Overestimation, Overfitting, and Plasticity in Actor-Critic: the Bitter Lesson of Reinforcement Learning

Michal Nauman [*1 2] **Michał Bortkiewicz** [*3] **Piotr Miłoś** [1 2 4] **Tomasz Trzciński** [1 3 5] **Mateusz Ostaszewski** [† 3] **Marek Cygan** [† 2 6]

Abstract

Recent advancements in off-policy Reinforcement Learning (RL) have significantly improved sample efficiency, primarily due to the incorporation of various forms of regularization that enable more gradient update steps than traditional agents. However, many of these techniques have been tested in limited settings, often on tasks from single simulation benchmarks and against well-known algorithms rather than a range of regularization approaches. This limits our understanding of the specific mechanisms driving RL improvements. To address this, we implemented over 60 different off-policy agents, each integrating established regularization techniques from recent state-of-the-art algorithms. We tested these agents across 14 diverse tasks from 2 simulation benchmarks, measuring training metrics related to overestimation, overfitting, and plasticity loss — issues that motivate the examined regularization techniques. Our findings reveal that while the effectiveness of a specific regularization setup varies with the task, certain combinations consistently demonstrate robust and superior performance. Notably, a simple Soft Actor-Critic agent, appropriately regularized, reliably finds a better-performing policy within the training regime, which previously was achieved mainly through model-based approaches.

* Main authors contributed equally to this work; † Senior authors contributed equally to this work. [1]Ideas NCBR [2]University of Warsaw [3]Warsaw University of Technology [4]Polish Academy of Sciences [5]Tooploox [6]Nomagic. Correspondence to: Michal Nauman <nauman.mic@gmail.com>, Michal Bortkiewicz <michalbortkiewicz8@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

In recent years, substantial improvements have been made in the domain of deep reinforcement learning, as evidenced by breakthroughs such as mastering complex games like Dota 2 (OpenAI et al., 2019), Go (Silver et al., 2017) and achieving control over nuclear fusion plasma (Degrave et al., 2022). In particular, off-policy RL has witnessed a surge of approaches reporting state-of-the-art results (Li et al., 2022; Hafner et al., 2023; Lee et al., 2023), including application to real robots (Smith et al., 2022). In general, those approaches build upon Soft Actor-Critic (SAC) algorithm with increased number of gradient steps per environment steps (Replay Ratio (RR)) used in conjunction with some form of regularization that stabilizes the learning in high RR setting (Janner et al., 2019; Chen et al., 2020; Hiraoka et al., 2021; Nikishin et al., 2022; Li et al., 2022; D'Oro et al., 2022; Cetin & Celiktutan, 2023). These approaches for regularization encompass considerations of reducing overfitting (Li et al., 2022) (*network regularization*), reducing critic overestimation (Cetin & Celiktutan, 2023) (*critic regularization*) or reducing the rate of plasticity loss (Lee et al., 2023) (*plasticity regularization*).

Despite significant advancements, the understanding of how different regularization techniques synergistically improve off-policy agent performance is still limited (Hiraoka et al., 2021; Lee et al., 2023). Moreover, most methods are tested in narrow contexts, mainly in locomotion or manipulation tasks, often restricted to a single simulation benchmark (Fujimoto et al., 2018; Haarnoja et al., 2018; Chen et al., 2020; Moskovitz et al., 2021; D'Oro et al., 2022), leading to questions about their broad applicability and robustness. In this study, our goal is to consolidate these lessons and address the following research questions: *Which regularization techniques lead to robust performance improvements across diverse tasks and agent designs? Can generic regularization techniques outperform domain-specific RL techniques that directly use the MDP structure?*. We extend the scope of prior research by examining over 60 design choices implemented within the Soft Actor-Critic framework. We test a diverse array of tasks, including both locomotion and manipulation, within two simulation benchmarks and two replay

ratio regimes. This comprehensive approach offers a deeper understanding of the effectiveness of these regularization techniques in various settings.

Our main result is a bitter lesson: across varied tasks, general neural network regularizers significantly outperform most RL-specific algorithmic improvements in terms of agent performance. Specifically, we find general methods that are motivated by stabilization of gradient-based learning significantly outperform RL-specific algorithmic improvements across a variety of environments. Such emphasis on generality is in line with the celebrated "Bitter Lesson" essay (Sutton, 2019). Notably, network regularization enables agents to find effective policies on tasks previously impossible for model-free agents, such as those in the dog domain. Our findings also show that layer normalisation is more effective in reducing overestimation than techniques specifically designed for mitigating Q-value overestimation in critic networks. Consequently, we show that replacing the ubiquitous Clipped Double Q-learning with network regularization techniques leads to significant performance gains. Our research further explores the impact of overestimation, overfitting, and plasticity loss on agent performance in a unified experimental setup. We examine the correlation between these factors and agent performance, showing a strong negative correlation for value overestimation and agent plasticity metrics. These influences vary significantly across environments, underscoring the complex nature of their effects on learning. A key observation is the environment-dependent performance of various methods. Strategies excelling in locomotion tasks may falter in manipulation scenarios and vice versa. Comparisons between experiments on the DeepMind Control Suite (Tassa et al., 2018) and MetaWorld (Yu et al., 2019) demonstrate the necessity for diverse benchmarking in research, highlighting the value of expansive experimental setups.

1. Our study presents an extensive empirical analysis of various regularization techniques in off-policy RL. We evaluate the effectiveness, robustness, and generality of 12 SAC design choices derived from recent literature, examining their diverse interactions. This encompasses testing 64 model designs across 14 tasks from two benchmarks under two replay ratio regimes.

2. Our findings show that combining well-established network regularization techniques with methods that prevent plasticity loss effectively addresses the value estimation problem, eliminating the need for critic regularization. Specifically, we observe that in network/plasticity regularized agents using critic regularization often leads to significant performance degradation. Leveraging these insights, we demonstrate that integrating specific regularization methods into the basic Soft Actor-Critic framework leads to state-of-the-art performance in dog domain tasks for model-free approaches.

3. Our study investigates the correlation between overestimation, overfitting, and plasticity proxies, and their impact on agent performance. We discover that interventions aimed at one type of issue, such as full-parameter resets, significantly affect proxies for issues other than plasticity such as overestimation and overfitting, often more than interventions specifically designed for those other issues. This suggests that RL agents encounter a range of complex problems that collectively affect the learning process.

2. Background

We consider a Markov Decision Process (MDP) (Puterman, 2014; Sutton & Barto, 2018) which is described via a tuple (S, A, r, p, γ), where states S and actions A are continuous, $r(s, a)$ is the transition reward, $p(s'|s, a)$ is a transition mapping, p_0 is the starting state distribution and $\gamma \in (0, 1]$ is the discount factor. Policy, denoted as $\pi(a|s)$ is a state-conditioned action distribution. Maximum Entropy Reinforcement Learning (MaxEnt RL) objective (Ziebart et al., 2008; Haarnoja et al., 2017) is to find a policy that maximizes the expected sum of discounted returns and policy entropies, or equivalently expected initial state values according to $\pi^* = \arg\max \mathbb{E}_{p_0} V^\pi(s_0)$. The Q-value is defined as $Q^\pi(s, a) = r(s, a) + \gamma V^\pi(s')$. State value is defined by $V^\pi(s) = \mathbb{E}_\pi(Q^\pi(s, a) - \alpha \log \pi(a|s))$, where $\alpha \log \pi(a|s)$ is the maximum entropy term. In actor-critic, policy and Q-value functions are represented by parameterized function approximators (Silver et al., 2014). Policy parameters θ are updated to maximize the value approximation at sampled states s from an off-policy replay buffer $\mathcal{D}$ (Fujimoto et al., 2018; Haarnoja et al., 2018):

$$\theta^* = \arg\max_\theta \mathop{\mathbb{E}}_{\mathcal{D}} Q_\phi(s, a) - \alpha \log \pi_\theta(a|s),\ a \sim \pi_\theta. \quad (1)$$

The critic parameters ϕ are updated by minimizing the temporal-difference (Silver et al., 2014):

$$\phi^* = \arg\min_\phi \mathop{\mathbb{E}}_{\mathcal{D}} \big(Q_\phi(s, a) - r_(s, a) - \gamma \bar{V}_\phi(s')\big)^2, \quad (2)$$

where $\bar{V}_\phi(s')$ is the target network (Mnih et al., 2015).

2.1. Overestimation

Q-learning methods employing function approximation have been observed to exhibit a bias toward overestimation, a phenomenon critical to the training process (Thrun & Schwartz, 2014; Fujimoto et al., 2018). Positive bias stems from the policy being trained to locally maximize action-value estimates, leading its actions to exploit potential model errors for higher scores. Modern actor-critic

algorithms leverage a variety of countermeasures to overestimation of Q-value targets, with Clipped Double Q-learning (**CDQ**) (Fujimoto et al., 2018) being most used by many other algorithms (Haarnoja et al., 2018; Chen et al., 2020; Hiraoka et al., 2021). In CDQ, the algorithm maintains two critics and uses their minimum as an approximate Q-value lower bound. The CDQ was generalized to the following pessimistic objective (Ciosek et al., 2019; Moskovitz et al., 2021; Cetin & Celiktutan, 2023):

$$Q^{\beta}_{\phi}(s,a) = Q^{\mu}_{\phi}(s,a) - \beta Q^{\sigma}_{\phi}(s,a). \tag{3}$$

We denote the level of pessimism as β, and the critic ensemble mean and standard deviation as Q^{μ}_{ϕ} and Q^{σ}_{ϕ} respectively. In particular, for $\beta = 1$, the above rule is exactly equal to the CDQ minimum (Ciosek et al., 2019; Cetin & Celiktutan, 2023). The success of pessimistic updates led to various methods for adjusting β online. A recent approach, Generalized Pessimism Learning (**GPL**) (Cetin & Celiktutan, 2023), estimates the critic approximation error and modifies β accordingly. A different strategy, Tactical Optimism and Pessimism (**TOP**) (Moskovitz et al., 2021), adjusts pessimism independent of the estimated approximation error. Specifically, TOP uses an external bandit controller to maximize online episodic rewards. Whereas this controller is aligned with the RL objective, it only allows for discrete values of pessimism.

2.2. Overfitting

Overfitting, while not commonly scrutinized in reinforcement learning, has gained attention in recent discussions (Li et al., 2022) as a phenomenon correlated with performance decline in models characterized by a high ratio of updates to data. To evaluate overfitting in agents, Li et al. (2022) utilizes a validation dataset that consists of samples gathered using the same policy as the canonical replay buffer. The validation buffer is established to provide an unbiased assessment of the critic error in experiences that were not used in the learning. Although there are many strategies to deal with overfitting in supervised learning, only a few of them were applied in the context of RL. To this end, application of Weight Decay (**WD**) (Schwarzer et al., 2023), Layer Normalization (**LN**) (Ball et al., 2023) or Spectral Normalization (**SN**) (Cetin & Celiktutan, 2023) was shown to greatly effect the performance of the underlying agent.

2.3. Plasticity

Plasticity, in the context of models, refers to their ability to learn new information. The concept of plasticity loss has recently gained prominence in the deep learning community, particularly in supervised learning (Achille et al., 2017; Ash & Adams, 2020; Dohare et al., 2021) and RL (Nikishin et al., 2022; Dohare et al., 2021; Lyle et al., 2023; Lee et al., 2023; Kumar et al., 2023; Nikishin et al., 2023). Numerous hypotheses have been proposed regarding the sources of plasticity loss, including dead or dormant units, rank collapse, and divergence due to large weight magnitudes (Lyle et al., 2022; Sokar et al., 2023; Kumar et al., 2020; Dohare et al., 2021). However, none of these mechanisms alone is sufficient to explain the phenomenon of plasticity loss. Whereas the cause of plasticity loss remains to be discovered, various approaches for regularizing the model plasticity have been proposed. For example, full-parameter **resets** of actor-critic modules were shown to greatly improve the agent's ability to learn (Nikishin et al., 2022; D'Oro et al., 2022). The problem of plasticity was also tackled at the level of the activation function with Concatenated ReLU (**CRLU**) (Abbas et al., 2023) or the optimizer with the Sharpness-Aware Minimization (**SAM**) (Foret et al., 2020).

