A Theoretical Justification for Asymmetric Actor-Critic Algorithms

Mila RL Sofa - December 20th, 2024 (updated May 31st, 2025)

Gaspard Lambrechts, Damien Ernst and Aditya Mahajan

Outline

Asymmetric Observability 3
Asymmetric Learning 9
A Theoretical Justification for Asymmetric Actor-Critic Algorithms 14
Conclusion 27

Asymmetric Observability

Partial observability

A **POMDP** is described by a model $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, O, T, R, \gamma)$.

- States $s_t \in \mathcal{S}$,
- Actions $a_t \in \mathcal{A}$,
- Observations $o_t \in \mathcal{O}$,
- Initialization $s_0 \sim P(\cdot)$,
- Perception $o_t \sim O(\cdot\,|s_t)$,
- Transition $s_{t+1} \sim T(\cdot\,|s_t, a_t)$,
- Reward $r_t \sim R(\cdot\,|s_t, a_t)$,
- Discount $\gamma \in [0, 1)$.

The **history** at time t is $h_t = (o_0, a_0, ..., o_t) \in \mathcal{H}$.

Definition 1: History-dependent stochastic policy.

A history-dependent stochastic policy $\pi \in \Pi = \mathcal{H} \to \Delta(\mathcal{A})$ is a mapping from histories to distributions over the actions, with density $\pi(a|h)$.

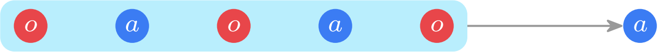


Fig. 1: History-dependent policy.

Learning under partial observability

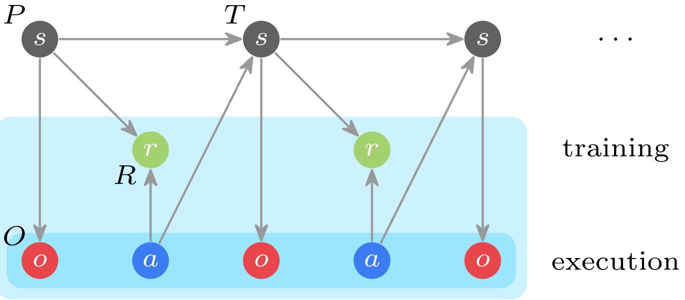


Fig. 2: Bayesian graph of a POMDP.

The problem of **RL in POMDP** is to find an optimal history-dependent policy

$$\pi^* \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right]$$

from samples $(o_0, a_0, r_0, \ldots, o_t)$.

Asymmetric observability

Decision process	Execution	Training	Generality
MDP	s	s	**Too optimistic.**
POMDP	o	o	**Too pessimistic.**
Privileged POMDP	o	s	**Too optimistic.**
Informed POMDP	o	i	**Just right?**

Examples: simulator state, trajectory in hindsight, additional sensors, additional viewpoints, observations of other agents, etc.

Learning under asymmetric observability

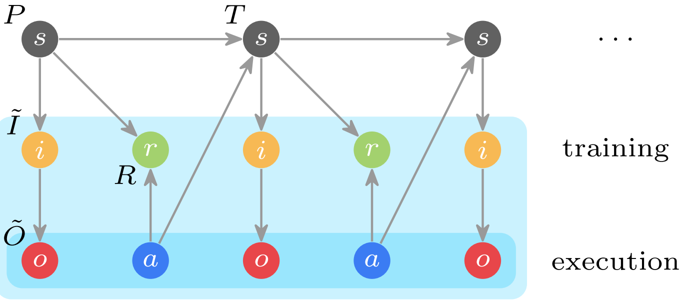


Fig. 3: Bayesian graph of an informed POMDP.

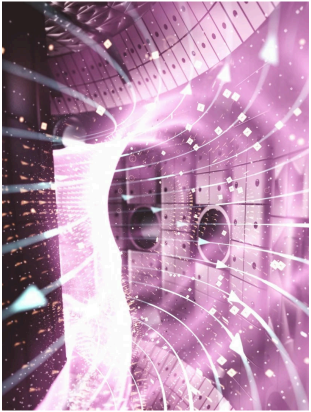
The problem of **RL in POMDP** is to find an optimal history-dependent policy

$$\pi^* \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right]$$

from samples $(i_0, o_0, a_0, r_0, \dots, i_t, o_t)$.

