Towards Understanding the Impact of Plasticity Loss on Reinforcement Learning in Stochastic Environments

Philipp Bordne
University of Freiburg
bordnep@cs.uni-freiburg.de

André Biedenkapp*
Karlsruhe Institute of Technology
biedenkapp@kit.edu

Abstract

Deep reinforcement learning (RL) is a powerful framework for sequential decision-making. Despite many high-profile successes in recent years, deep RL is known to be brittle. An emerging research direction into the causes of this brittleness concerns the loss of plasticity of neural networks. However, plasticity loss has so far been understudied in stochastic environments. We study this interplay systematically, training SAC agents on three DeepMind Control tasks augmented with controlled levels of observation and action noise. We find that plasticity-preserving methods increase learning robustness on the two harder tasks, with their advantage growing as observation noise increases, while combining resets with layer normalization underperforms resets alone at every observation noise level tested. Further, stochasticity can accelerate the loss of plasticity in harder environments, whereas in easier settings noise can act as a regularizer that sustains plasticity. Lastly, plasticity loss is most pronounced in the early stages of training, and in the most challenging stochastic settings TD instability–rather than gradual plasticity loss–becomes the dominant failure mode.

1 Introduction

Over the last decade, deep Reinforcement Learning (RL) [Sutton and Barto, 2018] has achieved impressive results from game playing [Mnih et al., 2015, Grooten et al., 2026] over energy markets [Harder et al., 2023] to control of complex real world systems [Bellemare et al., 2020, Degrave et al., 2022, Kaufmann et al., 2023]. Predominantly however, RL research still is focused on simulated environments with deterministic transition dynamics [Todorov et al., 2012, Brockman et al., 2016, Tassa et al., 2018]. However, stochasticity is very much inherent in real-world applications, where sensors are noisy, actuators imprecise, and dynamics uncertain. A natural question is therefore how well modern deep RL algorithms learn under such conditions, and what the root cause is for the observed brittleness of RL [Henderson et al., 2018] in the presence of noise.

Multiple lines of research have emerged to understand and mitigate the brittleness of RL. On the side of mitigation, many works address this by automating design choices present in RL pipelines [Parker-Holder et al., 2022]. Some work in this direction is more concerned with augmenting environments, e.g., via curriculum learning [OpenAI et al., 2019, Klink et al., 2020] to expose agents to more diverse experiences, while others are trying to address the brittleness via hyperparameter tuning [Eimer et al., 2023]. While such works are able to produce more stable and performative RL agents, they provide little insight into the ultimate causes.

On the side of understanding the causes, attention has shifted towards the *loss of plasticity* of neural networks in deep RL pipelines: agents are known to overfit to transitions collected early in training,

*Work done while at the University of Freiburg.

19th European Workshop on Reinforcement Learning (EWRL 2026).

when the policy is still far from optimal [Nikishin et al., 2022, Igl et al., 2021], gradually losing the ability to update their representations in response to new experience [Lyle et al., 2023]. Interventions that counteract this, such as periodic resets and normalization, measurably improve learning on standard benchmarks [Nikishin et al., 2022, Nauman et al., 2024]. Closely related work has attributed the primacy bias to temporal difference (TD) instability, i.e. the early divergence of Q-values [Hussing et al., 2024].

The interplay between plasticity loss and environment stochasticity remains largely unexamined, although two considerations suggest it warrants study. The first is observational: in a benchmarking study of robust RL algorithms, Glossop et al. [2022] find that the presence of noise *during training* is considerably more detrimental than the same noise at deployment. The second is mechanistic: RL algorithms optimize an *expected* return, and the samples used to estimate it are collected throughout training. Zero-mean noise should therefore average out, provided realizations are weighted equally regardless of when they were collected. Plasticity loss might break this, inducing a bias towards early samples. Accordingly, our object of study is *learning* robustness, the ability of an algorithm to learn a task at all from noisy samples, rather than robustness in the more common sense of a deployed controller that remains stable or safe under disturbances [Brunke et al., 2022].

This motivates our central question: *can preserving plasticity improve learning robustness?* We investigate this with a systematic study of Soft Actor-Critic (SAC) [Haarnoja et al., 2018] on three proprioceptive control tasks from the DeepMind Control suite (DMC suite) [Tassa et al., 2018], augmented with controlled levels of observation and action noise. We do not verify the mechanism directly, but our setting yields a testable prediction: if plasticity loss is more damaging under noise, the advantage of plasticity-preserving methods should grow with the noise level. We probe plasticity throughout training to verify that the tested methods indeed preserve it and to establish plasticity as a potential factor in learning robustness under noise. We additionally track maximum Q-values to separate the effects of plasticity loss and TD instability [Hussing et al., 2024].

Our key findings are:

- Plasticity-preserving methods — in particular periodic resets [Nikishin et al., 2022] — improve learning robustness under observation noise on the two harder tasks, with their advantage growing at higher noise levels.
- The relationship between noise and plasticity loss is bidirectional as noise accelerates plasticity loss, and low plasticity in turn compounds learning failure. On easier tasks, however, noise can act as a regularizer that actually sustains plasticity.
- Plasticity loss is most pronounced in the early stages of training and is exacerbated by noise, but in the most challenging settings the dominant failure mode is TD instability rather than gradual plasticity loss and stochastic environments make this catastrophic failure mode more likely to occur.

2 Related Work

Controlled Noise In Simulated Learning Tasks Previous work has augmented simulators with observation and action noise to reflect stochastic dynamics in real world applications, with the objective of obtaining policies that bridge the sim-to-real gap [Peng et al., 2018] or to study the robustness of learned policies under disturbances [Yuan et al., 2022]. Farebrother et al. [2024] compare categorical and MSE regression on the Arcade-Learning-Environment [Machado et al., 2018] with (default) and without action-noise. They find that categorical regressions' improved performance over MSE is contingent upon the presence of action noise and attribute it to being less prone to overfitting on noisy targets. Glossop et al. [2022] study how different noise types (dynamics, observation, action) affect the score achievable by PPO [Schulman et al., 2017] and SAC [Haarnoja et al., 2018] on the cart-pole environment and find noise on dynamics and observations to be much more detrimental to performance. Interestingly, they find that the presence of observation noise during training is already detrimental to performance under any noise level at test time and lower than performance if trained without noise. Our work starts exactly here, aiming to study how noise affects the learning dynamics.

RL that Matters To mitigate the known and well documented brittleness of RL [Henderson et al., 2018, Engstrom et al., 2020, Andrychowicz et al., 2021], various research directions are actively

being explored. Under the umbrella-term AutoRL, Parker-Holder et al. [2022] identify multiple dimensions that are (automatically) adapted to improve robustness of learning pipelines. For example, various methods and techniques are proposed to automatically adapt the hyperparameters of RL pipelines, to increase the final performance [see, e.g., Jaderberg et al., 2017, Franke et al., 2021, Parker-Holder et al., 2020], as well as improve robustness [Falkner et al., 2018, Eimer et al., 2023]. Further, environment design approaches propose to augment different elements of the environment, including reward function augmentation [Marthi, 2007, Faust et al., 2019, Hu et al., 2020], action spaces [Mankowitz et al., 2018, Farahani and Mozayani, 2019, Bagaria et al., 2021] as well as observation space augmentation [Raileanu et al., 2021, Yarats et al., 2021]. The latter is of particular interest to our study as such work considers automatic selection of transformations (such as crop, rotate and flip) to image observations, such that policies are invariant to these transformations while maximizing rewards. In contrast to these approaches, which primarily prioritize maximizing final rewards, our study shifts the focus toward deepening the understanding of reinforcement learning in noisy environments and exploring how plasticity-preserving methods can enhance robustness under such conditions. Thus, we focus on settings in which we draw random samples, rather than randomly perturbing samples.

3 Background

3.1 MDPs, RL and the Effect of different Sources of Noise

The goal of RL is to find the optimal policy in a Markov Decision Process (MDP). An MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ defines the transition probability distribution, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function and $\gamma \in [0, 1]$ is the discount factor. An optimal policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ then needs to maximize the expected discounted return $\mathbb{E}_{\tau(\pi)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$.

Perturbing the observations of the agent at time t as $\tilde{\mathrm{o}}_t \sim \nu(\cdot|s_t)$ breaks the Markov property and gives rise to a Partially Observable Markov Decision Process (POMDP) $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \nu, \mathcal{O})$ where $\mathcal{O}$ is the observation space. Hence, solving the POMDP generally involves a state estimation problem to maintain a belief over the true state [Kaelbling et al., 1998].

In contrast, perturbations on the actions as $\tilde{\mathrm{a}}_t \sim \rho(\cdot|a_t)$ yield a transformed MDP with a transition probability function which "absorbs" the action noise:

$$P'(s_{t+1}|s_t, a_t) = \int_{\mathcal{A}} P(s_{t+1}|s_t, \tilde{a})\rho(\tilde{a}|a_t)d\tilde{a} \tag{1}$$

3.2 Quantifying Plasticity Loss

The loss of plasticity of a neural network describes its decreasing ability to fit a changing data distribution or to "respond to different *possible* learning signals" [Lyle et al., 2023]. Lyle et al. [2023] propose to quantify it as the ability of a (trained) neural network parameterized by θ_t to fit a high frequency transformation of the output of a randomly initialized model from the same class: $y(x) = b + \sin\left(M \cdot f_{\theta_0}(x)\right)$, where $M = 10^5$ and b is a bias to compensate for the offset of the mean prediction of f_{θ_t}. This definition focuses on the ability to reduce the *training* loss, whereas other work has identified similar phenomena in terms of *generalization* capabilities of neural networks [Ash and Adams, 2020, Igl et al., 2021].

A broad range of work explains plasticity loss by a reduced representational and signal propagation capacity when compared to the time of initialization. Prominent measures to detect such capacity loss are dormant neurons [Sokar et al., 2023], a low effective rank of feature representations [Kumar et al., 2021] or linearization of activation functions [Lyle et al., 2024]. Further, optimization pathologies such as loss landscape sharpness and a degenerate neural-tangent kernel [Lyle et al., 2024] have been connected to the cause of plasticity loss.

