EFFICIENT CONTEXTUAL PARTIAL MONITORING WITH GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite stochastic partial monitoring (PM) games with side information, a paradigm of sequential learning with partial feedback. On each round $t \in [T]$, both player and environment observe a context and player chooses an action and environment simultaneously chooses an outcome. The player neither observes the outcome nor the loss; instead they only observe a feedback symbol which is partially informative about the outcome. We utilize the Estimation-to-Decisions (E2D) meta-algorithm of Foster et al. (2021) to design E2D.PMSIDE, a computationally-efficient framework for contextual partial monitoring, which uses an online regression oracle to efficiently estimate the outcome distributions based on historical (context, action, feedback symbol) tuples. Since the exploration rule given by the original E2D algorithm is intractable to compute, we propose two computationally-efficient exploration rules that achieve optimal regret orders in T in globally and locally observable settings, respectively. We show that a variant of E2D.PMSIDE named OE2D.PMSIDE can be combined with offline regression oracles to give an efficient algorithm with optimal regret when the contexts are drawn independently from a fixed distribution. We evaluate our algorithm on a synthetic label-efficient prediction game and an LLM routing task, in which a learner decides from prompt, uncertainty and internal state features whether to escalate a query from a weak model to a strong one. We show that our method has superior performance than prior approaches.

1 INTRODUCTION

The contextual stochastic partial monitoring problem (Bartók & Szepesvári, 2012) is an online learning problem where the learner repeatedly takes actions so as to minimize its loss based on the environment's partial feedback. At the beginning of the game, the learner and the environment are given a pair $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ of $k \times d$ matrices, where k is the number of actions available to the learner and d the number of outcomes available to the environment. The *loss matrix* $\boldsymbol{L} \in [0,1]^{k \times d}$ records the loss L_{ao} incurred when the learner plays action a and the environment's outcome is o. The *feedback matrix* $\boldsymbol{\Phi} \in \Sigma^{k \times d}$ records what the learner is shown instead: playing a against outcome o reveals the symbol Φ_{ao} from a finite alphabet $\Sigma = \{1, \dots, m\}$. The interaction between the learner and the environment proceeds over T rounds. At round t the learner observes a context x_t (which may be chosen adversarially), samples an action $A_t \sim p_t(\cdot \mid x_t)$ according to a policy $p_t : \mathcal{X} \to \Delta(k)$, and the environment draws an outcome O_t from some stochastic distribution $q^*(x_t)$ independently of the learner's randomization. The learner then observes the feedback symbol $\sigma_t = \Phi_{A_t O_t}$ and incurs the loss $L_{A_t O_t}$, which remains hidden. The goal of the learner is to minimize its regret, which is the difference between its expected cumulative loss and that of the best policy in hindsight:

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \left(p_t^\top \boldsymbol{L} q^*(x_t) - \min_{\lambda \in \Delta(k)} \lambda^\top \boldsymbol{L} q^*(x_t) \right).$$

Contextual stochastic partial monitoring has many applications, such as learning from one-sided feedback (Helmbold et al., 2000; Sculley, 2007), dynamic pricing (Kleinberg & Leighton, 2003), data procurement in crowdsourcing (Singla et al., 2014), and label-efficient prediction (Helmbold et al., 1997). Below we give two examples:

Example 1: LLM routing. Modern Large language models (LLM)-based question-answering service optimizes its cost-efficiency by operating in a hybrid mode (An et al., 2026): it keeps two LLMs – one weak but cheap and one strong but expensive; and given a prompt, it judiciously decides

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.

which LLMs to route the prompt to, obtaining its response. At time step t, the service receives a new prompt represented by context x_t, and the weak model generates a response at negligible cost, and the router then either serves the weak response (a_1) or queries the strong model and serves its response (a_2). The environment's (hidden) outcome is whether the weak response agrees with the strong one (o_1) or does not agree (o_2). We take the strong model to be correct and to cost 1 per query, and we incur cost of $\ell_{01} > 1$ when a served weak model's response is wrong. Partial monitoring enters through the feedback: serving the weak response never reveals whether it was correct, so both outcomes yield the same symbol $\odot$, and only querying the strong model distinguishes them. The game's loss matrix and feedback matrix are summarized below:

$$\boldsymbol{L} = \begin{matrix} \\ a_1 \\ a_2 \end{matrix} \begin{matrix} \begin{matrix} o_1 & o_2 \end{matrix} \\ \begin{bmatrix} 0 & \ell_{01} \\ 1 & 1 \end{bmatrix} \end{matrix}, \qquad \boldsymbol{\Phi} = \begin{matrix} \\ a_1 \\ a_2 \end{matrix} \begin{matrix} \begin{matrix} o_1 & o_2 \end{matrix} \\ \begin{bmatrix} \odot & \odot \\ \perp & \wedge \end{bmatrix} \end{matrix}.$$

This game was also referred to as the Apple Tasting game Helmbold et al. (2000) in earlier works.

Example 2: contextual label efficient prediction game (Lattimore & Szepesvári, 2020; Helmbold et al., 1997). Suppose we would like to learn a spam filter online, which predicts whether an email is spam (o_1) or not (o_2) based on mail features (x_t) and directs it to the inbox (a_2) or spam folder (a_3). There is also a third option where the classifier can ask the recipient/expert for classification, which is accurate but incurs a fixed cost. Only when the learner chooses action a_1 does it get to see the true label of the email. The game's loss matrix and feedback matrix are summarized below:

$$\boldsymbol{L} = \begin{matrix} \\ a_1 \\ a_2 \\ a_3 \end{matrix} \begin{matrix} \begin{matrix} o_1 & o_2 \end{matrix} \\ \begin{bmatrix} c & c \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix}, \qquad \boldsymbol{\Phi} = \begin{matrix} \\ a_1 \\ a_2 \\ a_3 \end{matrix} \begin{matrix} \begin{matrix} o_1 & o_2 \end{matrix} \\ \begin{bmatrix} \perp & \wedge \\ \odot & \odot \\ \odot & \odot \end{bmatrix} \end{matrix}.$$

Depending on the feedback structure, finite stochastic partial-monitoring games are classified into four categories (classification theorem): trivial, easy, hard, and hopeless games (Bartók et al., 2011; Lattimore & Szepesvári, 2020). In hopeless games, the feedback is too uninformative to support meaningful learning, and the regret might grow linearly with the time horizon ($\Omega(T)$). In contrast, in trivial games, there exists an action which is optimal for every outcome; consequently, they can be solved with zero regret. The main cases requiring more refined analysis are easy ($\Theta(\sqrt{T})$) and hard games ($\Theta(T^{2/3})$). Both classes satisfy global observability, which we define shortly. In addition, easy games satisfy local observability, whereas hard games do not.

Related work: partial monitoring with side information. We adopt the contextual partial monitoring setting of Bartók et al. (2012a); Heuillet et al. (2024); Komiyama et al. (2015), and use the terms *contextual partial monitoring* and *partial monitoring with side information* interchangeably for it. The classification of easy and hard games established for finite stochastic non-contextual partial monitoring (Bartók et al., 2011; Lattimore & Szepesvári, 2020) extends to this setting, at least in its dependence on T. First, the non-contextual lower bounds of Bartók et al. (2011) transfer immediately, since a non-contextual game can be viewed as a contextual game with a dummy context. Second, for the upper bounds, Foster et al. (2022) match the non-contextual rates by relating the decision-estimation coefficient (DEC) to the exploration-by-optimization objective of Lattimore & Szepesvári (2020) (ExO), and the contextual form of that framework (Foster et al., 2021) carries the same guarantees over. These optimal statistical guarantees come at a computational cost, however: the algorithm attaining them selects each action distribution by solving a minimax problem at every round, which is not known to be tractable (see Section 2.3 for further discussion).

Meanwhile, computationally efficient algorithms have been proposed for contextual partial monitoring, by imposing strong assumptions on the function class that models the true outcome distribution $q^\star$. Utilizing generalized linear models for modeling (functions of) $q^\star$, Bartók et al. (2012a) proposed CBPSide, a deterministic algorithm based on confidence intervals that extends UCB-style methods from the bandit setting; it applies only to easy games, where it achieves $O(\sqrt{T})$ regret. Heuillet et al. (2024) later introduced CBPSide* and its randomized counterpart RandCBPSide*, which cover both classes, attaining $O(\sqrt{T})$ regret on easy games and $O(T^{2/3})$ on hard ones. Kirschner et al. (2023) proposed the stochastic IDS-FW algorithm, which also covers both classes but whose computational cost grows quadratically in the number of rounds and which requires the context set to be finite and known in advance.

This motivates the question that drives our work: are there computationally efficient algorithms for contextual partial monitoring that support general function approximation and attain the optimal rates? In this paper, we answer this question affirmatively. Specifically, we make the following contributions:

Contributions. (1) we propose an algorithmic framework E2D.PMSide that reduces stochastic contextual partial monitoring to online regression. It is based on the E2D (Estimation to Decision) framework (Foster et al., 2021), where the algorithm aims at finding an exploration distribution that certifies a low Decision-Estimation Coefficient; our key observation is this can be done by optimizing a surrogate objective that trades off between greediness and coverage, which paves the way for the design of efficient algorithms for contextual partial monitoring. (2) Based on the above relaxed objective, we instantiate E2D.PMSide in the locally observable and globally observable games to obtain SquareCB.PMSide and ϵ-Greedy.PMSide algorithms respectively, achieving regret of optimal order in T in their respective settings. These, to our knowledge, are the first *computationally efficient* contextual partial monitoring algorithms with *general function approximation.* (3) When the contexts are iid, we propose a variant of E2D.PMSide named OE2D.PMSide, which is an efficient stochastic contextual partial monitoring algorithm based on offline regressions, a computational primitive that admits simpler implementations in practice. (4) We evaluate SquareCB.PMSide in two contextual partial monitoring tasks: label-efficient prediction and LLM routing, and show that it outperforms baselines.

2 Preliminaries

Basic Notation. We write $\mathbf{0}$ and $\mathbf{1}$ for the all-zeros and all-ones column vectors, $e_1, \dots, e_k$ for the standard basis vectors, and bold symbols for matrices. For a vector or matrix A, $\|A\|_\infty = \max_{ij} |A_{ij}|$ is the entrywise maximum norm, M_i denotes the vector generated by i-th row, and M_{ij} denotes the element at i-th row and j-th column. For two distributions u and v, $D^2_{\mathrm{H}}(u \,|\, v)$ denotes their squared Hellinger distance. Let $\mathrm{co}(\mathcal{S})$ denote the convex hull of a set $\mathcal{S}$.

Signal matrix. Given a feedback matrix $\mathbf{\Phi}$, it is convenient to describe the feedback of a single action as a linear map. The *signal matrix* of action a is the $m \times d$ matrix $\boldsymbol{S}_a$ with $\boldsymbol{S}_a[\sigma, o] = \mathbf{1}\{\Phi_{ao} = \sigma\}$. Writing $\Delta(n)$ for the probability simplex on n elements, if the outcome is drawn from $q \in \Delta(d)$ then $\boldsymbol{S}_a q \in \Delta(m)$ is precisely the distribution of the symbol observed on playing a. Two candidate outcome distributions q_1, q_2 are therefore distinguishable through action a exactly to the extent that $\boldsymbol{S}_a q_1$ and $\boldsymbol{S}_a q_2$ differ; and it is this quantity, rather than the gap between q_1 and q_2 themselves, that governs what the learner can hope to identify. Recall that in our stochastic contextual partial monitoring problem, each round is preceded by a context x_t drawn from a set $\mathcal{X}$, and the outcome distribution depends on it: outcomes follow $q^*(x_t)$ for an unknown $q^* : \mathcal{X} \to \Delta(d)$. The learner is given a model class $\mathcal{Q} \subseteq \{q : \mathcal{X} \to \Delta(d)\}$, which we assume is well-specified.

Assumption 2.1 (Realizability)**.** The model class $\mathcal{Q}$ contains the true outcome distribution q^*.

2.1 Structure of partial monitoring games

We now review some basic structures of partial monitoring games revealed in prior works.

Definition 2.1 (Cell decomposition, Bartók et al. (2012b))**.** The *cell* $\mathcal{C}_a$ of action a is defined as the subset of outcome distributions where action a is optimal. Formally, $\mathcal{C}_a = \{q \in \Delta(d) : (e_a - e_b)^\top \boldsymbol{L} q \le 0, \forall b \in [k]\}$. The collection $\{\mathcal{C}_a : a \in [k]\}$ is called the cell decomposition.

Neighborhood graph (Lattimore & Szepesvári, 2020). Cells that meet in a face of dimension $d-2$ correspond to actions that can be optimal simultaneously, and this adjacency structure organizes the analysis that follows. Call an action *Pareto optimal* when its cell is full dimensional, $\dim(\mathcal{C}_a) = d-1$, taking the dimension of a polytope to be that of its affine hull, and write Π for the set of such actions together with $\Delta(\Pi) := \{\lambda \in \Delta(k) : \lambda_b = 0 \text{ for all } b \notin \Pi\}$ for the corresponding face of $\Delta(k)$. The remaining actions are of two kinds: a is *degenerate* when $\mathcal{C}_a \neq \varnothing$ but $\dim(\mathcal{C}_a) \le d-2$, and *dominated* when $\mathcal{C}_a = \varnothing$. Since $\dim(\Delta(d)) = d-1$, neither kind is ever uniquely optimal; each is optimal only on a set of measure zero. A game with no degenerate actions is called *non-degenerate*, and two actions are *duplicates* when their loss rows coincide, $(e_a - e_b)^\top \boldsymbol{L} = \mathbf{0}$. Two Pareto optimal actions are *neighbors* when their cells meet in a face of the largest dimension short of coincidence, $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d-2$; the intersection reaches dimension $d-1$ only for duplicates. Writing $\mathcal{E}$ for the resulting edge set, $(\Pi, \mathcal{E})$ is the *neighborhood graph* of $\mathcal{G}$.

2.2 Observability

In partial-monitoring games, the learner observes feedback rather than the losses or rewards associated with its actions. Therefore, a central challenge is to infer (differences of) losses of actions from the

observed feedback. Whether this is possible depends on the interplay between the loss matrix and the feedback matrix.

Definition 2.2 (Global observability, Adapted from Bartók et al. (2012b) and Lattimore & Szepesvári (2020))**.** A partial monitoring game is *globally observable* if, for *every pair of actions* $(a, b) \in [k]\times[k]$, there exists a *witness function* $w_{(a,b)} : [k] \times \Sigma \to \mathbb{R}$ such that

$$L_{ao} - L_{bo} = \sum_{c=1}^{k} w_{(a,b)}(c, \Phi_{co}), \qquad \text{for all } o \in [d].$$

Definition 2.3 (Local observability, Lattimore & Szepesvári (2020))**.** A globally observable, *non-degenerate* partial monitoring game is *locally observable* if *for every edge* $e = (a, b) \in \mathcal{E}$ of the neighborhood graph the witness function w_e of Definition 2.2 may be chosen to be supported only on the endpoints, that is $w_e(c, \cdot) \equiv 0$ for all $c \notin \{a, b\}$, so that

$$L_{ao} - L_{bo} = \sum_{c\in\{a,b\}} w_e(c, \Phi_{co}), \qquad \text{for all } o \in [d].$$

By definition, every locally observable game is a globally observable game. Under global observability, the loss difference $L_{ao} - L_{bo}$ may depend on feedback from all actions $c \in [k]$, whereas under local observability, it can be expressed using only the feedback from the sequence of actions in $\mathcal{E}$ connecting a and b. Consequently, locally observable games permit more direct loss-difference estimation than globally observable games in general. Since trivial games and hopeless games are not interesting from an algorithm design perspective, for the remainder of the paper we focus on globally observable games.

Assumption 2.2 (Global observability and non-degeneracy; Lattimore & Szepesvári (2020))**.** We assume that the partial-monitoring game $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ is globally observable, non-degenerate, and contains no duplicate actions.

For simplicity, we maintain this assumption throughout the remainder of the analysis, following Lattimore & Szepesvári (2020). We discuss implications of non-degeneracy and possible relaxations in Appendix A.2.4.

2.3 The E2D meta-algorithm for interactive decision making

Foster et al. (2021) proposes E2D (Estimation-to-Decision), a meta-algorithm for interactive decision making, which reduces interactive decision making to online estimation. The E2D meta-algorithm yielded many practical algorithms for decision making, especially for contextual bandit problems (e.g., Foster & Rakhlin, 2020; Zhu et al., 2022; Zhu & Mineiro, 2022; Zhang et al., 2023). In a nutshell, at each round, it calls an online model estimation oracle to obtain an estimated model $\hat{q}$; based on $\hat{q}$ and context x, it solves a minimax optimization problem to obtain an exploration distribution p to take actions. Specialized to partial monitoring, the optimization problem can be written as:

$$\min_{p\in\Delta(k)} \underbrace{\sup_{q\in\mathcal{Q}_x,\lambda\in\Delta(k)} \left[p^\top \boldsymbol{L}\hat{q}(x) - \lambda^\top \boldsymbol{L}\hat{q}(x) - \gamma \cdot \mathbb{E}_{a\sim p}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}(x) \,\big|\, S_a q(x)\big)\right]\right]}_{\mathcal{V}^{\mathrm{dec}}_\gamma(\hat{q}(x),\mathcal{Q}_x,p)}, \tag{1}$$

where $\mathcal{Q}_x := \big\{q(x) : q \in \mathcal{Q}\big\} \subseteq \Delta(d)$ denotes the model class $\mathcal{Q}$ restricted to x. The objective explicitly trades off between regret and information gain: the first part $p^\top \boldsymbol{L}\hat{q}(x) - \lambda^\top \boldsymbol{L}\hat{q}(x)$ encourages p to have a low regret, while the second part $-\gamma \cdot \mathbb{E}_{a\sim p}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}(x) \,\big|\, S_a q(x)\big)\right]$ encourages p to explore diverse actions to correct $\hat{q}$ and gain new information about q^*. The minimax value of the objective is called the Decision-Estimation Coefficient (DEC) (Foster et al., 2021), denoted as $\mathrm{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x)$. To help the learner make decisions efficiently, E2D requires access to an oracle that can provide accurate estimation of the outcome distributions for all contexts online, which, when specialized to our stochastic contextual partial monitoring setting, is defined as:

Definition 2.4. An *online outcome estimation oracle* $\mathrm{Alg}_{\mathrm{On}}$ is an algorithm that estimates the unknown outcome distribution q^* from the data in a sequential fashion. That is, at each round $t \in [T]$, given the feedback symbols collected so far, the oracle produces an estimate

$$\hat{q}_t = \mathrm{Alg}^{(t)}_{\mathrm{On}}\Big(\big\{(A_s, x_s, \sigma_s)\big\}_{s=1}^{t-1}\Big) \in \mathrm{co}(\mathcal{Q}),$$

where $\mathrm{Alg}^{(t)}_{\mathrm{On}}$ denotes the map applied at round t and $\big\{(A_s, x_s, \sigma_s)\big\}_{s=1}^{t-1}$ is the history available at the beginning of round t (note that σ_s is an observed feedback symbol, but not the incurred loss).

$\mathrm{Alg}_{\mathrm{On}}$ is said to satisfy *a Hellinger regret bound* of $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)$ if its online Hellinger regret

$$\mathrm{Reg}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T) := \sum_{t=1}^{T} \mathbb{E}_{a\sim p_t(\cdot|x_t)}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_a q^*(x_t)\big)\right], \tag{2}$$

is at most $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)$ with probability $1-\delta$.

We note that instead of requiring $\hat{q}_t(x_t)$ to be pointwise close to $q^*(x_t)$ on average, Definition 2.4 only requires their associated feedback distributions to be close on average. This is reasonable since we can only use the feedback σ_s's to guide our online learning of $\hat{q}_t$; in the extreme case, it is possible that different q's results in the same feedback distribution, making the underlying q unidentifiable. Appendix A.1 gives two concrete examples of the online outcome estimation oracle. For example, For finite $\mathcal{Q}$, Vovk's aggregating algorithm (Vovk, 1995) attains $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta) = O(\log(|\mathcal{Q}|/\delta))$. Equipped with an online outcome estimation oracle, E2D is known to have a regret bound of

$$\mathrm{Reg}_T \le O\left(T\max_{x\in\mathcal{X}} \mathrm{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x) + \gamma\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)\right),$$

which translates to $O(T^{2/3})$ and $O(\sqrt{T})$ regrets with optimal tuning of γ in globally and locally observable settings, respectively (Foster et al., 2021). However, a key drawback of the naive E2D algorithm is that directly solving the minimax optimization problem (Equation (1)) is computationally intractable: although the objective is linear in p, it is not jointly concave in q, λ. As we will see, our algorithm design addresses this computational challenge.