3. Study Design

In this paper, we analyze the impact of various interventions on SAC performance across seven DeepMind Control Suite (Tassa et al., 2018) (DMC) tasks: acrobot-swingup, hopper-hop, humanoid-walk, humanoid-run, dog-trot, dog-run, quadruped-run and seven MetaWorld (Yu et al., 2019) (MW) tasks: Hammer, Push, Sweep, Coffee-Push, Stick-Pull, Reach, Hand-Insert. We chose a wide spectrum of tasks, ranging from easy (acrobot-swingup, Reach) to barely solvable (dog-run) for generic insights that are not overfitted to only a specific group. We choose tasks that are not easily solved by the baseline high replay SAC, as presented in D'Oro et al. (2022) and Hansen et al. (2022), with added dog tasks (which are generally unsolved in state-based representation). Finally, following (Li et al., 2022), we conduct experiments in low 2 and high replay regimes. Such experimental design allows us to pinpoint if specific regularization targets issues associated with high replay, or if it is universally applicable across varying replay regimes. Categorizing them based on current state-of-the-art methods, we identify three intervention groups:

- Critic Regularizations (CR),
  - Clipped Double Q-learning (CDQ) (Fujimoto et al., 2018),
  - Tactical Optimism Pessimism (TOP) (Moskovitz et al., 2021),
  - Generalized Pessimism Learning (GPL) (Cetin & Celiktutan, 2023),
- Network Regularizations (NR),
  - Layer Norm (LN) (Ba et al., 2016),
  - Spectral Norm (SN) (Miyato et al., 2018; Zhang et al., 2018; Brock et al., 2018),
  - Weight Decay (WD) (Loshchilov & Hutter, 2017),

- Plasticity Regularizations (PR),
  - Resets (Res) (Nikishin et al., 2022),
  - Concatenated ReLU activations (CRLU) (Abbas et al., 2023),
  - Sharpness-Aware Minimization Optimizer (SAM) (Foret et al., 2020).

To explore the interactions between interventions, we systematically run all possible combinations of methods across groups, ensuring that methods from the same group are not combined. Each configuration is evaluated on 10 seeds. In the results analysis, we categorize marginalization into three levels:

First-order marginalization combines all results for a specific intervention. For instance, the marginalized performance of layer norm will be computed as the average performance across all combinations with interventions from other groups (in this example, Critic Regularizations and Plasticity Regularizations).

Second-order marginalization involves evaluating the performance of fixed pairs of methods from two groups and marginalizing results from the third group.

Third-order results involve no intervention marginalization and represent the performance of a specific combination, including one method from each group. The only marginalization is over all tested environments (we present these results in Appendix C.1 due to space constraints). This cumulative result provides insight into the overall impact of a given intervention.

Furthermore, we conduct an analysis of various proxy metrics associated with the problems of overestimation, overfitting, and plasticity loss. For overestimation, we evaluate the state-action critic approximation error, denoted as $b_\phi(s,a)$, is quantified as the disparity between the critic output and the true on-policy Q-value according to $b_\phi(s,a) = Q_\phi(s,a) - Q^\pi(s,a)$, where Q_ϕ denotes the critic Q-value approximation and Q^π represents the on-policy Q-value which we estimate via a Monte-Carlo rollout with 5 samples. To calculate overfitting, we compare average TD errors on evaluation trajectories (which are not used for learning) to average TD errors observed in training according to $o_\phi = \frac{\mathbb{E}_{\mathcal{D}_v} TD_\phi}{\mathbb{E}_{\mathcal{D}} TD_\phi}$, where o_ϕ denotes critic overfitting, $\mathcal{D}_v$ denotes validation data, and TD_ϕ denotes the temporal difference loss. As such, the extent of overestimation is then quantified by the ratio of validation TD error to training TD error. We monitor plasticity loss by the rank of penultimate layer representations (Kumar et al., 2020), dormant neurons or dead units (Sokar et al., 2023), the L2 norm of weights (Nikishin et al., 2022; Lyle et al., 2023), and gradient norm (Nikishin et al., 2022; Lyle et al., 2023) as a proxy for plasticity loss.

4. Experiments

4.1. Combination of interventions – First-order marginalization

Study description: First-order marginalization provides insights into the robust impact of a given intervention on model performance, irrespective of what other type of regularization it is paired with. To measure such robustness, we compare the performance of the baseline SAC model augmented with one specific regularization (e.g., SAC + WD) to the performance of SAC augmented with this regularization paired with some other technique (e.g. SAC + WD + Resets).

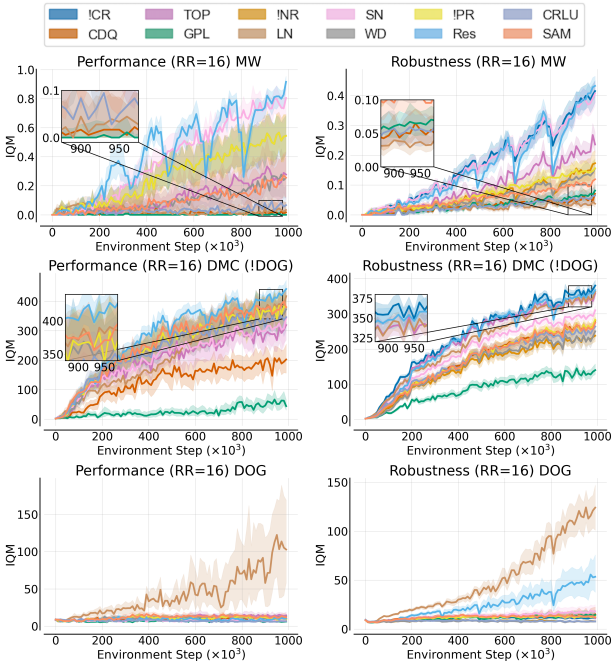

Figure 1. IQM performance of First-Order Marginalization. The left column presents results for baseline SAC augmented with a single regularization technique (and thus uses 10 seeds per task), and the right column presents the aggregate performance of a specific regularization technique when paired with other regularizations (and thus uses 640 seeds per task). Results are presented for MW (top row), DMC without Dog environments (middle row) and only Dog-run and Dog-trot (bottom row) benchmarks. 14 tasks.

Results: Examining the plots in Figure 1 shows that network and plasticity regularization techniques are generally more effective than critic regularization – dark blue line (!CR) on Robustness plots on MW and DMC (!DOC). We observe these results for both simple models using one regularization technique and more complex agents leveraging many regularizations at once. Most notably, avoiding the use of critic regularization interventions (!CR) proves advantageous for both DMC and MW, especially if some other type of regularization is used (such as layer norm or full-

parameter resets). This result is somewhat surprising, as critic regularization methods were designed specifically for off-policy actor-critic agents, whereas the network and plasticity regularization techniques are general. Notably, TOP (Moskovitz et al., 2021) emerges as an exception, particularly showcasing its effectiveness on the DMC benchmark. Conversely, the GPL intervention exhibits the least robust performance in both DMC and MW. Upon further analysis of the impact of CDQL presence (see Section C.4), it becomes apparent that in certain environments, such as Hopper Hop, the adverse effects of this intervention cannot be mitigated even with additional regularizations. Full-parameter resets (Nikishin et al., 2022), tailored for high replay ratio regimes, prove to be one of the most robust approaches in this RR regime. Further analysis reveals discrepancies in conclusions between benchmarks. Clearly, LN is the most effective approach in the DMC Dog environments (bottom row in Figure 1), but it also ranks among the top four interventions in the remaining DMC environments. However, it exhibits very poor performance on the MW benchmark. Therefore, we find SN to be more robust, significantly aiding the MW benchmark and providing moderate assistance in the DMC scenarios.

Takeaways:

- Critic regularization methods exhibit limited effectiveness in enhancing performance. When using network or plasticity regularization, critic regularization leads to reduced performance.
- Periodical network resetting is the most robust intervention across two benchmarks in a high replay ratio regime, and highly surpasses other plasticity regularization techniques in both robustness and performance.
- Layer norm is essential for Dog environments.
- When considering network regularization approaches, layer norm is generally recommended for DMC, while spectral norm is more effective for MW benchmarks. When considering a diverse range of tasks, we find spectral norm to be more robust than layer norm. Weight decay has generally low performance when used alone with SAC.

4.2. Combination of interventions – Second-order marginalization

Study description: This study delves into second-order marginalization to pinpoint the most effective combinations. Results are presented across various replay ratios (2 and 16) and benchmarks (DMC or MW). Given the limited impact of critic regularizations like CDQ or GPL in the first order experiments, our focus is on discerning the most advantageous combinations involving of regularization.

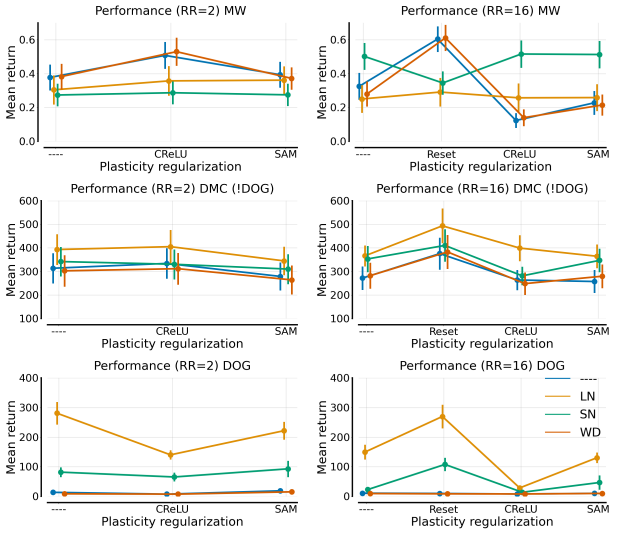

Figure 2. Second-order results marginalizing critic regularization methods. On the x-axis, we have different types of plasticity regularization, and each colour denotes network regularization. For better readability, points within one plasticity regularization are spaced slightly horizontally. Vertical lines indicate standard error.

Results: On the DMC with RR=2 and RR=16 (top row in Figure 2) a clear hierarchy of interventions is observed: layer norm and right below it, spectral norm consistently outperforms others in mean return, irrespective of the plasticity regularization (x-axis). Notably, the combination of layer norm and resets in RR=16 (middle and bottom-right plot in Figure 2) demonstrates exceptional performance across all critic regularization variations on DMC Dog and without Dog environments. In contrast, the hierarchy of interventions on the MW benchmark (top row in Figure 2) diverges significantly from the DMC setup. Furthermore, a higher replay ratio introduces shifts in training dynamics, as evidenced by SN transitioning from a lower position on RR=2 (bottom-left plot in Figure 2) to nearly the most versatile intervention on RR=16 (bottom-right plot in Figure 2). The results generally align with first-order marginalization findings, emphasizing the positive impact of SN, as well as using many different types of regularization at once in general. Deeper analysis (see Appendix C.2) reveals that on MW, indeed, the gradient norm in a higher RR regime is orders of magnitude bigger. The finding that most contrasts with the first-order experiments, is that we observe that weight decay can actually yield significant performance benefits, under the condition that it is paired with other specific methods, namely full-parameter resets. In particular, we observe that

this combination yields synergies surpassing using any of these methods alone.

Takeaways:

- The DMC benchmark can be largely trivialized by using high RR agents combined with layer norm and full-parameter resets.
- Spectral normalization intervention ranks best for the Meta World benchmark, but it's not universally applicable. Whereas weight decay does not perform when used alone, it seems to have high synergy with full-parameter resets.
- Resetting the network significantly outperforms other plasticity-inducing interventions such as CRLU and SAM.

4.3. A closer look on Dog environment performance

Study description: In this study, we delve into the intricacies of two challenging Dog tasks, Dog-Trot and Dog-Run, included in our DMC setup. These environments present considerable difficulties for model-free approaches relying on proprioceptive states, making them of particular interest within the research community. Due to the inherent difficulty of these tasks, we conducted additional experiments using the top three methods identified in Figure 23 (Appendix) for 4 million steps, akin to approaches used in model-based (Hansen et al., 2022) or pixel-based studies (Ji et al., 2024). For the rest of the detailed experimental information, please refer to Appendix A.

