Algorithm 4 BINSEARCH.UNBOUNDED.HIGH

Input: context-action pair (x, a), history H, radius $\beta > 0$, and precision $\alpha > 0$
Let $R(f) \coloneqq \sum_{(x',a',r')\in H}(f(x',a') - r')^2$.
Let $\widetilde{R}(f, w) \coloneqq R(f) + \frac{w}{2}(f(x,a) - 2)^2$
$w_{\mathrm{L}} \leftarrow 0,\ w_{\mathrm{H}} \leftarrow \beta/\alpha$
`// Invoke oracle twice`
$f_{\mathrm{L}} \leftarrow \arg\min_{f\in\mathcal{F}} \widetilde{R}(f, w_{\mathrm{L}}),\ z_{\mathrm{L}} \leftarrow f_{\mathrm{L}}(x,a)$
$f_{\mathrm{H}} \leftarrow \arg\min_{f\in\mathcal{F}} \widetilde{R}(f, w_{\mathrm{H}}),\ z_{\mathrm{H}} \leftarrow f_{\mathrm{H}}(x,a)$
$R_{\min} \leftarrow R(f_{\mathrm{L}})$
if $z_{\mathrm{L}} \geq 1$ or $R(f_{\mathrm{L}}) = R(f_{\mathrm{H}})$ **then return** 1
$\Delta \leftarrow \alpha\beta/(2 - z_{\mathrm{L}})^3$
while $|z_{\mathrm{H}} - z_{\mathrm{L}}| > \alpha$ and $|w_{\mathrm{H}} - w_{\mathrm{L}}| > \Delta$ **do**
    $w \leftarrow (w_{\mathrm{H}} + w_{\mathrm{L}})/2$
    `// Invoke oracle.`
    $f \leftarrow \arg\min_{\tilde{f}\in\mathcal{F}} \widetilde{R}(\tilde{f}, w),\ z \leftarrow f(x,a)$
    if $R(f) \geq R_{\min} + \beta$ **then**
        $w_{\mathrm{H}} \leftarrow w,\ z_{\mathrm{H}} \leftarrow z$
    else
        $w_{\mathrm{L}} \leftarrow w,\ z_{\mathrm{L}} \leftarrow z$
    end if
end while
return $\min\{z_{\mathrm{H}}, 1\}$.

A. Proofs

A.1. Proofs from Section 3.1

We prove the statement of Theorem 1 for BINSEARCH.UNBOUNDED.HIGH (Algorithm 4), which does not require the predictors in $\mathcal{F}$ to be bounded in $[0, 1]$. Note however that the actual rewards are still always bounded in $[0, 1]$, so that $f^\star(x, a)$ is always bounded by the realizability assumption. Compared with Algorithm 3, the algorithm includes some handling of special cases, which are automatically excluded in Algorithm 3 by the assumption about boundedness. The performance guarantee for BINSEARCH.UNBOUNDED.LOW (Algorithm 5) is analogous and therefore is omitted.

Lemma 1. Let $\mathcal{F}$ be convex and closed under pointwise convergence. Consider a run of Algorithm 4. Let $R(f)$ and $R_{\min}$ be defined as in Algorithm 4 and let

$$z^\star \coloneqq \max\big\{f(x,a) :\ f \in \mathcal{F} \text{ such that } R(f) \leq R_{\min} + \beta\big\}\ .$$

Then Algorithm 4 returns z such that $\big|z - \min\{z^\star, 1\}\big| \leq \alpha$ after at most $O\big(\log(1/\alpha) + \log(\max\{2 - z_0, 1\})\big)$ iterations, where $z_0 = f_{\min}(x,a)$ and $f_{\min} = \arg\min_f R(f)$.

Corollary 1. If $f(x,a) \in [0,1]$ for all $f \in \mathcal{F}$, $x \in \mathcal{X}$ and $a \in \mathcal{A}$, then Algorithm 4 returns z such that $|z - z^\star| \leq \alpha$ after at most $O(\log(1/\alpha))$ iterations.

Proof. The proof works by analyzing a univariate auxiliary function $\phi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, which maps $z \in \mathbb{R}$ to the smallest empirical error $R(f)$ among all functions that predict $f(x,a) = z$,

$$\phi(z) \coloneqq \begin{cases} \infty & \text{if } f(x,a) < z \text{ for all } f \in \mathcal{F} \\ \min\big\{R(f) :\ f \in \mathcal{F} \text{ and } f(x,a) = z\big\} & \text{otherwise.} \end{cases} \tag{2}$$

note that we do not need to worry about the case when $f(x,a)$ might take values both larger and smaller than z but not z exactly due to the assumed convexity of $\mathcal{F}$. We first show that this function is well-defined (i.e., the minimum in the definition is attained), convex and lower semicontinuous. We begin by embedding the least-squares optimization in a finite dimensional space. Let $H = \{(x_i, a_i, r_i)\}_{i=1}^n$ and define $x_{n+1} \coloneqq x$ and $a_{n+1} \coloneqq a$. We associate each f with a vector

Algorithm 5 BINSEARCH.UNBOUNDED.LOW

Input: context-action pair (x, a), history H, radius $\beta > 0$, and precision $\alpha > 0$
Let $R(f) \coloneqq \sum_{(x',a',r')\in H}(f(x',a') - r')^2$.
Let $\widetilde{R}(f, w) \coloneqq R(f) + \frac{w}{2}(f(x,a)+1)^2$
$w_{\mathrm{L}} \leftarrow 0,\ w_{\mathrm{H}} \leftarrow \beta/\alpha$
`// Invoke oracle twice`
$f_{\mathrm{L}} \leftarrow \arg\min_{f\in\mathcal{F}} \widetilde{R}(f, w_{\mathrm{L}}),\ z_{\mathrm{L}} \leftarrow f_{\mathrm{L}}(x,a)$
$f_{\mathrm{H}} \leftarrow \arg\min_{f\in\mathcal{F}} \widetilde{R}(f, w_{\mathrm{H}}),\ z_{\mathrm{H}} \leftarrow f_{\mathrm{H}}(x,a)$
$R_{\min} \leftarrow R(f_{\mathrm{L}})$
if $z_{\mathrm{L}} \le 0$ or $R(f_{\mathrm{L}}) = R(f_{\mathrm{H}})$ **then return** 0
$\Delta \leftarrow \alpha\beta/(1+z_{\mathrm{L}})^3$
while $|z_{\mathrm{H}} - z_{\mathrm{L}}| > \alpha$ and $|w_{\mathrm{H}} - w_{\mathrm{L}}| > \Delta$ **do**
  $w \leftarrow (w_{\mathrm{H}} + w_{\mathrm{L}})/2$
  `// Invoke oracle.`
  $f \leftarrow \arg\min_{\tilde{f}\in\mathcal{F}} \widetilde{R}(\tilde{f}, w),\ z \leftarrow f(x,a)$
  if $R(f) \ge R_{\min} + \beta$ **then**
    $w_{\mathrm{H}} \leftarrow w,\ z_{\mathrm{H}} \leftarrow z$
  else
    $w_{\mathrm{L}} \leftarrow w,\ z_{\mathrm{L}} \leftarrow z$
  end if
end while
return $\max\{z_{\mathrm{H}}, 0\}$.

$\mathbf{v}^f \in \mathbb{R}^{n+1}$ with entries $v_i^f = f(x_i, a_i)$. Let $\mathcal{V} \coloneqq \{\mathbf{v}^f : f \in \mathcal{F}\}$. Since $\mathcal{F}$ is closed under pointwise convergence and convex, the set $\mathcal{V}$ must also be closed and convex.

For $\mathbf{v} \in \mathbb{R}^{n+1}$, let

$$\rho(\mathbf{v}) \coloneqq \sum_{i=1}^{n}(v_i - r_i)^2 \ ,$$

where r_i are the rewards from H. Thus,

$$R(f) = \sum_{i=1}^{n}(f(x_i,a_i) - r_i)^2 = \rho(\mathbf{v}^f) \ ,$$

and therefore

$$\phi(z) = \min\{R(f) : f \in \mathcal{F} \text{ and } f(x) = z\} = \min\{\rho(\mathbf{v}) : \mathbf{v} \in \mathcal{V} \text{ and } v_{n+1} = z\} \ ,$$

where we use the convention that the minimum of an empty set equals ∞. The attainment of the minimum now follows by convexity and continuity of ρ along the affine space $\{v_{n+1} = z\}$. The convexity and lower semicontinuity of ϕ follows by Theorem 9.2 of Rockafellar (1970).

The upper confidence value $z^\star$ is then the largest z for which $\phi(z) \le R_{\min} + \beta$:

$$z^\star = \max\{z : \phi(z) \le R_{\min} + \beta\} \ .$$

Furthermore, for any $w \ge 0$, define

$$z_w \coloneqq \arg\min_{z\in\mathbb{R}}\left[\phi(z) + \frac{w}{2}(2-z)^2\right] \ .$$

Thus, $z_w = f(x,a)$ where $f = \arg\min_{\tilde{f}\in\mathcal{F}} \widetilde{R}(\tilde{f}, w)$ with $\widetilde{R}$ as defined in the algorithm. The algorithm maintains the identities $z_{\mathrm{L}} = z_{w_{\mathrm{L}}}$ and $z_{\mathrm{H}} = z_{w_{\mathrm{H}}}$, so it can be rewritten as follows:

if $z_0 \ge 1$ or $\phi(z_0) = \phi(z_{\beta/\alpha})$ **then return** 1
$w_{\mathrm{L}} \leftarrow 0,\ w_{\mathrm{H}} \leftarrow \beta/\alpha,\ \Delta \leftarrow \alpha\beta/(2-z_0)^3$
while $|z_{w_{\mathrm{H}}} - z_{w_{\mathrm{L}}}| > \alpha$ and $|w_{\mathrm{H}} - w_{\mathrm{L}}| > \Delta$ **do**

```
    w ← (w_H + w_L)/2
    if φ(z_w) > φ(z_0) + β then
        w_H ← w
    else
        w_L ← w
    end if
end while
return min{z_{w_H}, 1}.
```

Note that $z_0 = f_{\min}(x, a)$ where $f_{\min}$ is the minimizer of R, and therefore ϕ attains its minimum at z_0. If $z_0 \geq 1$, then the algorithm terminates and returns 1. Since $z^\star \geq z_0 \geq 1$, in this case the lemma holds.

Also, note that if $z^\star = z_0 < 1$ then the algorithm immediately terminates with $z_{w_\mathrm{L}} = z_{w_\mathrm{H}} = z_0$. This is because of the fact that $z^\star = z_0$, given $\beta > 0$, implies by lower semicontinuity that $\phi(z) = \infty$ for all $z > z_0$ and thus $z_w = z_0$ for all $w > 0$.

The final special case to consider is when $\phi(2) = \phi(z_0)$, i.e., there exist a minimizer $\tilde{f}_{\min}$ of R, which satisfies $\tilde{f}_{\min}(x, a) = 2$ and thus for any w, it also minimizes $\tilde{R}(f, w)$. This is exactly the case when $R(f_\mathrm{L}) = R(f_\mathrm{H})$ in Algorithm 4 and in this case the algorithm returns 1 and the lemma holds.

In the remainder of the proof we assume that $\phi(2) > \phi(z_0)$, $z_0 < 1$ and $z_0 < z^\star$. By convexity of ϕ, we know that ϕ is non-decreasing on $[z_0, \infty)$, and we will argue that by performing the binary search over w, the algorithm is also performing a binary search over z_w to find the point $z^\star$.

We begin by characterizing z_w and showing that $z_w < 2$ for all w. For any $w > 0$, by first-order optimality,

$$\phi'(z_w) - w(2 - z_w) = 0 \tag{3}$$

for some $\phi'(z_w) \in \partial\phi(z_w)$, where $\partial\phi$ denotes the subdifferential. First, note that $z_w \geq z_0$, because at any $z < z_0 \leq 1$, we have $w(2 - z) > 0$ while also $\phi'(z) \leq 0$, because ϕ is convex and minimized at z_0. Therefore, at $z < z_0$, we have $\phi'(z) - w(2 - z) < 0$, so Eq. (3) can only be satisfied by $z_w \geq z_0$. Rearranging, we obtain

$$w = \frac{\phi'(z_w)}{2 - z_w} \ . \tag{4}$$

Since $z_w \geq z_0$, the convexity of ϕ implies that $\phi'(z_w) \geq 0$. Since $w > 0$, we therefore must in fact have $\phi'(z_w) > 0$ and

$$z_w < 2 \text{ for all } w > 0. \tag{5}$$

Eq. (4) now implies that z_w is non-decreasing as a function of w.

Let $w^\star$ be such that $z_{w^\star} = z^\star$ (this can be obtained by Eq. 4). The remainder of the proof proceeds in two steps. The first step establishes that our initial setting $w_\mathrm{H} = \beta/\alpha$ is large enough to guarantee that the initial interval $[z_{w_\mathrm{L}}, z_{w_\mathrm{H}} + \alpha] = \left[z_0, z_{\beta/\alpha} + \alpha\right]$ contains the solution $\min\{z^\star, 1\}$. The execution of the algorithm then continues to maintain this condition, i.e., $\min\{z^\star, 1\} \in [z_{w_\mathrm{L}}, z_{w_\mathrm{H}} + \alpha]$, which we refer to as the *invariant*, while halving $|w_\mathrm{H} - w_\mathrm{L}|$. That the invariant holds can be seen as follows: First, if $z_0 \leq z^\star \leq z_{\beta/\alpha}$, then the update rule guarantees that $z_{w_\mathrm{L}} \leq z^\star \leq z_{w_\mathrm{H}}$ for every iteration. On the other hand, if $z^\star > z_{\beta/\alpha}$, then $z_{w_\mathrm{H}} = z_{\beta/\alpha}$ for every iteration, and so Step 1 below guarantees that $z^\star \in \left[z_{\beta/\alpha}, z_{\beta/\alpha} + \alpha\right] \supseteq [z_{w_\mathrm{L}}, z_{w_\mathrm{H}} + \alpha]$.

