A. Proofs

A.1. Regret Guarantees when Gradient Estimators are Used

For completeness, we show a proof of Proposition 1. As mentioned, it is an application of the Azuma-Hoeffding inequality for martingale difference sequences, which we now state (see, e.g., Theorem 3.14 of McDiarmid (1998) for a proof).

Theorem 1 (Azuma-Hoeffding inequality)**.** *Let $Y_1, \dots, Y_n$ be a martingale difference sequence with $a_k \leq Y_k \leq b_k$ for each k, for suitable constants a_k, b_k. Then for any $\tau \geq 0$,*

$$\mathbb{P}\Big[\sum Y_k \geq \tau\Big] \leq e^{-2\tau^2 / \sum (b_k - a_k)^2}.$$

Proposition 1. *Let M and $\tilde{M}$ be positive constants such that $|(\boldsymbol{\ell}^t)^\top (\boldsymbol{z} - \boldsymbol{z}')| \leq M$ and $|(\tilde{\boldsymbol{\ell}})^\top (\boldsymbol{z} - \boldsymbol{z}')| \leq \tilde{M}$ for all times $t = 1, \dots, T$ and all feasible points $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$. Then, for all $p \in (0, 1)$ and all $\boldsymbol{u} \in \mathcal{Z}$,*

$$\mathbb{P}\left[R^T(\boldsymbol{u}) \leq \tilde{R}^T(\boldsymbol{u}) + (M + \tilde{M})\sqrt{2T \log \frac{1}{p}}\right] \geq 1 - p.$$

Proof. As observed in the body, $d^t := (\boldsymbol{\ell}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u}) - (\tilde{\boldsymbol{\ell}}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})$ is a martingale difference sequence. Furthermore, at all times t,

$$\begin{aligned} |d^t| &= |(\boldsymbol{\ell}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u}) - (\tilde{\boldsymbol{\ell}}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})| \\ &\leq |(\boldsymbol{\ell}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})| + |(\tilde{\boldsymbol{\ell}}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})| \\ &\leq M + \tilde{M}, \end{aligned} \tag{8}$$

and therefore $-(M + \tilde{M}) \leq d^t \leq (M + \tilde{M})$ for each t.

Furthermore,

$$\sum_{t=1}^{T} d^t = \left(\sum_{t=1}^{T} (\boldsymbol{\ell}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})\right) - \left(\sum_{t=1}^{T} (\tilde{\boldsymbol{\ell}}^t)^\top (\boldsymbol{z}^t - \boldsymbol{u})\right) = R^T(\boldsymbol{u}) - \tilde{R}^T(\boldsymbol{u}).$$

So, using Theorem 1, for all $\tau \geq 0$

$$\begin{aligned} \mathbb{P}\Big[R^T(\boldsymbol{u}) \leq \tilde{R}^T(\boldsymbol{u}) + \tau\Big] &= \mathbb{P}\left[\sum_{t=1}^{T} d^t \leq \tau\right] \\ &= 1 - \mathbb{P}\left[\sum_{t=1}^{T} d^t \geq \tau\right] \\ &\geq 1 - \exp\left\{-\frac{2\tau^2}{\sum_{t=1}^{T} 4(M + \tilde{M})^2}\right\} \\ &= 1 - \exp\left\{-\frac{2\tau^2}{4T(M + \tilde{M})^2}\right\}. \end{aligned}$$

Finally, substituting $\tau = (M + \tilde{M})\sqrt{2T \log(1/p)}$ yields the statement. □

A.2. Properties of the Outcome Sampling Gradient Estimator

Let $\boldsymbol{w}^t \in \mathcal{X}$ be an arbitrary strategy for Player 1. Furthermore, let $\tilde{z}^t \in Z$ be a random variable such that for all $z \in Z$,

$$\mathbb{P}_t[\tilde{z}^t = z] = w^t[\sigma_1(z)] \cdot y^t[\sigma_2(z)] \cdot c[\sigma_{\mathtt{c}}(z)],$$

and let $\boldsymbol{e}_z$ be defined as the vector such that $e_z[\sigma_1(z)] = 1$ and $e_z[\sigma] = 0$ for all other $\sigma \in \Sigma_1, \sigma \neq \sigma_1(z)$.

Lemma 1. *The random vector*

$$\tilde{\boldsymbol{\ell}}_1^t := \frac{u_2(\tilde{z}^t)}{w^t[\sigma_1(\tilde{z}^t)]} \boldsymbol{e}_{\tilde{z}^t}$$

is such that $\mathbb{E}_t[\tilde{\boldsymbol{\ell}}_1^t] = \boldsymbol{\ell}_1^t$.

