Playing Against a DP-FTRL Learner in Repeated Zero-Sum Games

Yiheng Su

May 1, 2026

Abstract

We study the strategic consequences of differential privacy in repeated zero-sum games. The learner uses a tree-based differentially private Multiplicative Weights Update rule, equivalently private FTRL with the negative entropy regularizer, while the optimizer knows the learner's update rule and chooses the payoff sequence strategically. Using the regret guarantee of Agarwal and Singh [2017], we first translate the learner's privacy-aware regret bound into an upper bound on the optimizer's reward. We then show that, under a natural alternating-strategy condition on the game matrix, the optimizer can still obtain a matching $\Omega(\eta T)$ advantage over the game value against the noisy private dynamics. In particular, for the standard $\eta \asymp T^{-1/2}$ scaling, this gives an $\Omega(\sqrt{T})$ lower bound, up to the restrictions imposed by the privacy noise.

1 Introduction

Differential privacy is often studied as a constraint on the accuracy of an online learning algorithm. In full-information online linear optimization, Agarwal and Singh [2017] show that privacy can be added to FTRL with only a small additional cost: by releasing noisy prefix sums through a tree-based aggregation mechanism, one obtains an ϵ-differentially private learner with regret of order $O(\sqrt{T}) + \widetilde{O}(1/\epsilon)$. This result explains how privacy affects the learner's regret, but it leaves open a more game-theoretic question: how does the injected privacy noise change the ability of a strategic opponent to exploit the learning dynamics?

This project studies that question in repeated zero-sum matrix games. The learner observes full payoff vectors and updates with a private version of MWU, which is FTRL with the negative entropy regularizer. The optimizer knows the game matrix and the learner's update rule, and chooses a sequence of mixed strategies to maximize cumulative payoff. The benchmark is the value of the stage game, so the central quantity is the optimizer's excess reward above $T\,\mathrm{Val}(A)$.

The first step is to connect this strategic objective to the learner's regret. For any optimizer sequence, the optimizer's reward is at most the game-value benchmark plus the learner's regret. Therefore the tree-based DP-FTRL regret bound immediately gives an upper bound of

$$T\,\mathrm{Val}(A) + \frac{\log m}{2\eta} + \frac{\eta T}{2} + D_{\mathcal{D}'},$$

where $D_{\mathcal{D}'}$ is the additional error coming from the private tree-based perturbations.

The main lower-bound question is whether this upper bound is tight from the optimizer's perspective. We study an alternating optimizer strategy $x', x'', x', x'', \ldots$ whose average is a max-min strategy $x^\star$. For ordinary MWU, such alternating strategies are known to force regret of order

$\Omega(\eta T)$. The private setting adds a new difficulty: the learner reacts not to the exact cumulative payoff vector, but to a noisy tree-based prefix estimate. The analysis therefore controls the size of the tree perturbation and shows that, when the privacy noise is not too large relative to the step size, the same alternating construction still creates a positive variance term in the softmax path. This yields a lower bound of order $\Omega(\eta T)$, and hence $\Omega(\sqrt{T})$ for the usual choice $\eta \asymp T^{-1/2}$.

1.1 Preliminaries

Repeated Zero-Sum Matrix Games We consider a repeated zero-sum matrix game between an optimizer and a learner. Let $A \in \mathbb{R}^{n\times m}$ be the optimizer's payoff matrix. At each round $t = 1, \ldots, T$, the optimizer chooses a mixed strategy $x_t \in \Delta_n$, and the learner chooses a mixed strategy $y_t \in \Delta_m$, where Δ_n and Δ_m denote the probability simplices over the optimizer's and learner's action sets, respectively. In particular,

$$\Delta_d = \{x \in \mathbb{R}^d : x_i \geq 0,\ \sum_{i=1}^{d} x_i = 1\}.$$

The optimizer receives payoff $x_t^\top A y_t$, while the learner receives the opposite payoff. Equivalently, the learner's payoff matrix is $B = -A$. Thus, the game is zero-sum: any gain by the optimizer is exactly a loss for the learner.

The value of the game is defined as

$$\mathrm{Val}(A) = \max_{x\in\Delta_n} \min_{y\in\Delta_m} x^\top A y.$$

A max-min strategy $x^\star \in \Delta_n$ for the optimizer satisfies

$$(x^\star)^\top A y \geq \mathrm{Val}(A), \qquad \forall y \in \Delta_m.$$

In words, by playing $x^\star$, the optimizer guarantees a payoff of at least $\mathrm{Val}(A)$, no matter how the learner responds. For the learner, a min-max strategy $y^\star \in \Delta_m$ satisfies

$$x^\top A(y^\star) \leq \mathrm{Val}(A), \qquad \forall x \in \Delta_n.$$

In words, by playing $y^\star$, the learner guarantees that the optimizer's payoff is at most $\mathrm{Val}(A)$, no matter how the optimizer responds. Together, $x^\star$ and $y^\star$ form a Nash equilibrium. We use the value of the game as a benchmark to evaluate the performance of the optimizer's strategy.

