LEARNING STACKELBERG EQUILIBRIA FROM PREFERENCES OVER UNKNOWN MULTI-OBJECTIVE TRADE-OFFS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactions among self-interested agents often lead to socially inefficient outcomes. A central authority can counteract this by introducing incentives that steer agents toward more desirable outcomes, a setting naturally captured by a Stackelberg game in which the authority acts as the leader, the agents as followers, and the Stackelberg equilibrium characterizes the desired outcome. Defining what is "desirable", however, typically requires trading off multiple, possibly competing objectives. While each objective may be easy to evaluate on its own, assigning numerical weights that reflect their relative importance and combine them into a single scalar cost is often difficult. Existing methods for solving Stackelberg games typically assume the leader observes this scalar cost directly. We instead consider a leader who does not know the trade-off parameter, but observes each objective's value at an approximate follower equilibrium and receives a pairwise preference between consecutive realizations. We jointly estimate the equilibrium-induced objective maps via kernel ridge regression and the trade-off parameter via logistic maximum likelihood, combining their confidence bounds in an optimistic action-selection rule. This yields a high-probability regret bound that captures the objective and trade-off estimation errors together with the inexactness of the followers' response. We establish conditions under which the resulting policy achieves no regret and further translate these guarantees into a finite-time criterion for extracting an ϵ-Stackelberg policy-response pair from the interaction history. The framework is validated in a carbon-tax design study, showing that the leader can recover a socially desirable policy despite inexactness in the followers' responses.

1 INTRODUCTION

Many application domains, including competitive markets (Metcalf & Weisbach, 2009), communication systems (Kelly et al., 1998), and traffic networks (Maljkovic et al., 2023), consist of self-interested agents whose individually optimal decisions may not result in a desirable system-level outcome. A natural remedy is to introduce a central authority that can steer such agents toward a societally desirable outcome through interventions such as prices, taxes, tolls, or subsidies. This hierarchical interaction is naturally captured by a single-leader multi-follower Stackelberg game (Von Stackelberg, 1952), where the leader (central authority) commits to an action, the followers (agents) respond by playing a non-cooperative game parameterized by that action, and the quality of the intervention is evaluated at the resulting follower equilibrium.

What constitutes a desirable outcome, however, is often inherently multidimensional, requiring trade-offs among multiple, potentially competing objectives. For example, regulatory authorities may balance market efficiency against equity; transportation planners may seek to reduce congestion while improving spatial accessibility; environmental regulators may weigh emissions reductions against economic burden; and central banks must weigh competitiveness against safety when setting capital adequacy ratios for private banks. Although each objective can often be measured directly, assigning numerical weights that capture their

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.

relative importance and thereby combine them into a single scalar cost can be difficult. Instead, it may be more natural to express such trade-offs through comparisons. Moreover, the system-level outcome depends on how followers respond to the leader's intervention. Because followers' utilities are private, the leader cannot explicitly characterize how their equilibrium responses vary with its actions. Identifying desirable interventions is therefore challenging for two reasons: the weights defining the scalar cost are unknown, and the map from the leader's actions to the followers' equilibrium responses is unknown. This motivates learning socially desirable outcomes directly from repeated leader–follower interactions.

Existing approaches to learning Stackelberg equilibria differ substantially in the feedback available to the leader and the information required about the underlying game. When a differentiable model of the lower-level game is available, the leader can exploit equilibrium sensitivity and gradient information to optimize its induced objective (Dontchev & Rockafellar, 2014; Fiez et al., 2020; Grontas et al., 2024). Other approaches instead learn directly from interaction, therefore requiring less information about the followers. This includes learning in finite-action games from bandit observations (Bai et al., 2021), optimization from revealed follower responses (Roth et al., 2016), and continuous Stackelberg games in which the lower level is treated as a black box (Maheshwari et al., 2024; Maddux et al., 2025).

Our work builds on the latter line of research, particularly (Maddux et al., 2025), where the leader has a scalar cost and observes its evaluation at an approximate follower Nash equilibrium. We extend this work to more general settings in which the leader's cost captures an unknown trade-off among several competing objectives. In such settings, asking an expert or other evaluator to compare two realized outcomes may be more natural than to assign an absolute score to either one (Christiano et al., 2017; Zhu et al., 2021; Jerry Lin et al., 2022; Rafailov et al., 2023). Accordingly, we assume that the leader observes a vector of measurable objectives and receives pairwise preference feedback over consecutive realized outcomes, rather than direct observations of the underlying scalar objective.

Learning from pairwise comparisons is well studied in isolation. Probabilistic preference models such as Bradley–Terry (Bradley & Terry, 1952; Luce et al., 1959) underpin preference learning (Chu & Ghahramani, 2005), dueling bandits (Yue et al., 2012; Dudík et al., 2015), and preferential Bayesian optimization (González et al., 2017). Closely related to our setting, multi-attribute preferential optimization considers vector-valued outcomes whose unknown utility is learned from pairwise comparisons (Astudillo & Frazier, 2020; Jerry Lin et al., 2022). More recently, kernelized preference bandits have provided regret guarantees and confidence sets for latent preference functions (Pásztor et al., 2024), complemented by finite-sample confidence analyses for parametric preference models (Lee et al., 2025; Schlaginhaufen et al., 2025). The key distinction in our setting is that the quantities entering the preference model are not known functions of the leader's actions, but instead arise from an approximate equilibrium of an unknown follower game. The leader must therefore jointly learn each objective function evaluated not at the exact but at an approximate follower Nash equilibrium and the unknown scalarization that determines how each objective is traded off. These two estimation problems are further coupled through action selection as uncertainty about either component translates into uncertainty about the leader's equilibrium cost. This distinguishes our setting from both preference learning with directly parameterized utilities and previous scalar-feedback Stackelberg learning formulations, and leads to the central question considered in this work:

Can the leader learn a Stackelberg equilibrium from numerical observations of each performance objective evaluated at approximate equilibria and binary preferences over its own past interventions?

We answer this question affirmatively and summarize our main contributions as follows:

- In Section 2, we introduce a preference-based extension of follower-agnostic Stackelberg learning where the followers' game is unknown, for settings in which the leader's cost is an unknown linear scalarization of a known vector of measurable performance objectives.
- In Section 3, we jointly learn each objective function evaluated at an approximate follower equilibrium via kernel ridge regression and its unknown scalarization via constrained logistic maximum

likelihood under a Bradley–Terry preference model. Their confidence bounds are combined into an optimistic surrogate cost that accounts for preference uncertainty, feature-map estimation, and inexact follower responses, and is used for the leader's action selection mechanism.

- In Section 3, Theorem 3.4 establishes a high-probability regret bound separating these three sources of uncertainty. We show that for a properly chosen lower-level approximation accuracy, the regret scales as $\widetilde{O}(m\sqrt{\gamma_T T})$, yielding no regret whenever $\gamma_T = o(T)$. Corollary 3.5 further provides a computable criterion, given prior bounds on the structural game constants, for extracting an ϵ-Stackelberg policy-response pair from the interaction history with high probability. Section 4 illustrates the method in carbon-tax incentive design and examines the effect of lower-level inexactness.

Notation: For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$. For a positive semidefinite matrix V, we write $\|x\|_V = \sqrt{x^\top V x}$, and I_n denotes the $n \times n$ identity matrix. The indicator of event A is $\mathbb{1}\{A\}$. For a kernel k, $\mathcal{H}_k$ denotes its reproducing kernel Hilbert space with norm $\|\cdot\|_k$. $\tilde{O}(\cdot)$ suppresses logarithmic factors.

2 Problem Setup

We consider a Stackelberg game with $N + 1$ agents: a leader L and a set of followers $\mathcal{N} = [N]$. In a Stackelberg game, the leader first chooses an action $\pi \in \Pi$ from its action set $\Pi \subseteq \mathbb{R}^p$. The followers then simultaneously respond by choosing actions $x_i \in \mathcal{X}_i \subseteq \mathbb{R}^d$, where $\mathcal{X}_i$ denotes the action set of follower $i \in \mathcal{N}$. The leader has a cost function $J : \Pi \times \mathcal{X} \to \mathbb{R}$ and each follower has a utility function given by $U_i : \mathcal{X} \times \Pi \to \mathbb{R}$, where $\mathcal{X} = \Pi_{i=1}^N \mathcal{X}_i$ denotes the action space of the joint action profile $x \coloneqq (x_i)_{i \in \mathcal{N}}$.

Given the leader's action π, the followers' decision-making at the lower level corresponds to a parameterized game $\Gamma(\pi) = \{\mathcal{N}, \{\mathcal{X}_i\}_{i=1}^N, \{U_i\}_{i=1}^N; \pi\}$, where each agent aims to maximize its utility given the actions of the other agents and the action of the leader. A popular solution concept in such games is the Nash equilibrium at which no agent has an incentive to unilaterally deviate from its action.

Definition 2.1 (ϵ-Nash equilibrium)**.** For a given π, the joint action profile $x^\star \in \mathcal{X}$ is an ϵ-Nash equilibrium if $U_i(x_i^\star, x_{-i}^\star; \pi) \geq U_i(x_i, x_{-i}^\star; \pi) - \epsilon$ holds for all $i \in \mathcal{N}$ and $x_i \in \mathcal{X}_i$. Here $x_{-i} \coloneqq (x_j)_{j \in \mathcal{N} \setminus \{i\}}$ denotes the joint action of all agents except i. The action profile $x^\star$ is a Nash equilibrium if $\epsilon = 0$.

At the upper level, the leader aims to minimize its cost J given the followers' response $x^\star(\pi)$. When $x^\star(\pi)$ is unique, which is the case, for example, for concave and strongly monotone games (Rosen, 1965, Theorem 1 & 2), the resulting Stackelberg game can be defined as a nonsmooth, nonconvex optimization problem

$$\min_{\pi \in \Pi} J(\pi, x^\star(\pi)) \;\text{ s.t. }\; x_i^\star(\pi) = \arg\max_{x_i \in \mathcal{X}_i} U_i(x_i, x_{-i}^\star(\pi); \pi),\ \forall i \in \mathcal{N}. \tag{1}$$

To ensure the existence and uniqueness of $x^\star(\pi)$, we adopt the following two assumptions.

Assumption 2.2 (Concavity)**.** $\Gamma(\pi)$ is a concave game for every $\pi \in \Pi$. Namely, for each agent $i \in \mathcal{N}$ the set $\mathcal{X}_i$ is non-empty, compact, and convex, and the utility function $U_i(x_i, x_{-i}; \pi)$ is continuously differentiable in x, and concave in x_i for all $x_{-i} \in \Pi_{j \neq i} \mathcal{X}_j$ and all $\pi \in \Pi$.

Assumption 2.3 (Strong monotonicity)**.** For every $\pi \in \Pi$, $\Gamma(\pi)$ is α-strongly monotone, i.e., the game pseudogradient $v : \mathbb{R}^{Nd} \times \Pi \to \mathbb{R}^{Nd}$ defined as $v(x; \pi) = (v_i(x; \pi))_{i=1}^N$, where $v_i(x; \pi) = -\nabla_{x_i} U_i(x_i, x_{-i}; \pi)$, satisfies $\langle v(x; \pi) - v(x'; \pi), x - x' \rangle \geq \alpha \|x - x'\|^2$ for all $x, x' \in \mathcal{X}$ and some $\alpha > 0$.

In this work, we focus on the leader who aims to find a *Stackelberg equilibrium* of game (1), i.e., a stable outcome at which neither the leader nor the followers have an incentive to unilaterally deviate.

Definition 2.4 (ϵ-Stackelberg equilibrium)**.** A joint action profile $(\pi^\star, x^\star) \in \Pi \times \mathcal{X}$ is an ϵ-Stackelberg equilibrium if (i) $x^\star$ is an ϵ-Nash equilibrium of $\Gamma(\pi^\star)$ and (ii) $J(\pi^\star, x^\star) \leq J(\bar{\pi}, x^\star(\bar{\pi})) + \epsilon$ for all $\bar{\pi} \in \Pi$. The joint action profile $(\pi^\star, x^\star)$ is a Stackelberg equilibrium if $\epsilon = 0$.

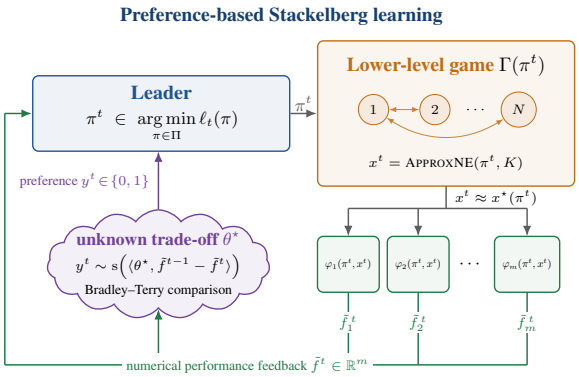

Figure 1: **Overview.** At round t, the leader selects π^t, which induces an approximate follower equilibrium x^t. The leader observes the m objective values $\varphi_j(\pi^t, x^t)$, while the corresponding scalar cost remains hidden. Instead, the binary preference y^t between consecutive realizations provides information about the unknown trade-off parameter $\theta^\star$. The leader uses these observations to learn the equilibrium maps f_j and $\theta^\star$, and combines their estimates in the optimistic acquisition function $\ell_t(\pi)$ in (7) to select the next action.