The algorithm terminates after at most

$$\log_2\left(\frac{\beta/\alpha}{\Delta}\right) = \log_2\left(\frac{(2 - z_0)^3}{\alpha^2}\right) = O\big(\log(1/\alpha) + \log(2 - z_0)\big)$$

iterations. If the reason for termination is that $|z_\mathrm{H} - z_\mathrm{L}| \leq \alpha$ then the lemma follows, thanks to the invariant. Otherwise, we must have $|w_\mathrm{H} - w_\mathrm{L}| \leq \Delta$, so our invariant together with the monotonicity of z_w in w implies that $w_\mathrm{H} \leq w^\star + \Delta$. Our second step below establishes that in this case we must also have $z_\mathrm{H} \leq z^\star + \alpha$. Our invariant separately also implies that $\min\{z^\star, 1\} \leq z_\mathrm{H} + \alpha$, so altogether we have $\min\{z_\mathrm{H}, 1\} - \alpha \leq \min\{z^\star, 1\} \leq \min\{z_\mathrm{H}, 1\} + \alpha$, proving the lemma. It remains to prove the two steps.

Step 1: $z_0 \le \min\{z^\star, 1\} \le z_{\beta/\alpha} + \alpha$**.** The first inequality is immediate from the definition of $z^\star$ and the fact that $z_0 < 1$. The second inequality holds if $z_{\beta/\alpha} \ge 1$, so it remains to consider $z_{\beta/\alpha} \le 1$. Let $w = \beta/\alpha$. Then by Eq. (4),

$$\frac{\beta}{\alpha} = w = \frac{\phi'(z_w)}{2 - z_w} \le \phi'(z_w) \ ,$$

where the last step follows because $z_w \le 1$. Now by convexity of ϕ, for any $\tilde{\alpha} > \alpha$

$$\phi(z_w + \tilde{\alpha}) \ge \phi(z_w) + \tilde{\alpha}\phi'(z_w) \ge \phi(z_w) + \tilde{\alpha} \cdot \frac{\beta}{\alpha} > \phi(z_0) + \beta \ ,$$

where the last step follows because $\phi(z_w) \ge \phi(z_0)$ and $\tilde{\alpha} > \alpha$. This shows that $z^\star \le z_w + \alpha$ and completes Step 1.

Step 2: $z_{w^\star + \Delta} \le z^\star + \alpha$**.** Let $w = w^\star + \Delta$. Then by convexity

$$\phi(z_0) \ge \phi(z^\star) + (z_0 - z^\star)\phi'(z^\star) \ ,$$

and since $z^\star > z_0$, we can rearrange this inequality to give

$$\phi'(z^\star) \ge \frac{\phi(z^\star) - \phi(z_0)}{z^\star - z_0} = \frac{\beta}{z^\star - z_0} \ge \frac{\beta}{2 - z_0} \ ,$$

where the last inequality follows by Eq. (5). By Eq. (4), we also have

$$w^\star = \frac{\phi'(z^\star)}{2 - z^\star} \ge \frac{\phi'(z^\star)}{2 - z_0}$$

because $z^\star > z_0$. Combining the two bounds yields

$$w^\star \ge \frac{\beta}{(2 - z_0)^2} \ . \tag{6}$$

Applying now Eq. (4) twice, and also using the monotonicity of ϕ', we obtain

$$w = \frac{\phi'(z_w)}{2 - z_w} \ge \frac{\phi'(z^\star)}{2 - z_w} = w^\star \cdot \frac{2 - z^\star}{2 - z_w} \ .$$

Therefore,

$$2 - z_w \ge \frac{w^\star}{w} \cdot (2 - z^\star)$$
$$z^\star - z_w \ge \frac{w^\star}{w} \cdot (2 - z^\star) - (2 - z^\star) \ .$$

Rearranging,

$$z_w - z^\star \le \frac{w - w^\star}{w} \cdot (2 - z^\star) = \frac{\Delta}{w} \cdot (2 - z^\star) \le \frac{\Delta}{w^\star} \cdot (2 - z_0) \ ,$$

where the final inequality uses the fact that $w \ge w^\star$ and $z^\star \ge z_0$. Finally, applying the bound (6) and the definition of Δ, we complete Step 2:

$$z_w - z^\star \le \frac{\Delta(2 - z_0)^3}{\beta} = \alpha \ .$$

□

A.2. Proof of Proposition 1

Proof of Proposition 1. Consider the following contextual bandit instance:

- Two actions a_g and a_b, so $K = 2$.
- $r_t(a_g) = 1 - \epsilon$ and $r_t(a_b) = 0$, regardless of context (there is no noise).

- N contexts $x^1, \ldots, x^N$. The context distribution $D_{\mathcal{X}}$ is uniform over these N contexts.
- Regressor class $\mathcal{F}$ contains the following $N+1$ predictors:
  - Ground truth regressor $f^\star$ defined by $f^\star(x, a_g) = 1-\epsilon,\ \forall x$ and $f^\star(x, a_b) = 0,\ \forall x$.
  - For each $i \in [N]$, f_i satisfying $f_i(x^i, a_g) = 0$, $f_i(x^i, a_b) = 1$, and $f_i(x^j, a_g) = 1-\epsilon$, $f_i(x^j, a_b) = 0$ for all $j \neq i$.

We can see that $\pi_{f^\star}$ has population reward $1-\epsilon$ and each π_{f_i} has population reward $(1-1/N)(1-\epsilon)$. Thus, each f_i has expected regret of $(1-\epsilon)/N$.

Suppose S is the set of contexts that have been observed by our algorithms at time t, and further assume $\beta_m = 0$ (as it will be clear that larger β_m can only make things worse), so that only regressors with zero square loss are considered. Observe that $f^\star \in \mathcal{F}_m$ and $f_i \in \mathcal{F}_m$ only if $x^i \notin S$.

Let x^i be the context observed at time t. If $x^i \in S$, then all regressors in $\mathcal{F}_m$ agree on it, so a_g will be played. Now, suppose $x^i \notin S$. Then we have $\textsc{High}_{\mathcal{F}_m}(x^i, a_g) = 1-\epsilon$ (obtained by $f^\star$), and $\textsc{Low}_{\mathcal{F}_m}(x^i, a_g) = 0$ (obtained by f_i). Likewise, $\textsc{High}_{\mathcal{F}_m}(x^i, a_b) = 1$ (from f_i) and $\textsc{Low}_{\mathcal{F}_m}(x^i, a_b) = 0$ (from $f^\star$).

We thus see that our algorithms will make a mistake and incur instantaneous regret of $(1-\epsilon)$ precisely at the time steps for which one of the N contexts is encountered for the first time. The regret of the algorithm after t steps can therefore be lower bounded as $\min\{N, \widetilde{\Omega}(t)\}$. □

A.3. Proofs from Section 4.1

We first set up some formal notations and then recall the definition of the disagreement coefficient for reader's convenience.

Definition 4. *For any $\varepsilon > 0$, the* policy-regret ball *of radius ε for $\mathcal{F}$ is defined as*

$$\mathcal{F}(\varepsilon) = \Big\{ f \in \mathcal{F} : \mathbb{E}[r(\pi_f(x))] \geq \mathbb{E}[r(\pi^\star(x))] - \varepsilon \Big\}.$$

Definition 5 (Reward width)**.** *For any predictor class $\mathcal{F}$, context x, and action a, the reward width is defined as*

$$W_{\mathcal{F}}(x, a) = \textsc{High}_{\mathcal{F}}(x, a) - \textsc{Low}_{\mathcal{F}}(x, a).$$

Definition 6 (Disagreement Region)**.** *For any predictor class $\mathcal{F}$, the disagreement region* $\mathrm{Dis}(\mathcal{F})$ *is defined as*[9]

$$\begin{aligned}&\mathrm{Dis}(\mathcal{F})\\ &= \Big\{ x \;\Big|\; \exists f, f' \in \mathcal{F} : \arg\max_{a \in \mathcal{A}} f(x, a) \neq \arg\max_{a \in \mathcal{A}} f'(x, a) \Big\}.\end{aligned}$$

Definition 7 (Disagreement set)**.** *For a predictor class $\mathcal{F}$ and a context x, the* disagreement set *at x is defined as*

$$A_{\mathcal{F}}(x) = \textstyle\bigcup_{f \in \mathcal{F}} \arg\max_{a \in \mathcal{A}} f(x, a).$$

With these preliminaries, the disagreement coefficient is defined as follows.

Definition (Disagreement Coefficient)**.** *The disagreement coefficient for $\mathcal{F}$ (with respect to $D_{\mathcal{X}}$) is defined as*

$$\theta_0 := \sup_{\delta > 0, \varepsilon > 0} \frac{\delta}{\varepsilon} \Pr_{D_{\mathcal{X}}} \Big[x \in \mathrm{Dis}(\mathcal{F}(\varepsilon)) \text{ and } \exists a \in A_{\mathcal{F}(\varepsilon)}(x) : W_{\mathcal{F}(\varepsilon)}(x, a) > \delta \Big].$$

In addition, the following condition on $f^\star$ is important to obtain fast rates, but it is not stated as an assumption because it is not strictly necessary for any of our algorithms.

Definition 8 (Massart noise condition)**.** *The distribution D satisfies the Massart noise condition if there exists $\gamma > 0$, called a* margin*, such that*

$$f^\star(x, \pi^\star(x)) \geq f^\star(x, a) + \gamma \quad \textit{for all } x \textit{ and } a \neq \pi^\star(x).$$

[9] When the maximizing action $\arg\max_{a \in \mathcal{A}} f(x, a)$ is not unique, the "$\neq$" in the disagreement set definition checks that the two argmax sets are identical.

For all subsequent analyses we will use the following filtration:

$$\mathcal{J}_t := \sigma((x_1, a_1, r_1), \ldots, (x_{t-1}, a_{t-1}, r_{t-1})).$$

Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{J}_t]$ and $\mathrm{Var}_t[\cdot] := \mathrm{Var}[\cdot \mid \mathcal{J}_t]$.

Lemma 2 (Freedman-type inequality e.g. (Agarwal et al., 2014))**.** For any real-valued martingale difference sequence $(Z_t)_{t\le T}$ with $|Z_t| \le R$ almost surely, it holds that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} Z_t \le \eta(e-2)\sum_{t=1}^{T}\mathbb{E}_t(Z_t)^2 + \frac{R\log(1/\delta)}{\eta} \tag{7}$$

for all $\eta \in [0, 1/R]$.

Recall that epoch schedule used by Algorithm 1 and Algorithm 2 is $\tau_m = 2^{m-1}$. Denote the length of epoch m by $T_m = \tau_{m+1} - \tau_m = 2^{m-1}$. In addition, we will use the notation $g_a^\star(x) := f^\star(x, a)$ where $f^\star$ as in the main text is the predictor that realizes the mean reward function, and also

$$M_t(g, a) = ((g(x_t) - r_t(a))^2 - (g_a^\star(x_t) - r_t(a))^2)\mathbf{1}\{a = a_t\}.$$

for any $g : \mathcal{X} \to [0, 1]$, and action $a \in \mathcal{A}$. When $f \in \mathcal{F} = \mathcal{G}^{\mathcal{A}}$ we will overload this notation by writing $M_t(f, a) := M_t(f(\cdot, a), a)$. Also define the class

$$\widetilde{\mathcal{G}}_m(\beta, a) = \left\{ g \in \mathcal{G} \mid \frac{1}{\tau_m - 1}\sum_{t=1}^{\tau_m - 1}\mathbb{E}_t[M_t(g, a)] \le \beta \right\}.$$

To prove the theorem, we make use of following lemmas.

Lemma 3. For any $g : \mathcal{X} \to [0, 1]$ and $a \in \mathcal{A}$ we have

$$\begin{aligned}\mathbb{E}_t[M_t(g, a)] &= \mathbb{E}_t\big[(g(x_t) - g_a^\star(x_t))^2\mathbf{1}\{a = a_t\}\big],\\ \mathrm{Var}_t[M_t(g, a)] &\le 4\,\mathbb{E}_t[M_t(g, a)].\end{aligned}$$

Proof. Note that a_t and r_t are conditionally independent given x_t and also $\mathbb{E}_{r_t}[r_t(a) \mid x_t] = g_a^\star(x_t)$. We thus have

$$\mathbb{E}_t[M_t(g, a)] = \mathbb{E}_t[(g(x_t) - g_a^\star(x_t))((g(x_t) + g_a^\star(x_t) - 2r_t(a))\mathbf{1}\{a = a_t\}] = \mathbb{E}_t\big[(g(x_t) - g_a^\star(x_t))^2\mathbf{1}\{a = a_t\}\big].$$

Similarly, since $((g(x_t) + g_a^\star(x_t) - 2r_t(a))^2 \le 4$ we have

$$\mathrm{Var}_t[M_t(g, a)] \le \mathbb{E}_t\big[M_t(g, a)^2\big] \le 4\,\mathbb{E}_t\big[(g(x_t) - g_a^\star(x_t))^2\mathbf{1}\{a = a_t\}\big] = 4\,\mathbb{E}_t[M_t(g, a)].$$

□

Definition 9 (Covering number)**.** *For a class $\mathcal{G}' \subseteq \{g : \mathcal{X} \to [0, 1]\}$, an empirical L_p-cover on a sequence $x_1, \ldots, x_T$ at scale ε is a set $V \subseteq \mathbb{R}^T$ such that*

$$\forall g \in \mathcal{G}'\ \exists v \in V \text{ s.t. } \left(\frac{1}{T}\sum_{t=1}^{T}(g(x_t) - v_t)^p\right)^{1/p} \le \varepsilon.$$

We define the covering number $\mathcal{N}_p(\mathcal{G}', \varepsilon, x_{1:T})$ to be the size of the smallest such cover.

Lemma 4. For any fixed class $\mathcal{G}' \subseteq \{g : \mathcal{X} \to [0, 1]\}$ and fixed $a \in \mathcal{A}$, with probability at least $1-\delta$, it holds that

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(g, a)] \le 2\sum_{t=\tau_1}^{\tau_2} M_t(g, a) + 16\log\left(\frac{|\mathcal{G}'|T^2}{\delta}\right) \tag{8}$$

for all $\tau_1 \le \tau_2$ and $g \in \mathcal{G}'$ when $\mathcal{G}'$ is finite and

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(g, a)] \le 2\sum_{t=\tau_1}^{\tau_2} M_t(g, a) + \inf_{\varepsilon>0}\left\{100\varepsilon T + 320\log\left(\frac{4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G}', \varepsilon, x_{1:T})T^2\log(T)}{\delta}\right)\right\} \tag{9}$$

for all $\tau_1 \le \tau_2$ and $g \in \mathcal{G}'$ in the general case.

