Journal of Machine Learning Research 25 (2024) 1-58 Submitted 11/21; Published 1/24

On the Sample Complexity and Metastability of Heavy-tailed Policy Search in Continuous Control

Amrit Singh Bedi* AMRITBEDI@UCF.EDU
Department of Computer Science,
University of Central FLorida,
Orlando, FL, USA

Anjaly Parayil* APARAYIL@MICROSOFT.COM
Microsoft India, Bengaluru

Junyu Zhang JUNYUZ@NUS.EDU.SG
Department of Industrial Systems Engineering and Management
National University of Singapore
Singapore, 119077

Mengdi Wang MENGDIW@PRINCETON.EDU
Department of Electrical Engineering
Center for Statistics and Machine Learning
Princeton University/Deepmind, Princeton, NJ 08544

Alec Koppel† ALEC.KOPPEL@JPMCHASE.COM
JP Morgan AI Research
383 Madison Ave, New York, NY 10017

Editor: Alexandre Proutiere

Abstract

Reinforcement learning is a framework for interactive decision-making with incentives sequentially revealed across time without a system dynamics model. Due to its scaling to continuous spaces, we focus on policy search where one iteratively improves a parameterized policy with stochastic policy gradient (PG) updates. In tabular Markov Decision Problems (MDPs), under *persistent exploration* and suitable parameterization, global optimality may be obtained. By contrast, in continuous space, the non-convexity poses a pathological challenge as evidenced by existing convergence results being mostly limited to stationarity or arbitrary local extrema. To close this gap, we step towards persistent exploration in continuous space through policy parameterizations defined by distributions of heavier tails defined by tail-index parameter α, which increases the likelihood of jumping in state space. Doing so invalidates smoothness conditions of the score function common to PG. Thus, we establish how the convergence rate to stationarity depends on the policy's tail index α, a Hölder continuity parameter, integrability conditions, and an exploration tolerance parameter introduced here for the first time. Further, we characterize the dependence of the set of local maxima on the tail index through an exit and transition time analysis of a suitably defined Markov chain, identifying that policies associated with Lévy Processes of a heavier tail converge to wider peaks. This phenomenon yields improved stability to perturbations

in supervised learning, which we corroborate also manifests in improved performance of policy search, especially when myopic and farsighted incentives are misaligned.

Keywords: Policy gradient algorithm, heavy-tailed distributions, non-convex optimization.

1. Introduction

In reinforcement learning (RL), an autonomous agent sequentially interacts with its environment and observes rewards incrementally across time (Sutton et al., 2017). This framework has gained attention in recent years for its successes in continuous control (Schulman et al., 2015b; Lillicrap et al., 2016; Bedi et al., 2022; Chakraborty et al., 2023), web services (Zou et al., 2019), personalized medicine (Kosorok and Moodie, 2015), among other contexts. Mathematically, it may be described by a Markov Decision Process (MDP) (Puterman, 2014), in which an agent seeks to select actions so as to maximize the long-term accumulation of rewards, known as the value. The key distinguishing point of RL from classical optimal control is its ability to discern control policies without a system dynamics model.

Algorithms for RL either operate by Monte Carlo tree search (Guo et al., 2014), approximately solve Bellman's equations (Bellman, 1957; Watkins and Dayan, 1992), or conduct direct policy search (Williams, 1992). While the first two approaches may have lower variance and converge faster (Even-Dar et al., 2003; Devraj and Meyn, 2017), they typically require representing a tree or Q-function for every state-action pair, which is intractable in continuous space. For this reason, we focus on policy gradient (PG).

Policy search hinges upon the Policy Gradient Theorem (Sutton et al., 2000), which expresses the gradient of the value function with respect to policy parameters as the expected value of the product of the score function of the policy and its associated Q function. Its performance has historically been understood only asymptotically (Konda and Borkar, 1999; Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009) via tools from dynamical systems (Kushner and Yin, 2003; Borkar, 2008). More recently, the non-asymptotic behavior of policy search has come to the fore. In continuous space, its finite-time performance has been linked to stochastic search over non-convex objectives (Bottou et al., 2018), whose $\mathcal{O}(1/\sqrt{k})$ convergence rate to stationarity is now clear (Bhatt et al., 2019; Zhang et al., 2020c). However, it is challenging to discern the quality of a given limit point under this paradigm.

