GENERALISED COUNTERACTIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has been the epicenter of immense scientific progress over the past decade, standing at the forefront of transformative advancements that enable agents to learn in high-dimensional spaces solely by interacting with an environment via trial and error without any supervision. Nonetheless, the learning process remains highly sensitive to the dimensions of the environment, as MDP complexity grows, the search space expands; agents continue to face a fundamental tension between complexity and policy success. Related to this disparity, recent work established a principled approach revealing that counteractive actions result in experiences with higher temporal differences. While influential, these results were founded on the assumption of discrete action spaces. In this work we focus on overcoming this tension and the agent's interaction with high-dimensional MDPs. We introduce a theoretically grounded foundational paradigm for reinforcement learning that accelerates the learning process without introducing any computational complexity. Our analysis demonstrates that learning with counteractive actions is a principled basis for scalable and efficient training. We conduct extensive experiments in high-dimensional complex MDPs and the empirical results verify the theoretical analysis provided in our paper demonstrating that generalized counteractive reinforcement learning results in accelerated and effective learning.

1 INTRODUCTION

What began as the approximation of state-action value functions via deep neural networks (Mnih et al., 2015) has evolved into agents capable of solving high-dimensional complex tasks (Hasselt et al., 2016) and driving frontier scientific and mathematical reasoning capabilities in modern large language models (Krishnamurthy et al., 2024). As the complexity of domain increases, the ability of current methods to effectively learn within an environment remains a primary bottleneck. Consequently, navigating and learning optimal policies within exponentially growing spaces remains a long-standing ambition of reinforcement learning research. In pursuit of this objective, significant research has been focused on foundational contributions that provide theoretical guarantees for efficient interaction and learning in tabular and bandit settings (Kearns and Singh, 2002; Fiechter, 1997; Kakade, 2003; Auer et al., 2008; Lu and Roy, 2019). Although these guarantees provide strong mathematical contributions, the assumptions they are built on involve tracking statistical estimates of every state-action pair, consequently in high-dimensional intractable search spaces these assumptions no longer capture the inherent complexity, leaving interaction and experience in deep reinforcement learning as an open and challenging research frontier, leading practitioners to still rely on simple and computationally tractable methods (Moore and Atkeson, 1993; Baird and Moore, 1998; Mnih et al., 2015; Hasselt et al., 2016; Wang et al., 2016; Dabney et al., 2018; Flennerhag et al., 2022; Kapturowski et al., 2023; Arnal et al., 2026).

As the space the agent interacts with grows, the computational complexity of what and how to learn from such an environment remains a foundational challenge in reinforcement learning. The number of samples required to learn a policy in exponentially expanding high-dimensional spaces is one of the primary bottlenecks of deep reinforcement learning research, rendering theoretically sound algorithms computationally restrictive in practice. These constraints limit reinforcement learning from being fully realized in domains ranging from the scientific and mathematical reasoning of large language models to physical intelligence; more critically, they represent the pivotal milestone that must be surmounted to achieve truly generalizable and adaptable intelligence capable of operating in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dynamic, non-stationary environments. Thus, in this paper we aim to seek answers for the following questions.

- *How can we have curiosity and novelty seeking interactions in exponentially growing high-dimensional environments without introducing computational complexity?*
- *What are the foundational principles and theoretical grounds for curiosity seeking zero-cost interaction behaviour that can accelerate learning?*

To be able to answer these questions, in our paper we focus on the core principles of reinforcement learning in high dimensional MDPs and make the following contributions.

Contributions. We introduce a principled theoretically founded paradigm for reinforcement learning based on actions that minimize the state-action values, that we call counteractive policy gradient, i.e. CoAct PG. We provide a theoretical analysis that establishes counteractive actions in high dimensional action spaces increase temporal differences in Section 3. By solely revisiting the fundamental principles of reinforcement learning, our method accelerates learning without introducing computational complexity. We first investigate CoAct PG in a motivating example, i.e. chain MDP, in Section 4, and the results verify the theoretical analysis and demonstrate that CoAct PG converges substantially faster and achieves significantly higher asymptotic performance. We then provide a large-scale experimental analysis in high-dimensional MDPs in Section 5 and CoAct PG improves performance by 60.98% in highly complex environments. The results provided in both the complex high dimensional MDPs and the motivating example verify the theoretical analysis, demonstrating that CoAct PG significantly accelerates learning and achieves substantially higher policy performance. Finally, we show that CoAct PG is a plug-and-play approach, does not introduce computational complexity and any learning method utilizing any kind of policy gradient algorithm can be immediately switched to counteractive policy gradient.

2 Background and Preliminaries

In reinforcement learning, the environment is formalized as a Markov Decision Process (MDP), defined by the tuple $\langle S, A, \mathcal{T}, r, \gamma, \rho_0 \rangle$ (Puterman, 1994). In this tuple, S represents the state space, A represents the action space, $\mathcal{T}$ represents the transition probability distribution on $S \times A \to \Delta(S)$, $r : S \times A \times S \to \mathbb{R}$ represents the reward function, and $\gamma \in (0, 1]$ represents the discount factor. The primary objective in reinforcement learning is to learn an optimal policy $\pi(s, a)$ that maximizes the expected cumulative discounted rewards $R = \mathbb{E}_{a_t \sim \pi(s_t, \cdot)} \sum_t \gamma^t r(s_t, a_t, s_{t+1})$. One way to achieve this objective is to learn a state-action value function that assigns values to each state-action pair, where these values represent the expected cumulative discounted rewards obtained by taking action $a \in A$ in state $s \in S$. The state-action value function is learnt via recursively updating the current state-action value with the immediate rewards and the discounted rewards of the next state (Watkins and Dayan, 1992)

$$\mathcal{Q}(s, a) = r(s, a, s') + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s, a) \max_{a'} \mathcal{Q}(s', a')$$

In particular, for a policy π the estimates of the state-action values are updated via the Bellman operator (Bellman, 1957)

$$(\Omega^\pi \mathcal{Q})(s, a) = \mathbb{E}_{s' \sim \mathcal{T}(s'|s,a)}[r(s, a, s') + \gamma \mathbb{E}_{a' \sim \pi} \mathcal{Q}(s', a')]$$

Once $\mathcal{Q}(s, a)$ is learnt, the policy that maximizes return is $\pi^*(s) = \arg\max_{a \in A} \mathcal{Q}(s, a)$. When the state-space is large, computing the precise values of each state-action pair is no longer computationally tractable; therefore, function approximators are used to learn the $\mathcal{Q}$-function. In particular, deep neural networks are used to approximate the state-action value function in deep reinforcement learning.

$$\theta_{t+1} = \theta_t + \alpha(r(s_t, a_t, s_{t+1}) + \gamma \mathcal{Q}(s_{t+1}, \arg\max_a \mathcal{Q}(s_{t+1}, a; \theta_t); \theta_t) - \mathcal{Q}(s_t, a_t; \theta_t)) \nabla_{\theta_t} \mathcal{Q}(s_t, a_t; \theta_t)$$

Although this is a powerful method, when the action space is large, identifying the argmax of $\mathcal{Q}(s, a)$ is no longer a computationally tractable operation compared to directly learning a parameterized policy $\pi_\theta(s, a)$. A canonical approach for learning in high-dimensional action and state spaces is to utilize a parameterized policy $\pi_\theta(s, a)$ which is responsible for taking actions, while separately

learning an approximation of the value function to assess the quality of the actions. The interaction between the actor and the critic is formally represented as (Crites and Barto, 1994)

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi[\nabla_\theta \log \pi_\theta(a|s)\mathcal{A}^\pi(s,a)]$$

where $\mathcal{A}^\pi(s,a)$ represents the advantage function $\mathcal{A}^\pi(s,a) = \mathcal{Q}^\pi(s,a) - V^\pi(s)$. Recent advancements by Korkmaz (2025) established a foundation for counteractive actions in discrete temporal difference learning; however, the computational tractability of extending this paradigm to vastly larger, high-dimensional action spaces has yet to be investigated. While influential, the results in Korkmaz (2025) remain confined to the discrete, low-dimensional nature of the Arcade Learning Environment (Wang et al., 2016; Hasselt et al., 2016; Flennerhag et al., 2022; Kapturowski et al., 2023), and lack the scalability required for the high-dimensional complexity inherent in training large-scale reasoning models (Krishnamurthy et al., 2024; Dwaracherla et al., 2024; Ye et al., 2025). In this paper, we address this gap and first we theoretically analyze and show that high dimensional counteractive actions lead to experiences that have higher temporal differences, and we introduce zero cost paradigms for computing counteractive actions in high dimensional action spaces in Section 3. We then provide experimental analysis that spans from motivating examples, such as the Chain MDP, to canonical methods in high-dimensional benchmarks. Our results demonstrate that counteractive policy gradient is a fundamental improvement over canonical approaches and a plug-and-play method that can be used with any baseline algorithm.

3 Counteractive High Dimensional Actions

How can we obtain novelty and curiosity seeking agents in high dimensional environments without introducing computational complexity?

