X-KB-UCB: Decentralized Cross Kriging-Believer UCB for Static and Time-Varying GP Bandits

Anonymous authors
Paper under double-blind review

Abstract

We study decentralized Gaussian process (GP) bandits under strict communication budgets and a shared reward function. We introduce *X-KB-UCB*, a gossip-based Upper-Confidence Bound (UCB) method in which agents periodically exchange only their most recent chosen arm and the observed reward. At gossip rounds, agents coordinate exploration through a cross-agent Kriging–Believer update, while between gossip rounds each agent follows the corresponding single-agent rule, GP-UCB for static rewards and TV-GP-UCB for time-varying rewards. We provide high-probability no-regret guarantees for augmented agents, using an agent-centric accounting that includes both locally collected and gossiped observations, in both the static setting and a time-varying setting modeled by a Markov-drift GP. The resulting bounds are expressed in terms of information gain and recover standard single-agent rates when gossip is absent. In the always-gossip regime, they match the centralized batch-selection rate of GP-BUCB, with an additional term reflecting drift. Experiments confirm that gossip yields consistent gains over independent agents and approaches a centralized baseline under the same evaluation budget.

1 Introduction

Decentralized learning and control arise in multi-robot teams, mobile sensor networks, and embedded platforms where bandwidth, energy, and latency limit communication (Boyd et al., 2006; Gielis et al., 2022; Halsted et al., 2021; Li et al., 2016; Fattah et al., 2020; Gussen et al., 2021; Imbert et al., 2022; Anderson et al., 2021). In these systems, agents act online in a potentially noisy environment, choosing actions often over continuous domains, without a central coordinator. Sample efficiency is critical because each new environmental measurement can be costly. Gaussian process (GP) bandits (Srinivas et al., 2012) provide a principled exploration exploitation tradeoff with finite-time guarantees under such environments and are used across a range of applications (Garnett, 2023a; Cheikh Melainine et al., 2025; Parker-Holder et al., 2020). Many applications are also nonstationary, which motivates time-varying GP models with controlled drift (Bogunovic et al., 2016).

We study a multi-agent GP bandit problem in which agents optimize a shared reward function under noisy observations and strict communication budgets. This setting is motivated by decentralized learning and control in multi-robot teams, mobile sensor networks, and embedded platforms, where full centralization is limited by bandwidth, energy constraints, and latency induced by hardware or the environment (Boyd et al., 2006; Gielis et al., 2022; Halsted et al., 2021; Li et al., 2016; Imbert et al., 2022; Anderson et al., 2021). In such systems, agents act online over continuous search spaces without a central coordinator, making sample efficiency important when each new measurement is costly. Under these constraints, independent policies can waste samples by duplicating evaluations, while centralized aggregation is often impractical. This setting matches systems with scheduled medium access and synchronized decision rounds, where small packet exchanges at predetermined time slots reduce contention and power consumption (Gielis et al., 2022; Halsted et al., 2021). These constraints are even stronger in underwater multi-agent systems, where acoustic links are low-bandwidth, high-latency, energy demanding, and unreliable (Fattah et al., 2020; Gussen et al., 2021). Centralized coordination is also difficult at scale because the subsea environment, which acts here as the communication medium, strongly weakens signals with distance and frequency (Thorp, 1967), while also

introducing signal distortions such as multipath propagation and Doppler effects that depend on distance and relative motion. In such contexts, limiting communication to one compact tuple per agent per gossip round is a realistic model of what the communication stack can support.

To coordinate exploration without centralized planning, we propose X-KB-UCB, a decentralized UCB policy that uses gossip to update uncertainty in a way that discourages redundant sampling across agents. At each gossip round, an agent freezes its GP posterior mean, incorporates the most recent arms received from neighbors into its design, recomputes the posterior variance through a cross-agent Kriging–Believer (KB) step (Garnett, 2023b), and then selects the next arm using a UCB rule. Between gossip rounds, each agent runs standard GP-UCB on its locally available data (Srinivas et al., 2012). For nonstationary rewards, we adopt the Markov-drift GP model with a separable space-time kernel (Bogunovic et al., 2016).

The contributions in this work are as follows.

1. **X-KB-UCB.** We propose X-KB-UCB, a decentralized UCB rule that uses a cross-agent KB step to coordinate exploration in continuous domains under gossip communication. Using a cross-agent KB step for coordination is first introduced within this work, to the best of the authors knowledge.
2. **Gossip for time-varying objectives.** We introduce an augmented-agent abstraction (Definition 1 in Section 5) that captures, for each agent, both locally collected data and tuples received through gossip. We use this tool to derive regret bounds for decentralized GP bandits with static and time-varying rewards over a gossip network in continuous domains.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the problem. Section 4 presents the proposed coordination method for decentralized agents. Theoretical guarantees are provided in Section 5. Experimental results are presented in Section 6, followed by concluding remarks in Section 7.

2 Related Work

GP bandits. (Srinivas et al., 2012) introduced GP-UCB and established sublinear regret in the single-agent sequential setting, with bounds controlled by the maximum information gain. In our framework, between gossip rounds, each agent reduces to GP-UCB on its locally available data in the static case.

Centralized batch selection. (Desautels et al., 2014) proposed GP-BUCB for static rewards with parallel or pending evaluations, using Kriging–Believer updates and a mutual-information inflation factor to control overconfidence during batch selection (see Section 5.3). (Parker-Holder et al., 2020) proposed PB2, an alternative batch selection strategy for time-varying objectives. Our setting differs because each agent selects a single arm per round and uses gossip to incorporate past arms observed by neighbors, without centralized batch planning.

Time-varying GP bandits. (Bogunovic et al., 2016) analyzed GP bandits with Markov-drift dynamics and a separable space-time kernel, and derived dynamic regret bounds for TV-GP-UCB in terms of a time-varying information gain. We adopt this model to handle nonstationary rewards in a decentralized network, and between gossip rounds we run the corresponding single-agent update on each agent's locally available data.

Decentralized and cooperative bandits. Prior work studies cooperation under limited communication and network constraints, including distributed and federated approaches for continuous domains with GP models (Chawla et al., 2020; Dubey & Pentland, 2020; Rai & Mou, 2025). (Chawla et al., 2020) consider homogeneous rewards over a discrete action set, which does not directly address continuous action spaces with GP priors. (Rai & Mou, 2025) communicate the most recent (x, y, t) but study independent static reward functions per agent and optimize the average reward across agents rather than a shared objective. (Dubey & Pentland, 2020) address static and contextual settings where rewards differ across agents through context, again differing from a common continuous-domain objective. In contrast, we consider a single noisy reward function shared by all agents over a continuous domain, either static or time-varying, and restrict communication to a constant-size tuple per agent per gossip round.

3 Problem and Notation

We consider multiple agents $a \in \mathcal{A}$ that cooperate via gossip (small messages exchanged among neighbors) to optimize a common time-indexed reward $f_t : D \to \mathbb{R}$ over a continuous action space $D \subset \mathbb{R}^d$.

Time proceeds in discrete iterations $t = 1, \ldots, T_{\text{iter}}$. At iteration t, agent a selects an arm $x_{a,t} \in D$ and observes an immediate noisy reward

$$y_{a,t} = f_t(x_{a,t}) + \varsigma_{a,t}, \qquad \varsigma_{a,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2). \tag{1}$$

Communication is available only at iterations that are multiples of a fixed period $t_g \in \{1, \ldots, T_{\text{iter}}\}$. These are the gossip rounds, during which each agent sends a constant-size message (its most recent $(x_{a,t-1}, y_{a,t-1}, t-1)$) to its neighbors. The parameter t_g controls the communication budget: smaller t_g yields more frequent gossip, larger t_g less frequent gossip.

At iterations $t \in [1, T_{\text{iter}}]$ with $t \bmod t_g = 0$, each agent broadcasts only the previous observation $(x_{a,t-1}, y_{a,t-1}, t-1)$. $N_g = \lfloor T_{\text{iter}}/t_g \rfloor$ is the number of gossip rounds. Let $P \in [0,1]^{|\mathcal{A}|\times|\mathcal{A}|}$ denote a (possibly directed) gossip matrix governing communication: at each gossip round t, agent a receives a tuple from agent v with probability $P(a, v)$. The number of tuples received by agent a in round t is modeled, within this work, as

$$\eta_{a,t} = \sum_{v \in \mathcal{A}, v \neq a} B_{a,v}, \quad B_{a,v} \sim \text{Bernoulli}(P(a,v)) \tag{2}$$

Denote the expected number received tuples per gossip round t by

$$N_a := \mathbb{E}[\eta_{a,t}] = \sum_{v \neq a, v \in \mathcal{A}} P(a,v). \tag{3}$$

The effective number of reward evaluations available to agent a is

$$T_\eta = T_{\text{iter}} + \sum_{k=1}^{N_g} \eta_{a,k}. \tag{4}$$

3.1 Reward model

Each agent searches for the best arm in a continuous domain $D \subset \mathbb{R}^d$. We formalize the two settings as assumptions.

Assumption 1. *The environment is static: $f_t \equiv f$ on D,*

$$f \sim \mathcal{GP}(0, k) \tag{5}$$

k being the kernel function.

Assumption 2. *The environment is time-varying and follows*

$$f_0(x) = g_0(x), \tag{6}$$

$\forall x \in D, \forall t \in \{1, \ldots, T_{iter}\}$,

$$f_t(x) = \sqrt{1-\epsilon} f_{t-1}(x) + \sqrt{\epsilon} g_t(x), \tag{7}$$

with $g_t \sim \mathcal{GP}(0, k)$ independent across t, forgetting factor $\epsilon \in [0,1]$, and k the kernel function (Bogunovic et al., 2016).

Under Assumption 1, each agent aims to identify

$$x^* = \arg\max_{x \in D} f(x). \tag{8}$$

Under Assumption 2, the goal is to track the instantaneous maximizer, $\forall t \in \{1, \ldots, T_{\text{iter}}\}$:

$$x_t^* = \arg\max_{x \in D} f_t(x), \tag{9}$$

4 X-KB-UCB

Our bandit method leverages GP-UCB (Srinivas et al., 2012) under Assumption 1 and TV-GP-UCB (Bogunovic et al., 2016) under Assumption 2. We first recall the GP surrogate used by each agent and the candidate arm selection technique (Section 4.1), then we describe our Kriging–Believer (KB) UCB decision rule (Section 4.2).