By contrast, for tabular MDPs, i.e., those with state and action spaces defined by finite discrete sets, stronger results have appeared (Bhandari and Russo, 2019; Zhang et al., 2020b; Agarwal et al., 2020c): linear convergence to *global* optimality for tabular or softmax parameterizations. A critical enabler of these recent innovations in finite MDPs is a persistent exploration condition: the initial distribution over the states is uniformly lower bounded away from null, under which the current policy may be shown to assign strictly positive likelihood to the optimal action over the entire state space (Mei et al., 2020b)[Lemma 9]. This concept of exploration is categorically different from notions common to bandits, i.e., optimism in the face of uncertainty (Thompson, 1933; Lai and Robbins, 1985; Jaksch et al., 2010; Russo and Van Roy, 2018), and instead echoes persistence of excitation in systems identification (Narendra and Annaswamy, 1987, 2012). Under this condition, then, a version

. *Equal contributions.

. †Work completed while at the U.S. Army Research Laboratory in Adelphi, MD 20783.

of gradient dominance (Luo and Tseng, 1993) (known also as Polyak-Łojasiewicz inequality (Lojasiewicz, 1963; Polyak, 1963)) holds, as derived in (Agarwal et al., 2019; Mei et al., 2020b,a). This result enables such global improvement bounds. Unfortunately, translating this condition to continuous space is elusive, as many common distributions in continuous space may fail to be integrable if their likelihood is lower bounded away from null over the entire (not necessarily compact) state space. Thus, the following question is our focus:

Can one nearly satisfy persistent exploration in MDPs over continuous spaces through appropriate policy parameterizations, and in doing so, mitigate the pathologies of non-convexity?

In this work, we step towards an answer by studying policy parameterizations defined by heavy-tailed distributions (Bryson, 1974; Focardi and Fabozzi, 2003), which includes the family of Lévy Processes common to fractal geometry (Hutchinson, 1981; Mandelbrot, 1982), finance (Taleb, 2007; Taylor and Williams, 2009), pattern formation in nature (Avnir et al., 1998), and networked systems (Clauset et al., 2009). By employing a heavy-tailed policy, the induced transition dynamics will be heavy-tailed, and hence at increased likelihood of jumping to non-adjacent states. That policies (Chou et al., 2017; Kobayashi, 2019) or stochastic policy gradient estimates (Garg et al., 2021) associated with heavy-tailed distributions exhibit improved coverage of continuous space is well-documented experimentally. Here we seek a more rigorous understanding of in what sense this impacts performance may be formalized through *metastability*, the study of how a stochastic process transitions between its equilibria. This marks a step towards persistent exploration in continuous space, but satisfying it precisely remains beyond our grasp.

Historically, heavy-tailed distributions have been recently employed in non-convex optimization to perturb stochastic gradient updates by α-stable Lévy noise (Gurbuzbalaban et al., 2020; Simsekli et al., 2020b), inspired by earlier approaches where instead Gaussian noise perturbations are used (Pemantle et al., 1990; Gelfand and Mitter, 1991). Doing so has notably been shown to yield improved stability to perturbations in parameter space since SGD perturbed by heavy-tailed noise can converge to local extrema with more volume, which in supervised learning is experimentally associated with improved generalization (Neyshabur et al., 2017; Zhu et al., 2018; Advani et al., 2020), and has given rise to a nascent generalization theory based on the tail index of the parameter estimate's limiting distribution (Simsekli et al., 2020a,b). Rather than perturbing stochastic gradient updates, we directly parameterize policies as heavy-tailed distributions, which induces heavy-tailed gradient noise. Doing so invalidates several aspects of existing analyses of PG in continuous spaces (Bhatt et al., 2019; Zhang et al., 2020c). Thus, our main results are:

