CONSERVATIVE DUELING BANDITS: PREFERENCE LEARNING UNDER BASELINE CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contextual dueling bandits learn from pairwise comparisons, as in fine-tuning language models from human preferences. Standard algorithms explore aggressively, however, and may show users two poor options. We require that, at all times, the cumulative score of the better displayed item stay above a $(1 - \alpha)$ fraction of a trusted baseline's. An item's score is its probability of beating the best available item, so the constraint needs no absolute rewards and is invariant to utility shifts. We show that directly transferring the safety test of the conservative bandit algorithm CLUCB can incur linear regret. For linear utilities, `MaxInP-BC` certifies safety through utility gaps and falls back to comparisons that pair the baseline with an informative partner. It is safe with high probability at all times, with regret $\tilde{\mathcal{O}}(d\sqrt{T}/\kappa)$ plus a horizon-independent cost of conservatism. For general utility classes, `BT-IGW-BC` builds safety certificates from an online logistic-regression oracle, achieving anytime safety, sublinear regret, and finitely many fallbacks. Both algorithms need only bounds on baseline scores. In experiments, both satisfy the constraint, which their unconstrained counterparts often violate, at a moderate cost in regret.

1 INTRODUCTION

Many learning systems that interact with people cannot rely on numerical rewards. When users are asked to rate a search result, a recommended item, or a piece of generated text on an absolute scale, the scores they give are noisy, poorly calibrated across individuals, and expensive to collect at scale. Asking the same users which of two options they prefer is easier for them and yields a more consistent signal. This has motivated a large body of work on learning from pairwise preferences (Bradley & Terry, 1952; Fürnkranz & Hüllermeier, 2010; Yue et al., 2012a; Wirth et al., 2017; Bengs et al., 2021). Its applications include online evaluation of search rankers (Joachims, 2002; Radlinski et al., 2008; Yue & Joachims, 2009; Chapelle et al., 2012; Hofmann et al., 2016), recommender systems (Jones et al., 2011; Kalloori et al., 2016), human-in-the-loop design (Brochu et al., 2007; González et al., 2017; Tucker et al., 2020), and reinforcement learning from human comparisons (Christiano et al., 2017), which has more recently become a standard way to fine-tune language models (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023).

The dueling bandit problem gives a clean abstraction of this interaction and has been studied for finitely many actions (Japkowicz & Shah, 2011; Yue et al., 2012b). In each round, the learner selects a pair of actions, presents them to a user, and observes only a noisy binary outcome indicating which one was preferred. Because no single observation reveals how good an action is, the learner must infer quality from patterns of wins and losses. A common way to add structure is to posit a probabilistic choice model. Under the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Hunter, 2004; Luce, 2005), each action has a latent utility, and the probability that one action beats another is a logistic function of their utility difference (cf. Definition 1). The Plackett–Luce model extends this to comparisons among more than two items (Zhao et al., 2020). Contextual dueling bandits (Saha, 2021) let the utility also depend on observed side information, such as a query, a user profile, or a prompt. They have led to algorithms with sublinear regret guarantees for linear and more general utility classes.

However, all of this work measures performance by regret against the best comparison in hindsight. The resulting algorithms therefore explore as aggressively as they can. This sits uneasily with how preference-based learners are deployed in practice, because every comparison is shown to a real

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

person. A user who is repeatedly offered two unappealing recommendations, or two unhelpful answers, has little reason to keep trusting the system, however informative those comparisons are for the learner. Moreover, a new learning system rarely starts from scratch. It usually replaces or augments an existing policy whose behavior is considered acceptable. In such settings, it is natural to require that algorithms avoid aggressive exploration, and ensure that the deployed policy performs at least comparably to an existing, trusted *baseline strategy*. Concretely, we ask that at every step, the better of the two displayed items be, in aggregate, at least a $(1-\alpha)$ fraction as good as what the baseline would have offered.

For bandits with numerical rewards, this kind of safe exploration is captured by conservative bandits (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020). These methods track a budget of permissible loss relative to the baseline and explore only when that budget allows. However, carrying this idea over to preference feedback is not straightforward. First, the learner never observes rewards, so it cannot track its own performance against the baseline by summing observed payoffs; it must instead reason about the scores of the items it displays through what it has learned about their utilities. Second, under the BTL model only utility differences affect observed preferences, so we state the constraint through preference scores, which makes it invariant to shifts of the utility function. Third, the learner must certify at every round that the constraint continues to hold, using only binary comparison outcomes. As we show, directly transferring the conservative linear bandit algorithm CLUCB to this setting can incur linear regret.

In this work, we study *Preference Learning under Baseline Constraints*. Our technical contributions are as summarized below.

1. **Framework:** We formulate conservative contextual dueling bandits under Bradley–Terry feedback, requiring the cumulative preference score of the better displayed item to remain above a $(1-\alpha)$ fraction of the baseline's at every prefix. This score-based constraint is invariant to utility shifts and requires no absolute reward observations.
2. **Linear utilities:** We show that a direct CLUCB transfer can incur linear regret and develop `MaxInP-BC`, which certifies scores through utility gaps and uses informative baseline fallbacks. We prove anytime high-probability safety and regret $\tilde{\mathcal{O}}\big(d\sqrt{T}/\kappa + \Delta_h d^2/(\kappa^2\alpha^2\underline{s}^2)\big)$, where $\underline{s} = \sigma(-\Delta_h)$. The additive conservatism term is independent of the horizon, and known score bounds suffice in place of exact baseline scores.
3. **General utility functions:** We develop C-BT-IGW, which converts online logistic-regression guarantees into cumulative score certificates and updates the oracle on both candidate and fallback comparisons. For realizable classes with oracle regret $\mathcal{R}_{\log}(T) = o(T)$, we prove anytime high-probability safety and regret $\tilde{\mathcal{O}}\big(\sqrt{KT[\mathcal{R}_{\log}(T)/\kappa + 1/\kappa^2]} + \Delta_h K[\mathcal{R}_{\log}(T)/\kappa + 1/\kappa^2]/(\alpha^2\underline{s}^2)\big)$.
4. **Experiments:** We evaluate both algorithms empirically, measuring their regret together with how reliably they satisfy the baseline constraint.

2 Problem Formulation

We consider a stochastic *Contextual Dueling Bandit* (CDB) problem. At each round $t \in [T]$, the learner is presented with a context set $\mathcal{S}_t = \{\mathbf{x}_{t,1}, \dots, \mathbf{x}_{t,K}\} \subset \mathcal{D}$ of K items (or arms), where each $\mathbf{x}_{t,i}$ is drawn i.i.d. from an unknown distribution $\mathcal{P}_{\mathcal{D}}$ over a decision space $\mathcal{D} \subset \mathbb{R}^d$. The agent selects an *ordered* pair $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{S}_t \times \mathcal{S}_t$ (repetitions are allowed), and receives noisy binary feedback $o_t \in \{0, 1\}$ indicating which item is preferred. In the context of LLMs, $(\mathbf{x}_t, \mathbf{y}_t)$ correspond to the two responses shown to the user, and o_t to the user's recorded preference. Let $\mathcal{H}_t^-$ denote all information available immediately before observing o_t, including all completed rounds, the current contexts and baseline information, and the selected pair. As in Saha (2021); Bengs et al. (2022); Li et al. (2024), we assume preferences are generated by a latent utility function $f : \mathcal{D} \to \mathbb{R}$, through the Bradley–Terry–Luce (BTL) model.

Assumption 1 (BTL Model) *There exists a utility function $f : \mathcal{D} \to \mathbb{R}$ such that the conditional preference probability satisfies*

$$\mathbb{P}(o_t = 1 \mid \mathcal{H}_t^-) = \frac{\exp(f(\mathbf{x}_t))}{\exp(f(\mathbf{x}_t)) + \exp(f(\mathbf{y}_t))} = \sigma\big(f(\mathbf{x}_t) - f(\mathbf{y}_t)\big), \tag{1}$$

where $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid function.

Two structural properties of (1) will be useful. The comparison law depends only on utility differences and is invariant to adding a constant to every utility. A self-duel has success probability $1/2$, so its outcome provides no information about f.

Assumption 2 (Bounded utility range) *There is a known constant $B > 0$ such that the absolute difference $|f(\mathbf{x}) - f(\mathbf{y})| \le B$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. In the linear case $f(\mathbf{x}) = \langle \theta^\star, \mathbf{x} \rangle$ with $\|\theta^\star\|_2 \le S$ and $\|\mathbf{x}\|_2 \le L$, one may take $B = 2LS$. We define the minimum slope of the logistic link on $[-B, B]$, $\kappa := \inf_{|z| \le B} \sigma'(z) = \sigma(B)\big(1 - \sigma(B)\big) > 0$.*

Let $\mathbf{x}_t^\star \in \arg\max_{\mathbf{x} \in \mathcal{S}_t} f(\mathbf{x})$ denote a best item in round t, ties broken by a fixed rule. Next we define the regret of an algorithm following Saha (2021); Verma et al. (2025). It captures the cumulative gap between the best possible utility and the average utility of the learner's choices over T rounds.

Regret. Suppose an algorithm chooses the pair $(\mathbf{x}_t, \mathbf{y}_t)$ at time $t \in [T]$. The instantaneous regret incurred at round t is the gap between the utility of the best item and the average utility of the compared pair, denoted $R_t := f(\mathbf{x}_t^\star) - \frac{1}{2}\big(f(\mathbf{x}_t) + f(\mathbf{y}_t)\big)$, and the cumulative T-round regret is

$$\mathtt{Reg}(T) := \sum_{t=1}^{T} R_t = \sum_{t=1}^{T} \big(f(\mathbf{x}_t^\star) - \tfrac{1}{2}\big(f(\mathbf{x}_t) + f(\mathbf{y}_t)\big)\big)\,. \tag{2}$$

Next we define a performance constraint that the algorithm must satisfy so that the overall user experience stays above a threshold set by an existing trusted baseline policy.

Definition 2.1 (Baseline Performance Constraint) *The* preference score *of an item $\mathbf{x} \in \mathcal{S}_t$ is its probability of beating the best item of the round: $s_t(\mathbf{x}) := \mathbb{P}\{\mathbf{x} \succ \mathbf{x}_t^\star\} = \sigma\big(f(\mathbf{x}) - f(\mathbf{x}_t^\star)\big) \in (0, \frac{1}{2}]$. Suppose the agent has access to baseline items $\mathbf{x}_{b_t} \in \mathcal{S}_t$, $t \in [T]$, with known scores $s_t(\mathbf{x}_{b_t})$. Given a risk parameter $\alpha \in (0, 1)$, a sequence of displayed pairs $(\mathbf{x}_i, \mathbf{y}_i), i \in [t]$ is* α-conservative *if*

$$\sum_{i=1}^{t} \max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} \ge (1 - \alpha) \sum_{i=1}^{t} s_i(\mathbf{x}_{b_i}), \quad \textit{for all } t \in [T]. \tag{3}$$

Assumption 3 (Baseline Gap and Bounds) *Let $\Delta_{t,b_t} := f(\mathbf{x}_t^\star) - f(\mathbf{x}_{b_t})$ denote the baseline gap. There are known constants $0 \le \Delta_l \le \Delta_h \le B$ with $\Delta_l \le \Delta_{t,b_t} \le \Delta_h$ for all $t \in [T]$; equivalently*

$$\underline{s} := \sigma(-\Delta_h) \;\le\; s_t(\mathbf{x}_{b_t}) \;\le\; \sigma(-\Delta_l) =: \bar{s} \;\le\; \tfrac{1}{2}.$$

Remark 2.1 (Why the maximum and why scores) The constraint in (3) compares cumulative preference probabilities along the learner's trajectory, rather than realized comparison labels. Using the maximum ensures that any pair containing the baseline has score at least that of the baseline, allowing its partner to be selected for information acquisition.

3 Linear Utilities: Conservative Maximum-Informative-Pair

Assumption 4 (Linear Utility) *There is an unknown $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \le S$ such that $f(\mathbf{x}) = \langle \theta^\star, \mathbf{x} \rangle$, and $\|\mathbf{x}\|_2 \le L$, for every $\mathbf{x} \in \mathcal{D}$. The learner optimises over $\Theta = \{\theta : \|\theta\|_2 \le S\}$, so Assumption 2 holds with $B = 2LS$ and $\kappa = \sigma(B)(1 - \sigma(B))$.*

In the unconstrained setting, MaxInP (Saha, 2021) maintains a set of items that may still be optimal and displays the most uncertain pair among them. Such a pair is informative, but nothing prevents both of its items from being poor. We therefore design `MaxInP-BC` around a safety budget. The algorithm fits $\theta^\star$ by regularized maximum likelihood on the observed comparisons and uses the resulting confidence ellipsoid to identify promising items. It then proposes the most informative pair and displays it only if the baseline constraint remains certifiably satisfied. Otherwise, it falls back to a pair that contains the baseline item and is still informative for learning.

Estimator. Each round the learner is given $\mathcal{S}_t$, the baseline item $\mathbf{x}_{b_t} \in \mathcal{S}_t$ and its score $s_t(\mathbf{x}_{b_t})$. The learner displays $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{S}_t \times \mathcal{S}_t$, and observes o_t. Write $\mathbf{z}_t := \mathbf{x}_t - \mathbf{y}_t$ for the displayed comparison vector, so $\|\mathbf{z}_t\|_2 \le 2L$. We define

$$V_t = \lambda I_d + \sum_{i<t} \mathbf{z}_i \mathbf{z}_i^\top, \qquad \hat{\theta}_t \in \arg\min_{\theta \in \Theta} \Big\{ \sum_{i<t} \big[\log(1 + e^{\mathbf{z}_i^\top \theta}) - o_i \mathbf{z}_i^\top \theta\big] + \tfrac{\lambda}{2}\|\theta\|_2^2 \Big\}, \tag{4}$$

with $\lambda > 0$; the objective is strongly convex, so $\hat{\theta}_t$ is unique. A self-duel has $\mathbf{z}_t = 0$ and contributes the constant $\log 2$ to the objective, leaving the design matrix and the minimizer unchanged. For $\delta \in (0,1)$, set

$$\beta_t = \frac{1}{\kappa}\Big[\tfrac{1}{2}\sqrt{d\log\big(1+\tfrac{4L^2(t-1)}{\lambda d}\big)+2\log\tfrac{1}{\delta}}+\sqrt{\lambda}S\Big]. \tag{5}$$

Lemma 3.1 (Anytime confidence ellipsoid) *Under Assumptions 1 and 4, $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ for the event $\mathcal{E} := \{\|\hat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t, \ \ \forall t \geq 1\}$.*

Preference residuals are conditionally $\frac{1}{2}$-sub-Gaussian; a mixture supermartingale and Ville's inequality give an anytime self-normalised bound, which first-order optimality and the curvature bound $\nabla^2 \succeq \kappa V_t$ convert into parameter concentration (Appendix A.3).

Promising items and the regret certificate. Let

$$\mathcal{C}_t = \big\{\mathbf{x} \in \mathcal{S}_t : \langle \hat{\theta}_t, \mathbf{x}-\mathbf{y}\rangle + \beta_t\|\mathbf{x}-\mathbf{y}\|_{V_t^{-1}} \geq 0, \ \ \forall \mathbf{y} \in \mathcal{S}_t\big\}, \tag{6}$$

and let the *candidate pair* be the maximum-width pair of promising items,

$$(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t) \in \operatorname*{arg\,max}_{\mathbf{x},\mathbf{y}\in\mathcal{C}_t} \|\mathbf{x}-\mathbf{y}\|_{V_t^{-1}}, \qquad w_t = \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{y}}_t\|_{V_t^{-1}}, \qquad c_t = \min\{B, 2\beta_t w_t\}. \tag{7}$$

Lemma 3.2 (Regret certificate) *$\mathcal{C}_t \neq \emptyset$ for all t and on $\mathcal{E}$, $\mathbf{x}_t^\star \in \mathcal{C}_t$ and $f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq c_t$ for every $\mathbf{y} \in \mathcal{C}_t$. In particular, the candidate pair has regret*

$$R_t^{\mathrm{cand}} := f(\mathbf{x}_t^\star) - \tfrac{1}{2}\big(f(\tilde{\mathbf{x}}_t) + f(\tilde{\mathbf{y}}_t)\big) \leq c_t.$$

We next convert this utility-gap certificate into a lower bound on the score performance in (3).

Lemma 3.3 (Score certificate) *If a pair $(\mathbf{x},\mathbf{y})$ satisfies $f(\mathbf{x}_t^\star) - \frac{1}{2}(f(\mathbf{x}) + f(\mathbf{y})) \leq c$ for some $c \geq 0$, then $\max\{s_t(\mathbf{x}), s_t(\mathbf{y})\} \geq \sigma(-c) \geq \frac{1}{2} - \frac{c}{4}$.*

Indeed, with $g_{\mathbf{x}} := f(\mathbf{x}_t^\star) - f(\mathbf{x}) \geq 0$ and $g_{\mathbf{y}}$ likewise, $\max\{s_t(\mathbf{x}), s_t(\mathbf{y})\} = \sigma(-\min\{g_{\mathbf{x}}, g_{\mathbf{y}}\})$ and $\min\{g_{\mathbf{x}}, g_{\mathbf{y}}\} \leq \frac{1}{2}(g_{\mathbf{x}} + g_{\mathbf{y}}) \leq c$; the second inequality is $\sigma' \leq \frac{1}{4}$. The anchor is $s_t(\mathbf{x}_t^\star) = \frac{1}{2}$, available without estimating any utility.

Informative fallback. When the safety test rejects the candidate, we retain the baseline item and select the promising partner with the largest comparison width:

$$\mathbf{y}_t^{\mathrm{fb}} \in \operatorname*{arg\,max}_{\mathbf{y}\in\mathcal{C}_t} \|\mathbf{x}_{b_t} - \mathbf{y}\|_{V_t^{-1}}, \qquad \text{displaying } (\mathbf{x}_{b_t}, \mathbf{y}_t^{\mathrm{fb}}). \tag{8}$$

Including the baseline guarantees that $\max\{s_t(\mathbf{x}_{b_t}), s_t(\mathbf{y}_t^{\mathrm{fb}})\} \geq s_t(\mathbf{x}_{b_t})$. The maximizing choice of partner also connects the displayed comparison to the candidate width. Since $\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t \in \mathcal{C}_t$, the triangle inequality gives

$$w_t \leq \|\tilde{\mathbf{x}}_t - \mathbf{x}_{b_t}\|_{V_t^{-1}} + \|\mathbf{x}_{b_t} - \tilde{\mathbf{y}}_t\|_{V_t^{-1}} \leq 2\|\mathbf{x}_{b_t} - \mathbf{y}_t^{\mathrm{fb}}\|_{V_t^{-1}}. \tag{9}$$

Thus, the displayed comparison retains at least half the candidate width, whether or not $\mathbf{x}_{b_t} \in \mathcal{C}_t$. Note that because the fallback pair contains the baseline, its one-round surplus under (3) is at least $\alpha s_t(\mathbf{x}_{b_t})$, regardless of its partner. The partner can therefore be chosen for information, and by (9) it retains at least half the candidate width. Its regret is also controlled: since $\mathbf{y}_t^{\mathrm{fb}} \in \mathcal{C}_t$, on $\mathcal{E}$,

$$R_t = \tfrac{1}{2}\Big(\Delta_{t,b_t} + f(\mathbf{x}_t^\star) - f(\mathbf{y}_t^{\mathrm{fb}})\Big) \leq \tfrac{1}{2}(\Delta_h + c_t).$$

The baseline contributes at most $\Delta_h/2$, and the partner's contribution is controlled by the same certificate c_t used for candidate decisions.

