How Not to Overpackage: AI for Sustainability in HelloFresh's Service Supply Chain

Evelyn Xiao-Yue Gong, Michael Johnson

Appendix A: Proof of Theorem 2

Proof of Theorem 2 Let $m = \lceil T^{1/3} \rceil$, and partition $[0,1]$ into intervals $I_1, \dots, I_m$ of length at most $h = 1/m$. For each interval, define the interval-averaged loss $\bar{\ell}_i(a) = \mathbb{E}[\ell(X,a) \mid X \in I_i]$ where intervals with zero probability mass may be ignored. The function $\bar{\ell}_i$ is convex in a, because it is an average of convex functions. Let

$$\mathrm{OPT}_i = \min_{a \in \mathcal{A}} \bar{\ell}_i(a), \qquad n_i = \sum_{t=1}^{T} \mathbf{1}\{x_t \in I_i\}.$$

On interval I_i, the algorithm runs a contextless bandit algorithm. Conditional on the arrival times to I_i, the observed losses for a fixed action are i.i.d. with mean $\bar{\ell}_i(a)$. Hence the result of Agarwal et al. (2011) gives, for $\mathcal{A} = [0,1]$,

$$\mathbb{E}\left[\sum_{t : x_t \in I_i} \bar{\ell}_i(a_t) - n_i \mathrm{OPT}_i \,\middle|\, n_i \right] \le \widetilde{O}(\sqrt{n_i}).$$

For $\mathcal{A} = \{0,1\}$, the same bound follows from the standard two-armed stochastic-bandit regret bound. Summing over intervals and using Jensen's inequality,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_{i_t}(a_t) - \sum_{i=1}^{m} n_i \mathrm{OPT}_i \right] \\ &\le \mathbb{E} \sum_{i=1}^{m} \widetilde{O}(\sqrt{n_i}) \le \widetilde{O}\left(\sqrt{mT}\right) = \widetilde{O}(T^{2/3}) \end{aligned} \tag{11}$$

where i_t denotes the index of the interval containing x_t. It remains to translate this interval-averaged benchmark into regret against the context-by-context optimum. Let L be a Lipschitz constant for ℓ in x. For any $x \in I_i$ and any a, $|\ell(x,a) - \bar{\ell}_i(a)| \le Lh$. Therefore, if $a^*(x) \in \arg\min_{a \in \mathcal{A}} \ell(x,a)$, then for every round t with $x_t \in I_{i_t}$,

$$\begin{aligned} \ell(x_t, a_t) - \ell(x_t, a^*(x_t)) &\le \bar{\ell}_{i_t}(a_t) - \bar{\ell}_{i_t}(a^*(x_t)) + 2Lh \\ &\le \bar{\ell}_{i_t}(a_t) - \mathrm{OPT}_{i_t} + 2Lh. \end{aligned}$$

Summing over t and taking expectations, Equation (11) contributes $\widetilde{O}(T^{2/3})$, while the discretization term contributes $2LTh = O(T^{2/3})$. This proves the claimed regret bound. □

Appendix B: Proof of Theorem 3

Proof of Theorem 3 Let $m = \lceil T^{1/3} \rceil$ and $h = 1/m$. We construct a family of instances indexed by independent signs $\sigma_1, \dots, \sigma_m \in \{-1, 1\}$. Let

$$I_j = [(j-1)h, jh], \qquad J_j = [(j - \tfrac{2}{3})h, (j - \tfrac{1}{3})h].$$

For each j, define $g_j^- = (j - \frac{2}{3})h$ and $g_j^+ = (j - \frac{1}{3})h$. The function $\gamma_\sigma : [0,1] \to [0,1]$ satisfies $\gamma_\sigma((j-1)h) = (j-1)h$, $\gamma_\sigma(jh) = jh$, and on the middle third J_j,

$$\gamma_\sigma(x) = \begin{cases} g_j^-, & \sigma_j = -1, \\ g_j^+, & \sigma_j = 1. \end{cases}$$

On the first and last thirds of I_j, interpolate linearly. Every γ_σ is nondecreasing and 2-Lipschitz.

Fix a small absolute constant $\lambda = 1/100$, and define the expected penalty

$$\ell_\sigma(x,a) = \lambda|a - \gamma_\sigma(x)|.$$

The optimal action is $a^*(x) = \gamma_\sigma(x)$, which is increasing in x. Assumption 2 also holds: if $x > x'$, then $\gamma_\sigma(x) \geq \gamma_\sigma(x')$, and the map

$$a \mapsto |a - \gamma_\sigma(x)| - |a - \gamma_\sigma(x')|$$

is nonincreasing. It remains to realize this expected penalty through monotone Bernoulli events. Let $\rho = 1/4$, and set

$$p_\sigma(x,a) = \frac{1}{2} + \rho x + \lambda|a - \gamma_\sigma(x)| - \lambda a.$$

Then $p_\sigma(x,a) \in [0,1]$; in fact $p_\sigma(x,a) \in [0.49, 0.76]$. For $\xi \sim \mathrm{Unif}([0,1])$, define

$$L_\sigma(x,a,\xi) = \lambda a - \rho x - \frac{1}{2} + \mathbf{1}\{\xi \leq p_\sigma(x,a)\}.$$

Thus $\mathbb{E}[L_\sigma(x,a,\xi)] = \ell_\sigma(x,a)$. The realized loss is uniformly bounded, so an affine rescaling places it in $[0,1]$ and changes regret only by a constant factor.

The failure probability p_σ is monotone in the correct directions. For fixed x,

$$p_\sigma(x,a) = \frac{1}{2} + \rho x + \begin{cases} \lambda\gamma_\sigma(x) - 2\lambda a, & a \leq \gamma_\sigma(x), \\ -\lambda\gamma_\sigma(x), & a \geq \gamma_\sigma(x), \end{cases}$$

so it is nonincreasing in a. For $x > x'$, using the 2-Lipschitz property of γ_σ,

$$\begin{aligned} p_\sigma(x,a) - p_\sigma(x',a) &\geq \rho(x - x') \\ &\quad - \lambda|\gamma_\sigma(x) - \gamma_\sigma(x')| \\ &\geq (\rho - 2\lambda)(x - x') \geq 0. \end{aligned}$$

Therefore the common-uniform realization $Y(x,a) = \mathbf{1}\{\xi \leq p_\sigma(x,a)\}$ satisfies Definition 1: failures propagate to harder contexts and weaker actions, while nonfailures propagate to easier contexts and stronger actions.

We now prove the regret lower bound. Let $X_t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1])$. For each j, define

$$M_j = \sum_{t=1}^{T} \mathbf{1}\{X_t \in I_j\}, \qquad N_j = \sum_{t=1}^{T} \mathbf{1}\{X_t \in J_j\}.$$

With probability $1 - o_T(1)$, simultaneously for all j,

$$M_j \leq \frac{2T}{m}, \qquad N_j \geq \frac{T}{6m}.$$

Condition on any context sequence satisfying these inequalities. Fix j and all signs except σ_j. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the transcript laws under $\sigma_j = 1$ and $\sigma_j = -1$. The conditional one-sided feedback is deterministic given the played pair and the binary failure outcome, so it does not add information beyond the binary observation. Moreover, only rounds with $X_t \in I_j$ can distinguish the two instances. For any $x \in I_j$ and action a,

$$\begin{aligned} |p_+(x,a) - p_-(x,a)| &\leq \lambda|\gamma_+(x) - \gamma_-(x)| \\ &\leq \lambda h. \end{aligned}$$

Since all Bernoulli means lie in $[0.49, 0.76]$, the one-step KL divergence is at most $6\lambda^2 h^2$. Writing $D_j = D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-)$, we have $D_j \le 6\lambda^2 M_j h^2 \le 12\lambda^2 \frac{T}{m^3} \le \frac{1}{8}$, where the last inequality uses $\lambda = 1/100$ and $m = \lceil T^{1/3} \rceil$. Pinsker's inequality gives $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \le 1/4$.

