FAIR BANDITS UNDER PRIVACY AND ADVERSARIAL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Differential privacy (DP) is widely used in contextual bandits to protect sensitive user data. In this paper, we show how an adversary can conceal fairness-targeted corruptions within DP noise while overcoming its smoothing effect on reward estimates, which otherwise makes non-smooth reward functions less sensitive to small perturbations. The attack can maximize group regret disparity under a KL budget relative to the DP noise. We analyze four methods for spending the attack budget: Variant I (diffuse) spreads the perturbation uniformly across all dimensions, Variant II (concentrated) aligns the perturbation with the group-boundary normal, Variant III (burst) concentrates the budget into a short temporal window, and Variant IV (adaptive) conditions the perturbation direction on the observed noisy context. Experiments on six bandit algorithms across three real-world datasets show that the concentrated attack causes the largest fairness damage at a modest regret cost, and DP with per-dimension monitoring protects group fairness.

1 INTRODUCTION

Contextual bandits deployed in privacy-sensitive domains can use local differential privacy (DP), where each user perturbs a context vector before transmission Dwork et al. (2006); Han et al. (2021); Tossou & Dimitrakakis (2016). When protected attributes are unavailable to a learner or auditor (because of restrictions such as GDPR and the Equal Credit Opportunity Act), group membership may instead be inferred from observed features (Chen et al., 2019; Höltgen & Oliver, 2025; Zhao et al., 2025; Deldjoo et al., 2024). We model this inference by a map applied to the received context. An adversary that alters transmitted contexts may change both the learner's decisions and the labels used to assess fairness. Privacy noise creates a security vulnerability where an adversary can hide group-targeted corruptions inside the DP noise, which raises serious concerns in domains such as healthcare, criminal justice, and credit scoring (Jang et al., 2024; Hendricks et al., 2018). On the other hand, the DP noise smooths the expected reward as a function of the clean context, which makes it provably insensitive to small perturbations when the reward function is non-smooth. An adversary therefore faces a two-sided problem: a DP noise that conceals corruption also diminishes its impact on expected rewards.

Related work with limitations. Existing bandits each cover only part of our setting. Private bandits (Zheng et al., 2020; Tossou & Dimitrakakis, 2016; Shariff & Sheffet, 2018; Han et al., 2021; Azize & Basu, 2024; Pavlovic et al., 2026) analyze the regret cost of privacy but assume no adversary. Fair bandits (Joseph et al., 2016; Patil et al., 2021; Grazzi et al., 2022; Baek & Farias, 2021) address group disparity in benign environments. Robust bandits (Lykouris et al., 2018; Gupta et al., 2019; He et al., 2022; Qi et al., 2024; Kang et al., 2023; Wang et al., 2025) defend against reward-space corruption, and adversarial attacks on bandits (Ma et al., 2018; Garcelon et al., 2020; Xu et al., 2021; Hosseini et al., 2026; Zeng et al., 2025) target aggregate regret rather than group-level harm. The intersection of privacy and adversarial corruption in bandits has been considered (Charisopoulos et al., 2023), but again in reward space and without a fairness objective.

Our threat model exploits a specific deployment scenario which is fairness under unawareness (Chen et al., 2019; Höltgen & Oliver, 2025; Zhao et al., 2025; Deldjoo et al., 2024), where the protected attribute is unavailable at prediction time and group membership must be inferred from observed features. Manipulation attacks on DP protocols themselves (Cheu et al., 2021; Cao et al., 2021) show that participants' noise can be abused by malicious ones, though the corruption sits upstream of the

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.

mechanism there rather than downstream as in our setting. Beyond bandits, adversarial fairness attacks in supervised learning (Solans et al., 2020; Mehrabi et al., 2021; Chai et al., 2025) demonstrate that static fairness is not preserved under poisoning; these methods target trained classifiers rather than an online learner under partial feedback.

Randomized smoothing certifies robustness to input perturbations by convolving a predictor with a noise density and reading a Lipschitz constant off the result in supervised learning (Cohen et al., 2019; Lecuyer et al., 2019; Pinot et al., 2019; Phan et al., 2020) and reinforcement learning (Kumar et al., 2022; Wu et al., 2022). We transfer the same mechanism to a context-level DP bandit; the certificate protects the reward channel but not the group channel that maps corrupted contexts to inferred labels. None of the papers above addresses an adversary that hides group-targeted corruption inside the DP noise of a contextual bandit, where the noise is imposed by the privacy mechanism rather than chosen for defence, the learner sees one noisy realisation per round rather than an average over injected noise, and the target is group-level disparity rather than aggregate accuracy or return.

Contributions. We formulate the problem as a budget-allocation game where the adversary hides a single perturbation inside the DP noise and chooses where to spend it to maximize fairness damage. We investigate two channels by which a perturbed context causes harm. On the *reward channel*, the perturbation moves reward estimates; for the expected reward, this effect is certifiably bounded when the reward is non-smooth. On the *group channel*, it changes the inferred group label, and no such certificate applies. Our contributions are summarized as follows:

1. We present the first adversarial fairness attack model for DP-protected contextual bandits, Group-Targeted Context Manipulation Attacks (GTCMA), that scrambles group assignments without altering rewards. We formulate it as a constrained optimization that maximizes group regret disparity subject to statistical indistinguishability from DP noise.
2. We show that a tight Lipschitz certificate for the smoothed expected reward can bound fairness damage. The certificate adds to the ordinary Lipschitz bound only for non-smooth rewards, and it transfers to the learner only if the learner's estimates track the smoothed reward.
3. We analyze four methods for spending the budget: diffuse (I), concentrated (II), burst (III), and adaptive (IV). We prove that the reward channel is robust to all variants, whereas the group channel remains vulnerable. The attack on the group boundary is more effective than distributing perturbations across all features.
4. Experiments with six bandit algorithms on three datasets show that the concentrated attack is the most damaging variant, that it increases regret far less than it increases disparity, and that it is neutralized by per-dimension detection.

2 Problem Formulation

2.1 Contextual Bandit Setting

Consider a K-armed contextual bandit problem over a horizon of T rounds. At each round $t \in [T]$, the learner observes a d-dimensional context set $\mathcal{X}_t = \{\mathbf{x}_{t,i} \in \mathbb{R}^d : i \in [K]\}$, selects an arm $a_t \in [K]$, and receives a stochastic reward

$$r_{t,a_t} = h(\mathbf{x}_{t,a_t}) + \xi_t, \tag{1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is an unknown bounded reward function ($\|h\|_\infty < \infty$) and ξ_t is zero-mean σ-sub-Gaussian noise (Valko et al., 2013; Bogunovic et al., 2022). Our analysis uses only boundedness of h and, where stated, its Lipschitz constant L_f. The learner minimizes cumulative pseudo-regret

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t})\big], \tag{2}$$

where $a_t^* = \arg\max_{i \in [K]} h(\mathbf{x}_{t,i})$ is the optimal arm at round t.

2.2 Demographic Structure

Let the function $F_{\mathrm{g}} : \mathbb{R}^d \to [G]$ map each context to one of G latent demographic groups. F_{g} depends on a subset $S \subset [d]$ of sensitive dimensions ($s = |S|$). We distinguish the *true* group $g_t = F_{\mathrm{g}}(\mathbf{x}_{t,a_t})$ from the *inferred* group $\hat{g}_t = F_{\mathrm{g}}(\hat{\mathbf{x}}_{t,a_t})$, computed on the context the learner actually receives. With $\tau_g = \{t : g_t = g\}$ and $n_g = |\tau_g|$, the group regret and the adversary's objective are

$$R_T^{(g)} = \sum_{t \in \tau_g} \mathbb{E}\big[h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t})\big], \qquad \Delta R_T^{(g)} \;=\; \frac{R_T^{(g)}}{n_g} \;-\; \frac{R_T}{T}. \tag{3}$$

A wrong inferred label causes *treatment harm* when a group-aware learner (e.g. FairBandit) conditions on $\hat{g}_t$, and *audit harm* when disparity is evaluated with inferred labels. For group-agnostic learners only the second route exists on the group channel; any change in true-label disparity must pass through the reward channel. We use merit-based group regret rather than demographic parity because a correct policy should treat groups differently when their merit differs. It also exposes a blind spot of aggregate monitoring: if $n_{g^\dagger} = o(T)$, the attacker can inflict $\Theta(1)$ per-capita regret while adding only $o(T)$ to R_T.

2.3 Context-Level Differential Privacy

Each user perturbs their own context vector via local DP (i.e, ε-DP) before transmission to the learner Dwork & Roth (2014):

$$\tilde{\mathbf{x}}_{t,i} = \mathbf{x}_{t,i} + \boldsymbol{\eta}_{t,i}, \qquad \boldsymbol{\eta}_{t,i} \sim \mathrm{Lap}^d(b), \qquad b = \Delta_1/\varepsilon, \tag{4}$$

where $\tilde{\mathbf{x}}_{t,i}$ is the protected context, $\mathrm{Lap}^d(b)$ denotes d independent Laplace draws with scale b, $\Delta_1 = \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|_1$ is the ℓ_1 sensitivity of the context vector, and $\varepsilon > 0$ is the privacy parameter. This mechanism satisfies ε-DP (Dwork & Roth, 2014): for any two contexts $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and any measurable set E, $\Pr[\tilde{\mathbf{x}} \in E \mid \mathbf{x}] \le e^{\varepsilon} \Pr[\tilde{\mathbf{x}} \in E \mid \mathbf{x}']$

2.4 Context-Space Adversary and Defender Model

An adversary corrupts the protected context before the learner updates, producing a corrupted context $\hat{\mathbf{x}}_{t,i} = \tilde{\mathbf{x}}_{t,i} + \boldsymbol{\delta}_{t,i}$. The adversary can target specific users, rounds, and feature dimensions via an intermediate relay, such as a data-collection gateway or client-side SDK. Because the learner trusts the relay and has no clean reference for the inherently random private context, it cannot separate DP noise $\boldsymbol{\eta}_{t,i}$ from corruption $\boldsymbol{\delta}_{t,i}$ Cheu et al. (2021); Cao et al. (2021). Unlike scalar reward corruption, context corruption is a *vector-valued* manipulation in $\mathbb{R}^d$, giving the adversary d degrees of freedom per round. The relay sees contexts but not rewards or model parameters (Assumption 1), so the attack must work through the group-assignment map rather than by poisoning rewards. Sending a signed group label separately would close this channel, but most systems infer group membership from features because regulation limits collecting demographic attributes directly.

Definition 1 (Group-Targeted Context Manipulation Attack — GTCMA)**.** *An adversary using GTCMA mechanism concentrates corruption on contexts belonging to group $g^\dagger$, setting $\boldsymbol{\delta}_{t,i}$ to shift sensitive attributes so that $F_{\mathrm{g}}(\hat{\mathbf{x}}_{t,i}) \neq g^\dagger$ (i.e., the corrupted context is classified into a different demographic group). GTCMA moves users across the group boundary and makes a well-performing group appear to have poor merit under a different label, violating merit-based fairness.*

Assumption 1 (Attacker capabilities)**.** *The adversary (i) observes $\{\tilde{\mathbf{x}}_{t,i}\}$ but not the learner's parameters θ_t, rewards r_{t,a_t}, or gradients; (ii) knows an estimate $\hat{\mathbf{w}}$ of the unit normal $\mathbf{w}$ of the boundary separating $g^\dagger$ from the group g' it wants users reassigned to; and (iii) does not observe or adapt to the defender's monitoring statistic.*

Defender model. The learner sees only $\hat{\mathbf{x}}_t = \mathbf{x}_t + \boldsymbol{\eta}_t + \boldsymbol{\delta}_t$. Without an attack, received contexts follow the true context distribution blurred by Laplace noise, $P_0 = P_{\mathbf{x}} * p_b$. The defender knows the noise but not $P_{\mathbf{x}}$, so it must estimate P_0 from trusted data, such as a small calibration panel of users who bypass the relay. Our experimental detectors are instead given the residual $\hat{\mathbf{x}}_t - \mathbf{x}_t$ and test it against $\mathrm{Lap}^d(b)$.

