policy-Similarity Gating

Until policies have separated, the unweighted average (6) pulls cross-cluster loss estimates into every update. Following attentive message passing in personalized federated learning (Huang et al., 2021; Ghosh et al., 2020), we reweight each neighbor j by a *policy-similarity gate*

$$g_{ij}^{(t)} = \exp\!\Big(-\beta \Big\|\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\Big\|\Big), \tag{12}$$

with concentration parameter $\beta \geq 0$, applied to both the loss average,

$$\bar{\boldsymbol{\ell}}_{\mathrm{gate}}^{(i,t)} := \frac{\hat{\boldsymbol{\ell}}^{(i,t)} + \sum_{j \in \mathcal{N}^{(i)}} g_{ij}^{(t)}\, \hat{\boldsymbol{\ell}}^{(j,t)}}{1 + \sum_{j \in \mathcal{N}^{(i)}} g_{ij}^{(t)}}, \tag{13}$$

and the coupling term, where $\operatorname{sign}(\cdot)$ in (9) (the Laplacian difference in (11)) is weighted by $g_{ij}^{(t)}$:

$$\mathbf{g}_{\text{adapt}}^{(i,t)} = \bar{\boldsymbol{\ell}}_{\text{gate}}^{(i,t)} + \frac{\lambda}{2} \sum_{j \in \mathcal{N}^{(i)}} g_{ij}^{(t)} \operatorname{sign}\big(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\big). \tag{14}$$

A burn-in of t_0 rounds without communication lets policies separate before the gate is applied. Setting $\beta = 0$, $t_0 = 0$ recovers the ungated algorithm and $\beta \to \infty$ local Tsallis-INF; pseudocode is Algorithm 1 in Appendix A. Both β and t_0 are tuned per experiment and reported in the corresponding figure captions, with sparser graphs favouring a sharper gate and a shorter burn-in. The exponential form of the gate, however, has two exact readings.

The gate as a Bayes factor. Suppose that a cluster-mate's policy is a perturbation of i's, $\mathbf{w}^{(j,t)} = \mathbf{w}^{(i,t)} + \boldsymbol{\delta}$ with $p(\boldsymbol{\delta}) \propto \exp(-\|\boldsymbol{\delta}\|/b)$, while a non-cluster-mate's policy follows a diffuse density independent of $\mathbf{w}^{(i,t)}$. The likelihood ratio of the two hypotheses is $\exp\big(-\big\|\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\big\|/b\big)$ up to a constant: the gate with $\beta = 1/b$. A product of Laplace densities gives the ℓ_1 gate, a multivariate Laplace density with the Euclidean norm the ℓ_2 gate. The normalization in (13) makes the weights a softmax over $\{i\} \cup \mathcal{N}^{(i)}$, the posterior under a uniform prior that exactly one member of this set holds an on-target estimate; the gated average is the posterior mean under this one-source model. The model also names the gate's cost: several cluster-mates are all under-weighted, which is the gap between gated and ungated pooling in the homogeneous setting (Appendix D).

The gated coupling as a saturating penalty. With $\boldsymbol{\delta} = \mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}$, the gated ℓ_1 term $\exp(-\beta\|\boldsymbol{\delta}\|_1)\operatorname{sign}(\boldsymbol{\delta})$ is the subgradient of $\psi_1(\boldsymbol{\delta}) = \beta^{-1}\big(1 - \exp(-\beta\|\boldsymbol{\delta}\|_1)\big)$ and the gated ℓ_2 term $\exp(-\beta\|\boldsymbol{\delta}\|_2)\,\boldsymbol{\delta}$ the gradient of $\psi_2(\boldsymbol{\delta}) = \beta^{-2}\big(1 - (1 + \beta\|\boldsymbol{\delta}\|_2)\exp(-\beta\|\boldsymbol{\delta}\|_2)\big)$. Both agree with the ungated measure for small gaps and saturate for large ones: a cross-cluster edge costs a bounded amount however far the policies separate, while a within-cluster edge is pulled tight. Gated GTV-TSALLIS is thus GTVMin with a truncated variation measure, the device of robust statistics for tolerating a minority of gross disagreements. Since b is unobserved, β is estimated rather than computed; a data-driven rule for (β, λ, t_0) remains open.

