COLLABORATIVE PURE EXPLORATION IN KERNEL BANDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a novel Collaborative Pure Exploration in Kernel Bandit model (CoPE-KB), where multiple agents collaborate to complete different but related tasks with limited communication. Our model generalizes prior CoPE formulation with the single-task and classic MAB setting to allow multiple tasks and general reward structures. We propose a novel communication scheme with an efficient kernelized estimator, and design optimal algorithms `CoKernelFC` and `CoKernelFB` for CoPE-KB with fixed-confidence and fixed-budget objectives, respectively. Nearly matching upper and lower bounds in both sampling and communication complexity are established to demonstrate the optimality of our algorithms. Our theoretical results explicitly quantify how task similarities influence learning speedup, and only depend on the effective dimension of feature space. Our novel techniques including an efficient kernelized estimator and linear structured instance transformation, which overcome the communication difficulty in high-dimensional feature space and derive communication round lower bounds, can be of independent interests.

1 INTRODUCTION

Pure exploration (Even-Dar et al., 2006; Kalyanakrishnan et al., 2012; Kaufmann et al., 2016) is a fundamental online learning problem in multi-armed bandits (Thompson, 1933; Lai & Robbins, 1985; Auer et al., 2002), where an agent chooses options (often called arms) and observes random feedback with the objective of identifying the best arm. This formulation has found many important applications, such as web content optimization (Agarwal et al., 2009) and online advertising (Tang et al., 2013). However, traditional pure exploration (Even-Dar et al., 2006; Kalyanakrishnan et al., 2012; Kaufmann et al., 2016) only considers single-agent decision making, and cannot be applied to prevailing distributed systems in real world, which often face a heavy computation load and require multiple parallel devices to process tasks, e.g., distributed web servers (Zhuo et al., 2003) and data centers (Liu et al., 2011).

To handle such distributed applications, prior works (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020) have developed the Collaborative Pure Exploration (CoPE) model, where multiple agents communicate and cooperate to identify the best arm with learning speedup. Yet, existing results focus only on the classic multi-armed bandit (MAB) setting with *single* task, i.e., all agents solve a common task and the rewards of arms are *individual* values (rather than generated by a reward function). However, in many distributed applications such as multi-task neural architecture search (Gao et al., 2020), different devices can face *different but related* tasks, and there exists similar dependency of rewards on option features among different tasks. Therefore, it is important to develop a more general CoPE model to allow heterogeneous tasks and structured reward dependency, and further theoretically understand how task correlation impacts learning.

Motivated by the above fact, we propose a novel Collaborative Pure Exploration in Kernel Bandit (CoPE-KB) model. Specifically, in CoPE-KB, each agent is given a set of arms, and the expected reward of each arm is generated by a task-dependent reward function in a high and possibly infinite dimensional Reproducing Kernel Hilbert Space (RKHS) (Wahba, 1990; Schölkopf et al., 2002). Each agent sequentially chooses arms to sample and observes random outcomes in order to identify the best arm. Agents can broadcast and receive messages to and from others in communication rounds, so that they can collaborate and exploit the task similarity to expedite learning processes.

Our CoPE-KB model is a novel generalization of prior CoPE problem (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020), which not only extends prior models from the single-task setting to multiple tasks, but also goes beyond classic MAB setting and allows general (linear or nonlinear) reward structures. CoPE-KB is most suitable for applications involving multiple tasks and complicated reward structures. For example, in multi-task neural architecture search (Gao et al., 2020), one wants to search for best architectures for different but related tasks on multiple devices, e.g., the object detection (Ghiasi et al., 2019) and object tracking (Yan et al., 2021) tasks in computer vision, which often use similar neural architectures. Instead of individually evaluating each possible architecture, one prefers to directly learn the relationship (reward function) between the accuracy results achieved and the features of used architectures (e.g., the type of neural networks), and exploit the similarity of reward functions among tasks to accelerate the search.

Our CoPE-KB generalization faces a unique challenge on *communication*. Specifically, in prior CoPE works with classic MAB setting (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020), agents only need to learn *scalar* rewards, which are easy to transmit. However, under the kernel model, agents need to estimate a *high or even infinite dimensional* reward parameter, which is inefficient to directly transmit. Also, if one naively adapts existing reward estimators for kernel bandits (Srinivas et al., 2010; Camilleri et al., 2021) to learn this high-dimensional reward parameter, he/she will suffer an expensive communication cost dependent on the number of samples $N^{(t)}$, since the reward estimators there need *all raw sample outcomes* to be transmitted. To tackle this challenge, we develop an efficient kernelized estimator, which only needs *average outcomes* on nV arms and reduces the required transmitted messages from $O(N^{(t)})$ to $O(nV)$. Here V is the number of agents, and n is the number of arms for each agent. The number of samples $N^{(t)}$ depends on the inverse of the minimum reward gap, and is often far larger than the number of arms nV.

Under the CoPE-KB model, we study two popular objectives, i.e., Fixed-Confidence (FC), where we aim to minimize the number of samples used under a given confidence, and Fixed-Budget (FB), where the goal is to minimize the error probability under a given sample budget. We design two novel algorithms `CoKernelFC` and `CoKernelFB`, which adopt an efficient kernelized estimator to simplify the required data transmission and enjoy a $O(nV)$ communication cost, instead of $O(N^{(t)})$ as in adaptions of existing kernel bandit algorithms (Srinivas et al., 2010; Camilleri et al., 2021). We provide near-optimal sampling and communication guarantees, and also interpret them by standard kernel measures, e.g., maximum information gain and effective dimension. Our results rigorously quantify the influences of task similarities on learning acceleration, and hold for both finite and infinite dimensional feature space.

The contributions of this paper are summarized as follows:

- We formulate a collaborative pure exploration in kernel bandit (CoPE-KB) model, which generalizes prior single-task CoPE formulation to allow multiple tasks and general reward structures, and consider two objectives, i.e., fixed-confidence (FC) and fixed-budget (FB).

- For the FC objective, we propose a near-optimal algorithm `CoKernelFC`, which adopts an efficient kernelized estimator to simplify the required data transmission and enjoys only a $O(nV)$ communication cost. We derive near-optimal sample complexity $\tilde{O}(\frac{\rho^*(\xi)}{V}\log\delta^{-1})$ and communication rounds $O(\log\Delta_{\min}^{-1})$. Here ξ is the regularization parameter, and $\rho^*(\xi)$ is the problem hardness (see Section 4.3).