Remark 1. *Equation (9) implies (8), but with weaker constants.*

Corollary 2. Define

$$C_\delta = \min\left\{16\log\left(\frac{2|\mathcal{G}|KT^2}{\delta}\right), \inf_{\varepsilon>0}\left\{100\varepsilon T + 320\log\left(\frac{8\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G},\varepsilon,x_{1:T})KT^2\log(T)}{\delta}\right)\right\}\right\}.$$

With probability at least $1-\delta/2$, it holds that

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(g,a)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(g,a) + C_\delta, \tag{10}$$

for all $g\in\mathcal{G}$, $a\in\mathcal{A}$, and $\tau_1,\tau_2\in[T]$.

Proof of Lemma 4. We first prove the inequality in the finite class case.

For any fixed $g\in\mathcal{G}'$, $a\in\mathcal{A}$, and $\tau_1,\tau_2\in[T]$, since $Z_t = \mathbb{E}_t[M_t(g,a)] - M_t(g,a)$ forms a martingale different sequence with $|Z_t|\le 1$, applying Lemma 2 and Lemma 3 we have with probability $1-\delta$,

$$\sum_{t=\tau_1}^{\tau_2}(\mathbb{E}_t[M_t(g,a)] - M_t(g,a)) \le 4\eta(e-2)\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(g,a)] + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right).$$

This implies

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(g,a)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(g,a) + 16\log\left(\frac{1}{\delta}\right)$$

after setting $\eta = 1/8$ and rearranging. Finally, we apply a union bound over all $g\in\mathcal{G}'$ and $\tau\le\tau_2\in[T]$ to get the result.

For the infinite class case, we appeal to Theorem 9 of (Krishnamurthy et al., 2017) (see page 36 specifically — we do not use the final theorem statement but rather an intermediate result that is the consequence of their Lemmas 7, 8, 9, and 10).

Let τ_1 and τ_2 be fixed. Then the result of (Krishnamurthy et al., 2017) implies that for any class $\mathcal{G}$, any fixed $\varepsilon>0$, $\nu>0$ and $a\in\mathcal{A}$, letting $c=1/8$,

$$\Pr\left(\sup_{g\in\mathcal{G}}\left\{\sum_{t=\tau_1}^{\tau_2}\frac{1}{2}\mathbb{E}_t[M_t(g,a)] - M_t(g,a)\right\} > 4\nu + 16T(1+c)\varepsilon\right) \le 4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G},\varepsilon,x_{1:T})\exp\left(-\frac{2c}{(3+c)^2}\nu\right).$$

Rearranging, this implies that with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left\{\sum_{t=\tau_1}^{\tau_2}\frac{1}{2}\mathbb{E}_t[M_t(g,a)] - M_t(g,a)\right\} \le 18\varepsilon T + 160\log(4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G},\varepsilon,x_{1:T})/\delta). \tag{11}$$

Now consider a grid $\varepsilon_i := e^i/T$ for $i\in[\log(T)]$. By union bound, (11) implies that with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left\{\sum_{t=\tau_1}^{\tau_2}\frac{1}{2}\mathbb{E}_t[M_t(g,a)] - M_t(g,a)\right\} \le 18\varepsilon_i T + 160\log(4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G},\varepsilon_i,x_{1:T})\log(T)/\delta) \quad \forall i\in[\log(T)].$$

This implies that with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left\{\sum_{t=\tau_1}^{\tau_2}\frac{1}{2}\mathbb{E}_t[M_t(g,a)] - M_t(g,a)\right\} \le \inf_{\varepsilon>0}\{50\varepsilon T + 160\log(4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{G},\varepsilon,x_{1:T})\log(T)/\delta)\}.$$

To see that this inequality is implied by the preceeding inequality, first observe that the infimum over ε above may be restricted to $[1/T,1]$ without loss of generality. This holds because M_t lies in $[-1,1]$ and $\mathcal{N}_1(\mathcal{G},1,x_{1:T})\le 1$, which both follow from the fact that the range of $\mathcal{G}$ lies in $[0,1]$. Now let $\varepsilon^\star$ obtain the infimum and let $i^\star = \min\{i \mid \varepsilon_i \ge \varepsilon^\star\}$. Then $\mathcal{N}_1(\mathcal{G},\varepsilon_{i^\star},x_{1:T}) \le \mathcal{N}_1(\mathcal{G},\varepsilon^\star,x_{1:T})$ and $18\varepsilon_{i^\star}T \le 18e\varepsilon^\star T \le 50\varepsilon^\star T$.

To conclude, we take a union bound over all $\tau_1<\tau_2\in[T]$. □

Lemma 5. Conditioned on the event of Corollary 2, it holds that

1. $g_a^\star \in \widehat{\mathcal{G}}_m\left(\frac{C_\delta}{2(\tau_m-1)}, a\right)$ for all $m \in [M]$ and $a \in \mathcal{A}$.
2. For all $\beta \geq 0$, $m \in [M]$, and $a \in \mathcal{A}$,
$$\widehat{\mathcal{G}}_m(\beta, a) \subseteq \widetilde{\mathcal{G}}_m\left(2\beta + \frac{C_\delta}{\tau_m - 1}, a\right).$$
3. For all $\beta \geq 0$, $m \in [M]$, $k \in [m]$, and $a \in \mathcal{A}$,
$$\widehat{\mathcal{G}}_m(\beta, a) \subseteq \widehat{\mathcal{G}}_k\left(\frac{\tau_m - 1}{\tau_k - 1}\beta + \frac{C_\delta}{\tau_k - 1}, a\right).$$
4. With $\beta_m = \frac{(M-m+1)C_\delta}{\tau_m - 1}$, we have for any $m \in [M]$, $f^\star \in \mathcal{F}_m$ and also $\mathcal{F}_m \subseteq \mathcal{F}_{m-1} \subseteq \cdots \subseteq \mathcal{F}_1$.

Proof. Each claim in the lemma statement will be handled separately.
First claim. From (10) and nonnegativity of $\mathbb{E}_t[M_t(g,a)]$, we have that
$$\min_{g \in \mathcal{G}}\left\{2\sum_{t=1}^{\tau_m - 1} M_t(g,a)\right\} + C_\delta \geq 0.$$
Expanding out $M_t(g,a)$ and rearranging, this gives $\widehat{\mathcal{R}}_m(g_a^\star, a) - \min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_m(g,a) \leq \frac{C_\delta}{2(\tau_m - 1)}$, which implies $g_a^\star \in \widehat{\mathcal{G}}_m\left(\frac{C_\delta}{2(\tau_m-1)}, a\right)$.

Second claim. For any $g \in \widehat{\mathcal{G}}_m(\beta, a)$, we have by definition
$$\frac{1}{\tau_m - 1}\sum_{t=1}^{\tau_m - 1} M_t(g,a) = \widehat{\mathcal{R}}_m(g,a) - \widehat{\mathcal{R}}_m(g_a^\star, a) \leq \widehat{\mathcal{R}}_m(g,a) - \min_{g' \in \mathcal{G}} \widehat{\mathcal{R}}_m(g', a) \leq \beta. \tag{12}$$
Therefore applying (10) leads to
$$\frac{1}{\tau_m - 1}\sum_{t=1}^{\tau_m - 1} \mathbb{E}_t[M_t(g,a)] \leq \frac{2}{\tau_m - 1}\sum_{t=1}^{\tau_m - 1} M_t(g,a) + \frac{C_\delta}{\tau_m - 1} \leq 2\beta + \frac{C_\delta}{\tau_m - 1},$$
which implies $g \in \widetilde{\mathcal{G}}_m\left(2\beta + \frac{C_\delta}{\tau_m - 1}, a\right)$.

Third claim. For any $g \in \widehat{\mathcal{G}}_m(\beta, a)$, we have for any $k \in [m]$,
$$\begin{aligned}
(\tau_k - 1)\left(\widehat{\mathcal{R}}_k(g,a) - \min_{g' \in \mathcal{G}} \widehat{\mathcal{R}}_k(g', a)\right) &\leq (\tau_k - 1)\left(\widehat{\mathcal{R}}_k(g,a) - \widehat{\mathcal{R}}_k(g_a^\star, a)\right) + C_\delta/2 && \text{(by the first claim)} \\
&= \sum_{t=1}^{\tau_m - 1} M_t(g,a) - \sum_{t=\tau_k}^{\tau_m - 1} M_t(g,a) + C_\delta/2 \\
&\leq (\tau_m - 1)\beta - \frac{\sum_{t=\tau_k}^{\tau_m - 1} \mathbb{E}_t[M_t(g,a)]}{2} + C_\delta && \text{(by (12) and (10))} \\
&\leq (\tau_m - 1)\beta + C_\delta, && \text{(by nonnegativity of } \mathbb{E}_t[M_t(g,a)])
\end{aligned}$$
which implies $g \in \widehat{\mathcal{G}}_k\left(\frac{\tau_m - 1}{\tau_k - 1}\beta + \frac{C_\delta}{\tau_k - 1}, a\right)$.

Fourth claim. The value of β_m ensures that $\frac{C_\delta}{2(\tau_m - 1)} \leq \beta_m$ for any $m \in [M]$, and also for any $k < m$,
$$\frac{\tau_m - 1}{\tau_k - 1}\beta_m + \frac{C_\delta}{\tau_k - 1} = \frac{(M - m + 2)C_\delta}{\tau_k - 1} \leq \beta_k.$$
Therefore by the first and the third statement we have the claimed conclusions. □

Proposition 2. *For any two classes $\mathcal{F} \subseteq \mathcal{F}'$ and any context x, $A_{\mathcal{F}}(x) \subseteq A_{\mathcal{F}'}(x)$.*

Lemma 6. Algorithm 1 with OPTION I ensures that for any $m \in [M]$ and $t \in \{\tau_m, \dots, \tau_{m+1} - 1\}$,

$$A_t = \mathcal{A}_{\mathcal{F}_m}(x_t) = \bigcup_{f \in \mathcal{F}_m} \arg\max_{a \in \mathcal{A}} f(x_t, a).$$

Proof. For any $f \in \mathcal{F}_m$ and any $a \in \arg\max_{a' \in \mathcal{A}} f(x_t, a')$, we have by definitions

$$\textsc{High}_{\mathcal{F}_m}(x_t, a) \geq f(x_t, a) = \max_{a'} f(x_t, a') \geq \max_{a'} \textsc{Low}_{\mathcal{F}_m}(x_t, a'),$$

which implies $a \in A_t$. On the other hand, for each $a \in A_t$, there exists $g_a \in \widehat{\mathcal{G}}(\beta_m, a)$ such that $g_a(x_t) \geq \max_{a'} \min_{g \in \widehat{\mathcal{G}}(\beta_m, a')} g(x_t)$, which further implies that for any $a' \neq a$, there exists $g_{a'} \in \widehat{\mathcal{G}}(\beta_m, a')$ such that $g_a(x_t) \geq g_{a'}(x_t)$. Therefore, we can construct an f so that $f(\cdot, a) = g_a(\cdot)$ and $f(\cdot, a') = g_{a'}(\cdot)$ for all $a' \neq a$, so that clearly $f \in \mathcal{F}_m$ and $a \in \arg\max_{a' \in \mathcal{A}} f(x_t, a')$. This proves the lemma. □

Lemma 7. Conditioned on the event of Corollary 2, Algorithm 1 with OPTION I and $\beta_m = \frac{(M-m+1)C_\delta}{\tau_m - 1}$ ensures that for any $m \in [M]$, we have $\mathcal{F}_m \subseteq \mathcal{F}(\varepsilon_m)$ with

$$\varepsilon_m = \inf_{\eta > 0} \left\{ \eta P_\eta + \frac{4K^2}{\eta(\tau_m - 1)} (2M - 2m + 3) C_\delta \right\},$$

where $P_\eta = \Pr_x\big(f^\star(x, \pi^\star(x)) - \max_{a \neq \pi^\star(x)} f^\star(x, a) < \eta\big)$.

Proof. We first prove that for any $t < \tau_m$ and $f \in \mathcal{F}_m$, the following holds

$$\mathbb{E}_{x,r}[r(\pi^\star(x)) - r(\pi_f(x))] \leq \inf_{\eta > 0} \left\{ \eta P_\eta + \frac{4K}{\eta} \sum_{a \in \mathcal{A}} \mathbb{E}_t[M_t(f, a)] \right\}. \tag{13}$$

Indeed, note that for any $\eta > 0$, with realizability we have

$$\begin{aligned}
&\mathbb{E}_{x,r}[r(\pi^\star(x)) - r(\pi_f(x))] \\
&= \mathbb{E}_x[f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x))] \\
&\leq \eta \Pr_x(f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x)) < \eta \text{ and } \pi^\star(x) \neq \pi_f(x)) + \frac{1}{\eta} \mathbb{E}_x(f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x)))^2 \\
&\leq \eta P_\eta + \frac{1}{\eta} \mathbb{E}_x(f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x)))^2.
\end{aligned}$$

By the definition of π_f we also have for any x, $f(x, \pi_f(x)) - f(x, \pi^\star(x)) \geq 0$ and thus

$$\begin{aligned}
\mathbb{E}_x(f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x)))^2 &\leq \mathbb{E}_x(f^\star(x, \pi^\star(x)) - f^\star(x, \pi_f(x)) + f(x, \pi_f(x)) - f(x, \pi^\star(x)))^2 \\
&\leq 2\,\mathbb{E}_x(f^\star(x, \pi^\star(x)) - f(x, \pi^\star(x)))^2 + 2\,\mathbb{E}_x(f(x, \pi_f(x)) - f^\star(x, \pi_f(x)))^2.
\end{aligned}$$

Now suppose round t is in epoch k. Since both $f \in \mathcal{F}_m \subseteq \mathcal{F}_k$ and $f^\star \in \mathcal{F}_k$, we have $\pi_f(x_t), \pi^\star(x_t) \in A_t$ by Lemma 6. Therefore, the fact that a_t is drawn uniformly from A_t implies

$$\mathbb{E}_x(f^\star(x, \pi^\star(x)) - f(x, \pi^\star(x)))^2 \leq K\,\mathbb{E}_{x,a_t}(f^\star(x, a_t) - f(x, a_t))^2,$$

and likewise

$$\mathbb{E}_x(f^\star(x, \pi_f(x)) - f(x, \pi_f(x)))^2 \leq K\,\mathbb{E}_{x,a_t}(f^\star(x, a_t) - f(x, a_t))^2.$$

Lastly, plugging the equality

$$\mathbb{E}_{x,a_t}(f^\star(x_t, a_t) - f(x_t, a_t))^2 = \sum_{a \in \mathcal{A}} \mathbb{E}_t[M_t(f, a)]$$

proves Eq. (13). Averaging over $t = 1, \dots, \tau_m - 1$ then gives

$$\mathbb{E}_{x,r}[r(\pi^\star(x)) - r(\pi_f(x))] \leq \inf_{\eta > 0} \left\{ \eta P_\eta + \frac{4K}{\eta(\tau_m - 1)} \sum_{a \in \mathcal{A}} \sum_{t=1}^{\tau_m - 1} \mathbb{E}_t[M_t(f, a)] \right\}.$$

Using the second statement of Lemma 5 we have $\sum_{t=1}^{\tau_m-1}\mathbb{E}_t[M_t(f,a)] \le 2(\tau_m-1)\beta_m + C_\delta = (2M-2m+3)C_\delta$ and thus

$$\mathbb{E}_{x,r}[r(\pi^\star(x)) - r(\pi_f(x))] \le \inf_{\eta>0}\left\{\eta P_\eta + \frac{4K^2}{\eta(\tau_m-1)}(2M-2m+3)C_\delta\right\} = \varepsilon_m,$$

completing the proof by the definition of $\mathcal{F}(\varepsilon_m)$. □

We are now ready to prove Theorem 2, which is restated below with an extra result under the Massart condition.

