PARTIALLY LINEAR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

In contextual bandit problems, traditional methods for modeling the context–reward relationship typically employ either fully parametric or nonparametric approaches. Parametric approaches are computationally efficient and simple to implement, but they rely on restrictive assumptions that may not hold in complex real-world scenarios. Nonparametric methods offer greater flexibility, but may not fully leverage clear underlying patterns that could otherwise improve prediction. In this work, we bridge these two paradigms by proposing a partially linear model for the expected reward, in which a designated subset of contextual features influences the reward linearly while the remaining features act on it through an unknown function belonging to a reproducing kernel Hilbert space. We introduce PartLinUCB, an algorithm designed for this partially linear reward model. PartLinUCB naturally encompasses both linear and kernelized bandits as special cases, yielding a unified regret guarantee that recovers the established rates for each when the context–reward relationship is fully linear or fully nonparametric, respectively. We further establish that PartLinUCB is nearly minimax optimal for the radial basis function (RBF) kernel by deriving a lower bound that matches our regret upper bound up to logarithmic factors. Empirical studies demonstrate that PartLinUCB outperforms representative baseline methods and highlight the practical utility of the partially linear approach.

1 INTRODUCTION

Sequential decision-making systems are fundamental to many modern applications such as personalized recommendation systems and adaptive clinical trials (Li et al., 2010; Liao et al., 2019). The contextual bandits framework provides a formal mathematical model for these sequential decision-making problems. In the standard setting, an agent must make decisions using contextual information while optimizing a specific reward metric. Specifically, at each round t, an agent observes contexts $\{c_{t,a}\}_{a\in\mathcal{A}}$ associated with each available action $a \in \mathcal{A}$ in the action space $\mathcal{A}$. Based on this contextual information, the agent selects an action a_t and receives a corresponding reward r_{t,a_t}. The agent's primary objective is to maximize cumulative reward over a sequence of rounds. Accurate prediction of action outcomes can be improved by effectively incorporating contextual information into the bandit model. Therefore, appropriate modeling of the context-reward relationship is crucial for optimal performance of sequential decision-making systems.

Parametric reward models have been widely studied, with the linear model being the most fundamental: the expected reward is assumed to be a linear function of the context features. (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Agrawal & Goyal, 2014). Algorithms designed for this setting focus on learning the parameters of the linear model. With the parameters accurately estimated, the agent can make near-optimal decisions. However, the linearity assumption does not often hold in real-world applications. When the true relationship between context and reward is non-linear, this model misspecification can lead to degraded performance and potentially linear regret (Lattimore et al., 2020). To address these, researchers have introduced various nonparametric modeling approaches, where the reward is assumed to be an unknown, complex function of the context vector (Valko et al., 2013; Guan & Jiang, 2018; Zhou et al., 2020). While non-parametric methods offer greater flexibility in capturing intricate relationships, they may not efficiently exploit the underlying structural properties of the data when such structure exists, potentially resulting in higher sample complexity and a greater computational burden for learning the unknown function. A middle ground between these extremes has been explored under the name of *semiparametric* contextual bandits, which augment a

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

linear term with a nonparametric component. In existing formulations, however, the nonparametric component enters as a covariate-*independent* adjustment such as a global, time-varying bias (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim & Paik, 2019) and an arm-specific offset (Ou et al., 2019; Peng et al., 2019). Such models capture unobserved confounding and temporal drift, but not a nonlinear dependence of the reward on the features.

Despite their differences, these approaches all commit the *entire* context vector to one regime: fully linear, fully unstructured, or linear up to a covariate-free adjustment. This leads us to a natural question:

when the effects of some features are well understood to be linear in the reward model while those of others are not, can an agent retain the efficiency of linear bandits on the former while remaining fully flexible on the latter?

We answer this question by introducing a partially linear bandit whose expected reward decomposes into a linear effect of the designated features and an nonlinear effect of the remaining ones. This formulation builds on the partially linear model, which has been extensively studied in statistics (Engle et al., 1986; Robinson, 1988), econometrics (Schmalensee & Stoker, 1999), and biomedicine (Zeger & Diggle, 1994).

We summarize our main contributions as follows:

Contributions.

