15-888 Computational Game Solving (Fall 2023) Released on: Oct. 7, 2023

Homework 2

Student: [*** Your Andrew ID here ***] **Due on: Oct. 28, 2023, beginning of class**

Instructions Please submit your homework on Gradescope. The Gradescope entry code can be found on Piazza. On Gradescope, please submit a single `pdf` file containing **both** your written solutions **and** your code attached to the end (for example, using a `verbatim` environment). You can (but are not required to) typeset your written solutions in the `.tex` file provided in the homework `.zip`.

Bonus points This homework contains 3 problems (each split into smaller tasks), for a total of 125 points. The maximum score for the homework will be capped at 100. In other words, your score will be set to the minimum between 100 and the sum of the points you scored on all questions, including the bonus questions.

1 No-Regret Learning: OMWU, FTRL, OMD (35 points + 12 bonus)

Multiplicative weights update (MWU) and its predictive variant called *optimistic multiplicative weights update (OMWU)* are popular regret minimization algorithms for the probability simplex

$$\Delta^n \coloneqq \{(x_1, \dots, x_n) \in \mathbb{R}^n_{\geq 0} : x_1 + \cdots + x_n = 1\}. \tag{1}$$

They enjoy many strong theoretical properties, and were involved in a series of important papers in game theory. In this question, you will derive and analyze MWU and OMWU from first principles.

1.1 The negative entropy regularizer (10 points)

OMWU is just a special name for the predictive FTRL algorithm when the regularizer $\omega : \Delta^n \to \mathbb{R}$ is chosen to be the *negative entropy regularizer*

$$\omega(\boldsymbol{x}) \coloneqq \sum_{i=1}^{n} x_i \log(x_i).$$

To avoid annoying issues with the logarithm of 0, we will only ever evaluate and differentiate ω in the relative interior of Δ^n, that is the set

$$\operatorname{relint} \Delta^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n_{>0} : x_1 + \cdots + x_n = 1\}$$

(note the strict inequality $\mathbb{R}_{>0}$, as opposed to $\mathbb{R}_{\geq 0}$ in (1)).

For ω to be a valid choice of regularizer in predictive FTRL, we need to check that ω is 1-strongly convex with respect to some norm. In particular, it turns out that ω is 1-strongly convex both with respect to the Euclidean norm

$$\|\boldsymbol{x}\|_2 \coloneqq \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^n$$

and with respect to the ℓ_1 norm

$$\|\boldsymbol{x}\|_1 \coloneqq \sum_{i=1}^{n} |x_i| \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

The easiest way to verify strong convexity in this case passes through the following well-known characterization.

Lemma 1. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set, $f : \mathcal{X} \to \mathbb{R}$ be a twice-differentiable function with Hessian matrix $\nabla^2 f(\boldsymbol{x})$ at every $\boldsymbol{x} \in \mathcal{X}$, and $\|\cdot\|$ be a norm. If

$$\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x})\, \boldsymbol{u} \ge \|\boldsymbol{u}\|^2 \qquad \forall\, \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{x} \in \mathcal{X},$$

then f is 1-strongly convex on $\mathcal{X}$ with respect to norm $\|\cdot\|$.

In the next two exercises you will use Lemma 1 to verify that ω is 1-strongly convex on relint Δ^n with respect to $\|\cdot\|_2$ and $\|\cdot\|_1$.

Problem 1.1 (5 points)**.** Apply Lemma 1 for $\mathcal{X} = \operatorname{relint} \Delta^n$, $f = \omega$, and $\|\cdot\| = \|\cdot\|_2$ and conclude that ω is 1-strongly convex with respect to the Euclidean norm on relint Δ^n.

★ Hint: the Hessian matrix of ω is particularly nice. Start by working that out first.
★ Hint: at some point, it might be useful to argue that $1/x_i \ge 1$ for any $i \in \{1, \dots, n\}$ whenever $\boldsymbol{x} \in \operatorname{relint} \Delta^n$.

Solution. [*** Your solution here ***] □

Problem 1.2 (5 points)**.** Apply Lemma 1 for $\mathcal{X} = \operatorname{relint} \Delta^n$, $f = \omega$, and $\|\cdot\| = \|\cdot\|_1$ and conclude that ω is 1-strongly convex with respect to the ℓ_1 norm on relint Δ^n.

★ Hint: The Cauchy-Schwarz inequality asserts that for any pair of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$,

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right). \tag{2}$$

Now, let $\boldsymbol{x} \in \operatorname{relint} \Delta^n$ and $\boldsymbol{u} \in \mathbb{R}^n$, and consider the vectors $\boldsymbol{a} \coloneqq (|u_1|/\sqrt{x_1}, \dots, |u_n|/\sqrt{x_n})$ and $\boldsymbol{b} \coloneqq (\sqrt{x_1}, \dots, \sqrt{x_n})$. What happens if you plug them into (2)? Don't forget that $x_1 + \cdots + x_n = 1$ since $\boldsymbol{x} \in \operatorname{relint} \Delta^n$.

Solution. [*** Your solution here ***] □

1.2 Gradient of ω and of its conjugate (15 points)

In this subsection, you will derive a formula for the gradient of ω and for the gradient of its convex conjugate. Let's start with the gradient.