Proof. For all $\boldsymbol{x} \in \mathbb{R}^{|\Sigma_1|}$,

$$\begin{aligned}
\mathbb{E}_t[\boldsymbol{\ell}_1^t]^\top \boldsymbol{x} &= \left(\sum_{z \in Z} \mathbb{P}[\tilde{z}^t = z] \cdot \frac{u_1(z)}{w^t[\sigma_1(z)]} \boldsymbol{e}_z \right)^\top \boldsymbol{x} \\
&= \left(\sum_{z \in Z} u_2(z) \cdot y^t[\sigma_2(z)] \cdot c[\sigma_{\mathtt{c}}(z)] \cdot \boldsymbol{e}_z \right)^\top \boldsymbol{x} \\
&= \sum_{z \in Z} u_2(z) \cdot y^t[\sigma_2(z)] \cdot c[\sigma_{\mathtt{c}}(z)] \cdot (\boldsymbol{e}_z^\top \boldsymbol{x}) \\
&= \sum_{z \in Z} u_2(z) \cdot y^t[\sigma_2(z)] \cdot c[\sigma_{\mathtt{c}}(z)] \cdot x[\sigma_1(z)] \\
&= u_2(\boldsymbol{x}, \boldsymbol{y}^t, \boldsymbol{c}) = \boldsymbol{\ell}_1^\top \boldsymbol{x}.
\end{aligned}$$

Since the equality holds for all $\boldsymbol{x} \in \mathbb{R}^{|\Sigma_1|}$, we conclude $\mathbb{E}_t[\tilde{\boldsymbol{\ell}}_1^t] = \boldsymbol{\ell}_1$. □

Furthermore,

Lemma 2. *For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,*

$$(\tilde{\boldsymbol{\ell}}_1)^\top (\boldsymbol{x} - \boldsymbol{x}') \leq \Delta \cdot \max_{\sigma \in \Sigma_1} \frac{1}{w^t[\sigma]}.$$

Proof. Using the definition of $\tilde{\boldsymbol{\ell}}_1$,

$$(\tilde{\boldsymbol{\ell}}_1)^\top (\boldsymbol{x} - \boldsymbol{x}') = \frac{u_2(\tilde{z}^t)}{w^t[\sigma_1(\tilde{z}^t)]} \Big(x[\sigma_1(\tilde{z}^t)] - x'[\sigma_1(\tilde{z}^t)] \Big).$$

Since each entry of $\boldsymbol{x}$ and $\boldsymbol{x}'$ is in the interval $[0, 1]$, the quantity $x[\sigma_1(\tilde{z}^t)] - x'[\sigma_1(\tilde{z}^t)]$ has absolute value in $[0, 1]$ as well. Hence,

$$\left| (\tilde{\boldsymbol{\ell}}_1)^\top (\boldsymbol{x} - \boldsymbol{x}') \right| \leq \max_{z \in Z} \left| \frac{u_2(z)}{w^t[\sigma_1(z)])} \right| \leq \Delta \cdot \max_{\sigma \in \Sigma_1} \frac{1}{w^t[\sigma]}$$

as we wanted to show. □

A.3. Exploration-Balanced Strategy

We now describe the construction of the *exploration-balanced strategy* $\boldsymbol{w}^*$. Given $\sigma \in \Sigma_1$, we let $\mathcal{C}_\sigma$ be the set of information sets $I in \mathcal{I}_1$ such that $\sigma_1(I) = \sigma$. Furthermore, let m_σ, for $\sigma \in \Sigma_1$, be the number of terminal sequences in the subtree rooted under σ; formally, m_σ is defined recursively as

$$m_\sigma = \begin{cases} 1 & \text{if } \mathcal{C}_\sigma = \varnothing; \\ \displaystyle\sum_{I \in \mathcal{C}_\sigma} \sum_{a \in A_I} m_{(I,a)} & \text{otherwise.} \end{cases}$$

Clearly, $m_\sigma \leq |\Sigma_1| - 1$, since the empty sequence is never terminal (assuming Player 1 acts at least once). With that, we define $\boldsymbol{w}^*$ such that $w^*[\varnothing] = 1$ and that for all $\sigma = (I, a) \in \Sigma_1$,

$$w^*[\sigma] = \frac{m_\sigma}{\sum_{a' \in A_I} m_{(I,a')}} w^*[\sigma_1(I)].$$

It is immediate to verify that $\boldsymbol{w}^*$ is indeed a valid sequence-form strategy. Furthermore, since for all $I \in \mathcal{I}_1$, $I \in \mathcal{C}_{\sigma_1(I)}$, we have

$$\sum_{a' \in A_I} m_{(I,a')} \leq m_{\sigma(I)}.$$

So,

$$w^*[\sigma] \geq \frac{m_\sigma}{m_{\sigma_1(I)}} w^*[\sigma_1(I)].$$

By recursively expanding the definition of $w^*[\sigma_1(I)]$ on the right-hand side until $\sigma_1(I) = \varnothing$, we ultimately obtain

$$w^*[\sigma] \geq \frac{1}{m_\varnothing} \geq \frac{1}{|\Sigma_1| - 1}$$

for all σ, as we wanted to show.

A.4. Proposition 3

As mentioned in the body of the paper, Proposition 3 is a direct consequence of the concentration result for martingale difference sequences of Bartlett et al. (2008), which we state next.