Seeking novel and curious experiences is a long-standing pursuit of reinforcement learning (Schmidhuber, 1999; 1991; Barto and Duff, 1993; Doya, 1995), as it foundationally leads to higher temporal differences (Moore and Atkeson, 1993). Quite recent work of (Korkmaz, 2025) showed that actions that minimize the state-action value results in higher temporal differences.

$$\mathbb{E}_{s^{\text{min}}_{t+1}\sim\mathcal{T}(s_t,a^{\text{min}},\cdot),\theta\sim\Theta}[r(s_t,a^{\text{min}}) + \gamma \max_a Q_\theta(s^{\text{min}}_{t+1},a) - Q_\theta(s_t,a^{\text{min}})]$$
$$> \mathbb{E}_{a_t\sim\mathcal{U},(\mathcal{A})s_{t+1}\sim\mathcal{T}(s_t,a_t,\cdot),\theta\sim\Theta}[r(s_t,a_t) + \gamma \max_a Q_\theta(s_{t+1},a) - Q_\theta(s_t,a_t)] + \mathcal{D}(s_t) - 2\delta - \eta$$

Here $\mathcal{D}(s_t)$ represents the disadvantage gap, while η and δ account for the uninformedness and smoothness of the state-action value function, respectively. Although influential, this study remains constrained to discrete sets of actions and it is not clear how to validate this paradigm in vastly complex high-dimensional action spaces. Given the scale of reinforcement learning training of large reasoning models, the low dimensional action set assumption of prior work lacks the essential scalability required for modern reinforcement learning training regimes (Krishnamurthy et al., 2024). In this paper, we address these limitations and introduce counteractive actions for complex high dimensional action spaces. In this section, we will introduce the theoretical analysis for our proposed algorithm counteractive policy gradient, designed specifically for learning in these complex spaces. In Section 5, we will provide experimental results that verify the theoretical analysis provided in Section 3. Let us now formalize the core principles and dynamics of learning in high dimensional MDPs. In deep reinforcement learning prior to convergence during training, when obtaining new experiences, the actions taken by the policy π_ϕ are uncorrelated with the immediate rewards and the state-action value function Q_θ. Let us now formalize this intuition and let $\Lambda_\phi(\cdot \mid s)$ to denote a distribution over actions $a^{\text{min}} \in \mathcal{A}$ that have lower value than those drawn from the current policy $\pi_\phi(\cdot \mid s)$.

$$\mathbb{E}_{a^{\text{min}}\sim\Lambda_\phi(\cdot|s)}[Q_\theta(s,a^{\text{min}})] < \mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[Q_\theta(s,a)] \tag{1}$$

For completeness, we prove how counteractive high dimensional actions result in higher temporal differences in Theorem A.4 and in Theorem A.6. Now here we will show how to identify counteractive actions in high dimensional action spaces.

Computing the Minimum in High Dimensional Action Spaces. While Theorem A.4 and Theorem A.6 established that counteractive actions fundamentally increase temporal differences, their practical

utility depends on efficient computation—particularly in high dimensional action spaces where exhaustive sampling is intractable. In this section, we introduce the theoretical foundations for identifying counteractive actions without introducing additional computational complexity. We demonstrate that the optimization of the policy objective induces a distribution π_ϕ whose density concentrates around the local maxima of the current Q_θ-function estimate. By leveraging this convergence property, we show that counteractive actions can be efficiently recovered by sampling from the low-density tails of the existing policy. These samples inherently correspond to regions of lower Q_θ-values, providing a natural contrastive signal without introducing any computational complexity. Let us now formalize the learning dynamics that govern high-dimensional action spaces. In high dimensional action spaces, the policy gradient family of algorithms typically aims to learn a policy $\pi_\phi : S \to A \subset \mathbb{R}^d$ that maps states $s \in S$ to actions $a \in A$ via the gradient

$$\nabla J(\phi) = \nabla_a Q_\theta(s,a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s). \tag{2}$$

Here $\nabla_\phi \pi_\phi$ denotes the $d \times n$ Jacobian matrix of the policy, where n is the number of parameters in ϕ. From the chain-rule we have that $J(\phi) = Q_\theta(s, \pi_{\phi(s)})$ is an anti-derivative of the gradient in equation 2. Let us now have a policy π_ϕ that is fully representative, i.e. in any state there is a setting of the weights that is fully representative of any action.

Definition 3.1 (*Non-singular policy*)**.** The policy π_ϕ is non-singular if the $d \times n$ Jacobian matrix $\nabla_\phi \pi_\phi(s)$ has rank d for every $s \in S$.

Now we prove that all local maxima of the objective $J(\phi)$ correspond to local maxima of Q_θ. To do so we appeal to a standard fact that non-singular maps preserve local maxima. We provide a proof for completeness.

Lemma 3.2 (*Preservation of Critical Points*)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^d$ and let the Jacobian matrix $\nabla g(x)$ have rank r for all x. Then the critical points x^* of $f(g)$ are in one-to-one correspondence with the critical points $y^* = g(x^*)$ of f. Further, x^* and y^* have the same type i.e. are both local maxima or both local minima.*

Proof. See full proof in appendix A.7 □

Lemma 3.3. *Let π_ϕ be a non-singular policy. Suppose that ϕ^* is a local maximum of $J(\phi)$ and $a^* = \pi_{\phi^*}(s)$. Then a^* is a local maximum of $Q_\theta(s,a)$.*

Proof. See full proof in appendix A.9 □

Algorithms utilizing the deterministic policy gradient equation 2 perform gradient ascent on $J(\phi)$, and hence Lemma 3.3 implies that such algorithms move toward policies $\pi_\phi(s)$ that output actions a^* locally maximizing $Q_\theta(s,a)$. Now let us try to formalize the structure of the state-action value function Q_θ in high-dimensional MDPs.

Definition 3.4 (*(M,λ)-regularity*)**.** Let $M, \lambda > 0$. The function $Q_\theta(s,a)$ is (M,λ)-regular if for every local maximum a^* of $Q_\theta(s,a)$ the Hessian $\nabla_a^2 Q(s,a)|_{a=a^*}$ has all eigenvalues less than $-\lambda$, and the operator norm of the tensor of third derivatives is at most M i.e. for any vectors $u, v, w \in \mathbb{R}^d$ we have $\left| \sum_{i,j,k} (\nabla_a^3 Q_\theta(s,a))_{ijk} u_i v_j w_k \right| \leq M\|u\|\|v\|\|w\|$.

In general, sufficiently small deviations from local maxima of any smooth function cannot have higher values. Here, we explicitly quantify the magnitude of the deviations given the regularity parameters, such that the perturbed points from the local maximum of Q_θ will have strictly lower values.

Proposition 3.5 (*Counteractive Values*)**.** *Let Q_θ be (M,λ)-regular, let $\epsilon < \frac{3\lambda}{M}$, let a^* be a local maximum of Q_θ, and let a be an action with $\|a - a^*\| < \epsilon$. Then we have $Q_\theta(s,a) < Q_\theta(s,a^*)$.*

Proof. See full proof in appendix A.10 □

These results provide a formal basis for the policy's behavior and we have shown in Lemma 3.3 that optimizing $J(\phi)$ leads to a policy π_ϕ that outputs actions a^* that are local maxima of $Q_\theta(s,a^*)$. We have further proven in Proposition 3.5 that actions that corresponds to the tails of the policy at local maxima of $Q_\theta(s,a)$ will lead to strictly lower values. Taken together, these results directly imply that actions that correspond to the tails of π_ϕ will tend to have lower values under Q_θ. We state this as the main result of this section.

Theorem 3.6 (*Counteractive Values of Policy Tails*). *Let π_ϕ be a non-singular policy and Q_θ have (M, λ)-strict local maxima. Let $\epsilon < \frac{3\lambda}{M}$. Let ϕ^* be a local maximum of $J(\phi)$, and $a^* = \pi_{\phi^*}(s)$. Let a be any action with $\|a - a^*\| \leq \epsilon$. Then $Q_\theta(s, a) < Q_\theta(s, a^*) - \frac{\lambda}{2}\|a - a^*\|$.*

Proof. See full proof in appendix A.11 □

Now let us focus on the stochastic policy optimization objectives and show that solutions of this objective have the property that actions that correspond to the tails of the policy distribution result in lower Q-values. The main intuition is that the KL-divergence objective is mode-seeking, implying that the action distribution generated by the policy will lock-on to one of the modes of Q. Hence, deviations from the mean of this action distribution will have lower Q-values. In particular, a neural network is trained to output the mean and standard deviation vectors $\mu, \sigma \in \mathbb{R}^d$, while sampling and computing probabilities under $\pi_\phi(\cdot|s)$ is done with the explicit formula for the Gaussian distribution with mean μ and covariance $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. The policy is updated by optimizing the loss

$$J(\phi) = D_{\mathrm{KL}}\left(\pi_\phi(\cdot|s) \,\middle\|\, \frac{\exp(Q_\theta(s, \cdot))}{Z_\theta(s)}\right) \tag{3}$$

where $Z_\theta(s) = \int \exp(Q_\theta(s, a))da$ is a normalizing constant so that the right-hand term is a probability density. The loss $J(\phi)$ is the reverse-KL, and hence is mode-seeking. Intuitively, this implies that the Gaussian distribution determined by π_ϕ will attempt to align with a mode of the distribution given by the soft-max $\exp(Q_\theta(s, \cdot))/Z_\theta(s)$. Modes of a soft-max distribution correspond to local maxima of the underlying function Q_θ. Thus, we expect that the policy π_ϕ will output a mean μ close to a local maximum of Q_θ. Now let us provide the formal statements and proofs for the stochastic loss. Let $\gamma_{\mu,\sigma}$ denote the density of $\mathcal{N}(\mu, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2))$. Our analysis will rely on properties of the function $F(\mu, \sigma) = -D_{\mathrm{KL}}(\gamma_{\mu,\sigma} \parallel \exp Q_\theta(s, \cdot)/Z_\theta(s))$.