Problem 1.3 (5 points)**.** Give an expression for the gradient of ω at any point $\boldsymbol{x} \in \operatorname{relint} \Delta^n$.

Solution. [*** Your solution here ***] □

Now, let's focus on the gradient of the convex conjugate of ω, that is, the solution to the optimization problem

$$\nabla\omega^*(\boldsymbol{g}) := \operatorname*{arg\,max}_{\hat{\boldsymbol{x}}\in\operatorname{relint}\Delta^n} \{\boldsymbol{g}^\top\hat{\boldsymbol{x}} - \omega(\hat{\boldsymbol{x}})\}. \tag{3}$$

Problem (3) is a *constrained* optimization problem, since the optimization variable $\hat{\boldsymbol{x}}$ is constrained to satisfy $\hat{\boldsymbol{x}} \in \operatorname{relint}\Delta^n$. Call $\boldsymbol{x}^*$ the optimal solution to (3). As a result of an important theorem in optimization theory (the Lagrange multiplier theorem), there exists a constant (called *Lagrange multiplier*) $\alpha \in \mathbb{R}$ such that

$$\boldsymbol{g} - \nabla\omega(\boldsymbol{x}^*) = \alpha\mathbf{1}, \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones.

Problem 1.4 (5 points). Plug in the expression for the gradient of ω that you developed in Problem 1.3 into (4). Note that (4) is a vector equation, and therefore it is equivalent to a system of n scalar equations. Isolate and solve for x_i^* for every $i \in \{1,\ldots,n\}$, and show that

$$x_i^* = e^{-1-\alpha}\cdot e^{g_i} \qquad \forall\, i \in \{1,\ldots,n\}. \tag{5}$$

Solution. [*** Your solution here ***] □

Problem 1.5 (5 points). Use (5) together with the fact that the sum of the entries of $\boldsymbol{x}^* \in \operatorname{relint}\Delta^n$ must be 1 to solve for the value of the Lagrange multiplier α. Then, plug in the value of α to conclude that for any $\boldsymbol{g} \in \mathbb{R}^n$, $\nabla\omega^*(\boldsymbol{g})$—that is, the solution to the argmax in (3)—satisfies

$$x_i^* = \frac{e^{g_i}}{\sum_{j=1}^n e^{g_j}} \qquad \forall\, i \in \{1,\ldots,n\}.$$

Solution. [*** Your solution here ***] □

1.3 OMWU as predictive FTRL (10 points)

Now that we verified that ω is 1-strongly convex, we can safely run predictive FTRL with ω as a regularizer. As a reminder, predictive FTRL is the algorithm recalled in Algorithm 1. In our case, $\mathcal{X}$ will be the relative interior $\operatorname{relint}\Delta^n$ of the probability simplex Δ^n, the regularizer φ will be the negative entropy function ω, and $\eta > 0$ will be a generic stepsize. The resulting algorithm is called OMWU.

Problem 1.6 (6 points). Use the characterization of $\nabla\omega^*(\boldsymbol{g})$ given in the statement of Problem 1.5 to prove that at times $t = 2, 3, \ldots,$ for all $i \in \{1,\ldots,n\}$, the strategies $\boldsymbol{x}^t \in \Delta^n$ produced by OMWU satisfy[a]

$$x_i^t = \frac{x_i^{t-1}\,\exp\{\eta(\ell_i^{t-1} - m_i^{t-1} + m_i^t)\}}{\sum_{j=1}^n x_j^{t-1}\,\exp\{\eta(\ell_j^{t-1} - m_j^{t-1} + m_j^t)\}},$$

[a] For readability we used the notation $\exp\{\bullet\}$ to mean $e^{\bullet}$.

```
Algorithm 1: Predictive FTRL
  Data: $\mathcal{X} \subseteq \mathbb{R}^n$ convex domain
        $\varphi : \mathcal{X} \to \mathbb{R}_{\geq 0}$ strongly convex regularizer
        $\eta > 0$ step-size parameter
  $\boldsymbol{L}^0 \leftarrow \mathbf{0} \in \mathbb{R}^n$
  function NextStrategy($\boldsymbol{m}^t$)
      [▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version]
      return $\arg\max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \left\{ (\boldsymbol{L}^{t-1} + \boldsymbol{m}^t)^\top \hat{\boldsymbol{x}} - \frac{1}{\eta}\varphi(\hat{\boldsymbol{x}}) \right\}$
  function ObserveUtility($\boldsymbol{\ell}^t$)
      $\boldsymbol{L}^t \leftarrow \boldsymbol{L}^{t-1} + \boldsymbol{\ell}^t$
```

```
Algorithm 2: Predictive OMD
  Data: $\mathcal{X} \subseteq \mathbb{R}^n$ convex domain
        $\varphi : \mathcal{X} \to \mathbb{R}_{\geq 0}$ strongly convex regularizer
        $\eta > 0$ step-size parameter
  $\boldsymbol{z}^0 \leftarrow \arg\min_{\boldsymbol{z} \in \mathcal{X}} \omega(\boldsymbol{z})$
  function NextStrategy($\boldsymbol{m}^t$)
      [▷ Set $\boldsymbol{m}^t = \mathbf{0}$ for non-predictive version]
      return $\arg\max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \left\{ (\boldsymbol{m}^t)^\top \hat{\boldsymbol{x}} - \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{z}^{t-1}) \right\}$
  function ObserveUtility($\boldsymbol{\ell}^t$)
      $\boldsymbol{z}^t \leftarrow \arg\max_{\hat{\boldsymbol{z}} \in \mathcal{X}} \left\{ (\boldsymbol{\ell}^t)^\top \hat{\boldsymbol{z}} - \frac{1}{\eta} D_\varphi(\hat{\boldsymbol{z}} \,\|\, \boldsymbol{z}^{t-1}) \right\}$
```