Asymmetric learning is successful

- Magnetic Control of Tokamak Plasma through Deep RL (Degrave et al., 2022).
- Champion-Level Drone Racing using Deep RL (Kaufmann et al., 2023).
- A Super-Human Vision-Based RL Agent in Gran Turismo (Vasco et al., 2024).



The international journal of science / 31 August 2023
nature
DRONE RACER
AI pilot beats human champions in aerial contest
Offset agreement
Overhaul pricing of carbon credits to help fund climate projects
Dining companions
Corals devour algal partners when food supplies run low
Sentence synthesis
Brain implants show promise in rendering speech from thoughts

The international journal of science / 10 February 2022
nature
DRIVING FORCE
AI algorithm outcompetes human champions in *Gran Turismo* racing game
RACE TOGETHER 2021

Image credits: first, second, third.

Asymmetric Learning

1. Imitation learning approaches

The idea consists of imitating an expert policy (Choudhury et al., 2018):

1. Learn a policy for the MDP: $\mu(a|s)$,
2. Imitate the policy: $\pi(a|h) \approx \mu(a|s)$.

It is known to be suboptimal: the policy corresponds to the greedy policy with respect to the Q-MDP approximation (Littman et al., 1995):

$$\hat{V}_{\text{POMDP}}(h) = \mathbb{E}[V_{\text{MDP}}(S)|H = h].$$

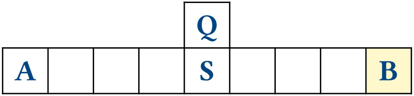


Fig. 4: Environment with random unobserved goal where imitation is optimal.

Recent works have constrained the expert policy such that its imitation results in an optimal history-dependent policy (Warrington et al., 2021).

2. Asymmetric critic approaches

The idea comes from the lack of need of the critic at execution (Pinto et al., 2018):

1. The policy is conditioned on the history: $\pi(a|h)$.
2. The critic is conditioned on the state: $Q(s, a)$.
3. The policy gradient is approximated using: $\log \pi(a|h) Q(s, a)$.

It is known to be biased or even ill-defined: the environment state s is not a Markovian state of the future execution of the environment and policy $\pi(a|h)$.

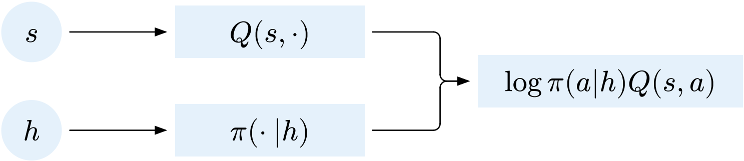


Fig. 5: Illustration of the asymmetric actor-critic.

Recent works have proposed a well-defined and unbiased asymmetric actor-critic by introducing the history-state critic $Q(h, s, a)$ (Baisero & Amato, 2022).

3. Representation learning approaches

The idea comes from the sufficiency of the belief (Nguyen et al., 2021).

1. The history is compressed into a statistic: $z = f(h)$.
2. The statistic is trained to be predictive of the belief: $\hat{b}(s|z) \approx p(s|h)$.
3. The policy is conditioned on that statistic: $\pi(a|h) = g(a|z)$.

It is known to be unrealistic: computing the belief $p(s|h)$ requires the dynamics to be known and is in general intractable.

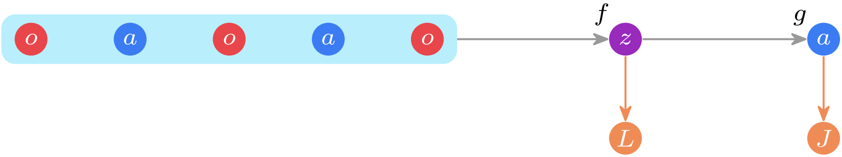


Fig. 6: Representation learning in POMDP.

Recent works have proposed to learn belief representations from the sample states that are observed at training time (Wang et al., 2023).