Previous work suggests that preserving the plasticity of the critic is more important in SAC [Ma et al., 2024, Nauman et al., 2024]. This likely is due to the fact that the phenomena treated under the term plasticity loss have also been studied more specifically in the context of TD learning. Kumar et al. [2021] studied and explained the decrease of the effective rank in deep RL as a direct result of TD

learning. And Hussing et al. [2024] attributed the primacy-bias not to overfitting on early data but to early value divergence due to limited action-space coverage of samples at the beginning of training.

We will refer to Q-value divergence as TD instability as one specific mechanism that may cause severe plasticity loss, in accordance with the findings by Lyle et al. [2024] that regression on high offset targets results in particularly severe loss of plasticity.

4 Maintaining Comparable Plasticity Probing Throughout Training

To assess the ability of a neural network implementing a parametric function $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$ to learn from non-stationary data which was gathered in a stochastic environment, we aim to directly quantify the plasticity of its parameters θ. To this end we adopt the definition of plasticity as proposed by Lyle et al. [2023]:

Definition 1 (Optimization procedure)**.** *Given dataset $\boldsymbol{D}$ and a loss function $L(\cdot \mid \boldsymbol{D})$, we define an optimization (or update) procedure $\mathcal{U} : \Theta \times \mathbb{N}_+ \to \Theta$ as a function that returns updated parameters $\theta' = \mathcal{U}(\theta, T)$, by performing $T \in \mathbb{N}_+$ steps minimizing $L(\cdot \mid \boldsymbol{D})$ starting from $\theta \in \Theta$.*

Together with a neural network architecture, θ implements the function f_θ. The optimization procedure $\mathcal{U}$ may have further updatable parameters, such as dynamic learning rates or gradient moment statistics. For the sake of a clear presentation, we subsume these parameters into θ and note that they have no immediate effect on the computation of $L(\theta \mid \boldsymbol{D})$[1]. Likewise the optimization procedure is implicitly specified by the choice of the loss function L and the dataset $\boldsymbol{D}$ and we omit the explicit dependence for the sake of brevity.

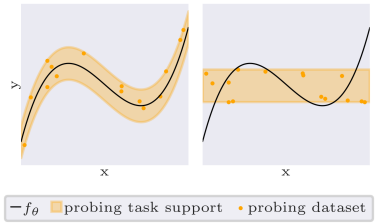

Figure 1: Illustration of the probing task. As implemented in this study (left); as originally proposed by Lyle et al. [2023] (right).

Definition 2 (Neural network plasticity)**.** *Given an optimization procedure $\mathcal{U}$ as in definition 1, a neural network parameterized by θ, a probing dataset $\boldsymbol{D}$ and a number of probing steps T we define plasticity as:*

$$\mathcal{P}(\theta|\boldsymbol{D}) = \frac{L(\theta|\boldsymbol{D}) - L\left(\mathcal{U}(\theta, T)|\boldsymbol{D}\right)}{L(\theta|\boldsymbol{D})} \tag{2}$$

That is, plasticity is the fraction of the initial loss $L\left(\theta|\boldsymbol{D}\right)$ that can be removed by the update procedure $\mathcal{U}$ on the task $L\left(\cdot|\boldsymbol{D}\right)$ within T probing steps.

Following Lyle et al. [2023], we define the probing task as an MSE-regression task, however with an important distinction: We generate the probing task by adding randomly-sampled and sine-transformed offsets to the current network's predictions. I.e., given a sample $\boldsymbol{x}$ we set the target as $\boldsymbol{y} = f_\theta(\boldsymbol{x}) + \sin\left(\epsilon\right)$, where $\epsilon \sim \Pi_{i=1}^{m} \text{Unif}\left([0, 2\pi]\right)$ is sampled from a uniform distribution and has the same dimensionality as the output of f_θ.

The central difference to Lyle et al. [2023] is that we offset $\boldsymbol{y}$ of the probing samples specific to the respective input $\boldsymbol{x}$ instead of using a global offset $\boldsymbol{b}$, which they set to the mean prediction of f_θ over $\boldsymbol{D}$. By considering the "shape" of f_θ we keep the initial probing loss $L\left(\theta|\boldsymbol{D}\right)$ approximately constant since all probing samples lie within an ϵ-ball around f_θ. We illustrate the difference between the original and our approach in Figure 1. We can think of this as probing the ability of the model to fit perturbations in arbitrary directions.

5 Experiments

5.1 Setup and Evaluation Protocol

Stochastic environments: We study additive observation and action noise for two reasons: (1) both can be implemented by wrapping existing DMC suite environments (humanoid-walk, quadruped-run

[1] However, these optimizer statistics can have a significant impact on adaptation to distribution shifts and hence plasticity loss [Lyle et al., 2023].

and walker-run) without modifying the physics simulation, making the approach straightforward to extend to other benchmarks [Rajan et al., 2023]; (2) they affect learning in fundamentally different ways: observation noise breaks the Markov property and gives rise to a POMDP, whereas action noise induces a transformed MDP that adds stochasticity to otherwise deterministic dynamics (Section 3.1). Following Rajan et al. [2023], the wrapper transforms the agent's action before passing it to the environment as $\tilde{\boldsymbol{a}}_t = \boldsymbol{a}_t + \epsilon_{\boldsymbol{a},t},\ \epsilon_{\boldsymbol{a},t} \sim \mathcal{N}(0, \sigma_{\boldsymbol{a}})$ and the next-state observation to $\tilde{\boldsymbol{o}}_{t+1} = \boldsymbol{o}_{t+1} + \epsilon_{\boldsymbol{o},t},\ \epsilon_{\boldsymbol{o},t} \sim \mathcal{N}(0, \sigma_{\boldsymbol{o}})$ before returning it to the agent.[2]

Learning algorithm: We train SAC agents [Haarnoja et al., 2018, 2019] for one million environment steps with default hyperparameters (see Appendix A). We vary the replay ratio (gradient steps per environment step) over $\{1, 4, 8\}$ to investigate whether higher update frequencies exacerbate overfitting under noise [Nikishin et al., 2022, Igl et al., 2021]. When adding layer normalization (LN) or resets we follow the implementation details of the respective papers [Nikishin et al., 2022, Lyle et al., 2023, 2024]: LN is inserted before each ReLU activation and resets are applied to all training networks (actor, critics) and Adam optimizer statistics every $200k$ steps.

Evaluation protocol: We report performance as the mean over 5 evaluation episodes in a separately seeded evaluation environment. Increasing noise levels reduce the performance attainable by any method (Figure 2), compressing the range of achievable scores toward the random-agent baseline. As a result, we deem an equal absolute performance gap between two methods to carry more weight under high noise than in a noise-free setting. To account for this, we normalize scores per task and noise level using the Normalized Score of Agarwal et al. [2021]:

$$\text{Normalized Score} = \frac{\text{Score} - \text{Minimal Score}}{\text{Target Score} - \text{Minimal Score}} \tag{3}$$

In absence of a clearly defined reference performance, such as a game solved in ATARI, we set the target score as the mean over the best configuration per task and training regime, and the minimal score is given by the performance of a random agent in the environment. This normalization ensures that method comparisons are expressed relative to the difficulty of each noise condition, making differences across noise levels and tasks commensurable. For reporting per method we select the best replay ratio based on mean training performance over the last 5000 environment steps, decoupling hyperparameter selection from evaluation. We report the raw learning curves in Appendix B.

5.2 The Effect of Noise during Training on SAC

Glossop et al. [2022] show that even low levels of observation noise *during training* are far more detrimental than the same noise applied only at evaluation time, suggesting that noise actively disrupts the learning process rather than merely making the task harder to execute. We confirm this finding on proprioceptive control tasks from DMC suite. Figure 2 shows that SAC is already strongly impaired by mild observation noise on the challenging tasks `quadruped-run` and `humanoid-walk`, whereas action noise causes a noticeable drop only at higher magnitudes.

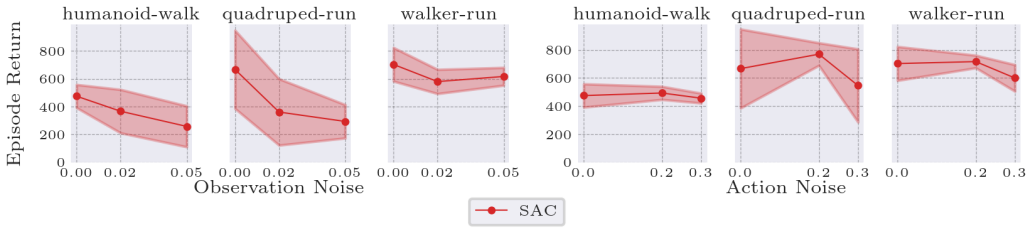

Figure 2: Final performance of SAC under increasing observation noise (left) and action noise (right) after 10^6 training steps.

[2]To facilitate better reproducibility of our research, we open-source our code and research artifacts. The code to reproduce our experiments, as well as links to the generated data, can be accessed via `https://github.com/PhilippBordne/noisy-plasticity`.

Note that noise magnitudes across types are not directly comparable. The action noise is first absorbed by the environment's dynamics before affecting the next state, so its effective impact on the state distribution can differ substantially from the nominal σ_a (see Appendix C). This makes it difficult to attribute the stronger impact of observation noise unambiguously to the Markov property violation versus the perturbation of observation dimensions critical for action selection.

5.3 Can Plasticity-Preserving Methods Improve Training Robustness under Noise?

Following the motivation laid out in section 1, we ask whether methods proposed to preserve plasticity and counteract overfitting to early transitions [Nikishin et al., 2022, Igl et al., 2021] also improve robustness when learning under noise.