Lemma 3 (Lemma 2 of Bartlett et al. (2008))**.** *Suppose $X^1, \dots, X^T$ is a martingale difference sequence with $|X^t| \le b$. Let*

$$\operatorname{Var}_t X^t := \operatorname{Var}[X^t \mid X^1, \dots, X^{t-1}].$$

Let $V := \sum_{t=1}^T \operatorname{Var}_t X^t$ be the sum of conditional variances of X^t's. Further, let $\sigma := \sqrt{V}$. Then we have, for any $\delta < 1/e$ and $T \ge 4$,

$$\mathbb{P}\left[\sum_{t=1}^T X^t > 2\max\{2\sigma, b\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta)}\right] \le \log(T)\delta.$$

Proposition 3. *Let $T \ge 4$, and let M and $\tilde{M}$ be positive constants such that $|(\boldsymbol{\ell}^t)^\top(\boldsymbol{z} - \boldsymbol{u})| \le M$ and $|(\tilde{\boldsymbol{\ell}})^\top(\boldsymbol{z} - \boldsymbol{u})| \le \tilde{M}$ for all times $t = 1, \dots, T$ and all feasible points $\boldsymbol{z}, \boldsymbol{u} \in \mathcal{X}$. Furthermore, let $\sigma := \sqrt{\sum_{t=1}^T \operatorname{Var}[d^t \mid \tilde{\boldsymbol{\ell}}^1, \dots, \tilde{\boldsymbol{\ell}}^{t-1}]}$ be the square root of the sum of conditional variances of the random variables d^t introduced in* (5)*. Then, for all $p \in (0, 1/2]$ and all $\boldsymbol{u} \in \mathcal{X}$,*

$$\mathbb{P}\Big[R^T(\boldsymbol{u}) \le \tilde{R}^T(\boldsymbol{u}) + 4\max\{\sigma\beta, (M + \tilde{M})\beta^2\}\Big] \ge 1 - p,$$

where

$$\beta := \sqrt{\log\left(\frac{\log T}{p}\right)}.$$

Proof. We apply Lemma 3 to the martinagle difference sequence $X^t = d_t$. As argued in (8), $|X^t| \le (M + \tilde{M})$ at all times t, so the constant $b = M + \tilde{M}$ satisfies the requirements of Lemma 3. Finally, we set $\delta = p/\log(T)$ in Lemma 3, so that

$$\sqrt{\log(1/\delta)} = \sqrt{\log\left(\frac{\log T}{p}\right)} = \beta.$$

Furthermore, since by hypothesis $T \ge 4$ and $p \le 1/2$, $\delta = p/\log(T) \le 1/(2\log 4) \le 1/e$, so all hypotheses of Lemma 3 are satisfied. Hence, we have

$$\begin{aligned}
\mathbb{P}\Big[R^T(\boldsymbol{u}) - \tilde{R}^T(\boldsymbol{u}) \le 4\max\{\sigma\beta, (M + \tilde{M})\beta^2\}\Big] &= \mathbb{P}\left[\sum_{t=1}^T X^t \le 4\max\{\sigma\beta, b\beta^2\}\right] \\
&= \mathbb{P}\left[\sum_{t=1}^T X^t \le 4\max\{\sigma\sqrt{\log(1/\delta)}, b\log(1/\delta)\}\right] \\
&= \mathbb{P}\left[\sum_{t=1}^T X^t \le 2\max\{2\sigma, 2b\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta)}\right] \\
&\ge \mathbb{P}\left[\sum_{t=1}^T X^t \le 2\max\{2\sigma, b\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta)}\right] \\
&\ge 1 - \log(T)\delta = 1 - p,
\end{aligned}$$

where the last inequality follows from Lemma 3. □

B. Description of the Game Instances Used in the Experiments

We run our experiments on four different games, each described below.

Leduc poker is a standard benchmark in the EFG-solving community (Southey et al., 2005). Our variant, Leduc 13, has a deck of 13 unique cards, with two copies of each card. The game consists of two rounds. In the first round, each player places an ante of 1 in the pot and receives a single private card. A round of betting then takes place with a two-bet maximum, with Player 1 going first. A public shared card is then dealt face up and another round of betting takes place. Again, Player 1 goes first, and there is a two-bet maximum. If one of the players has a pair with the public card, that player wins. Otherwise, the player with the higher card wins. All bets in the first round are 1, while all bets in the second round are 2. This game has 166336 nodes and 6007 sequences per player.

Goofspiel The variant of Goofspiel (Ross, 1971) that we use in our experiments is a two-player card game, employing three identical decks of 4 cards each. At the beginning of the game, each player receives one of the decks to use it as its own hand, while the last deck is put face down between the players, with cards in increasing order of rank from top to bottom. Cards from this deck will be the prizes of the game. In each round, the players privately select a card from their hand as a bet to win the topmost card in the prize deck. The selected cards are simultaneously revealed, and the highest one wins the prize card. In case of a tie, the prize card is discarded. Each prize card's value is equal to its face value, and at the end of the game the players' score are computed as the sum of the values of the prize cards they have won. This game has 54421 nodes and 21329 sequences per player.