In general, however, the leader lacks access to the followers' utility functions and therefore cannot evaluate the equilibrium response map $\pi \mapsto x^\star(\pi)$ in closed form. This makes direct computation of a Stackelberg equilibrium impossible, which motivates learning it through repeated interactions.

Specifically, we consider repeated interactions between the leader and the followers over several rounds. At round t, the leader first announces an action π^t. The followers then run an equilibrium-seeking subroutine $\text{APPROXNE}(\pi^t, K)$ for K iterations and return $x^t(\pi^t) \in \mathcal{X}$. We treat the lower-level procedure as a *black box* and require only a rate at which its output approaches the unique Nash equilibrium.

Assumption 2.5 (Lower-level approximation)**.** The subroutine $\text{APPROXNE}(\pi^t, K)$ satisfies $\mathbb{E}\big[\|x^\star(\pi^t) - x^t(\pi^t)\|_2^2\big] \le C(\pi^t)K^{-c}$, for constants $c \in \mathbb{R}_{>0}$ and a function $C(\cdot)$ with $C = \max_{\pi\in\Pi} C(\pi) < \infty$.

This assumption does not require the leader to know the internal structure of followers' learning dynamics. It only quantifies the accuracy attained after a finite lower-level computation and is satisfied by several equilibrium-seeking dynamics for strongly monotone games (Bravo et al., 2018; Drusvyatskiy et al., 2022).

2.1 Information structure and regularity

In this work, we consider learning the Stackelberg equilibrium in settings where the leader's cost function is a trade-off between several interpretable performance measures. Let $\varphi = (\varphi_1, \ldots, \varphi_m) : \Pi \times \mathcal{X} \to \mathbb{R}^m$ denote a known performance map, whose components represent the objectives the leader can observe, evaluated at the response of the followers $x^t(\pi^t)$. While the leader can measure each φ_j, the trade-off that aggregates them into a scalar societal cost is unknown. Consequently, the leader neither knows its equilibrium-induced cost function nor directly observes its realized scalar value.

Assumption 2.6 (Linear multi-objective leader cost)**.** There exists an unknown weight vector $\theta^\star \in \mathbb{R}^m$ with $\|\theta^\star\|_2 \le S$ such that $J(\pi, x) = \langle \theta^\star, \varphi(\pi, x) \rangle$.

At the exact Nash equilibrium of the followers $x^\star(\pi)$, define the *equilibrium feature map*

$$f(\pi) = \varphi(\pi, x^\star(\pi)) \in \mathbb{R}^m, \tag{2}$$

so that, with a slight abuse of notation, $J(\pi) \coloneqq J(\pi, x^\star(\pi)) = \langle \theta^\star, f(\pi) \rangle$. To relate the performance maps at $x^t(\pi^t)$ to $f(\pi^t)$, we impose the following regularity condition.

Assumption 2.7 (Lipschitz objective)**.** $\|\varphi(\pi, x) - \varphi(\pi, x')\|_2 \leq L_\varphi \|x - x'\|$ for all $x, x' \in \mathcal{X}$ and $\pi \in \Pi$.

Note that for $m = 1$, our setup recovers the Stackelberg learning setting of (Maddux et al., 2025). For $m > 1$, all equilibrium features f_j are themselves unknown since the equilibrium response map $\pi \mapsto x^\star(\pi)$ must be learned. We therefore impose the standard reproducing kernel Hilbert space (RKHS) regularity on each f_j.

Assumption 2.8 (RKHS regularity of the equilibrium features)**.** Let Π be compact and, for each $j \in [m]$, $f_j \in \mathcal{H}_{k_j}(\Pi)$ with $\|f_j\|_{k_j} \leq B_j$, where k_j is a positive semidefinite kernel normalized with $k_j(\pi, \pi) \leq 1$.

Let $B = \max_{j \in [m]} B_j$ and $B_f^2 = \sum_{j=1}^m B_j^2$. Then $|f_j(\pi)| = |\langle f_j, k_j(\pi, \cdot) \rangle_{k_j}| \leq B_j \sqrt{k_j(\pi, \pi)} \leq B_j$, so $\|f(\pi)\|_2 \leq B_f$ and $\|\varphi(\pi, x)\|_2 \leq B_f + L_\varphi M = L$, where $M = \operatorname{diam}(\mathcal{X}) = \max_{x, x' \in \mathcal{X}} \|x - x'\|$.

Assumption 2.8 is an operational regularity condition standard in kernelized bandit literature. Although it is stated directly in terms of the composed equilibrium features, it can be related to regularity of the performance map φ on the joint space $\Pi \times \mathcal{X}$ and sensitivity of $x^\star(\pi)$, as discussed in Appendix B. First, RKHS regularity of φ_j on $\Pi \times \mathcal{X}$ induces RKHS regularity of f_j through an equilibrium embedding and the associated pullback kernel (Proposition B.1). Second, if the lower-level game's pseudogradient $v(x; \pi)$ and φ are also Lipschitz in π, standard sensitivity arguments imply that each f_j is Lipschitz (Proposition B.2), and hence continuous on the compact set Π. Such features can therefore be uniformly approximated to arbitrary accuracy by RKHS functions associated with universal kernels, including the Matérn class (Sriperumbudur et al., 2011).

Figure 1 summarizes the complete interaction structure. After each round t, we assume the leader observes numerical measurements of each realized performance objective, generalizing the scalar cost feedback of (Maheshwari et al., 2024; Maddux et al., 2025). In addition, to learn about the trade-off between different performance objectives, the leader queries an oracle, such as an expert or automated evaluator, for a pairwise comparison between the current action π^t and π^{t-1}. Such feedback is natural when absolute evaluations are unreliable or costly, while pairwise judgments are easier to elicit (Rafailov et al., 2023). We model these comparisons using the well-established Bradley–Terry model (Bradley & Terry, 1952; Luce et al., 1959).

Assumption 2.9 (Leader feedback)**.** At the end of round t, the leader observes $\tilde{f}^t \coloneqq \varphi(\pi^t, x^t(\pi^t)) \in \mathbb{R}^m$, and for $t \geq 2$, also a preference signal $y^t \in \{0, 1\}$ indicating whether the current action is preferred over the previous one. Let $\Delta^t = \tilde{f}^{t-1} - \tilde{f}^t$ and $\mathcal{G}_t$ be the information available after $\tilde{f}^t$ is observed and before y^t is revealed. Then, $\mathbb{P}\big(y^t = 1 \mid \mathcal{G}_t\big) = \mathrm{s}(\langle \theta^\star, \Delta^t \rangle)$, where $\mathrm{s}(x) = 1/(1 + e^{-x})$ is the sigmoid function.

Note that the sign convention in Δ^t reflects that J is a cost, so a realization with lower cost is more likely to be preferred. The history $\mathcal{G}_t$ collects all information available before y^t is revealed, ensuring that Δ^t is $\mathcal{G}_t$-measurable. This formulation will later allow us to establish the confidence bound for $\theta^\star$ in Lemma 3.3.

2.2 Learning objective

We evaluate the leader's learning policy via its cumulative regret over T iterations, measured relative to the best fixed leader action under the exact lower-level Nash equilibrium

$$R^T = \sum_{t=1}^{T} J(\pi^t, x^t(\pi^t)) - T \min_{\pi \in \Pi} J(\pi, x^\star(\pi)). \tag{3}$$

Regret R_T captures two sources of suboptimality: the leader not knowing $\pi^* \in \arg\min_{\pi \in \Pi} J(\pi, x^\star(\pi))$ in advance, and the followers playing an approximate equilibrium $x^t(\pi^t)$ rather than $x^\star(\pi^t)$.

Algorithm 1 Stackelberg learning from preferences

Input: horizon T, inner-loop length K, kernels $\{k_j\}_{j=1}^m$, $\sigma > 0$, $\lambda > 0$, confidence level δ.
Set ρ; set $V^1 = \lambda I_m$; pick $\tilde{\theta}^1$ such that $\|\tilde{\theta}^1\|_2 \leq S$ and $\pi^1 \in \Pi$ arbitrarily.
Announce π^1; followers return $x^1 = \textsc{ApproxNE}(\pi^1, K)$; observe $\tilde{f}^{\,1} = \varphi(\pi^1, x^1)$.
for $t = 2, \ldots, T$ **do**
    Update the estimates $\tilde{\mu}_j^{\,t-1}, \sigma_j^{t-1}$ for $j \in [m]$ via equations 4 and 5 using $\{(\pi^\tau, \tilde{f}^{\,\tau})\}_{\tau=1}^{t-1}$.
    Select $\pi^t \in \arg\min_{\pi \in \Pi} \ell_t(\pi)$ with ℓ_t as in equation 7.
    Announce π^t; followers return $x^t = \textsc{ApproxNE}(\pi^t, K)$.
    Observe $\tilde{f}^{\,t} = \varphi(\pi^t, x^t)$ and preference $y^t \sim \mathrm{Bernoulli}\big(\mathrm{s}(\langle \theta^\star, \tilde{f}^{\,t-1} - \tilde{f}^{\,t} \rangle)\big)$.
    Set $\Delta^t = \tilde{f}^{\,t-1} - \tilde{f}^{\,t}$, $V^t = V^{t-1} + \Delta^t (\Delta^t)^\top$, and compute $\tilde{\theta}^t$ via equation 6.
end for

For the previously introduced Stackelberg game, we aim to design a *no-regret* leader policy satisfying $R^T/T \to 0$ as $T \to \infty$. As will be shown in Section 3.1, this property further translates into a computable rule to extract a single policy-response pair from the interaction history that is an ϵ-Stackelberg equilibrium.

3 Learning Stackelberg equilibria

Attaining no-regret and, in turn, learning Stackelberg equilibria requires estimating both components of the unknown scalar cost $J(\pi) = \langle \theta^\star, f(\pi) \rangle$: the equilibrium feature maps f_j from numerical observations and the trade-off parameter $\theta^\star$ from pairwise preferences. The two estimation problems are coupled, since both estimates and their uncertainty jointly affect how the leader selects its next action. We propose Algorithm 1 that estimates f_j and $\theta^\star$ via kernel ridge regression and logistic maximum likelihood, respectively, and based on both constructs a surrogate cost function used to select the next leader action.

For each $j \in [m]$, we decompose $\tilde{f}_j^\tau = f_j(\pi^\tau) + \varepsilon_j^\tau$, where ε_j^τ is the error induced by the followers responding with $x^\tau(\pi^\tau)$ instead of $x^\star(\pi^\tau)$. Hence, each numerical observation can be viewed as a corrupted evaluation of the RKHS-regular f_j. For $\mathcal{D}_j^t = \{(\pi^\tau, \tilde{f}_j^\tau)\}_{\tau \leq t}$, let $k_j^t(\pi) = (k_j(\pi^\tau, \pi))_{\tau=1}^t$, $K_j^t = (k_j(\pi^\tau, \pi^{\tau'}))_{\tau,\tau'=1}^t$, and $\tilde{F}_j^t = (\tilde{f}_j^\tau)_{\tau=1}^t$. We construct the kernel ridge regression estimate $\tilde{\mu}_j^{\,t}$ of f_j, and the uncertainty $\sigma_j^t(\pi)$ as:

$$\tilde{\mu}_j^{\,t}(\pi) = k_j^t(\pi)^\top \big(K_j^t + \sigma^2 I_t\big)^{-1} \tilde{F}_j^t, \tag{4}$$

$$\big(\sigma_j^t\big)^2(\pi) = k_j(\pi, \pi) - k_j^t(\pi)^\top \big(K_j^t + \sigma^2 I_t\big)^{-1} k_j^t(\pi), \tag{5}$$

where for all $j \in [m]$, we jointly set $\tilde{\mu}^{\,t}(\pi) = (\tilde{\mu}_j^{\,t}(\pi))_{j=1}^m$ and $\|\sigma^t(\pi)\|_2^2 = \sum_{j=1}^m (\sigma_j^t)^2(\pi)$.

Furthermore, given data $\mathcal{D}_{\mathrm{pr}}^t = \{(\Delta^\tau, y^\tau)\}_{\tau=2}^t$, we compute the *logistic maximum likelihood estimate* of $\theta^\star$:

$$\tilde{\theta}^t \in \underset{\|\theta\|_2 \leq S}{\arg\min}\ \mathcal{L}_t(\theta), \ \text{ with } \ \mathcal{L}_t(\theta) = \sum_{(\Delta^\tau, y^\tau) \in \mathcal{D}_{\mathrm{pr}}^t} h\big(\langle \theta, \Delta^\tau \rangle, y^\tau\big), \tag{6}$$

with logistic loss $h(z, y) = -(y \log \mathrm{s}(z) + (1 - y) \log\big(1 - \mathrm{s}(z)\big))$ (Shalev-Shwartz & Ben-David, 2014). As a convex problem with norm constraints, (6) can be solved efficiently via Sequential Least Squares Programming (SLSQP) (Kraft, 1988) or projected gradient methods (Blondel et al., 2022).

Since the available preference feedback compares consecutive realized outcomes, it provides information about relative rather than the absolute scalar leader cost. This motivates an optimistic action-selection rule based on the relative equilibrium cost $J(\pi) - J(\pi^{t-1})$ underlying the next preference observation.