4. Model-based learning approaches

Several concurrent approaches to replace the word model $q(r, o'|h, a)$:

- **Bisimulation** of belief world model (Avalos et al., 2024):
  1. Predict the belief: $b = f(h)$.
  2. Learn a belief world model: $b' = g(b, a, o')$.
- **Representation learning** with asymmetric model (Lambrechts et al., 2024):
  1. Predict the next information: $q(r, i'|h, a)$.
  2. Use latent policy to learn from the informed world model.
- **Distillation** of a privileged world-model (Hu et al., 2024):
  1. Learn a privileged world model: $q(r, o'|h^+, a)$.
  2. Distillate the world model: $q(r, o'|h, a)$.

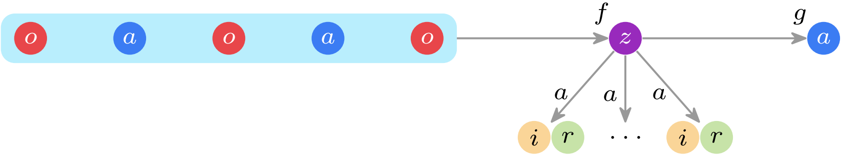


Fig. 7: Informed world model.

A Theoretical Justification for Asymmetric Actor-Critic Algorithms

Lack of justification

To sum up, while early approaches were **heuristic**, a recent line of work has focused on proposing **theoretically grounded** objectives:

- They provide optimal history-dependent policies when satisfied,
- They make use of the additional state information.

But there are still **no theoretical justification for the benefits**. While at optimality policies are equivalent, asymmetric learning should converge faster.

Goal of this work: justification for the asymmetric actor-critic algorithm.

NB: Some explanations exist in the literature (Baisero & Amato, 2022; Sinha & Mahajan, 2023). Recently, an asymmetric actor-critic relying on learning beliefs was shown more efficient than symmetric learning (Cai et al., 2024).

Asymmetric actor-critic algorithm

We make the following assumptions:

- **Discrete space:**
  - State space $\mathcal{S}$, observation space $\mathcal{O}$, action space $\mathcal{A}$.
  - Agent state space $\mathcal{Z}$.
- **Finite state policy:**
  - Agent state process $z_{t+1} \sim U(\cdot\,|z_t, a_t, o_{t+1})$.
  - Policy $a_t \sim \pi(\cdot\,|z_t)$.
- **Finite state Q-functions:**
  - Asymmetric $\mathcal{Q}^\pi(s, z, a) = \mathbb{E}^\pi\left[\sum_{t=0} \gamma^t R_t | S_0 = s, Z_0 = z, A_0 = a\right]$.
  - Symmetric $Q^\pi(z, a) = \mathbb{E}^\pi\left[\sum_{t=0} \gamma^t R_t | Z_0 = z, A_0 = a\right]$.
- **Linear Q-functions approximations:**
  - Asymmetric $\hat{\mathcal{Q}}^\pi_\beta(\cdot) = \langle \beta, \varphi(\cdot) \rangle$ with $\varphi : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathbb{R}^d$.
  - Symmetric $\hat{Q}^\pi_\beta(\cdot) = \langle \beta, \chi(\cdot) \rangle$ with $\chi : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}^d$.
- **Log-linear policy:**
  - $\pi_\theta(a|z) \propto \exp(\langle \theta, \psi(z, a) \rangle)$.

Asymmetric actor-critic algorithm (ii)

We use the abbreviations $\hat{\mathcal{Q}}^{\pi}_{k,i} = \hat{\mathcal{Q}}^{\pi}(s_{k,i}, z_{k,i}, a_{k,i})$ and $\hat{Q}^{\pi}_{k,i} = \hat{Q}^{\pi}(z_{k,i}, a_{k,i})$.

The **asymmetric semi-gradient** is,

$$g_k = \left(\sum_{i=0}^{m-1} \gamma^i r_{k,i} + \gamma^m \hat{\mathcal{Q}}^{\pi}_{k,m} - \hat{\mathcal{Q}}^{\pi}_{k,0} \right) \nabla_{\beta} \hat{\mathcal{Q}}^{\pi}_{k,0}. \tag{1}$$

and the **symmetric semi-gradient** is,

$$g_k = \left(\sum_{i=0}^{m-1} \gamma^i r_{k,i} + \gamma^m \hat{Q}^{\pi}_{k,m} - \hat{Q}^{\pi}_{k,0} \right) \nabla_{\beta} \hat{Q}^{\pi}_{k,0}. \tag{2}$$

Interestingly, the asymmetric Q-function $\mathcal{Q}(s, z, a)$ is the solution of its Bellman equation, while the symmetric Q-function $Q(z, a)$ is not!