2.5 THE ATTACKER'S OPTIMIZATION PROBLEM

Let $P_{\mathrm{DP},t}$ denote the distribution of the protected context $\tilde{\mathbf{x}}_{t,i}$ and $P_{\mathrm{obs},t}$ the distribution actually observed by the learner after corruption. The adversary maximizes disparity while keeping the observed distribution close to the protected one:

$$\max_{\{\boldsymbol{\delta}_{t,i}\}} \Delta R_T^{(g^\dagger)} \qquad \text{s.t.} \qquad \sum_{t=1}^{T} \mathrm{KL}\big(P_{\mathrm{obs},t} \,\|\, P_{\mathrm{DP},t}\big) \;\le\; A, \tag{5}$$

where $A > 0$ is the total divergence budget over the horizon. Under Laplace noise, each per-round KL term decomposes coordinate-wise and, when the budget is allocated uniformly ($\alpha = A/T$ per round), the constraint reduces to first order to the camouflage set $\mathcal{B}_\alpha = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le b\sqrt{2\alpha}\}$. Constraining only the per-round divergence describes an adversary that is undetectable at each instant yet reliably caught within a few hundred rounds.

3 CAMOUFLAGE TRADE-OFF

An essential issue for a fixed-shift adversary is how large the perturbation $\boldsymbol{\delta}$ can be while remaining undetectable.

Lemma 1 (Detectability cost of a perturbation)**.** *Let $\boldsymbol{\eta} \sim \mathrm{Lap}^d(b)$ be the privacy noise and $\boldsymbol{\delta} \in \mathbb{R}^d$ be a perturbation with coordinates $\delta_1, \ldots, \delta_d$, and write $P_{\boldsymbol{\eta}}$ and $P_{\boldsymbol{\eta}+\boldsymbol{\delta}}$ for the distributions of the noise before and after the shift. Then*

$$\mathrm{KL}(P_{\boldsymbol{\eta}+\boldsymbol{\delta}} \,\|\, P_{\boldsymbol{\eta}}) = \sum_{j=1}^{d} \psi\left(\frac{|\delta_j|}{b}\right), \qquad \psi(u) = u + e^{-u} - 1. \tag{6}$$

where $u = |\delta_j|/b$ is the shift-to-noise ratio in dimension j. The cost function ψ has two regimes: for small shifts ($|\delta_j| \ll b$), $\psi(u) \approx u^2/2$, so cost grows as the square of the shift; for large shifts ($|\delta_j| \gg b$), $\psi(u) \approx u - 1$, so cost grows linearly. The total cost across all d dimensions simply adds up because the noise in each dimension is independent.

Proposition 1 (A conservative Euclidean approximation)**.** *For a per-round divergence budget $\alpha \le 1/3$, the set of feasible perturbations is*

$$\mathcal{B}_\alpha = \{\boldsymbol{\delta} : \Psi(\boldsymbol{\delta}) \le \alpha\}, \qquad \mathcal{E}_\alpha = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le R_\alpha\}, \qquad R_\alpha = b\sqrt{2\alpha}. \tag{7}$$

Then $\mathcal{E}_\alpha \subseteq \mathcal{B}_\alpha \subseteq (1 - \sqrt{3\alpha}/3)^{-1/2}\mathcal{E}_\alpha$. The first inclusion holds for every $\alpha \ge 0$. The exact set is generally not a Euclidean ball and its relative radial approximation error is $O(\sqrt{\alpha})$ as $\alpha \downarrow 0$. With equal weights on $s = 3$ coordinates and $\alpha = 0.05$, the exact radius is approximately $0.326b$, compared with $R_\alpha \approx 0.316b$.

4 CAMOUFLAGE CHANNEL ATTACK

This section investigates four variants of an attack model based on how an adversary can allocate its budget A to maximize fairness damage under privacy noise and detection constraints.

Definition 2 (Allocated GTCMA)**.** *An allocation is specified by a schedule $\{\alpha_t\}$ with $\sum_t \alpha_t \le A$ and unit directions $\hat{\boldsymbol{\delta}}_t$; the adversary injects $\boldsymbol{\delta}_t = b\sqrt{2\alpha_t}\,\hat{\boldsymbol{\delta}}_t$. Let $\mathbf{w}$ be the unit boundary normal supported on the sensitive coordinates S. We study four allocations:* Attack I *(diffuse), $\hat{\boldsymbol{\delta}}_t = \mathbf{v}/\sqrt{d}$ with $\mathbf{v} \in \{\pm 1\}^d$ and $v_j = \mathrm{sign}(w_j)$ for $j \in S$, α_t constant;* Attack II *(concentrated), $\hat{\boldsymbol{\delta}}_t = \mathbf{w}$, α_t constant;* Attack III *(burst), α_t supported on $W_{\mathrm{att}} \ll T$ rounds;* Attack IV *(adaptive), $\hat{\boldsymbol{\delta}}_t = \hat{\boldsymbol{\delta}}_t(\tilde{\mathbf{x}}_t)$.*

4.1 REWARD CHANNEL

Let $\tilde{h}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\eta}}[h(\mathbf{x}+\boldsymbol{\eta})] = (h * p_b)(\mathbf{x})$ be the noise-averaged reward at clean context $\mathbf{x}$. This is not the conditional expected reward $h(\mathbf{x})$ in Eq. (1), nor is it automatically the learner's fitted predictor. If h is L_f-Lipschitz, then $|\tilde{h}(\mathbf{x}+\boldsymbol{\delta}) - \tilde{h}(\mathbf{x})| \le L_f\|\boldsymbol{\delta}\|_1$. Smoothing improves this bound only when

$\|h\|_\infty/b < L_f$, such as threshold rewards. The connection between noise and certified robustness motivates this function-level result (Lecuyer et al., 2019; Cohen et al., 2019; Pinot et al., 2019; Phan et al., 2020). We formalize it in the bandit setting.

Theorem 2 (Smoothing certificate under context-level DP)**.** *Let h be measurable with $\|h\|_\infty < \infty$ and let $\tilde{h} = h * p_b$ with $b = \Delta_1/\varepsilon$. Then $\tilde{h}$ is differentiable and*

$$\mathrm{Lip}_{\ell_1}(\tilde{h}) \;\le\; L_{\mathrm{eff}} \;\triangleq\; \min\left\{L_f,\; \frac{\|h\|_\infty}{b}\right\} = \min\left\{L_f,\; \frac{\|h\|_\infty\,\varepsilon}{\Delta_1}\right\}, \tag{8}$$

with $L_f = \infty$ when h is not Lipschitz. Hence $|\tilde{h}(\mathbf{x}+\boldsymbol{\delta}) - \tilde{h}(\mathbf{x})| \le L_{\mathrm{eff}}\|\boldsymbol{\delta}\|_1$ for every $\boldsymbol{\delta}$. The constant $\|h\|_\infty/b$ is tight (Proposition 5, Appendix B).

Remark 1 (Scope of the certificate)**.** *Theorem 2 bounds how much the auxiliary noise-averaged function $\tilde{h}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\eta}}[h(\mathbf{x}+\boldsymbol{\eta})]$ can shift when the input is perturbed, but the learner never computes this average. Three limits follow. First, if $L_f \le \|h\|_\infty/b$ the minimum is attained by L_f and privacy adds nothing. Second, $\tilde{h}$ averages over fresh noise, but the learner sees one noisy realisation per round and never computes that average. Third, the learner's estimates are fitted online to corrupted data, so perturbations can also act through training, the usual poisoning route, which the certificate does not cover. We therefore use the certificate only through Assumption 2 and describe the reward channel as* conditionally *bounded.*

Remark 2 (Two regimes of the reward-channel bound)**.** *Substituting the ℓ_1 size of the diffuse allocation, $\Gamma_\alpha = \sqrt{d}\,R_\alpha = \sqrt{d}\,(\Delta_1/\varepsilon)\sqrt{2\alpha}$, into Eq. (8) gives*

$$L_{\mathrm{eff}}\,\Gamma_\alpha \;=\; \sqrt{2\alpha d}\;\min\left\{\frac{L_f\Delta_1}{\varepsilon},\; \|h\|_\infty\right\}, \qquad \varepsilon^\star \;\triangleq\; \frac{L_f\Delta_1}{\|h\|_\infty}. \tag{9}$$

For $\varepsilon < \varepsilon^\star$, the smoothing term binds and the certified reward-channel damage is $\|h\|_\infty\sqrt{2\alpha d}$ and strengthening privacy enlarges the camouflage radius and weakens the certificate. For $\varepsilon > \varepsilon^\star$ the Lipschitz term binds and certified damage decays as $1/\varepsilon$. The bound is informative only when $2\alpha d < 4$, since the range of h is at most $2\|h\|_\infty$; at $\alpha = 0.05$ and $d = 12$ it equals $1.1\|h\|_\infty$.

Assumption 2 (Estimates track the smoothed reward)**.** *For a specified learner and a common sequence of clean contexts and true labels, suppose a known finite $e_T \ge 0$ controls the compared histories. Perturbing contexts by $\|\boldsymbol{\delta}_t\|_1 \le \Gamma$ changes each arm's estimate by at most $L_{\mathrm{eff}}\Gamma + e_T$, that is, $\|\hat{\boldsymbol{\mu}}'_t - \hat{\boldsymbol{\mu}}_t\|_\infty \le L_{\mathrm{eff}}\Gamma + e_T$ uniformly over rounds and the histories being compared.*

Definition 3 (λ-smooth selection)**.** *A policy is λ-smooth if $\|\pi(\hat{\boldsymbol{\mu}}) - \pi(\hat{\boldsymbol{\mu}}')\|_1 \le \lambda\,\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}'\|_\infty$ for all reward-estimate vectors. For softmax at temperature $\kappa > 0$, $\lambda = 2/\kappa$ is valid. Ordinary ϵ-greedy with $\epsilon_{\mathrm{expl}} < 1$ is discontinuous at score ties and is not globally λ-smooth.*

Proposition 2 (Fairness bound on the reward channel)**.** *Under Assumption 2, a λ-smooth policy and $\|\boldsymbol{\delta}_t\|_1 \le \Gamma$, with disparity measured on fixed, action-independent true labels and a common clean reward sequence,*

$$\left|\Delta R^{(g)}_{T,\mathrm{att}} - \Delta R^{(g)}_{T,\mathrm{clean}}\right| \;\le\; 2\lambda\big(L_{\mathrm{eff}}\Gamma + e_T\big)\Delta_{\max}, \qquad \Delta_{\max} = \sup_{t,i,j}|h(\mathbf{x}_{t,i}) - h(\mathbf{x}_{t,j})|. \tag{10}$$

For deterministic selection, let $v_{t,i}$ be the complete score used to choose an arm. If $\|v'_t - v_t\|_\infty \le B$ and the clean best-score gap exceeds $2B$ at every compared round, the selected arms are unchanged. For UCB this must control the confidence bonuses and history as well as the estimated means.

4.2 Variant I (diffuse) vs. Variant II (concentrated)

The group channel has no smoothing certificate, and a user is reassigned once the perturbation pushes the inferred context across the boundary. Reassignment depends on the displacement $\langle \mathbf{w}, \boldsymbol{\delta}_t\rangle$, while stealth constrains $\|\boldsymbol{\delta}_t\|_2 \le R_\alpha$. **Variant I** spreads the budget across all d dimensions, which weakens boundary displacement as d grows because most dimensions are irrelevant to the boundary. **Variant II** targets the s sensitive dimensions and maximizes displacement under the same budget.

Theorem 3 (Optimal spatial allocation)**.** *With a per-round budget α and $\mathbf{w}$ as the unit normal of the group boundary supported on S with $|S| = s$, among all $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_2 \le R_\alpha$,*

$$\max_{\|\boldsymbol{\delta}\|_2 \le R_\alpha} \langle \mathbf{w}, \boldsymbol{\delta} \rangle = R_\alpha = b\sqrt{2\alpha}, \qquad \textit{attained at } \boldsymbol{\delta}^\star = R_\alpha \mathbf{w}. \tag{11}$$

An adversary that spreads its budget over all d dimensions with signs matched to $\mathbf{w}$ achieves $\langle \mathbf{w}, \boldsymbol{\delta} \rangle = R_\alpha \|\mathbf{w}\|_1 / \sqrt{d}$. Alignment is therefore worth

$$\frac{\textit{aligned reach}}{\textit{uniform reach}} \;=\; \frac{\sqrt{d}}{\|\mathbf{w}\|_1} \;\ge\; \sqrt{\frac{d}{s}}, \tag{12}$$

with equality exactly when $\mathbf{w}$ places equal weight on all s sensitive dimensions.