- **Novel reward model.** We introduce a new reward model, which we call the *partially linear contextual bandit*. To our knowledge, this is the first bandit model in which the nonparametric component is an explicit function of a designated subset of features, rather than a covariate-free adjustment. The model embraces fully linear and fully nonparametric contextual bandits as special cases. Prior knowledge about individual features plays a key role in the feature partition of the partially linear bandit.
- **Algorithmic design.** We propose PARTLINUCB (Algorithm 1), which estimates the linear and nonparametric components jointly through a single ridge-type criterion with per-block regularization so that each component is regularized at its own scale. Since the standard self-normalized inequality does not apply to the coupled estimator, we develop a compensated self-normalized argument to control its estimation error. This argument yields componentwise confidence sets for the linear parameter and the nonparametric component, together with a joint confidence set that drives optimistic action selection.
- **Regret guarantee.** We prove that PARTLINUCB attains a regret bound of $\widetilde{O}\big(\sqrt{T\max\{p, d_{\text{eff}}\}(p + d_{\text{eff}})}\big)$ (Theorem 3), where p is the dimension of the linear feature block, d_{eff} is the effective dimension of the kernel at horizon T, and $\widetilde{O}(\cdot)$ suppresses polylogarithmic factors. The guarantee specializes, at the two extremes, to the rates of the models that our setting generalizes. When $q = 0$, the bound recovers the near-optimal rate of algorithms developed for linear contextual bandits (Abbasi-Yadkori et al., 2011), while when $p = 0$, it recovers the corresponding rate for kernelized contextual bandits (Valko et al., 2013; Chowdhury & Gopalan, 2017).
- **Regret lower bound.** We establish a lower bound on the worst-case regret of any algorithm under the partially linear reward model for the Radial Basis Function (RBF) and Matérn kernels (Theorem 4). The bound is the sum of the lower bounds for a p-dimensional linear bandit and a kernelized bandit with the same kernel, showing that the difficulties of learning the two components contribute additively to the minimax regret. The remaining gap between our upper and lower bounds is inherited from the existing gap in kernelized-bandit theory: it is only polylogarithmic for the RBF kernel and of order $\sqrt{d_{\text{eff}}}$ for the Matérn kernel. Hence, PARTLINUCB is nearly minimax optimal for the RBF kernel, and jointly learning the two components incurs no additional loss beyond the individual costs of learning each component.
- **Numerical performance.** Experiments on both synthetic environments and real-world data show that PARTLINUCB consistently outperforms linear, nonparametric, and semiparametric baselines when the reward is partially linear. Moreover, in the two extreme cases, PARTLINUCB matches the specialized baselines, performing the same as LINUCB when the reward is fully linear ($q = 0$) and the same as KERNELUCB when the reward is fully nonparametric ($p = 0$).

2 Related Works

Contextual Bandits. The contextual bandits literature features a rich body of work (Auer et al., 2002; Langford & Zhang, 2007; Li et al., 2010; Tewari & Murphy, 2017; Lattimore & Szepesvári, 2020). One of the leading design paradigms employs upper confidence bound (UCB) algorithms, which implement the "optimism in the face of uncertainty" (OFU) principle for action selection (Lai & Robbins, 1985). This principle guides algorithms to favor actions with both high uncertainty and high estimated rewards, thereby naturally balancing exploration and exploitation. Within this paradigm, `LinRel` (Auer, 2003) was among the first UCB-based algorithms designed for linear reward settings and the method used for the linear model constructs confidence sets from self-normalized concentration inequalities to attain near-optimal regret (Chu et al., 2011; Abbasi-Yadkori et al., 2011). Beyond linearity, the same optimism template has been extended to kernelized reward models (Srinivas et al., 2010; Valko et al., 2013; Chowdhury & Gopalan, 2017), neural function classes (Allesiardo et al., 2014; Zhou et al., 2020), and other nonparametric or hybrid formulations (Guan & Jiang, 2018; Lee et al., 2024), where the regret typically scales with an effective dimension or information gain of the underlying function class.

Semiparametric Reward Models. The semiparametric contextual bandit problem has been studied to overcome the limitations of parametric models in practical settings. This line of work was initiated by Greenewald et al. (2017), who modeled the reward as the sum of a linear function of the context and a time-varying nonlinear term. Subsequent works (Krishnamurthy et al., 2018; Kim & Paik, 2019; Choi et al., 2022; Chowdhury et al., 2023; Kim et al., 2025) built on this foundation while retaining the nonlinear component as a global bias term that operates independently of the chosen action's features. Other formulations incorporate the nonparametric component as an arm-specific bias added to a parametric predictor (Ou et al., 2019; Peng et al., 2019). Similarly, misspecification models introduce an arm-indexed deviation from linearity (Ghosh et al., 2017). In all of these approaches, the covariate-independence of the nonlinear term is an essential ingredient of the analysis: a bias that is common to all actions offered in a round can be removed from the estimating equations through action centering or reward differencing (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim & Paik, 2019), allowing the linear parameter to be estimated as if the nonlinear term were absent.

Our model departs from existing semiparametric reward models by treating the nonlinear component as an explicit function of a pre-designated subset of the chosen action's features, rather than as a global or arm-specific adjustment. Because the nonlinear term varies across actions within a round, it cannot be eliminated through centering and must instead be estimated jointly with the linear component. As a result, the agent learns and explores with respect to both components, requiring theoretical techniques that differ from the centering-based analyses above.

Partially Linear Models. Among estimation methods for the partially linear model (Engle et al., 1986; Robinson, 1988; Speckman, 1988; Härdle et al., 2000), the closest to ours is the double-penalized objective of Müller & van de Geer (2015) and its distributed extension by Lv & Lian (2022). In both, the two components are estimated jointly by least squares with an ℓ_1 penalty on the linear coefficients and an RKHS-norm penalty on the nonparametric component. Related formulations (Xie & Huang, 2009; Ni et al., 2009; Cheng et al., 2015) also pair smoothness penalties on the nonlinear part with sparsity penalties on the linear part to perform variable selection in a high-dimensional regime, where the number of linear covariates may grow with or exceed the sample size.