On J_j, the two possible optimal actions are g_j^- and g_j^+, separated by $h/3$. Couple the two transcript laws so that, with probability at least $3/4$, the algorithm chooses the same action under the two signs. On this coupled event, for any chosen action a,

$$\lambda |a - g_j^-| + \lambda |a - g_j^+| \ge \frac{\lambda h}{3}.$$

Therefore, averaging over the two equally likely signs, the expected regret on each round with $X_t \in J_j$ is at least $\lambda h/8$. The conditional expected regret from interval j is at least

$$\frac{\lambda h}{8} N_j \ge c_0 \frac{T}{m^2}$$

for an absolute constant $c_0 > 0$. Summing over $j = 1, \dots, m$ yields $\Omega(T/m) = \Omega(T^{2/3})$ regret on the high-probability context event, and averaging over contexts proves the theorem. □

Appendix C: Proof of Theorem 4

Proof of Theorem 4 We give a construction with two actions $0 < 1$, where action 1 is stronger. Again let $m = \lceil T^{1/3} \rceil$ and $h = \frac{1}{m}$, and let $I_j = [(j-1)h, jh]$, $j = 1, \dots, m$. Let $J_j = [(j - \frac{2}{3})h, (j - \frac{1}{3})h]$ be the middle third of I_j. For $x \in I_j$, define

$$r_j^-(x) = \frac{3(x - (j-1)h)}{h}, \qquad r_j^+(x) = \frac{3(jh - x)}{h},$$

and

$$b_j(x) = h\big(1 \wedge r_j^-(x) \wedge r_j^+(x)\big),$$

and set $b_j(x) = 0$ outside I_j. Thus $b_j = h$ on J_j, b_j is supported on I_j, and b_j is 3-Lipschitz. Let $\sigma_1, \dots, \sigma_m$ be independent uniform signs in $\{-1, 1\}$. Fix small absolute constants

$$\alpha = \frac{1}{32}, \qquad \delta = \frac{1}{256}, \qquad c = \frac{1}{8}.$$

The material cost is $M(0) = 0$, $M(1) = c$. The failure probabilities are

$$p_\sigma(x, 0) = \frac{1}{2} + \alpha x,$$

and, for $x \in I_j$,

$$p_\sigma(x, 1) = \frac{1}{2} + \alpha x - c + \delta \sigma_j b_j(x).$$

These probabilities lie in $[0, 1]$. They are increasing in x, because $\alpha > 3\delta$, and $p_\sigma(x, 1) \le p_\sigma(x, 0)$, because $\delta b_j(x) \le \delta < c$. Hence, using the common-uniform realization $Y(x, a) = \mathbf{1}\{\xi \le p_\sigma(x, a)\}$, the instance has the conditional one-sided feedback property in Definition 1: a failure propagates to harder contexts and weaker actions, while a nonfailure propagates to easier contexts and stronger actions.

The realized penalty is $L(x, a, \xi) = M(a) + Y(x, a)$ so the expected penalties are $\ell_\sigma(x, 0) = \frac{1}{2} + \alpha x$, and, for $x \in I_j$, $\ell_\sigma(x, 1) = \frac{1}{2} + \alpha x + \delta \sigma_j b_j(x)$. Thus on J_j, the suboptimality gap is δh, and the optimal action is 0 if $\sigma_j = 1$ and 1 if $\sigma_j = -1$. The expected penalty functions are uniformly Lipschitz in x.

It remains to show that the signs cannot be learned quickly enough. Let $X_t \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$. For each j, define $M_j = \sum_{t=1}^{T} \mathbf{1}\{X_t \in I_j\}$ and $N_j = \sum_{t=1}^{T} \mathbf{1}\{X_t \in J_j\}$. With probability $1 - o_T(1)$, simultaneously for all j, we have $M_j \le \frac{2T}{m}$ and $N_j \ge \frac{T}{6m}$. Condition on any context sequence satisfying these inequalities. Fix j and all signs except σ_j. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the transcript laws under $\sigma_j = 1$ and $\sigma_j = -1$. Only rounds with $X_t \in I_j$ and action 1 have transcript probabilities that depend on σ_j. For $x \in I_j$, writing $u = \delta b_j(x) \le \delta h$, the Bernoulli KL divergence between the two possible observations is at most $C_0 u^2$ for an absolute constant C_0. Writing $D_j = D_{\text{KL}}(\mathbb{P}_+ \| \mathbb{P}_-)$, we have

$$D_j \le C_0 M_j \delta^2 h^2 \le 2C_0 \delta^2 \frac{T}{m^3} \le \frac{1}{8}$$

after decreasing the absolute constant δ, if necessary. Pinsker's inequality gives $\text{TV}(\mathbb{P}_+, \mathbb{P}_-) \le 1/4$.

At any time t with $X_t \in J_j$, action 1 is suboptimal under $\sigma_j = 1$, while action 0 is suboptimal under $\sigma_j = -1$. Hence

$$\mathbb{P}_+(A_t = 1) + \mathbb{P}_-(A_t = 0) \ge 1 - \text{TV}(\mathbb{P}_+, \mathbb{P}_-) \ge \frac{3}{4}.$$

Averaging over the two equally likely signs, the probability of choosing the suboptimal action on J_j is at least $3/8$. The conditional expected regret from interval j is therefore at least $\frac{3}{8} \cdot \delta h \cdot N_j \ge c_0 \frac{T}{m^2}$ for an absolute constant $c_0 > 0$. Summing over $j = 1, \dots, m$, and writing $R_T = \mathbb{E}[\text{Regret}(T) \mid X_1, \dots, X_T]$, yields

$$R_T \ge c_0 \frac{T}{m} = \Omega(T^{2/3})$$

on the high-probability context event. Averaging over the context sequence proves the claimed lower bound. □

Appendix D: Proof of Theorem 5

Proof of Theorem 5 Fix $d \ge 2$. Write $z(s) = (s, 1-s, 1/2, \dots, 1/2) \in [0,1]^d$ for $s \in [0,1]$, where the final $d-2$ coordinates are omitted when $d = 2$. The contexts are generated as $x_t = z(S_t)$, where $S_t \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$. If $s < s'$, then the first coordinate of $z(s)$ is smaller than the first coordinate of $z(s')$, while the second coordinate of $z(s)$ is larger than the second coordinate of $z(s')$. Hence neither $z(s) \preceq z(s')$ nor $z(s') \preceq z(s)$ under the coordinatewise Pareto order. Therefore the support D is an antichain, and any monotonicity condition that only compares Pareto-ordered contexts imposes no restriction on the values of the loss function on $D := \{z(s) : s \in [0,1]\}$.

We now construct a hard family of instances on this antichain. Again let $m = \lceil T^{1/3} \rceil$ and $h = \frac{1}{m}$, and partition $[0,1]$ into intervals

$$I_j = [(j-1)h, jh], \qquad j = 1, \dots, m.$$

Let

$$J_j = \left[(j - \tfrac{2}{3})h, (j - \tfrac{1}{3})h\right]$$

denote the middle third of I_j. For $s \in I_j$, define $b_j(s) = h\left(1 \wedge \frac{3(s-(j-1)h)}{h} \wedge \frac{3(jh-s)}{h}\right)$ and set $b_j(s) = 0$ for $s \notin I_j$. Thus $b_j(s) = h$ on J_j, b_j is supported on I_j, and b_j is uniformly Lipschitz.

Let $\sigma = (\sigma_1, \dots, \sigma_m) \in \{-1, 1\}^m$ be a vector of signs. For $s \in I_j$, define the expected losses of the two actions by

$$\ell^\sigma(z(s), 0) = \frac{1}{2} - \sigma_j \frac{b_j(s)}{16},$$
$$\ell^\sigma(z(s), 1) = \frac{1}{2} + \sigma_j \frac{b_j(s)}{16}.$$

The realized loss is Bernoulli with the corresponding mean. Since $b_j(s) \le h \le 1$, all means lie in $[0,1]$. Also, because b_j vanishes at the endpoints of I_j and is uniformly Lipschitz, the functions $s \mapsto \ell^\sigma(z(s), a)$ are uniformly Lipschitz on D. Since $\|z(s) - z(s')\|_2 = \sqrt{2}|s - s'|$, they are uniformly Lipschitz as functions on the line segment D. By the McShane extension theorem, followed by truncation to $[0,1]$, these functions may be extended to uniformly Lipschitz functions on all of $[0,1]^d$. Because the context distribution is supported on D, this extension does not change either the observations or the regret.

For $s \in I_j$, the optimal action is determined by σ_j: action 0 is optimal if $\sigma_j = 1$, and action 1 is optimal if $\sigma_j = -1$. On the middle third J_j, the gap between the two actions is

$$|\ell^\sigma(z(s), 1) - \ell^\sigma(z(s), 0)| = \frac{h}{8} = \frac{1}{8m}.$$

Thus every time $S_t \in J_j$, choosing the wrong action incurs instantaneous regret $1/(8m)$. Since D is an antichain, no two distinct support points are comparable under the Pareto order. Consequently any action map on D, including the optimal-action map induced by the signs σ_j, is monotone with respect to the Pareto partial order on the support.