As in the case of deterministic policy gradient, the goal will be to show that the optimum of the loss corresponds to a local maximum of Q, and hence perturbations away from this optimum decrease Q values. For the KL objective equation 3, the optima will not be precisely equal, but only approximately equal. Under certain analytic regularity conditions we will show that this is sufficient for our purposes. To simplify notation in this section we will drop the θ subscript i.e. just write $Q(s, a)$ instead of $Q_\theta(s, a)$. Furthermore, all gradients and higher derivatives of $Q(s, a)$ will be taken with respect to the a coordinates i.e. we will use the notation $\nabla Q(s, b) = \nabla_a Q(s, a)|_{a=b}$. To begin our analysis, we apply Lemma 3.2 to relate the stationary points of $J(\phi)$ to those of $F(\mu, \sigma)$.

Lemma 3.7. *Let π_ϕ be a non-singular policy. Let μ, σ be the mean and covariance vectors for the policy π_ϕ. If the point (μ, σ) is a stationary point of $F(\mu, \sigma)$, then ϕ is a stationary point of $J(\phi)$.*

Proof. Let $g(\phi)$ be the function mapping $\phi \mapsto (\mu, \sigma)$ where μ and σ are the respective mean and standard deviation vectors of π_ϕ. Then $-J(\phi) = F(g(\phi))$. Applying Lemma 3.2 completes the proof. □

We next rewrite F in a more convenient form. Let $Z = \int \exp(Q(s, a))da$ and $H(\sigma) = \frac{d}{2}\log(2\pi e) + \sum_{i=1}^d \log \sigma_i$ be the entropy of $\gamma_{\mu,\sigma}$. Using the fact that $D_{\mathrm{KL}}(p \parallel q) = -H(p) - \mathbb{E}_p[\log q]$ we have

$$F(\mu, \sigma) = H(\sigma) + \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}[Q(s, a)] - \log Z. \tag{4}$$

To understand the critical points of F, we compute the first-order stationarity conditions.

Lemma 3.8. *The derivatives of F are given by*

$$\nabla_\mu F(\mu, \sigma) = \mathbb{E}_{a\sim\gamma_{\mu^*,\sigma^*}}[\nabla Q(s, a)] \text{ and } \frac{\partial}{\partial\sigma_i}F(\mu, \sigma) = \frac{1}{\sigma_i} + \sigma_i \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}\left[\nabla_{ii}^2 Q(s, a)\right] \tag{5}$$

Proof. See full proof in appendix A.13 □

The next step in the argument will be to show that for any local maximum a^* of $Q(s, a)$, there is a compact set K which contains a local optimum (μ^*, σ^*) where μ^* is close to a^*. The candidate set will take the form

$$K = \left\{(\mu, \sigma) \,\middle|\, \|\mu - a^*\| < \rho,\ \frac{1}{2L} \leq \sigma_i^2 \leq \frac{4}{\lambda}\right\} \tag{6}$$

Algorithm 1 CoAct PG: Counteractive Policy Gradient

Input: In MDP $\mathcal{M}$ with $\gamma \in (0, 1]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ with $Q_\theta(s, a)$ function parametrized by θ, $\pi_\phi(\cdot \mid s)$ is the policy parametrized by ϕ, ϵ dithering parameter, $\mathcal{N}$ is the training learning steps.

Interactions:
for s_t in e **do**
  Sample $\kappa \sim U(0, 1)$
  if $\kappa < \epsilon$ **then**
    $a^{min} \sim \Lambda_\phi(\cdot \mid s)$, $s^{min}_{t+1} \sim \mathcal{T}(s_t, a^{min}, \cdot)$
    $\mathcal{B} \leftarrow (r(s_t, a^{min}), s_t, s^{min}_{t+1}, a^{min})$
  else
    $a^{\max} \sim \pi_\phi(\cdot \mid s)$, $s_{t+1} \sim \mathcal{T}(s_t, a^{\max}, \cdot)$
    $\mathcal{B} \leftarrow (r(s_t, a^{\max}), s_t, s_{t+1}, a^{\max})$
  end if
end for

Learning:
for n in $\mathcal{N}$ **do**
  Interactions produced by the policy π_ϕ:
  $\langle s_t, a_t, r(s_t, a_t), s_{t+1} \rangle$
  Thus, $\mathcal{TD}$ receives update with probability ϵ
  $\mathcal{TD} = r(s_t, a^{min}) + \gamma \mathbb{E}_{a \sim \pi_\phi} Q(s^{min}_{t+1}, a) - Q(s_t, a^{min})$
  $\mathcal{TD}$ receives update with probability $1 - \epsilon$:
  $\mathcal{TD} = r(s_t, a^{\max}) + \gamma \mathbb{E}_{a \sim \pi_\phi} Q(s_{t+1}, a) - Q(s_t, a^{\max})$
  Update π_ϕ via $\nabla J_\pi(\phi)$ with the updated Q_θ
end for

where the relationship between L, ρ and λ required for our results will be set below.

We will now prove in a series of lemmas that the gradient of $F(\mu, \sigma)$ points inwards at the boundary of this set.

Lemma 3.9. *Let Q be (M, λ)-regular with $\lambda^4 \geq 8M^2$. Let the radius ρ defining K be given by $\rho = \frac{\lambda}{2M}$. Let (μ, σ) be a point on the boundary of K, then $(\mu - a^*)^\top \nabla_\mu F(\mu, \sigma) < 0$*

Proof. See full proof in appendix A.14 □

We next prove that the components of the gradient with respect to σ all point inwards at the boundary of K.

Lemma 3.10. *Let Q be (M, λ)-regular with $\lambda > 4M\sqrt{d}$. Let $\rho = \lambda/2M$. Let $L > 2\lambda + \frac{2M\sqrt{d}}{\sqrt{\lambda}}$ and assume $\nabla_{ii} Q(s, a^*) \geq -L$. Then for each i the gradient of F points inward at the boundary of K i.e.*

$$\partial_{\sigma_i} F(\mu, 2/\sqrt{\lambda}) < 0 \text{ and } \partial_{\sigma_i} F(\mu, 1/(\sqrt{2L})) > 0$$

Proof. See full proof in appendix A.15 □

We now apply the results of Lemma 3.9 and Lemma 3.10 to prove that there exists a stationary point μ^*, σ^* within K.

Lemma 3.11. *Let Q be (M, λ)-regular with $\lambda > \max\{2\sqrt{M}, 4M\sqrt{d}\}$. Assume $\nabla_{ii} Q(s, a^*) \geq -L$ for $L > 2\lambda + \frac{2M\sqrt{d}}{\sqrt{\lambda}}$. Let $\rho = \lambda/(2M)$. Then there exists a stationary point (μ^*, σ^*) of $F(\mu, \sigma)$ within K.*

Proof. See full proof in appendix A.16 □

Having proven that there is a stationary point $(\mu^*, \sigma^*) \in K$, we proceed to compute upper bounds on $\nabla Q(s, \mu^*)$ and $\nabla^2 Q(s, \mu^*)$.

Lemma 3.12. *Let Q be (M, λ)-regular and let ρ, λ, M be as in Lemma 3.11. Let $(\mu^*, \sigma^*) \in K$ be a stationary point of F. Then we have that all eigenvalues of $\nabla^2 Q(s, \mu^*) < -\lambda/2$ and*

$$|\nabla_i Q(s, \mu^*)| \leq \frac{2dM}{\lambda}$$

Proof. See complete proof in appendix A.17. □

Finally, we are able to prove the main result of this section. This shows that near every local maximum of Q there is a stationary point of J_ϕ such that deviations proportional to σ from the mean μ of π_ϕ decrease the Q value.

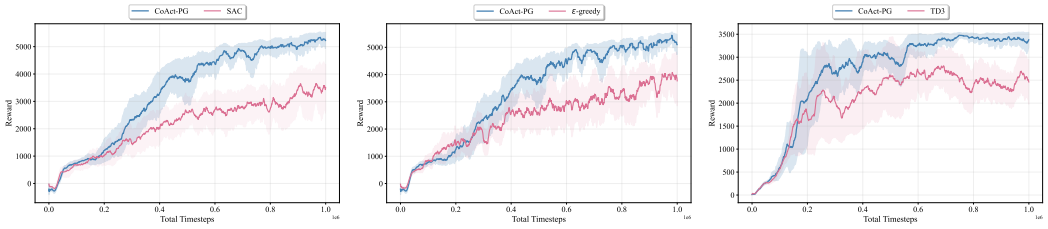


Figure 1: Learning curves of our proposed algorithm CoAct-PG that maximizes temporal differences, and canonical baseline algorithms, SAC, TD3 and ϵ-greedy in high dimensional MDPs Hopper and Ant.