5 NUMERICAL EXPERIMENTS

Full configurations and the COOP-UCB2 baseline are described in Appendix A. Defaults: $K = 10$, $T = 10{,}000$, $\eta_t = 4/\sqrt{t}$ (RV), $\lambda = 1/\sqrt{T}$, SBM edge probabilities $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.1$. Regret curves of Experiment 1 average 100 Monte Carlo runs with ± 1 standard deviation bands; Experiments 2 to 4 use 20 runs and the p_{out} sweep of Appendix B 10 runs with $\pm$SEM bars. We compare local TSALLIS-INF ($\lambda = 0$, no communication); ungated GTV-TSALLIS (ℓ_1 and ℓ_2; $\beta = 0$, $t_0 = 0$), evaluated in the homogeneous setting; gated GTV-TSALLIS (ℓ_1 and ℓ_2; β and t_0 tuned per experiment and given in each figure caption), the variants for the heterogeneous setting; COOP-UCB2 (Landgren et al., 2021), a consensus UCB method for the stochastic homogeneous setting; and DFTRL (Yi & Vojnović, 2023), decentralized FTRL with a hybrid Tsallis/log-barrier regularizer for the adversarial homogeneous setting. All Tsallis-based variants use the RV estimator and instantiate Algorithm 1.

5.1 EXPERIMENT 1: HOMOGENEOUS REGRET COMPARISON

One cluster of $N = 10$ agents shares $\boldsymbol{\mu} = \operatorname{linspace}(0.1, 0.9, K)$ with arm 0 optimal, so every neighbor's observation helps and the ungated variants are the right comparison; this isolates the variance-reduction gain before the heterogeneous experiments add contamination. Under stochastic losses (Figure 2, left) both GTV-TSALLIS variants beat local Tsallis-INF, the ℓ_1 variant reaching a final regret of 51.9 against 117.6 (56% reduction; ℓ_2: 53.2), consistent with pooling over the neighborhood; the ℓ_1 edge over ℓ_2 may reflect its exact within-cluster agreement. COOP-UCB2 also gains from consensus but converges more slowly, UCB exploration being less sample-efficient than RV Tsallis-INF here. Under adversarial losses (right), the phased protocol of Zimmert & Seldin (2021, Sec. 8) with phase lengths growing by a factor 1.6 (Appendix A), the Tsallis-INF family stays sublinear across all phase switches and GTV-TSALLIS (ℓ_1) is again lowest (62.9 against 144.4, 56% reduction; ℓ_2: 68.4), while DFTRL grows near-linearly, more than a decade above the rest (logarithmic axis), so its design assumptions do not transfer to this setting.

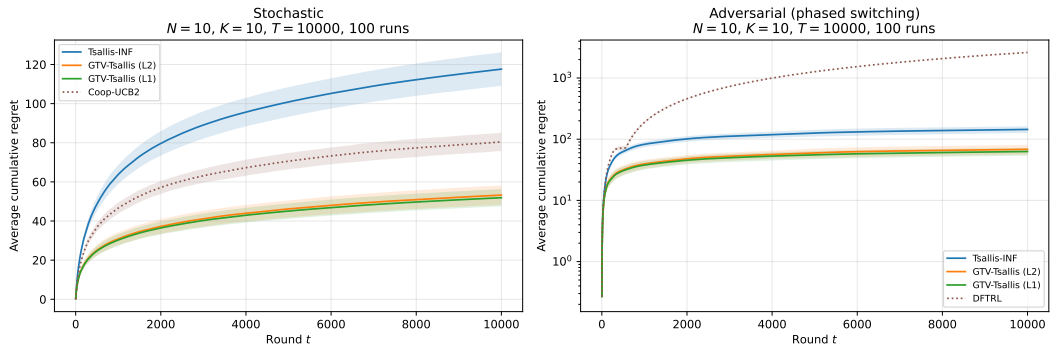

Figure 2: Homogeneous setting, average per-agent cumulative regret ($N = 10$, $K = 10$, $T = 10{,}000$, 100 runs, $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.1$; cooperative variants ungated, $\beta = 0$, $t_0 = 0$). *Left*: stochastic losses, linear axes. *Right*: phased-switching adversarial losses (phase factor 1.6, $\delta = 0.3$), logarithmic vertical axis, with DFTRL as the adversarial baseline.

