



POLITECNICO
MILANO 1863

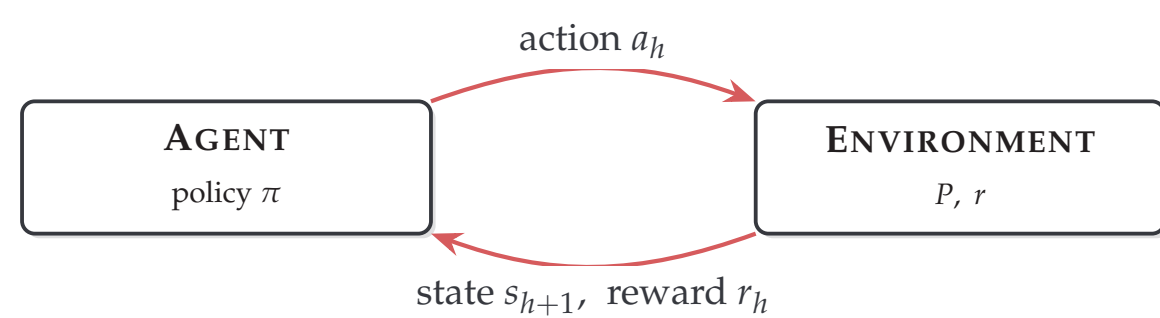
TRADING-OFF STATISTICAL AND COMPUTATIONAL EFFICIENCY VIA W -STEP MARKOV DECISION PROCESSES: A POLICY GRADIENT APPROACH

G. TEDESCHI, M. MUSSI, A. M. METELLI, M. RESTELLI

{gianmarco.tedeschi, marco.mussi, albertomaria.metelli, marcello.restelli}@polimi.it



SETTING: REINFORCEMENT LEARNING



- MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \gamma, H, \mu_0)$
- Policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, a distribution over actions
- Goal: maximise the discounted return:
 $J(\pi) = \mathbb{E}_\pi[\sum_h \gamma^h r(s_h, a_h)]$

MOTIVATION: THE CURSE OF HORIZON

Both **statistical** and **computational** cost grow with the horizon H (Liu et al., 2018). Two classical remedies buy speed, each in a different currency:

Action persistence (Metelli et al., 2020)

Repeat *one* action for several steps — **rigid**: a single action per window.

Open-loop control (Bubeck and Munos, 2010)

Plan a whole sequence, ignore feedback — **suboptimal** under stochastic dynamics.

Our idea. Act in **windows of W steps**: the whole action sequence is drawn *from the state*, so control is open-loop *within* a window and closed-loop *across* windows. A single knob W sweeps the range.

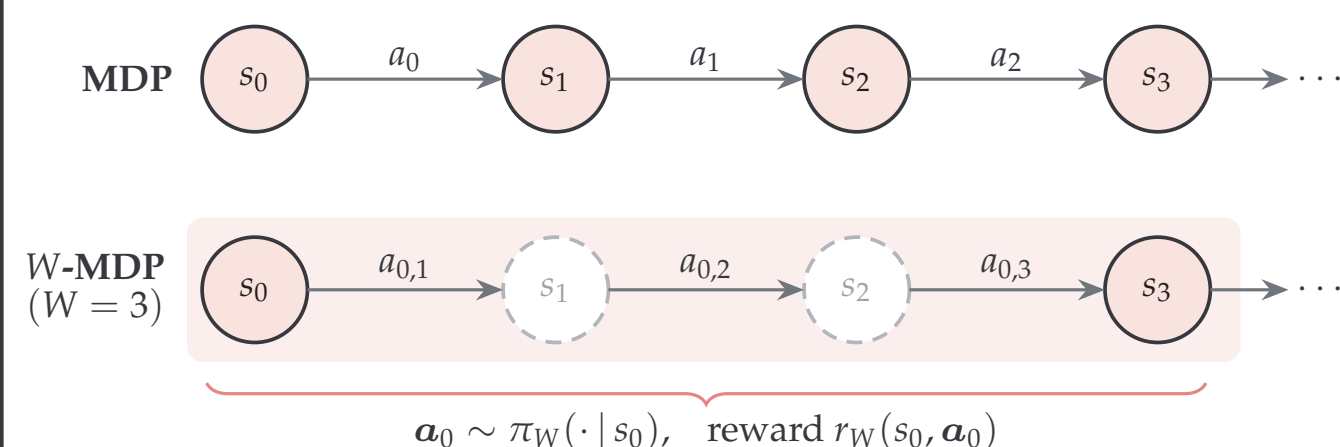
THE W -STEP MDP

DEFINITION 1 W -step MDP

Given an MDP $\mathcal{M}$ and a window $W \in \mathbb{N}_{\geq 1}$,

$$\mathcal{M}_W := (\mathcal{S}, \mathcal{A}^W, P_W, r_W, \bar{H}, \bar{\gamma}, \mu_0),$$

with $\mathcal{A}^W = \{\mathbf{a} = (a_1, \dots, a_W) : a_i \in \mathcal{A}\}$, horizon $\bar{H} = \lceil H/W \rceil$ and discount $\bar{\gamma} = \gamma^W$.



