A Study of the Weighted Multi-step Loss Impact on the Predictive Error and the Return in MBRL

Abdelhakim Benechehab[12]

[1]Huawei Paris Noah's Ark Lab
[2]Department of Data Science, EURECOM
[3]Statistics Program, KAUST

abdelhakim.benechehab1@huawei.com

Albert Thomas[1], Giuseppe Paolo[1], Maurizio Filippone[3], Balázs Kégl[1]

August 8, 2024

Problem Setup

Goal: given a dataset of real system trajectories, learn a parametric model of its transition function.

- Input $[s_t, a_t] \in \mathbb{R}^{d_s+d_a}$, target $s_{t+1} \in \mathbb{R}^{d_s}$,
- Training set of N trajectories $\mathcal{D} = \{(s_0^i, a_0^i, s_1^i, \ldots)\}_{i=1}^N$,
- Train a model $\hat{p}_{\boldsymbol{\theta}} : \mathbb{R}^{d_s+d_a} \to \mathbb{R}^{d_s}$ that minimizes the MSE loss (or NLL).

Single-step error

99% R2

Multi-step error

Please help, errors are compounding!

Solution: Weighted Multi-Step Loss

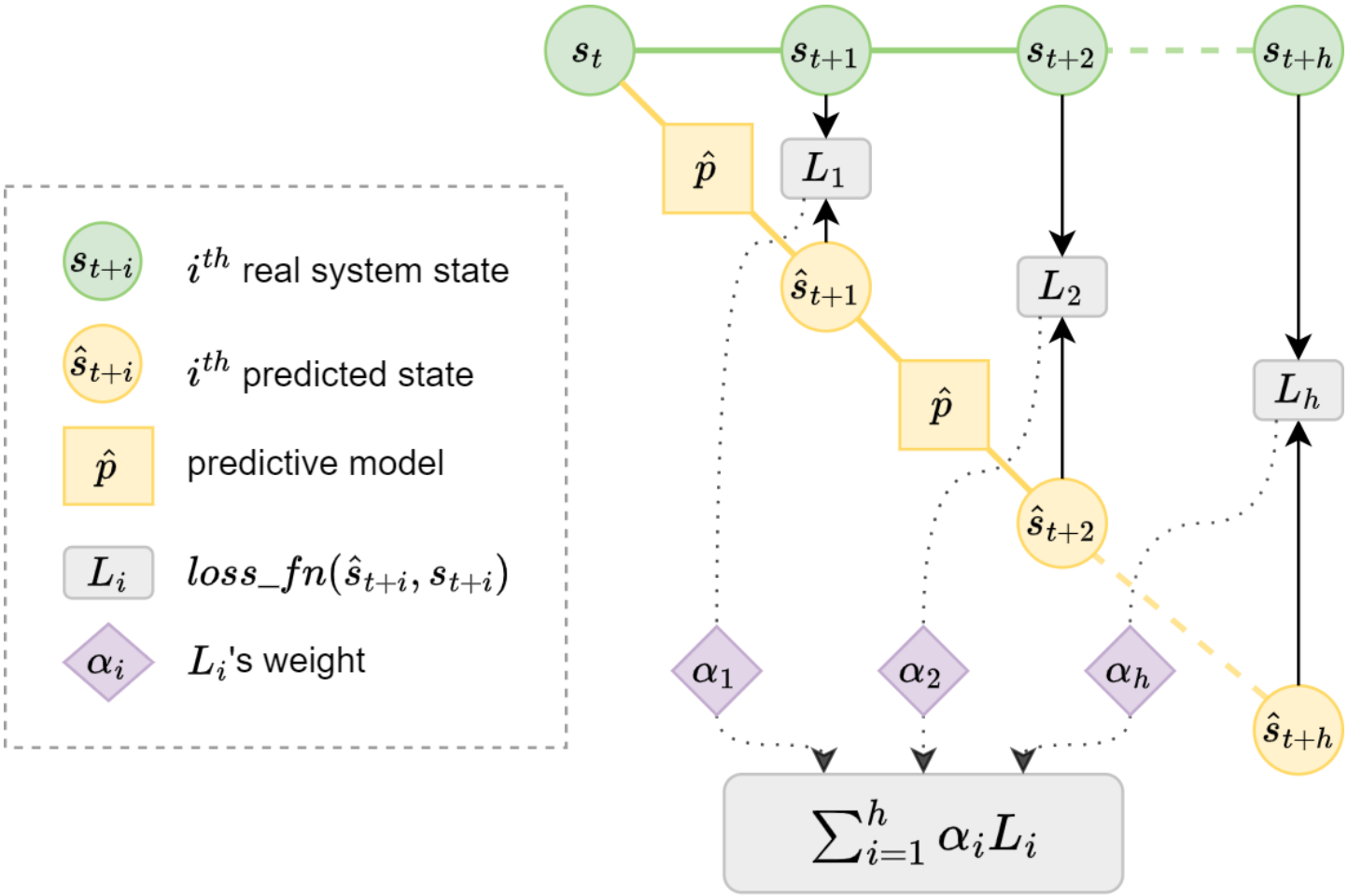
- horizon-dependent **weights** $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_h)$ with $\sum_{i=1}^{h} \alpha_i = 1$,
- a single-step **loss function** L (MSE),
- an initial state s_t, an action sequence $\mathbf{a}_\tau = \mathbf{a}_{t:t+h-1}$, and the real (ground truth) visited states $\mathbf{s}_\tau = \mathbf{s}_{t+1:t+h}$,

