Achieving $\widetilde{\mathcal{O}}(1/N)$ Optimality Gap in Weakly-Coupled Markov Decision Processes through Gaussian Approximation

Chen Yan* **Weina Wang†** **Lei Ying***

Abstract

We study finite-horizon weakly-coupled Markov decision processes (WCMDPs) with N homogeneous agents, where each agent is modeled as an MDP. Prior work has shown that linear-programming-based (LP-based) policies, derived from the fluid approximation that captures the system's mean dynamics, achieve an $\mathcal{O}(1/\sqrt{N})$ optimality gap per agent. In this paper, we present instances where this gap is in fact $\Theta(1/\sqrt{N})$.[3] We further propose a novel stochastic-programming-based (SP-based) policy that, under a mild uniqueness assumption, achieves an $\widetilde{\mathcal{O}}(1/N)$ optimality gap per agent. Our approach constructs a Gaussian stochastic system centered around the fluid-optimal trajectory, capturing both the mean and the variance of the WCMDP dynamics. This results in a more accurate approximation than the fluid approximation. The policy is then obtained by solving a linear Gaussian stochastic program for this system. To the best of our knowledge, this is the first result to establish an $\widetilde{\mathcal{O}}(1/N)$ optimality gap for WCMDPs under a uniqueness assumption.

1 Introduction

WCMDPs model the control and decision-making of systems with N agents (each an MDP) under global resource constraints. Agents evolve independently given their local states/actions, but the actions are globally coupled due to resource constraints. WCMDPs arise in machine maintenance, healthcare, target tracking, and ride-hailing empty-car routing [1, 2, 5–7, 9, 10, 12, 14, 15, 17, 19, 23].

A classical approach replaces the stochastic transition by its expectation, yielding a deterministic *fluid LP* and a family of LP-based policies. These are asymptotically optimal per agent ($o(1)$) and have been shown to achieve $\mathcal{O}(1/\sqrt{N})$ gaps [3, 4, 8, 11, 20–22, 25]. Under additional non-degeneracy assumptions (the exact definition of non-degeneracy varies by work [4, 6, 24]), the optimality gap reduces to $\mathcal{O}(1/N)$.

The non-degeneracy assumptions are quite restrictive for many practical applications. In particular, when the constraints of the WCMDPs are saturated, e.g., when the system is operating in a resource-

*Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor. Email: chenyaa@umich.edu, leiying@umich.edu

†Computer Science Department, Carnegie Mellon University. Email: weinaw@cs.cmu.edu

[3]We adopt standard asymptotic notation throughout. Specifically, for functions $f(N)$ and $g(N)$, we write $f(N) = \mathcal{O}(g(N))$ if there exist positive constants C and N_0 such that $|f(N)| \leq C|g(N)|$ for all $N \geq N_0$. Similarly, $f(N) = \Omega(g(N))$ if $g(N) = \mathcal{O}(f(N))$, and $f(N) = \Theta(g(N))$ if both $f(N) = \mathcal{O}(g(N))$ and $f(N) = \Omega(g(N))$ hold. We use $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ to suppress logarithmic factors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

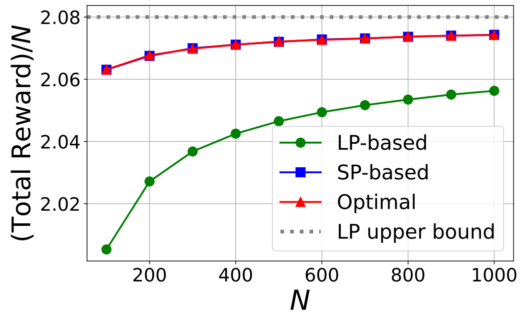

Figure 1: Empty-car routing (WCMDP): SP vs. LP-based policies across N.

constrained scenario such that some resources have to be fully utilized to maximize the total reward, the non-degeneracy assumption is unlikely to hold.