As a canonical example, consider the rock-paper-scissors game with payoff matrix

$$A_{\mathrm{RPS}} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

One can verify that the unique equilibrium strategy is the uniform distribution

$$x^\star = y^\star = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right),$$

and the value of the game is $\mathrm{Val}(A_{\mathrm{RPS}}) = 0$. Therefore, in rock-paper-scissors, if both players use the uniform strategy, neither player can obtain a positive expected advantage.

The Learner The learner employs a fixed update Multiplicative Weights Update (MWU) rule which is a family member of the Follow-the-Regularized-Leader (FTRL) template with the negative entropy regularizer $h(y) = \sum_{i=1}^{m} y_i \log y_i$ [Arora et al., 2012].

The learner receives the full payoff vector for each round $g_t = B^\top x_t$ and keep record of the cumulative payoff vector $L_t = \sum_{s=1}^{t-1} g_s$. Then, at each round t, the learner updates her strategy according to the MWU rule:

$$y_t = \arg\max_{y \in \Delta_m} \{\eta \langle L_t, y \rangle - h(y)\} = \text{softmax}(\eta L_t), \qquad y_{t,i} = \frac{e^{\eta L_{t,i}}}{\sum_{j=1}^{m} e^{\eta L_{t,j}}}.$$

Now, the learner instead of using the exact cumulative payoff vector L_t, uses a noisy estimate $\widetilde{L}_t$ of L_t to update her strategy. In particular, we will analyze the tree-based DP-FTRL algorithm of Agarwal and Singh [2017], where the private prefix-sum estimates can be written as $\widetilde{L}_t = L_t + Z_t$, where Z_t is the tree-based private prefix-sum perturbation. We list the DP-MWU algorithms 1 here which are from Agarwal and Singh [2017, Algorithm 1 and 2] where we need to substitute the choice map $Q_h(u) = \arg\max_{y \in \Delta_m} \{\langle u, y \rangle - h(y)\}$ for the negative entropy regularizer h to get the DP-MWU algorithm. Agarwal and Singh [2017] showed that Algorithm 1 is ϵ-differentially private when we choose the distribution $\mathcal{D}$ to be the Laplace distribution with scale parameter $\lambda = O(\frac{\log T}{\epsilon})$.

Algorithm 1: DP-MWU for the Learner in the Full-Information Zero-Sum Game

Input: Noise distribution $\mathcal{D}$, regularizer $h(y)$, step size $\eta > 0$, horizon T
Initialize an empty binary tree $\mathcal{B}$ to compute differentially private estimates of $\sum_{s=1}^{t-1} B^\top x_s$.
Sample $z_0^1, \ldots, z_0^{\lceil \log T \rceil}$ independently from $\mathcal{D}$.
Set $\widetilde{L}_1 \leftarrow \sum_{i=1}^{\lceil \log T \rceil} z_0^i$.
for $t = 1$ **to** T **do**
  Choose $y_t = Q_h(\eta \widetilde{L}_t)$.
  Observe the optimizer's strategy $x_t \in \Delta_n$, and hence the payoff vector $B^\top x_t$.
  Receive payoff $x_t^\top B y_t$.
  Update $(\widetilde{L}_{t+1}, \mathcal{B}) \leftarrow \textsc{TreeBasedAgg}(B^\top x_t, \mathcal{B}, t, \mathcal{D}, T)$.

Algorithm 2: $\textsc{TreeBasedAgg}(B^\top x_t, \mathcal{B}, t, \mathcal{D}, T)$

Input: Current payoff vector $B^\top x_t$, binary tree $\mathcal{B}$, round t, noise distribution $\mathcal{D}$, horizon T
$(\widetilde{L}'_{t+1}, \mathcal{B}) \leftarrow \textsc{PrivateSum}(B^\top x_t, \mathcal{B}, t, \mathcal{D}, T)$ (Algorithm 5 Jain et al. [2012]), whether internal or leaf, is sampled independently from $\mathcal{D}$.
$s_t \leftarrow$ the binary representation of t as a string.
Find the minimum set S of already populated nodes in $\mathcal{B}$ that can compute $\sum_{s=1}^{t} B^\top x_s$.
Set $P \leftarrow |S| \le \lceil \log T \rceil$.
Set $r_t \leftarrow \lceil \log T \rceil - P$.
Sample $z_t^1, \ldots, z_t^{r_t}$ independently from $\mathcal{D}$.
Set $\widetilde{L}_{t+1} \leftarrow \widetilde{L}'_{t+1} + \sum_{i=1}^{r_t} z_t^i$.
Output: $(\widetilde{L}_{t+1}, \mathcal{B})$

The learner is measuring her performance through the notion of regret, which compares her cumulative payoff to the best fixed strategy in hindsight. Formally, the learner's regret after T

rounds is defined as

$$\text{Regret}_T^{\text{learner}} = \mathbb{E}\left[\max_{y\in\Delta_m} \sum_{t=1}^{T} x_t^\top B y - \sum_{t=1}^{T} x_t^\top B y_t\right]. \tag{1}$$