Definition

Given the elements above, we define the weighted multi-step loss of horizon h as:

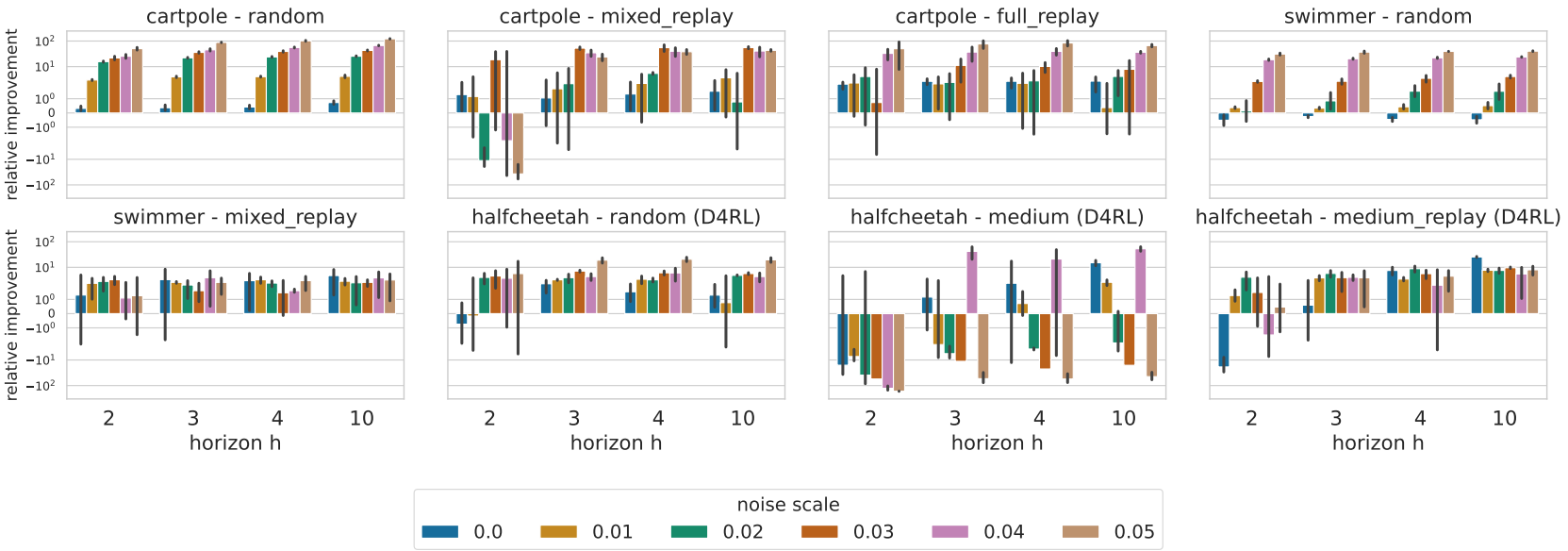
$$L^h_{\boldsymbol{\alpha}}\big(\mathbf{s}_\tau, \hat{p}_\theta(s_t, \mathbf{a}_\tau)\big) = \sum_{j=1}^{h} \alpha_j L\big(s_{t+j}, \hat{p}^j_\theta(s_t, \mathbf{a}_{t:t+j-1})\big)$$