3 Efficient algorithm for contextual partial monitoring

In this section we present E2D.PMSide (Estimation-to-Decision for Partial Monitoring with Side information, Algorithm 1), which can be viewed as a specialization of the E2D meta-algorithm. At every round t, E2D.PMSide obtains an outcome distribution estimate $\hat{q}_t(\cdot)$ (line 4) from an online outcome estimation oracle with Hellinger regret guarantee. Then it uses $\hat{q}_t(\cdot)$ and observed context x_t to construct an exploration distribution p_t that approximately minimizes an objective function $\mathcal{V}_\gamma(\hat{q}_t(x_t), p)$, a relaxation of the DEC objective $\mathcal{V}^{\mathrm{dec}}_\gamma(\hat{q}(x_t), \mathcal{Q}_{x_t}, p)$ (line 5). The choice of p_t balances two competing demands. The first term in the objective measures the loss of p_t relative to a comparator policy λ under the estimated outcome distribution $\hat{q}_t(x_t)$, and is small when p_t exploits what has been learned. The second term is a *coverage* term (which we will define shortly) that measures the amenability of p's data in evaluating λ's loss; the smaller the coverage value the better p is for exploration. The exploration parameter γ fixes the rate of exchange between the two. We now define coverage formally.

Definition 3.1 (Coverage)**.** We define the coverage of action distribution $p \in \Delta(k)$ for action distribution $\lambda \in \Delta(k)$ for partial monitoring game $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ as

$$\mathrm{Coverage}(p, \lambda) := \sup_{q_1, q_2 \in \Delta(d)} \frac{\Big((\lambda - p)^\top \boldsymbol{L}\,(q_1 - q_2)\Big)^2}{\mathbb{E}_{a\sim p}\left[D^2_{\mathrm{H}}\big(S_a q_1\,\big|\,S_a q_2\big)\right]}. \qquad \text{(with convention } \tfrac{0}{0} = 0\text{)}$$

To make sense of this definition it may be helpful to think about the following hypothetical offline policy evaluation problem in non-contextual stochastic partial monitoring: suppose the outcome distribution q^* is unknown; we use a behavior policy p to take action and collect feedback symbols, so as to estimate the expected loss of λ (more precisely, the expected loss difference between λ and p); how much utility does each data point carry for such evaluation? Let q_1 be q^* and q_2 be $\hat{q}$, our estimate of q^* based on data. The numerator thus measures the squared policy evaluation error; the denominator measures the in-sample estimation error, which we expect to decrease via standard concentration arguments. A small $\mathrm{Coverage}(p, \lambda)$ implies that feedback signals via following p is useful in estimating p and λ's loss difference.

Our coverage in the above definition bears some similarities with the coverage defined in the contextual bandit setting by Qin & Zhang (2026b), but is different in two important ways: (1) here, the numerator and the denominator are with respect to the loss matrix (evaluation error) and the signal matrix (information gain), respectively, where in the bandit setting they are defined both using the loss function; (2) it is well-known that for some globally observable partial monitoring problems the absolute value of a loss of an action is not identifiable (Luo (2019), Example 1.1), and only loss

Algorithm 1 E2D.PMSide

1: **parameters:**
Online estimation oracle $\mathrm{Alg}_{\mathrm{On}}$
Exploration parameter $\gamma > 0$
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Receive context x_t
4: Compute estimate $\hat{q}_t = \mathrm{Alg}_{\mathrm{On}}^{(t)}(\{A_s, x_s, \sigma_s\}_{s=1}^{t-1})$
5: Find p_t that approximately minimizes

$$\mathcal{V}_\gamma(\hat{q}_t(x_t), p) := \sup_{\lambda \in \Delta(\Pi)} \underbrace{p^\top \boldsymbol{L}\hat{q}_t(x_t) - \lambda^\top \boldsymbol{L}\hat{q}_t(x_t)}_{\text{exploit}} + \frac{1}{\gamma} \cdot \underbrace{\mathrm{Coverage}(p, \lambda)}_{\text{explore}}.$$

using ϵ-Greedy.PMSide (Alg. 2) or SquareCB.PMSide (Alg. 3) for globally and locally observable settings, respectively.
6: Choose action $A_t \sim p_t$ and update the estimation oracle with (A_t, x_t, σ_t)
7: **end for**

Algorithm 2 ϵ-Greedy.PMSide for globally observable games

1: **parameters:**
2: Context x with predicted outcome distribution $\hat{q}(x)$
3: Exploration parameter $\gamma > 16k^2\beta^2$
4: Let $\hat{b} \in \arg\min_{a \in \Pi} e_a^\top \boldsymbol{L}\hat{q}(x)$
5: **return** $p = (1-\epsilon)e_{\hat{b}} + \epsilon\frac{1}{k}\mathbf{1}$ for $\epsilon = k\beta/\sqrt{\gamma}$

differences between action pairs is identifiable; thus, the numerator in the coverage measures the difference of losses of two policies. We now present the regret guarantee of E2D.PMSide:

Theorem 3.1 (Regret of E2D.PMSide)**.** Suppose $\mathrm{Alg}_{\mathrm{On}}$ has a Hellinger regret bound $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)$. Fix $\gamma > 0$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathrm{Reg}_T \le T \cdot \max_{t=1}^{T} \mathcal{V}_\gamma\big(\hat{q}_t(x_t), p_t\big) + \gamma \cdot \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta).$$

Its proof is deferred to Appendix A.2.1. In a nutshell, our theorem follows from the meta-regret theorem of Foster et al. (2021); as long as all action distributions p_t certify that $\mathcal{V}^{\mathrm{dec}}_\gamma(\hat{q}(x_t), \mathcal{Q}_{x_t}, p_t)$ has an upper bound, E2D has a small regret in terms of that bound. Indeed, our p_t here certifies that $\mathcal{V}^{\mathrm{dec}}_\gamma(\hat{q}(x_t), \mathcal{Q}_{x_t}, p_t)$ is upper bounded by $\mathcal{V}_\gamma(\hat{q}_t(x_t), p_t)$; and ideally, we should choose exploration distribution p_t's that make $\mathcal{V}_\gamma(\hat{q}_t(x_t), p_t)$'s as small as possible.

Unfortunately, finding p_t that exactly minimizes $\mathcal{V}_\gamma(\hat{q}_t(x_t), p)$ may be computationally intractable; to see this, note that the definition of coverage involves supremum over a pair of outcome distributions q_1, q_2, which is not a convex optimization problem. This motivates the question: can we find computationally efficient exploration rules $x \mapsto p(x)$ that *approximately minimize* $\mathcal{V}_\gamma\big(\hat{q}_t(x), p(x)\big)$ for all x? Indeed, we show in the following two subsections that such efficient exploration rule exist and yield efficient implementations of E2D.PMSide.

3.1 Efficient algorithm for globally observable games

We show that in the globally observable setting ϵ-Greedy.PMSide (Alg. 2), yields a policy p satisfying $\mathcal{V}_\gamma(\hat{q}(x), p(x)) \le O(\sqrt{1/\gamma})$, which gives an $O(T^{2/3})$ regret bound when plugged into E2D.PMSide. ϵ-Greedy.PMSide identifies the empirical best action $\hat{b}$ (Line 4) and assigns it probability $1-\epsilon$, spreading the remaining ϵ uniformly over all k actions (Line 5).

Lemma 3.1 (Guarantee for Alg. 2 for globally observable games)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be a globally observable partial monitoring game in the sense of Definition 2.2, and set $\beta := \max_{(a,b) \in [k] \times [k]} \|w_{(a,b)}\|_\infty$. We have for any $\hat{q} \in \Delta(d)$,

$$\mathcal{V}_\gamma\big(\hat{q}, p\big) \le \max_{\lambda \in \Delta(\Pi)} p^T \boldsymbol{L}\hat{q} - \lambda^T \boldsymbol{L}\hat{q} + \frac{16\beta^2}{\gamma} \sum_{c=1}^{k} \frac{1}{p_c}.$$

Thus, for every $\gamma > 16k^2\beta^2$, every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \le O\left(k\beta\sqrt{\frac{1}{\gamma}}\right).$$

Proof is deferred to Appendix A.2.2. Furthermore, by combining Theorem 3.1 and Lemma 3.1, our algorithm achieves the following minimax rate for globally observable games.

Corollary 1 (Regret of E2D.PMSide for globally observable games)**.** *Suppose* $\mathrm{Alg}_{\mathrm{On}}$ *has a Hellinger regret bound* $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)$*. For a globally observable partial monitoring game* $(\boldsymbol{L}, \boldsymbol{\Phi})$*, set* $\beta = \max_{(a,b)\in[k]\times[k]} \|w_{(a,b)}\|_\infty$*. Then, for any* $\delta \in (0,1)$*, and for any large enough* T *that satisfies* $\gamma > 16k^2\beta^2$ *with* $\gamma = (Tk\beta \,/\, \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta))^{2/3}$*, with probability of at least* $1-\delta$*, E2D.PMSide run with Alg 2 as the exploration rule satisfies,*

$$\mathrm{Reg}_T \le O\left(\sqrt[3]{T^2k^2\beta^2\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)}\right).$$

In the setting when $\mathcal{Q}$ is finite, setting the estimation oracle to be Vovk's aggregating algorithm, the above regret bound is of order $O(\sqrt[3]{T^2k^2\beta^2\log|\mathcal{Q}|})$. In the noncontextual setting, when $\mathcal{Q} = \Delta(d)$, we can set the estimation oracle to be any efficient exp-concave optimization algorithm (Hazan et al., 2007), which has $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta) = O(d\ln T)$, implying E2D.PMSide has a regret of $\tilde{O}(\sqrt[3]{T^2k^2\beta^2 d})$. This matches the lower bound of non-contextual stochastic partial monitoring (Bartók et al., 2011) with a simple computationally efficient approach.

3.2 Efficient algorithm for locally observable games

In the locally observable setting, we give an efficient exploration rule (Alg. 3) that finds p such that $\mathcal{V}_\gamma(\hat{q}_t(x), p(x)) \le O\left(\frac{1}{\gamma}\right)$, which yields a $O\left(\sqrt{T}\right)$ regret when plugged into E2D.PMSide. Under global observability, to estimate the loss difference between two actions (a, b), its associated witness function $w_{a,b}$ may be supported on all actions (Definition 2.2), so recovering the difference between a single pair of actions can require playing all k of them; this is why the certifying policy of Algorithm 2 mixes with the uniform distribution over $[k]$. In contrast, under local observability, the loss difference between two *neighboring* Pareto optimal actions can be expressed using the feedback of those two actions alone (Definition 2.3), and the difference between an arbitrary pair of Pareto optimal actions is then recovered by chaining such local relations along a path joining them in the neighborhood graph, so that only the actions encountered along that path need be played.

3.2.1 Sparse relaxation of $\mathcal{V}_\gamma(\hat{q}, p)$

We aim to find tractable relaxations of V_γ which sheds lights on finding efficient exploration strategies for approximately minimizing V_γ. In particular, we aim to exploit the structured feedback mechanism of locally observable games to restrict exploratory actions to those lying along paths that connect pairs of actions whose loss differences are of interest. The neighborhood graph $(\Pi, \mathcal{E})$ is connected. Assign an arbitrary action $\mathrm{root}_\mathcal{T}$ as the root and construct an in-tree $\mathcal{T}$ such that, there is a unique directed $\mathrm{path}_\mathcal{T}(a)$ from any action $a \in \Pi$ to $\mathrm{root}_\mathcal{T} \in \Pi$. Now, we can relax the $\mathcal{V}_\gamma(\hat{q}, p)$ as follows,

Lemma 3.2. Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be a locally observable partial monitoring game in the sense of Definition 2.3. Let $\mathcal{T}$ be an in-tree of the neighborhood graph. For all $p \in \Delta(\Pi)$, we have

$$\mathcal{V}_\gamma(\hat{q}(x), p) \le \max_{\lambda\in\Delta(\Pi)} p^T\boldsymbol{L}\hat{q}(x) - \lambda^T\boldsymbol{L}\hat{q}(x) + \frac{32}{\gamma}\max_{e\in\mathcal{E}}\|w_e\|_\infty^2 \sum_{(b,c):c\in\mathrm{path}_\mathcal{T}(b)}\left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c}\right). \tag{3}$$

Proof is deferred to Appendix A.2.3. Above result is true for any in-tree with an arbitrary root. However, in general, we do not want the paths to be arbitrary: for a "bad" action c and a "good" action b (i.e., $e_c^\top L\hat{q}$ is large while $e_b^\top L\hat{q}$ is small), we would like p_c to be small and p_b to be large to ensure $p^T\boldsymbol{L}\hat{q}(x) - \lambda^T\boldsymbol{L}\hat{q}(x)$ to be small; however, if c happens to lie in $\mathrm{path}_\mathcal{T}(b)$, that can cause the $\frac{\lambda_b}{p_c}$ term to be very large. How should the in-tree $\mathcal{T}$ be chosen so that the actions lying on each path are ones we are willing to play? Our answer is to let the tree follow the estimated losses. We take $\mathcal{T}$ to be $\boldsymbol{L}\hat{q}(x)$-decreasing (that is for a directed edge $e = (b, c)$, $e_b^\top\boldsymbol{L}\hat{q}(x) \ge e_c^\top\boldsymbol{L}\hat{q}(x)$), so that estimated loss decreases along every path toward the root and, in particular, the empirical best action sits at

Algorithm 3 SquareCB.PMSide for Locally observable games

1: **parameters:**
2: Context x with predicted outcome distribution $\hat{q}(x)$
3: Exploration parameter $\gamma > 0$, $\mu = k$ and $\kappa = \frac{\gamma}{8km^2}$
4: For each action $a \in \Pi$, compute $\hat{\ell}_a = e_a^\top \boldsymbol{L}\hat{q}(x)$ and let $b = \arg\min_{a\in\Pi} \hat{\ell}_{t,a}$
5: For each $a \in \Pi \setminus \{b\}$, define $p_a = \frac{1}{\mu+\kappa(\hat{\ell}_a - \hat{\ell}_b)}$; let $p_b = 1 - \sum_{a\in\Pi\setminus\{b\}} p_a$.
6: **return** p

the root. It then remains to allocate probability so that it increases toward the root and the familiar inverse-gap-weighting (IGW) rule (similar to Foster & Rakhlin (2020) for bandits), already does so. Accordingly, SquareCB.PMSide assigns probability by inverse gap (Line 5), and an action nearer the root of an $\boldsymbol{L}\hat{q}(x)$-decreasing in-tree receives more probability. Now we state the main guarantee of SquareCB.PMSide.

Lemma 3.3 (Guarantee for Alg. 3 for locally observable games)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be a locally observable partial monitoring game in the sense of Definition 2.3. For every $\gamma > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 3 run with $\kappa = \frac{\gamma}{8km^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \le O\left(\frac{m^2k^2}{\gamma}\right).$$

Proof is deferred to Appendix A.2.3. Furthermore, by combining Theorem 3.1 and Lemma 3.3, we can achieve the following minimax rate for locally observable games.

Corollary 2 (Regret of E2D.PMSide for locally observable games)**.** *Suppose* $\mathrm{Alg}_{\mathrm{On}}$ *has a Hellinger regret bound* $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)$*. Let Assumption 2.2 hold, and suppose in addition that the game is locally observable. Then, for any* $\delta \in (0,1)$*, with probability of at least* $1-\delta$*, E2D.PMSide run with Alg. 3 as the exploration rule (with* $\kappa = \frac{\gamma}{8km^2}$ *and* $\mu = k$*), with* $\gamma = (Tm^2k^2 \,/\, \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta))^{1/2}$ *satisfies*

$$\mathrm{Reg}_T \le O\left(\sqrt{T \cdot m^2k^2 \cdot \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta)}\right).$$

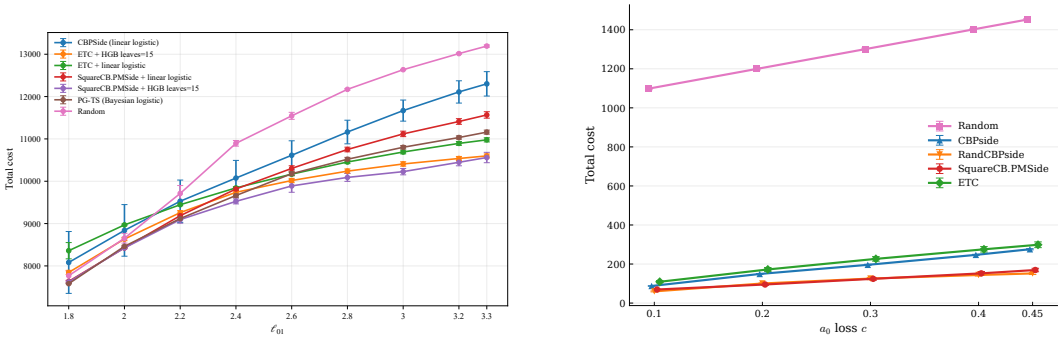
Furthermore, in the setting when $\mathcal{Q}$ is finite, setting the estimation oracle to be Vovk's aggregating algorithm, the above regret bound is of order $O(\sqrt{T \cdot m^2k^2 \cdot \log|\mathcal{Q}|})$. In the noncontextual setting, when $\mathcal{Q} = \Delta(d)$, we can set the estimation oracle to be the any efficient exp-concave optimization algorithm (Hazan et al., 2007), which has $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta) = O(d \ln T)$, implying E2D.PMSide has a regret of $\tilde{O}(\sqrt{T \cdot m^2k^2d})$. This matches the lower bound of non-contextual stochastic partial monitoring (Bartók et al., 2011) with a simple computationally efficient approach.

Offline oracle-efficient algorithms In Appendix A.3, we present a close variant of E2D.PMSide, namely OE2D.PMSide (Offline Estimation to Decisions for Partial Monitoring with Side Information), that efficiently reduces contextual partial monitoring to *offline* estimation. Our OE2D.PMSide generalizes FALCON (Simchi-Levi & Xu, 2022) and its linear version (Xu & Zeevi, 2020). Similar to E2D.PMSide, OE2D.PMSide has two versions, one that employs Alg. 2 for the globally observable setting, the other employs Alg. 3 for the locally observable setting, which enjoys $O(T^{2/3})$ and $O(\sqrt{T})$ regret in the respective settings. Compared with E2D.PMSide that makes $O(T)$ calls to the online outcome estimation oracle, OE2D.PMSide only makes $O(\log T)$ calls to the offline outcome estimation oracle, making it suitable for applications with low computational cost requirements.

4 Empirical studies

We evaluate SquareCB.PMSide in two settings we discussed in Section 1 (Examples 1 & 2): LLM routing and contextual label-efficient prediction game.

LLM routing. We evaluate the routing problem on a cached version of the BoolQ benchmark containing 12,648 valid responses from both models. We use `Qwen2.5-0.5B-Instruct` as the weak model and `Llama-2-13B-Chat` as the strong model. The routing outcome records whether their cached responses disagree. Each prompt is represented by a 138-dimensional context comprising 10



(a) LLM routing: Total cost vs ℓ_{01}

(b) LE game: Total cost vs c

Figure 1: Comparison of algorithms for and LLM routing and Label efficient prediction game

weak-model uncertainty statistics, 64 PCA components of its hidden states, and 64 PCA components of a context-free prompt embedding.

We compare linear-logistic CBPSide (Bartók et al., 2012a; An et al., 2026), SquareCB.PMSide (ours), ETC (Explore-Then-Commit adapted for apple tasting) (Garivier et al., 2016) with both linear logistic regression and a nonlinear 15-leaf histogram gradient-boosting (HGB) model, Bayesian-logistic PG-TS (Grant & Leslie, 2021), and a random policy matched to the routing rate of SquareCB.PMSide with gradient boosting. We consider nine values of ℓ_{01} between 1.8 and 3.3 and evaluate each configuration over 20 paired random permutations. Figure 1a reports the mean realized total cost. For each ℓ_{01}, we select a hyperparameter multiplier $s \in \{0.1, 0.3, 1, 3, 10\}$ separately for each algorithm: s scales the confidence radius in CBPSide, the exploration parameter (κ) in both SquareCB.PMSide variants, initial exploration budget ($s \cdot T^{2/3}$) in ETC, and the prior standard deviation in PG-TS. Further implementation details and results are provided in Appendix A.4.

Across the evaluated loss settings, both the linear-logistic and nonlinear HGB instantiations of SquareCB.PMSide achieve lower mean realized total cost than CBPSide, while the nonlinear HGB variant also outperforms PG-TS. These results illustrate a practical advantage of the oracle-efficient SquareCB.PMSide formulation: it can incorporate flexible nonlinear prediction models without modifying the surrounding routing policy. Furthermore, generally ETC is expected to perform worse due to $O(T^{2/3})$ regret, however here due to limited test data size 12,648 compared to larger dimension 138, it performs competitively. See Figure 1a.

Contextual label-efficient prediction. We conduct an analogous study for the contextual label-efficient prediction game. Local observability requires the expert-query cost to satisfy $c < 0.5$; accordingly, we evaluate five values of c in the range $0.1 \le c < 0.5$. Because PG-TS is designed specifically for apple-tasting games and is thus inapplicable to this problem, we replace it with RandCBPSide (Heuillet et al., 2024). We generate the data synthetically from a linear logistic model with 10-dimensional context vectors. Further, details are provided in Appendix A.4. Our method consistently outperforms ETC and CBPSide across the evaluated settings and remains competitive with RandCBPSide, outperforming it for some of the expert-query costs. Moreover, RandCBPSide requires comparatively expensive confidence-width calculations, whereas our method is computationally cheaper. See Figure 1b.