- For the FB objective, we design a communication-optimal algorithm `CoKernelFB` with error probability $\tilde{O}(\exp(-\frac{TV}{\rho^*(\xi)})n^2V)$ and communication rounds $O(\log(\omega(\xi,\tilde{\mathcal{X}})))$. Here T is the sample budget, $\tilde{\mathcal{X}}$ is the set of arms, and $\omega(\xi,\tilde{\mathcal{X}})$ is the principle dimension of data projections in $\tilde{\mathcal{X}}$ to RKHS (see Section 5.1).

- Our algorithms offer an efficient communication scheme for information exchange in high dimensional feature space. Our results explicitly quantify how task similarities impact learning acceleration, and only depend on the effective dimension of feature space.

Due to the space limit, we defer all proofs and experiments to the appendix.

2 Related Work

Below we review the most related works, and defer a complete literature review to Appendix B.

Collaborative Pure Exploration (CoPE). Hillel et al. (2013); Tao et al. (2019) initiate the CoPE literature with the single-task and classic MAB setting, where all agents solve a common classic best arm identification problem (without reward structures). Karpov et al. (2020) further extend the formulation in (Hillel et al., 2013; Tao et al., 2019) to best m arm identification. Our CoPE-KB generalizes prior CoPE works (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020) to allow multiple tasks and general reward structures, and faces a unique challenge on communication, since the reward parameter we need to learn is high-dimensional and inefficient to transmit.

Kernel Bandits. For the regret minimization objective, Srinivas et al. (2010); Valko et al. (2013); Scarlett et al. (2017); Li & Scarlett (2022) investigate the single-agent kernel bandit problem and establish regret bounds dependent on maximum information gain. Krause & Ong (2011); Deshmukh et al. (2017) study multi-task kernel bandits, which consider a composite kernel constituted by two sub-kernels with respect to tasks and items. Dubey et al. (2020); Li et al. (2022) consider multi-agent kernel bandits with local-broadcast and client-server communication protocols, respectively, which are different from our communication round protocol. There are only a few works studying kernel bandits with the pure exploration objective, and all of them consider the single-agent formulation. Camilleri et al. (2021) propose a robust inverse propensity score estimator for high-dimensional experimental design, and Zhu et al. (2021) use neural networks to approximate nonlinear reward functions. Scarlett et al. (2017); Vakili et al. (2021) study the simple regret setting in pure exploration, which aims to minimize the suboptimality (simple regret) of the selected arm after T timesteps. This setting is different from our best arm identification problem. The above kernel bandit works consider either regret minimization or single-agent formulation, which do not involve our challenges on round-speedup analysis and communication.

Recently, Dubey & Pentland (2020); Huang et al. (2021); Li & Wang (2022) investigate (low-dimensional) federated linear bandits, where agents directly communicate the whole reward parameter. These works cannot be adapted to resolve our challenges, because in the kernel model, the reward parameter is high-dimensional and expensive to explicitly calculate or transmit.

3 Collaborative Pure Exploration in Kernel Bandit (CoPE-KB)

In this section, we define the Collaborative Pure Exploration in Kernel Bandit (CoPE-KB) problem.

Agents and Rewards. There are V agents indexed by $[V] := \{1, \dots, V\}$, who collaborate to solve different but possibly related instances (tasks) of the Pure Exploration in Kernel Bandit (PE-KB) problem. For each agent $v \in [V]$, she is given a set of n arms $\mathcal{X}_v = \{x_{v,1}, \dots, x_{v,n}\} \subseteq \mathbb{R}^{d_\mathcal{X}}$, where $x_{v,i}$ ($i \in [n]$) describes the arm feature, and $d_\mathcal{X}$ is the dimension of arm feature vectors. The expected reward of each arm $x \in \mathcal{X}_v$ is $f_v(x)$, where $f_v : \mathcal{X}_v \mapsto \mathbb{R}$ is an unknown reward function. Let $\mathcal{X} := \cup_{v\in[V]} \mathcal{X}_v$. At each timestep t, each agent v pulls an arm $x_v^t \in \mathcal{X}_v$ and observes a random reward $y_v^t = f_v(x_v^t) + \eta_v^t$. Here η_v^t is a zero-mean and 1-sub-Gaussian noise, and it is independent among different t and v. We assume that the best arms $x_{v,*} := \operatorname{argmax}_{x\in\mathcal{X}_v} f_v(x)$ are unique for all $v \in [V]$, which is a common assumption in the pure exploration literature (Even-Dar et al., 2006; Audibert et al., 2010; Kaufmann et al., 2016).

Multi-Task Kernel Composition. We assume that the functions f_v are parametric functionals of a global function $F : \mathcal{Z} \times \mathcal{X} \mapsto \mathbb{R}$, which satisfies that, for each agent $v \in [V]$, there exists a task feature vector $z_v \in \mathcal{Z}$ such that

$$f_v(x) = F(z_v, x), \ \forall x \in \mathcal{X}_v. \tag{1}$$

Here $\mathcal{Z}$ and $\mathcal{X}$ denote the task feature space and arm feature space, respectively. Note that Eq. (1) allows tasks to be different for agents (by having different task features), whereas prior CoPE works (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020) restrict the tasks ($\mathcal{X}_v$ and f_v) to be the same for all agents $v \in [V]$.

We denote the overall input space of F by $\tilde{\mathcal{X}} := \mathcal{Z} \times \mathcal{X}$ and the overall input by $\tilde{x} := (z_v, x)$. As a standard assumption in kernel bandits (Krause & Ong, 2011; Deshmukh et al., 2017; Dubey et al., 2020), we assume that F has a bounded norm in a high (possibly infinite) dimensional Reproducing

Kernel Hilbert Space (RKHS) $\mathcal{H}$ specified by the kernel $k : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \mapsto \mathbb{R}$, and there exists a feature mapping $\phi : \tilde{\mathcal{X}} \mapsto \mathcal{H}$ and an unknown parameter $\theta^* \in \mathcal{H}$ such that[1]

$$F(\tilde{x}) = \phi(\tilde{x})^\top \theta^*, \ \forall \tilde{x} \in \tilde{\mathcal{X}},$$
$$k(\tilde{x}, \tilde{x}') = \phi(\tilde{x})^\top \phi(\tilde{x}'), \ \forall \tilde{x}, \tilde{x}' \in \tilde{\mathcal{X}}.$$

Here $k(\cdot, \cdot)$ is a product composite kernel, which satisfies that for any $z, z' \in \mathcal{Z}$ and $x, x' \in \mathcal{X}$,

$$k((z, x), (z', x')) = k_{\mathcal{Z}}(z, z') \cdot k_{\mathcal{X}}(x, x'),$$

where $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ is the task feature kernel which measures the similarity of functions f_v, and $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is the arm feature kernel which depicts the feature structure of arms. The computation of kernel function $k(\cdot, \cdot)$ operates only on low-dimensional input data. By expressing all calculations via $k(\cdot, \cdot)$, it enables us to avoid maintaining high-dimensional $\phi(\cdot)$ and attain computation efficiency. In addition, the composite structure of $k(\cdot, \cdot)$ allows us to model the similarity among tasks.

