Universität Konstanz

A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance

Axel Friedrich Wolter

Joint Work with Tobias Sutter, XVIIth ICSP, 30.07.2025

Motivation & Challenges

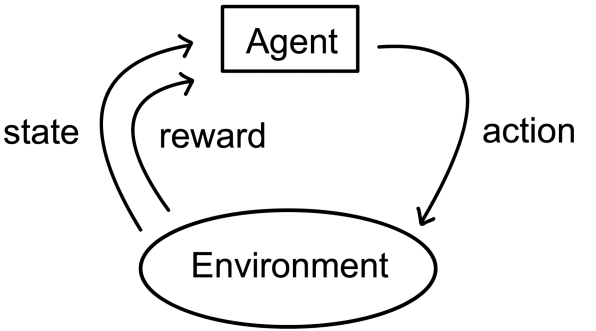
- Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$
- $\mathcal{S}$ and $\mathcal{A}$ finite state and action space
- $r(s, a) \in [0, 1]$ and $\gamma \in (0, 1)$

Infinite Horizon Discounted Reward

Find $\pi^\star$ such that

$$V_{ur}^\star(s) = \max_{\pi \in \Pi} V_{ur}^\pi(s),$$

where $V_{ur}^\pi(s) := \mathbb{E}_s^\pi \left[\sum_{k=0}^\infty \gamma^k r(s_k, a_k)\right]$



Reinforcement Learning Loop

RL as a Saddle Point Problem

Linear Programming formulation

$V_{ur}^{\star}$ is the minimizer to the LP

$$\begin{cases} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} & \mu^{\top} V \\ \text{s.t.} & 0 \geq \underbrace{-V(s) + r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) V(s')}_{:=\Delta[V](s,a)} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \end{cases}$$

with $\mu \in \mathbb{R}_{+}^{|\mathcal{S}|}$

[1]

$$\max_{\rho \geq 0} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L_{ur}(V, \rho) := \mu^{\top} V + \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \rho(s,a) \Delta[V](s,a), \quad \pi^{\rho}(a|s) = \frac{\rho(s,a)}{\sum_{a' \in \mathcal{A}} \rho(s,a')}$$

[1] Manne (1960); Borkar (2002); Puterman (1994); Chen and Wang (2016); Lee and He (2019)

Regularized Lagrangian Formulation

- Entropy regularization[2] and convex modification

$$\max_{\rho \geq 0} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L_{ur}(V, \rho) := \mu^\top V + \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \rho(s,a) \Delta[V](s,a)$$

$$\max_{\rho \geq 0} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L(V, \rho) := \frac{\eta_V}{2} \|\boldsymbol{V}\|_2^2 + \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \rho(s,a) \left(\Delta[V](s,a) - \boldsymbol{\eta_\rho} \log \left(\frac{\boldsymbol{\rho(s,a)}}{\sum_{a' \in \mathcal{A}} \boldsymbol{\rho(s,a')}} \right) \right)$$

3

[2] Haarnoja et al. (2018); Zhan et al. (2022)
[3] Geist et al. (2019); Neu et al. (2017); Ying and Zhu (2020); Li et al. (2024)

Properties of Regularized RL Objective

- Exploration and robustness[4] at the cost of **bounded suboptimality**[5]

$$V_{ur}^{\star}(s) - \eta_{\rho}\frac{\log|\mathcal{A}|}{1-\gamma} \leq V_{ur}^{\pi_r^{\star}}(s) \leq V_{ur}^{\star}(s)$$

- **Unique stochastic optimal policy** $\pi_r^{\star}$ induced by optimal dual $\rho^{\star} \in H \subset \mathbb{R}_{>0}^{|\mathcal{S}|\cdot|\mathcal{A}|}$
- No effect of strongly convex modification in V on optimal regularized primal solution[6]

[4] Derman et al. (2021)
[5] Geist et al. (2019, Theorem 2)
[6] Li et al. (2024, Theorem 2.1)

Overview of PGDA-RL — Generative Model Access

Synchronous Algorithm (PGDA-RL)

1. **Sample state-transitions** $s' \sim \mathcal{P}(\cdot|s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

↓

2. **Primal update:** $V_{k+1} \leftarrow V_k - \boldsymbol{\alpha_k} \hat{\nabla}_V L(V_k, \rho_k)$

↓

3. **Dual update:** $\rho_{k+1} \leftarrow \Pi_H \left[\rho_k + \boldsymbol{\beta_k} \hat{\nabla}_\rho L(V_k, \rho_k)\right]$

↓

Repeat for $k = 1, 2, \ldots, K$

Output: final iterates (V_K, ρ_K)

Convergence of PGDA-RL

- Stochastic gradients $\hat{\nabla}_V L$ and $\hat{\nabla}_\rho L$ estimated via generative model
- Step sizes on two timescales: $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$,
  with $\sum_{k\in\mathbb{N}} a_k = \infty$, $\sum_{k\in\mathbb{N}} a_k^2 < \infty$, $a \in \{\alpha, \beta\}$