4.1 Arm selection

At each time step t, every agent chooses an arm $x_{a,t}$. Each agent builds its own surrogate of the reward function from its history: $\boldsymbol{X}_{a,t-1} = [x_{a,1}, \dots, x_{a,t-1}]$ and $\boldsymbol{Y}_{a,t-1} = [y_{a,1}, \dots, y_{a,t-1}]$.

Under Assumption 1, agent a's surrogate is a Gaussian process, fully specified by its posterior mean $\mu_{a,t}$ and variance $\sigma^2_{a,t}$, for any $x \in D$:

$$\mu_{a,t}(x) = \boldsymbol{k}_{t-1}(x)^\top (\boldsymbol{K}_{t-1} + \sigma_n^2 \boldsymbol{I}_{t-1})^{-1} \boldsymbol{Y}_{a,t-1}, \tag{10}$$

$$\sigma^2_{a,t}(x) = k(x) - \boldsymbol{k}_{t-1}(x)^\top (\boldsymbol{K}_{t-1} + \sigma_n^2 \boldsymbol{I}_{t-1})^{-1} \boldsymbol{k}_{t-1}(x). \tag{11}$$

Here, σ_n^2 is the noise variance, $\boldsymbol{k}_{t-1}(x) = [k(x_i, x)]_{i=1}^{t-1}$, $\boldsymbol{K}_{t-1} = [k(x, x')]_{x,x' \in \boldsymbol{X}_{a,t-1}}$, and $\boldsymbol{I}_{t-1}$ is the $(t-1) \times (t-1)$ identity. We assume $k(x, x) \leq 1$ for all $x \in D$. We use two standard kernels for analysis, for all $(x, x') \in D^2$:

- Squared exponential (SE):
$$k_{SE}(x, x') = \exp\Big(- \tfrac{\|x - x'\|^2}{2l^2} \Big), \tag{12}$$
- Matérn:
$$k_{M^q}(x, x') = \frac{2^{1-q}}{\Gamma(q)} \Big(\frac{\sqrt{2q}\|x - x'\|}{l} \Big)^q B_q \Big(\frac{\sqrt{2q}\|x - x'\|}{l} \Big), \tag{13}$$

where $l > 0$ is the lengthscale and $q > 0$ controls smoothness. Γ is the Gamma function and B_q denotes the modified Bessel function (Rasmussen & Williams, 2006; Garnett, 2023a).

(Bogunovic et al., 2016) extend equation 10–equation 11 to the time-varying setting of Assumption 2 using a separable space–time kernel:

$$k^{dyn}(x_i, x_j) = k(x_i, x_j) \cdot k_s(i, j), \tag{14}$$

where $k_s(i, j) = (1 - \epsilon)^{|i-j|/2}$,, and $i, j \geq 1$

When $t \bmod t_g \neq 0$, the next arm is selected by maximizing the UCB:

$$x_{a,t} = \arg\max_{x \in D} \mu_{a,t}(x) + \alpha_t^{1/2} \sigma_{a,t}(x). \tag{15}$$

The schedule α_t balances exploration and exploitation and is chosen to guarantee high-probability regret bounds (see Section 5).

4.2 Cross-Agent Kriging–Believer

In a multi-agent setting, some mechanism is needed to coordinate exploration across agents. Indeed, if all agents share the same GP prior, observe the same history through gossip, and use the same UCB schedule, then perfect maximization of the acquisition function would make them select the same query points at each iteration. In that case, the agents would collect redundant observations and gossip would provide little additional benefit. For this reason, we incorporate an explicit cross-agent coordination mechanism via a cross-agent KB construction, which promotes diverse queries.

At a gossip round, agent a receives

$$S_{a,t} = \{(x_{v,t-1}, y_{v,t-1}, t-1)\}_{v \in \mathcal{A}_a \subset \mathcal{A}},$$

Algorithm 1 X-KB-UCB (stationary/time-varying reward)

Input: Domain D, gossip period t_g, schedules α_t, β_t, gossip matrix P
Initialize $\mathcal{D}_a \leftarrow \emptyset$ for each agent $a \in \mathcal{A}$
for $t = 1, 2, \ldots, T_{\text{iter}}$ **do**
  if $t \bmod t_g = 0$ **then**
    / gossip round */*
    All agents broadcast $(x_{a,t-1}, y_{a,t-1}, t-1)$
    Agent a receives $\{(x_{v,t-1}, y_{v,t-1}, t-1)\}_{v \in \mathcal{A}_a}$ and forms $S_{a,t}$
    Compute the candidate $x_{a,t}$ using equation 17
    Sample $y_{a,t} = f(x_{a,t}) + \varsigma_{a,t}$
    $\mathcal{D}_{a,t} \leftarrow \mathcal{D}_{a,t} \cup S_{a,t} \cup \{(x_{a,t}, y_{a,t}, t)\}$
  else
    Compute the candidate $x_{a,t}$ using equation 15
    Sample $y_{a,t} = f(x_{a,t}) + \varsigma_{a,t}$
    $\mathcal{D}_{a,t} \leftarrow \mathcal{D}_{a,t} \cup \{(x_{a,t}, y_{a,t}, t)\}$
  end if
  Update the mean with $D_{a,t}$ using equation 10
end for=0

where $\mathcal{A}_a$ is the subset of agents whose messages reached a (as determined by P).

KB is widely used to select multiple nonredundant candidates in batch BO (Desautels et al., 2014; Parker-Holder et al., 2020). We adapt it to the gossip setting: freeze the mean $\mu_{a,t}$, insert the received arms into the design, recompute the variance, and then pick the next arm. More specifically, at a gossip round

$$\boldsymbol{X}_{a,t} = [x_{a,1}, \ldots, x_{a,t-1}] \cup [x_{v,t-1}]_{v \in \mathcal{A}_a}, \tag{16}$$

compute $\widehat{\sigma}^2_{a,t}$ from equation 11 using $\boldsymbol{X}_{a,t}$ (with the mean frozen), and maximize

$$x_{a,t} = \arg\max_{x \in D} \mu_{a,t}(x) + \beta_t^{1/2} \widehat{\sigma}_{a,t}(x). \tag{17}$$

This encourages $x_{a,t}$ to differ from the received arms in $S_{a,t}$ while preserving UCB-driven exploration. The schedule β_t is adapted to gossip rounds to compensate for KB overconfidence (Section 5).

In summary, as shown in Algorithm 1, at a gossip round t, agent a (i) freezes its posterior mean $\mu_{a,t}$, (ii) recomputes the posterior variance $\widehat{\sigma}_{a,t}$ by appending the arms of the received tuples $S_{a,t} = \{(x_{v,t-1}, y_{v,t-1}, t-1)\}_{v \in \mathcal{A}_a}$ to its design, and (iii) selects its candidate $x_{a,t}$ via equation 17. Crucially, after selecting $x_{a,t}$, agent a updates its mean fuction using equation 10 after assimilating both the received tuples and its new sample. After acquisition, the received tuples $S_{a,t} = \{(x_{v,t-1}, y_{v,t-1}, t-1)\}$ are inserted with their original timestamps, and the new tuple $(x_{a,t}, y_{a,t}, t)$ is added, thus $\mathcal{D}_a \leftarrow \mathcal{D}_a \cup S_{a,t} \cup \{(x_{a,t}, y_{a,t}, t)\}$.

Discussion. Cross-Agent UCB (X-KB-UCB) is a novel algorithm introduced for the first time in this paper, to the best of the authors' knowledge. It fundamentally differs from centralized batch selection methods such as GP-BUCB (Desautels et al., 2014) and PB2 (Parker-Holder et al., 2020). In batch BO, one plans a set of B arms and uses KB to collapse posterior variance around those same future candidates within the batch, explicitly pushing the next picks away from them. In X-KB-UCB, at time t we pick a single arm while incorporating a set of received arms from time $t-1$. We then freeze the mean and update only the variance using those received arms.

Under Assumption 2, the time indices matter: received arms correspond to $t-1$, and the Markov-drift prior equation 7 reintroduces uncertainty at time t. The temporal innovation yields a nonzero variance floor at time t even if an arm x was just observed at $t-1$. Hence X-KB-UCB updates (but does not collapse) the uncertainty around received points; larger drift (larger ϵ) naturally increases this residual uncertainty. This contrasts with PB2/GP-BUCB-style intra-batch variance collapse.

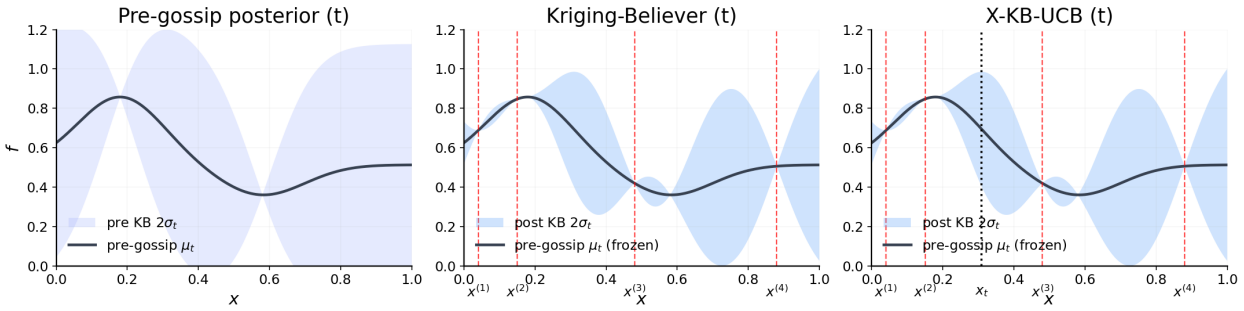


Figure 1: Illustration of the gossip update used by X-KB-UCB at a fixed time t (here $\varepsilon = 0$, no temporal carryover). **Left:** pre-gossip posterior with mean μ_t (gray) and uncertainty band $\pm 2\sigma_t$ (light blue). At the **middle (KB):** Kriging–Believer update using the received evaluations $\{x^{(1)}, \ldots, x^{(4)}\}$: the mean μ_t is frozen and only the standard deviation is updated (blue band). **Right (X-KB-UCB):** action selection with UCB computed from the frozen μ_t and the post-gossip standard deviation, yielding the next query x_t.

An illustration of the X-KB-UCB is shown in Figure 1, which summarizes the technique described in this section using a toy example.