Assumption 3 (Regular margin)**.** *Let $M = \langle \mathbf{w}, \tilde{\mathbf{x}}_S \rangle - \vartheta$ denote the signed margin of a target-group context, after privacy noise, to the group boundary at offset ϑ. The law of M admits a density p_M that is continuous and strictly positive on a neighbourhood of the origin.*

Corollary 4 (Amplification in group reassignment)**.** *Under Assumption 3, a displacement $q = \langle \mathbf{w}, \boldsymbol{\delta} \rangle \to 0^+$ reassigns a target-group user with probability*

$$\Pr[\textit{flip}] \;=\; \Pr\big[0 < M \le q\big] \;=\; p_M(0)\, q + o(q). \tag{13}$$

Consequently, the concentrated and diffuse allocations of Theorem 3 satisfy

$$\frac{\Pr[\textit{flip} \mid \text{II}]}{\Pr[\textit{flip} \mid \text{I}]} \;=\; \frac{\sqrt{d}}{\|\mathbf{w}\|_1} + o(1) \;\ge\; \sqrt{\frac{d}{s}} + o(1), \tag{14}$$

At a finite budget, if p_M is non-increasing on $(0, R_\alpha]$, the ratio lies in $\big[1,\ \sqrt{d}/\|\mathbf{w}\|_1\big]$, so the first-order value is also an upper bound. If each reassignment contributes a constant c_{fair} to the disparity (Eq. (3)*), expected fairness damage inherits the same factor.*

The first-order ratio is exact as the budget vanishes and both allocations induce $O(R_\alpha)$ displacement and the common factor $p_M(0)$ cancels. For finite budgets, $F(q) = \Pr[0 < M \le q]$ with $F(0) = 0$ implies $F(q)/q$ cannot increase, so measured amplification should fall below $\sqrt{d/s}$ as budget grows. Fairness damage requires c_{fair} to be approximately constant across reassigned users.

4.3 Variants III and IV: Timing and Adaptivity

Attacks I and II fix the per-round budget $\alpha_t = A/T$ and the location for the perturbation. We now ask *when* to spend it (Variant III) and whether conditioning on the realised noise helps (Variant IV).

Variant III: Temporal burst. A burst attack concentrates the entire budget A into a short window of $W_{\text{att}} \ll T$ rounds, spending $\alpha_t = A/W_{\text{att}}$ inside the window. The per-round budget inside the window can therefore far exceed the uniform level A/T, giving each perturbation a larger reach $R_{\alpha_t} = b\sqrt{2\alpha_t}$. Let f be the margin density of Assumption 3 and $g(\alpha) = c_{\text{fair}} \int_0^{b\sqrt{2\alpha}} f(m)\, dm$ be the expected per-round damage of a concentrated perturbation at per-round budget α. We consider schedules with $\sum_t \alpha_t \le A$ and $\alpha_t \le \alpha_{\max} \le 1/3$, so that Proposition 1 applies on every round.

Theorem 5 (When to spread and burst)**.** *Let $r_{\max} = b\sqrt{2\alpha_{\max}}$ and $A/T \le \alpha_{\max}$.*

(i) If f is non-increasing on $(0, r_{\max}]$, then g is concave and $\sum_t g(\alpha_t) \le T\, g(A/T)$, with equality if and only if $\alpha_t = A/T$ for every t. For a constant boundary density, a burst on W_{att} rounds achieves only a fraction $\sqrt{W_{\text{att}}/T}$ of the uniform damage.

(ii) If $rf'(r) \ge f(r)$ on $(0, r_{\max}]$ (e.g. $f(m) \propto m^k$, $k \ge 1$), then g is convex and $\sum_t g(\alpha_t) \le (A/\alpha_{\max})\, g(\alpha_{\max})$, attained by spending $\alpha_{\max}$ on $A/\alpha_{\max}$ rounds, i.e., a burst.

In Case (i), many users lie near the boundary, which creates diminishing returns; thus, spreading the budget evenly maximizes reassignment. In Case (ii), few users are near the boundary, so the attacker must exceed a threshold and should spend the maximum per round until the budget is exhausted. Detection does not alter these optima: gradual spending remains best in Case (i), while bursting maximizes damage in Case (ii).

Variant IV: Adaptive. An adversary that observes the protected context $\tilde{\mathbf{x}}_t$ might try to adapt its perturbation to the realized noise, for example, spending more when the noise has already pushed the context near the boundary. Under a pathwise constraint, $\boldsymbol{\delta}_t \in \mathcal{B}_\alpha$ on every realization, this does not help. The adaptive rule picks from the same set of vectors as a fixed rule on every round, so neither the reward-channel bound of Theorem 2 nor the group-channel bound of Theorem 3 can be exceeded.

Figure 1 depicts the four allocations using a 3D example. **Attack I (isotropic/diffuse)** spreads the budget equally over all three coordinates, so its projection onto $\mathbf{w} = \mathbf{e}_1$ is only $R_\alpha/\sqrt{d}$ and its arrow tip stops short of the boundary. **Attack II (concentrated)** places the whole budget on x_1, and reaches across the boundary. **Attack III (burst)** spends several rounds' budget at once, so $\|\boldsymbol{\delta}\|_2 = 4.49$ lies far outside $\mathcal{B}_\alpha$. It flips a boundary user quickly but is detectable within a few rounds by an oracle test (Theorem 6). **Attack IV (adaptive)** conditions on the realised noise but, under the pathwise constraint, attains the same directional reach as Attack II (Proposition 3). The three in-budget attacks satisfy the exact KL constraint $\sum_j \psi(|\delta_j|/b) \approx \alpha$ but have slightly different ℓ_2 norms.

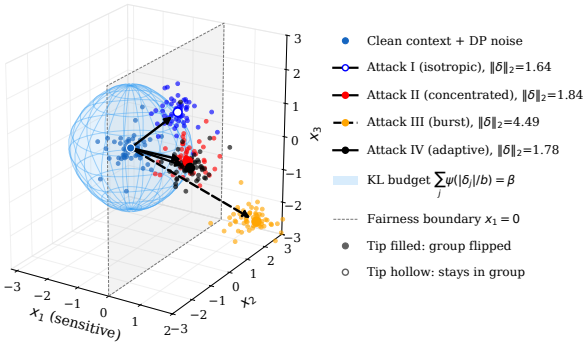


Figure 1: Per-round geometry of Attacks I–IV in a 3D example ($d = 3$, $s = 1$, x_1 sensitive, $b = 1$, $\beta = 1$). The blue ball is the camouflage set $\mathcal{B}_\alpha$ (Proposition 1).

5 Detection Delay

Theorem 6 (Oracle sequential detection)**.** *Let a defender observe noise residuals and know $\boldsymbol{\delta}$, with per-round divergence $\alpha \le 1/3$. (i) The pooled log-likelihood-ratio test fires after $\Theta(\log(1/\delta_{\mathrm{fp}})/\alpha)$ rounds for every allocation, and no test has a better exponent. (ii) Per-dimension tests with Bonferroni correction fire after $\Theta(s\log(d/\delta_{\mathrm{fp}})/\alpha)$ rounds against Attack II with equal weights and $\Theta(d\log(d/\delta_{\mathrm{fp}})/\alpha)$ against Attack I.*

An oracle defender cannot be evaded by choosing where to place the budget, and per-dimension monitors are d/s times faster against concentrated than diffuse attacks but never faster than the oracle. A deployed monitor must estimate $P_{\mathbf{x}} * p_b$ and does not know $\boldsymbol{\delta}$. A direction-agnostic pooled statistic, such as the energy of residuals over all d dimensions, averages a signal carried by s dimensions against noise from all d, and so loses power against concentrated attacks; monitoring sensitive dimensions separately recovers it. Figure 2 illustrates this on a synthetic example. The per-dimension monitor crosses the threshold within ∼40 rounds, while the pooled monitor needs ∼290, a $\sim d/s$ ratio. Spending α_b per round, an attacker is detected after $n \approx \log(1/\delta_{\mathrm{fp}})/\alpha_b$ rounds and accrues damage about $\log(1/\delta_{\mathrm{fp}})\, g(\alpha_b)/\alpha_b$. In case (i) of Theorem 5 $g(\alpha)/\alpha$ decreases, so a detection-aware attacker should spend slowly; in case (ii) it increases, so bursting also maximizes damage before detection.

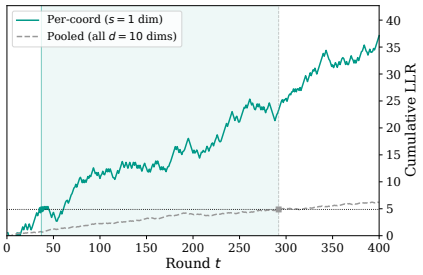


Figure 2: Detection asymmetry under Attack II (concentrated) on a synthetic example with $d = 10$, $s = 1$.

6 Experimental Evaluation

6.1 Experimental Setup

Benchmarks and algorithms. We evaluate on three real-world datasets adapted to the contextual bandit setting: *Adult Income* (d=12, s=2, G=4, K=3), *COMPAS* (d=12, s=2, G=2, K=3) , and *German Credit* (d=20, s=3, G=2, K=3). All benchmarks use T=10,000 rounds, N=50 runs,

Table 1: Average ΔDP ($\times 10^{-2}$) across Adult, COMPAS, and German. The final column is the relative increase under Attack II over the no-attack baseline.

Algorithm	No Attack	Attack I	Attack II	Attack III	Attack IV	Rel.↑ (II)
NeuralLinUCB	15.6±0.3	19.5±0.3	**22.9±0.3**	19.6±0.3	19.8±0.4	20.5%
NeuralUCB	17.9±0.5	17.6±0.6	**22.7±0.6**	18.7±0.5	18.7±0.6	26.8%
FairBandit	8.5±0.4	9.9±0.4	**12.8±0.4**	9.1±0.4	9.6±0.4	**50.6%**
RobustBandit	10.7±0.3	10.0±0.4	**11.5±0.3**	8.2±0.3	11.2±0.4	**55.4%**
DP-LinUCB	15.3±0.3	15.9±0.3	**19.4±0.3**	16.7±0.3	15.9±0.3	26.8%
DP-NCB	14.4±0.5	18.0±0.5	**21.3±0.5**	18.5±0.6	19.1±0.5	20.3%

ε=1.0, and Δ_1=1.0. Victims are NeuralUCB Zhou et al. (2020), NeuralLinUCB Xu et al. (2022), RobustBandit Ding et al. (2022), FairBandit (Patil et al., 2021), DP-LinUCB Shariff & Sheffet (2018), and DP-NCB Sarkar et al. (2026). DP algorithms apply context-level DP with Laplace noise at the specified ε; non-DP algorithms receive clean contexts with corruption applied directly. **Attacks and detectors.** All attacks use an ℓ_1 budget $\|\boldsymbol{\delta}_t\|_1 \leq \beta db$, $\beta \in \{0.5, 1, 1.5\}$. Attack II places a fraction $\gamma = 0.9$ on the s sensitive dimensions; Attack III uses $W_{\text{att}} = 0.2T$, $\gamma_{\text{burst}} \in \{2, 5, 10\}$. Both detectors read residuals against a $\text{Lap}(b)$ null (the oracle defender): Aniso runs a Bonferroni-corrected sequential test per coordinate, and Iso pools a direction-agnostic statistic over all d. **Budgets are matched in ℓ_1, not KL.** Because ψ is convex, concentrating fixed ℓ_1 mass costs more divergence. Attack II spends 2.0× the per-round KL of Attack I at $\beta = 1$ and about 2.8× at $\beta = 0.5$, and at $\beta = 1$ the diffuse attack already shifts each coordinate by one noise scale.