The Optimizer The optimizer is a strategic agent with *strategic foresight*. Unlike the learner, who reacts to past payoffs, the optimizer knows the game matrix A and the learner's update rule. The optimizer's goal is to select a sequence of strategies $(x_t)_{t=1}^T$ that maximizes his total reward

$$\text{Reward}_T^{\text{optimizer}} = \mathbb{E}\left[\sum_{t=1}^{T} x_t^\top A y_t\right]. \tag{2}$$

2 Related Work

Our work is closely related to regret analysis in adversarial online learning, especially for Follow-the-Regularized-Leader (FTRL) dynamics [Kwon and Mertikopoulos, 2017, Assos et al., 2024, Su and Vlatakis-Gkaragkounis, 2026]. A standard regret upper bound for an FTRL learner is given by Kwon and Mertikopoulos [2017]. For a strongly convex regularizer h, the learner's regret satisfies

$$\text{Regret}_T^{\text{learner}} \le \frac{h_{\max} - h_{\min}}{\eta} + \frac{\eta}{2\alpha}\sum_{t=1}^{T} \|g_t\|_*^2, \tag{3}$$

where

$$h_{\max} = \max_{y\in\Delta_m} h(y), \qquad h_{\min} = \min_{y\in\Delta_m} h(y),$$

and α is the strong convexity parameter of h.

For the private setting, Agarwal and Singh [2017] analyze a tree-based DP-FTRL algorithm. In this algorithm, the learner updates using private prefix-sum estimates rather than the exact cumulative payoff vectors. Their regret bound takes the form

$$\text{Regret}_T^{\text{learner}} \le \frac{h_{\max} - h_{\min}}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \max_{y\in\Delta_m} \|g_t\|^{*2}_{\nabla^2 h(y)} + D_{\mathcal{D}'}, \tag{4}$$

where $D_{\mathcal{D}'}$ captures the additional error introduced by the private tree-based perturbations.

Our work is also related to lower-bound results for adversarial play against learning dynamics. Assos et al. [2024] show that, when the learner uses ordinary Multiplicative Weights Update (MWU), equivalently FTRL with the negative entropy regularizer, the learner's regret can be lower bounded by $\Omega\left(\max\left\{\frac{1}{\eta}, \eta T\right\}\right)$ against a suitable adversarial optimizer. More recently, Su and Vlatakis-Gkaragkounis [2026] extend this type of lower bound beyond MWU to a broader class of FTRL dynamics with strongly convex regularizers, again showing a lower bound of the form $\Omega\left(\max\left\{\frac{1}{\eta}, \eta T\right\}\right)$.

These prior works provide regret lower bounds for ordinary, non-private FTRL learners. In contrast, our goal is to understand whether similar adversarial lower-bound arguments continue to hold when the learner uses tree-based DP-FTRL. The main technical challenge is that the learner

no longer follows the exact FTRL trajectory: each update is computed from a noisy private prefix-sum estimate. Therefore, the lower-bound argument must account for the interaction between the adversarial payoff sequence and the additional tree-based perturbations introduced for differential privacy.

3 Results

3.1 Upper Bound

In this section, we analyze the optimizer's reward from the optimizer's viewpoint, as defined in Equation (2). We first connect the learner's regret to the optimizer's reward and show that the optimizer's reward is upper bounded by the learner's regret.

Proposition 1. *For any sequence of strategies $(x_t)_{t=1}^T$ played by the optimizer, the optimizer's expected reward is upper bounded by*

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \le T \cdot \mathrm{Val}(A) + \mathrm{Regret}_T^{\mathrm{learner}}. \tag{5}$$

Proof. Let $y^\star \in \Delta_m$ be a min-max strategy of the learner. Then

$$x^\top A y^\star \le \mathrm{Val}(A), \qquad \forall x \in \Delta_n.$$

Hence, for any optimizer sequence $(x_t)_{t=1}^T$,

$$\min_{y \in \Delta_m} \sum_{t=1}^T x_t^\top A y \le \sum_{t=1}^T x_t^\top A y^\star \le T\,\mathrm{Val}(A).$$

Since $B = -A$, the learner's regret can be rewritten as

$$\begin{aligned}\mathrm{Regret}_T^{\mathrm{learner}} &= \max_{y \in \Delta_m} \sum_{t=1}^T x_t^\top B y - \sum_{t=1}^T x_t^\top B y_t \\ &= \sum_{t=1}^T x_t^\top A y_t - \min_{y \in \Delta_m} \sum_{t=1}^T x_t^\top A y.\end{aligned}$$

Therefore,

$$\sum_{t=1}^T x_t^\top A y_t = \mathrm{Regret}_T^{\mathrm{learner}} + \min_{y \in \Delta_m} \sum_{t=1}^T x_t^\top A y \le \mathrm{Regret}_T^{\mathrm{learner}} + T\,\mathrm{Val}(A).$$

Taking expectation over the learner's randomness gives

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \le T\,\mathrm{Val}(A) + \mathrm{Regret}_T^{\mathrm{learner}}.$$

□

Then, we can directly apply Agarwal and Singh [2017, Theorem 3.1], which gives an upper bound on $\mathrm{Regret}_T^{\mathrm{learner}}$, to obtain an upper bound on the optimizer's reward when the learner uses tree-based DP-MWU in Algorithm 1.

