# Near-Optimal Last-Iterate Convergence for Zero-Sum Games with Bandit Feedback and Opponent Actions

Soumita Hait[1], Ping Li[2], Haipeng Luo[1], and Mengxiao Zhang[3]

[1]University of Southern California, {hait,haipengl}@usc.edu
[2]Shanghai University of Finance and Economics, pinglee@stu.sufe.edu.cn
[3]University of Iowa, mengxiao-zhang@uiowa.edu

Last-iterate convergence of learning dynamics in games has recently emerged as a central question in online learning and game theory, with growing relevance to AI alignment and preference learning. In two-player zero-sum games with bandit feedback—where players only observe the loss of the played action pair—recent work of Fiegel et al. [1] established a striking separation between average-iterate and last-iterate convergence in duality gap: while the optimal $t^{-1/2}$ rate is achievable for average iterates via standard no-regret algorithms, last iterates cannot converge faster than $t^{-1/3}$ in expectation or $t^{-1/4}$ with high probability.

However, in many practical settings, players observe not only their loss but also the opponent's action. For example, preference learning is often modeled as learning over a skew-symmetric preference game matrix where a central algorithm decides which pair of actions to recommend to the user, in which case the action information is directly available. Such additional information intuitively facilitates estimation and mitigates the exploration burden, leading to a natural question: *does observing the opponent's actions enable faster last-iterate convergence in two-player zero-sum games?*

We answer this question affirmatively in Hait et al. [2], showing that $t^{-1/2}$ last-iterate convergence is achievable with high probability under opponent-action feedback. Our algorithm is efficient and updates its strategy infrequently by solving an estimated log-barrier regularized game. Technically, we identify fundamental obstacles preventing standard analysis for multi-armed bandits (the single-player case) from generalizing to games, and develop a novel analysis to overcome them. Our results also improve the state of the art for dueling bandits, a canonical model for preference learning. Experiments confirm that our algorithm indeed converges faster than naive baselines and prior methods that do not exploit opponent-action feedback.

## References

[1] Côme Fiegel, Pierre Menard, Tadashi Kozuno, Michal Valko, and Vianney Perchet. The harder path: Last iterate convergence for uncoupled learning in zero-sum games with bandit feedback. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[2] Soumita Hait, Ping Li, Haipeng Luo, and Mengxiao Zhang. Near-optimal last-iterate convergence for zero-sum games with bandit feedback and opponent actions. *arXiv preprint arXiv:2605.09363*, 2026. URL https://arxiv.org/pdf/2605.09363.