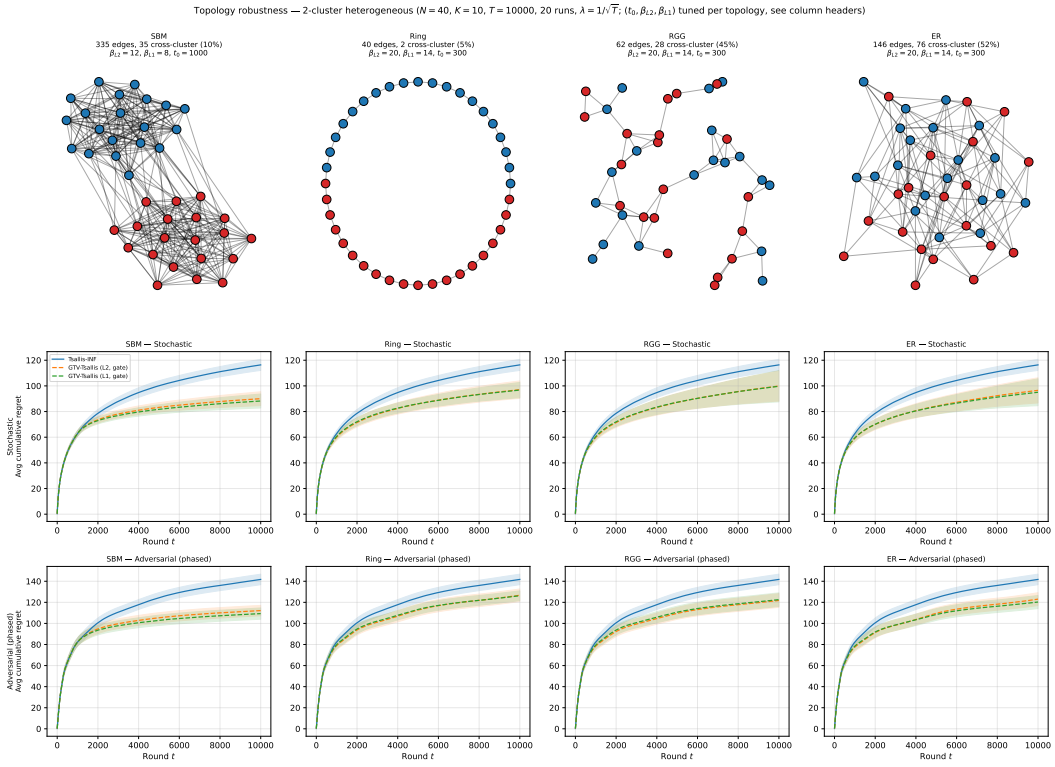

Figure 3: Heterogeneous two-cluster setting on four graph families ($N = 40$, $K = 10$, $T = 10{,}000$, 20 runs, $\lambda = 1/\sqrt{T}$; adversarial panels use phase base 1.6, $\delta = 0.3$; gate settings SBM $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 1000$, others $\beta_{\ell_2} = 20$, $\beta_{\ell_1} = 14$, $t_0 = 300$). *Top*: one graph realization per family, colored by cluster, with its cross-cluster edge fraction. *Middle*: average cumulative regret in the stochastic regime. *Bottom*: the same under the phased-switching adversarial protocol. Both gated variants improve on local TSALLIS-INF in every family and in both regimes.

5.2 EXPERIMENT 2: HETEROGENEOUS CLUSTERS ACROSS GRAPH TOPOLOGIES

With $C = 2$ clusters of different optimal arms, the gate must suppress cross-cluster contamination while preserving within-cluster pooling. Since its benefit depends on how much cluster information the graph carries, we use four graph families ($N = 40$, 20 runs): an SBM, whose edges depend

on cluster identity, and three that ignore it, a ring (degree 2), a random geometric graph (RGG; uniform positions in $[0,1]^2$, fixed radius), and an Erdős–Rényi (ER) graph. Their cross-cluster edge fractions range from 5% (ring) to 52% (ER, which carries no cluster information at all). SBM uses $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 1000$; the sparser families a sharper gate and shorter burn-in ($\beta_{\ell_2} = 20$, $\beta_{\ell_1} = 14$, $t_0 = 300$), since they offer each node fewer cluster-correlated neighbors. Both gated variants improve on local TSALLIS-INF in every family (Figure 3), so the benefit is not an artifact of the SBM's within-cluster edge bias. Under the phased adversarial protocol with the same settings (Figure 3, bottom row) local Tsallis-INF reaches a final regret of 141.8 throughout, the best gated variant 109.3 on SBM (23% less), 126.3 on ring (11%), 121.8 on RGG (14%), and 120.3 on ER (15%). Three further tests are in the appendix: with clusters forming an independent set, so that every edge crosses clusters, the gated variants track local Tsallis-INF instead of degrading (Appendix C); a direct gated-versus-ungated comparison in both settings and regimes (Appendix D); and a sweep of p_{out}, over which both gated variants stay below local Tsallis-INF (Appendix B).

5.3 EXPERIMENT 3: NUMBER OF CLUSTERS

Each agent's usable signal comes from its same-cluster neighbors, so increasing C at fixed N shrinks every cluster and raises the fraction of contaminating edges. Sweeping $C \in \{2, 3, 5, 10\}$ at $N = 100$ (SBM defaults, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 1000$, one distinct optimal arm per cluster, both regimes), Figure 4 shows the margin between the gated variants and local Tsallis-INF narrowing as clusters proliferate, as expected when there is progressively less same-cluster signal to pool.