Theorem 3.13. *Suppose the regularity conditions of Lemma 3.11 hold. Then for every local maximum a^* of $Q(s,a)$ and $c > 0$, there is a value of λ such that when Q is (M,λ) regular, there exists a stationary point of ϕ^* of $J(\phi)$, such that $Q(s,\mu^* + c\sigma^*) < Q(s,\mu^*)$ where μ^*, σ^* are the mean and variance vectors of the policy π_{ϕ^*}.*

Proof. First by Lemma 3.7 the critical points of ϕ correspond to the critical points of F, so it is sufficient to consider those. Let $(\mu^*, \sigma^*) \in K$ be the critical point of F guaranteed to exist by Lemma 3.11. Let $c > 0$ and let $v = c\sigma^*$. By Taylor's theorem we have that there exists a point ξ on the line from μ^* to $\mu^* + v$ such that

$$Q(s,\mu^* + c\sigma^*) - Q(s,\mu^*) = \nabla Q(s,\mu^*) \cdot v + \frac{1}{2} v^\top \nabla^2 Q(s,\mu^*) v + \frac{1}{6} \sum_{i,j,k=1}^{d} (\nabla^3_{ijk} Q_\theta(s,\xi)) v_i v_j v_k.$$

We next upper bound the terms in the Taylor expansion one at a time, starting by applying Lemma 3.12 to the linear term,

$$\nabla Q(s,\mu^*) \cdot v = \sum_i \nabla_i Q v_i \leq \frac{2dM}{\lambda} \sum_i v_i \leq \frac{4cd^2 M}{\lambda^{3/2}}. \tag{7}$$

Next we apply Lemma 3.12 to the quadratic term to obtain

$$\frac{1}{2} v^\top \nabla^2 Q(s,\mu^*) v \leq \frac{-\lambda}{4} \|v\|^2 \leq -dc^2 \tag{8}$$

Finally the degree three term is bounded by

$$\frac{1}{6} \sum_{i,j,k=1}^{d} (\nabla^3_{ijk} Q_\theta(s,\xi)) v_i v_j v_k \leq M \|v\|^3 \leq c^3 M \left(\frac{4d}{\lambda}\right)^{3/2} \tag{9}$$

Both equation 7 and equation 9 have dependence on λ of the form $\lambda^{-3/2}$, while the negative term equation 8 is $-c^2 d$ has no λ dependence. Thus, if λ is large enough relative to M,d, and c we conclude that $Q(s,\mu^* + c\sigma^*) < Q(s,\mu^*)$. □

We describe counteractive policy gradient in Algorithm 1 based on prioritizing actions that target low state-action values as theoretically analyzed in detail in Section 3. Note that counteractive policy gradient is modular and can be used with any baseline algorithm.

CoAct-PG is modular: A drop-in replacement for any baseline algorithm.

4 Motivating Example

Before we dive into high dimensional MDPs, we first develop intuition for the counterintuitive approach through a motivating example. In the chain MDP, the environment is structured as a sequence of n states, denoted as $S = \{1, 2, \ldots, n\}$. Within each state i, the agent can choose from four distinct

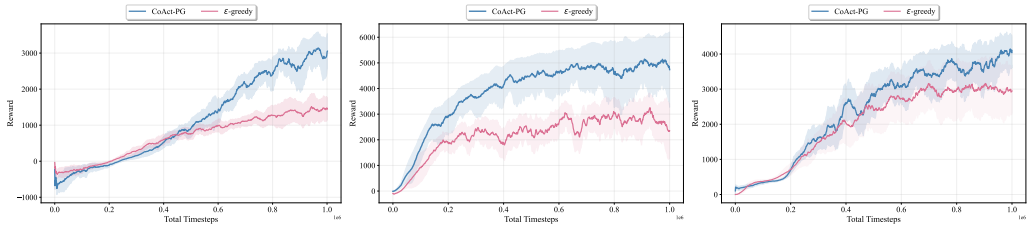


Figure 3: Average return of our proposed algorithm CoAct-PG and baseline canonical methods ϵ-greedy in high dimensional MDPs: Walker2D, HalfCheetah and Ant with learning curves.

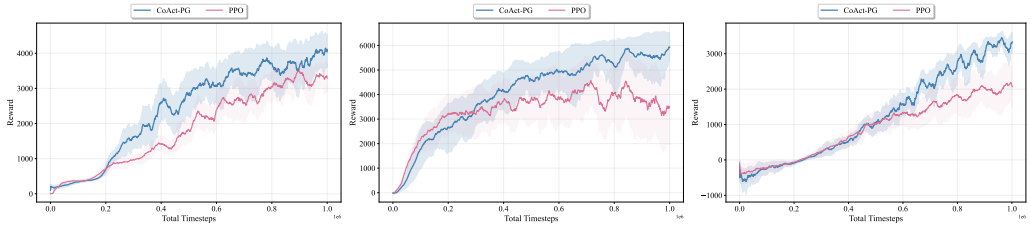


Figure 4: Learning curves of our proposed algorithm counteractive policy gradient, i.e. CoAct-PG, and baseline proximal policy gradient, i.e. PPO, in Walker2D, HalfCheetah and Ant.

actions: one action moves the agent forward to the next state, $i + 1$, i.e. forward progression, two separate actions allow the agent to jump directly to state 2 or state 3, regardless of its current position, and one action returns the agent to the very beginning of the sequence at state 1, i.e. reset. The agent is only able to see a reward when it moves from the final state n back to state 1, which provides the only positive reward in the MDP. To be able to learn in such a MDP, the agent must learn a policy that successfully takes $n - 2$ consecutive steps up the chain, and then one final step to reset and get a reward. The chain MDP experiments are run 10K steps with 10 episodes of length 100, and evaluation in every 200 steps. Figure 2 reports results for our proposed algorithm CoAct-PG and baseline SAC algorithm in the Chain MDP. These results verify the theoretical predictions of Section 3 and demonstrate that counteractive policy gradient learns substantially faster and reaches a higher performing policy.

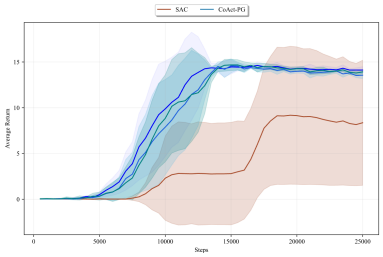


Figure 2: Average return results with CoAct-PG and SAC in the Chain MDP.

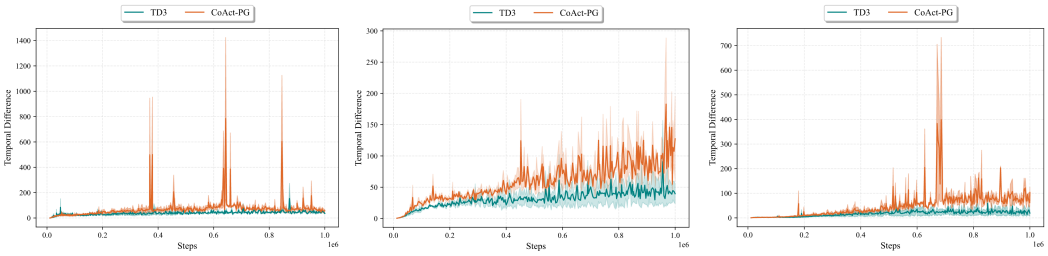
5 Experimental Analysis

The experiments are conducted in high-dimensional action and state space learning environments in MuJoCo. We conduct analysis and evaluate with multiple canonical baseline algorithms including Deep Deterministic Policy Gradient, i.e. DDPG, Proximal Policy Gradient, i.e. PPO, Twin Delayed Deep Deterministic policy gradient and Soft Actor Critic algorithm. The experimental results are conducted with 5 trials, 10 runs and all of the results are reported with the standard error of the mean.

Figure 1 and Figure 6 report results for learning curves in the Walker2D, Ant, HalfCheetah and Hopper environments with our proposed algorithm CoAct-PG and canonical baseline algorithms Twin Delayed Deep Deterministic policy gradient, Soft-Actor Critic and ϵ-greedy algorithm over 1 million steps. Figure 4 and Figure 3 report learning curves for PPO, ϵ-greedy and our algorithm CoAct-PG, demonstrating the clear impact of increasing temporal differences via low-value actions. Table 1 reports average return obtained by our proposed algorithm CoAct-PG based on maximizing temporal differences via counteractive

Environments	PPO	Co-Act PG
HalfCheetah	3498.526±891.673	**5932.317±346.573**
Swimmer	307.686±0.1092	**343.503±1.567**
Walker2D	3292.636±221.272	**4041.492±255.618**
Hopper	1988.637±225.686	**3195.907±17.775**
Ant	2068.536±320.126	**3330.026±151.320**

Table 1: Performance of our proposed algorithm CoAct-PG and canonical baseline algorithm in HalfCheetah, Swimmer, Walker2D, Hopper, and Ant.



Walker2D HalfCheetah Ant

Figure 5: Temporal differences obtained by our proposed algorithm CoAct-PG and canonical baseline methods in Walker2D, HalfCheetah and Ant environments in high dimensional MDPs.