Search is a security-inspired pursuit-evasion game. The game is played on the graph shown in Figure 5.

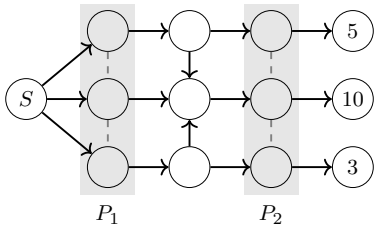

Figure 5. The graph on which the search game is played.

It is a simultaneous-move game (which can be modeled as a turn-taking EFG with appropriately chosen information sets). The defender controls two patrols that can each move within their respective shaded areas (labeled P1 and P2). At each time step the controller chooses a move for both patrols. The attacker is always at a single node on the graph, initially the leftmost node labeled S. The attacker can move freely to any adjacent node (except at patrolled nodes, the attacker cannot move from a patrolled node to another patrolled node). The attacker can also choose to wait in place for a time step in order to clean up their traces. If a patrol visits a node that was previously visited by the attacker, and the attacker did not wait to clean up their traces, they can see that the attacker was there. If the attacker reaches any of the rightmost nodes they receive the respective payoff at the node (5, 10, or 3, respectively). If the attacker and any patrol are on the same node at any time step, the attacker is captured, which leads to a payoff of -1 for the attacker and a payoff of 1 for the defender. Finally, the game times out after k simultaneous moves, in which case both players defender receive payoffs 0. Search-4 (Search-5) has 21613 (87,927) nodes, 2029 (11,830) defender sequences, and 52 (69) attacker sequences.

Our search game is a zero-sum variant of the one used by Kroer et al. (2018). A similar search game considered by Bošanskỳ et al. (2014) and Bošanskỳ & Čermák (2015).

Battleship is a parametric version of a classic board game, where two competing fleets take turns shooting at each other (Farina et al., 2019c). At the beginning of the game, the players take turns at secretly placing a set of ships on separate grids (one for each player) of size 3×2. Each ship has size 2 (measured in terms of contiguous grid cells) and a value of 1, and must be placed so that all the cells that make up the ship are fully contained within each player's grids and do not overlap with any other ship that the player has already positioned on the grid. After all ships have been placed. the players take turns at firing at their opponent. Ships that have been hit at all their cells are considered sunk. The game continues until either one player has sunk all of the opponent's ships, or each player has completed r shots. At the end of the game, each player's payoff is calculated as the sum of the values of the opponent's ships that were sunk, minus the sum of the values of ships which that player has lost. The game has 732607 nodes, 73130 sequences for player 1, and 253940 sequences for player 2.

C. Additional Experimental Results

C.1. External Sampling

The Search-5 plot omitted from the main paper is shown here.

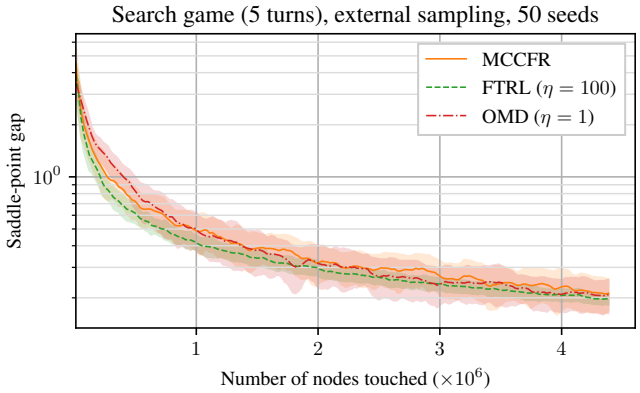

Figure 6. Performance of MCCFR, FTRL, and OMD with external sampling on Search-5.

Figures 7 through 11 show the performance of FTRL and OMD for all four stepsizes that we tried on each game: $\eta = 0.1, 1, 10, 100$.