6.2 Results

Fairness damage. Table 1 shows that Attack II introduces the highest damage for the victim, raising ΔDP by 20–55% although it also spends about twice compared with Attack I. FairBandit and RobustBandit have the lowest absolute disparity but the largest relative increase (50.6% and 55.4%); defenses against reward corruption do not address context corruption that changes group membership. The absolute increase is similar with and without privacy: +4.1 and +3.6 points for DP-LinUCB and DP-NCB against +3.9 to +4.8 for the others. Privacy noise thus offers no visible protection on the group channel, which the certificate does not cover (per-dataset results in Appendix D.4).

Detection. Table 2 presents the detection analysis across three datasets under Attack II with aggregate (Iso) and per-dimension (Aniso) monitoring. For the DP-protected learners, Aniso detects the attack after 113–118 rounds, whereas Iso requires 702–733 rounds, yielding a detection-time ratio of approximately 6.2. Aniso monitoring also reduces the average post-attack ΔDP from 21.0 to 16.7 for DP-LinUCB and from 23.6 to 19.2 for DP-NCB. These values remain 1.4 and 4.8 points above their respective clean baselines, but are substantially lower than those under Iso monitoring.

Privacy level. Figure 3 shows the effect of the privacy parameter ε on fairness damage ΔDP and cumulative regret R_T, for DP-LinUCB (a) and DP-NCB (b), averaged across the three datasets. For

Table 2: Average detection analysis for Attack II (concentrated) across Adult, COMPAS, and German under aggregate (Iso) and per-coordinate (Aniso) monitoring.

Algorithm	Clean	Iso ΔDP	Aniso ΔDP	Flip	Det(Iso)	Det(Aniso)	Ratio
NeuralLinUCB	15.6±0.08	20.8±0.11	20.8±0.07	0.20	N/D	N/D	—
NeuralUCB	17.9±0.10	21.8±0.04	21.8±0.06	0.15	N/D	N/D	—
FairBandit	8.5±0.08	15.3±0.05	11.3±0.06	0.22	N/D	N/D	—
RobustBandit	10.8±0.05	21.3±0.08	21.3±0.09	0.10	N/D	N/D	—
DP-LinUCB	15.3±0.06	21.0±0.24	16.7±0.09	0.21	702	**113**	6.2
DP-NCB	14.4±0.07	23.6±0.10	19.2±0.11	0.21	733	**118**	6.2

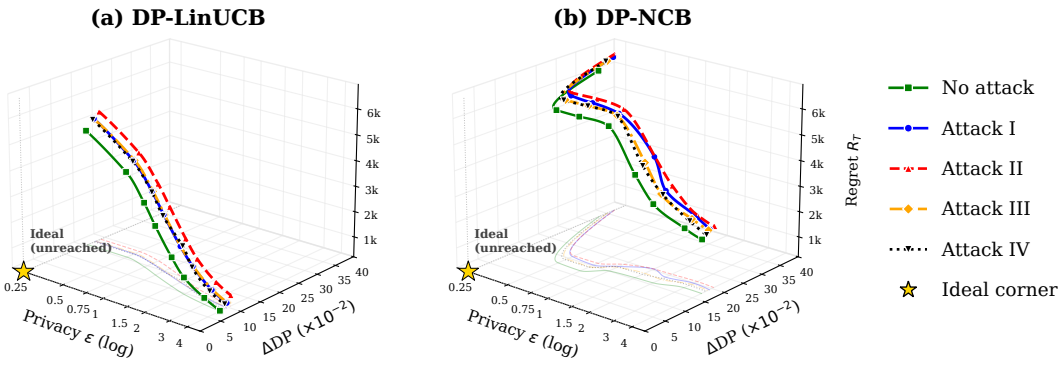


Figure 3: Privacy–utility trade-off for DP-LinUCB (a) and DP-NCB (b), averaged across Adult, COMPAS, and German. Panels (1): fairness damage Δ_{DP}, Panels (2): cumulative regret R_T.

DP-LinUCB, both ΔDP and R_T decrease steadily with ε, as the camouflage radius $R_\alpha = b\sqrt{2\alpha}$ shrinks and the private-noise estimation cost falls. DP-NCB preserves the same ordering across attack variants in ΔDP, but its regret peaks near $\varepsilon \approx 0.75$–1 before falling. We attribute this to the neural learner's sensitivity to the noise scale in that range. Attack II dominates the other variants in ΔDP at every ε on both algorithms, and its regret cost is modest.

Attack budget. Figure 4 shows fairness damage (panel 1) and cumulative regret (panel 2) against the per-round KL budget β for DP-LinUCB. Every variant sits above the no-attack baseline already at β=0.1. Attacks I and II climb steeply from $\beta \approx 0.5$, while the burst and adaptive variants stay nearly flat until $\beta \approx 2$, consistent with case (i) of Theorem 5. Over $\beta \in [0.1, 4]$ Attack II raises ΔDP by 19.8% but regret by only 5.8%, so aggregate regret registers the corruption but understates the group-level harm roughly threefold (per-dataset results in Appendix D.3).

7 Conclusion

We studied adversarial fairness attacks that hide group-targeted context corruption within DP noise and formulated them as a budget-allocation problem. Aligning perturbations with the group boundary increases reassignment probability by up to $\sqrt{d/s}$ over diffuse spreading. Steady spending works best when many users are near the boundary, while bursting is preferable when groups are well separated; detection does not change these regimes. Laplace smoothing protects expected rewards only for non-smooth reward functions, not inferred group labels. Experiments across three datasets and six learners show that concentrated attacks cause the most harm and that per-coordinate monitoring detects them several times faster than direction-agnostic pooling.

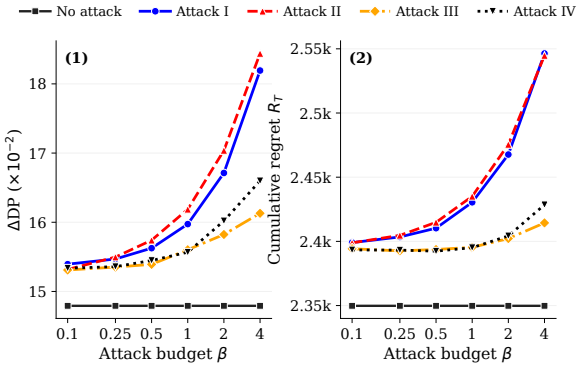


Figure 4: Budget sensitivity for DP-LinUCB, averaged across Adult, COMPAS, and German. (1) Fairness damage ΔDP vs. attack budget β (ℓ_1 scale). (2) Cumulative regret R_T vs. β.

Several questions remain open. (1) The per-capita regret differential A_{fair} that converts a group reassignment into fairness damage is left as a property of the learner. (2) Our adaptivity result assumes the attacker respects its budget on every round, and whether an attacker constrained only on average could do better is unresolved. (3) Whether the d/s detection advantage is the best any detector can achieve is unknown, and settling it requires a matching lower bound. (4) The critical privacy level $\varepsilon^\star$ covers the reward channel only; no analogue is known for the group channel.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and assist with implementing parts of the experiments under supervision. For the experiments, these tools helped fix coding errors, revise experimental scripts, and prepare plotting code. The authors specified the experimental methods and evaluation procedures, reviewed the AI-assisted code, tested its execution, and checked the resulting outputs. We did not use these tools to develop the novelty of the work, discover literature, draft sections of the paper, generate synthetic datasets, or formulate or prove mathematical claims. All AI-assisted text and code were reviewed, tested, and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work investigates an adversarial threat to fairness in DP-protected contextual bandit systems, including deployments in healthcare, credit scoring, and criminal justice referenced in the Introduction. We publish the attack analysis to inform defence. Theorem 6 and Section 5 give a per-coordinate residual monitor that detects the concentrated attack at a d/s speedup over pooled monitoring and reduces post-attack disparity to within a few points of the clean baseline. The datasets we use (Adult, COMPAS, German Credit) are public research benchmarks with well-known fairness properties; we do not release any personally identifying data. We considered the disclosure trade-off between publishing an attack that could be reproduced by a malicious relay and providing a defence that would not exist otherwise and concluded that per-coordinate monitoring is available to any defender who reads this paper, whereas the attack requires estimating the group boundary from external data, a strictly higher bar. We do not use human subjects, and the work does not require IRB approval.

REPRODUCIBILITY STATEMENT

Full proofs of Theorems 2, 3, 5, and 6, together with the assumptions they rely on, are given in Appendices A and B. Experimental setup (datasets, victim algorithms, attack budgets, detector configurations, and sample counts) is documented in Section 6.1; per-dataset breakdowns are in Appendix D. Code to reproduce all experiments and figures will be released upon acceptance; an anonymized code is included in the supplementary material.

REFERENCES

Achraf Azize and Debabrota Basu. Concentrated differential privacy for bandits. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 78–109. IEEE, 2024.

Jackie Baek and Vivek Farias. Fair exploration via axiomatic bargaining. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 22034–22045, 2021.

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23951–23964, 2022.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning attacks to local differential privacy protocols. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 947–964, 2021.

Junyi Chai, Taeuk Jang, Jing Gao, and Xiaoqian Wang. On the alignment between fairness and accuracy: from the perspective of adversarial robustness. In *International Conference on Machine Learning (ICML)*, 2025.

Vasileios Charisopoulos, Hossein Esfandiari, and Vahab Mirrokni. Robust and private stochastic linear bandits. In *International Conference on Machine Learning*, pp. 4096–4115. PMLR, 2023.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. In *2021 IEEE symposium on security and privacy (SP)*, pp. 883–900. IEEE, 2021.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pp. 1310–1320. PMLR, 2019.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions: Y. deldjoo et al. *User Modeling and User-Adapted Interaction*, 34(1):59–108, 2024.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123. PMLR, 2022.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 14362–14373, 2020.

Riccardo Grazzi, Arya Akhavan, John Isak Texas Falk, Leonardo Cella, and Massimiliano Pontil. Group meritocratic fairness in linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

Yuxuan Han, Zhipeng Liang, Yang Wang, and Jiheng Zhang. Generalized linear bandits with local differential privacy. In *Advances in neural information processing systems (NeurIPS)*, volume 34, pp. 26511–26522, 2021.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in neural information processing systems (NeurIPS)*, volume 35, pp. 34614–34625, 2022.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 771–787, 2018.

Benedikt Höltgen and Nuria Oliver. Reconsidering fairness through unawareness from the perspective of model multiplicity. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 241–255, 2025.

Seyed Mohammad Hadi Hosseini, Amir Najafi, and Mahdieh Baghshah. Efficient adversarial attacks on high-dimensional offline bandits. In *International Conference on Learning Representations (ICLR)*, 2026.

Taeuk Jang, Xiaoqian Wang, and Heng Huang. Adversarial fairness network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22159–22166, 2024.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in neural information processing systems (NeurIPS)*, volume 29, 2016.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 10897–10908, 2023.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security & Privacy (SP)*, 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. Data poisoning attacks in contextual bandits. In *International Conference on Decision and Game Theory for Security*, pp. 186–204. Springer, 2018.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174):1–31, 2021.

Nikola Pavlovic, Sudeep Salgia, and Qing Zhao. Differential privacy in kernelized contextual bandits via random projections. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning (ICML)*, pp. 7683–7694. PMLR, 2020.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in neural information processing systems (NeurIPS)*, volume 32, 2019.

Yunzhe Qi, Yikun Ban, Arindam Banerjee, and Jingrui He. Robust neural contextual bandit against adversarial corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 19378–19446, 2024.

Dhruv Sarkar, Nishant Pandey, and Sayak Ray Chowdhury. Dp-ncb: Privacy preserving fair bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25192–25199, 2026.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 162–177. Springer, 2020.

Aristide Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *AAAI conference on artificial intelligence*, volume 30, 2016.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Xuchuang Wang, Maoli Liu, Jinhang Zuo, Xutong Liu, John Lui, and Mohammad Hajiesmaili. Stochastic bandits robust to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2025.

Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations (ICLR)*, 2022.

Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. In *International Conference on Learning Representations (ICLR)*, 2022.

Yinglun Xu, Bhuvesh Kumar, and Jacob D Abernethy. Observation-free attacks on stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 22550–22561, 2021.

Qirun Zeng, Eric He, Richard Hoffmann, Xuchuang Wang, and Jinhang Zuo. Practical adversarial attacks on stochastic bandits via fake data injection. *arXiv preprint arXiv:2505.21938*, 2025.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: a survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–28, 2025.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12300–12310, 2020.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning (ICML)*, pp. 11492–11502. PMLR, 2020.