Safety budget. The resulting algorithm, `MaxInP-BC`, is given in Algorithm 1. It maintains a computable quantity Λ_t, initialized at $\Lambda_0 = 0$, that lower-bounds the cumulative safety surplus on $\mathcal{E}$. By Lemma 3.3, the candidate increment is $\frac{1}{2} - \frac{c_t}{4} - (1-\alpha)s_t(\mathbf{x}_{b_t})$. A fallback is credited with $\alpha s_t(\mathbf{x}_{b_t})$, which lower-bounds its actual surplus because the displayed pair contains the baseline. Each update is determined by the history, current contexts, baseline score, and chosen pair before

Algorithm 1 `MaxInP-BC` (Conservative Maximum-Informative-Pair)

Require: risk level α, confidence δ, regularisation λ, bounds S, L

Initialise $\Lambda_0 \leftarrow 0$, $V_1 \leftarrow \lambda I_d$
for $t = 1, 2, \ldots$ **do**
  Observe $\mathcal{S}_t$, baseline $\mathbf{x}_{b_t}$ and its known score $s_t(\mathbf{x}_{b_t})$; compute $\hat{\theta}_t, \beta_t$ by (4)–(5)
  Form $\mathcal{C}_t$ by (6); compute $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t)$, w_t, c_t by (7)
  if $\Lambda_{t-1} + \frac{1}{2} - \frac{c_t}{4} - (1-\alpha)s_t(\mathbf{x}_{b_t}) \geq 0$ **then**
    Display $(\mathbf{x}_t, \mathbf{y}_t) \leftarrow (\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t)$ and set $\Lambda_t \leftarrow \Lambda_{t-1} + \frac{1}{2} - \frac{c_t}{4} - (1-\alpha)s_t(\mathbf{x}_{b_t})$ {candidate}
  else
    Display $(\mathbf{x}_t, \mathbf{y}_t) \leftarrow (\mathbf{x}_{b_t}, \mathbf{y}_t^{\mathrm{fb}})$ from (8) and set $\Lambda_t \leftarrow \Lambda_{t-1} + \alpha s_t(\mathbf{x}_{b_t})$ {fallback}
  end if
  Observe o_t; set $\mathbf{z}_t \leftarrow \mathbf{x}_t - \mathbf{y}_t$ and $V_{t+1} \leftarrow V_t + \mathbf{z}_t\mathbf{z}_t^\top$
end for

the comparison outcome is observed. The candidate is accepted exactly when its proposed update leaves Λ_t nonnegative; otherwise, the fallback preserves nonnegativity. On $\mathcal{E}$, these per-round lower bounds therefore imply the cumulative safety constraint.

Let $\mathcal{A}_T, \mathcal{B}_T$ be the candidate and fallback rounds up to T, with $M_T = |\mathcal{A}_T|$ and $N_T = |\mathcal{B}_T|$. Since $\Lambda_{t-1} \geq 0$ by construction and $s_t(\mathbf{x}_{b_t}) \leq \frac{1}{2}$, the candidate increment is at least $\frac{\alpha}{2} - \frac{c_t}{4}$. Consequently, every round with $c_t \leq 2\alpha$ is accepted, and any rejection requires $c_t > 2\alpha$. This criterion applies separately at each round: the certificate can change with the contexts, estimates, and confidence radius, so acceptance at one round does not imply acceptance at all subsequent rounds.

Main result. With $\underline{s} := \sigma(-\Delta_h)$ and $\bar{\beta}_t := \max\{B/2, \beta_t\}$, define certificate envelope and the fallback bound

$$G(t) := 4\bar{\beta}_t \sqrt{2dt \log\left(1 + \frac{4L^2 t}{\lambda d}\right)}, \qquad N_{\mathrm{cons}} := \sup\left\{t \in \mathbb{N}_{\geq 1} : 4\alpha \underline{s}\, t < G(t)\right\}, \tag{10}$$

with $G(0) = 0$ and $\sup \emptyset := 0$. The function G is nondecreasing. For fixed problem parameters and confidence level, $G(t) = O(\sqrt{t}\log(et)) = o(t)$, so $N_{\mathrm{cons}} < \infty$.

Theorem 3.1 (Anytime safety and regret) *Let Assumptions 1, 2, 3, and 4 hold, and run Algorithm 1 with $\alpha, \delta \in (0,1)$ and $\lambda > 0$. With probability at least $1-\delta$, simultaneously for every $T \geq 1$, the displayed pairs are α-conservative in the sense of (3), the fallback count satisfies $N_T \leq \min\{T, N_{\mathrm{cons}}\}$, and $\mathtt{Reg}(T) \leq G(T) + \frac{1}{2}\Delta_h N_{\mathrm{cons}}$. When L, S, λ are treated as fixed constants, suppressing logarithmic factors in the remaining parameters gives*

$$\mathtt{Reg}(T) = \tilde{\mathcal{O}}\left(\frac{d\sqrt{T}}{\kappa} + \frac{\Delta_h d^2}{\kappa^2 \alpha^2 \underline{s}^{\,2}}\right).$$

Proof sketch. Work on $\mathcal{E}$, which has probability at least $1-\delta$ by Lemma 3.1; all guarantees hold simultaneously on this event.

(a) Envelope. On a candidate round, $c_t = \min\{B, 2\beta_t \|\mathbf{z}_t\|_{V_t^{-1}}\}$. On a fallback round, (9) gives $c_t \leq \min\{B, 4\beta_t \|\mathbf{z}_t\|_{V_t^{-1}}\}$. Thus, in both cases, $c_t \leq 4\bar{\beta}_t \min\{1, \|\mathbf{z}_t\|_{V_t^{-1}}\}$. Cauchy–Schwarz and the elliptical-potential bound $\sum_{i=1}^t \min\{1, \|\mathbf{z}_i\|^2_{V_i^{-1}}\} \leq 2d \log\left(1 + \frac{4L^2 t}{\lambda d}\right)$ yield $\sum_{i=1}^t c_i \leq G(t)$. This includes the certificate computed on a rejected round, since the displayed fallback comparison controls that certificate through (9).

(b) Safety. On candidate rounds, Lemma 3.3 shows that the true one-round surplus is at least the budget increment. On fallback rounds, including the baseline guarantees a true surplus of at least $\alpha s_t(\mathbf{x}_{b_t})$, again the budget increment. Induction therefore shows that Λ_t lower-bounds the true cumulative surplus. Candidate updates preserve $\Lambda_t \geq 0$ by the acceptance test, and fallback updates add a positive quantity, proving safety.

(c) Fallbacks. Expanding the failed acceptance test at a fallback round τ, and using $\frac{1}{2} - (1-\alpha)s_i(\mathbf{x}_{b_i}) \geq \alpha/2$ and $\alpha s_i(\mathbf{x}_{b_i}) \geq \alpha\underline{s}$, gives

$$\frac{\alpha}{2}(M_{\tau-1}+1) + \alpha\underline{s}N_{\tau-1} < \frac{1}{4}\left(\sum_{i\in\mathcal{A}_{\tau-1}} c_i + c_\tau\right) \leq \frac{1}{4}G(\tau).$$

Because $\underline{s} \leq 1/2$ and $M_{\tau-1} + N_{\tau-1} + 1 = \tau$, the left side is at least $\alpha\underline{s}\tau$. Hence every fallback time satisfies $4\alpha\underline{s}\tau < G(\tau)$ and is at most N_{cons}, proving $N_T \leq \min\{T, N_{\mathrm{cons}}\}$.

(d) Regret. Lemma 3.2 gives $R_t \leq c_t$ on candidate rounds. On fallback rounds, $\mathbf{y}_t^{\mathrm{fb}} \in \mathcal{C}_t$ gives $R_t \leq (\Delta_h + c_t)/2$. Therefore,

$$\mathtt{Reg}(T) \leq \sum_{t=1}^{T} c_t + \frac{\Delta_h}{2}N_T \leq G(T) + \frac{\Delta_h}{2}N_{\mathrm{cons}}.$$

For fixed L, S, λ, the simplified rate follows from $\bar{\beta}_t = \tilde{\mathcal{O}}(\sqrt{d}/\kappa)$ and the defining inequality for N_{cons}. Details are given in Appendix A. □

Remark 3.1 (An additive cost of conservatism) Under the normalization convention above, the leading term has the same dependence on d and T as unconstrained MaxInP (Saha, 2021), while the additional term $\Delta_h N_{\mathrm{cons}}/2$ is independent of T for fixed parameters.

Remark 3.2 (Why not transfer CLUCB directly) For bandits with numerical rewards, CLUCB (Kazerouni et al., 2017) certifies safety by lower-bounding cumulative candidate rewards with a confidence set, and it credits known baseline rewards separately. Under preference feedback, the design matrix in (4) accumulates comparison vectors $\mathbf{z}_t = \mathbf{x}_t - \mathbf{y}_t$, whereas average utilities depend on $(\mathbf{x}_t + \mathbf{y}_t)/2$. A comparison-based confidence set can therefore retain uncertainty in a common utility component that comparisons never reveal. Proposition A.2 (Appendix A.9) gives an i.i.d. two-dimensional instance on which this transfer incurs $\mathtt{Reg}(T) = \Omega(T)$ with high probability, even with exact baseline utilities and arbitrary candidate selection. The obstruction concerns the unchanged certification rule: known baseline utilities would also permit a baseline-centered cardinal certificate. `MaxInP-BC` instead certifies scores through utility gaps (Lemmas 3.2 and 3.3), which depend only on pairwise differences. Through (9), it also ties the uncertainty in this certificate to the comparisons displayed on both candidate and fallback rounds.

Remark 3.3 (Relaxing baseline-score knowledge) For ease of presentation, we assume that baseline scores are known exactly. Appendix A.10 relaxes this assumption by modifying the safety budget to use lower and upper score bounds together with the known range $[\underline{s}, 1/2]$. Deterministic bounds preserve the guarantees of Theorem 3.1. Statistical bounds with simultaneous failure probability at most δ_b give the same safety and regret bounds with success probability at least $1 - \delta - \delta_b$.

4 General Utility Functions

Linear utilities can be restrictive when items are described by rich features. We now let f lie in a general class $\mathcal{F}$, accessed only through an online logistic-regression oracle. In the unconstrained setting, Sekhari et al. (2023) sample comparisons from an inverse-gap-weighted (IGW) distribution built from the oracle's predictions. Without a confidence ellipsoid, we cannot certify each candidate individually as in Lemma 3.2. `BT-IGW-BC` therefore certifies safety cumulatively: the oracle guarantee yields a computable envelope on total regret, and a candidate is displayed only if the accumulated score credits cover it. Otherwise, the algorithm falls back to a baseline-containing pair whose partner is still drawn by IGW.

Assumption 5 (Realizability and logistic oracle) *The utility f belongs to a class $\mathcal{F}$ whose functions satisfy $|g(\mathbf{x}) - g(\mathbf{y})| \leq B$ on $\mathcal{D}$. Before its m-th comparison $(\mathbf{x}_m, \mathbf{y}_m)$ is drawn, the oracle outputs $\hat{f}_m$ with the same range bound; $\hat{f}_m$ may use the current item set and need not lie in $\mathcal{F}$. With $\ell_m(g) = \log(1 + e^{g(\mathbf{x}_m)-g(\mathbf{y}_m)}) - o_m[g(\mathbf{x}_m) - g(\mathbf{y}_m)]$, every fresh instance satisfies, on every adaptive sequence and for all $n \geq 1$,*

$$\sum_{m=1}^{n} \ell_m(\hat{f}_m) - \inf_{g\in\mathcal{F}} \sum_{m=1}^{n} \ell_m(g) \leq \mathcal{R}_{\log}(n), \tag{11}$$

where $\mathcal{R}_{\log}$ is known, nonnegative, nondecreasing, and $o(n)$.

Algorithm 2 BT-IGW-BC (Bradley–Terry Inverse-Gap-Weighting Baseline-Conservative)

Require: $\alpha, \delta \in (0,1)$, B, $\mathcal{R}_{\log}$, and an oracle satisfying Assumption 5

Initialise $Q_0 \leftarrow 0$
for $t = 1, 2, \ldots$ **do**
  If $t = 2^e$, restart the oracle and set ι_e, E_e, C_e by (13)
  Observe $\mathcal{S}_t$, $\mathbf{x}_{b_t}$, $s_t(\mathbf{x}_{b_t})$; obtain $\hat{f}_t$; form p_t with γ_t; set $h_t \leftarrow \frac{1}{2} - (1-\alpha)s_t(\mathbf{x}_{b_t})$
  if $Q_{t-1} + h_t \geq G_{\mathrm{IGW}}(t)/4$ **then**
    Draw $I_t, J_t \sim p_t$ independently; display $(\mathbf{x}_{t,I_t}, \mathbf{x}_{t,J_t})$; $Q_t \leftarrow Q_{t-1} + h_t$ {candidate}
  else
    Draw $J_t \sim p_t$; display $(\mathbf{x}_{b_t}, \mathbf{x}_{t,J_t})$; $Q_t \leftarrow Q_{t-1} + \alpha s_t(\mathbf{x}_{b_t})$ {fallback}
  end if
  Observe o_t and update the oracle with this comparison
end for

IGW with a fixed partner. Given $\hat{f}_t$, let $\hat{a}_t \in \arg\max_a \hat{f}_t(\mathbf{x}_{t,a})$ and

$$p_t(a) = \frac{1}{K + \gamma_t[\hat{f}_t(\mathbf{x}_{t,\hat{a}_t}) - \hat{f}_t(\mathbf{x}_{t,a})]} \quad (a \neq \hat{a}_t), \qquad p_t(\hat{a}_t) = 1 - \sum_{a \neq \hat{a}_t} p_t(a). \tag{12}$$

Write $\Delta_t(\mathbf{x}) = f(\mathbf{x}_t^\star) - f(\mathbf{x})$ and $\varepsilon_t(\mathbf{x}, \mathbf{y}) = (\hat{f}_t - f)(\mathbf{x}) - (\hat{f}_t - f)(\mathbf{y})$ for the margin error.

Lemma 4.1 (IGW with a fixed partner) *For every $\gamma_t > 0$ and $\mathbf{u} \in \mathcal{S}_t$, $\mathbb{E}_{a\sim p_t}\Delta_t(\mathbf{x}_{t,a}) \leq \frac{2K}{\gamma_t} + \frac{\gamma_t}{4}\mathbb{E}_{a\sim p_t}\varepsilon_t(\mathbf{x}_{t,a}, \mathbf{u})^2$.*

This is the standard IGW bound applied after subtracting the values at $\mathbf{u}$. The free partner lets the error be measured on the displayed pair: for a candidate $(\mathbf{x}_{t,I_t}, \mathbf{x}_{t,J_t})$ with $I_t, J_t \sim p_t$ independent, take $\mathbf{u}$ to be the other item; for a fallback $(\mathbf{x}_{b_t}, \mathbf{x}_{t,J_t})$, take $\mathbf{u} = \mathbf{x}_{b_t}$. This plays the role of (9): the fallback partner's gap is controlled exactly as a candidate's regret.

An anytime envelope. We run the oracle in epochs $e = 0, 1, \ldots$ of length $n_e = 2^e$, restarting it at each, with confidence $\delta_e = \delta/[(e+1)(e+2)]$. For local round m, set

$$\iota_e = \log\frac{3n_e}{\delta_e}, \quad E_e = \frac{8\mathcal{R}_{\log}(n_e)}{\kappa} + \frac{5\iota_e}{\kappa^2}, \quad \gamma_{e,m} = \sqrt{\frac{8Km}{E_e}}, \quad C_e = \frac{3}{\sqrt{2}}\sqrt{KE_e} + B\sqrt{2\iota_e}. \tag{13}$$

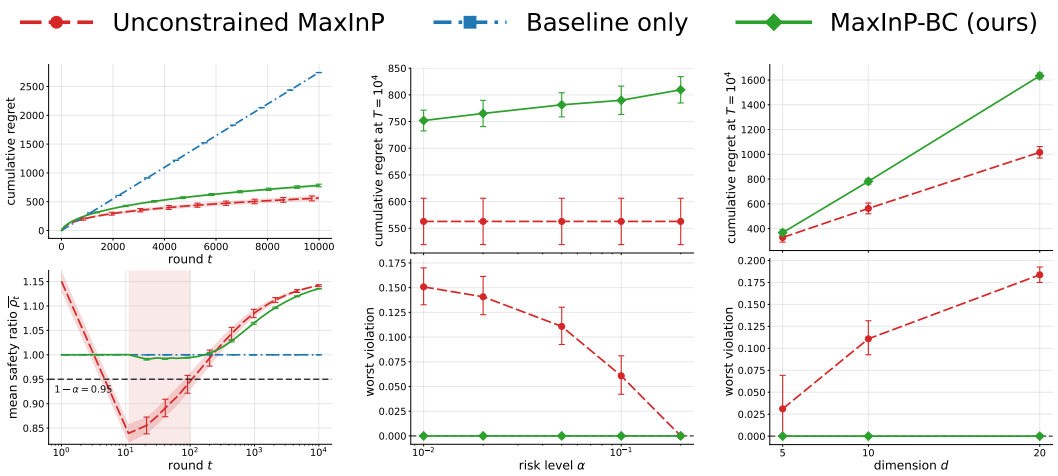
At time t, let $e(t) = \lfloor \log_2 t \rfloor$, $m(t) = t - 2^{e(t)} + 1$, $\gamma_t = \gamma_{e(t),m(t)}$, and $G_{\mathrm{IGW}}(t) = \sum_{j<e(t)} C_j\sqrt{2^j} + C_{e(t)}\sqrt{m(t)}$, $G_{\mathrm{IGW}}(0) = 0$. Let $r_t^\circ = R_t$ on candidate rounds and $r_t^\circ = \Delta_t(\mathbf{y}_t)$ on fallback rounds.

Lemma 4.2 (Anytime regret envelope) *Under Assumptions 1, 2, and 5, with probability at least $1-\delta$, Algorithm 2 satisfies $\sum_{i\leq t} r_i^\circ \leq G_{\mathrm{IGW}}(t)$ for all $t \geq 1$. Moreover, with $A(t) := \mathcal{R}_{\log}(t)/\kappa + 1/\kappa^2$, $G_{\mathrm{IGW}}(t) = \tilde{\mathcal{O}}(\sqrt{KtA(t)}) = o(t)$.*

The proof applies Lemma 4.1, the oracle bound and martingale concentration within each epoch (Appendix B).

Safety credits. Instead of charging each candidate its own certificate, BT-IGW-BC charges the envelope cumulatively. It keeps credits Q_t, with $Q_0 = 0$ and no reset between epochs: a candidate earns $h_t := \frac{1}{2} - (1-\alpha)s_t(\mathbf{x}_{b_t})$ and a fallback earns $\alpha s_t(\mathbf{x}_{b_t})$. By Lemma 3.3 with $c = R_t$, a candidate's true surplus is at least $h_t - R_t/4$; a fallback's is at least $\alpha s_t(\mathbf{x}_{b_t})$ because it contains the baseline. The true cumulative surplus is therefore at least $Q_t - \frac{1}{4}\sum_{i\in\mathcal{A}_t} R_i$, where $\mathcal{A}_t$ is the set of candidate rounds, and by Lemma 4.2 the charge is at most $G_{\mathrm{IGW}}(t)/4$, which the acceptance test in Algorithm 2 requires the credits to cover.

Main result. Since $s_t(\mathbf{x}_{b_t}) \in [\underline{s}, \frac{1}{2}]$ with $\underline{s} = \sigma(-\Delta_h)$, every credit is at least $\alpha\underline{s}$, so a rejection at time t requires $4\alpha\underline{s}\,t < G_{\mathrm{IGW}}(t)$. Let N_T count fallbacks up to T, and let $N_{\mathrm{IGW}} := \sup\{t \geq 1 : 4\alpha\underline{s}\,t < G_{\mathrm{IGW}}(t)\}$, which is finite since $G_{\mathrm{IGW}}(t) = o(t)$.



(a) $d = 10$, $\alpha = 0.05$ (b) Varying α ($d = 10$) (c) Varying d ($\alpha = 0.05$)

Figure 1: Linear utilities. Top: cumulative regret. Bottom: mean safety ratio $\bar{\rho}_t$ in (a) and worst violation in (b, c). Dashed lines mark the bar $1-\alpha$ in (a) and zero violation in (b, c); the vertical band in (a) marks the rounds where MaxInP is below the bar. Baseline-only is shown in (a) only. Intervals are 95% confidence intervals over 50 seeds. `MaxInP-BC` has zero violation in every setting.