Asymmetric actor-critic algorithm (iii)

Algorithm 1: m-step TD learning.

```
1. For $k = 0, \ldots, K - 1$:
   1. Sample $s_{k,0}, z_{k,0}$ from the discounted visitation measure $d^{\pi}(\cdot)$.
   2. For $i = 0, \ldots, m - 1$:
      1. Take action $a_{k,i} \sim \pi\left(\cdot \mid z_{k,i}\right)$.
      2. Observe $r_{k,i+1}, s_{k,i+1}, o_{k,i+1}, z_{k,i+1}$ according to $R, T, O, U$.
   3. Sample last action $a_{k,m} \sim \pi\left(z_{k,m}\right)$.
   4. Compute semi-gradient $g_k$ using (1) or (2).
   5. Update parameters: $\beta_{k+1} = \Gamma_{\mathcal{B}(0,B)}(\beta_k + \alpha g_k)$.
2. Return average estimate $\overline{Q}^{\pi}(\cdot) = \langle \overline{\beta}, \varphi(\cdot) \rangle$ or $\overline{Q}^{\pi}(\cdot) = \langle \overline{\beta}, \chi(\cdot) \rangle$ with average parameter $\overline{\beta} = \frac{1}{K} \sum_{k=0}^{K-1} \beta_k$.
```

Asymmetric actor-critic algorithm (iv)

Let us define the asymmetric and symmetric advantage functions:

$$\hat{\mathcal{A}}^{\pi}(s, z, a) = \hat{\mathcal{Q}}^{\pi}(s, z, a) - \sum_{a \in \mathcal{A}} \hat{\mathcal{Q}}^{\pi}(s, z, a)$$

$$\hat{A}^{\pi}(z, a) = \hat{Q}^{\pi}(z, a) - \sum_{a \in \mathcal{A}} \hat{Q}^{\pi}(z, a).$$

We use the abbreviations $\hat{\mathcal{A}}^{\pi} = \hat{\mathcal{A}}^{\pi}(s, z, a)$ and $\hat{A}^{\pi} = \hat{A}^{\pi}(z, a)$.

Finally, the **asymmetric natural gradient loss** is,

$$v_{t,n} = \nabla_w \Big(\langle \nabla_\theta \log \pi_\theta(a_{t,n} | z_{t,n}), w_{t,n} \rangle - \overline{\mathcal{A}}_{t,n} \Big)^2. \tag{3}$$

and the **symmetric natural gradient loss** is,

$$v_{t,n} = \nabla_w \Big(\langle \nabla_\theta \log \pi_\theta(a_{t,n} | z_{t,n}), w_{t,n} \rangle - \overline{A}_{t,n} \Big)^2. \tag{4}$$

Asymmetric actor-critic algorithm (v)

Algorithm 2: Natural actor critic.

```
1. For $t = 0, ..., T - 1$:
    1. Obtain $\overline{\mathcal{Q}}^{\pi_t}$ or $\overline{Q}^{\pi_t}$ using Algorithm 1.
    2. For $n = 0, ..., N - 1$:
        1. Sample $s_{k,n}, z_{k,n}$ from the discounted visitation measure $d^{\pi}(\cdot)$.
        2. Take action $a_{k,n} \sim \pi(\cdot \mid z_{k,n})$.
        3. Compute gradient $v_{t,n}$ of the natural policy gradient using (3) or (4).
        4. Update natural policy gradient: $w_{t,n+1} = \Gamma_{\mathcal{B}(0,B)}(w_{t,n} + \zeta v_{t,n})$.
    3. Estimate natural policy gradient: $\bar{w}_t = \frac{1}{N} \sum_{n=0}^{N-1} w_{t,n}$.
    4. Update parameters $\theta_{t+1} = \theta_t + \eta \bar{w}_t$.
2. Return final policy $\pi_T = \pi_{\theta_T}$.
```

Previous finite-time bounds

Based on **previous bounds** for TD learning and NAC algorithms:

- Convergence of linear TD learning in MDP (Tsitsiklis & Van Roy, 1996).
- Finite-time analysis of linear TD learning in MDP (Bhandari et al., 2021).
- Finite-time analysis of log-linear NAC in MDP (Agarwal et al., 2021).