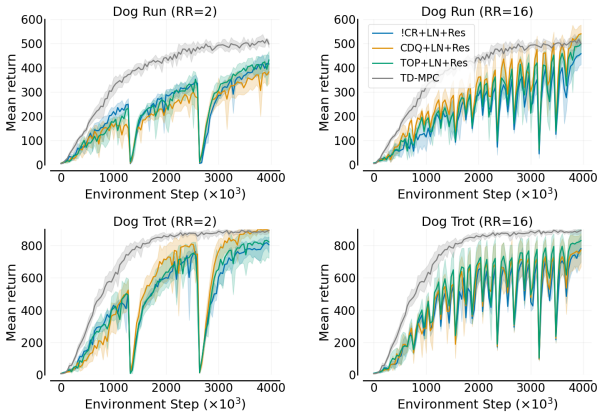

Figure 3. Mean return evolution across 4 million timesteps for Dog-Run (top row) and Dog-Trot (bottom row) environments. Gray plot depicts model-based agent performance. Each plot showcases the top three combinations.

Results: Specific intervention combinations effectively tackle the challenges posed by the Dog environment, as depicted in Figure 3. Analyzing the top three approaches for Dog-Run and Dog-Trot tasks reveals the prevalence of layer norm in nearly all combinations. Additionally, each critic regularization approach contributes to the leading group. Notably, in scenarios with high replay ratios, resets emerge as a crucial intervention. These observations are further substantiated by the analysis of second-order marginalization IQM plots (see 4). Indeed, layer norm without critic regularization excels in RR=2, and layer norm with resets outperforms all others convincingly. This achievement is particularly notable as, to our best knowledge, no model-free agent has previously find a better-performing policy within the training regime the Dog environments using proprioceptive states. Notably, there is a recent study(Ji et al., 2024) where a model-free agent achieved comparable results on the Dog environments but using pixel-based inputs instead. Additionally, our results demonstrate that while a model-based approach on proprioceptive states (Hansen et al., 2022) outperforms slightly, the above model-free approach with simple regularization techniques achieves performance very close to that of the model-based approach. This suggests the efficacy and competitiveness of our approach in challenging environments.

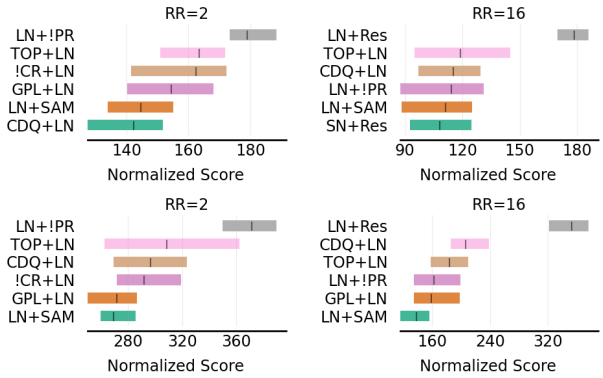

Figure 4. IQM performance of the top six intervention pair combinations based on 1 million steps experiments. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point. Results come from 1 million steps experiments. Top row: Dog-Run. Bottom row: Dog-Trot.

Takeaways:

- Well-established network regularization techniques such as layer norm and spectral norm enable finding high-performing policies for Dog-Trot and Dog-Run effectively.
- The choice of domain-specific RL critic regularization has little significance in dog environments when layer norm and resetting interventions are employed.

4.4. Correlation of Overestimation, Overfitting and Plasticity metrics with Performance

Study description: This study analyses the relationships between Overestimation, Overfitting, Plasticity, and model performance. Overestimation is quantified as an approximation error, overfitting as the ratio of TD error on the validation set to TD error on the training set. Expressing Plasticity loss is challenging, so we utilize proxy metrics, including the percentage of Dormant neurons (Sokar et al., 2023), representations rank (Kumar et al., 2020), gradient norm (Nikishin et al., 2022; Lyle et al., 2023), and parameters norm (Nikishin et al., 2022; Lyle et al., 2023).

We employ a Spearman correlation matrix to scrutinize these dependencies. This statistic is chosen because we observe non-linear yet monotonic dependencies between the mentioned metrics. We employ it on the data from all performed experiments, i.e., form runs with different combinations of interventions. Moreover, we do not have a division into RR=2 and RR=16, only the results from both setups are combined, and analyses are made on them. Overestimation, and gradient norm, and parameters norm are analyzed in a logarithmic scale for precision. We exclude metric pairs where the p-value of correlation is above 5% by whitening tiles in the correlation matrix.

Spearman correlation: In Figure 5, we investigate the relations between plasticity loss, overestimation and overfitting metrics and agents return, separately for every benchmark with special separation for Dog environments. Notably, overestimation exhibits the strongest correlation with agent returns on both benchmarks, offering insights into the findings of previous sections regarding the limited robustness of critic regularization methods designed to minimize overestimation. Interestingly, as shown in Figure 6, layer norm and spectral norm effectively mitigate approximation errors, outperforming CR methods. However, further investigation on the DMC benchmark (Figure 5) uncovers a strong negative correlation between the percentage of dormant neurons and performance, closely tied to the rank of representations on the critic's penultimate layer. What is more, the overestimation is not the best predictor for Dog environments where we observe the highest values of overestimation (Figure 5). We hypothesize that overestimation becomes a good predictor of performance only when more fundamental issues, such as plasticity, are mitigated, indicating a multifaceted learning problem in harder environments.

As an observation that supports this hypothesis, we refer to the critic gradient norm, which exhibits the most monotonic relation with the return in dog environments, as indicated by Spearman correlation (Figure 5). Analysing Figure 10 one can see, that Dog environments especially with high replay ratio experience exploding gradients. The high gradient

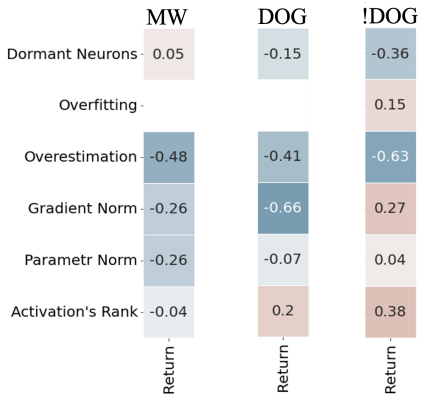

Figure 5. Explanatory metrics correlations for three different groups of environment, namely: MetaWorld, DMC Dog environments, and DMC environments without Dog environments. It's important to observe that not only does the main explanatory metric, gradient norm, vary for dog environments, but the remaining DMC environments also exhibit a different correlation sign for this metric.

norm directly points to high curvature of the loss landscape, which, as indicated by (Lee et al., 2023), describes low input plasticity. For this reason, layer norm primarily smoothens the activation distribution and plays a critical role in making SAC work in dog environments. Environments from the MW benchmark also encounter challenges with high curvature loss landscapes (as indicated by the Spearman correlation between gradient norm and return). This suggests a resemblance between MW and DMC dog environments.

Moreover, on the MW benchmark (Figure 5), the second-best correlated metrics with performance are the critic gradient norm and the critic parameters norm. This aligns with the results from section 4.2, highlighting the significant performance boost provided by spectral norm and weight decay, particularly in the RR=16 setup. An in-depth analysis of how RR increases the negative correlation of gradient norm and return can be found in sections C.2 and C.5 of Appendix.

Takeaways:

- There are distinctive correlations between plasticity loss, overestimation, overfitting metrics, and agent returns in various benchmark suites. These underscore the importance of considering environment-specific factors when assessing model performance and designing effective regularization strategies.

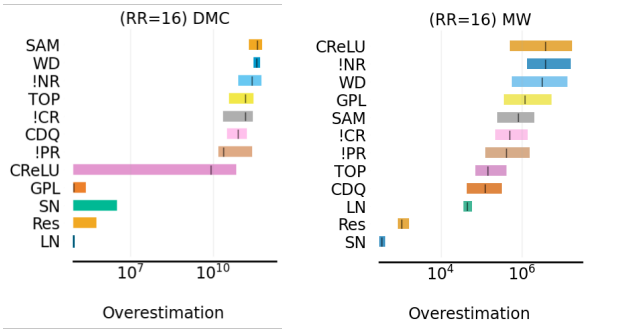

Figure 6. IQM overestimation in logarithmic scale. Each plot presents sorted results. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point. Left Figure: DMC benchmark. Right Figure: MW benchmark. Colours indicate hierarchy on the plot, not specific names.

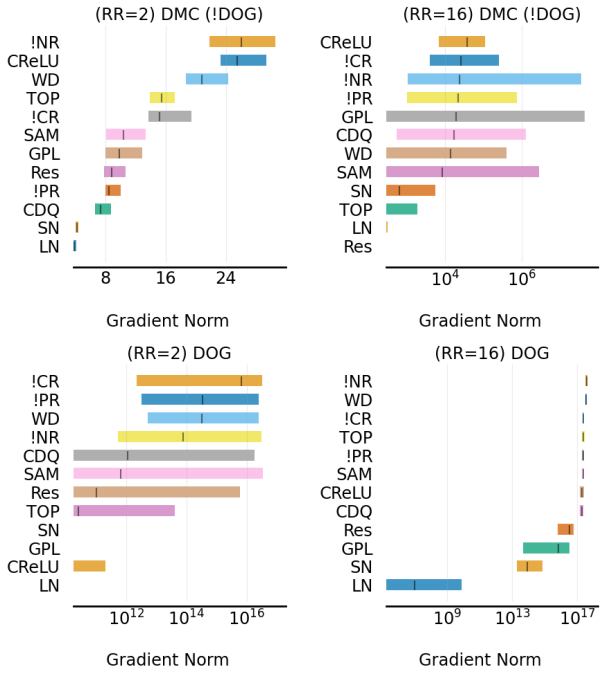

Figure 7. IQM gradient norm of first-order results. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point.

- Techniques like layer or spectral norm and resets are particularly effective in mitigating overestimation also compared to methods specifically designed for that purpose.
- The negative correlation of the critic's gradient norm and return becomes more apparent in challenging environments and mainly in a high replay ratio regime.

5. Related Works

The literature on deep reinforcement learning has long explored various factors contributing to performance challenges, approaching the issue from different perspectives. A notable study, akin to our pragmatic approach, investigates the impact of diverse design choices in the training process of on-policy methods (Andrychowicz et al., 2021).

In the realm of off-policy methods, numerous hypotheses have been proposed regarding crucial factors. One key focus is on addressing the overestimation problem, with attempts to harness its potential benefits (Ciosek et al., 2019) and, more prominently, to mitigate the phenomenon by introducing novel loss functions (Fujimoto et al., 2018; Moskovitz et al., 2021; Cetin & Celiktutan, 2023).

Regularization schemes have proven effective in enhancing deep reinforcement learning methods. Neural network regularizations, such as Spectral Norm (Gogianu et al., 2021; Bjorck et al., 2021), Layer Norm (Bjorck et al., 2021; Ball et al., 2023), or weight decay (Liu et al., 2020), have yielded significant improvements in results. Notably, periodic resets of critic weights proposed by (Nikishin et al., 2022) constitute a strong baseline in robotics control. Several regularization schemes have been proposed to address the issue of discarding knowledge caused by fully resetting the critic network. Of particular interest is Shrink and Perturb (Ash & Adams, 2020) and L2 Init (Kumar et al., 2023). In both of these methods the benefit comes from the regularization towards the distribution of "freshly" initialized weights. A concurrent work finds that using unit-ball normalization allows for learning with a high-replay ratio without full-parameter resets (Hussing et al., 2024). An ensemble approach was also suggested to address the challenge of determining the optimal number of gradients per environment step, where decisions are based on the validation TD error (Li et al., 2022).

6. Limitations

In this work, we found crucial choices that drive SAC effectiveness in a wide range of control tasks from two popular benchmarks. Through extensive experiments, we uncovered perplexities concerning explanatory metrics correlations and complex dynamics of overestimation, which is successfully mitigated by widely used regularizations. Nevertheless, our study has certain constraints. Our empirical evaluations were limited to proprioceptive tasks on DMC and Meta-World benchmarks and only for SAC method.