Theorem 5 (Full version of Theorem 2)**.** *With $\beta_m = \frac{(M-m+1)C_\delta}{\tau_m-1}$ and C_δ as in Corollary 2, Algorithm 1 with Option I ensures that with probability at least $1-\delta$,*

$$\mathrm{Reg}_T = O\left(T^{\frac{3}{4}}C_\delta{}^{\frac{1}{4}}\sqrt{\theta_0 K\log T} + \log(1/\delta)\right). \tag{14}$$

In particular, for finite classes regret is bounded as $\widetilde{O}\left(T^{\frac{3}{4}}(\log|\mathcal{G}|)^{\frac{1}{4}}\sqrt{\theta_0 K}\right)$.
Furthermore, if the Massart noise condition (Definition 8) is satisfied with parameter γ, then Algorithm 1 configured as above with $\delta = 1/T$ enjoys an in-expectation regret bound of

$$\mathbb{E}\left[\sum_{t=1}^T r_t(\pi^\star(x_t)) - \sum_{t=1}^T r_t(a_t)\right] = O\left(\frac{\theta_0 K^2 C_{1/T}\log^2 T}{\gamma^2}\right), \tag{15}$$

which for finite classes is upper bounded by $\widetilde{O}\left(\frac{\theta_0 K^2\log(|\mathcal{G}|T)}{\gamma^2}\right)$.

Remark 2. *This theorem and the subsequent regret bounds based on moment conditions (Theorem 6 and Theorem 7) give a high-probability empirical regret bound in the general case, but only give an in-expectation regret bound under the Massart condition. This is because one incurs an extra $O(\sqrt{T})$ factor in going from a (conditional) expected regret bound to an empirical regret bound, which is a low order term in the general case but may dominate in the Massart case.*

Proof. We will first provide a bound on

$$\sum_{t=1}^T \mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)],$$

then relate this quantity to the left-hand-side of (14) and (15) at the end.

This proof conditions on the above event and the events of Corollary 2, which happen with probability at least $1-\delta/2$, and bounds the conditional expected regret terms $\mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)]$ individually.

For any $\eta' > 0$, we recall the definition used in the proof of Lemma 7: $P_{\eta'} = \Pr_x\left(f^\star(x,\pi^\star(x)) - \max_{a\neq\pi^\star(x)} f^\star(x,a) < \eta'\right)$. Further define two events:

$$E_1 = \{\exists a \in A_t : f^\star(x,a) < f^\star(x,\pi^\star(x))\}$$
$$E_2 = \{\exists a \in A_t : f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a) \ge \eta'\}.$$

We then have

$$\begin{aligned}
\mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)] &= \mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t) \mid E_1]\Pr_{x_t}(E_1)\\
&= \mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t) \mid E_1,\neg E_2]\Pr_{x_t}(E_1,\neg E_2)\\
&\quad + \mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t) \mid E_1,E_2]\Pr_{x_t}(E_1,E_2)\\
&\le \eta'\Pr_{x_t}(E_1,\neg E_2) + \Pr_{x_t}(E_1,E_2)\\
&\le \eta' P'_\eta + \Pr_{x_t}(E_1,E_2).
\end{aligned}$$

Next we argue two facts (suppose round t is in epoch m): E_1 implies $x_t \in \mathrm{Dis}(\mathcal{F}_m)$, and E_2 implies that there exists $a' \in A_t$ such that $W_{\mathcal{F}_m}(x,a') > \eta'/2$. Indeed, with a being the action stated in event E_1, we know that by Lemma 6 there exists $f \in \mathcal{F}_m$ such that $a \in \arg\max_{a'} f(x_t,a)$. However, clearly a is not in $\arg\max_{a'} f^\star(x_t,a)$, and thus by $f^\star \in \mathcal{F}_m$ and the

definition of disagreement region we have $x_t \in \mathrm{Dis}(\mathcal{F}_m)$. On the other hand, with a being the action stated in event E_2, we have

$$\begin{aligned}\eta' &\le f^\star(x_t, \pi^\star(x_t)) - f^\star(x_t, a)\\ &\le \textsc{High}_{\mathcal{F}_m}(x_t, \pi^\star(x_t)) - \textsc{Low}_{\mathcal{F}_m}(x_t, a)\\ &\le \textsc{High}_{\mathcal{F}_m}(x_t, \pi^\star(x_t)) - \textsc{Low}_{\mathcal{F}_m}(x_t, \pi^\star(x_t)) + \textsc{High}_{\mathcal{F}_m}(x_t, a) - \textsc{Low}_{\mathcal{F}_m}(x_t, a)\\ &= W_{\mathcal{F}_m}(x_t, \pi^\star(x_t)) + W_{\mathcal{F}_m}(x_t, a)\end{aligned}$$

where the last inequality is by the fact $a \in A_t$ and the definition of A_t. The last inequality thus implies that there exists $a' \in A_t$ such that $W_{\mathcal{F}_m}(x, a') > \eta'/2$. We therefore continue with

$$\begin{aligned}\Pr_{x_t}(E_1, E_2) &\le \Pr_{x_t}(x_t \in \mathrm{Dis}(\mathcal{F}_m) \text{ and } \exists a \in A_t : W_{\mathcal{F}_m}(x, a) > \eta'/2)\\ &\le \Pr_{x_t}(x_t \in \mathrm{Dis}(\mathcal{F}_m) \text{ and } \exists a \in A_{\mathcal{F}_m}(x_t) : W_{\mathcal{F}_m}(x, a) > \eta'/2)\\ &\le \Pr_{x_t}(x_t \in \mathrm{Dis}(\mathcal{F}(\varepsilon_m)) \text{ and } \exists a \in A_{\mathcal{F}(\varepsilon_m)}(x_t) : W_{\mathcal{F}(\varepsilon_m)}(x, a) > \eta'/2) \quad \text{(by Lemma 7 and Proposition 2)}\\ &\le \frac{2\theta_0 \varepsilon_m}{\eta'}. \qquad \text{(by the definition of } \theta_0)\end{aligned}$$

Combining everything we arrive at for any $\eta, \eta' > 0$,

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_t[f^\star(x_t, \pi^\star(x_t)) - f^\star(x_t, a_t)] &\le \eta' T P'_\eta + \frac{2\theta_0}{\eta'}\left(\eta T P_\eta + \frac{4K^2 C_\delta}{\eta}\sum_{m=1}^M \frac{T_m(2M - 2m + 3)}{\tau_m - 1}\right)\\ &\le \eta' T P'_\eta + \frac{2\theta_0}{\eta'}\left(\eta T P_\eta + \frac{8K^2 C_\delta}{\eta}(M^2 + 2M)\right).\end{aligned}$$

In the general case we simply bound P_η and $P_{\eta'}$ by 1 and choose the optimal η and η' to arrive at a regret bound of order $O\left(T^{\frac{3}{4}} C_\delta^{\frac{1}{4}} \sqrt{\theta_0 K \log T} + \log(1/\delta)\right)$. On the other hand, under the Massart condition (Definition 8) one can pick $\eta = \eta' = \gamma$ so that $P_\eta = P_{\eta'} = 0$ and obtain a regret bound of order $O\left(\frac{\theta_0 K^2 C_\delta \log^2 T}{\gamma^2}\right)$.

Lastly, we relate the sum of conditional expected instantaneous regrets to the left-hand side of (14) and (15). In the general case, since instantaneous regret lies in $[-1, 1]$, Azuma-Hoeffding implies that

$$\sum_{t=1}^T r_t(\pi^\star(x_t)) - r_t(a_t) \le \sum_{t=1}^T \mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)] + O(\sqrt{T \log(1/\delta)})$$

with probability at least $1 - \delta/2$. By union bound, the theorem statement holds with probability at least $1 - \delta$.

In the Massart case, the law of total expectation implies

$$\mathbb{E}\left[\sum_{t=1}^T r_t(\pi^\star(x_t)) - r_t(a_t)\right] \le O\left(\frac{\theta_0 K^2 C_{1/T} \log^2 T}{\gamma^2}\right) + \frac{1}{T} \cdot T,$$

where the second term uses boundedness of regret along with the fact that the events of Corollary 2 hold with probability at least $1 - 1/T$. □

A.4. Proofs from Section 4.2

Similarly to the notation $M_t(g, a)$ for the case $\mathcal{F} = \mathcal{G}^{\mathcal{A}}$, for a general predictor class $\mathcal{F}$ we define for any $f \in \mathcal{F}$

$$M_t(f) = (f(x_t, a_t) - r_t(a_t))^2 - (f^\star(x_t, a_t) - r_t(a_t))^2.$$

and also the class

$$\widetilde{\mathcal{F}}_m(\beta) = \left\{f \in \mathcal{F} \mid \frac{1}{\tau_m - 1}\sum_{t=1}^{\tau_m - 1} \mathbb{E}_t[M_t(f)] \le \beta\right\}.$$

Finally, for any $a \in \mathcal{A}$ we define a class

$$\mathcal{F}|_a = \{x \mapsto f(x,a) \mid f \in \mathcal{F}\}.$$

We establish several lemmas similar to those in Appendix A.3.

Lemma 8. For any $f \in \mathcal{F}$ we have

$$\mathbb{E}_t[M_t(f)] = \mathbb{E}_t\big[(f(x_t,a_t) - f^\star(x_t,a_t))^2\big],$$
$$\mathrm{Var}_t[M_t(f)] \le 4\,\mathbb{E}_t[M_t(f)].$$

Lemma 9. Define

$$C'_\delta = \min\left\{16\log\left(\frac{2|\mathcal{F}|T^2}{\delta}\right), \inf_{\varepsilon>0}\left\{100\varepsilon KT + 320\sum_{a\in\mathcal{A}}\log\left(\frac{8\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{F}|_a,\varepsilon,x_{1:T})KT^2\log(T)}{\delta}\right)\right\}\right\}. \tag{16}$$

With probability at least $1-\delta/2$, it holds that

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(f)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(f) + C'_\delta,$$

for all $f \in \mathcal{F}$ and $\tau_1,\tau_2 \in [T]$.

Proof of Lemma 9. We first prove the inequality in the finite class case. For any fixed $f \in \mathcal{F}$, and $\tau_1 \le \tau_2 \in [T]$, $Z_t := \mathbb{E}_t[M_t(f)] - M_t(f)$ forms a martingale different sequence with $|Z_t| \le 1$. Applying Lemma 2 and Lemma 8 we have with probability $1-\delta$,

$$\sum_{t=\tau_1}^{\tau_2}(\mathbb{E}_t[M_t(f)] - M_t(f)) \le 4\eta(e-2)\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(f)] + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right),$$

which implies after setting $\eta = 1/8$ and rearranging.

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(f)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(f) + 16\log\left(\frac{1}{\delta}\right)$$

We apply a union bound over all $f \in \mathcal{F}$ and $\tau_1 \le \tau_2 \in [T]$ to get the result.

To handle the infinite class case we invoke Lemma 4. In particular, for any fixed a, the lemma with $\mathcal{G}' = \mathcal{F}|_a$ implies that with probability at least $1-\delta$,

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(f(\cdot,a),a)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(f(\cdot,a),a) + \inf_{\varepsilon>0}\left\{100\varepsilon T + 320\log\left(\frac{4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{F}|_a,\varepsilon,x_{1:T})T^2\log(T)}{\delta}\right)\right\}$$

for all $f \in \mathcal{F}$ and $\tau_1 \le \tau_2$. Observe that $M_t(f) = \sum_{a\in\mathcal{A}}M_t(f(\cdot,a),a)$. Taking a union bound and then summing over all actions, the preceding statement therefore implies that with probability at least $1-\delta$,

$$\sum_{t=\tau_1}^{\tau_2}\mathbb{E}_t[M_t(f)] \le 2\sum_{t=\tau_1}^{\tau_2}M_t(f) + \sum_{a\in\mathcal{A}}\inf_{\varepsilon>0}\left\{100\varepsilon T + 320\log\left(\frac{4\,\mathbb{E}_{x_{1:T}}\mathcal{N}_1(\mathcal{F}|_a,\varepsilon,x_{1:T})KT^2\log(T)}{\delta}\right)\right\}$$

for all $f \in \mathcal{F}$ and $\tau_1 \le \tau_2$. The final result follows from superadditivity of the infimum. □

Lemma 10. Conditioned on the event of Lemma 9, it holds that

1. $f^\star \in \widehat{\mathcal{F}}_m\left(\frac{C'_\delta}{2(\tau_m-1)}\right)$ for all $m \in [M]$.
2. For all $\beta \ge 0$ and $m \in [M]$,

$$\widehat{\mathcal{F}}_m(\beta) \subseteq \widetilde{\mathcal{F}}_m\left(2\beta + \frac{C'_\delta}{\tau_m - 1}\right).$$

Consequently, we have $\mathbb{E}_{\tau_{m-1}}[M_{\tau_{m-1}}(f)] \le \frac{2\beta(\tau_m-1)+C'_\delta}{T_{m-1}}$ for any $f \in \widehat{\mathcal{F}}_m(\beta)$.