○ observed by the learner

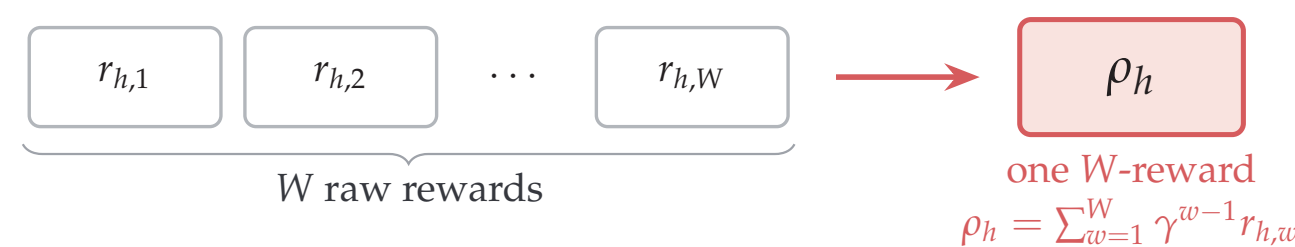
○ traversed, never observed

METASTEP: INTERACTION PROTOCOL

METASTEP leaves the learner untouched and changes only the **agent–environment interface**.

Input: horizon H , window W , policy π_W
for $h = 0, \dots, \lceil H/W \rceil - 1$ **do**
 Observe state s_h
 Sample $\mathbf{a}_h = (a_{h,1}, \dots, a_{h,W}) \sim \pi_W(\cdot | s_h)$
for $w = 1, \dots, W$ **do**
 | Play $a_{h,w}$, collect reward $r_{h,w}$
end
end

The W raw rewards collapse into a single **W -reward**, which the learner treats as one ordinary reward signal:



HOW MUCH DO WE LOSE?

THEOREM 1 Performance bias

For every $s \in \mathcal{S}$:

$$V^*(s) - V_W^*(s) \leq 4R_{\max} D(P) \frac{\gamma(1 - \gamma^{W-1})}{(1 - \gamma)^2(1 - \gamma^W)},$$

$$D(P) := \max_{(s,a)} \min_{s'} \{1 - P(s' | s, a)\}.$$

- $W = 1 \Rightarrow$ bias = 0: the W -MDPs reduces to MDPs
- The gap grows **linearly** in $D(P)$
- P deterministic $\Rightarrow D(P) = 0 \Rightarrow$ **no bias at any W**



The price of the window is set by the **environment**, not by the algorithm.

METASTEP ON GPOMDP

The GPOMDP estimator (Baxter and Bartlett, 2001) instantiated on the W -MDP:

$$\hat{\nabla}_\theta J(\theta; W) = \frac{1}{N} \sum_{l=1}^N \sum_{h=0}^{\lceil H/W \rceil - 1} \left(\sum_{u=0}^h \nabla_\theta \log \pi_{W,\theta}(a_{l,u} | s_{l,u}) \right) \gamma^{Wh} \rho_{l,h}$$

GPOMDP $\xrightarrow{H \text{ terms per trajectory}}$
 METASTEP $\xrightarrow{\lceil H/W \rceil \text{ terms } (W=3) \text{ — policy queried } \approx W \times \text{less often}}$

LEMMA 1 Variance

With $\Phi_H(W) := 1 - \frac{H}{W}(\gamma^H - \gamma^{H+W}) - \gamma^H$, for every W :

$$\text{Var}[\hat{\nabla}_\theta J(\theta; W)] \leq \frac{V(W)}{N}, \quad V(W) = \frac{R_{\max}^2 G^2 (1 - \gamma^H)}{(1 - \gamma)^2 (1 - \gamma^W)} \Phi_H(W).$$

$V(W)$ is **non-increasing in W** : **cheaper, less noisy gradients**.