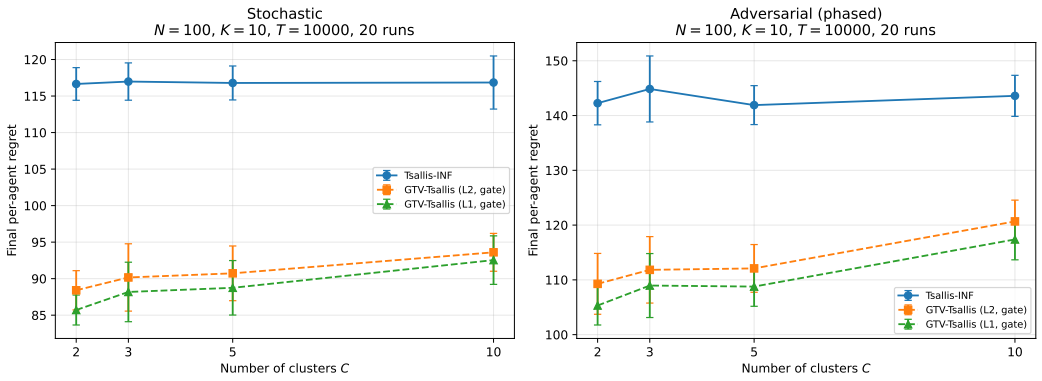

Figure 4: Final per-agent regret at $T = 10{,}000$ against the number of clusters $C \in \{2, 3, 5, 10\}$ at $N = 100$ ($K = 10$, 20 runs, $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.1$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 1000$, $\lambda = 1/\sqrt{T}$; bars ± 1 standard deviation). *Left*: stochastic. *Right*: phased-switching adversarial.

5.4 EXPERIMENT 4: REAL-WORLD DATA

Two datasets test whether the gain survives when both losses and graph come from real data. Agents are users, arms a fixed item set, losses 1 minus a normalized rating or listening weight, and clusters the two largest groups of users sharing a personal-optimal arm, so every agent in a cluster genuinely shares an optimal arm; $T = 10{,}000$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_1} = \beta_{\ell_2} = 4$, $t_0 = 50$, 20 runs, no per-dataset tuning. On MovieLens-100k ($N = 232$ users after best-arm filtering, $K = 10$ most-rated movies, 8-nearest-neighbor graph from cosine similarity of rating profiles), GTV-TSALLIS (ℓ_1, gated) reduces final regret by 29.6% relative to local Tsallis-INF (60.5 ± 2.1 vs. 85.9 ± 2.0; ℓ_2: 64.3 ± 3.3; Figure 5, top). Since that graph is itself a similarity construction, we repeat the experiment on HetRec 2011 LastFM-2k, whose user–user friendship network is real ($N = 228$ users, $K = 20$ most-listened artists, 37.5% of edges cross clusters): the gated ℓ_1 variant reduces final regret by 16.9% (266.6 ± 5.5 vs. 320.8 ± 5.1; ℓ_2: 272.9 ± 12.0; Figure 5, bottom).