Let $K_{\mathcal{Z}} := [k_{\mathcal{Z}}(z_v, z_{v'})]_{v,v' \in [V]}$ denote the kernel matrix of task features. $\texttt{Rank}(K_{\mathcal{Z}})$ characterizes how much the tasks among agents are similar. For example, if agents solve a common task, i.e., the arm set $\mathcal{X}_v$ and reward function f_v are the same for all agents, we have that the task feature z_v is the same for all $v \in [V]$, and $\texttt{Rank}(K_{\mathcal{Z}}) = 1$; If all tasks are totally different, $\texttt{Rank}(K_{\mathcal{Z}}) = V$.

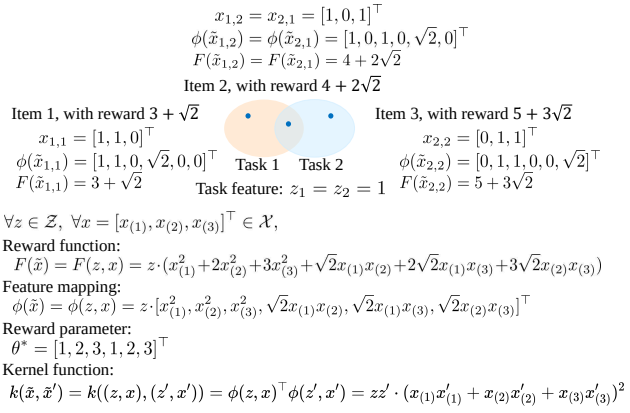

Figure 1: Illustrating example for CoPE-KB.

To better illustrate the model, we provide a 2-agent example in Figure 1. There are Items $1, 2, 3$ with expected rewards $3 + \sqrt{2}, 4 + 2\sqrt{2}, 5 + 3\sqrt{2}$, respectively. Agent 1 is given Items $1, 2$, denoted by $\mathcal{X}_1 = \{x_{1,1}, x_{1,2}\}$, and Agent 2 is given Items $2, 3$, denoted by $\mathcal{X}_2 = \{x_{2,1}, x_{2,2}\}$, where both $x_{1,2}$ and $x_{2,1}$ refer to Item 2. The task feature $z_1 = z_2 = 1$, which means that the expected rewards of all items are generated by a common function. The reward function F is nonlinear with respect to x, but can be represented in a linear form in a higher-dimensional feature space, i.e., $F(\tilde{x}) = F(z, x) = \phi(\tilde{x})^\top \theta^*$ with $\tilde{x} := (z, x)$ for any $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Here $\phi(\tilde{x})$ is the feature embedding, and θ^* is the reward parameter. The computation of kernel function $k(\tilde{x}, \tilde{x}')$ only involves low-dimensional input data $\tilde{x}$ and $\tilde{x}'$, rather than higher-dimensional feature embedding $\phi(\cdot)$. F, ϕ, θ^* and k are specified in Figure 1. The two agents can share the learned information on θ^* to accelerate learning.

Communication. Following the popular communication protocol in the CoPE literature (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020), we allow these V agents to exchange information via *communication rounds*, in which each agent can broadcast and receive messages from others. Following existing CoPE works (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020), we restrict the length of each message within $O(n)$ bits for practicability, where n is the number of arms for each agent, and we consider the number of bits for representing a real number as a constant. We want agents to cooperate and complete all tasks using as few communication rounds as possible.

Objectives. We consider two objectives of the CoPE-KB problem, one with Fixed-Confidence (FC) and the other with Fixed-Budget (FB). In the FC setting, given a confidence parameter $\delta \in (0, 1)$, the agents aim to identify $x_{v,*}$ for all $v \in [V]$ with probability at least $1 - \delta$, and minimize the average number of samples used per agent. In the FB setting, the agents are given an overall TV sample budget (T average samples per agent), and aim to use at most TV samples to identify $x_{v,*}$ for all $v \in [V]$ and minimize the error probability. In both FC and FB settings, the agents are requested to minimize the number of communication rounds and control the length of each message within $O(n)$ bits, as in the CoPE literature (Hillel et al., 2013; Tao et al., 2019; Karpov et al., 2020).

[1]For any $h, h' \in \mathcal{H}$, we denote their inner-product by $h^\top h' := \langle h, h' \rangle_{\mathcal{H}}$ and denote the norm of h by $\|h\| := \sqrt{h^\top h}$. For any $h \in \mathcal{H}$ and matrix M, we denote $\|h\|_{M^{-1}} := \sqrt{h^\top M^{-1} h}$.

Algorithm 1 Collaborative Multi-agent Algorithm `CoKernelFC`: for Agent $v \in [V]$