SAMPLE COMPLEXITY AND TRADE-OFFS

THEOREM 2 Global Convergence

A suitable step size guarantees:

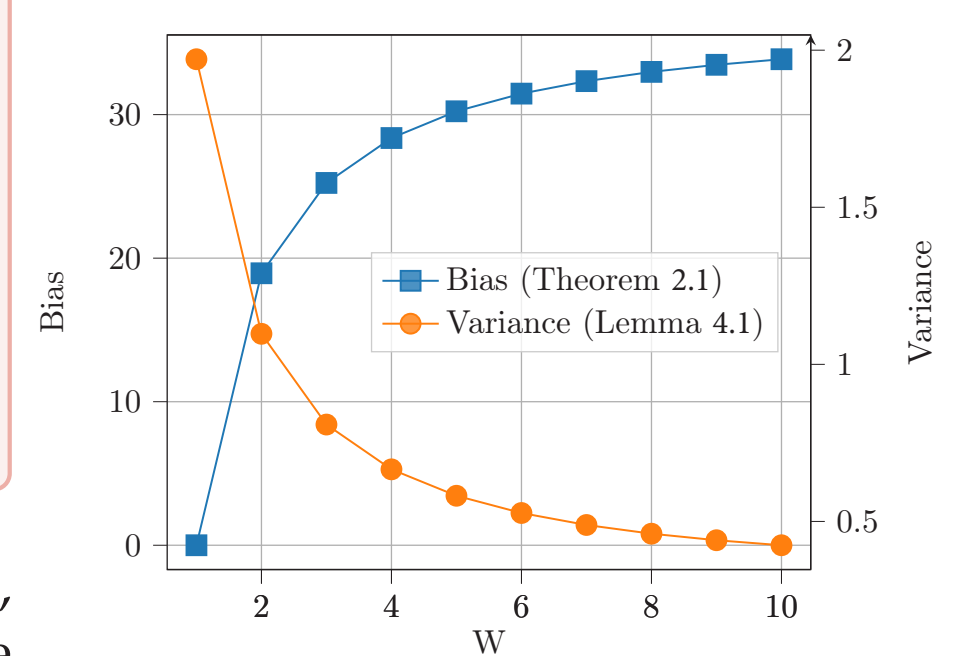
$$\mathbb{E}[J^*(W) - J(\theta_K; W)] \leq \epsilon + \beta$$

with:

$$NK \leq \tilde{O}\left(\frac{R_{\max}^3 (1 - \gamma^H) \Phi_H(W)^2}{\epsilon^3 (1 - \gamma)^3 (1 - \gamma^W)^2}\right).$$

The rate in ϵ is standard (Montenegro et al., 2024); since $\gamma^W \leq \gamma$, the gain sits in the **horizon-dependent constants**.

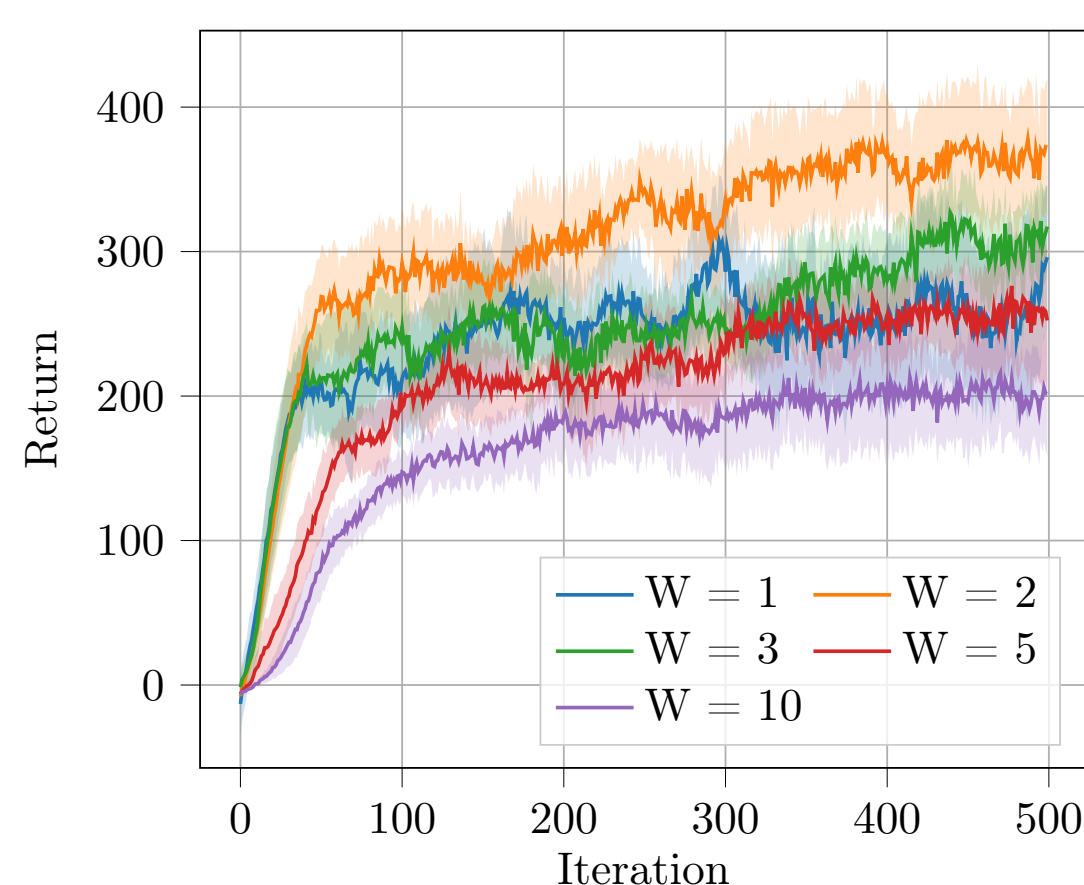
Variance falls with W (Lem. 1) while bias grows (Thm. 1): **intermediate W s** are optimal.