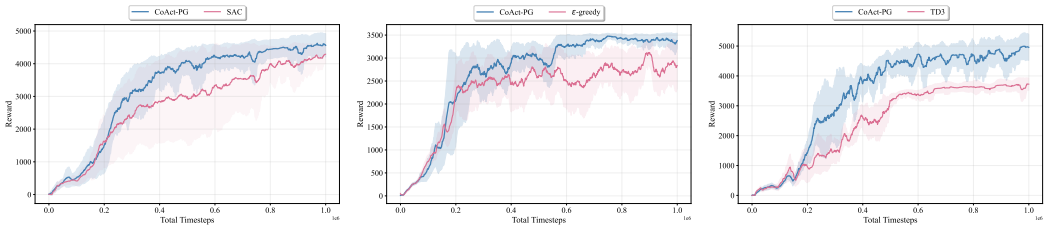


Figure 6: Learning curves of our proposed algorithm CoAct-PG that maximizes temporal differences, and canonical baseline algorithms, SAC, TD3 and ϵ-greedy in high dimensional MDPs Walker2D, Hopper and Ant.

actions and baseline methods in HalfCheetah, Swimmer, Hopper, Walker2D and Ant. These results demonstrate that CoAct-PG not only achieves substantially accelerated convergence but further reaches a superior asymptotic performance.

In highly complex environments like Ant, characterized by inherent instability, high-dimensional state-action spaces, with interdependence of joints that need to be coordinated in time and space, CoAct-PG learns a substantially superior policy compared to canonical baseline methods. In particular, the results demonstrate that CoAct-PG achieves a **60.98% higher** performance over the baseline without introducing any computational complexity.

Theorem 3.6 and Theorem A.4 provide the formal basis for our approach, proving that counteractive sampling in high-dimensional action spaces yields transitions with significantly larger temporal differences. To evaluate this theoretical grounding empirically, we compare the temporal differences of CoAct-PG against the baseline in Figure 5. The empirical results in Figure 5 provide a striking validation of our Section 3 analysis. Specifically, the counteractive policy consistently generates higher-magnitude temporal differences throughout the entire learning regime.

Not only does the counteractive policy gradient accelerate learning without introducing computational complexity and yields in learning substantially higher performing policies, but further it is an **extremely modular** algorithm requiring **only two lines of additional code** that can be used with any baseline algorithm.

6 Conclusion

In this paper we focused on the following questions: *(i) How can we have curiosity seeking agents in high-dimensional environments without introducing computational complexity? (ii) What are the foundations for zero-cost curiosity seeking reinforcement learning paradigm that accelerates learning in high dimensional spaces?* To be able to answer these questions, we introduce a fundamental paradigm for reinforcement learning based on counteractive actions that targets lower state-action values. We provide theoretical analysis showing that counteractive actions increase temporal differences obtained from experiences, thereby creating more unique experiences. We then introduce a theoretically well-founded zero-cost method for computing counteractive actions in high dimensional continuous MDPs. We first demonstrate in a motivating example, i.e. the chain MDP, that counteractive policy gradient achieves substantially higher performance and faster convergence compared to standard baselines. We then extend this paradigm and conduct extensive experiments in high-dimensional complex MDPs and CoAct-PG learns significantly faster and converges to substantially higher-performing policies in continuous spaces.

AI USE STATEMENT

This paper does not contain any AI generated content and does not use any AI tools in any shape or form.

REFERENCES

C. Arnal, V. Cabannes, T. Cohen, J. Kempe, and R. Munos. Efficient rl training for llms with experience replay. In *International Conference on Machine Learning*. ICML, 2026.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 89–96. Curran Associates, Inc., 2008.

L. Baird and A. Moore. Gradient descent for general reinforcement learning. In *Annual Conference on Neural Information Processing Systems*, 1998.

A. G. Barto and M. O. Duff. Monte carlo matrix inversion and reinforcement learning. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6, 7th NIPS Conference, Denver, Colorado, USA, 1993*, pages 687–694. Morgan Kaufmann, 1993.

R. E. Bellman. Dynamic programming. In *Princeton, NJ: Princeton University Press*, 1957.

R. H. Crites and A. G. Barto. An actor/critic algorithm that equivalent to q-learning. *Advances in Neural Information Processing Systems 7 (NIPS 1994)*, 1994.

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2892–2901. AAAI Press, 2018.

K. Doya. Temporal difference learning in continuous time and space. In D. S. Touretzky, M. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8, NIPS, Denver, , 1995*, pages 1073–1079, 1995.

V. Dwaracherla, S. M. Asghari, B. Hao, and B. V. Roy. Efficient exploration for llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 12215–12227. PMLR / OpenReview.net, 2024.

C.-N. Fiechter. Expected mistake bound model for on-line reinforcement learning. In *International Conference on Machine Learning (ICML)*, 1997.

S. Flennerhag, Y. Schroecker, T. Zahavy, H. van Hasselt, D. Silver, and S. Singh. Bootstrapped meta-learning. *10th International Conference on Learning Representations, ICLR*, 2022.

H. v. Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

S. Kakade. On the sample complexity of reinforcement learning. In *PhD Thesis: University College London*, 2003.

S. Kapturowski, V. Campos, R. Jiang, N. Rakicevic, H. van Hasselt, C. Blundell, and A. P. Badia. Human-level atari 200x faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/pdf?id=JtC6yOHRoJJ`.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 2002.

E. Korkmaz. Counteractive RL: Rethinking core principles for efficient and scalable deep reinforcement learning. *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025 [Spotlight Presentation]*, 2025.

A. Krishnamurthy, K. Harris, D. J. Foster, C. Zhang, and A. Slivkins. Can large language models explore in-context? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

X. Lu and B. V. Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2458–2466, 2019.

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

A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13:103–130, 1993.

M. L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. *John Wiley and Sons, Inc*, 1994.

J. Schmidhuber. Curious model building control systems. *International Joint Conference on Neural Networks*, 1991.

J. Schmidhuber. Artificial curiosity based on discovering novel algorithmic predictability through coevolution. In *Proceedings of the 1999 Congress on Evolutionary Computation, CEC 1999, Washington, DC, USA July 6-9, 1999*, pages 1612–1618. IEEE, 1999. doi: 10.1109/CEC.1999.785467. URL https://doi.org/10.1109/CEC.1999.785467.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, page 1995–2003, 2016.

C. Watkins and P. Dayan. Learning from delayed rewards. volume 8. Machine Learning, 1992.

Z. Ye, R. Agarwal, T. Liu, R. Joshi, S. Velury, Q. V. Le, Q. Tan, and Y. Liu. Reward-guided prompt evolving in reinforcement learning for llms. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025*, Proceedings of Machine Learning Research, 2025.

A APPENDIX

Definition A.1 (*η-uninformed*)**.** The policy π_ϕ and state-action value function Q_θ are η-uninformed in state s if for $a \sim \pi_\phi(\cdot \mid s)$ and $a^{\min} \sim \Lambda_\phi(\cdot \mid s)$ we have

$$\left|\mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[r(s,a)] - \mathbb{E}_{a^{\min}\sim\Lambda_\phi(\cdot|s)}[r(s,a^{\min})]\right| < \eta.$$

Definition A.2 (δ*-smooth*)**.** Let s be a state and let $\hat{a}$ be either an action sampled from $\pi_\phi(\cdot \mid s)$ or from $\Lambda_\phi(\cdot \mid s)$. The policy π_ϕ and state-action value function Q_θ are δ-smooth in state s if

$$\left|\mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[Q_\theta(s,a)] - \mathbb{E}_{s'\sim\mathcal{T}(s,\hat{a},\cdot);a'\sim\pi_\phi(\cdot|s')}[Q(s',a')]\right| < \delta$$

Definition A.3 (*Disadvantage Gap*)**.** Let s be a state. The disadvantage gap in state s is given by

$$\mathcal{D}(s) = \mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[Q_\theta(s,a)] - \mathbb{E}_{a^{\min}\sim\Lambda_\phi(\cdot|s)}[Q_\theta(s,a^{\min})].$$

The following theorem captures the intuition that choosing lower-valued actions will lead to larger temporal difference.