Solution. [*** Your solution here ***] □

Since OMWU is just predictive FTRL, we can use the known regret bound for predictive FTRL we saw in class—and the following proposition recalls—to give a regret guarantee for OMWU.

Proposition 1. Consider the predictive FTRL algorithm given in Algorithm 1. Let Ω denote the range of φ over $\mathcal{X}$, that is, $\Omega := \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}}\{\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x}')\}$. For any T, the regret accumulated up to time T satisfies

$$R^T \leq \frac{\Omega}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_*^2 - \frac{1}{4\eta} \sum_{t=2}^{T} \|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|^2,$$

where $\|\cdot\|$ is any norm with respect to which φ is 1-strongly convex, and $\|\cdot\|_*$ is the dual of the norm $\|\cdot\|$.

Proposition 1 was stated in general for any instantiation of FTRL. In the particular case of OMWU, the negative entropy function ω was proven to be 1-strongly convex with respect to both the Euclidean norm (Problem 1.1) and the ℓ_1 norm (Problem 1.2). So, in principle, either norm can be used in Proposition 1. However, one choice dominates the other.

Problem 1.7 (2 points). The negative entropy function ω is 1-strongly convex with respect to both the Euclidean norm (Problem 1.1) and the ℓ_1 norm (Problem 1.2). So, in principle, either norm can be used when invoking Proposition 1. Which norm do you think leads to a stronger regret bound, and why?

Solution. [*** Your solution here ***] □

Problem 1.8 (2 points). Prove that the range $\Omega = \sup_{\boldsymbol{x},\boldsymbol{x}' \in \operatorname{relint} \Delta^n} \{\omega(\boldsymbol{x}) - \omega(\boldsymbol{x}')\}$ of ω on $\operatorname{relint} \Delta^n$ is $\Omega = \log n$. Then, use Proposition 1—which was stated in general for any instantiation of FTRL—to argue that OMWU for the simplex Δ^n satisfies the regret bound

$$R^T \le \frac{\log n}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}^t - \boldsymbol{m}^t\|_\infty^2 - \frac{1}{4\eta} \sum_{t=2}^{T} \|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_1^2.$$

★ Hint: The minimizer $\boldsymbol{x}^*$ of ω over $\operatorname{relint} \Delta^n$ is $\nabla\omega^*(\mathbf{0})$. You already know how to compute this from Problem 1.5.
★ Hint: The supremum of ω over $\operatorname{relint} \Delta^n$ is 0 (you should prove this).
★ Hint: You can take for granted the fact that $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$.

Solution. [*** Your solution here ***] □

1.4 OMWU as predictive OMD (4 points, bonus)

Optimistic mirror descent (OMD) is a different framework for building no-regret learning algorithms that is closely related, but not quite equivalent to, FTRL. The general formulation of OMD is given in Algorithm 2. It turns out that, in the special case when the regularizer is negative entropy, OMD and FTRL coincide, and both produce MWU. Your task is to show this.

As a reminder, the Bregman divergence $D_\varphi(\cdot \,\|\, \cdot)$ is defined with respect to any regularizer φ and any two points $\boldsymbol{x}, \boldsymbol{c}$ as

$$D_\varphi(\boldsymbol{x} \,\|\, \boldsymbol{c}) \coloneqq \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{c}) - \nabla\varphi(\boldsymbol{c})^\top (\boldsymbol{x} - \boldsymbol{c}).$$

Problem 1.9 (4 points). Consider the predictive OMD algorithm (Algorithm 2), where $\mathcal{X}$ is set to be the relative interior $\operatorname{relint} \Delta^n$ of the n-simplex, the regularizer φ is set to be the negative entropy function ω, and $\eta > 0$ is a generic stepsize. Prove that the iterates produced by that algorithm coincide with those produced by OMWU as defined in Section 1.3.