Theorem (Almost Sure Convergence)

Under the above assumptions on step sizes and bounded noise, the iterates of PGDA-RL converge almost surely:

$$\lim_{K\to\infty} (V_K, \rho_K) = (V^\star, \rho^\star) \quad \text{a.s.}$$

Proof: Two-timescale ODE method from stochastic approximation theory (Borkar, 1997)

ODE Approach to Stochastic Approximation

- Classic SA update with martingale difference noise sequence (MDS)[7]:

$$x_{k+1} = x_k + \alpha_k \left(h(x_k) + M_k \right)$$

- Noisy Euler discretization of the ODE:

$$\dot{x}(t) = h(x(t))$$

- For $\sum_{k\in\mathbb{N}} \alpha_k = \infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < \infty$, well-behaved MDS M_k, and $\sup_k \|x_k\| < \infty$ a.s.:

$$x_k \to x^\star \quad \text{a.s.}, \quad \text{where } h(x^\star) = 0$$

[7] Kushner and Yin (2003); Borkar (2023)

Two-Timescale Structure and Limiting Dynamics

Stochastic Recursion

$$V_k = V_{k-1} - \alpha_k \left(\nabla_V L + M_k^{(1)} \right)$$
$$\rho_k = \Pi_H \left[\rho_{k-1} + \beta_k \left(\nabla_\rho L + M_k^{(2)} \right) \right]$$

Limiting ODEs

$$\dot{V}(t) = -\nabla_V L(V(t), \rho) \quad \text{(fast)}$$
$$\lambda(\rho) \in \arg\min_V L(V, \rho)$$
$$\dot{\rho}(t) = \nabla_\rho L(\lambda(\rho(t)), \rho(t)) + \zeta(t)$$

- $\alpha_k \gg \beta_k$ (fast primal, slow dual)
- $M_k^{(1)}$, $M_k^{(2)}$ are MDS
- $\zeta(t)$ projects flow to stay within H

Coupled stochastic recursion tracks a globally asymptotically stable two-timescale dynamical system.

From Synchronous to Asynchronous PGDA-RL

Motivation

- No access to generative model — learn from a single trajectory
- Only visited (s, a) pairs are updated per iteration
- Use the current dual iterate to guide on-policy exploration

Key Differences to Synchronous Setting

	Synchronous	**Asynchronous**
Gradient Estimation	Per-step simulator samples	Replay-based estimates
Policy Exploration	-	Adaptive via ρ_k
Gradient Noise	I.I.D. across all (s, a)	Asymptotically unbiased
Step Sizes	Uniform schedule	Entry-specific steps

Convergence via Stochastic Differential Inclusion

Theorem Convergence of Asynchronous PGDA-RL (Informal)

Under suitable step-size conditions and the replay-based gradient structure, the iterates of Asynchronous PGDA-RL converge almost surely:

$$\lim_{K\to\infty} (V_K, \rho_K) = (V^\star, \rho^\star) \quad \text{a.s.}$$

Main Change in the Proof — Stochastic Differential Inclusion Analysis[8]

Captures non-uniform, time-varying update structure

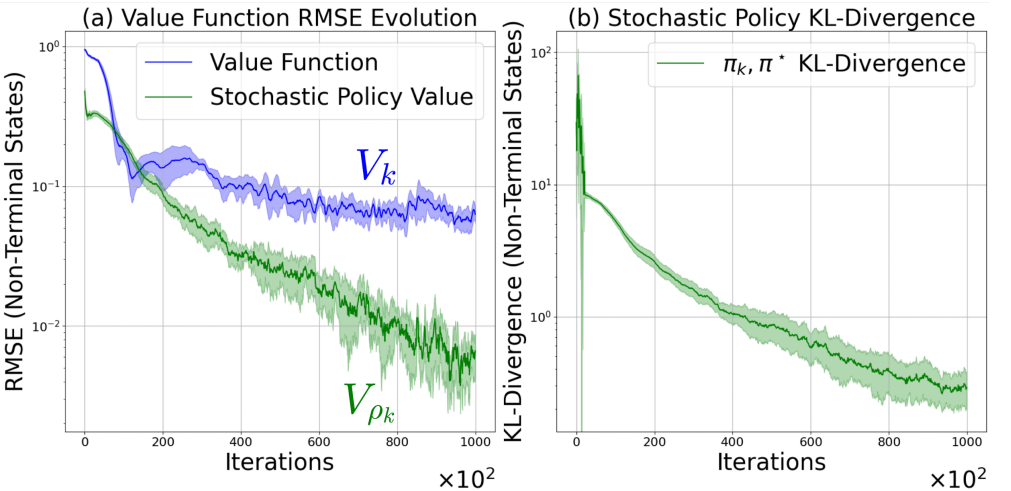
$$\dot{V}(t) \in -\Omega^\epsilon_{\mathcal{S}} \cdot \nabla_V L, \quad \dot{\rho}(t) \in \Omega^\epsilon_{\mathcal{S}\times\mathcal{A}} \cdot \nabla_\rho L + \zeta(t)$$