1: **Input:** δ, $\tilde{\mathcal{X}}$, $k(\cdot,\cdot) : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \mapsto \mathbb{R}$, regularization parameter ξ, rounding procedure `ROUND`, rounding approximation parameter $\varepsilon = \frac{1}{10}$
2: **Initialization:** $\delta_t := \frac{\delta}{2t^2}$ for all $t \geq 1$. $\mathcal{B}_{v'}^{(1)} \leftarrow \tilde{\mathcal{X}}_{v'}$ for all $v' \in [V]$.
3: **for** $t = 1, 2, \ldots$ **do**
4: Let λ_t^* and ρ_t^* be the optimal solution and optimal value of $\min_{\lambda \in \triangle_{\tilde{\mathcal{X}}}} \max_{\tilde{x}_i, \tilde{x}_j \in \mathcal{B}_{v'}^{(t)}, v' \in [V]} \|\phi(\tilde{x}_i) - \phi(\tilde{x}_j)\|^2_{(\xi I + \sum_{\tilde{x} \in \tilde{\mathcal{X}}} \lambda_{\tilde{x}} \phi(\tilde{x}) \phi(\tilde{x})^\top)^{-1}}$ // compute the optimal sample allocation
5: $N^{(t)} \leftarrow \max\{\lceil 32(2^t)^2(1+\varepsilon)^2 \rho_t^* \log\left(2n^2V/\delta_t\right) \rceil,\ r(\xi, \lambda_t^*, \varepsilon)\}$, where $r(\xi, \lambda_t^*, \varepsilon)$ is the number of samples needed by `ROUND`
6: $(\tilde{s}_1, \ldots, \tilde{s}_{N^{(t)}}) \leftarrow$ `ROUND`$(\xi, \lambda_t^*, N^{(t)}, \varepsilon)$
7: Extract a sub-sequence $\tilde{\boldsymbol{s}}_v^{(t)}$ from $(\tilde{s}_1, \ldots, \tilde{s}_{N^{(t)}})$ which only contains the arms in $\tilde{\mathcal{X}}_v$
8: Sample the arms in $\tilde{\boldsymbol{s}}_v^{(t)}$ and observe random rewards $\boldsymbol{y}_v^{(t)}$
9: Let $N_{v,i}^{(t)}$ and $\bar{y}_{v,i}^{(t)}$ be the number of samples and the average sample outcome on arm $\tilde{x}_{v,i}$
10: Broadcast $\{(N_{v,i}^{(t)}, \bar{y}_{v,i}^{(t)})\}_{i \in [n]}$, and receive $\{(N_{v',i}^{(t)}, \bar{y}_{v',i}^{(t)})\}_{i \in [n]}$ from all other agents $v' \neq v$
11: $k_t(\tilde{x}) \leftarrow [\sqrt{N_1^{(t)}} k(\tilde{x}, \tilde{x}_1), \ldots, \sqrt{N_{nV}^{(t)}} k(\tilde{x}, \tilde{x}_{nV})]^\top$ for any $\tilde{x} \in \tilde{\mathcal{X}}$. $K^{(t)} \leftarrow [\sqrt{N_i^{(t)} N_j^{(t)}} k(\tilde{x}_i, \tilde{x}_j)]_{i,j \in [nV]}$. $\bar{y}^{(t)} \leftarrow [\sqrt{N_1^{(t)}} \bar{y}_1^{(t)}, \ldots, \sqrt{N_{nV}^{(t)}} \bar{y}_{nV}^{(t)}]^\top$ // organize overall sample information
12: **for** all $v' \in [V]$ **do**
13: $\hat{\Delta}_t(\tilde{x}, \tilde{x}') \leftarrow (k_t(\tilde{x}) - k_t(\tilde{x}'))^\top (K^{(t)} + N^{(t)} \xi I)^{-1} \bar{y}^{(t)}$, $\forall \tilde{x}, \tilde{x}' \in \mathcal{B}_{v'}^{(t)}$ // use a kernelized estimator (described in Section 4.2) to estimate reward gaps
14: $\mathcal{B}_{v'}^{(t+1)} \leftarrow \mathcal{B}_{v'}^{(t)} \setminus \{\tilde{x} \in \mathcal{B}_{v'}^{(t)} \mid \exists \tilde{x}' \in \mathcal{B}_{v'}^{(t)} : \hat{\Delta}_t(\tilde{x}', \tilde{x}) \geq 2^{-t}\}$ // discard sub-optimal arms
15: **end for**
16: **if** $\forall v' \in [V], |\mathcal{B}_{v'}^{(t+1)}| = 1$, **return** $\mathcal{B}_1^{(t+1)}, \ldots, \mathcal{B}_V^{(t+1)}$
17: **end for**

To sum up, in CoPE-KB, we let agents collaborate to simultaneously complete multiple related best arm identification tasks using few communication rounds. In particular, when all agents solve the same task and $\mathcal{X}$ is the canonical basis, our CoPE-KB reduces to existing CoPE with classic MAB setting (Hillel et al., 2013; Tao et al., 2019).

4 Fixed-Confidence CoPE-KB

We start with the fixed-confidence setting. We present algorithm `CoKernelFC` equipped with an efficient kernelized estimator, and provide near-optimal guarantees in sampling and communication.

4.1 Algorithm `CoKernelFC`

`CoKernelFC` (Algorithm 1) is an elimination-based multi-agent algorithm, and the procedure for each agent v is as follows. Agent v maintains candidate arm sets $\mathcal{B}_{v'}^{(t)}$ for all $v' \in [V]$. In each phase t, she solves a global min-max optimization (Line 4) to find the optimal sample allocation $\lambda_t^* \in \triangle_{\tilde{\mathcal{X}}}$, which can achieve the minimum estimation error. Here $\triangle_{\tilde{\mathcal{X}}}$ denotes the collection of all distributions on $\tilde{\mathcal{X}}$, and ρ_t^* is the factor of optimal estimation error. In practice, the high-dimensional feature embedding $\phi(\tilde{x})$ is only implicitly maintained, and this optimization can be efficiently solved by kernelized gradient descent (Camilleri et al., 2021) (see Appendix D.1). After solving this optimization, agent v uses ρ_t^* to compute the number of samples $N^{(t)}$ to ensure the estimation error of reward gaps to be smaller than $2^{-(t+1)}$ (Line 5).

Next, we call a rounding procedure `ROUND`$(\xi, \lambda_t^*, N^{(t)}, \varepsilon)$ (Allen-Zhu et al., 2021; Camilleri et al., 2021), which rounds a weighted sample allocation $\lambda_t^* \in \triangle_{\tilde{\mathcal{X}}}$ into a discrete sample sequence $(\tilde{s}_1, \ldots, \tilde{s}_{N^{(t)}}) \in \tilde{\mathcal{X}}^{N^{(t)}}$, and ensures the rounding error within ε (Line 6). This rounding pro-

cedure requires the number of samples $N^{(t)} \geq r(\xi, \lambda_t^*, \varepsilon) = O(\frac{\tilde{d}(\xi, \lambda_t^*)}{\varepsilon^2})$ (Line 5). Here $r(\xi, \lambda_t^*, \varepsilon)$ is the number of samples needed by ROUND. $\tilde{d}(\xi, \lambda_t^*)$ is the number of the eigenvalues of matrix $\sum_{\tilde{x} \in \tilde{\mathcal{X}}} \lambda_t^*(\tilde{x}) \phi(\tilde{x}) \phi(\tilde{x})^\top$ which are greater than ξ. It stands for the effective dimension of feature space (see Appendix C.1 for details of this rounding procedure).