5 Conclusion and Future Work

Our work gives the first efficient algorithms for stochastic contextual partial monitoring with optimal regret with general function approximation. Experiments on LLM routing and a modified label-efficient game support these guarantees. Our analysis is restricted to non-degenerate games. Lemma A.10, which underlies the $O(1/\gamma)$ bound for locally observable games, relies on non-degeneracy. We conjecture that our SquareCB.PMSide might work for degenerate games as well with minor modification. However, their analysis need a careful approach which is left to future work. Another interesting question is, can our algorithm be generalized to the setting when the loss matrix and the feedback matrix is context-dependent? This seems practically relevant for e.g. LLM routing applications, since the cost of wrongly following the weak model may depend on the importance of the prompt. Furthermore, extending our work to linear partial monitoring (Kirschner et al., 2023) and partial monitoring with infinite outcomes (Lattimore, 2022) are also interesting future directions.

AI USE STATEMENT

In this work, we used generative AI tools for coding for experiments, and polishing the writing. We have not used generative AI tools for writing proof or related works, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops learning algorithms for routing queries among large language models. Such methods may reduce inference costs and computational resource consumption while improving access to capable models. However, routing decisions may propagate or amplify biases present in the underlying models and data, potentially producing unequal performance across demographic groups, languages, or application domains. Furthermore, queries may contain sensitive information, particularly when routed to externally hosted models. Our methods optimize a specified performance objective and do not independently guarantee fairness, privacy, reliability, or safety. Practical deployments should therefore incorporate appropriate data protection, subgroup evaluation, human oversight, and fallback mechanisms.

REPRODUCIBILITY STATEMENT

We provide complete definitions, assumptions, algorithm descriptions, and regret analyses in the main paper and appendices. The experimental section reports the datasets, preprocessing procedures, model configurations, baselines, evaluation metrics, and hyperparameters used in our study. We will release the implementation, experiment configurations, and evaluation scripts later to facilitate reproduction of our results, subject to the licenses and access conditions of the underlying datasets and language models. We also report results across multiple random seeds to account for experimental variability.

REFERENCES

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Byeongwoo An, Kapilan Balagopalan, Hyowon Wi, Sehwa Jeong, Jiun Jeong, Kyoungseok Jang, Noseong Park, Gi-Soo Kim, and Kwang-Sung Jun. Efficient algorithms for contextual apple tasting with log-loss. In *ICML 2026 DEMO workshop*, 2026.

Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pp. 305–319. Springer, 2012.

Gábor Bartók et al. Partial monitoring with side information. In *In Proc. ALT*, 2012a.

Gábor Bartók et al. An adaptive algorithm for finite stochastic partial monitoring. *In Proc. ICML*, 2012b.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In Sham M. Kakade and Ulrike von Luxburg (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 133–154, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL `https://proceedings.mlr.press/v19/bartok11a.html`.

Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The Statistical Complexity of Interactive Decision Making. *CoRR*, abs/2112.1, 2021.

Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35:35404–35417, 2022.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29, pp. 784–792. Curran Associates, Inc., 2016.

James A Grant and David S Leslie. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification. *arXiv preprint arXiv:2109.14412*, 2021.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic Regret Algorithms for Online Convex Optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

David Helmbold et al. Some label efficient learning results. In *In Proc. CoLT*, 1997.

David P Helmbold, Nicholas Littlestone, and Philip M Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

Maxime Heuillet, Ola Ahmad, and Audrey Durand. Randomized confidence bounds for stochastic partial monitoring, 2024.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision-making: Algorithms, regret bounds and applications. *JMLR*, 2023.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pp. 1141–1154. PMLR, 2015.

Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *Conference on Learning Theory*, pp. 1547–1575. PMLR, 2022.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pp. 2488–2515. PMLR, 2020.

David A Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2 edition, 2017. ISBN 978-1-4704-2962-1.

Haipeng Luo. Lecture 10: Partial monitoring — lower bounds; reinforcement learning — episodic Markov decision processes, Bellman equations, Q-learning. Lecture notes, CSCI 699: Theoretical Machine Learning, University of Southern California, 2019. Available at `https://haipeng-luo.net/courses/CSCI699_2019/lecture10.pdf`, accessed 24 September 2026.

Hao Qin and Chicheng Zhang. Optimal and efficient contextual combinatorial semi-bandits with general function approximation, 2026a. URL `https://arxiv.org/abs/2607.13686`.

Hao Qin and Chicheng Zhang. Taming the monster every context: Complexity measure and unified framework for offline-oracle efficient contextual bandits, 2026b. URL `https://arxiv.org/abs/2602.09456`.

David Sculley. Practical learning from one-sided feedback. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 609–618, 2007.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Adish Singla, Ian Lienert, Gábor Bartók, and Andreas Krause. Contextual procurement in online crowdsourcing markets. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pp. 58–59, 2014.

A. Tsybakov. Introduction to nonparametric estimation. In *Springer Series in Statistics*, 2008. URL `https://api.semanticscholar.org/CorpusID:42933599`.

Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 51–60, 1995.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Mengxiao Zhang, Yuheng Zhang, Olga Vrousgou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. *Advances in Neural Information Processing Systems*, 36: 30592–30617, 2023.

Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, pp. 27574–27590. PMLR, 2022.

Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 27428–27453, 2022.

A Appendix

Contents

A.1 Online and offline estimation oracles 14
A.1.1 Online Oracles 14
A.1.2 Offline estimation oracle 16
A.2 Analysis of E2D.PMSide 18
A.2.1 Proof of Theorem 3.1 18
A.2.2 Proofs for Algorithm 2 (ϵ-Greedy.PMSide) 19
A.2.3 Proofs for Algorithm 3 (SquareCB.PMSide) 21
A.2.4 Non-degeneracy assumption 25
A.3 Contextual partial monitoring: reduction to offline regression 25
A.3.1 OE2D.PMSide 25
A.3.2 Regret analysis 27
A.4 Experiments 32
A.5 Algorithm for $\boldsymbol{m}Lq$-decreasing tree construction 37
A.6 Auxiliary Lemmata 38

A.1 Online and offline estimation oracles

A.1.1 Online Oracles

In this section, we provide examples for the online estimation oracles $\mathrm{Alg}_{\mathrm{On}}$ and prove their bounds for both finite and infinite model classes.

Lemma A.1 (Finite-class online Hellinger guarantee via Vovk's algorithm)**.** Suppose that the candidate conditional outcome-model class

$$\mathcal{Q} \subseteq \left\{q : \mathcal{X} \to \Delta(d)\right\}$$

is finite and realizable, so that $q^* \in \mathcal{Q}$ and the oracles are allowed to output predictors in $\mathrm{co}(\mathcal{Q})$. Then there exists a contextual online estimation oracle $\mathrm{Alg}_{\mathrm{On}}$ such that,

$$\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta) = 2\log\left(\frac{|\mathcal{Q}|}{\delta}\right).$$

The contextual online estimation oracle can be instantiated by applying Vovk's (Vovk, 1995) aggregating algorithm to the conditional outcome models in $\mathcal{Q}$.

Proof. Define $\mathcal{G} \subseteq \Delta(m)$ as a set of functions $g_q : \mathcal{X} \times [k] \to \Delta(m)$ induced by q, formally

$$\mathcal{G} = \{g_q : g_q(\sigma \mid x, a) = S_a q(x)[\sigma] \;\; \forall q \in \mathcal{Q}\}$$

Hence,

$$|\mathcal{G}| \le |\mathcal{Q}|$$

Specifically,

$$\begin{aligned} g_{q^*}(\sigma \mid x_t, A_t) &= \mathbb{P}\left(\sigma_t = \sigma \mid x_t, A_t\right) \\ &= S_{A_t} q^*(x_t)[\sigma] \end{aligned}$$

Vovk's algorithm calculate the $g_{\hat{q}_t}$ in the following way.

Let,

$$\begin{aligned} \ell^{(t)}_{\log}(g_q) &:= \log\left(\frac{1}{g_q(\sigma_t \mid x_t, A_t)}\right) \\ &= \log\left(\frac{1}{\left(S_{A_t} q(x_t)\right)[\sigma_t]}\right). \end{aligned}$$

Then $\forall g_q \in \mathcal{G}$, the algorithm calculates a probability simplex $b_t \in \Delta(|\mathcal{G}|)$ as follows.

$$b_t(g_q) \propto \exp\left(-\sum_{s=1}^{t-1} \ell^{(s)}_{\log}(g_q)\right)$$

Estimate $g_{\hat{q}_t}$

$$g_{\hat{q}_t}(\cdot) = \mathbb{E}_{g \sim b_t}\left[g(\cdot)\right]$$

Let the corresponding $\hat{q}_t$ such that,

$$\hat{q}_t(\cdot) = \mathbb{E}_{g \sim b_t}\left[q(\cdot)\right].$$

Then Vovk's algorithm guarantees,

$$\sum_{t=1}^{T} \log \frac{1}{g_{\hat{q}_t}(\sigma_t \mid x_t, a_t)} - \sum_{t=1}^{T} \log \frac{1}{g_{q^*}(\sigma_t \mid x_t, a_t)} \le \log |\mathcal{G}|$$

Then, by applying Lemma A.14 from Foster et al. (2021), the output estimator of Vovk's aggregating algorithm $g_{\hat{q}_t}$ satisfies, with probability of at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D^2_{\mathrm{H}}\left(g_{\hat{q}_t}(\cdot \mid x_t, A_t) \mid g_{q^*}(\cdot \mid x_t, A_t)\right)\right] \le 2\log\frac{|\mathcal{G}|}{\delta} \tag{4}$$

transforming this to our case, with probability at least $1-\delta$, $\hat{q}_t$ satisfies,

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\left(S_a\hat{q}_t(x_t) \mid S_a q^*(x_t)\right)\right] \le 2\log\frac{|\mathcal{Q}|}{\delta}.$$

Remark 1. Define filtration $\mathcal{H}$ as,

$$\mathcal{H}_{t-1} = \sigma\left(x_1, A_1, \sigma_1, \cdots, x_{t-1}, A_{t-1}, \sigma_{t-1}\right)$$

Equation equation 4 also implies that, if $x_1, \cdots, x_T \overset{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and $p_t, \hat{q}_t$ are $\mathcal{H}_{t-1}$ measurable, then with probability of at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{x\sim\mathcal{D}_x, a\sim p_t(\cdot|x)}\left[D^2_{\mathrm{H}}\left(S_a\hat{q}_t(x) \mid S_a q^*(x)\right)\right] \le 2\log\frac{|\mathcal{Q}|}{\delta}.$$

□

Lemma A.2 (Infinite-class online Hellinger guarantee via contextual smoothed aggregation)**.** Suppose that the candidate conditional outcome-model class is,

$$\mathcal{Q} \subseteq \left\{q : \mathcal{X} \to \Delta(d)\right\},$$

is possibly infinite but has finite covering numbers under the uniform Hellinger metric and that the model is realizable $q^* \in \mathcal{Q}$ and the oracles are allowed to output predictors in $\mathrm{co}(\mathcal{Q})$.

Define $\mathcal{G} \subseteq \Delta(m)$ as a set of functions $g_q : \mathcal{X} \times [k] \to \Delta(m)$ induced by q, formally

$$\mathcal{G} = \{g_q : g_q(\sigma \mid x, a) = S_a q(x)[\sigma] \;\; q \in \mathcal{Q}\}$$

Moreover, let us define an ϵ-cover $\mathcal{J}$, such that,

$$\forall g_q \in \mathcal{G} \quad \exists g'_q \in \mathcal{J} \text{ such that } \sup_{x\in\mathcal{X}, a\in[k]} D^2_{\mathrm{H}}\left(g'_q(\cdot \mid x, a) \mid g_q(\cdot \mid x, a)\right) \le \epsilon^2$$

Furthermore, let $\mathcal{N}(\mathcal{G}, \epsilon)$ denote the size of the smallest such cover. Define,

$$\mathrm{est}(\mathcal{G}, T) := \inf_{\epsilon\ge 0}\{\log(\mathcal{N}(\mathcal{G}, \epsilon)) + \epsilon^2 T\}$$

Then there exists a contextual online estimation oracle $\mathrm{Alg}_{\mathrm{On}}$ such that,

$$\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q}, T, \delta) \le O\left(\log(2m^2T)\cdot\mathrm{est}(\mathcal{G}, T) + \log^2(2m^2T)\log\frac{1}{\delta}\right).$$

The contextual online estimation oracle can be instantiated by applying an online aggregation algorithm Vovk (1995); Foster et al. (2021) to a suitably discretized and smoothed approximation of the parameter space $\mathcal{Q}$.

Proof. The regularity assumptions required in Assumption A.2 of Foster et al. (2021) are already satisfied by setting $B = \max\{m, e\}$

1. Number of possible outcome for a context is upper bounded by $m \le B$
2. $g_q(\sigma \mid x, a)$ is a probability simplex ($1 \le B$).

Smoothed Vovk's algorithm calculate the $g_{\hat{q}_t}$ in the following way, let,

$$h(\sigma \mid x, a) = (1-\alpha)g_q(\sigma \mid x, a) + \frac{\alpha}{m}$$

then $\forall g_q \in \mathcal{J}$ calculate the a probability simplex $b_t \in \Delta(|\mathcal{J}|)$ as follows

$$b_t(g_q) \propto \exp\left(-\sum_{s=1}^{t-1} \ell_{\log}^{(s)}(h)\right)$$

Estimate $\hat{g}_t$

$$g_{\hat{q}_t} = \mathbb{E}_{g_q \sim b_t}\left[g_q\right]$$

Hence,

$$\hat{q}_t = \mathbb{E}_{q \sim b_t}\left[q\right]$$

Then by applying Lemma A.16 from Foster et al. (2021), with probability of at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D_{\mathrm{H}}^2\left(g_{\hat{q}_t}(\cdot \mid x_t, A_t) \mid g_{q^*}(\cdot \mid x_t, A_t)\right)\right] \le O\left(\log(2m^2T)\cdot \mathrm{est}(\mathcal{G}, T) + \log^2(2m^2T)\log\frac{1}{\delta}\right) \tag{5}$$

Let $\hat{q}_t \in \mathcal{Q}$ be the corresponding solution, with probability of at least $1-\delta$, $\hat{q}_t$ satisfies,

$$\sum_{t=1}^{T} \mathbb{E}_{a \sim p_t}\left[D_{\mathrm{H}}^2\left(S_a\hat{q}_t(x_t) \mid S_a q^*(x_t)\right)\right] \le O\left(\log(2m^2T)\cdot \mathrm{est}(\mathcal{G}, T) + \log^2(2m^2T)\log\frac{1}{\delta}\right).$$

Remark 2. Define filtration $\mathcal{H}$ as,

$$\mathcal{H}_{t-1} = \sigma\left(x_1, A_1, \sigma_1, \cdots, x_{t-1}, A_{t-1}, \sigma_{t-1}\right)$$

Equation equation 4 also implies that, if $x_1, \cdots, x_T \overset{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and $p_t, \hat{q}_t$ are $\mathcal{H}_{t-1}$ measurable, then with probability of at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{x \sim \mathcal{D}_x, a \sim p_t(\cdot|x)}\left[D_{\mathrm{H}}^2\left(S_a\hat{q}_t(x) \mid S_a q^*(x)\right)\right] \le O\left(\log(2m^2T)\cdot \mathrm{est}(\mathcal{G}, T) + \log^2(2m^2T)\log\frac{1}{\delta}\right).$$

□

A.1.2 Offline estimation oracle

In this section, first we define *offline estimation oracle* and we provide examples for the offline estimation oracles $\mathrm{Alg}_{\mathrm{Off}}$ and prove their estimation bounds.

Definition A.1 (Offline estimation oracle)**.** An *offline estimation oracle* $\mathrm{Alg}_{\mathrm{Off}}$ is an algorithm that estimates the unknown parameter $q^* \in \mathcal{Q}$ from the observed data in a batch fashion. That is, after observing t rounds, given the feedback symbols collected so far, the oracle produces an estimate

$$\hat{q} = \mathrm{Alg}_{\mathrm{Off}}\left(\left\{(A_s, x_s, \sigma_s)\right\}_{s=1}^{t}\right).$$

Assumption A.1 (Hellinger-closeness of the offline estimation oracle)**.** For any $t \in \mathbb{N}$, let $x_1, \ldots, x_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and let $\hat{q}$ denote the estimate returned by $\mathrm{Alg}_{\mathrm{Off}}$ on the resulting sample. Define $\mathrm{Reg}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t)$, such that

$$\mathbb{E}_{x \sim \mathcal{D}_x,\, a \sim p(\cdot|x)}\left[D_{\mathrm{H}}^2\big(S_a\hat{q}(x) \,\|\, S_a q^*(x)\big)\right] = \frac{\mathrm{Reg}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t)}{t}$$

where p denotes the behavioral policy used to collect the sample, and the expectation is taken over a fresh draw (x, a), conditionally on $\hat{q}$. We assume that the offline estimation oracle $\mathrm{Alg}_{\mathrm{Off}}$ guarantees, for any $\delta \in (0,1)$, that with probability at least $1-\delta$,

$$\mathrm{Reg}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t) \le \overline{\mathrm{Reg}}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t, \delta),$$

where $\overline{\mathrm{Reg}}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t, \delta)$ is a known, deterministic upper bound.

Having established in Appendix A.1.1 the existence of online oracles with favorable guarantees for both finite and infinite model classes, we can always obtain offline oracles for both by online-to-batch conversion. Beyond this generic construction, we also show that the conditional maximum-likelihood estimator satisfies Assumption A.1 with a favorable bound $\overline{\mathrm{Reg}}_{\mathrm{Off}}^{\mathrm{H}}(\mathcal{Q}, t, \delta)$.

Lemma A.3 (Existence of a contextual offline estimation oracle for a finite class)**.** Suppose that the candidate conditional outcome-model class

$$\mathcal{Q} \subseteq \left\{q : \mathcal{X} \to \Delta(d)\right\}$$

is finite and realizable, so that $q^* \in \mathcal{Q}$. Then there exists a contextual offline estimation oracle $\mathrm{Alg}_{\mathrm{Off}}$ such that Assumption A.1 is satisfied with

$$\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, t, \delta) = 2\log\left(\frac{|\mathcal{Q}|}{\delta}\right).$$

The contextual offline estimation oracle can be instantiated using the conditional maximum-likelihood estimator

$$\hat{q}_t \in \arg\max_{q \in \mathcal{Q}} \sum_{s=1}^{t} \log\Big(\big(S_{A_s} q(x_s)\big)[\sigma_s]\Big).$$

Proof. Let

$$z_s = (x_s, A_s) \qquad \text{(Where } x_s \sim \mathcal{X}\text{(i.i.d) and } A_s \sim p_s\left(z_{1:s-1}, y_{1:s-1}, x_s\right)\text{)}$$
$$y_s = \sigma_s$$

Hence,

$$z_s \sim \mathcal{D}_s\left(z_{1:s-1}, y_{1:s-1}\right) \qquad \text{(for some distribution } \mathcal{D}_s\text{)}$$

Define

$$\begin{aligned} f^*(z_s, y_s) &= \mathbb{P}(y_s \mid z_s) \\ &= \mathbb{P}(\sigma_s \mid A_s, x_s) \\ &= \mathbb{P}(\sigma_s \mid S_{A_s} q^*(x_s)) \\ &= S_{A_s} q^*(x_s)[\sigma] \end{aligned}$$

Furthermore, define

$$\mathcal{F} = \{f : f(z, y) = S_a q(x)[\sigma] \;\; \forall q \in \mathcal{Q}\}$$

Hence,

$$|\mathcal{F}| \le |\mathcal{Q}|$$

Following the work of Agarwal et al. (2020) (section E)

$$\begin{aligned} \hat{f} &:= \arg\max_{f \in \mathcal{F}} \sum_{s=1}^{t} \log f(z_s, y_s) \\ &= \arg\max_{q} \sum_{s=1}^{t} \mathbb{P}(\sigma_s \mid S_{A_s} q(x_s)) \end{aligned}$$

Therefore, let

$$\begin{aligned} \hat{q}_t &= \arg\max_{q} \sum_{s=1}^{t} \log\Big(\mathbb{P}\left(\sigma_s \mid x_s, A_s, q\right)\Big) \\ &= \arg\max_{q} \sum_{s=1}^{t} \log\left(S_{A_s} \cdot q(x_s)[\sigma_s]\right) \end{aligned}$$

then,

$$\hat{f}(z, y) = S_a \hat{q}_t(x)$$

By Theorem 21 of Agarwal et al. (2020), this maximum likelihood solution, satisfies the following,

$$\sum_{s=1}^{t} \mathbb{E}_{z \sim D_s} \left\| \hat{f}(z, \cdot) - f^*(z, \cdot) \right\|^2_{\mathrm{TV}} \le 2\log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

Although Theorem 21 is stated in terms of squared total variation distance, Remark 23 observes that its proof directly establishes the corresponding squared-Hellinger bound.