[8] Perkins and Leslie (2013)

Simulation Results on Frozen Lake

Frozen Lake[a]

[a] Towers et al. (2024)



Parameters:

Asynchronous PGDA-Agent, 10 runs à 10^5 steps,
$\{\gamma, \eta_\rho, \eta_V\} = \{0.9, 0.1, 0.1\}$, $\alpha_k = \frac{1}{\tilde{k}}, \beta_k = \frac{1}{1+\tilde{k}\log(\tilde{k})}$,

Comparison to Related Work

Paper	Model Access	Function Appr.	Convergence
Chen and Wang (2016)	Generator	Tabular	PAC
Dai et al. (2018)	Markovian	Non-linear	Sketch
Lee and He (2019)	Markovian	Tabular	PAC
Gabbianelli et al. (2024)	Offline	Linear	PAC
Li et al. (2024)	Full Model	Tabular	Det. Asympt.
Our paper	Markovian	Tabular	Almost Sure

Conclusion and Key Takeaways

- Proposed PGDA-RL: primal-dual algorithm with almost sure convergence
- Asynchronous learning using replay buffers (on-policy exploration and off-policy data)
- Two-timescale updates ensure stable value and dual variable learning

Reference Paper:

arXiv:2505.04494

References I

V. S. Borkar. Stochastic approximation with two time scales. Systems & Control Letters, 29(5):291–294, 1997.

V. S. Borkar. Convex Analytic Methods in Markov Decision Processes, pages 347–375. Springer US, Boston, MA, 2002. ISBN 978-1-4615-0805-2. doi: 10.1007/978-1-4615-0805-2_11.

V. S. Borkar. Stochastic approximation: a dynamical systems viewpoint. Springer Nature Singapore, second edition, 2023.

Y. Chen and M. Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. arXiv preprint arXiv:1612.02516, 2016.

B. Dai, A. Shaw, N. He, L. Li, and L. Song. Boosting the actor with dual critic. In International Conference on Learning Representations, 2018.

E. Derman, M. Geist, and S. Mannor. Twice regularized mdps and the equivalence between robustness and regularization. Advances in Neural Information Processing Systems, 34:22274–22287, 2021.

G. Gabbianelli, G. Neu, M. Papini, and N. M. Okolo. Offline primal-dual reinforcement learning for linear MDPs. In International Conference on Artificial Intelligence and Statistics, pages 3169–3177, 2024.

M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In International Conference on Machine Learning, pages 2160–2169, 2019.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International Conference on Machine Learning, pages 1861–1870, 2018.

H. J. Kushner and G. G. Yin. Convergence with probability one: Correlated noise. Stochastic Approximation and Recursive Algorithms and Applications, pages 161–212, 2003.

References II

D. Lee and N. He. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In 2019 American Control Conference (ACC), pages 4897–4902, 2019.

H. Li, H.-F. Yu, L. Ying, and I. S. Dhillon. Accelerating primal-dual methods for regularized Markov decision processes. SIAM Journal on Optimization, 34(1):764–789, 2024.

A. S. Manne. Linear programming and sequential decisions. Management Science, 6(3):259–267, 1960. ISSN 00251909, 15265501.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. arXiv preprint arXiv:1705.07798, 2017.

S. Perkins and D. S. Leslie. Asynchronous stochastic approximation with differential inclusions. Stochastic Systems, 2(2):409–446, 2013.

M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., 1st edition, 1994. ISBN 0471619779.

M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulao, A. Kallinteris, M. Krimmel, A. KG, et al. Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032, 2024.

L. Ying and Y. Zhu. A note on optimization formulations of Markov decision processes. arXiv preprint arXiv:2012.09417, 2020.

W. Zhan, B. Huang, A. Huang, N. Jiang, and J. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In Conference on Learning Theory, pages 2730–2775, 2022.

Structured Experience Replay Buffer

Buffer Construction

Trajectory step: (s_{k-1}, a_{k-1}, s_k)

↓

Append s_k to $\mathcal{D}_k(s_{k-1}, a_{k-1})$

↓

Add (s_{k-1}, a_{k-1}) to $\mathcal{D}_{\text{inc},k}(s_k)$

Sampling for Gradient Estimation

Current state-action: $(\bar{s}, \bar{a})$

↓

Sample m transitions from $\mathcal{D}_k(\bar{s}, \bar{a})$

↓

Estimate $\hat{\mathcal{P}}_k^{(2)}(s'|\bar{s}, \bar{a})$ via empirical freq.

↓

Compute $\hat{\nabla}_\rho L(V, \rho)(\bar{s}, \bar{a})$