Theorem 4.1 (Anytime safety and regret) *Under Assumptions 1, 2, 3, and 5, with probability at least $1-\delta$, simultaneously for every $T \geq 1$, Algorithm 2 is α-conservative in the sense of (3), $N_T \leq \min\{T, N_{\mathrm{IGW}}\}$, and $\mathtt{Reg}(T) \leq G_{\mathrm{IGW}}(T) + \frac{1}{2}\Delta_h N_T$. Suppressing log factors in T and $1/\delta$,*

$$\mathtt{Reg}(T) = \tilde{\mathcal{O}}\left(\sqrt{KT\,A(T)} + \frac{\Delta_h K\,A(T)}{\alpha^2 \underline{s}^{\,2}}\right). \tag{14}$$

Proof sketch. Work on the event of Lemma 4.2. *Safety:* if $\tau \leq t$ is the last candidate round, then since $r_i^\circ \geq 0$ and credits are positive, $\sum_{i\in\mathcal{A}_t} R_i \leq \sum_{i\leq\tau} r_i^\circ \leq G_{\mathrm{IGW}}(\tau) \leq 4Q_\tau \leq 4Q_t$, so the true surplus is nonnegative. *Fallbacks:* every fallback time satisfies $4\alpha\underline{s}\,\tau < G_{\mathrm{IGW}}(\tau)$, so $\tau \leq N_{\mathrm{IGW}}$; at the last fallback $\tau \leq T$, $G_{\mathrm{IGW}}(\tau) = \tilde{\mathcal{O}}(\sqrt{K\tau A(T)})$ gives $N_T = \tilde{\mathcal{O}}(KA(T)/(\alpha^2\underline{s}^{\,2}))$. *Regret:* $R_t = r_t^\circ$ on candidate rounds and $R_t \leq \frac{1}{2}(\Delta_h + r_t^\circ)$ on fallbacks, so $\mathtt{Reg}(T) \leq G_{\mathrm{IGW}}(T) + \frac{1}{2}\Delta_h N_T$. Details are in Appendix B.6. □

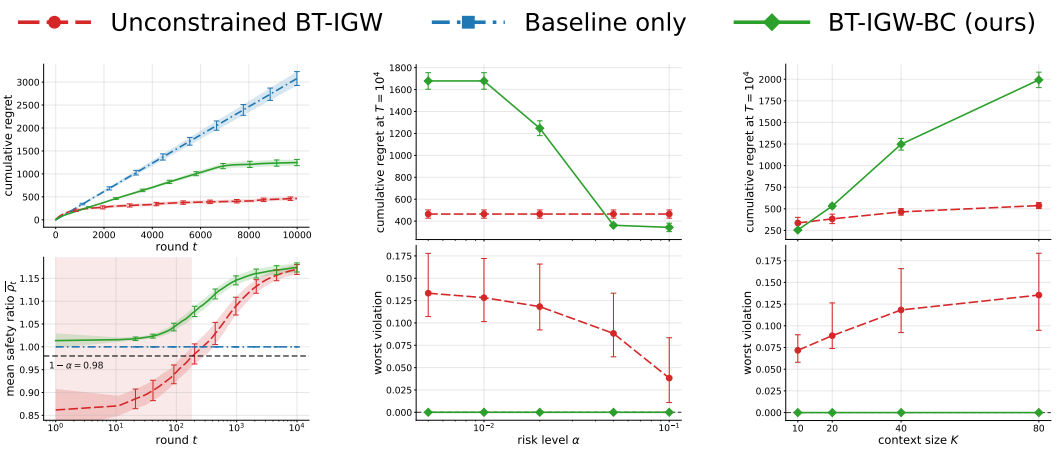
Remark 4.1 (Finite classes) The oracle in Appendix B.7 has $\mathcal{R}_{\log}(n) = e^B \log|\mathcal{F}|$, so $A(T) \equiv A_{\mathcal{F}} := e^B \log|\mathcal{F}|/\kappa + 1/\kappa^2$ and, as in the linear case, the cost of conservatism is independent of T.

Remark 4.2 (From C-SquareCB to preference feedback) C-SquareCB (Deb et al., 2025) certifies safety by calibrating cumulative predictions of cardinal costs. Our oracle controls only comparison margins, so we certify scores through cumulative regret (Lemmas 3.3 and 4.2), and Lemma 4.1 lets fallbacks keep learning. Appendix B.1 explains why the calibration argument does not transfer.

5 Experiments

We compare each conservative algorithm with its unconstrained counterpart, MaxInP (Saha, 2021) or BT-IGW (Sekhari et al., 2023), and with *baseline-only*, which selects the baseline item.

Protocol. Each configuration runs for $T = 10^4$ rounds on 50 independent instances. We report cumulative regret and the *safety ratio* $\rho_t = \sum_{i\leq t} \max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} / \sum_{i\leq t} s_i(\mathbf{x}_{b_i})$; the constraint (3) holds at prefix t if and only if $\rho_t \geq 1-\alpha$. Sweeps report the *worst violation* $\max_t[(1-\alpha) - \bar{\rho}_t]_+$ of the cross-seed mean $\bar{\rho}_t$. Intervals are 95% confidence intervals over seeds. The baseline is the item at fractional utility rank q, the fraction of items with higher utility. We tune only the multipliers on β_t, $\gamma_{e,m}$ and G_{IGW}, by grid search on seeds disjoint from those reported: unconstrained



(a) $K = 40$, $\alpha = 0.02$ (b) Varying α ($K = 40$) (c) Varying K ($\alpha = 0.02$)

Figure 2: General utilities ($d = 5$, $|\mathcal{F}| = 64$, $q = 0.10$), laid out as in Figure 1 with BT-IGW in place of MaxInP. BT-IGW-BC stays above the bar in (a), and its extra regret over BT-IGW arises mainly where BT-IGW violates the constraint.

learners minimise regret, and conservative learners minimise regret subject to the constraint. Tuned multipliers fall outside the theory, so the safety reported here is empirical.

Linear utilities: We take $d = 10$, $K = 50$, $\alpha = 0.05$ and $q = 0.14$, with β-multipliers 0.02 for MaxInP and 0.05 for MaxInP-BC. MaxInP attains the lowest regret, but its mean safety ratio drops to 0.839 and stays below the bar for roughly the first 100 rounds, because its most-uncertain pair can contain two poor items. MaxInP-BC never drops below 0.991, at $1.39\times$ the regret of MaxInP, while baseline-only is safe but incurs linear regret, $3.5\times$ that of MaxInP-BC (Figure 1(a)). The sweeps confirm this across settings: MaxInP-BC has zero violation for every α and d, whereas MaxInP violates the constraint in all but the loosest setting.

General utilities: Here $d = 5$, $K = 40$, $\alpha = 0.02$ and $q = 0.10$, with a realizable class of $|\mathcal{F}| = 64$ random quadratics and an exponential-weights oracle, for which $\mathcal{R}_{\log}(n) = \log|\mathcal{F}|$. BT-IGW has the lowest regret (464.5), but its mean safety ratio starts at 0.862 and stays below the bar for the first $\approx$190 rounds, whereas BT-IGW-BC stays at or above 1.014 at $2.69\times$ the regret, and baseline-only is a further $2.47\times$ worse (Figure 2(a)). Across the sweeps, BT-IGW-BC has zero violation while BT-IGW violates the constraint in every setting. The cost of conservatism is largest when the constraint is tight (small α, large K), and when it is loose ($\alpha \geq 0.05$ or $K = 10$), BT-IGW-BC matches or beats BT-IGW.

6 Conclusion

We introduced conservative contextual dueling bandits, which require the cumulative score of the better displayed item to stay above a $(1-\alpha)$ fraction of a trusted baseline's at all times. Because the constraint is stated through preference scores rather than rewards, it is invariant to shifts of the utility function and can be enforced from binary comparisons alone. We showed that this setting is not a routine extension of conservative bandits with numerical rewards, since transferring CLUCB directly can incur linear regret. MaxInP-BC and BT-IGW-BC avoid this failure by certifying the constraint through utility gaps instead of observed payoffs and by continuing to learn from the comparisons made when they fall back to the baseline. Both are safe at all times with high probability, for linear and general utility classes respectively. In both cases, conservatism adds only an additive term to the regret. Over long horizons, the price of safety is therefore negligible compared with the cost of learning itself. Future work includes lower bounds that would establish whether the cost of conservatism is tight, extensions to feedback over more than two items, and experiments with human preference data.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, support qualitative and thematic data analysis, interpret results, and propose or refine hypotheses. We did not use generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, or clean and reformat data sets. Assistance with translation is not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures and to polish the writing of the paper. We have reviewed all AI-assisted work: we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and LLM-generated code was verified and tested for correctness by one author. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided our code to generate our results as a zip file with all the instructions. The hyperparameters are provided in the code.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Eric Brochu, Nando de Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2007.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems*, 30(1):6:1–6:41, 2012.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SThJXvucjQ.

Johannes Fürnkranz and Eyke Hüllermeier (eds.). *Preference Learning*. Springer, 2010.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *AAAI*, 2020.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence. Preferential Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1282–1291, 2017.

Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1):1–117, 2016.

David R Hunter. Mm algorithms for generalized bradley–terry models. *Annals of Statistics*, 32(1): 384–406, 2004.

Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, 2002.

Nicolas Jones, Armelle Brun, and Anne Boyer. Comparisons instead of ratings: Towards more stable preferences. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pp. 451–456, 2011. doi: 10.1109/WI-IAT.2011.13.

Saikishore Kalloori, Francesco Ricci, and Marko Tkalčič. Pairwise preferences based matrix factorization and nearest neighbor recommendation techniques. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 2016.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/bdc4626aa1d1df8e14d80d345b2a442d-Paper.pdf.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29406–29426. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24co.html.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation, 2005.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 43–52, 2008.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30050–30062. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fc3cf452d3da8402bebb765225ce8c0e-Paper.pdf.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2567c95fd41459a98a73ba893775d22a-Paper-Conference.pdf.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 3008–3021, 2020.

Maegan Tucker, Ellen Novoseller, Claudia Kann, Yanan Sui, Yisong Yue, Joel W. Burdick, and Aaron D. Ames. Preference-based learning for exoskeleton gait optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2351–2357, 2020.

Arun Verma, Zhongxiang Dai, Xiaoqiang Lin, Patrick Jaillet, and Bryan Kian Hsiang Low. Neural dueling bandits: Preference-based optimization with human feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VELhv9BBfn.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari. Conservative bandits. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/wu16.html.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012a.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, September 2012b. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.028. URL https://doi.org/10.1016/j.jcss.2011.12.028.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models, 2020. URL https://arxiv.org/abs/1603.07323.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Proofs for Section 3

Throughout this appendix Assumptions 1, 2, 3 and 4 are in force, and $\lambda > 0$, $\alpha, \delta \in (0,1)$ are the inputs of Algorithm 1.

A.1 Notation and filtration

Recall $\mathbf{z}_t = \mathbf{x}_t - \mathbf{y}_t$ is the comparison vector of the pair actually displayed at round t, that $V_1 = \lambda I_d$ and $V_{t+1} = V_t + \mathbf{z}_t \mathbf{z}_t^\top$, and write

$$\varepsilon_t \;:=\; o_t - \sigma(\mathbf{z}_t^\top \theta^\star), \qquad \mathbf{S}_t \;:=\; \sum_{i=1}^{t} \varepsilon_i \mathbf{z}_i, \qquad v_t \;:=\; \|\mathbf{z}_t\|_{V_t^{-1}}.$$

Let $\mathcal{F}_{t-1}$ be the σ-algebra generated by all contexts, displayed pairs and outcomes of rounds $1, \dots, t-1$, together with the round-t context set $\mathcal{S}_t$, the baseline item $\mathbf{x}_{b_t}$ and its score, and the round-t displayed pair. Every quantity the algorithm computes at round t — V_t, $\hat{\theta}_t$, β_t, $\mathcal{C}_t$, $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t)$, w_t, c_t, Λ_{t-1}, the acceptance decision, and hence $\mathbf{z}_t$ — is $\mathcal{F}_{t-1}$-measurable, and $(\mathcal{F}_t)_{t \ge 0}$ is a filtration with $\mathbb{E}[o_t \mid \mathcal{F}_{t-1}] = \sigma(\mathbf{z}_t^\top \theta^\star)$ by Assumption 1. Unlike analyses that discard uninformative baseline rounds, no random time change is needed here.

Remark A.1 (Degenerate rounds) A round with $\mathbf{z}_t = 0$ contributes $\varepsilon_t \mathbf{z}_t = 0$ to $\mathbf{S}_t$, leaves $V_{t+1} = V_t$, and adds the constant $\log 2$ to the objective in (4); it therefore affects neither the estimator nor any argument below. Such a round can only be a candidate round: by (9), if $\max_{\mathbf{y} \in \mathcal{C}_t} \|\mathbf{x}_{b_t} - \mathbf{y}\|_{V_t^{-1}} = 0$ then $w_t = 0$, hence $c_t = 0$, and the acceptance test passes because $\Lambda_{t-1} \ge 0$ and $\frac{1}{2} - (1-\alpha) s_t(\mathbf{x}_{b_t}) \ge \frac{\alpha}{2} > 0$ (Lemma A.1). Consequently every fallback round has $v_t \ge w_t / 2 > 0$.

A.2 Elementary bounds on scores

Lemma A.1 *For every t and every $\mathbf{x} \in \mathcal{S}_t$, $\sigma(-B) \le s_t(\mathbf{x}) \le \frac{1}{2}$, and $s_t(\mathbf{x}_t^\star) = \frac{1}{2}$. Moreover $\underline{s} := \sigma(-\Delta_h)$ satisfies $\underline{s} \le s_t(\mathbf{x}_{b_t}) \le \frac{1}{2}$, so that*

$$\tfrac{1}{2} - (1-\alpha) s_t(\mathbf{x}_{b_t}) \;\ge\; \tfrac{\alpha}{2} \qquad \text{and} \qquad \alpha\, s_t(\mathbf{x}_{b_t}) \;\ge\; \alpha \underline{s}.$$

Proof. By definition $s_t(\mathbf{x}) = \sigma\big(f(\mathbf{x}) - f(\mathbf{x}_t^\star)\big)$ with $f(\mathbf{x}) - f(\mathbf{x}_t^\star) \in [-B, 0]$ by Assumption 2 and the definition of $\mathbf{x}_t^\star$; monotonicity of σ and $\sigma(0) = \frac{1}{2}$ give the first claim. For the baseline, $f(\mathbf{x}_t^\star) - f(\mathbf{x}_{b_t}) = \Delta_{t,b_t} \in [\Delta_l, \Delta_h]$ by Assumption 3, so $s_t(\mathbf{x}_{b_t}) \in [\sigma(-\Delta_h), \sigma(-\Delta_l)] \subseteq [\underline{s}, \frac{1}{2}]$. Finally $\frac{1}{2} - (1-\alpha) s_t(\mathbf{x}_{b_t}) \ge \frac{1}{2} - (1-\alpha)\frac{1}{2} = \frac{\alpha}{2}$. □

Note for later use that $\underline{s} \le \frac{1}{2}$, since $\Delta_h \ge 0$.

A.3 Anytime logistic confidence ellipsoid

This subsection proves Lemma 3.1. The argument is self-contained.

Lemma A.2 (Hoeffding's lemma) *Let X be a random variable supported on an interval of length $b - a$. Then $\mathbb{E} \exp\big(\eta(X - \mathbb{E}X)\big) \le \exp\big(\eta^2 (b-a)^2 / 8\big)$ for every $\eta \in \mathbb{R}$.*

Proof. After the affine rescaling $a = 0$, $b = 1$ it suffices to treat $X \in [0,1]$; write $\mu = \mathbb{E}X$. Convexity of the exponential gives $e^{\eta x} \le (1-x)e^0 + x e^\eta$ for $x \in [0,1]$, so taking expectations and centring,

$$\mathbb{E}\, e^{\eta(X-\mu)} \;\le\; (1-\mu) e^{-\mu\eta} + \mu e^{(1-\mu)\eta}. \tag{15}$$

Let $h(\eta) = \log\big((1-\mu)e^{-\mu\eta} + \mu e^{(1-\mu)\eta}\big)$ and $q(\eta) = \mu e^{(1-\mu)\eta} / \big((1-\mu)e^{-\mu\eta} + \mu e^{(1-\mu)\eta}\big)$. Differentiating, $h'(\eta) = q(\eta) - \mu$ and $h''(\eta) = q(\eta)(1-q(\eta)) \le \frac{1}{4}$, while $h(0) = 0$ and $h'(0) = q(0) - \mu = 0$. Taylor's formula with integral remainder, valid for either sign of η, gives

$$h(\eta) = \eta^2 \int_0^1 (1-s) h''(s\eta)\, \mathrm{d}s \;\le\; \frac{\eta^2}{8}.$$

Exponentiating and combining with (15) proves the normalised case; for a general interval apply it to $(X-a)/(b-a)$ with parameter $\eta(b-a)$. □

Lemma A.3 (Conditional sub-Gaussianity of preference residuals) *For every $t \geq 1$, $\mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[e^{\eta\varepsilon_t} \mid \mathcal{F}_{t-1}] \leq e^{\eta^2/8}$ for all $\eta \in \mathbb{R}$. That is, ε_t is conditionally R-sub-Gaussian with $R = \frac{1}{2}$ in the convention $\mathbb{E}[e^{\eta\varepsilon} \mid \mathcal{F}_{t-1}] \leq e^{R^2\eta^2/2}$.*

Proof. Conditionally on $\mathcal{F}_{t-1}$ the outcome o_t is Bernoulli with mean $\sigma(\mathbf{z}_t^\top \theta^\star)$, since $\mathbf{z}_t$ is $\mathcal{F}_{t-1}$-measurable; it is supported on an interval of length one. Apply Lemma A.2 conditionally. □

Lemma A.4 (Ville's inequality) *Let $(M_n)_{n\geq 0}$ be a nonnegative supermartingale with $\mathbb{E}\, M_0 \leq 1$. Then $\mathbb{P}(\sup_{n\geq 0} M_n \geq c) \leq 1/c$ for every $c > 0$.*

Proof. Fix a deterministic N and let $\tau_N = \min\{n \leq N : M_n \geq c\}$, with $\tau_N = N$ if the set is empty. Since $\{\tau_N \geq n\} \in \mathcal{F}_{n-1}$,

$$\mathbb{E}\, M_{\tau_N} = \mathbb{E}\, M_0 + \sum_{n=1}^{N} \mathbb{E}\left[\mathbb{I}\{\tau_N \geq n\}\, \mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}]\right] \;\leq\; \mathbb{E}\, M_0 \;\leq\; 1.$$

On the event $\max_{0\leq n\leq N} M_n \geq c$ the stopped value is at least c, so $c\,\mathbb{P}(\max_{n\leq N} M_n \geq c) \leq \mathbb{E}\, M_{\tau_N} \leq 1$. Let $N \to \infty$ and use continuity from below. □

Lemma A.5 (Method-of-mixtures self-normalised bound) *Let $(\varepsilon_i, z_i)_{i\geq 1}$ be adapted with $z_i \in \mathbb{R}^d$ being $\mathcal{F}_{i-1}$-measurable and $\mathbb{E}[e^{\eta\varepsilon_i} \mid \mathcal{F}_{i-1}] \leq e^{R^2\eta^2/2}$ for all $\eta \in \mathbb{R}$. Let $V_0 \succ 0$, $S_n = \sum_{i\leq n} \varepsilon_i z_i$ and $V_n' = V_0 + \sum_{i\leq n} z_i z_i^\top$. Then with probability at least $1-\delta$, simultaneously for every $n \geq 0$,*

$$\|S_n\|^2_{(V_n')^{-1}} \;\leq\; 2R^2 \log\left(\frac{\det(V_n')^{1/2}}{\det(V_0)^{1/2}\,\delta}\right). \tag{16}$$