$$\sum_{s=1}^{t} \mathbb{E}_{z\sim D_s}\left[D^2_{\mathrm{H}}\left(\hat{f}(z,\cdot)\middle|f^*(z,\cdot)\right)\right] \le 2\log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

Therefore, in our case

$$\sum_{s=1}^{t} \mathbb{E}_{x\sim\mathcal{X},a\sim p_s}\left[D^2_{\mathrm{H}}\left(S_a\cdot\hat{q}_t(x)\middle|S_a\cdot q^*(x)\right)\right] \le 2\log\left(\frac{|\mathcal{Q}|}{\delta}\right)$$

□

A.2 Analysis of E2D.PMSide

A.2.1 Proof of Theorem 3.1

Theorem 3.1 (Regret of E2D.PMSide)**.** Suppose $\mathrm{Alg}_{\mathrm{On}}$ has a Hellinger regret bound $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q},T,\delta)$. Fix $\gamma>0$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\mathrm{Reg}_T \le T\cdot\max_{t=1}^{T}\mathcal{V}_\gamma\big(\hat{q}_t(x_t),p_t\big)+\gamma\cdot\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{On}}(\mathcal{Q},T,\delta)\,.$$

Proof. Let

$$\lambda^*_{x_t}=\arg\min_{p'\in\Delta(\Pi)}(p')^\top \boldsymbol{L}q^*(x_t) \qquad \text{(The best policy for context } x_t\text{)}$$

$$\begin{aligned}\mathrm{Reg}_T &= \sum_{t=1}^{T} p_t^\top \boldsymbol{L}q^*(x_t)-(\lambda^*_{x_t})^\top \boldsymbol{L}q^*(x_t)\\ &= \sum_{t=1}^{T} p_t^\top \boldsymbol{L}\hat{q}_t(x_t)-(\lambda^*_{x_t})^\top \boldsymbol{L}\hat{q}_t(x_t)+\big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}(\hat{q}_t(x_t)-q^*(x_t))\big)\end{aligned}$$

$$\begin{aligned}&\big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}\hat{q}_t(x_t)-(\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}q^*(x_t)\big)^2\\ =\;& \big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}(\hat{q}_t(x_t)-q^*(x_t))\big)^2\\ =\;& \frac{\big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}(\hat{q}_t(x_t)-q^*(x_t))\big)^2}{\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]}\cdot\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right)\\ \le\;& \max_{q_1,q_2\in\Delta(d)}\frac{\big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}(q_1-q_2)\big)^2}{\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_aq_1\,\big|\,S_aq_2\big)\right]}\cdot\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right).\end{aligned}$$

Therefore,

$$\big((\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}\hat{q}_t(x_t)-(\lambda^*_{x_t}-p_t)^\top \boldsymbol{L}q^*(x_t)\big)^2 \le \mathrm{Coverage}(p_t,\lambda^*_{x_t})\cdot\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right).$$

$$\begin{aligned}\mathrm{Reg}_T &\le \sum_{t=1}^{T} p_t^\top \boldsymbol{L}\hat{q}_t(x_t)-(\lambda^*_{x_t})^\top \boldsymbol{L}\hat{q}_t(x_t)+\sqrt{\mathrm{Coverage}(p_t,\lambda^*_{x_t})\cdot\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right)}\\ &\le \sum_{t=1}^{T} p_t^\top \boldsymbol{L}\hat{q}_t(x_t)-(\lambda^*_{x_t})^\top \boldsymbol{L}\hat{q}_t(x_t)+\frac{1}{\gamma}\mathrm{Coverage}_\epsilon(p_t,\lambda^*_{x_t})+\gamma\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right)\\ &\qquad\qquad\text{(AM-GM)}\\ &\le \sum_{t=1}^{T}\sup_{\lambda\in\Delta(\Pi)}\left[p_t^\top \boldsymbol{L}\hat{q}_t(x_t)-\lambda^\top \boldsymbol{L}\hat{q}_t(x_t)+\frac{1}{\gamma}\mathrm{Coverage}(p_t,\lambda)\right]+\sum_{t=1}^{T}\gamma\left(\mathbb{E}_{a\sim p_t}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}_t(x_t)\,\big|\,S_aq^*(x_t)\big)\right]\right)\end{aligned}$$

$$\leq \sum_{t=1}^{T} \sup_{\lambda \in \Delta(\Pi)} \left[p_t^\top \boldsymbol{L} \hat{q}_t(x_t) - \lambda^\top \boldsymbol{L} \hat{q}_t(x_t) + \frac{1}{\gamma} \cdot \mathrm{Coverage}(p_t, \lambda) \right] + \gamma \left(\mathrm{Reg}_{\mathrm{On}}^{\mathrm{H}}(\mathcal{Q}, T) \right).$$

Therefore the regret can be bounded as,

$$\begin{aligned}
\mathrm{Reg}_T &\leq \sum_{t=1}^{T} \sup_{\lambda \in \Delta(\Pi)} \left[p_t^\top \boldsymbol{L} \hat{q}_t(x_t) - \lambda^\top \boldsymbol{L} \hat{q}_t(x_t) + \frac{1}{\gamma} \cdot \mathrm{Coverage}(p_t, \lambda) \right] + \gamma \left(\mathrm{Reg}_{\mathrm{On}}^{\mathrm{H}}(\mathcal{Q}, T) \right) \\
&= \sum_{t=1}^{T} \mathcal{V}_\gamma\big(\hat{q}_t(x_t), p_t\big) + \gamma \left(\mathrm{Reg}_{\mathrm{On}}^{\mathrm{H}}(\mathcal{Q}, T) \right) \\
&\leq T \cdot \max_{t=1}^{T} \mathcal{V}_\gamma\big(\hat{q}_t(x_t), p_t\big) + \gamma \left(\mathrm{Reg}_{\mathrm{On}}^{\mathrm{H}}(\mathcal{Q}, T) \right).
\end{aligned}$$

Hence, with probability of at least $1 - \delta$,

$$\mathrm{Reg}_T \leq T \cdot \max_{t=1}^{T} \mathcal{V}_\gamma\big(\hat{q}_t(x_t), p_t\big) + \gamma \cdot \overline{\mathrm{Reg}}_{\mathrm{On}}^{\mathrm{H}}(\mathcal{Q}, T, \delta).$$

□

A.2.2 Proofs for Algorithm 2 (ϵ-Greedy.PMSide)

First of all, we simplify the coverage as follows,

Lemma A.4 (Decomposition of coverage about a reference action)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be an arbitrary partial monitoring game and let $p, \lambda \in \Delta(k)$. Then for every reference action $a_0 \in [k]$,

$$\mathrm{Coverage}(p, \lambda) \leq \underbrace{8 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\sum_{b=1}^{k} \lambda_b \left(L_b - L_{a_0} \right) (q_1 - q_2) \right)^2}{\sum_{a=1}^{k} p_a \| \boldsymbol{S}_a (q_1 - q_2) \|_1^2}}_{=: \mathrm{Cov}_1(p, \lambda, a_0)} + \underbrace{8 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\sum_{b=1}^{k} p_b \left(L_b - L_{a_0} \right) (q_1 - q_2) \right)^2}{\sum_{a=1}^{k} p_a \| \boldsymbol{S}_a (q_1 - q_2) \|_1^2}}_{=: \mathrm{Cov}_2(p, a_0)},$$

with the convention $0/0 = 0$.

Furthermore, to abstract the repetitive proof steps, we use the following lemma.

Lemma A.5 (Witness bound for the coverage terms)**.** Let $a_0 \in [k]$ and let $\mu, p \in \Delta(k)$. Suppose that for every $b \in [k]$ there is a function $h_{(b,a_0)} : [k] \times \Sigma \to \mathbb{R}$ such that

$$L_{bo} - L_{a_0 o} = \sum_{c=1}^{k} h_{(b,a_0)}(c, \Phi_{co}), \qquad \text{for all } o \in [d],$$

and write $\| h_{(b,a_0)}(c, \cdot) \|_\infty := \max_{\sigma \in \Sigma} |h_{(b,a_0)}(c, \sigma)|$. Assume $p_c > 0$ whenever $h_{(b,a_0)}(c, \cdot) \not\equiv 0$ for some b. Then

$$\sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\sum_{b=1}^{k} \mu_b \left(L_b - L_{a_0} \right) (q_1 - q_2) \right)^2}{\sum_{a=1}^{k} p_a \| \boldsymbol{S}_a (q_1 - q_2) \|_1^2} \leq \left(\sum_{b=1}^{k} \mu_b \sqrt{\sum_{c=1}^{k} \frac{\| h_{(b,a_0)}(c, \cdot) \|_\infty^2}{p_c}} \right)^2 \leq \sum_{b=1}^{k} \mu_b \sum_{c=1}^{k} \frac{\| h_{(b,a_0)}(c, \cdot) \|_\infty^2}{p_c},$$

with the convention $0/0 = 0$.

The proofs of Lemma A.4 and Lemma A.5 are deferred to Appendix A.6.

Lemma 3.1 (Guarantee for Alg. 2 for globally observable games)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be a globally observable partial monitoring game in the sense of Definition 2.2, and set $\beta := \max_{(a,b) \in [k] \times [k]} \| w_{(a,b)} \|_\infty$. We have for any $\hat{q} \in \Delta(d)$,

$$\mathcal{V}_\gamma\big(\hat{q}, p\big) \leq \max_{\lambda \in \Delta(\Pi)} p^T \boldsymbol{L} \hat{q} - \lambda^T \boldsymbol{L} \hat{q} + \frac{16 \beta^2}{\gamma} \sum_{c=1}^{k} \frac{1}{p_c}.$$

Thus, for every $\gamma > 16k^2\beta^2$, and every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \le O\left(k\beta\sqrt{\frac{1}{\gamma}}\right).$$

Proof. By Lemma A.4, we already have, for any $a_0 \in [k]$,

$$\mathrm{Coverage}(p,\lambda) \le \underbrace{8\max_{q_1,q_2}\frac{\left(\sum_{b=1}^k \lambda_b(L_b - L_{a_0})(q_1-q_2)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_1(p,\lambda,a_0)} + \underbrace{8\max_{q_1,q_2}\frac{\left(\sum_{b=1}^k p_b(L_b - L_{a_0})(q_1-q_2)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_2(p,a_0)}.$$

Furthermore, By Lemma A.5, by setting, $\mu_b \leftarrow \lambda_b$ and $h_{b,a_0} \leftarrow w_e = w_{(b,a_0)}$,

$$\begin{aligned}
\mathrm{Cov}_1(p,\lambda,a) &\le 8\sum_{b=1}^k\sum_{c=1}^k \frac{\lambda_b}{p_c}\|w_{(b,a_0)}(c,\cdot)\|_\infty^2 \\
&\le 8\max_{(a,b)\in[k]\times[k]}\|w_{(a,b)}\|_\infty^2 \sum_{b=1}^k\sum_{c=1}^k \frac{\lambda_b}{p_c} \\
&\le 8\max_{(a,b)\in[k]\times[k]}\|w_{(a,b)}\|_\infty^2 \sum_{c=1}^k \frac{1}{p_c}
\end{aligned}$$

We can do the same thing to $\mathrm{Cov}_2(p,a_0)$ as well. Hence,

$$\begin{aligned}
\mathrm{Coverage}(p,\lambda) &\le 16\max_{(a,b)\in[k]\times[k]}\|w_{(a,b)}\|_\infty^2 \sum_{c=1}^k \frac{1}{p_c} \\
&= 16\beta^2\sum_{c=1}^k \frac{1}{p_c} && (\beta = \max_e \|w_e\|_\infty^2)
\end{aligned}$$

This finishes the proof of the first part of the lemma.

Let $\hat{a}$ be the empirical best arm based on our current estimate $\hat{q}$. That is,

$$\hat{a} := \arg\min_{a\in\Pi} e_a^\top \boldsymbol{L}\hat{q}(x).$$

Set,

$$p = (1-\epsilon)e_{\hat{a}} + \epsilon\frac{\mathbf{1}}{k},$$

so that $p_c \ge \frac{\epsilon}{k}$ for every $c \in [k]$, and hence

$$\sum_{c=1}^k \frac{1}{p_c} \le \frac{k^2}{\epsilon}.$$

$$\begin{aligned}
\mathcal{V}_\gamma(\hat{q}(x),p) &= \sup_{\lambda\in\Delta(\Pi)}\left[p^\top\boldsymbol{L}\hat{q}(x) - \lambda^\top\boldsymbol{L}\hat{q}(x) + \frac{1}{\gamma}\cdot\mathrm{Coverage}(p,\lambda)\right] \\
&\le p^\top\boldsymbol{L}\hat{q}(x) - e_{\hat{a}}^\top\boldsymbol{L}\hat{q}(x) + \frac{1}{\gamma}16\beta^2\sum_{c=1}^k\frac{1}{p_c} \\
&\le \left((1-\epsilon)e_{\hat{a}} + \epsilon\frac{\mathbf{1}}{k}\right)^\top\boldsymbol{L}\hat{q}(x) - e_{\hat{a}}^\top\boldsymbol{L}\hat{q}(x) + \frac{16k^2\beta^2}{\gamma\epsilon} \\
&\le \epsilon\,\tfrac{1}{k}\mathbf{1}^\top\boldsymbol{L}\hat{q}(x) + \frac{16k^2\beta^2}{\gamma\epsilon} \\
&\le \alpha + \frac{16k^2\beta^2}{\gamma\epsilon} && (\boldsymbol{L}\in[0,1]^{k\times d})
\end{aligned}$$

$$= O\left(k\beta\sqrt{\frac{1}{\gamma}}\right). \qquad \text{(by choosing } \epsilon = k\beta\sqrt{\tfrac{1}{\gamma}} < 1\text{)}$$

This finishes the proof of the second part of the lemma. □

A.2.3 Proofs for Algorithm 3 (SquareCB.PMSide)

Define, $G_{\mathcal{T}} : [k] \times \Sigma \to \mathbb{R}^{\Pi}$ $\forall a \in \Pi$ such that,

$$G_a^{\mathcal{T}}(c,\sigma) := \sum_{e \in \mathrm{path}_{\mathcal{T}}(a)} w_e(c,\sigma), \qquad a \in \Pi,$$

where w_e denotes the witness function associated with the edge e by local observability (Definition 2.3). Here it is noteworthy that, let $e = (p,q)$ then $w_e(c,\sigma) = 0$ $\forall c \notin \{p,q\}$. Therefore,

$$G_a^{\mathcal{T}}(c,\sigma) = 0 \text{ if } c \notin \mathrm{path}_{\mathcal{T}}(a). \tag{6}$$

Therefore, the local observability condition induces a sparse $G_a^{\mathcal{T}}(\cdot,\sigma)$ function, which will become handy for us to allocate probability only for the actions populating it. Furthermore, we can also show that, $\forall a \in \Pi$

$$\sum_{c\in\Pi} G_a^{\mathcal{T}}(c,\Phi_{co}) = \sum_{c\in\Pi}\sum_{e\in\mathrm{path}_{\mathcal{T}}(a)} w_e(c,\Phi_{co}) = \sum_{e\in\mathrm{path}_{\mathcal{T}}(a)}\sum_{c\in\Pi} w_e(c,\Phi_{co}) = L_{ao} - L_{\mathrm{root}_{\mathcal{T}}o}$$

Therefore,

Lemma A.6. For a non-degenerate locally observable partial monitoring game $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$, for any in-tree $\mathcal{T}$ constructed from any $\mathrm{root}_{\mathcal{T}}$,

$$\sum_{c\in\Pi}\left(\sum_{\sigma\in\Sigma} G_a^{\mathcal{T}}(c,\sigma)\cdot(S_c)_{\sigma o} - \sum_{\sigma\in\Sigma} G_b^{\mathcal{T}}(c,\sigma)\cdot(S_c)_{\sigma o}\right) = L_{ao} - L_{bo}\,.$$

Proof.

$$\begin{aligned} G_a^{\mathcal{T}}(c,\Phi_{co}) &= \sum_{\sigma\in\Sigma} G_a^{\mathcal{T}}(c,\sigma)\cdot\mathbb{1}\{\sigma = \Phi_{co}\} \\ &= \sum_{\sigma\in\Sigma} G_a^{\mathcal{T}}(c,\sigma)\cdot(S_c)_{\sigma o} \qquad \text{(by the definition of } S_c\text{)} \end{aligned}$$

Hence,

$$\begin{aligned} \sum_{c\in\Pi}\left(\sum_{\sigma\in\Sigma} G_a^{\mathcal{T}}(c,\sigma)\cdot(S_c)_{\sigma o} - \sum_{\sigma\in\Sigma} G_b^{\mathcal{T}}(c,\sigma)\cdot(S_c)_{\sigma o}\right) &= \sum_{c\in\Pi}\left(G_a^{\mathcal{T}}(c,\Phi_{co}) - G_b^{\mathcal{T}}(c,\Phi_{co})\right) \\ &= L_{ao} - L_{\mathrm{root}_{\mathcal{T}}o} - (L_{bo} - L_{\mathrm{root}_{\mathcal{T}}o}) \\ &= L_{ao} - L_{bo}. \end{aligned}$$

□

Now we relax $\mathcal{V}_\gamma\big(\hat{q},p\big)$ with sparse $G_a^{\mathcal{T}}(\cdot,\sigma)$.

Lemma 3.2. Let $\mathcal{G} = (\boldsymbol{L},\boldsymbol{\Phi})$ be a locally observable partial monitoring game in the sense of Definition 2.3. Let $\mathcal{T}$ be an in-tree of the neighborhood graph and $G^{\mathcal{T}}$ be its associated loss estimator function. For all $p \in \Delta(\Pi)$, we have

$$\begin{aligned} \mathcal{V}_\gamma\big(\hat{q}(x),p\big) &\le \max_{\lambda\in\Delta(\Pi)} p^T\boldsymbol{L}\hat{q}(x) - \lambda^T\boldsymbol{L}\hat{q}(x) + \frac{8}{\gamma}\sum_{b\in\Pi}\sum_{c\in\Pi}\left(\frac{\lambda_b}{p_c}+\frac{p_b}{p_c}\right)\|G^{\mathcal{T}}(c,\cdot)_b\|_\infty^2 \\ &\le \max_{\lambda\in\Delta(\Pi)} p^T\boldsymbol{L}\hat{q}(x) - \lambda^T\boldsymbol{L}\hat{q}(x) + \frac{32}{\gamma}\max_{e\in\mathcal{E}}\|w_e\|_\infty^2 \sum_{(b,c):c\in\mathrm{path}_{\mathcal{T}}(b)}\left(\frac{\lambda_b}{p_c}+\frac{p_b}{p_c}\right). \end{aligned}$$

Proof. For an in-tree $\mathcal{T} \subseteq \mathcal{E}$ with an arbitrary arm $\mathrm{root}_{\mathcal{T}} \in \Pi$ as it's root, by Lemma A.4, we already have,

$$\mathrm{Coverage}(p,\lambda) \le 8 \underbrace{\sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \lambda_b (L_b - L_{\mathrm{root}_{\mathcal{T}}})(q_1-q_2)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_1(p,\lambda,\mathrm{root}_{\mathcal{T}})} + 8 \underbrace{\frac{\left(\sum_{b=1}^k p_b (L_b - L_{\mathrm{root}_{\mathcal{T}}})(q_1-q_2)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_2(p,\mathrm{root}_{\mathcal{T}})}.$$

Moreover, we have $p, \lambda \in \Delta(\Pi)$, hence all the terms with $p_a, \lambda_a \notin [k] \setminus \Pi$ will be zero. Therefore,

$$\mathrm{Coverage}(p,\lambda) \le 8 \underbrace{\sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b\in\Pi} \lambda_b (L_b - L_{\mathrm{root}_{\mathcal{T}}})(q_1-q_2)\right)^2}{\sum_{a\in\Pi} p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_1(p,\lambda,\mathrm{root}_{\mathcal{T}})} + 8 \underbrace{\frac{\left(\sum_{b\in\Pi} p_b (L_b - L_{\mathrm{root}_{\mathcal{T}}})(q_1-q_2)\right)^2}{\sum_{a\in\Pi} p_a \|S_a(q_1-q_2)\|_1^2}}_{\mathrm{Cov}_2(p,\mathrm{root}_{\mathcal{T}})}.$$

Furthermore,

$$\begin{aligned}
\mathrm{Cov}_1(p,\lambda,\mathrm{root}_{\mathcal{T}}) &= \max_{\xi} \frac{\left(\sum_{b\in\Pi} \lambda_b (L_b - L_{\mathrm{root}})\xi\right)^2}{\sum_{a\in\Pi} p_a \|S_a\xi\|_1^2} \\
&= \max_{\xi} \frac{\left(\sum_{b\in\Pi} \lambda_b (\sum_{c\in\Pi} G_b^{\mathcal{T}}(c,:) S_c\xi\right)^2}{\sum_{a\in\Pi} p_a \|S_a\xi\|_1^2} \\
&\le \max_{\xi} \frac{\left(\sum_{c\in\Pi} \|S_c\xi\|_1 (\sum_{b\in\Pi} \lambda_b \|G_b^{\mathcal{T}}(c,:)\|_\infty)\right)^2}{\sum_{a\in\Pi} p_a \|S_a\xi\|_1^2} \\
&\le \sum_{c\in\Pi} \frac{1}{p_c} \left(\sum_{b\in\Pi} \lambda_b \|G_b^{\mathcal{T}}(c,:)\|_\infty\right)^2 \\
&\le \sum_{c\in\Pi} \sum_{b\in\Pi} \frac{\lambda_b}{p_c} \|G_b^{\mathcal{T}}(c,:)\|_\infty^2 \\
&= \sum_{b\in\Pi} \sum_{c\in\Pi} \frac{\lambda_b}{p_c} \|G_b^{\mathcal{T}}(c,:)\|_\infty^2.
\end{aligned}$$

where the first equality is by a change of variables: $\xi = q_1 - q_2$; the second equality is from Lemma A.6; the first inequality is by Holder's inequality that $G_b^{\mathcal{T}}(c,:)S_c\xi \le \|G_b^{\mathcal{T}}(c,:)\|_\infty \|S_c\xi\|_1$ and algebra; the second inequality is by Cauchy-Schwarz; the last inequality is by the power-mean inequality that $(\sum_{b=1}^k \lambda_b \|G_b^{\mathcal{T}}(c,:)\|_\infty)^2 \le \sum_{b=1}^k \lambda_b \|G_b^{\mathcal{T}}(c,:)\|_\infty^2$.