A Deferred statements

Remark 3 (Estimated boundary normal)**.** *Aligning with an estimate* $\hat{\mathbf{w}}$ *reaches* $R_\alpha\langle \mathbf{w}, \hat{\mathbf{w}}\rangle$*, so concentration beats the diffuse allocation if and only if* $\langle \mathbf{w}, \hat{\mathbf{w}}\rangle > \|\mathbf{w}\|_1/\sqrt{d}$*, which for equal weights reads* $\cos\angle(\mathbf{w}, \hat{\mathbf{w}}) > \sqrt{s/d}$*. For* $d = 12$ *and* $s = 2$ *this allows an error of up to* 66°*.*

Proposition 3 (Pathwise-constrained adaptivity gives no advantage)**.** *Let* $\boldsymbol{\delta}_t = \boldsymbol{\delta}_t(\tilde{\mathbf{x}}_t)$ *be measurable with* $\boldsymbol{\delta}_t \in \mathcal{B}_\alpha$ *almost surely, and* $\Gamma_\alpha = \sup_{\boldsymbol{\delta}\in\mathcal{B}_\alpha} \|\boldsymbol{\delta}\|_1$*. Then* $|\tilde{h}(\mathbf{x} + \boldsymbol{\delta}_t) - \tilde{h}(\mathbf{x})| \le L_{\text{eff}}\Gamma_\alpha$ *and* $\mathbb{E}\langle \mathbf{w}, \boldsymbol{\delta}_t\rangle \le R_\alpha(1 + \mathcal{O}(\sqrt{\alpha}))$*, and both bounds are attained to first order by a fixed allocation.*

Proposition 4 (No free evasion, for additive attacks)**.** *For fixed* $\boldsymbol{\delta} \neq \mathbf{0}$*,* $\mathrm{KL}(P_{\boldsymbol{\eta}+\boldsymbol{\delta}} \| P_{\boldsymbol{\eta}}) > 0$*, so an additive adversary is detected within* $\Theta(\log(1/\delta_{\text{fp}})/\alpha)$ *rounds by the oracle test of Theorem 6; evasion buys delay, not immunity. Measure-preserving maps can have zero divergence with nonzero displacement (Appendix C).*

B Proofs

B.1 Proof of Proposition 1

Let $u_j = |\delta_j|/b$ and $\Psi(\boldsymbol{\delta}) = \sum_j \psi(u_j)$, so $\mathcal{B}_\alpha = \{\boldsymbol{\delta} : \Psi(\boldsymbol{\delta}) \le \alpha\}$. Taylor's theorem with Lagrange remainder, using $\psi(0) = \psi'(0) = 0$, $\psi''(0) = 1$ and $\psi'''(u) = -e^{-u} \in [-1, 0)$, gives

$$\frac{u^2}{2} - \frac{u^3}{6} \;\le\; \psi(u) \;\le\; \frac{u^2}{2}, \qquad u \ge 0. \tag{15}$$

Summing the upper bound, $\Psi(\boldsymbol{\delta}) \le \|\boldsymbol{\delta}\|_2^2/(2b^2)$, so $\|\boldsymbol{\delta}\|_2 \le R_\alpha := b\sqrt{2\alpha}$ implies $\Psi(\boldsymbol{\delta}) \le \alpha$:

$$\big\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le R_\alpha\big\} \subseteq \mathcal{B}_\alpha. \tag{16}$$

This inclusion is exact rather than asymptotic, so the displacement R_α used in Theorem 3 is always attainable and the $\sqrt{d/s}$ amplification does not inherit the approximation error.

Conversely, let $u_{\max} = \max_j u_j$. Summing the lower bound in (15) and using $\sum_j u_j^3 \le u_{\max} \sum_j u_j^2$,

$$\Psi(\boldsymbol{\delta}) \;\ge\; \frac{\|\boldsymbol{\delta}\|_2^2}{2b^2}\Big(1 - \frac{u_{\max}}{3}\Big).$$

Since ψ is increasing, $\psi(u_{\max}) \le \alpha$, and (15) gives $\psi(u) \ge u^2/3$ for $u \le 1$, whence $u_{\max} \le \sqrt{3\alpha}$. Therefore every $\boldsymbol{\delta} \in \mathcal{B}_\alpha$ satisfies $\|\boldsymbol{\delta}\|_2 \le R_\alpha(1 - u_{\max}/3)^{-1/2} = R_\alpha(1 + \mathcal{O}(\sqrt{\alpha}))$, and combining with (16),

$$\big\{\|\boldsymbol{\delta}\|_2 \le R_\alpha\big\} \;\subseteq\; \mathcal{B}_\alpha \;\subseteq\; \big\{\|\boldsymbol{\delta}\|_2 \le R_\alpha(1+\mathcal{O}(\sqrt{\alpha}))\big\},$$

which is the claimed relative error in the radius.

As a check, at $\alpha = 0.05$ with equal weight on $s = 3$ coordinates the budget binds when $\psi(u) = \alpha/3$, giving $u^\star = 0.1884$ and an exact radius $u^\star b\sqrt{3} = 0.326\,b$ against $R_\alpha = 0.316\,b$: a 3.1% error, exceeding the approximation as (16) requires. □

B.2 Tightness of the certificate

Proposition 5 (The bound is attained)**.** *There exists a bounded h with $|\tilde{h}(\mathbf{x}+\boldsymbol{\delta}) - \tilde{h}(\mathbf{x})| = (\|h\|_\infty/b)\|\boldsymbol{\delta}\|_1 + \mathcal{O}(\|\boldsymbol{\delta}\|_1^2)$, so the certificate of Theorem 2 is tight. Likewise, $\boldsymbol{\delta}^\star = R_\alpha \mathbf{w}$ attains* (11)*.*

Proof. Take $h(\mathbf{x}) = M\,\mathrm{sign}(x_1)$, which is bounded with $\|h\|_\infty = M$ and has $L_f = \infty$ (it is discontinuous at $x_1 = 0$). Convolving with the Laplace density,

$$\tilde{h}(\mathbf{x}) \;=\; \mathbb{E}\big[h(\mathbf{x}+\boldsymbol{\eta})\big] \;=\; M\big(\Pr[\eta_1 > -x_1] - \Pr[\eta_1 < -x_1]\big) \;=\; M\big(1 - 2F_b(-x_1)\big),$$

where F_b is the univariate Laplace CDF with scale b. Then $\partial_{x_1}\tilde{h}(\mathbf{x}) = 2M\,p_b(-x_1)$, and at the origin

$$\partial_{x_1}\tilde{h}(\mathbf{0}) \;=\; 2M\,p_b(0) \;=\; 2M\cdot\frac{1}{2b} \;=\; \frac{M}{b} \;=\; \frac{\|h\|_\infty}{b}.$$

A displacement $\boldsymbol{\delta} = \delta\,\mathbf{e}_1$ therefore achieves $|\tilde{h}(\boldsymbol{\delta}) - \tilde{h}(\mathbf{0})| = (M/b)\|\boldsymbol{\delta}\|_1 + \mathcal{O}(\|\boldsymbol{\delta}\|_1^2)$, so the constant $\|h\|_\infty/b$ in Theorem 2 cannot be improved. The second claim is the equality case of Theorem 3. □

B.3 Proof of Theorem 2

Let $\tilde{h}(\mathbf{x}) = \int h(\mathbf{u})\,p_b(\mathbf{u}-\mathbf{x})\,d\mathbf{u}$, where $p_b(\mathbf{z}) = \prod_{j=1}^d (2b)^{-1}e^{-|z_j|/b}$ is the d-dimensional Laplace density. Since h is bounded and p_b is integrable with partial derivatives, dominated convergence permits differentiation under the integral sign.

Away from the measure-zero set $\{z_j = 0\}$, $\partial_{z_j} p_b(\mathbf{z}) = -b^{-1}\,\mathrm{sign}(z_j)\,p_b(\mathbf{z})$, and the chain rule gives $\partial_{x_j} p_b(\mathbf{u}-\mathbf{x}) = b^{-1}\,\mathrm{sign}(u_j - x_j)\,p_b(\mathbf{u}-\mathbf{x})$. Hence

$$\big|\partial_{x_j}\tilde{h}(\mathbf{x})\big| \;=\; \Big|\,b^{-1}\int h(\mathbf{u})\;\mathrm{sign}(u_j - x_j)\,p_b(\mathbf{u}-\mathbf{x})\,d\mathbf{u}\Big| \;\le\; b^{-1}\|h\|_\infty\int p_b \;=\; \frac{\|h\|_\infty}{b},$$

so $\|\nabla\tilde{h}\|_\infty \le \|h\|_\infty/b$. By Hölder's inequality applied along the segment from $\mathbf{x}$ to $\mathbf{x}+\boldsymbol{\delta}$,

$$\big|\tilde{h}(\mathbf{x}+\boldsymbol{\delta}) - \tilde{h}(\mathbf{x})\big| \;\le\; \|\nabla\tilde{h}\|_\infty\,\|\boldsymbol{\delta}\|_1 \;\le\; \frac{\|h\|_\infty}{b}\,\|\boldsymbol{\delta}\|_1.$$

When h is L_f-Lipschitz in ℓ_1, the bound $|\tilde{h}(\mathbf{x}+\boldsymbol{\delta}) - \tilde{h}(\mathbf{x})| \le L_f\|\boldsymbol{\delta}\|_1$ holds by averaging the pointwise Lipschitz bound against p_b. Taking the smaller of the two gives $L_{\text{eff}} = \min\{L_f,\ \|h\|_\infty/b\}$. □

B.4 Proof of Proposition 2

In Definition 3, a policy π is λ-smooth if

$$\|\pi(\hat{\boldsymbol{\mu}}) - \pi(\hat{\boldsymbol{\mu}}')\|_1 \;\le\; \lambda\,\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}'\|_\infty \tag{17}$$

for all pairs of reward-estimate vectors.

The total-variation form on the left is deliberate: bounding each arm's probability separately and summing over arms introduces a spurious factor of K. For softmax at temperature κ the Jacobian is $\partial\pi_i/\partial\mu_j = \kappa^{-1}(\pi_i\delta_{ij} - \pi_i\pi_j)$, so for any perturbation direction v,

$$\|\Delta\boldsymbol{\pi}\|_1 \;\le\; \kappa^{-1}\sum_i \pi_i\,|v_i - \langle\boldsymbol{\pi}, v\rangle| \;\le\; 2\kappa^{-1}\|v\|_\infty,$$

giving $\lambda = 2/\kappa$ with no dependence on K. For ϵ-greedy with randomised tie-breaking, a single change of $\arg\max$ moves at most $2(1 - \epsilon_{\text{expl}})$ of the probability mass, again independent of K.

Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}'$ denote the arm-value estimates under the clean and perturbed contexts. Assumption 2 gives $\|\hat{\boldsymbol{\mu}}' - \hat{\boldsymbol{\mu}}\|_\infty \le L_{\text{eff}}\Gamma + e_T$, and the λ-smoothness assumption (17) yields

$$\|\boldsymbol{\pi}' - \boldsymbol{\pi}\|_1 \;\le\; \lambda\,(L_{\text{eff}}\Gamma + e_T).$$

Per-capita group regret is an average of terms $\sum_i \pi_i(\mu_{a^*} - \mu_i)$. By Hölder,

$$\Big|\sum_i (\pi'_i - \pi_i)(\mu_{a^*} - \mu_i)\Big| \;\le\; \|\boldsymbol{\pi}' - \boldsymbol{\pi}\|_1 \max_i |\mu_{a^*} - \mu_i| \;\le\; \lambda\,(L_{\text{eff}}\Gamma + e_T)\,\Delta_{\max},$$

so each group's per-capita regret moves by at most $\lambda(L_{\text{eff}}\Gamma + e_T)\Delta_{\max}$.

For the $\arg\max$ policy, suppose the gap between the best and second-best arm exceeds $2(L_{\text{eff}}\Gamma + e_T)$ at every round. Since each estimate moves by at most $L_{\text{eff}}\Gamma + e_T$, the top two arms cannot exchange order: the leader's estimate falls by at most $L_{\text{eff}}\Gamma + e_T$ and the runner-up's rises by at most $L_{\text{eff}}\Gamma + e_T$, a total closure of $2(L_{\text{eff}}\Gamma + e_T)$, which is strictly less than the gap. The selected arm is therefore unchanged at every round, so both group regrets are identical and the disparity is exactly zero.