For fixed $\delta \in (0,1)$ and $\lambda > 0$, the leader combines $\tilde{\theta}^{t-1}$ and $\tilde{\mu}^{t-1}$ to construct the surrogate

$$\ell_t(\pi) \;=\; \langle \tilde{\theta}^{t-1}, \tilde{\mu}^{t-1}(\pi)\rangle - \beta_{t-1}(\delta/2)\,\big\|\tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})\big\|_{V_{t-1}^{-1}} - S q_{t-1}(\pi), \tag{7}$$

where

$$\rho = B + L_\varphi\sqrt{2TCK^{-c}}/(\sigma\sqrt{\delta}),\;\; q_t(\pi) = \rho\|\sigma^t(\pi)\|_2,\;\; V_t = \lambda I_m + \sum_{\tau=2}^{t}\Delta^\tau(\Delta^\tau)^\top \tag{8}$$

and $\beta_t(\delta)^2 = \kappa\left[\log\left(\frac{1}{\delta}\right) + m\log\left(\max\left\{e, \frac{4eSL(t-1)}{m}\right\}\right)\right] + 4\lambda S^2$ with $\kappa = \sup_{\|\theta\|_2\le S,\ \|u\|_2\le 2L} 1/\mathrm{s}'(\langle\theta,u\rangle)$.

The second and third terms in (7) capture uncertainty and lower-level inexactness in estimating θ^* and f, respectively. Proposition C.1 in Appendix C establishes also that $\ell_t(\pi) - c_t \le J(\pi) - J(\pi^{t-1})$ holds with probability at least $1-\delta$ for $c_t = \langle\tilde{\theta}^{t-1}, \tilde{\mu}^{t-1}(\pi^{t-1})\rangle + S q_{t-1}(\pi^{t-1})$. Thus, $\ell_t(\pi) - c_t$ is a lower confidence bound on the relative equilibrium cost. Since c_t is independent of π, minimizing ℓ_t is equivalent to minimizing this lower bound, and the leader can select $\pi^t \in \arg\min_{\pi\in\Pi}\ell_t(\pi)$.

3.1 Regret analysis

The regret analysis relies on three lemmas, each bounding one source of errors. First, Lemma 3.1 converts the mean-square guarantee of Assumption 2.5 into a single high-probability event that controls all lower-level approximation errors $e_t = \|x^\star(\pi^t) - x^t(\pi^t)\|_2$. Second, to deal with these lower-level errors corrupting the numerical observations used to learn f_j, we leverage the Lipschitz property of φ in Lemma 3.2 to translate them into a confidence bound for the learned equilibrium feature map. Third, learning the unknown trade-off θ^* from observed pairs (Δ^t, y^t) obeys a well-specified logistic model in the design vector Δ^t, which we exploit in Lemma 3.3 to characterize the uncertainty of $\theta^\star$ through the time-uniform confidence bound based on (Schlaginhaufen et al., 2025, Lemma 3.1) and (Lee et al., 2025, Theorem 3.1)

Lemma 3.1 (Aggregate approximation error)**.** *Let $e_t = \|x^\star(\pi^t) - x^t(\pi^t)\|_2$ and define event $\mathcal{E}_1 = \big\{\sum_{t=1}^T e_t^2 \le 2TCK^{-c}/\delta\big\}$. Then $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta/2$. It further holds that $\sum_{t=1}^T e_t \le T\sqrt{2CK^{-c}/\delta}$.*

Lemma 3.2 (Kernel ridge confidence)**.** *Fix $j\in[m]$, $t\ge 1$, and define $\varepsilon_j^t = (\tilde{f}_j^\tau - f_j(\pi^\tau))_{\tau=1}^t$. Under Assumption 2.8, $\big|f_j(\pi) - \tilde{\mu}_j^t(\pi)\big| \le \big(B_j + \frac{1}{\sigma}\|\varepsilon_j^t\|_2\big)\,\sigma_j^t(\pi)$ holds for all $\pi\in\Pi$. Consequently, under Assumption 2.7, on the event $\mathcal{E}_1$ of Lemma 3.1, $\|f(\pi) - \tilde{\mu}^t(\pi)\|_2 \le q_t(\pi)$ holds for all $t\le T$ and $\pi\in\Pi$.*

Lemma 3.3 (Preference confidence)**.** *Let $\mathcal{E}_2$ be the event that $\|\tilde{\theta}^t - \theta^\star\|_{V_t} \le \beta_t(\delta/2)$ holds for all $t\in[T]$ simultaneously. Under Assumptions 2.6 and 2.9 it holds that $\mathbb{P}(\mathcal{E}_2)\ge 1-\delta/2$.*

We now state our high-probability bound on R^T defined in (3). Let $\gamma_{T,j} = \max_{A\subset\Pi,|A|=T}\frac{1}{2}\log\det(I_T + \sigma^{-2}K_j^A)$ be the maximum information gain for f_j and set $\gamma_T = \max_{j\in[m]}\gamma_{T,j}$ and $C_\sigma = 2/\log(1+\sigma^{-2})$.

Theorem 3.4 (Regret bound)**.** *Let Assumptions 2.2, 2.3, 2.5 and 2.6-2.9 hold, let $\delta\in(0,1)$, $\lambda>0$, $Q = 12mL^2/(\lambda\log 2)\cdot\log\big(1+4L^2/(\lambda\log 2)\big)$. Then, with probability at least $1-\delta$, Algorithm 1 attains*

$$\begin{aligned} R^T \;\le\; \underbrace{2SB_f}_{\textit{initialization}} \;+\; \underbrace{2\beta_T(\delta/2)\sqrt{(T-1)\Big(2m\log\Big(1+\tfrac{4(T-1)L^2}{m\lambda}\Big)+Q\Big)}}_{\textit{(I) unknown trade-off } \theta^\star} \\ +\; \underbrace{4\Big(S+\tfrac{\beta_T(\delta/2)}{\sqrt{\lambda}}\Big)\rho\sqrt{C_\sigma\, m\,\gamma_T\, T}}_{\textit{(II) unknown feature map } f} \;+\; \underbrace{\Big(S+\tfrac{4\beta_T(\delta/2)}{\sqrt{\lambda}}\Big)L_\varphi\, T\sqrt{\tfrac{2CK^{-c}}{\delta}}}_{\textit{(III) inexact lower-level equilibrium}}. \end{aligned} \tag{9}$$

The bound in (9) separates the effects of learning the unknown trade-off (I), estimating the equilibrium feature map from corrupted observations (II), and interacting with inexact follower responses (III). Lower-level inexactness enters through ρ in (II) and directly through (III), showing that follower accuracy must

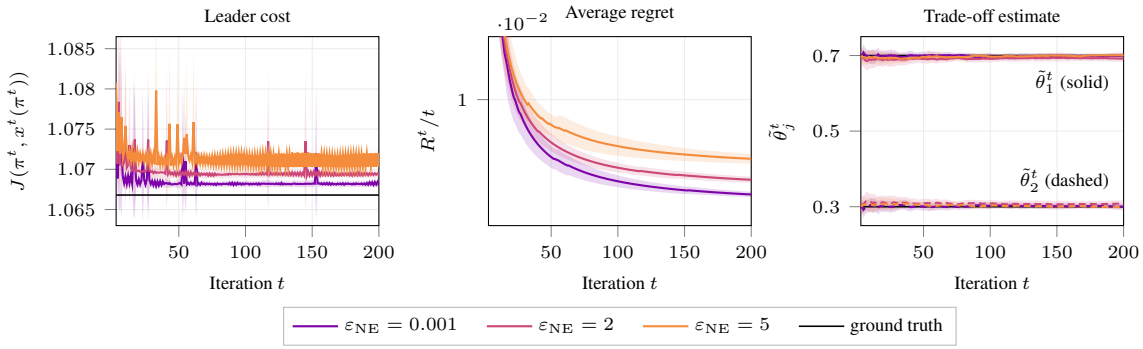

Figure 2: **Learning a Stackelberg equilibrium.** Each curve summarizes 10 runs of Algorithm 1 over $T = 200$ rounds, differing only in the equilibrium accuracy $\varepsilon_{\mathrm{NE}}$. Curves show the means across seeds and shaded bands 95% confidence intervals for the mean.

increase with the learning horizon for the average regret to vanish. This is made explicit in Corollary D.1 of Appendix D, where we show that choosing $K = \lceil T^{1/c} \rceil$ yields $R^T = \widetilde{O}\big(m\sqrt{\kappa \gamma_T T}\big)$ and hence $R^T/T \to 0$ whenever $\gamma_T = o(T)$, as for standard linear and squared-exponential kernels.

No regret alone, however, does not directly identify a near-optimal policy-response pair from the interaction history. To this end, given prior bounds on the structural constants entering the confidence sets, we construct a computable confidence-based certificate for each iterate and select the one with the smallest certificate. Under Lipschitz follower utilities, this yields the following finite-time guarantee.

Corollary 3.5 (Convergence to an ϵ-Stackelberg equilibrium)**.** *Assume, in addition to the assumptions of Theorem 3.4, that each follower utility $U_i(\cdot;\pi)$ is L_{U_i}-Lipschitz in $x \in \mathcal{X}$ uniformly over $\pi \in \Pi$, and set $L_U = \max_{i \in \mathcal{N}} L_{U_i}$. Define*

$$t^\star \in \operatorname*{arg\,min}_{2 \le t \le T} 2\beta_{t-1}(\delta/2)\, \big\|\tilde{\mu}^{t-1}(\pi^t) - \tilde{\mu}^{t-1}(\pi^{t-1})\big\|_{V_{t-1}^{-1}} + 2S\big(q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})\big). \tag{10}$$

Fix $\epsilon \in (0,1)$ and let $L_{\mathrm{SE}} = \max\{L_U, SL_\varphi\}$. It then suffices to choose $T = \tilde{\Theta}(\kappa m^2 \gamma_T/\epsilon^2)$ and $K = \Theta((L_{\mathrm{SE}}^2 CT/\delta\epsilon^2)^{1/c})$, so that, with probability at least $1-\delta$, $(\pi^{t^\star}, x^{t^\star}(\pi^{t^\star}))$ is an ϵ-Stackelberg equilibrium.

Corollary 3.5 extends finite-time equilibrium guarantees from scalar-feedback settings to the case where the leader receives pairwise preferences over multi-objective outcomes. The key difficulty is that the scalar leader cost is never observed, so historical actions cannot be ranked directly by their realized costs as in (Maddux et al., 2025; Maheshwari et al., 2024). Instead, (10) assigns each iterate a confidence-based certificate that jointly accounts for uncertainty in the scalarization and equilibrium feature map, enabling identification of an ϵ-Stackelberg policy-response pair from the interaction history.

4 Numerical Study: Carbon-Tax Incentive Design

We evaluate Algorithm 1 on a carbon-tax incentive-design problem in which a government sets a domestic carbon tax τ, a foreign carbon tax b, and an abatement subsidy s. The followers are three domestic and two foreign firms, competing in a common Cournot market by choosing a production quantity q_i and abatement level a_i. The firms' utilities U_i account for market revenue, production and abatement costs, emissions charges, and, for domestic firms, the abatement subsidy. The government evaluates the resulting market

outcome through two measurable objectives: aggregate net emissions φ_1 and an economic-burden index that captures market prices, subsidy expenditure, and the cost of aggressive policy interventions φ_2. A complete specification, together with the structural properties of the resulting game, is provided in Appendix A, with the numerical calibration reported in Appendix E.

At round t, the government announces $\pi^t = (\tau^t, b^t, s^t)$, and the firms approximately solve the induced Nash equilibrium. The government then observes the emissions and economic-burden objectives and receives a binary preference against the previous realization, while both $\theta^\star$ and the corresponding scalar cost remain hidden. For evaluation, we compute a full-information Stackelberg benchmark offline using an extensive grid search over the policy domain $\Pi = [0, 30] \times [0, 40] \times [0, 20]$ with $\theta^\star = (0.7, 0.3)$. This yields $\pi^\star = (7.92, 22.96, 1.85)$ with $J(\pi^\star) = 1.067$[1]. We then run Algorithm 1 for $T = 200$ rounds and ten random seeds, under three fixed equilibrium-accuracy levels $\varepsilon_{\mathrm{NE}} \in \{0.001, 2, 5\}$ satisfying $\|x^t(\pi^t) - x^\star(\pi^t)\|_2 \leq \varepsilon_{\mathrm{NE}}$, as illustrated in Figure 4. Unlike the increasing inner-loop effort prescribed by Corollary D.1, keeping $\varepsilon_{\mathrm{NE}}$ fixed deliberately introduces persistent follower-response error and allows us to isolate its effect on learning. Moreover, in order to stabilize the early estimates of $\theta^\star$, the preference estimator is warm-started with ten random policy pairs.

Figure 2 shows that the average regret decreases in all three configurations, reaching 2.5×10^{-3} under nearly exact follower responses and 5.3×10^{-3} under the largest equilibrium error. Averaged over the last 50 rounds, the realized leader cost settled at 1.0683, 1.0694 and 1.0711 for increasing levels of equilibrium error ϵ_{NE} (compared to $J(\pi^\star) = 1.067$). These values indicate that the leader cost remains close to the benchmark, although less accurate follower responses lead to a modest degradation in policy quality. In contrast, the estimated trade-off is considerably less sensitive to lower-level inexactness. Across all accuracy levels, the mean environmental weight approaches its true value 0.7, with nearly overlapping trajectories. This is consistent with Assumption 2.9, under which preferences are generated directly from the same realized objective vectors used to estimate the trade-off. Thus, lower-level inexactness primarily affects policy quality rather than $\theta^\star$ estimation. Appendix E provides additional results and implementation details.