- We present a few heavy-tailed policy parameterizations that may be used in lieu of a Gaussian policy for continuous space, which can prioritize selecting actions far from the distribution's center (Sec. 3), and discuss how policy search manifests for this setting (Sec. 4);

- We establish the attenuation rate of the expected gradient norm of the value function when the score function is Hölder continuous, and may be unbounded but whose moment is integrable with respect to the policy (Theorem 5.4). This statement generalizes previous results that break for non-compact spaces (Bhatt et al., 2019; Zhang et al., 2020c), and further requires introducing an exploration tolerance parameter (Definition

5.1) to quantify the subset of the action space where the score function is absolutely bounded;

- In sec. 5.2, by rewriting the PG under a heavy-tailed policy as a discretization of a Lévy Process, we establish that the time required to exit a (possibly spurious) local extrema decreases polynomially with heavier tails (smaller α), and the width of a peak's neighborhood (Theorem 5.11). Further, the proportion of time required to transition from one local extrema to another depends polynomially on its width, which decreases for smaller tail index (Theorem 5.12). By contrast, lighter-tailed policies exhibit transition times depending exponentially on the volume of an extrema's neighborhood;

- Experimentally, we observe that policies associated with heavy-tailed distributions converge more quickly in problems that are afflicted with multiple spurious stationary points, which are especially common when myopic and farsighted incentives are in conflict with one another (Sec. 6).

2. Additional Context and Related Work

Efforts to circumvent the necessity of persistent exploration and obtain rates to global optimality have been considered in both finite and continuous space. In tabular settings, one may incorporate proximal-style updates in order to leverage a performance-difference lemma (Kakade and Langford, 2002), which has given rise to recent analyses of natural policy gradient (Schulman et al., 2017; Tomar et al., 2020; Lan, 2021). Translating these results to the continuum remains an open problem.

Alternatively, in continuous space, one may hypothesize the policy parameterization is a neural network whose size grows unbounded with the number of samples processed (Wang et al., 2019; Liu et al., 2019). Doing so belies the fact that typically a parameterization has fixed dimension during training. Alternatively, one may impose a "transferred compatible function approximation error" condition that mandates the ability to sample from the occupancy measure of the optimal policy to ensure sufficient state space coverage (Agarwal et al., 2019; Liu et al., 2020), which is difficult to perform in practice.

Two additional lines of effort are pertinent to the objective of this work. The first is state aggregation (Singh et al., 1995), in which one hypothesizes a large but finite space admits a representation in terms of low-dimensional features, such as tile coding (Sutton et al., 2017) or interpolators (Tsitsiklis and Van Roy, 1996). A long history of works seeks to discern such state aggregations adaptively (Bertsekas and Castanon, 1989; Singh et al., 1995; Dean and Givan, 1997; Jiang et al., 2015; Duan et al., 2019; Misra et al., 2020).

Such representations can be used in, e.g., policy search (Agarwal et al., 2020a; Russo, 2020) or value iteration (Arumugam and Van Roy, 2020) to obtain refined convergence behavior that depends only on the properties of the representation rather than the underlying state or action spaces. Finding this representation is itself not necessarily easier than solving the original MDP, however. See, for instance, (Agarwal et al., 2020b; Modi et al., 2021), where a variety of structural assumptions and representations are discussed. In this work, we assume such a feature map is fixed at the outset of training as part of one's specification of a policy parameterization.

The other research thrust broadly related to this work is information-theoretic exploration that seeks comprehensive state-space coverage. The simplest way to achieve this goal is to simply replace the cumulative return with an objective that prioritizes state-space coverage, such as the entropy of the occupancy measure induced by a policy (Savas et al., 2018; Hazan et al., 2019; Zhang et al., 2020a). This goal does not necessarily result in good performance with respect to the cumulative return, however. Alternatively, exploration bonuses in the form of upper-confidence bound (Lai and Robbins, 1985; Jaksch et al., 2010), Thompson sampling (Thompson, 1933), information-directed sampling (Russo and Van Roy, 2018), among other strategies (see (Russo et al., 2018) for a thorough review), have percolated into RL in various forms.