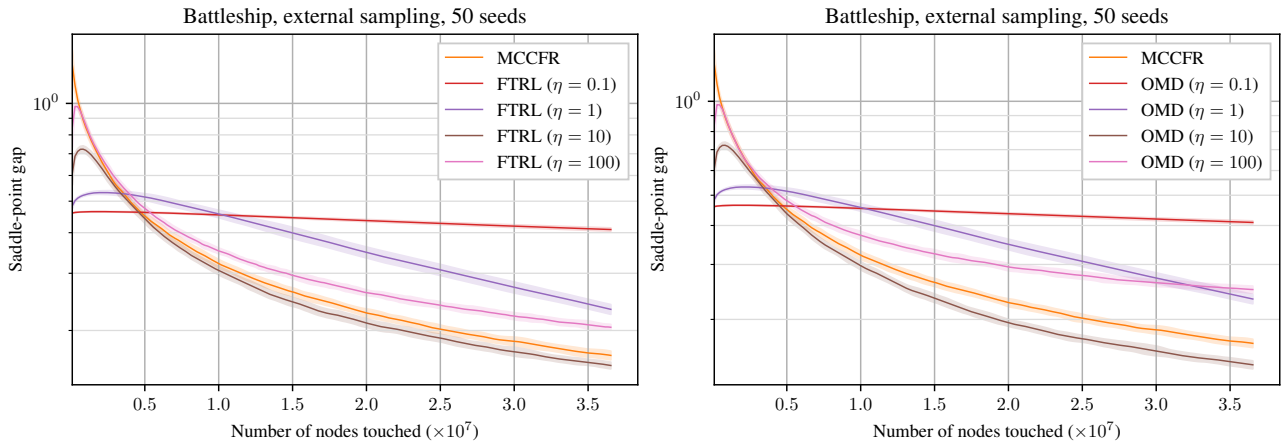

Figure 7. Performance of FTRL and OMD with four stepsizes on Battleship with external sampling. MCCFR shown for reference

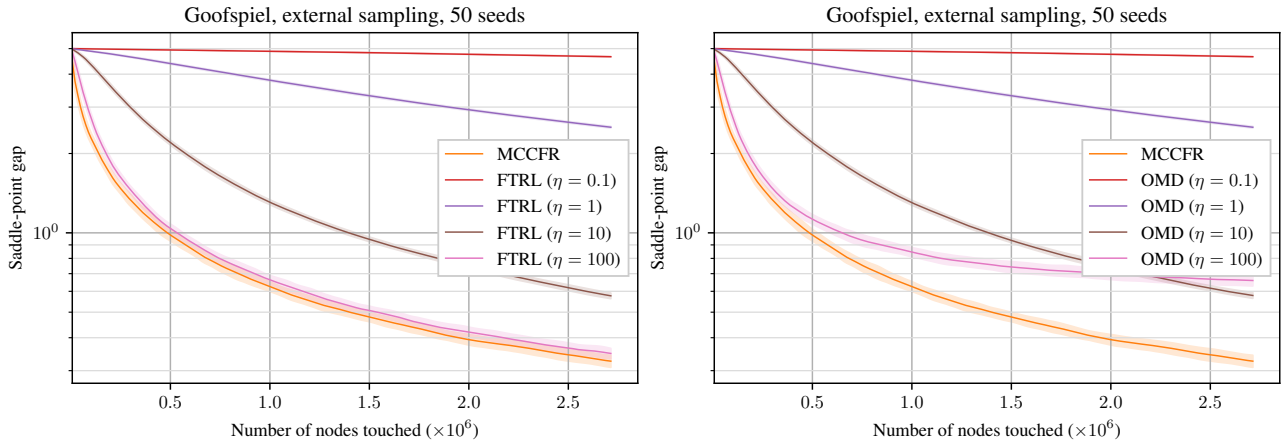

Figure 8. Performance of FTRL and OMD with four stepsizes on Goofspiel with external sampling. MCCFR shown for reference

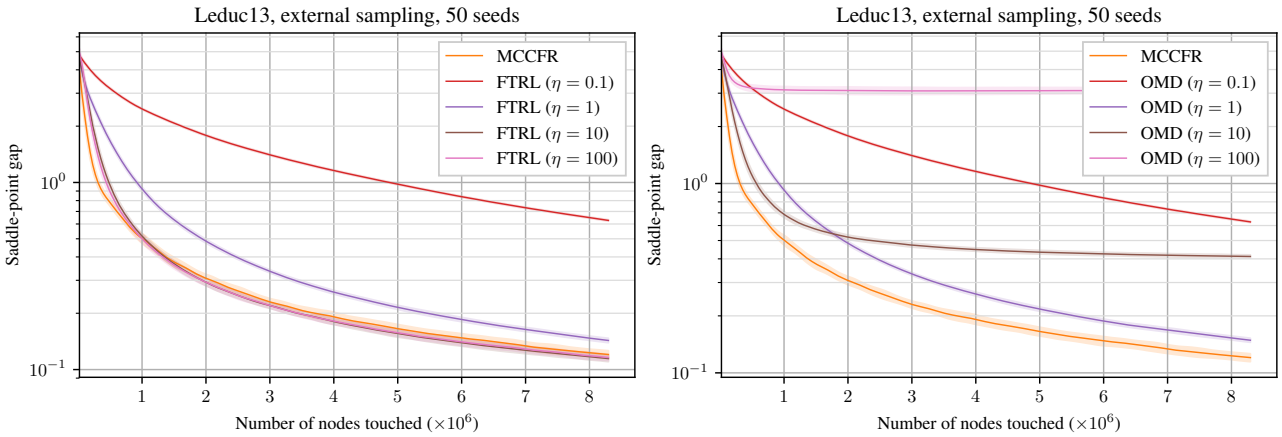

Figure 9. Performance of FTRL and OMD with four stepsizes on Leduc 13 with external sampling. MCCFR shown for reference