5 Conclusion

We studied Stackelberg learning when the leader's scalar cost trades off multiple objectives, while neither this trade-off nor the followers' game are known. Instead, we leveraged that through repeated interactions the leader observes numerical performance objectives at approximate follower equilibria together with binary preferences over realized outcomes. By jointly learning the equilibrium objective maps and their unknown scalarization, the proposed approach enables the leader to optimize its scalar cost and learn a Stackelberg equilibrium without observing the cost directly. Our analysis quantifies the distinct effects of preference uncertainty, objective-map estimation, and lower-level inexactness, establishes conditions under which the average regret vanishes, and provides a computable criterion for extracting an ε-Stackelberg policy-response pair from the interaction history. The carbon-tax study illustrates these results, showing that the leader can learn a policy close to the full-information benchmark despite inexact follower responses.

A central limitation of our framework is that it assumes the relevant aspects of an outcome can be represented by a vector of numerically measurable objectives, with preferences determined by their scalarization. In some applications, however, an evaluator may be able to compare two outcomes without being able to express all relevant considerations through such measurable objectives. In this case, the current preference model would be misspecified and our theoretical guarantees would no longer apply directly. Extending the framework to preference models that can capture such richer notions of desirable societal outcome, as well as relaxing the structural assumptions on the lower-level game, provides an important direction for future work.

[1] Existing preferential and follower-agnostic Stackelberg methods operate under different feedback structures, making direct comparison nontrivial without additional adaptations (see Appendix E.5).

AI USE STATEMENT

Large language models (LLMs) were used exclusively to improve the clarity and style of the writing, under careful supervision by the authors. No scientific content, experimental design, data analysis, or intellectual contributions were generated by LLMs, and their use did not influence any reported results.

REPRODUCIBILITY STATEMENT

We provide the assumptions, learning protocol, and complete specification of the proposed algorithm in the main text. All theoretical results are accompanied by detailed proofs in the appendices. For the numerical study, the complete carbon-tax game specification and its theoretical properties are provided in Appendix A, while Appendix E reports the implementation details, numerical parameters used in the experiments, and other additional results. The source code used to generate the numerical results will be made publicly available upon publication.

REFERENCES

Raul Astudillo and Peter Frazier. Multi-attribute bayesian optimization with interactive preference learning. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4496–4507. PMLR, 26–28 Aug 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in Neural Information Processing Systems*, 35:5230–5242, 2022.

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. *Advances in neural information processing systems*, 2021.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 2018.

E. Brochu, M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-23, UBC, 2009.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 844–853. JMLR.org, 2017.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 137–144, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102369.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition, 2014. doi: 10.1007/978-1-4939-1037-3.

Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587, Paris, France, 03–06 Jul 2015. PMLR.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer New York, NY, 2003. ISBN 978-0-387-21814-4. doi: 10.1007/b97543.

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3133–3144. PMLR, 13–18 Jul 2020.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence. Preferential Bayesian optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1282–1291. PMLR, 06–11 Aug 2017.

Panagiotis D. Grontas, Giuseppe Belgioioso, Carlo Cenedese, Marta Fochesato, John Lygeros, and Florian Dörfler. Big hype: Best intervention in games via distributed hypergradient descent. *IEEE Transactions on Automatic Control*, 69(12):8338–8353, 2024. doi: 10.1109/TAC.2024.3410890.

Zhiyuan Jerry Lin, Raul Astudillo, Peter Frazier, and Eytan Bakshy. Preference exploration for efficient bayesian optimization with multiple outcomes. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4235–4258. PMLR, 28–30 Mar 2022.

F P Kelly, A K Maulloo, and D K H Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, 1998. doi: 10.1057/palgrave.jors.2600523.

Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits, 2025.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Anna Maddux, Marko Maljkovic, Nikolas Geroliminis, and Maryam Kamgarpour. No-regret learning in stackelberg games with an application to electric ride-hailing. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 4009–4014, 2025. doi: 10.1109/CDC57313.2025.11312071.

Chinmay Maheshwari, James Cheng, Shankar Sastry, Lillian Ratliff, and Eric Mazumdar. Follower agnostic learning in stackelberg games. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 222–228, 2024. doi: 10.1109/CDC56724.2024.10885984.

Marko Maljkovic, Gustav Nilsson, and Nikolas Geroliminis. Learning how to price charging in electric ride-hailing markets. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 3136–3141, 2023. doi: 10.1109/CDC49753.2023.10383750.

Gilbert E. Metcalf and David A. Weisbach. The design of a carbon tax. *Harvard Environmental Law Review*, 33(2):499–559, 2009.

Barna Pásztor, Parnian Kassraie, and Andreas Krause. Bandits with preference feedback: a stackelberg game perspective. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Vern I. Paulsen. An introduction to the theory of reproducing kernel hilbert spaces. Lecture notes, 2009. Available online.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, 1965.

Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Watch and learn: optimizing from revealed preferences feedback. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pp. 949–962, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325. doi: 10.1145/2897518.2897579.

Andreas Schlaginhaufen, Reda Ouhamma, and Maryam Kamgarpour. Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *arXiv preprint arXiv:2506.09508*, 2025.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *J. Mach. Learn. Res.*, 12(null):2389–2410, July 2011. ISSN 1532-4435.

Heinrich Von Stackelberg. *The theory of the market economy*. William Hodge, 1952.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.028. JCSS Special Issue: Cloud Computing 2011.

Mengjia Zhu, Alberto Bemporad, and Dario Piga. Preference-based mpc calibration. In *2021 European Control Conference (ECC)*, pp. 638–645, 2021. doi: 10.23919/ECC54610.2021.9654900.

A CARBON-TAX INCENTIVE DESIGN: GAME SPECIFICATION AND PROPERTIES

To ground the proposed framework in a concrete setting, we consider a stylized carbon-tax design problem inspired by (Metcalf & Weisbach, 2009) and illustrated in Figure 3. A government regulator acts as the Stackelberg leader, while a set of domestic firms, $\mathcal{D} = \{D_1, D_2, D_3\}$, and foreign firms, $\mathcal{F} = \{F_1, F_2\}$, compete in a common domestic Cournot market. The government sets

$$\pi = (\tau, b, s) \in \Pi = [0, \bar{\tau}] \times [0, \bar{b}] \times [0, \bar{s}],$$

where τ is a domestic carbon tax, b is a carbon charge on foreign producers, and s is an abatement subsidy for domestic firms.

After observing π, each firm $i \in \mathcal{N} = \mathcal{D} \cup \mathcal{F}$ chooses a product output quantity q_i and an abatement level a_i, collected in $x_i = (q_i, a_i)$, subject to

$$\mathcal{X}_i = \left\{(q_i, a_i) \in \mathbb{R}^2_+ : 0 \le q_i \le \bar{q}_i,\ 0 \le a_i \le \nu_i q_i,\ \bar{q}_i, \nu_i > 0\right\},$$

where $\bar{q}_i > 0$ denotes the production capacity and $\nu_i > 0$ denotes the per-unit production emissions intensity. Aggregate output is $Q(x) = \sum_{i \in \mathcal{N}} q_i$, and the price is determined by consumers' inverse demand $P(Q(x)) = A - BQ(x)$ for some $A, B > 0$. The net emissions of firm i are $e_i(x_i) = \nu_i q_i - a_i$.

The government-set policy enters the firms' payoffs through the cost of net emissions and the reward for abatement. Define the emissions charge $t_i(\pi) = \tau \mathbb{1}\{i \in \mathcal{D}\} + b\mathbb{1}\{i \in \mathcal{F}\}$ and the subsidy indicator $u_i(\pi) = s\mathbb{1}\{i \in \mathcal{D}\}$. Moreover, assume that firm $i \in \mathcal{N}$ has constant marginal production cost $c_i > 0$ and quadratic abatement cost $\frac{\kappa_i}{2} a_i^2$ with $\kappa_i > 0$. Then firm i's utility is of the form

$$\begin{aligned} U_i(x_i, x_{-i}; \pi) &= (P(Q(x)) - c_i)q_i - \tfrac{\kappa_i}{2} a_i^2 - t_i(\pi) e_i(x_i) + u_i(\pi) a_i \\ &= \big(A - BQ(x) - c_i - t_i(\pi)\nu_i\big) q_i - \tfrac{\kappa_i}{2} a_i^2 + \big(t_i(\pi) + u_i(\pi)\big) a_i. \end{aligned}$$

Note that the emissions charge increases the cost of an additional unit of output by $t_i(\pi)\nu_i$, and that both the emissions charge and the abatement subsidy increase the marginal incentive to abate. For fixed competitors' actions, U_i is strictly concave in $x_i = (q_i, a_i)$, since

$$\frac{\partial^2 U_i}{\partial q_i^2} = -2B < 0, \qquad \frac{\partial^2 U_i}{\partial a_i^2} = -\kappa_i < 0, \qquad \frac{\partial^2 U_i}{\partial q_i \partial a_i} = 0.$$

Moreover, each $\mathcal{X}_i$ is nonempty, compact, and convex, so Assumption 2.2 is satisfied. The following proposition further establishes Assumption 2.3 and provides an explicit strong monotonicity constant.

Proposition A.1. *(Strong monotonicity of the carbon-tax incentive design) For every $\pi \in \Pi$, the mapping $v(\cdot; \pi)$ is α-strongly monotone on $\mathcal{X} = \prod_{i \in \mathcal{N}} \mathcal{X}_i$, with $\alpha = \min\{B, \min_{i \in \mathcal{N}} \kappa_i\} > 0$.*

Proof. The individual components of the pseudogradient $v(x; \pi)$ are given by

$$\begin{aligned} v_{q_i}(x; \pi) &= c_i + t_i(\pi)\nu_i - A + BQ(x) + Bq_i, \\ v_{a_i}(x; \pi) &= \kappa_i a_i - t_i(\pi) - u_i(\pi). \end{aligned}$$

Let $x = (q, a)$ and $y = (r, d)$ belong to $\mathcal{X}$, and set $\Delta q = q - r$ and $\Delta a = a - d$. The affine policy terms cancel in $v(x; \pi) - v(y; \pi)$, giving

$$\begin{aligned} \langle v(x; \pi) - v(y; \pi), x - y \rangle &= B\Big(\sum_{i \in \mathcal{N}} \Delta q_i\Big)^2 + B \sum_{i \in \mathcal{N}} (\Delta q_i)^2 + \sum_{i \in \mathcal{N}} \kappa_i (\Delta a_i)^2 \\ &\ge B\|\Delta q\|_2^2 + \big(\min_i \kappa_i\big)\|\Delta a\|_2^2 \\ &\ge \alpha \|x - y\|_2^2. \end{aligned}$$

Thus, $v(\cdot; \pi)$ is α-strongly monotone with $\alpha = \min\{B, \min_{i \in \mathcal{N}} \kappa_i\} > 0$. □

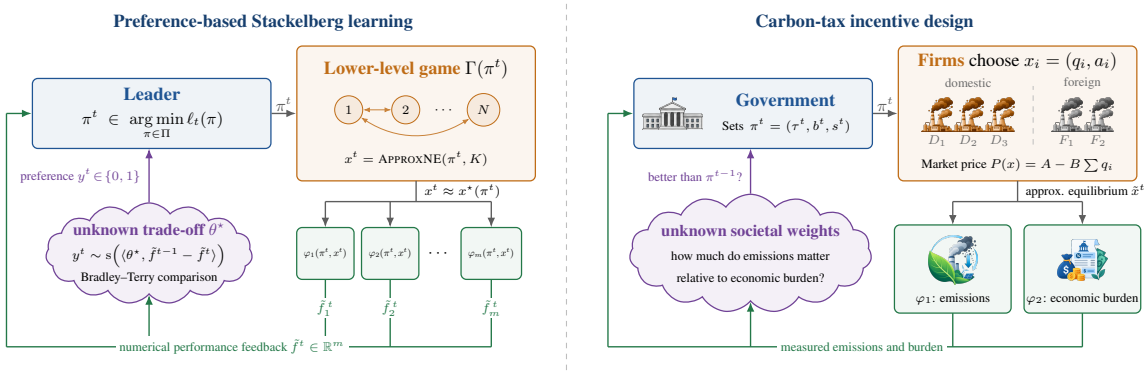

Figure 3: *Left:* proposed learning framework introduced in Section 2. *Right:* corresponding carbon-tax case study equivalent. The government selects a domestic carbon tax τ, a foreign carbon tax b, and an abatement subsidy s, while firms choose output q_i and abatement level a_i. The observed performance objectives are net emissions φ_1 and economic burden φ_2.