We next show that the learner cannot identify the signs σ_j quickly enough. Let

$$M_j = \sum_{t=1}^{T} \mathbf{1}\{S_t \in I_j\}, \qquad N_j = \sum_{t=1}^{T} \mathbf{1}\{S_t \in J_j\}.$$

By standard Chernoff bounds, with probability $1 - o_T(1)$, simultaneously for all $j \in [m]$,

$$M_j \le \frac{2T}{m}, \qquad N_j \ge \frac{T}{6m}.$$

Condition on any context sequence satisfying these inequalities.

Fix an interval j and fix all signs except σ_j. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the laws of the full transcript under $\sigma_j = 1$ and $\sigma_j = -1$, respectively. Only observations with $S_t \in I_j$ have distributions depending on σ_j. Moreover, for any action and any $s \in I_j$, writing $u = b_j(s)/16$, we have

$$D_{\mathrm{KL}}\left(\mathrm{Ber}\left(\frac{1}{2} + u\right) \middle\| \mathrm{Ber}\left(\frac{1}{2} - u\right)\right) \le 16u^2 \le \frac{1}{16m^2}.$$

Therefore, by the chain rule for KL divergence,

$$D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-) \le \frac{M_j}{16m^2} \le \frac{T}{8m^3} \le \frac{1}{8}.$$

Pinsker's inequality gives $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \le \frac{1}{4}$.

At any time t with $S_t \in J_j$, action 1 is suboptimal under $\sigma_j = 1$, while action 0 is suboptimal under $\sigma_j = -1$. Hence, if a_t is the action chosen at time t, we have $\mathbb{P}_+(a_t = 1) + \mathbb{P}_-(a_t = 0) \ge 1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \ge \frac{3}{4}$. Averaging over the two equally likely values of σ_j, the probability of choosing the suboptimal action at such a time is therefore at least $3/8$. Consequently, conditional on the context sequence and on the other signs, the expected regret contributed by the middle third J_j is at least $\frac{3}{8} \cdot \frac{1}{8m} \cdot N_j \ge c\,\frac{T}{m^2}$ for an absolute constant $c > 0$. Summing over the m intervals gives

$$\mathbb{E}_\sigma[\mathrm{Regret}(T) \mid S_1, \ldots, S_T] \ge c\,\frac{T}{m} = \Omega(T^{2/3})$$

on the high-probability event above. Since regret is nonnegative, averaging over the context sequence yields

$$\mathbb{E}_\sigma[\mathrm{Regret}(T)] = \Omega(T^{2/3}).$$

This lower bound is for a uniform prior over the sign vectors $\sigma \in \{-1, 1\}^m$. Therefore, for every algorithm, there exists at least one fixed sign vector σ for which the expected regret is $\Omega(T^{2/3})$. This proves the theorem. □

Appendix E: Finite-Action Minimax Lower Bound

We prove the lower-bound assertion used in Theorem 1. The construction is a one-context, two-action instance of the packaging model, so it also lower-bounds the richer contextual problem.

Proof of the lower-bound claim in Theorem 1 Fix the single context x_0 and two ordered actions $a_1 < a_2$, where a_2 is the stronger packaging action. Let $G(a)$ denote the known part $E(a) + M(a)$ of the penalty and set

$$G(a_1) = 0, \qquad G(a_2) = \Delta, \qquad Q(a_1) = Q(a_2) = \frac{1}{2},$$

where $\Delta > 0$ will be chosen below. The realized penalty from action a is $L(a) = G(a) + \frac{1}{2}Y(a)$ where $Y(a)$ is the failure indicator. We compare two possible instances. In instance $+$,

$$p_+(a_1) = \frac{1}{2}, \qquad p_+(a_2) = \frac{1}{2} - 4\Delta,$$

while in instance $-$,

$$p_-(a_1) = p_-(a_2) = \frac{1}{2} - 4\Delta.$$

For $\Delta \le 1/16$, these are valid probabilities and both instances satisfy the action monotonicity of the failure probability, $p(a_1) \ge p(a_2)$.

The expected penalties are

	a_1	a_2
$+$	$1/4$	$1/4 - \Delta$
$-$	$1/4 - 2\Delta$	$1/4 - \Delta$.

Thus a_2 is uniquely optimal in instance $+$, a_1 is uniquely optimal in instance $-$, and the suboptimality gap is always Δ.

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the laws of the full transcript induced by an arbitrary possibly randomized algorithm under the two instances. The algorithm's own randomization can be included in the transcript. Conditional on the same past history, the action-selection rule is the same under both instances; the only difference in the observation distribution occurs when the algorithm chooses a_1. By the chain rule for KL divergence in adaptive experiments,

$$D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-) \le T\, D_{\mathrm{KL}}\left(\mathrm{Ber}\left(\frac{1}{2}\right) \middle\| \mathrm{Ber}\left(\frac{1}{2} - 4\Delta\right)\right).$$

For $\Delta \le 1/16$, the Bernoulli KL divergence is at most $64\Delta^2$. Choosing $\Delta = 1/(32\sqrt{T})$ gives $D_{\mathrm{KL}}(\mathbb{P}_+ \| \mathbb{P}_-) \le \frac{1}{16}$ and $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \le \frac{1}{4}$ by Pinsker's inequality. Let A_t be the action chosen at time t. The regret in instance $+$ is Δ whenever $A_t = a_1$, and the regret in instance $-$ is Δ whenever $A_t = a_2$. Therefore

$$\begin{aligned} R_+ + R_- &= \Delta \sum_{t=1}^{T} \left[\mathbb{P}_+(A_t = a_1) + \mathbb{P}_-(A_t = a_2)\right] \\ &\ge \Delta \sum_{t=1}^{T} \left(1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)\right) \ge \frac{3}{4}\,\Delta T. \end{aligned}$$

Hence at least one of the two instances has expected regret at least $\frac{3}{8}\Delta T = \Omega(\sqrt{T})$. This proves the claimed minimax lower bound in the dependence on T. □

Appendix F: EXP4 Baseline for Adversarial Contexts

Our main algorithm uses random context arrivals to build uniform confidence bounds over intervals of the context order. If contexts are instead chosen adversarially, this particular analysis no longer applies. A generic way to recover a sublinear guarantee is to add a weak continuity assumption and reduce the problem to adversarial contextual bandits with a finite class of "expert" policies, then apply EXP4 (Auer et al. 1995). The relevant expert class is not the class of all context-to-action maps: Corollary 1 implies that it suffices to compete with monotone maps, which is crucial to obtain the statistical rate $\widetilde{O}(\sqrt{T})$. However, the resulting method is *computationally* infeasible for large action sets: its time and space requirements are far larger than those of our method, of order exponential in K (recall that $K = 50$ in our applications). Additional advantages of our method include infrequent updates to the policy and clear certificates for all elimination decisions.

Consider the finite-action case $\mathcal{A} = \{1, \dots, K\}$ and suppose the expected penalty functions $\ell(x, a)$ are uniformly β-Hölder in the scalar ordered context $x \in [0, 1]$: for some $\beta \in (0, 1]$ and $\Lambda < \infty$, and for all $x, y \in [0, 1]$ and $a \in \mathcal{A}$,

$$|\ell(x, a) - \ell(y, a)| \le \Lambda |x - y|^{\beta}.$$

The Lipschitz case corresponds to $\beta = 1$, but the argument only needs $\beta > 0$. The contexts $x_1, \dots, x_T$ may be arbitrary, and the realized penalties are bounded in $[0, 1]$ with conditional means $\ell(x_t, a)$. Partition $[0, 1]$ into m equal intervals $\mathcal{I}_1, \dots, \mathcal{I}_m$. Let $\Pi_m^{\uparrow}$ be the class of monotone step policies on this grid: each $\pi \in \Pi_m^{\uparrow}$ assigns one action to each interval and these assigned actions are nondecreasing across intervals. The number of such policies is

$$\begin{aligned} |\Pi_m^{\uparrow}| = \binom{m + K - 1}{K - 1} &\le \left(\frac{e(m + K - 1)}{K - 1} \right)^{K-1} \\ &\le O(e(m + K)^K). \end{aligned} \tag{12}$$

Thus $\log |\Pi_m^{\uparrow}| = O(K \log(m + K))$; by contrast one gets $m \log K$ for the unrestricted class of non-monotone policies (which would eventually yield a much larger regret of order $T^{2/3}$ after a similar analysis).