3. For all $\beta \geq 0$, $m \in [M]$, and $k \in [m]$,
$$\widehat{\mathcal{F}}_m(\beta) \subseteq \widehat{\mathcal{F}}_k\left(\frac{\tau_m - 1}{\tau_k - 1}\beta + \frac{C'_\delta}{\tau_k - 1}\right).$$
4. With $\beta_m = \frac{(M-m+1)C_\delta}{\tau_m - 1}$, we have for any $m \in [M]$, $f^\star \in \mathcal{F}_m$ and also $\mathcal{F}_m \subseteq \mathcal{F}_{m-1} \subseteq \cdots \subseteq \mathcal{F}_1$.

Proof. The proof of this lemma is essentially the same as that of Lemma 5 in Appendix A.3. The only new statement is the second statement of the second claim in Lemma 10. This holds because for any $f \in \widehat{\mathcal{F}}_m(\beta) \subseteq \widetilde{\mathcal{F}}_m\left(2\beta + \frac{C'_\delta}{\tau_m - 1}\right)$, we have
$$\sum_{t=\tau_{m-1}}^{\tau_m - 1} \mathbb{E}_t[M_t(f)] \leq \sum_{t=1}^{\tau_m - 1} \mathbb{E}_t[M_t(f)] \leq 2\beta(\tau_m - 1) + C'_\delta,$$
and also by the epoch schedule of the algorithm $\mathbb{E}_t[M_t(f)]$ remains the same for all $t \in \{\tau_{m-1}, \ldots, \tau_m - 1\}$ and thus $T_{m-1}\,\mathbb{E}_{\tau_{m-1}}[M_{\tau_{m-1}}(f)] = \sum_{t=\tau_{m-1}}^{\tau_m - 1} \mathbb{E}_t[M_t(f)] \leq 2\beta(\tau_m - 1) + C'_\delta$, proving the statement. □

We are now ready to prove the main theorems, which are again restated with extra results under the Massart condition.

Theorem 6 (Full version of Theorem 3)**.** *With $\beta_m = \frac{(M-m+1)C'_\delta}{\tau_m - 1}$ and C'_δ as in Lemma 9, Algorithm 1 with Option II ensures that with probability at least $1 - \delta$,*
$$\mathrm{Reg}_T = O\left(\sqrt{TL_{2,0}C'_\delta}\log T + \log(1/\delta)\right).$$
In particular, for finite classes regret is bounded as $\widetilde{O}\left(\sqrt{TL_{2,0}\log|\mathcal{F}|}\right)$.

Furthermore, if the Massart noise condition Definition 8 is satisfied with parameter γ, then Algorithm 1 configured as above with $\delta = 1/T$ enjoys an in-expectation regret bound of
$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(\pi^\star(x_t)) - r_t(a_t)\right] = O\left(\frac{L_{2,0}C'_{1/T}\log^2 T}{\gamma}\right),$$
which for finite classes is bounded as $\widetilde{O}\left(\frac{L_{2,0}\log|\mathcal{F}|}{\gamma}\right)$.

Proof. Similar to the proof of Theorem 2, we condition on the events of Lemma 9, which happen with probability at least $1 - \delta/2$.

With m denoting the epoch to which round t belongs and $P_\eta = \Pr_x\left(f^\star(x, \pi^\star(x)) - \max_{a \neq \pi^\star(x)} f^\star(x,a) < \eta\right)$ for any $\eta > 0$, we have
$$\begin{aligned}
&\mathbb{E}_t[f^\star(x_t, \pi^\star(x_t)) - f^\star(x_t, a_t)] \\
&\leq \eta P_\eta + \frac{1}{\eta}\mathbb{E}_t\left[(f^\star(x_t, \pi^\star(x_t)) - f^\star(x_t, a_t))^2\right] \\
&\leq \eta P_\eta + \frac{1}{\eta}\mathbb{E}_t\left[(f^\star(x_t, \pi^\star(x_t)) - \textsc{Low}_{\mathcal{F}_m}(x_t, \pi^\star(x_t)) + \textsc{High}_{\mathcal{F}_m}(x_t, a_t) - f^\star(x_t, a_t))^2)^2\right] && (a_t \in A_t) \\
&\leq \eta P_\eta + \frac{2}{\eta}\mathbb{E}_t\left[(f^\star(x_t, \pi^\star(x_t)) - \textsc{Low}_{\mathcal{F}_m}(x_t, \pi^\star(x_t)))^2\right] + \frac{2}{\eta}\mathbb{E}_t\left[(\textsc{High}_{\mathcal{F}_m}(x_t, a_t) - f^\star(x_t, a_t))^2\right] \\
&\leq \eta P_\eta + \frac{2}{\eta}\mathbb{E}_t\left[\sup_{f \in \mathcal{F}_m}(f^\star(x_t, \pi^\star(x_t)) - f(x_t, \pi^\star(x_t)))^2\right] + \frac{2}{\eta}\mathbb{E}_t\left[\sup_{f \in \mathcal{F}_m}(f(x_t, a_t) - f^\star(x_t, a_t))^2\right] \\
&\leq \eta P_\eta + \frac{4}{\eta}\sup_{x \in \mathcal{X}, a \in \mathcal{A}}\sup_{f \in \mathcal{F}_m}\left\{(f^\star(x,a) - f(x,a))^2\right\} \\
&= \eta P_\eta + \frac{4}{\eta}\sup_{f \in \mathcal{F}_m}\sup_{x \in \mathcal{X}, a \in \mathcal{A}}\left\{(f^\star(x,a) - f(x,a))^2\right\} \\
&\leq \eta P_\eta + \frac{4L_{2,0}}{\eta}\sup_{f \in \mathcal{F}_m}\mathbb{E}_{x \sim D_{\mathcal{X}}}\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}\left[\mathbf{1}\{x \in U_0(a)\}(f^\star(x,a) - f(x,a))^2\right] && \text{(by Definition 3)} \\
&\leq \eta P_\eta + \frac{4L_{2,0}}{\eta}\sup_{f \in \mathcal{F}_m}\mathbb{E}_{x \sim D_{\mathcal{X}}}\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}\left[\mathbf{1}\{a \in A_{\tau_{m-1}}\}(f^\star(x,a) - f(x,a))^2\right],
\end{aligned}$$

where the last step holds because $x \in U_0(a)$ along with the fact $f^\star \in \mathcal{F}_{m-1}$ implies

$$\textsc{High}_{\mathcal{F}_{m-1}}(x,a) \geq f^\star(x,a) = \max_{a'} f^\star(x,a') \geq \max_{a'} \textsc{Low}_{\mathcal{F}_{m-1}}(x,a'),$$

and thus by definition $a \in A_{\tau_{m-1}}$. We continue with

$$\begin{aligned}
\mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)] &\leq \eta P_\eta + \frac{4L_{2,0}}{\eta} \sup_{f\in\mathcal{F}_m} \mathbb{E}_{x\sim D_\mathcal{X}} \mathbb{E}_{a\sim\mathrm{Unif}(A_{\tau_{m-1}})}\left[(f^\star(x,a) - f(x,a))^2\right] \\
&= \eta P_\eta + \frac{4L_{2,0}}{\eta} \sup_{f\in\mathcal{F}_m} \mathbb{E}_{\tau_{m-1}}[M_{\tau_{m-1}}(f)] \\
&\leq \eta P_\eta + \frac{4L_{2,0}}{\eta} \cdot \frac{2\beta_m(\tau_m - 1) + C'_\delta}{T_{m-1}} \qquad \text{(by the second claim of Lemma 10)} \\
&= \eta P_\eta + \frac{4L_{2,0}(2M - 2m + 3)C'_\delta}{\eta T_{m-1}}.
\end{aligned}$$

Summing over $t = 1,\dots,T$, we arrive at

$$\sum_{t=1}^T \mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)] = \eta T P_\eta + \sum_{m=1}^M T_m \frac{4L_{2,0}(2M-2m+3)C'_\delta}{\eta T_{m-1}} = \eta T P_\eta + \frac{8L_{2,0}(M^2+2M)C'_\delta}{\eta}.$$

Finally in the general case we bound P_η by 1 and pick the optimal η to arrive at a conditional expected regret bound of order $O(\sqrt{TL_{2,0}C'_\delta}\log T + \log(1/\delta))$, while under the Massart condition (Definition 8) one can pick $\eta = \gamma$ so that $P_\eta = 0$ and obtain a conditional expected regret bound of order $O\left(\frac{L_{2,0}C'_\delta \log^2 T}{\gamma}\right)$.

As in the proof of Theorem 5, we relate the sum of conditional expected instantaneous regrets back to the quantities in the theorem statement differently in the general case and the Massart case. In the general case we have

$$\sum_{t=1}^T r_t(\pi^\star(x_t)) - r_t(a_t) \leq \sum_{t=1}^T \mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)] + O(\sqrt{T\log(1/\delta)})$$

with probability at least $1 - \delta/2$.

In the Massart case, the law of total expectation implies

$$\mathbb{E}\left[\sum_{t=1}^T r_t(\pi^\star(x_t)) - r_t(a_t)\right] \leq O\left(\frac{L_{2,0}C'_{1/T}\log^2 T}{\gamma}\right) + \frac{1}{T}\cdot T.$$

□

Theorem 7 (Full version of Theorem 4)**.** *With $\beta_m = \frac{(M-m+1)C'_\delta}{\tau_m - 1}$, where C'_δ is as in Lemma 9, and $M_0 = 2 + \left\lfloor \log_2\left(1 + \frac{(2M+3)L_1 C'_\delta}{\lambda^2}\right)\right\rfloor$ for any $\lambda \in (0,1)$, Algorithm 2 ensures that with probability at least $1-\delta$,*

$$\mathrm{Reg}_T = O\left(\frac{L_1 C'_\delta \log T}{\lambda^2} + \sqrt{TL_{2,\lambda}C'_\delta}\log T\right),$$

which for finite classes is bounded by $\widetilde{O}\left(\frac{L_1\log|\mathcal{F}|}{\lambda^2} + \sqrt{TL_{2,\lambda}\log|\mathcal{F}|}\right)$.
Furthermore, if the Massart noise condition (Definition 8) is satisfied with parameter γ, then Algorithm 2 configured as above with $\delta = 1/T$ enjoys an expected regret bound of

$$\mathbb{E}\left[\sum_{t=1}^T r_t(\pi^\star(x_t)) - r_t(a_t)\right] = O\left(\frac{L_1 C'_{1/T}\log T}{\lambda^2} + \frac{L_{2,\lambda}C'_{1/T}\log^2 T}{\gamma}\mathbf{1}\{\lambda > \gamma\}\right),$$

which for finite classes is bounded by $\widetilde{O}\left(\frac{L_1\log|\mathcal{F}|}{\lambda^2} + \frac{L_{2,\lambda}\log|\mathcal{F}|}{\gamma}\mathbf{1}\{\lambda > \gamma\}\right)$.

Proof. We condition on the same events of Lemma 9, which hold with probability at least $1-\delta/2$. By the second claim of Lemma 10, we have for any $f \in \mathcal{F}_{M_0}$,

$$\sum_{t=1}^{\tau_{M_0}-1} \mathbb{E}_t[M_t(f)] \le 2\beta_{M_0}(\tau_{M_0}-1) + C'_\delta = 2\frac{(M-M_0+1)C'_\delta}{\tau_{M_0}-1}(\tau_{M_0}-1) + C'_\delta \le (2M+3)C'_\delta.$$

Since Algorithm 2 performs pure exploration for any t before epoch M_0, we conclude that

$$\mathbb{E}_t[M_t(f)] = \mathbb{E}_{x\sim D}\,\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}(f(x,a)-f^\star(x,a))^2 \le \frac{(2M+3)C'_\delta}{\tau_{M_0}-1},$$

and therefore together with Definition 2, we have for any $f \in \mathcal{F}_{M_0}$, $x \in \mathcal{X}$, and $a \in \mathcal{A}$,

$$(f(x,a)-f^\star(x,a))^2 \le L_1\,\mathbb{E}_{x'\sim D_\mathcal{X}}\,\mathbb{E}_{a'\sim\mathrm{Unif}(\mathcal{A})}(f(x',a')-f^\star(x',a'))^2 \le \frac{(2M+3)L_1C'_\delta}{\tau_{M_0}-1} < \lambda^2, \tag{17}$$

where the last step holds by the choice of M_0. Next we claim that for any $t \ge \tau_{M_0}$, if $x_t \in U_\lambda(a)$ for some a, then it must be the case $a_t = a = \pi^\star(x_t)$. To begin, we have that $a = \pi^\star(x_t)$, which is by the definition of $U_\lambda(a)$. Moreover, with m being the epoch to which t belongs and $a' = \arg\max_{a\neq\pi^\star(x_t)} \textsc{High}_{\mathcal{F}_m}(x_t,a)$, we have

$$\begin{aligned}
&\textsc{High}_{\mathcal{F}_m}(x_t,a) - \textsc{High}_{\mathcal{F}_m}(x_t,a') \\
&= \underbrace{f^\star(x_t,a) - f^\star(x_t,a')}_{\ge\lambda} + \underbrace{\textsc{High}_{\mathcal{F}_m}(x_t,a) - f^\star(x_t,a)}_{\ge 0} + \underbrace{f^\star(x_t,a') - \textsc{High}_{\mathcal{F}_m}(x_t,a')}_{>-\lambda} \\
&> \lambda + 0 - \lambda = 0,
\end{aligned}$$

where the inequality is by $x_t \in U_\lambda(a)$, $f^\star \in \textsc{High}_{\mathcal{F}_m}$, and Eq. (17). By the optimistic strategy of Algorithm 2, this implies $a_t = a$.