Proof. Fix $q \in \mathbb{R}^d$. Conditional sub-Gaussianity with $\eta = q^\top z_i$ gives $\mathbb{E}\left[\exp\left(q^\top z_i\varepsilon_i - \frac{R^2}{2}(q^\top z_i)^2\right) \mid \mathcal{F}_{i-1}\right] \leq 1$, so

$$M_n(q) = \exp\left(q^\top S_n - \tfrac{R^2}{2} q^\top (V_n' - V_0) q\right)$$

is a nonnegative supermartingale with $M_0(q) = 1$. Let Q be an independent Gaussian vector with density $f(q) = R^d \det(V_0)^{1/2}(2\pi)^{-d/2}\exp(-\frac{R^2}{2}q^\top V_0 q)$, which integrates to one because $\int e^{-\frac{1}{2}q^\top A q}\mathrm{d}q = (2\pi)^{d/2}\det(A)^{-1/2}$ with $A = R^2 V_0$. By Tonelli's theorem and the conditional supermartingale property for each fixed q, the mixture $M_n = \int M_n(q) f(q)\mathrm{d}q$ is a nonnegative supermartingale with $M_0 = 1$, and

$$M_n = \frac{R^d \det(V_0)^{1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(q^\top S_n - \tfrac{R^2}{2} q^\top V_n' q\right)\mathrm{d}q.$$

Completing the square with $a = R^{-2}(V_n')^{-1} S_n$ gives

$$q^\top S_n - \tfrac{R^2}{2} q^\top V_n' q = -\tfrac{R^2}{2}(q-a)^\top V_n'(q-a) + \tfrac{1}{2R^2}\|S_n\|^2_{(V_n')^{-1}},$$

and evaluating the Gaussian integral yields

$$M_n = \left(\frac{\det V_0}{\det V_n'}\right)^{1/2} \exp\left(\frac{1}{2R^2}\|S_n\|^2_{(V_n')^{-1}}\right). \tag{17}$$

Lemma A.4 with $c = 1/\delta$ bounds $\sup_n M_n < 1/\delta$ with probability at least $1-\delta$; taking logarithms in (17) and rearranging gives (16). □

Lemma A.6 (Trace–determinant bound) *If $\|z_i\|_2 \le 2L$ for all i and $W_m = \lambda I_d + \sum_{i=1}^{m-1} z_i z_i^\top$, then*

$$\log \frac{\det W_m}{\det(\lambda I_d)} \;\le\; d \log\Big(1 + \frac{4L^2(m-1)}{\lambda d}\Big).$$

Proof. Let $A = \lambda^{-1} \sum_{i<m} z_i z_i^\top \succeq 0$ with eigenvalues $\rho_1, \dots, \rho_d \ge 0$, so that $\det W_m / \det(\lambda I_d) = \prod_j (1+\rho_j)$. By the arithmetic–geometric mean inequality, $\prod_j (1+\rho_j) \le \big(1 + \mathrm{Tr}(A)/d\big)^d$, and $\mathrm{Tr}(A) = \lambda^{-1} \sum_{i<m} \|z_i\|_2^2 \le 4L^2(m-1)/\lambda$. Take logarithms. □

Lemma A.7 (Global curvature of the regularised likelihood) *Let $F_t(\theta) = \sum_{i<t} \big[\log(1 + e^{\mathbf{z}_i^\top \theta}) - o_i \mathbf{z}_i^\top \theta\big] + \frac{\lambda}{2}\|\theta\|_2^2$. For all $\theta, \theta' \in \Theta$,*

$$\big\langle \nabla F_t(\theta) - \nabla F_t(\theta'),\, \theta - \theta' \big\rangle \;\ge\; \kappa\, \|\theta - \theta'\|_{V_t}^2.$$

Proof. For $\vartheta \in \Theta$ we have $\nabla^2 F_t(\vartheta) = \sum_{i<t} \sigma'(\mathbf{z}_i^\top \vartheta) \mathbf{z}_i \mathbf{z}_i^\top + \lambda I_d$. Since $\|\mathbf{z}_i\|_2 \le 2L$ and $\|\vartheta\|_2 \le S$, $|\mathbf{z}_i^\top \vartheta| \le 2LS = B$, so $\sigma'(\mathbf{z}_i^\top \vartheta) \ge \kappa$ and the fact that σ' is even and decreasing in $|s|$. As $0 < \kappa \le \frac{1}{4} < 1$,

$$\nabla^2 F_t(\vartheta) \;\succeq\; \kappa \sum_{i<t} \mathbf{z}_i \mathbf{z}_i^\top + \lambda I_d \;\succeq\; \kappa \Big(\sum_{i<t} \mathbf{z}_i \mathbf{z}_i^\top + \lambda I_d \Big) = \kappa V_t.$$

Writing $\mathbf{u} = \theta - \theta'$ and noting Θ is convex, the fundamental theorem of calculus gives $\langle \nabla F_t(\theta) - \nabla F_t(\theta'), \mathbf{u} \rangle = \int_0^1 \mathbf{u}^\top \nabla^2 F_t(\theta' + s\mathbf{u}) \mathbf{u}\, \mathrm{d}s \ge \kappa\, \mathbf{u}^\top V_t \mathbf{u}$. □

Lemma 3.1 (restated). *With probability at least $1 - \delta$, the event $\mathcal{E} = \{\|\hat\theta_t - \theta^\star\|_{V_t} \le \beta_t \ \forall t \ge 1\}$ holds, with β_t as in (5).*

Proof. Set $\mathbf{u}_t = \hat\theta_t - \theta^\star$. Since $\hat\theta_t$ minimises the differentiable convex F_t over the closed convex Θ and $\theta^\star \in \Theta$, first-order optimality gives $\langle \nabla F_t(\hat\theta_t), \theta^\star - \hat\theta_t \rangle \ge 0$, i.e. $\langle \nabla F_t(\hat\theta_t), \mathbf{u}_t \rangle \le 0$. Lemma A.7 with $\theta = \hat\theta_t$, $\theta' = \theta^\star$ then yields

$$\kappa \|\mathbf{u}_t\|_{V_t}^2 \;\le\; \big\langle \nabla F_t(\hat\theta_t) - \nabla F_t(\theta^\star), \mathbf{u}_t \big\rangle \;\le\; -\big\langle \nabla F_t(\theta^\star), \mathbf{u}_t \big\rangle. \tag{18}$$

The gradient at the truth is $\nabla F_t(\theta^\star) = \sum_{i<t} \big(\sigma(\mathbf{z}_i^\top \theta^\star) - o_i\big)\mathbf{z}_i + \lambda \theta^\star = -\mathbf{S}_{t-1} + \lambda \theta^\star$, so by Cauchy–Schwarz in the V_t geometry,

$$-\big\langle \nabla F_t(\theta^\star), \mathbf{u}_t \big\rangle = \langle \mathbf{S}_{t-1}, \mathbf{u}_t \rangle - \lambda \langle \theta^\star, \mathbf{u}_t \rangle \;\le\; \Big(\|\mathbf{S}_{t-1}\|_{V_t^{-1}} + \lambda \|\theta^\star\|_{V_t^{-1}} \Big) \|\mathbf{u}_t\|_{V_t}.$$

Since $V_t \succeq \lambda I_d$ we have $V_t^{-1} \preceq \lambda^{-1} I_d$, hence $\lambda \|\theta^\star\|_{V_t^{-1}} \le \sqrt{\lambda}\, \|\theta^\star\|_2 \le \sqrt{\lambda}\, S$. If $\mathbf{u}_t = 0$ the claim is trivial; otherwise divide (18) by $\kappa \|\mathbf{u}_t\|_{V_t}$ to obtain

$$\|\mathbf{u}_t\|_{V_t} \;\le\; \frac{1}{\kappa} \Big(\|\mathbf{S}_{t-1}\|_{V_t^{-1}} + \sqrt{\lambda}\, S \Big). \tag{19}$$

Apply Lemma A.5 with $R = \frac{1}{2}$, $V_0 = \lambda I_d$, $z_i = \mathbf{z}_i$ and $n = t - 1$, so that $V'_{t-1} = V_t$: with probability at least $1 - \delta$, simultaneously for every $t \ge 1$,

$$\|\mathbf{S}_{t-1}\|_{V_t^{-1}} \;\le\; \frac{1}{2} \sqrt{2 \log\Big(\frac{\det(V_t)^{1/2}}{\det(\lambda I_d)^{1/2} \delta} \Big)} \;=\; \frac{1}{2} \sqrt{\log \frac{\det V_t}{\det(\lambda I_d)} + 2 \log \frac{1}{\delta}}\,.$$

Bounding the determinant ratio by Lemma A.6 with $m = t$ and substituting into (19) gives exactly β_t of (5). □

For the remainder of the appendix we work on $\mathcal{E}$, which by Lemma 3.1 has probability at least $1 - \delta$. All statements below are deterministic consequences of $\mathcal{E}$; in particular no further probabilistic argument is attached to the safety budget.

A.4 THE PROMISING SET AND THE REGRET CERTIFICATE

Write $U_t(\mathbf{x}, \mathbf{y}) = \langle \hat{\theta}_t, \mathbf{x} - \mathbf{y} \rangle + \beta_t \|\mathbf{x} - \mathbf{y}\|_{V_t^{-1}}$, so that $\mathcal{C}_t = \{\mathbf{x} \in \mathcal{S}_t : U_t(\mathbf{x}, \mathbf{y}) \geq 0 \ \forall \mathbf{y} \in \mathcal{S}_t\}$ as in (6).

Lemma A.8 (Nonemptiness) *For every round t and every value of $\hat{\theta}_t$, $\mathcal{C}_t \neq \emptyset$.*

Proof. Let $\hat{\mathbf{x}}_t \in \arg\max_{\mathbf{x} \in \mathcal{S}_t} \langle \hat{\theta}_t, \mathbf{x} \rangle$. For every $\mathbf{y} \in \mathcal{S}_t$ we have $\langle \hat{\theta}_t, \hat{\mathbf{x}}_t - \mathbf{y} \rangle \geq 0$, and the confidence bonus is nonnegative, so $U_t(\hat{\mathbf{x}}_t, \mathbf{y}) \geq 0$. Hence $\hat{\mathbf{x}}_t \in \mathcal{C}_t$. □

Lemma A.9 (The optimal item is promising) *On $\mathcal{E}$, $\mathbf{x}_t^\star \in \mathcal{C}_t$ for every t.*

Proof. Fix $\mathbf{y} \in \mathcal{S}_t$ and put $\mathbf{z} = \mathbf{x}_t^\star - \mathbf{y}$. On $\mathcal{E}$, Cauchy–Schwarz gives $\langle \hat{\theta}_t - \theta^\star, \mathbf{z} \rangle \geq -\beta_t \|\mathbf{z}\|_{V_t^{-1}}$, whence

$$U_t(\mathbf{x}_t^\star, \mathbf{y}) = \langle \hat{\theta}_t, \mathbf{z} \rangle + \beta_t \|\mathbf{z}\|_{V_t^{-1}} \ \geq \ \langle \theta^\star, \mathbf{z} \rangle = f(\mathbf{x}_t^\star) - f(\mathbf{y}) \ \geq \ 0.$$

□

Lemma A.10 (Gap of a promising item) *On $\mathcal{E}$, every $\mathbf{x} \in \mathcal{C}_t$ satisfies $f(\mathbf{x}_t^\star) - f(\mathbf{x}) \leq 2\beta_t \|\mathbf{x}_t^\star - \mathbf{x}\|_{V_t^{-1}}$.*

Proof. Since $\mathbf{x} \in \mathcal{C}_t$ and $\mathbf{x}_t^\star \in \mathcal{S}_t$, the definition of $\mathcal{C}_t$ with $\mathbf{y} = \mathbf{x}_t^\star$ gives $U_t(\mathbf{x}, \mathbf{x}_t^\star) \geq 0$, i.e.

$$\langle \hat{\theta}_t, \mathbf{x}_t^\star - \mathbf{x} \rangle \ \leq \ \beta_t \|\mathbf{x}_t^\star - \mathbf{x}\|_{V_t^{-1}}. \tag{20}$$

On $\mathcal{E}$, Cauchy–Schwarz also gives

$$\langle \theta^\star - \hat{\theta}_t, \mathbf{x}_t^\star - \mathbf{x} \rangle \ \leq \ \beta_t \|\mathbf{x}_t^\star - \mathbf{x}\|_{V_t^{-1}}. \tag{21}$$

Adding (20) and (21) yields $\langle \theta^\star, \mathbf{x}_t^\star - \mathbf{x} \rangle \leq 2\beta_t \|\mathbf{x}_t^\star - \mathbf{x}\|_{V_t^{-1}}$. □

Lemma 3.2 (restated). *$\mathcal{C}_t \neq \emptyset$ always. On $\mathcal{E}$, $\mathbf{x}_t^\star \in \mathcal{C}_t$ and $f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq c_t$ for every $\mathbf{y} \in \mathcal{C}_t$; in particular the candidate pair satisfies $R_t = f(\mathbf{x}_t^\star) - \frac{1}{2}\big(f(\tilde{\mathbf{x}}_t) + f(\tilde{\mathbf{y}}_t)\big) \leq c_t$.*

Proof. Nonemptiness is Lemma A.8 and $\mathbf{x}_t^\star \in \mathcal{C}_t$ is Lemma A.9. Fix $\mathbf{y} \in \mathcal{C}_t$. Both $\mathbf{x}_t^\star$ and $\mathbf{y}$ lie in $\mathcal{C}_t$, and $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t)$ maximises $\|\cdot - \cdot\|_{V_t^{-1}}$ over $\mathcal{C}_t \times \mathcal{C}_t$ by (7), so $\|\mathbf{x}_t^\star - \mathbf{y}\|_{V_t^{-1}} \leq w_t$. Lemma A.10 then gives $f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq 2\beta_t w_t$. Moreover $f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq B$ by Assumption 2. Hence $f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq \min\{B, 2\beta_t w_t\} = c_t$. Applying this to $\mathbf{y} = \tilde{\mathbf{x}}_t$ and $\mathbf{y} = \tilde{\mathbf{y}}_t$, both of which lie in $\mathcal{C}_t$, and averaging the two inequalities gives $R_t \leq c_t$. □

Lemma 3.3 (restated). *If a pair $(\mathbf{x}, \mathbf{y})$ satisfies $f(\mathbf{x}_t^\star) - \frac{1}{2}(f(\mathbf{x}) + f(\mathbf{y})) \leq c$ for some $c \geq 0$, then $\max\{s_t(\mathbf{x}), s_t(\mathbf{y})\} \geq \sigma(-c) \geq \frac{1}{2} - \frac{c}{4}$.*

Proof. Write $g_{\mathbf{x}} = f(\mathbf{x}_t^\star) - f(\mathbf{x}) \geq 0$ and $g_{\mathbf{y}} = f(\mathbf{x}_t^\star) - f(\mathbf{y}) \geq 0$, so that $s_t(\mathbf{x}) = \sigma(-g_{\mathbf{x}})$ and $s_t(\mathbf{y}) = \sigma(-g_{\mathbf{y}})$. Since σ is increasing,

$$\max\{s_t(\mathbf{x}), s_t(\mathbf{y})\} = \sigma\big(-\min\{g_{\mathbf{x}}, g_{\mathbf{y}}\}\big), \qquad \min\{g_{\mathbf{x}}, g_{\mathbf{y}}\} \leq \tfrac{1}{2}(g_{\mathbf{x}} + g_{\mathbf{y}}) \leq c,$$

which gives the first inequality. For the second, $\frac{1}{2} - \sigma(-c) = \sigma(0) - \sigma(-c) = \int_{-c}^{0} \sigma'(u)\,\mathrm{d}u \leq c/4$ because $\sigma' \leq \frac{1}{4}$ everywhere. □

Width transfer (9). *For every round t, $w_t \leq 2\|\mathbf{x}_{b_t} - \mathbf{y}_t^{\mathrm{fb}}\|_{V_t^{-1}}$.*

Proof. $V_t \succ 0$, so $\|\cdot\|_{V_t^{-1}}$ is a norm. Since $\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t \in \mathcal{C}_t$,

$$w_t = \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{y}}_t\|_{V_t^{-1}} \leq \|\tilde{\mathbf{x}}_t - \mathbf{x}_{b_t}\|_{V_t^{-1}} + \|\mathbf{x}_{b_t} - \tilde{\mathbf{y}}_t\|_{V_t^{-1}} \leq 2 \max_{\mathbf{y} \in \mathcal{C}_t} \|\mathbf{x}_{b_t} - \mathbf{y}\|_{V_t^{-1}} = 2\|\mathbf{x}_{b_t} - \mathbf{y}_t^{\mathrm{fb}}\|_{V_t^{-1}},$$

the last step by the definition (8) of $\mathbf{y}_t^{\mathrm{fb}}$. Note that $\mathbf{x}_{b_t} \in \mathcal{C}_t$ is not required. □

A.5 Elliptical potential and the certificate envelope

Lemma A.11 (Matrix determinant recursion) *If $W' = W + zz^\top$ with $W \succ 0$, then* $\det W' = \det W \left(1 + \|z\|^2_{W^{-1}}\right)$.

Proof. Factoring $W^{1/2}$ from both sides, $\det(W + zz^\top) = \det(W)\det(I_d + ww^\top)$ with $w = W^{-1/2}z$. The rank-one matrix $ww^\top$ has eigenvalues $\|w\|_2^2, 0, \ldots, 0$, so $\det(I_d + ww^\top) = 1 + \|w\|_2^2 = 1 + z^\top W^{-1} z$. □

Lemma A.12 (Elliptical potential) *Let $W_1 = \lambda I_d$ and $W_{i+1} = W_i + z_i z_i^\top$ with $\|z_i\|_2 \le 2L$. Then for every $m \ge 1$,*

$$\sum_{i=1}^{m} \min\left\{1, \|z_i\|^2_{W_i^{-1}}\right\} \;\le\; 2d \log\Big(1 + \frac{4L^2 m}{\lambda d}\Big).$$

Proof. First, $\min\{1, x\} \le 2\log(1+x)$ for every $x \ge 0$: if $0 \le x \le 1$ then $\log(1+x) \ge x - x^2/2 \ge x/2$; if $x \ge 1$ then $2\log(1+x) \ge 2\log 2 > 1$. Writing $\omega_i^2 = \|z_i\|^2_{W_i^{-1}}$ and applying Lemma A.11,

$$\sum_{i=1}^{m} \min\{1, \omega_i^2\} \le 2\sum_{i=1}^{m} \log(1 + \omega_i^2) = 2\log \frac{\det W_{m+1}}{\det W_1}.$$

Lemma A.6 applied with $m+1$ in place of m bounds the right side by $2d\log(1 + 4L^2 m/(\lambda d))$. □

Lemma A.13 (Pointwise certificate bound) *Let $\bar\beta_t = \max\{B/2, \beta_t\}$. For every round t of Algorithm 1,*

$$c_t \;\le\; 4\bar\beta_t \min\{1, v_t\}, \qquad v_t = \|\mathbf{z}_t\|_{V_t^{-1}}.$$

The same bound holds for the certificate c_τ of a rejected proposal with v_τ replaced by $w_\tau = \|\tilde{\mathbf{x}}_\tau - \tilde{\mathbf{y}}_\tau\|_{V_\tau^{-1}}$.

Proof. Recall $c_t = \min\{B, 2\beta_t w_t\}$ and $2\bar\beta_t \ge B$, $\bar\beta_t \ge \beta_t$. We distinguish the two round types.

Candidate round. Here $\mathbf{z}_t = \tilde{\mathbf{x}}_t - \tilde{\mathbf{y}}_t$, so $v_t = w_t$. If $v_t \le 1$ then $c_t \le 2\beta_t v_t \le 2\bar\beta_t v_t \le 4\bar\beta_t \min\{1, v_t\}$. If $v_t > 1$ then $c_t \le B \le 2\bar\beta_t \le 4\bar\beta_t = 4\bar\beta_t \min\{1, v_t\}$.

Fallback round. Here $\mathbf{z}_t = \mathbf{x}_{b_t} - \mathbf{y}_t^{\mathrm{fb}}$, and (9) gives $w_t \le 2v_t$, so $c_t \le \min\{B, 4\beta_t v_t\}$. If $v_t \le 1$ then $c_t \le 4\beta_t v_t \le 4\bar\beta_t \min\{1, v_t\}$; if $v_t > 1$ then $c_t \le B \le 2\bar\beta_t \le 4\bar\beta_t$.