The EXP4 algorithm maintains a weight $w_{\pi,t}$ for every $\pi \in \Pi_m^{\uparrow}$. After observing x_t, it forms the induced action distribution

$$\begin{aligned} q_t(a) &= \frac{\sum_{\pi \in \Pi_m^{\uparrow} : \pi(x_t) = a} w_{\pi,t}}{\sum_{\pi \in \Pi_m^{\uparrow}} w_{\pi,t}}, \\ p_t(a) &= (1 - \gamma) q_t(a) + \gamma / K, \end{aligned}$$

samples $a_t \sim p_t$, and observes the realized penalty $L_t(a_t)$. It then uses the importance-weighted estimate

$$\widehat{L}_t(a) = \frac{L_t(a_t) \mathbf{1}\{a = a_t\}}{p_t(a)}$$

and assigns expert π the estimated penalty $\widehat{L}_t(\pi(x_t))$. The weights are updated as $w_{\pi,t+1} = w_{\pi,t} \exp\{-\eta \widehat{L}_t(\pi(x_t))\}$. With the standard choices of η, γ, EXP4 guarantees

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} L_t(a_t) \right] - \min_{\pi \in \Pi_m^{\uparrow}} \mathbb{E}\left[\sum_{t=1}^{T} \ell(x_t, \pi(x_t)) \right] \\ &\le O\left(\sqrt{T K \log |\Pi_m^{\uparrow}|} \right) = O\left(K \sqrt{T \log(m + K)} \right). \end{aligned}$$

The discretization contributes a second error term. Fix an increasing optimal action selection $a^*(x) \in \arg\min_{a\in\mathcal{A}} \ell(x,a)$, whose existence follows from Corollary 1. For each interval $\mathcal{I}_j$, let c_j be a representative point and define the grid policy $\pi_m(x) = a^*(c_j)$ for $x \in \mathcal{I}_j$. This policy belongs to $\Pi_m^{\uparrow}$. Moreover, for every $x \in \mathcal{I}_j$, the Hölder condition gives $\ell(x,\pi_m(x)) - \min_{a\in\mathcal{A}} \ell(x,a) \le O(\Lambda m^{-\beta})$. Therefore

$$\min_{\pi\in\Pi_m^{\uparrow}} \sum_{t=1}^{T} \ell(x_t,\pi(x_t)) - \sum_{t=1}^{T} \min_{a\in\mathcal{A}} \ell(x_t,a) \le O(\Lambda T m^{-\beta}).$$

Combining the EXP4 and discretization terms gives $O\left(K\sqrt{T\log(m+K)} + \Lambda T m^{-\beta}\right)$. Taking $m = \lceil (1+\Lambda)^{1/\beta} T^{1/(2\beta)} \rceil$ makes the discretization term $O(\sqrt{T})$ and gives

$$O_\beta\left(K\sqrt{T\log M_{\Lambda,T}} + \sqrt{T}\right)$$

where $M_{\Lambda,T} = K + (1+\Lambda)^{1/\beta} T^{1/(2\beta)}$. Thus the dependence on the Hölder constant enters only logarithmically through the EXP4 expert-class term. If K, Λ, and β are treated as constants, this is $\widetilde{O}(\sqrt{T})$ regret against the context-by-context optimal action, even for adversarial context arrivals. The monotonicity reduction is what makes this possible under very weak continuity: the expert-class term depends on $\log m$, not on m.

This logarithmic dependence on the Lipschitz/Hölder constant is not an artifact of the proof, but is the optimal dependence. The following construction makes this point formal.

THEOREM 6. *Fix $\beta \in (0,1]$. For every $n \le T$ and every possibly randomized algorithm, there exists a two-action contextual bandit instance with a monotone optimal-action map and β-Hölder expected losses with constant at most $C_\beta 2^{\beta n}$, for a constant C_β depending only on β, such that an oblivious context adversary forces expected regret at least $n/2$. Consequently, any adversarial-context regret bound that is uniform over β-Hölder instances must have at least logarithmic dependence on the Hölder constant. If no continuity is imposed, the same construction gives $\Omega(T)$ regret.*

Proof of Theorem 6 Let $N = 2^n$, let the possible threshold locations be

$$\Theta_N = \left\{\theta_j = \frac{j+1/2}{N} : j = 0,\dots,N-1\right\}.$$

We put a uniform prior on Θ_N and study the expected (Bayes) regret. Set $\delta = 1/(4N)$. Define the clipped ramp

$$g_\delta(z) = \begin{cases} 0, & z \le -\delta, \\ (z+\delta)/(2\delta), & -\delta < z < \delta, \\ 1, & z \ge \delta. \end{cases}$$

For threshold θ, define deterministic losses $\ell_\theta(x,1) = g_\delta(x-\theta)$ and $\ell_\theta(x,2) = 1 - g_\delta(x-\theta)$. The losses lie in $[0,1]$. Moreover, g_δ is $1/(2\delta)$-Lipschitz and bounded by 1, so for any $r \in [0,1]$, we have $\min\{1, r/(2\delta)\} \le (r/(2\delta))^\beta$. Thus each $\ell_\theta(\cdot,a)$ is β-Hölder with constant at most $(2\delta)^{-\beta} = (2N)^\beta$. The optimal action is action 1 to the left of the threshold and action 2 to the right of the threshold, with arbitrary tie-breaking inside the transition region, so there is a monotone optimal-action selection.

We now describe an oblivious context adversary. After θ is drawn, but before the learning process begins, the adversary fixes the full sequence $x_1,\dots,x_n$ as the dyadic search path to θ. Specifically, set $I_0 = [0,1]$. For $t = 1,\dots,n$, let x_t be the midpoint of I_{t-1}, and let I_t be the left or right half of I_{t-1} that contains θ. The full context sequence is therefore fixed in advance as a deterministic function of the instance θ; it does not depend on the algorithm's realized actions or observations.

At the beginning of round t, after observing the context prefix $x_1, \ldots, x_t$ but before observing the loss at x_t, the posterior support is exactly the set of threshold points in I_{t-1}, and it is uniform on this set. The current query x_t bisects this support into two equal sub-blocks. Since all thresholds are cell midpoints and all query points are cell boundaries, every remaining $\theta \in \Theta_N$ satisfies $|x_t - \theta| \geq 1/(2N) > \delta$. Hence at x_t, the loss gap between the two actions is exactly 1: action 1 is optimal if $\theta > x_t$, and action 2 is optimal if $\theta < x_t$.

Condition on any history before the algorithm acts at round t. The posterior puts probability $1/2$ on each side of x_t. If the algorithm chooses action 1 with probability p_t, its conditional expected regret is $p_t \cdot \frac{1}{2} + (1 - p_t) \cdot \frac{1}{2} = \frac{1}{2}$. After the action is chosen and its loss is observed, the learner also learns which side of x_t contains θ: for either chosen action, the observed loss is 0 on one side and 1 on the other. The next context x_{t+1}, which was already fixed before the game, is the midpoint of that selected half-block. Therefore the induction continues. Summing over $t = 1, \ldots, n$, the Bayes regret under the uniform prior is $n/2$, so the expected regret is at least $n/2$ for some instance.