In this paper, we propose a second-order method based on *stochastic-programming* (SP-based), which considers both mean and variance near the fluid optimum. Our contributions:

- We construct a Gaussian stochastic system centered at a fluid-optimal solution $\mathbf{y}^*$; optimizing within its $\widetilde{\Theta}(1/\sqrt{N})$ neighborhood yields a policy implementable in the N-system via integer rounding.
- Under a uniqueness assumption for the fluid LP, the SP-based policy attains an $\widetilde{\mathcal{O}}(1/N)$ optimality gap (Theorem 4.1). This result does not require the problem to be non-degenerate.
- We present a class of problems under which LP policies have $\Theta(1/\sqrt{N})$ optimality gap and furthermore, the LP bound upper bound is $\Theta(1/\sqrt{N})$ larger than the true optimum (Theorem 4.2).

Illustration. In an empty-car routing WCMDP, our SP-based policy is near-optimal while LP-based policies exhibit clear gaps (Figure 1).

2 Model

We consider N statistically identical MDPs (agents) indexed by $n \in \{1, \dots, N\}$ with finite state space $\mathcal{S} = \{1, \dots, S\}$ and action space $\mathcal{A} = \{0, 1, \dots, A-1\}$. At step $h \in \{1, \dots, H\}$, the joint state is $\mathbf{s}_h = (s_{1,h}, \dots, s_{N,h}) \in \mathcal{S}^N$ and a feasible joint action is $\mathbf{a}_h = (a_{1,h}, \dots, a_{N,h}) \in \mathcal{A}^N$. Each agent evolves independently given $(\mathbf{s}_h, \mathbf{a}_h)$, so that $\mathbf{P}_h(\mathbf{s}_{h+1} \mid \mathbf{s}_h, \mathbf{a}_h) = \prod_{n=1}^N \mathbf{P}_h(s_{n,h+1} \mid s_{n,h}, a_{n,h})$. The immediate reward $\sum_{n=1}^N r_h(s_{n,h}, a_{n,h})$ is additive, and there are J resource types with per-epoch budgets $b_j N$. Let $C_j(s,a) \geq 0$ denote the consumption of resource j by (s,a) and assume $C_j(s,0) = 0$. Feasibility at time h requires $\sum_{n=1}^N C_j(s_{n,h}, a_{n,h}) \leq b_j N$, for each $1 \leq j \leq J$. The objective is to find a policy π^N mapping $\mathbf{s}_h$ to $\mathbf{a}_h$ that maximizes the per-agent total reward:

$$V_{\text{opt}}^N(\mathbf{x}_{\text{ini}}, 1) \;:=\; \max_{\pi^N} \sum_{h=1}^H \frac{1}{N} \mathbb{E}\left[\sum_{n=1}^N r_h(s_{n,h}, a_{n,h})\right].$$

Aggregated state–action representation. Let $\mathbf{X}_h \in \Delta^S$ collect the fractions of agents in each state and let $\mathbf{Y}_h \in \Delta^{SA}$ collect the fractions taking each (s,a). Write $\mathbf{r}_h = (r_h(s,a))_{s,a}$ and let $\mathbf{C}$ be the $J \times SA$ matrix with columns $\mathbf{C}(s,a)$; denote the resource vector by $\mathbf{b} \in \mathbb{R}^J$. Feasibility and the objective become

$$\sum_{s,a} Y_h(s,a)\, \mathbf{C}(s,a)^\top \leq \mathbf{b}, \qquad \sum_a \mathbf{Y}_h(\cdot, a) = \mathbf{X}_h, \qquad 1 \leq h \leq H, \tag{1}$$

$$V_{\text{opt}}^N(\mathbf{x}_{\text{ini}}, 1) \;=\; \max_{\pi} \sum_{h=1}^H \mathbb{E}\left[\mathbf{r}_h\, \mathbf{Y}_h^\top\right], \qquad \mathbf{X}_1 = \mathbf{x}_{\text{ini}}. \tag{2}$$

This aggregated model is an equivalent representation of the WCMDP for homogeneous systems, but greatly simplifies the notation.