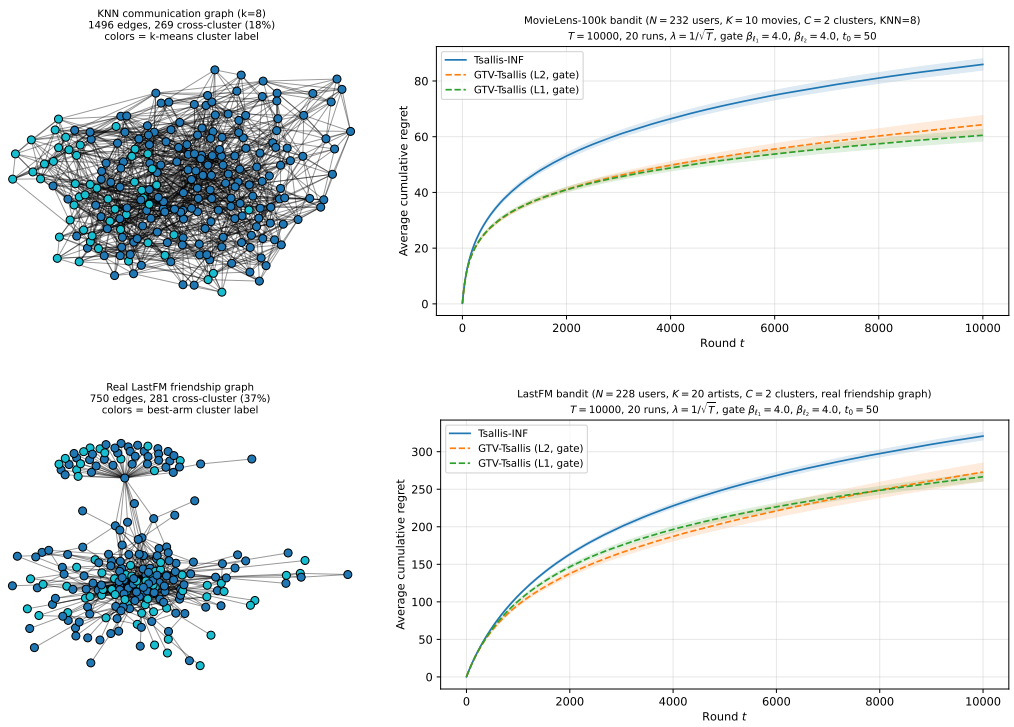

Figure 5: Real-world data, both at $C = 2$ clusters, $T = 10{,}000$, 20 runs, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_1} = \beta_{\ell_2} = 4$, $t_0 = 50$; each row pairs the communication graph, colored by best-arm cluster, with the average cumulative regret, bands ± 1 standard deviation. *Top*: MovieLens-100k ($N = 232$ users, $K = 10$ movies) on an 8-nearest-neighbor graph from cosine similarity of rating profiles. *Bottom*: HetRec 2011 LastFM ($N = 228$ users, $K = 20$ artists) on the real friendship graph, 37.5% of whose edges cross clusters. Both gated GTV-TSALLIS variants improve on local TSALLIS-INF on both datasets.

6 CONCLUSION

Ungated GTV-TSALLIS more than halves the regret of local Tsallis-INF when all agents share an optimal arm, the ℓ_1 variant slightly ahead of ℓ_2 and both ahead of COOP-UCB2 and DFTRL. In the heterogeneous setting the gated variants decouple across clusters after the burn-in and keep most of the within-cluster gain where ungated pooling collapses; the advantage holds on four graph families, under stochastic and adversarial losses, and on two real datasets, one with an unmodified friendship graph. The method is fully decentralized, needs no server and no knowledge of the clusters, and (β, t_0) interpolates between fully cooperative and fully local play. Its limitation is tuning: β, λ and t_0 are set by hand, and a data-driven rule for them is the main direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, generating synthetic data sets, and cleaning and reformatting data sets. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or interpreting results, and assisting with translation and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code, summarizing existing literature, drafting parts of the paper, and editing the paper to improve readability. We have reviewed all AI-assisted work. AI-assisted code was read and run by the authors, and all reported numbers come from our own runs of that code. We take responsi-

bility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The supplementary material contains the implementation of GTV-Tsallis and of the baselines, together with one script per figure, listed in its README. The algorithm is stated as Algorithm 1 in Appendix A, which also gives the software versions, the construction of the loss vectors, the phased adversarial protocol and the Coop-UCB2 baseline in the loss setting; the gate sharpness β, the burn-in t_0, the coupling λ, the horizon and the number of runs are given in the caption of each figure. The two datasets of Section 5.4 are public and are downloaded and preprocessed by their scripts.

References

Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. *arXiv preprint arXiv:1907.03346*, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, Dec. 2012. doi: 10.1561/2200000024.

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

C. Gentile, S. Li, and G. Zapella. Online clustering of bandits. In *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, volume 32, Beijing, China, 2014.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19586–19597, 2020.

D. Hallac, J. Leskovec, and S. Boyd. Network lasso: Clustering and optimization in large graphs. In *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 387–396, Aug. 2015. doi: 10.1145/2783258.2783313.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-IID data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7865–7873, 2021. doi: 10.1609/aaai.v35i9.16960.

Nathan Korda, Balazs Szorenyi, and Shuai Li. Distributed clustering of linear bandits in peer to peer networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. *Automatica*, 125:109445, 2021. doi: 10.1016/j.automatica.2020.109445.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proc. of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy, Jul. 2016.

Yasmin SarcheshmehPour, Yu Tian, Linli Zhang, and Alexander Jung. Clustered federated learning via generalized total variation minimization. *IEEE Trans. Signal Process.*, 71:4240–4256, 2023. doi: 10.1109/TSP.2023.3322848.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. In *Advances in Neural Information Processing Systems*, 2023.