Theorem A.4 (*Counteractive High Dimensional Actions Increase Temporal Difference*)**.** *Let $\eta, \delta > 0$ and let π_ϕ and Q_θ be η-uninformed and δ-smooth at state $s \in \mathcal{S}$. Let $a^{min} \sim \Lambda_\phi(\cdot \mid s)$ and $a \sim \pi_\phi(\cdot \mid s)$. Let $s^{min} \sim \mathcal{T}(s, a^{min}, \cdot)$ and let $s' \sim \mathcal{T}(s, a, \cdot)$. Then*

$$\begin{aligned}&\mathbb{E}_{s^{min}\sim\mathcal{T}(s,a^{min},\cdot);\tilde{a}\sim\pi_\phi(\cdot|s^{min})}\left[r(s,a^{min}) + \gamma Q_\theta(s^{min},\tilde{a}) - Q_\theta(s,a^{min})\right]\\ &\qquad > \mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[r(s,a) + \gamma Q_\theta(s',a') - Q_\theta(s,a)\right] + \mathcal{D}(s) - 2\delta - \eta\end{aligned}$$

Proof. By δ-smoothness we have

$$\begin{aligned}&\mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot);\tilde{a}\sim\pi_\phi(\cdot|s^{\min})}\left[\gamma Q_\theta(s^{\min},\tilde{a}) - Q_\theta(s,a^{\min})\right]\\ &\qquad > \mathbb{E}_{a\sim\pi_\phi(\cdot|s)}\left[\gamma Q_\theta(s,a)\right] - \delta - \mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot)}\left[Q_\theta(s,a^{\min})\right]\\ &\qquad > \mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[\gamma Q_\theta(s',a')\right] - 2\delta - \mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot)}\left[Q_\theta(s,a^{\min})\right]\\ &\qquad > \mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[\gamma Q_\theta(s',a') - Q_\theta(s,a)\right] + \mathcal{D}(s) - 2\delta\end{aligned}$$

where the final line uses Definition A.3. Finally, because Q_θ is η-uninformed we have

$$\mathbb{E}_{a^{\min}\sim\Lambda_\phi(\cdot|s)}[r(s,a^{\min})] > \mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[r(s,a)] - \eta.$$

Combining this with the previous inequality completes the proof. □

Theorem A.4 shows that high dimensional counteractive actions, i.e. actions that correspond to low state-action values, achieve higher temporal difference. Many standard policy gradient algorithms use two state-action value function estimators $Q_{\theta_1}, Q_{\theta_2}$ to reduce overestimation bias. We now show that sampling lower-valued actions still increases temporal difference in the two Q-function setting. First, we must adapt the definition of δ-smoothness to the new setting.

Definition A.5 (δ*-smooth for Double-Q*)**.** Let s be a state and let $\hat{a}$ be either an action sampled from $\pi_\phi(\cdot \mid s)$ or from $\Lambda_\phi(\cdot \mid s)$. The policy π_ϕ and state-action value functions $Q_{\theta_1}, Q_{\theta_2}$ are δ-smooth in state s if

$$\left|\mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[Q_\theta(s,a)] - \mathbb{E}_{s'\sim\mathcal{T}(s,\hat{a},\cdot);a'\sim\pi_\phi(\cdot|s')}\left[\min_{i=1,2} Q_{\theta_i}(s',a')\right]\right| < \delta$$

Let us now prove that sampling lower-valued actions increases temporal difference in the two Q-function setting.

Theorem A.6. *Let $\eta, \delta > 0$ and let π_ϕ and Q_{θ_i} be η-uninformed and δ-smooth at state $s \in \mathcal{S}$. Let $a^{min} \sim \Lambda_\phi(\cdot \mid s)$ and $a \sim \pi_\phi(\cdot \mid s)$. Let $s^{min} \sim \mathcal{T}(s, a^{min}, \cdot)$ and let $s' \sim \mathcal{T}(s, a, \cdot)$. Then*

$$\begin{aligned}&\mathbb{E}_{s^{min}\sim\mathcal{T}(s,a^{min},\cdot);\tilde{a}\sim\pi_\phi(\cdot|s^{min})}\left[r(s,a^{min}) + \gamma \min_{i=1,2} Q_{\theta_i}(s^{min},\tilde{a}) - Q_\theta(s,a^{min})\right]\\ &\qquad > \mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[r(s,a) + \gamma \min_{i=1,2} Q_{\theta_i}(s',a') - Q_\theta(s,a)\right] + \mathcal{D}(s) - 2\delta - \eta.\end{aligned}$$

Proof. By δ-smoothness we have

$$\begin{aligned}
&\mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot);\tilde{a}\sim\pi_\phi(\cdot|s^{\min})}\left[\gamma\min_{i=1,2}Q_{\theta_i}(s^{\min},\tilde{a})-Q_\theta(s,a^{\min})\right]\\
&\qquad>\mathbb{E}_{a\sim\pi_\phi(\cdot|s)}\left[\gamma Q_\theta(s,a)\right]-\delta-\mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot)}\left[Q_\theta(s,a^{\min})\right]\\
&\qquad>\mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[\gamma\min_{i=1,2}Q_{\theta_i}(s',a')\right]-2\delta-\mathbb{E}_{s^{\min}\sim\mathcal{T}(s,a^{\min},\cdot)}\left[Q_\theta(s,a^{\min})\right]\\
&\qquad>\mathbb{E}_{s'\sim\mathcal{T}(s,a,\cdot);a'\sim\pi_\phi(\cdot|s')}\left[\gamma\min_{i=1,2}Q_{\theta_i}(s',a')-Q_\theta(s,a)\right]+\mathcal{D}(s)-2\delta
\end{aligned}$$

where the final line uses Definition A.3. Finally, because Q_θ is η-uninformed we have

$$\mathbb{E}_{a^{\min}\sim\Lambda_\phi(\cdot|s)}[r(s,a^{\min})]>\mathbb{E}_{a\sim\pi_\phi(\cdot|s)}[r(s,a)]-\eta.$$

Combining this with the previous inequality completes the proof. □

Lemma A.7 (*Preservation of Critical Points*)**.** *Let $f:\mathbb{R}^d\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}^d$ and let the Jacobian matrix $\nabla g(x)$ have rank r for all x. Then the critical points x^* of $f(g)$ are in one-to-one correspondence with the critical points $y^*=g(x^*)$ of f. Further, x^* and y^* have the same type i.e. are both local maxima or both local minima.*

Proof. First, by the chain rule we have

$$\nabla f(g)(x)=\nabla_y f(y)|_{y=g(x)}\nabla g(x)$$

The Jacobian matrix ∇g has full row-rank d. Hence the matrix-vector product $\nabla_y f(y)|_{y=g(x)}\nabla g(x)$ is zero if an only if $\nabla_y f(y)|_{y=g(x)}=0$. Thus, $y^*=g(x^*)$ is a first-order stationary point of f if and only if x^* is a first-order stationary point of $f(g)$. To prove that any local optimum y^* has the same type as x^* we compute the Hessian of $f(g)$. To do so, we let $\nabla f(y^*)_k$ denote the k-th coordinate of $\nabla f(y^*)$ and $\nabla^2 g_k(x^*)$ denote the Hessian of the function $g_k(x)$ given by the k-th coordinate of $g(x)$.

With this notation the Hessian of $f(g)$ is given by:

$$\begin{aligned}
\nabla^2 f(g)(x^*)&=(\nabla g(x^*))^\top\nabla^2 f(y^*)\nabla g(x^*)+\underbrace{\sum_{k=1}^{d}\nabla f(y^*)_k\nabla^2 g_k(x^*)}_{=0}\\
&=(\nabla g(x^*))\nabla^2 f(y^*)\nabla g(x^*)
\end{aligned}$$

where the equality follows because the first-order stationary conditions imply $\nabla f_k(y^*)=0$ for all k. First, assume that x^* is a local maximum of $f(g)$, which implies that $\nabla^2 f(g)(x^*)$ is negative semi-definite i.e. $v^\top\nabla^2 f(g)(x^*)v\le 0$ for all $v\in\mathbb{R}^n$. Plugging in the above formula for $\nabla^2 f(g)(x^*)$ we conclude that for all $v\in\mathbb{R}^n$

$$v^\top(\nabla g(x^*))\nabla^2 f(y^*)\nabla g(x^*)v\le 0. \tag{10}$$

Finally, by assumption the $d\times n$ Jacobian matrix $\nabla g(x^*)$ has rank d, which implies that for all $w\in\mathbb{R}^d$ there exists $v\in\mathbb{R}^n$ such that $w=\nabla g(x^*)v$. Applying this to equation 11 we have that every $w\in\mathbb{R}^d$ satisfies $w^\top\nabla^2 f(y^*)w\le 0$ i.e. $\nabla^2 f(y^*)$ is negative semi-definite. This is the required second-order condition to conclude that y^* is a local maximum of f. The argument that local minima of $f(g)$ correspond to local minima of f is identical, except that we use the fact that the Hessian is positive-semidefinite. □

Lemma A.8. *Let π_ϕ be a non-singular policy. Suppose that ϕ^* is a local maximum of $J(\phi)$ and $a^*=\pi_{\phi^*}(s)$. Then a^* is a local maximum of $Q_\theta(s,a)$.*

Proof. Observe that $J(\phi)=Q(s,\pi_\phi(s))$. Applying Lemma 3.2 with $f(a)=Q(s,a)$ and $g(\phi)=\pi_\phi(s)$ yields the desired result. □

Lemma A.9. *Suppose that ϕ^* is a local maximum of $J(\phi)$ and $a^* = \pi_{\phi^*}(s)$. Then a^* is a local maximum of $Q_\theta(s, a)$.*

Proof. By equation 2 we have

$$0 = \nabla J(\phi^*) = \nabla_a Q_\theta(s, a)|_{a=a^*} \nabla_\phi \pi_\phi(s)|_{\phi=\phi^*}.$$

Next by Definition 3.1 the Jacobian matrix $\nabla_\phi \pi_\phi(s)|_{\phi=\phi^*}$ has full row-rank d. Hence in order for the matrix-vector product $\nabla_a Q_\theta(s, a)|_{a=a^*} \nabla_\phi \pi_\phi(s)|_{\phi=\phi^*}$ to be zero, it must be that $\nabla_a Q_\theta(s, a)|_{a=a^*} = 0$. Thus, a^* is a first-order stationary point of $Q(s, a)$. To prove that a^* is a local maximum we compute the Hessian of $J(\phi)$. To do so, we introduce the following notation:

$$\begin{aligned} H_Q &= \nabla_a^2 Q_\theta(s, a)|_{a=a^*} \\ \nabla Q_k &= (\nabla_a Q_\theta(s, a)|_{a=a^*})_k \\ D_\pi &= \nabla_\phi \pi_\phi(s)|_{\phi=\phi^*} \\ H_{\pi,k} &= \nabla_\phi^2 (\pi_\phi(s)|_{\phi=\phi^*})_k. \end{aligned}$$