Two interventions stand out as particularly effective on the DMC suite [Nauman et al., 2024], each targeting a distinct mechanism. LN [Ba et al., 2016] has been motivated from two complementary angles. On the one hand, Lyle et al. [2024] show that distribution shifts can cause large pre-activation changes that push neurons into saturation and thereby reduce plasticity, and that LN mitigates this. On the other hand, constraining feature activation norms stabilizes TD learning [Bhatt et al., 2024, Gallici et al., 2025, Hussing et al., 2024]. Periodic parameter resets [Nikishin et al., 2022] directly counteract overfitting to early samples by periodically reinitializing network weights, erasing the temporal ordering bias accumulated in the replay buffer [Igl et al., 2021].

Adding LN, periodic resets, or both improves final performance on the two harder tasks but not on `walker-run` (Figure 3), with the largest gains on `quadruped-run` where baseline SAC degrades rapidly under observation noise. On `humanoid-walk`, however, all four configurations converge at $\sigma_o = 0.05$. Aggregated normalized scores (Figure 4, left) confirm that the relative advantage of periodic resets grows with increasing observation noise. Resetting alone outperforms resets combined with LN at all three observation noise levels, while baseline SAC and LN alone fare progressively worse. Under action noise (Figure 4, right), reset-based methods retain their advantage over baseline SAC, but it no longer grows with the noise level and the ranking between the two reset variants flips.

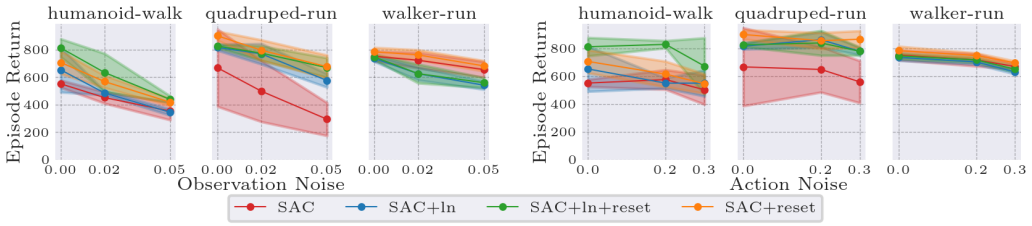

Figure 3: Episode returns from final evaluation after 10^6 steps. Each point aggregates to mean and std over 5 seeds per method and its best replay ratio.

5.4 Can plasticity preservation explain robustness under noise?

The results in Section 5.3 show that plasticity-preserving methods improve robustness under noise, but cannot fully prevent performance degradation. This is unsurprising in part as observation noise turns the MDP into a POMDP and thus some performance loss is inevitable regardless of plasticity. Beyond this, we use our results to better understand how noise and plasticity loss interact during learning. Specifically, we ask whether this relationship is unidirectional, with noise impairing plasticity which then harms performance, or whether learning failure and plasticity loss mutually reinforce each other.

To this end, Figure 5 plots mean critic plasticity across training against final normalized score, for each combination of method, environment, and observation noise level. We use mean plasticity as a proxy for the agent's sustained ability to adapt its representations in response to the learning signal.[3]

Hard tasks: noise and plasticity loss reinforce each other. On `quadruped-run`, most runs that fail to improve over the random baseline show a catastrophic loss of plasticity, whereas runs that

[3]Plasticity as defined in equation 2 can occasionally be negative due to destructive gradient interference across probing batches; we clip values to $[0, 1]$ before averaging.

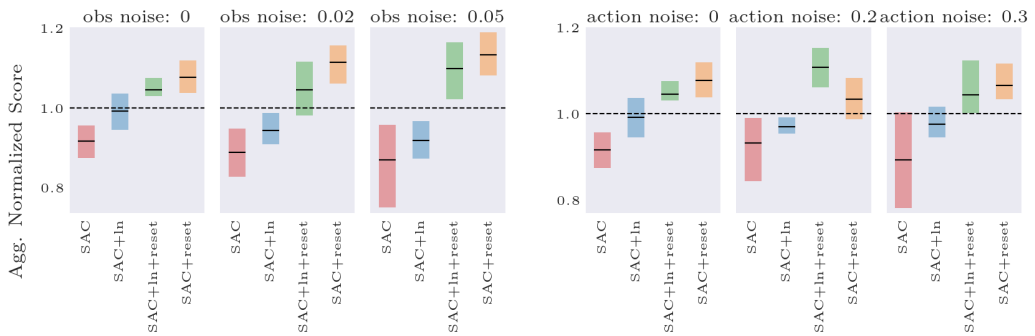

Figure 4: Aggregated normalized scores (stratified bootstrapping, inter-quartile mean (IQM) with 95% CI [Agarwal et al., 2021]) across environments. Per method, environment and noise level the results for the best replay ratio were selected for aggregation. *Left* (observation noise): the relative effectiveness of periodic resets grows with noise magnitude; resets alone consistently outperform resets combined with LN, while baseline SAC and LN alone deteriorate at higher noise levels. *Right* (action noise): no consistent ranking emerges and the relative performance of SAC becomes increasingly uncertain.

retain higher levels of plasticity tend to reach a better final return (Figure 5). Crucially, higher noise also shifts the distribution of learners toward low-plasticity outcomes, indicating the relationship is not unidirectional. Noise does not merely make the task harder for an otherwise intact agent; it also accelerates plasticity loss, which then further undermines learning. On `humanoid-walk` the same trend appears but is less pronounced. Agents with only mild plasticity loss frequently fail under high noise, and some failure is already observed in the noise-free setting, which shows that plasticity loss is only one of several factors making this particular task difficult.

Easy tasks: noise can prevent overfitting. Compared to the previous observations, `walker-run` shows the opposite pattern. Adding observation noise leads to *more* plastic learners on average, and final performance is weakly negatively correlated with plasticity. We hypothesize that noise acts as a mild regularizer in this setting. By making the task slightly harder, it prevents early convergence to simple behaviors that exhibit reduced plasticity.

Action noise. We observe qualitatively similar trends under the highest investigated level of action noise ($\sigma_{\boldsymbol{a}} = 0.3$), but not at lower magnitudes, consistent with the smaller absolute performance drops seen at those levels (Figure 3; see also Appendix D). These findings indicate that an environment's stochasticity itself is a factor reducing learning robustness, and that this cannot be traced back to partial observability alone.

5.5 When Does Plasticity Loss Occur and Are Early Dynamics Critical?

In Section 5.3 we motivated periodic resets as a remedy for overfitting to the noise-corrupted early transitions that dominate the replay buffer. This raises a more specific question: "During which training phase does plasticity loss actually occur, and is the early phase particularly critical?" If so, this would motivate more targeted interventions that apply resets aggressively early but less frequently later, trading plasticity preservation for learning speed.

To investigate, we focus on the most challenging settings in our study, `quadruped-run` and `humanoid-walk` under high observation noise ($\sigma_{\boldsymbol{o}} = 0.05$), and plot mean critic plasticity throughout training (Figure 6). We use a replay ratio of 4 to amplify any effects on plasticity and TD stability.

Consistent with prior work [Hussing et al., 2024], plasticity loss is most pronounced in the very early stages of training and is further exacerbated by observation noise (Figure 6). For methods with periodic resets, the plasticity drop following each subsequent reset becomes progressively less severe. Notably, combining resets with LN does not fully eliminate this pattern: a persistent plasticity drop follows each reset, whereas resets alone eventually yield a regime with negligible plasticity loss between cycles.

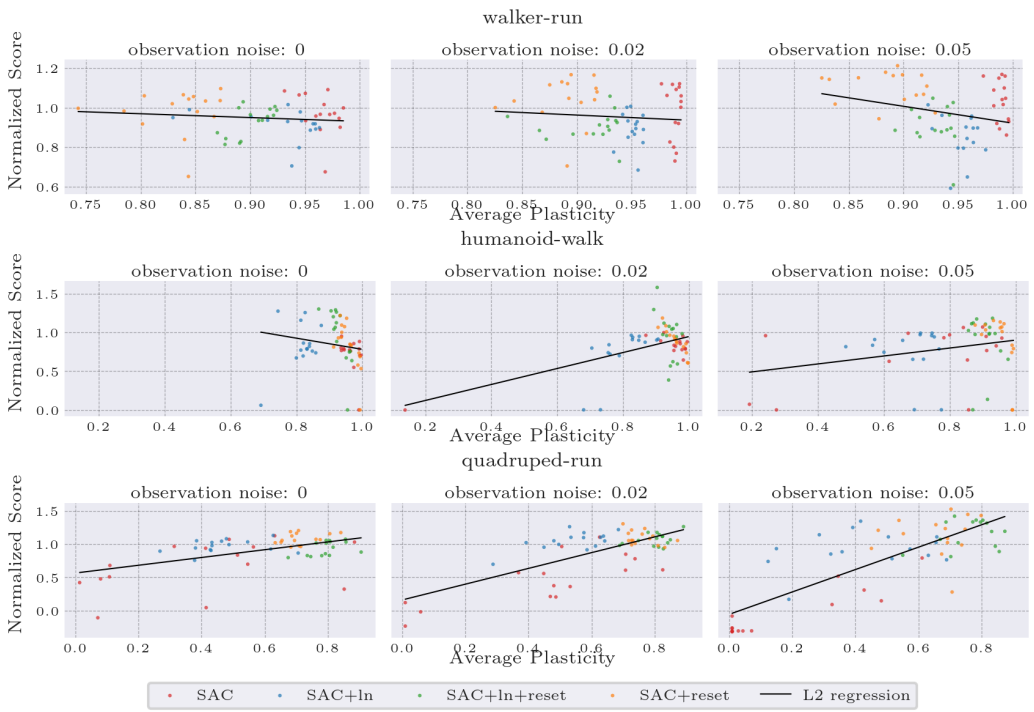

Figure 5: Mean critic plasticity over training vs. final normalized score, per environment and observation noise level (σ_o). Per plot, an individual point is a combination of (method, seed, replay ratio). An $L2$-regression line is fit across all runs per plot.