Mengfan Xu, Liren Shan, Fatemeh Ghaffari, Xuchuang Wang, Xutong Liu, and Mohammad Hajiesmaili. Heterogeneous multi-agent multi-armed bandits on stochastic block models, 2025.

L. Zhang Y. SarcheshmehPour, Y. Tian and A. Jung. Networked federated learning. *arXiv e-prints*, 2022.

Jialin Yi and Milan Vojnović. On regret-optimal cooperative nonstochastic multi-armed bandits. *arXiv preprint arXiv:2211.17154*, 2023.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A Experimental Details

Computational environment. All experiments were run on an Apple M1 Pro (ARM64) with Python 3.9.6, NumPy 2.0.2, Matplotlib 3.8.4, NetworkX 2.8.8, and scikit-learn 1.6.1. Parallelism across Monte Carlo runs used `multiprocessing.Pool`.

SBM graphs are generated with `networkx.stochastic_block_model`; agents observe Bernoulli losses $\ell^{(i,t)} \sim \mathrm{Ber}(\mu^{(c(i))}_{A^{(i,t)}})$. In the heterogeneous setup, Cluster c has loss vector equal to a one-position cyclic shift of $\mathrm{linspace}(0.1, 0.9, K)$, so Cluster c has arm c optimal. Random seeds are fixed for reproducibility.

Adversarial protocol. Following Zimmert & Seldin (2021), mean losses alternate between phase A, $\boldsymbol{\mu}^A_k = 1 - \delta \cdot \mathbb{1}\{k = 0\}$, and phase B, $\boldsymbol{\mu}^B_k = \delta \cdot \mathbb{1}\{k \neq 0\}$, with $\delta = 0.3$. Arm 0 is always optimal; phase lengths grow exponentially with factor 1.6.

Coop-UCB2 (loss setting). Each agent maintains running consensus estimates of mean reward $r = 1 - \ell$ and sample count per arm; arm selection uses the standard UCB index of (Landgren et al., 2021) applied to the consensus reward estimates.

Algorithm. The pseudocode for GTV-Tsallis is given below.

B Robustness to Inter-Cluster Connectivity

We fix $p_{\mathrm{in}} = 0.8$, $C = 2$ clusters of $N = 10$, and sweep $p_{\mathrm{out}} \in \{0.20, \ldots, 0.70\}$; burn-in $t_0 = 1000$ is fixed at a conservative value that works across the full range (at high p_{out} agents need more local-play rounds to separate policies before the gate becomes informative). Both gated variants stay below the local Tsallis-INF baseline across the full range in both stochastic and adversarial settings (Figure 6). Per-run regret is sensitive to the realized graph and to the (un-tuned) heuristic β, t_0, so the experiment shows a trend in the mean rather than a tight guarantee at each p_{out}.



Figure 6: Final cumulative regret at $T = 10{,}000$ vs. inter-cluster edge probability p_{out} ($p_{\mathrm{in}} = 0.8$, $N = 20$, $K = 10$, $\beta_{\mathrm{L2}} = 5$, $\beta_{\mathrm{L1}} = 3$, $t_0 = 1000$, 10 runs per point). *Left*: stochastic heterogeneous setting. *Right*: adversarial heterogeneous setting. Dashed line: local Tsallis-INF (RV) reference (independent of p_{out}). Error bars show ± 1 SEM over runs. Both gated variants are on average below the local baseline; variability reflects sensitivity to the specific graph realization.

Algorithm 1 GTV-TSALLIS (RV)

Require: Graph $\mathcal{G}$, arms K, coupling λ, gate $\beta \geq 0$, burn-in $t_0 \geq 0$