Schematic representation



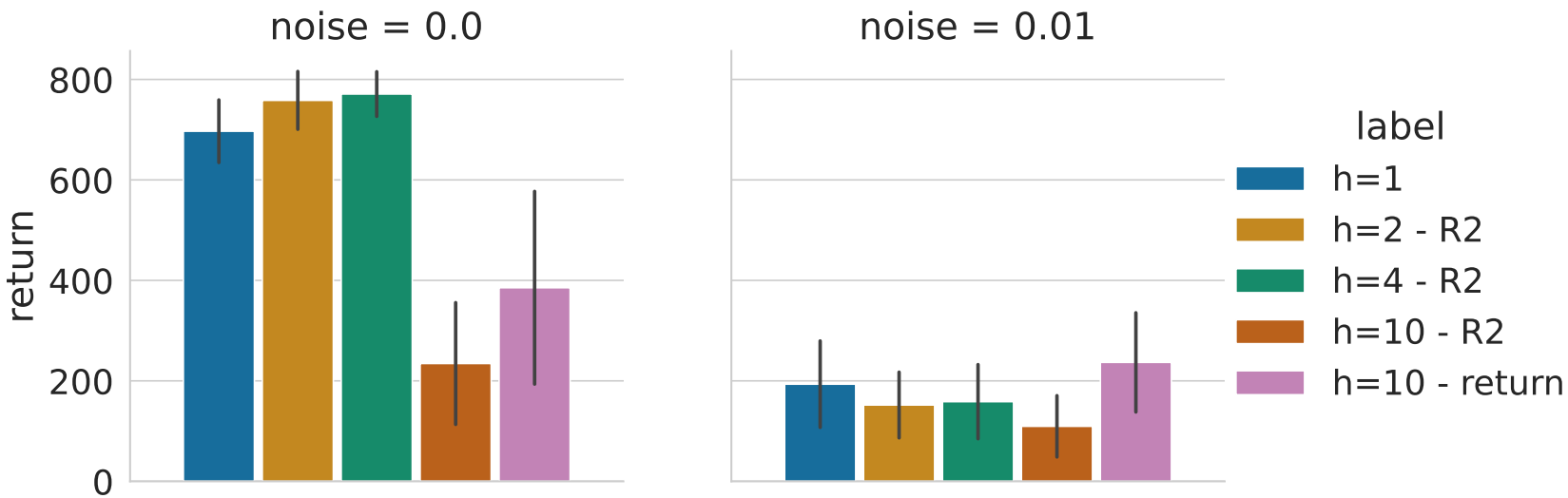
Results: Predictive error on offline datasets

- **setup.** noisy observations $o_t = s_t + \epsilon_t$ with $\epsilon_t \sim \mathbf{N}(0, \sigma^2)$
- **Metric.** aggregated R2 score $\overline{R2}(H) = \frac{1}{H} \sum_{h=1}^{H} \text{R2}(h)$
- **Benchmark.** Environments (Cartpole swingup, Halfcheetah, Swimmer), Datasets (random, medium, replay)



Results: Offline MBRL

- **agent.** *Dyna*-style using Soft Actor-Critic (SAC) a la MBPO
- $\mathbf{h = 1}$. the baseline
- $\mathbf{h}$ = **h - R2.** we select the optimal β value in grid search based on the $R2$ metric
- $\mathbf{h}$ = **h - return.** we select the optimal β value in grid search based on the return of the agent
- **task.** Cartpole swing-up mixed replay dataset, with two levels of noise 0% and 1%



Conclusion

- In MBRL, Models face compounding errors and a distribution mismatch at test time
- The **Weighted Multi-Step loss** is a way to solve this problem
- Although it improves the predictive error, it doesn't necessarily lead to better policies

Take Home Message

The **Weighted Multi-step loss** is useful to improve the **predictive error** down the horizon.
$\rightarrow$ But is this a good metric for model selection in MBRL ?!

[Benechehab et al., 2024]

References

Benechehab, A., Thomas, A., Paolo, G., Filippone, M., and Kégl, B. (2024).
A study of the weighted multi-step loss impact on the predictive error and the return in MBRL.
In *I Can't Believe It's Not Better Workshop: Failure Modes of Sequential Decision-Making in Practice (RLC 2024)*.