Finally we proceed exactly the same as the proof of Theorem 3 to arrive at for any $\eta > 0$, $\lambda \in (0,1)$, $m > M_0$, and t in epoch m,

$$\mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)] \le \eta P_\eta + \frac{4L_{2,\lambda}}{\eta}\sup_{f\in\mathcal{F}_m}\mathbb{E}_{x\sim D_\mathcal{X}}\,\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}\left[\mathbf{1}\{x\in U_\lambda(a)\}(f^\star(x,a)-f(x,a))^2\right].$$

With the fact established above, since $\tau_{m-1} \ge \tau_{M_0}$ we continue with

$$\begin{aligned}
\mathbb{E}_t[f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,a_t)] &\le \eta P_\eta + \frac{4L_{2,\lambda}}{K\eta}\sup_{f\in\mathcal{F}_m}\mathbb{E}_{x_{\tau_{m-1}}}\left[(f^\star(x_{\tau_{m-1}},a_{\tau_{m-1}}) - f(x_{\tau_{m-1}},a_{\tau_{m-1}}))^2\right] \\
&= \eta P_\eta + \frac{4L_{2,\lambda}}{\eta}\sup_{f\in\mathcal{F}_m}\mathbb{E}_{\tau_{m-1}}[M_{\tau_{m-1}}(f)] \\
&\le \eta P_\eta + \frac{4L_{2,\lambda}}{\eta}\cdot\frac{2\beta_m(\tau_m-1)+C'_\delta}{T_{m-1}} \qquad \text{(by the second claim of Lemma 10)} \\
&= \eta P_\eta + \frac{4L_{2,\lambda}(2M-2m+3)C'_\delta}{\eta T_{m-1}}.
\end{aligned}$$

Therefore, the regret bound is

$$\mathrm{Reg}_T \le \tau_{M_0+1} + \eta T P_\eta + \sum_{m=M_0+1}^{M} T_m\frac{4L_{2,\lambda}(2M-2m+3)C'_\delta}{\eta T_{m-1}} \le O\left(\frac{L_1C'_\delta\log T}{\lambda^2}\right) + \eta T P_\eta + \frac{8L_{2,\lambda}(M^2+2M)C'_\delta}{\eta}.$$

Again in general we bound P_η by 1 and pick the optimal η to arrive at

$$\sum_{t=1}^{T}\mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)] = O\left(\frac{L_1C'_\delta\log T}{\lambda^2} + \sqrt{TL_{2,\lambda}C'_\delta}\log T\right),$$

while under the Massart condition we pick $\eta = \gamma$ so that $P_\eta = 0$ and

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)] = O\left(\frac{L_1 C'_\delta \log T}{\lambda^2} + \frac{L_{2,\lambda} C'_\delta \log^2 T}{\gamma}\right).$$

Specifically, if we choose $\lambda \le \gamma$, then every x_t is in $U_\lambda(\pi^\star(x_t))$ and thus $a_t = \pi^\star(x_t)$ for all $t \ge \tau_{M_0}$ and the algorithm suffers no regret at all after the warm start, that is, $\mathrm{Reg}_T = O\left(\frac{L_1 C'_\delta \log T}{\lambda^2}\right)$.

To conclude we proceed as in the proof of Theorem 6: In the general case we have

$$\sum_{t=1}^{T} r_t(\pi^\star(x_t)) - r_t(a_t) \le \sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi^\star(x_t)) - r_t(a_t)] + O(\sqrt{T \log(1/\delta)})$$

with probability at least $1 - \delta/2$ by Azuma-Hoeffding.

In the Massart case, the law of total expectation implies

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(\pi^\star(x_t)) - r_t(a_t)\right] \le O\left(\frac{L_1 C'_{1/T} \log T}{\lambda^2} + \frac{L_{2,\lambda} C'_{1/n} \log^2 T}{\gamma}\right) + \frac{1}{T} \cdot T.$$

□

For the following proposition we recall the notation $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and $\mathcal{F} = \{(x,a) \mapsto w^\top \phi(x,a) \mid w \in \mathcal{W}\}$ for some $\mathcal{W} \subseteq \mathbb{R}^d$.

Proposition 3.

- *If $\|\phi(x,a)\|_2 \le 1$ and $\|w\|_2 \le 1$ then $L_{2,\lambda}$ is bounded by*

$$\frac{K}{\lambda_{\min}\left(\sum_{a \in \mathcal{A}} \mathbb{E}_x\left[\mathbf{1}\{x \in U_\lambda(a)\}\, \phi(x,a)\phi(x,a)^\top\right]\right)},$$

  where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of a matrix, and L_1 is bounded by

$$\frac{K}{\lambda_{\min}\left(\sum_{a \in \mathcal{A}} \mathbb{E}_x[\phi(x,a)\phi(x,a)^\top]\right)}.$$

- *In the sparse high-dimensional setting with $\|\phi(x,a)\|_\infty \le 1$, $\|w\|_\infty \le 1$, and $\|w\|_0 \le s$, then $L_{2,\lambda}$ is bounded by*

$$\frac{2Ks}{\psi_{\min}\left(\sum_{a \in \mathcal{A}} \mathbb{E}_x\left[\mathbf{1}\{x \in U_\lambda(a)\}\, \phi(x,a)\phi(x,a)^\top\right]\right)},$$

  where $\psi_{\min}(A) \coloneqq \min_{w \neq 0:\, \|w\|_0 \le 2s} w^\top A w \,/\, w^\top w$. The coefficient L_1 is bounded by

$$\frac{2Ks}{\psi_{\min}\left(\sum_{a \in \mathcal{A}} \mathbb{E}_x\left[\phi(x,a)\phi(x,a)^\top\right]\right)}.$$

Proof of Proposition 3. For this proof we will adopt the shorthand $\mathbf{1}\{U_\lambda(a)\} \coloneqq \mathbf{1}\{x \in U_\lambda(a)\}$.

We first consider the ℓ_2 case. In this case (using $w \in \mathbb{R}^d$ as a stand-in for $f - f^\star$ and $\mathcal{W}^\star \coloneqq \mathcal{W} - w^\star \setminus \{0\}$) it is sufficient to

take

$$\begin{aligned}
L_{2,\lambda} &\leq \sup_{w\in\mathcal{W}^\star} \frac{\sup_{x\in\mathcal{X},a}(w^\intercal\phi(x,a))^2}{\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x(w^\intercal\phi(x,a)\mathbf{1}\{U_\lambda(a)\})^2} \\
&\leq \sup_{w\in\mathcal{W}^\star} \frac{\|w\|_2^2}{\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x(w^\intercal\phi(x,a)\mathbf{1}\{U_\lambda(a)\})^2} \\
&\leq \sup_{w\in\mathcal{W}^\star} \frac{\|w\|_2^2}{\|w\|_2^2\lambda_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\mathbf{1}\{U_\lambda(a)\}\right)} \\
&= \frac{1}{\lambda_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\mathbf{1}\{U_\lambda(a)\}\right)}.
\end{aligned}$$

In the sparse high-dimensional setting we have

$$\begin{aligned}
L_{2,\lambda} &\leq \sup_{w\in\mathcal{W}^\star} \frac{\sup_{x\in\mathcal{X},a}(w^\intercal\phi(x,a))^2}{\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x(w^\intercal\phi(x,a)\mathbf{1}\{U_\lambda(a)\})^2} \\
&\leq \sup_{w\in\mathcal{W}^\star} \frac{2s\|w\|_2^2}{\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x(w^\intercal\phi(x,a)\mathbf{1}\{U_\lambda(a)\})^2} \\
&\leq \sup_{w\in\mathcal{W}^\star} \frac{2s\|w\|_2^2}{\|w\|_2^2\psi_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\mathbf{1}\{U_\lambda(a)\}\right)} \\
&= \frac{2s}{\psi_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\mathbf{1}\{U_\lambda(a)\}\right)}.
\end{aligned}$$

As remarked in the main body, in general it holds that $L_1 \leq L_{2,0}$. Nonetheless, it is also possible to directly bound L_1 using similar reasoning to the proof above:

$$L_1 \leq \frac{1}{\lambda_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\right)}$$

for the ℓ_2 example and

$$L_1 \leq \frac{2s}{\psi_{\min}\left(\frac{1}{K}\sum_{a\in\mathcal{A}}\mathbb{E}_x\,\phi(x,a)\phi(x,a)^\intercal\right)},$$

for the sparsity example. □

B. Experimental Details

B.1. Datasets

We evaluated on datasets for learning-to-rank and for multiclass classification.

The learning-to-rank datasets, which were previously used for evaluating contextual semibandits in (Krishnamurthy et al., 2016), are as follows:

- Microsoft Learning to Rank (Qin & Liu, 2010). We use the `MSLR-WEB30K` variant available at https://www.microsoft.com/en-us/research/project/mslr/. This dataset has $T = 31278$, $d = 136$. We limit the choices to $K = 10$ documents (actions) per query. The MSLR repository comes partitioned into five segments, each with $T = 31278$ queries and a varying number of documents. We use the first three segments for the documents in our training dataset and use documents from the fourth segment for validation.
- Yahoo! Learning to Rank Challenge V2.0 (Chapelle & Chang, 2011) (variant C14B at https://webscope.sandbox.yahoo.com/catalog.php?datatype=c). The dataset has $T = 33850$, $d = 415$, and $K = 6$. We hold out 7000 examples for test.

Each learning-to-rank dataset contains over $30,000$ queries, with the number of documents varying. In both datasets feedback each document is labeled with relevance score in $\{0,\dots,4\}$. We transform this to a contextual bandit problem by presenting K documents as actions and their relevance scores as the rewards, so that the goal of the learner is to choose the document with the highest relevance each time it is presented with a query.

The multiclass classification datasets are taken from the UCI repository (Lichman, 2013) summarized in Table 1. This collection was previously used for evaluating contextual bandit learning in (Dudík et al., 2011). Each context is labeled with the index in $[K]$ of the class to which the context belongs, and the reward for selecting a class is 1 if correct, 0 otherwise.

Validation Validation is performed by simulating the algorithm's predictions on examples from a holdout set without allowing the algorithm to incorporate these examples. The validation error at round t therefore approaches the instantaneous expected reward $\mathbb{E}_{x_t,a_t}[f^\star(x_t,a_t)]$ at a rate determined by uniform convergence for the class $\mathcal{F}$. We also plot the validation reward of a "supervised" baseline obtained by training the oracle (either Linear or GB5) on the entire training set at once (including rewards for all actions).

Noisy dataset variants For all of the multiclass datasets we also create an alternate version with real-valued costs by constructing a reward matrix $R_t \in \mathbb{R}^{K\times K}$ and returning $R_t(a,a^\star)$ as the reward for selecting action a when $a^\star$ is the correct label at time t. R_t is constructed as a (possibly asymmetric) matrix with all ones on the diagonal ($R_t(a,a) = 1$) and random values in the range $[0,1]$ for each off-diagonal entry. The off diagonal elements are generated through the following process: 1) For each off-diagonal pair (a,a') draw a "mean" $\mu(a,a') \in [0,1]$ uniformly at random. This value of μ is held constant across all timesteps and all repetitions. 2) At time t, sample $R_t(a,a')$ as a Bernoulli random variable with bias $\mu(a,a')$. The reward matrices that were sampled are included in Section B.7 for reference.

Table 1. UCI datasets (before validation split).

Dataset	T	d	K
letter	20000	17	26
optdigits	5620	65	10
adult	45222	105	2
page-blocks	5473	11	5
pendigits	10992	17	10
satimage	6430	37	6
vehicle	846	19	4
yeast	1479	9	9

B.2. Benchmark algorithms

We compared with the following benchmark algorithms:

- ϵ-Greedy (Langford & Zhang, 2008). Policy is updated on a doubling schedule: Every $2^{i/2}$ rounds. We use an exploration probability of $\max\{1/\sqrt{t},\epsilon\}$ at time t, then tune ϵ as described in the main paper.
- ILOVETOCONBANDITS (Agarwal et al., 2014): Updated every $2^{i/2}$ rounds. We tune the constant in front of the parameter μ_m described in Algorithm 1 in (Agarwal et al., 2014).
- Bootstrap-TS (Dimakopoulou et al., 2017): At each epoch, the algorithm draws N bootstrap replicates of the dataset so far, then fits a predictor in $\mathcal{F}$ to each replicate, giving a collection of predictors $(f_i)_{i\in[N]}$. To predict on a new context x we compute the mean $\mathrm{Mean}(a)$ and variance $\mathrm{Var}(a)$ of the predictions $f_i(x,a)$. To pick an action we first sample $z_a \sim \mathcal{N}(\mathrm{Mean}(a), \beta\cdot\mathrm{Var}(a))$ for each $a\in[K]$, then set $a_t = \arg\max_a z_a$. The parameter $\beta > 0$ is tuned.

  We also experimented with a UCB variant of the Bootstrap algorithm (Dimakopoulou et al., 2017) in which the algorithm picks the action maximizing $\mathrm{Mean}(a) + \sqrt{\beta\cdot\mathrm{Var}(a)}$. The results of this experiment are omitted as we found the UCB variant to be dominated by the Thompson sampling variant.
- As discussed in the main body, we tune the parameter $\beta = \beta_m$ for both RegCB variants.

- We also compared against a pure exploitation strategy but found that this was uniformly outperformed by RegCB and Bootstrap-TS. These results are omitted for space.

For each algorithm we tried 8 different values of the relevant parameter coming from a logarithmically spaced grid ranging from 10^2 to 10^{-8} for the confidence interval-based algorithms (RegCB and Bootstrap-TS) and range 10^{-1} to 10^{-8} for ϵ-Greedy and ILTCB.

Each algorithm must be supplied with a model class $\mathcal{F}$ and an optimization oracle for this class. Both the model class $\mathcal{F}$ and the oracle implementation are hyperparameters. How to choose the oracle once the class $\mathcal{F}$ is been fixed is discussed below.

B.3. Oracle implementation

All of the oracle-based algorithms require optimization oracles, for either predictor classes or policy classes. We consider the following three types of basic oracles.