For instance, incorporating randomized perturbations/exploration bonuses into value iteration (Osband et al., 2016), Q-learning (Jin et al., 2018), or augmenting a policy's variance hyper-parameters in policy search in a manner reminiscent of line-search for step-size selection (Papini et al., 2020). Alternative approaches based on Thompson sampling (Gopalan and Mannor, 2015; Osband and Van Roy, 2017) and various Bayesian models of the value function (Bellemare et al., 2017; Azizzadenesheli et al., 2018) have been considered, as well as approaches which subsume exploration goals into the choice of the aforementioned state aggregator (Agarwal et al., 2020a; Modi et al., 2021). Our approach contrasts with approaches that inject suitably scaled randomness into an RL update, by searching over a policy class that is itself more inherently random.

Notations: All the norms $\|\cdot\|$ are Euclidean norm unless otherwise stated.

3. Markov Decision Problems

In RL, an agent evolves through states $s \in \mathcal{S}$ selecting actions $a \in \mathcal{A}$, which causes transitions to another state s' to occur according to a Markov transition density $\mathbb{P}(s'|s,a)$ and a reward $r(s,a)$ is revealed by the environment to inform its merit. Formally, an MDP (Puterman, 2014) consists of the tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where continuous state $\mathcal{S} \subseteq \mathbb{R}^q$ and action spaces may be unbounded, i.e., Euclidean space in the appropriate dimension. We hypothesize that actions $a_t \sim \pi(\cdot|s_t)$ are selected according to a time-invariant distribution $\pi(a|s) := \mathbb{P}(a_t = a|s_t = s)$ called a policy determining the probability of action a when in state s. Define the value as the average long-term accumulation of reward (Bertsekas and Shreve, 2004):

$$V^\pi(s) = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, \pi\bigg]. \tag{3.1}$$

Moreover, γ is a discount factor that trades off the future relative to the present, s_0 denotes the initial state along trajectory $\{s_u, a_u, r_u\}_{u=0}^{\infty}$, and we abbreviate the instantaneous reward as $r_t = r(s_t, a_t)$. In (3.1), the expectation is with respect to randomized policy $a_t \sim \pi(\cdot|s_t)$ and state transition dynamics $s_{t+1} \sim \mathbb{P}(.|s_t, a_t)$ over times $t \geq 0$. We further define the action-value (known also as Q) function $Q^\pi(s,a)$ as the value conditioning on an initially selected action: $Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a, \pi\right]$. We focus on policy search over policies $\pi_{\boldsymbol{\theta}}(\cdot|s_t)$ parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$, which we estimate via maximizing the

expected cumulative returns (Sutton et al., 2017):

$$\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) := V^{\pi_{\boldsymbol{\theta}}}(s_0) \tag{3.2}$$

One difficulty in RL is that (3.2) is non-convex in parameters $\boldsymbol{\theta}$. Thus, finding a global optimizer is challenging even if the problem were deterministic. However, in the present context, the search procedure also interacts with the transition dynamics $\mathbb{P}(s'|s,a)$. Before delving into how one may iteratively and approximately solve (3.2), we present a few representative policy parameterizations.

Example 1 (Gaussian policy) The Gaussian policy is written as $\pi_{\boldsymbol{\theta}}(a|s) = \mathcal{N}(a|\phi(s)^\top \mathbf{x}, e^y)$, where the parameters $\boldsymbol{\theta} = [\mathbf{x}, y]$ determine the mean (centering) of a Gaussian distribution at $\phi(s)^\top \mathbf{x}$, and $e^y \geq \delta_0$ is the variance parametrized by y for some $\delta_0 > 0$. Here, $\phi(s)$ represents a feature map with $\|\phi(s)\| \leq S < \infty$ which maps continuous state s to a higher-dimension, i.e., $\phi : \mathcal{S} \to \mathbb{R}^d$.