5 Theoretical results

A key property of GP-based UCB methods in continuous domains is *no regret*, i.e., the average cumulative regret vanishes with time (Srinivas et al., 2012):

$$\frac{R_T}{T} := \frac{1}{T} \sum_{t=1}^{T} \big(f(x^*) - f(x_t) \big) \xrightarrow[T \to \infty]{} 0, \tag{18}$$

where x^* is the optimal arm.

We proceed our regret analysis as follows. First, in Section 5.1 we introduce a purely analytical device used to quantify the theoretical benefit of gossip between agents referred to as the ***augmented-agent***. Next, in Section 5.2 we review the maximum information gain, a central quantity in GP regret analyses. We then discuss the overconfidence issue induced by Kriging–Believer in Section 5.3. Finally, we state the main theorems in Section 5.4.

5.1 Augmented agent regrets

We depart from the classic single-agent sequential (GP-UCB/TV-GP-UCB) and batch (GP-BUCB/PB2) settings and adopt an agent-centric notion of feedback that counts both locally played tuples and tuples received via gossip. In this abstraction, we define the following object.

Definitions 1. *An **augmented agent** is an agent whose information set is augmented by the data it receives from its neighbors.*

The augmented agent is responsible for both what it actively queries and what it passively receives. If an agent receives from all neighbors at all iterations, its augmented version operates close to full information and, in a static scenario, is close in spirit to a batch GP method such as GP-BUCB, while its time-varying counterpart is closer to a time-varying batch variant such as PB2 (Parker-Holder et al., 2020). Conversely, if the agent receives no tuples from other agents, the augmented agent reduces to the purely local setting. This construction is meaningful in our framework because the agent is assumed to accept any tuples delivered by gossip whenever communication allows it. The augmented-agent view is a mathematical device, introduced in this work, that enables a regret analysis under gossip sharing. Throughout the remainder of this work, all regret statements are with respect to augmented agents.

$$\mathcal{T}_g := \big\{ t \in \{1, \ldots, T_{\text{iter}}\} : t \bmod t_g = 0 \big\}. \tag{19}$$

At a gossip round $t \in \mathcal{T}_g$, agent a receives

$$S_{a,t} := \{ (v, x_{v,t-1}, t-1) : v \in \mathcal{A}_a(t) \} \tag{20}$$

with $\eta_{a,t}$, as defined in equation 2 is $|S_{a,t}|$. Define the agent's tuple collection

$$\mathcal{Z}_a := \{(a, x_{a,\tau}, \tau) : \tau = 1, \dots, T_{\text{iter}}\} \cup \bigcup_{t \in \mathcal{T}_g} S_{a,t}. \tag{21}$$

Its size is denoted T, and satisfies :

$$\mathbb{E}\big[\,|\mathcal{Z}_a|\,\big] = T_{\text{iter}} + \sum_{t \in \mathcal{T}_g} \mathbb{E}[\eta_{a,t}] = T_{\text{iter}} + N_g N_a := T. \tag{22}$$

Under Assumption 1,

$$R^{st}_{a,T} := \mathbb{E}\left[\sum_{(i,x,\tau) \in \mathcal{Z}_a} \big(f(x^*) - f(x)\big) \right]. \tag{23}$$

Under Assumption 2,

$$R^{dyn}_{a,T} := \mathbb{E}\left[\sum_{(i,x,\tau) \in \mathcal{Z}_a} \big(f_\tau(x^*_\tau) - f_\tau(x)\big) \right]. \tag{24}$$

These are augmented agents' cumulative regrets: they include tuples obtained via gossip as well as locally decided ones, showcasing the benefit of gossip with an expected N_a received tuples per gossip round.

We focus on the upper bounds on this cumulative regrets in equation 23 and equation 24 that imply no-regret for our X-KB-UCB method.

To establish these upper bounds, we have the following assumption:

Assumption 3. *There exist $q_1, q_2 \geq 0$ such that for all $L \geq 0$,*

$$\Pr\Big(\sup_{x \in D} \Big| \frac{\partial f}{\partial x^{(j)}} \Big| \geq L \Big) \leq q_1 \exp\Big(- \big(\tfrac{L}{q_2}\big)^2 \Big), \quad j = 1, \dots, d, \tag{25}$$

with $f \sim \mathcal{GP}(0, k)$. Here $\frac{\partial f}{\partial x^{(j)}}$ refers to the partial derivative of f with respect to its jth elements.

Assumption 3 is a standard smoothness tail bound (used in (Srinivas et al., 2012; Bogunovic et al., 2016)). The SE and Matérn kernels (with $q > 2$) in equation 12–equation 13 satisfy this assumption.

Before the main theorems, we recall two quantities that drive the regret bounds: (i) the *maximum information gain* and (ii) the *KB overconfidence* adjustment at gossip rounds.

5.2 Maximum information gain

The first key quantity is the maximum information gain after B evaluations:

$$\gamma_B = \max_{X \subset D, |X| = B} I(\boldsymbol{f}_X; \boldsymbol{y}_X), \tag{26}$$

where $\boldsymbol{f}_X = [f(x)]_{x \in X}$ and $\boldsymbol{y}_X = [f(x) + \varsigma_x]_{x \in X}$ with independent noise $\varsigma_x \sim \mathcal{N}(0, \sigma_n^2)$. For Gaussian processes,

$$I(\boldsymbol{f}_X; \boldsymbol{y}_X) = \tfrac{1}{2} \log\det\big(\boldsymbol{I} + \sigma_n^{-2} \boldsymbol{K}_X\big) \tag{27}$$

with $\boldsymbol{K}_X = [k(x, x')]_{x, x' \in X}$ and k the kernel (k^{dyn} under Assumption 2). Under Assumption 2, the same identity holds with the space–time Gram matrix $\boldsymbol{K}^{dyn}_X [k^{dyn}(x, x')]_{x, x' \in X}$.

5.3 KB overconfidence

The sequential GP confidence band is

$$CI_t(x) = \big[\mu_t(x) \pm \alpha_t^{1/2}\sigma_t(x)\big]. \tag{28}$$

KB reduces posterior variance using received arms at gossip rounds. The resulting band,

$$CI_{t_g}(x) = \big[\mu_t(x) \pm \beta_t^{1/2}\widehat{\sigma}_t(x)\big], \tag{29}$$

can be narrower, potentially breaking the inclusion

$$f(x) \in CI_t(x) \Rightarrow f(x) \in CI_{t_g}(x) \tag{30}$$

needed to inherit both the sequential GP-UCB and TV-GP-UCB proofs. Following (Desautels et al., 2014), we prevent this by inflating β_t so that $CI_t(x) \subseteq CI_{t_g}(x)$ at gossip rounds. For all $x \in D$,

$$I\big(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}\big) = \tfrac{1}{2}\log\Big(\frac{\sigma_t^2(x)}{\widehat{\sigma}_t^2(x)}\Big). \tag{31}$$

Thus, if $I(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}) \le C_a$ then $\sigma_t(x) \le e^{C_a}\widehat{\sigma}_t(x)$. Choosing

$$\beta_t = e^{2C_a}\alpha_t$$

at gossip rounds yields $\alpha_t^{1/2}\sigma_t(x) \le \beta_t^{1/2}\widehat{\sigma}_t(x)$ and restores the inclusion. An upper bound on C_a is presented in Lemma. 1, and the proof is provided in the supplementary material.

Lemma 1. *Let $\mathcal{A}$ denote the set of agents. At gossip round t, let $S_{a,t}$ be the set of tuples received by agent a, let $\boldsymbol{y}_{S_{a,t}}$ denote the corresponding noisy rewards, with noise variance σ_n^2, and let $\boldsymbol{y}_{1:t-1}$ denote the rewards observed by agent a up to time $t-1$ (including previously gossiped data). Then, for all $x \in D$,*

$$I\big(f_t(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}\big) \le \gamma_{n_{a,t}} \tag{32}$$

$$\le \frac{|\mathcal{A}|-1}{2\sigma_n^2}. \tag{33}$$

where $n_{a,t} = |S_{a,t}|$. In particular, choosing $C_a = \frac{|\mathcal{A}|-1}{2\sigma_n^2}$ guarantees the band inclusion at gossip rounds.

We use schedules that certify the uniform confidence event

$$\Pr\Big(\max_{1\le t\le T}\sup_{x\in D}|f_t(x) - \mu_{a,t}(x)| \le \sqrt{\alpha_t}\,\sigma_{a,t}(x)\Big) \ge 1-\delta, \tag{34}$$

obtained via the standard discretization and time–grid union bound (Srinivas et al., 2012; Bogunovic et al., 2016). The specific information-gain bounds are schedule-independent. At gossip rounds we inflate widths to preserve band inclusion.

One can keep the uniform cap C_a from Lemma 1 and use the inflation factor $\exp(2C_a)$. Alternatively, let $S_{a,t}$ be the received set at round t. It can be shown that the exact mutual information is (see Appendix A.1)

$$C_a \le I(f_{S_{a,t}}; \boldsymbol{y}_{S_{a,t}}) = \frac{1}{2}\log\det(I + \sigma_n^{-2}K_{S_{a,t}}) = C_{a,t}. \tag{35}$$

Using $C_{a,t}$ in place of a uniform C_a leads to a tight, per-round inflation factor $\exp(2C_{a,t})$. Since $|S_{a,t}| \approx N_a$, this can be computed efficiently via incremental Cholesky updates. If one wants a single worst-case budgeted cap for sets of size at most m, one can approximate $\max_{|S|\le m} I(f_S; \boldsymbol{y}_S)$ via greedy maximization, which provides a $(1-1/e)$-approximation due to submodularity, as in the information-gain analysis of GP-UCB (Srinivas et al., 2012).

Remark. C_a inflates α_t at gossip rounds to ensure the theoretical band inclusion. Our implementation uses a common schedule across all iterations and sets $\beta_t = \alpha_t$. This simplifies tuning, while the theoretical guarantees correspond to the inflated version.

5.4 Main theorems

We now present regret bounds for decentralized GP bandits running X-KB-UCB, for both stationary rewards (Thm. 1) and time-varying rewards (Thm. 2). The bounds apply to a fully decentralized setting and account for rewards that drift over time. They are stated for augmented agents, meaning each agent is analyzed together with the tuples it receives through gossip, which makes explicit how performance shifts from the no-gossip regime toward the near-centralized regime as gossip becomes more informative.