The final claim is the candidate-round computation with w_τ in place of v_τ, which does not use that the pair is displayed. □

Lemma A.14 (Cumulative certificate envelope) *On any trajectory of Algorithm 1, for every t,*

$$\sum_{i=1}^{t} c_i \;\le\; G(t). \tag{22}$$

Moreover, if the proposal at round τ is rejected, then

$$\sum_{i=1}^{\tau-1} c_i + c_\tau \;\le\; G(\tau). \tag{23}$$

Proof. Since $t \mapsto \beta_t$ is nondecreasing, so is $t \mapsto \bar\beta_t$, and hence $\bar\beta_i \le \bar\beta_t$ for $i \le t$. By Lemma A.13 and Cauchy–Schwarz,

$$\sum_{i=1}^{t} c_i \;\le\; 4\bar\beta_t \sum_{i=1}^{t} \min\{1, v_i\} \;\le\; 4\bar\beta_t \sqrt{t \sum_{i=1}^{t} \min\{1, v_i\}^2} \;=\; 4\bar\beta_t \sqrt{t \sum_{i=1}^{t} \min\{1, v_i^2\}},$$

using $\min\{1, v\}^2 = \min\{1, v^2\}$ for $v \geq 0$. Applying Lemma A.12 with $z_i = \mathbf{z}_i$, which satisfies $\|\mathbf{z}_i\|_2 \leq 2L$ by Assumption 4, bounds the inner sum by $2d \log(1 + 4L^2 t/(\lambda d))$, giving (22) by the definition (10) of G.

For (23), apply the same computation to the length-τ vector sequence $\mathbf{z}_1, \ldots, \mathbf{z}_{\tau-1}, \tilde{\mathbf{z}}_\tau$ with $\tilde{\mathbf{z}}_\tau := \tilde{\mathbf{x}}_\tau - \tilde{\mathbf{y}}_\tau$. Its associated matrices are $V_1, \ldots, V_\tau$, which are exactly those the algorithm formed, since the first $\tau - 1$ vectors are the displayed ones; the τ-th width in this sequence is $\|\tilde{\mathbf{z}}_\tau\|_{V_\tau^{-1}} = w_\tau$, and $\|\tilde{\mathbf{z}}_\tau\|_2 \leq 2L$. Lemma A.13 bounds the first $\tau - 1$ certificates and, by its final claim, also c_τ; Lemma A.12 with $m = \tau$ needs only $V_1, \ldots, V_\tau$ together with $\|\tilde{\mathbf{z}}_\tau\|_2 \leq 2L$. Hence the argument applies without $\tilde{\mathbf{z}}_\tau$ ever being displayed or added to V. □

A.6 Safety, fallback count, and regret

Let $\mathcal{A}_t, \mathcal{B}_t$ be the candidate and fallback rounds up to t, with $M_t = |\mathcal{A}_t|$ and $N_t = |\mathcal{B}_t|$, so $M_t + N_t = t$. Iterating the update in Algorithm 1,

$$\Lambda_t = \sum_{i \in \mathcal{A}_t} \Big[\tfrac{1}{2} - \tfrac{c_i}{4} - (1-\alpha) s_i(\mathbf{x}_{b_i})\Big] + \sum_{i \in \mathcal{B}_t} \alpha\, s_i(\mathbf{x}_{b_i}), \tag{24}$$

and we write the true cumulative surplus as

$$\Sigma_t = \sum_{i=1}^{t} \Big[\max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} - (1-\alpha) s_i(\mathbf{x}_{b_i})\Big]. \tag{25}$$

Lemma A.15 (Budget invariant) *On $\mathcal{E}$, $\Sigma_t \geq \Lambda_t \geq 0$ for every $t \geq 0$.*

Proof. Induction on t; both quantities vanish at $t = 0$. Suppose $\Sigma_{t-1} \geq \Lambda_{t-1} \geq 0$.

Candidate round. On $\mathcal{E}$, Lemma 3.2 gives $R_t \leq c_t$ for the displayed pair, so Lemma 3.3 with $c = c_t$ yields $\max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} \geq \frac{1}{2} - \frac{c_t}{4}$. Hence the true one-round increment of (25) is at least $\frac{1}{2} - \frac{c_t}{4} - (1-\alpha) s_t(\mathbf{x}_{b_t})$, which is exactly the budget increment, and $\Sigma_t \geq \Lambda_t$. Nonnegativity of Λ_t is precisely the acceptance test.

Fallback round. The displayed pair is $(\mathbf{x}_{b_t}, \mathbf{y}_t^{\mathrm{fb}})$, so $\max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} \geq s_t(\mathbf{x}_{b_t})$ and the true increment is at least $s_t(\mathbf{x}_{b_t}) - (1-\alpha) s_t(\mathbf{x}_{b_t}) = \alpha s_t(\mathbf{x}_{b_t})$, again the budget increment; so $\Sigma_t \geq \Lambda_t$. Since $s_t(\mathbf{x}_{b_t}) > 0$ and $\Lambda_{t-1} \geq 0$, we get $\Lambda_t \geq 0$. □

Lemma A.16 (Fallback rounds are early) *On $\mathcal{E}$, every fallback round τ satisfies $4\alpha \underline{s}\, \tau < G(\tau)$. Consequently $N_T \leq \min\{T, N_{\mathrm{cons}}\}$ for every $T \geq 1$, with N_{cons} as in (10).*

Proof. Let τ be a fallback round, so the acceptance test failed:

$$\Lambda_{\tau-1} + \tfrac{1}{2} - \tfrac{c_\tau}{4} - (1-\alpha) s_\tau(\mathbf{x}_{b_\tau}) < 0.$$

Expanding $\Lambda_{\tau-1}$ by (24) and rearranging,

$$\sum_{i \in \mathcal{A}_{\tau-1}} \Big[\tfrac{1}{2} - (1-\alpha) s_i(\mathbf{x}_{b_i})\Big] + \sum_{i \in \mathcal{B}_{\tau-1}} \alpha s_i(\mathbf{x}_{b_i}) + \Big[\tfrac{1}{2} - (1-\alpha) s_\tau(\mathbf{x}_{b_\tau})\Big] \;<\; \frac{1}{4}\Big(\sum_{i \in \mathcal{A}_{\tau-1}} c_i + c_\tau\Big).$$

By Lemma A.1 the first bracket is at least $\frac{\alpha}{2}$ for each i, the second summand is at least $\alpha \underline{s}$ for each i, and the third bracket is at least $\frac{\alpha}{2}$. On the right, certificates are nonnegative, so $\sum_{i \in \mathcal{A}_{\tau-1}} c_i + c_\tau \leq \sum_{i=1}^{\tau-1} c_i + c_\tau \leq G(\tau)$ by (23). Therefore

$$\tfrac{\alpha}{2} M_{\tau-1} + \alpha \underline{s}\, N_{\tau-1} + \tfrac{\alpha}{2} \;<\; \tfrac{1}{4} G(\tau).$$

Since $\underline{s} \leq \frac{1}{2}$ we have $\frac{\alpha}{2} \geq \alpha \underline{s}$, so the left side is at least $\alpha \underline{s}\,(M_{\tau-1} + N_{\tau-1} + 1) = \alpha \underline{s}\, \tau$. Hence $4\alpha \underline{s}\, \tau < G(\tau)$.

Every fallback time therefore belongs to $\{t \geq 1 : 4\alpha \underline{s}\, t < G(t)\}$ and so is at most the supremum N_{cons} of that set. Fallback times are distinct positive integers in $\{1, \ldots, N_{\mathrm{cons}}\}$, giving $N_T \leq N_{\mathrm{cons}}$; and $N_T \leq T$ trivially. □

Lemma A.17 (Finiteness of N_{cons}) *$G(t) = o(t)$, and consequently $N_{\text{cons}} < \infty$.*

Proof. From (5), $\beta_t = O\big(\sqrt{d \log t + \log(1/\delta)}/\kappa + \sqrt{\lambda} S/\kappa\big)$, so $\bar{\beta}_t = O(\sqrt{d \log(et)})$ for fixed $d, L, S, \lambda, \delta, \kappa$, and hence

$$G(t) = 4\bar{\beta}_t \sqrt{2dt \log\Big(1 + \frac{4L^2 t}{\lambda d}\Big)} = O\big(\sqrt{t}\, \log(et)\big) = o(t).$$

Thus there is t_0 with $G(t) < 4\alpha \underline{s}\, t$ for all $t \geq t_0$, so the set in (10) is contained in $\{1, \ldots, t_0 - 1\}$ and its supremum is finite. Note $\underline{s} \geq \sigma(-B) > 0$ by Assumption 2 and $\Delta_h \leq B$, so the threshold is strictly positive on every instance. □

Theorem 3.1 (restated). *With probability at least $1 - \delta$, simultaneously for every $T \geq 1$: the displayed pairs are α-conservative in the sense of (3); $N_T \leq \min\{T, N_{\text{cons}}\}$; and $\texttt{Reg}(T) \leq G(T) + \frac{1}{2}\Delta_h N_{\text{cons}}$.*

Proof. By Lemma 3.1, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$; all three claims are established on $\mathcal{E}$, hence hold simultaneously on a single event.

Safety. By Lemma A.15, $\Sigma_T \geq 0$ for every T, and expanding (25) this is exactly $\sum_{t \leq T} \max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} \geq (1 - \alpha) \sum_{t \leq T} s_t(\mathbf{x}_{b_t})$, i.e. (3). Since the invariant holds at every t, the guarantee is anytime.

Fallback count. Lemma A.16, with $N_{\text{cons}} < \infty$ by Lemma A.17.

Regret. Split by round type. On a candidate round, $R_t \leq c_t$ by Lemma 3.2. On a fallback round the displayed pair is $(\mathbf{x}_{b_t}, \mathbf{y}_t^{\text{fb}})$ with $\mathbf{y}_t^{\text{fb}} \in \mathcal{C}_t$, so Lemma 3.2 gives $f(\mathbf{x}_t^\star) - f(\mathbf{y}_t^{\text{fb}}) \leq c_t$, while $f(\mathbf{x}_t^\star) - f(\mathbf{x}_{b_t}) = \Delta_{t,b_t} \leq \Delta_h$ by Assumption 3; averaging,

$$R_t = \tfrac{1}{2}\Big[\big(f(\mathbf{x}_t^\star) - f(\mathbf{x}_{b_t})\big) + \big(f(\mathbf{x}_t^\star) - f(\mathbf{y}_t^{\text{fb}})\big)\Big] \leq \tfrac{1}{2}\big(\Delta_h + c_t\big).$$

Summing and using $c_t \geq 0$,

$$\texttt{Reg}(T) \leq \sum_{t \in \mathcal{A}_T} c_t + \sum_{t \in \mathcal{B}_T} \frac{c_t}{2} + \frac{\Delta_h}{2} N_T \leq \sum_{t=1}^{T} c_t + \frac{\Delta_h}{2} N_T \leq G(T) + \frac{\Delta_h}{2} N_{\text{cons}},$$

by (22) and Lemma A.16. □

A.7 The simplified rate

Lemma A.18 *Let $H(t) := G(t)/\sqrt{t} = 4\bar{\beta}_t \sqrt{2d \log\big(1 + 4L^2 t/(\lambda d)\big)}$, which is nondecreasing. Then*

$$N_{\text{cons}} \leq \frac{H(N_{\text{cons}})^2}{16\, \alpha^2 \underline{s}^{\,2}} = \frac{2d\, \bar{\beta}^2_{N_{\text{cons}}} \log\big(1 + 4L^2 N_{\text{cons}}/(\lambda d)\big)}{\alpha^2 \underline{s}^{\,2}}, \tag{26}$$

and, suppressing logarithmic factors in $T, d, L, S, \lambda, 1/\delta$,

$$N_{\text{cons}} = \tilde{\mathcal{O}}\Big(\frac{d^2}{\kappa^2 \alpha^2 \underline{s}^{\,2}}\Big), \qquad G(T) = \tilde{\mathcal{O}}\Big(\frac{d\sqrt{T}}{\kappa}\Big), \qquad \texttt{Reg}(T) = \tilde{\mathcal{O}}\Big(\frac{d\sqrt{T}}{\kappa} + \frac{\Delta_h d^2}{\kappa^2 \alpha^2 \underline{s}^{\,2}}\Big).$$

Proof. If $N_{\text{cons}} = 0$ there is nothing to prove. Otherwise N_{cons} belongs to the set in (10), so $4\alpha \underline{s}\, N_{\text{cons}} < G(N_{\text{cons}}) = H(N_{\text{cons}})\sqrt{N_{\text{cons}}}$, whence $\sqrt{N_{\text{cons}}} < H(N_{\text{cons}})/(4\alpha \underline{s})$ and (26) follows by squaring and substituting the definition of H.

For the orders, (5) gives $\bar{\beta}_t = \tilde{\mathcal{O}}(\sqrt{d}/\kappa)$, so $H(t) = \tilde{\mathcal{O}}(d/\kappa)$ and $G(T) = H(T)\sqrt{T} = \tilde{\mathcal{O}}(d\sqrt{T}/\kappa)$. Substituting $H(N_{\text{cons}}) = \tilde{\mathcal{O}}(d/\kappa)$ into (26) gives $N_{\text{cons}} = \tilde{\mathcal{O}}\big(d^2/(\kappa^2 \alpha^2 \underline{s}^{\,2})\big)$, where the logarithmic factors involving N_{cons} are themselves $\tilde{\mathcal{O}}\big(\log(d/(\kappa\alpha\underline{s}))\big)$ by back-substitution. Combining with Theorem 3.1 gives the regret bound. The bound in Theorem 3.1 is the exact nonasymptotic statement; the display above is an order statement only. □

A.8 Equivalence of latent and preference regret

Proposition A.1 *Define the preference regret of a displayed pair as* $R_t^{\mathrm{BT}} := s_t(\mathbf{x}_t^\star) - \frac{1}{2}\big(s_t(\mathbf{x}_t) + s_t(\mathbf{y}_t)\big)$. *Then for every pair,*

$$\kappa\, R_t \;\le\; R_t^{\mathrm{BT}} \;\le\; \tfrac{1}{4} R_t.$$

Proof. With $g_{\mathbf{x}} = f(\mathbf{x}_t^\star) - f(\mathbf{x}_t)$ and $g_{\mathbf{y}} = f(\mathbf{x}_t^\star) - f(\mathbf{y}_t)$, both in $[0, B]$ by Assumption 2, we have $R_t = \frac{1}{2}(g_{\mathbf{x}} + g_{\mathbf{y}})$ and, using $s_t(\mathbf{x}_t^\star) = \sigma(0) = \frac{1}{2}$,

$$R_t^{\mathrm{BT}} = \tfrac{1}{2}\Big[\big(\tfrac{1}{2} - \sigma(-g_{\mathbf{x}})\big) + \big(\tfrac{1}{2} - \sigma(-g_{\mathbf{y}})\big)\Big].$$

For $g \in [0, B]$, $\frac{1}{2} - \sigma(-g) = \int_{-g}^{0} \sigma'(u)\mathrm{d}u$, and σ' on $[-B, 0]$ satisfies $\kappa \le \sigma' \le \frac{1}{4}$; hence $\kappa g \le \frac{1}{2} - \sigma(-g) \le \frac{g}{4}$. Apply this to $g_{\mathbf{x}}$ and $g_{\mathbf{y}}$, add, and divide by two. □

A.9 Failure of Direct CLUCB Certification under Preference Feedback

We examine a direct transfer of CLUCB's cumulative safety test (Kazerouni et al., 2017) to preference feedback. The result concerns the use of this test with confidence sets constructed from comparisons. Known baseline utilities permit alternative, baseline-centered certificates, which we discuss below.

Throughout this subsection, the learner is additionally supplied with the exact baseline utilities

$$\mu_t^0 := f(\mathbf{x}_{b_t}).$$

These are distinct from the baseline preference scores used in the main formulation. Providing this additional information makes the comparison with CLUCB explicit.

Definition A.1 (Direct cumulative CLUCB certification) *Let $\mathcal{A}_t$ and $\mathcal{B}_t$ denote the admitted candidate rounds and fallback rounds through time t, respectively, and write*

$$M_t := |\mathcal{A}_t|, \qquad N_t := |\mathcal{B}_t|, \qquad \mathbf{m}_i := \tfrac{1}{2}(\mathbf{x}_i + \mathbf{y}_i).$$

At round t, let $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t) \in \mathcal{S}_t \times \mathcal{S}_t$ be an arbitrary proposed pair, with

$$\widetilde{\mathbf{m}}_t := \tfrac{1}{2}(\tilde{\mathbf{x}}_t + \tilde{\mathbf{y}}_t).$$

Construct the confidence set

$$\mathcal{K}_t := \Big\{\theta \in \Theta : \|\theta - \hat{\theta}_t\|_{V_t} \le \beta_t\Big\}, \tag{27}$$

where $\hat{\theta}_t$, V_t, and β_t are defined by (4)–(5), using the observed comparisons. The algorithm admits the candidate only if

$$L_t := \inf_{\theta \in \mathcal{K}_t} \left\langle \theta, \sum_{i \in \mathcal{A}_{t-1}} \mathbf{m}_i + \widetilde{\mathbf{m}}_t \right\rangle, \tag{28}$$

$$L_t + \sum_{i \in \mathcal{B}_{t-1}} \mu_i^0 \ge (1 - \alpha) \sum_{i=1}^{t} \mu_i^0. \tag{29}$$

Otherwise, it displays a pair containing $\mathbf{x}_{b_t}$ and credits μ_t^0 for that round in subsequent applications of (29). Candidate pairs and fallback partners may be selected by arbitrary nonanticipating rules.

The baseline utilities are credited separately in (29); the confidence set (27) is not intersected with the equalities $\langle \theta, \mathbf{x}_{b_i} \rangle = \mu_i^0$.

This definition retains CLUCB's use of the current confidence set to certify cumulative candidate performance, while replacing scalar-reward regression by the comparison estimator. The baseline self-duel $(\mathbf{x}_{b_t}, \mathbf{x}_{b_t})$ reproduces the usual baseline fallback and justifies crediting μ_t^0. We also allow other baseline-containing fallbacks: on the instance constructed below, their average utility is always at least μ_t^0, so the same credit is valid. On a general instance, such a credit for a non-self fallback would require a separate justification.

Proposition A.2 (Failure of the direct cumulative transfer) *Fix $\lambda > 0$ and $\alpha, \delta \in (0,1)$, and write $p := 1 - \alpha$. There is a two-dimensional instance with $K = 2$, satisfying the itemwise i.i.d. context assumption, Assumptions 1, 2, 3, and 4, and the normalization $f(\cdot) \in [0,1]$, with the following properties.*

Every possible displayed pair satisfies both the cardinal average-utility constraint and the score constraint (3). Nevertheless, every algorithm satisfying Definition A.1 obeys, on its confidence event $\mathcal{E}$,

$$N_T \geq \frac{p}{2} T \qquad \text{for every } T \geq 1. \tag{30}$$

Moreover, for every $\eta \in (0,1)$ and every fixed integer

$$T \geq \frac{32}{p^2} \log \frac{1}{\eta},$$

$$\mathbb{P}\left\{ \mathtt{Reg}(T) \geq \frac{p^3}{512} T \right\} \geq 1 - \delta - \eta. \tag{31}$$

If all fallbacks are baseline self-duels, the denominator in (31) improves from 512 to 256.

On the same instance, `MaxInP-BC` satisfies, with probability at least $1 - \delta$, simultaneously for every $T \geq 1$,

$$N_T \leq N_{\mathrm{cons}}, \qquad \mathtt{Reg}(T) \leq G(T) + \frac{p}{8} N_{\mathrm{cons}} = \tilde{\mathcal{O}}(\sqrt{T}),$$

where the rate treats the instance parameters and α, δ, λ as fixed.

Proof. The instance. Set

$$g := \frac{p}{4}, \qquad \theta^\star := \begin{pmatrix} 1/2 \\ g \end{pmatrix}, \qquad \mathcal{D} := \left\{ (1,u)^\top : u \in [0,1] \right\}.$$

Independently over rounds and items, draw

$$U_{t,1}, U_{t,2} \sim \mathrm{Unif}[0,1], \qquad \mathbf{x}_{t,i} := (1, U_{t,i})^\top.$$

Thus each item is sampled independently from the same distribution on $\mathcal{D}$, and the two items are distinct almost surely.