1: Initialize $\mathbf{w}^{(i,1)} = \mathbf{1}/K$, $\hat{\boldsymbol{L}}^{(i,0)} = \mathbf{0}_K$ for all i
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad \eta_t \leftarrow 4/\sqrt{t}$ **(RV)**
4: $\quad$ **for** each agent $i \in \mathcal{V}$ **do** ▷ in parallel
5: $\quad\quad$ Sample $A^{(i,t)} \sim \mathbf{w}^{(i,t)}$; observe loss $\ell^{(i,t)}$
6: $\quad\quad$ Compute $\hat{\boldsymbol{\ell}}^{(i,t)}$ via (5)
7: $\quad\quad$ **if** $t < t_0$ **then** ▷ burn-in: local Tsallis-INF only
8: $\quad\quad\quad \hat{\boldsymbol{L}}^{(i,t)} \leftarrow \hat{\boldsymbol{L}}^{(i,t-1)} + \hat{\boldsymbol{\ell}}^{(i,t)}$
9: $\quad\quad$ **else** ▷ cooperative phase (all rounds if $t_0 = 0$)
10: $\quad\quad\quad$ Receive $(\mathbf{w}^{(j,t)}, \hat{\boldsymbol{\ell}}^{(j,t)})$ from $j \in \mathcal{N}^{(i)}$
11: $\quad\quad\quad g_{ij} \leftarrow \exp(-\beta \left\|\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)}\right\|)$ for all $j \in \mathcal{N}^{(i)}$ ▷ $\beta = 0 \Rightarrow g_{ij} = 1$
12: $\quad\quad\quad \bar{\boldsymbol{\ell}}^{(i,t)} \leftarrow \dfrac{\hat{\boldsymbol{\ell}}^{(i,t)} + \sum_j g_{ij} \hat{\boldsymbol{\ell}}^{(j,t)}}{1 + \sum_j g_{ij}}$ ▷ gate-weighted loss average (13)
13: $\quad\quad\quad \mathbf{g}_{\text{adapt}}^{(i,t)} \leftarrow \bar{\boldsymbol{\ell}}^{(i,t)} + \frac{\lambda}{2} \sum_j g_{ij} \operatorname{sign}(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)})$ ▷ gate-weighted coupling (14)
14: $\quad\quad\quad \hat{\boldsymbol{L}}^{(i,t)} \leftarrow \hat{\boldsymbol{L}}^{(i,t-1)} + \mathbf{g}_{\text{adapt}}^{(i,t)}$
15: $\quad\quad$ **end if**
16: $\quad\quad \mathbf{w}^{(i,t+1)} \leftarrow$ solution of (4) with $\hat{\boldsymbol{L}}^{(i,t)}, \eta_t$
17: $\quad$ **end for**
18: **end for**

For the ℓ_2 variant, replace $\operatorname{sign}(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)})$ *with* $2(\mathbf{w}^{(i,t)} - \mathbf{w}^{(j,t)})$ *in line 13.*

C INDEPENDENT-SET CLUSTER CONFIGURATION

As a further stress test beyond Appendix B, we configure clusters as an independent set: agents in the same cluster share no direct edges, so all communication is via cross-cluster edges ($p_{\text{cross}} = 0.4$, $t_0 = 2000$, otherwise the same settings as the topology-robustness experiment of Section 5.2). Here standalone reaches 118.0 ± 4.0 and the gated variants reach 119.4 ± 7.0 (ℓ_2) and 119.4 ± 6.8 (ℓ_1) at $T = 10{,}000$, statistically indistinguishable from standalone (Figure 7). With no within-cluster edges there is no useful signal to share, and every available edge is purely contaminating; the gate correctly suppresses it rather than producing a spurious gain, matching standalone instead of degrading below it.

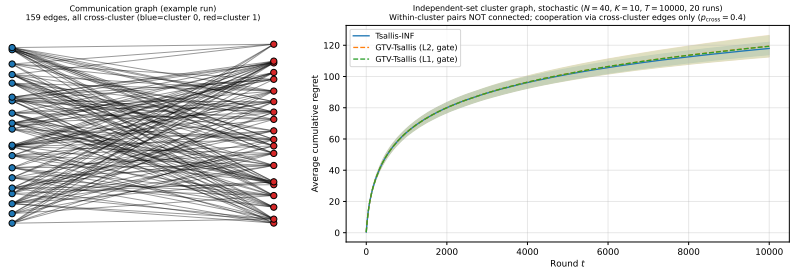

Figure 7: Independent-set cluster configuration: within-cluster agents share no direct edges, so cooperation can only occur across cluster boundaries ($N = 40$, $K = 10$, $T = 10{,}000$, 20 runs, $p_{\text{cross}} = 0.4$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 2000$). *Left*: example communication graph (blue = cluster 0, red = cluster 1); every edge crosses clusters. *Right*: both gated variants track standalone TSALLIS-INF rather than degrading, showing the gate suppresses purely contaminating edges instead of producing a spurious gain.

Repeating this under the phased-switching adversarial protocol gives the same conclusion, more sharply: standalone reaches 141.0 ± 5.4 and the gated variants reach 141.0 ± 5.4 (ℓ_2) and 141.1 ± 5.4 (ℓ_1) at $T = 10{,}000$ (Figure 8), indistinguishable from standalone to within noise. The gate's

response to a graph with no usable signal is the same whether the environment is stochastic or switches phases: it suppresses the purely contaminating edges rather than being confused by them.