Since the Hölder constant in this construction is at most $C_\beta 2^{\beta n}$, taking $n = \left\lfloor \beta^{-1} \log_2(\Lambda / C_\beta) \right\rfloor \wedge T$ gives an $\Omega(\min\{T, \beta^{-1} \log \Lambda\})$ lower bound as a function of the continuity constant. Finally, if no continuity bound is imposed, take the limit $g_0(z) = \mathbf{1}\{z > 0\}$ and set $n = T$. The same binary-search argument then yields expected regret $T/2$, ruling out a uniform sublinear-regret guarantee against oblivious adversarial contexts without continuity. □

Appendix G: Proof of Lemma 1

Proof of Lemma 1 Let $n_s = |S_{\mathcal{E}}(x_s)|$. We note that under any policy, the quadruples (x_s, a_s, n_s, L_s) are i.i.d. (within a fixed epoch $\mathcal{E}$), and the policy used is constant within epoch $\mathcal{E}$. We will apply the multi-dimensional Dvoretzky—Kiefer—Wolfowitz—Massart inequality, see, e.g., Naaman (2021). This inequality states that the multi-dimensional cumulative distribution function $F(x, a, n, L) = \mathbb{P}[x_t \leq x, a_t \leq a, n_t \leq n, L_t \leq L]$ is well-approximated by the empirical estimate

$$\hat{F}(x, a, n, L) = \frac{1}{t - \tau_{\mathcal{E}} + 1} \sum_{s=\tau_{\mathcal{E}}}^{t} 1_{x_s \leq x, a_s \leq a, n_s \leq n, L_s \leq L}$$

in the sense that at time t, for any $\delta > 0$:

$$\begin{aligned} &\mathbb{P}[\sup_{x,a,n,\ell} |\hat{F}(x, a, n, \ell) - F(x, a, n, \ell)| \geq \delta] \\ &\leq 4(t - \tau_{\mathcal{E}} + 1) e^{-2(t - \tau_{\mathcal{E}} + 1)\delta^2} \\ &\leq 4T e^{-2(t - \tau_{\mathcal{E}} + 1)\delta^2}. \end{aligned} \tag{13}$$

(Note $t - \tau_{\mathcal{E}} + 1$ is the number of samples obtained by time t in epoch $\mathcal{E}$, i.e. the number of i.i.d. samples in the multi-dimensional DKWM inequality.) As in the inclusion-exclusion principle, a similar bound holds for any product of intervals $I^{(4)} = [x^{(1)}, x^{(2)}] \times [a^{(1)}, a^{(2)}] \times [n^{(1)}, n^{(2)}] \times [L^{(1)}, L^{(2)}]$.

Define the population rectangle probability $F_{\text{bdd}}(x^{(1)}, x^{(2)}, a^{(1)}, a^{(2)}, n^{(1)}, n^{(2)}, L^{(1)}, L^{(2)}) = \mathbb{P}\{(x_t, a_t, n_t, L_t) \in I^{(4)}\}$ and define its empirical counterpart by $\widehat{F}_{\text{bdd}}(x^{(1)}, x^{(2)}, a^{(1)}, a^{(2)}, n^{(1)}, n^{(2)}, L^{(1)}, L^{(2)}) = \frac{1}{t - \tau_{\mathcal{E}} + 1} \sum_{s=\tau_{\mathcal{E}}}^{t} \mathbf{1}\{(x_s, a_s, n_s, L_s) \in I^{(4)}\}$. Then with $\vec{i} = (i_x, i_a, i_n, i_L)$, the alternating sum of 16 terms

$$\sum_{\vec{i} \in \{1,2\}^4} (-1)^{i_x + i_a + i_n + i_L} F(x^{(i_x)}, a^{(i_a)}, n^{(i_n)}, L^{(i_L)})$$

gives the corresponding value for F_{bdd} corresponding to the half-open intervals $(x^{(1)}, x^{(2)}] \times (a^{(1)}, a^{(2)}] \times (n^{(1)}, n^{(2)}] \times (L^{(1)}, L^{(2)}]$, and similarly for $\hat{F}$. Applying the triangle inequality to (13) then shows the same bound for the product

$(x^{(1)}, x^{(2)}] \times (a^{(1)}, a^{(2)}] \times (n^{(1)}, n^{(2)}] \times (L^{(1)}, L^{(2)}]$, with error term $64(T+1)e^{-2(t-\tau_{\mathcal{E}}+1)\,\delta^2}$. If the context distribution is free of atoms, this finishes the proof. In general, since this bound is uniform in the interval endpoints, we get the same bound for $[x^{(1)}, x^{(2)}] \times [a^{(1)}, a^{(2)}] \times [n^{(1)}, n^{(2)}] \times [L^{(1)}, L^{(2)}]$. This is since for any probability measure μ on $\mathbb{R}^4$, $\lim_{x^1 \uparrow x^{(1)}, a^1 \uparrow a^{(1)}, n^1 \uparrow n^{(1)}, L^1 \uparrow L^{(1)}} \mu\Big((x^1, x^{(2)}] \times (a^1, a^{(2)}] \times (n^1, n^{(2)}] \times (L^1, L^{(2)}]\Big) = \mu\Big([x^{(1)}, x^{(2)}] \times [a^{(1)}, a^{(2)}] \times [n^{(1)}, n^{(2)}] \times [L^{(1)}, L^{(2)}]\Big)$.

We apply this bound for each $[x^{(1)}, x^{(2)}] = [A, B]$ by fixing $a^{(1)} = a^{(2)} = a_i$ for action a_i and using the loss interval $[L, 1]$. Define $I_{A,B,s,i,n,L}$ to be 1 if $\{x_s \in [A, B], a_s = a_i, n_s \geq n, L_s \in [L, 1]\}$, and 0 otherwise. The conclusion is that with probability $1 - O(Te^{-2(t-\tau_{\mathcal{E}}+1)\,\delta^2})$, simultaneously over $[A, B]$, n, and L:

$$\Big| \sum_{s=\tau_{\mathcal{E}}}^{t} I_{A,B,s,i,n,L} - \sum_{s=\tau_{\mathcal{E}}}^{t} \mathbb{E}[\mathbf{1^4}(s)] \Big| \leq \delta(t - \tau_{\mathcal{E}} + 1),$$

where $\mathbf{1^4}(s) := 1_{x_s \in [A,B]} \cdot 1_{a_s = a_i} \cdot 1_{n_s \geq n} \cdot 1_{L_s \in [L,1]}$. Integrating over $L \in [0, 1]$ replaces the indicator $1_{L_s \in [L,1]}$ by the value L_s in both terms, with the same error bound. Thus

$$\Big| \sum_{s=\tau_{\mathcal{E}}}^{t} I_{A,B,s,i,n}\, L_s - \sum_{s=\tau_{\mathcal{E}}}^{t} \mathbb{E}[\ell_i(x_s)\, \mathbf{1^3}(s)] \Big| \leq \delta(t - \tau_{\mathcal{E}} + 1), \tag{14}$$

where $I_{A,B,s,i,n} = 1_{\{x_s \in [A,B], a_s = a_i, n_s \geq n\}}$ and $\mathbf{1^3}(s) := 1_{x_s \in [A,B]} \cdot 1_{a_s = a_i} \cdot 1_{n_s \geq n}$.

Then summing (14) over $n = 1, 2, \ldots, K$ gives an upper bound of $K\delta(t - \tau_{\mathcal{E}} + 1)$ on the error between the true expectation and the importance-weighted estimate. Indeed, $\sum_{n=1}^{K} \sum_{s=\tau_{\mathcal{E}}}^{t} I_{A,B,s,i,n} \cdot L_s$ exactly gives the numerator of the importance-weighted estimate in (4), while $\sum_{n=1}^{K} \mathbb{E}[\ell_i(x_s) 1_{x_s \in [A,B]} \cdot 1_{a_s = a_i} \cdot 1_{n_s \geq n}]$ gives the expected true penalty on a_i from contexts in $[A, B]$. This is because the event $n_s = j$ is counted j times, balancing the $1/j$ probability to choose a_i. Taking $\delta = \frac{10\sqrt{\log(KT)}}{\sqrt{t - \tau_{\mathcal{E}} + 1}}$ finishes the proof. □

Appendix H: Proof of Lemma 3

Proof of Lemma 3 Fix $\mathcal{E}, i, t, A, B$, and write $m = t - \tau_{\mathcal{E}} + 1$. For simplicity, we suppress $\mathcal{E}, i, A, B$ and set

$$Z_s = \mathbf{1}\{x_s \in [A, B]\}\mathbf{1}\{a_s = a_i\} n_{\mathcal{E}}(x_s) L_s.$$

Since $L_s \in [0, 1]$ and $n_{\mathcal{E}}(x_s) \leq K$, we have $0 \leq Z_s \leq K$. Moreover,

$$\widehat{\ell}_{\mathcal{E},t,i}(A, B) = \frac{1}{m} \sum_{s=\tau_{\mathcal{E}}}^{t} Z_s = \frac{m-1}{m} \widehat{\ell}_{\mathcal{E},t-1,i}(A, B) + \frac{1}{m} Z_t.$$