We **adapt existing bounds** for TD and NAC in POMDP (Cayci et al., 2024):

- It **does not assume** a stationary distribution nor full rank feature matrices.
- It **does assume** to sample i.i.d. from the discounted visitation measure.

These adaptations resulted in the following **contributions**:

- Fixing a few typos and errors in the original proofs.
- Adapting it to $z_t \sim f(\cdot\,|h_t)$ instead of $z_t \sim f(\cdot\,|h_{t-1}, a_{t-1})$.
- Generalizing these bounds to the asymmetric learning setting.

We define the belief $b(s|h) = \Pr(s|h)$ and approximate belief $\hat{b}(s|z) = \Pr(s|z)$.

Finite-time bound for the critics

Theorem 1: Finite-time bound for symmetric and asymmetric Q-functions.

For any $\pi \in \Pi_{\mathcal{M}}$, and any $m \in \mathbb{N}$, we have for Algorithm 1 with $\alpha = \frac{1}{K}$,

$$\sqrt{\mathbb{E}\left[\left\|\mathcal{Q}^\pi - \overline{\mathcal{Q}}^\pi\right\|_{d^\pi}^2\right]} \leq \varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}}$$

$$\sqrt{\mathbb{E}\left[\left\|Q^\pi - \overline{Q}^\pi\right\|_{d^\pi}^2\right]} \leq \varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}} + \varepsilon_{\text{alias}}.$$

$$\varepsilon_{\text{td}} = \sqrt{\frac{4B^2 + \left(\frac{1}{1-\gamma} + 2B\right)^2}{2\sqrt{K}(1-\gamma^m)}}$$

$$\varepsilon_{\text{app}} = \frac{1+\gamma^m}{1-\gamma^m} \min_{f \in \mathcal{F}_\varphi^B} \left\|f - Q^\pi\right\|_{d^\pi}$$

$$\varepsilon_{\text{shift}} = \left(B + \frac{1}{1-\gamma}\right)\sqrt{\frac{2\gamma^m}{1-\gamma^m}\sqrt{\left\|d_m^\pi \otimes \pi - d^\pi \otimes \pi\right\|_{\text{TV}}}}$$

$$\varepsilon_{\text{alias}} = \frac{2}{1-\gamma}\left\|\mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^{km}\left\|\hat{b}_{k,m} - b_{k,m}\right\|_{\text{TV}} \,\middle|\, Z_0 = \cdot, A_0 = \cdot\right]\right\|_{d^\pi}.$$

Sketch of the critic proof

We can easily show that for any $l \in \mathbb{R}$,

$$\sqrt{\mathbb{E}\left[\left\|Q-\overline{Q}\right\|_d^2\right]} \leq \left\|Q-\tilde{Q}\right\|_d^2 + \underbrace{\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(\underbrace{\mathbb{E}\left[\sqrt{\left\|\tilde{Q}-\hat{Q}_k\right\|_d^2}\right]}_{\Delta_k} - l\right)^2}}_{(*)} + l.$$

We bound $(*)$ for $l = \frac{1+\gamma^m}{1-\gamma^m}\left(\left\|\hat{Q}_* - Q\right\|_d + \left\|Q-\tilde{Q}\right\|_d\right)$ by bounding the drift,

$$\mathbb{E}\left[\left\|\beta_* - \beta_{k+1}\right\|_d^2 - \left\|\beta_* - \beta_k\right\|_d^2\right] \leq -(\ldots)(\Delta_k - l)^2 + (\ldots)l^2 + (\ldots).$$

By summing all Lyapounov drifts and rearranging,

$$\frac{1}{K}\sum_{k=0}^{K-1}(\Delta_k - l)^2 \leq -\frac{1}{K(\ldots)}\mathbb{E}\left[\left\|\beta_* - \beta_K\right\|_d^2 - \left\|\beta_* - \beta_0\right\|_d^2\right] + (\ldots)l^2 + (\ldots).$$

Substituting and setting $\alpha = \frac{1}{K}$, we obtain the bound.