With this notation the Hessian of $J(\phi)$ is given by:

$$\begin{aligned} \nabla^2 J(\phi) &= D_\pi^\top H_Q D_\pi + \underbrace{\sum_{k=1}^{d} \nabla Q_k H_{\pi,k}}_{=0} \\ &= D_\pi^\top H_Q D_\pi \end{aligned}$$

where the equality follows because the first-order stationary conditions imply $\nabla Q_k = 0$ for all k. By assumption we have that ϕ^* is a local maximum of $J(\phi)$, which implies that $\nabla^2 J(\phi)$ is negative semi-definite i.e. $v^\top \nabla^2 J(\phi) v \leq 0$ for all $v \in \mathbb{R}^n$. Plugging in the above formula for $\nabla^2 J(\phi)$ we conclude that for all $v \in \mathbb{R}^n$

$$v^\top D_\pi^\top H_Q D_\pi v \leq 0. \tag{11}$$

Finally, by Definition 3.1 we know that the $d \times n$ Jacobian matrix D_π has rank d, which implies that for all $w \in \mathbb{R}^d$ there exists $v \in \mathbb{R}^n$ such that $w = D_\pi v$. Applying this to equation 11 we have that every $w \in \mathbb{R}^d$ satisfies $w^\top H_Q w \leq 0$ i.e. H_Q is negative semi-definite. This is the required second-order condition to conclude that a^* is a local maximum of $Q_\theta(s, a)$. □

Proposition A.10. *Let Q_θ be (M, λ)-regular, let $\epsilon < \frac{3\lambda}{M}$, let a^* be a local maximum of Q_θ, and let a be an action with $\|a - a^*\| < \epsilon$. Then we have $Q_\theta(s, a) < Q_\theta(s, a^*)$.*

Proof. Let $v = a^* - a$. By Taylor's theorem we have that there exists a point b on the line from a^* to a such that

$$\begin{aligned} Q_\theta(s, a) &= Q_\theta(s, a^*) + \nabla_a Q_\theta(s, a)|_{a=a^*} + \frac{1}{2} v^\top \nabla_a^2 Q_\theta(s, b)|_{a=a^*} v + \frac{1}{6} \sum_{i,j,k=1}^{d} (\nabla_a^3 Q_\theta(s, b))_{ijk} v_i v_j v_k \\ &\leq Q_\theta(s, a^*) - \frac{\lambda}{2}\|v\|^2 + \frac{M}{6}\|v\|^3 \\ &= Q_\theta(s, a^*) - \frac{\epsilon^2 \lambda}{2} + \frac{\epsilon^3 M}{6}. \end{aligned}$$

where the inequality follows from the first-order stationarity conditions at a^* and Definition 3.4. Plugging in $\epsilon < \frac{3\lambda}{M}$ completes the proof by ensuring that the negative degree-two term dominates the positive degree three-term. □

Theorem A.11 (*Counteractive Values of Policy Tails*)**.** *Let π_ϕ be a non-singular policy and Q_θ have (M, λ)-strict local maxima. Let $\epsilon < \frac{3\lambda}{M}$. Let ϕ^* be a local maximum of $J(\phi)$, and $a^* = \pi_{\phi^*}(s)$. Let a be any action with $\|a - a^*\| \leq \epsilon$. Then $Q_\theta(s, a) < Q_\theta(s, a^*) - \frac{\lambda}{2}\|a - a^*\|$.*

Proof. By Lemma 3.3 $a^* = \pi_{\phi^*}(s)$ is a local maximum of $Q_\theta(s, a)$. Hence, by Proposition 3.5 we have $Q_\theta(s, a) < Q_\theta(s, a^*)$ as desired. □

Lemma A.12. *Let π_ϕ be a non-singular policy. Let μ, σ be the mean and covariance vectors for the policy π_ϕ. If the point (μ, σ) is a stationary point of $F(\mu, \sigma)$, then ϕ is a stationary point of $J(\phi)$.*

Proof. Let $g(\phi)$ be the function mapping $\phi \mapsto (\mu, \sigma)$ where μ and σ are the respective mean and standard deviation vectors of π_ϕ. Then $-J(\phi) = F(g(\phi))$. Applying Lemma 3.2 completes the proof. □

Lemma A.13. *The derivatives of F are given by*

$$\nabla_\mu F(\mu, \sigma) = \mathbb{E}_{a\sim\gamma_{\mu^*,\sigma^*}}[\nabla Q(s, a)]. \tag{12}$$

$$\frac{\partial}{\partial \sigma_i} F(\mu, \sigma) = \frac{1}{\sigma_i} + \sigma_i \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}\left[\nabla^2_{ii} Q(s, a)\right] \tag{13}$$

Proof. We first compute the gradient with respect to μ, using the fact that in the formula equation 4 only the term $\mathbb{E}_{\gamma_{\mu,\sigma}}[Q(s, a)]$ depends on μ. To do so we use the fact that a Gaussian vector $a \sim \gamma_{\mu,\sigma}$ can equivalently be sampled by drawing $z \sim \mathcal{N}(0, I)$ and setting $a = \mu + \sigma \odot z$ where $\odot$ denotes the entrywise product.

$$\begin{aligned}
\nabla_\mu F(\mu) &= \nabla_\mu \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}[Q(s, a)] \\
&= \nabla_\mu \mathbb{E}_{z\sim\mathcal{N}(0,I)}[Q(s, \mu + \sigma \odot z)] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I)}[\nabla Q(s, \mu + \sigma \odot z)] \\
&= \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}[\nabla Q(s, a)]
\end{aligned}$$

Setting the gradient equal to zero yields equation 5.

For the gradient with respect to σ_i we will need Stein's lemma, which states that for $x \sim \mathcal{N}(0, 1)$ we have $\mathbb{E}[xf(x)] = \mathbb{E}[f'(x)]$.

$$\begin{aligned}
\frac{\partial}{\partial \sigma_i} F(\mu, \sigma) &= \frac{1}{\sigma_i} + \frac{\partial}{\partial \sigma_i} \mathbb{E}_{z\sim\mathcal{N}(0,I)}[Q(s, \mu + \sigma \odot z)] \\
&= \frac{1}{\sigma_i} + \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\frac{\partial}{\partial \sigma_i} Q(s, \mu + \sigma \odot z)\right] \\
&= \frac{1}{\sigma_i} + \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[z_i \nabla_i Q(s, \mu + \sigma \odot z)\right] \\
&= \frac{1}{\sigma_i} + \sigma_i \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}\left[\nabla^2_{ii} Q(s, a)\right]
\end{aligned}$$

where the final equality follows from Stein's lemma. Setting the above formula to zero and rearranging gives us equation 13. □

Lemma A.14. *Let Q be (M, λ)-regular with $\lambda^4 \geq 8M^2$. Let the radius ρ defining K be given by $\rho = \frac{\lambda}{2M}$. Let (μ, σ) be a point on the boundary of K, then*

$$(\mu - a^*)^\top \nabla_\mu F(\mu, \sigma) < 0$$

Proof. By Equation (5) we have

$$\begin{aligned}
\nabla_\mu F(\mu, \sigma) &= \mathbb{E}_{a\sim\gamma_{\mu,\sigma}}[\nabla Q(s, a)] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I)}[\nabla Q(s, \mu + \sigma \odot z)]
\end{aligned}$$

Let $v = \mu + \sigma \odot z - a^*$ and $w = \mu - a^*$. Next, we apply Taylor's theorem to $\nabla Q(s, \mu + \sigma \odot z)$ to obtain

$$\begin{aligned}
w^\top \nabla Q(s, \mu + \sigma \odot z) &= \underbrace{\nabla Q(s, a^*)}_{=0} + w^\top \nabla^2 Q(s, a^*) v + \frac{1}{2} \sum_{i,j,k} \nabla_{ijk} Q(s, \xi) w_i v_j v_k \\
&\leq w^\top \nabla^2 Q(s, a^*) v + \frac{M}{2} \|w\| \|v\|^2.
\end{aligned} \tag{14}$$

Note that $\mathbb{E}[z] = 0$ and the coordinates of z are independent. Hence $\mathbb{E}[v] = \mu - a^* = w$. Additionally $\mathbb{E}[\|v\|^2] = \|\sigma\|^2 + \|w\|^2$. Plugging these expectations into equation 14 yields

$$\begin{aligned} w^\top \nabla_\mu F(\mu, \sigma) &\leq w^\top \nabla^2 Q(s, a^*) w + \frac{M}{2}\|w\|\mathbb{E}[\|v\|^2] \\ &\leq -\lambda\|w\|^2 + \frac{M}{2}\|w\|\left(\|\sigma\|^2 + \|w\|^2\right) \\ &\leq -\lambda\rho^2 + M\rho\left(\frac{2}{\lambda^2} + \rho^2\right). \end{aligned}$$