Example 2 (Moderate-tailed policy) *A distribution whose tail decays at a slower rate than the Gaussian may better explore environment, which we define as* $\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sigma A_\alpha} \exp\left(-|a - \phi(s)^T\boldsymbol{\theta}|^\alpha/\sigma^\alpha\right)$ *with normalizing constant* $A_\alpha := \int \exp\left(-|x|^\alpha\right) dx < \infty$*, tail index* $\alpha \in [1,2]$ *determining the likelihood of tail events, and scale parameter* $\sigma > 0$*.*

We next introduce heavy-tailed policies, specifically, Lévy processes called α-stable distributions, which are historically associated with fractal geometry (Hutchinson, 1981), finance (Focardi and Fabozzi, 2003), and network science (Barabási et al., 2003).

Example 3 (Lévy Process Policy) Symmetric α stable, $\mathcal{S}\alpha\mathcal{S}$ distributions generalize Gaussians with $\alpha \in (0,2]$ as the tail index determining the decay rate of the distribution's tail (Nguyen et al., 2019b). Denote random variable $\mathbf{X} \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ with associated characteristic function $\mathbb{E}\left[e^{i\omega\mathbf{X}}\right] = e^{-\sigma|\omega|^\alpha}$ and scale parameter $\sigma \in (0,\infty)$. For non-integer (fractional) value of α, there is no closed form expression but the density decays at a rate $1/|a|^{1+\alpha}$, and is referred to as fractal (Mandelbrot, 1982). In finance, $\mathcal{S}\alpha\mathcal{S}$ distributions have been associated with "black swan" events (Taleb, 2007; Taylor and Williams, 2009). For $\alpha = 2$, it reduces to a Gaussian, and for $\alpha = 1$ it is a Cauchy whose parametric form is: $\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{e^y \pi(1+((a-\phi(s)^\top \mathbf{x})/e^y)^2)}$, where, $\boldsymbol{\theta} = [\mathbf{x}, y]$, $\phi(s)^\top \mathbf{x}$ is the mode of the distribution and e^y is the scaling parameter.

With a few policy choices of detailed, we delve into their relative merits and drawbacks. Intuitively, policies that select actions far from a learned mean parameter over actions may better explore the space, which exhibits outsize importance when near and long-term incentives of the MDP are misaligned (Misra et al., 2020). More formally, persistent exploration has been identified in *tabular* MDPs as key to the ability to converge to the optimal policy using first-order methods (Agarwal et al., 2019; Mei et al., 2020b,a) and avoid spurious behavior. Persistent exploration formally ensures that under any initial distribution over s_0 in (3.1), the current policy assigns strictly positive likelihood to the optimal action over the entire state space (Mei et al., 2020b)[Lemma 9], under which a version of gradient dominance (akin to strong convexity) holds (Lemma 8). Interestingly, these results echo classical persistence of excitation in systems identification (Narendra and Annaswamy,

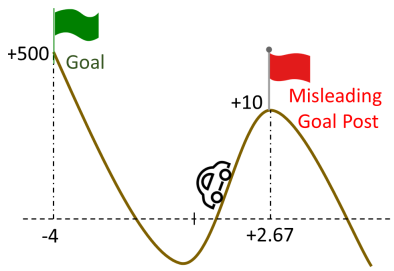


Figure 1: A continuous pathological Mountain Car environment with a low reward state and a bonanza atop a higher hill. Policies that do not incentivize exploration get stuck at the misleading goal.

1987, 2012). The stumbling block in translating these conditions from finite to continuous spaces is that many common distributions over unbounded continuous space may fail to be integrable if their likelihood is lower bounded away from null. As a step towards satisfying this condition, we seek to ensure that the induced transition dynamics under a policy are heavy-tailed, which increases the likelihood of jumping to cover more of the state space. Doing so may be accomplished by specifying a heavy-tailed policy (Example 2 - 3), whose likelihood approaches null slowly while still defining a valid distribution. That continuous space necessitates exploration to eventuate in suitable behavior may be illuminated through the Pathological Mountain Car (PMC) (cf. Fig. 1) introduced next, where a car is between two mountains.