Solution. [*** Your solution here ***] □

1.5 The quadratic regularizer and projected gradient descent (8 points, bonus)

Now we will consider the quadratic regularizer,

$$\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2.$$

Problem 1.10 (3 points). Show that, with the quadratic regularizer, predictive FTRL produces the iterates

$$\boldsymbol{x}^t = \Pi_{\mathcal{X}}(\eta(\boldsymbol{L}^{t-1} + \boldsymbol{m}^t)).$$

where $\Pi_{\mathcal{X}}$ is the ℓ_2-projection onto $\mathcal{X}$:

$$\Pi_{\mathcal{X}}(\boldsymbol{z}) := \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{z}\|_2$$

Solution. [*** Your solution here ***] □

Problem 1.11 (3 points)**.** Show that, with the quadratic regularizer, predictive OMD produces the iterates

$$\boldsymbol{z}^t = \Pi_{\mathcal{X}}(\boldsymbol{z}^{t-1} + \eta \boldsymbol{\ell}^{t-1}); \quad \boldsymbol{x}^t = \Pi_{\mathcal{X}}(\boldsymbol{z}^{t-1} + \eta \boldsymbol{m}^t).$$

This algorithm is known as *(optimistic) projected gradient descent*, for obvious reasons.

Solution. [*** Your solution here ***] □

Problem 1.12 (2 points)**.** In a few sentences, qualitatively describe the difference between FTRL and OMD when using the quadratic regularizer.

Solution. [*** Your solution here ***] □

2 Safe Abstraction (30 points + 7 bonus)

Consider a two-player zero-sum extensive-form game Γ that proceeds as follows.

Step 1:

- Each player $i \in \{1, 2\}$ privately rolls a fair n-sided die, and observes its outcome $\theta_i \in \{1, \dots, n\}$.
- Both players simultanoeusly pick actions a_1, a_2 from some action set A of size m.

Step 2:

- Each player $i \in \{1, 2\}$ privately rolls another fair n-sided die, and observes its outcome $\theta_i' \in \{1, \dots, n\}$.
- Both players again simultanoeusly pick actions $a_1, a_2 \in A$.

The utility of P1[1] depends only on the sums $\theta_1 + \theta_1'$ and $\theta_2 + \theta_2'$ and the actions a_1, a_2, a_1', a_2'. That is, each player only cares about the *sum* of their two dice, not on their individual values.

Note that we can identify a decision point (information set) of a player i by specifying either θ_i (for Step 1), or a tuple $(\theta_i, \boldsymbol{a}, \theta_i')$ (for Step 2) (where $\boldsymbol{a} = (a_1, a_2)$).

2.1 Warm-up

Problem 2.1 (3 points)**.** In terms of m and n, how many *terminal nodes* does this game have? How many *information sets* does each player have? How about *sequences* (including the empty sequence)? (No proofs required; just state the answers.)

[1] This is a zero-sum game: P2's utility is the negative of P1's.

> *Solution.* [*** Your solution here ***] □

2.2 Symmetric strategies and CFR

Let $J_i = (\theta_i, a_i, a_{-i}, \theta'_i)$ and $\tilde{J}_i = (\tilde{\theta}_i, a_i, a_{-i}, \tilde{\theta}'_i)$ be two infosets of the same player i in Step 2. We say that J_i and $\tilde{J}_i$ are *equivalent* if $\theta_i + \theta'_i = \tilde{\theta}_i + \tilde{\theta}'_i$. (note that the notation also forces the actions a_i, a_{-i} to be the same in both infosets)

Call a behavioral strategy $\pi_i : \mathcal{J}_i \to \Delta(A)$, where $\mathcal{J}_i$ is player i's information partition, *symmetric* if equivalent infosets have identical action distributions, that is, $\pi_i(\cdot|J_i) = \pi_i(\cdot|J'_i)$ whenever J_i, J'_i are equivalent.

Suppose both players run CFR with simultaneous (as opposed to alternating) iterates. Let π_i^t be the behavioral strategy played by player i at time t. Recall that, in CFR for a generic extensive-form game, the counterfactual utility given to the regret minimizer at infoset $J_i \in \mathcal{J}_i$ at time t is given by

$$u_i^t(a_i|J_i) = \sum_{\substack{h \in J_i \\ z \succeq ha}} \pi_i^t(z|ha_i) \cdot \pi_{-i}^t(z) \cdot c(z) \cdot u_i(z) \tag{6}$$

where

- ha_i is the child of node h reached by playing action a_i,
- the sum is over terminal nodes z that are descendants of nodes ha_i for $h \in J_i$,
- $\pi_i(z|ha_i)$ is the probability that player i plays all actions on the $ha_i \to z$ path,
- $\pi_{-i}(z)$ (resp. $c(z)$) is the probability that the opponent (resp. chance) plays all actions on the path from the root node to z, and
- $u_i(z)$ is the utility of terminal node z for player i.

Problem 2.2 (8 points)**.** Let $\pi_i^t : \mathcal{J}_i \to \Delta(A)$ be the behavioral strategy played by player i at time t in CFR. Show that if $\pi_{-i}^1, \dots, \pi_{-i}^{t-1}$ are all symmetric, then $\pi_i^1, \dots, \pi_i^t$ are also symmetric.

★ Hint: Given two equivalent infosets $(\theta_i, \boldsymbol{a}, \theta'_i)$, $(\tilde{\theta}_i, \boldsymbol{a}, \tilde{\theta}'_i)$, show that these two infosets will observe the same counterfactual utility $u_i^t(\cdot|J)$ at every timestep t.