Finite-time bound for the actors

Theorem 2: Finite-time bound for symmetric and asymmetric NAC.

For any $(\mathcal{Z}, U)$, we have for Algorithm 2 with $\alpha = \frac{1}{K}$, $\zeta = \frac{R\sqrt{1-\gamma}}{\sqrt{2N}}$, $\eta = \frac{1}{\sqrt{T}}$,

$$(1-\gamma) \min_{0 \leq t < T} \mathbb{E}[J(\pi^*) - J(\pi_t)] \leq \varepsilon_{\text{nac}} + \varepsilon_{\text{inf}} + \overline{C}_\infty \left(\varepsilon_{\text{actor}} + 2\varepsilon_{\text{grad}} + 2\sqrt{6}\frac{1}{T}\sum_{t=0}^{T-1} \varepsilon_{\text{critic}}^{\pi_t} \right),$$

$$\varepsilon_{\text{nac}} = \frac{R^2 + 2\log|A|}{2\sqrt{T}}$$

$$\varepsilon_{\text{actor}} = \sqrt{\frac{(2-\gamma)R}{(1-\gamma)\sqrt{N}}}$$

$$\varepsilon_{\text{inf,asym}} = 0 \qquad \varepsilon_{\text{inf,sym}} = 2\mathbb{E}^{\pi^*}\left[\sum_{k=0}^{\infty} \gamma^k \left\| \hat{b}_k - b_k \right\|_{\text{TV}}\right]$$

$$\varepsilon_{\text{grad,asym}} = \sup_{0 \leq t < T} \sqrt{\min_w \mathcal{L}_t(w)} \qquad \varepsilon_{\text{grad,sym}} = \sup_{0 \leq t < T} \sqrt{\min_w L_t(w)}$$

The term $\varepsilon_{\text{critic}}^{\pi_t}$ is given by Theorem 1, and $\sup_{0 \leq t < T} \mathbb{E}\left[\frac{d^{\pi^*}(S,Z,A)}{d^{\pi_t}(S,Z,A)}\right] \leq \overline{C}_\infty$.

Sketch of the actor proof

Let us first give the following performance difference lemma (Cayci et al., 2024),

$$V^{\pi^*}(z) - V^{\pi}(z) \leq \frac{1}{1-\gamma}\mathbb{E}^{d^{\pi^*}}[A^{\pi}(Z,A) \mid Z_0 = z] + \frac{2}{1-\gamma}\varepsilon_{\text{inf}}^{\pi^*}(z).$$

We start from the Lyapounov function $\Lambda(\pi) = \sum_z d^{\pi^*}(z)\ \mathrm{KL}(\pi^*(\cdot\,|z) \parallel \pi(\cdot\,|z))$, for which we can show, given than log-linear policies are 1-smooth,

$$\Lambda(\pi_{t+1}) - \Lambda(\pi_t) \leq \frac{\eta^2}{2}R^2 - \eta\sum_{z,a} d^{\pi^*}(z,a)A^{\pi}(z,a) + \eta\sum_{z,a} d^{\pi^*}(z,a)\sqrt{L(\bar{w}_t, z, a)}$$

After a few manipulation on $L(\bar{w}_t, z, a)$ by bounding $\|\bar{w} - w_*\|_2$ using SGD results for convex functions, we have using the lemma,

$$\Lambda(\pi_{t+1}) - \Lambda(\pi_t) \leq \frac{\eta^2}{2}R^2 - \eta(1-\gamma)\mathbb{E}[J(\pi^*) - J(\pi_t)] + 2\eta\varepsilon_{\text{inf}}^{\pi^*}(P)$$
$$+\eta\overline{C}_\infty\left(\sqrt{2}\varepsilon_{\text{actor}} + 2\varepsilon_{\text{grad}}^{\pi_t} + 2\sqrt{6}\varepsilon_{\text{critic}}^{\pi_t}\right)$$

By summing all Lyapounov drifts, rearranging, noting that $\Lambda(\pi_0) \leq \log|\mathcal{A}|$, and setting $\eta = \frac{1}{\sqrt{T}}$, we obtain the bound.