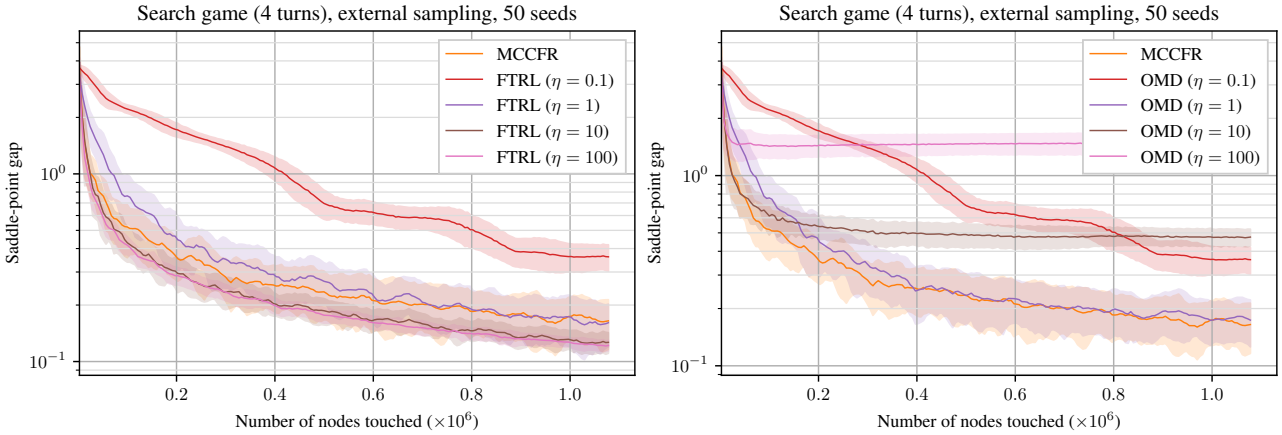

Figure 10. Performance of FTRL and OMD with four stepsizes on Search-4 with external sampling. MCCFR shown for reference

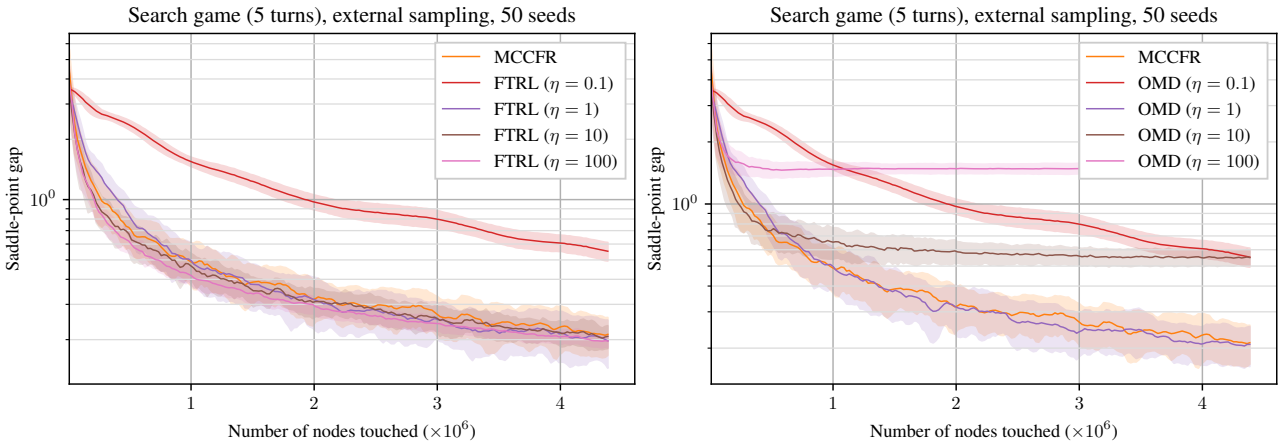

Figure 11. Performance of FTRL and OMD with four stepsizes on Search-5 with external sampling. MCCFR shown for reference

C.2. Exploration-Balanced Outcome Sampling

The Search-4 plot omitted from the main paper is shown here.

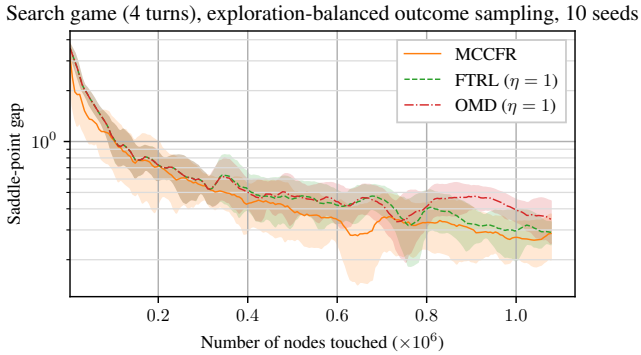

Figure 12. Performance of MCCFR, FTRL, and OMD with outcome sampling on Search-4.

Figure 12 shows the performance on Search-4 and Search-5 with outcome sampling. In Search-4 we find that MCCFR performs better than FTRL and OMD, though FTRL is comparable at later iterations.