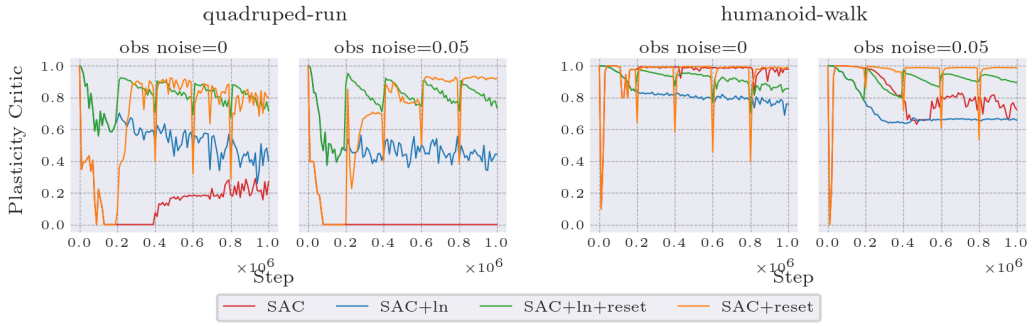

Figure 6: Evolution of critic plasticity throughout training on the `quadruped-run` and `humanoid-walk` tasks. Each curve shows the mean over 5 seeds. Plasticity values have been clipped to $[0, 1]$ before averaging.

To further distinguish plasticity loss from TD instability, which Hussing et al. [2024] identified as the primary failure mode, we track the maximum Q-value within training batches (Figure 7). For baseline SAC on `quadruped-run`, Q-values diverge dramatically under high observation noise, indicating that TD updates become unstable. On `humanoid-walk` no comparable overestimation occurs, consistent with the smaller performance differences between methods seen on that task. These results suggest that in stochastic environments, TD instability is a particularly important failure mode to prevent. While TD instability is catastrophic in any setting, noise makes it more likely to occur whereas plasticity loss and overfitting to early samples alone tend to cause only gradual drops in performance.

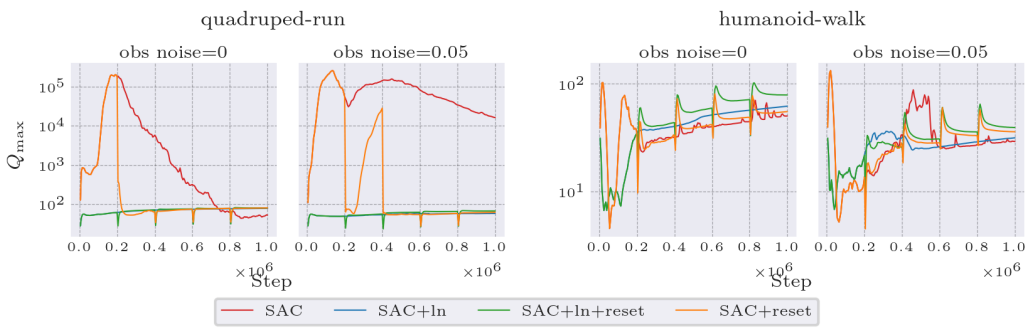

Figure 7: Evolution of the maximum Q-value within training batches throughout training on the `quadruped-run` and `humanoid-walk` tasks. Each curve shows the mean over 5 seeds.

6 Limitations

Due to the computational overhead of evaluation over different noise types and levels we had to limit our experiments to a subset of benchmarking environments. Nevertheless, in terms of the evaluation setting we see a bigger limitation in the characteristics of the studied noise types, in particular applying the same level of noise indiscriminately of the observation dimension. We also believe that studying noisy transition dynamics is of great interest to better understand the interplay between plasticity loss and potential overfitting to early, noisy samples without directly transforming the setting from an MDP into a POMDP. We expect its implementation to be more difficult in relative terms as it will be necessary to implement this on the simulation level and can not be achieved by a simple wrapper. In line with prior works, we opted to apply unrealistically high levels of action noise. However, Glossop et al. [2022] suggest that true transition noise would result in a noticeable impact earlier. Lastly, while plasticity loss and TD stability have been established as important factors for the successful application of deep RL, they are only two of many and studying their interplay with other aspects, such as exploration, is required to gain a holistic understanding of the effectiveness of the evaluated methods.

7 Conclusion

Aggregated across the three DMC suite tasks studied, we find that plasticity-preserving methods improve learning robustness in stochastic environments: reset-based agents achieve a higher aggregate normalized score than baseline SAC at every noise type and level tested. Under observation noise this advantage grows with the noise level, meeting the criterion set out in section 1, whereas under action noise it is persistent but independent of the noise level. Moreover, we observe an increased loss of plasticity in these settings, and in the most drastic case (baseline SAC on `quadruped-run` under $\sigma_{\boldsymbol{o}} = 0.05$) observation noise led to severe value overestimation from which training did not recover. Since stochasticity makes environments inherently more difficult, it is unsurprising that the methods investigated can mitigate but not fully prevent the decrease in performance. Overall, our findings suggest that for SAC in stochastic environments, maintaining TD stability is more critical than preventing overfitting to perturbed, early samples. We found LN to be primarily effective in preventing value divergence while permitting some loss of plasticity. Under observation noise, resetting was the most effective of the four configurations tested, providing both long-term TD stability and plasticity preservation. Somewhat surprisingly, combining resetting with LN underperformed resetting alone at every observation noise level tested and led to an ongoing, albeit slow, loss of plasticity throughout training.

We should expect data collected in the real world to be stochastic by nature and hence we identify a need for benchmark environments that allow the assessment of robust learning under such conditions for the holistic assessment of an algorithm's capabilities. While we resorted to a simple extension of existing benchmark environments we envision benchmarks with more realistic non-deterministic characteristics that we should expect in the real world such as stochastic environment dynamics, random observation delays or observation noise that is well-calibrated with regard to the overall

signal magnitude on the respective channels. Once access to such simulators exists, we should aim to further understand the interplay between environment stochasticity, plasticity loss and further central components of RL algorithms. Lastly, evaluation in such environments should inform the further development and refinement of algorithms in general and plasticity preserving methods in particular.

Acknowledgments

The authors acknowledge support by the state of Baden - Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. The authors further acknowledge funding from The Carl Zeiss Foundation through the research network "Responsive and Scalable Learning for Robots Assisting Humans" (ReScaLe) of the University of Freiburg. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 539134284, through EFRE (FEIH_2698644) and the state of Baden-Württemberg.

References

R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. Liang, J. Vaughan, and Y. Dauphin, editors, *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, pages 29304–29320. Curran Associates, 2021.

M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, S. Gelly, and O. Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR, 2021.

J. Ash and R. P. Adams. On warm-starting neural network training. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*, pages 3884–3894. Curran Associates, 2020.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv:1607.06450 [stat.ML]*, 2016.

A. Bagaria, J. K. Senthil, M. Slivinski, and G. Konidaris. Robustly learning composable options in deep reinforcement learning. In Z. Zhou, editor, *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, pages 2161–2169, 2021.

M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

A. Bhatt, D. Palenicek, B. Belousov, M. Argus, A. Amiranashvili, T. Brox, and J. Peters. Crossq: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity. In *The Twelfth International Conference on Learning Representations (ICLR'24)*. ICLR, 2024.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI gym. *arXiv:1606.01540*, 2016.

L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. D. L. Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

T. Eimer, M. Lindauer, and R. Raileanu. Hyperparameters in reinforcement learning and how to tune them. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *The Eigth International Conference on Learning Representations (ICLR'20)*. ICLR, 2020.

S. Falkner, A. Klein, and F. Hutter. BOHB: Robust and efficient Hyperparameter Optimization at scale. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, volume 80, pages 1437–1446. Proceedings of Machine Learning Research, 2018.

M. D. Farahani and N. Mozayani. Automatic construction and evaluation of macro-actions in reinforcement learning. *Applied Soft Computing*, 82, 2019.

J. Farebrother, J. Orbay, Q. H. Vuong, A. A. Taiga, Y. Chebotar, T. Xiao, A. Irpan, S. Levine, P. S. Castro, A. Faust, A. Kumar, and R. Agarwal. Stop regressing: Training value functions via classification for scalable deep rl. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, volume 251 of *Proceedings of Machine Learning Research*. PMLR, 2024.

A. Faust, A. Francis, and D. Mehta. Evolving rewards to automate reinforcement learning. In K. Eggensperger, M. Feurer, F. Hutter, and J. Vanschoren, editors, *ICML workshop on Automated Machine Learning (AutoML workshop 2019)*, 2019.

J. Franke, G. Köhler, A. Biedenkapp, and F. Hutter. Sample-efficient automated deep reinforcement learning. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR, 2021.

M. Gallici, M. Fellows, B. Ellis, B. Pou, I. Masmitja, J. Foerster, and M. Martin. Simplifying deep temporal difference learning. In *The Thirteenth International Conference on Learning Representations (ICLR'25)*, pages 78148–78190. ICLR, 2025.

C. Glossop, J. Panerati, A. Krishnan, Z. Yuan, and A. P. Schoellig. Characterising the robustness of reinforcement learning for continuous control using disturbance injection. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

B. Grooten, P. MacAlpine, K. Subramanian, P. R. Wurman, and P. Stone. Out-of-distribution generalization with a sparc: Racing 100 unseen vehicles with a single policy. In S. Koenig, C. Jenkins, and M. E. Taylor, editors, *Proceedings of the Fourtieth AAAI Conference on Artificial Intelligence (AAAI'26)*. AAAI Press, 2026.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, volume 80, pages 1861–1870. Proceedings of Machine Learning Research, 2018.

T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine. Soft actor-critic algorithms and applications. *arXiv:1812.05905 [cs.LG]*, 2019.

N. Harder, R. Qussous, and A. Weidlich. Fit for purpose: Modeling wholesale electricity markets realistically with multi-agent deep reinforcement learning. *Energy and AI*, 14, 2023.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In S. McIlraith and K. Weinberger, editors, *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI'18)*. AAAI Press, 2018.