Theorem 2. *Assume that the learner uses tree-based DP-MWU. If* $\|A^\top x_t\|_\infty \le 1, \forall t \in [T]$, *then, for any optimizer strategy sequence* $(x_t)_{t=1}^T$,

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \le T\,\mathrm{Val}(A) + \frac{\log m}{2\eta} + \frac{\eta T}{2} + D_{\mathcal{D}'},$$

where

$$D_{\mathcal{D}'} = \mathbb{E}_{Z\sim\mathcal{D}'}\left[\max_{i\in[m]} Z_i - \min_{i\in[m]} Z_i\right].$$

Remark 1. *The term* $\frac{\log m}{2\eta} + \frac{\eta T}{2}$ *is actually the regret of a MWU learner without noise or privacy. The last term* $D_{\mathcal{D}'}$ *captures the additional error introduced by the private tree-based perturbations. If we choose the noise distribution* $\mathcal{D}$ *to be the Laplace distribution with scale parameter* $\lambda = O(\frac{\log T}{\epsilon})$, *then* $D_{\mathcal{D}'}$ *is of order* $\widetilde{O}(\frac{1}{\epsilon})$. *Choosing the standard optimal step size* $\eta \asymp (\frac{1}{\sqrt{T}})$ *gives the final upper bound* $T\,\mathrm{Val}(A) + O(\sqrt{T\log m}) + \widetilde{O}(\frac{1}{\epsilon})$.

3.2 Lower Bound

The next natural question would be whether this upper bound is tight. In particular, we want to show that the optimizer can achieve a reward of order $T\,\mathrm{Val}(A) + \Omega(\sqrt{T})$ against the DP-MWU learner, which would match the upper bound up to logarithmic factors and the privacy term. To establish such a lower bound, we will need to construct an adversarial optimizer strategy sequence $(x_t)_{t=1}^T$ that exploits the learner's update rule and the structure of the tree-based perturbations. This is the main technical challenge of our project.

We use the following alternating-strategy assumption.

Assumption 1. *Let* $x^\star \in \Delta_n$ *be a max-min strategy of the optimizer. There exist two optimizer strategies* $x', x'' \in \Delta_n$ *such that*

$$\frac{x' + x''}{2} = x^\star.$$

Moreover, there exist two learner best responses

$$i, j \in \mathrm{br}(x^\star) := \arg\min_{k\in[m]} (A^\top x^\star)_k$$

such that

$$(A^\top x')_i \neq (A^\top x')_j.$$

Equivalently, after relabeling i, j *if necessary,*

$$(A^\top x')_i > (A^\top x')_j, \qquad (A^\top x'')_i < (A^\top x'')_j.$$

For random payoff matrices with independent entries $A_{ij} \sim \mathrm{Unif}[-1, 1]$, Su and Vlatakis-Gkaragkounis [2026, Lemma E.4] shows that such a construction exists with probability

$$1 - \frac{n!m!}{(n+m-1)!}.$$

Hence, in the lower-bound analysis below, we assume that the payoff matrix admits such a pair x', x'' and best-response actions i, j. Since the padded tree perturbation is a sum of $\lceil \log T \rceil$ independent Laplace random variables in each coordinate, Bernstein's inequality for sub-exponential random variables gives the following sup-norm bound.

Lemma 3. *For each coordinate $k \in [m]$, suppose*

$$Z_k = \sum_{\ell=1}^{\lceil \log T \rceil} \xi_{\ell,k}, \qquad \xi_{\ell,k} \overset{\text{i.i.d.}}{\sim} \text{Lap}(\lambda).$$

Then there exists a universal constant $C > 0$ such that, for every $\rho \in (0,1)$, with

$$R_\rho = C\lambda \max\left\{ \sqrt{\lceil \log T \rceil \log \frac{2m}{\rho}}, \log \frac{2m}{\rho} \right\},$$

we have

$$\mathbb{P}\left(\|Z\|_\infty \le R_\rho \right) \ge 1 - \rho.$$

Proof. Fix a time k. Since

$$Z_k = \sum_{\ell=1}^{L} \xi_{\ell,k}, \qquad \xi_{\ell,k} \overset{\text{i.i.d.}}{\sim} \text{Lap}(\lambda),$$

Z_k is a sum of $\lceil \log T \rceil$ independent centered sub-exponential random variables. Equivalently, using the moment generating function of a centered Laplace random variable

$$\mathbb{E}e^{\theta \xi_{\ell,k}} = \frac{1}{1 - \lambda^2\theta^2}, \qquad |\theta| < \frac{1}{\lambda}.$$