Theorem 1. *Let $D \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}$, $r > 0$. Let k be a kernel such that $f \sim \mathcal{GP}(0,k)$ satisfies Assumptions 1 (static reward) and 3 (smoothness tail bound). Fix $\delta \in (0,1)$ and $C_a \geq 0$ with $I(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}) \leq C_a$ at gossip rounds (all agents are* peers*: they all run X-KB-UCB with the same schedules and exchange their previous UCB-selected* arms*). Let α_t be the exploration schedule defined as*

$$\alpha_t^{\mathrm{UCB}} = 2\log\Big(\frac{4\pi t^2}{3\delta}\Big) + 2d\log\Big(t^2\, d\, q_2\, r\,\sqrt{\log\Big(\frac{8d\, q_1}{\delta}\Big)}\Big),$$

, and set

$$\beta_t = \begin{cases} \alpha_t, & t \bmod t_g \neq 0, \\ e^{2C_a}\alpha_t, & t \bmod t_g = 0. \end{cases}$$

Then, for each agent $a \in \mathcal{A}$, with probability at least $1-\delta$,

$$R_{T,a}^{st} \leq \sqrt{2C_1\Big([T+\phi_g]\alpha_{T+\phi_g}\gamma_{T+\phi_g}\big(1+e^{2C_a}\big)\Big)} + 4, \tag{36}$$

where $T = T_{iter} + N_a N_g$, $C_1 = \frac{8}{\log(1+\sigma_n^{-2})}$, and $\phi_g = N_a(T_{iter} - N_g)$ is the gossip gap. This regret bound yields the no-regret property of X-KB-UCB in static environments.

Here, N_a and N_g are defined around equation 4 and equation 3. The proof is provided in the supplementary material. The key point is that, by the Gaussian Mutual Information identity equation 27 the classic and batch confidence intervals are contained in our inflated band at gossip rounds. Also, at non-gossip rounds $\beta_t = \alpha_t$ and the KB variance does not enlarge the band. Therefore we can upper bound the non-gossip rounds part of the regret by the GP-UCB bound (with horizon T) and the gossiped part by the GP-BUCB bound (with horizon $T_B = T_{\mathrm{iter}} + N_a T_{\mathrm{iter}} = T + \phi_g$).

Baselines: if $N_a = 0$ (no information sharing, $P = [0]^{|\mathcal{A}|\times|\mathcal{A}|}$) then $C_a = 0$ (since $S_{a,t} = \emptyset$) and $T = T_{\mathrm{iter}}$, recovering the classic GP-UCB upper bound. If $N_g = 1$ (one gossip), then $T + \phi_g = (T_{\mathrm{iter}} + 1)N_a$, i.e., the largest gap between a single agent GP-UCB on T_{iter} arms and a one-shot GP-BUCB with batch size $1 + N_a$. If $N_g = T_{\mathrm{iter}}$ (gossip every round), then $\phi_g = 0$ and the upper bound behaves like that of GP-BUCB with batch $1 + N_a$. The term $\gamma_{T+\phi_g}$ is the maximum information gain; the bound inherits its standard rates (Srinivas et al., 2012).

Theorem 2. *Let $D \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}$, $r > 0$. Let k be a kernel such that $f \sim \mathcal{GP}(0,k)$ satisfies Assumptions 2 (time-varying reward) and 3. Fix $\delta \in (0,1)$ and $C_a \geq 0$ with $I(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_t) \leq C_a$. Define*

$$\alpha_t := \max\big\{\, \alpha_t^{\mathrm{UCB}},\ \alpha_t^{\mathrm{TV}} \,\big\}. \tag{37}$$

where

$$\alpha_t^{\mathrm{UCB}} = 2\log\Big(\frac{4\pi t^2}{3\delta}\Big) + 2d\log\Big(t^2\, d\, q_2\, r\,\sqrt{\log\Big(\frac{8d\, q_1}{\delta}\Big)}\Big), \tag{38}$$

$$\alpha_t^{\mathrm{TV}} = 2\log\Big(\frac{\pi^2 t^2}{2\delta}\Big) + 2d\log\Big(r\, d\, q_2\, t^2\,\sqrt{\log\Big(\frac{d\, q_1\, \pi^2 t^2}{2\delta}\Big)}\Big), \tag{39}$$

and use the same schedule $\beta_t = \alpha_t$ when $t \bmod t_g \neq 0$ and $\beta_t = e^{2C_a}\alpha_t$ when $t \bmod t_g = 0$. Then, for each agent $a \in \mathcal{A}$, with probability at least $1-\delta$, and for any $\tilde{N} \in \{1,\dots,T\}$,

$$R_{a,T}^{dyn} \leq \sqrt{2C_1\Big(\gamma_{\tilde{N}} + \tilde{N}^3\epsilon\Big)\Big[M_T + M_{T+\phi_g}\Big]} + 4 \tag{40}$$

with

$$M_T = T\alpha_T\Big(\frac{T}{\tilde{N}} + 1\Big) \tag{41}$$

$$M_{T+\phi_g} = (T+\phi_g)e^{2C_a}\alpha_{T+\phi_g}\Big(\frac{T+\phi_g}{\tilde{N}} + 1\Big) \tag{42}$$

where $T = T_{iter} + N_a N_g$, $C_1 = \frac{8}{\log(1+\sigma_n^{-2})}$, $\phi_g = N_a(T_{iter} - N_g)$ *is the gossip gap, and* ϵ *is the forgetting factor in equation 7.*

The proof is similar to that of Theorem 1, and leverages the regret bounds derived in (Bogunovic et al., 2016) and (Desautels et al., 2014). In the time-varying GP model of Bogunovic et al. (2016, Cor. 4.1), the average regret vanishes whenever the drift decreases polynomially with the horizon: if $\epsilon \leq T^{-p}$ for some $p > 0$, then $R^{\text{dyn}}_{a,T}/T \to 0$. Conversely, for any fixed $\epsilon > 0$, the average regret does not vanish.

For the SE kernel, $\gamma_T = \mathcal{O}\big((\log T)^{d+1}\big)$; for Matérn-3/2, $\gamma_T = \mathcal{O}\big(T^{\frac{d(d+1)}{3+d(d+1)}}\log T\big)$ (Srinivas et al., 2012).

6 Experimental Results

This section studies the effect of three system parameters, the number of agents N, the gossip period t_g, and the expected number of received tuples per round N_a, in both static and time-varying settings. All results are averaged over 30 Monte Carlo runs, and the reported regrets are augmented-agent regrets, averaged over N. The gossip matrix P is sampled uniformly and then rescaled to match a target N_a, where larger N_a corresponds to higher connectivity. With $N = 1$, the method reduces to GP-UCB in the static case and to TV-GP-UCB in the time-varying case. We use $\beta_t = \log(t)$ and inflate the kernel output scale to widen posterior bands. Additional details and experiments are provided in Appendix A.4.

Section 6.1 reports a static synthetic study on the four-dimensional Styblinski–Tang function equation 75 with $T_{\text{iter}} = 200$. Section 6.2 reports a controlled time-varying synthetic study with $T_{\text{iter}} = 100$ and includes the effect of mismatches between the true dynamics and the TV-GP surrogate. Section 6.3 reports results on the Intel sensor temperature dataset (Madden, 2004) in a time-varying setting, following (Bogunovic et al., 2016).

6.1 Synthetic static reward function

Throughout this section, The hyperparameters of the GPs (the lengthscale and the outputscale) of each agent are learned with every iteration $t \in [1, \ldots, T_{iter}]$ using maximum log-likelihood (see Rasmussen & Williams (2006)). The iteration budget is fixed to T_{iter}. With N agents in parallel, each agent runs $\lfloor T_{\text{iter}}/N \rfloor$ iterations, so wall-clock time effectively scales by N.

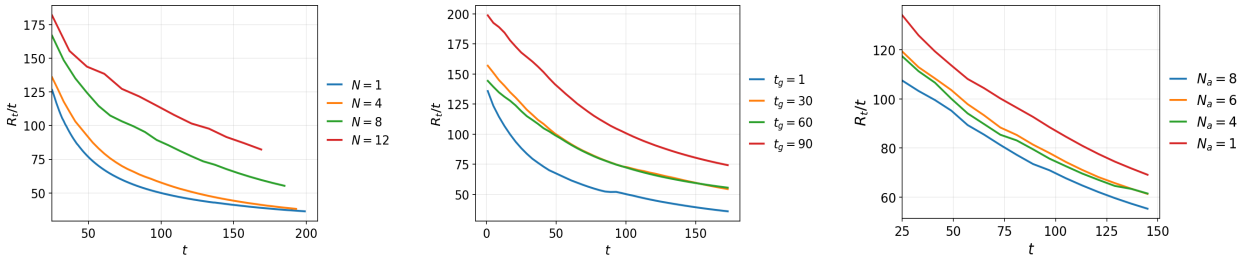
Figure 2 reports performance using the per-agent average regret $R^{\text{st}}_{a,T}/T$ (lower is better), averaged across agents. The number of points in the displayed curves are lower that T_{iter} due to the fact that the gossip matrix P is random and that the there is no gossip at $t = 1$.

Figure 2a considers $t_g = 1$ (gossip at every iteration) with a fully connected gossip matrix $P = \mathbf{1}^{|\mathcal{A}|\times|\mathcal{A}|}$. Because work is spread across agents, the speedup is close to linear: with N agents we nearly match an N-times faster single-agent run. A small gap appears as N increases, which we attribute to the Kriging–Believer step shrinking variance around received arms and slightly tempering exploration.

Figure 2b fixes $N = 4$ and a randomly generated P with $N_a = N$, and varies the gossip period t_g. More frequent communication (smaller t_g) consistently reduces the average regret.

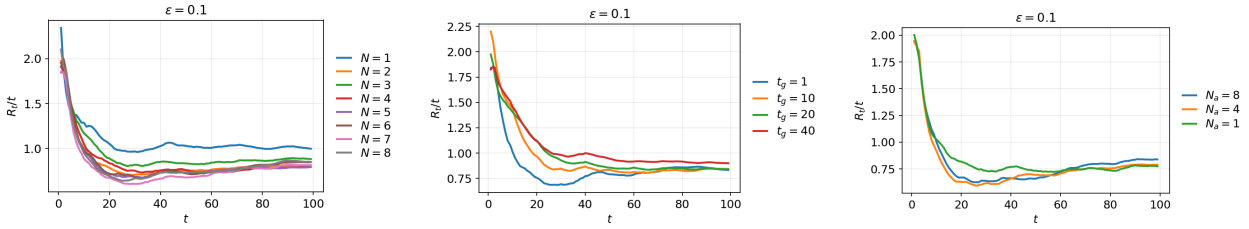
Figure 2c varies the effective connectivity of the gossip network through N_a, the expected number of received tuples per round. Larger N_a improves coordination among agents and lowers regret.