Our work brings the partially linear model from the fixed-sample statistical setting to sequential decision-making with adaptively collected data. For regret analysis in the sequential setting, the estimation error must be controlled throughout the run rather than once at a final sample size, and this error bound must remain valid even though the design depends on the past noise. Neither guarantee is provided by the fixed-sample oracle inequalities developed for existing estimators. While we retain the RKHS-norm penalty on the nonparametric component, these demands motivate us to replace the sparsity penalty on the linear component with a ridge penalty, yielding coupled estimators whose errors are explicit linear functions of the noise. This representation allows us to apply self-normalized martingale concentration to obtain sequential confidence bounds. Extending the self-normalized analysis of linear contextual bandits (Abbasi-Yadkori et al., 2011) to our coupled

estimators is not immediate, because the weight on each noise variable depends on the kernel matrix of all rounds observed so far and the standard inequality no longer applies. Our analysis therefore residualizes both the features and the noise by kernel ridge prediction from the previous rounds, which restores the required pairing between predictable weights and noise, and corrects the resulting predictable bias of the residual noise through a compensated self-normalized argument. The resulting estimator reduces to ridge regression in the fully linear case and to kernel ridge regression in the fully nonparametric case, so that the bandit algorithm built on it recovers `LinUCB`-type and `KernelUCB`-type methods at the two extremes.

3 Partially linear contextual bandits

3.1 Notations and definitions

For a positive integer n, we write $[n] := \{1, \dots, n\}$. For a real Hilbert space $\mathcal{H}$, we write $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ for its inner product and norm, dropping the subscript in the Euclidean case. For a bounded, self-adjoint, positive definite operator A on $\mathcal{H}$, we define the weighted norm of $g \in \mathcal{H}$ as $\|g\|_A := \sqrt{\langle g,\, Ag \rangle_{\mathcal{H}}}$. For self-adjoint operators (in particular, symmetric matrices) A and B on the same space, $A \succeq B$ means that $A - B$ is positive semidefinite. We use standard asymptotic notation, and $\widetilde{O}(\cdot)$ suppresses polylogarithmic factors. Additional notation is collected in Appendix A.

3.2 Problem Setting

We consider a contextual bandit environment specified by a context space $\mathcal{X} \subseteq \mathbb{R}^{p+q}$ and an action space $\mathcal{A}$, which may be finite or infinite. Every context vector consists of p features that affect the expected reward through an unknown linear function and q features that affect it through an unknown nonlinear function. This partition is fixed over time and known to the agent, and without loss of generality we take the linear features to be the first p coordinates. The context of action a at round t then takes the form

$$c_{t,a} = \begin{pmatrix} x_{t,a}^\top & z_{t,a}^\top \end{pmatrix}^\top \in \mathbb{R}^{p+q}, \qquad x_{t,a} \in \mathbb{R}^p, \quad z_{t,a} \in \mathbb{R}^q.$$

Learning proceeds over T rounds. In each round t, the agent (i) observes the contexts $\{c_{t,a}\}_{a \in \mathcal{A}}$, (ii) selects an action $a_t \in \mathcal{A}$, and (iii) receives the corresponding reward r_{t,a_t}. Rewards follow the partially linear model

$$r_{t,a} = \langle x_{t,a}, \theta_* \rangle + f_*(z_{t,a}) + \varepsilon_t, \tag{1}$$

where $\theta_* \in \mathbb{R}^p$ is an unknown parameter vector, $f_* : \mathbb{R}^q \to \mathbb{R}$ is an unknown function, and ε_t is zero-mean random noise. We model f_* as an element of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated with a prespecified continuous kernel $k : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$. For $z \in \mathbb{R}^q$, let $\phi(z) := k(z, \cdot) \in \mathcal{H}_k$ denote the canonical feature at z. Then, by the reproducing property, every $f \in \mathcal{H}_k$ satisfies $f(z) = \langle f,\, \phi(z) \rangle_{\mathcal{H}_k}$.

The agent aims to maximize the cumulative reward or, equivalently, to minimize the *cumulative regret*

$$\mathrm{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \left(r_{t,a_t^*} - r_{t,a_t} \right) \right], \tag{2}$$

where $a_t^* := \arg\max_{a \in \mathcal{A}} \mathbb{E}[r_{t,a}]$ denotes the optimal action at round t.

Throughout the algorithm and the subsequent analysis, we write r_t, x_t, and z_t for the reward and the context vectors associated with the chosen action a_t, dropping the action subscript for notational simplicity.

Remark 1 (Relation to kernelized bandits)**.** The partially linear reward model can be viewed as a kernelized bandit on the full context $c_{t,a}$ using the additive kernel

$$k_+\big((x,z),(x',z')\big) = \langle x, x' \rangle + k(z, z'),$$

whose RKHS is the direct sum of the linear functions on $\mathbb{R}^p$ and $\mathcal{H}_k$. Generic kernel-bandit algorithms could, in principle, be applied using k_+. Folding all $p + q$ coordinates into a single generic kernel would charge the linear features at a nonparametric rate. In contrast, our regret bound matches the sum of the complexity of a p-dimensional linear bandit and that of a kernel bandit on z alone.

Following prior work on contextual bandits (Lattimore & Szepesvári, 2020; Chu et al., 2011; Abbasi-Yadkori et al., 2011), we make the following boundedness and sub-gaussian assumption.