Bias (Thm. 1) grows, variance (Lem. 1) falls.

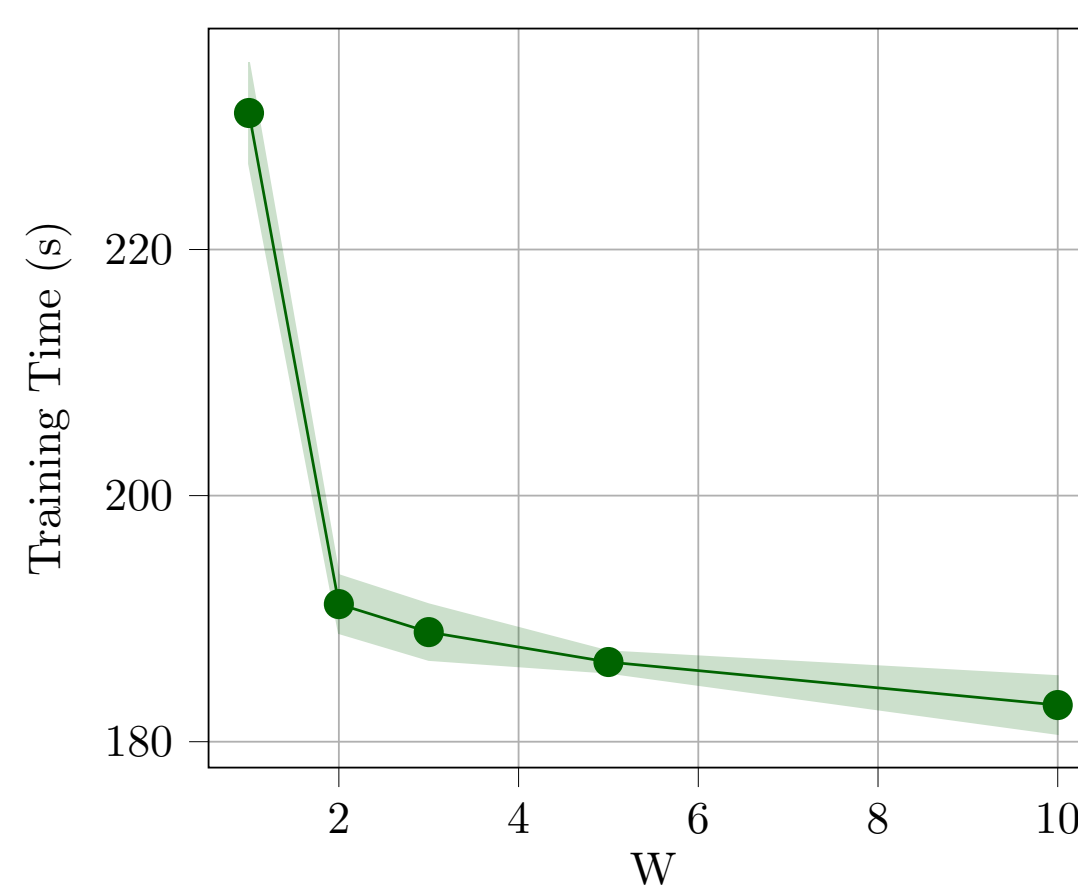
EXPERIMENTAL VALIDATION: HALF-CHEETAH

Half-Cheetah | $d_S = 17, d_A = 6$ | $H = 200, \gamma = 0.999$ | $N = 10, K = 500$ | $W \in \{1, 2, 3, 5, 10\}$, $W = 1$ vanilla GPOMDP | 10 runs, mean $\pm$ 95% C.I.