Obtaining sample sequence $(\tilde{s}_1, \ldots, \tilde{s}_{N^{(t)}})$, agent v extracts a sub-sequence $\tilde{\boldsymbol{s}}_v^{(t)}$ which only contains the arms in her arm set $\tilde{\mathcal{X}}_v$. Then, she sample the arms in $\tilde{\boldsymbol{s}}_v^{(t)}$ and observe sample outcomes $\boldsymbol{y}_v^{(t)}$. After sampling, she only communicates the number of samples $N_{v,i}^{(t)}$ and *average* sample outcome $\bar{y}_{v,i}^{(t)}$ on each arm $\tilde{x}_{v,i}$ with other agents (Line 10). Receiving overall sample information, she uses a kernelized estimator (discussed shortly) to estimate the reward gap $\hat{\Delta}_t(\tilde{x}, \tilde{x}')$ between any two arms $\tilde{x}, \tilde{x}' \in \mathcal{B}_{v'}^{(t)}$ for all $v' \in [V]$, and discards sub-optimal arms (Lines 13,14). In Line 11, for any $i \in [nV]$, we use $\tilde{x}_i$, $\bar{y}_i^{(t)}$ and $N_i^{(t)}$ with a single subscript to denote the i-th arm in $\tilde{\mathcal{X}}$, the average sample outcome on this arm and the number of samples allocated to this arm in phase t, respectively.

4.2 Kernelized Estimator, Communication and Computation

Now we introduce the kernelized estimator (Line 13), which boosts communication and computation efficiency of CoKernelFC. First note that, our CoPE-KB generalization faces a unique challenge on communication, i.e., how to let agents efficiently share their learned information on the high-dimensional reward parameter θ^*. Naively adapting existing federated linear bandit or kernel bandit algorithms (Dubey & Pentland, 2020; Huang et al., 2021; Dubey et al., 2020), which transmit the *whole* estimated reward parameter or *all raw sample outcomes*, will suffer a $O(\dim(\mathcal{H}))$ or $O(N^{(t)})$ communication cost, respectively. Here, the number of samples $N^{(t)} = \tilde{O}(d_{\text{eff}}/\Delta_{\min}^2)$, where d_{eff} is the effective dimension of feature space and $\Delta_{\min}$ is the minimum reward gap. Thus, $N^{(t)}$ is far larger than the number of arms nV when $\Delta_{\min}$ is small.

Kernelized Estimator. To handle the communication challenge, we develop a novel kernelized estimator (Eq. (3)) to significantly simplify the required transmitted data. Specifically, we make a key observation: since the sample sequence $(\tilde{s}_1, \ldots, \tilde{s}_{N^{(t)}})$ are constituted by arms $\tilde{x}_1, \ldots, \tilde{x}_{nV}$, one can *merge repetitive computations* for same arms. Then, for all $i \in [nV]$, we merge the repetitive feature embeddings $\phi(\tilde{x}_i)$ for same arms $\tilde{x}_i$, and condense the $N^{(t)}$ raw sample outcomes to *average outcome* $\bar{y}_i^{(t)}$ on each arm $\tilde{x}_i$. As a result, we express $\hat{\theta}_t$ in a simplified kernelized form, and use it to estimate the reward gap $\hat{\Delta}(\tilde{x}, \tilde{x}')$ between any two arms $\tilde{x}, \tilde{x}'$ as

$$\hat{\theta}_t := \Phi_t^\top \left(N^{(t)} \xi I + K^{(t)}\right)^{-1} \bar{y}^{(t)}, \tag{2}$$

$$\hat{\Delta}(\tilde{x}, \tilde{x}') := \left(\phi(\tilde{x}) - \phi(\tilde{x}')\right)^\top \hat{\theta}_t = \left(k_t(\tilde{x}) - k_t(\tilde{x}')\right)^\top \left(N^{(t)} \xi I + K^{(t)}\right)^{-1} \bar{y}^{(t)}. \tag{3}$$

Here $\Phi_t := [\sqrt{N_1^{(t)}} \phi(\tilde{x}_1)^\top; \ldots; \sqrt{N_{nV}^{(t)}} \phi(\tilde{x}_{nV})^\top]$, $K^{(t)} := [\sqrt{N_i^{(t)} N_j^{(t)}} k(\tilde{x}_i, \tilde{x}_j)]_{i,j \in [nV]}$, $k_t(\tilde{x}) := \Phi_t \phi(\tilde{x}) = [\sqrt{N_1^{(t)}} k(\tilde{x}, \tilde{x}_1), \ldots, \sqrt{N_{nV}^{(t)}} k(\tilde{x}, \tilde{x}_{nV})]^\top$ and $\bar{y}^{(t)} := [\sqrt{N_1^{(t)}} \bar{y}_1^{(t)}, \ldots, \sqrt{N_{nV}^{(t)}} \bar{y}_{nV}^{(t)}]^\top$ stand for the feature embeddings, kernel matrix, correlations and average outcomes of the nV arms, respectively, which *merge* repetitive information on same arms. We refer interested reader to Appendix D.1, D.2 for a detailed derivation of our kernelized estimator and a comparison with existing estimators (Dubey et al., 2020; Camilleri et al., 2021).

Communication. Thanks to the kernelized estimator, we only need to transmit the nV average outcomes $\bar{y}_1^{(t)}, \ldots, \bar{y}_{nV}^{(t)}$ (Line 10), instead of the whole $\hat{\theta}_t$ or all $N^{(t)}$ raw outcomes as in existing federated linear bandit or kernel bandit algorithms (Dubey & Pentland, 2020; Dubey et al., 2020). This significantly reduces the number of transmission bits from $O(\dim(\mathcal{H}))$ or $O(N^{(t)})$ to only $O(nV)$, and satisfies the $O(n)$-bit per message requirement.

Computation. In CoKernelFC, $\phi(\tilde{x})$ and $\hat{\theta}_t$ are maintained implicitly, and all steps (e.g., Lines 4, 13) can be implemented efficiently by only querying kernel function $k(\cdot, \cdot)$ (see Appendix D.1 for implementation details). Thus, the computation complexity for reward estimation is only $\text{Poly}(nV)$, instead of $\text{Poly}(\dim(\mathcal{H}))$ as in prior kernel bandit algorithms (Zhou et al., 2020; Zhu et al., 2021).

We remark that while we focus on the collaborative setting, the communication challenge we tackle is *fundamental* in distributed bandits (Liu & Zhao, 2010; Hillel et al., 2013; Dubey et al., 2020). Our algorithm offers an efficient communication scheme for information exchange in high (possibly infinite) dimensional RKHS, which can also be applied to other multi-agent environments, e.g., the master-slave setting (Kong et al., 2017; Wang et al., 2017), where a master agent does all computations and sends sampling commands, and the slave agents conduct sampling in local.