The government evaluates the induced market outcome through two measurable performance objectives:

$$\varphi_1(\pi, x) = \frac{1}{S_E}\sum_{i\in\mathcal{N}} e_i(x_i), \text{ and } \varphi_2(\pi, x) = \frac{P(Q(x))}{S_P} + \omega_G\frac{G(\pi, x)}{S_G} + \omega_b\left(\frac{b}{\bar{b}}\right)^2 + \omega_\tau\left(\frac{\tau}{\bar{\tau}}\right)^2 + \omega_s\left(\frac{s}{\bar{s}}\right)^2.$$

The first captures environmental performance through normalized aggregate net emissions, with $S_E > 0$ a fixed scaling constant. The second represents the economic burden index associated with the policy, combining the normalized market price, the subsidy-expenditure term $G(\pi, x) = s\sum_{i\in\mathcal{D}} a_i$, and convex penalties associated with aggressive carbon-pricing instruments. The precise scaling constants and coefficients used in the numerical study are reported in Appendix E.

The price term reflects the burden of reduced aggregate supply on consumers, the expenditure term prevents abatement subsidies from acting as a costless policy instrument, and the quadratic terms provide smooth reduced-form proxies for the increasing implementation, compliance, and distributional costs associated with aggressive carbon taxation and border adjustments. Importantly, all these coefficients are part of the known definitions of the leader's performance measures. The unknown quantity is instead the relative weighting $\theta^\star \in \mathbb{R}^2$ between environmental performance and economic burden. Thus, the regulator can measure φ_1 and φ_2 but does not directly observe the scalar value $J = \langle\theta^\star, \phi(\pi, x)\rangle$, consistent with the information structure introduced in Section 2.1.

B Justification of Assumption 2.8

Assumption 2.8 is a regularity condition on the *composition* of the performance measure map and the equilibrium map, and deserves a comment, since the natural assumption is regularity of φ on the joint space $\Pi \times \mathcal{X}$. The following pullback argument shows that the composition indeed lies in an RKHS over Π.

Proposition B.1 (Pullback). *Let $\phi : \Pi \to \Pi \times \mathcal{X}$ be the equilibrium embedding $\phi(\pi) = (\pi, x^\star(\pi))$. Let $\varphi_j \in \mathcal{H}_{\bar{k}}(\Pi \times \mathcal{X})$ for a positive semidefinite kernel $\bar{k}$ on $\Pi \times \mathcal{X}$, with $\|\varphi_j\|_{\bar{k}} \le \bar{B}_j$. Define the pullback kernel $k^\phi(\pi, \pi') = \bar{k}(\phi(\pi), \phi(\pi'))$. Then k^ϕ is a positive semidefinite kernel on Π, its RKHS is $\mathcal{H}_{k^\phi}(\Pi) = \{g \circ \phi : g \in \mathcal{H}_{\bar{k}}(\Pi \times \mathcal{X})\}$, and in particular $f_j \in \mathcal{H}_{k^\phi}(\Pi)$ with $\|f_j\|_{k^\phi} \le \bar{B}_j$.*

Proof. Under Assumptions 2.2 and 2.3, $\phi(\pi)$ is well defined as $\Gamma(\pi)$ admits a unique Nash equilibrium for every $\pi \in \Pi$. Since $\bar{k}$ is a kernel function on $\Pi \times \mathcal{X}$, and $\phi : \Pi \to \Pi \times \mathcal{X}$, based on (Paulsen, 2009, Proposition 5.13), $(\bar{k} \circ \phi)(\pi, \pi') = \bar{k}(\phi(\pi), \phi(\pi'))$ is a positive semidefinite kernel on Π. The description of $\mathcal{H}_{k^\phi}(\Pi)$ as the set of compositions $g \circ \phi$ follows from (Paulsen, 2009, Theorem 5.14), together with the fact that for every $u \in \mathcal{H}_{k^\phi}(\Pi)$, it holds that $\|u\|_{k^\phi} = \inf\{\|g\|_{\bar{k}} : u = g \circ \phi\}$. Since $\varphi_j \in \mathcal{H}_{\bar{k}}(\Pi \times \mathcal{X})$ by assumption, and $f_j = \varphi_j \circ \phi$ by construction, it is clear that $f_j \in \mathcal{H}_{k^\phi}(\Pi)$. Furthermore, by plugging $g = \varphi_j$, which is feasible, we immediately get $\|f_j\|_{k^\phi} = \inf\{\|g\|_{\bar{k}} : f_j = g \circ \phi\} \leq \|\varphi_j\|_{\bar{k}} \leq \bar{B}_j$. □

Proposition B.1 establishes that regularity of the performance measure map on the joint action space induces an RKHS structure on the equilibrium features. However, the resulting pullback kernel k^ϕ depends on the unknown equilibrium map and therefore provides an existence result rather than a directly implementable kernel for the leader. A complementary justification can be obtained from standard sensitivity properties of strongly monotone variational inequalities (Dontchev & Rockafellar, 2014).

Proposition B.2 (Lipschitz regularity of the equilibrium features)**.** *Suppose that in addition to Assumptions 2.3 and 2.7, there exist $L_{v,\pi}, L_{\varphi,\pi} > 0$ such that*

$$\|v(x;\pi) - v(x;\pi')\|_2 \leq L_{v,\pi}\|\pi - \pi'\|_2 \;\; \textit{and} \;\; \|\varphi(\pi, x) - \varphi(\pi', x)\|_2 \leq L_{\varphi,\pi}\|\pi - \pi'\|_2 \tag{11}$$

hold for all $x \in \mathcal{X}$ and $\pi, \pi' \in \Pi$. Then the equilibrium feature map $f(\pi)$ is Lipschitz continuous with

$$\|f(\pi) - f(\pi')\|_2 \leq \big(L_{\varphi,\pi} + \tfrac{L_\varphi L_{v,\pi}}{\alpha}\big)\|\pi - \pi'\|_2. \tag{12}$$

Proof. Let us fix $\pi, \pi' \in \Pi$ and write $x = x^\star(\pi)$, $x' = x^\star(\pi')$. Since Nash equilibria are solutions of the corresponding variational inequalities, we have

$$\langle v(x;\pi), x' - x\rangle \geq 0 \;\text{ and }\; \langle v(x';\pi'), x - x'\rangle \geq 0. \tag{13}$$

Adding the two inequalities gives $\langle v(x;\pi) - v(x';\pi'), x' - x\rangle \geq 0$, i.e. $\langle v(x;\pi) - v(x';\pi'), x - x'\rangle \leq 0$. Hence, by Assumption 2.3,

$$\begin{aligned}
\alpha\,\|x - x'\|^2 \;&\leq\; \langle v(x;\pi) - v(x';\pi), x - x'\rangle \\
&= \underbrace{\langle v(x;\pi) - v(x';\pi'), x - x'\rangle}_{\leq 0} + \langle v(x';\pi') - v(x';\pi), x - x'\rangle \\
&\leq \|v(x';\pi') - v(x';\pi)\|_2\|x - x'\|_2 \qquad (14)\\
&\leq L_{v,\pi}\|\pi - \pi'\|_2\|x - x'\|_2, \qquad (15)
\end{aligned}$$

where (14) was obtained using Cauchy–Schwarz inequality and (11). Hence, for $\pi \neq \pi'$,

$$\|x^\star(\pi) - x^\star(\pi')\|_2 \leq \tfrac{L_{v,\pi}}{\alpha}\|\pi - \pi'\|_2. \tag{16}$$

Finally, by combining (16) with the Lipschitz condition on $\varphi(\pi, x)$ in (11), we have

$$\begin{aligned}
\|f(\pi) - f(\pi')\|_2 &= \|\varphi(\pi, x^\star(\pi)) - \varphi(\pi', x^\star(\pi'))\|_2 \\
&\leq \|\varphi(\pi, x^\star(\pi)) - \varphi(\pi', x^\star(\pi))\|_2 + \|\varphi(\pi', x^\star(\pi)) - \varphi(\pi', x^\star(\pi'))\|_2 \\
&\leq L_{\varphi,\pi}\|\pi - \pi'\|_2 + L_\varphi\|x^\star(\pi) - x^\star(\pi')\|_2 \\
&\leq \big(L_{\varphi,\pi} + \tfrac{L_\varphi L_{v,\pi}}{\alpha}\big)\|\pi - \pi'\|_2,
\end{aligned}$$

which completes the proof. □

Since Lipschitz functions are continuous, Proposition B.2 implies that each f_j is continuous on Π, i.e., $f_j \in C(\Pi)$. This observation also motivates practical kernel choices beyond the implicit pullback kernel. For any universal kernel (Sriperumbudur et al., 2011) k on compact Π, its RKHS $\mathcal{H}_k$ is dense in $C(\Pi)$ under the uniform norm, i.e., for every $f_j \in C(\Pi)$ and $\varepsilon > 0$, there exists $g_j \in \mathcal{H}_k$ such that $\sup_{\pi\in\Pi}|f_j(\pi)-g_j(\pi)| \le \varepsilon$. This is precisely the definition of c-universality introduced in (Sriperumbudur et al., 2011), which also shows that, on compact domains, the notions of c-, cc-, and c_0-universality coincide. Since the Matérn class is c_0-universal on $\mathbb{R}^p$, Matérn kernels therefore provide a practical choice whose RKHS can uniformly approximate continuous equilibrium features to arbitrary accuracy.

C Auxiliary Results

Proposition C.1 (Exact per-round regret)**.** *On the joint event $\mathcal{E}_1 \cap \mathcal{E}_2$, each defined in Lemmas 3.1 and 3.3, respectively, for every $t \ge 2$, the instantaneous exact regret $r_e^t = J(\pi^t) - J(\pi^\star)$ satisfies*

$$r_e^t \;\le\; 2\beta_{t-1}(\delta/2)\;\big\|\tilde{\mu}^{t-1}(\pi^t) - \tilde{\mu}^{t-1}(\pi^{t-1})\big\|_{V_{t-1}^{-1}} + 2S\big(q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})\big). \tag{17}$$

Proof. We start by rewriting the instantaneous exact regret as

$$\begin{aligned} r_e^t = J(\pi^t) - J(\pi^\star) &= \langle\theta^\star, f(\pi^t) - f(\pi^{t-1})\rangle - \langle\theta^\star, f(\pi^\star) - f(\pi^{t-1})\rangle \\ &= g_t(\pi^t) - g_t(\pi^\star), \end{aligned} \tag{18}$$

for $g_t(\pi) = \langle\theta^\star, f(\pi) - f(\pi^{t-1})\rangle$. Moreover, let us define $\hat{g}_t(\pi) = \langle\tilde{\theta}_{t-1}, \tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})\rangle$. Then

$$\begin{aligned} g_t(\pi) - \hat{g}_t(\pi) = \langle\theta^\star - \tilde{\theta}_{t-1}, \tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})\rangle \\ + \langle\theta^\star, f(\pi) - \tilde{\mu}^{t-1}(\pi)\rangle + \langle\theta^\star, \tilde{\mu}^{t-1}(\pi^{t-1}) - f(\pi^{t-1})\rangle. \end{aligned}$$

Using the Cauchy–Schwarz inequality and $\|\theta^\star\|_2 \le S$, we can bound the difference $|g_t(\pi) - \hat{g}_t(\pi)|$ as

$$\begin{aligned} |g_t(\pi) - \hat{g}_t(\pi)| \le \|\theta^\star - \tilde{\theta}_{t-1}\|_{V_t-1}\|\tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})\|_{V_{t-1}^{-1}} \\ + S\|f(\pi) - \tilde{\mu}^{t-1}(\pi)\|_2 + S\|\tilde{\mu}^{t-1}(\pi^{t-1}) - f(\pi^{t-1})\|_2. \end{aligned}$$

Based on Lemma 3.2, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we directly have

$$|g_t(\pi) - \hat{g}_t(\pi)| \le \underbrace{\beta_{t-1}(\delta/2)\|\tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})\|_{V_{t-1}^{-1}} + S(q_{t-1}(\pi) + q_{t-1}(\pi^{t-1}))}_{=:\omega_t(\pi)}. \tag{19}$$

To bound (18), we first observe that (19) yields

$$\begin{aligned} g_t(\pi^t) &\le \hat{g}_t(\pi^t) + \omega_t(\pi^t) \\ g_t(\pi^\star) &\ge \hat{g}_t(\pi^\star) - \omega_t(\pi^\star). \end{aligned}$$

However, since $\pi^t \in \arg\min_{\pi\in\Pi} \ell_t(\pi)$, it also holds that $\pi^t \in \arg\min_{\pi\in\Pi} \hat{g}_t(\pi) - \omega_t(\pi)$ because

$$\hat{g}_t(\pi) - \omega_t(\pi) = \ell_t(\pi) - \langle\tilde{\theta}_{t-1}, \tilde{\mu}^{t-1}(\pi^{t-1})\rangle - Sq_{t-1}(\pi^{t-1}), \tag{20}$$

and the added terms are constant. Therefore, $\hat{g}_t(\pi^t) - \omega_t(\pi^t) \le \hat{g}_t(\pi^\star) - \omega_t(\pi^\star)$, and

$$\begin{aligned} r_e^t = g_t(\pi^t) - g_t(\pi^\star) &\le \hat{g}_t(\pi^t) - \hat{g}_t(\pi^\star) + \omega_t(\pi^t) + \omega_t(\pi^\star) \\ &\le \omega_t(\pi^t) - \omega_t(\pi^\star) + \omega_t(\pi^t) + \omega_t(\pi^\star) \\ &= 2\omega_t(\pi^t) \\ &= 2\beta_{t-1}(\delta/2)\;\big\|\tilde{\mu}^{t-1}(\pi^t) - \tilde{\mu}^{t-1}(\pi^{t-1})\big\|_{V_{t-1}^{-1}} + 2S\big(q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})\big). \end{aligned}$$

which completes the proof. □

Lemma C.2 (Elliptical potential). *For $\Delta^2, \dots, \Delta^T \in \mathbb{R}^m$ defined in Assumption 2.9, the following holds:*

$$\sum_{t=2}^{T} \left\|\Delta^t\right\|_{V_{t-1}^{-1}} \leq \sqrt{(T-1)\Big(2m \log\Big(1 + \tfrac{4(T-1)L^2}{m\lambda}\Big) + \tfrac{12mL^2}{\lambda \log 2} \log\Big(1 + \tfrac{4L^2}{\lambda \log 2}\Big)\Big)}. \tag{21}$$