Some insights

When using an asymmetric actor-critic algorithm:

- The **critic error** has a **smaller upper bound**.
  - Because the asymmetric critic is the solution of a Bellman equation.
- The **actor suboptimality** has a **smaller upper bound**.
  - This benefit comes from the smaller upper bound on the critic error.

Some limitations:

- The analysis assumes a **fixed agent state process**.
  - Shed light on the effect of an aliased agent state (e.g., RNN at initialization).
  - It can easily be extended to learnable agent state processes: $\mathcal{A}^+ = \mathcal{A} \times \mathcal{Z}$.
- Requires **samples from the discounted visitation measure**.
  - But it is still feasible without assumption on the mixing time.
  - Reveal an interesting term $\varepsilon_{\text{shift}}$ when not assuming stationary distribution.

Conclusion

Take-home message

Don’t make the problem harder than it is.

Consider all available information at training.

Bonus: we start to see some theoretical justifications in the literature.

References

Agarwal, A., Kakade, S. M., Lee, J. D., & Mahajan, G. (2021). On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. *Journal of Machine Learning Research*.

Avalos, R., Delgrange, F., Nowe, A., Perez, G., & Roijers, D. M. (2024). The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models. *Proceedings of the 12th International Conference on Learning Representations*.

Baisero, A., & Amato, C. (2022). Unbiased Asymmetric Reinforcement Learning under Partial Observability. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*.

Bhandari, J., Russo, D., & Singal, R. (2021). A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation. *Operations Research*.

References (ii)

Cai, Y., Liu, X., Oikonomou, A., & Zhang, K. (2024). Provable Partially Observable Reinforcement Learning with Privileged Information. *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*.

Cayci, S., He, N., & Srikant, R. (2024). Finite-Time Analysis of Natural Actor-Critic for POMDPs. *SIAM Journal on Mathematics of Data Science*.

Choudhury, S., Bhardwaj, M., Arora, S., Kapoor, A., Ranade, G., Scherer, S., & Dey, D. (2018). Data-Driven Planning via Imitation Learning. *The International Journal of Robotics Research*.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., Las Casas, D. de, & others. (2022). Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning. *Nature*.

References (iii)

Hu, E. S., Springer, J., Rybkin, O., & Jayaraman, D. (2024). Privileged Sensing Scaffolds Reinforcement Learning. *Proceedings of the 12th International Conference on Learning Representations.*

Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Champion-Level Drone Racing using Deep Reinforcement Learning. *Nature.*

Lambrechts, G., Bolland, A., & Ernst, D. (2024). Informed POMDP: Leveraging Additional Information in Model-Based RL. *Reinforcement Learning Journal.*

Littman, M. L., Cassandra, A. R., & Kaelbling, L. P. (1995). *Learning Policies for Partially Observable Environments: Scaling Up.*

Nguyen, H., Daley, B., Song, X., Amato, C., & Platt, R. (2021). Belief-Grounded Networks for Accelerated Robot Learning under Partial Observability. *Conference on Robot Learning.*

References (iv)

Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., & Abbeel, P. (2018). Asymmetric Actor Critic for Image-Based Robot Learning. *14th Robotics: Science and Systems.*

Sinha, A., & Mahajan, A. (2023). Asymmetric Actor-Critic with Approximate Information State. *62nd IEEE Conference on Decision and Control.*

Tsitsiklis, J., & Van Roy, B. (1996). Analysis of Temporal-Difference Learning with Function Approximation. *Advances in Neural Information Processing Systems.*

Vasco, M., Seno, T., Kawamoto, K., Subramanian, K., Wurman, P. R., & Stone, P. (2024). A Super-Human Vision-Based Reinforcement Learning Agent for Autonomous Racing in Gran Turismo. *Reinforcement Learning Journal.*

Wang, A., Li, A. C., Klassen, T. Q., Icarte, R. T., & McIlraith, S. A. (2023). Learning Belief Representations for Partially Observable Deep RL. *International Conference on Machine Learning.*

References (v)

Warrington, A., Lavington, J. W., Scibior, A., Schmidt, M., & Wood, F. (2021). Robust Asymmetric Learning in POMDPs. *Proceedings of the 38th International Conference on Machine Learning*.

A warm thank to Aditya Mahajan for his precious supervision throughout this research project.