4.3 Theoretical Performance of CoKernelFC

To formally state our results, we define the speedup and hardness as in the literature (Hillel et al., 2013; Tao et al., 2019; Fiez et al., 2019).

For a CoPE-KB instance $\mathcal{I}$, let $T_{\mathcal{A}_M,\mathcal{I}}$ denote the average number of samples used per agent in a multi-agent algorithm $\mathcal{A}_M$ to identify all best arms. Let $T_{\mathcal{A}_S,\mathcal{I}}$ denote the average number of samples used per agent, by replicating V copies of a single-agent algorithm $\mathcal{A}_S$ to complete all tasks (without communication). Then, the speedup of $\mathcal{A}_M$ on instance $\mathcal{I}$ is defined as

$$\beta_{\mathcal{A}_M,\mathcal{I}} = \inf_{\mathcal{A}_S} \frac{T_{\mathcal{A}_S,\mathcal{I}}}{T_{\mathcal{A}_M,\mathcal{I}}}. \tag{4}$$

It holds that $1 \leq \beta_{\mathcal{A}_M,\mathcal{I}} \leq V$. When all tasks are the same, $\beta_{\mathcal{A}_M,\mathcal{I}}$ can approach V. When all tasks are totally different, communication brings no benefit and $\beta_{\mathcal{A}_M,\mathcal{I}} = 1$. This speedup can be similarly defined for error probability results (see Section 5.2), by taking $T_{\mathcal{A}_M,\mathcal{I}}$ and $T_{\mathcal{A}_S,\mathcal{I}}$ as the smallest numbers of samples needed to meet the confidence constraint.

The hardness for CoPE-KB is defined as

$$\rho^*(\xi) = \min_{\lambda \in \triangle_{\tilde{\mathcal{X}}}} \max_{\tilde{x} \in \tilde{\mathcal{X}}_v \setminus \{\tilde{x}_{v,*}\}, v \in [V]} \frac{\|\phi(\tilde{x}_{v,*}) - \phi(\tilde{x})\|^2_{A(\xi,\lambda)^{-1}}}{(F(\tilde{x}_{v,*}) - F(\tilde{x}))^2}, \tag{5}$$

where $\xi \geq 0$ is a regularization parameter, and $A(\xi, \lambda) := \xi I + \sum_{\tilde{x} \in \tilde{\mathcal{X}}} \lambda_{\tilde{x}} \phi(\tilde{x}) \phi(\tilde{x})^\top$. This definition of $\rho^*(\xi)$ is adapted from prior linear bandit work (Fiez et al., 2019). Here $F(\tilde{x}_{v,*}) - F(\tilde{x})$ is the reward gap between the best arm $\tilde{x}_{v,*}$ and a suboptimal arm $\tilde{x}$, and $\|\phi(\tilde{x}_{v,*}) - \phi(\tilde{x})\|^2_{A(\xi,\lambda)^{-1}}$ is a dimension-related factor of estimation error. Intuitively, $\rho^*(\xi)$ indicates how many samples it takes to make the estimation error smaller than the reward gap under regularization parameter ξ.

Let $\Delta_{\min} := \min_{\tilde{x} \in \tilde{\mathcal{X}}_v \setminus \{\tilde{x}_{v,*}\}, v \in [V]} (F(\tilde{x}_{v,*}) - F(\tilde{x}))$ denote the minimum reward gap between the best arm and suboptimal arms among all tasks. Let S denote the average number of samples used by each agent, i.e., per-agent sample complexity. Below we present the performance of CoKernelFC.

Theorem 1 (Fixed-Confidence Upper Bound)**.** *Suppose that ξ satisfies $\sqrt{\xi} \max_{\tilde{x}_i, \tilde{x}_j \in \tilde{\mathcal{X}}_v, v \in [V]} \|\phi(\tilde{x}_i) - \phi(\tilde{x}_j)\|_{A(\xi,\lambda_1^*)^{-1}} \leq \frac{\Delta_{\min}}{32(1+\varepsilon)\|\theta^*\|}$. With probability at least $1-\delta$,* CoKernelFC *returns the best arms $\tilde{x}_{v,*}$ for all $v \in [V]$, with per-agent sample complexity*

$$S = O\left(\frac{\rho^*(\xi)}{V} \cdot \log \Delta_{\min}^{-1} \left(\log \left(\frac{nV}{\delta} \right) + \log \log \Delta_{\min}^{-1} \right) + \frac{\tilde{d}(\xi, \lambda_1^*)}{V} \cdot \log \Delta_{\min}^{-1} \right)$$

and communication rounds $O(\log \Delta_{\min}^{-1})$.

Remark 1. The condition on regularization parameter ξ implies that CoKernelFC needs a small regularization parameter such that the bias due to regularization is smaller than $\frac{\Delta_{\min}}{2}$. Such conditions are similarly needed in prior kernel bandit work (Camilleri et al., 2021), and can be dropped in the extended PAC setting (allowing a gap between the identified best arm and true best arm). The $\tilde{d}(\xi, \lambda_1^*) \log \Delta_{\min}^{-1} / V$ term is a cost for using rounding procedure ROUND. This is a second order term when the reward gaps $\Delta_{v,i} := F(\tilde{x}_{v,*}) - F(\tilde{x}_{v,i}) < 1$ for all $\tilde{x}_{v,i} \in \tilde{\mathcal{X}}_v \setminus \{\tilde{x}_{v,*}\}$ and $v \in [V]$, which is the common case in pure exploration (Fiez et al., 2019; Zhu et al., 2021).

When all tasks are the same, Theorem 1 achieves a V-speedup, since replicating V copies of single-agent algorithms (Camilleri et al., 2021; Zhu et al., 2021) to complete all tasks without communication will cost $\tilde{O}(\rho^*(\xi))$ samples per agent. When tasks are totally different, there is no speedup in Theorem 1, since each copy of single-agent algorithm solves a task in her own sub-dimension and costs $\tilde{O}(\frac{\rho^*(\xi)}{V})$ samples ($\rho^*(\xi)$ stands for the effective dimension of *all* tasks). This result matches the restriction of speedup.

Regarding the optimality, when $\{\phi(\tilde{x})\}_{\tilde{x}\in\tilde{\mathcal{X}}}$ spans $\mathcal{H}$ and $\xi = 0$, Theorem 1 nearly matches the sample complexity lower bound $\Omega(\rho^*(0)\log\delta^{-1})$ (presented in Appendix D.4.1). In addition, when all tasks are the same, Theorem 1 matches the communication round-speedup lower bound (Theorem 2) up to double-logarithmic factors.