> *Solution.* [*** Your solution here ***] □

It follows by induction that both players will play symmetric strategies at all timesteps. You may use this fact for the remainder of the problem.

Problem 2.3 (8 points)**.** Show that, in this game, CFR can be implemented in $O(n^2m^4)$ time per iteration. Briefly discuss: how does this time complexity compare to that of the naive implementation of CFR, whose runtime would be linear in the number of nodes in the tree?

★ Hint: Because of the symmetry condition that you have just proven, many infosets are "copies" of each other, and you don't need to do the work of storing multiple copies of the same infoset. The naive computation will not work—you have to be a bit clever about saving reusable work across infosets.
★ Hint: Don't forget to also show that the updates at the first layer of infosets in Step 1 can also be performed efficiently. For this, it may be helpful to note that, if J_i is an infoset that is never the last infoset before any terminal node, then

$$u_i^t(a_i|J_i) = \sum_{J_i a_i \to J'_i} u_i^t(J'_i) \quad \text{where} \quad u_i^t(J'_i) := \mathop{\mathbb{E}}_{a'_i \sim \pi_i^t(a'_i|J_{-i})} u_i^t(a'_i|J'_i)$$

and the first sum is taken over infosets J_i' whose parent sequence is $J_i a_i$.

Solution. [*** Your solution here ***] □

2.3 Imperfect-recall abstraction

Now consider the imperfect-recall abstraction of Γ in which equivalent information sets have been merged into larger infosets. That is, infosets in Step 2 are now identified by tuples $(\theta_i^*, \boldsymbol{a})$ where $\theta_i^* = \theta_i + \theta_i'$. Call this game $\hat{\Gamma}$. We define CFR in imperfect-recall games using the formula (6).

CFR on imperfect-recall zero-sum games is *not* generally guaranteed to converge to Nash equilibrium, even in averages, but you will show that *in this particular class of games* CFR works.

Problem 2.4 (8 points)**.** Show that, when the local regret minimizer is regret matching, running CFR on $\hat{\Gamma}$ produces the same iterates as running CFR on Γ.

★ Hint: Read the next problem before solving this one.

Solution. [*** Your solution here ***] □

Problem 2.5 (3 points)**.** Your proof in the previous problem should have used a property specific to (variants of) regret matching. What is that property? Suppose that we were to instead run CFR in Γ with FTRL at each infoset and regularization parameter η. What would we need to do to η in $\hat{\Gamma}$ to recover the same iterates? (No proof required—just state the change)

Solution. [*** Your solution here ***] □

In some sense, the fact that CFR produces symmetric iterates (Problem 2.2) is what makes the above imperfect-recall abstraction "safe". As we discussed in class, some abstractions used in practice (*e.g.*, for poker) are safe in this sense, and some are not.

Problem 2.6 (7 points, bonus)**.** Now suppose that the dice are weighted according to some (common-knowledge) distribution $p \in \Delta(n)$. Revise your solutions to Problems 2.2 through 2.5 so that they can handle this case as well.

★ Hint: You may assume that the local regret minimizer is regret matching, and you probably want to use the same property that you discussed in Problem 2.5

Solution. [*** Do not write a solution to this problem. Instead, revise your solutions to Problems 2.2 through 2.5 to account for the non-uniform weighting. ***] □

3 Counterfactual Regret Minimization (35 points + 6 bonus)

In this problem, you will implement the CFR regret minimizer for sequence-form decision problems.

You will run your CFR implementation on three games: rock-paper-superscissors (a simple variant of rock-paper-scissors, where beating paper with scissors gives a payoff of 2 instead of 1) and two well-known poker variants: Kuhn poker [Kuhn, 1950] and Leduc poker [Southey et al., 2005]. A description of each game is given in the zip of this homework, according to the format described in Section 3.1. The zip of the homework also contains a stub Python file to help you set up your implementation.

3.1 Format of the game files

Each game is encoded as a json file with the following structure.

- At the root, we have a dictionary with three keys: `decision_problem_pl1`, `decision_problem_pl2`, and `utility_pl1`. The first two keys contain a description of the tree-form sequential decision problems faced by the two players, while the third is a description of the bilinear utility function for Player 1 as a function of the sequence-form strategies of each player. Since both games are zero-sum, the utility for Player 2 is the opposite of the utility of Player 1.
- The tree of decision points and observation points for each decision problem is stored as a list of nodes. Each node has the following fields

  `id` is a string that represents the identifier of the node. The identifier is unique among the nodes for the same player.

  `type` is a string with value either `decision` (for decision points) or `observation` (for observation points).

  `actions` (only for decision points). This is a set of strings, representing the actions available at the decision node.

  `signals` (only for observation points). This is a set of strings, representing the signals that can be observed at the observation node.

  `parent_edge` identifies the parent edge of the node. If the node is the root of the tree, then it is `null`. Else, it is a pair `(parent_node_id, action_or_signal)`, where the first member is the `id` of the parent node, and `action_or_signal` is the action or signal that connects the node to its parent.

  `parent_sequence` (only for decision points). Identifies the parent sequence p_j of the decision point, defined as the last sequence (that is, decision point-action pair) encountered on the path from the root of the decision process to j.