Therefore $\widehat{\ell}_{\mathcal{E},t,i}(A, B) - \widehat{\ell}_{\mathcal{E},t-1,i}(A, B) = \frac{Z_t - \widehat{\ell}_{\mathcal{E},t-1,i}(A,B)}{m}$. Because $\widehat{\ell}_{\mathcal{E},t-1,i}(A, B)$ is an average of terms in $[0, K]$, it also lies in $[0, K]$. Hence $\Big|\widehat{\ell}_{\mathcal{E},t,i}(A, B) - \widehat{\ell}_{\mathcal{E},t-1,i}(A, B)\Big| \leq \frac{K}{m} = \frac{K}{t - \tau_{\mathcal{E}} + 1}$. This proves the claim. □

Appendix I: Proof of Lemma 4

Proof of Lemma 4 On the event of Lemma 1, the frontier interval I_j is in $F_{\mathcal{E},j}$, so $F_{\mathcal{E},i} \cap I_j$ is a valid interval to compare actions j and i. At any time t with $n = t - \tau_{\mathcal{E}} + 1$ satisfying $\sqrt{\frac{CK^2 \log(KT)}{n}} \leq \frac{\varepsilon}{12}$, Lemma 1 gives

$$\begin{aligned}
&\hat{\ell}_{\mathcal{E},t,i}(F_{\mathcal{E},i} \cap I_j) - \hat{\ell}_{\mathcal{E},t,j}(F_{\mathcal{E},i} \cap I_j) \\
&\geq \varepsilon - 2\sqrt{\frac{CK^2 \log(KT)}{n}} \\
&\geq 10\sqrt{\frac{CK^2 \log(KT)}{n}}.
\end{aligned}$$

Hence Equation (5) would trigger with action j as the better action and action i as the worse action. This proves $\tau_{\mathcal{E}+1} - \tau_{\mathcal{E}} \le \widetilde{O}(K^2/\varepsilon^2)$. Now let $t = \tau_{\mathcal{E}+1} - 1$ be the first elimination time, let $n = t - \tau_{\mathcal{E}} + 1$, and let (A', B', i', j') be the comparison that triggers. Since it triggers, Lemma 1 implies $\int_{A'}^{B'} \ell_{j'}(x) - \ell_{i'}(x)\,\mathrm{d}\mu(x) \ge 8\sqrt{\frac{CK^2\log(KT)}{n}}$. Conversely, the comparison on $F_{\mathcal{E},i} \cap I_j$ had not triggered at the preceding time. Using Lemma 1 at that time gives

$$\varepsilon \le 12\sqrt{\frac{CK^2\log(KT)}{n-1}} \le 17\sqrt{\frac{CK^2\log(KT)}{n}},$$

with the first-round case handled by increasing C. Combining the last two displays proves the elimination condition in Lemma 4. In particular, the eliminated action is truly worse on the eliminated interval.

Finally, for the midpoint M', Equation (6) and Lemma 1 imply

$$\left|\int_{A'}^{M'} \ell_{j'}(x) - \ell_{i'}(x)\,\mathrm{d}\mu(x) - \int_{M'}^{B'} \ell_{j'}(x) - \ell_{i'}(x)\,\mathrm{d}\mu(x)\right| \le \frac{5K}{n} + 4\sqrt{\frac{CK^2\log(KT)}{n}} \le \frac{9}{2}\sqrt{\frac{CK^2\log(KT)}{n}},$$

where the last inequality follows by taking C large enough. Since the sum of the two midpoint integrals is at least $8\sqrt{CK^2\log(KT)/n}$, each integral is at least $\varepsilon/10$. This proves Equation (8). □

Appendix J: Proof of Lemma 5

Proof of Lemma 5 Assuming $\sum_{i=1}^{K} R_{\mathcal{E},i} \ge \varepsilon$, there must exist $i \in [K]$ such that $R_{\mathcal{E},i} \ge \varepsilon/K$. Therefore by Lemma 2, there exists $j \in [K]$ and an interval $F_{\mathcal{E},i} \cap I_j$ such that

$$\int_{x \in F_{\mathcal{E},i} \cap I_j} \ell_i(x) - \ell_j(x)\,\mathrm{d}\mu(x) \ge \varepsilon/K^2.$$

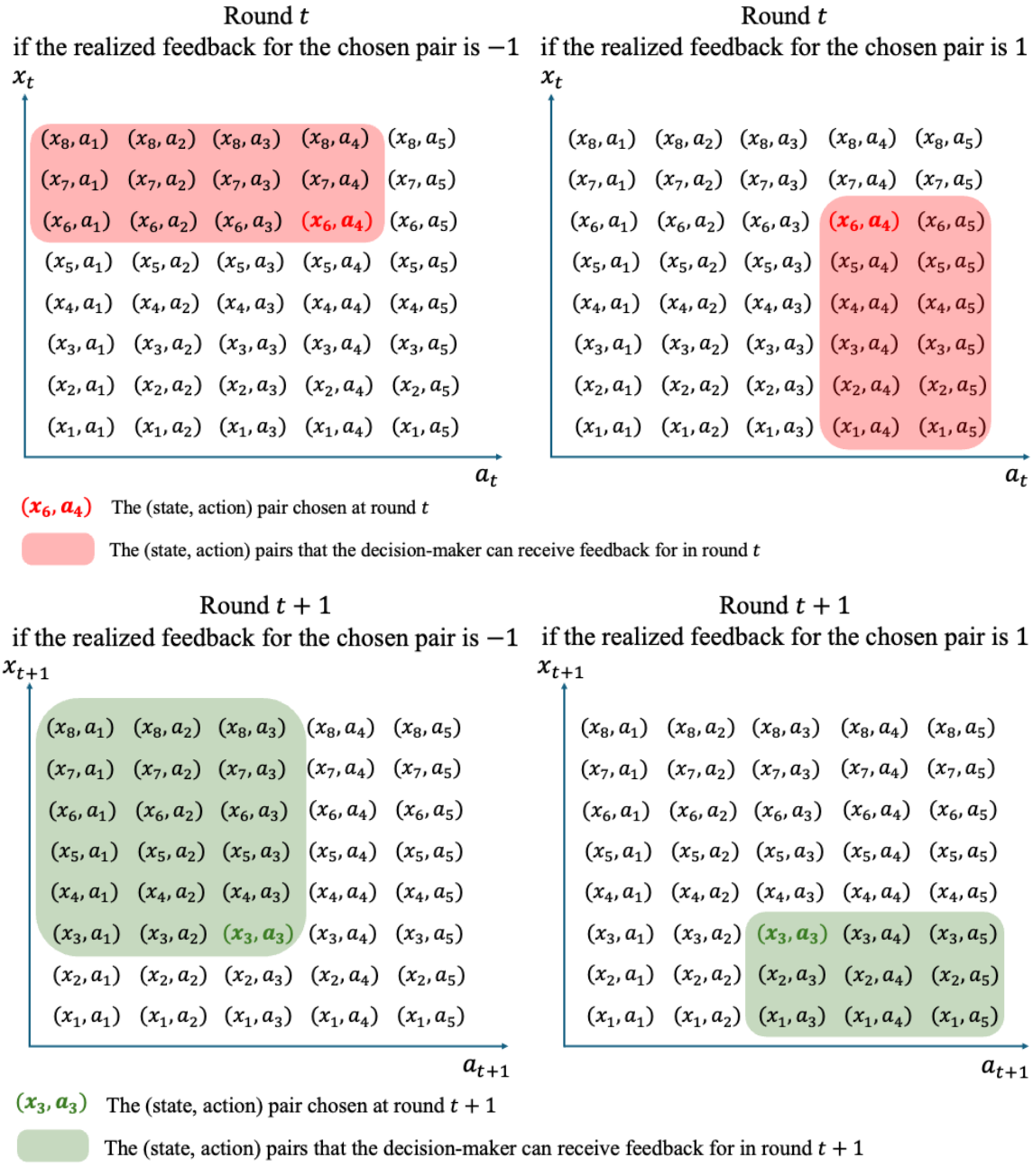
The claims now follow by applying Lemma 4 with ε replaced by ε/K^2; note that $K^2/(\varepsilon/K^2)^2 = K^6/\varepsilon^2$. It follows from Lemma 4 and Equation (8) (via Equation (6)) that the next elimination phase removes $\Omega(\varepsilon/K^2)$ of the total active suboptimality (as defined in Equation (7)). With $\varepsilon = \sum_{i=1}^{K} R_{\mathcal{E},i}$ chosen so equality holds, this implies the desired result (the equality here is without loss of generality). The final assertion follows from (8) and Assumption 2. □

