Robust Linear Dueling Bandits with Post-serving Context under Unknown Delays and Adversarial Corruptions

Youngmin Oh
InfiniTree, Republic of Korea
youngmin0.oh@gmail.com

ICML 2026

Why this problem?

Preference-based learning is the workhorse of modern AI.

- RLHF for LLMs, recommendation, ad ranking — users *compare*, not rate.
- Standard model: contextual (linear) dueling bandits [Yue & Joachims '09; Saha '21].

Three challenges that real deployments hit together:

- Post-serving context — useful side info (e.g. render time, watch time, page-load latency) is observed *after* the action.
- Delayed feedback — clicks/votes arrive late; the delay can be stochastic (typical case) *or* adversarial (rare bursty case). The learner does not know which.
- Adversarial corruption — a fraction of preference labels are arbitrarily flipped (bots, spam, adversaries).

Goal. One algorithm that is robust to *all three*, with a sublinear regret guarantee, and *agnostic* to the delay regime.

Setting in one picture

Feature. $z_{t,k} = (x_{t,k}, y_{t,k}) \in \mathbb{R}^d$ with $d = d_x + d_y$; $y_{t,k} = \phi_*(x_{t,k}) + \epsilon_{t,k}$, $\mathbb{E}[\epsilon] = 0$.

Utility (linear) & BTL preference. $u_{t,k} = \langle \Theta_*, z_{t,k} \rangle$, $\mathbb{E}[l_t \mid a_t, b_t] = g(\langle \Theta_*, z_{t,a_t} - z_{t,b_t} \rangle)$, $g =$ logistic.

Observed outcome o_t may differ from true label l_t (corruption) and is revealed at round $t + \tau_t$ (delay).

Regret. $R_T = \sum_{t=1}^{T} \frac{1}{2} \big(\langle \Theta_*, z^*_{t,k^*_t} - z^*_{t,a_t} \rangle + \langle \Theta_*, z^*_{t,k^*_t} - z^*_{t,b_t} \rangle \big)$, with $z^*_{t,k} = (x_{t,k}, \phi_*(x_{t,k}))$.

Notation — what budgets are we fighting?

Symbol	Meaning
T	time horizon (number of rounds)
K	number of arms; $d = d_x + d_y$ total feature dim
$\Theta_* \in \mathbb{R}^d$	unknown utility parameter, $\|\Theta_*\|_2 \le M$
$\phi_* : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$	ground-truth post-serving map; learned by $\widehat{\phi}_t$
$\mathcal{C}$ (corruption budget)	total label flips: $\sum_{t=1}^{T} \|l_t - \gamma_t\| \le \mathcal{C}$, where γ_t is what the adversary submits
τ_t (delay)	feedback for round t arrives at round $t + \tau_t$
Λ (adv. delay budget)	adversarial regime: $\sum_{t=1}^{T} \tau_t \le \Lambda$
μ_τ, σ^2 (stoch. delay)	stochastic regime: τ_t i.i.d. σ^2-sub-Gaussian, mean μ_τ
$\mathcal{D} = \max(\mu_\tau, \sqrt{\Lambda})$	*delay complexity* — single quantity capturing both regimes
$\mathcal{H}_t$	observed-feedback history: $\{s \le t : s + \tau_s \le t\}$
$\widetilde{V}_t, \widetilde{W}_t$	full-info vs. observed weighted design matrices (next slide)

The *regime* (stochastic vs. adversarial) is unknown to the learner; $\mathcal{C}, \Lambda, \mu_\tau$ are problem-dependent quantities used in the analysis, not inputs to the algorithm.

RCDP-UCB: algorithm

(1) Learn the post-serving map. Train $\widehat{\phi}_t$ on the buffer of seen (x, y) pairs and form $\widehat{z}_{t,k} = (x_{t,k}, \widehat{\phi}_t(x_{t,k}))$.

(2) Pick a pair (optimism in face of uncertainty).

$$a_t = \arg\max_k \langle \Theta_{t-1}, \widehat{z}_{t,k} \rangle, \qquad b_t = \arg\max_k \Big\{ \langle \Theta_{t-1}, \widehat{z}_{t,k} \rangle + c_t \, \|\Delta \widehat{z}_{t,k,a_t}\|_{\widetilde{V}_{t-1}^{-1}} \Big\}.$$

(3) Adaptive sample weight (the key knob):

$$\omega_s \;=\; \min\left(1,\; \frac{\alpha}{\|\Delta z_s\|_{\widetilde{V}_{s-1}^{-1}}}\right), \qquad \boxed{\alpha = \frac{\sqrt{d}}{\mathcal{C} + \mathcal{D}}}$$

(4) Update. Maintain $\widetilde{V}_t = \lambda I + \kappa \sum_{s \le t} \omega_s \, \Delta z_s \Delta z_s^\top$ (*all* contexts — no delay) and $\widetilde{W}_t = \lambda I + \kappa \sum_{s \in \mathcal{H}_t} \omega_s \, \Delta z_s \Delta z_s^\top$ (only arrived outcomes).
Estimate $\Theta_t = \arg\min_\Theta \big\{ \frac{\lambda}{2} \|\Theta\|^2 - \sum_{s \in \mathcal{H}_{t-1}} \omega_s \log g((-1)^{1-o_s} \langle \Theta, \Delta z_s \rangle) \big\}$.

Key idea — weight on $\widetilde{V}_t$, not $\widetilde{W}_t$

Standard delay analysis bounds $\|\Delta z_s\|_{\widetilde{W}_t^{-1}} \lesssim \|\Delta z_s\|_{\widetilde{V}_t^{-1}} + C(\tau_t)\|\Delta z_s\|^2_{\widetilde{V}_t^{-1}}$, which makes corruption bias multiply by delay $\Rightarrow$ ugly $\widetilde{\mathcal{O}}(\mathcal{C}\mathcal{D})$ cross-term.