7. Conclusions

This study explored different common RL design choices, as well as interactions thereof, evaluating their impact on

agent learning. Specifically, we consider three types of regularization families: critic regularization (motivated by value overestimation); network regularization (motivated by model overfitting); and plasticity regularization (motivated by plasticity loss). Our analysis revealed that generic network regularization methods such as layer normalization, especially when paired with full-parameter resets, can have a vastly greater impact on the final performance than domain-specific RL approaches. To this end, the same network regularization methods can lead to strong performing policies on domains previously solved by model-based agents, such as the dog domain. Furthermore, we studied a variety of metrics that were shown to co-occur with the deterioration of learning in low and high replay regimes. Our analysis revealed the complex interactions of the considered metrics with agent performance. Surprisingly, we found that interventions motivated by a specific problem, for example, overfitting, can have a pronounced impact on the metrics associated with overestimation or plasticity. Finally, we found that the effectiveness of considered critic, network, and plasticity regularization techniques is not only highly dependent on the simulation benchmark but also type of simulated task. The most prominent example is Clipped Double Q-learning, a technique used in a majority of modern actor-critic algorithms, which is effective in DMC locomotion tasks, but leads to significant performance deterioration on the MetaWorld manipulation tasks. To this end, we highlighted the need to test new algorithms on a diverse set of tasks, preferably stemming from more than one suite.

Acknowledgements

Marek Cygan was partially supported by an NCBiR grant POIR.01.01.01-00-0433/20. Mateusz Ostaszewski was funded by the National Science Center Poland under the grant agreement 2020/39/B/ST6/01511. Michał Bortkiewicz was funded by the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. This research was also supported by National Science Centre, Poland grant no 2020/39/B/ST6/01511 and grant no 2022/45/B/ST6/02817. This work was partially funded by the European Union under the Horizon Europe grant OMINO (grant number 101086321) and Horizon Europe Program (HORIZON-CL4-2022-HUMAN-02) under the project "ELIAS: European Lighthouse of AI for Sustainability", GA no. 101120237. We also gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016783.

Impact Statement

This paper focuses on the issue of interplay between different design choices within Reinforcement Learning (RL) algorithms. While the successful application of RL has the potential to influence society in many ways, our work, focused primarily on a technical advancement in RL algorithms, does not introduce novel ethical considerations beyond those already inherent in the broader field of RL.

References

Abbas, Z., Zhao, R., Modayil, J., White, A., and Machado, M. C. Loss of plasticity in continual deep reinforcement learning. *arXiv preprint arXiv: 2303.07507*, 2023.

Achille, A., Rovere, M., and Soatto, S. Critical learning periods in deep neural networks. *arXiv preprint arXiv: 1711.08856*, 2017.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. What matters in on-policy reinforcement learning? a large-scale empirical study. In *ICLR 2021-Ninth International Conference on Learning Representations*, 2021.

Ash, J. and Adams, R. P. On warm-starting neural network training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3884–3894. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/288cd2567953f06e460a33951f55daaf-Paper.pdf.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Ball, P. J., Smith, L., Kostrikov, I., and Levine, S. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv: 2302.02948*, 2023.

Bjorck, J., Gomes, C. P., and Weinberger, K. Q. Towards deeper deep reinforcement learning with spectral normalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8242–8255, 2021. URL https://proceedings.

neurips.cc/paper/2021/hash/4588e674d3f0faf985047d4c3f13ed0d-Abstract.html.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2018.

Cetin, E. and Celiktutan, O. Learning pessimism for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6971–6979, 2023.

Chen, X., Wang, C., Zhou, Z., and Ross, K. W. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2020.

Ciosek, K., Vuong, Q., Loftin, R., and Hofmann, K. Better exploration with optimistic actor critic. *Advances in Neural Information Processing Systems*, 32, 2019.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., Noury, S., Pesamosca, F., Pfau, D., Sauter, O., Sommariva, C., Coda, S., Duval, B., Fasoli, A., Kohli, P., Kavukcuoglu, K., Hassabis, D., and Riedmiller, M. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602 (7897):414—419, February 2022. ISSN 0028-0836. doi: 10.1038/s41586-021-04301-9.

Dohare, S., Sutton, R. S., and Mahmood, A. R. Continual backprop: Stochastic gradient descent with persistent randomness. *arXiv preprint arXiv: 2108.06325*, 2021.

D'Oro, P., Schwarzer, M., Nikishin, E., Bacon, P.-L., Bellemare, M. G., and Courville, A. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations*, 2022.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv: 2010.01412*, 2020.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Gogianu, F., Berariu, T., Rosca, M. C., Clopath, C., Busoniu, L., and Pascanu, R. Spectral normalisation for deep reinforcement learning: an optimisation perspective. In *International Conference on Machine Learning*, pp. 3734–3744. PMLR, 2021.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hansen, N., Wang, X., and Su, H. Temporal difference learning for model predictive control. In *International Conference on Machine Learning, PMLR*, 2022.

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., and Tsuruoka, Y. Dropout q-functions for doubly efficient reinforcement learning. In *International Conference on Learning Representations*, 2021.

Hussing, M., Voelcker, C., Gilitschenski, I., Farahmand, A.-m., and Eaton, E. Dissecting deep rl with high update ratios: Combatting value overestimation and divergence. *arXiv preprint arXiv:2403.05996*, 2024.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Ji, T., Luo, Y., Sun, F., Zhan, X., Zhang, J., and Xu, H. Seizing serendipity: Exploiting the value of past success in off-policy actor-critic, 2024.

Kumar, A., Agarwal, R., Ghosh, D., and Levine, S. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. *arXiv preprint arXiv: 2010.14498*, 2020.

Kumar, S., Marklund, H., and Roy, B. V. Maintaining plasticity in continual learning via regenerative regularization. *arXiv preprint arXiv: 2308.11958*, 2023.

Lee, H., Cho, H., Kim, H., Gwak, D., Kim, J., Choo, J., Yun, S.-Y., and Yun, C. Plastic: Improving input and label plasticity for sample efficient reinforcement learning. *NEURIPS*, 2023.

Li, Q., Kumar, A., Kostrikov, I., and Levine, S. Efficient deep reinforcement learning requires regulating overfitting. In *The Eleventh International Conference on Learning Representations*, 2022.

Liu, Z., Li, X., Kang, B., and Darrell, T. Regularization matters in policy optimization-an empirical study on continuous control. In *International Conference on Learning Representations*, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2017.

Lyle, C., Rowland, M., and Dabney, W. Understanding and preventing capacity loss in reinforcement learning. *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2204.09560.

Lyle, C., Zheng, Z., Nikishin, E., Avila Pires, B., Pascanu, R., and Dabney, W. Understanding plasticity in neural networks. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 23-29 Jul 2023.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Moskovitz, T., Parker-Holder, J., Pacchiano, A., Arbel, M., and Jordan, M. Tactical optimism and pessimism for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12849–12863, 2021.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., and Courville, A. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pp. 16828–16847. PMLR, 2022.

Nikishin, E., Oh, J., Ostrovski, G., Lyle, C., Pascanu, R., Dabney, W., and Barreto, A. Deep reinforcement learning with plasticity injection. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. URL https://openreview.net/forum?id=O9cJADBZT1.

OpenAI, :, Berner, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv: 1912.06680*, 2019.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Schwarzer, M., Ceron, J. S. O., Courville, A., Bellemare, M. G., Agarwal, R., and Castro, P. S. Bigger, better, faster: Human-level atari with human-level efficiency. In *International Conference on Machine Learning*, pp. 30365–30380. PMLR, 2023.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Smith, L., Kostrikov, I., and Levine, S. A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. *arXiv preprint arXiv:2208.07860*, 2022.

Sokar, G., Agarwal, R., Castro, P. S., and Evci, U. The dormant neuron phenomenon in deep reinforcement learning. *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2302.12902.

Sutton, R. The bitter lesson. *Incomplete Ideas (blog)*, 13(1), 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Thrun, S. and Schwartz, A. Issues in using function approximation for reinforcement learning. In *Proceedings of the 1993 connectionist models summer school*, pp. 255–263. Psychology Press, 2014.

Yarats, D., Kostrikov, I., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*, 2020.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Kaelbling, L. P., Kragic, D., and Sugiura, K. (eds.), *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 2019. URL http://proceedings.mlr.press/v100/yu20a.html.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. *arXiv preprint arXiv: 1805.08318*, 2018.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A. Details of experiments

Results reporting Interquartile mean (IQM) are based on 500 bootstrapping points as calculated by *rliable* package (Agarwal et al., 2021). The final performance is defined as the average of the last 10 policy evaluations.

B. Architecture details

In all experiments, the Actor and Critic are represented by three-layer MLP networks, each containing 256 neurons in the hidden layers, utilizing the ReLU activation function (except for the CReLU variant, which effectively doubles the number of activations).

In the scenario involving Layer Norm, it is applied to each hidden layer (Li et al., 2022; Ball et al., 2023), while the spectral norm is applied exclusively to the last hidden layer (Gogianu et al., 2021; Li et al., 2022). Weight decay is uniformly applied across all layers (Li et al., 2022). It's important to note that all network regularizations are exclusively applied to the Critic network.

B.1. Hyperparameters

All hyperparameters are taken from original papers introducing the given intervention.

Table 1. Hyperparameter values used in the experiments.

HYPERPARAMETER	NOTATION	VALUE
JOINT		
NETWORK SIZE	NA	$(256, 256)$
ACTION REPEAT	NA	1
OPTIMIZER	NA	ADAM
LEARNING RATE	NA	$3e-4$
BATCH SIZE	B	256
DISCOUNT	γ	0.99
INITIAL TEMPERATURE	α_0	1.0
INITIAL STEPS	NA	10000
TARGET ENTROPY	$\mathcal{H}^*$	$\lvert\mathcal{A}\rvert/2$
POLYAK WEIGHT	τ	0.005
TOP		
PESSIMISM VALUES	β	$\{0, 1\}$
BANDIT LEARNING RATE	NA	0.1
GPL		
PESSIMISM LEARNING RATE	NA	$1e-5$

C. Further Experiments

C.1. Third-order marginalization

Examining the plots without marginalization (Figure 8) provides further insights into the conclusions drawn from the previous experiment. Specifically, most combinations without critic regularization (red points) consistently perform well across all setups. Additionally, the results for GPL (pink points) affirm the overall subpar performance of this method. On the DMC with RR=2 plot (top-left in Figure 8), layer norm (points labeled "L") consistently outperforms others in mean return, irrespective of the Plasticity regularization (x-axis) or Critic Regularization (color). Notably, the combination of layer norm and resets in RR=16 (top-right plot in Figure 8) demonstrates exceptional performance across all critic regularization variations.

For the MW benchmark with RR=16 (bottom-right plot in Figure 8), the results align with first-order marginalization findings, highlighting the positive impact of Spectral Normalization. Particularly interesting is the role of Weight Decay, forming

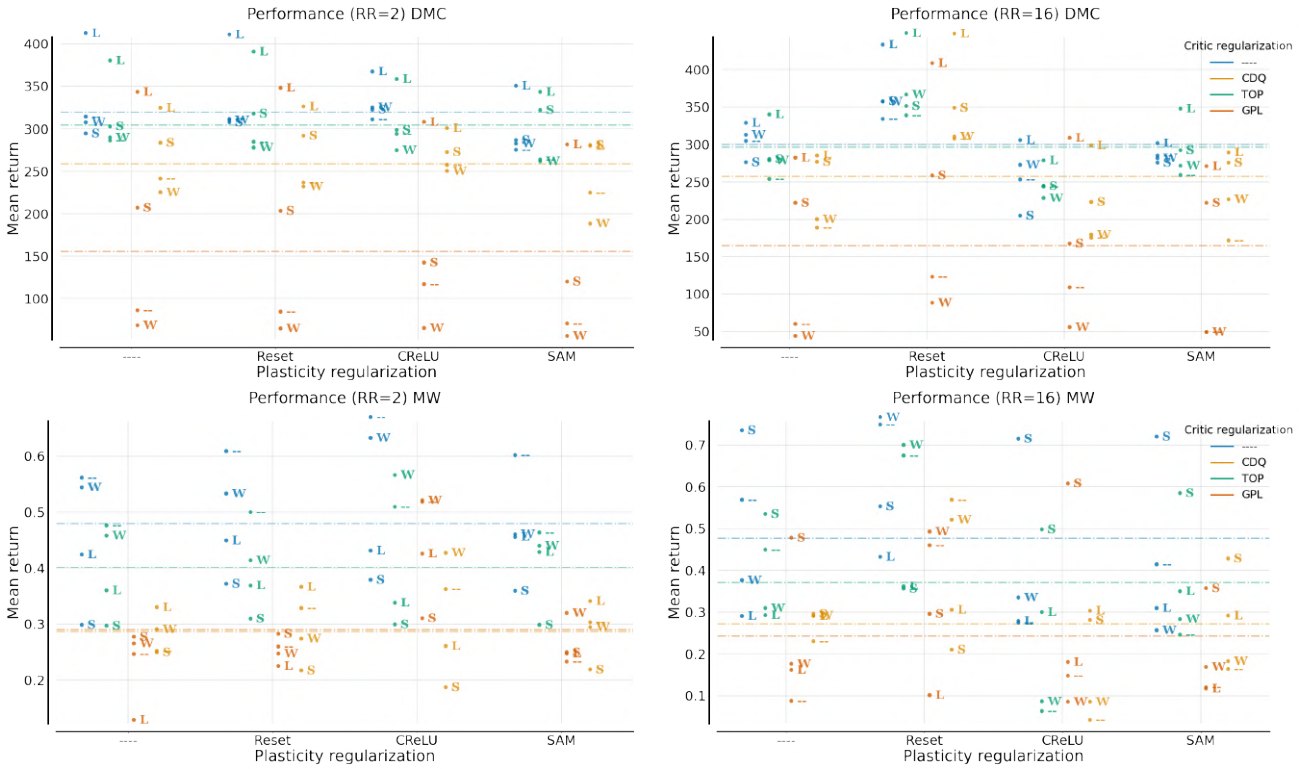

Figure 8. DMC (top) MW (bottom).

optimal combinations with the Reset method on RR=16 and consistently performing well across various combinations on RR=2 (bottom-left plot in Figure 8).