Y. Hu, W. Wang, H. Jia, Y. Wang, Y. Chen, J. Hao, F. Wu, and C. Fan. Learning to utilize shaping rewards: A new approach of reward shaping. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*. Curran Associates, 2020.

S. Huang, R. F. J. Dossa, C. Ye, J. Braga, D. Chakraborty, K. Mehta, and J. G. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

M. Hussing, C. A. Voelcker, I. Gilitschenski, A. Farahmand, and E. Eaton. Dissecting deep RL with high update ratios: Combatting value divergence. *Reinforcement Learning Journal*, 2:995–1018, 2024.

M. Igl, G. Farquhar, J. Luketina, W. Boehmer, and S. Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR, 2021.

M. Jaderberg, V. Dalibard, S. Osindero, W. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu. Population based training of neural networks. *arXiv:1711.09846*, 2017.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

P. Klink, C. D'Eramo, J. Peters, and J. Pajarinen. Self-paced deep reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*, pages 9216–9227. Curran Associates, 2020.

A. Kumar, R. Agarwal, D. Ghosh, and S. Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR, 2021.

C. Lyle, Z. Zheng, E. Nikishin, B. A. Pires, R. Pascanu, and W. Dabney. Understanding plasticity in neural networks. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202 of *Proceedings of Machine Learning Research*, pages 23190–23211. PMLR, 2023.

C. Lyle, Z. Zheng, K. Khetarpal, H. V. Hasselt, R. Pascanu, J. Martens, and W. Dabney. Disentangling the causes of plasticity loss in neural networks. In V. Lomonaco, S. Melacci, T. Tuytelaars, S. Chandar, and R. Pascanu, editors, *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274, pages 750–783. Proceedings of Machine Learning Research, 2024.

G. Ma, L. Li, S. Zhang, Z. Liu, Z. Wang, Y. Chen, L. Shen, X. Wang, and D. Tao. Revisiting plasticity in visual reinforcement learning: Data, modules and training stages. In *The Twelfth International Conference on Learning Representations (ICLR'24)*. ICLR, 2024.

M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. J. Hausknecht, and M. Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

D. J. Mankowitz, T. A. Mann, P. Bacon, D. Precup, and S. Mannor. Learning robust options. In S. McIlraith and K. Weinberger, editors, *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI'18)*, pages 6409–6416. AAAI Press, 2018.

B. Marthi. Automatic shaping and decomposition of reward functions. In Z. Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, pages 601–608. Omnipress, 2007.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

M. Nauman, M. Bortkiewicz, P. Miłoś, T. Trzcinski, M. Ostaszewski, and M. Cygan. Overestimation, overfitting, and plasticity in actor-critic: the bitter lesson of reinforcement learning. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, volume 251 of *Proceedings of Machine Learning Research*, pages 37342–37364. PMLR, 2024.

E. Nikishin, M. Schwarzer, P. D'Oro, P. Bacon, and A. C. Courville. The primacy bias in deep reinforcement learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16828–16847. PMLR, 2022.

OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand. *arXiv:1910.07113*, 2019.

J. Parker-Holder, V. Nguyen, and S. J. Roberts. Provably efficient online Hyperparameter Optimization with population-based bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*. Curran Associates, 2020.

J. Parker-Holder, R. Rajan, X. Song, A. Biedenkapp, Y. Miao, T. Eimer, B. Zhang, V. Nguyen, R. Calandra, A. Faust, F. Hutter, and M. Lindauer. Automated reinforcement learning (AutoRL): A survey and open problems. *Journal of Artificial Intelligence Research*, 74:517–568, 2022.

X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA'18)*, pages 1–8, 2018.

R. Raileanu, M. Goldstein, D. Yarats, I. Kostrikov, and R. Fergus. Automatic data augmentation for generalization in reinforcement learning. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. Liang, J. Vaughan, and Y. Dauphin, editors, *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, pages 5402–5415. Curran Associates, 2021.

R. Rajan, J. L. B. Diaz, S. Guttikonda, F. Ferreira, A. Biedenkapp, J. O. V. Hartz, and F. Hutter. MDP playground: An analysis and debug testbed for reinforcement learning. *Journal of Machine Learning Research*, 77:821–890, 2023.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347 [cs.LG]*, 2017.

G. Sokar, R. Agarwal, P. S. Castro, and U. Evci. The dormant neuron phenomenon in deep reinforcement learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction, 2nd Edition*. MIT Press, 2018.

Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. D. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. P. Lillicrap, and M. A. Riedmiller. Deepmind control suite. *arXiv:1801.00690*, 2018.

E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS'12)*, pages 5026–5033. ieeecis, IEEE, 2012.

D. Yarats, I. Kostrikov, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR, 2021.

Z. Yuan, A. W. Hall, S. Zhou, L. Brunke, M. Greeff, J. Panerati, and A. P. Schoellig. Safe-control-gym: A unified benchmark suite for safe learning-based control and reinforcement learning in robotics. *IEEE Robotics and Automation Letters*, 7(4):11142–11149, 2022.

A Hyperparameter Settings Soft Actor Critic

For our study we use the SAC implementation of CleanRL [Huang et al., 2022]. We keep the default hyperparameters and only vary the replay ratio. We report the hyperparameters in Table 1. Further details about hyperparameters can be seen in our code repository.

Table 1: Hyperparameters for SAC.

optimizer	Adam
batch size	256
β_1 (Adam)	0.9
β_2 (Adam)	0.999
learning rate	`3e-4`
buffer size	`1e6`
γ (discount factor)	0.99
τ (target smoothing coefficient)	0.005
α (entropy coefficient, fixed)	0.2
train frequency	1
target network frequency	1
replay ratio	[1, 4, 8]
learning starts	1000

B Learning Curves and Normalized Scores Per Environment

We first provide the raw reward curves per considered replay ratio, environment and observation noise (Figure 8) as well as action noise (Figure 9). We further provide the plasticity (Figures 10 and 11) and $\mathcal{Q}_{max}$ evolution (Figures 12 and 13) plots.

Raw Reward Curves Across observation noise levels.

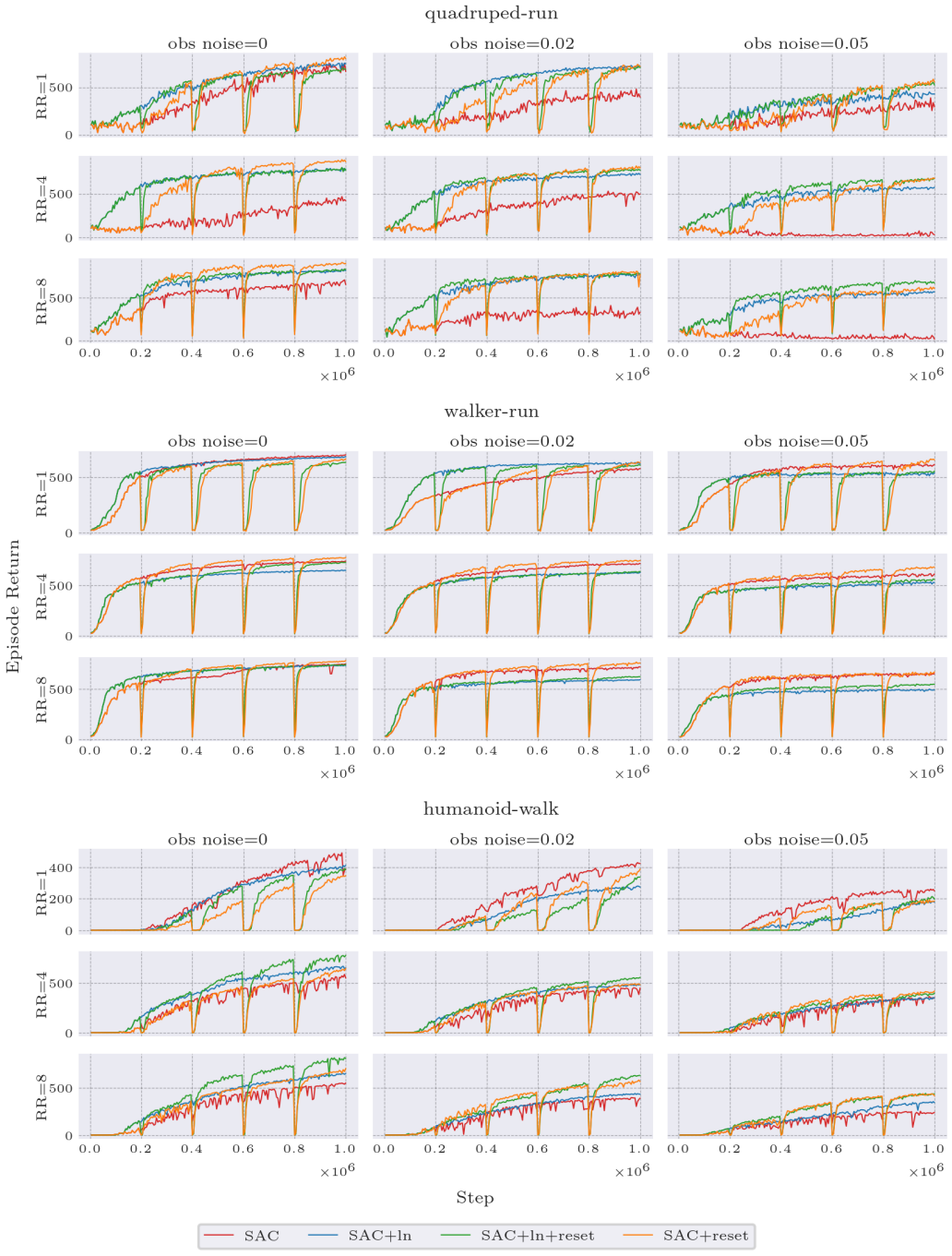

Figure 8: Learning curves for the raw reward across replay ratios and observation noise level

Across action noise levels.