B.5 Proof of Lemma 1

Consider first $d = 1$, with density $p_b(z) = (2b)^{-1}e^{-|z|/b}$. Writing the divergence as an expectation under the shifted law and substituting $Y = Z - \delta$ so that $Y \sim p_b$,

$$\mathrm{KL}\big(P_{\eta+\delta}\,\|\,P_\eta\big) \;=\; \mathbb{E}_{Z\sim P_{\eta+\delta}}\left[\frac{|Z| - |Z-\delta|}{b}\right] \;=\; \frac{1}{b}\,\mathbb{E}_{Y\sim p_b}\big[\,|Y+\delta| - |Y|\,\big].$$

Assume $\delta > 0$ without loss of generality (the Laplace density is symmetric). Splitting the integral at $y = -\delta$ and $y = 0$, where the two absolute values change form, the integrand equals $-\delta$ on $(-\infty, -\delta)$, equals $2y + \delta$ on $(-\delta, 0)$, and equals δ on $(0, \infty)$. Evaluating the three pieces gives

$$\mathbb{E}_Y\big[|Y+\delta| - |Y|\big] \;=\; -\tfrac{\delta}{2}e^{-\delta/b} \;+\; \tfrac{1}{2}\big[\delta e^{-\delta/b} - 2b + 2b\,e^{-\delta/b} + \delta\big] \;+\; \tfrac{\delta}{2} \;=\; \delta + b\big(e^{-\delta/b} - 1\big).$$

Dividing by b yields $\delta/b + e^{-\delta/b} - 1 = \psi(|\delta|/b)$ with $\psi(u) = u + e^{-u} - 1$, as claimed.

For general d, the coordinates of $\boldsymbol{\eta}$ are independent, so the joint KL divergence is the sum of the marginal divergences, giving (6).

Finally, $\psi'(u) = 1 - e^{-u}$ and $\psi''(u) = e^{-u} > 0$, so ψ is strictly convex and strictly increasing on $[0, \infty)$ with $\psi(0) = 0$. The Taylor expansion at 0 gives $\psi(u) = u^2/2 - u^3/6 + \mathcal{O}(u^4)$, and as $u \to \infty$ the term e^{-u} vanishes, leaving $\psi(u) = u - 1 + o(1)$. □

B.6 Proof of Theorem 3

Since $\|\mathbf{w}\|_2 = 1$ and $\|\boldsymbol{\delta}\|_2 \le R_\alpha$ for every $\boldsymbol{\delta} \in \mathcal{B}_\alpha$, Cauchy–Schwarz gives

$$\langle \mathbf{w}, \boldsymbol{\delta}\rangle \;\le\; \|\mathbf{w}\|_2\,\|\boldsymbol{\delta}\|_2 \;\le\; R_\alpha,$$

with equality if and only if $\boldsymbol{\delta} = R_\alpha \mathbf{w}$. This is the concentrated allocation, and it attains the maximum displacement R_α along the boundary normal.

For the diffuse allocation, let $\in \{\pm 1\}^d$ be the sign vector that agrees with $\mathrm{sign}(w_j)$ for $j \in S$ and is chosen arbitrarily on $[d] \setminus S$, and set $\boldsymbol{\delta} = (R_\alpha/\sqrt{d})\,\mathbf{v}$. Then $\|\boldsymbol{\delta}\|_2 = R_\alpha$, so the allocation is feasible and saturates the budget, and since $\mathbf{w}$ vanishes off S,

$$\langle \mathbf{w}, \boldsymbol{\delta}\rangle \;=\; \frac{R_\alpha}{\sqrt{d}} \sum_{j\in S} |w_j| \;=\; \frac{R_\alpha}{\sqrt{d}}\,\|\mathbf{w}\|_1.$$

Applying Cauchy–Schwarz a second time on the support, $\|\mathbf{w}\|_1 \le \sqrt{s}\,\|\mathbf{w}\|_2 = \sqrt{s}$, with equality if and only if $|w_j| = s^{-1/2}$ for all $j \in S$. Therefore the diffuse allocation achieves displacement at most $R_\alpha\sqrt{s/d}$, and the ratio of the two displacements is

$$\frac{R_\alpha}{R_\alpha\sqrt{s/d}} \;=\; \sqrt{\frac{d}{s}},$$

which is (12). □

B.7 Proof of Theorem 5

Let $r(\alpha) = b\sqrt{2\alpha}$ be the camouflage radius, so that $r^2 = 2b^2\alpha$ and $dr/d\alpha = b^2/r$. By Corollary 4, the expected per-round damage at budget α is $g(\alpha) = A_{\text{fair}} \int_0^{r(\alpha)} f(m)\, dm$, the mass of target-group users within reach of the boundary, and the chain rule gives $g'(\alpha) = A_{\text{fair}}\, f(r)\, b^2/r$. Differentiating once more,

$$g''(\alpha) \;=\; A_{\text{fair}}\, \frac{d}{dr}\left[\frac{f(r)\, b^2}{r}\right] \frac{dr}{d\alpha} \;=\; A_{\text{fair}}\, \frac{b^4}{r^3}\big(r f'(r) - f(r)\big). \tag{18}$$

If f is non-increasing on $(0, r(A)]$ then $f'(r) \le 0$ and $f(r) \ge 0$, so the right-hand side of (18) is nonpositive and g is concave on $(0, A]$.

Fix a total budget $A = \sum_{t=1}^{T} \alpha_t$. By Jensen's inequality applied to the concave g,

$$\sum_{t=1}^{T} g(\alpha_t) \;\le\; T\, g\left(\frac{1}{T}\sum_{t=1}^{T} \alpha_t\right) \;=\; T\, g\left(\frac{A}{T}\right),$$

with equality if and only if $\alpha_t = A/T$ for every t; the uniform schedule is therefore optimal. If the schedule is supported on only W_{att} rounds, the same argument with T replaced by W_{att} gives total damage at most $W_{\text{att}}\, g(A/W_{\text{att}})$.

Specializing to a locally constant boundary density $f \equiv \rho_0$ gives $g(\alpha) = \rho_0\, b\sqrt{2\alpha}$, and the ratio of burst to uniform damage is

$$\frac{W_{\text{att}}\ \rho_0 b\sqrt{2A/W_{\text{att}}}}{T\ \rho_0 b\sqrt{2A/T}} \;=\; \frac{\sqrt{W_{\text{att}}}}{\sqrt{T}} \;=\; \sqrt{\frac{W_{\text{att}}}{T}} \;\le\; 1.$$

Bursting is thus strictly suboptimal whenever $W_{\text{att}} < T$, and the loss is by the factor $\sqrt{W_{\text{att}}/T}$. Note that no detector appears in this argument: the penalty is a consequence of the concavity of $\alpha \mapsto \sqrt{\alpha}$, not of increased detectability. The damage decomposition follows by pairing each group reassignment with A_{fair} and adding the reward-channel bound from Proposition 2. □

B.8 Proof of Theorem 6

Consider the sequential test that accumulates the log-likelihood ratio between the attacked and clean observation models. Under the attacked model, the expected increment per round is exactly the KL divergence between the two laws, so after n rounds the statistic has mean $n\,\alpha_{\text{obs}}$, where α_{obs} is the per-round divergence carried by the channel being monitored. A test that declares an attack when the statistic exceeds $\log(1/\delta_{\text{fp}})$ has false-positive probability at most δ_{fp} by Ville's inequality, and by the Chernoff–Stein lemma this threshold is attained after

$$n \;=\; \Theta\left(\frac{\log(1/\delta_{\text{fp}})}{\alpha_{\text{obs}}}\right)$$

rounds, which is optimal in the exponent.

Pooled monitoring. By the additivity in Lemma 1, the pooled per-round divergence is $\sum_j \psi(|\delta_j|/b) = \alpha_{\text{obs}}$ regardless of how the budget is distributed across coordinates. The pooled detection time is therefore $\Theta(\log(1/\delta_{\text{fp}})/\alpha_{\text{obs}})$ for every allocation: pooled monitoring cannot distinguish a concentrated attack from a diffuse one.

Per-coordinate monitoring. Running d separate tests, one per coordinate, and rejecting when any statistic crosses its threshold requires a Bonferroni correction: each test uses threshold $\log(d/\delta_{\text{fp}})$ to preserve the family-wise false-positive rate. The first test to fire is the one with the largest per-round divergence $\kappa_j = \psi(|\delta_j|/b)$, so the detection time is

$$n \;=\; \Theta\left(\frac{\log(d/\delta_{\text{fp}})}{\max_j \kappa_j}\right).$$

For the concentrated allocation the budget sits on s coordinates, so $\max_j \kappa_j = \alpha_{\text{obs}}/s$; for the diffuse allocation it is spread over all d, so $\max_j \kappa_j = \alpha_{\text{obs}}/d$. The ratio of per-coordinate detection times is therefore d/s: a concentrated attack is caught d/s times faster by per-coordinate testing than a diffuse one, at the cost of only a $\log d$ factor inside the logarithm. □

B.9 Proof of Proposition 3

Both the smoothing certificate of Theorem 2 and the Cauchy–Schwarz bound of Theorem 3 are *pointwise* in $\boldsymbol{\delta}$: they hold for every perturbation vector in $\mathcal{B}_\alpha$, with no reference to how that vector was chosen. In particular, they hold for a vector selected as a measurable function of the observed protected context $\tilde{\mathbf{x}}_t$.

The camouflage constraint $\boldsymbol{\delta}_t \in \mathcal{B}_\alpha$ binds *pathwise* — on every realisation, not merely in expectation — so an adaptive rule cannot borrow budget from favourable rounds to spend on unfavourable ones. Writing $\mathcal{D}_{\mathrm{ad}}$ for the class of adaptive rules and $\mathcal{D}_{\mathrm{fix}} \subset \mathcal{D}_{\mathrm{ad}}$ for the fixed allocations, the pointwise bounds give

$$\sup_{\boldsymbol{\delta} \in \mathcal{D}_{\mathrm{ad}}} \mathrm{Damage}(\boldsymbol{\delta}) \;\le\; B(\alpha),$$

where $B(\alpha)$ is the non-adaptive bound. Proposition 5 exhibits fixed allocations attaining $B(\alpha)$, so $\sup_{\mathcal{D}_{\mathrm{fix}}} \mathrm{Damage} = B(\alpha)$. Since a fixed allocation is a degenerate adaptive rule, $\mathcal{D}_{\mathrm{fix}} \subset \mathcal{D}_{\mathrm{ad}}$ implies $\sup_{\mathcal{D}_{\mathrm{fix}}} \le \sup_{\mathcal{D}_{\mathrm{ad}}}$. Combining the two inequalities,

$$\sup_{\mathcal{D}_{\mathrm{ad}}} \mathrm{Damage} \;=\; \sup_{\mathcal{D}_{\mathrm{fix}}} \mathrm{Damage} \;=\; B(\alpha)\,:$$

conditioning on the realised noise buys the adversary nothing. □

B.10 Proof of Proposition 4

Lemma 1 gives $\mathrm{KL}(P_{\boldsymbol{\eta}+\boldsymbol{\delta}} \,\|\, P_{\boldsymbol{\eta}}) = \sum_j \psi(|\delta_j|/b)$. Since $\psi(0) = 0$ and $\psi'(u) = 1 - e^{-u} > 0$ for $u > 0$, each summand is nonnegative and vanishes only at $\delta_j = 0$. Consequently, the total divergence is zero if and only if $\boldsymbol{\delta} = \mathbf{0}$.