In a scenario with a high replay ratio, the resets are the most important intervention. They work best in combination with Layer norm, but other neural network regularization methods combined with resets are at the forefront regarding performance in the rr=16 scenario (see the top of Figure 8). We present more results in Figures 22 and 23.

C.2. Gradient Norm Analysis

Drawing insights from the findings in section 4.2, where we highlighted the significance of spectral norm in enhancing agent performance on the MW benchmark, we now delve deeper into the behavior of the critic's gradient norm. In Figure 9, one can compare orders of magnitude of IQM of gradient norm with respect to different replay ratio (RR) regimes and on different benchmarks. Referring to results from section 4.2, Figure 9 underscores that the gradient norm on MetaWorld, particularly in the RR=16 setup, exhibits orders of magnitude higher values compared to the DMC benchmark.

A similar phenomenon can be observed on the DMC benchmark, but the layer norm proves more robust in mitigating exploding gradients than the Spectral Norm. Interestingly, training on very complex environments such as Dog causes enormous gradient explosion, even in a small replay ratio regime 10.

C.3. Comparison of ReDO to other plasticity-inducing methods

The ReDO method, as proposed by Sokar et al. (Sokar et al., 2023), is a technique for inducing network plasticity. It shares similarities with the full reset approach but involves more targeted interventions within the network. In particular, ReDo does not reset the full network; it only resets weights connected to dormant neurons. Specifically, incoming weights to dormant neurons are initialized as in full reset, but outgoing weights from dormant neurons are zeroed out, resulting in less severe network output changes.

Figure 11 presents results from Figure2 updated with the ReDo method for both MetaWorld and DMC environments for

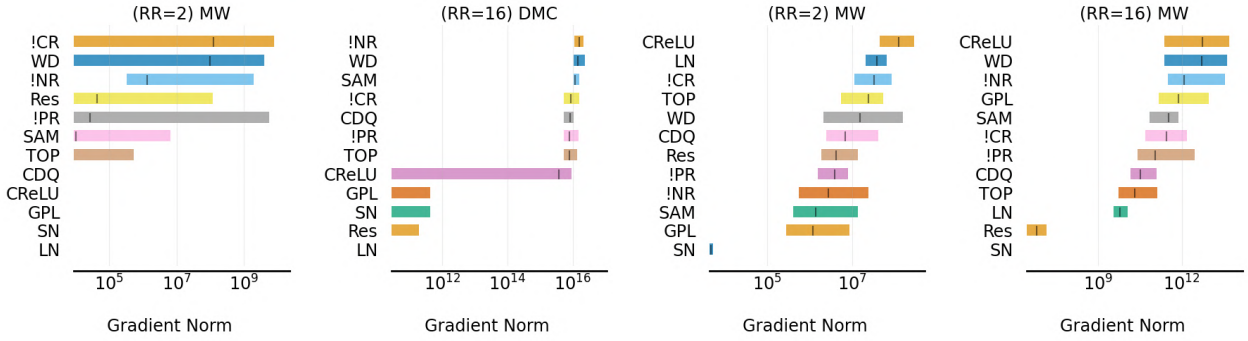

Figure 9. IQM gradient norm of first-order results. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point.

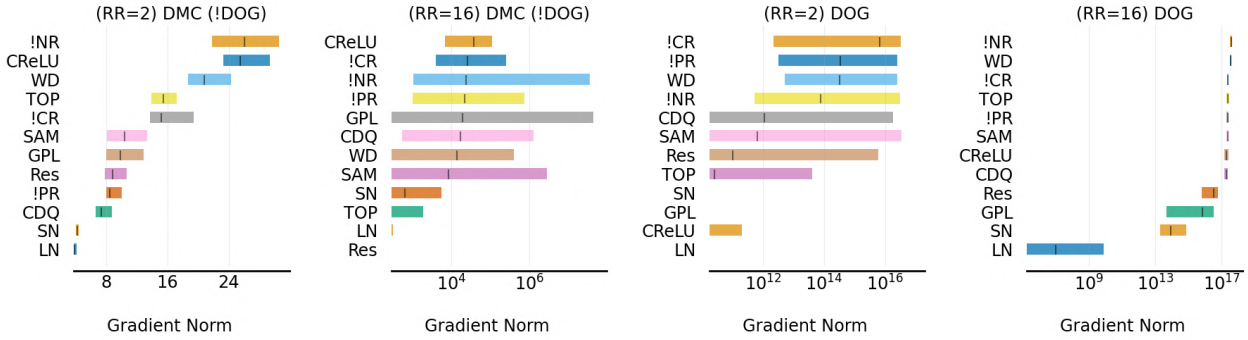

Figure 10. IQM gradient norm of first-order results. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point.

a high replay ratio regime. ReDo does not perform as well as resets in the Metaworld and DMC suites. However, both methods effectively reduce the critic gradient norm, overestimation and the number of dormant neurons for both Metaworld and DMC without dog benchmarks, as shown in Figure 12. In dog environments, we observed that ReDo was unstable for runs without SN or LN, and some runs crashed. We report results for the last ten timesteps before the crash for these runs. Interestingly, all methods except SN and LN cannot reduce the critic gradient norm, as shown in the bottom right plot in Figure 12.

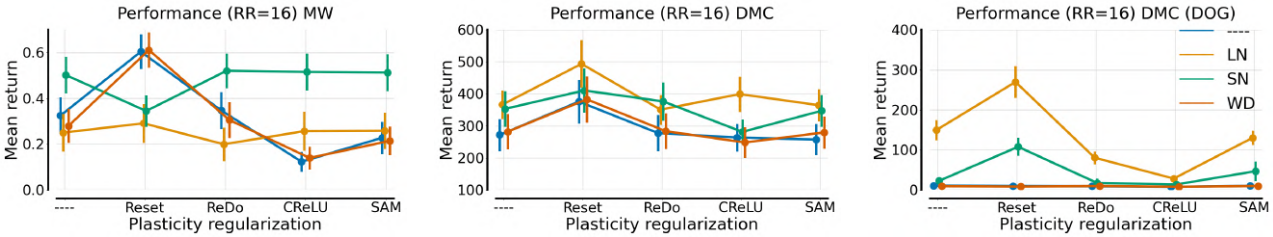

Figure 11. Second-order results marginalizing critic regularization methods with ReDO.

C.4. Closer look on CDQL performance on Hopper and Quadruped environments

Figure 13 illustrates the outcomes of applying Clipped Double Q-learning (CDQL) (Haarnoja et al., 2018; Fujimoto et al., 2018; Hansen et al., 2022), a widely used and straightforward critic regularization method, across various environments. In the case of the hopper hop environment or some MW environments (bottom row), CDQL exhibits a detrimental effect on performance, with additional regularization techniques such as resets or layer normalization failing to alleviate this effect. Similarly, in the quadruped run task, CDQL demonstrates a comparable negative impact, although the application of

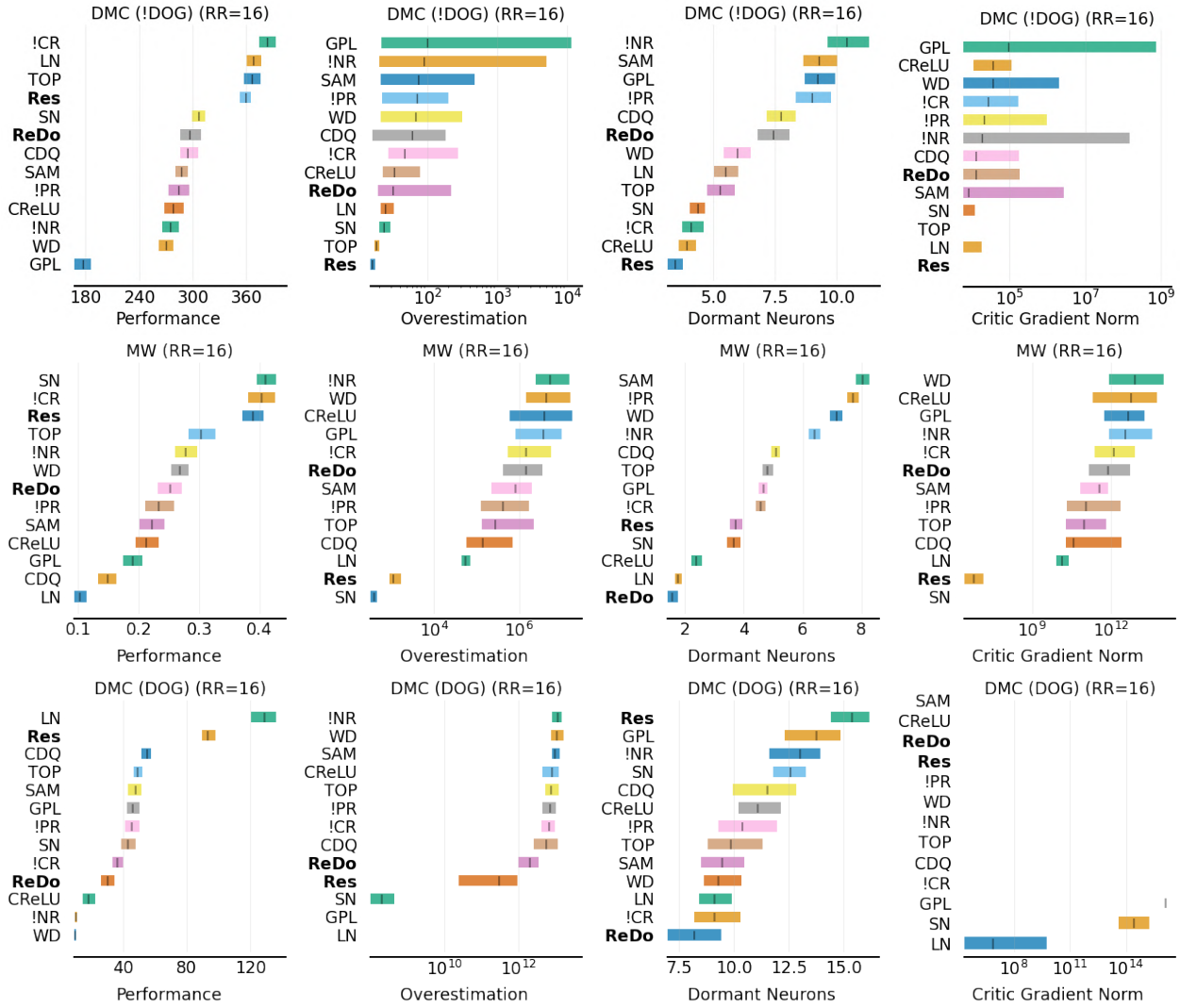

Figure 12. IQM of performance, overestimation, dormant neurons, and gradient norm, of first-order results. The IQM is calculated based on the average of the last ten evaluation points in each run, not the last evaluation point.

additional regularization methods successfully mitigates it.