Let the baseline select the item with the smaller second coordinate, breaking ties arbitrarily. Since $g > 0$, the optimal item has the larger second coordinate. Consequently,

$$\begin{aligned} f(\mathbf{x}_{t,i}) &= \frac{1}{2} + g U_{t,i}, \\ \mu_t^0 &= \frac{1}{2} + g \min\{U_{t,1}, U_{t,2}\}, \\ \Delta_{t,b_t} &= g Q_t, \qquad Q_t := |U_{t,1} - U_{t,2}|. \end{aligned}$$

In particular,

$$\frac{1}{2} \leq \mu_t^0 \leq \frac{1}{2} + g, \qquad 0 \leq \Delta_{t,b_t} \leq g, \qquad \frac{1}{2} \leq f(\mathbf{x}) \leq \frac{3}{4} \quad \text{for every } \mathbf{x} \in \mathcal{D}.$$

The boundedness and linearity assumptions hold with $S = 1$, $L = \sqrt{2}$, and $B = 2LS = 2\sqrt{2}$. The baseline-gap assumption holds with $\Delta_l = 0$ and $\Delta_h = g$. Generate comparison outcomes according to the BTL model with parameter $\theta^\star$. The exact baseline utility is supplied as stipulated above; the baseline score is

$$s_t(\mathbf{x}_{b_t}) = \sigma(-g Q_t).$$

Because the baseline is a least-utility item in $\mathcal{S}_t$, every pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}_t \times \mathcal{S}_t$ satisfies

$$\frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \geq \mu_t^0, \qquad \max\{s_t(\mathbf{x}), s_t(\mathbf{y})\} \geq s_t(\mathbf{x}_{b_t}). \tag{32}$$

Summing these inequalities proves the asserted safety properties at every prefix. It also justifies the fallback credit for every possible baseline-containing pair on this instance.

The confidence set retains a lower utility offset. Every comparison vector has zero first coordinate. Hence, regardless of the algorithm's choices,

$$V_t = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda + W_t \end{pmatrix}, \qquad W_t := \sum_{i<t} (\mathbf{z}_i)_2^2 \geq 0.$$

Every possible pair average is $\mathbf{m} = (1, v)^\top$ with $v \in [0, 1]$, so

$$\|\mathbf{m}\|^2_{V_t^{-1}} = \frac{1}{\lambda} + \frac{v^2}{\lambda + W_t} \geq \frac{1}{\lambda}. \tag{33}$$

The likelihood term in (4) does not depend on the first coordinate of θ. Setting that coordinate to zero preserves feasibility in the Euclidean ball Θ and reduces the regularization penalty unless the coordinate was already zero. Therefore,

$$\hat{\theta}_t = (0, \hat{g}_t)^\top \qquad \text{for every } t.$$

Consider the alternative parameter

$$\theta^\circ := (0, g)^\top.$$

On the event $\mathcal{E}$ of Lemma 3.1,

$$\begin{aligned} \|\theta^\circ - \hat{\theta}_t\|^2_{V_t} &= (\lambda + W_t)(g - \hat{g}_t)^2 \\ &\leq \frac{\lambda}{4} + (\lambda + W_t)(g - \hat{g}_t)^2 \\ &= \|\theta^\star - \hat{\theta}_t\|^2_{V_t} \leq \beta_t^2. \end{aligned}$$

Since $\|\theta^\circ\|_2 = g < 1$, this proves $\theta^\circ \in \mathcal{K}_t$ for every t on $\mathcal{E}$. Moreover,

$$0 \leq \langle \theta^\circ, \mathbf{x} \rangle \leq g \qquad \text{for every } \mathbf{x} \in \mathcal{D}.$$

Thus the inclusion also survives intersecting $\mathcal{K}_t$ with the utility-normalization constraints $0 \leq \langle \theta, \mathbf{x} \rangle \leq 1$ for all $\mathbf{x} \in \mathcal{D}$.

For every possible pair average $\mathbf{m} = (1, v)^\top$,

$$\langle \theta^\circ, \mathbf{m} \rangle = gv \leq g.$$

Consequently, on $\mathcal{E}$,

$$L_t \leq \left\langle \theta^\circ, \sum_{i \in \mathcal{A}_{t-1}} \mathbf{m}_i + \widetilde{\mathbf{m}}_t \right\rangle \leq g(M_{t-1} + 1). \tag{34}$$

This bound holds for arbitrary proposed pairs, including optimal self-duels.

A constant fraction of rounds must be fallbacks. At an admitted round t, we have $M_t = M_{t-1} + 1$ and $N_t = N_{t-1}$. Combining (29), (34), and the bounds on μ_i^0 gives

$$gM_t + \left(\frac{1}{2} + g\right) N_t \geq \frac{p}{2} t. \tag{35}$$

This inequality also persists through a fallback round: its left side increases by $\frac{1}{2} + g$, while its right side increases by $p/2$, and

$$\frac{1}{2} + g - \frac{p}{2} = \frac{\alpha}{2} + g > 0.$$

Since both sides are zero at time zero, (35) holds for every t on $\mathcal{E}$.

Substituting $M_t = t - N_t$ and $g = p/4$ yields

$$\frac{p}{4} t + \frac{1}{2} N_t \geq \frac{p}{2} t,$$

which proves (30). Lemma 3.1 gives $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

A linear number of fallbacks implies linear regret here. The baseline gaps are not uniformly bounded away from zero, so we must account for the possibility that the algorithm selects small-gap rounds for fallback.

Set $\varepsilon := p/8$ and define

$$H_T := \sum_{t=1}^{T} \mathbf{1}\{Q_t < \varepsilon\}.$$

The random variables Q_t are i.i.d., with

$$\mathbb{P}\{Q_t < \varepsilon\} = 2\varepsilon - \varepsilon^2 \leq 2\varepsilon = \frac{p}{4}.$$

Hoeffding's inequality therefore gives, for each fixed T,

$$\mathbb{P}\left\{ H_T \leq \frac{p}{4}T + \sqrt{\frac{T}{2}\log\frac{1}{\eta}} \right\} \geq 1 - \eta.$$

If $T \geq 32p^{-2}\log(1/\eta)$, the square-root term is at most $pT/8$. On this event and $\mathcal{E}$, the number of fallback rounds with $Q_t \geq \varepsilon$ is at least

$$N_T - H_T \geq \frac{p}{2}T - \frac{3p}{8}T = \frac{p}{8}T.$$

Any pair containing the baseline incurs regret at least $\Delta_{t,b_t}/2 = gQ_t/2$. Since all other rounds have nonnegative regret, we obtain

$$\mathtt{Reg}(T) \geq \frac{p}{8}T \cdot \frac{g\varepsilon}{2} = \frac{p^3}{512}T.$$

A union bound proves (31); independence between $\mathcal{E}$ and the context-counting event is not required. A baseline self-duel incurs regret gQ_t, which doubles the lower bound.

Performance of `MaxInP-BC`. The same instance satisfies every assumption of Theorem 3.1, with

$$\Delta_h = g = \frac{p}{4}, \qquad \underline{s} = \sigma(-g) > 0.$$

Applying that theorem to the trajectory of `MaxInP-BC` gives, with probability at least $1 - \delta$, simultaneously for every $T \geq 1$,

$$N_T \leq N_{\mathrm{cons}}, \qquad \mathtt{Reg}(T) \leq G(T) + \frac{g}{2}N_{\mathrm{cons}} = G(T) + \frac{p}{8}N_{\mathrm{cons}}.$$

For fixed parameters, $G(T) = O(\sqrt{T}\log(eT)) = o(T)$ and $N_{\mathrm{cons}} < \infty$ by (10). This proves the stated sublinear regret bound without requiring linear growth of W_t or pointwise convergence of c_t. □

Remark A.2 (Scope of the separation) Proposition A.2 concerns the specific certification rule in Definition A.1. The parameters $\theta^\star$ and θ° induce identical comparison probabilities, but they do not agree with the same supplied absolute baseline utilities. The latter information is not imposed on the confidence set (27).

Known baseline utilities permit the identity

$$f(\mathbf{m}_i) = \mu_i^0 + \langle \theta^\star, \mathbf{m}_i - \mathbf{x}_{b_i} \rangle.$$

Consequently, for baseline self-duel fallbacks, a valid alternative cumulative test is

$$\alpha \sum_{i=1}^{t} \mu_i^0 + \inf_{\theta \in \mathcal{K}_t} \left\langle \theta, \sum_{i \in \mathcal{A}_{t-1}} (\mathbf{m}_i - \mathbf{x}_{b_i}) + \widetilde{\mathbf{m}}_t - \mathbf{x}_{b_t} \right\rangle \geq 0. \tag{36}$$

On $\mathcal{E}$, the left side lower-bounds the cumulative cardinal surplus after admitting the candidate. The same conclusion holds for other fallbacks whose average utilities are certified to be at least their baseline utilities.

On the constructed instance, every centered vector in (36) has zero first coordinate, so the common utility component disappears. Incorporating the known-baseline equalities into the confidence set is another possible modification. The proposition therefore establishes a limitation of the direct cumulative transfer, rather than an impossibility of conservative cardinal certification under preference feedback.

Remark A.3 (Score certification through comparisons) Scores can also be certified directly from pairwise differences. On $\mathcal{E}$, for every $\mathbf{y} \in \mathcal{S}_t$,

$$s_t(\mathbf{y}) \geq \sigma\left(-\max_{\mathbf{x}\in\mathcal{S}_t}\left\{\langle\hat{\theta}_t, \mathbf{x}-\mathbf{y}\rangle + \beta_t\|\mathbf{x}-\mathbf{y}\|_{V_t^{-1}}\right\}\right). \tag{37}$$

Indeed, each expression inside the maximum upper-bounds $f(\mathbf{x}) - f(\mathbf{y})$, and

$$s_t(\mathbf{y}) = \sigma\left(-\max_{\mathbf{x}\in\mathcal{S}_t}\{f(\mathbf{x}) - f(\mathbf{y})\}\right).$$

No estimate of an absolute utility is needed.

Our construction uses the uniform optimality-gap certificate c_t from Lemma 3.2 and converts it into a score guarantee through Lemma 3.3. The candidate selection and informative fallback ensure, by (9), that

$$c_t \leq 4\bar{\beta}_t \min\{1, \|\mathbf{z}_t\|_{V_t^{-1}}\}, \qquad \bar{\beta}_t := \max\{B/2, \beta_t\}.$$

Thus its cumulative certification error is bounded by $G(T)$ along the comparisons actually observed. This connection between certification and information acquisition yields the finite fallback bound in Theorem 3.1.

A.10 Relaxing knowledge of baseline scores

The baseline constraint (3) depends only on utility differences and is therefore invariant to additive shifts of f. Indeed, for every $a \in \mathbb{R}$,

$$\sigma\big((f(\mathbf{x}) + a) - (f(\mathbf{x}_t^\star) + a)\big) = \sigma\big(f(\mathbf{x}) - f(\mathbf{x}_t^\star)\big).$$

Consequently, exact knowledge of $s_t(\mathbf{x}_{b_t})$ is an optional source of information, rather than a requirement for resolving an additive utility offset. This subsection proves that valid lower and upper bounds on the baseline score suffice.

Available baseline-score bounds. Retain Assumptions 1, 2, 3, and 4, but replace the exact-score input to Algorithm 1 by two bounds L_t, U_t. These bounds must be available before the round-t pair is selected and satisfy, on every realization,

$$\underline{s} \leq L_t \leq U_t \leq \tfrac{1}{2}, \qquad \underline{s} = \sigma(-\Delta_h) > 0. \tag{38}$$

The BTL assumption is understood conditionally on the full pre-feedback information, including any side information used to construct these bounds.

For some $\delta_b \in [0, 1)$, assume simultaneous coverage:

$$\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b, \qquad \mathcal{E}_b := \bigcap_{t\geq 1}\{L_t \leq s_t(\mathbf{x}_{b_t}) \leq U_t\}. \tag{39}$$

Deterministically valid bounds correspond to $\delta_b = 0$. No independence between $\mathcal{E}_b$ and the model-confidence event is required.

Modified safety budget. Compute $\hat{\theta}_t$, β_t, $\mathcal{C}_t$, $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t)$, w_t, and c_t using (4)–(7), and define the fallback partner by (8). Initialize $\Lambda_0 = 0$ and set

$$q_t := \frac{1}{2} - \frac{c_t}{4} - (1-\alpha)U_t. \tag{40}$$

The displayed pair and budget update are

$$(\mathbf{x}_t, \mathbf{y}_t) = \begin{cases} (\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t), & \Lambda_{t-1} + q_t \geq 0, \\ (\mathbf{x}_{b_t}, \mathbf{y}_t^{\mathrm{fb}}), & \Lambda_{t-1} + q_t < 0, \end{cases} \tag{41}$$

$$\Lambda_t = \begin{cases} \Lambda_{t-1} + q_t, & \Lambda_{t-1} + q_t \geq 0, \\ \Lambda_{t-1} + \alpha L_t, & \Lambda_{t-1} + q_t < 0. \end{cases} \tag{42}$$

After displaying the pair, observe o_t, set $\mathbf{z}_t = \mathbf{x}_t - \mathbf{y}_t$, and update the design matrix by $V_{t+1} = V_t + \mathbf{z}_t \mathbf{z}_t^\top$. The estimator continues to use all displayed comparisons.

The upper bound U_t appears in the subtracted baseline term of the candidate test, whereas the lower bound L_t determines the guaranteed credit from a fallback round. The unknown quantity $s_t(\mathbf{x}_{b_t})$ is not used by this decision rule.

Let $\mathcal{A}_T, \mathcal{B}_T$ denote the candidate and fallback rounds of this variant, and write $M_T = |\mathcal{A}_T|$ and $N_T = |\mathcal{B}_T|$. Use β_t from (5), with model-confidence parameter δ, and G and N_{cons} from (10).

Theorem A.1 (Guarantees with baseline-score bounds) *Suppose the assumptions above hold, with $\alpha, \delta \in (0,1)$, $\lambda > 0$, and $\delta + \delta_b < 1$. Run the variant (40)– (42). With probability at least $1 - \delta - \delta_b$, simultaneously for every $T \geq 1$,*

$$\sum_{t=1}^{T} \max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} \geq (1-\alpha) \sum_{t=1}^{T} s_t(\mathbf{x}_{b_t}), \tag{43}$$

$$N_T \leq \min\{T, N_{\mathrm{cons}}\}, \tag{44}$$

$$\mathtt{Reg}(T) \leq G(T) + \frac{\Delta_h}{2} N_{\mathrm{cons}}. \tag{45}$$

In particular, deterministically valid bounds preserve the exact guarantees and confidence level of Theorem 3.1.

Proof. We first establish the relevant model-confidence event. The proof of Lemma 3.1 applies to any adaptively selected comparison vectors that are measurable before their feedback is observed. The modified safety test preserves this property. Hence

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta, \qquad \mathcal{E} := \left\{ \|\hat{\theta}_t - \theta^\star\|_{V_t} \leq \beta_t \text{ for every } t \geq 1 \right\}.$$

The union bound gives

$$\mathbb{P}(\mathcal{E} \cap \mathcal{E}_b) \geq 1 - \delta - \delta_b. \tag{46}$$

We prove all three conclusions on this single event.

Certificates and their cumulative envelope. On $\mathcal{E}$, Lemma 3.2 gives

$$f(\mathbf{x}_t^\star) - f(\mathbf{y}) \leq c_t \qquad \text{for every } \mathbf{y} \in \mathcal{C}_t. \tag{47}$$

In particular, the candidate pair has instantaneous regret at most c_t.

Write

$$v_t := \|\mathbf{z}_t\|_{V_t^{-1}}, \qquad \bar{\beta}_t := \max\{B/2, \beta_t\},$$

where $\mathbf{z}_t$ is the comparison vector actually displayed. On a candidate round, $w_t = v_t$. On a fallback round, the triangle inequality and the definition of $\mathbf{y}_t^{\mathrm{fb}}$ give

$$\begin{aligned} w_t &= \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{y}}_t\|_{V_t^{-1}} \\ &\leq \|\tilde{\mathbf{x}}_t - \mathbf{x}_{b_t}\|_{V_t^{-1}} + \|\mathbf{x}_{b_t} - \tilde{\mathbf{y}}_t\|_{V_t^{-1}} \\ &\leq 2\|\mathbf{x}_{b_t} - \mathbf{y}_t^{\mathrm{fb}}\|_{V_t^{-1}} = 2v_t. \end{aligned}$$

Thus, on either type of round,

$$c_t \leq \min\{B, 4\beta_t v_t\} \leq 4\bar{\beta}_t \min\{1, v_t\}. \tag{48}$$

For the last inequality, if $v_t \leq 1$, use $c_t \leq 4\beta_t v_t$. If $v_t > 1$, use $c_t \leq B \leq 2\bar{\beta}_t$.

The matrix determinant lemma gives

$$\det(V_{t+1}) = \det(V_t)(1 + v_t^2).$$

Since $\min\{1, u\} \le 2\log(1+u)$ for every $u \ge 0$,

$$\begin{aligned}\sum_{t=1}^{T} \min\{1, v_t^2\} &\le 2\sum_{t=1}^{T} \log(1 + v_t^2) \\ &= 2\log\frac{\det(V_{T+1})}{\det(V_1)} \\ &\le 2d\log\left(1 + \frac{4L^2T}{\lambda d}\right). \end{aligned} \tag{49}$$

The final inequality follows from the arithmetic–geometric mean inequality applied to the eigenvalues of V_{T+1}, together with

$$\operatorname{tr}(V_{T+1}) = \lambda d + \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 \le \lambda d + 4L^2T.$$

Because $\bar{\beta}_t$ is nondecreasing, Cauchy–Schwarz and $\min\{1, v\}^2 = \min\{1, v^2\}$ yield

$$\begin{aligned}\sum_{t=1}^{T} c_t &\le 4\bar{\beta}_T \sum_{t=1}^{T} \min\{1, v_t\} \\ &\le 4\bar{\beta}_T \sqrt{T\sum_{t=1}^{T} \min\{1, v_t^2\}} \\ &\le 4\bar{\beta}_T \sqrt{2dT\log\left(1 + \frac{4L^2T}{\lambda d}\right)} = G(T). \end{aligned} \tag{50}$$

This bound includes c_t on fallback rounds, even though it is the certificate of the rejected candidate. The width-transfer inequality controls that certificate through the comparison vector actually displayed.

Safety. Define the true cumulative surplus by

$$\Sigma_t := \sum_{i=1}^{t} \left[\max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} - (1-\alpha)s_i(\mathbf{x}_{b_i})\right], \qquad \Sigma_0 = 0. \tag{51}$$

We show that

$$\Sigma_t \ge \Lambda_t \ge 0 \qquad \text{for every } t \ge 0. \tag{52}$$

On a candidate round, Lemmas 3.2 and 3.3 imply

$$\max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} \ge \frac{1}{2} - \frac{c_t}{4}.$$

Since $s_t(\mathbf{x}_{b_t}) \le U_t$ on $\mathcal{E}_b$,

$$\begin{aligned}\Sigma_t - \Sigma_{t-1} &\ge \frac{1}{2} - \frac{c_t}{4} - (1-\alpha)s_t(\mathbf{x}_{b_t}) \\ &\ge \frac{1}{2} - \frac{c_t}{4} - (1-\alpha)U_t \\ &= q_t = \Lambda_t - \Lambda_{t-1}. \end{aligned}$$

The acceptance test ensures $\Lambda_t \ge 0$.

On a fallback round, the displayed pair contains $\mathbf{x}_{b_t}$, so

$$\begin{aligned}\Sigma_t - \Sigma_{t-1} &\ge s_t(\mathbf{x}_{b_t}) - (1-\alpha)s_t(\mathbf{x}_{b_t}) \\ &= \alpha s_t(\mathbf{x}_{b_t}) \\ &\ge \alpha L_t = \Lambda_t - \Lambda_{t-1}. \end{aligned}$$

Moreover, $L_t \ge \underline{s} > 0$, so a fallback preserves nonnegativity of the budget. Starting from $\Sigma_0 = \Lambda_0 = 0$, induction proves (52). Expanding $\Sigma_T \ge 0$ gives (43).