In summary, for static environments, we recommend keeping the gossip period t_g small whenever budget permits. Connectivity N_a is dependent on the environment and is independent of user choices. Further results on other synthetic functions are presented in Appendix A.4.1.



(a) Impact of number of agents N. (b) Impact of gossip period t_g. (c) Impact of N_a.

Figure 2: Impact of system parameters on the average regret over the Styblinski-Tang function. Each panel isolates a factor: (a) number of agents, (b) gossip period, and (c) average number of received tuples N_a.



(a) Impact of number of agents N. (b) Impact of gossip period t_g. (c) Impact of N_a.

Figure 3: Time-varying setting: impact of system parameters on the average cumulative regret. Each panel isolates a factor: (a) number of agents, (b) gossip period, and (c) average number of received tuples per round.

6.2 Synthetic Time-Varying reward function

We consider a time-varying reward that follows the Markov model in Assumption 2. We generate a two-dimensional function with lengthscale 0.2 and unit output scale; the drift parameter is fixed to $\epsilon = 0.1$. For simplicity, the agents are assumed to know ϵ, the lengthscale, and the output scale.

Similarly to Section 6.1, we first study the effect of the number of agents N. In Figure 3a, agents communicate every iteration ($t_g = 1$) over a fully connected network $P = \mathbf{1}^{|\mathcal{A}|\times|\mathcal{A}|}$. Increasing N substantially improves tracking compared to the single-agent case, reflecting the benefit of parallel exploration under frequent gossip.

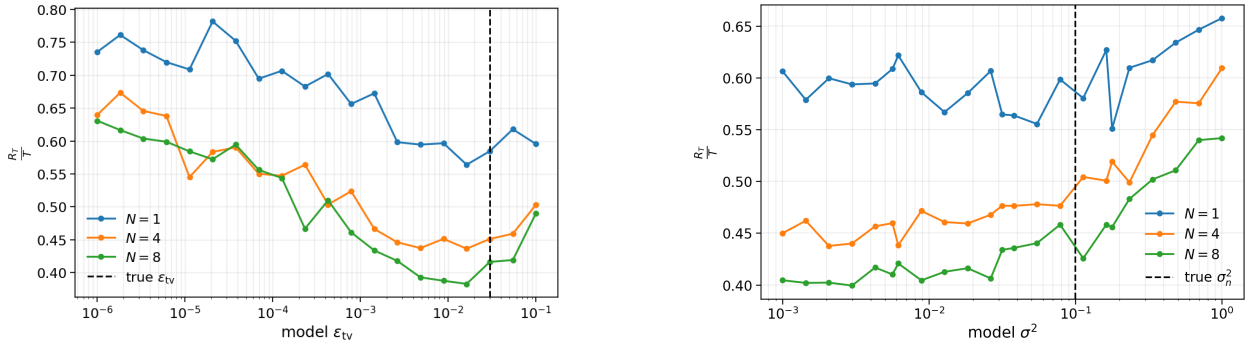
Next, Figure 3b shows the sensitivity to the gossip period t_g with $N = 8$ and a fully connected network $P = \mathbf{1}^{|\mathcal{A}|\times|\mathcal{A}|}$. More frequent gossip (smaller t_g) consistently improves tracking, with a pronounced gain relative to the single-agent baseline.

Finally, Figure 3c examines the effect of network connectivity for $N = 8$ by varying the expected number of received tuples per round N_a. The curves are close, indicating a limited sensitivity to N_a in this setup; in other words, performance is relatively robust to the particular connectivity pattern encoded by P.

Another important property of the proposed technique is its resilience to mismatches between the drifting reward and the surrogate model used by each agent or the observation noise variance of the reward σ_n^2. In practice, each agent can tune its own hyperparameters by maximizing the marginal likelihood of its GP surrogate, as discussed in the supplementary material of (Bogunovic et al., 2016), provided that enough observations are available. Throughout this section, we assumed a match between the reward function and the agents' surrogate models.

To test robustness to mismatch, Figure 4 reports an example involving two hyperparameters, the observation noise σ_n^2 in equation 1 and the drift parameter ϵ_{TV}. Panel (a) shows the final average augmented regret obtained when the true reward drifts at rate $\epsilon_{\text{TV}} = 0.03$ while the agent surrogate models use different values of ϵ_{TV}. We repeat the experiment for different numbers of agents N over 30 Monte Carlo runs. Across all

cases, the final average augmented regret degrades smoothly and remains robust under moderate mismatch, in line with (Bogunovic et al., 2016). For simplicity, the impact of mismatch in other modeling choices, such as the kernel type and GP kernel hyperparameters, is left for future work.



(a) Drift parameter mismatch.

(b) Observation noise mismatch.

Figure 4: Final average augmented regret for different numbers of agents, namely $N = 1$, $N = 4$, and $N = 8$, under mismatch in the model drift parameter and the observation noise.

6.3 Real-data reward function

In this section, the proposed algorithm is tested in both a static scenario and a time-varying scenario. We highlight the impact of gossip when the expected number of received tuples per agent is $N_a = 0.8$, for three values of the gossip period $t_g \in \{1, 10, 100\}$.

We use the Intel Lab dataset (Madden, 2004), collected from sensors deployed in the Intel Berkeley Research Lab. As in (Bogunovic et al., 2016), we restrict the study to the first 46 sensors, since beyond sensor 46 the optimum becomes fixed and the problem becomes trivial. The task is then to track over time the sensor index with the highest temperature among these 46 sensors. Each agent uses a time-varying GP surrogate with drift parameter $\epsilon_{\mathrm{TV}} = 0.006$, which we found to work well empirically. The remaining surrogate hyperparameters, using the squared exponential kernel k_{SE} in equation 12, are learned by maximum likelihood on the first 500 datapoints. The algorithm is then run on the next 1500 datapoints (iterations) with these learned hyperparameters. Figure 5 reports the average cumulative regret over time, averaged across agents.As in the static case, the cumulative regret decreases as cooperation increases, with a clear gain relative to the single-agent baseline.

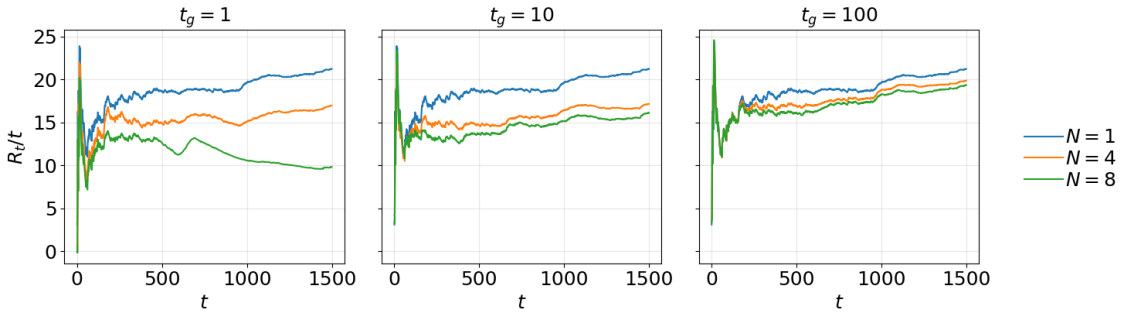


Figure 5: X-KB-UCB performance on the sensor temperature dataset for three networks of agents $N = 1$, $N = 4$, and $N = 8$. Each panel corresponds to a different gossip period t_g, with $N_a = 0.8$.

7 Conclusion

We studied decentralized GP bandits under strict communication budgets, in both static and time-varying settings. We introduced X-KB-UCB, a decentralized UCB policy that assimilates gossiped observations through a cross-agent KB step to coordinate exploration while keeping decision making local. To analyze the effect of gossip, we relied on an augmented-agent abstraction that makes explicit what information an agent has after receiving tuples from its neighbors, and it lets us compare regimes ranging from isolated learning (no gossip) to near-centralized information sharing (frequent gossip). We proved high-probability no-regret guarantees in the static case and extended them to time-varying rewards under a Markov-drift model. Experiments supported the analysis, showing that gossip improves sample efficiency on static objectives and improves tracking performance when the maximizer drifts.

Several avenues merit further work. First, it would be useful to study mismatches between the true reward kernel and the kernel assumed by the agents' surrogates. Second, throughout this work, we assumed homogeneous agents using GP surrogates with the same kernel. Allowing heterogeneity, for example agents using different kernels or operating under different drift rates, is a natural extension. Third, the cross-agent KB coordination step could be tested in settings with non-identical reward functions across agents and under communication constraints induced by privacy. Finally, we considered synchronous communication, and extending the analysis to asynchronous gossip with delays or dropped messages is an important direction.

References

J. Anderson, M. Eich, M. T. Wolf, and S. Rock. Communication planning for cooperative terrain-based navigation of multiple auvs. *Sensors*, 21(5):1675, 2021. doi: 10.3390/s21051675. URL `https://www.mdpi.com/1424-8220/21/5/1675`.

Ilija Bogunovic, Jonathan Scarlett, and Volkan Cevher. Time-varying gaussian process bandit optimization. In *Artificial Intelligence and Statistics*, pp. 314–323. PMLR, 2016.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert–eliminate algorithm for multi-agent bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, 2020. URL `https://proceedings.mlr.press/v108/chawla20a/chawla20a.pdf`.

Cheikh M. Cheikh Melainine, François-Xavier Socheleau, and Arnaud Jarrot. Sequential optimization of decision feedback equalizer hyperparameters in mobile acoustic channels. In *OCEANS 2025 Brest*, pp. 1–6, 2025. doi: 10.1109/OCEANS58557.2025.11104411.

Thomas Desautels, Andreas Krause, and Joel W Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *J. Mach. Learn. Res.*, 15(1):3873–3923, 2014.

Abhimanyu Dubey and Alex Pentland. Kernel methods for cooperative multi-agent contextual bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL `https://proceedings.mlr.press/v119/dubey20b.html`.