Appendix K: Conditional One-Sided Feedback

Definition 1 *An instance of the contextual packaging problem has* ***conditional one-sided feedback*** *with respect to the ordering functions ψ and ϕ if, when any $(\mathbf{x}_t, \mathbf{a}_t)$ is played, the realized failure outcome determines one of two feedback orthants. If a failure occurs, then all failure outcomes for pairs $(\mathbf{x}, \mathbf{a})$ with $\psi(\mathbf{x}) \ge \psi(\mathbf{x}_t)$ and $\phi(\mathbf{a}) \le \phi(\mathbf{a}_t)$ are observed in that round t; these are harder contexts and weaker packaging actions. If no failure occurs, then all nonfailure outcomes for pairs $(\mathbf{x}, \mathbf{a})$ with $\psi(\mathbf{x}) \le \psi(\mathbf{x}_t)$ and $\phi(\mathbf{a}) \ge \phi(\mathbf{a}_t)$ are observed.*

Toy Counterexample. Suppose in a simple bandit-learning problem, there are three arms to choose from repeatedly over a horizon of $T = 1000$ rounds. The penalty received for pulling an arm at round t is either 0 or 1. The true average penalties of arms $1, 2, 3$ are $0.49, 0.5$, and 0.51, respectively. We set up the problem so that there exists conditional one-sided feedback in this problem: penalties occur when ξ_t is respectively at most $0.49, 0.5, 0.51$. In particular, if at round t, Arm 2 is pulled and gives penalty 0, then if we had pulled Arm 1 at round t, we would have also received penalty 0; if at round t, Arm 2 is pulled and gives penalty 1, then if we had pulled Arm 3 at round t, we

Figure 8 **Example of Conditional One-Sided Feedback in two rounds t (red) and $t+1$ (green).** The (state, action)-pairs chosen by the decision-maker are bolded and colored. In the left diagrams, the highlighted regions indicate which (state, action)-pairs the decision-maker receives feedback for when the realized feedback is -1. The right diagrams indicate those pairs the decision-maker receives feedback for when the realized feedback is 1. For ease of presentation, in this example, the contexts and the actions are already indexed by their orderings.



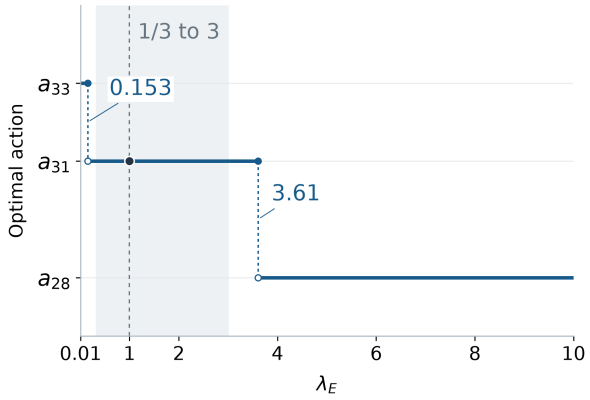
would have also received penalty 1. In other words, in this example, Arm 1 is stochastically dominated by Arm 2; Arm 2 is stochastically dominated by Arm 3. This is a simple toy example of conditional one-sided feedback.

Now suppose an algorithm uses this extra feedback to update the penalty estimates for the three arms, and pulls Arm 2 for all $T = 1000$ rounds. Then our penalty estimate for Arm 2 would converge correctly to approximately 0.5, while our penalty estimates with high confidence for Arms 1 and 3 would converge to 0 and 1 respectively, which would be far from the true expected penalties. This is because we only update our estimate for Arm 1 when the outcome for Arm 2 is 0, and only update for Arm 3 when the outcome for Arm 2 is 1. By conducting updates for different arms conditioned on the realized outcomes, an algorithm can introduce severe biases into the estimates for the arms.

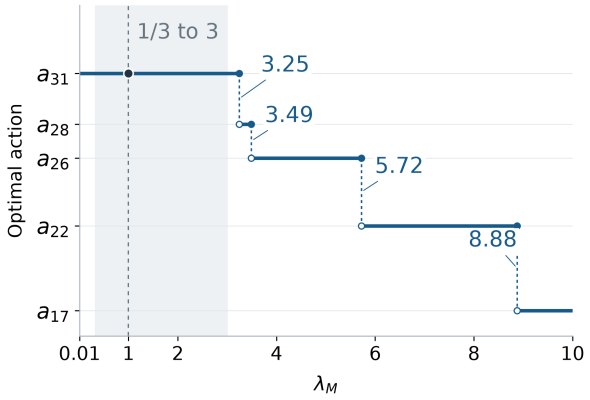
Appendix L: Sensitivity Analysis Details

Figure 9 **Sensitivity of the optimal packaging action for an illustrative beef-containing shipment described in this appendix. Each plot varies one coefficient while holding the other two constant at 1. Actions $a_0, \ldots, a_{49}$ are indexed by increasing thermal insulation strength, with a fixed ordering for ties; vertical axis ranges differ across panels. In panels (a)–(c), shading marks multipliers from $1/3$ to 3, and the black point on the dashed vertical line marks the baseline optimal action. Panel (d) enlarges the leftmost part of panel (c).**

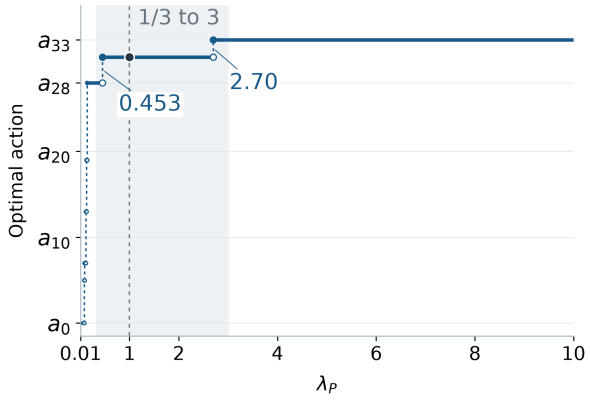
(a) Sensitivity to the Environmental Coefficient λ_E.



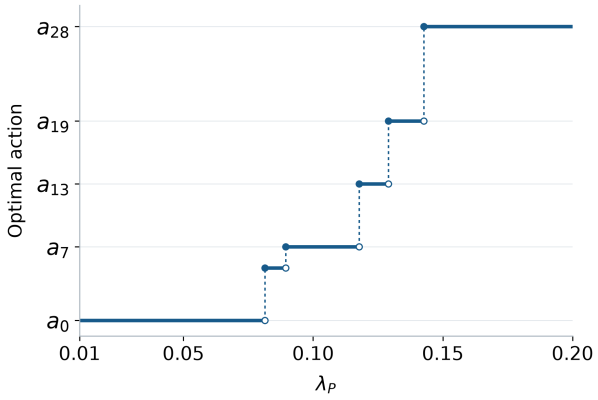
(b) Sensitivity to the Material-Cost Coefficient λ_M.



(c) Sensitivity to the Failure-Penalty Coefficient λ_P.



(d) Sensitivity to the Failure-Penalty Coefficient: Expanded View for Small λ_P Values.



To examine how the relative weights of the penalty components affect packaging decisions, we conduct a numerical sensitivity analysis. Figure 9 shows an illustrative example of the sensitivity of the optimal packaging action to changes in the relative weights assigned to environmental impact E, material cost M, and expected failure penalty $\overline{P}$, respectively. The context in Figure 9 encodes a forecast maximum temperature of 75°F, 68 hours in transit, and no active refrigeration. The shipment uses a $16\times16\times16$-inch box containing 1 kg of raw minced beef, 2 kg of vegetables, and 1 kg of grains and sauces, with FedEx Home Delivery as the ground-service scenario and an assumed road distance of 3,500 km.

Panels 9a–9c vary one coefficient from 0.01 to 10 times baseline while holding the others fixed; panel 9d shows the failure coefficient varying over 0.01–0.20, enlarging the leftmost part of panel (c). In panels (a)–(c), shading marks multipliers from $1/3$ to 3, and the black point on the dashed vertical line marks the baseline optimal action.