For the second term, the same reasoning gives

$$\mathrm{Cov}_2(p,\mathrm{root}_{\mathcal{T}}) \le \sum_{c\in\Pi} \sum_{b\in\Pi} \frac{p_b}{p_c} \|G_b^{\mathcal{T}}(a,\cdot)\|_\infty^2.$$

Therefore,

$$\begin{aligned}
\mathrm{Coverage}(p,\lambda) &\le 8 \sum_{b=1}^k \sum_{c=1}^k \frac{\lambda_b}{p_c} \|G_b^{\mathcal{T}}(c,\cdot)\|_\infty^2 + 8 \sum_{b=1}^k \sum_{c=1}^k \frac{p_b}{p_c} \|G_b^{\mathcal{T}}(c,\cdot)\|_\infty^2 \\
&= 8 \sum_{b\in\Pi} \sum_{c\in\Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c}\right) \|G_b^{\mathcal{T}}(c,\cdot)\|_\infty^2.
\end{aligned}$$

This finishes the proof of the first inequality of the lemma.

Furthermore, we have

$$G_b^{\mathcal{T}}(c,\cdot) = \sum_{e\in\mathrm{path}_{\mathcal{T}}(b)} w_e(c,\cdot),$$

Let $e = (e_1, e_2)$, then by Definition 2.3, $w_e(c, \sigma) = 0 \ \ \forall c \neq e_1, e_2$. Hence,

$$\begin{aligned}
\text{Coverage}(p, \lambda) &\leq 8 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \| G_b^{\mathcal{T}}(c, \sigma) \|_\infty^2 \\
&= 8 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \| \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(c, \cdot) \mathbb{1}\left\{ c \in \{e_1, e_2\} \right\} \|_\infty^2 \\
&= 8 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \| \sum_{e \in \text{path}_{\mathcal{T}}(b)} \mathbb{1}\left\{ c \in \{e_1, e_2\} \right\} \|_\infty^2 \\
&= 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b,c): c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \\
&\qquad \text{(one vertex will be a part of maximum 2 edges in an in-tree path)} \\
&\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b,c): c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right)
\end{aligned}$$

This finishes the proof of the second inequality of the lemma. □

Above result is true for any in-tree with an arbitrary root. However, in general, we do not want the paths to be arbitrary: for a "bad" action c and a "good" action b (i.e., $e_c^\top L\hat{q}$ is large while $e_b^\top L\hat{q}$ is small), we would like p_c to be small and p_b to be large to ensure $p^T \boldsymbol{L}\hat{q}(x) - \lambda^T \boldsymbol{L}\hat{q}(x)$ to be small; however, if c happens to lie in $\text{path}_{\mathcal{T}}(b)$, that can cause the $\frac{\lambda_b}{p_c}$ term to be very large. To address this challenge, Lattimore & Szepesvári (2020) follows a unique construction. They construct an $\boldsymbol{L}\hat{q}(x)$-decreasing in-tree $\mathcal{T}$ (that is, for every directed edge (b, c) in $\mathcal{T}$, $e_b^\top \boldsymbol{L}\hat{q}(x) \geq e_c^\top \boldsymbol{L}\hat{q}(x)$) with $\text{root}_{\mathcal{T}} = \arg\min_a e_a^\top \boldsymbol{L}\hat{q}(x)$ and allocate probabilities using a special operator called water-transfer (Lemma A.2.3) so that along every path, the predicted losses are decreasing, and thus it is always that "good" action covers "bad" actions. However, it turns out that the simple inverse gap weighting (IGW) based probability allocation (Line 5) can complete the same task. This is because, since IGW is based on loss, it assigns higher probability towards the root in $\boldsymbol{L}\hat{q}(x)$-decreasing in-tree so that along every path, the predicted losses are decreasing. Using this simple construction, we also save a $\sqrt{k}$ factor on the regret bound compared to that of Lattimore & Szepesvári (2020).

However, we wish to point out that our optimization problem is fundamentally different from Lattimore & Szepesvári (2020)'s; our problem has the min player being the exploration distribution p and max player being another action distribution; their optimization problem has min player being exploration distribution and max player being outcome distribution, and when they try to bound the minimax value of their optimization problem, they use the Sion's minimax theorem which turns the objective to one that coincides with our Equation (3).

Before moving on to prove Lemma 3.3, we define

Definition A.2 ($\mathcal{T}$-decreasing in-tree (Lattimore & Szepesvári, 2020))**.** Let $\mathcal{T}$ be an in-tree over $[\Pi]$. A vector $y \in \mathbb{R}^\Pi$ is called $\mathcal{T}$-decreasing if $y_a \geq y_b$ for all directed edge $(a, b) \in \mathcal{T}$. That is, $a \mapsto y_a$ is decreasing towards the root of $\mathcal{T}$.

Now we state the $\mathcal{V}_\gamma(\hat{q}, p)$ bound for SquareCB.PMSide.

Lemma 3.3 (Guarantee for Alg. 3 for locally observable games)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be a locally observable partial monitoring game in the sense of Definition 2.3. For every $\gamma > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 2 run with $\kappa = \frac{\gamma}{8km^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \leq O\left(\frac{m^2 k^2}{\gamma} \right).$$

Proof. Recall that we define in Alg. 3 that, $\hat{b} = \arg\min_{a \in \Pi} e_a^\top \boldsymbol{L}\hat{q}(x)$ and $\forall a \in \Pi \setminus \{\hat{b}\}$, $p(a) = \frac{1}{\mu + \kappa (e_a - e_{\hat{b}})^\top L\hat{q}(x)}$ and $p(\hat{b}) = 1 - \sum_{a \in \Pi \setminus \{\hat{b}\}} p(a)$.

We choose $\mathcal{T}$ to be $\boldsymbol{L}\hat{q}(x)$-decreasing in-tree in the neighborhood graph; such a $\mathcal{T}$ is guaranteed to exist by Lattimore & Szepesvári (2020) (Lemma 19) we have that for any (b, c) such that $c \in \mathrm{path}_{\mathcal{T}}(b)$, $p_c \geq p_b$. This is true because for any $c \in \mathrm{path}(b)$ of $\boldsymbol{L}\hat{q}(x)$-decreasing in-tree, estimated loss of c is lower than estimated loss of b. Hence, by construction of IGW based policy p, $p_c \geq p_b$.

By Lemma 3.2 we have,

$$\mathrm{Coverage}(p, \lambda) \leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b,c): c \in \mathrm{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right).$$

Since for every c in $\mathrm{path}_{\mathcal{T}}(b)$, $p_c \geq p_b$

$$\begin{aligned}
\mathrm{Coverage}(p, \lambda) &\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{b \in \Pi} \left(\frac{\lambda_b}{p_b} + \frac{p_b}{p_b} \right) \\
&\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2 + 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b}
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sup_{\lambda \in \Delta(\Pi)} \left\{ p^\top \boldsymbol{L}\hat{q}(x) - \lambda^\top \boldsymbol{L}\hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b} \right\} \\
&\leq \sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} (p_b - \lambda_b) \boldsymbol{L}_{b,:}\hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \lambda_b \left(\mu + \kappa \left(e_b - e_{\hat{b}} \right)^\top L\hat{q}(x) \right) \right]
\end{aligned}$$

In the above inequality, we use the observation that $p_{\hat{b}}$ is lower bounded by $\frac{1}{k}$. This is true, since $\mu = k$.

$$\begin{aligned}
&\sup_{\lambda \in \Delta(\Pi)} \left\{ p^\top \boldsymbol{L}\hat{q}(x) - \lambda^\top \boldsymbol{L}\hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b} \right\} \\
&\leq \sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} \big(p(b) - \lambda(b) \big) \boldsymbol{L}_{b,:}\hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2 + \sum_{b \in \Pi} \lambda_b \left(e_b - e_{\hat{b}} \right)^\top L\hat{q}(x) \right] \\
&\qquad\qquad (\mu = k, \kappa = \tfrac{\gamma}{32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k})
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} (p_b - \lambda_b) \boldsymbol{L}_{b,:}\hat{q}(x) + \sum_{b \in \Pi} \lambda_b \left(e_b - e_{\hat{b}} \right)^\top L\hat{q}(x) \right] \\
&= \sum_{b \in \Pi} p_b \cdot e_b^\top \boldsymbol{L}\hat{q}(x) - e_{\hat{b}}^\top \boldsymbol{L}\hat{q}(x) \\
&= \sum_{b \in \Pi \setminus \{\hat{b}\}} p_b \cdot e_b^\top \boldsymbol{L}\hat{q}(x) - e_{\hat{b}}^\top \boldsymbol{L}\hat{q}(x) \\
&= \sum_{b \in \Pi \setminus \{\hat{b}\}} \frac{1}{\mu + \kappa \left(e_a - e_{\hat{b}} \right)^\top L\hat{q}(x)} \cdot e_b^\top \boldsymbol{L}\hat{q}(x) - e_{\hat{b}}^\top \boldsymbol{L}\hat{q}(x) \\
&\leq \frac{k}{\kappa} = \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2
\end{aligned}$$

Hence, $\forall \hat{q} \in \Delta(d)$

$$\mathcal{V}_\gamma(\hat{q}, p) \le \frac{1}{\gamma} \cdot 96 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2.$$

Furthermore, revisiting Definition 2.3,

$$\begin{aligned} \boldsymbol{L}_{ao} - \boldsymbol{L}_{bo} &= \sum_{c \in \{a,b\}} w_e(c, \Phi_{co}), \qquad \text{for all } o \in [d]. \\ &= w_e(a, \Phi_{co}) + w_e(b, \Phi_{co}). \end{aligned}$$

By Lemma A.10, one can always find, a witness function such that, $\|w_e\|_\infty \le \frac{m}{2}$, Hence, $\forall \hat{q} \in \Delta(d)$

$$\mathcal{V}_\gamma(\hat{q}, p) = O\left(\frac{m^2 k^2}{\gamma}\right).$$

This completes the proof. □

A.2.4 Non-degeneracy assumption

Definition 2.3 is applicable only when the game is non-degenerate. When there is degenerate action in $[k]$, the definition for local observability becomes,

Definition A.3 (Local observability, general form; Bartók et al. (2012b))**.** A globally observable partial monitoring game is *locally observable* if for every edge $e = (a, b) \in \mathcal{E}$ of the neighborhood graph the witness function w_e of Definition 2.2 may be chosen supported on

$$N_{a,b}^+ := \{c \in [k] : \mathcal{C}_a \cap \mathcal{C}_b \subseteq \mathcal{C}_c\},$$

that is, $w_e(c, \cdot) \equiv 0$ for all $c \notin N_{a,b}^+$, so that

$$L_{ao} - L_{bo} = \sum_{c \in N_{a,b}^+} w_e(c, \Phi_{co}), \qquad \text{for all } o \in [d].$$

The set $N_{a,b}^+$ contains the two endpoints, but may also contain degenerate actions whose cells lie inside the shared face $\mathcal{C}_a \cap \mathcal{C}_b$. For $d = 3$, for instance, a degenerate action with $\dim(\mathcal{C}_c) = 1$ is optimal only along a line segment, whereas a Pareto optimal action is optimal on a two-dimensional region, yet such an action can still be a necessary source of information.

Two parts of our development break under this more general condition. First, the exploration rule must assign probability to the degenerate actions as well, which requires a modification of SquareCB.PMSide. Second, Lemma A.10 no longer applies and must be replaced by Lemma A.11 of Lattimore & Szepesvári (2020), which introduces a factor exponential in d. The modification to the algorithm is straightforward, and we conjecture that SquareCB.PMSide continues to guarantee $O(\sqrt{T})$ regret, but the analysis requires care. We leave this extension to future work.

A.3 Contextual partial monitoring: reduction to offline regression

For definitions and examples of offline outcome estimation oracles, Please see Section A.1.2.

A.3.1 OE2D.PMSide

The bounds on $\mathcal{V}_\gamma(\hat{q}(x), p(x))$ established for globally observable games (Lemma 3.1) and for locally observable games (Lemma 3.3), under the exploration strategies ϵ-Greedy.PMSide and SquareCB.PMSide respectively, remain valid in the offline setting, since both hold for every $\hat{q} \in \Delta(d)$ rather than for a particular estimate.

What does not carry over is the regret decomposition of Theorem 3.1. Inspecting the guarantee of Assumption A.1, an offline oracle's Hellinger bound makes two demands of the sampling policy: it must be held fixed over the rounds to which the bound is applied, and it must be the policy that collected the data from which $\hat{q}$ was estimated.

The first is obtained by freezing the policy within each epoch (Line 8). The second is where coverage does the work. Because $\mathrm{Coverage}(p, \lambda)$ is a supremum over pairs of models in the class and never

references $\hat{q}$, a policy with small coverage distinguishes *any* two members of $\mathcal{Q}$, in particular $\hat{q}$ and q^*, whichever estimate the oracle happens to return. Lemma A.7 makes this precise.

The DEC of Foster et al. (2021) lacks this property: its exploration term $-\gamma \cdot \mathbb{E}_{a\sim p}\left[D^2_{\mathrm{H}}\big(S_a\hat{q}(x)\,\big|\,S_a q\big)\right]$ is anchored at $\hat{q}$. In an epoch-based algorithm (Lines 7–14), the estimation error for a given $\hat{q}$ therefore cannot be controlled using a policy from a different epoch, which played no part in producing it.

Now we present our algorithm OE2D.PMSIDE.

Algorithm 4 OE2D.PMSIDE

1: **parameters:**
   Offline estimation oracle $\mathrm{Alg}_{\mathrm{Off}}$
   Exploration parameters $\{\gamma_\ell\}_{\ell=1}^{L}$
   Error rates $\{\delta_\ell\}_{\ell=1}^{L}$
   Epoch schedule $0 = \tau_0 < \tau_1 < \cdots < \tau_L$
2: **for** $t = 1$ **to** τ_1 **do**
3: Observe the context x_t
4: Find p_t that approximately minimizes,

$$\sup_{\lambda\in\Delta(\Pi)} \mathrm{Coverage}(p,\lambda).$$

5: Choose action $A_t \sim p_t$ and observe feedback σ_t
6: **end for**
7: **for** $\ell = 2, 3, \cdots, L$ **do**
8: Compute estimate $\hat{q}_\ell = \mathrm{Alg}_{\mathrm{Off}}\left(\{A_s, x_s, \sigma_s\}_{s=\tau_{\ell-2}+1}^{\tau_{\ell-1}}\right)$
9: **for** $t = \tau_{\ell-1} + 1$ **to** τ_ℓ **do**
10: Receive context x_t
11: Find p_t that approximately minimizes

$$\mathcal{V}_{\gamma_\ell}(\hat{q}_\ell(x_t), p) := \sup_{\lambda\in\Delta(\Pi)} \underbrace{p^\top \boldsymbol{L}\hat{q}_\ell(x_t) - \lambda^\top \boldsymbol{L}\hat{q}_\ell(x_t)}_{\text{exploit}} + \frac{1}{\gamma_\ell}\cdot \underbrace{\mathrm{Coverage}(p,\lambda)}_{\text{explore}}.$$

12: Choose action $A_t \sim p_t$ and observe feedback σ_t.
13: **end for**
14: **end for**

Furthermore, we can employ Algorithm 2 for globally observable games, such that for every $\gamma_\ell > 16k^2\beta^2$, and every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_{\gamma_\ell}\big(\hat{q}, p\big) \le O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right). \qquad \text{(Lemma 3.1)}$$

and for locally observable games, we can employ Algorithm 3, such that For every $\gamma_\ell > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 2 run with $\kappa = \frac{\gamma}{8k^2m^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_{\gamma_\ell}\big(\hat{q}, p\big) \le O\left(\frac{m^2k^3}{\gamma_\ell}\right). \qquad \text{(Lemma 3.3)}$$

So in both cases, we have algorithms to find policies that establishes an upper bound for our objective function. Let us denote it by V_ℓ. So,

$$V_\ell = O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right) \qquad \text{(for globally observable games)}$$

$$V_\ell = O\left(\frac{m^2k^3}{\gamma_\ell}\right) \qquad \text{(for locally observable games)}$$

Furthermore, within a single epoch ℓ for all round $t \in [\tau_{\ell-1}+1, \tau_\ell]$, the estimator does not change($\hat{q}_\ell(\cdot)$), only the context x_t changes. So, within an epoch ℓ,

$$x_{t_1} = x_{t_2} \implies \hat{q}_\ell(x_{t_1}) = \hat{q}_\ell(x_{t_2}) \implies p_{t_1} = p_{t_2} \quad \forall t_1, t_2 \in [\tau_{\ell-1}+1, \tau_\ell]$$

So, we can always think that within an epoch, every p_t is sampled from a fixed distribution $p_\ell(\cdot \mid x_t)$, that is, we can denote,

$$p_t \leftarrow p_\ell(\cdot \mid x_t) \quad \forall t \in [\tau_{\ell-1}+1, \tau_\ell].$$

A.3.2 Regret analysis

Next, for readability and reduce the notations clutter we define the following,

Definition A.4. Define:

- $\mathcal{L}^\star(\lambda \mid x) = \lambda^T \boldsymbol{L} q^*(x)$: The loss incurred by policy λ under true outcome distribution q^* for a context x.
- $\widehat{\mathcal{L}}_\ell(\lambda \mid x) = \lambda^T \boldsymbol{L} \hat{q}_\ell(x)$: The loss incurred by policy λ under estimated distribution $\hat{q}_\ell$ for a context x.
- $\lambda^* = \arg\min_{\lambda \in \Delta(\Pi)} \mathcal{L}^\star(\lambda \mid x)$: The true best policy for context x.
- $\hat{\lambda}_\ell = \arg\min_{\lambda \in \Delta(\Pi)} \widehat{\mathcal{L}}_\ell(\lambda \mid x)$: The estimated best policy for context x.
- $\mathrm{Reg}(\lambda \mid x) = \mathcal{L}^\star(\lambda \mid x) - \mathcal{L}^\star(\lambda^\star \mid x)$: The regret incurred by policy λ under true outcome distribution q^* for a context x.
- $\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) = \widehat{\mathcal{L}}_\ell(\lambda \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda} \mid x)$: The regret incurred by policy λ under estimated distribution $\hat{q}_\ell$ for a context x.
- $C(p_\ell, \lambda) = \mathrm{Coverage}(p_\ell, \lambda,)$
- $E_\ell = \mathbb{E}_{a \sim p_{\ell-1}} \left[D^2_{\mathrm{H}} \left(S_a \hat{q}_\ell(x) \mid S_a q^*(x) \right) \right]$

In all the above definitions, we might drop $(\cdot \mid x)$, when it is evident from the context. For example, we might denote $\mathcal{L}^\star(\lambda \mid x) \to \mathcal{L}^\star(\lambda)$.

Next we present two important properties of our policies and estimators.

Certification:

This property is a direct consequence of our algorithms finding the policies which bounds our objective function by V_ℓ.