Assumption 1 (Boundedness)**.** There exist constants $R_\theta, R_f, L_x, L_z > 0$ such that $\|\theta_*\| \le R_\theta$, $\|f_*\|_{\mathcal{H}_k} \le R_f$, $\|x_{t,a}\| \le L_x$, and $\|\phi(z_{t,a})\|_{\mathcal{H}_k} \le L_z$ for all $a \in \mathcal{A}$ and all $t \in [T]$.

Assumption 2 (Sub-Gaussian noise)**.** For all $t \in [T]$, the noise ε_t is conditionally R-sub-Gaussian: $\mathbb{E}[\exp(\alpha\varepsilon_t) \mid \mathcal{F}_{t-1}] \le \exp(\alpha^2 R^2/2)$ for all $\alpha \in \mathbb{R}$, where $\mathcal{F}_{t-1}$ is the σ-algebra generated by the history $\{x_{i,a_i}, z_{i,a_i}, r_{i,a_i}\}_{i \le t-1}$ together with the contexts $\{x_{t,a}, z_{t,a}\}_{a \in \mathcal{A}}$ observed at round t.

4 Main Results

4.1 The PartLinUCB algorithm

In this section, we introduce `PartLinUCB`, an algorithm for the partially linear model of Section 3.2 built on the OFU principle. Each round consists of three steps. The algorithm first computes joint ridge estimates of the two reward components, then constructs a confidence set around each estimate, and finally selects the action whose optimistic reward is maximized. The complete procedure is summarized in Algorithm 1.

Algorithm 1 `PartLinUCB`

Require: regularization parameters λ, γ, kernel k, confidence level δ

1: **for** $t = 1, \dots, T$ **do**
2: Observe the contexts $\{c_{t,a}\}_{a \in \mathcal{A}}$.
3: Compute the estimators $\widehat{\theta}_t$ and $\widehat{f}_t$ using (4).
4: Construct the confidence sets $\mathcal{C}_t^\theta$ and $\mathcal{C}_t^f$ with the radii E_t^δ given in (6) and (7).
5: Select the action

$$a_t \in \arg\max_{a \in \mathcal{A}} \left\{ x_{t,a}^\top \hat{\theta}_t + \hat{f}_t(z_{t,a}) + b_t^\delta(a) \right\} \text{ where } b_t^\delta(a) \text{ is given in (8)}$$

6: Observe the reward r_t and update the history $H_t = H_{t-1} \cup \{(x_t, z_t, r_t)\}$.
7: **end for**

Reward estimation

For t observations, let $\mathbf{r}_t = (r_1, \dots, r_t)^\top, X_t = (x_1, \dots, x_t)^\top \in \mathbb{R}^{n \times p}$. The evaluation operator $\Psi_t : \mathcal{H}_k \to \mathbb{R}^n$ and its adjoint satisfy $\Psi_t f = (f(z_1), \dots, f(z_t))^\top, \Psi_t^* a = \sum_{s=1}^n a_s \phi(z_s)$. The corresponding Gram matrix is $K_t = \Psi_t \Psi_t^*$ with entries $(K_t)_{ij} = k(z_i, z_j)$.

For fixed regularization parameters $\lambda, \gamma > 0$, the estimators are computed by solving the double-penalized least squares problem

$$(\widehat{\theta}_t, \widehat{f}_t) = \arg\min_{\theta \in \mathbb{R}^p,\, f \in \mathcal{H}_k} \left\{ \|\mathbf{r}_{t-1} - (X_{t-1}\theta + \Psi_{t-1} f)\|^2 + \lambda \|\theta\|^2 + \gamma \|f\|_{\mathcal{H}_k}^2 \right\}. \tag{3}$$

Theorem 1 (Closed form of estimators)**.** *For every $t > 1$, problem (3) has a unique solution*

$$\widehat{\theta}_t = V_{t-1}^{-1} X_{t-1}^\top \gamma (K_{t-1} + \gamma I)^{-1} \mathbf{r}_{t-1}, \qquad \widehat{f}_t = \Psi_{t-1}^* M_{t-1}^{-1} \mathbf{r}_{t-1}, \tag{4}$$

where

$$\begin{aligned} V_{t-1} &= X_{t-1}^\top \gamma (K_{t-1} + \gamma I)^{-1} X_{t-1} + \lambda I_p, \\ W_{t-1} &= \Psi_{t-1}^\top \lambda (X_{t-1} X_{t-1}^\top + \lambda I)^{-1} \Psi_{t-1} + \gamma I, \\ M_{t-1} &= \tfrac{\gamma}{\lambda} X_{t-1} X_{t-1}^\top + K_{t-1} + \gamma I. \end{aligned} \tag{5}$$

At $t = 1$, both estimates are zero.

Proof. See Appendix B.1. □

Remark 2 (Recovering `LinUCB` and `KernelUCB`)**.** Our framework recovers two well-known algorithms as special cases. When no linear features are present, $\widehat{f}_t$ reduces to

$$\widehat{f}_t(\cdot) = \phi(\cdot)^\top \Psi_{t-1}^\top \left(K_{t-1} + \gamma I\right)^{-1} \mathbf{r}_{t-1},$$

which is exactly the kernelized bandit estimator of `KernelUCB` (Valko et al., 2013). When no nonlinear features are present, $\widehat{\theta}_t$ reduces to the ridge regression estimator of `LinUCB` (Li et al., 2010). The confidence sets and their radii in the uncertainty modeling also reduce to those of the respective algorithms in both cases.