The packaging menu combines five liner options with ten configurations of up to three freezer packs, each weighing 3 or 5 lb. We evaluate all 50 packaging actions, indexed in increasing order of approximate thermal-insulation strength as $a_0, a_1, \ldots, a_{49}$, where a_0 denotes no liner material and no freezer packs. The ordering does not imply that consecutive actions become optimal in sequence. For example, a_{31} uses a light liner with three 3-lb packs, a_{32} uses a medium liner with one 3-lb and one 5-lb pack, and a_{33} uses a light liner with two 5-lb packs. At the selected context, the sum of environmental and material penalties of a_{32} at baseline coefficients is slightly higher than that of a_{33}, although a_{32} provides less thermal protection. Consequently, when the failure coefficient reaches approximately 2.70 times baseline, the optimum switches directly from a_{31} to a_{33}, skipping a_{32}. More generally, all 50 configurations are evaluated, but only those that minimize the weighted objective for some coefficient value appear on the optimal-action curves. Skipped configurations are generally less favorable in their environmental and material consequences (given the relative coefficients) among the options that provide a given level of thermal insulation.

As λ_P varies, a crossover between actions a and b with distinct failure penalties must occur at

$$\lambda_P^* = \frac{\lambda_E[E(b)-E(a)] + \lambda_M[M(b)-M(a)]}{\overline{P}(a)-\overline{P}(b)}. \tag{15}$$

For positive coefficients, every optimal action is Pareto optimal with respect to $(E, M, \overline{P})$.

Appendix M: (Unconditional) One-Sided Feedback

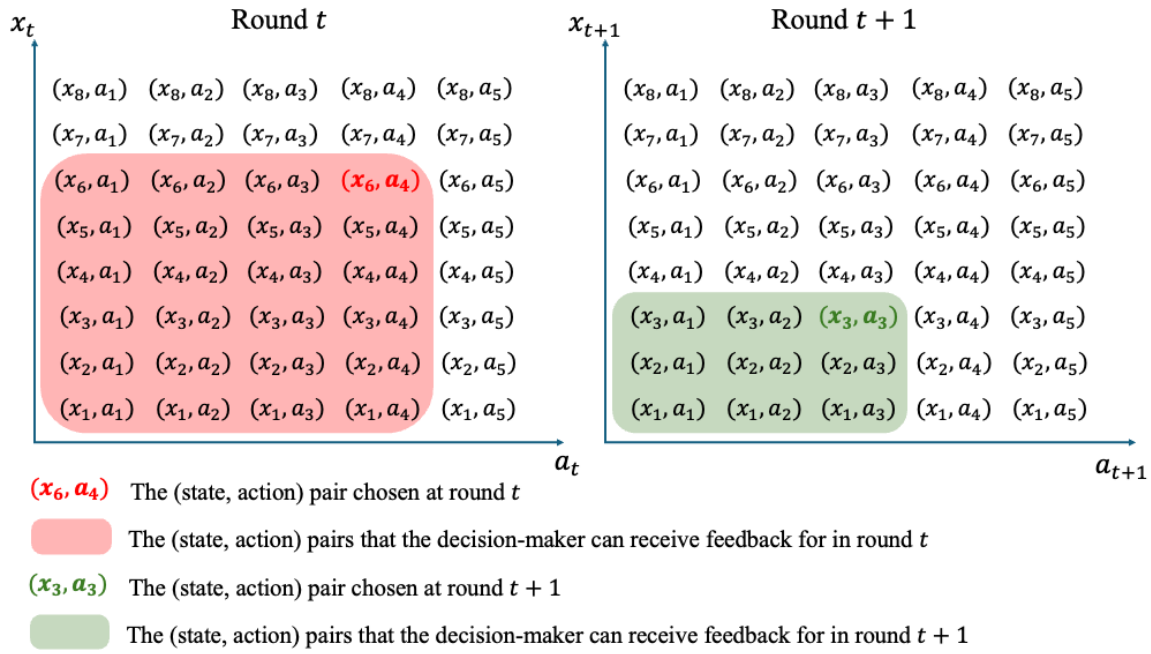
We provide a definition of (unconditional) one-sided feedback, adapted to the contextual bandit setting based on Gong and Simchi-Levi (2024). We also include in this Appendix an $\widetilde{O}(\sqrt{T})$ regret algorithm for this setting, which exploits the random arrival order of the contexts.

Definition 2 *A problem has* one-sided feedback *if there exist ordering functions ψ and ϕ such that, when any $(\mathbf{x}_t, \mathbf{a}_t)$ is played, all rewards (penalties) for pairs $(\mathbf{x}, \mathbf{a})$ with $\psi(\mathbf{x}) \le \psi(\mathbf{x}_t)$ and $\phi(\mathbf{a}) \le \phi(\mathbf{a}_t)$ are observed in that round t.*

The idea of Algorithm 2 is similar to *Elimination-Based Half Q-Learning* from Gong and Simchi-Levi (2024), but crucially exploits the fact that contexts are i.i.d., hence in random order. We define the confidence bound $\mathrm{CB}_1(n) := 100\sqrt{\frac{\log(T)}{n+1}}$.

THEOREM 7. *For one-sided feedback with i.i.d. contexts, the regret of Algorithm 2 is $\widetilde{O}(\sqrt{KT})$.*

Figure 10 **An example of One-Sided Feedback in two rounds t and $t+1$. For ease of presentation, the contexts and the actions are already indexed by their orderings.**



Proof As explained in the pseudocode, Algorithm 2 chooses the largest action in the running set at each time. The running set $\mathcal{R}(x_t)$ consists of all actions that have not yet been eliminated as possibly optimal, based on online confidence intervals. The idea is that at most times, there will be substantial previous relevant feedback because the stochastic contexts arrive in random order. First we inductively establish validity of the running sets. We say the running set $\mathcal{R}(x_t)$ is *valid* if it contains the frontier action for x_t.

By a standard Chernoff bound for each fixed t, with probability $1-\frac{1}{T^2}$, we have

$$|f(a)-\hat{f}(a)| \le CB_1(n_t) \le \widetilde{O}(1/\sqrt{n_t+1}) \tag{16}$$

for all $a \le a^*_{x_t}$ (i.e. all $a \in \mathcal{R}(x_t)$), on the event that $a^*_{x_t} \le a^*_{x_s}$ holds for all $s \le t$ with $x_s \ge x_t$. This in fact follows by the definition of u_{x_t}. Moreover, it follows by the triangle inequality that all running sets are indeed valid with probability $1-\frac{1}{T}$. Hence on this high-probability event, the total regret is at most $\sum_{t=1}^{T}\widetilde{O}(1/\sqrt{n_t+1})$. Next, observe that for each fixed t, the value n_t will be a uniformly random integer between 0 and $t-1$ inclusive, thanks to the i.i.d. arrivals. (Unless there are repeated contexts, which only makes n_t larger. Even in this case, n_t is still uniform if random tie-breaking is used for context order, as in the analysis of Algorithm 1.) Hence the expected instantaneous regret at time t is $\frac{\sqrt{K}}{t}\sum_{s=1}^{t}\widetilde{O}(s^{-1/2}) = \widetilde{O}\left(\sqrt{K/t}\right)$. Summing over t, the overall expected regret is $\widetilde{O}(\sqrt{KT})$. □

Algorithm 2 Elimination-Based Contextual Bandit Algorithm

for time $t = 1, 2, \ldots$ **do**
  Receive context x_t.
  Let $\mathcal{T}(x_t) := \{s < t : x_s \geq x_t\}$. Let $C_t := \{x_s : x_s \geq x_t, s < t\}$. Let $n_t := |C_t|$.
  Let u_{x_t} be the minimum of the upper endpoints of the running sets $\mathcal{R}(x_s)$, over $x_s \in C_t$:

$$u_{x_t} = \min_{x_s \in C_t} \sup \mathcal{R}(x_s).$$

  if the running set $\mathcal{R}(x_t)$ for context x_t does not already exist **then**
    Initiate $\mathcal{R}(x_t) := \mathcal{A} \cap [0, u_{x_t}]$.
  else
    Update $\mathcal{R}(x_t) \leftarrow \mathcal{A} \cap [0, u_{x_t}]$.
  end if
  for action $a \in \mathcal{R}(x_t)$ **do**
    Compute the empirical mean estimate $\hat{f}(a)$ with the n_t data points from $\mathcal{T}(x_t)$.
  end for
  Let $a^*_{x_t} \leftarrow \arg\min_{a \in \mathcal{R}(x_t)} \hat{f}(a)$.
  Update $\mathcal{R}(x_t) \leftarrow \{a \in \mathcal{R}(x_t) : \left|\hat{f}(a^*_{x_t}) - \hat{f}(a)\right| \leq 2\,\mathrm{CB}_1(n_t)\}$.
  Take action $a_t \leftarrow \max \mathcal{R}(x_t)$.
end for