Figures 13 through 17 show the performance of FTRL and OMD with outcome sampling for all four stepsizes that we tried on each game: $\eta = 0.1, 1, 10, 100$.

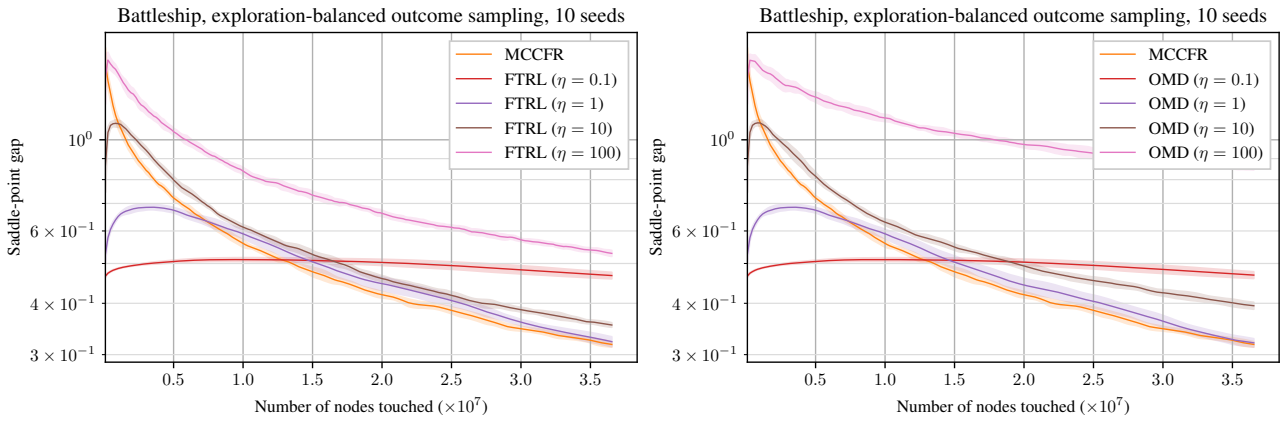

Figure 13. Performance of FTRL and OMD with four stepsizes on Battleship with outcome sampling. MCCFR shown for reference

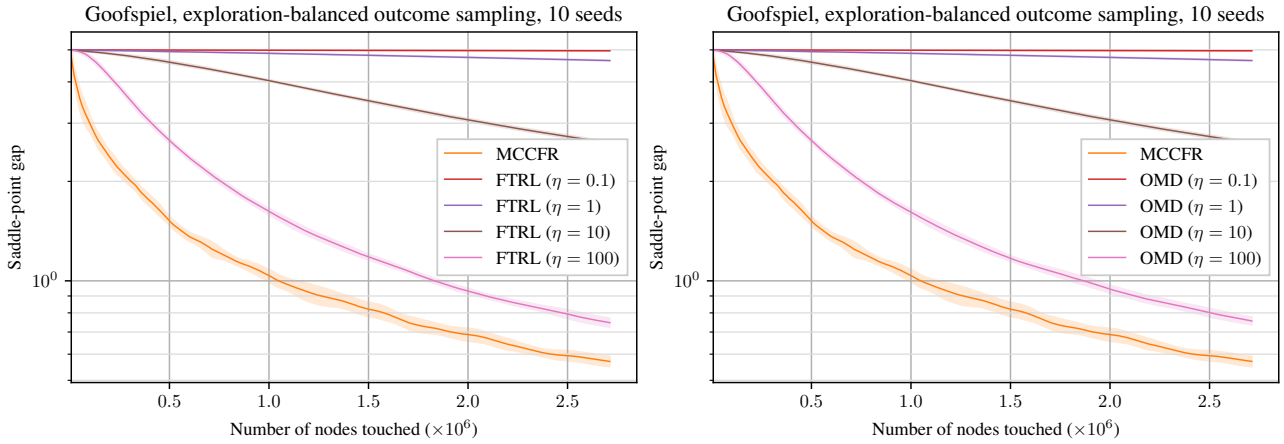

Figure 14. Performance of FTRL and OMD with four stepsizes on Goofspiel with outcome sampling. MCCFR shown for reference