For $|\theta| \le c/\lambda$, this implies

$$\mathbb{E}e^{\theta \xi_{\ell,k}} \le \exp(C_0 \lambda^2 \theta^2)$$

for a universal constant $C_0 > 0$. Therefore,

$$\mathbb{E}e^{\theta Z_k} \le \exp(C_0 \lceil \log T \rceil \lambda^2 \theta^2), \qquad |\theta| \le \frac{c}{\lambda}.$$

By the Chernoff bound, there exists a universal constant $c_0 > 0$ such that, for every $u \ge 0$,

$$\mathbb{P}(|Z_k| \ge u) \le 2 \exp\left(-c_0 \min\left\{ \frac{u^2}{\lceil \log T \rceil \lambda^2}, \frac{u}{\lambda} \right\} \right).$$

Applying the union bound over $k \in [m]$, we get

$$\mathbb{P}(\|Z\|_\infty \ge u) \le \sum_{k=1}^{m} \mathbb{P}(|Z_k| \ge u) \le 2m \exp\left(-c_0 \min\left\{ \frac{u^2}{\lceil \log T \rceil \lambda^2}, \frac{u}{\lambda} \right\} \right).$$

Now choose

$$u = R_\rho = C\lambda \max\left\{ \sqrt{\lceil \log T \rceil \log \frac{2m}{\rho}}, \log \frac{2m}{\rho} \right\},$$

with $C > 0$ large enough. Then

$$\min\left\{ \frac{R_\rho^2}{\lceil \log T \rceil \lambda^2}, \frac{R_\rho}{\lambda} \right\} \ge \frac{1}{c_0} \log \frac{2m}{\rho}.$$

Hence, $\mathbb{P}(\|Z\|_\infty \ge R_\rho) \le \rho$. Equivalently, $\mathbb{P}(\|Z\|_\infty \le R_\rho) \ge 1 - \rho$. □

With this lemma, we can show the following lower bound for the optimizer's reward against the DP-MWU learner.

Theorem 4. *Assume that T is even and that the learner uses tree-based DP-MWU. Assume also that the tree-based mechanism is padded so that each perturbation Z_t has the same marginal law $\mathcal{D}'$. Let x', x'' and i, j satisfy Assumption 1, and define $\Delta_{ij} := \left|(A^\top x')_i - (A^\top x')_j\right| > 0$.*

For any $\rho \in (0,1)$, define

$$R_\rho = C\lambda \max\left\{\sqrt{\lceil \log T \rceil \log \frac{2m}{\rho}}, \log \frac{2m}{\rho}\right\}$$

from Lemma 3. If we choose the step size

$$\eta \le \frac{1}{2m\left(R_\rho + \|A^\top x'\|_\infty\right)},$$

then the reward of the optimizer is lower bounded by

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \ge T\,\mathrm{Val}(A) + \frac{\eta T}{8m}(1-\rho)\Delta_{ij}^2.$$

Proof. Recall that the learner's payoff vector is

$$g_t = B^\top x_t = -A^\top x_t, \qquad L_t = \sum_{\tau=1}^{t-1} g_\tau.$$

Under Assumption 1, the optimizer plays

$$x_{2s+1} = x', \qquad x_{2s+2} = x'', \qquad s = 0, 1, \ldots, T/2 - 1.$$

Since $\frac{x'+x''}{2} = x^\star$, we have

$$A^\top x' + A^\top x'' = 2A^\top x^\star.$$

Therefore, $L_{2s+1} = -2sA^\top x^\star$, and $L_{2s+2} = -2sA^\top x^\star - A^\top x'$.

The tree-based DP-MWU learner plays $y_t = \mathrm{softmax}\big(\eta(L_t + Z_t)\big)$. Hence

$$y_{2s+1} = \mathrm{softmax}\Big(-2\eta s A^\top x^\star + \eta Z_{2s+1}\Big) \quad \text{and} \quad y_{2s+2} = \mathrm{softmax}\Big(-2\eta s A^\top x^\star - \eta A^\top x' + \eta Z_{2s+2}\Big).$$

We first lower bound the reward over one pair of rounds. Using $A^\top x'' = 2A^\top x^\star - A^\top x'$, we get

$$\begin{aligned} x'^\top A y_{2s+1} + x''^\top A y_{2s+2} &= \langle A^\top x', y_{2s+1}\rangle + \langle A^\top x'', y_{2s+2}\rangle \\ &= 2\langle A^\top x^\star, y_{2s+2}\rangle + \langle A^\top x', y_{2s+1}\rangle - \langle A^\top x', y_{2s+2}\rangle. \end{aligned}$$

Since $x^\star$ is max-min,

$$\langle A^\top x^\star, y\rangle = (x^\star)^\top A y \ge \mathrm{Val}(A), \qquad \forall y \in \Delta_m.$$