Uncertainty modeling

At each round, the algorithm maintains one confidence set for each component, centered at the estimators $(\hat{\theta}_t, \hat{f}_t)$. The following theorem shows that both sets contain the true components uniformly over time.

Theorem 2 (Confidence sets)**.** *For every $\lambda, \gamma > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, it holds simultaneously for all $t \geq 1$ that $\theta_* \in \mathcal{C}_t^\theta$ and $f_* \in \mathcal{C}_t^f$, where*

$$\mathcal{C}_t^\theta = \left\{\theta : \left\|\theta - \widehat{\theta}_t\right\|_{V_{t-1}} \leq E_t^\delta\right\}, \qquad \mathcal{C}_t^f = \left\{f \in \mathcal{H}_k : \left\|f - \widehat{f}_t\right\|_{W_{t-1}} \leq E_t^\delta\right\}, \tag{6}$$

with the radii

$$E_t^\delta = \sqrt{\lambda} R_\theta + \sqrt{\gamma} R_f + R\sqrt{\log\det\left(\frac{M_{t-1}}{\gamma}\right) + 2\log(\frac{1}{\delta})}. \tag{7}$$

Proof. See Appendix B.2. □

Discussion of Theorem 2. The radius in (7) consists of a regularization bias and a self-normalized noise term. The bias part makes the price of joint estimation visible. The nonparametric component enters the radius for θ_* through $\sqrt{\gamma} R_f$ and the linear component enters the radius for f_* through $\sqrt{\lambda} R_\theta$. The noise part measures the complexity of the two components jointly through the log-determinant.

The difficulty in the proof is that the standard self-normalized inequality does not apply to this noise term. In our estimator the weight on each noise variable mixes the features of all rounds observed so far and is therefore correlated with the noise it multiplies, whereas the inequality requires each weight to be fixed before its noise is revealed. Our sequential transformation replaces each feature and each noise variable by its residual after kernel ridge prediction from the previous rounds, which pairs every residual noise with a weight known at the start of its round. The residual noise is then biased by its own prediction, and our compensated self-normalized argument corrects this by adding back the noise energy captured by the kernel ridge fit, which is always nonnegative. The compensated quantity equals the self-normalized noise energy of the joint problem in which both components are estimated together, so the super-martingale argument yields the joint log-determinant in (7) and one event covers both confidence sets without a union bound. We believe this compensated self-normalized argument may be of independent interest. The detailed proof is provided in Appendix B.2.

Action selection

On the event of Theorem 2, the mean reward of every action is estimated to within the exploration bonus

$$b_t^\delta(a) := E_t^\delta \left(\left\| x_{t,a} - X_{t-1}^\top (K_{t-1} + \gamma I_{t-1})^{-1} \Psi_{t-1} \phi(z_{t,a}) \right\|_{V_{t-1}^{-1}} + \|\phi(z_{t,a})\|_{W_{t-1}^{-1}} \right), \tag{8}$$

for all $t \geq 1$ and $a \in \mathcal{A}$. `PartLinUCB` selects its maximizer (Line 5 of Algorithm 1).

4.2 Regret Upper Bound

We establish a regret upper bound of the proposed algorithm in Theorem 3. The rate depends on the kernel component only through its effective dimension (Hastie et al., 2001)

$$d_{\text{eff}}(\gamma, T) := \operatorname{Tr}\left(K_T(K_T + \gamma I_T)^{-1}\right), \tag{9}$$

Theorem 3 (Regret upper bound of `PartLinUCB`). *For any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $T \geq 1$,*

$$\mathit{Reg}_T \leq 2\max\{L_x R_\theta + L_z R_f,\ \sqrt{2}\, E_{T+1}^{\delta}\}\sqrt{\frac{T}{\log 2}\left(p\log\left(1+\tfrac{TL_x^2}{\lambda p}\right) + d_{\text{eff}}\log\left(e+\tfrac{eTL_z^2}{\gamma}\right)\right)},$$

so that $\mathit{Reg}_T = \widetilde{O}\left(\sqrt{T\ \max\{p, d_{\text{eff}}\}\,(p+d_{\text{eff}})}\right)$.

Remark 3 (Interpolation). The regret decomposes additively into a linear contribution of order $\widetilde{O}(p\sqrt{T})$ and a kernel contribution of order $\widetilde{O}(d_{\text{eff}}\sqrt{T})$, so `PartLinUCB` pays no asymptotic price for modeling the two components jointly while recovering the regret rates of `LinUCB` and `KernelUCB` at the two extremes.