3 Second-order Gaussian approximation and SP-based policy

Fluid LP (first order) as the baseline. Replacing the stochastic transition by its expectation yields the *fluid LP*:

$$\overline{V}_{\mathrm{LP}}(\mathbf{x}_{\mathrm{ini}},1)=\max_{(\mathbf{x}_h,\mathbf{y}_h)}\sum_{h=1}^{H}\mathbf{r}_h\mathbf{y}_h^\top \tag{3}$$
$$\text{s.t.}\ \sum_{s,a}y_h(s,a)\mathbf{C}(s,a)^\top\le\mathbf{b},\ \sum_a\mathbf{y}_h(\cdot,a)=\mathbf{x}_h,\ \mathbf{x}_{h+1}=\sum_{s,a}y_h(s,a)\mathbf{P}_h(\cdot|s,a),$$

with $\mathbf{x}_1=\mathbf{x}_{\mathrm{ini}}$, $\mathbf{y}_h\ge 0$. Let $(\mathbf{x}^*,\mathbf{y}^*)$ be an optimal solution. It is known that $V_{\mathrm{opt}}^N\le\overline{V}_{\mathrm{LP}}$ [4], and LP-based policies derived from $(\mathbf{x}^*,\mathbf{y}^*)$ are efficient yet may incur $\Theta(1/\sqrt{N})$ gaps, see Theorem 4.2 below.

Gaussian surrogate (second order). To capture variance, we define a Gaussian system on Δ^S with the same initial state $\mathbf{x}_{\mathrm{ini}}$. For action $\mathbf{y}_h$, consider

$$\widetilde{\mathbf{X}}_{h+1}=\mathrm{Proj}_{\Delta^S}\left(\sum_{s,a}y_h(s,a)\mathbf{P}_h(\cdot|s,a)+\mathbf{Z}_h/\sqrt{N}\right), \tag{4}$$

where $\mathbf{Z}_h$ is zero-mean with covariance matching the N-system if applying action $\mathbf{y}_h^*$ (action-*independent* noise near $\mathbf{y}_h^*$ keeps optimization tractable). Projection is rarely active (with probability $1-\widetilde{\mathcal{O}}(N^{-\log N})$). This second-order correction is of scale $1/\sqrt{N}$ (CLT regime) and is accurate in an $\widetilde{\mathcal{O}}(1/\sqrt{N})$-neighborhood of $\mathbf{y}_h^*$.

Neighborhood policy class and Gaussian SP. Let $\delta_N=2\log N/\sqrt{N}$, fix $\kappa>0$, a fallback policy $\pi^\perp$, and a sequence z_h independent of N. Define

$$\Pi_{\delta_N}(\mathbf{y}^*)\ =\ \Big\{\pi:\ \|\pi(\mathbf{x}_h,h)-\mathbf{y}_h^*\|_\infty\le\kappa z_h\delta_N\ \text{if}\ \|\mathbf{x}_h-\mathbf{x}_h^*\|_\infty\le z_h\delta_N;\ \text{else}\ \pi=\pi^\perp\Big\}. \tag{5}$$

The $\widetilde{\Theta}(1/\sqrt{N})$ radius is tight: larger neighborhoods inflate second-order error; smaller ones risk excluding the N-optimal policy.

We then optimize over $\Pi_{\delta_N}(\mathbf{y}^*)$ on the following Gaussian stochastic system:

$$\max_{\pi\in\Pi_{\delta_N}(\mathbf{y}^*)}\sum_{h=1}^{H}\mathbb{E}\left[\mathbf{r}_h\,\widetilde{\mathbf{Y}}_h^\top\right] \tag{6}$$
$$\text{s.t.}\quad\sum_{s,a}\widetilde{Y}_h(s,a)\mathbf{C}(s,a)^\top\le\mathbf{b},\ \sum_a\widetilde{\mathbf{Y}}_h(\cdot,a)=\widetilde{\mathbf{X}}_h,\ \text{system transition follows (4).}$$

Let $\widetilde{\pi}^{N,*}\in\arg\max_{\pi\in\Pi_{\delta_N}(\mathbf{y}^*)}$ be an SP-optimal policy on the Gaussian system. The SP-based policy applied to the N-system is described in Algorithm 1.

Implementation & complexity. Solving (6) amounts to a low-variance stochastic program restricted to a small-radius neighborhood, with i.i.d. Gaussian noise; in practice one can construct in the $1/\sqrt{N}$ scale an N-independent and projection-free SP from (6), and use well-established algorithm such as SDDP [16, 18] and EDDP [13] to solve its sample-average approximation. In addition, we round the action so that $N\mathbf{Y}_h\in\mathbb{N}^{SA}$, as in Lines 6 and 8 of Algorithm 1.

4 Main results

Let V_π^N (resp. $\widetilde{V}_\pi^N$) denote the value of π in the N-system (resp. Gaussian system), and define the Q-functions analogously.