Fallback count. Expanding the modified budget gives

$$\Lambda_t = \sum_{i\in\mathcal{A}_t}\left[\frac{1}{2} - \frac{c_i}{4} - (1-\alpha)U_i\right] + \sum_{i\in\mathcal{B}_t}\alpha L_i. \tag{53}$$

Let τ be a fallback round. By the rejection condition,

$$\Lambda_{\tau-1} + \frac{1}{2} - \frac{c_\tau}{4} - (1-\alpha)U_\tau < 0.$$

Substituting (53) and rearranging yields

$$\begin{aligned}\sum_{i\in\mathcal{A}_{\tau-1}}\left[\frac{1}{2} - (1-\alpha)U_i\right] + \sum_{i\in\mathcal{B}_{\tau-1}}\alpha L_i + \frac{1}{2} - (1-\alpha)U_\tau \\ < \frac{1}{4}\left(\sum_{i\in\mathcal{A}_{\tau-1}} c_i + c_\tau\right).\end{aligned} \tag{54}$$

The endpoint restrictions (38) imply

$$\frac{1}{2} - (1-\alpha)U_i \geq \frac{\alpha}{2}, \qquad \alpha L_i \geq \alpha\underline{s}.$$

Since all certificates are nonnegative, (50) gives

$$\sum_{i\in\mathcal{A}_{\tau-1}} c_i + c_\tau \leq \sum_{i=1}^{\tau} c_i \leq G(\tau).$$

Therefore,

$$\frac{\alpha}{2}M_{\tau-1} + \alpha\underline{s}N_{\tau-1} + \frac{\alpha}{2} < \frac{1}{4}G(\tau). \tag{55}$$

Using $\underline{s} \leq \frac{1}{2}$ and $M_{\tau-1} + N_{\tau-1} = \tau - 1$, the left-hand side is at least $\alpha\underline{s}\tau$. Hence every fallback time satisfies

$$4\alpha\underline{s}\tau < G(\tau).$$

By the definition of N_{cons} in (10), every fallback time is at most N_{cons}. Because fallback times are distinct positive integers,

$$N_T \leq \min\{T, N_{\text{cons}}\}.$$

For completeness, N_{cons} is finite. For fixed problem parameters and δ, (5) gives $\bar{\beta}_t = O(\sqrt{\log(et)})$, and consequently

$$G(t) = O\big(\sqrt{t}\log(et)\big) = o(t).$$

Since $\alpha\underline{s} > 0$, the inequality $4\alpha\underline{s}t < G(t)$ can hold only for bounded t.

Regret. On a candidate round, Lemma 3.2 gives $R_t \leq c_t$. On a fallback round, $\mathbf{y}_t^{\text{fb}} \in \mathcal{C}_t$, so (47) and Assumption 3 give

$$\begin{aligned}R_t &= \frac{1}{2}\left[f(\mathbf{x}_t^\star) - f(\mathbf{x}_{b_t}) + f(\mathbf{x}_t^\star) - f(\mathbf{y}_t^{\text{fb}})\right] \\ &\leq \frac{1}{2}(\Delta_h + c_t).\end{aligned}$$

It follows that

$$\begin{aligned}\texttt{Reg}(T) &\leq \sum_{t\in\mathcal{A}_T} c_t + \frac{1}{2}\sum_{t\in\mathcal{B}_T} c_t + \frac{\Delta_h}{2}N_T \\ &\leq \sum_{t=1}^{T} c_t + \frac{\Delta_h}{2}N_T \\ &\leq G(T) + \frac{\Delta_h}{2}N_{\text{cons}}.\end{aligned}$$

All conclusions hold simultaneously for every T on $\mathcal{E}\cap\mathcal{E}_b$. Equation (46) therefore proves the stated probability guarantee. □

A variant requiring no baseline-score estimates. Assumption 3 and the definition of the preference score give the deterministic bounds

$$\underline{s} \le s_t(\mathbf{x}_{b_t}) \le \tfrac{1}{2}.$$

Thus one may set

$$L_t = \underline{s}, \qquad U_t = \tfrac{1}{2} \qquad \text{for every } t. \tag{56}$$

In this case, the candidate is accepted precisely when

$$\Lambda_{t-1} + \frac{\alpha}{2} - \frac{c_t}{4} \ge 0,$$

and the budget increments are

$$\Lambda_t - \Lambda_{t-1} = \begin{cases} \dfrac{\alpha}{2} - \dfrac{c_t}{4}, & t \in \mathcal{A}_T, \\ \alpha \underline{s}, & t \in \mathcal{B}_T, \end{cases} \qquad t \le T. \tag{57}$$

Here $\mathcal{E}_b$ holds surely, so $\delta_b = 0$ and Theorem A.1 gives the same exact bounds and confidence level as Theorem 3.1. The learner needs the baseline item and the known gap bound Δ_h, but no exact or estimated baseline score.

If the lower gap bound Δ_l is available, the tighter deterministic upper bound $U_t = \sigma(-\Delta_l)$ may instead be used. It satisfies $U_t \le \frac{1}{2}$ and therefore preserves the same guarantees.

Statistical bounds and confidence accounting. Statistical intervals must cover the actual context-dependent scores $s_t(\mathbf{x}_{b_t})$ simultaneously over time. A confidence interval for a marginal average win rate does not by itself establish (39).

For example, suppose a procedure produces intervals $[\widetilde{L}_t, \widetilde{U}_t]$ before the round-t decision and guarantees, under the adaptive interaction protocol,

$$\mathbb{P}\left(s_t(\mathbf{x}_{b_t}) \notin [\widetilde{L}_t, \widetilde{U}_t]\right) \le \delta_{b,t}, \qquad \sum_{t=1}^{\infty} \delta_{b,t} \le \delta_b.$$

The union bound then gives simultaneous coverage with probability at least $1 - \delta_b$. One possible allocation is

$$\delta_{b,t} = \frac{6\delta_b}{\pi^2 t^2}.$$

Alternatively, a procedure may provide a simultaneous confidence sequence directly.

To enforce (38), intersect each supplied interval with $[\underline{s}, \frac{1}{2}]$. If the intersection is empty, use $[\underline{s}, \frac{1}{2}]$ instead. Whenever the supplied interval contains the true score, its intersection with this known range is nonempty and still contains that score. Thus this operation preserves the simultaneous coverage guarantee while ensuring that the algorithm is well-defined on every realization.

The model-confidence and baseline-coverage failure probabilities add:

$$\mathbb{P}(\mathcal{E} \cap \mathcal{E}_b) \ge 1 - \delta - \delta_b.$$

For a desired total failure probability $\delta_{\mathrm{tot}} \in (0, 1)$, one may use $\delta = \delta_b = \delta_{\mathrm{tot}}/2$. In that case, β_t, G, and N_{cons} are evaluated using the model-confidence parameter $\delta_{\mathrm{tot}}/2$. For the deterministic choice (56), no split is necessary: take $\delta = \delta_{\mathrm{tot}}$ and $\delta_b = 0$.

B Analysis for General Utility Functions

This appendix proves Theorem 4.1, constructs a finite-class oracle, and treats baseline-score bounds. Throughout, $B > 0$, $K \ge 2$, and $\kappa = \sigma(B)(1 - \sigma(B))$. All contexts and baseline information are available before pair sampling, and the BTL law is conditioned on all information available before the comparison outcome. The analysis allows arbitrary predictable context sets, and therefore includes the i.i.d. model of the main paper. No absolute baseline utility is used.

B.1 Adapting C-SquareCB to preference feedback

This subsection explains the changes to certification and fallback sampling relative to C-SquareCB (Deb et al., 2025). The concentration bounds and proof of Theorem 4.1 are given in the subsequent subsections.

The quantity controlled by the oracle changes. C-SquareCB observes noisy cardinal costs and uses a square-loss oracle to calibrate cumulative predictions of those costs; its safety test adds a correction for their cumulative estimation error. In our setting, the conditional mean of the label is $\sigma(f(\mathbf{x}_t) - f(\mathbf{y}_t))$. Accordingly, the logistic oracle controls the margin error

$$\varepsilon_t(\mathbf{x}_t, \mathbf{y}_t) = \big[\hat{f}_t(\mathbf{x}_t) - \hat{f}_t(\mathbf{y}_t)\big] - \big[f(\mathbf{x}_t) - f(\mathbf{y}_t)\big], \tag{58}$$

as established in Appendix B.3. This does not control individual utility errors: adding the same constant to every prediction leaves both the logistic loss and (58) unchanged. Thus C-SquareCB's cardinal calibration cannot be invoked with these predictions. Scores themselves are shift invariant, however, so this observation is not an impossibility of score certification.

Pairwise fit alone does not calibrate plug-in scores. A separate example shows why simply substituting estimated scores into a cardinal test also requires an additional argument. For any utility function g, define its plug-in score on a context set $\mathcal{S}$ by

$$s^g(\mathbf{x}) := \sigma\Big(g(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{S}} g(\mathbf{y})\Big).$$

Consider a realized context set containing four distinct items $\mathcal{S} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}_b, \mathbf{x}^\star\}$ and two utility functions with values

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}_b$	$\mathbf{x}^\star$
f	0	0	$\log 2$	$\log 3$
g	$\log 2$	$\log 2$	$\log 2$	$\log 3$

Both have utility range at most $B = \log 3$, the same best item, and the same baseline score $s^f(\mathbf{x}_b) = s^g(\mathbf{x}_b) = 2/5$. On the displayed pair $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, both margins are zero. Hence the predicted and true comparison probabilities coincide, the margin error vanishes, and the two logistic losses agree for either outcome. Nevertheless,

$$\max_{j\in\{1,2\}} s^f(\mathbf{x}^{(j)}) = \frac{1}{4}, \qquad \max_{j\in\{1,2\}} s^g(\mathbf{x}^{(j)}) = \frac{2}{5}. \tag{59}$$

For $\alpha = 1/4$, the required baseline fraction is $3/10$, so the plug-in score exceeds this threshold while the true pair score falls below it. Repeating this pair n times leaves the observed margin errors zero but gives a cumulative score overestimate of $3n/20$. This is an algebraic failure of calibration from observed margin errors alone, conditional on the displayed pairs; it is not a regret lower bound for an IGW algorithm or a claim about every possible comparison. Indeed, comparing $\mathbf{x}^{(1)}$ against the baseline distinguishes the models: its win probabilities are $1/3$ under f and $1/2$ under g. The example shows why a score guarantee must also account for which comparisons the learner collects.

The replacement couples certification to the sampling rule. Algorithm 2 uses $p_t \otimes p_t$ on candidate rounds and a baseline paired with one draw from p_t on fallback rounds. The branch is selected before sampling, so these are the conditional sampling laws. Let $\mathbb{E}_t$ condition on all information before pair sampling, including the branch decision, and put $v_t := \mathbb{E}_t[\varepsilon_t(\mathbf{x}_t, \mathbf{y}_t)^2]$. For r_t° defined in (72), Lemma 4.1 gives on both branches

$$\mathbb{E}_t[r_t^\circ] \le \frac{2K}{\gamma_t} + \frac{\gamma_t}{4} v_t. \tag{60}$$

On candidates, the fixed-partner inequality is averaged over an independent IGW partner; on fallbacks, the partner is fixed to the baseline. In both cases its error term is that of the comparison actually observed. The centering argument underlying this step appears in the proof of Lemma 12 of Sekhari et al. (2023); Appendix B.2 verifies the constants used here. Our construction applies it to both conservative branches and updates the oracle with both types of comparison. A fallback may still be a self-duel; informative feedback is not guaranteed on every individual round.

Cumulative regret replaces cardinal calibration. Lemma B.3 and Appendix B.5 combine (60) with the oracle guarantee to establish the envelope (77). The score inequality (82) then yields

$$\sum_{i=1}^{t}\left[\max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} - (1-\alpha)s_i^b\right] \geq Q_t - \frac{1}{4}\sum_{i\in\mathcal{A}_t} R_i, \tag{61}$$

where $\mathcal{A}_t$ contains the candidate rounds and Q_t is the credit total in (81). This explains the charge $G_{\mathrm{IGW}}(t)/4$ when considering a candidate. Including the baseline makes each fallback safe directly, while (60) also controls its sampled partner's gap. Appendix B.6 proves prefix safety, bounds the fallback count, and derives the regret guarantee. These are changes to the certificate and its connection to observed feedback; they do not assert that a regret certificate is the only possible way to enforce the score constraint.

B.2 The fixed-partner IGW inequality

We first establish an elementary IGW inequality, including its constants.

Lemma B.1 (Single-action IGW inequality) *Let $v, \hat{v} \in \mathbb{R}^K$, $a^\star \in \arg\max_a v_a$, $\hat{a} \in \arg\max_a \hat{v}_a$, and $\gamma > 0$. Define $p_a = (K + \gamma(\hat{v}_{\hat{a}} - \hat{v}_a))^{-1}$ for $a \neq \hat{a}$ and put the remaining mass on $\hat{a}$. Then p is a probability distribution, $p_{\hat{a}} \geq 1/K$, and*

$$\sum_a p_a(v_{a^\star} - v_a) \leq \frac{2K}{\gamma} + \frac{\gamma}{4}\sum_a p_a(\hat{v}_a - v_a)^2. \tag{62}$$

Proof. Every action other than $\hat{a}$ has probability in $(0, 1/K]$, so the residual probability is at least $1/K$. Let $e_a = \hat{v}_a - v_a$, $u_a = \hat{v}_{\hat{a}} - \hat{v}_a \geq 0$, $\eta = \gamma/4$, and $p_\star = p_{a^\star} > 0$. Subtracting the quadratic term in (62), and using $e - \eta e^2 \leq 1/(4\eta)$ for actions other than $a^\star$, gives

$$\begin{aligned} D &:= \sum_a p_a(v_{a^\star} - v_a) - \eta\sum_a p_a e_a^2 \\ &\leq \sum_{a\neq a^\star} p_a(\hat{v}_{a^\star} - \hat{v}_a) - (1-p_\star)e_{a^\star} - \eta p_\star e_{a^\star}^2 + \frac{1-p_\star}{4\eta}. \end{aligned} \tag{63}$$

Completing the square yields $-(1-p_\star)e - \eta p_\star e^2 \leq (1-p_\star)^2/(4\eta p_\star)$. Hence

$$D \leq \sum_a p_a u_a - u_{a^\star} + \frac{1-p_\star}{\gamma p_\star} \leq \sum_a p_a u_a - u_{a^\star} + \frac{1}{\gamma p_\star}.$$

Also,

$$\sum_a p_a u_a = \sum_{a\neq\hat{a}} \frac{u_a}{K + \gamma u_a} \leq \frac{K-1}{\gamma}.$$

If $a^\star = \hat{a}$, then $u_{a^\star} = 0$ and $p_\star \geq 1/K$. Otherwise, $p_\star^{-1} = K + \gamma u_{a^\star}$. In either case, $-u_{a^\star} + 1/(\gamma p_\star) \leq K/\gamma$. Thus $D \leq (2K-1)/\gamma \leq 2K/\gamma$, proving the claim. The argument requires no restriction on the absolute levels of v or $\hat{v}$. □

Proof of Lemma 4.1. Fix $\mathbf{u} \in \mathcal{S}_t$. Apply Lemma B.1 with $v_a = f(\mathbf{x}_{t,a}) - f(\mathbf{u})$ and $\hat{v}_a = \hat{f}_t(\mathbf{x}_{t,a}) - \hat{f}_t(\mathbf{u})$. Subtracting the fixed partner's value leaves both the estimated maximizer and every estimated gap unchanged, so the resulting distribution is exactly p_t. The true regret is $v_{a^\star} - v_a = \Delta_t(\mathbf{x}_{t,a})$, and $\hat{v}_a - v_a = \varepsilon_t(\mathbf{x}_{t,a}, \mathbf{u})$. Equation (62) gives the bound in the lemma. □

B.3 From logistic regret to predictable margin error

Fix an epoch of length n and failure budget $\delta' \in (0,1)$, and reindex its rounds by $m = 1, \ldots, n$. The oracle is fresh at the start of this epoch. All probability bounds below also hold conditional on the history before the epoch. Let $\mathbb{E}_m$ denote conditional expectation given all information immediately

before sampling the current pair, including the current context, baseline information, oracle prediction, and candidate/fallback decision. The label law, after conditioning additionally on the sampled pair, is BTL. Define

$$\Lambda = \log \frac{3n}{\delta'}, \qquad E = \frac{8\mathcal{R}_{\log}(n)}{\kappa} + \frac{5\Lambda}{\kappa^2}, \tag{64}$$

$$d_m = f(\mathbf{x}_m) - f(\mathbf{y}_m), \qquad \hat{d}_m = \hat{f}_m(\mathbf{x}_m) - \hat{f}_m(\mathbf{y}_m),$$
$$e_m = \hat{d}_m - d_m, \qquad X_m = e_m^2, \qquad v_m = \mathbb{E}_m[X_m]. \tag{65}$$

The true and predicted margins lie in $[-B, B]$, so $0 \le X_m \le 4B^2$. A useful elementary bound is

$$\kappa^{-1} = 2 + 2\cosh B \ge 4 + B^2 \ge 4B, \qquad B\kappa \le \tfrac{1}{4}. \tag{66}$$

Lemma B.2 (Uniform predictable error bound) *For any predictable comparison-sampling rule, with probability at least $1 - 2\delta'/3$, simultaneously for every $j \le n$,*

$$\sum_{m=1}^{j} v_m \le E. \tag{67}$$

Proof. Let $\Phi(d) = \log(1+e^d)$, $\pi_m = \sigma(d_m)$, and $\xi_m = e_m(o_m - \pi_m)$. The logistic loss difference decomposes as

$$\ell_m(\hat{f}_m) - \ell_m(f) = D_m - \xi_m, \qquad D_m := \Phi(\hat{d}_m) - \Phi(d_m) - \Phi'(d_m)e_m. \tag{68}$$

Since the segment between d_m and $\hat{d}_m$ lies in $[-B, B]$ and $\Phi'' = \sigma' \ge \kappa$ there,

$$D_m = e_m^2 \int_0^1 (1-u)\Phi''(d_m + ue_m)\, du \ge \frac{\kappa}{2} X_m.$$

Realizability and the prefix oracle guarantee therefore imply, for every $j \le n$,

$$\frac{\kappa}{2} \sum_{m=1}^{j} X_m \le \mathcal{R}_{\log}(n) + \sum_{m=1}^{j} \xi_m. \tag{69}$$

Label noise. Conditional on the information before the label, o_m is Bernoulli with mean π_m, and the two possible values of ξ_m lie in an interval of length $|e_m|$. Conditional Hoeffding's lemma gives

$$\mathbb{E}\left[\exp\left(a\xi_m - \frac{a^2 X_m}{8} \right) \middle| \text{information before the label} \right] \le 1 \qquad (a \in \mathbb{R}).$$

Taking a further conditional expectation before pair sampling shows that $\exp(a\sum_{m\le j}\xi_m - a^2\sum_{m\le j} X_m/8)$ is a nonnegative supermartingale starting at one. Ville's inequality with $a = 2\kappa$ implies, outside an event of probability at most $e^{-\Lambda}$, simultaneously for every $j \le n$,

$$\sum_{m=1}^{j} \xi_m \le \frac{\kappa}{4} \sum_{m=1}^{j} X_m + \frac{\Lambda}{2\kappa}.$$

Combining this with (69) gives

$$\sum_{m=1}^{j} X_m \le \frac{4\mathcal{R}_{\log}(n)}{\kappa} + \frac{2\Lambda}{\kappa^2}. \tag{70}$$

Pair sampling. Put $c = (1 - e^{-1})/(4B^2)$. Convexity of $x \mapsto e^{-x/(4B^2)}$ implies the chord bound $e^{-x/(4B^2)} \le 1 - cx$ on $[0, 4B^2]$. Thus

$$\mathbb{E}_m[e^{-X_m/(4B^2)}] \le 1 - cv_m \le e^{-cv_m}.$$

It follows that $\exp(c\sum_{m\le j} v_m - \sum_{m\le j} X_m/(4B^2))$ is another nonnegative supermartingale starting at one. A second application of Ville's inequality gives, with additional failure probability at most $e^{-\Lambda}$,

$$\sum_{m=1}^{j} v_m \le \frac{1}{1-e^{-1}} \sum_{m=1}^{j} X_m + \frac{4B^2\Lambda}{1-e^{-1}} \le 2\sum_{m=1}^{j} X_m + 7B^2\Lambda. \tag{71}$$

On the intersection of the two events, (70), (71), and (66) yield

$$\sum_{m=1}^{j} v_m \le \frac{8\mathcal{R}_{\log}(n)}{\kappa} + \frac{4\Lambda}{\kappa^2} + 7B^2\Lambda \le \frac{8\mathcal{R}_{\log}(n)}{\kappa} + \frac{5\Lambda}{\kappa^2} = E.$$

The total failure probability is at most $2e^{-\Lambda} = 2\delta'/(3n) \le 2\delta'/3$. No independence between the chosen pairs across rounds is required. □

B.4 An epoch envelope including fallback comparisons

At each round, the branch is chosen before sampling. A candidate displays two independent draws from p_m; a fallback displays the baseline and one draw from p_m. Define

$$r_m^\circ = \begin{cases} R_m, & \text{candidate,} \\ f(\mathbf{x}_m^\star) - f(\mathbf{y}_m), & \text{fallback.} \end{cases} \tag{72}$$

Thus $0 \le r_m^\circ \le B$. This quantity is used only in the analysis.