Discussion of Theorem 3. The proof follows the standard optimism template, but the classical elliptical potential lemma does not apply. The lemma charges the width of each selected feature to the growth of the matrix that accumulates the feature, so the matrix must grow by one rank-one term per round. Our matrix V_t does not evolve this way in the raw features, because each new round updates $\gamma(K_t + \gamma I_t)^{-1}$ and thereby reweights every past feature. Hence the widths $\|x_t\|_{V_{t-1}^{-1}}$ do not telescope into a log-determinant. The resolution is the same residualization that underlies Theorem 2. The matrix V_t grows by exactly one rank-one term per round in the residual features, and the bonus in (8) was chosen to measure precisely these residual widths. By contrast, the seemingly natural bonus $E_t^{\delta}\left(\|x_{t,a}\|_{V_{t-1}^{-1}} + \|\phi(z_{t,a})\|_{W_{t-1}^{-1}}\right)$ also bounds the prediction error on the same event. Its widths are not the ones the design matrices accumulate, so no telescoping identity is available for them. The potential argument then holds as an identity rather than an inequality, since the widths of the selected actions telescope exactly into the log-determinant of (7). Each round of exploration is therefore charged once, to the same complexity that widens the confidence sets, and no coupling constant between the two components enters the bound. The detailed proof is provided in Appendix C.

4.3 Regret Lower Bound

We now present a lower bound for the worst-case expected regret with the RBF and Matérn kernels. The nonlinear feature ranges over the compact domain $\mathcal{Z} = [0,1]^q$ and the linear feature ranges over the scaled sign cube $\mathcal{X}_p := \{L_x\chi/\sqrt{p} : \chi \in \{-1,+1\}^p\}$. The noise is $\mathcal{N}(0, R^2)$. We write $\mathrm{Reg}_T^{\pi}(\theta, f)$ for the expected regret of a policy π under the parameters (θ, f). The worst-case expected regret of π is the supremum of $\mathrm{Reg}_T^{\pi}(\theta, f)$ over $\|\theta\|_2 \leq R_\theta$ and $\|f\|_{\mathcal{H}_k} \leq R_f$.

Theorem 4 (Regret lower bound). *Let $C > 0$ be a sufficiently large constant depending only on q, the kernel parameters, and (R, R_θ, R_f, L_x). For $T \geq Cp^2$, the worst-case expected regret of any policy π is at least*

1. *$\Omega\left(p\sqrt{T} + \sqrt{T}(\log T)^{q/4}\right)$ when $k = k_{\text{RBF}}$, and*
2. *$\Omega\left(p\sqrt{T} + T^{(\nu+q)/(2\nu+q)}\right)$ when $k = k_\nu$ is a Matérn kernel with $\nu > 0$,*

where the constants hidden in $\Omega(\cdot)$ depend only on the same quantities.

Discussion of Theorem 4. The bound is the sum of two known lower bounds. The first is the lower bound for linear bandits in dimension p (Dani et al., 2008). The second is the lower bound for a kernelized bandit with the same kernel and domain, which coincides with the best known lower bounds on compact Euclidean domains (Scarlett et al., 2018; Cai & Scarlett, 2021). Hence the partially linear problem is at least as hard as each of its two special cases and the two difficulties add. This additive form does not follow from the two existing arguments because a policy in our setting

observes both components through a single reward. Hence, a hard instance for one component does not control what the policy learns about the other. We therefore construct a single family of $2^p m$ instances indexed by a sign vector for θ and a peak location for f. Then, we average the regret over both indices at once. On the product action set the regret splits exactly into a linear part and a kernel part. Averaging over the sign vector for a fixed peak yields the linear term and averaging over the peak for a fixed sign vector yields the kernel term. This argument requires no independence between the observations of the two components. The kernel part follows the construction of Scarlett et al. (2017), which places bounded-norm bumps at well-separated locations. We simplify this construction in two ways. First, the bumps are products of squared sinc functions, so their overlap is controlled by Parseval's identity rather than by decay estimates. Second, the relative entropy between two instances is evaluated exactly along the queried actions rather than through a quantization of the domain into cells. As in the lower-bound literature for kernelized bandits, the theorem is stated on the compact domain $[0, 1]^q$ for the RBF and Matérn kernels.

Appendix D derives the theorem from a general lower bound that applies to any kernel and domain admitting a family of well-separated functions in the RKHS ball. The appendix also treats two further cases. On a finite domain with m points, the kernel term becomes $\Omega(\sqrt{mT})$ for every kernel with a positive definite Gram matrix. On $\mathbb{R}^q$, every policy suffers linear regret for every stationary kernel whose spectral density is bounded away from zero near the origin. The compactness of $\mathcal{Z}$ is therefore necessary for sublinear regret. The linear growth of γ_T on unbounded domains makes Theorem 3 vacuous there, and the lower bound shows that this reflects the problem rather than the analysis. Compared with Theorem 3, the linear term matches up to logarithmic factors. The kernel term matches up to the gap that already separates confidence-bound analyses of kernelized bandits from their lower bounds. This gap is polylogarithmic for the RBF kernel and of order $\sqrt{\gamma_T}$ for the Matérn kernel. The detailed proof is provided in Appendix D.

5 Numerical Experiments

We evaluate the empirical performance of `PartLinUCB` against `LinUCB` (Li et al., 2010), `KernelUCB` (Valko et al., 2013), the neural contextual bandit `EE-Net` (Ban et al., 2022), and a uniform-random baseline that serves as a naive performance floor. `EE-Net` is run with the authors' default architecture and learning rates (Appendix E). We additionally report two oracle baselines based on `PartLinUCB`: the *f-oracle*, which is given the nonlinear component and learns only θ, and the *θ-oracle*, which is given the linear component and learns only f (Appendix E).