Assumption 4.1 (Uniqueness)**.** *The fluid LP* (3) *has a unique optimal solution* $\mathbf{y}^*$.

Algorithm 1 Stochastic-programming-based (SP-based) policy for the N-system

Input: Fluid-optimal $\mathbf{y}^*$ from (3); constants z_h, δ_N, κ; fallback policy $\pi^\perp$
Solve (6) to obtain $\widetilde{\pi}^{N,*} \in \Pi_{\delta_N}(\mathbf{y}^*)$
for $h = 1$ to H **do**
    Observe N-system state $\mathbf{X}_h$
    if $\|\mathbf{X}_h - \mathbf{x}_h^*\|_\infty \le z_h \delta_N$ **then**
        $\mathbf{Y}_h \leftarrow \mathrm{round}\big(\widetilde{\pi}^{N,*}(\mathbf{X}_h, h)\big)$
    else
        $\mathbf{Y}_h \leftarrow \mathrm{round}\big(\pi^\perp(\mathbf{X}_h, h)\big)$
    end if
    Apply integer action $\mathbf{Y}_h$ (so $N\mathbf{Y}_h$ has integer entries)
end for

Theorem 4.1 (SP-based policy is $\widetilde{\mathcal{O}}(1/N)$-optimal)**.** *Under the Uniqueness Assumption 4.1, the SP-based policy in Algorithm 1 satisfies*

$$V_{\mathrm{opt}}^N(\mathbf{x}_{\mathrm{ini}}, 1) - V_{\widetilde{\pi}^{N,*}}^N(\mathbf{x}_{\mathrm{ini}}, 1) = \widetilde{\mathcal{O}}(1/N).$$

Sketch. *(i) An optimal policy for the N-system belongs to $\Pi_{\delta_N}(\mathbf{y}^*)$ under the Uniqueness Assumption 4.1; (ii) local second-order accuracy of the Gaussian surrogate in a $\widetilde{\Theta}(1/\sqrt{N})$ neighborhood of $(\mathbf{x}^*, \mathbf{y}^*)$; (iii) rounding loss is $\widetilde{\mathcal{O}}(1/N)$.*

LP baselines are tightly $\Theta(1/\sqrt{N})$. Consider the LP-based policies class

$$\Pi_{\text{fluid}}(\mathbf{y}^*) = \Big\{\pi : \ \|\pi(\mathbf{x}_h, h) - \mathbf{y}_h^*\|_\infty \le \kappa \|\mathbf{x}_h - \mathbf{x}_h^*\|_\infty, \ \forall h\Big\}. \tag{7}$$

Theorem 4.2 (Tight lower bounds)**.** *There exist WCMDPs where every $\pi \in \Pi_{\mathit{fluid}}(\mathbf{y}^*)$ satisfies $V_{\mathrm{opt}}^N(\mathbf{x}_{\mathrm{ini}}, 1) - V_\pi^N(\mathbf{x}_{\mathrm{ini}}, 1) = \Theta(1/\sqrt{N})$, and the LP bound is loose: $\overline{V}_{\mathrm{LP}}(\mathbf{x}_{\mathrm{ini}}, 1) - V_{\mathrm{opt}}^N(\mathbf{x}_{\mathrm{ini}}, 1) = \Theta(1/\sqrt{N})$.*

Discussion & insights. The $\widetilde{\Theta}(1/\sqrt{N})$ neighborhood is not just a technical convenience: it is the *CLT scale* at which random fluctuations of the N-system live. Optimizing the policy over a larger neighborhood would increase second–order approximation error, which would become dominate in the optimality gap analysis, while reducing the neighborhood may exclude the N-optimal policy. This explains why our policy class $\Pi_{\delta_N}(\mathbf{y}^*)$ is both *necessary* (to retain optimal policies under uniqueness) and *sufficient* (to keep the approximation error small enough).

Role of uniqueness. When the fluid LP has a unique optimum, the value surface near $(\mathbf{x}^*, \mathbf{y}^*)$ behaves like a well-conditioned local summit: the N-optimal policy concentrates within a $\widetilde{\Theta}(1/\sqrt{N})$ tube around $\mathbf{y}^*$. In that regime, variance-aware corrections are reliably beneficial and the global gap contracts to $\widetilde{\mathcal{O}}(1/N)$ (Theorem 4.1).