$\forall x \in \mathcal{X}$ and $\forall \ell \geq 1$,

$$\mathcal{V}_{\gamma_\ell}(\hat{q}_\ell(x), p_\ell) = \sup_{\lambda \in \Delta(\Pi)} p_\ell^\top \boldsymbol{L} \hat{q}_\ell(x) - \lambda^\top \boldsymbol{L} \hat{q}_\ell(x) + \frac{1}{\gamma_\ell} \cdot \mathrm{Coverage}(p_\ell, \lambda) \leq V_\ell$$

Hence,

$\forall x \in \mathcal{X}$, $\forall \ell \geq 1$ and, $\forall \lambda \in \Pi$

$$p_\ell^\top \boldsymbol{L} \hat{q}_\ell(x) - \lambda^\top \boldsymbol{L} \hat{q}_\ell(x) + \frac{1}{\gamma_\ell} \cdot \mathrm{Coverage}(p_\ell, \lambda) \leq V_\ell$$
$$\implies \widehat{\mathcal{L}}_\ell(p_\ell \mid x) - \widehat{\mathcal{L}}_\ell(\lambda \mid x) + \frac{1}{\gamma_\ell} \cdot C(p_\ell, \lambda) \leq V_\ell$$

Since, this is true for all $x \in \mathcal{X}$, we take the liberty to drop it,

$$\widehat{\mathcal{L}}_\ell(p_\ell) - \widehat{\mathcal{L}}_\ell(\lambda) + \frac{1}{\gamma_\ell} \cdot C(p_\ell, \lambda) \leq V_\ell. \qquad \text{(certification)}$$

Off-policy evaluation (OPE):

Lemma A.7 (Off-policy evaluation lemma)**.** With all the definitions in Definition A.4, $\forall \lambda$, $\forall \ell \geq 2$, and $\forall x$,

$$\left|\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)\right| \leq \frac{1}{6\gamma_\ell} \cdot C(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell. \qquad \text{(Off-policy evaluation)}$$

Proof.

$$\begin{aligned}
&\left(\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)\right)^2 \\
&= \left((\lambda - p_{\ell-1})^T \boldsymbol{L} \hat{q}_\ell(x) - (\lambda - p_{\ell-1})^T \boldsymbol{L} q^*(x)\right)^2 \\
&= \frac{\left((\lambda - p_{\ell-1})^T \boldsymbol{L}(\hat{q}_\ell(x) - q^*(x))\right)^2}{\mathbb{E}_{a\sim p_{\ell-1}}\left[D^2_{\mathrm{H}}\left(S_a \hat{q}_\ell(x) \mid S_a q^*(x)\right)\right]} \cdot \mathbb{E}_{a\sim p_{\ell-1}}\left[D^2_{\mathrm{H}}\left(S_a \hat{q}_\ell(x) \mid S_a q^*(x)\right)\right] \\
&\leq \mathrm{Coverage}(p_{\ell-1}, \lambda) \cdot \mathbb{E}_{a\sim p_{\ell-1}}\left[D^2_{\mathrm{H}}\left(S_a \hat{q}_\ell(x) \mid S_a q^*(x)\right)\right] \\
&\left|\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)\right| \\
&\leq \sqrt{\mathrm{Coverage}(p_{\ell-1}, \lambda) \cdot \mathbb{E}_{a\sim p_{\ell-1}}\left[D^2_{\mathrm{H}}\left(S_a \hat{q}_\ell(x) \mid S_a q^*(x)\right)\right]} \\
&\leq \frac{1}{6\gamma_\ell} \cdot \mathrm{Coverage}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} \mathbb{E}_{a\sim p_{\ell-1}}\left[D^2_{\mathrm{H}}\left(S_a \hat{q}_\ell(x) \mid S_a q^*(x)\right)\right] \qquad \text{(AM-GM)} \\
&= \frac{1}{6\gamma_\ell} \cdot C(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.
\end{aligned}$$

□

The second term is the generalization error of the offline oracle: $\hat{q}_\ell$ is estimated from data collected in the previous epoch under $p_{\ell-1}$, so Assumption A.1 (Hellinger boundedness) applies to it directly. This coupling between the estimate and the policy that produced it is what makes the offline analysis more delicate than the online one, where the estimator is updated as the data arrive.

The first term involves no estimate at all: $\mathrm{Coverage}(p_{\ell-1}, \lambda)$ depends only on the model class, on λ and on $p_{\ell-1}$. A policy that makes it small therefore controls the evaluation error for every $\hat{q} \in \mathrm{co}(\mathcal{Q}_x)$ at once, in particular for whichever estimate the oracle returns in epoch ℓ.

Now we present an abstract regret theorem about epoch-based offline oracle-efficient interactive learning algorithm from Qin & Zhang (2026a).

Theorem A.1 (Generalization of regret concentration theorem from Qin & Zhang (2026a))**.** Let Λ be a set of policies, $\mathcal{L}(\cdot \mid x) : \Lambda \to [0, 1]$ a value function and $\mathcal{L}^\star(\cdot \mid x)$ and $\widehat{\mathcal{L}}_\ell(\cdot \mid x)$, $\ell \in [L]$ are value functions measured under the true model and estimated models respectively. $\lambda^\star \in \Lambda$ and $\hat{\lambda}_\ell \in \Lambda$ are respectively the optimal policies under respective true and estimated models. Let $C_\ell : \Lambda \times \Lambda \to \mathbb{R}_+$ be coverage maps, $p_\ell \in \Lambda$ are decisions, $0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_L$ learning rates, and $V_\ell, E_\ell \geq 0$ numbers such that,

(C1) (certificate) for every $\ell \geq 1$ and every $\lambda \in \Lambda$,

$$\widehat{\mathcal{L}}_\ell(p_\ell \mid x) - \widehat{\mathcal{L}}_\ell(\lambda \mid x) + \frac{1}{\gamma_\ell} C_\ell(p_\ell, \lambda) \leq V_\ell.$$

(C2) (off-policy evaluation) for every $\ell \geq 2$ and every $\lambda \in \Lambda$,

$$\left|\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)\right| \leq \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.$$

Define $G_1 \coloneqq 1$ and, for $\ell \geq 2$,

$$\gamma_\ell G_\ell \coloneqq \gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell, \qquad \text{i.e.} \qquad G_\ell = \frac{1}{\gamma_\ell}\left[\gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i)\right].$$

Then, for every $\ell \in [L]$ and every $\lambda \in \Lambda$,

$$\mathrm{Reg}(\lambda \mid x) \le 2\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + G_\ell \qquad \text{and} \qquad \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) \le 2\mathrm{Reg}(\lambda \mid x) + G_\ell.$$

Consequently $\mathrm{Reg}(p_\ell \mid x) \le 2V_\ell + G_\ell$ for every ℓ, and

$$\sum_{\ell=1}^{L}(\tau_\ell - \tau_{\ell-1})\,\mathrm{Reg}(p_\ell \mid x) \le \tau_1 + \sum_{\ell=2}^{L}\frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1}V_{i-1} + 9\gamma_i^2 E_i)\right].$$

Proof is deferred to Appendix A.6. We can apply this theorem directly and obtain a per-context regret bound. Now we state the main theorem that bounds the regret with high probability for any epoch schedule.

Theorem A.2 (Regret of OE2D.PMSIDE)**.** Suppose $\mathrm{Alg}_{\mathrm{Off}}$ has a Hellinger regret bound $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T, \delta)$. For any non-degenerate globally observable partial monitoring game without duplicate actions, assume that $x_t \overset{i.i.d}{\sim} \mathcal{D}_x$, then for any $\delta \in (0,1)$, OE2D.PMSIDE run with

1. exploration rule Algorithm 2 for globally observable game (with $\gamma_\ell > 16k^2\beta^2 \ \forall \ell \in [L]$).
2. exploration rule Algorithm 3 for locally observable game (with $\kappa = \frac{\gamma_\ell}{8k^2m^2} \ \forall \ell \in [L]$ and $\mu = k$)
3. exploration parameters such that, $0 < \gamma_1 \le \gamma_2 \le \cdots \le \gamma_L$

satisfies,

$$\widetilde{\mathrm{Reg}}_T \le L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}\left(\gamma_{i-1}V_{i-1} + 9\gamma_i^2 \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta_i)\right)\right).$$

where,

- $\widetilde{\mathrm{Reg}}_T = \mathbb{E}_{x_1,\ldots,x_T \overset{i.i.d}{\sim} \mathcal{D}_x}\left[\mathrm{Reg}_T\right]$
- $V_\ell = O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right)$ for globally observable games
- $V_\ell = O\left(\frac{m^2k^3}{\gamma_\ell}\right)$ for locally observable games

Proof. From Theorem A.1, the regret per-context is bounded by,

$$\begin{aligned}\sum_{\ell=1}^{L}(\tau_\ell - \tau_{\ell-1})\,\mathrm{Reg}(p_\ell \mid x) &\le \tau_1 + \sum_{\ell=2}^{L}(\tau_\ell - \tau_{\ell-1})(2V_\ell + G_\ell) && (\mathrm{Reg}(p_1 \mid x) \le 1)\\ &= \tau_1 + \sum_{\ell=2}^{L}\frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1}V_{i-1} + 9\gamma_i^2 E_i)\right].\end{aligned}$$

Let us denote,

$$\widetilde{\mathrm{Reg}}_T = \mathbb{E}_{x \overset{i.i.d}{\sim} \mathcal{D}_x}\left[\mathrm{Reg}_T\right]$$

Here, note that, even though the randomness w.r.t to contexts are removed, $\widetilde{\mathrm{Reg}}_T$, is still a random variable in terms of policy. Hence,

$$\begin{aligned}\widetilde{\mathrm{Reg}}_T = \mathbb{E}_{x \overset{i.i.d}{\sim} \mathcal{D}_x}\left[\mathrm{Reg}_T\right] &= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{t=\tau_{\ell-1}}^{\tau_\ell}\mathrm{Reg}(p_\ell \mid x_t)\right]\\ &\le \mathbb{E}\left[\sum_{\ell=1}^{L}\tau_1 + \sum_{\ell=2}^{L}\frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1}V_{i-1} + 9\gamma_i^2 E_i)\right)\right]\\ &\le L \cdot \tau_1 + 2L \max_{\ell \in [L]}\frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}\left(\gamma_{i-1}V_{i-1} + 9\gamma_i^2\,\mathbb{E}_{x \overset{i.i.d}{\sim} \mathcal{D}_x}\left[E_i\right]\right)\right).\end{aligned}$$

If the contexts are i.i.d such that, $x_t \overset{i.i.d}{\sim} \mathcal{D}_x$, then,

$$\widetilde{\mathrm{Reg}}_T \le L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \, \mathbb{E}_{x \overset{i.i.d}{\sim} \mathcal{D}_x} \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2 \left(S_a \hat{q}_\ell(x) \mid S_a q^*(x) \right) \right] \right) \right)$$

$$\le L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} D_{i-1} + 9\gamma_i^2 \frac{\mathrm{Reg}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i - \tau_{i-1})}{\tau_i - \tau_{i-1}} \right) \right).$$

We know that, with probability of at least 1- δ',

$$\mathrm{Reg}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i - \tau_{i-1}) \le \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta)$$

Set $\delta_\ell = \frac{\delta}{L}$, with a probability at least $1 - \delta$,

$$\widetilde{\mathrm{Reg}}_T \le L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \frac{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta_i)}{\tau_i - \tau_{i-1}} \right) \right).$$

□

The final regret depends on epoch schedule. We choose a doubling doubling schedule.

Corollary 3. *Suppose* $\mathrm{Alg}_{\mathrm{Off}}$ *has a Hellinger regret bound* $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T, \delta)$. *For any non-degenerate globally observable partial monitoring game without duplicate actions, assume that* $x_1, \ldots, x_T \overset{i.i.d}{\sim} \mathcal{D}_x$, *then for any* $\delta \in (0,1)$, *the algorithm OE2D.PMSide run with,*

1. *exploration rule Algorithm 2 for globally observable game (with* $\gamma_\ell > 16k^2\beta^2 \ \forall \ell \in [L]$*).*
2. *exploration rule Algorithm 3 for locally observable game (with* $\kappa = \frac{\gamma_\ell}{8k^2m^2} \ \forall \ell \in [L]$ *and* $\mu = k$*)*
3. *doubling schedule, such that* $\tau_\ell = 2^\ell$
4. *exploration parameters such that,* $0 < \gamma_1 \le \gamma_2 \le \cdots \le \gamma_L$
5. *with assumption* $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, t, \delta)$ *does not depend on* t. *Generally this is the case.*
6. *with* $\delta_\ell = \frac{\delta}{L}$

satisfies the following guarantees,

1. *For globally observable games, with* $\gamma_\ell = \left(\frac{C \cdot \tau_\ell \cdot k\beta}{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_\ell/2, \delta/L)} \right)^{\frac{2}{3}}$ *for some constant* C,

$$\widetilde{\mathrm{Reg}}_T \le O\left(T^{\frac{2}{3}} \log^2(T) (k\beta)^{\frac{2}{3}} (\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T, \delta/\log(T)))^{\frac{1}{3}} \right)$$

2. *For locally observable games, with* $\gamma_\ell = \sqrt{\frac{D \cdot \tau_\ell \cdot k^3 m^2}{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_\ell/2, \delta_\ell)}}$ *for some constant* D,

$$\widetilde{\mathrm{Reg}}_T = O\left(\log^2(T) \sqrt{T \cdot k^3 m^2 \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T/2, \delta/\log(T))} \right)$$

where, $\widetilde{\mathrm{Reg}}_T = \mathbb{E}_{x \overset{i.i.d}{\sim} \mathcal{D}_x} \left[\mathrm{Reg}_T \right]$.

Proof. First of all, for doubling epoch schedule,

$$\tau_L = O(T)$$
$$\tau_\ell - \tau_{\ell-1} = \frac{\tau_\ell}{2}$$
$$L = O(\log T)$$

$$\widetilde{\mathrm{Reg}}_T \le L\cdot\tau_1 + 2L\max_{\ell\in[L]}\frac{\tau_\ell-\tau_{\ell-1}}{\gamma_\ell}\left(2\gamma_\ell V_\ell+\gamma_1+\sum_{i=2}^{\ell}\left(\gamma_{i-1}V_{i-1}+9\gamma_i^2\frac{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i-\tau_{i-1},\delta_i)}{\tau_i-\tau_{i-1}}\right)\right).$$

$$= L\cdot\tau_1 + 2L\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(2\gamma_\ell V_\ell+\gamma_1+\sum_{i=2}^{\ell}\left(\gamma_{i-1}V_{i-1}+9\gamma_i^2\frac{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)}{\tau_i/2}\right)\right).$$

$$\le L\cdot\tau_1 + 2L\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+\sum_{i=2}^{\ell}\gamma_i\left(2V_i+18\gamma_i\frac{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)}{\tau_i}\right)\right).$$

$$\le L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+\max_{i\in[\ell]}\gamma_i\left(2V_i+18\gamma_i\frac{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)}{\tau_i}\right)\right).$$

$$= L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+\max_{i\in[\ell]}\frac{\gamma_i}{\tau_i}\left(2\tau_iV_i+18\gamma_i\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)\right)\right).$$

For globally observable games,

$$V_i = C\cdot k\beta\sqrt{\frac{1}{\gamma_i}} \qquad \text{(for some constant } C\text{)}$$

$$\widetilde{\mathrm{Reg}}_T \le \quad = L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+\max_{i\in[\ell]}\frac{\gamma_i}{\tau_i}\left(2\tau_iC\cdot k\beta\sqrt{\frac{1}{\gamma_i}}+18\gamma_i\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)\right)\right).$$

set $\gamma_i = \left(\frac{C\tau_i\cdot k\beta}{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i)}\right)^{\frac{2}{3}}$

$$\widetilde{\mathrm{Reg}}_T$$

$$\le L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+\max_{i\in[\ell]}\frac{\gamma_i}{\tau_i}\left(2\cdot(C\cdot\tau_ik\beta)^{\frac{2}{3}}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{\frac{1}{3}}+18(C\cdot\tau_ik\beta)^{\frac{2}{3}}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{\frac{1}{3}}\right)\right)$$

$$= L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+20\max_{i\in[\ell]}\frac{\gamma_i}{\tau_i}(C\cdot\tau_ik\beta)^{\frac{2}{3}}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{\frac{1}{3}}\right).$$

$$= L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+20\max_{i\in[\ell]}\frac{1}{\tau_i}(C\cdot\tau_ik\beta)^{\frac{4}{3}}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{-\frac{1}{3}}\right).$$

$$\le L\cdot\tau_1 + 2L^2\max_{\ell\in[L]}\frac{\tau_\ell}{2\gamma_\ell}\left(\gamma_1+20\tau_\ell^{\frac{1}{3}}(C\cdot k\beta)^{\frac{4}{3}}(\max_{i\in[\ell]}\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{-\frac{1}{3}}\right).$$

$$= L\cdot\tau_1 + L^2\max_{\ell\in[L]}\frac{(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_\ell/2,\delta_\ell))^{\frac{2}{3}}\tau_\ell^{\frac{1}{3}}}{(C\cdot k\beta)^{\frac{2}{3}}}\left(\gamma_1+20\tau_\ell^{\frac{1}{3}}(C\cdot k\beta)^{\frac{4}{3}}(\max_{i\in[\ell]}\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_i/2,\delta_i))^{-\frac{1}{3}}\right).$$

$$= O\left(L^2(C\cdot k\beta)^{\frac{2}{3}}\max_{\ell\in[L]}\tau_\ell^{\frac{2}{3}}\frac{\max_{\ell\in[L]}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_\ell/2,\delta_\ell))^{\frac{2}{3}}}{\min_{\ell\in[L]}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},\tau_\ell/2,\delta_\ell))^{\frac{1}{3}}}\right)$$

$$= O\left(T^{\frac{2}{3}}\log^2(T)(k\beta)^{\frac{2}{3}}(\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},T,\delta/\log(T)))^{\frac{1}{3}}\right)$$

(Here we assume $\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q},t,\delta)$ do not depend on t. Generally this is the case.)

And for locally observable games,

$$V_i = D \cdot \frac{m^2 k^3}{\gamma_i} \qquad \text{(for some constant } D\text{)}$$

$$\widetilde{\mathrm{Reg}}_T \leq \quad = L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(2\tau_i D \cdot m^2 k^3 \frac{1}{\gamma_i} + 18\gamma_i \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i/2, \delta_i) \right) \right).$$

set $\gamma_i = \sqrt{\frac{D \cdot \tau_i \cdot k^3 m^2}{\overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i/2, \delta_i)}}$

$$\begin{aligned}
\widetilde{\mathrm{Reg}}_T &\leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20 \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(\sqrt{D \cdot \tau_i \cdot k^3 m^2 \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_i/2, \delta_i)} \right) \right) \\
&= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20D \cdot k^3 m^2 \right) \\
&= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\sqrt{\tau_\ell \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, \tau_\ell/2, \delta_\ell)}}{\sqrt{D \cdot k^3 m^2}} \left(\gamma_1 + 20D \cdot k^3 m^2 \right) \\
&= O\left(L^2 \sqrt{T \cdot k^3 m^2 \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T/2, \delta/L)} \right) \\
&= O\left(\log^2(T) \sqrt{T \cdot k^3 m^2 \overline{\mathrm{Reg}}^{\mathrm{H}}_{\mathrm{Off}}(\mathcal{Q}, T/2, \delta/\log(T))} \right)
\end{aligned}$$

□

A.4 Experiments

Additional LLM-routing details. The cached BoolQ dataset contains 12,697 prompts, of which 12,648 have parseable responses from both `Qwen2.5-0.5B-Instruct`, used as the weak model, and `Llama-2-13B-Chat`, used as the strong reference model. The binary routing outcome indicates whether their cached responses disagree; BoolQ gold labels are not used as routing outcomes. Each prompt is represented by a 138-dimensional context consisting of 10 weak-model uncertainty and probability statistics, 64 whitened principal components of the weak model's hidden states, and 64 whitened principal components of a context-free prompt embedding. All 12,648 eligible examples are presented online, and outcomes are revealed only after querying the strong model. Each policy makes a routing decision at every round, whereas its estimator is updated only once in 32 rounds, using all feedback revealed before that round. CBPSide (Bartók et al., 2012a; An et al., 2026) uses a linear logistic disagreement model and the confidence radius

$$\beta_t(x_t) = \min\left\{ 0.5 s \sqrt{x_t^\top V_{t-1}^{-1} x_t},\ 0.5 \right\},$$

where s is the tuned multiplier. The two SquareCB.PMSide variants use $\mu = 2$, $\kappa = s\sqrt{T}$, and inverse-propensity weights capped at 10. Their only intended difference is the disagreement-probability estimator. The linear variant fits an L_2-regularized logistic model with $C = 1$, the `lbfgs` solver, and a weighted standardization step. The nonlinear variant uses scikit-learn's histogram gradient-boosting (HGB) classifier with binary log loss, 50 boosting iterations, learning rate 0.05, at most 15 leaves per tree, a minimum of 20 samples per leaf, L_2 regularization 1, and no early stopping. At each eligible update, both estimators are refitted using their complete accumulated inverse-propensity-weighted feedback. SquareCB.PMSide (Grant & Leslie, 2021) uses a Bayesian logistic disagreement model with

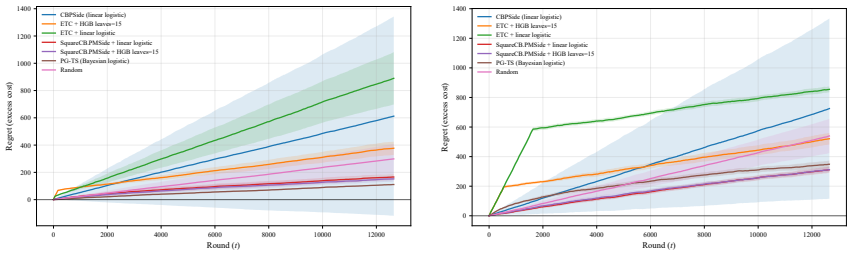
$$\theta \sim \mathcal{N}\left(0, (s\sigma_0)^2 I\right), \qquad \sigma_0 = 1,$$

and performs 15 Pólya–Gamma Gibbs transitions whenever its posterior is updated. For each policy and each ℓ_{01}, we select $s \in \{0.1, 0.3, 1, 3, 10\}$ by minimizing mean total cost over 20 paired random permutations of the online arrival order. Random is computed analytically by matching the routing

rate of the selected nonlinear SquareCB.PMSide policy. Figure 1a reports mean realized total cost with pointwise 95% Student-t confidence intervals. The appendix learning curves report cumulative excess cost relative to a fixed five-fold cross-fitted HGB reference policy, together with its per-round average (R_t/t). Each reference probability is produced by an HGB model that excludes the evaluated example and is used only for evaluation, never as online feedback. Because multiplier selection and evaluation use the same permutations, all selected-policy results are interpreted as exploratory tuned comparisons.