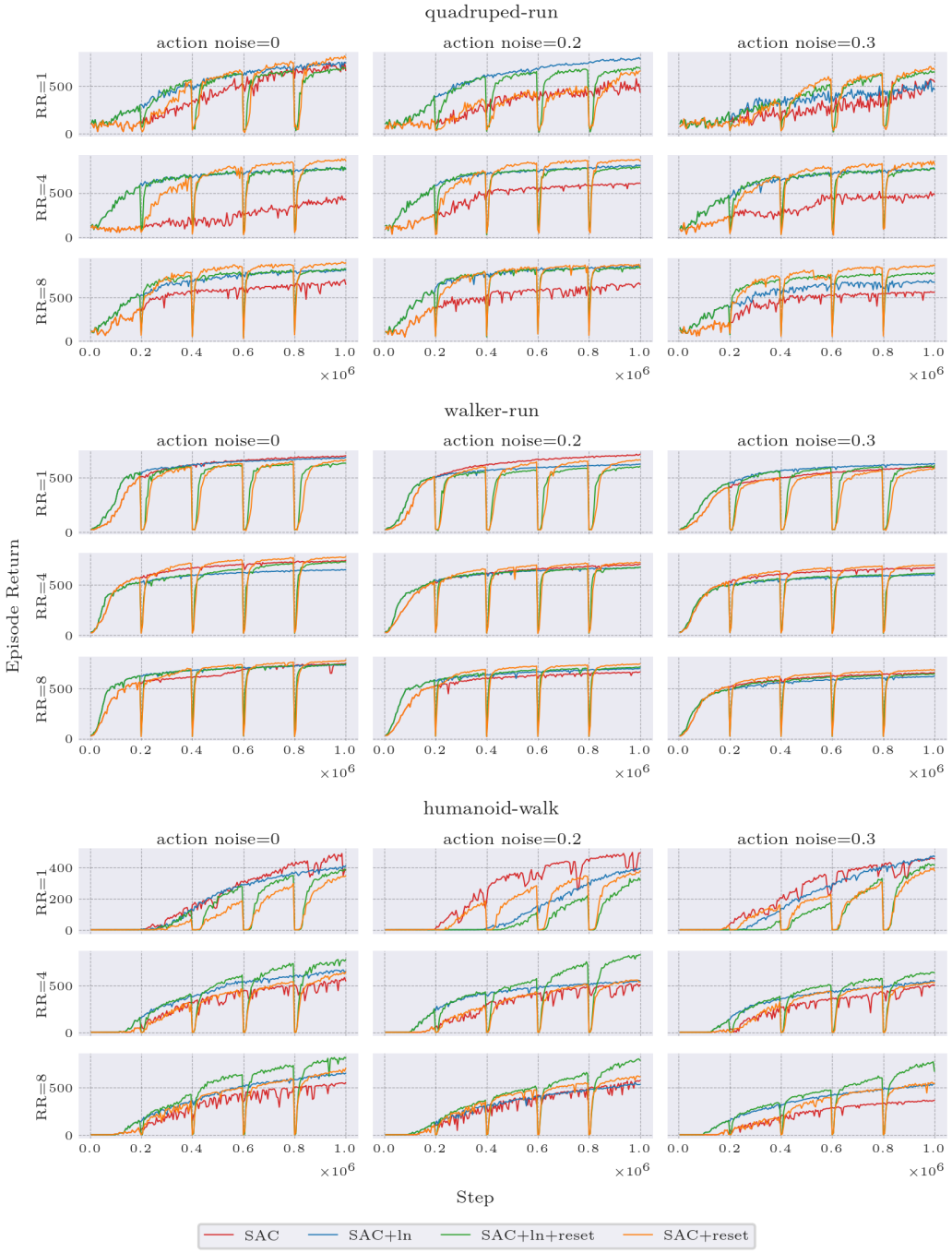

Figure 9: Learning curves for the raw reward across replay ratios and action noise level

Plasticity Evolution Across observation noise levels.

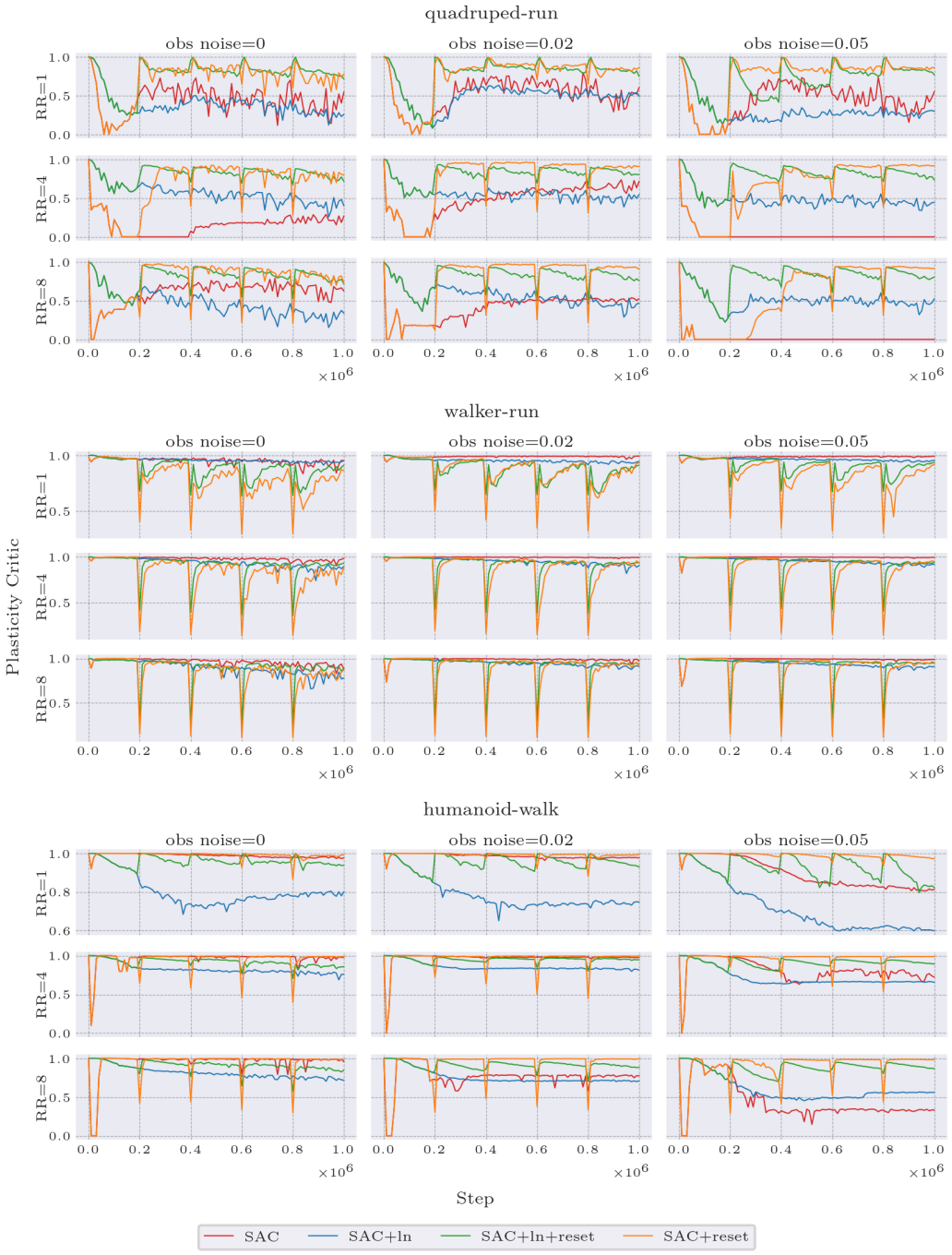

Figure 10: Learning curves for the plasticity evolution across replay ratios and observation noise level

Across action noise levels.