Salmah Fattah, Abdullah Gani, Ismail Ahmedy, Mohd Yamani Idna Idris, and Ibrahim Abaker Targio Hashem. A survey on underwater wireless sensor networks: Requirements, taxonomy, recent advances, and open research challenges. *Sensors*, 20(18):5393, 2020.

Roman Garnett. *Modeling with Gaussian Processes*, pp. 45–66. Cambridge University Press, Cambridge, 2023a.

Roman Garnett. *Common Bayesian Optimization Policies*, pp. 123–156. Cambridge University Press, Cambridge, 2023b.

Jennifer Gielis, Ajay Shankar, and Amanda Prorok. A critical review of communications in multi-robot systems. *Current robotics reports*, 3(4):213–225, 2022.

Camila M. G. Gussen, Christophe Laot, François-Xavier Socheleau, Benoît Zerr, Thomas Le Mézo, Raphaël Bourdon, and Céline Le Berre. Optimization of acoustic communication links for a swarm of auvs: The comet and nemosens examples. *Applied Sciences*, 11(17):8200, 2021. doi: 10.3390/app11178200. URL `https://www.mdpi.com/2076-3417/11/17/8200`.

Trevor Halsted, Ola Shorinwa, Javier Yu, and Mac Schwager. A survey of distributed optimization methods for multi-robot systems. *arXiv preprint arXiv:2103.12840*, 2021.

Stephane Imbert, Christophe Laot, Abdel-Ouahab Boudraa, and Jean-Jacques Szkolnik. Performance optimization of underwater communication links at different ranges for ais relay to auv. *Applied Sciences*, 12 (9):4166, 2022.

Ning Li, José-Fernán Martínez, Juan Manuel Meneses Chaus, and Martina Eckert. A survey on underwater acoustic sensor network routing protocols. *Sensors*, 16(3):414, 2016.

Samuel Madden. Intel lab data. `https://db.csail.mit.edu/labdata/labdata.html`, 2004. Data collected from 54 sensors deployed in the Intel Berkeley Research Lab, February 28–April 5, 2004. Accessed: 2026-03-23.

Jack Parker-Holder, Vu Nguyen, and Stephen J. Roberts. Provably efficient online hyperparameter optimization with population-based bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/c7af0926b294e47e52e46cfebe173f20-Paper.pdf`.

Ayush Rai and Shaoshuai Mou. Distributed optimization via kernelized multi-armed bandits. *IEEE Transactions on Automatic Control*, pp. 1–16, 2025. doi: 10.1109/TAC.2025.3574032.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012. doi: 10.1109/TIT.2011.2182033.

William H. Thorp. Analytic description of the low-frequency attenuation coefficient. *The Journal of the Acoustical Society of America*, 42(1):270–270, 1967. doi: 10.1121/1.1910566.

A Appendix

We work under the standing assumptions in the main text: $D \subset [0, r]^d$ compact and convex; $k(x, x) \leq 1$; i.i.d. Gaussian noise with variance σ_n^2 within noisy observations. All agents are *peers*: they run the same X-KB-UCB schedules and broadcast their previous UCB-selected *arms* at gossip.

A.1 Proof of Lemma 1

Fix a gossip round t and agent a, and write $S := S_{a,t}$, $\boldsymbol{f}_S := [f_{t-1}(x)]_{(x,t-1)\in S}$, and $y_S = \boldsymbol{f}_S + \varsigma_S$ with $\varsigma_S \sim \mathcal{N}(0, \sigma_n^2 I)$ independent of everything. We have:

$$I\big(f_t(x); y_S \mid \boldsymbol{y}_{1:t-1}\big) \leq I\big(f_t(x), \boldsymbol{f}_S; y_S \mid \boldsymbol{y}_{1:t-1}\big) = I\big(\boldsymbol{f}_S; y_S \mid \boldsymbol{y}_{1:t-1}\big). \tag{43}$$

because, augmenting the first argument cannot decrease MI. Moreover $y_S \perp (f(x,t), \boldsymbol{y}_{1:t-1}) \mid \boldsymbol{f}_S$ because $\boldsymbol{y}_S = \boldsymbol{f}_S + \varsigma_S$. Using $I(A; B \mid C) = H(B \mid C) - H(B \mid A, C)$ and $H(B \mid C) \leq H(B)$,

$$I\big(\boldsymbol{f}_S; y_S \mid \boldsymbol{y}_{1:t-1}\big) \leq I\big(\boldsymbol{f}_S; y_S\big). \tag{44}$$

$\forall x \in D, k(x,x) \leq 1$ and $k^{dyn}(x,x) \leq 1$. By equation 27 we have : $I\big(\boldsymbol{f}_S; y_S\big) = \frac{1}{2}\log\det\Big(I + \sigma_n^{-2}K_S\Big)$. Use $\log\det(I+M) \leq \mathrm{tr}(M)$ for $M \succeq 0$ and $\mathrm{tr}(K_S) \leq |S|$ and $|S| \leq |\mathcal{A}| - 1$:

$$\frac{1}{2}\log\det\Big(I + \sigma_n^{-2}K_S\Big) \leq \frac{1}{2}\sigma_n^{-2}\mathrm{tr}(K_S) \leq \frac{|S|}{2\sigma_n^2} \leq \frac{|\mathcal{A}|-1}{2\sigma_n^2}. \tag{45}$$

□

A.2 Theorem 1 (Stationary reward)

Consider f under assumptions 1 and 3. We first recall the high-probability confidence sets for GP-UCB (Srinivas et al., 2012) and the batch inflation argument of GP-BUCB (Desautels et al., 2014). Using C_a provided with Lemma 1,

$$I\big(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}\big) = \tfrac{1}{2}\log\frac{\sigma^2_{a,t}(x)}{\widehat{\sigma}^2_{a,t}(x)} \leq C_a \tag{46}$$

implies the band inclusion

$$\alpha_t^{1/2}\sigma_{a,t}(x) \leq (e^{2C_a}\alpha_t)^{1/2}\widehat{\sigma}_{a,t}(x) = \beta_t^{1/2}\widehat{\sigma}_{a,t}(x), \quad \forall x \in D, \tag{47}$$

so the classic and batch UCB bands are contained in our X-KB-UCB band at gossip rounds. At non-gossip rounds $\beta_t = \alpha_t$, and the KB variance does not enlarge the band, hence the inclusion. This inclusion enables the inheritance of regret bounds from both the GP-UCB and BUCB as highlighted in (Desautels et al., 2014).

A.2.1 GP-UCB and GP-BUCB

The single-agent GP-UCB bound the regret $R_T = \sum_{t=1}^{T}(f(x^*) - f(x_t))$ as follows:

Theorem 3 (Srinivas et al. (2012))**.** *Let $D \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}$, $r > 0$. Let k be a kernel such that $f \sim \mathcal{GP}(0,k)$ satisfies Assumptions 1 and 3. Fix $\delta \in (0,1)$, and define $\alpha_t = 2\log\left(\frac{4\pi t^2}{3\delta}\right) + 2d\log\left(t^2 dq_2 r\sqrt{\log\frac{8dq_1}{\delta}}\right)$,then with probability at least $1-\delta/2$:*

$$R_T \leq \sqrt{C_1 T\alpha_T\gamma_T} + 2 \tag{48}$$

where $C_1 = \frac{8}{\log(1+\sigma_n^{-2})}$.

The centralized GP-BUCB, where at each iteration t, we choose a batch of B *arms* using KB and parallelize reward computation, bounds $R_{T_B} = \sum_{t=1}^{T_B}(f(x^*) - f(x_t))$, $T_B = B \cdot T$, as follows.

Theorem 4 (Desautels et al. (2014))**.** *Let $D \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}$, $r > 0$. Let k be a kernel such that $f \sim \mathcal{GP}(0,k)$ satisfies Assumptions 1 and 3. Fix $\delta \in (0,1)$, and $C_a > 0$ with $I(f(x); \boldsymbol{y}_{S_{a,t}} \mid \boldsymbol{y}_{1:t-1}) \leq C_a$. Define $\alpha_t = 2\log\left(\frac{4\pi t^2}{3\delta}\right) + 2d\log\left(t^2 dq_2 r\sqrt{\log\frac{8dq_1}{\delta}}\right)$, then with probability at least $1-\delta/2$:*

$$R_{T_B} \leq \sqrt{C_1 T_B e^{2C_a}\alpha_{T_B}\gamma_{T_B}} + 2. \tag{49}$$

A.2.2 Proof of Theorem 1

X-KB-UCB uses the single-agent GP-UCB rule at non-gossip rounds ($t \bmod t_g \neq 0$) and applies KB at gossip rounds ($t \bmod t_g = 0$). Using the regret of an augmented agent equation 23, we decompose the static regret into the sum over locally played tuples at non-gossip times and received tuples at gossip times:

$$Z_{ng} := \{(a, x_{a,t}, t) : t \in \{1,\dots,T_{\mathrm{iter}}\},\ t \bmod t_g \neq 0\}, \qquad Z_g := \bigcup_{t\in\{1,\dots,T_{\mathrm{iter}}\},\ t \bmod t_g = 0} S_{a,t}. \tag{50}$$

$$R^{st}_{a,T} = \sum_{(i,x,\tau)\in Z_{ng}} \big(f(x^*) - f(x)\big) + \mathbb{E}\Bigg[\sum_{(i,x,\tau)\in Z_g} \big(f(x^*) - f(x)\big)\Bigg]. \tag{51}$$

The expectation $\mathbb{E}$ is over the number of received gossip tuples. The expected count

$$T := \mathbb{E}[|\mathcal{Z}_a|] = T_{\text{iter}} + N_g N_a, \tag{52}$$

and defining

$$T_B = T_{\text{iter}} + N_a T_{\text{iter}} = T + N_a(T_{\text{iter}} - N_g) =: T + \phi_g, \tag{53}$$

the band inclusion equation 47 implies that the classic GP-UCB confidence set is contained in our band at non-gossip rounds and the batch-inflated one is contained at gossip rounds. Therefore, under the events of Theorems 3 and 4 (each holding with probability $1-\delta/2$), and using that cumulative regret is non-decreasing in the horizon,

$$E_1: \quad \sum_{(i,x,\tau)\in Z_{ng}} \big(f(x^*) - f(x)\big) \le R_T \le \sqrt{C_1 T \alpha_T \gamma_T} + 2, \tag{54}$$

$$E_2: \quad \mathbb{E}\Bigg[\sum_{(i,x,\tau)\in Z_g} \big(f(x^*) - f(x)\big)\Bigg] \le R_{T_B} \le \sqrt{C_1 T_B e^{2C_a} \alpha_{T_B} \gamma_{T_B}} + 2. \tag{55}$$

Using $P(E_1 \cap E_2) \ge P(E_1) + P(E_2) - 1$, with probability at least $1-\delta$,

$$R^{st}_{T,a} \le \sqrt{C_1 T \alpha_T \gamma_T} + \sqrt{C_1 T_B e^{2C_a} \alpha_{T_B} \gamma_{T_B}} + 4. \tag{56}$$

Finally, by Cauchy-Schwarz inequality,

$$R^{st}_{T,a} \le \sqrt{2C_1\Big(T\alpha_T\gamma_T + (T+\phi_g)e^{2C_a}\alpha_{T+\phi_g}\gamma_{T+\phi_g}\Big)} + 4, \tag{57}$$

and since α_t and γ_t are non-decreasing, replacing both terms by their values at $T+\phi_g$ yields

$$R^{st}_{T,a} \le \sqrt{2C_1\Big([T+\phi_g]\alpha_{T+\phi_g}\gamma_{T+\phi_g}\big(1+e^{2C_a}\big)\Big)} + 4. \tag{58}$$

This proves the theorem. □

A.3 Theorem 2 (Time-Varying reward)

Consider f_t under assumptions 2 and 3. We first recall the high-probability upper bound of the regret in TV-GP-UCB (Bogunovic et al., 2016).