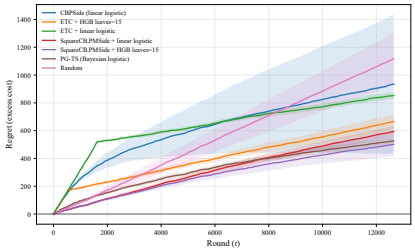
Table 1: Meaning of the multiplier s for each algorithm.

Algorithm	Scaled quantity	Effect of larger s
CBPside	confidence width sw_i	more exploration
RandCBPside	sampled confidence width sw_i	more exploration
SquareCB.PMSide	IGW rate $\kappa = s\sqrt{T}$	less exploration
ETC	exploration length $s\,T^{2/3}$	more exploration
PG-TS	prior variance $s\sigma_0$	more exploration
Random	— (no hyperparameter)	—

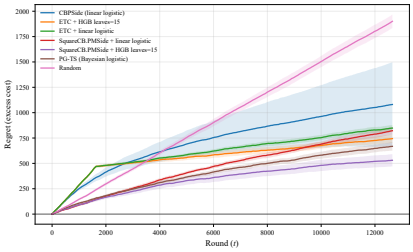


(a) $\ell_{01} = 1.8$

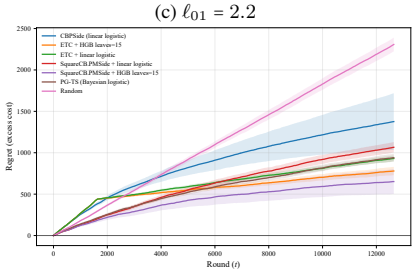
(b) $\ell_{01} = 2.0$



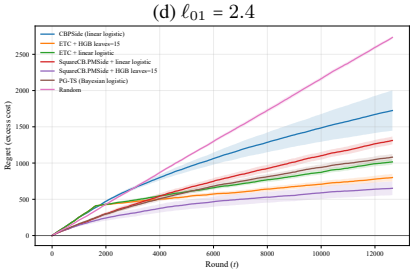
(c) $\ell_{01} = 2.2$



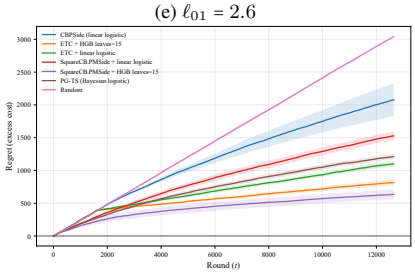
(d) $\ell_{01} = 2.4$



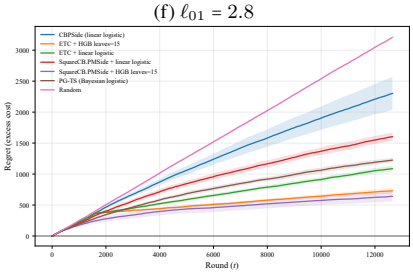
(e) $\ell_{01} = 2.6$



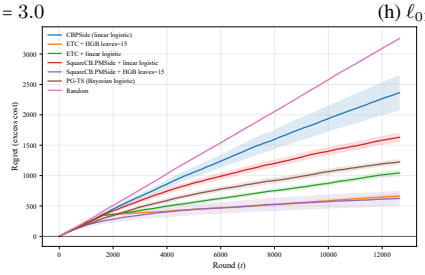
(f) $\ell_{01} = 2.8$



(g) $\ell_{01} = 3.0$

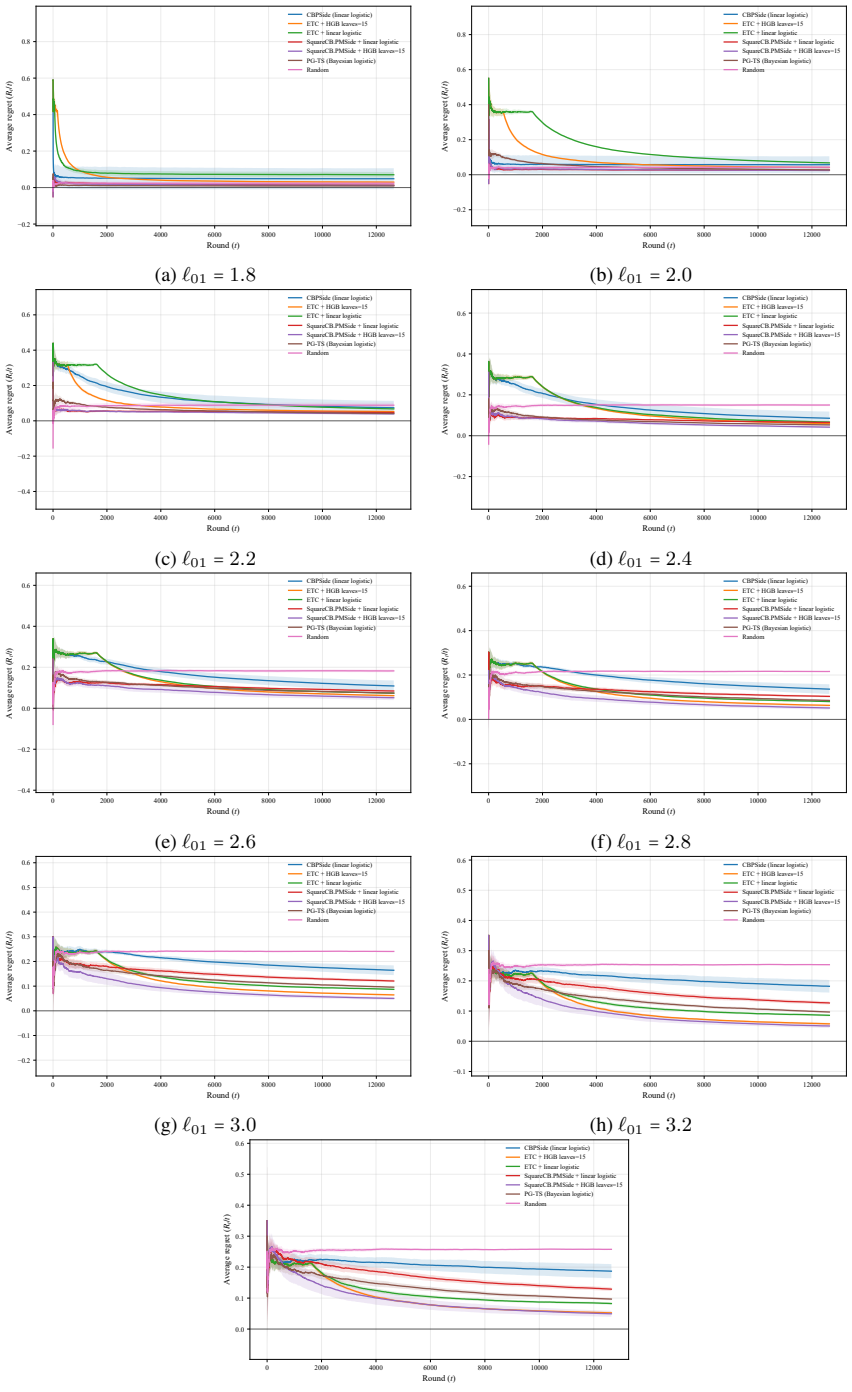


(h) $\ell_{01} = 3.2$



(i) $\ell_{01} = 3.3$

Figure 2: Cumulative empirical regret, measured as excess realized cost relative to the fixed five-fold cross-fitted HGB reference policy, for each value of ℓ_{01}. Curves show means over 20 paired random arrival orders, and shaded regions are pointwise 95% Student-t confidence intervals. Lower values are better; negative regret indicates that the online policy incurs lower realized cost than the fixed reference policy.



(a) $\ell_{01} = 1.8$ (b) $\ell_{01} = 2.0$

(c) $\ell_{01} = 2.2$ (d) $\ell_{01} = 2.4$

(e) $\ell_{01} = 2.6$ (f) $\ell_{01} = 2.8$

(g) $\ell_{01} = 3.0$ (h) $\ell_{01} = 3.2$

(i) $\ell_{01} = 3.3$

Figure 3: Per-round average empirical regret R_t/t, relative to the fixed five-fold cross-fitted HGB reference policy, for each value of ℓ_{01}. Curves show means over 20 paired random arrival orders, and shaded regions are pointwise 95% Student-t confidence intervals. A curve approaching zero indicates diminishing average excess cost relative to the fixed reference; it does not by itself establish a theoretical regret rate.

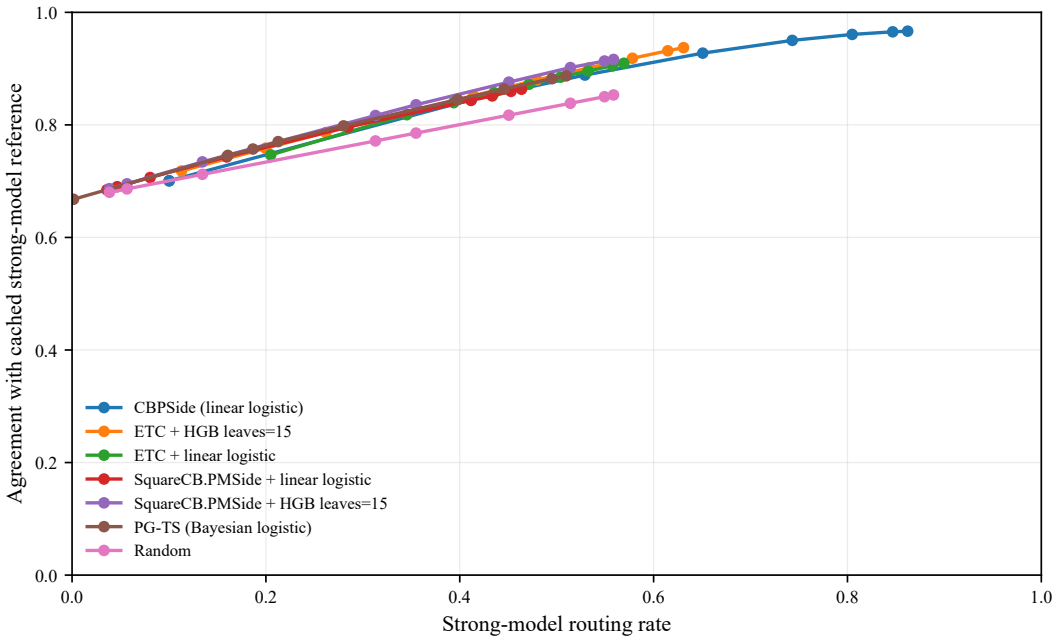
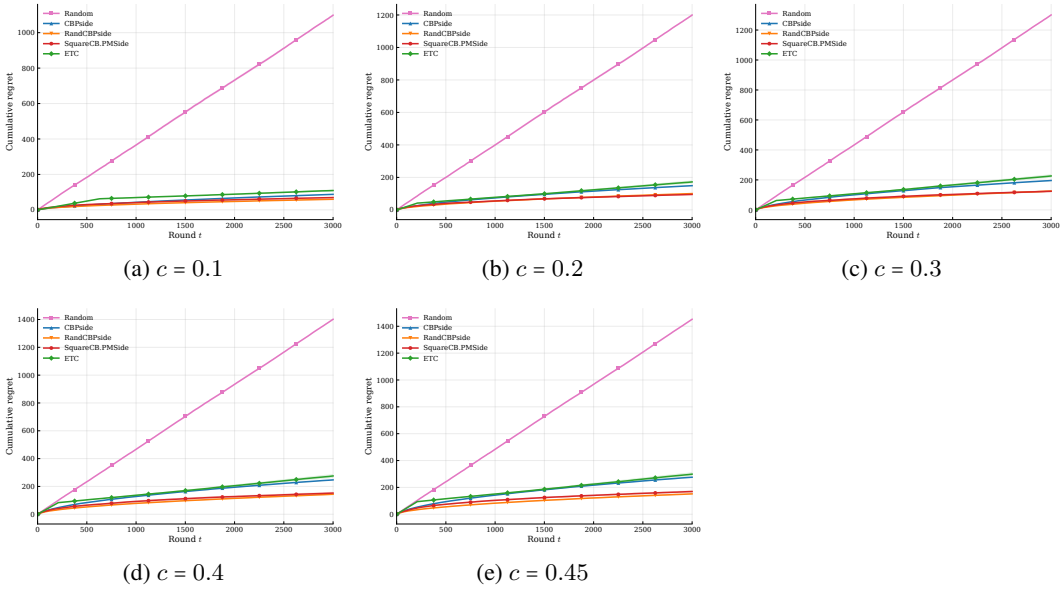


Figure 4: Routing–accuracy trade-off on BoolQ. The horizontal axis reports the fraction of prompts routed to the strong model, while the vertical axis reports agreement with the cached strong-model response. Thus, points closer to the upper-left corner are preferable, as they achieve higher agreement while making fewer strong-model queries. Each marker is the mean over 20 shuffled online orders, with confidence intervals omitted for visual clarity. Along each curve, the connected markers follow increasing values $\ell_{01} \in \{1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.3\}$. Increasing ℓ_{01} assigns a larger penalty to an incorrect weak-model response and therefore generally moves the policy toward more strong-model routing and higher agreement.



(a) $c = 0.1$ (b) $c = 0.2$ (c) $c = 0.3$

(d) $c = 0.4$ (e) $c = 0.45$

Figure 5: Cumulative regret against round t for each value of the information cost c, with every algorithm at its best multiplier. Shaded bands are 95% Student-t confidence intervals for the mean over 20 seeds.

Additional contextual label-efficient prediction experimental details. Contexts are drawn i.i.d. as $x_t \sim \mathrm{Unif}\big([0,1]^{10}\big)$, and a single weight vector w is sampled once from $\mathrm{Unif}(0, 0.1)^{10}$ and normalized so that $w_i \geq 0$ and $\sum_i w_i = 1$; consequently $w^\top x$ is a convex combination of the coordinates of x and lies in $[0, 1]$ by construction. The binary outcome is the deterministic threshold $y_t = \mathbb{1}\big\{w^\top x_t \geq 0.5\big\}$. We evaluate $c \in \{0.1, 0.2, 0.3, 0.4, 0.45\}$; we stop short of 0.5 because the cell of the expert action collapses to a single point there, the pair of prediction actions becomes a neighboring pair, and the game is no longer structurally comparable to the other settings. Each experiment uses horizon $T = 3000$ and is repeated over 20 independent seeds, and for each policy and each c we select $s \in \{0.1, 0.3, 1, 3, 10\}$ by minimizing mean total cost over those 20 paired seeds. Each method has exactly one exploration parameter. To place them on a common footing we scale each by a shared multiplier s, sweeping $s \in \{0.1, 0.3, 1, 3, 10\}$ as summarised in Table 1.

A.5 Algorithm for $\boldsymbol{L}q$-decreasing tree construction

Given a locally observable game $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ and an outcome distribution $q \in \Delta(d)$, we denote $\ell_a := e_a^\top \boldsymbol{L} q$ for the loss of action a under the outcome distribution q. Since a partial monitoring game can be represented as a neighborhood graph $(\Pi, \mathcal{E})$, we can construct a spanning in-tree $\mathcal{T} \subseteq \mathcal{E}$ rooted at the action with the smallest loss $\hat{a} \in \arg\min_{a \in \Pi} \ell_a$. The following algorithm constructs an $\boldsymbol{L}q$-decreasing in-tree by selecting each action's best neighbor to be its parent, then constructing the in-tree $\mathcal{T}$ with the selected neighbors as edges.

Algorithm 5 Constructing an $\boldsymbol{L}q$-decreasing in-tree by selecting each action's parent

Input: A locally observable game $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$, its neighborhood graph $(\Pi, \mathcal{E})$, and $q \in \Delta(d)$.
Output: A spanning in-tree $\mathcal{T} \subseteq \mathcal{E}$ with root $\mathrm{root}_\mathcal{T} = \hat{a}$ such that $\boldsymbol{L}q$ is $\mathcal{T}$-decreasing.
For each $a \in \Pi$, compute $\ell_a = e_a^\top \boldsymbol{L} q$, and choose the root $\hat{a} \in \arg\min_{a \in \Pi} \ell_a$.
Choose $q^\circ \in \mathrm{ri}(\mathcal{C}_{\hat{a}})$, which is the relative interior of the cell $\mathcal{C}_{\hat{a}}$.
Initialize $\mathcal{T} \leftarrow \varnothing$.
for each $a \in \Pi \setminus \{\hat{a}\}$ **do**
    Find the best neighbors $B_a \leftarrow \arg\min_{b:\,(a,b) \in \mathcal{E}} \ell_b$.
    Choose $\mathrm{par}_\mathcal{T}(a) \in \arg\min_{b \in B_a} e_b^\top \boldsymbol{L} q^\circ$, breaking any remaining ties arbitrarily.
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(a, \mathrm{par}_\mathcal{T}(a))\}$.
end for
return the in-tree $\mathcal{T}$ with $\mathrm{root}_\mathcal{T} = \hat{a}$.

The proof of such an $\boldsymbol{L}q$-decreasing in-tree construction is provided in Lattimore & Szepesvári (2020, Lemma 19).

A.6 Auxiliary Lemmata

Lemma A.8 (Pareto optimal comparators suffice)**.** For every $q \in \Delta(d)$,

$$\min_{\lambda \in \Delta(\Pi)} \lambda^\top \boldsymbol{L} q = \min_{a \in \Pi} e_a^\top \boldsymbol{L} q = \min_{a \in [k]} e_a^\top \boldsymbol{L} q = \min_{\lambda \in \Delta(k)} \lambda^\top \boldsymbol{L} q.$$

Proof. A linear function on a simplex attains its minimum at a vertex, which gives the outer two equalities, so it suffices to show $\bigcup_{a \in \Pi} \mathcal{C}_a = \Delta(d)$. The cells cover $\Delta(d)$ by Definition 2.1, and $\bigcup_{a \notin \Pi} \mathcal{C}_a$ is a finite union of polytopes of dimension at most $d-2$, hence has empty interior relative to the $(d-1)$-dimensional set $\Delta(d)$. So $\bigcup_{a \in \Pi} \mathcal{C}_a$ is dense in $\Delta(d)$; being a finite union of closed sets it is closed, and therefore equals $\Delta(d)$. □

Lemma A.4 (Decomposition of coverage about a reference action)**.** Let $\mathcal{G} = (\boldsymbol{L}, \boldsymbol{\Phi})$ be an arbitrary partial monitoring game and let $p, \lambda \in \Delta(k)$. Then for every reference action $a_0 \in [k]$,

$$\begin{aligned}
&\mathrm{Coverage}(p, \lambda) \\
&\le \underbrace{8 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\sum_{b=1}^k \lambda_b \left(L_b - L_{a_0} \right) (q_1 - q_2) \right)^2}{\sum_{a=1}^k p_a \| \boldsymbol{S}_a (q_1 - q_2) \|_1^2}}_{=: \mathrm{Cov}_1(p, \lambda, a_0)} + \underbrace{8 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\sum_{b=1}^k p_b \left(L_b - L_{a_0} \right) (q_1 - q_2) \right)^2}{\sum_{a=1}^k p_a \| \boldsymbol{S}_a (q_1 - q_2) \|_1^2}}_{=: \mathrm{Cov}_2(p, a_0)},
\end{aligned}$$

with the convention $0/0 = 0$.