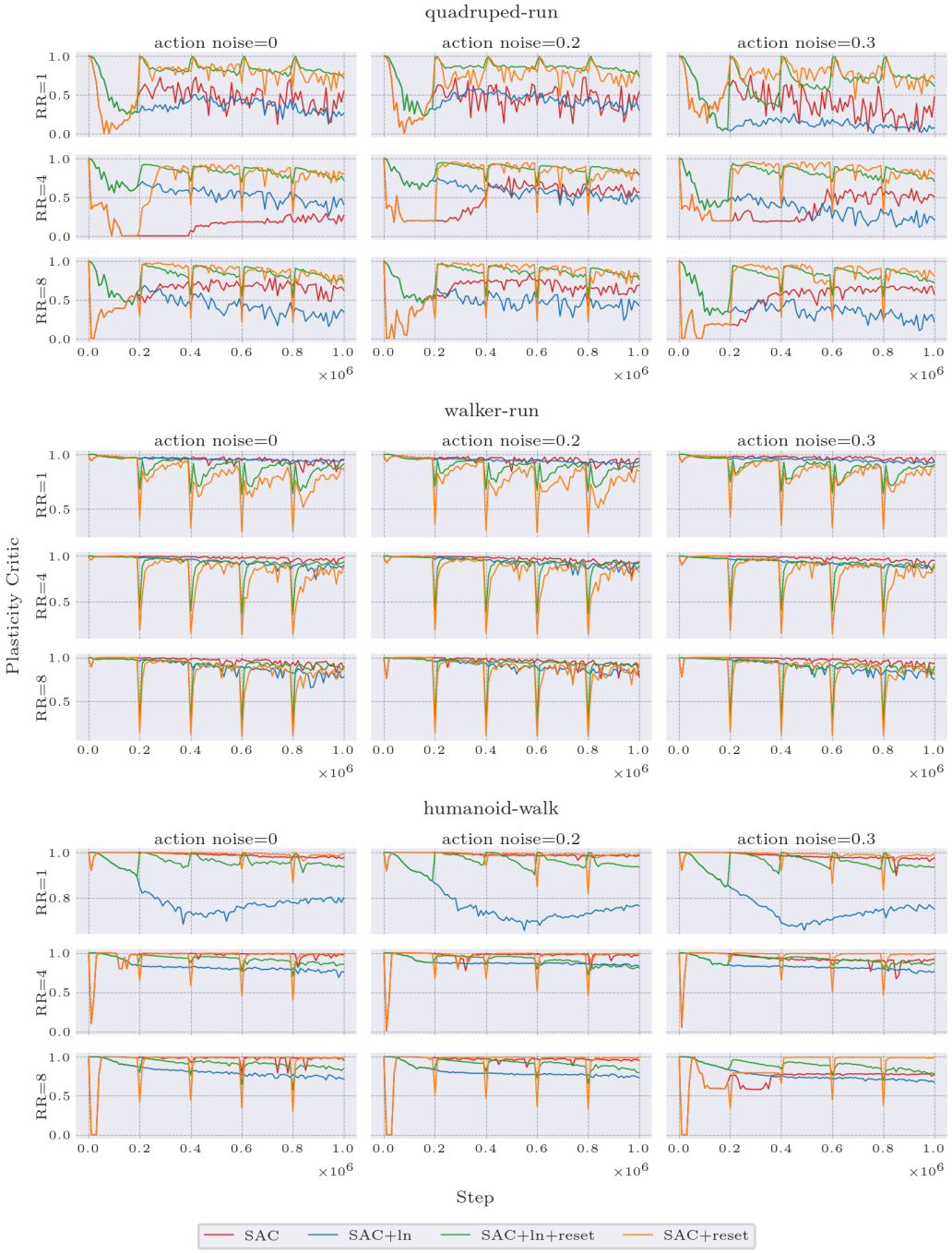

Figure 11: Learning curves for the plasticity evolution across replay ratios and action noise level

Maximum $\mathcal{Q}$-value Across observation noise levels.

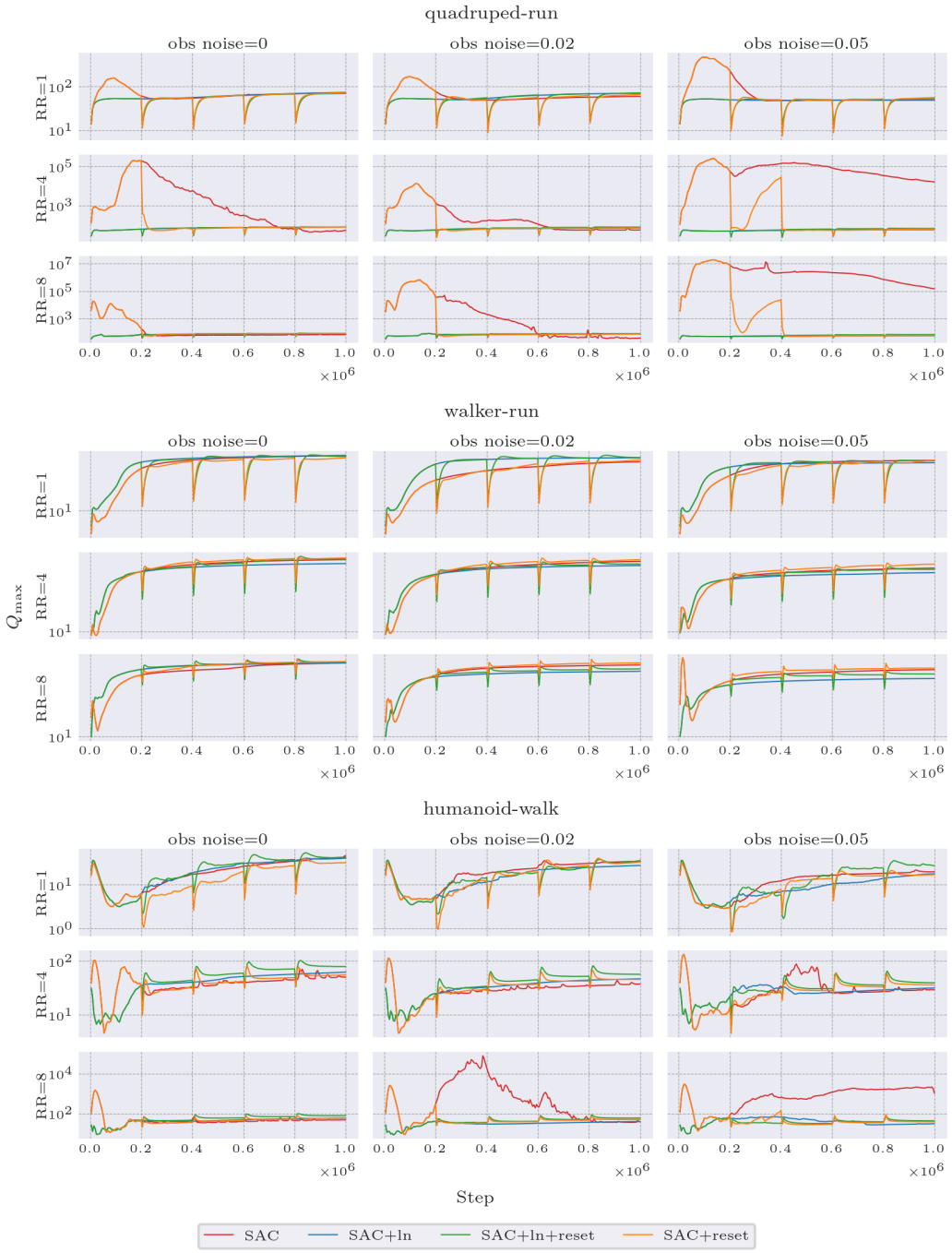

Figure 12: Learning curves for the $\mathcal{Q}_{max}$ across replay ratios and observation noise level

Across action noise levels.

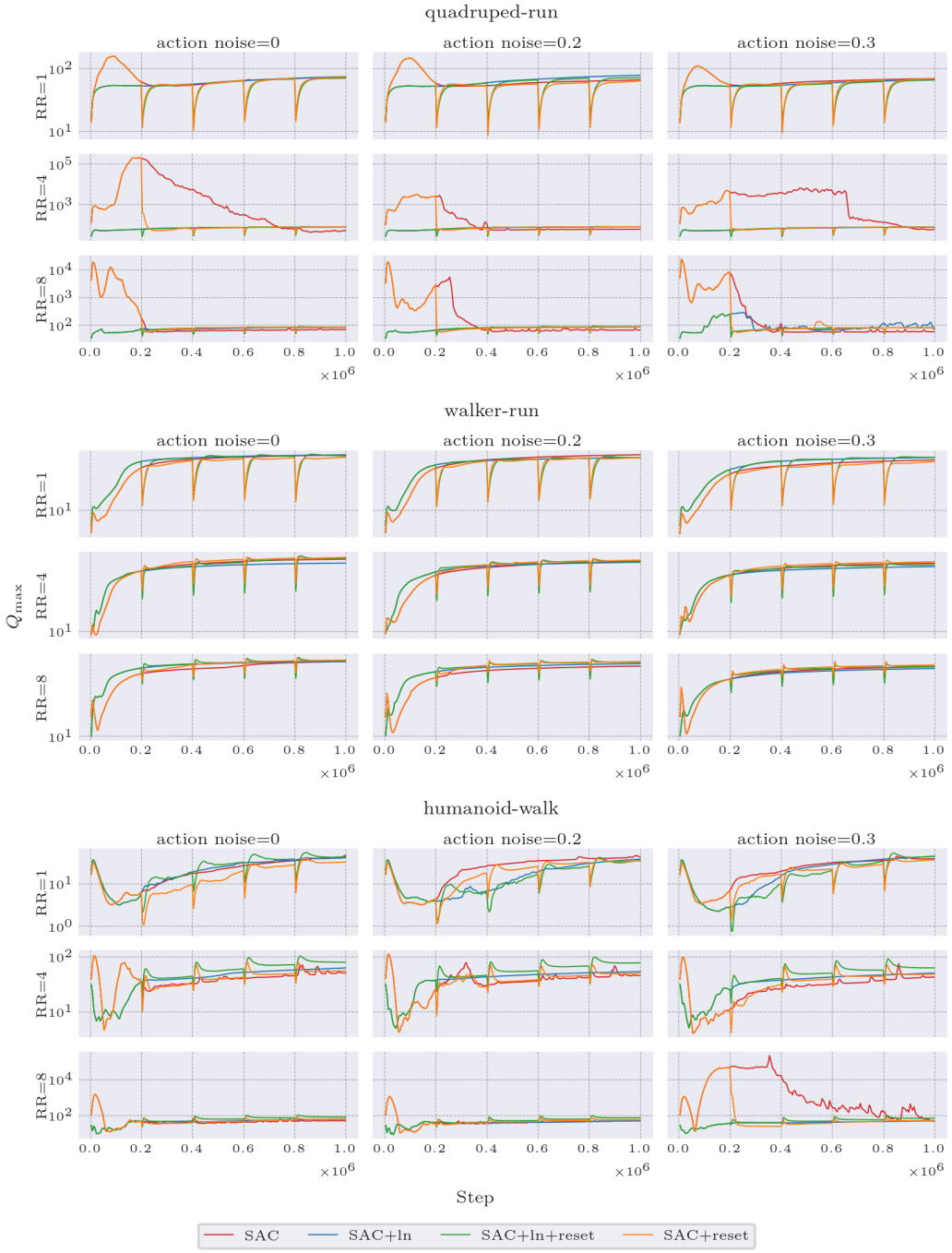

Figure 13: Learning curves for the $\mathcal{Q}_{max}$ across replay ratios and action noise level

Normalized Scores Here we report the final episode returns for the best replay ratio across all prior combinations.