1. Weighted regression onto single action

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} w_t (f(x_t, a_t) - r_t(a_t))^2.$$

2. Weighted regression onto all actions

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} w_{t,a} (f(x_t, a) - r_t(a))^2.$$

3. Weighted multiclass classification

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} w_t \mathbf{1}\{\pi_f(x_t) \neq a_t\}.$$

Oracles for importance-weighted observations One of the most common datasets one needs to optimize over to implement oracle-based contextual bandit algorithms is an importance weighted history of interactions. That is, $H_T = \{(x_t, a_t, r_t(a_t), p_t(a_t))\}_{t=1}^{T}$, where x_t and $r_t(a_t)$ are the unmodified context and reward provided by nature, a_t is the action selected by a randomized contextual bandit algorithm, and $p_t(a_t)$ is the (positive) probability that a_t was selected. The core optimization problem that must be solved for such a dataset (e.g., in ϵ-Greedy) is

$$\arg\max_{f\in\mathcal{F}} \sum_{t=1}^{T} \frac{r_t(a_t)}{p_t(a_t)} \mathbf{1}\{\pi_f(x_t) = a_t\}. \tag{18}$$

This problem most naturally reduces to weighted multiclass classification, but under the realizability assumption in Assumption 1 it can also be reduced to regression in a number of principled ways. The full list of possible reductions we consider is as follows:

- Unweighted regression

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} (f(x_t, a_t) - r_t(a_t))^2. \tag{A}$$

- Importance-weighted regression

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} \frac{1}{p_t(a_t)} (f(x_t, a_t) - r_t(a_t))^2. \tag{B}$$

- Regression with importance weighted targets

$$\arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} (f(x_t, a_t) - r_t(a_t)/p_t(a_t))^2 + \sum_{a\neq a_t} (f(x_t, a))^2. \tag{C}$$

- Importance-weighted multiclass
$$\arg\min_{\pi\in\Pi} \sum_{t=1}^{T} \frac{r_t(a_t)}{p_t(a_t)} \mathbf{1}\{\pi(x_t) \neq a_t\}. \tag{D}$$
Note that in this case the policy class Π is not necessarily induced by a predictor class $\mathcal{F}$, though when it is it may be possible to further reduce this optimization problem to one of the first three problems.

The minimizer of (D) corresponds to the maximizer of (18). Reductions (A), (B), and (C) all have the property that if the conditional expectation version of the loss (e.g. $\sum_{t=1}^{T} \mathbb{E}_{(x_t,r_t)\sim\mathcal{D}} \mathbb{E}_{a_t|x_t,H_{t-1}} \left[(f(x_t,a_t) - r_t(a_t))^2 \right]$ for (A)) is used, then the Bayes predictor $f^\star(x,a) = \mathbb{E}[r(a) \mid x]$ is the minimizer when $f^\star \in \mathcal{F}$, which (via uniform convergence) justifies the use of the empirical versions.

Oracle choices for benchmark algorithms Depending on the needs of each benchmark algorithm, (A), (B), (C), or (D) as well as other oracles may be possible to use or required. We discuss the choices for each benchmark

- ϵ-Greedy: This strategy only needs to solve an importance weighted argmax of the form (18), so all of (A), (B), (C), and (D) can be used under realizability. Note that since actions are sampled uniformly in the non-greedy rounds, (A) and (B) are equivalent under this strategy. In experiments we use (B).
- Bootstrap-TS: Like RegCB, this strategy is tailored to the realizable regression setting, so (A) suffices. While (B) and (C) could also be used, we expect them to have higher variance.
- ILOVETOCONBANDITS: This algorithm requires two different oracles. First, it requires the optimization problem (18) to be solved on the unmodified reward/context sequence. Second, it requires a bonafide *cost-sensitive classification* optimization oracle of the form
$$\arg\max_{f\in\mathcal{F}} \sum_{t=1}^{T} r_t(\pi_f(x_t))$$
for an artificial sequence of rewards which may not be realizable even when the rewards given by nature are. As in ϵ-Greedy, the first oracle can use (A), (B), (C), and (D). The second oracle is more complicated. Cost-sensitive classification is typically not implemented directly and instead is reduced to either weighted multiclass (D) or multi-output regression, for which (C) is a special case. Note that (D) can further be reduced to (A), (B), (C), but because we do not expect realizability to hold it is more natural to use the direct reduction to (C) in this case. In experiments we used (B) for empirical regret minimizer and (C) for the cost-sensitive classifier to solve the optimization problem OP in Agarwal et al. (2014).

Label-dependent features For different datasets we consider different instantiations of the general predictor class setup described in the main paper. We assume there is a base context space $\mathcal{Z}$ and predictor class $\mathcal{G} : \mathcal{Z} \to \mathbb{R}$. Give such a class there are two natural ways to build a class of predictors over the joint context-action space depending on how the dataset is featurized.

- **Label-dependent features** For the MSLR and Yahoo datasets, each context comes with a distinct set of features for each action. This is captured by our abstraction by defining a fixed feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathcal{Z}$, then defining the class $\mathcal{F}$ via $\mathcal{F} = \{(x,a) \mapsto g(\phi(x,a)) \mid g \in \mathcal{G}\}$.
- **Label-independent features** When the contexts do not have label-dependent features, we use one instance of the base real-valued predictor class $\mathcal{G}$ for each action, i.e. we set $\mathcal{Z} = \mathcal{X}$ and take $\mathcal{F} = \left\{(x,a) \mapsto g_a(x) \mid g = (g_a)_{a\in\mathcal{A}} \in \mathcal{G}^{\mathcal{A}}\right\}$.

Predictor class and base oracle implementation We use real-valued predictors from the scikit-learn library (Pedregosa et al., 2011). The two predictor classes used were

- `sklearn.linear_model.Ridge(alpha=1)`
- `sklearn.tree_model.GradientBoostingRegressor(max_depth=5, n_estimators=100)`.

Each of the scikit-learn predictor classes handles this real-valued output case directly, via the `fit()` function for each class. In the label-dependent feature case we use a single oracle for $\mathcal{G}$, and in the label-independent feature case we use the oracle for $\mathcal{G}$, then take $\mathcal{F} = \mathcal{G}^{\mathcal{A}}$, so that there are actually $|\mathcal{A}|$ oracle instances.

Incremental implementation for RegCB As mentioned in the main body, we restrict the optimization for the gradient boosting oracle when used with RegCB. At the beginning of each epoch m, we find best regression tree ensemble on the dataset so far (with respect to $\widehat{R}_m$). For each round within the epoch, we keep this tree structure fixed for the call to ORACLE(H), so that only the ensemble and leaf weights need to be re-optimized.

B.4. Holdouts and multiple trials

Each dataset shuffled via random permutation, then presented to the learner in order.

Each (algorithm, parameter configuration, dataset) tuple was run for 5 repetitions. For a given trial we distinguish between two sources of randomness: Randomness from the dataset, which may come from the random ordering or from randomness in the labels as described in the dataset section, and randomness in the contextual bandit algorithm's decisions. We control for randomness in the dataset across different (algorithm, parameter) configurations by giving each repetition an index k and using the same random seed to select the dataset randomness across all configurations. This means that when k is fixed, all variance is induced by the algorithm's action distribution.

Validation reward was computed every $T/15$ steps.

B.5. Additional details for disagreement plots

For Figure 3 and Figure 6 all plots are averaged using a sliding window of length 20. The set A_t is well-defined for both RegCB-Opt and Bootstrap-TS even through neither algorithm instantiates it explicitly. For the `yahoo` and `mslr` datasets $|A_t|$ is technically a lower bound on the true disagreement set size $|A_{\mathcal{F}_m}(x_t)|$ because our classes $\mathcal{F}$ do not have product structure on these datasets—see discussion in Section 4.1.

B.6. Full collection of plots

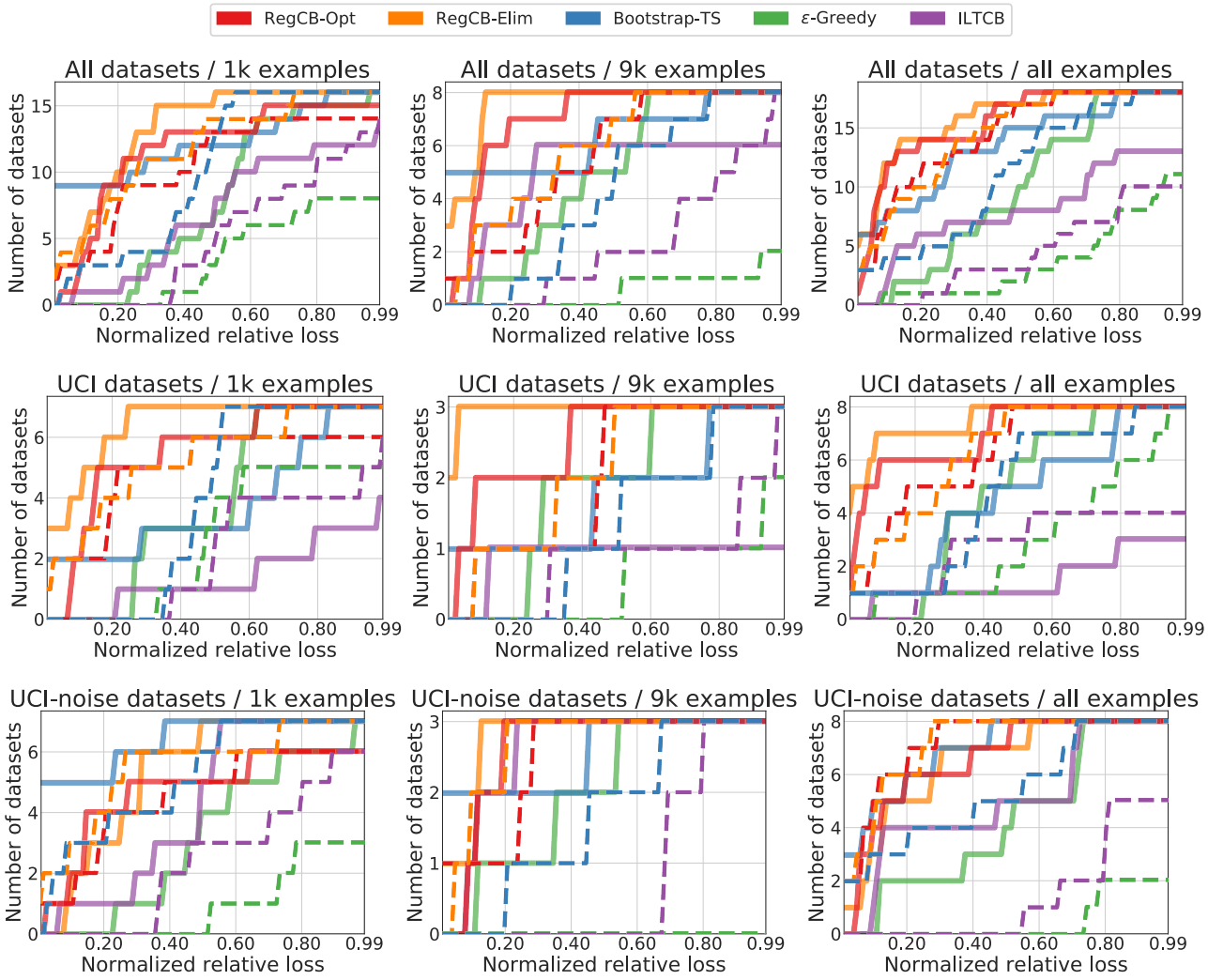

Figure 4. Cumulative performance across all data sets at various sample sizes.

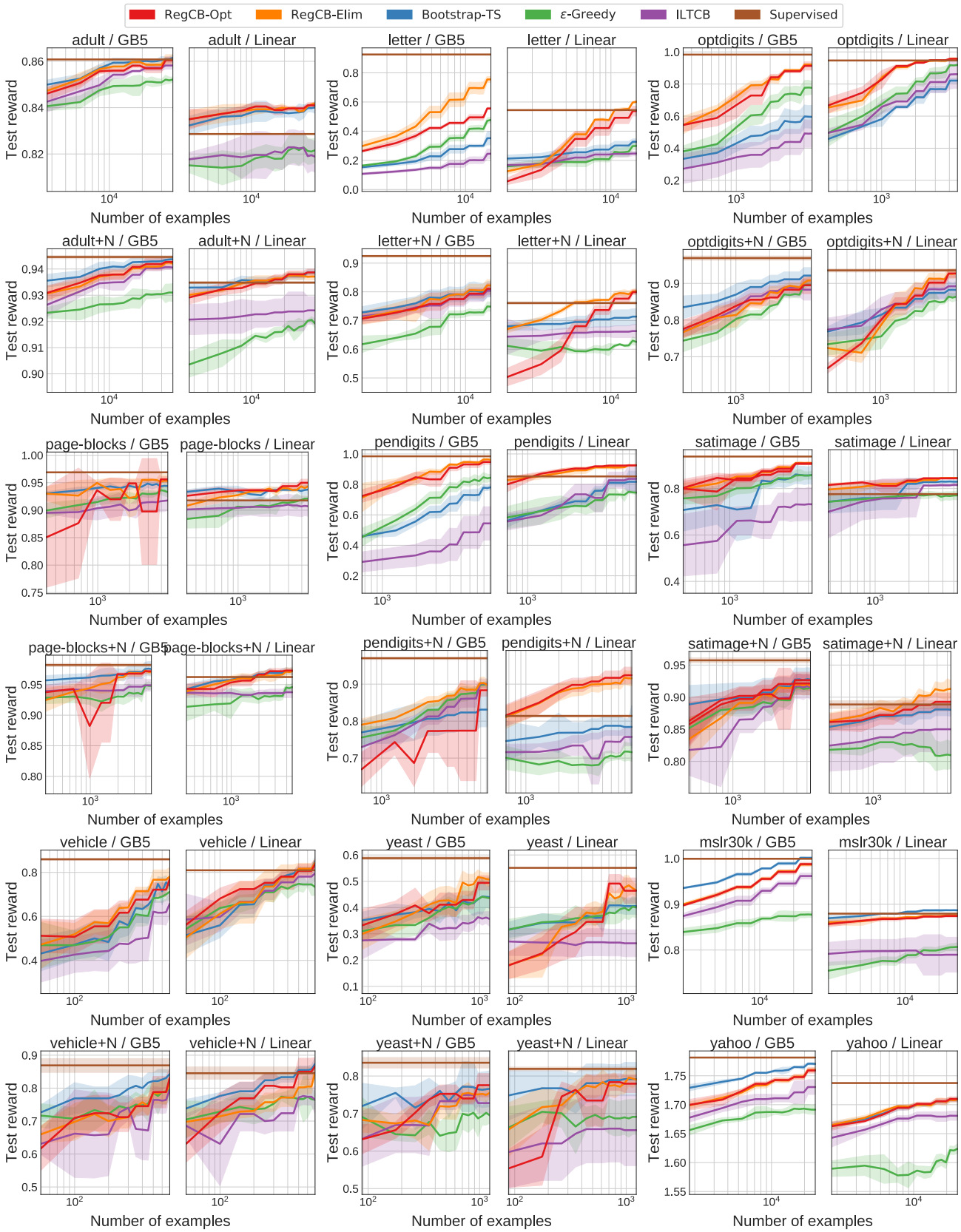

Figure 5. Performance on individual datasets.