> **Remark 1.** The list of nodes of the tree-form sequential decision process is given in top-down traversal order. The bottom-up traversal order can be obtained by reading the list of nodes backwards.

- The bilinear utility function for Player 1 is given through the payoff matrix $\boldsymbol{A}$ such that the (expected) utility of Player 1 can be written as

$$u_1(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y},$$

  where $\boldsymbol{x}$ and $\boldsymbol{y}$ are sequence-form strategies for Players 1 and 2 respectively. We represent $\boldsymbol{A}$ in the file as a list of all non-zero matrix entries, storing for each the row index, column index, and value. Specifically, each entry is an object with the fields

  `sequence_pl1` is a pair `(decision_pt_id_pl1, action_pl1)` which represents the sequence of Player 1 (row of the entry in the matrix).

  `sequence_pl2` is a pair `(decision_pt_id_pl2, action_pl2)` which represents the sequence of Player 2 (column of the entry in the matrix).

  `value` is the non-zero float value of the matrix entry.

Example: Rock-paper-superscissors In the case of rock-paper-superscissors the decision problem faced by each of the players has only one decision points with three actions: playing rock, paper, or superscissors. So, each tree-form sequential decision process only has a single node, which is a decision node. The payoff matrix of the game[2] is

$$\begin{array}{c} \\ \mathrm{r} \\ \mathrm{p} \\ \mathrm{s} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{r} & \mathrm{p} & \mathrm{s} \end{array} \\ \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -2 \\ -1 & 2 & 0 \end{pmatrix} \end{array}.$$

So, the game file in this case has content:

```
{
  "decision_problem_pl1": [
    {"id": "d1_pl1", "type": "decision", "actions": ["r", "p", "s"],
     "parent_edge": null, "parent_sequence": null}
  ],
  "decision_problem_pl2": [
    {"id": "d1_pl2", "type": "decision", "actions": ["r", "p", "s"],
     "parent_edge": null, "parent_sequence": null}
  ],
  "utility_pl1": [
    {"sequence_pl1": ["d1_pl1", "r"], "sequence_pl2": ["d1_pl2", "p"], "value": -1},
    {"sequence_pl1": ["d1_pl1", "r"], "sequence_pl2": ["d1_pl2", "s"], "value": 1},
    {"sequence_pl1": ["d1_pl1", "p"], "sequence_pl2": ["d1_pl2", "r"], "value": 1},
    {"sequence_pl1": ["d1_pl1", "p"], "sequence_pl2": ["d1_pl2", "s"], "value": -2},
    {"sequence_pl1": ["d1_pl1", "s"], "sequence_pl2": ["d1_pl2", "r"], "value": -1},
    {"sequence_pl1": ["d1_pl1", "s"], "sequence_pl2": ["d1_pl2", "p"], "value": 2}
  ]
}
```

3.2 Learning to best respond

Let Q_1 and Q_2 be the sequence-form strategy polytopes corresponding to the tree-form sequential decision problems faced by Players 1 and 2 respectively. A good smoke test when implementing regret minimization algorithms is to verify that they learn to best respond. In particular, you will verify that your implementation of CFR applied to the decision problem of Player 1 learns a best response against Player 2 when Player 2 plays the *uniform* strategy, that is, the strategy that at each decision points picks any of the available actions with equal probability.

Let $\boldsymbol{u} \in Q_2$ be the sequence-form representation of the strategy for Player 2 that at each decision point selects each of the available actions with equal probability. When Player 2 plays according to that strategy, the utility vector for Player 1 is given by $\boldsymbol{\ell} := \boldsymbol{A}\boldsymbol{u}$, where $\boldsymbol{A}$ is the payoff matrix of the game.

For each of the three games, take your CFR implementation for the decision problem of Player 1, and let it output strategies $\boldsymbol{x}^t \in Q_1$ while giving as feedback at each time t the same utility vector $\boldsymbol{\ell}$. As $T \to \infty$, the average strategy

$$\bar{\boldsymbol{x}}^T := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}^t \in Q_1 \tag{7}$$

will converge to a best response to the uniform strategy $\boldsymbol{u}$, that is,

$$\lim_{T\to\infty} (\bar{\boldsymbol{x}}^T)^\top \boldsymbol{A}\boldsymbol{u} = \max_{\hat{\boldsymbol{x}}\in Q_1} \hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{u}.$$

[2] A Nash equilibrium of the game is reached when all players play rock with probability 1/2, paper with probability 1/4 and superscissors with probability 1/4. Correspondingly, the game value is 0.

If the above doesn't happen empirically, something is wrong with your implementation.

Problem 3.1 (10 points)**.** In each of the three games, apply your CFR implementation to the tree-form sequential decision problem of Player 1, using as local regret minimizer at each decision point the regret matching algorithm (Lecture 4). At each time t, give as feedback to the algorithm the same utility vector $\boldsymbol{\ell} = \boldsymbol{A}\boldsymbol{u}$, where $\boldsymbol{u} \in Q_2$ is the uniform strategy for Player 2. Run the algorithm for 1000 iterations. After each iteration $T = 1, \ldots, 1000$, compute the value of $v^T := (\bar{\boldsymbol{x}}^T)^\top \boldsymbol{A}\boldsymbol{u}$ where $\bar{\boldsymbol{x}}^T \in Q_1$ is the average strategy output so far by CFR, as defined in (7).