Pathological Mountain Car. The environment consists of two goal posts, a less-rewarding goal at $s = 2.667$ with a reward of 10 and a bonanza at $s = -4.0$ of 500 units of reward. In Fig. 1, it is possible to get stuck at the lower peak and never reach the jackpot without sufficient exploration. Its potential pitfalls are illuminated experimentally in Sec. 6.

With the motivation clarified, we shift to illuminating that heavy-tailed policies, while encouraging actions far from the mean, may cause policy search directions to possibly be unbounded and non-smooth. These issues are the focus of Section 4.

4. Policy Gradient Methods

Policy gradient (PG) is an RL algorithm in which policy parameters in $\mathbb{R}^d$ are iteratively updated as approximate gradient ascent with respect to the value function (3.1). Its starting point is the Policy Gradient Theorem (Sutton et al., 2017), which expresses search directions in parameter space:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma} \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}}}(s,a) \cdot \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \,|\, s) \cdot \rho_{\boldsymbol{\theta}}(s,a) \cdot dsda \tag{4.1}$$

where $\rho_{\boldsymbol{\theta}}(s,a) = \rho_{\pi_{\boldsymbol{\theta}}}(s) \cdot \pi_{\boldsymbol{\theta}}(a \,|\, s)$ is a distribution called the *discounted state-action occupancy measure* defined as the product of the discounted state occupancy measure $\rho_{\pi_{\boldsymbol{\theta}}}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(s_k = s \,|\, s_0, \pi_{\boldsymbol{\theta}})$ and policy $\pi_{\boldsymbol{\theta}}(a \,|\, s)$. In (Sutton et al., 2000), both $\rho_{\pi_{\boldsymbol{\theta}}}(s)$ and $\rho_{\boldsymbol{\theta}}(s,a)$ are established as valid distributions. Despite this fact, the integral in (4.1) may

Algorithm 1 **H**eavy-tailed **P**olicy **G**radient (HPG)

1: **Initialize** : policy parameters $\boldsymbol{\theta}_0$, discount γ, step-size η, gradient $\mathbf{g}_0 = \mathbf{0}$, starting point (s_0, a_0)
Repeat for $k = 1, \ldots$
2: Starting from (s_0, a_0) , generate B_k trajectories $\tau_{k,i} = (s_0, a_0, s_1, a_1, \ldots s_{T_{k,i}}, a_{T_{k,i}})$ of length $T_{k,i} \sim \text{Geom}(1-\gamma^{1/2})$ with actions $a_u \sim \pi_{\boldsymbol{\theta}_k}(.|s_u)$
3: Compute policy gradient estimate $\mathbf{g}_k$ and update parameters $\boldsymbol{\theta}_k$:

$$\mathbf{g}_k \leftarrow \frac{1}{B_k}\sum_{i=0}^{B_k}\left[\sum_{t=0}^{T_{k,i}}\gamma^{t/2}\cdot R(s_t,a_t)\cdot\left(\sum_{\tau=0}^{t}\nabla\log\pi_{\boldsymbol{\theta}_k}(a_\tau\,|\,s_\tau)\right)\right], \quad \boldsymbol{\theta}_{k+1}\leftarrow\boldsymbol{\theta}_k+\eta\mathbf{g}_k$$

4: $k \leftarrow k+1$ **Until Convergence**
5: **Return:** $\boldsymbol{\theta}_k$

not exist due to the heavy-tailed nature of policy $\pi_{\boldsymbol{\theta}}(a\,|\,s)$. Therefore, we first present some preliminaries regarding (4.1)

Assumption 4.1 *The absolute value of the reward is uniformly bounded,* $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}|R(s,a)| \leq U_R < \infty$.