Information asymmetry we exploit:

- Outcomes o_t are *delayed*, but contexts Δz_t are *instantaneous*.
- So we can build the weight ω_s from $\widetilde{V}_{s-1}$ (full-information geometry) at decision time.

Consequence — additive decoupling: partition past rounds into

$$[t-1] = \underbrace{\mathcal{A}_t \cap \mathcal{E}^c}_{\text{arrived \& clean}} \sqcup \underbrace{\mathcal{A}_t \cap \mathcal{E}}_{\text{arrived \& corrupted}} \sqcup \underbrace{\mathcal{A}_t^c}_{\text{pending}},$$

and the estimation error splits cleanly:

$$\|\Theta_t - \Theta_*\|_{\widetilde{V}_t} \le \underbrace{\widetilde{\mathcal{O}}(\sqrt{d})}_{\text{statistical}} + \underbrace{\alpha\mathcal{C}}_{\text{corruption}} + \underbrace{\alpha\mathcal{D}}_{\text{delay}}.$$

Choosing $\alpha = \sqrt{d}/(\mathcal{C}+\mathcal{D})$ balances them.

Regret bound

Theorem (upper bound).

Under standard regularity (link function $\dot{g} \geq \kappa$, $\|\Theta_*\| \leq M$, $\|z\| \leq 1/2$) and the learnability of ϕ_* with rate $a \in (0, 1/2]$, with $\alpha = \sqrt{d}/(\mathcal{C} + \mathcal{D})$ and $c_t = 2\beta_t$,

$$R_T = \widetilde{\mathcal{O}}\Big(\underbrace{d\sqrt{T}}_{\text{exploration}} + \underbrace{d\,\mathcal{C}}_{\text{corruption}} + \underbrace{d\,\mathcal{D}}_{\text{delay}} + \underbrace{T^{1-a} d_x^a (1 + \sqrt{d})}_{\phi_* \text{ approximation}} \Big).$$

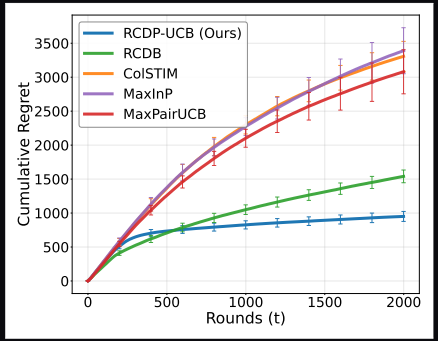
- Recall $\mathcal{D} = \max(\mu_\tau, \sqrt{\Lambda})$ — same bound, two regimes, no algorithm switch.
- Parametric ϕ_* ($a = 1/2$): the ϕ-term is $\widetilde{\mathcal{O}}(\sqrt{T})$, fully absorbed.
- Lower bound (no post-serving): $\Omega((d\sqrt{T} + d\mathcal{C} + \mathcal{D}')/\kappa)$ — matches up to $\sqrt{d}$ in the adversarial-delay term.

Setting	**Prior best**	**Ours**
Dueling, no corruption, no delay	$\widetilde{\mathcal{O}}(d\sqrt{T})$	$\widetilde{\mathcal{O}}(d\sqrt{T})$
Dueling, corruption only (RCDB)	$\widetilde{\mathcal{O}}(d\sqrt{T} + d\mathcal{C})$	$\widetilde{\mathcal{O}}(d\sqrt{T} + d\mathcal{C})$
Dueling, corruption + delay + post-serving	—	$\widetilde{\mathcal{O}}(d(\sqrt{T} + \mathcal{C} + \mathcal{D}))$

Experiments — synthetic dueling

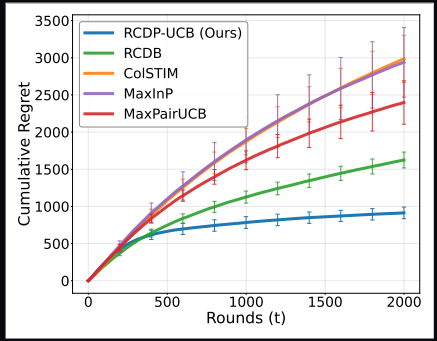
Strategic (adversarial) delay

$\Lambda = 10^4$, $\mathcal{C} = 25$, 10 runs



Stochastic delay

$\mu_\tau = 100$, $\sigma = 10$, $\mathcal{C} = 25$, 10 runs



- RCDP-UCB beats RCDB and delay-aware baselines under *both* regimes with the *same* hyperparameters.
- Ablation: swapping the $\widetilde{V}_t$-anchored clip for $\widetilde{W}_t$ destroys robustness — the *mechanism*, not $\widehat{\phi}$, drives the gain.

Takeaway

- First contextual dueling-bandit algorithm robust to post-serving contexts, unknown delays, and adversarial corruption simultaneously.
- **Mechanism.** Anchor the adaptive weight to the full-information geometry $\widetilde{V}_t$ — one clip absorbs both corruption and delay bias additively.
- **Guarantee.** $R_T = \widetilde{\mathcal{O}}\big(d(\sqrt{T} + \mathcal{C} + \mathcal{D})\big)$, $\mathcal{D} = \max(\mu_\tau, \sqrt{\Lambda})$ — regime-agnostic, nearly tight.
- **Empirically.** Consistent wins over RCDB and delay-aware baselines on synthetic and real datasets (Magic, Statlog, Spambase).

Thank you!

Youngmin Oh | InfiniTree | youngmin0.oh@gmail.com