5.1 Synthetic Environments

For each round, the context vectors are drawn uniformly from the unit sphere in $\mathbb{R}^{p+q}$, where the first p coordinates enter the reward linearly and the remaining q coordinates enter nonlinearly. We cover regimes from predominantly linear to predominantly nonlinear by varying (p, q) over $\{4, 8, 12\}^2$. The number of actions is fixed to $K = 10$. We consider two ground-truth reward models,

$$\text{Polynomial: } x^\top \theta_* + z^\top A z, \qquad \text{Periodic: } x^\top \theta_* + \cos\langle z,\, w_* \rangle\,,$$

where θ_* and w_* are sampled uniformly from the unit sphere and the sphere of radius 5, respectively, and $A_{ij} \sim \mathcal{N}(0, 1)$ i.i.d. We use the RBF kernel $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$ with $\sigma = 1$, and set $\lambda = \gamma = 1$. We conducted 20 independent runs of $T = 2000$ rounds for each configuration.

As shown in Figure 1, `PartLinUCB` outperforms all feasible baselines in every configuration, confirming that partitioning the context into linear and nonlinear components is more sample-efficient than committing to a single model class. `PartLinUCB` also closely tracks the combined regret of the two oracles, mirroring the additive structure of the bound in Theorem 3. As shown in Appendix E, the regret of both `KernelUCB` and `PartLinUCB` increases with p at fixed q, but the increase is markedly smaller for `PartLinUCB`.

5.2 MovieLens

For the real-data evaluation, we adapt the protocol of Ban et al. (2022) to MovieLens 1M (Harper & Konstan, 2015). Each round presents $K = 10$ user–movie pairs: one with a rating of at most 2

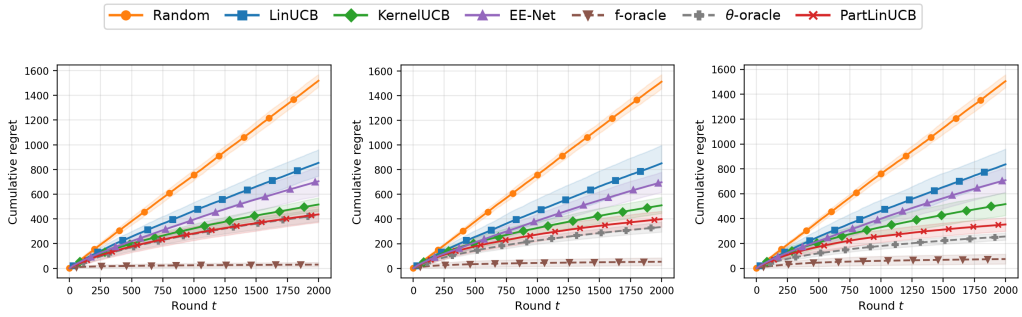

Figure 1: Cumulative regret (mean ± 2 standard deviations over 20 runs) in the Periodic environment for representative configurations $(p, q) \in \{(4, 12), (8, 8), (12, 4)\}$. Complete results for all nine (p, q) configurations and both reward models are reported in Appendix E.

and nine with ratings above 2. The reward is 1 only when the low-rated pair is selected. Contexts concatenate the user's one-hot demographics ($p = 30$) with 10-dimensional SVD-based user and movie representations ($q = 20$). We report mean cumulative regret over 20 runs of $T = 3000$ rounds.

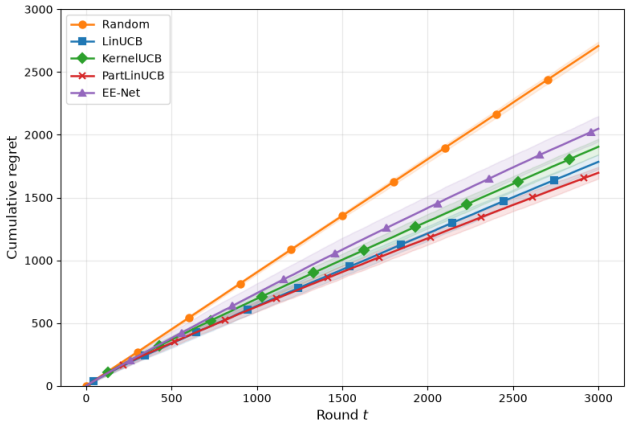

Figure 2: Cumulative regret on MovieLens over $T = 3000$ rounds (mean ± 2 standard deviations over 20 seeds).

In Figure 2, `PartLinUCB` achieves the smallest cumulative regret. The partially linear structure of `PartLinUCB` is beneficial, since the demographic attributes behave linearly, while the latent preference factors captured by the SVD embeddings are better modeled nonlinearly. `PartLinUCB` successfully adapts to both linear and nonlinear modeling, lowering the regret.

6 Conclusion

We introduced a partially linear contextual bandit model and proposed PartLinUCB with high-probability confidence sets, regret bounds, and estimators combining a linear ridge component with a kernelized component. Across synthetic and real-world datasets, PartLinUCB consistently outperforms strong baselines, showing that the hybrid semiparametric models improve data efficiency without sacrificing flexibility. Natural extensions of the model include nonstationary rewards and additional structure on the two components, such as sparsity in the linear part or additivity in the nonparametric part, which would replace the dimensions in our regret bounds with smaller structure-dependent quantities.

