Heterogeneous Multi-Agent Bandits with Parsimonious Hints

AMIRMAHDI MIRFAKHAR, UMass Amherst, USA
XUCHUANG WANG, UMass Amherst, USA
JINHANG ZUO, City UHK, Hong Kong
YAIR ZICK, UMass Amherst, USA
MOHAMMAD HAJIESMAILI, UMass Amherst, USA

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Amirmahdi Mirfakhar, Xuchuang Wang, Jinhang Zuo, Yair Zick, and Mohammad Hajiesmaili. 2018. Heterogeneous Multi-Agent Bandits with Parsimonious Hints. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Abstract

In competitive resource allocation, agents must learn to secure optimal resources while adapting to the presence of others. In communication networks, radio stations seek interference-free channels, as collisions can corrupt messages and incur high costs. Similarly, in job markets, strategic applications determine employment prospects, where strong applicants can unintentionally crowd out others. A key challenge in such settings is efficient learning under competition, where excessive exploration leads to costly misallocations. To mitigate this, less critical stations should refrain from transmitting on high-quality channels that serve as the best options for critical transmissions. Likewise, applicants must target positions that are less likely to attract significantly stronger candidates. We investigate the role of low-cost supplementary information—*hints*—that assess action quality without direct execution, such as test signals for channel evaluation or interviews for refining application strategies. These hints enable agents to converge more efficiently, reducing costly trial-and-error exploration and facilitating quicker adaptation to equilibrium.

We introduce the *Hinted Heterogeneous Multi-Agent Multi-Armed Bandits* (HMA2B) problem, a novel framework in which agents can query low-cost hints alongside taking actions, enabling them to assess action quality before direct execution. Pulling an arm in this setting corresponds to committing to an action, such as selecting a transmission channel or applying for a position, where strategic choices are crucial for avoiding costly misallocations. In this setting, M agents with unique reward distributions over N arms interact over T rounds, where arm rewards are observed only in the

Authors' Contact Information: Amirmahdi Mirfakhar, smirfakhar@umass.edu, UMass Amherst, Amherst, MA, USA; Xuchuang Wang, UMass Amherst, Amherst, MA, USA, xuchuangw@gmail.com; Jinhang Zuo, City UHK, Kowloon Tong, Kowloon, Hong Kong, jinhangzuo@gmail.com; Yair Zick, UMass Amherst, Amherst, MA, USA, yzick@umass.edu; Mohammad Hajiesmaili, UMass Amherst, Amherst, MA, USA, hajiesmaili@cs.umass.edu.

Manuscript submitted to ACM

absence of collisions. Agents aim to maximize total reward while minimizing hint queries, ensuring time-independent regret. Specifically, agents seek to learn the allocation that maximizes utilitarian social welfare or, equivalently, the maximum weighted matching.

We analyze HMA2B in both *centralized* and *decentralized* settings. In the centralized case, we design an adaptive hinting strategy that achieves time-independent, $O(1)$, regret while using only $O(\log T)$ hints, substantially improving upon prior approaches that suffered $O(\log T)$ regret without hints. This method balances learning and decision-making by adaptively querying hints only when necessary. By separating exploitation and exploration—conducting the former through arm-pulling and the latter through adaptive hinting—agents can efficiently learn the optimal matching early without requiring all hints upfront, a non-trivial and insightful result that highlights the power of selective hinting in structured decision-making. In the decentralized setting, we develop a collision-based communication protocol where agents act independently, relying on uniform hint inquiries to guide their choices until a stopping condition is met for their set of active actions. While this hint inquiry resembles the observation-gathering phase of Explore-then-Commit (ETC), our approach asks for $O(\log T)$ hints to achieve $O(1)$ regret, significantly improving upon ETC's $O(\log T)$ regret. To establish the fundamental limits of our approach, we provide tight lower bounds and demonstrate that our methods achieve optimal performance.

Beyond these core results, our framework can be extended to general linear programs where feasible solutions correspond to matchings, including applications in two-sided markets such as hiring or school admissions. A notable extension involves two-sided markets, where arms have preferences over agents and resolve conflicts by selecting their most preferred agent. This setting introduces additional strategic complexity, as agents must not only compete for resources but also account for the preferences of the resources themselves in order to reach a stable matching, further highlighting the versatility of our approach. In these settings, hints—whether in the form of structured interviews or prompts from large language models—help agents efficiently identify their best stable match, enabling fully decentralized decision-making without explicit communication.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009