The same holds for the quantity the learner monitors. An additive $\boldsymbol{\delta}$ translates the whole observation law, so $\alpha_{\mathrm{obs}} = \mathrm{KL}(\tau_{\boldsymbol{\delta}\#} P_0 \,\|\, P_0)$ with $P_0 = P_{\mathbf{x}} * p_b$, and this vanishes only if P_0 is invariant under translation by $\boldsymbol{\delta}$, which no probability measure on $\mathbb{R}^d$ is. Hence $\alpha_{\mathrm{obs}} > 0$ for every $\boldsymbol{\delta} \neq \mathbf{0}$, and the log-likelihood statistic of Theorem 6 drifts at this positive rate and crosses $\log(1/\delta_{\mathrm{fp}})$ after $\Theta(\log(1/\delta_{\mathrm{fp}})/\alpha_{\mathrm{obs}})$ rounds. Shrinking $\boldsymbol{\delta}$ lengthens this time without bound but never makes it infinite: evasion buys delay, not immunity. The argument uses additivity only through the translation step; for a general map $T_\# P_0 = P_0$ is possible with $T(\mathbf{z}) \neq \mathbf{z}$ (Appendix C). □

C Pathwise versus divergence constraints

Proposition 3 constrains the adversary pathwise, requiring $\|\boldsymbol{\delta}_t\|_2 \le R_\alpha$ on every realisation. This is not the same as the divergence constraint of (5). When $\boldsymbol{\delta}$ depends on $\tilde{\mathbf{x}}$, the observed variable is $T(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}} + \boldsymbol{\delta}(\tilde{\mathbf{x}})$ and its law is the pushforward $T_\# P$, not a translate of P, so Lemma 1 does not apply and the correct cost is $\mathrm{KL}(T_\# P \,\|\, P)$. That cost can be smaller than $\sup_{\mathbf{x}} \sum_j \psi(|\delta_j(\mathbf{x})|/b)$, sometimes much smaller, because displacement that moves mass within the support along the level sets of the density is cheap. The pathwise constraint is therefore strictly stronger than the divergence constraint, and Proposition 3 shows only that adaptivity buys nothing among rules obeying the stronger one.

Whether a divergence-constrained adaptive rule can beat the best fixed allocation is open, and we believe it can. The transport map $T(\mathbf{z}) = -\mathbf{z}$ satisfies $T_\# P = P$ exactly for symmetric Laplace noise while displacing every point, so zero divergence in the residual coordinate does not force zero displacement. Turning this into an attack requires a map that preserves the law of the privatised *context* rather than of the noise, which is a real constraint since the context distribution is not symmetric about the group boundary; we have no argument that it cannot be done. This is listed as open question (2) in Section 7.

D Additional experiments

D.1 Fixed versus adaptive allocation

Table 3 isolates the prediction of Proposition 3 by comparing Attack I, which fixes its direction in advance, against Attack IV, which chooses the direction as a function of the observed privatised

context. Across all six algorithms, the two variants differ by less than one standard error, with paired t tests giving $p > 0.6$ throughout, so the data provide no evidence that conditioning on the realised noise helps the adversary. This is what the theorem predicts: the camouflage constraint binds on every realisation, so an adaptive rule cannot borrow budget from rounds where the noise happens to be favourable and spend it on rounds where it is not, and the pointwise bounds of Theorems 2 and 3 apply to any perturbation regardless of how it was selected. For the two DP algorithms, the certified bound of Proposition 2 evaluates to 0.083, and both variants fall well inside it, at 0.051 and 0.049 for DP-LinUCB and 0.043 and 0.041 for DP-NCB. The bound is not available for the non-private algorithms, since L_{eff} is defined through the smoothing induced by the privacy noise.

Table 3: Fairness damage (ΔDP) under Attack I (fixed direction) vs. Attack IV (adaptive) for $\varepsilon = 1.0$, $\beta = 1.0$. Under Proposition 3, adaptivity cannot exceed the best fixed allocation; we test the stronger null of no effect and find no evidence that conditioning on the realized noise helps ($p > 0.6$ throughout).

Algorithm	**Attack I**	**Attack IV**	p**-value**	**Certified bound**
NeuralLinUCB	0.142 ± 0.008	0.139 ± 0.009	0.81	N/A (no DP)
NeuralUCB	0.168 ± 0.011	0.171 ± 0.010	0.85	N/A (no DP)
DP-LinUCB	0.051 ± 0.004	0.049 ± 0.004	0.74	0.083
FairBandit	0.187 ± 0.012	0.191 ± 0.013	0.83	N/A (no DP)
RobustBandit	0.156 ± 0.009	0.153 ± 0.010	0.82	N/A (no DP)
DP-NCB	0.043 ± 0.003	0.041 ± 0.003	0.67	0.083

D.2 Sensitivity Analysis: Privacy Parameter ε

Figure 5 shows fairness damage against the privacy parameter for DP algorithms on each dataset separately. For DP-LinUCB, ΔDP decreases with ε, as Proposition 1 requires, since the camouflage radius $R_\alpha = (\Delta_1/\varepsilon)\sqrt{2\alpha}$ shrinks as privacy weakens and leaves the adversary less room to displace contexts across the group boundary. Attack II dominates the other variants on Adult and COMPAS throughout the range, consistent with the spatial allocation result of Theorem 3; on German the concentrated and adaptive variants are closer; German has the largest d/s ratio (≈ 6.7) and hence the largest predicted advantage for concentration, so the closeness is not explained by the spatial theory and may reflect the different context distribution. The separation between attacked and clean curves narrows as ε grows on all three datasets, so the attack is most effective exactly in the strong-privacy regime that motivates deploying DP in the first place. DP-NCB preserves the ordering of attacks, but where damage falls steeply up to $\varepsilon \approx 0.75$ and then rises again on Adult, and shows a pronounced peak at $\varepsilon = 1$ on German. We attribute this to the interaction between noise scale and the neural model's representation learning.

Figure 6 depicts cumulative regret against the privacy parameter. Regret falls steeply in ε for both algorithms and all three datasets. The clean baseline falls at the same rate, so this is the estimation cost of the Laplace noise and not an effect of the attack. Attack II is uppermost, and Variants I, III and IV are close to one another. The regret penalty is modest in relative terms, at $\varepsilon = 0.25$. Attack II raises R_T by 10–13% over the clean run, and should be compared against the 20–55% increase in ΔDP the same attack produces. The attack therefore buys roughly four times as much fairness damage as it pays in visible performance loss, which is the precise sense in which it is stealthy: not that aggregate monitoring registers nothing, but that it registers far less than the group-level harm being inflicted. DP-NCB (bottom row) shows a plateau below $\varepsilon \approx 1$ followed by a sharp decline, the same threshold at which its fairness damage changes behaviour in Figure 5.

D.3 Sensitivity Analysis: Adversarial Parameter

Figure 7 presents fairness damage against the per-round KL budget for DP-LinUCB at $\varepsilon = 1$, separately by dataset. Damage increases with β for all variants. The growth is consistent with the prediction of Corollary 4: reassignment probability is first-order proportional to the displacement, which grows linearly in β since β scales the ℓ_1 budget. Over the range $\beta \in [0.1, 4]$ this predicts a factor of 40 in excess damage; the measured factors are 6.9 on German, 10 on COMPAS, and 4.9 on Adult, bracketing the prediction. On COMPAS and German, Attack II dominates from $\beta \approx 0.25$

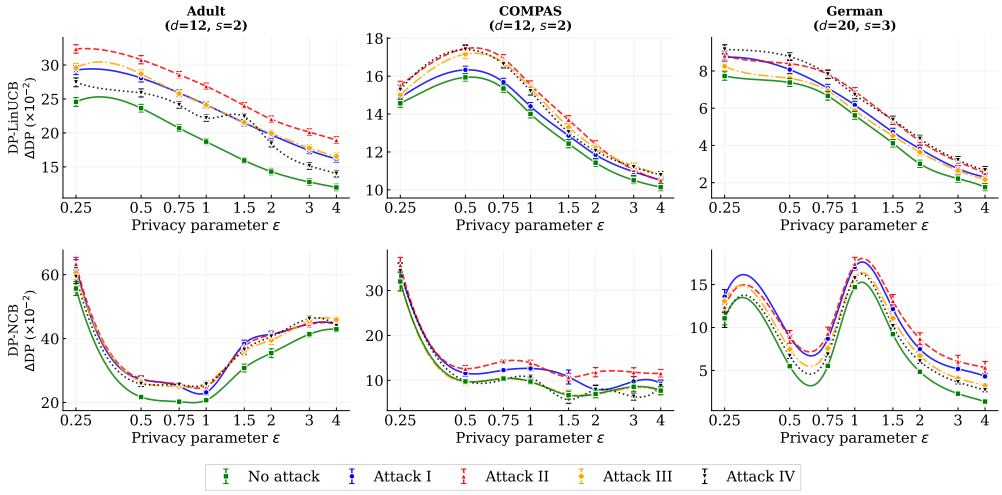


Figure 5: Fairness damage (ΔDP) vs. privacy parameter ε under all attack variants, for DP-LinUCB (top row) and DP-NCB (bottom row) on each of the three datasets. Attack II (Concentrated) dominates the other variants across most of the range, and the gap between attacked and clean curves narrows as ε grows, so the attack is most effective in the strong-privacy regime.

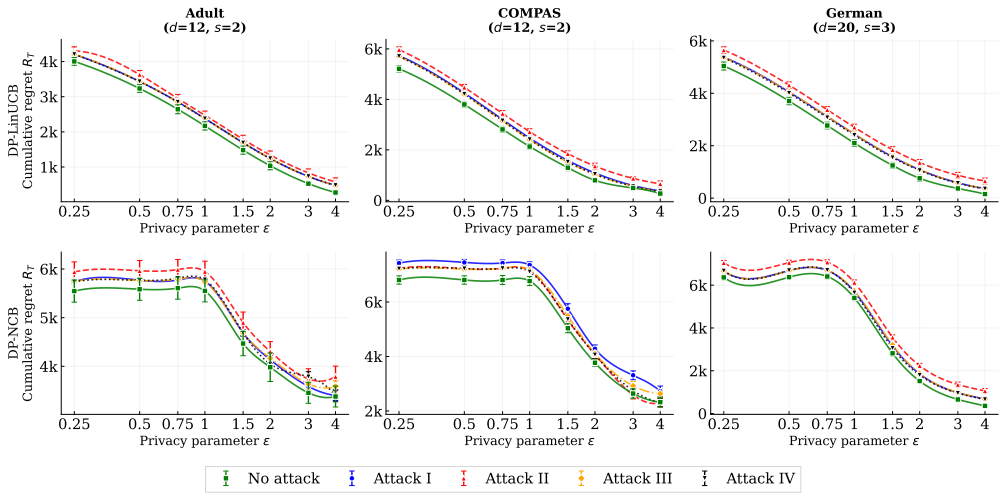


Figure 6: Global regret R_T vs. privacy parameter ε under all attack variants, for DP-LinUCB (top row) and DP-NCB (bottom row) on each of the three datasets. Attacked curves lie slightly above the clean baseline, with the largest gap under Attack II; the regret penalty is modest (10–13% at $\varepsilon = 0.25$) compared to the 20–55% increase in ΔDP the same attack produces.

upward, with Attack I following and the burst and adaptive variants well below, matching Theorem 3. On Adult, Attack IV leads at large β, with Attack II weakest below $\beta = 1$. Adult also has the highest baseline disparity of the three (23.1 against 15.0 and 6.3), so the attack operates against a group structure that is already substantially unequal and the marginal effect of any allocation is correspondingly smaller.

Figure 8 shows cumulative regret against the per-round KL budget for DP-LinUCB at $\varepsilon = 1$. Two regimes are visible. For $\beta \lesssim 0.5$ every variant tracks the no-attack baseline closely, with excess regret below 2% on all three datasets, so an operator monitoring aggregate performance has essentially nothing to observe. Beyond $\beta \approx 1$, the concentrated and diffuse variants separate upward (on COMPAS, Attack II reaches 2.62k against a baseline of 2.37k at $\beta = 4$) while the burst and adaptive variants remain near-flat. Comparing against Figure 7 quantifies the trade the attacker faces. At

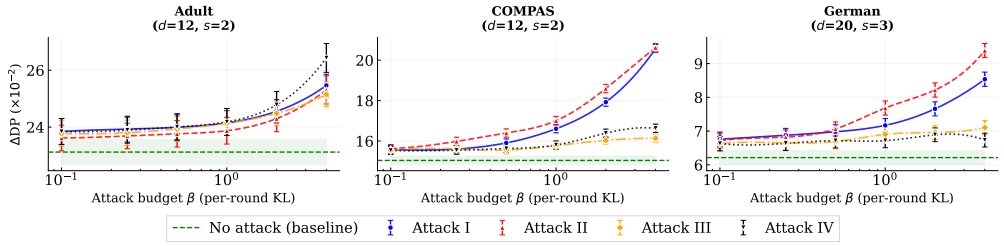


Figure 7: Fairness damage (ΔDP) vs. attack budget β for DP-LinUCB at ε=1.0.