Proof. Using the arithmetic mean–quadratic mean (AM–QM) inequality, we have

$$\Big(\sum_{t=2}^{T} \|\Delta^t\|_{V_{t-1}^{-1}}\Big)^2 \leq (T-1) \sum_{t=2}^{T} \|\Delta^t\|^2_{V_{t-1}^{-1}}. \tag{22}$$

Now by invoking (Schlaginhaufen et al., 2025, Lemma D.4) with $\|\Delta^t\|_2 \leq 2L$, we get

$$\sum_{t=2}^{T} \|\Delta^t\|^2_{V_{t-1}^{-1}} \leq 2m \log\Big(1 + \tfrac{4(T-1)L^2}{m\lambda}\Big) + \tfrac{12mL^2}{\lambda \log 2} \log\Big(1 + \tfrac{4L^2}{\lambda \log 2}\Big)$$

which completes the proof. □

Lemma C.3 (Information gain). *If $k_j(\pi, \pi) \leq 1$ and $C_\sigma = 2/\log(1 + \sigma^{-2})$, then for any $k < T$ and any sequence $\{\pi^t\}_{t=k}^T$,*

$$\sum_{t=k}^{T} \left\|\sigma^{t-1}(\pi^t)\right\|_2 \leq \sqrt{C_\sigma \, m \, \gamma_T \, (T-k+1)}. \tag{23}$$

Proof. Based on (Srinivas et al., 2010, Lemma 5.3), for each $j \in [m]$, the maximum information gain after observing T noisy samples can be expressed in terms of predictive GP variances as

$$\gamma_{T,j} = \max_{\{\pi^1, \cdots, \pi^T\} \subset \Pi} \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2}(\sigma_j^{t-1})^2(\pi^t)). \tag{24}$$

Using the (AM)–(QM) inequality, we get

$$\Big(\sum_{t=k}^{T} \left\|\sigma^{t-1}(\pi^t)\right\|_2\Big)^2 \leq (T-k+1) \sum_{t=k}^{T} \sum_{j \in [m]} (\sigma_j^{t-1})^2(\pi^t) \leq (T-k+1) \sum_{j \in [m]} \sum_{t=1}^{T} (\sigma_j^{t-1})^2(\pi^t). \tag{25}$$

Now let us define $C_1 = \sigma^{-2}/\log(1 + \sigma^{-2})$ and observe that $\sigma^{-2}(\sigma_j^{t-1})^2(\pi^t) \leq \sigma^{-2} k_j(\pi^t, \pi^t) \leq \sigma^{-2}$. Then, because $s^2 \leq C_1 \log(1 + s^2)$ for $s \in [0, \sigma^{-2}]$ (Srinivas et al., 2010, Lemma 5.4), we can write

$$\sum_{t=1}^{T} (\sigma_j^{t-1})^2(\pi^t) = \sigma^2 \sum_{t=1}^{T} \sigma^{-2}(\sigma_j^{t-1})^2(\pi^t) \leq \sigma^2 \sum_{t=1}^{T} C_1 \log(1 + \sigma^{-2}(\sigma_j^{t-1})^2(\pi^t)) \leq 2\gamma_{T,j}\sigma^2 C_1.$$

Finally, we have

$$\sum_{t=k}^{T} \left\|\sigma^{t-1}(\pi^t)\right\|_2 \leq \sqrt{(T-k+1) \sum_{j \in [m]} 2\sigma^2 C_1 \gamma_{T,j}} \leq \sqrt{C_\sigma m \gamma_T (T-k+1)} \tag{26}$$

which completes the proof. □

D Regret Analysis: Proofs

D.1 Proof of Lemma 3.1

Proof. Combining Markov's inequality and Assumption (2.5), we immediately have

$$\mathbb{P}\left(\sum_{t=1}^{T} e_t^2 \geq \frac{2TCK^{-c}}{\delta}\right) \leq \frac{\delta \mathbb{E}\left[\sum_{t=1}^{T} e_t^2\right]}{2TCK^{-c}} = \frac{\delta \sum_{t=1}^{T} \mathbb{E}[e_t^2]}{2TCK^{-c}} \leq \frac{\delta TCK^{-c}}{2TCK^{-c}} = \frac{\delta}{2}. \tag{27}$$

Therefore, $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta/2$. On $\mathcal{E}_1$, using the (AM)–(QM) inequality we get

$$\sum_{t=1}^{T} e_t \leq T\sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \leq T\sqrt{2CK^{-c}/\delta}\,,$$

which completes the proof. □

D.2 Proof of Lemma 3.2

Proof. To ease the notation, let $A = K_j^t + \sigma^2 I_t$, $k = k_j^t(\pi)$, $F = (f_j(\pi^\tau))_{\tau=1}^t$ and $\varepsilon = \varepsilon_j^t = \tilde{F}_j^t - F$. Moreover, let $\mu_j^t(\pi) = k^\top A^{-1} F$ be the (unavailable) posterior mean based on uncorrupted targets and $\tilde{\mu}_j^t(\pi) = k^\top A^{-1}\tilde{F}_j^t$. Then we can write

$$\begin{aligned} |f_j(\pi) - \tilde{\mu}_j^t(\pi)| &= |f_j(\pi) - \mu_j^t(\pi) + \mu_j^t(\pi) - \tilde{\mu}_j^t(\pi)| \\ &= |(f_j(\pi) - \mu_j^t(\pi)) - k^\top A^{-1}(\tilde{F}_j^t - F)| \\ &= |(f_j(\pi) - \mu_j^t(\pi)) - k^\top A^{-1}\varepsilon| \\ &\leq |f_j(\pi) - \mu_j^t(\pi)| + |k^\top A^{-1}\varepsilon|, \end{aligned} \tag{28}$$

and proceed to bound the two pieces separately.

To bound $|f_j(\pi) - \mu_j^t(\pi)|$, define $\alpha = A^{-1}k \in \mathbb{R}^t$, recall that $\forall \pi \in \Pi$, $k_j(\pi, \cdot) \in \mathcal{H}_{k_j}(\Pi)$, and rewrite

$$f_j(\pi) - \mu_j^t(\pi) = f_j(\pi) - k^\top A^{-1} F = f_j(\pi) - \alpha^\top F = f_j(\pi) - \sum_{\tau \leq t} \alpha_\tau f_j(\pi^\tau).$$

Then, by the reproducing property, the following holds:

$$f_j(\pi) - \mu_j^t(\pi) = \langle f_j, k_j(\pi, \cdot)\rangle_{k_j} - \sum_{\tau \leq t} \alpha_\tau \langle f_j, k_j(\pi^\tau, \cdot)\rangle_{k_j} = \langle f_j, k_j(\pi, \cdot) - \sum_{\tau \leq t} \alpha_\tau k_j(\pi^\tau, \cdot)\rangle_{k_j}.$$

Hence, if $g = k_j(\pi, \cdot) - \sum_{\tau=1}^t \alpha_\tau k_j(\pi^\tau, \cdot)$, we can write $|f_j(\pi) - \mu_j^t(\pi)| = |\langle f_j, g\rangle_{k_j}|$, so Cauchy–Schwarz inequality gives $|f_j(\pi) - \mu_j^t(\pi)| \leq \|f_j\|_{k_j}\|g\|_{k_j} \leq B_j\|g\|_{k_j}$. We have $\|g\|_{k_j}^2 = \langle g, g\rangle_{k_j}$ and hence

$$\|g\|_{k_j}^2 = \langle k_j(\pi, \cdot), k_j(\pi, \cdot)\rangle_{k_j} - 2\sum_{\tau \leq t} \alpha_\tau \langle k_j(\pi^\tau, \cdot), k_j(\pi, \cdot)\rangle_{k_j} + \sum_{\tau \leq t}\sum_{\tau' \leq t} \alpha_\tau \alpha_{\tau'} \langle k_j(\pi^\tau, \cdot), k_j(\pi^{\tau'}, \cdot)\rangle_{k_j}.$$

Therefore, using the reproducing property and $K_j^t = A - \sigma^2 I_t$, we can write

$$\begin{aligned} \|g\|_{k_j}^2 &= k_j(\pi, \pi) - 2\alpha^\top k + \alpha^\top K_j^t \alpha = k_j(\pi, \pi) - 2k^\top A^{-1}k + k^\top A^{-1}\big(A - \sigma^2 I\big)A^{-1}k \\ &= k_j(\pi, \pi) - k^\top A^{-1}k - \sigma^2 \left\|A^{-1}k\right\|_2^2 = \left(\sigma_j^t\right)^2(\pi) - \sigma^2 \left\|A^{-1}k\right\|_2^2. \end{aligned} \tag{29}$$

It is clear now that $\|g\|_{k_j}^2 \leq \left(\sigma_j^t\right)^2(\pi)$, giving $|f_j(\pi) - \mu_j^t(\pi)| \leq B_j \sigma_j^t(\pi)$.

To bound $|k^\top A^{-1}\varepsilon|$, we start by applying Cauchy–Schwarz inequality, i.e., $|k^\top A^{-1}\varepsilon| \leq \|A^{-1}k\|_2 \|\varepsilon\|_2$. Because $\|g\|_{k_j}^2 \geq 0$, from equation (29) we have $\|A^{-1}k\|_2 \leq \frac{1}{\sigma}\sigma_j^t(\pi)$, resulting in

$$|f_j(\pi) - \tilde{\mu}_j^t(\pi)| \leq B_j \sigma_j^t(\pi) + \tfrac{1}{\sigma}\sigma_j^t(\pi)\|\varepsilon\|_2 = \big(B_j + \tfrac{1}{\sigma}\|\varepsilon\|_2\big)\sigma_j^t(\pi).$$

which completes the first part of the proof.

For the second part, due to Assumption 2.5, on event $\mathcal{E}_1$, the following holds

$$\|\varepsilon_j^t\|_2^2 = \sum_{\tau \leq t} |\tilde{f}_j^\tau - f_j(\pi^\tau)|^2 \leq \sum_{\tau \leq t} \|\tilde{f}^\tau - f(\pi^\tau)\|_2^2 \leq L_\varphi^2 \sum_{\tau \leq t} e_\tau^2 \leq L_\varphi^2 \frac{2TCK^{-c}}{\delta}.$$

Therefore, $B_j + \frac{1}{\sigma}\|\varepsilon_j^t\|_2 \leq \max_j B_j + L_\varphi \sqrt{2TCK^{-c}}/(\sigma\sqrt{\delta}) = \rho$, which yields

$$|f_j(\pi) - \tilde{\mu}_j^t(\pi)| \leq \rho \sigma_j^t(\pi).$$

Finally, by squaring and summing over all $j \in [m]$, we get that on $\mathcal{E}_1$

$$\|f(\pi) - \tilde{\mu}^t(\pi)\|_2^2 \leq \sum_{j \in [m]} \rho^2 (\sigma_j^t)^2(\pi) = \rho^2 \|\sigma^t(\pi)\|_2^2, \tag{30}$$

which yields $\|f(\pi) - \tilde{\mu}^t(\pi)\|_2 \leq \rho \|\sigma^t(\pi)\|_2 = q_t(\pi)$. □

D.3 Proof of Lemma 3.3

Proof. For the constrained MLE estimate $\tilde{\theta}_t$ of $\theta^\star$ obtained using $\mathcal{D}_{\mathrm{pr}}^t$, by (Schlaginhaufen et al., 2025, Lemma 3.1), the following deterministic inequality holds

$$\|\theta^\star - \tilde{\theta}_t\|_{V_t}^2 \leq \kappa(\mathcal{L}_t(\theta^\star) - \mathcal{L}_t(\tilde{\theta}_t)) + \lambda\|\theta^\star - \tilde{\theta}_t\|_2^2, \tag{31}$$

where $V_t = \lambda I_m + \sum_{\tau=2}^t \Delta^\tau (\Delta^\tau)^\top$. It therefore remains to control the likelihood-ratio term. Applying (Lee et al., 2025, Theorem 3.1) to $\mathcal{L}_t(\theta)$, with Lipschitz constant L_t, yields, for every $\delta \in (0,1)$,

$$\mathbb{P}(\mathcal{L}_t(\theta^\star) - \mathcal{L}_t(\tilde{\theta}_t) \leq \mathrm{B}(\delta)) \geq 1 - \delta, \tag{32}$$

where

$$\mathrm{B}(\delta) = \log \tfrac{1}{\delta} + m \log(\max\{e, \tfrac{2eSL_t}{m}\}).$$