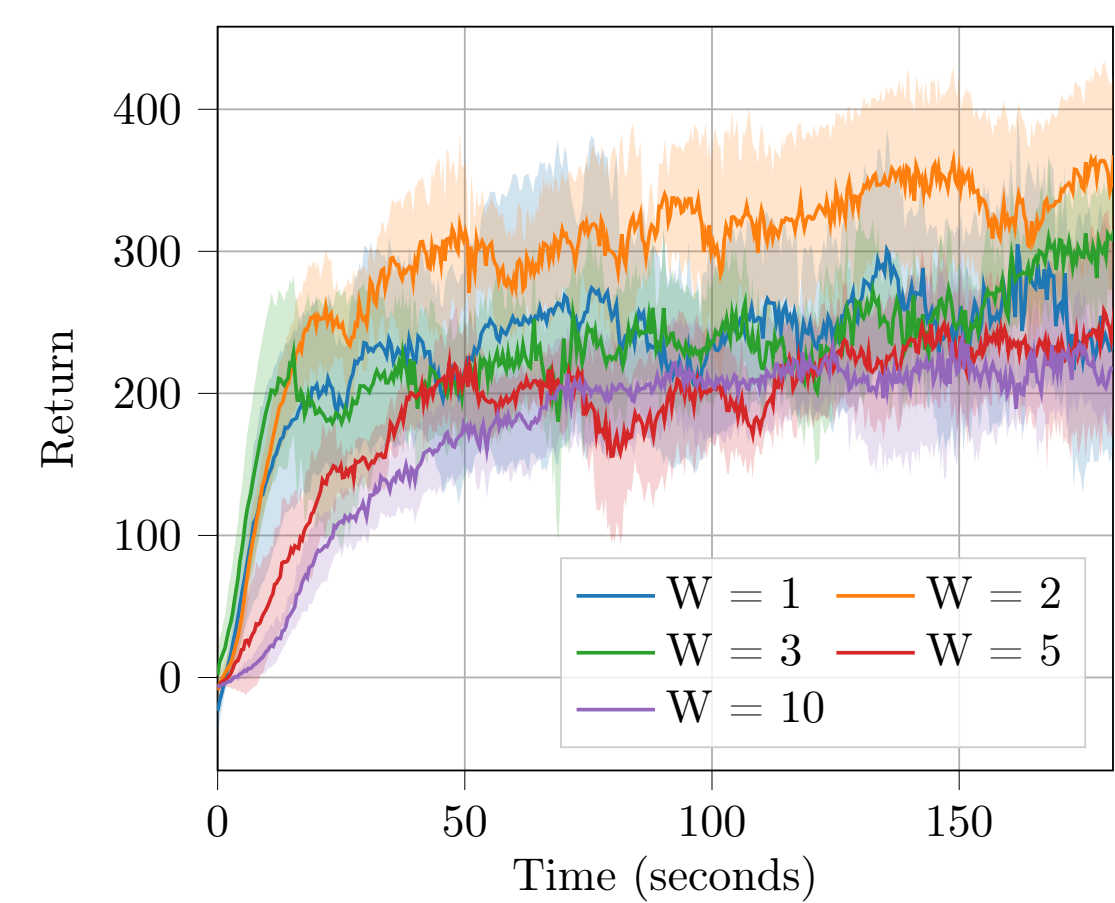
(a) Return over iterations.

An **intermediate W** wins: $W = 2, 3$ beat $W = 1$.



(b) Training time vs. W .

Wall-clock cost **falls monotonically** with W .



(c) Return over wall-clock time.

The ranking holds **per unit of compute**.

REFERENCES

- J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 2001.
- S. Bubeck and R. Munos. Open loop optimistic planning. In *Annual Conference on Learning Theory (COLT)*, 2010.
- Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- A. M. Metelli, F. Mazzolini, L. Bisi, L. Sabbioni, and M. Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020.
- A. Montenegro, M. Mussi, A. M. Metelli, and M. Papini. Learning optimal deterministic policies with stochastic policy gradients. In *International Conference on Machine Learning (ICML)*, 2024.