C.5. Regression and Spearman correlation analysis

Deepening analysis from section 4.4, we investigate the coefficients of Ordinary Least Square regression for overestimation, overfitting, dormant neurons and critic's gradient norm with return. Except for dormant neurons, we take the symmetrical logarithm of every metric. We perform regression fitting for every variable separately, which results in four different regression analyses. As previously stated, because these relations are highly non-linear, however, often monotonic, goodness of fit coefficients such as R^2 for linear regression are low. Besides regression coefficients (β), we report a standard error of regression (se), R^2, p-value p in the plot legend. In the axis titles, the Spearman correlation and its p-value are shown.

In figures 14 and 15, we observe consistent negative coefficients, except for the Coffee environment for overestimation and return. Interestingly, for the MetaWorld benchmark, most environments exhibit a symmetrical degree of overestimation and underestimation as two distinct clusters of well-performing runs are mirrored by the Y axis. For the DMC environments, we observe a much higher amount of runs in which overestimation is present.

For the gradient norm, we observe a higher value of this metric for runs with RR=16 both for DMC and MW benchmarks

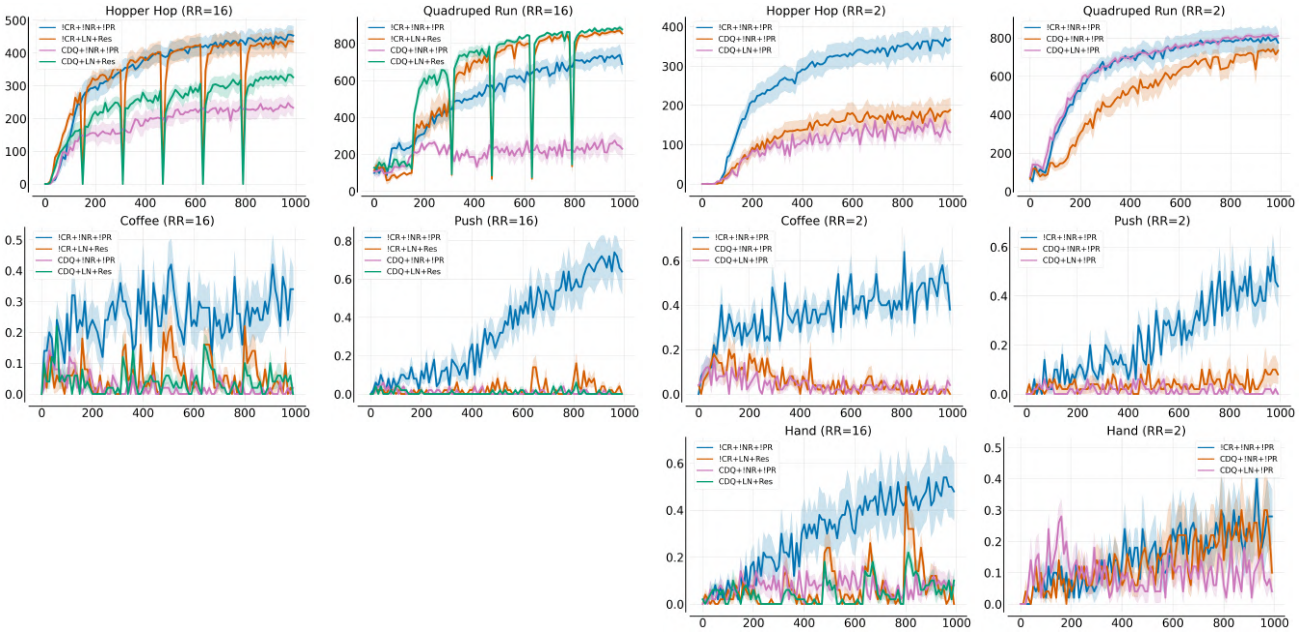

Figure 13. Examining the influence of CDQL on performance in the Hopper Hop and Quadruped Run environments within the DMC benchmark (top row), as well as the Push, Coffee, and Hand environments within the MW benchmark (bottom row).

(Figures 16 and 17. There is consistency in coefficients for MW; however, in DMC, we observe that only four out of 7 environments exhibit negative signs of the coefficient. However, for most of the extremely high values of gradient norm, we systematically observe low returns regardless of the environment.

Dormant neurons, presented in Figures 18 and 19, clearly correlate negatively with the return, especially for DMC environments. However, there are cases, such as the Reach environment from MetaWorld, where a high percentage of dormant neurons benefit the agent, probably because this particular environment is especially easy.

Overfitting is the metric that exhibits the highest variance in coefficient signs across environments and benchmarks. As shown in Figures 20 and 21, the best-performing runs are located for low absolute values of overfitting. In addition, RR=16 clearly results in higher overfitting values.

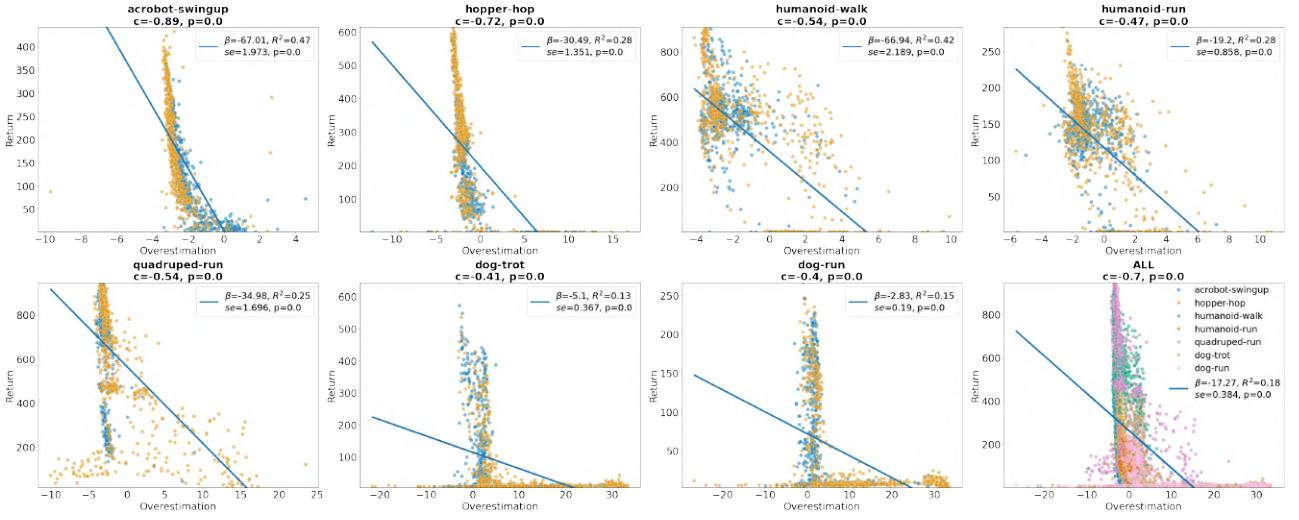

Figure 14. Overestimation logarithm scatter plots with regression line for DMC environments.

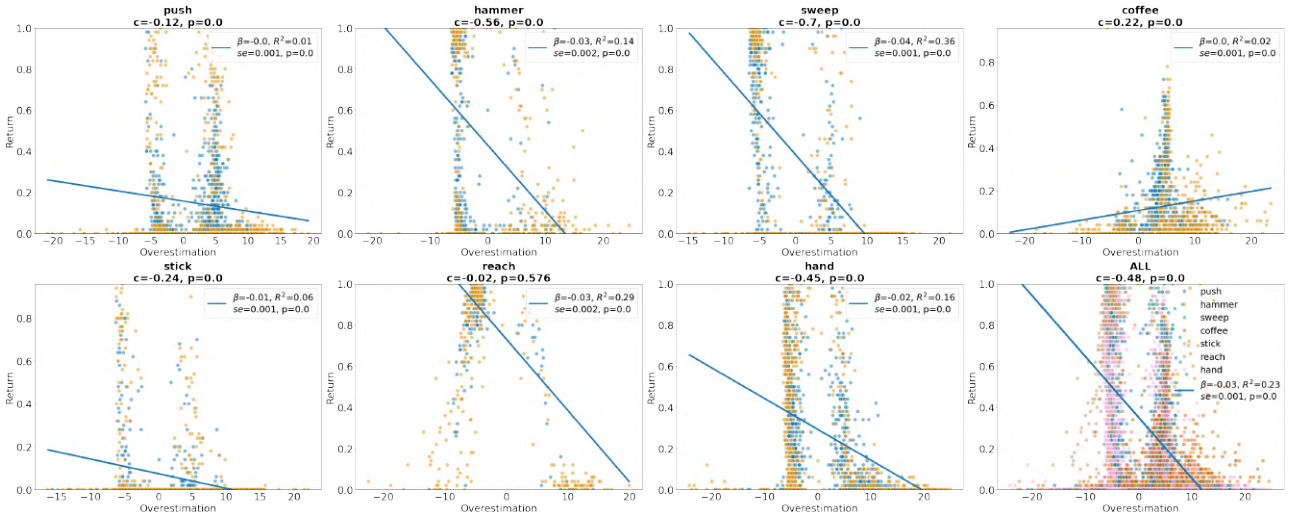

Figure 15. Overestimation logarithm scatter plots with regression line for MW environments.

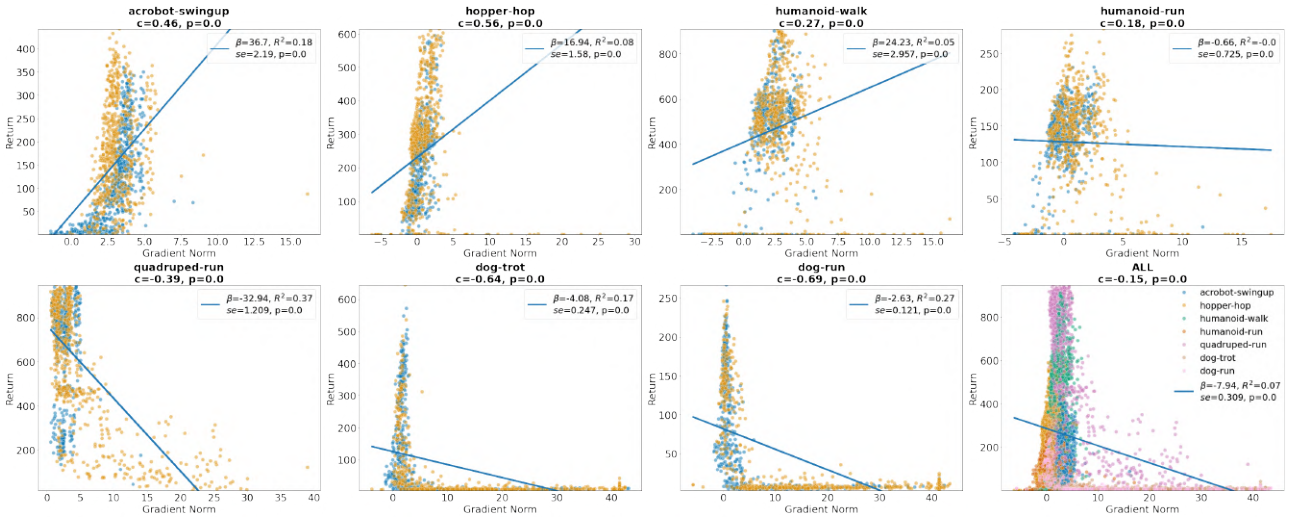

Figure 16. Gradient norm logarithm scatter plots with regression line for DMC environments.