To bound L_t, note that $y^\tau \in \{0,1\}$, $\mathrm{s}(x) \in (0,1)$, and consequently $-1 \leq \mathrm{s}(\langle\theta, \Delta^\tau\rangle) - y^\tau \leq 1$. Hence

$$\|\nabla_\theta \mathcal{L}(\theta)\|_2 = \Big\| \sum_{\tau=2}^t \big(\mathrm{s}(\langle\theta, \Delta^\tau\rangle) - y^\tau\big)\Delta^\tau \Big\|_2 \leq \sum_{\tau=2}^t \|\Delta^\tau\|_2 \leq 2L(t-1) = L_t, \tag{33}$$

since $\|\Delta^\tau\|_2 \leq 2L$. Finally, by substituting (33) into $\mathrm{B}(\delta)$, and observing that $\|\theta^\star - \tilde{\theta}_t\|_2 \leq 2S$, we get that

$$\|\theta^\star - \tilde{\theta}_t\|_{V_t}^2 \leq \kappa \left[\log\left(\tfrac{1}{\delta}\right) + m \log\left(\max\left\{e, \tfrac{4eSL(t-1)}{m}\right\}\right)\right] + 4\lambda S^2 = \beta_t(\delta)^2, \tag{34}$$

holds with probability at least $1 - \delta$. □

D.4 Proof of Theorem 3.4

We work on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ which holds with probability $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \delta$. Let us denote $\beta = \beta_T(\delta/2)$ and observe that $\beta_t(\delta/2) \leq \beta_T(\delta/2)$ for all $t \leq T$ and $\delta \in (0, 1)$. We start by regrouping the regret R_T

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} J(\pi^t, x^t(\pi^t)) - J(\pi^\star, x^\star(\pi^\star)) \\
&= \sum_{t=1}^{T} J(\pi^t, x^t(\pi^t)) - J(\pi^t, x^\star(\pi^t)) + J(\pi^t, x^\star(\pi^t)) - J(\pi^\star, x^\star(\pi^\star)) \\
&= \sum_{t=1}^{T} \langle \theta^\star, \tilde{f}^t - f(\pi^t) \rangle + \sum_{t=1}^{T} r_e^t \\
&\leq \sum_{t=1}^{T} \|\theta^\star\|_2 \|\tilde{f}^t - f(\pi^t)\|_2 + \sum_{t=1}^{T} r_e^t \\
&\leq S L_\varphi \sum_{t=1}^{T} e_t + \sum_{t=1}^{T} r_e^t,
\end{aligned}$$

where $r_e^t = J(\pi^t) - J(\pi^\star)$ and the last two lines that follow from applying the Cauchy–Schwarz inequality and Assumption 2.7. The round $t = 1$ contributes $r_1^e = J(\pi^1) - J(\pi^\star) \leq 2B_f S$ by Assumption 2.8. For $t \geq 2$, we can apply Proposition C.1 to upper bound r_e^t. In particular, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$R_T \leq 2B_f S + S L_\varphi \sum_{t=1}^{T} e_t + 2\beta \sum_{t=2}^{T} \|u_t(\pi^t)\|_{V_{t-1}^{-1}} + 2S \sum_{t=2}^{T} (q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})), \tag{35}$$

where $u_t(\pi) = \tilde{\mu}^{t-1}(\pi) - \tilde{\mu}^{t-1}(\pi^{t-1})$. Recall that $\Delta^t = \tilde{f}^{t-1} - \tilde{f}^t$ and observe that

$$u_t(\pi^t) = (\tilde{\mu}^{t-1}(\pi^t) - \tilde{f}^t) - (\tilde{\mu}^{t-1}(\pi^{t-1}) - \tilde{f}^{t-1}) - \Delta^t.$$

Therefore, using the triangle inequality we get

$$\|u_t(\pi^t)\|_{V_{t-1}^{-1}} \leq \|\Delta^t\|_{V_{t-1}^{-1}} + \|\tilde{\mu}^{t-1}(\pi^t) - \tilde{f}^t\|_{V_{t-1}^{-1}} + \|\tilde{\mu}^{t-1}(\pi^{t-1}) - \tilde{f}^{t-1}\|_{V_{t-1}^{-1}}. \tag{36}$$

Since $V_{t-1} \succeq \lambda I_m$, we have $V_{t-1}^{-1} \preceq \lambda^{-1} I_m$ and $\|\cdot\|_{V_{t-1}^{-1}} \leq \lambda^{-1/2} \|\cdot\|_2$, which yields

$$\begin{aligned}
\|u_t(\pi^t)\|_{V_{t-1}^{-1}} &\leq \|\Delta^t\|_{V_{t-1}^{-1}} + \lambda^{-1/2} \|\tilde{\mu}^{t-1}(\pi^t) - \tilde{f}^t\|_2 + \lambda^{-1/2} \|\tilde{\mu}^{t-1}(\pi^{t-1}) - \tilde{f}^{t-1}\|_2 \\
&\leq \|\Delta^t\|_{V_{t-1}^{-1}} + \lambda^{-1/2} \|\tilde{\mu}^{t-1}(\pi^t) - f(\pi^t)\|_2 + \lambda^{-1/2} \|f(\pi^t) - \tilde{f}^t\|_2 + \\
&\quad + \lambda^{-1/2} \|\tilde{\mu}^{t-1}(\pi^{t-1}) - f(\pi^{t-1})\|_2 + \lambda^{-1/2} \|f(\pi^{t-1}) - \tilde{f}^{t-1}\|_2 \\
&\leq \|\Delta^t\|_{V_{t-1}^{-1}} + \lambda^{-1/2} L_\varphi (e_t + e_{t-1}) + \lambda^{-1/2} (q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})),
\end{aligned}$$

due to Assumption 2.7 and Lemma 3.2. The summation then gives

$$\sum_{t=2}^{T} \|u_t(\pi^t)\|_{V_{t-1}^{-1}} \leq \sum_{t=2}^{T} \|\Delta^t\|_{V_{t-1}^{-1}} + \frac{1}{\sqrt{\lambda}} \Big(2L_\varphi \sum_{t=1}^{T} e_t + \sum_{t=2}^{T} q_{t-1}(\pi^t) + \sum_{t=2}^{T} q_{t-1}(\pi^{t-1}) \Big). \tag{37}$$

By substituting (37) into (35), we get

$$R_T \leq 2B_f S + 2\beta \sum_{t=2}^{T} \|\Delta^t\|_{V_{t-1}^{-1}} + 2\Big(S + \frac{\beta}{\sqrt{\lambda}} \Big) \sum_{t=2}^{T} (q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})) + \Big(S + \frac{4\beta}{\sqrt{\lambda}} \Big) L_\varphi \sum_{t=1}^{T} e_t. \tag{38}$$

Based on Lemma C.3, we also have

$$\sum_{t=2}^{T} q_{t-1}(\pi^t) \le \rho\sqrt{C_\sigma m\gamma_T(T-1)}.$$

Observe now that for all $j \in [m]$, the posterior variance is non-increasing in the data (Chowdhury & Gopalan, 2017, Appendix F), i.e., $\sigma_j^{t-1}(\pi^{t-1}) \le \sigma_j^{t-2}(\pi^{t-1})$. Hence,

$$\sum_{t=2}^{T} q_{t-1}(\pi^{t-1}) \le \rho\sum_{t=2}^{T}\|\sigma^{t-2}(\pi^{t-1})\|_2 = \rho\sum_{t=1}^{T-1}\|\sigma^{t-1}(\pi^t)\|_2 \le \rho\sqrt{C_\sigma m\gamma_T T},$$

and we can bound

$$\sum_{t=2}^{T}(q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})) \le 2\rho\sqrt{C_\sigma m\gamma_T T}. \tag{39}$$

The two remaining sums in (38) can directly be bounded using Lemmas C.2 and 3.1, i.e.,

$$\sum_{t=2}^{T}\left\|\Delta^t\right\|_{V_{t-1}^{-1}} \le \sqrt{(T-1)\Big(2m\log\Big(1+\tfrac{4(T-1)L^2}{m\lambda}\Big) + \tfrac{12mL^2}{\lambda\log 2}\log\Big(1+\tfrac{4L^2}{\lambda\log 2}\Big)\Big)},$$

and $\sum_{t=1}^{T} e_t \le T\sqrt{2CK^{-c}/\delta}$, which, when combined with (39), completes the proof.

D.5 Sufficient condition for achieving no-regret

Corollary D.1 (No regret)**.** *Under the assumptions of Theorem 3.4, choose $K = \lceil T^{1/c}\rceil$. Then $R^T = \widetilde{O}\big(\kappa\, m\sqrt{\gamma_T T}\big)$ with probability at least $1-\delta$. In particular, $R^T/T \to 0$ whenever $\gamma_T = o(T)$, which holds for the linear kernel with $\gamma_T = O(p\log T)$, and the squared-exponential kernel with $\gamma_T = O(\log^{p+1} T)$.*

With $K = \lceil T^{1/c}\rceil$ we have $K^{-c} \le T^{-1}$, so $\rho \le B + \frac{L_\varphi}{\sigma}\sqrt{2C/\delta}$ and $T\sqrt{2CK^{-c}/\delta} \le \sqrt{2CT/\delta} = O(\sqrt{T})$. Substituting into Theorem 3.4 and using $\beta_T(\delta/2) = O\big(\sqrt{\kappa}(\sqrt{m\log T} + \sqrt{\lambda}S)\big)$ gives the displayed bound, where the final $\tilde{O}$ expression follows by absorbing logarithmic factors and the problem constants $S, B, L_\varphi, \sigma, \lambda, C, \delta$, and by $\gamma_T \ge 1$.

D.6 Proof of Corollary 3.5

By Proposition C.1, on $\mathcal{E}_1 \cap \mathcal{E}_2$, which has probability at least $1-\delta$, for every $t \ge 2$, $J(\pi^t, x^\star(\pi^t)) - J(\pi^\star, x^\star(\pi^\star)) \le c_t$, where $c_t = 2\omega_t(\pi^t)$. In particular, given the definition of $t^\star$ in (10), we have

$$J(\pi^{t^\star}, x^\star(\pi^{t^\star})) - \min_{\pi\in\Pi} J(\pi, x^\star(\pi)) \le c_{t^\star} \le \frac{1}{T-1}\sum_{t=2}^{T} c_t.$$

By the definition of c_t, letting $\beta = \beta_T(\delta/2)$, we have

$$\sum_{t=2}^{T} c_t \le 2\beta\sum_{t=2}^{T}\|u_t(\pi^t)\|_{V_{t-1}^{-1}} + 2S\sum_{t=2}^{T}(q_{t-1}(\pi^t) + q_{t-1}(\pi^{t-1})),$$

which, when combined with upper-bounds (37) and (39), and Lemmas C.2 and 3.1, directly gives

$$c_{t^\star} \le 2\beta\sqrt{\frac{2m\log\Big(1+\frac{4(T-1)L^2}{m\lambda}\Big)+Q}{(T-1)}} + \frac{4(S+\frac{\beta}{\sqrt{\lambda}})\rho\sqrt{C_\sigma m\gamma_T T}}{T-1} + \frac{4\beta L_\varphi T}{\sqrt{\lambda}(T-1)}\sqrt{\frac{2CK^{-c}}{\delta}}. \tag{40}$$

Observe that for $K^c \geq 8L_{\mathrm{SE}}^2 CT/\delta\epsilon^2$ it holds that $\sqrt{2TCK^{-c}/\delta} \leq \epsilon/(2L_{\mathrm{SE}})$ and $K^{-c} = O(T^{-1})$. Consequently, for $\beta_T(\delta/2) = O\big(\sqrt{\kappa}(\sqrt{m\log T} + \sqrt{\lambda}S)\big)$, we get

$$c_{t^\star} = \tilde{O}\left(m\sqrt{\frac{\kappa\gamma_T}{T}}\right). \tag{41}$$

Therefore, taking $T = \tilde{\Theta}(\kappa m^2\gamma_T/\epsilon^2)$ with a sufficiently large hidden constant ensures $c_{t^\star} \leq \epsilon/2$.