A.3.1 Single agent TV-GP-UCB

Regret in the single agent TV-GP-UCB framework is defined analogously to equation 24,

$$R^{single}_T = \sum_{t=1}^{T} f_t(x^*_t) - f_t(x_t) \tag{59}$$

highlighting that this is a tracking problem in contrast with Assumption 1.

Theorem 5 (Bogunovic et al. (2016)). *Let $D \subset [0,r]^d$ be compact and convex. Assume the time-varying model in Assumption 2 and the smoothness tail bound in Assumption 3 with constants (q_1, q_2). Fix $\delta \in (0,1)$, and define $\alpha_T = 2\log\left(\frac{\pi^2 T^2}{\delta}\right) + 2d\log\left(rdq_2T^2\sqrt{\log\frac{dq_1\pi^2T^2}{\delta}}\right)$. Let $C_1 = \frac{8}{\log(1+\sigma_n^{-2})}$. Then the TV-GP-UCB algorithm satisfies, with probability at least $1-\delta/2$,*

$$R^{single}_T \le \sqrt{C_1 T \alpha_T \tilde{\gamma}_T} + 2, \tag{60}$$

and moreover, for any $\tilde{N} \in \{1, \dots, T\}$,

$$R_T^{single} \leq \sqrt{C_1 T \alpha_T \Big(\frac{T}{\tilde{N}} + 1\Big)\Big(\gamma_{\tilde{N}} + \tilde{N}^3 \epsilon\Big)} + 2. \tag{61}$$

Here

$$\tilde{\gamma}_T = \max_{X \subset D, |X| = T} \frac{1}{2} \log\det\Big(\boldsymbol{I}_T + \sigma_n^{-2} \boldsymbol{K}_T^{dyn}\Big), \tag{62}$$

with $\boldsymbol{K}_T^{dyn} = \big[k(x_i, x_j)(1-\epsilon)^{|i-j|/2}\big]_{(i,j)\in[1,T]^2, (x_i,x_j)\in X^2}$, *and* $\gamma_{\tilde{N}}$ *defined as in equation 26.*

A.3.2 Proof of Theorem 2

Both (Bogunovic et al., 2016) and (Desautels et al., 2014) follow the same steps for bounding the regret as in (Srinivas et al., 2012): (i) build a time-dependent ε_t-net for D using the derivative tail bound (Assumption 3); (ii) apply Gaussian concentration on the grid and union-bound jointly over time and grid points. This is where the expression of the schedule differ. (iii) lift the grid bound to the continuum; and (iv) convert $\sum_t \sigma_t^2(\cdot)$ into an information-gain term (static γ_T or time-varying $\tilde{\gamma}_T$). The exploration schedule is used only to certify the uniform confidence event $|f_t(x) - \mu_t(x,t)| \leq \alpha_t^{1/2} \sigma_t(x,t)$; the maximum-information-gain bounds are schedule-independent.

$$\alpha_t := \max\big\{\, \alpha_t^{\mathrm{UCB}},\ \alpha_t^{\mathrm{TV}} \,\big\}. \tag{63}$$

where

$$\alpha_t^{\mathrm{UCB}} = 2\log\Big(\frac{4\pi t^2}{3\delta}\Big) + 2d\log\Big(t^2\, d\, q_2\, r\, \sqrt{\log\Big(\frac{8d\, q_1}{\delta}\Big)}\Big), \tag{64}$$

$$\alpha_t^{\mathrm{TV}} = 2\log\Big(\frac{\pi^2 t^2}{2\delta}\Big) + 2d\log\Big(r\, d\, q_2\, t^2\, \sqrt{\log\Big(\frac{d\, q_1\, \pi^2 t^2}{2\delta}\Big)}\Big), \tag{65}$$

Since α_t is the pointwise maximum of two valid schedules, its band contains both, and by a union bound (splitting δ if desired) the same high-probability event holds. Consequently, all downstream steps (variance-to-information-gain conversion and MI inflation at gossip) go through unchanged, and we inherit the regret guarantees while paying only constant factors through $\sqrt{\alpha_T}$.

Similarly with section A.2.2, we decompose the cumulative regret in equation 24 into gossip and non-gossip parts.

$$Z_{ng} := \{(x_{a,t}, t) : t \in \{1, \dots, T_{\mathrm{iter}}\},\ t \bmod t_g \neq 0\}, \qquad Z_g := \bigcup_{t\in\{1,\dots,T_{\mathrm{iter}}\},\ t \bmod t_g = 0} S_{a,t}. \tag{66}$$

$$R_{a,T}^{dyn} = \mathbb{E}\Bigg[\sum_{(x,\tau)\in Z_a} \big(f_\tau(x^*) - f_\tau(x)\big)\Bigg] = \sum_{(x,\tau)\in Z_{ng}} \big(f_\tau(x^*) - f_\tau(x)\big) + \mathbb{E}\Bigg[\sum_{(x,\tau)\in Z_g} \big(f_\tau(x^*) - f_\tau(x)\big)\Bigg]. \tag{67}$$

We use T and ϕ_g equation 53 from section A.2.2. Under Assumptions 2 and 3, and using the schedule α_t we have, with probability at least $1 - \delta/2$:

$$E_1 : \sum_{(x,\tau)\in Z_{ng}} \big(f_\tau(x^*) - f_\tau(x)\big) \leq R_T^{single} \leq \sqrt{C_1 T \alpha_T \tilde{\gamma}_T} + 2 \tag{68}$$

and

$$E_2 : \mathbb{E}\Bigg[\sum_{(x,\tau)\in Z_g} \big(f_\tau(x^*) - f_\tau(x)\big)\Bigg] \leq R_{T_B} \leq \sqrt{C_1 T_B e^{2C_a} \alpha_{T_B} \tilde{\gamma}_{T_B}} + 2. \tag{69}$$

Note that the maximum mutual information $\tilde{\gamma}_T$ and $\tilde{\gamma}_{T_B}$ have the exact same expressions as equation 26. What changes is the kernel matrix in equation 27, where K_T^{dyn} and $K_{T_B}^{dyn}$ are used to compute the Mutual information. In (Bogunovic et al., 2016) They provide an upper bound of such quantities using the maximum

mutual information at each in the stationary domain, hence equation 61. This gives us, probability at least $1-\delta/2$:

$$E_1 : \sum_{(x,\tau)\in Z_{ng}} \big(f_\tau(x^*) - f_\tau(x)\big) \leq R_T^{single} \leq \sqrt{C_1 T \alpha_T \Big(\frac{T}{\tilde{N}} + 1\Big)\Big(\gamma_{\tilde{N}} + \tilde{N}^3\epsilon\Big)} + 2, \qquad \forall \tilde{N} \in [1, T] \tag{70}$$

and

$$E_2 : \mathbb{E}\left[\sum_{(x,\tau)\in Z_g} \big(f_\tau(x^*) - f_\tau(x)\big)\right] \leq R_{T_B} \leq \sqrt{C_1 T_B e^{2C_a} \alpha_{T_B} \Big(\frac{T_B}{\tilde{N}} + 1\Big)\Big(\gamma_{\tilde{N}} + \tilde{N}^3\epsilon\Big)} + 2, \qquad \forall \tilde{N} \in [1, T_B]. \tag{71}$$

Hence, with probability at least $1-\delta$, and choosing the same $\tilde{N}$ for both events, we get :

$$R_{a,T}^{dyn} \leq \sqrt{2C_1\Big(\gamma_{\tilde{N}} + \tilde{N}^3\epsilon\Big)\left[T\alpha_T\Big(\frac{T}{\tilde{N}} + 1\Big) + (T + \phi_g)e^{2C_a}\alpha_{T+\phi_g}\Big(\frac{T+\phi_g}{\tilde{N}} + 1\Big)\right]} + 4, \qquad \forall \tilde{N} \in [1, T]. \tag{72}$$

□

A.4 Experimental Details

This section provides additional information complementing Section 6.

A.4.1 Synthetic Static Reward

The following benchmark functions are used to evaluate X-KB-UCB in static environments:

- **Rastrigin (d=4)** :

$$f_{\mathrm{Ras}}(x) = 10d + \sum_{i=1}^{d}\big(x_i^2 - 10\cos(2\pi x_i)\big), x \in [-5.12,\, 5.12]^4. \tag{73}$$

- **Rosenbrock (d=3)**:

$$f_{\mathrm{Ros}}(x) = \sum_{i=1}^{d-1}\Big(100\big(x_{i+1} - x_i^2\big)^2 + (1 - x_i)^2\Big), x \in [-5,\, 10]^3. \tag{74}$$

- **Styblinski-Tang (d=4)**:

$$f_{\mathrm{ST}}(x) = \tfrac{1}{2}\sum_{i=1}^{d}\big(x_i^4 - 16x_i^2 + 5x_i\big), \quad x \in [-5,\, 5]^4. \tag{75}$$

- **Schwefel 2.26 (d=3)**:

$$f_{\mathrm{Sch}}(x) = 418.9829\, d - \sum_{i=1}^{d} x_i \sin\big(\sqrt{|x_i|}\big), x \in [-500,\, 500]^3. \tag{76}$$

- **Powell (singular, d=4)**:

$$f_{\mathrm{Pow}}(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4, x \in [-5,\, 5]^4. \tag{77}$$

Further results with a number of agents $N = 4$, $N = 8$, and $N = 12$ respectively displayed in Table. 1 and Table. 3. These results further confirm the remarks in Section. **??**.