Plot v^T as a function of T. Empirically, what is the limit you observe v^T is converging to?

★ Hint: represent vectors on $\mathbb{R}^{|\Sigma|}$ (including the sequence-form strategies output by CFR and utility vectors given to CFR) in memory as dictionaries from sequences (tuples `(decision_point_id, action)`) to floats.
★ Hint: in rock-paper-superscissor, v^T should approach the value $1/3$. In Kuhn poker, the value $1/2$. In Leduc poker, the value 2.0875.

> *Solution.* [*** Your solution here. Your solution should include three plots (one for each game), and three values. Don't forget to turn in your implementation. ***] □

3.3 Learning a Nash equilibrium

Now that you are confident that your implementation of CFR is correct, you will use CFR to converge to Nash equilibrium using the self-play idea described in Lecture 3 and recalled next.

The idea behind using regret minimization to converge to Nash equilibrium in a two-player zero-sum game is to use *self play*. We instantiate two regret minimization algorithms, $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$, for the domains of the maximization and minimization problem, respectively. At each time t the two regret minimizers output strategies $\boldsymbol{x}^t$ and $\boldsymbol{y}^t$, respectively. Then, they receive as feedback the vectors $\boldsymbol{\ell}^t_\mathcal{X}, \boldsymbol{\ell}^t_\mathcal{Y}$ defined as

$$\boldsymbol{\ell}^t_\mathcal{X} := \boldsymbol{A}\boldsymbol{y}^t, \qquad \boldsymbol{\ell}^t_\mathcal{Y} := -\boldsymbol{A}^\top \boldsymbol{x}^t, \tag{8}$$

where $\boldsymbol{A}$ is Player 1's payoff matrix.

We summarize the process pictorially in Figure 1.

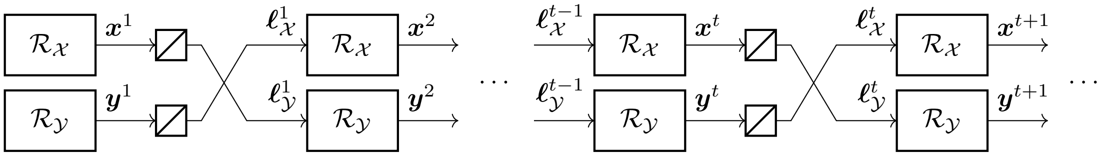
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^1$
$\boldsymbol{\ell}^1_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^2$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^1$
$\boldsymbol{\ell}^1_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^2$
$\cdots$
$\boldsymbol{\ell}^{t-1}_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^t$
$\boldsymbol{\ell}^t_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^{t+1}$
$\boldsymbol{\ell}^{t-1}_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^t$
$\boldsymbol{\ell}^t_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^{t+1}$
$\cdots$

Figure 1: The flow of strategies and utilities in regret minimization for games. The symbol ⧄ denotes computation/construction of the utility vector.

A well known folk theorem establish that the pair of average strategies produced by the regret minimizers up to any time T converges to a Nash equilibrium, where convergence is measured via the *saddle point gap*

$$0 \le \epsilon(\boldsymbol{x}, \boldsymbol{y}) := \left(\max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \{\hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{y}\} - \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y}\right) + \left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} - \min_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \{\boldsymbol{x}^\top \boldsymbol{A}\hat{\boldsymbol{y}}\}\right) = \max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \{\hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{y}\} - \min_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \{\boldsymbol{x}^\top \boldsymbol{A}\hat{\boldsymbol{y}}\}.$$

A point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ has zero saddle point gap if and only if it is a Nash equilibrium of the game.

Theorem 1. Consider the self-play setup summarized in Figure 1, where $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$ are regret minimizers for the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $R^T_\mathcal{X}$ and $R^T_\mathcal{Y}$ be the (sublinear) regret accumulated by $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$, respectively, up to time T, and let $\bar{\boldsymbol{x}}^T$ and $\bar{\boldsymbol{y}}^T$ denote the average of the strategies produced up to time T, that is,

$$\bar{\boldsymbol{x}}^T \coloneqq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}^t, \qquad \bar{\boldsymbol{y}}^T \coloneqq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}^t. \tag{9}$$

Then, the saddle point gap $\epsilon(\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T)$ of $(\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T)$ satisfies

$$\epsilon(\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T) \le \frac{R^T_\mathcal{X} + R^T_\mathcal{Y}}{T} \to 0 \qquad \text{as } T \to \infty.$$

Problem 3.2 (15 points)**.** Let the CFR implementation (using regret matching as the local regret minimizer at each decision point) for Player 1's and Player 2's tree-form sequential decision problems play against each other in self play, as described above.

Plot the saddle point gap and the expected utility (for Player 1) of the average strategies $\epsilon(\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T)$ as a function of the number of iterations $T = 1, \dots, 1000$.