$\beta = 4$ on COMPAS, Attack II inflicts 37% additional fairness damage for 10.5% additional regret, a ratio of 3.5; on German the same attack inflicts 50% additional damage for 7.6% additional regret, a ratio of 6.5. The corresponding ratios for Attack IV are 2.9 and 3.4. Attack II is therefore not merely the most damaging allocation but the most *efficient* one, extracting more unfairness per unit of visible performance loss than any other variant, which is the operationally relevant sense in which concentration is optimal, and which Theorem 3 predicts, since budget aligned with $\mathbf{w}$ contributes to boundary displacement while budget spent elsewhere contributes only to estimation error.

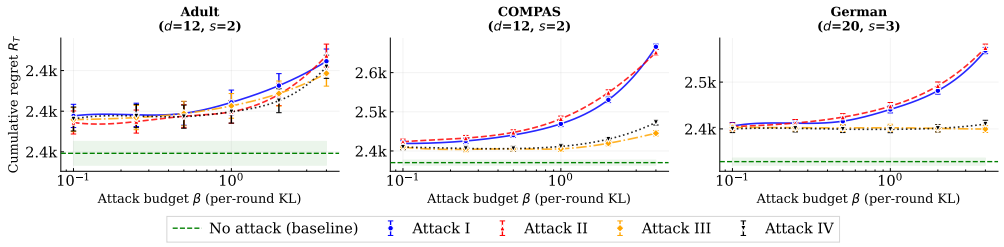


Figure 8: Global regret vs. attack budget β for DP-LinUCB at ε=1.0. Regret departs from the clean baseline for $\beta \gtrsim 1$.

D.4 Detection Analysis

Table 4 breaks the detection analysis down by dataset. Both detectors fire within the horizon on every DP-protected algorithm, but at very different speeds. Per-dimension monitoring detects the attack in 95–140 rounds, roughly one percent of $T = 10{,}000$, while aggregate monitoring needs 637–840. The ratio is the quantity Theorem 6 predicts, and it matches closely: 6.0 on Adult and COMPAS against $d/s = 6$, and 6.7 on German against $d/s \approx 6.7$. The agreement holds for both DP-LinUCB and DP-NCB on all three datasets. Pooling dilutes a perturbation concentrated on s coordinates across all d, so the pooled statistic accumulates evidence d/s times more slowly than the per-coordinate tests. We do not read an $s \log(d/\delta_{\text{fp}})$ dependence off these three datasets: German carries the largest sensitive subspace (s=3) yet is detected fastest, so with G, K and the context distribution varying alongside s, the three points do not isolate that term. The detection times are independent of the victim algorithm, since the detector reads the residual stream, which depends on the attack and the privacy noise but not on how the learner uses the contexts. Appendix D.5 shows the pooled test firing but slowing from 218 to 2830 rounds as d/s grows, while per-coordinate monitoring stays roughly constant—recovering the ds speedup Theorem 6 predicts. The two settings are therefore not directly comparable, and only the Iso/Aniso ratio transfers between them. The non-private algorithms are reported as N/D in the detection-time columns because both detectors take $\text{Lap}(b)$ as their null, which is undefined without privacy noise. Their Iso and Aniso ΔDP columns are populated because the flagging step still runs and marks whichever coordinates the pooled or per-coordinate statistic exceeds an empirical quantile, but the null is not calibrated by DP noise, so these values should be read as the effect of an uncalibrated flagging rule rather than as a comparison of two detection schemes.

Table 4: Per-dataset detection analysis for Attack II (concentrated, $\gamma = 0.9$) under aggregate (Iso) and per-coordinate (Aniso) monitoring. The Iso/Aniso ratio tracks d/s on every DP-protected row (Theorem 6). N/D: detector not applicable (no DP noise calibration)

.

Algorithm	**Clean** ΔDP	**Iso** ΔDP	**Aniso** ΔDP	**Flip rate**	**Det(Iso)**	**Det(Aniso)**	**Ratio**
Adult Income ($d = 12$, $s = 2$, $d/s = 6$)							
NeuralLinUCB	15.9±0.12	22.5±0.21	22.5±0.11	0.18	N/D	N/D	—
NeuralUCB	43.2±0.26	48.1±0.08	48.1±0.17	0.12	N/D	N/D	—
DP-LinUCB	23.7±0.16	29.4±0.15	25.1±0.25	0.15	810	135	6.0
FairBandit	16.5±0.23	24.8±0.13	24.8±0.15	0.20	N/D	N/D	—
RobustBandit	11.7±0.08	34.2±0.05	34.2±0.22	0.08	N/D	N/D	—
DP-NCB	15.7±0.11	31.3±0.14	27.0±0.24	0.15	840	140	6.0
COMPAS Recidivism ($d = 12$, $s = 2$, $d/s = 6$)							
NeuralLinUCB	19.5±0.20	24.8±0.24	24.8±0.16	0.20	N/D	N/D	—
NeuralUCB	8.2±0.14	11.5±0.09	11.5±0.07	0.14	N/D	N/D	—
DP-LinUCB	15.5±0.05	21.2±0.12	17.0±0.12	0.18	660	110	6.0
FairBandit	7.0±0.06	13.6±0.08	7.2±0.08	0.22	N/D	N/D	—
RobustBandit	9.5±0.1	12.8±0.14	12.8±0.11	0.10	N/D	N/D	—
DP-NCB	9.7±0.12	14.9±0.16	11.2±0.16	0.18	690	115	6.0
German Credit ($d = 20$, $s = 3$, $d/s \approx 6.7$)							
NeuralLinUCB	11.5±0.07	15.2±0.12	15.2±0.02	0.22	N/D	N/D	—
NeuralUCB	2.3±0.007	5.8±0.009	5.8±0.04	0.18	N/D	N/D	—
DP-LinUCB	6.6±0.05	12.4±0.7	8.1±0.03	0.31	637	95	6.7
FairBandit	1.9±0.008	7.5±0.006	1.9±0.005	0.25	N/D	N/D	—
RobustBandit	11.1±0.08	17.0±0.18	17.0±0.13	0.12	N/D	N/D	—
DP-NCB	17.7±0.15	24.5±0.21	19.3±0.14	0.31	670	100	6.7

D.5 SYNTHETIC d/s SWEEP

To isolate the effect of ambient dimension on the spatial-allocation advantage, we fix s=3 and sweep $d \in \{6, 12, 24, 48, 96\}$, giving $d/s \in \{2, 4, 8, 16, 32\}$. All other parameters are held constant (ε=1, β=1, T=10,000, algorithm = DP-LinUCB, N=50 runs). Because the group boundary normal is supported on the s sensitive coordinates with equal weight, the margin distribution stays fixed across d values and Corollary 4 predicts a flip-rate ratio of exactly $\sqrt{d/s}$.

Table 5 presents the flip rates under Attacks I and II, their ratio, and detection times under aggregate (Iso) and per-coordinate (Ani) monitoring. As d/s increases from 2 to 32, the flip-rate ratio rises from 1.43 to 4.67, broadly following the predicted $\sqrt{d/s}$ trend, although the empirical amplification increasingly falls below the theoretical value at larger dimensions. This gap is consistent with the finite-budget setting, in which the asymptotic amplification factor need not be attained exactly. Meanwhile, per-coordinate detection remains fast, decreasing from 108 to 88 rounds, whereas pooled detection slows substantially from 218 to 2830 rounds. Thus, concentrating the perturbation on the s sensitive coordinates both increases flip effectiveness and creates a sharper coordinate-level signature, while its signal becomes increasingly diluted under aggregate monitoring as the ambient dimension grows.

Table 5: Synthetic d/s sweep for flip amplification (Corollary 4) and detection speedup (Theorem 6). Values are means ± standard errors over N=50 runs.

			Flip amplification				*Detection speedup*	
d	s	d/s	flip(I)	flip(II)	ratio	$\sqrt{d/s}$	det(Iso)	det(Ani)
6	3	2	0.201±0.011	0.287±0.013	1.43	1.41	218±24	108±9
12	3	4	0.149±0.009	0.291±0.012	1.95	2.00	413±31	103±8
24	3	8	0.108±0.008	0.284±0.014	2.63	2.83	782±47	97±7
48	3	16	0.079±0.006	0.279±0.015	3.53	4.00	1510±89	92±6
96	3	32	0.058±0.005	0.271±0.016	4.67	5.66	2830±130	88±5

Figure 9 shows the synthetic d/s sweep on log-log axes with s=3 fixed and $d \in \{6, 12, 24, 48, 96\}$. Panel (a) confirms that the flip-rate ratio between Attack II (concentrated) and Attack I (diffuse) grows with d/s in the direction predicted by Corollary 4, but sits well below the $\sqrt{d/s}$ reference over the whole sweep, rising from 1.43 at $d/s = 2$ to only 4.67 at $d/s = 32$. Corollary 4 is a first-order expansion in $q \to 0^+$; at β=1 the displacements are large enough for the margin density to be no longer locally constant, and the finite-budget bound in Corollary 4 predicts exactly this saturation. Panel (b) shows that the $\Delta\mathrm{DP(II)}/\Delta\mathrm{DP(I)}$ ratio also grows with d/s, sitting slightly above the $\sqrt{d/s}$ reference and between the two dashed guides. The concentration advantage therefore carries from flip rates through to aggregate disparity, the step Corollary 4 leaves conditional on A_{fair}, with each reassignment contributing a comparable amount to the group-level gap. Panel (c) shows that the detection-time ratio Iso/Aniso tracks the $\propto d/s$ prediction of Theorem 6 closely across the whole sweep, growing from 2.02 at $d/s = 2$ to 32.2 at $d/s = 32$ and lying essentially on the reference line within the standard errors of Table 5. Pooling dilutes a signal concentrated on s coordinates across all d, and per-coordinate monitoring recovers the lost power at exactly the ratio d/s; the same concentration that amplifies group-label flips also sharpens the coordinate-level detection signature.

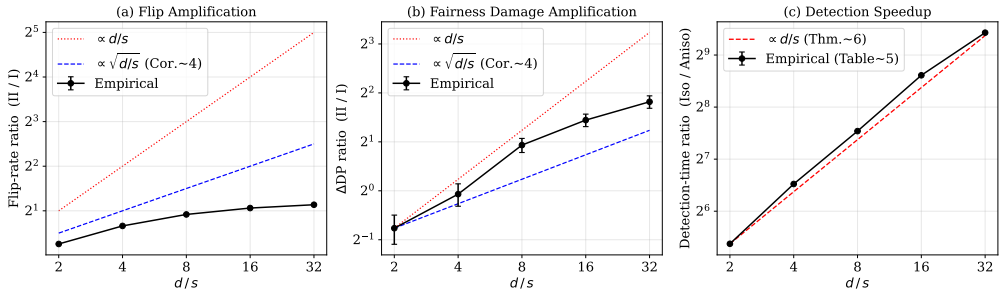


Figure 9: Synthetic d/s sweep on log-log axes with $s = 3$ fixed and $d \in \{6, 12, 24, 48, 96\}$. (a) Flip-rate ratio (II/I) vs. d/s: empirical growth (black) is in the predicted direction but below the $\sqrt{d/s}$ first-order value (blue dashed), which is an upper bound at finite budget (Corollary 4). (b) Fairness-damage ratio ΔDP(II)/ΔDP(I) vs. d/s: the concentration advantage carries through from flip rates to disparity, tracking between the $\sqrt{d/s}$ and d/s reference lines. (c) Detection-time ratio (Iso/Aniso) vs. d/s: the empirical ratio (black) tracks the $\propto d/s$ prediction of Theorem 6 (red dashed) closely across the whole sweep, growing from 2.02 at $d/s = 2$ to 32.2 at $d/s = 32$ and lying essentially on the reference line within the standard errors of Table 5.