Therefore,

$$x'^\top A y_{2s+1} + x''^\top A y_{2s+2} \ge 2\,\mathrm{Val}(A) + \langle A^\top x', y_{2s+1}\rangle - \langle A^\top x', y_{2s+2}\rangle.$$

It remains to lower bound the final difference in expectation. Define

$$F_s(z) := \langle A^\top x', y_{2s+1}\rangle = \Big\langle A^\top x', \mathrm{softmax}\Big(-2\eta s A^\top x^\star + \eta z\Big)\Big\rangle$$

and

$$G_s(z) := \langle A^\top x', y_{2s+2}\rangle = \Big\langle A^\top x', \mathrm{softmax}\Big(-2\eta s A^\top x^\star - \eta A^\top x' + \eta z\Big)\Big\rangle.$$

By the padding assumption on the tree-based mechanism, Z_{2s+1} and Z_{2s+2} have the same marginal law $\mathcal{D}'$. Hence

$$\mathbb{E}\Big[\langle A^\top x', y_{2s+1}\rangle - \langle A^\top x', y_{2s+2}\rangle\Big] = \mathbb{E}[F_s(Z_{2s+1})] - \mathbb{E}[G_s(Z_{2s+2})] = \mathbb{E}_{Z\sim\mathcal{D}'}[F_s(Z) - G_s(Z)].$$

For fixed Z, define the interpolation path

$$y_{s,Z}(r) := \mathrm{softmax}\Big(-2\eta s A^\top x^\star + \eta Z - r\eta A^\top x'\Big), \qquad r \in [0,1].$$

Then

$$F_s(Z) - G_s(Z) = \langle A^\top x', y_{s,Z}(0)\rangle - \langle A^\top x', y_{s,Z}(1)\rangle.$$

For the negative entropy regularizer, differentiating the softmax path gives

$$\frac{d}{dr}\langle A^\top x', y_{s,Z}(r)\rangle = -\eta \,\mathrm{Var}_{I\sim y_{s,Z}(r)}\Big((A^\top x')_I\Big).$$

Therefore,

$$F_s(Z) - G_s(Z) = \eta \int_0^1 \mathrm{Var}_{I\sim y_{s,Z}(r)}\Big((A^\top x')_I\Big)\, dr.$$

Now define the event $E_\rho := \{\|Z\|_\infty \le R_\rho\}$. By Lemma 3, $\mathbb{P}_{Z\sim\mathcal{D}'}(E_\rho) \ge 1-\rho$.

We next lower bound the variance term on E_ρ. By Assumption 1, the coordinates i, j belong to the learner's best-response set against $x^\star$. Equivalently, i and j are maximizers of $-A^\top x^\star$. Hence

$$\mathrm{softmax}(-2\eta s A^\top x^\star)_i \ge \frac{1}{m}, \qquad \mathrm{softmax}(-2\eta s A^\top x^\star)_j \ge \frac{1}{m}.$$

Moreover, the softmax map is 1-Lipschitz from ℓ_∞ to ℓ_1. Thus, for $k \in \{i,j\}$,

$$y_{s,Z,k}(r) \ge \mathrm{softmax}(-2\eta s A^\top x^\star)_k - \eta\|Z - rA^\top x'\|_\infty \ge \frac{1}{m} - \eta\Big(\|Z\|_\infty + \|A^\top x'\|_\infty\Big).$$

On E_ρ, this implies

$$y_{s,Z,k}(r) \ge \frac{1}{m} - \eta\Big(R_\rho + \|A^\top x'\|_\infty\Big), \qquad k \in \{i,j\}.$$

By the step-size condition

$$\eta \le \frac{1}{2m\,(R_\rho + \|A^\top x'\|_\infty)},$$

we obtain

$$y_{s,Z,i}(r) \ge \frac{1}{2m}, \qquad y_{s,Z,j}(r) \ge \frac{1}{2m}, \qquad \forall r \in [0,1],$$

on the event E_ρ.

For any probability vector $p \in \Delta_m$, if $p_i, p_j \ge 1/(2m)$, then

$$\mathrm{Var}_{k\sim p}\Big((A^\top x')_k\Big) \ge \frac{p_i p_j}{p_i + p_j}\Big((A^\top x')_i - (A^\top x')_j\Big)^2 \ge \frac{1}{4m}\Delta_{ij}^2.$$

Therefore, on E_ρ,

$$\mathrm{Var}_{I\sim y_{s,Z}(r)}\Big((A^\top x')_I\Big) \geq \frac{1}{4m}\Delta_{ij}^2, \qquad \forall r \in [0,1].$$

Consequently,

$$\begin{aligned}\mathbb{E}\Big[\langle A^\top x', y_{2s+1}\rangle - \langle A^\top x', y_{2s+2}\rangle\Big] &= \eta \mathbb{E}_{Z\sim\mathcal{D}'} \int_0^1 \mathrm{Var}_{k\sim y_{s,Z}(r)}\Big((A^\top x')_k\Big)\, dr \\ &\geq \eta \mathbb{P}_{Z\sim\mathcal{D}'}(E_\rho)\frac{1}{4m}\Delta_{ij}^2 \geq \frac{\eta(1-\rho)}{4m}\Delta_{ij}^2.\end{aligned}$$