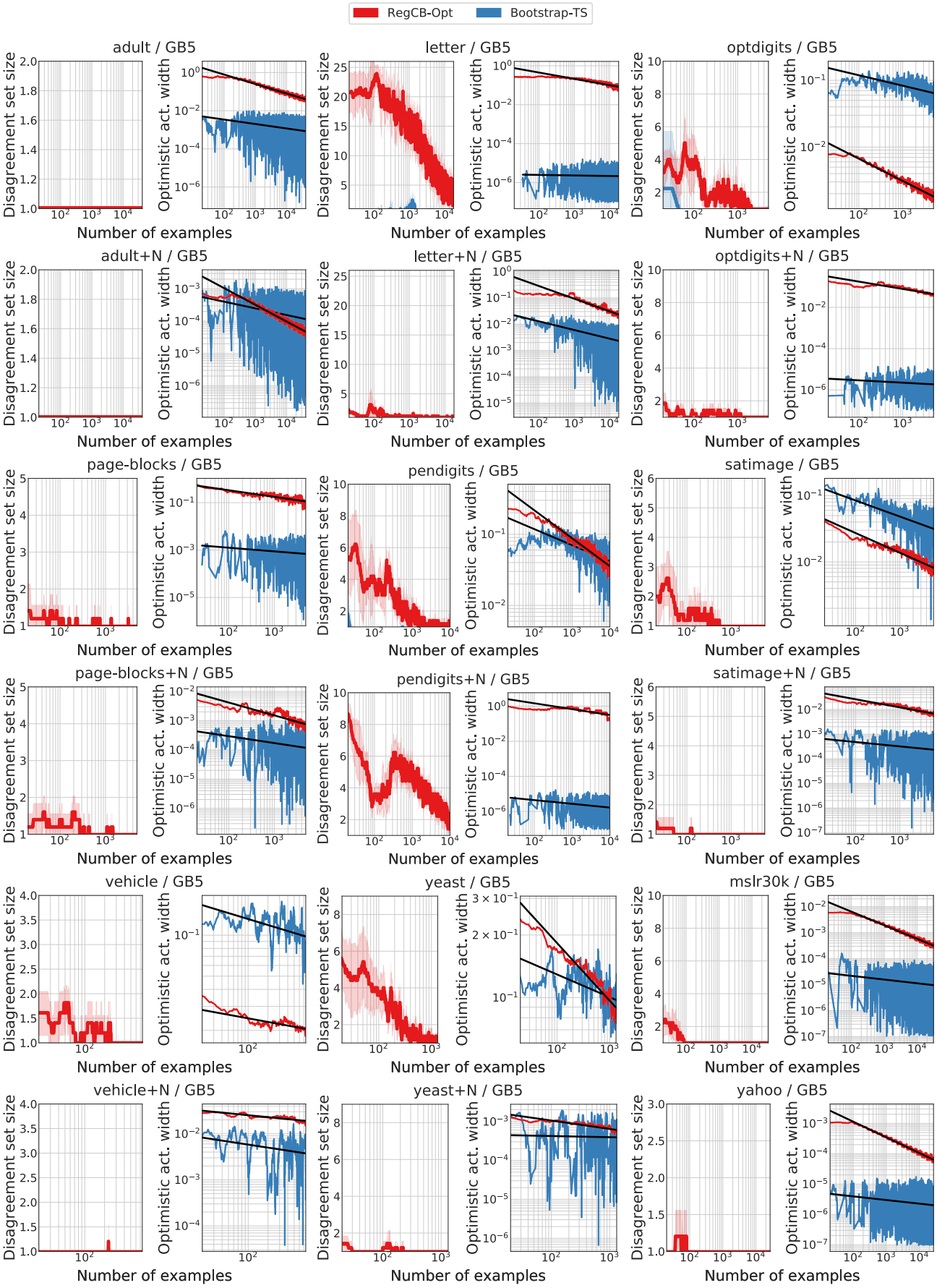

Figure 6. Size of disagreement set and confidence width.

B.7. UCI Reward Matrices

Table 2. Mean reward matrix: yeast

1.00	0.02	0.26	0.89	0.20	0.31	0.97	0.34	0.39
0.25	1.00	0.65	0.29	0.03	0.52	0.30	0.10	0.09
0.49	0.92	1.00	0.61	0.33	0.84	0.22	0.37	0.62
0.21	0.67	0.13	1.00	0.27	0.69	0.39	0.97	0.15
0.26	0.63	0.11	0.39	1.00	0.25	0.34	0.17	0.69
0.78	0.11	0.22	0.16	0.22	1.00	0.68	0.39	0.55
0.40	0.71	0.40	0.45	0.23	0.48	1.00	0.99	0.43
0.41	0.48	0.85	0.87	0.22	0.46	0.33	1.00	0.96
0.07	0.76	0.49	0.74	0.44	0.85	0.09	0.23	1.00

Table 3. Mean reward matrix: letter

1.00	0.33	0.82	0.04	0.11	0.60	0.53	0.42	0.34	0.62	0.44	0.74	0.52	0.58	0.65	0.99	0.82	0.41	0.88	0.82	0.05	0.72	0.80	0.74	0.71	0.54
0.12	1.00	0.40	0.22	0.72	0.99	0.26	0.67	0.60	0.72	0.94	0.35	0.25	0.40	0.75	0.72	0.41	0.99	0.45	0.37	0.71	0.08	0.40	0.77	0.76	0.28
0.19	0.47	1.00	0.73	0.19	0.33	0.84	0.62	0.89	0.98	0.84	0.18	0.62	0.48	0.40	0.74	0.83	0.68	0.14	0.70	0.06	0.19	0.92	0.41	0.15	0.68
0.16	0.65	0.25	1.00	0.96	0.07	0.51	0.34	0.66	0.84	0.60	0.59	0.12	0.71	0.20	0.49	0.04	0.32	0.86	0.56	0.55	0.37	0.83	0.28	0.13	0.56
0.27	0.78	0.18	0.78	1.00	0.04	0.56	0.67	0.94	0.79	0.75	0.50	0.04	0.82	0.01	0.55	0.57	0.11	0.06	0.57	0.49	0.30	0.04	0.63	0.12	0.01
0.28	0.30	0.18	0.07	0.78	1.00	0.25	0.52	0.25	0.85	0.48	0.62	0.97	0.35	0.22	0.98	0.59	0.98	0.97	0.71	0.02	0.61	0.25	0.13	0.37	0.20
0.77	0.93	0.03	0.26	0.27	0.14	1.00	0.25	0.36	0.05	0.24	0.88	0.96	0.66	0.30	0.06	0.86	0.16	0.27	0.55	0.25	0.84	0.50	0.48	0.91	0.92
0.24	0.02	0.67	0.27	0.01	0.10	0.42	1.00	0.21	0.75	0.46	0.11	0.22	0.93	0.01	0.64	0.64	0.68	0.58	0.78	0.82	0.65	0.18	0.73	0.28	0.84
0.57	0.09	0.91	0.46	0.94	0.04	0.11	0.76	1.00	0.45	0.82	0.42	0.19	0.84	0.11	0.29	0.22	0.46	0.32	0.91	0.79	0.71	0.14	0.61	0.85	0.92
0.66	0.26	0.28	0.64	0.72	0.31	0.68	0.51	0.83	1.00	0.91	0.12	0.84	0.95	0.57	0.00	0.03	0.41	0.46	0.48	0.68	0.75	0.82	0.35	0.61	0.39
0.73	0.56	0.59	0.39	0.63	0.87	0.65	0.13	0.09	0.68	1.00	0.31	0.89	0.86	0.81	0.36	0.64	0.60	0.24	0.59	1.00	0.05	0.24	0.33	0.80	0.44
0.06	0.32	0.83	0.74	0.28	0.73	0.32	0.15	0.98	0.26	0.61	1.00	0.64	0.43	0.40	0.05	0.08	0.45	0.92	0.23	0.87	0.81	0.17	0.31	0.43	0.86
0.63	0.82	0.50	0.58	0.45	0.26	0.62	0.58	0.87	0.92	0.57	0.69	1.00	0.68	1.00	0.94	0.14	0.94	0.04	0.03	0.18	0.31	0.98	0.94	0.76	0.62
0.97	0.57	0.21	0.13	0.76	0.53	0.82	0.79	0.67	0.78	0.69	0.43	0.83	1.00	0.78	0.09	0.95	0.48	0.89	0.08	0.94	0.31	0.42	0.69	0.09	0.21
0.58	0.39	0.11	0.01	0.90	0.67	0.32	0.89	0.97	0.08	0.26	0.53	0.92	0.23	1.00	0.90	0.34	0.23	0.18	0.05	0.96	0.15	0.96	0.34	0.06	0.82
0.80	0.46	0.77	0.75	0.45	0.28	0.14	0.91	0.08	0.73	0.08	0.67	0.06	0.11	0.48	1.00	0.03	0.64	0.90	0.48	0.84	0.71	0.93	0.97	0.59	0.95
0.71	0.46	0.92	0.58	0.24	0.39	0.42	0.16	0.02	0.05	0.68	0.25	0.15	0.20	0.82	0.89	1.00	0.74	0.58	0.49	0.64	0.95	0.80	0.41	0.25	0.00
0.29	0.98	0.42	0.54	0.06	0.14	0.99	0.54	0.22	0.64	0.73	0.50	0.33	0.72	0.13	0.72	0.45	1.00	0.63	0.86	0.32	0.70	0.12	0.44	0.72	0.89
0.56	0.63	0.53	0.35	0.85	0.57	0.26	0.80	0.83	0.45	0.68	0.09	0.72	0.34	0.02	0.71	0.55	0.83	1.00	0.99	0.33	0.13	0.04	0.32	0.21	0.57
0.96	0.22	0.33	0.27	0.27	0.69	0.89	0.58	0.40	0.43	0.55	0.31	0.26	0.91	0.51	0.12	0.57	0.25	0.01	1.00	0.36	0.68	0.61	0.17	0.30	0.72
0.43	0.13	0.17	0.73	0.62	0.56	0.06	0.39	0.45	0.58	0.70	0.72	0.59	0.27	0.41	0.78	0.47	0.40	0.85	1.00	1.00	0.63	0.91	0.15	0.29	0.65
0.18	0.28	0.94	0.31	0.10	0.50	0.08	0.25	0.96	0.84	0.15	0.25	0.05	0.20	0.81	0.91	0.62	0.09	0.50	0.67	0.11	1.00	0.76	0.39	0.83	0.17
0.26	0.80	0.68	0.78	0.18	0.95	0.18	0.70	0.31	0.51	0.91	0.78	0.75	0.11	0.91	0.90	0.98	0.11	0.38	0.27	0.85	0.90	1.00	0.22	0.05	0.88
0.95	0.75	0.82	0.31	0.13	0.10	0.67	0.14	0.92	0.24	0.75	0.61	0.34	0.63	0.02	0.76	0.17	0.61	0.12	0.57	0.73	0.80	0.14	1.00	0.41	0.40
0.83	0.19	0.76	0.74	0.42	0.14	0.70	0.88	0.18	0.12	0.21	0.44	0.46	0.76	0.16	0.90	0.52	0.28	0.02	0.59	0.20	0.44	0.96	0.20	1.00	0.84
0.03	0.67	0.47	0.34	0.50	0.43	0.56	0.11	0.36	0.93	0.50	0.64	0.47	0.97	0.12	0.35	0.68	0.79	0.40	0.74	0.37	0.10	0.02	0.14	0.99	1.00

Table 4. Mean reward matrix: optdigits

1.00	0.56	0.12	0.40	0.78	0.51	0.18	0.85	0.96	0.98
0.19	1.00	0.23	0.03	0.95	0.92	0.29	0.17	0.40	0.51
0.31	0.43	1.00	0.56	0.83	1.00	0.33	0.09	0.77	0.15
0.73	0.96	0.07	1.00	0.84	0.15	0.77	0.78	0.68	0.13
0.04	0.66	0.25	0.99	1.00	0.06	0.70	0.63	0.90	0.16
0.61	0.32	0.76	0.16	0.93	1.00	0.83	0.23	0.11	0.67
0.58	0.88	1.00	0.28	0.74	0.28	1.00	0.49	0.87	0.16
0.97	0.05	0.70	0.65	0.05	0.20	0.33	1.00	0.37	0.53
0.35	0.51	0.26	0.85	0.62	0.30	0.78	0.90	1.00	0.86
0.82	0.87	0.38	0.61	0.42	0.24	0.06	0.82	0.38	1.00

Table 5. Mean reward matrix: page-blocks

1.00	0.38	0.66	0.16	0.96
0.35	1.00	0.24	0.59	0.41
0.14	0.54	1.00	0.77	0.93
0.09	0.20	0.99	1.00	0.24
0.63	0.73	0.69	0.03	1.00

Table 6. Mean reward matrix: pendigits

1.00	0.37	0.56	0.96	0.74	0.82	0.10	0.93	0.61	0.60
0.09	1.00	0.66	0.44	0.55	0.70	0.59	0.05	0.56	0.77
0.91	0.09	1.00	0.46	0.45	1.00	0.16	0.71	0.16	0.81
0.04	0.53	0.17	1.00	0.05	0.24	0.67	0.78	0.70	0.33
0.49	0.52	0.30	0.46	1.00	0.50	0.40	0.73	0.86	0.03
0.29	0.79	0.46	0.01	0.42	1.00	0.60	0.32	0.98	0.59
0.13	0.52	0.36	0.01	0.10	0.78	1.00	0.20	0.62	0.64
0.27	0.13	0.47	0.39	0.41	0.38	0.29	1.00	0.43	0.78
0.70	0.78	0.29	0.21	0.50	0.13	0.17	0.25	1.00	0.23
0.63	0.63	0.53	0.74	0.82	0.37	0.80	0.88	0.59	1.00

Table 7. Mean reward matrix: satimage

1.00	0.06	0.12	0.79	0.98	0.27
0.87	1.00	0.64	0.78	0.63	0.13
1.00	0.63	1.00	0.62	0.34	0.76
0.11	0.52	0.63	1.00	0.11	0.29
0.07	0.67	0.23	0.52	1.00	0.45
0.73	0.97	0.20	0.72	0.79	1.00

Table 8. Mean reward matrix: vehicle

1.00	0.36	0.18	0.52
0.01	1.00	0.80	0.76
0.67	0.03	1.00	0.40
0.19	0.77	0.62	1.00

Table 9. Mean reward matrix: adult

1.00	0.61
0.66	1.00