Since $e_{t^\star}^2 \leq \sum_{t=1}^T e_t^2$, on $\mathcal{E}_1$ it holds that $e_{t^\star} \leq \sqrt{2TCK^{-c}/\delta} \leq \epsilon/(2L_{\mathrm{SE}})$ regardless of how $t^\star$ was selected. Write $\bar{x} = x^{t^\star}(\pi^{t^\star})$, $\bar{\pi} = \pi^{t^\star}$ and $\bar{x}^\star = x^\star(\bar{\pi})$. For any $i \in \mathcal{N}$ and any $x_i \in \mathcal{X}_i$, using L_U-Lipschitzness twice and the Nash property of $\bar{x}^\star$,

$$U_i(x_i, \bar{x}_{-i}; \bar{\pi}) \leq U_i(x_i, \bar{x}^\star_{-i}; \bar{\pi}) + L_U e_{t^\star} \leq U_i(\bar{x}^\star; \bar{\pi}) + L_U e_{t^\star} \leq U_i(\bar{x}; \bar{\pi}) + 2L_U e_{t^\star}. \tag{42}$$

Because $L_U \leq L_{\mathrm{SE}}$, we have that $2L_U e_{t^\star} \leq \epsilon$, making $\bar{x}$ an ϵ-Nash equilibrium of $\Gamma(\bar{\pi})$. Similarly, for the leader condition, we have

$$J(\bar{\pi}, \bar{x}) - J(\bar{\pi}, \bar{x}^\star) = \langle \theta^\star, \varphi(\bar{\pi}, \bar{x}) - \varphi(\bar{\pi}, \bar{x}^\star)\rangle \leq SL_\varphi e_{t^\star} \leq \epsilon/2,$$

since $SL_\varphi \leq L_{\mathrm{SE}}$. Finally, we have

$$\begin{aligned} J(\bar{\pi}, \bar{x}) &\leq J(\bar{\pi}, \bar{x}^\star) + \epsilon/2 \leq \min_{\pi\in\Pi} J(\pi, x^\star(\pi)) + c_{t^\star} + \epsilon/2 \\ &\leq \min_{\pi\in\Pi} J(\pi, x^\star(\pi)) + \epsilon \leq J(\pi, x^\star(\pi)) + \epsilon, \end{aligned}$$

for all $\pi \in \Pi$. Thus, $(\pi^{t^\star}, x^{t^\star}(\pi^{t^\star}))$ is an ϵ-Stackelberg equilibrium with probability at least $1 - \delta$.

E Carbon-Tax Incentive Design: Implementation Details And Additional Results

This appendix provides the numerical parameters and implementation details used for the carbon-tax experiment in Section 4. The game specification and its structural properties are given in Appendix A. We first describe the numerical calibration and then detail the implementation of the lower-level equilibrium solver, the models used to learn the equilibrium feature maps, the preference estimator, and the optimization of the leader's acquisition function.

E.1 Case-study calibration

Tables 1 and 2 report the parameters used throughout the experiments. The calibration is chosen to induce heterogeneous strategic responses across firms. For example, F1 has relatively low production costs but high emissions intensity, whereas D3 has cleaner production and lower abatement costs at the expense of a higher production cost. The latent leader weights are set to $\theta^\star = (0.7, 0.3)$, meaning that the leader assigns greater importance to emissions reduction. The coefficient values 0.05, 0.5, 0.5, and 0.5 are part of the case-study calibration and should be interpreted as known performance objective-construction parameters.

The resulting scalar cost is used only to evaluate the algorithm. Neither J nor $\theta^\star$ is revealed to the leader during learning. The corresponding full-information Stackelberg solution, computed offline for evaluation purposes using an extensive grid search procedure is $\pi^\star = (7.92,\ 22.96,\ 1.85)$ with $J(\pi^\star) = 1.067$.

E.2 Lower-level equilibrium computation

For each announced policy π^t, the followers solve the induced game using projected pseudogradient iterations. Strong monotonicity established in Proposition A.1 ensures that these iterations converge to the unique Nash equilibrium (Facchinei & Pang, 2003, Ch.12.1).

Table 1: Market, policy, scaling, and leader-objective parameters.

Parameter	Value	Interpretation
A	220	Inverse-demand intercept
B	6	Inverse-demand slope
$\bar{\tau}$	30	Maximum domestic carbon tax
$\bar{b}$	40	Maximum foreign carbon tax
$\bar{s}$	20	Maximum domestic abatement subsidy
S_E	19.7625	Emissions scale
S_P	87.5	Price scale
S_G	100.0	Subsidy expenditure scale
ω_G	0.05	Coefficient multiplying G/S_G in φ_2
ω_b	0.5	Coefficient multiplying $(b/\bar{b})^2$
ω_τ	0.5	Coefficient multiplying $(\tau/\bar{\tau})^2$
ω_s	0.5	Coefficient multiplying $(s/\bar{s})^2$
$\theta^\star$	$(0.7, 0.3)$	Hidden environmental/burden weights

Table 2: Firm-level parameters. Firms D1–D3 are domestic and F1–F2 are foreign.

Firm	c_i	ν_i	κ_i	$\bar{q}_i$
D_1	55	0.95	12	45
D_2	62	0.75	9	40
D_3	72	0.55	6	35
F_1	50	1.20	16	55
F_2	66	0.70	10	45

As described in Section 4, to isolate the effect of persistent lower-level inexactness, we consider three prescribed equilibrium-accuracy levels $\varepsilon_{\mathrm{NE}} \in \{0.001, 2, 5\}$. Rather than increasing the inner-loop accuracy with the horizon as indicated in Corollary D.1 for the asymptotic no-regret guarantee, each experiment maintains approximately the same response accuracy throughout the leader runs. The resulting realized equilibrium errors are reported in Figure 4. For $\varepsilon_{\mathrm{NE}} = 0.001$, the numerical solver reaches the reference equilibrium with negligible error, whereas the remaining runs maintain errors close to their prescribed levels.

E.3 Learning the equilibrium feature maps

The two equilibrium feature maps are learned independently using Gaussian-process regression implemented in GPyTorch. The kernel ridge regression expressions (4) and (5) coincide with the posterior mean and variance of a zero-mean Gaussian process (GP) with covariance kernel k_j and noise variance σ^2 (Rasmussen & Williams, 2005; Srinivas et al., 2010). In the current setting, however, the errors ε_j^τ arise from inexact lower-level equilibria and need not be zero-mean sub-Gaussian, resulting in a misspecified setting similar to (Bogunovic & Krause, 2021; Maddux et al., 2025). The GP posterior is used computationally as the kernel-ridge estimator and associated uncertainty measure entering the acquisition function.

For each objective $j \in \{1, 2\}$, we use a zero-mean GP with a squared-exponential/RBF covariance kernel,

$$k_j(\pi, \pi') = \exp\left(-\frac{\|\pi - \pi'\|_2^2}{2l_j^2}\right), \tag{43}$$

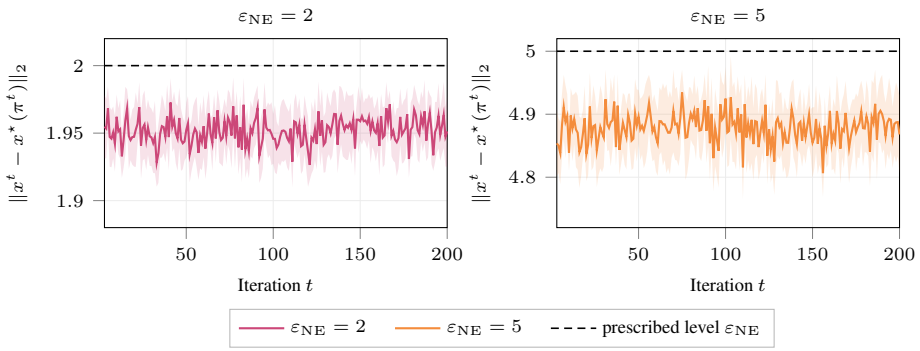

Figure 4: **Realized lower-level error.** Distance between the follower profile actually returned by APPROXNE and the exact Nash equilibrium $x^\star(\pi^t)$: the mean across seeds, with a 95% confidence interval, for the same runs as in Figure 2. In both, the realised error sits just below its prescribed level $\varepsilon_{\mathrm{NE}}$ (black dashed) for the whole horizon rather than decaying, so the accuracy of the lower-level response is held fixed while the leader learns. The case $\varepsilon_{\mathrm{NE}} = 0.001$ is omitted as its recorded error is numerically zero at every round. Note the narrow vertical ranges, as each panel spans about 10% of its own level.

where $l_j > 0$ denotes the learned lengthscale. The GP mean is set to zero. Prior to the online experiment, we collect a preliminary set of leader-action/observation pairs and use these data solely to calibrate the GP hyperparameters. For each $j \in \{1, 2\}$, the hyperparameters are estimated by maximizing the exact GP marginal likelihood. Equivalently, we minimize the negative exact marginal log likelihood (Rasmussen & Williams, 2005). Optimization is performed for 1250 iterations using Adam with learning rate 0.1. We use a Gaussian likelihood whose noise parameter is learned jointly with the kernel hyperparameters and constrained to satisfy $\sigma_j^2 \geq 0.01$. After this calibration step, the learned hyperparameters are kept fixed and in the experiments we used a common $\sigma = \max_j \sigma_j$. During the execution of Algorithm 1, the GP training set is updated only with the accumulated online observations and the posterior mean and standard deviation are recomputed without re-optimizing the kernel hyperparameters.

E.4 PREFERENCE LEARNING AND ACTION SELECTION

The preference weights were estimated through constrained logistic maximum likelihood using the Bradley–Terry model. The optimization was carried out using the Sequential Least Squares Programming (SLSQP) routine implemented in `scipy`. To improve numerical efficiency, we supplied the analytic gradient, initialized the optimization at $\tilde{\theta} = 1/m \cdot \mathbb{1}$, and constrained to satisfy $\tilde{\theta}_j \geq 0$ and $\mathbb{1}^\top \tilde{\theta} = 1$.

At each online round, the leader selects the next policy by minimizing the optimistic acquisition function in (7) over $\Pi = [0, 30] \times [0, 40] \times [0, 20]$. Since this objective is generally nonconvex, we approximate its global minimum using the standard `DIRECT` heuristic approach from (Brochu et al., 2009). This avoids relying on a particular initialization of a local solver and is inexpensive in the low-dimensional policy space considered in our numerical case study. Although the confidence parameter $\beta_t(\delta)$ proposed in Theorem 3.4 provides theoretical guarantees, in practice we empirically found that a fixed schedule $\beta_t = 0.05\sqrt{\log(t+1)}$ enables the leader to find a high-quality policy for all values of $\varepsilon_{\mathrm{NE}}$, while avoiding the computational burden of finding the problem-specific constants.

Figure 5 reports the three components of the announced action $\pi^t = (\tau^t, b^t, s^t)$ for the same configurations and the same seeds as Figure 2, aggregated across seeds at each round, i.e. representing the mean across seeds, with a 95% confidence interval.

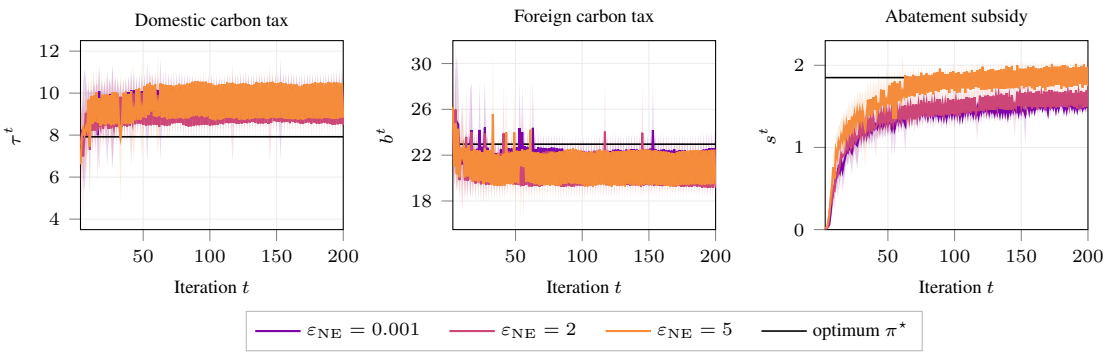

Figure 5: **Leader policy over the execution horizon.** The three components of the announced action $\pi^t = (\tau^t, b^t, s^t)$ for the same configurations as Figure 2, aggregated across the 10 seeds: mean across seeds, with a 95% confidence interval. Black lines are the components of the Stackelberg-optimal policy $\pi^\star = (7.92,\ 22.96,\ 1.85)$ located by grid search with full knowledge of the followers and of $\theta^\star$.

The round-to-round variation in the mean action indicates that the fluctuations are not solely driven by differences across seeds and is consistent with continued exploration by the optimistic acquisition rule. Thus, the announced actions continue to vary even after the leader has localized a near-optimal Π. The mean action ends at $(9.26,\ 20.61,\ 1.48)$, $(8.48,\ 19.16,\ 1.49)$ and $(8.83,\ 19.42,\ 1.74)$ for $\varepsilon_{\mathrm{NE}} = 0.001$, 2 and 5, i.e., the carbon tax component remains generally above $\tau^\star$ and the border adjustment below $b^\star$, so the leader reaches a region of near-optimal cost rather than $\pi^\star$ itself.

E.5 Alternative baselines

Our numerical study is intended primarily to validate the behavior predicted by the analysis and to isolate the effect of lower-level inexactness, rather than to establish empirical superiority over existing learning methods. A direct comparison with prior algorithms is not entirely straightforward because the closest approaches assume different feedback structures.

Preferential Bayesian optimization (González et al., 2017) and kernelized preference bandits (Pásztor et al., 2024) optimize an unknown scalar reward using pairwise comparisons between queried actions, without exploiting intermediate vector-valued objectives. Applying them here would therefore require treating the composition of the leader action with the unknown follower equilibrium as a single preference-generating function and discarding the observed objective vectors. Conversely, follower-agnostic Stackelberg methods (Maheshwari et al., 2024; Maddux et al., 2025) accommodate unknown lower-level games, but assume that the leader can evaluate its scalar objective at the observed follower response. This feedback is unavailable in our setting because the scalarization is unknown. Thus, meaningful comparisons with these approaches require nontrivial adaptations and are left for future empirical study.