The hyperparameters of the GPs (the lengthscale and the outputscale) of each agent are learned with every iteration $t \in [1, \ldots, T_{iter}]$ using maximum log-likelihood (see Rasmussen & Williams (2006)). The acquisition function is maximized using a randoms sampling over 100000 candidates.

	Block A: vary N_a $N = 4$, $t_g = 1$			**Block B: vary t_g** $N = 4$, $N_a = N$		
	$N_a = N$	$0.8N$	$0.5N$	$t_g = 1$	$t_g = 30$	$t_g = 60$
rastrigin (d=4)	**60.140**	61.078	62.077	**60.140**	64.683	65.008
rosenbrock (d=3)	52.287	**49.284**	49.538	**52.287**	99.017	97.713
styblinski_tang (d=4)	**35.853**	37.522	36.509	**35.853**	54.346	55.485
schwefel (d=3)	**1190.122**	1204.829	1206.759	1190.122	1185.093	**1181.585**
powell (d=4)	3309.859	3021.473	**2951.404**	**3309.859**	6677.535	6457.541

Table 1: Average instantaneous regret $rt\_$ (lower is better) for $N = 4$ agents under varying gossip connectivity and communication budget. **Block A** varies the expected number of received tuples per round (N_a) at fixed $t_g = 1$. **Block B** varies the gossip period ($t_g \in \{1, 30, 60\}$) at fixed $N_a = N$. Per row, the best value *within each block* is shown in **bold**.

	Block A: vary N_a $N = 8$, $t_g = 1$			**Block B: vary t_g** $N = 8$, $N_a = N$		
	$N_a = N$	$0.8N$	$0.5N$	$t_g = 1$	$t_g = 30$	$t_g = 60$
rastrigin (d=4)	64.391	**63.432**	63.999	**64.391**	72.906	72.204
rosenbrock (d=3)	**85.623**	93.018	92.992	**85.623**	196.546	197.215
styblinski_tang (d=4)	**50.667**	61.470	61.311	**50.667**	103.180	93.570
schwefel (d=3)	120.060	1214.844	**1211.386**	**1210.060**	1219.815	1220.090
powell (d=4)	5460.591	5010.299	**4942.631**	**5460.591**	11958.663	11074.484

Table 2: Average instantaneous regret $rt\_$ (lower is better) for $N = 8$ agents under varying gossip connectivity and communication budget. **Block A** varies the expected number of received tuples per round (N_a) at fixed $t_g = 1$. **Block B** varies the gossip period ($t_g \in \{1, 30, 60\}$) at fixed $N_a = N$. Per row, the best value *within each block* is shown in **bold**.

	Block A: vary N_a $N = 12$, $t_g = 1$			**Block B: vary t_g** $N = 12$, $N_a = N$		
	$N_a = N$	$0.8N$	$0.5N$	$t_g = 1$	$t_g = 30$	$t_g = 60$
rastrigin (d=4)	66.878	67.252	**66.840**	**66.878**	73.274	73.221
rosenbrock (d=3)	144.638	135.101	**132.080**	**144.638**	294.918	306.498
styblinski_tang (d=4)	**91.435**	92.701	98.948	**91.435**	134.189	128.418
schwefel (d=3)	**1224.084**	1237.489	1231.136	**1224.084**	1229.465	1228.699
powell (d=4)	**7687.572**	7791.842	7736.240	**7687.572**	20149.818	19473.558

Table 3: Average instantaneous regret $rt\_$ (lower is better) for $N = 12$ agents under varying gossip connectivity and communication budget. **Block A** varies the expected number of received tuples per round (N_a) at fixed $t_g = 1$. **Block B** varies the gossip period ($t_g \in \{1, 30, 60\}$) at fixed $N_a = N$. Per row, the best value *within each block* is shown in **bold**.

A.4.2 Synthetic Time-Varying Reward

Figure 6 illustrates the reward function, modeled according to Assumption 2, at three different time steps ($t = 10, 50, 90$) under drift levels $\epsilon \in \{0.001, 0.01, 0.03\}$, in a two-dimentional space. The plots highlight how the reward landscape evolves over time and demonstrate the pronounced sensitivity of the function to the drift parameter ϵ.

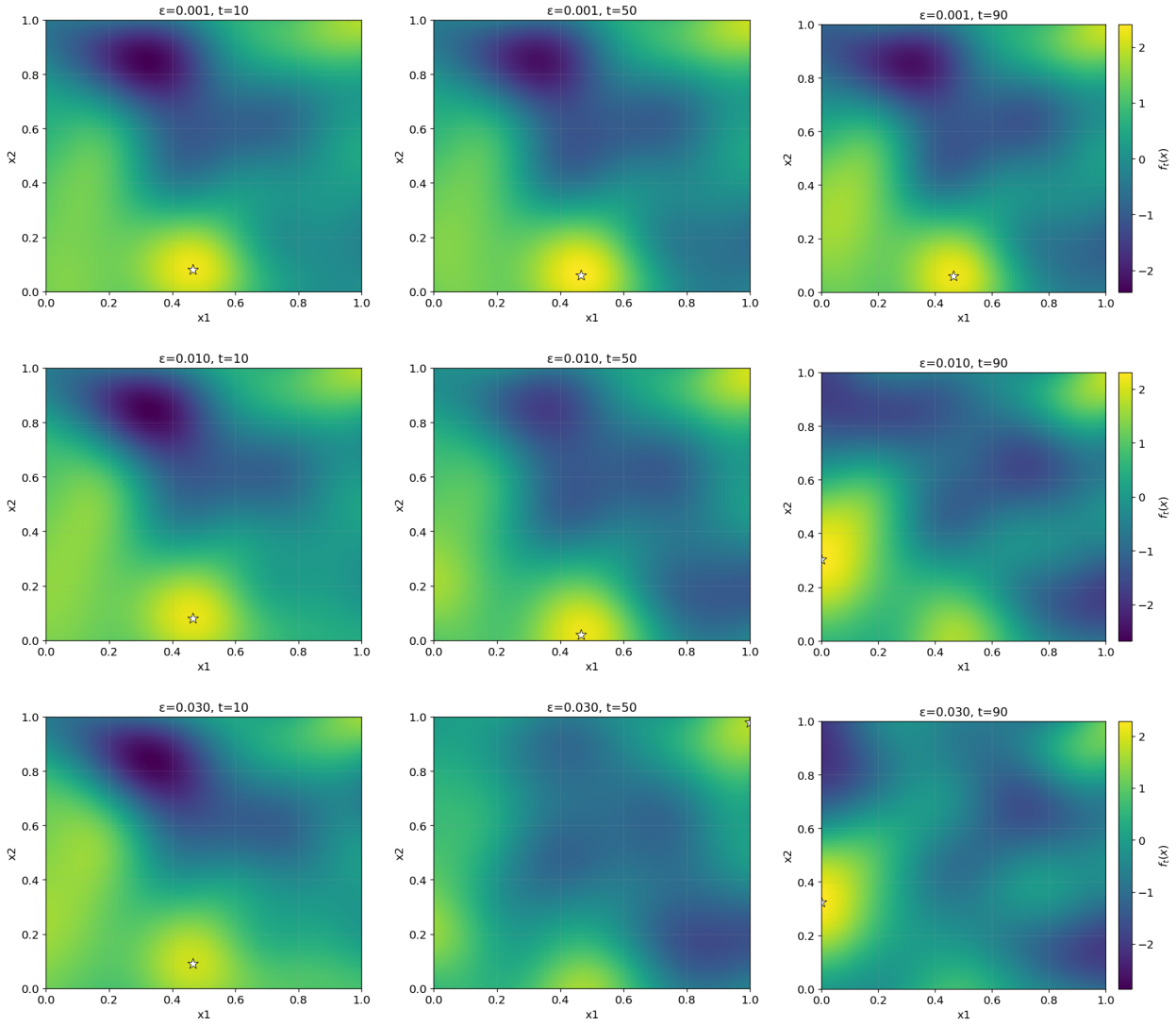
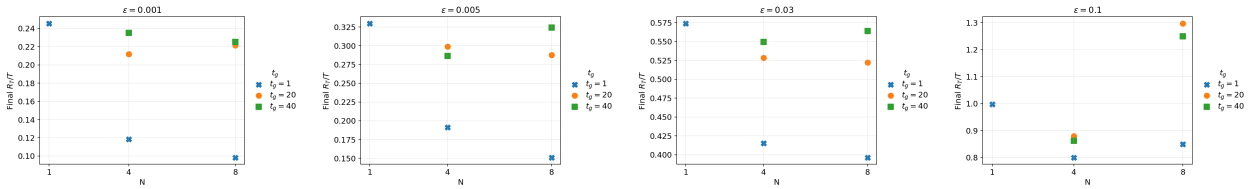


Figure 6: Time-varying reward snapshots for drift levels $\varepsilon \in \{0.001, 0.01, 0.03\}$ at times $t \in \{10, 50, 90\}$. Each row fixes ε (top to bottom: 0.001, 0.01, 0.03) and each column fixes t (left to right: 10, 50, 90). The white star marks the instantaneous maximizer $x_t^\star$ on the rendering grid in each panel.

The acquisition function is optimized via grid sampling, using 50 points per dimension. The selected arm corresponds to the grid point that attains the highest acquisition value. While the optimal reward is the highest actual reward in the function. The observations are noisy, with a noise variance of $\sigma_n^2 = 0.01$.

Figure 7 reports the final average cumulative regret for $N_a = N$ across several drift levels ϵ. As expected, larger drift makes tracking harder. Across all ϵ, performance consistently improves with more agents and smaller gossip periods t_g. In short, more frequent communication with a larger team yields the lowest regret in the time-varying setting.



(a) $\epsilon = 0.001$ (b) $\epsilon = 0.005$ (c) $\epsilon = 0.03$ (d) $\epsilon = 0.1$

Figure 7: Time-varying experiments with $N_a = N$: final average regret as a function of the number of agents and the gossip period t_g for different drift levels ϵ. Lower is better.