Proof.

$$\begin{aligned}
&\mathrm{Coverage}(p, \lambda) \\
&= \sup_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda - p)^\top L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&\le \sup_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda - p)^\top L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&= \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\lambda - e_{a_0} \right)^\top L (q_1 - q_2) - \left(p - e_{a_0} \right)^\top L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&\le 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\lambda - e_{a_0} \right)^\top L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} + 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(p - e_{a_0} \right)^\top L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&= 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k \lambda_b L_b (q_1 - q_2) \right) - L_{a_0} (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} + 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k p_b L_b (q_1 - q_2) \right) - L_{a_0} (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&= 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k \lambda_b (L_b - L_{a_0}) (q_1 - q_2) \right) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} + 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k p_b (L_b - L_{a_0}) (q_1 - q_2) \right) \right)^2}{\mathbb{E}_{a \sim p} \left[D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right) \right]} \\
&= 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k \lambda_b (L_b - L_{a_0}) (q_1 - q_2) \right) \right)^2}{\sum_{a=1}^k p_a D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right)} + 2 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left(\left(\sum_{b=1}^k p_b (L_b - L_{a_0}) (q_1 - q_2) \right) \right)^2}{\sum_{a=1}^k p_a D_{\mathrm{H}}^2 \left(S_a q_1 \,\|\, S_a q_2 \right)}
\end{aligned}$$

$$\leq \underbrace{8 \max_{q_1,q_2} \frac{\left(\left(\sum_{b=1}^k \lambda_b (L_b - L_{a_0})(q_1 - q_2)\right)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1 - q_2)\|_1^2}}_{\mathrm{Cov}_1} + \underbrace{8 \max_{q_1,q_2} \frac{\left(\left(\sum_{b=1}^k p_b (L_b - L_{a_0})(q_1 - q_2)\right)\right)^2}{\sum_{a=1}^k p_a \|S_a(q_1 - q_2)\|_1^2}}_{\mathrm{Cov}_2}$$

(because $\frac{1}{4}\|S_a(q_1 - q_2)\|_1^2 \leq D_{\mathrm{H}}^2\left(S_a q_1 \,\|\, S_a q_2\right)$, Lemma A.9)

The right-hand side depends on neither x nor $\mathcal{Q}$, and the root is arbitrary. So the bound holds uniformly over contexts and model classes, and the root may be chosen freely by the caller. □

Lemma A.5 (Witness bound for the coverage terms)**.** Let $a_0 \in [k]$ and let $\mu, p \in \Delta(k)$. Suppose that for every $b \in [k]$ there is a function $h_{(b,a_0)} : [k] \times \Sigma \to \mathbb{R}$ such that

$$L_{bo} - L_{a_0 o} = \sum_{c=1}^k h_{(b,a_0)}(c, \Phi_{co}), \qquad \text{for all } o \in [d],$$

and write $\|h_{(b,a_0)}(c,\cdot)\|_\infty := \max_{\sigma\in\Sigma}|h_{(b,a_0)}(c,\sigma)|$. Assume $p_c > 0$ whenever $h_{(b,a_0)}(c,\cdot) \not\equiv 0$ for some b. Then

$$\sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \left(L_b - L_{a_0}\right)(q_1 - q_2)\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \leq \left(\sum_{b=1}^k \mu_b \sqrt{\sum_{c=1}^k \frac{\|h_{(b,a_0)}(c,\cdot)\|_\infty^2}{p_c}}\right)^2 \leq \sum_{b=1}^k \mu_b \sum_{c=1}^k \frac{\|h_{(b,a_0)}(c,\cdot)\|_\infty^2}{p_c},$$

with the convention $0/0 = 1$.

Proof.

$$\begin{aligned}
&\sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \left(L_b - L_{a_0}\right)(q_1 - q_2)\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b (L_b - L_{a_0})(q_1 - q_2)\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{o\in[d]} (L_{bo} - L_{a_0 o})(q_1[o] - q_2[o])\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&\qquad\qquad (q_1[o] \text{ denotes the } o\text{-th index of } q_1) \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{o\in[d]} (\sum_{c\in[k]} h_{(b,a_0)}(c, \Phi_{co}))(q_1[o] - q_2[o])\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{o\in[d]} (\sum_{c\in[k]} \sum_{\sigma\in\Sigma} h_{(b,a_0)}(c,\sigma) \boldsymbol{S}_c[\sigma, o])(q_1[o] - q_2[o])\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&\qquad\qquad (\text{By definition } \boldsymbol{S}_c[\sigma, o] = \mathbb{1}\{\Phi_{co} = \sigma\}) \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{c\in[k]} \sum_{\sigma\in\Sigma} h_{(b,a_0)}(c,\sigma) \sum_{o\in[d]} \boldsymbol{S}_c[\sigma, o](q_1[o] - q_2[o])\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{c\in[k]} \sum_{\sigma\in\Sigma} h_{(b,a_0)}(c,\sigma) \left(\boldsymbol{S}_c(q_1 - q_2)\right)[\sigma]\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \\
&\leq \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{c\in[k]} \|h_{(b,a_0)}(c,:)\|_\infty \|\boldsymbol{S}_c(q_1 - q_2)\|_1\right)^2}{\sum_{a=1}^k p_a \|\boldsymbol{S}_a(q_1 - q_2)\|_1^2} \quad \text{(Holder's inequality w.r.t } (1,\infty))
\end{aligned}$$

$$= \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sum_{c\in[k]} \frac{\|h_{(b,a_0)}(c,:)\|_\infty}{\sqrt{p_c}} \left(\sqrt{p_c}\|\boldsymbol{S}_c(q_1-q_2)\right)\|_1\right)^2}{\sum_{a=1}^k p_a\|\boldsymbol{S}_a(q_1-q_2)\|_1^2}$$

$$\leq \sup_{q_1,q_2\in\Delta(d)} \frac{\left(\sum_{b=1}^k \mu_b \sqrt{\sum_{c\in[k]} \frac{\|h_{(b,a_0)}(c,:)\|_\infty^2}{p_c}} \sqrt{\sum_{c\in[k]} p_c\|\boldsymbol{S}_c(q_1-q_2)\|_1^2}\right)^2}{\sum_{a=1}^k p_a\|\boldsymbol{S}_a(q_1-q_2)\|_1^2}$$

(Holder's inequality w.r.t $(2,2)$)

$$\leq \left(\sum_{b=1}^k \mu_b \sqrt{\sum_{c\in[k]} \frac{\|h_{(b,a_0)}(c,:)\|_\infty^2}{p_c}}\right)^2$$

$$= \sum_{b=1}^k \sum_{c=1}^k \frac{\mu_b}{p_c}\|h_{(b,a_0)}(c,:)\|_\infty^2. \qquad \text{(Jensen's Inequality)}$$

□

Theorem A.1 (Regret concentration generalized from Qin & Zhang (2026a))**.** Let Λ be a set of policies, $\mathcal{L}(\cdot \mid x) : \Lambda \to [0,1]$ a value function and $\mathcal{L}^\star(\cdot \mid x)$ and $\widehat{\mathcal{L}}_\ell(\cdot \mid x)$, $\ell \in [L]$ are value functions measured under the true model and estimated models respectively. $\lambda^\star \in \Lambda$ and $\hat{\lambda}_\ell \in \Lambda$ are respectively the optimal policies under respective true and estimated models. Let $C_\ell : \Lambda \times \Lambda \to \mathbb{R}_+$ be coverage maps, $p_\ell \in \Lambda$ are decisions, $0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_L$ learning rates, and $V_\ell, E_\ell \geq 0$ numbers such that,

(C1) (certificate) for every $\ell \geq 1$ and every $\lambda \in \Lambda$,

$$\widehat{\mathcal{L}}_\ell(p_\ell \mid x) - \widehat{\mathcal{L}}_\ell(\lambda \mid x) + \frac{1}{\gamma_\ell} C_\ell(p_\ell, \lambda) \leq V_\ell.$$

(C2) (off-policy evaluation) for every $\ell \geq 2$ and every $\lambda \in \Lambda$,

$$\left|\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)\right| \leq \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.$$

Define $G_1 \coloneqq 1$ and, for $\ell \geq 2$,

$$\gamma_\ell G_\ell \coloneqq \gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell, \qquad \text{i.e.} \qquad G_\ell = \frac{1}{\gamma_\ell}\left[\gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i)\right].$$

Then, for every $\ell \in [L]$ and every $\lambda \in \Lambda$,

$$\mathrm{Reg}(\lambda \mid x) \leq 2\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + G_\ell \qquad \text{and} \qquad \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) \leq 2\mathrm{Reg}(\lambda \mid x) + G_\ell. \tag{7}$$

Consequently $\mathrm{Reg}(p_\ell \mid x) \leq 2V_\ell + G_\ell$ for every ℓ, and

$$\sum_{\ell=1}^{L}(\tau_\ell - \tau_{\ell-1})\,\mathrm{Reg}(p_\ell \mid x) \leq \tau_1 + \sum_{\ell=2}^{L} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell}\left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell}(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i)\right]. \tag{8}$$

Proof. We prove Eq. equation 7 by induction on ℓ. For $\ell = 1$ both inequalities hold with $G_1 = 1$, since Reg and $\widehat{\mathrm{Reg}}_1$ take values in $[0,1]$. Fix $\ell \geq 2$ and assume Eq. equation 7 to hold at $\ell - 1$.

First, we establish that, for every λ, our policy and estimator guarantees Low Regret (LR) and Good Coverage(GC).

That is, from (C1) for epoch $\ell - 1$, we have

$$\widehat{\mathcal{L}}_{\ell-1}(p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_{\ell-1}(\lambda \mid x) + \frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) \leq V_{\ell-1}$$

$$\widehat{\mathcal{L}}_{\ell-1}(p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_{\ell-1}(\hat{\lambda}_{\ell-1} \mid x) - (\widehat{\mathcal{L}}_{\ell-1}(\lambda \mid x) - \widehat{\mathcal{L}}_{\ell-1}(\hat{\lambda}_{\ell-1} \mid x)) + \frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) \leq V_{\ell-1}$$

$$\widehat{\mathrm{Reg}}_{\ell-1}(p_{\ell-1} \mid x) + \left(\frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) - \widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x) \right) \le V_{\ell-1}$$

Above statement is true for all λ, including the the case where $\widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x) = 0$ and $C_\ell(p_{\ell-1}, \lambda)$ is always ≥ 0. Hence,

$$\widehat{\mathrm{Reg}}_{\ell-1}(p_{\ell-1} \mid x) \le V_{\ell-1}. \qquad \text{(Low Regret (LR))}$$

Also,

$$\left(\frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) - \widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x) \right) \le V_{\ell-1}.$$

Hence,

$$C_{\ell-1}(p_{\ell-1}, \lambda) \le \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x) \right]. \qquad \text{(Good Coverage (GC))}$$

Moreover,

$$\begin{aligned}
&\mathrm{Reg}(\lambda \mid x) - \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) \\
&= \mathcal{L}^\star(\lambda \mid x) - \mathcal{L}^\star(\lambda^* \mid x) - \left(\widehat{\mathcal{L}}_\ell(\lambda \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell \mid x) \right) \\
&= \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x) - \left(\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) \right) \\
&= \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \left(\mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) \right) \\
&\le \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \left(\mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\lambda^* - p_{\ell-1} \mid x) \right) \\
&\qquad (\widehat{\mathcal{L}}_\ell(\lambda^* \mid x) \ge \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell \mid x)) \\
&\le |\mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x)| + |\mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\lambda^* - p_{\ell-1} \mid x)| \\
&\le \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda^*) + \frac{3\gamma_\ell}{2} E_\ell \qquad \text{(C2)} \\
&\le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x) \right] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\mathrm{Reg}}_{\ell-1}(\lambda^\star \mid x) \right] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad (\text{ equation Good Coverage (GC)}) \\
&\le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + 2\mathrm{Reg}(\lambda \mid x) + G_{\ell-1} \right] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + 2\mathrm{Reg}(\lambda^\star \mid x) + G_{\ell-1} \right] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad (\text{ equation 7 true for } \ell - 1) \\
&\le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[D_{\ell-1} + 2\mathrm{Reg}(\lambda \mid x) + G_{\ell-1} \right] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + G_{\ell-1} \right] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad (\mathrm{Reg}(\lambda^\star \mid x) = 0) \\
&= \frac{1}{3\gamma_\ell} \left(\gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell \right) + \frac{\gamma_{\ell-1}}{3\gamma_\ell} \mathrm{Reg}(\lambda \mid x) \\
&\le \frac{1}{3} \mathrm{Reg}(\lambda \mid x) + \frac{1}{3} G_\ell \qquad (\gamma_{\ell-1} \le \gamma_\ell)
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Reg}(\lambda \mid x) - \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) &\le \frac{1}{3} \mathrm{Reg}(\lambda \mid x) + \frac{1}{3} G_\ell \\
\frac{2}{3} \mathrm{Reg}(\lambda \mid x) &\le \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + \frac{1}{3} G_\ell \\
\mathrm{Reg}(\lambda \mid x) &\le \frac{3}{2} \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + \frac{1}{2} G_\ell \\
&\le 2\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + G_\ell \qquad \text{(Step **)}
\end{aligned}$$

Furthermore,

$$
\begin{aligned}
&\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) - \mathrm{Reg}(\lambda \mid x) \\
&\quad = \widehat{\mathcal{L}}_\ell(\lambda \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell \mid x) - \big(\mathcal{L}^\star(\lambda \mid x) - \mathcal{L}^\star(\lambda^* \mid x)\big) \\
&\quad = \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) - \big(\mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x)\big) \\
&\quad = \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \Big(\widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda^* - p_{\ell-1} \mid x)\Big) \\
&\quad \le \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x) - \Big(\widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) - \mathcal{L}^\star(\hat{\lambda}_\ell - p_{\ell-1} \mid x)\Big) \\
&\qquad\qquad (\mathcal{L}^\star(\hat{\lambda}_\ell \mid x) \ge \mathcal{L}^\star(\lambda^\star \mid x)) \\
&\quad \le |\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^\star(\lambda - p_{\ell-1} \mid x)| + |\widehat{\mathcal{L}}_\ell(\hat{\lambda}_\ell - p_{\ell-1} \mid x) - \mathcal{L}^\star(\hat{\lambda}_\ell - p_{\ell-1} \mid x)| \\
&\quad \le \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \hat{\lambda}_\ell) + \frac{3\gamma_\ell}{2} E_\ell \qquad \text{(C2)} \\
&\quad \le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + \widehat{\mathrm{Reg}}_{\ell-1}(\lambda \mid x)\Big] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + \widehat{\mathrm{Reg}}_{\ell-1}(\hat{\lambda}_\ell \mid x)\Big] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad\qquad (\text{ equation Good Coverage (GC)}) \\
&\quad \le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + 2\mathrm{Reg}(\lambda \mid x) + G_{\ell-1}\Big] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + 2\mathrm{Reg}(\hat{\lambda}_\ell \mid x) + G_{\ell-1}\Big] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad\qquad (\text{ equation 7 true for } \ell - 1) \\
&\quad \le \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + 2\mathrm{Reg}(\lambda \mid x) + G_{\ell-1}\Big] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + 4\widehat{\mathrm{Reg}}_\ell(\hat{\lambda}_\ell \mid x) + 2G_\ell + G_{\ell-1}\Big] \\
&\qquad + \frac{3\gamma_\ell}{2} E_\ell \qquad\qquad (\text{ equation 7 true for } \ell \text{ (Step **)}) \\
&\quad = \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \Big[V_{\ell-1} + 2\mathrm{Reg}(\lambda \mid x) + G_{\ell-1}\Big] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} \left[V_{\ell-1} + 2G_\ell + G_{\ell-1}\right] + \frac{3\gamma_\ell}{2} E_\ell \\
&\qquad\qquad (\widehat{\mathrm{Reg}}_\ell(\hat{\lambda}_\ell \mid x) = 0) \\
&\quad = \frac{1}{3\gamma_\ell} \left(\gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell\right) + \frac{\gamma_{\ell-1}}{3\gamma_\ell} \mathrm{Reg}(\lambda \mid x) + \frac{\gamma_{\ell-1}}{3\gamma_\ell} G_\ell \\
&\quad \le \frac{1}{3} \mathrm{Reg}(\lambda \mid x) + \frac{2}{3} G_\ell \qquad\qquad (\gamma_{\ell-1} \le \gamma_\ell)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) - \mathrm{Reg}(\lambda \mid x) &\le \frac{1}{3} \mathrm{Reg}(\lambda \mid x) + \frac{2}{3} G_\ell \\
\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) &\le \frac{4}{3} \mathrm{Reg}(\lambda \mid x) + \frac{2}{3} G_\ell \\
\mathrm{Reg}(\lambda \mid x) &\le \frac{3}{2} \widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + \frac{1}{2} G_\ell \\
&\le 2\widehat{\mathrm{Reg}}_\ell(\lambda \mid x) + G_\ell
\end{aligned}
$$

Finally, we prove the consequences. Apply (C1) at $\lambda = \hat{\lambda}_\ell$, we have

$$
\begin{aligned}
&\widehat{\mathrm{Reg}}_\ell(p_\ell \mid x) \le V_\ell && \text{(equation Low Regret (LR))} \\
&\Rightarrow \mathrm{Reg}(p_\ell \mid x) \le 2\widehat{\mathrm{Reg}}_\ell(p_\ell \mid x) + G_\ell \le 2V_\ell + G_\ell. && \text{(By equation 7)}
\end{aligned}
$$

Summing over epochs, we have

$$
\sum_{\ell=1}^{L} (\tau_\ell - \tau_{\ell-1}) \mathrm{Reg}(p_\ell \mid x) \le \tau_1 + \sum_{\ell=2}^{L} (\tau_\ell - \tau_{\ell-1})(2V_\ell + G_\ell) \qquad (\mathrm{Reg}(p_1 \mid x) \le 1)
$$

$$= \tau_1 + \sum_{\ell=2}^{L} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right],$$

(By the closed form of G_ℓ)

which is Eq. equation 8. □

Lemma A.9. Let P, Q are two probability simplices. $D_{\mathrm{H}}(P\|Q)$, $D_{\mathrm{KL}}(P\|Q)$, $\|P-Q\|_{\mathrm{TV}}$, $\|P-Q\|_1$ be the *Hellinger* distance, *Kullback* divergence, *total variation* distance and level-1 norm between P, Q respectively, then the following holds,

$$\frac{\|P-Q\|_1}{2} = \|P-Q\|_{\mathrm{TV}} \le D_{\mathrm{H}}(P\|Q)$$

and,

$$\frac{\|P-Q\|_1}{2} = \|P-Q\|_{\mathrm{TV}} \le \sqrt{D_{\mathrm{KL}}(P\|Q)}$$

Proof.

$$\|P-Q\|_{\mathrm{TV}} \le D_{\mathrm{H}}(P\|Q) \qquad \text{(Tsybakov (2008), Lemma 2.3)}$$

and,

$$\|P-Q\|_{\mathrm{TV}} \le \sqrt{D_{\mathrm{KL}}(P\|Q)} \qquad \text{(Tsybakov (2008), Lemma 2.5)}$$

Furthermore,

$$\frac{\|P-Q\|_1}{2} = \|P-Q\|_{\mathrm{TV}} \qquad \text{(Levin \& Peres (2017), Proposition 4.2)}$$

□

Lemma A.10 (Local observability, Lattimore & Szepesvári (2020))**.** If $\mathcal{G} = (\Phi, \boldsymbol{L})$ is locally observable and non-degenerate and actions $(a, b) \in \mathcal{E}$ are neighbours, then there exist functions $w_a, w_b : \Sigma \to \mathbb{R}$ such that $\|w_a\|_\infty \le m/2$ and $\|w_b\|_\infty \le m/2$ and

$$\boldsymbol{L}_{ax} - \boldsymbol{L}_{bx} = w_a(\Phi_{ax}) + w_b(\Phi_{bx}) \text{ for all } x \in [d]\,.$$

In Lemma A.10, the assumption of non-degeneracy is important here for $\|w_a\|_\infty \le m/2$ and $\|w_b\|_\infty \le m/2$ to hold. The Proposition A.11 bounds $\|w_e\|_\infty$ for more general case.

Lemma A.11 (Lattimore & Szepesvári (2020))**.** Suppose that (L, Φ) is globally observable and a and b are neighbours. Then there exists a function $w_e : [k] \times \Sigma \to \mathbb{R}$ such that for all $x \in [d]$,

$$\sum_{c=1}^{k} w_e(c, \Phi_{cx}) = L_{ax} - L_{bx}\,.$$

Furthermore, w_e can be chosen so that $\|w_e\|_\infty \le d^{1/2} k^{d/2}$.

Lemma A.12 (Water-transfer operator adapted from Lattimore & Szepesvári (2020))**.** Suppose that $\mathcal{G}$ is non-degenerate and locally observable and $q \in \Delta(d)$ Assume that we have an in-tree $\mathcal{T}$ that is Lq-decreasing. Then there exists a function $W_q^{\mathcal{T}} : \Delta(\Pi) \to \Delta(\Pi)$ such that the following hold for all $p \in \Delta(\Pi)$:

(a) The expected loss does not increase: $\left(W_q(p) - p\right)^\top \boldsymbol{L} q \le 0$.

(b) Action probabilities are not too small: $W_q(p)[a] \ge p[a]/k$ for all $a \in [k]$.

(c) Probabilities increase towards the root of the in-tree $\mathcal{T}$. That is, $W_q(p)[a] \le W_q(p)[b]$ for all $e = (a, b) \in \mathcal{T}$.