AI USE STATEMENT

We use LLMs as a tool to refine our writing and correct grammatical errors and typos.

REFERENCES

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

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2014.

Robin Allesiardo, Raphael Feraud, and Djallel Bouneffouf. A neural networks committee for the contextual bandit problem. *arXiv preprint arXiv:1409.8191*, 2014.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003.

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

Yikun Ban, Yuchen Yan, Arindam Banerjee, and Jingrui He. EE-Net: Exploitation-exploration neural networks in contextual bandits. In *International Conference on Learning Representations (ICLR)*, 2022.

Xu Cai and Jonathan Scarlett. On lower bounds for standard and robust Gaussian process bandit optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1216–1226. PMLR, 2021. URL https://proceedings.mlr.press/v139/cai21f.html.

Guang Cheng, Hao Helen Zhang, and Zuofeng Shang. Sparse and efficient estimation for partial spline models with increasing dimension. *Annals of the Institute of Statistical Mathematics*, 67 (1):93–127, 2015.

Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik, and Myunghee Cho Paik. Semi-parametric contextual bandits with graph-Laplacian regularization. *arXiv preprint arXiv:2205.08295*, 2022.

Mubarrat Chowdhury, Elkhan Ismayilzada, Khalequzzaman Sayem, and Gi-Soo Kim. GBOSE: Generalized bandit orthogonalized semiparametric estimation. *arXiv preprint arXiv:2301.08781*, 2023.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, 2008. URL https://www.learningtheory.org/colt2008/papers/80-Dani.pdf.

Robert F. Engle, C. W. J. Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. *arXiv preprint arXiv:1704.06880*, 2017.

Kristjan Greenewald, Ambuj Tewari, Predrag Klasnja, and Susan Murphy. Action centered contextual bandits. *arXiv preprint arXiv:1711.03596*, 2017.

Melody Y. Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. *arXiv preprint arXiv:1801.01750*, 2018.

Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially Linear Models*. Contributions to Statistics. Physica-Verlag, Heidelberg, 2000. doi: 10.1007/978-3-642-57700-0.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA, 2001.

Shogo Iwazaki. Tighter regret lower bound for Gaussian process bandits with squared exponential kernel in hypersphere. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2602.17940.

Gi-Soo Kim and Myunghee Cho Paik. Contextual multi-armed bandit algorithm for semiparametric reward model. *arXiv preprint arXiv:1901.11221*, 2019.

Seok-Jin Kim, Gi-Soo Kim, and Min-hwan Oh. Experimental design for semiparametric bandits. In *Proceedings of the Thirty Eighth Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3215–3252. PMLR, 2025.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. *arXiv preprint arXiv:1803.04204*, 2018.

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

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. *arXiv preprint arXiv:1911.07676*, 2020.

Kyungbok Lee, Myunghee Cho Paik, Min-hwan Oh, and Gi-Soo Kim. Mixed-effects contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13409–13417, 2024.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010.

Peng Liao, Kristjan H. Greenewald, Predrag V. Klasnja, and Susan A. Murphy. Personalized HeartSteps: A reinforcement learning algorithm for optimizing physical activity. *arXiv preprint arXiv:1909.03539*, 2019.

Shaogao Lv and Heng Lian. Debiased distributed learning for sparse partial linear models in high dimensions. *Journal of Machine Learning Research*, 23(2):1–32, 2022.

Patric Müller and Sara van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42(2):580–608, 2015.

Xiao Ni, Hao Helen Zhang, and Daowen Zhang. Automatic model selection for partially linear models. *Journal of Multivariate Analysis*, 100(9):2100–2111, 2009. doi: 10.1016/j.jmva.2009.06.009.

Mingdong Ou, Nan Li, Cheng Yang, Shenghuo Zhu, and Rong Jin. Semi-parametric sampling for stochastic bandits with many arms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7933–7940, 2019.

Yi Peng, Miao Xie, Jiahao Liu, Xuying Meng, Nan Li, Cheng Yang, Tao Yao, and Rong Jin. A practical semi-parametric contextual bandit. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3246–3252, 2019.

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, pp. 931–954, 1988.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1723–1742. PMLR, 2017. URL https://proceedings.mlr.press/v65/scarlett17a.html.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. *arXiv preprint arXiv:1706.00090*, 2018. URL https://arxiv.org/pdf/1706.00090v3. Corrected version 3, May 31, 2018; Theorem 2.

Richard Schmalensee and Thomas M. Stoker. Household gasoline demand in the United States. *Econometrica*, 67(3):645–662, 1999.

Paul Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(3):413–436, 1988.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010. URL https://icml.cc/Conferences/2010/papers/422.pdf.

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications*. Springer, 2017.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 654–663. AUAI Press, 2013. URL https://auai.org/uai2013/prints/papers/161.pdf.

Huiliang Xie and Jian Huang. SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696, 2009. doi: 10.1214/07-AOS580.

S. L. Zeger and Peter Diggle. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3):689–699, 1994.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. *arXiv preprint arXiv:1911.04462*, 2020.