Lemma B.3 (Uniform epoch envelope) *With E, Λ from (64), use $\gamma_m = \sqrt{8Km/E}$ and define*

$$C(n,\delta') := \frac{3}{\sqrt{2}}\sqrt{KE} + B\sqrt{2\Lambda}. \tag{73}$$

For any predictable sequence of candidate/fallback decisions, with probability at least $1-\delta'$, simultaneously for every $j \le n$,

$$\sum_{m=1}^{j} r_m^\circ \le C(n,\delta')\sqrt{j}. \tag{74}$$

Proof. On a candidate round, independence and identical distribution of the two draws give $\mathbb{E}_m[R_m] = \mathbb{E}_{a\sim p_m}\Delta_m(\mathbf{x}_{m,a})$. Average Lemma 4.1 over an independent partner from p_m to obtain

$$\mathbb{E}_m[r_m^\circ] \le \frac{2K}{\gamma_m} + \frac{\gamma_m}{4} v_m.$$

On a fallback round, apply the same lemma with the baseline as the fixed partner. The squared error of the observed comparison equals the squared error with the pair reversed, so this inequality again holds, now with $r_m^\circ = \Delta_m(\mathbf{y}_m)$. The error v_m is always that of the comparison actually displayed.

On the event of Lemma B.2, monotonicity of γ_m and $\sum_{m=1}^{j} m^{-1/2} \le 2\sqrt{j}$ imply

$$\sum_{m=1}^{j} \mathbb{E}_m[r_m^\circ] \le 2K\sum_{m=1}^{j}\frac{1}{\gamma_m} + \frac{\gamma_j}{4}\sum_{m=1}^{j} v_m \le \sqrt{2KjE} + \frac{1}{\sqrt{2}}\sqrt{KjE} = \frac{3}{\sqrt{2}}\sqrt{KjE}. \tag{75}$$

The differences $r_m^\circ - \mathbb{E}_m[r_m^\circ]$ have conditional mean zero and absolute value at most B. For each j, Azuma's inequality bounds the probability that their sum exceeds $B\sqrt{2j\Lambda}$ by $e^{-\Lambda}$. A union bound over $j \le n$ gives additional failure probability at most $ne^{-\Lambda} = \delta'/3$. Combining with Lemma B.2 proves (74). □

Lemma B.4 (Envelope constants) *The constant in (73) satisfies* $C(n, \delta') \le \frac{9}{4}\sqrt{KE}$.

Proof. Since $E \ge 5\Lambda/\kappa^2$, $K \ge 2$, and $B\kappa \le 1/4$,

$$\frac{B\sqrt{2\Lambda}}{\sqrt{KE}} \le B\kappa\sqrt{\frac{2}{5K}} \le \frac{1}{4\sqrt{5}} < \frac{1}{8}.$$

Use $3/\sqrt{2} + 1/8 < 9/4$ in (73). □

B.5 Combining epochs and controlling growth

Epoch e comprises rounds $2^e, \dots, 2^{e+1} - 1$. Let $\mathcal{E}_e$ be its simultaneous event from Lemma B.3, with length $n_e = 2^e$ and failure budget $\delta_e = \delta/[(e+1)(e+2)]$. The oracle is restarted, while the credit total is preserved. The conditional guarantees of that lemma apply despite dependence on all previous epochs. Since $\sum_{e\ge 0} \delta_e = \delta$,

$$\mathbb{P}(\mathcal{E}_{\mathrm{IGW}}) \ge 1 - \delta, \qquad \mathcal{E}_{\mathrm{IGW}} := \bigcap_{e\ge 0} \mathcal{E}_e. \tag{76}$$

Summing the completed-epoch bounds and the current partial-epoch bound gives, on this event,

$$\sum_{i=1}^{t} r_i^{\circ} \le G_{\mathrm{IGW}}(t) \qquad \text{for every } t \ge 1. \tag{77}$$

This explains why fallback labels enter the oracle: their errors control their partners' gaps through the same inequality as on candidate rounds.

To make the growth rate explicit, define for $t \ge 1$

$$\mathcal{L}(t) := \log \frac{3t(1+\log_2 t)(2+\log_2 t)}{\delta}, \qquad \overline{E}(t) := \frac{8\mathcal{R}_{\log}(t)}{\kappa} + \frac{5\mathcal{L}(t)}{\kappa^2}, \tag{78}$$
$$c_G := \frac{9(2+\sqrt{2})}{4}.$$

Both E_e and C_e are nondecreasing in e. For $e = e(t)$, we have $2^e \le t$, $E_e \le \overline{E}(t)$, and

$$\begin{aligned} G_{\mathrm{IGW}}(t) &\le C_e \left(\sum_{j<e} \sqrt{2^j} + \sqrt{m(t)} \right) \\ &\le (2+\sqrt{2}) C_e \sqrt{t} \le c_G \sqrt{Kt\overline{E}(t)}. \end{aligned} \tag{79}$$

Here $\sum_{j<e} 2^{j/2} \le (1+\sqrt{2})2^{e/2}$, and the last step uses Lemma B.4. The envelope is nondecreasing, including at epoch boundaries. Since $\mathcal{R}_{\log}(t) = o(t)$ and $\mathcal{L}(t) = O(\log(t/\delta) + \log\log(t+2))$, (79) implies $G_{\mathrm{IGW}}(t) = o(t)$.

B.6 Safety, fallback count, and regret

Let $\mathcal{A}_t$ and $\mathcal{B}_t$ denote the candidate and fallback rounds through t, and let $N_t = |\mathcal{B}_t|$. For the uniform bound on the total number of fallbacks, define

$$N_{\mathrm{IGW}} := \sup\{n \in \mathbb{N}_{\ge 1} : 4\alpha \underline{s}\, n < G_{\mathrm{IGW}}(n)\}, \tag{80}$$

with $\sup \emptyset := 0$. This quantity is finite because $G_{\mathrm{IGW}}(n) = o(n)$. It depends on the problem parameters and the oracle-regret function, but not on the stopping horizon, and is not needed by the algorithm. The credit total satisfies

$$Q_t = \sum_{i\in\mathcal{A}_t} [\tfrac{1}{2} - (1-\alpha)s_i^b] + \sum_{i\in\mathcal{B}_t} \alpha s_i^b. \tag{81}$$

Proof of Theorem 4.1. Work on $\mathcal{E}_{\mathrm{IGW}}$ from (76).

Safety. Put $g_{t,\mathbf{x}} = f(\mathbf{x}_t^\star) - f(\mathbf{x}_t)$ and $g_{t,\mathbf{y}} = f(\mathbf{x}_t^\star) - f(\mathbf{y}_t)$. Because $R_t = (g_{t,\mathbf{x}} + g_{t,\mathbf{y}})/2$,

$$\max\{s_t(\mathbf{x}_t), s_t(\mathbf{y}_t)\} = \sigma(-\min\{g_{t,\mathbf{x}}, g_{t,\mathbf{y}}\}) \geq \sigma(-R_t) \geq \tfrac{1}{2} - \tfrac{1}{4}R_t, \tag{82}$$

where the final inequality follows from $\sigma' \leq 1/4$. A fallback contains the baseline, so its true surplus is at least αs_t^b. Therefore the cumulative true surplus satisfies

$$\begin{aligned} S_t &:= \sum_{i=1}^{t} \left[\max\{s_i(\mathbf{x}_i), s_i(\mathbf{y}_i)\} - (1-\alpha)s_i^b\right] \\ &\geq Q_t - \tfrac{1}{4}\sum_{i\in\mathcal{A}_t} R_i. \end{aligned} \tag{83}$$

Let $\tau_t = \max(\mathcal{A}_t \cup \{0\})$ be the last candidate time, using zero if there has been no candidate. Nonnegativity of r_i° and (77) give

$$\sum_{i\in\mathcal{A}_t} R_i \leq \sum_{i=1}^{\tau_t} r_i^\circ \leq G_{\mathrm{IGW}}(\tau_t),$$

where empty sums are zero and $G_{\mathrm{IGW}}(0) = 0$. Thus $\Lambda_t := Q_t - G_{\mathrm{IGW}}(\tau_t)/4$ is a lower certificate for S_t. Initially $\Lambda_0 = 0$. On a candidate round, $\tau_t = t$ and the acceptance test ensures $\Lambda_t \geq 0$. On a fallback round, $\tau_t = \tau_{t-1}$ and $\Lambda_t = \Lambda_{t-1} + \alpha s_t^b \geq 0$. This remains true across oracle restarts because the credits are preserved. Hence $S_t \geq 0$ at every prefix.

Fallbacks. Since $s_t^b \leq 1/2$, the candidate credit satisfies $h_t \geq \alpha/2 \geq \alpha\underline{s}$. The fallback credit satisfies $\alpha s_t^b \geq \alpha\underline{s}$. Consequently, $Q_{t-1} \geq \alpha\underline{s}(t-1)$, and rejection at time t implies

$$\alpha\underline{s}\, t \leq Q_{t-1} + h_t < G_{\mathrm{IGW}}(t)/4. \tag{84}$$

Every fallback time therefore belongs to the finite set in (80). It follows that $N_T \leq \min\{T, N_{\mathrm{IGW}}\}$ for every T, and the total number of fallbacks is finite.

For an explicit bound in terms of the oracle regret at T, suppose $N_T > 0$ and let $\tau = \max \mathcal{B}_T$ be the last fallback through time T. Combining (84) with (79) yields

$$4\alpha\underline{s}\,\tau < G_{\mathrm{IGW}}(\tau) \leq c_G\sqrt{K\tau\overline{E}(\tau)}.$$

Squaring and dividing by $\tau > 0$ gives

$$\tau < \frac{c_G^2 K\overline{E}(\tau)}{16\alpha^2\underline{s}^{\,2}} \leq \frac{c_G^2 K\overline{E}(T)}{16\alpha^2\underline{s}^{\,2}},$$

where the second inequality uses $\tau \leq T$ and monotonicity of $\overline{E}$. Since $N_T \leq \tau$, we obtain

$$N_T \leq \min\left\{T,\, N_{\mathrm{IGW}},\, \frac{c_G^2 K\overline{E}(T)}{16\alpha^2\underline{s}^{\,2}}\right\}. \tag{85}$$

The same bound is immediate when $N_T = 0$.

Regret. On candidates, $R_t = r_t^\circ$. On fallbacks,

$$R_t = \tfrac{1}{2}\left[\Delta_{t,b_t} + \Delta_t(\mathbf{y}_t)\right] \leq \tfrac{1}{2}(\Delta_h + r_t^\circ).$$

Summing and using (77) gives

$$\begin{aligned} \mathtt{Reg}(T) &\leq \sum_{t\in\mathcal{A}_T} r_t^\circ + \tfrac{1}{2}\sum_{t\in\mathcal{B}_T} r_t^\circ + \tfrac{1}{2}\Delta_h N_T \\ &\leq G_{\mathrm{IGW}}(T) + \tfrac{1}{2}\Delta_h N_T. \end{aligned} \tag{86}$$

Equations (79) and (85) therefore imply the explicit bound

$$\mathtt{Reg}(T) \leq c_G\sqrt{KT\overline{E}(T)} + \frac{c_G^2\Delta_h K\overline{E}(T)}{32\alpha^2\underline{s}^{\,2}}. \tag{87}$$

Substituting $\overline{E}(T) = 8\mathcal{R}_{\log}(T)/\kappa + 5\mathcal{L}(T)/\kappa^2$ from (78) and suppressing logarithmic factors in T and $1/\delta$ gives

$$\mathtt{Reg}(T) = \tilde{\mathcal{O}}\left(\sqrt{KT\left[\frac{\mathcal{R}_{\log}(T)}{\kappa} + \frac{1}{\kappa^2}\right]} + \frac{\Delta_h K}{\alpha^2 \underline{s}^{\,2}}\left[\frac{\mathcal{R}_{\log}(T)}{\kappa} + \frac{1}{\kappa^2}\right]\right),$$

which is (14). In particular, $\overline{E}(T) = o(T)$ implies $\mathtt{Reg}(T) = o(T)$ directly from (87). We also retain the uniform bound

$$\mathtt{Reg}(T) \le G_{\mathrm{IGW}}(T) + \tfrac{1}{2}\Delta_h N_{\mathrm{IGW}}, \tag{88}$$

which is useful when a growth bound on $\mathcal{R}_{\log}$ allows an explicit horizon-independent estimate of N_{IGW}. All statements hold simultaneously for every T on the same event of probability at least $1-\delta$. □

B.7 A FINITE-CLASS ORACLE AND EXPLICIT SPECIALIZATION

Suppose $1 \le |\mathcal{F}| < \infty$ and each function has range diameter at most B. Within each epoch initialize $w_1(g) = 1/|\mathcal{F}|$, let $\eta = e^{-B}$, and predict and update according to

$$\hat{f}_m(\mathbf{x}) = \sum_{g\in\mathcal{F}} w_m(g) g(\mathbf{x}), \tag{89}$$

$$w_{m+1}(g) = \frac{w_m(g) e^{-\eta \ell_m(g)}}{\sum_{h\in\mathcal{F}} w_m(h) e^{-\eta \ell_m(h)}}. \tag{90}$$

The common weights produce one predicted utility function, and $|\hat{f}_m(\mathbf{x}) - \hat{f}_m(\mathbf{y})| \le \sum_g w_m(g) B = B$.

Proposition B.1 (Finite-class logistic oracle) *The oracle in (89)–(90) satisfies Assumption 5 with $\mathcal{R}_{\log}(n) = e^B \log|\mathcal{F}|$ for every n.*

Proof. For $\ell(d,o) = \log(1+e^d) - od$,

$$\frac{\partial_d^2 \ell(d,o)}{(\partial_d \ell(d,o))^2} = \begin{cases} e^d, & o = 1, \\ e^{-d}, & o = 0, \end{cases} \qquad |d| \le B.$$

This ratio is at least $e^{-B} = \eta$, so $d \mapsto e^{-\eta\ell(d,o)}$ is concave on $[-B,B]$. Because the predicted margin is the weighted average of the functions' margins, Jensen's inequality yields

$$e^{-\eta \ell_m(\hat{f}_m)} \ge Z_m, \qquad Z_m := \sum_g w_m(g) e^{-\eta \ell_m(g)}.$$

The weight update gives $\prod_{m=1}^n Z_m = |\mathcal{F}|^{-1} \sum_g \exp(-\eta \sum_{m=1}^n \ell_m(g))$. Taking logarithms and comparing the sum with any one of its terms therefore gives

$$\sum_{m=1}^n \ell_m(\hat{f}_m) \le \min_{g\in\mathcal{F}} \sum_{m=1}^n \ell_m(g) + \frac{\log|\mathcal{F}|}{\eta}.$$

The bound is deterministic for every sequence and every prefix, and so permits adaptive contexts and comparisons. The range and timing requirements also hold by construction. □

For an explicit fallback bound, let

$$a_0 := \frac{c_G^2 K}{16\alpha^2 \underline{s}^{\,2}}\left[\frac{8e^B \log|\mathcal{F}|}{\kappa} + \frac{5}{\kappa^2}\log\frac{12}{\delta}\right], \tag{91}$$

$$a_1 := \frac{15 c_G^2 K}{16\alpha^2 \underline{s}^{\,2}\kappa^2}. \tag{92}$$

For $t \geq 1$, $1 + \log_2 t \leq 2t$ and $2 + \log_2 t \leq 2t$, so $\mathcal{L}(t) \leq \log(12/\delta) + 3\log t$. Combining (79) with the defining inequality for N_{IGW} shows that each integer in that set satisfies $t < a_0 + a_1 \log t$. Concavity of the logarithm gives $\log t \leq t/(2a_1) + \log(2a_1) - 1$, whence

$$N_{\mathrm{IGW}} \leq 2a_0 + 2a_1 \log\bigl(\max\{e, 2a_1\}\bigr). \tag{93}$$

Combining (93) with (88), and writing $A_{\mathcal{F}} := e^B \log|\mathcal{F}|/\kappa + 1/\kappa^2$, gives

$$\mathtt{Reg}(T) = \tilde{\mathcal{O}}\left(\sqrt{KA_{\mathcal{F}}T} + \frac{\Delta_h K A_{\mathcal{F}}}{\alpha^2 \underline{s}^2}\right). \tag{94}$$

Here the suppressed factors are logarithmic in T, $1/\delta$, K, $A_{\mathcal{F}}$, and $1/(\alpha\underline{s})$. The additive contribution uses (93) and can be bounded independently of the stopping horizon. For $B = 1$, κ is a numerical constant and $A_{\mathcal{F}} = \mathcal{O}(\log|\mathcal{F}| + 1)$. Computing p_t takes $\mathcal{O}(K)$ arithmetic after the K score evaluations; the oracle's own computational cost is separate.

B.8 Relaxing baseline-score knowledge

Suppose L_t, U_t are available before pair sampling and always satisfy $\underline{s} \leq L_t \leq U_t \leq 1/2$. Assume that, with probability at least $1 - \delta_b$, they simultaneously cover the baseline scores at every round: $L_t \leq s_t^b \leq U_t$. Replace the candidate credit and fallback credit, respectively, by

$$h_t = \tfrac{1}{2} - (1-\alpha)U_t, \qquad \text{fallback credit} = \alpha L_t. \tag{95}$$

Use these credits in the same acceptance test and updates. On the coverage event, a candidate's true surplus is at least $h_t - R_t/4$, and a fallback's true surplus is at least αL_t. Thus the same safety argument applies with these credits. Moreover, $h_t \geq \alpha/2$ and $\alpha L_t \geq \alpha\underline{s}$, so the fallback-count argument and all regret bounds are unchanged. The overall probability becomes at least $1 - \delta - \delta_b$, without an independence requirement. Deterministic bounds have $\delta_b = 0$; in particular, $(L_t, U_t) = (\underline{s}, 1/2)$ removes exact baseline-score knowledge entirely while retaining the exact guarantees of Theorem 4.1.

B.9 A known-horizon version without oracle restarts

If a horizon T is given, use one oracle instance, the parameters in (64) with $(n, \delta') = (T, \delta)$, $\gamma_t = \sqrt{8Kt/E}$, and the envelope $G_T(t) = C(T,\delta)\sqrt{t}$. The same test and sampling rules then satisfy, with probability at least $1 - \delta$, simultaneously for every $t \leq T$, the max-score constraint and

$$N_t \leq \min\left\{t, \frac{C(T,\delta)^2}{16\alpha^2\underline{s}^2}\right\}, \tag{96}$$

$$\mathtt{Reg}(t) \leq C(T,\delta)\sqrt{t} + \tfrac{1}{2}\Delta_h N_t. \tag{97}$$

Indeed, Lemma B.3 gives the required envelope on the entire run, and (84) forces every fallback time to be smaller than $C(T,\delta)^2/(16\alpha^2\underline{s}^2)$. The safety and regret arguments are otherwise identical. Lemma B.4 yields the explicit terminal bound

$$\mathtt{Reg}(T) \leq \frac{9}{4}\sqrt{KTE} + \frac{81\Delta_h KE}{512\alpha^2\underline{s}^2}. \tag{98}$$

This version uses the known horizon in its parameters and avoids oracle restarts. The epoch construction additionally admits the uniform total-fallback bound (80).