C.6. Image-based DeepMind Control

We test whether the results achieved on proprioceptive state representation transfer to image-based control. To this end, we run 4 versions of the DrQ agent (Yarats et al., 2020):

1. Vanilla DrQ (DrQ)
2. DrQ with layer normalization on the critic network (DrQ + LN)
3. DrQ with full-parameter resets every 200k environment steps (DrQ + Res)
4. DrQ with both normalization on the critic network and full-parameter resets every 200k environment steps (DrQ + LN + Res)

We run these variations on 6 tasks from the DeepMind Control benchmark: Acrobot Swingup, Cheetah Run, Hopper Hop, Humanoid Run, Humanoid Stand, and Humanoid Walk. We run the humanoid tasks for 3mln frames and the other tasks for

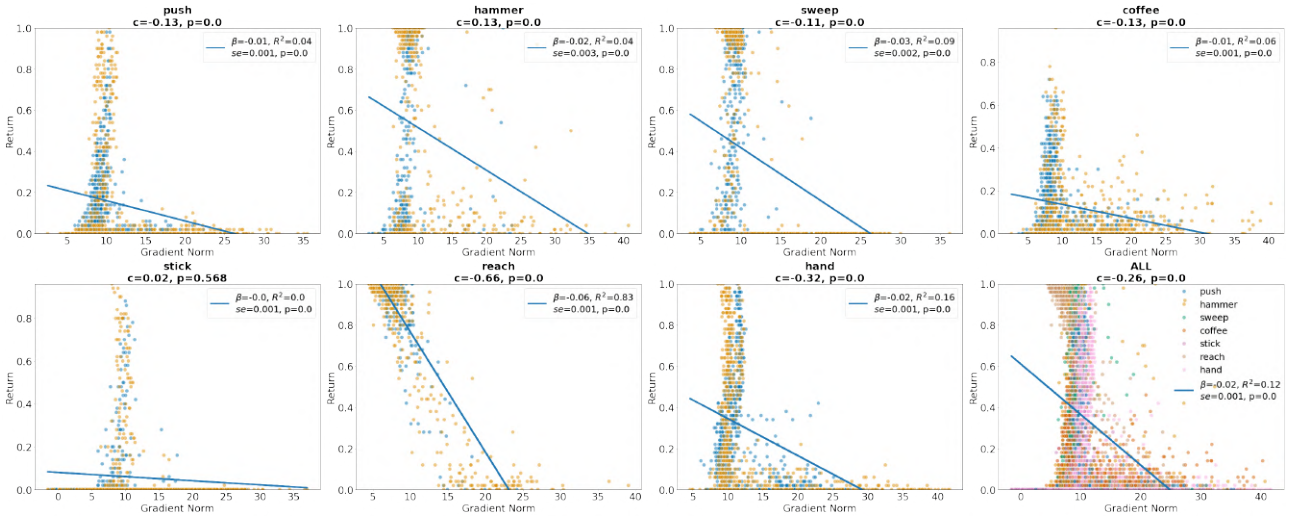

Figure 17. Gradient norm logarithm scatter plots with regression line for MW environments.

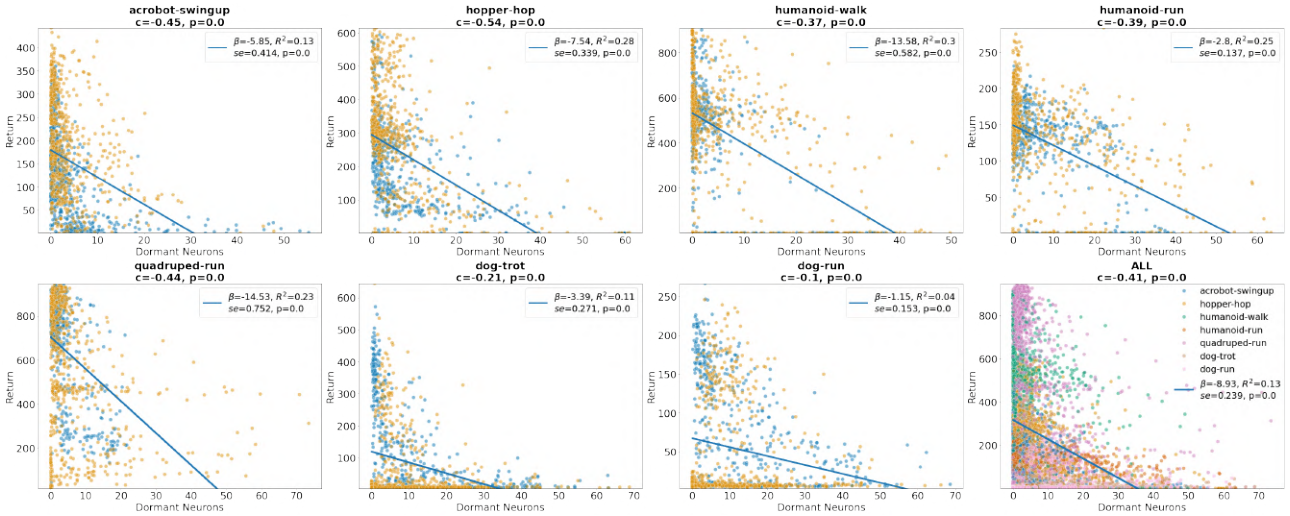

Figure 18. Dormant neurons scatter plots with regression line for DMC environments.

1mln frames with a replay ratio of 1. We calculate the relationship between frames and environment steps according to the methodology presented in Yarats et al. (2020). We present the results in the table below.

Unfortunately, the humanoid agents were mostly unable to achieve non-random policies in the budget of 3mln frames. Interestingly, the proprioceptive results do not seem to directly transfer to image-based agents with a low-replay ratio. As such, we believe the image-based benchmark requires further studies.

C.7. Best combinations of intervention performance plots

C.8. Other

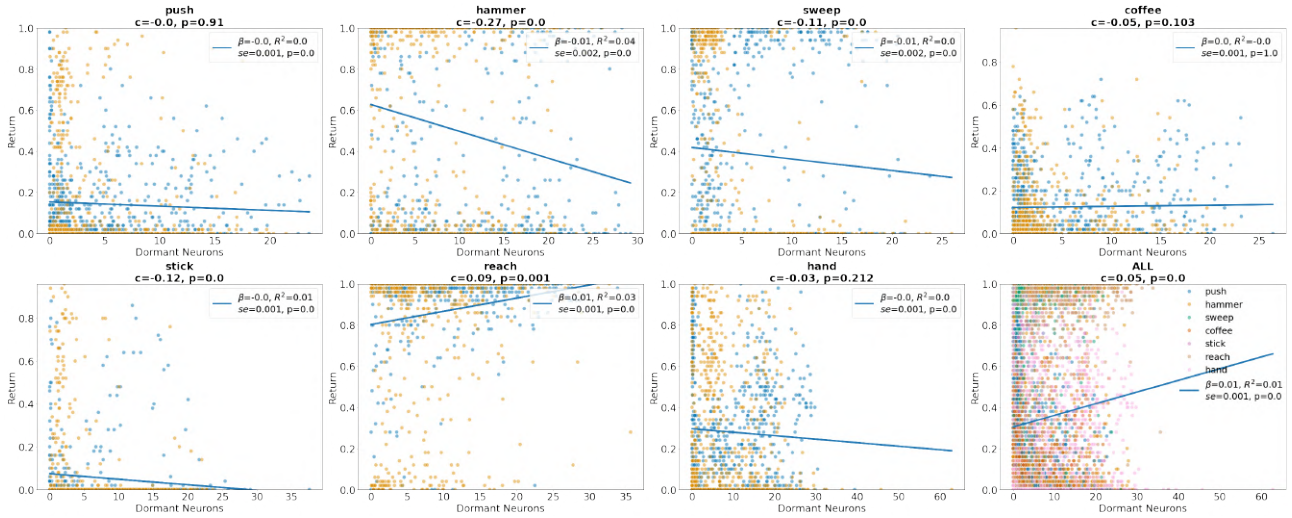

Figure 19. Dormant neurons scatter plots with regression line for MW environments.

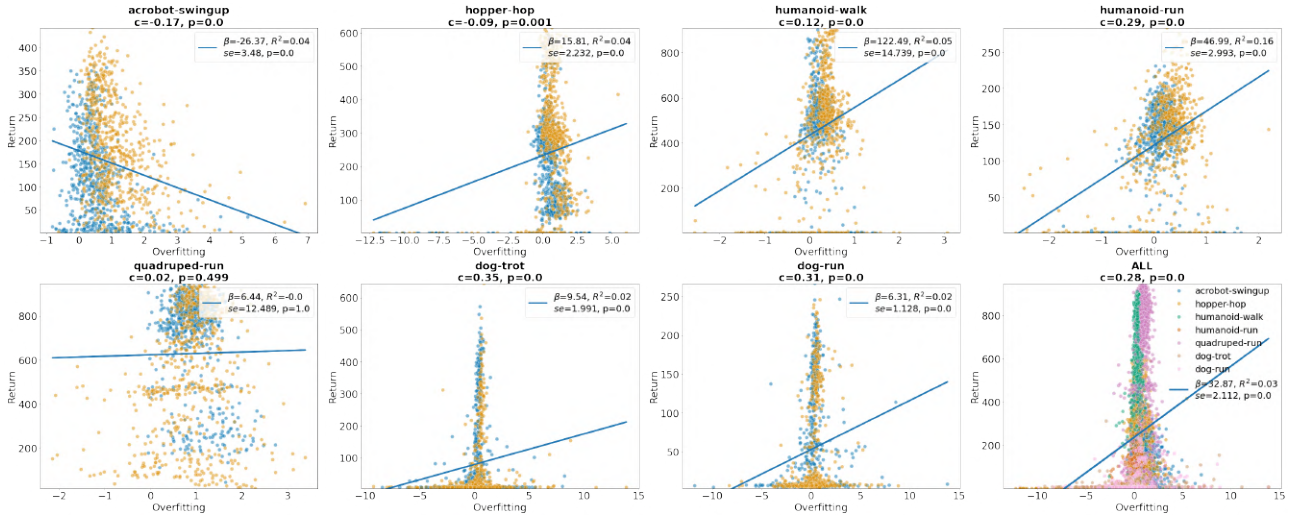

Figure 20. Overfitting logarithm scatter plots with regression line for DMC environments.

Table 2. Final performance in image-based environments. 3 seeds per task.

	DrQ	**DrQ + LN**	**DrQ + Res**	**DrQ + LN + Res**
Acrobot Swingup	172.9 ± 22.1	87.6 ± 19.7	52.3 ± 11.2	88.8 ± 8.2
Cheetah Run	727.4 ± 7.3	680.0 ± 2.3	715.6 ± 7.1	653.9 ± 2.1
Hopper Hop	74.6 ± 34.9	116.3 ± 27.2	135.7 ± 24.1	167.3 ± 14.3
Humanoid Stand	7.8 ± 0.2	8.1 ± 0.2	7.5 ± 0.4	7.8 ± 0.4

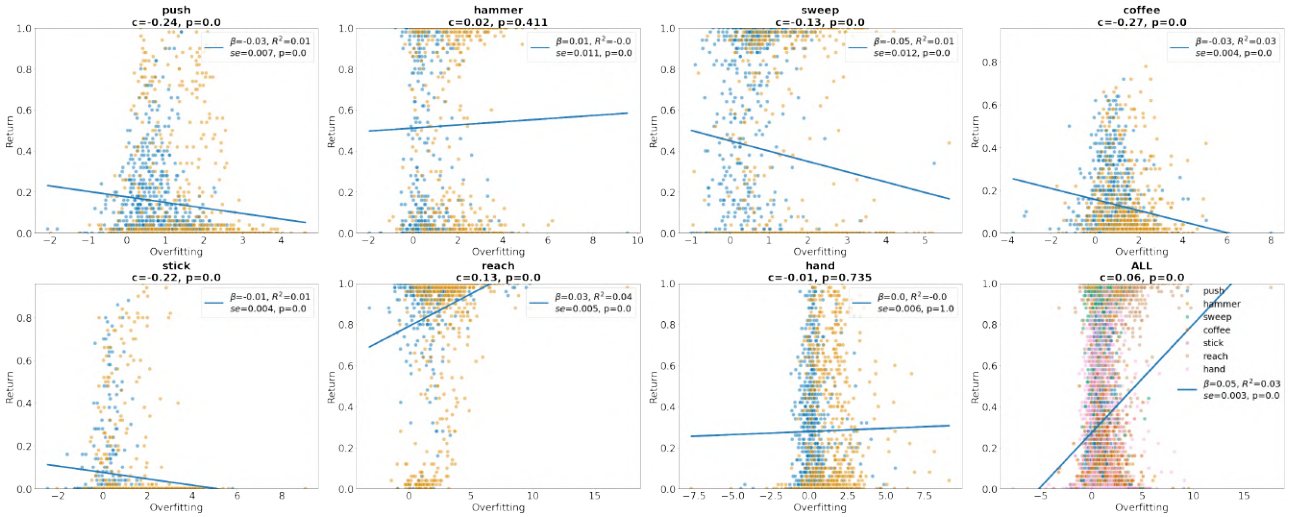

Figure 21. Overfitting logarithm scatter plots with regression line for MW environments.