5 Conclusion

We developed a second-order SP-based policy for finite-horizon WCMDPs by optimizing a Gaussian surrogate in a $\widetilde{\Theta}(1/\sqrt{N})$ neighborhood of the fluid optimum. Under LP uniqueness, we established that the policy achieved an $\widetilde{\mathcal{O}}(1/N)$ optimality gap, and we showed that the fluid LP bound was $\Theta(1/\sqrt{N})$ loose. Complete proofs and additional experiments are provided in the extended version of this work.

References

[1] Peyman Ashkrof, Gonçalo Homem de Almeida Correia, Oded Cats, and Bart Van Arem. On the relocation behavior of ride-sourcing drivers. *Transportation Letters*, 16(4):330–337, 2024.

[2] Anton Braverman, Jim Dai, Xin Liu, and Lei Ying. Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452, 2019.

[3] David B. Brown and James E. Smith. Index policies and performance bounds for dynamic selection problems. *Manag. Sci.*, 66:3029–3050, 2020.

[4] David B. Brown and Jingwei Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research*, 2023.

[5] Jim Dai, Manxi Wu, and Zhanhao Zhang. Atomic proximal policy optimization for electric robo-taxi dispatch and charger allocation. *arXiv preprint arXiv:2502.13392*, 2025.

[6] Nicolas Gast, Bruno Gaujal, and Chen Yan. Reoptimization nearly solves weakly coupled Markov decision processes. *arXiv preprint arXiv:2211.01961*, 2024.

[7] Yasin Gocgun and Archis Ghate. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers and Operations Research*, 39(10):2323–2336, 2012. ISSN 0305-0548.

[8] Diego Goldsztajn and Konstantin Avrachenkov. Asymptotically optimal policies for weakly coupled Markov decision processes. *arXiv preprint arXiv:2406.04751*, 2024.

[9] Xiaotong Guo, Nicholas S. Caros, and Jinhua Zhao. Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand. *Transportation Research Part B: Methodological*, 150:161–189, 2021.

[10] Jeffrey Thomas Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

[11] D. J. Hodge and K. D. Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Adv. in Appl. Probab.*, 47(3):652–667, 09 2015.

[12] Yan Jiao, Xiaocheng Tang, Zhiwei Tony Qin, Shuaiji Li, Fan Zhang, Hongtu Zhu, and Jieping Ye. Real-world ride-hailing vehicle repositioning using deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 130:103289, 2021.

[13] Guanghui Lan. Complexity of stochastic dual dynamic programming. *Mathematical Programming*, 191(2):717–754, 2022.

[14] Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L. Dean, and Craig Boutilier. Solving very large weakly coupled Markov decision processes. In *AAAI/IAAI*, pages 165–172, 1998.

[15] Jonathan Patrick, Martin L. Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525, 2008.

[16] M.V.F. Pereira and L.M.V.G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical programming*, 52:359–375, 1991.

[17] Mahshid Salemi Parizi. *Approximate dynamic programming for weakly coupled Markov decision processes with perfect and imperfect information*. PhD thesis, 2018.

[18] Alexander Shapiro. Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209(1):63–72, 2011.

[19] Jiahui Sun, Haiming Jin, Zhaoxing Yang, Lu Su, and Xinbing Wang. Optimizing long-term efficiency and fairness in ride-hailing via joint order dispatching and driver repositioning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3950–3960, 2022.

[20] Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016.

[21] Guojun Xiong, Jian Li, and Rahul Singh. Reinforcement learning for finite-horizon restless multi-armed multi-action bandits. *arXiv preprint arXiv:2109.09855*, 2021.

[22] Guojun Xiong, Shufan Wang, and Jian Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35:17911–17925, 2022.

[23] Zhengtian Xu, Yafeng Yin, Xiuli Chao, Hongtu Zhu, and Jieping Ye. A generalized fluid model of ride-hailing systems. *Transportation Research Part B: Methodological*, 150:587–605, 2021.

[24] Jingwei Zhang. Leveraging nondegeneracy in dynamic resource allocation. *Available at SSRN*, 2024.

[25] Xiangcheng Zhang, Yige Hong, and Weina Wang. Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs. *arXiv preprint arXiv:2502.06072*, 2025.