Interpretation. Now, we interpret Theorem 1 by standard measures in kernel bandits and a decomposition with respect to task similarities and task features. Below we first introduce the definitions of maximum information gain and effective dimension, which are adapted from kernel bandits with regret minimization (Srinivas et al., 2010; Valko et al., 2013) to the pure exploration setting.

We define the *maximum information gain* as $\Upsilon := \max_{\lambda\in\triangle_{\tilde{\mathcal{X}}}} \log\det\left(I + \xi^{-1}K_\lambda\right)$, where $K_\lambda := [\sqrt{\lambda_i\lambda_j}k(\tilde{x}_i,\tilde{x}_j)]_{i,j\in[nV]}$ denotes the kernel matrix under sample allocation λ. Υ stands for the maximum mutual information between reward parameter θ^* and the samples generated according to sample allocation λ.

Let $\lambda^* := \operatorname{argmax}_{\lambda\in\triangle_{\tilde{\mathcal{X}}}} \log\det\left(I + \xi^{-1}K_\lambda\right)$ denote the sample allocation which achieves the maximum information gain, and let $\alpha_1 \geq \cdots \geq \alpha_{nV}$ denote the eigenvalues of K_{λ^*}. Then, we define the *effective dimension* as $d_{\text{eff}} := \min\{j \in [nV] : j\xi\log(nV) \geq \sum_{i=j+1}^{nV}\alpha_i\}$. d_{eff} stands for the number of principle directions that data projections in RKHS spread.

Recall that $K_{\mathcal{Z}} := [k_{\mathcal{Z}}(z_v, z_{v'})]_{v,v'\in[V]}$ denotes the kernel matrix of task similarities. Let $K_{\mathcal{X},\lambda^*} := [\sqrt{\lambda_i^*\lambda_j^*}k_{\mathcal{X}}(x_i,x_j)]_{i,j\in[nV]}$ denote the kernel matrix of arm features under sample allocation λ^*.

Corollary 1. *The per-agent sample complexity S of algorithm* `CoKernelFC` *can be bounded by*

$$\text{(a) } S = \tilde{O}\left(\frac{\Upsilon}{\Delta_{\min}^2 V}\right), \quad \text{(b) } S = \tilde{O}\left(\frac{d_{\text{eff}}}{\Delta_{\min}^2 V}\right), \quad \text{(c) } S = \tilde{O}\left(\frac{\texttt{Rank}(K_{\mathcal{Z}})\cdot\texttt{Rank}(K_{\mathcal{X},\lambda^*})}{\Delta_{\min}^2 V}\right),$$

where $\tilde{O}(\cdot)$ omits the rounding cost term $\tilde{d}(\xi,\lambda_1^*)\log\Delta_{\min}^{-1}/V$ and logarithmic factors.

Remark 2. Corollaries 1(a),(b) show that, our sample complexity can be bounded by the maximum information gain, and only depends on the effective dimension of kernel representation, which avoids the curse of dimensionality.

Corollary 1(c) reveals that the more tasks are similar (i.e., the smaller $\texttt{Rank}(K_{\mathcal{Z}})$ is), the fewer samples agents need. For example, when all tasks are the same, i.e., $\texttt{Rank}(K_{\mathcal{Z}}) = 1$, each agent only needs a $\frac{1}{V}$ fraction of the samples required by single-agent algorithms (Camilleri et al., 2021; Zhu et al., 2021) (which need $\tilde{O}(\frac{\texttt{Rank}(K_{\mathcal{X},\lambda^*})}{\Delta_{\min}^2})$ samples). Conversely, when all tasks are totally different, i.e., $\texttt{Rank}(K_{\mathcal{Z}}) = V$, no advantage can be obtained from communication, since the information from other agents is useless for solving local tasks. Our experimental results also reflect this relationship between task similarities and speedup, which match our theoretical bounds (see Appendix A). We note that these theoretical results hold for both finite and infinite RKHS.

4.4 Lower Bound for Fixed-Confidence Setting

We now give a lower bound for communication round-speedup to validate the optimality of `CoKernelFC`. We say an algorithm is δ-correct, if it returns the best arms $\tilde{x}_{v,*}$ for all $v \in [V]$ with probability at least $1-\delta$. In addition, we say an algorithm is fair, if when all tasks are the same, it assigns the same number of samples to each agent. `CoKernelFC` is a fair algorithm.

Theorem 2 (Fixed-Confidence Round-Speedup Lower Bound)**.** *There exists an instance of fixed-confidence CoPE-KB, where all tasks are the same, and a $o(1)$-correct, V-speedup, and fair algorithm $\mathcal{A}$ must use*

$$\Omega\left(\frac{\log\Delta_{\min}^{-1}}{\log\log\Delta_{\min}^{-1}}\right)$$

communication rounds.

Theorem 2 shows that logarithmic communication rounds are indispensable for attaining full speedup, which validates the communication optimality of `CoKernelFC`. Moreover, when our CoPE-KB problem reduces to classic MAB setting (Hillel et al., 2013; Tao et al., 2019), i.e., all agents solve a common classic best arm identification task, our upper and lower bounds (Theorems 1,2) match the state-of-the-art results (Tao et al., 2019).

Novel Lower Bound Analysis. Different from prior CoPE work (Tao et al., 2019) which only analyzes a 2-*armed* case *without* reward structures, in our lower bound analysis, we investigate general multi-armed instances with high-dimensional linear reward structures. Our novel techniques include *linear structured progress and instance transformation lemmas* (Lemmas 3, 5 in Appendix D.4.2), which handle the challenges due to different possible sample allocation on multiple arms and the change of probability measure in instance transformation with structured rewards.

5 FIXED-BUDGET COPE-KB

For the fixed-budget objective, we propose algorithm `CoKernelFB` and error probability guarantees. Due to space limit, we defer the pseudo-code and detailed description to Appendix C.2.

5.1 ALGORITHM `CoKernelFB`

`CoKernelFB` is also an elimination-based algorithm as `CoKernelFC`. However, different from `CoKernelFC`, to efficiently use given sample budget, `CoKernelFB` pre-determines the number of phases and the number of samples according to the principle dimension of arms, and successively cut down candidate arms to a half based on principle dimension, rather than based on reward gaps as in `CoKernelFC`. `CoKernelFB` also adopts the efficient kernelized estimator (in Section 4.2) to estimate the rewards of arms, so that agents only need to transmit average outcomes rather than all raw sample outcomes or the whole estimated reward parameter. Thus, `CoKernelFB` only requires a $O(nV)$ communication cost, instead of $O(N^{(t)})$ or $O(\dim(\mathcal{H}))$ as in adaptions of prior single-agent and fixed-budget algorithm (Katz-Samuels et al., 2020).