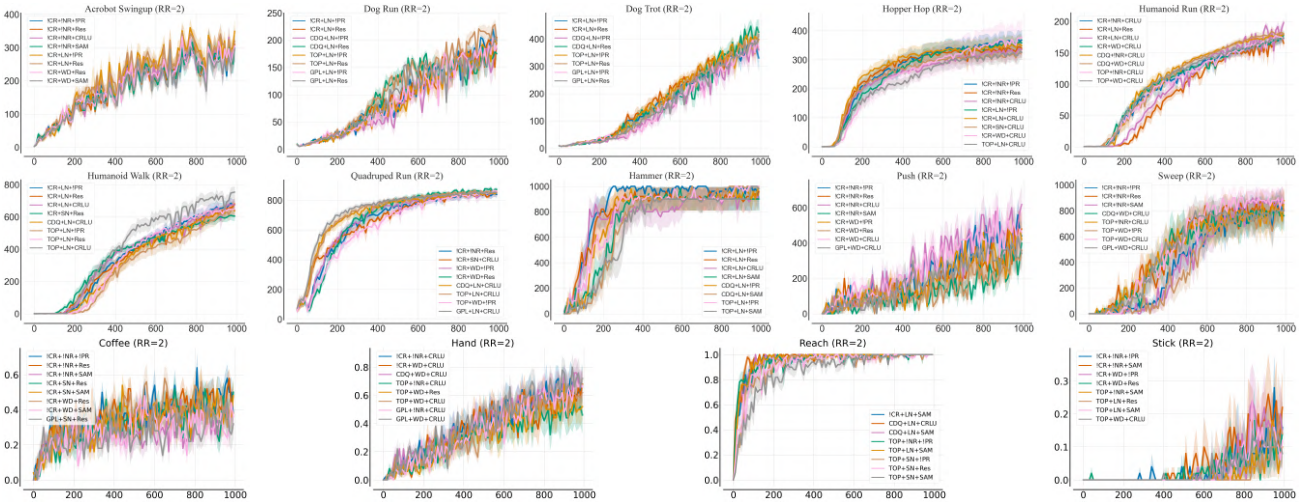

Figure 22. Top performing configuration in the low replay regime. 10 seeds per task per algorithm.

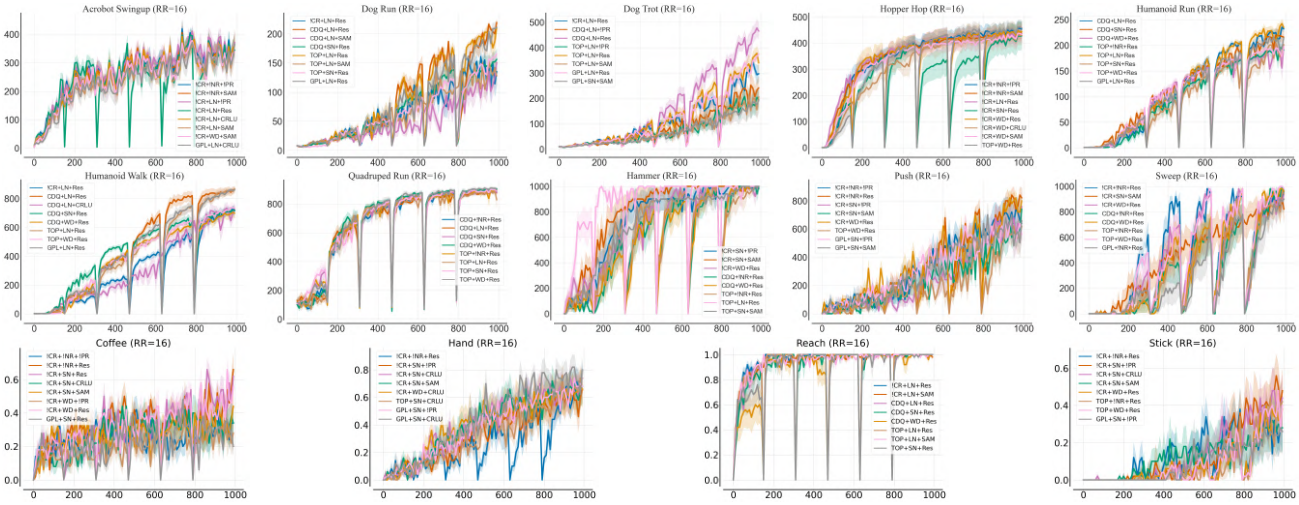

Figure 23. Top performing configuration in the high replay regime. 10 seeds per task per algorithm.

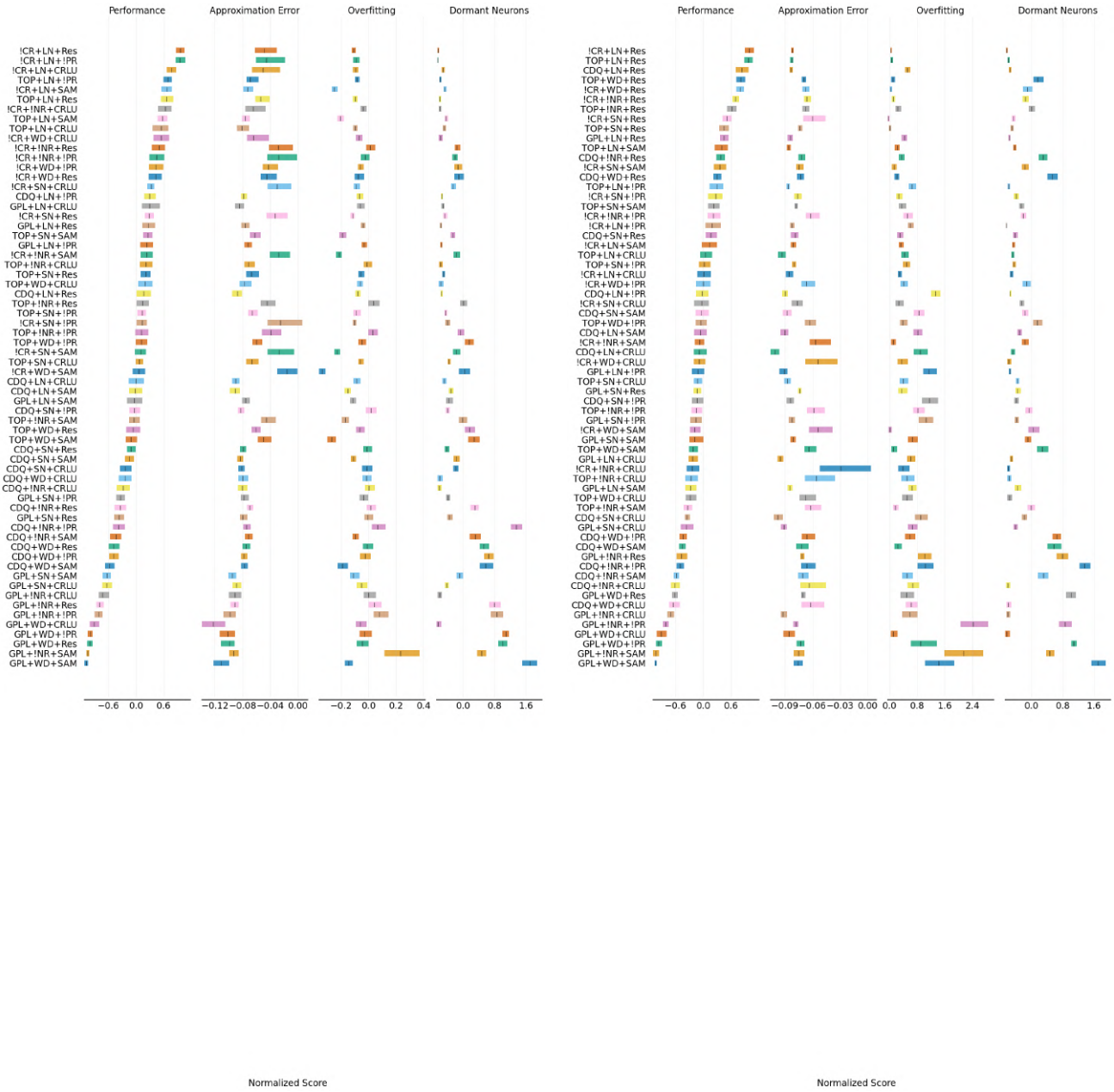

Figure 24. Top performing configuration in the low (left) and high (right) replay regime. 10 seeds per task per algorithm.

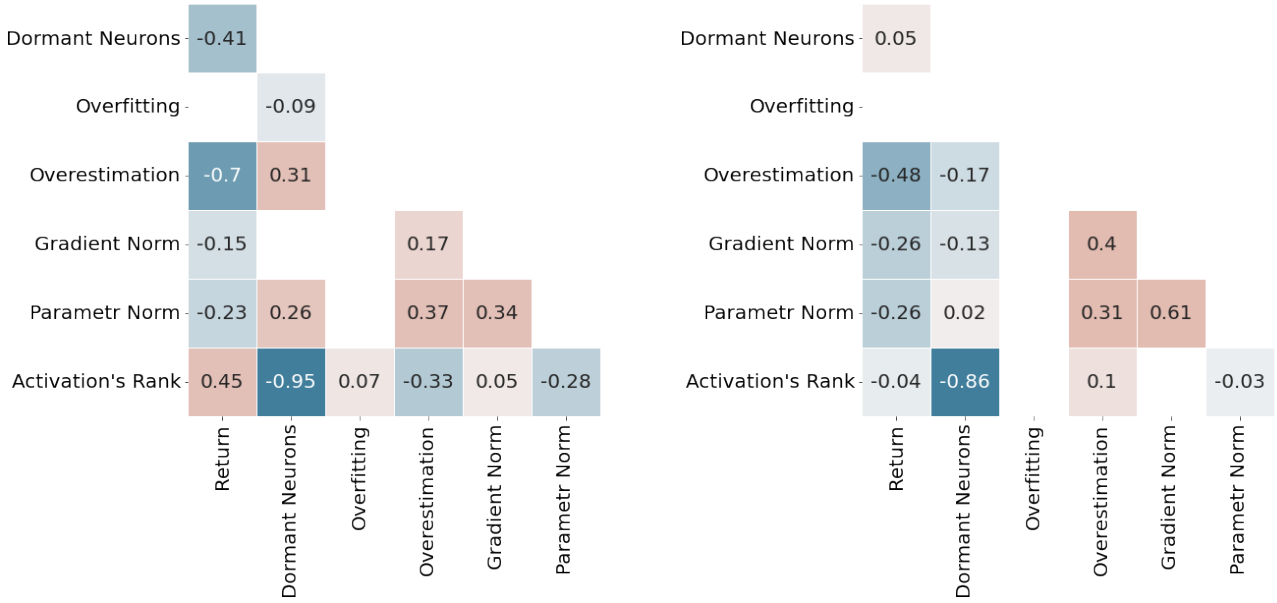

Figure 25. Spearman correlation matrix for explanatory metrics on DMC benchmark (left) and on MetaWorld benchmark (roght plot). Blank spaces are correlations that do not meet the p-value.

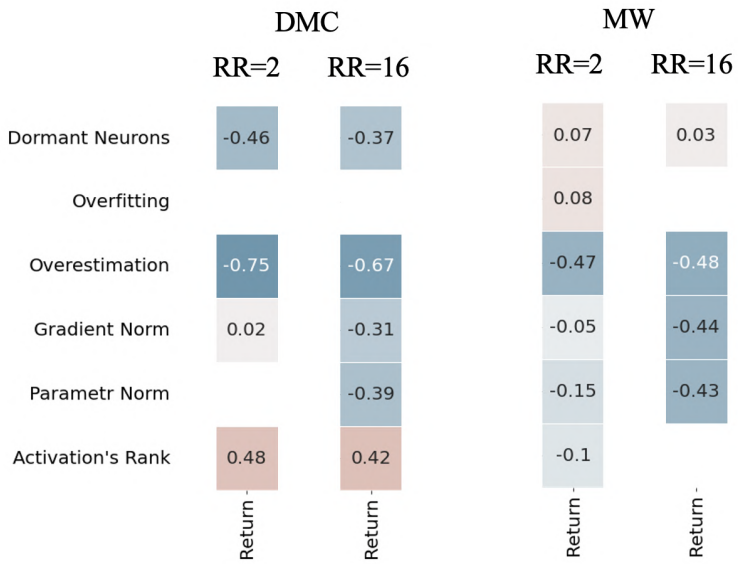

Figure 26. Spearman correlation for different replay ratios.