Let $p(\rho) = -\lambda\rho^2 + M\rho\left(\frac{2}{\lambda^2} + \rho^2\right)$. We have shown that $w^\top \nabla_\mu F(\mu, \sigma) \leq p(\rho)$. Observe that $p(\rho) = \rho\left(M\rho^2 - \lambda\rho + \frac{2M}{\lambda^2}\right)$. This polynomial has a root at $\rho = 0$, as well as two more given by the quadratic formula applied to the inner degree two factor:

$$\rho = \frac{\lambda^2 \pm \sqrt{\lambda^4 - 8M^2}}{2M\lambda}.$$

These roots will both be real as long as $\lambda^4 \geq 8M^2$. In this case, $p(\rho) < 0$ between the two roots. Hence, setting $\rho = \frac{\lambda}{2M}$ guarantees that $w^\top \nabla_\mu F(\mu, \sigma) < 0$ on the boundary of K. □

Lemma A.15. *Let Q be (M, λ)-regular with $\lambda > 4M\sqrt{d}$. Let $\rho = \lambda/2M$. Let $L > 2\lambda + \frac{2M\sqrt{d}}{\sqrt{\lambda}}$ and assume $\nabla_{ii} Q(s, a^*) \geq -L$. Then for each i the gradient of F points inward at the boundary of K i.e.*

$$\partial_{\sigma_i} F(\mu, 2/\sqrt{\lambda}) < 0 \text{ and } \partial_{\sigma_i} F(\mu, 1/(\sqrt{2L})) > 0$$

Proof. We begin with the right hand end of the interval defined by K. Let $\sigma_i^2 = \frac{4}{\lambda}$. First we bound the value of the second derivative of Q. Let $v = \mu - a^* + \sigma \odot z$. Then by Taylor's theorem we have

$$\begin{aligned} \nabla^2_{ii} Q(s, \mu + \sigma \odot z) &= \nabla^2_{ii} Q(s, a^*) + \nabla(\nabla^2_{ii} Q(s, \xi)) \cdot v \\ &\leq -\lambda + M\|v\| \\ &\leq -\lambda + M(\|\mu - a^*| + \|\sigma \odot z\|) \end{aligned}$$

Taking expectations over $z \sim \mathcal{N}(0, I)$ yields

$$\begin{aligned} \mathbb{E}\left[\nabla^2_{ii} Q(s, \mu + \sigma \odot z)\right] &\leq -\lambda + M(\rho + \mathbb{E}[\|\sigma \odot z\|] \\ &\leq -\lambda + M(\rho + \max_j \sigma_j \mathbb{E}[\|z\|] \\ &\leq -\lambda + M(\rho + \max_j \sigma_j \sqrt{\mathbb{E}[\|z\|^2]}) \\ &\leq -\lambda + M(\rho + \max_j \sigma_j \sqrt{d}) \end{aligned} \tag{15}$$

where the second-to-last inequality follows from Jensen's inequality applied to $\mathbb{E}[\|z\|] = \mathbb{E}[\sqrt{\|z\|^2}]$. By Equation (13) we have

$$\frac{\partial F}{\partial \sigma_i} = \frac{1}{\sigma_i} + \sigma_i \mathbb{E}_{z \sim \mathcal{N}(0,I)}\left[\nabla^2_{ii} Q(s, \mu + \sigma \odot z)\right]$$

Plugging equation 15 into the above formula, and recalling that $\sigma_i^2 = \frac{4}{\lambda}$, $\sigma_j^2 \leq \frac{4}{\lambda}$, and $\rho = \frac{\lambda}{2M}$ gives us

$$\begin{aligned} \frac{\partial F}{\partial \sigma_i} &\leq \frac{\sqrt{\lambda}}{2} - \sqrt{\lambda} + \frac{2}{\sqrt{\lambda}} M\sqrt{d} \\ &= -\frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} M\sqrt{d} \\ &< 0 \end{aligned}$$

where the final inequality follows from $\lambda > 4M\sqrt{d}$.

We now turn the left-hand side of the interval i.e. $\sigma_i = 1/(\sqrt{2L})$. As before by Taylor's theorem we have

$$\begin{aligned}\mathbb{E}[\nabla^2_{ii} Q(s, \mu + \sigma \odot z)] &= \nabla^2_{ii} Q(s, a^*) + \mathbb{E}[\nabla(\nabla^2_{ii} Q(s, \xi)) \cdot v] \\ &\geq -L - M\mathbb{E}[\|v\|] \\ &\geq -L - M(\rho + \max_j \sigma_j \sqrt{d}) \end{aligned} \tag{16}$$

Plugging equation 16 into equation 13 yields

$$\frac{\partial F}{\partial \sigma_i} \geq \frac{1}{\sigma_i} + \sigma_i \left(-L - 2\lambda - \frac{2M\sqrt{d}}{\sqrt{\lambda}} \right)$$

Using the fact that $L > 2\lambda + \frac{2M\sqrt{d}}{\sqrt{\lambda}}$ and $\sigma_i = \frac{1}{\sqrt{2L}}$ yields $\frac{\partial F}{\partial \sigma_i} > 0$ as desired. □

Lemma A.16. *Let Q be (M, λ)-regular with $\lambda > \max\{2\sqrt{M}, 4M\sqrt{d}\}$. Assume $\nabla_{ii} Q(s, a^*) \geq -L$ for $L > 2\lambda + \frac{2M\sqrt{d}}{\sqrt{\lambda}}$. Let $\rho = \lambda/(2M)$. Then there exists a stationary point (μ^*, σ^*) of $F(\mu, \sigma)$ within K.*

Proof. First observe that on the set K the gradient ∇F is continuous. This follows from the fact that convolution of the bounded, integrable function Q with a Gaussian leads to a C^∞ function so long as the variance is nonzero (which it is on K). By Lemma 3.9 and Lemma 3.10 the gradient of F points into the set K at the boundary of K. Now consider the map $G(\mu, \sigma) = (\mu, \sigma) + \epsilon \nabla F(\mu, \sigma)$. Since the ∇F points inward at the boundary of K we can set $\epsilon > 0$ sufficiently small such that G maps K to itself. Since ∇F is continuous, so is G. Thus we can apply the Brouwer Fixed Point Theorem to K to conclude that there exists a fixed point (μ^*, σ^*) of G in K. This directly implies that $(\mu^*, \sigma^*) + \epsilon \nabla F(\mu^*, \sigma^*) = (\mu^*, \sigma^*)$. Since $\epsilon > 0$, $\nabla F(\mu^*, \sigma^*) = 0$ as desired. □

Lemma A.17. *Let Q be (M, λ)-regular and let ρ, λ, M be as in Lemma 3.11. Let $(\mu^*, \sigma^*) \in K$ be a stationary point of F. Then we have that all eigenvalues of $\nabla^2 Q(s, \mu^*) < -\lambda/2$ and*

$$|\nabla_i Q(s, \mu^*)| \leq \frac{2dM}{\lambda}$$

Proof. Let $\delta(z) = \sigma \odot z$. By equation 5 we have that at a stationary point

$$0 = \nabla F(\mu^*, \sigma^*) = \mathbb{E}[\nabla Q(s, \mu^* + \delta(z))]$$

Computing a Taylor expansion for the i-th of μ^* and noting that $\mathbb{E}[z] = 0$ and $\mathbb{E}[z_i z_j] = \delta_{ij}$ we have

$$\begin{aligned} 0 &= \mathbb{E}[\nabla_i Q(s, \mu^* + \delta(z))] \\ &= \nabla_i Q(s, \mu^*) + \underbrace{\mathbb{E}\left[\sum_j \nabla^2_{ij} Q(s, \mu^*) z_j \sigma^*_j\right]}_{=0} + \frac{1}{2}\mathbb{E}\left[\sum_{j,k} \nabla^3_{ijk} Q(s, \xi) z_j z_k \sigma^*_j \sigma^* k\right] \\ &= \nabla_i Q(s, \mu^*) + \frac{1}{2}\sum_j \nabla^3_{ijj} Q(s, \xi)(\sigma^*_j)^2 \end{aligned}$$

Re-arranging we have the bound

$$\begin{aligned} |\nabla_i Q(s, \mu^*)| &\leq \frac{M}{2}\|\sigma^*\|^2 \\ &\leq \frac{2dM}{\lambda} \end{aligned}$$

where we have used the fact that $(\sigma^*_i)^2 \leq \frac{4}{\lambda}$ in the set K.

Next recall that because Q is (M, λ)-regular, for any unit vector v we have $v^\top \nabla^2 Q(s, a^*)v \leq -\lambda$. Thus, taking yet another Taylor expansion we get that for any unit vector v

$$\begin{aligned} v^\top \nabla^2 Q(s, \mu^*)v &= v^\top \nabla^2 Q(s, a^*)v \\ &\quad + \sum_{ijk} \nabla^3_{ijk} Q(s, \xi) v_i v_j (\mu^* - a^*)_k \\ &\leq -\lambda + M\|\mu^* - a^*\| \\ &\leq -\lambda + M\rho \\ &= -\frac{\lambda}{2} \end{aligned}$$

where we have used that $\rho = \frac{\lambda}{2M}$. □