★ Hint: represent vectors on $\mathbb{R}^{|\Sigma|}$ (including the sequence-form strategies output by CFR and utility vectors given to CFR) in memory as dictionaries from sequences (tuples `(decision_point_id, action)`) to floats.
★ Hint: to compute the saddle-point gap, feel free to use the function `gap(game, strategy_pl1, strategy_pl2)` provided in the Python stub file.
★ Hint: the saddle point gap shuold be going to zero. The expected utility of the average strategies in rock-paper-superscissor should approach the value 0. In Kuhn poker it should approach −0.055. In Leduc poker it should approach −0.085.

Solution. [*** Your solution here. Your solution should include six plots (two for each game—one for the saddle point gap and one for the utility). Don't forget to turn in your implementation. ***] □

3.4 CFR+, Alternating Iterates, and Linear Averaging

To achieve better performance in practice when learning Nash equilibria in two-player zero-sum games, people often make the following modifications to the setup of the previous subsection.

- Instead of regret matching, CFR is set up to use the regret matching plus algorithm (see Lecture 3) at each decision point.
- Instead of using the classical self-play scheme described in Figure 1, people *alternate* the iterates and feedback as described in Figure 2, where the utility vector $\boldsymbol{\ell}^t_\mathcal{X}$ is as defined in (8), whereas

  $$\tilde{\boldsymbol{\ell}}^t_\mathcal{Y} \coloneqq -\boldsymbol{A}^\top \boldsymbol{x}^{t+1}.$$

  (Note that at the very beginning, $\boldsymbol{x}^1$ does not participate in the computation of any utility vector).

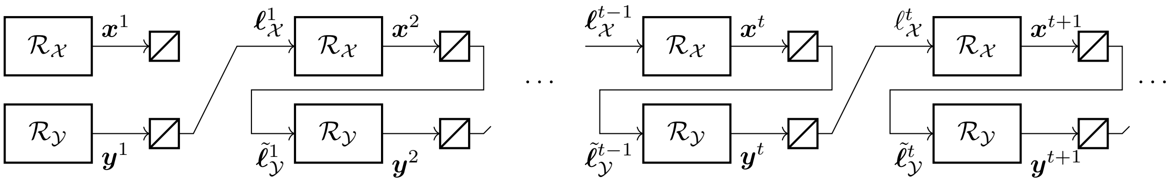
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^1$
$\boldsymbol{\ell}^1_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^2$
…
$\boldsymbol{\ell}^{t-1}_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^t$
$\boldsymbol{\ell}^t_\mathcal{X}$
$\mathcal{R}_\mathcal{X}$
$\boldsymbol{x}^{t+1}$
…
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^1$
$\tilde{\boldsymbol{\ell}}^1_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^2$
$\tilde{\boldsymbol{\ell}}^{t-1}_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^t$
$\tilde{\boldsymbol{\ell}}^t_\mathcal{Y}$
$\mathcal{R}_\mathcal{Y}$
$\boldsymbol{y}^{t+1}$

Figure 2: The alternation method for CFR in games. The symbol ⧄ denotes computation/construction of the utility vector.

- Finally, the output is weighted average of the strategies,

$$\bar{\bar{\boldsymbol{x}}}^T := \frac{1}{\sum_{t=1}^T t^\gamma} \sum_{t=1}^T t^\gamma \boldsymbol{x}^t, \qquad \bar{\bar{\boldsymbol{y}}}^T := \frac{1}{\sum_{t=1}^T t^\gamma} \sum_{t=1}^T t^\gamma \boldsymbol{y}^t$$

  is considered instead of the regular averages (9) when computing the saddle point gap.

Collectively, the modified setup we just described is referred to as “running CFR+”.

Problem 3.3 (10 points + 6 bonus). Modify your implementation of CFR so that it also supports each of the following variants.

(a) (10 points) CFR+, with alternating iterates and $\gamma = 1$,
(b) (2 points, bonus) DCFR, with alternating iterates and parameters $\alpha = 1.5, \beta = 0, \gamma = 2$,
(c) (2 points, bonus) PCFR+, with alternating iterates and $\gamma = 2$, and
(d) (2 points, bonus) PCFR+, with alternating iterates and $\gamma = \infty$ (*i.e.,* do not average; keep only the last iterate).

For each variant, run it for 1000 iterations, plotting the expected utility for Player 1 and the saddle point gap of the linear averages $\epsilon(\bar{\bar{\boldsymbol{x}}}^T, \bar{\bar{\boldsymbol{y}}}^T)$ after each iteration T. Add these lines to your plots from Problem 3.2.

Solution. [*** Do not write a solution for this problem. Instead, modify your plots in Problem 3.2 so that they also contain each of these variants. If you complete all parts of this problem, your solution to Problem 3.2 should now contain six plots, each with five lines (CFR, CFR+, DCFR, PCFR+ $\gamma = 2$, and PCFR+ last-iterate). Make sure to specify which game and which algorithm each line refers to. ***] □

References

H. W. Kuhn. A simplified two-person poker. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 1 of *Annals of Mathematics Studies, 24*, pages 97–103. Princeton University Press, Princeton, New Jersey, 1950.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 550–558, 2005.