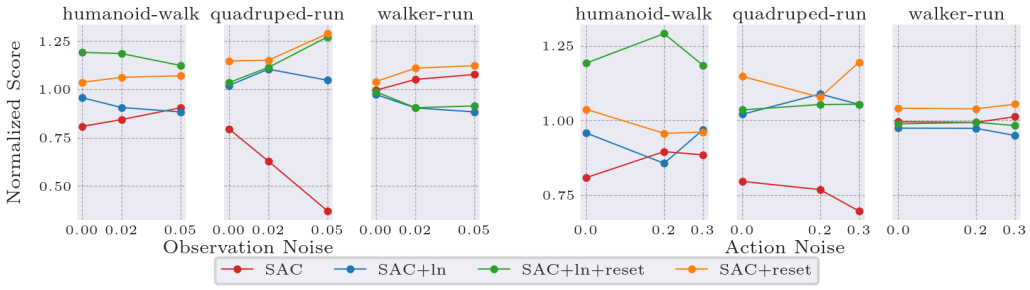

Figure 14: Normalized final episode return of the best replay ratio per method, environment and noise level.

C Noise Effect on Transition Dynamics

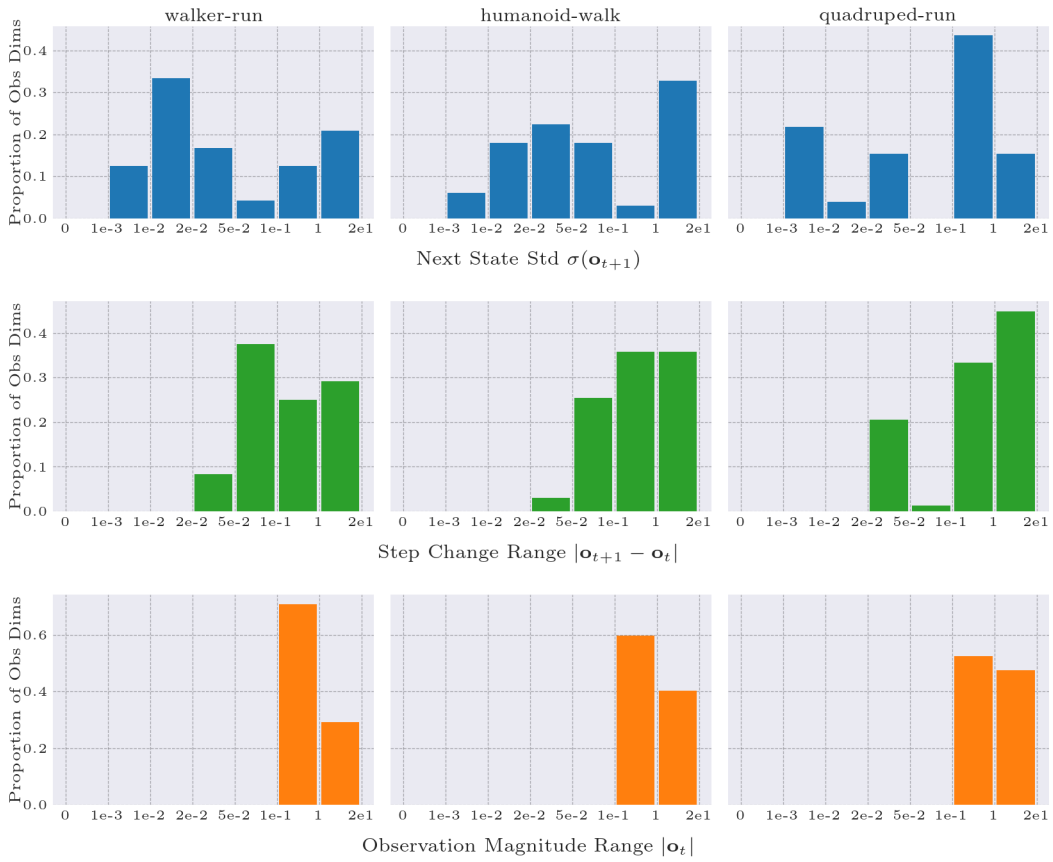

Figure 15: Distribution of per-dimension next-state observation sensitivity to action noise for three `dm_control` environments. For each step along a trajectory, $n = 10$ independent action–noise pairs $(\mathbf{a}, \mathbf{a} + \boldsymbol{\varepsilon})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ with $\sigma = 0.3$, are evaluated from the same physics state. The per-step standard deviation of the resulting observation errors $f(s, \mathbf{a} + \boldsymbol{\varepsilon}) - f(s, \mathbf{a})$ is computed for each observation dimension and averaged over steps (1000 steps $\times$ 3 seeds). Bars show the proportion of observation dimensions whose mean std falls within each magnitude bin.

D Plasticity vs Learning Robustness under Action Noise

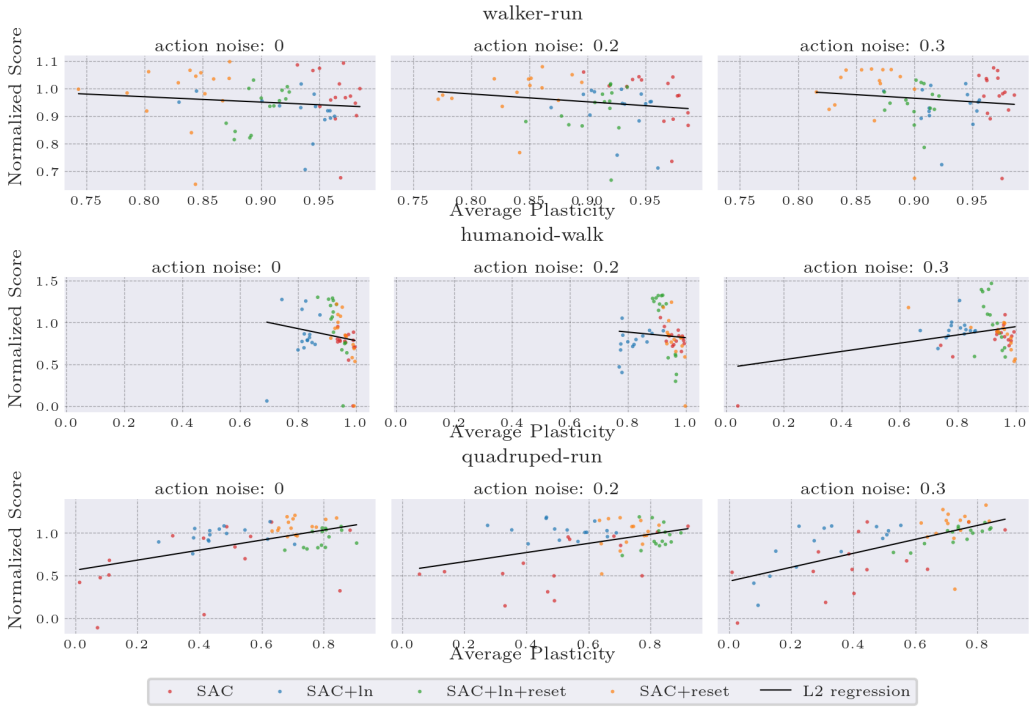

Figure 16: Mean critic plasticity over training vs. final normalized score, per environment and action noise level (σ_a). Within one plot each point is a combination of (method, seed, replay ratio). A $L2$-regression line is fit across all runs per plot.

E Replay-Ratio Sensitivity to Environment Noise

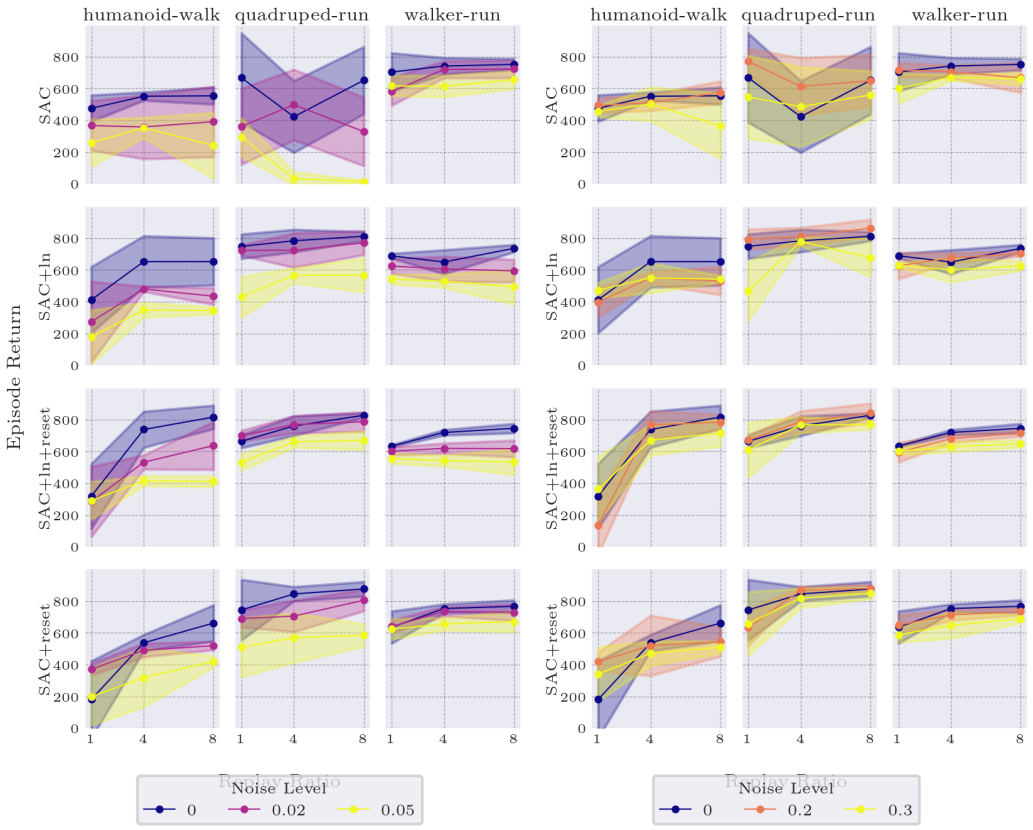

Figure 17: Episode returns from final evaluation after 10^6 steps per noise level and across replay ratios. The goal is to detect whether high replay ratios are more affected by high noise levels due to overfitting (which is generally not the case). Each point aggregates to mean and std over 5 seeds.