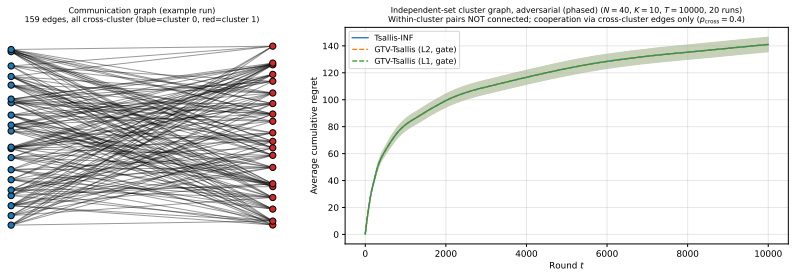

Figure 8: Independent-set cluster configuration under the phased-switching adversarial protocol (same setting as Figure 7: $N = 40$, $K = 10$, $T = 10{,}000$, 20 runs, $p_{\text{cross}} = 0.4$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 2000$, phase base 1.6, $\delta = 0.3$). Both gated variants track standalone TSALLIS-INF, as in the stochastic case.

D GATED VS. UNGATED COMPARISON

We compare the gated and ungated GTV-TSALLIS variants directly, in the homogeneous and heterogeneous settings, with $N = 40$, $K = 10$, $T = 10{,}000$, 20 runs, $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.4$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, and $t_0 = 2000$. In the stochastic setting (Figure 9), the homogeneous panel favors ungated cooperation (43–44 vs. the gated variants' 97–100), since pooling every neighbor's estimate is pure variance reduction when all agents share an optimal arm and the gate pays a mild cost for its burn-in caution; standalone reaches 118.8. The heterogeneous panel reverses this: ungated cooperation pools across the cluster boundary and collapses to 486–488, more than four times standalone (117.5), while the gated variants remain controlled at 99–101.

Repeating this under the phased-switching adversarial protocol (Figure 10) gives the same ordering in both panels. The homogeneous panel again favors ungated cooperation (45–46 vs. the gated variants' 120–123; standalone 142.1). The heterogeneous panel again shows the ungated variants collapsing, to 429–640 against standalone's 144.3, while the gated variants stay controlled at 124–126.

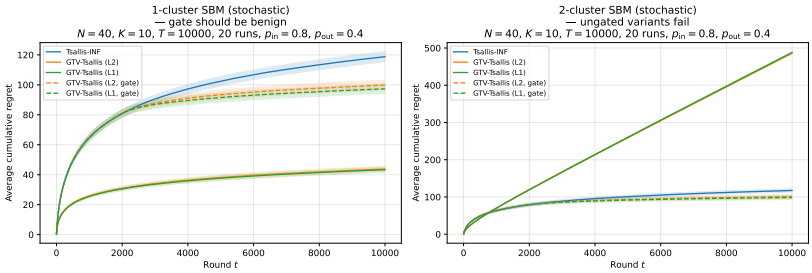

Figure 9: Gated vs. ungated GTV-TSALLIS, stochastic setting ($K = 10$, $T = 10{,}000$, 20 runs, $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.4$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 2000$). *Left*: homogeneous single-cluster ($N = 40$); ungated cooperation is best, as pooling is pure variance reduction when every agent shares an optimal arm. *Right*: heterogeneous two-cluster ($N = 40$); ungated cooperation pools across the cluster boundary and its regret grows well past standalone, while the gated variants remain below standalone throughout.

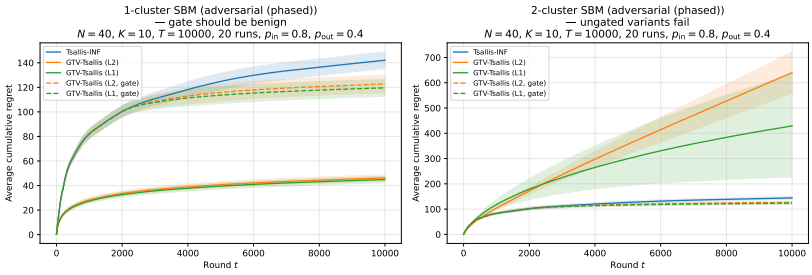

Figure 10: Gated vs. ungated GTV-TSALLIS under the phased-switching adversarial protocol (same setting as Figure 9: $K = 10$, $T = 10{,}000$, 20 runs, $p_{\text{in}} = 0.8$, $p_{\text{out}} = 0.4$, $\lambda = 1/\sqrt{T}$, $\beta_{\ell_2} = 12$, $\beta_{\ell_1} = 8$, $t_0 = 2000$, phase base 1.6, $\delta = 0.3$). *Left*: homogeneous single-cluster ($N = 40$); ungated cooperation attains the lowest regret. *Right*: heterogeneous two-cluster ($N = 40$); ungated cooperation collapses well past standalone, while the gated variants remain controlled.