Combining this with the pairwise reward lower bound gives

$$\mathbb{E}\Big[x'^\top A y_{2s+1} + x''^\top A y_{2s+2}\Big] \geq 2\,\mathrm{Val}(A) + \frac{\eta(1-\rho)}{4m}\Delta_{ij}^2.$$

Finally, summing over $s = 0, 1, \ldots, T/2-1$, and using the definition of $\mathrm{Reward}_T^{\mathrm{optimizer}}$ as the expected cumulative optimizer reward, we obtain

$$\begin{aligned}\mathrm{Reward}_T^{\mathrm{optimizer}} &= \sum_{s=0}^{T/2-1} \mathbb{E}\Big[x'^\top A y_{2s+1} + x''^\top A y_{2s+2}\Big] \\ &\geq T\,\mathrm{Val}(A) + \frac{T}{2}\cdot\frac{\eta(1-\rho)}{4m}\Delta_{ij}^2 \\ &= T\,\mathrm{Val}(A) + \frac{\eta T}{8m}(1-\rho)\Delta_{ij}^2.\end{aligned}$$

This proves the theorem. □

Remark 2. *This is the final remark on the lower bound above. For the standard tree mechanism, the Laplace scale is typically chosen as*

$$\lambda = O\left(\frac{\log T}{\varepsilon}\right).$$

Therefore,

$$R_\rho = O\left(\frac{\log T}{\varepsilon}\max\left\{\sqrt{\log T \log\frac{2m}{\rho}}, \log\frac{2m}{\rho}\right\}\right).$$

In particular, if m, ρ, ε are treated as constants, then

$$R_\rho = O\left(\frac{(\log T)^{3/2}}{\varepsilon}\right).$$

Hence $R_\rho + \|A^\top x'\|_\infty = o(\sqrt{T})$, and the step-size condition

$$\eta \leq \frac{1}{2m(R_\rho + \|A^\top x'\|_\infty)}$$

is satisfied by the standard choice $\eta \asymp T^{-1/2}$ for all sufficiently large T. Consequently, Theorem 4 gives

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \geq T\,\mathrm{Val}(A) + \frac{c(1-\rho)\Delta_{ij}^2}{8m}\sqrt{T} = T\,\mathrm{Val}(A) + \Omega(\sqrt{T}).$$

4 Experiments

We run two experiments to illustrate the theoretical results above.

The first experiment considers the Rock–Paper–Scissors game. The optimizer plays the max-min strategy $x^\star$, while the learner uses DP-MWU. In this case, the optimizer can only obtain the baseline reward $T\,\mathrm{Val}(A)$. Indeed, for Rock–Paper–Scissors, the payoff vector $A^\top x^\star$ assigns the same payoff $\mathrm{Val}(A)$ to every learner action. Therefore, regardless of the learner's mixed strategy y_t, we have

$$(x^\star)^\top A y_t = \mathrm{Val}(A), \qquad \forall t.$$

Thus the cumulative optimizer reward is exactly $T\,\mathrm{Val}(A)$. See Figure 1 for the experimental results, which are consistent with this theoretical prediction.

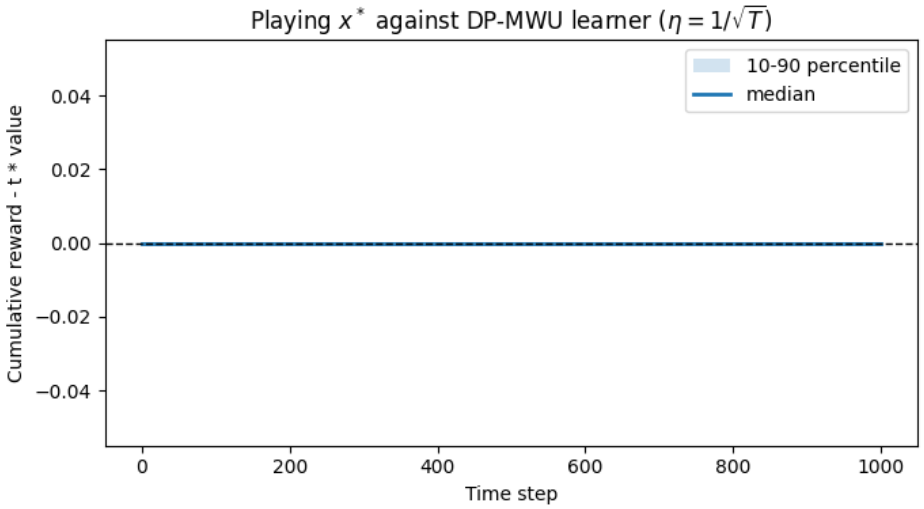

Figure 1: DP-MWU against the max-min optimizer strategy in the Rock–Paper–Scissors game. The optimizer reward remains at the baseline $T\,\mathrm{Val}(A)$.