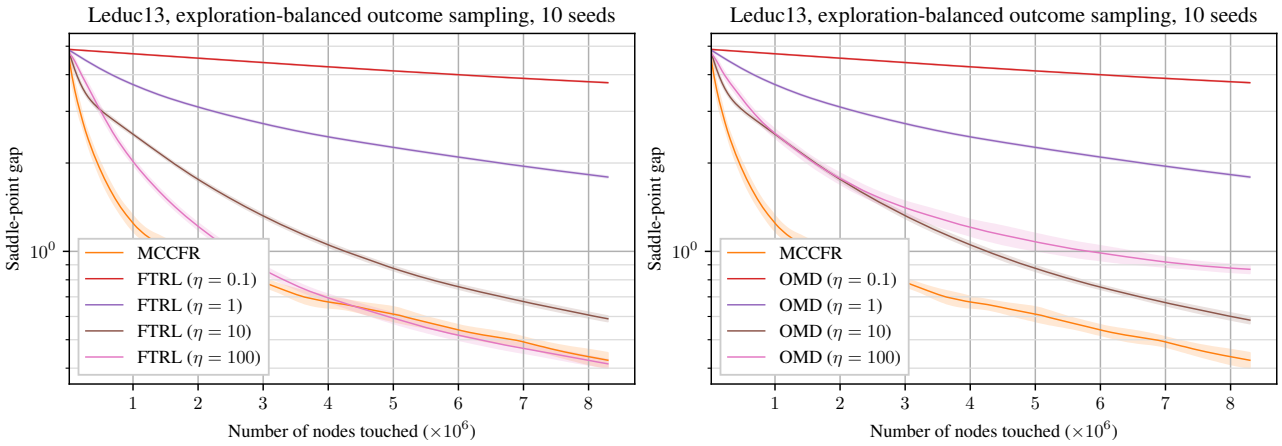

Figure 15. Performance of FTRL and OMD with four stepsizes on Leduc 13 with outcome sampling. MCCFR shown for reference

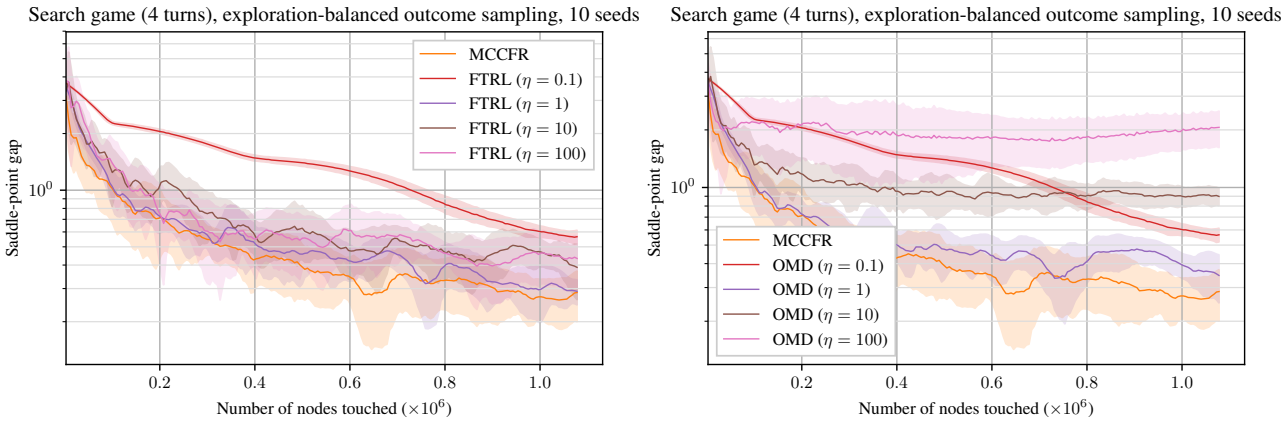

Figure 16. Performance of FTRL and OMD with four stepsizes on Search-4 with outcome sampling. MCCFR shown for reference

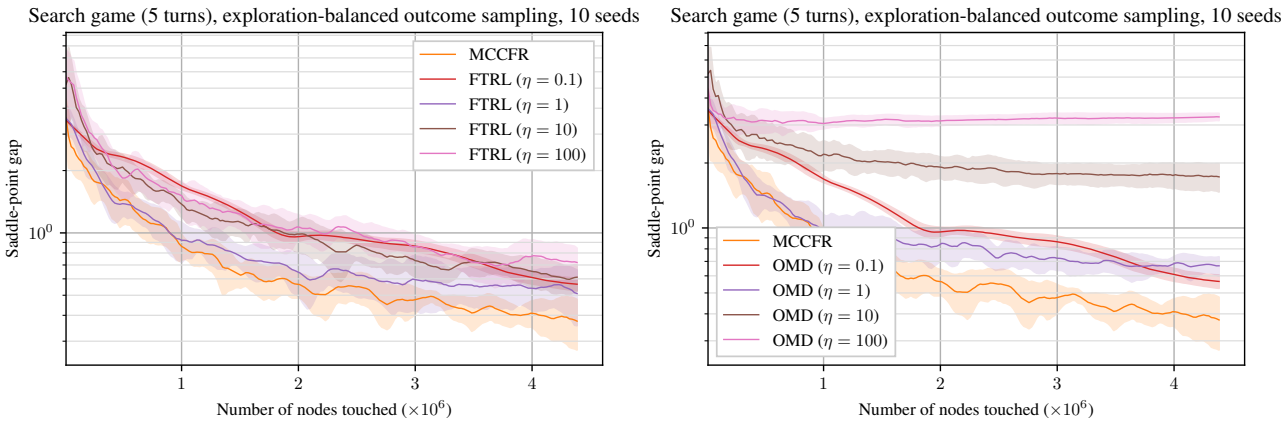

Figure 17. Performance of FTRL and OMD with four stepsizes on Search-5 with outcome sampling. MCCFR shown for reference