5.2 THEORETICAL PERFORMANCE OF `CoKernelFB`

Define the principle dimension of $\tilde{\mathcal{X}}$ as $\omega(\xi, \tilde{\mathcal{X}}) := \min_{\lambda \in \triangle_{\tilde{\mathcal{X}}}} \max_{\tilde{x}_i, \tilde{x}_j \in \tilde{\mathcal{X}}} \|\phi(\tilde{x}_i) - \phi(\tilde{x}_j)\|^2_{A(\xi,\lambda)^{-1}}$, where $A(\xi, \lambda) := \xi I + \sum_{\tilde{x} \in \tilde{\mathcal{X}}} \lambda_{\tilde{x}} \phi(\tilde{x})\phi(\tilde{x})^\top$. $\omega(\xi, \tilde{\mathcal{X}})$ represents the principle dimension of data projections in $\tilde{\mathcal{X}}$ to RKHS. Now we provide the error probability guarantee for `CoKernelFB`.

Theorem 3 (Fixed-Budget Upper Bound)**.** *Suppose that ξ satisfies $\xi \leq \frac{1}{16(1+\varepsilon)^2(\rho^*(\xi))^2\|\theta^*\|^2}$. Algorithm* `CoKernelFB` *uses at most T samples per agent, and returns the best arms $\tilde{x}_{v,*}$ for all $v \in [V]$ with error probability*

$$O\left(\exp\left(-\frac{TV}{\rho^*(\xi) \cdot \log(\omega(\xi, \tilde{\mathcal{X}}))}\right) \cdot n^2 V \log(\omega(\xi, \tilde{\mathcal{X}}))\right)$$

and $O(\log(\omega(\xi, \tilde{\mathcal{X}})))$ communication rounds.

It can be seen that, to guarantee an error probability δ, `CoKernelFB` only requires $\tilde{O}(\frac{\rho^*(\xi)}{V} \log \delta^{-1})$ sample budget, which attains full speedup when all tasks are the same. Theorem 3 can also be decomposed into components related to task similarities and arm features as in Corollary 1 (see Appendix E.1.2). Moreover, we also present a lower bound to demonstrate that when all tasks are the same, $\{\phi(\tilde{x})\}_{\tilde{x} \in \tilde{\mathcal{X}}}$ spans $\mathcal{H}$ and $\xi = 0$, to achieve full speedup, agents must use $\Omega(\frac{\log(\omega(0,\tilde{\mathcal{X}}))}{\log\log(\omega(0,\tilde{\mathcal{X}}))})$ communication rounds (see Appendix E.2). This corroborates the communication optimality of `CoKernelFB`.

6 CONCLUSION

In this paper, we propose a collaborative pure exploration in kernel bandit (CoPE-KB) model with fixed-confidence and fixed-budget objectives. CoPE-KB generalizes prior CoPE formulation from the single-task and classic MAB setting to allow multiple tasks and general reward structures. We propose novel algorithms with an efficient kernelized estimator and a novel communication scheme. Nearly matching upper and lower bounds are established to demonstrate the sampling and communication optimalities of our algorithms. Our results explicitly quantify the influences of task similarities on learning speedup, and only depend on the effective dimension of feature space.

7 REPRODUCIBILITY STATEMENT

To ensure that our results can be easily reproduced, we present complete proofs for all theoretical results in Appendix. Furthermore, the source code and implementation instructions for our experiments are also provided in the supplementary material.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Explore/exploit schemes for web content optimization. In *IEEE International Conference on Data Mining*, pp. 1–10. IEEE, 2009.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming*, 186(1): 439–478, 2021.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pp. 41–53. Citeseer, 2010.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pp. 1227–1237. PMLR, 2021.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Aniket Anand Deshmukh, Urun Dogan, and Clayton Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 4851–4859, 2017.

Yihan Du, Yuko Kuroki, and Wei Chen. Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7262–7270, 2021.

Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

Abhimanyu Dubey et al. Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, pp. 2740–2750, 2020.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6), 2006.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in Neural Information Processing Systems*, 32:10667–10677, 2019.

Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11543–11552, 2020.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7036–7045, 2019.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 26, pp. 854–862, 2013.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, volume 12, pp. 655–662, 2012.

Nikolai Karpov, Qin Zhang, and Yuan Zhou. Collaborative top distribution identifications with limited interaction. In *IEEE Annual Symposium on Foundations of Computer Science*, pp. 160–171, 2020.

Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Xiangyu Kong, Bo Xin, Fangchen Liu, and Yizhou Wang. Revisiting the master-slave architecture in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1712.07305*, 2017.

Andreas Krause and Cheng Soon Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pp. 2447–2455, 2011.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Chuanhao Li and Hongning Wang. Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 6529–6553. PMLR, 2022.

Chuanhao Li, Huazheng Wang, Mengdi Wang, and Hongning Wang. Communication efficient distributed learning for kernelized contextual bandits. *arXiv preprint arXiv:2206.04835*, 2022.

Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pp. 92–107. PMLR, 2022.

Keqin Liu and Qing Zhao. Decentralized multi-armed bandit with multiple distributed players. In *Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2010.

Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H Low, and Lachlan LH Andrew. Geographical load balancing with renewables. *ACM SIGMETRICS Performance Evaluation Review*, 39(3): 62–66, 2011.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pp. 1723–1742. PMLR, 2017.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 27, pp. 828–836, 2014.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pp. 1587–1594, 2013.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *IEEE Annual Symposium on Foundations of Computer Science*, pp. 126–146, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Da-shan Shiu. Optimal order simple regret for gaussian process bandits. *Advances in Neural Information Processing Systems*, 34:21202–21215, 2021.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Yue Wang, Eloy Garcia, David Casbeer, and Fumin Zhang. Cooperative control of multi-agent systems: Theory and applications. 2017.

Bin Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu, and Huchuan Lu. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15189, 2021.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

Yinglun Zhu, Dongruo Zhou, Ruoxi Jiang, Quanquan Gu, Rebecca Willett, and Robert Nowak. Pure exploration in kernel and neural bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Ling Zhuo, Cho-Li Wang, and Francis C.M. Lau. Document replication and distribution in extensible geographically distributed web servers. *Journal of Parallel and Distributed Computing*, 63 (10):927–944, 2003.