The second experiment considers the alternating optimizer strategy from Assumption 1. In this case, the optimizer can exploit the DP-MWU learner and obtain reward above the game-value baseline. The cumulative reward grows as

$$T\,\mathrm{Val}(A) + \Omega(\sqrt{T}),$$

when the step size is chosen as $\eta \asymp T^{-1/2}$. This behavior matches the lower bound in Theorem 4, which predicts an optimizer surplus of order $\Omega(\eta T) = \Omega(\sqrt{T})$. See Figure 2 for the experimental results, which are consistent with our theoretical prediction.

5 Future Directions

One thing I want to point out is that the upper bound on the reward from Theorem 2 is

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \le T\,\mathrm{Val}(A) + \frac{\log m}{2\eta} + \frac{\eta T}{2} + D_{\mathcal{D}'},$$

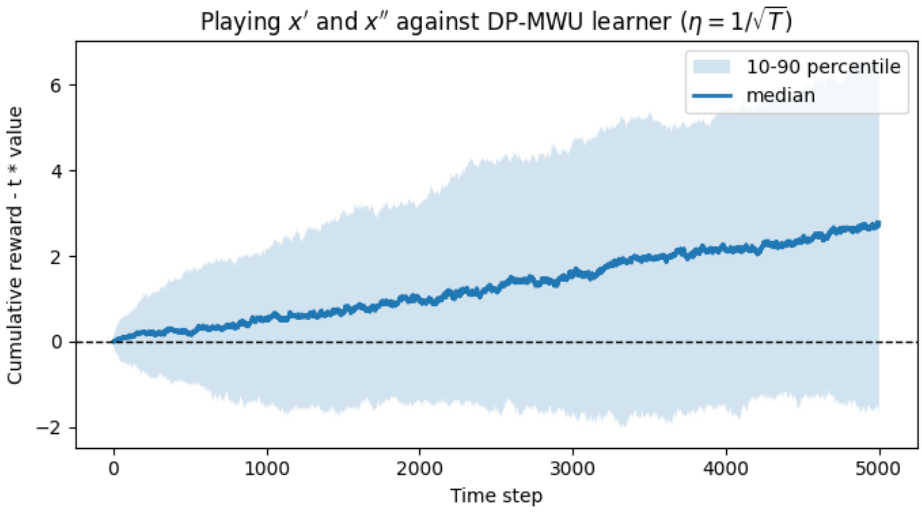

Figure 2: DP-MWU against the alternating optimizer strategy. The optimizer obtains positive surplus above $T\,\mathrm{Val}(A)$, consistent with the $\Omega(\sqrt{T})$ lower bound.

which has two parts that depend on the step size η. The optimal choice of η to minimize the upper bound is $\eta \asymp \sqrt{\frac{1}{T}}$. What we have shown is that the reward of the optimizer can be lower bounded by

$$\mathrm{Reward}_T^{\mathrm{optimizer}} \geq T\,\mathrm{Val}(A) + \Omega(\eta T).$$

If we choose $\eta \asymp \sqrt{\frac{1}{T}}$, then the lower bound matches the upper bound. However, if we want the step size to be of the form $\eta \asymp T^{-\beta}$ with $\beta \in [0, 1]$, we still need to show that the optimizer can also achieve a reward of order $T\,\mathrm{Val}(A) + \Omega(\frac{1}{\eta})$. Then, we can conclude that the optimizer can achieve a reward of order $T\,\mathrm{Val}(A) + \Omega(\max\{\eta T, \frac{1}{\eta}\})$ against the DP-MWU learner, which would match the upper bound up to logarithmic factors and the privacy term for all choices of η. This is one future direction that I want to explore.

The second point comes from Professor Brown's question: Can we come up with a more advanced DP-MWU learner that can avoid being exploited by the alternating optimizer strategy? This would be another interesting future direction to explore.

The last point is that we only consider the case where the learner uses DP-MWU. It would be more interesting if we could generalize the lower bound to the DP-FTRL family.

References

Naman Agarwal and Karan Singh. The price of differential privacy for online learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 32–40. PMLR, 06–11 Aug 2017. URL https://arxiv.org/pdf/1701.07953.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a

meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012. URL https://theoryofcomputing.org/articles/v008a006/.

Angelos Assos, Yuval Dagan, and Constantinos Daskalakis. Maximizing utility in multi-agent environments by anticipating the behavior of other learners, 2024. URL https://arxiv.org/abs/2407.04889.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 24.1–24.34, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/jain12.html.

Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017. ISSN 2164-6066. doi: 10.3934/jdg.2017008. URL https://arxiv.org/abs/1401.6956.

Yiheng Su and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the exploitability of ftrl dynamics, 2026. URL https://arxiv.org/abs/2604.05129.