Assumption 4.2 *For any* s, $\boldsymbol{\theta}$, $\int_{\mathcal{A}}\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a\,|\,s)\|^2\cdot\pi_{\boldsymbol{\theta}}(a\,|\,s)da \leq B < \infty$, *where* B *is a finite constant.*

Assumption 4.2 is weaker than the standard almost-sure boundedness of the score function assumed in prior work (Zhang et al., 2020c, Assumption 3.1), which is restrictive, and not valid even for Gaussians (Example 1). To see this, write the norm of its score function as $\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(s,a)\| \leq \mathcal{O}\left(\|a\|+\|a\|^2\right)$, which grows unbounded when the support of the action space is infinite. The score function also is unbounded in Example 2. These subtleties motivate the relaxed condition in Assumption 4.2 which is valid regardless of a policy's tail index (Examples 1 - 3).

Next, we shift towards detailing PG. Under Assumptions 4.1 - 4.2, we establish that the integral in (4.1) is finite (Lemma A.1 in Appendix A.1). Thus, we employ it to compute search directions, which requires unbiased estimates of (4.1). To realize this estimate, conduct a Monte Carlo rollout of length $T_k \sim \text{Geom}(1-\gamma^{1/2})$, collect trajectory data $\tau_k = (s_0, a_0, s_1, a_1, \ldots, s_{T_k}, a_{T_k})$, and form the PG estimate (akin to (Baxter and Bartlett, 2001; Liu et al., 2020), except with a randomized horizon):

$$\hat{\nabla}_{\boldsymbol{\theta}}J(\boldsymbol{\theta}_k) = \sum_{t=0}^{T_k}\gamma^{t/2}\cdot R(s_t,a_t)\cdot\left(\sum_{\tau=0}^{t}\nabla\log\pi_{\boldsymbol{\theta}_k}(a_\tau\,|\,s_\tau)\right), \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta\hat{\nabla}_{\boldsymbol{\theta}}J(\boldsymbol{\theta}_k)\,. \tag{4.2}$$

where the parameter update for $\boldsymbol{\theta}_k$ is defined according to stochastic gradient ascent with step-size $\eta > 0$. The procedure for policy search along a trajectory is summarized as Algorithm 1, where in the pseudo-code, we permit mini-batching with batch-size B_k, but subsequently assume $B_k = 1$.

Next, we establish that the stochastic gradient $\hat{\nabla}_{\boldsymbol{\theta}}J(\boldsymbol{\theta})$ is an unbiased estimate of the true gradient $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$. As previously mentioned, almost sure boundedness

(Zhang et al., 2020c, Assumption 3.1) of the score function $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \,|\, s)$ does not even hold for the Gaussian (Example 1), which motivates the moment condition in Assumption 4.2. This alternate condition is employed to establish unbiasedness of (4.2) formalized next (see Appendix A.2 for proof).

Lemma 4.3 *Under the Assumptions 4.1-4.2, it holds that* $\mathbb{E}[\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}] = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.

An additional condition which is called into question of existing analyses of policy search is Lipschitz continuity (Zhang et al., 2020c; Liu et al., 2020) of the score function. In particular, heavy-tailed policies such as Examples 3 - 2 necessitate generalizing smoothness conditions to Hölder-continuity, as formalized next.

Assumption 4.4 *The score function function* $\nabla \log \pi_{\boldsymbol{\theta}}(\cdot)$ *is Hölder continuous with constants* $M > 0$ *and* $0 < \beta \leq 1$*, which implies that* $\|\nabla \log \pi_{\boldsymbol{\theta}_1}(\cdot) - \nabla \log \pi_{\boldsymbol{\theta}_2}(\cdot)\| \leq M \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\beta}$ *for all* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$.

Observe that the policy parameterization in Example 2 is not Lipschitz but Hölder continuous. In the next section, we formalize the convergence of (4.2), discerning the convergence rate to stationarity and metastability characteristics: the proportion of time the algorithm's limit points spent at wider versus narrower local extrema as a function of the tail index.

5. Convergence Analysis

We analyze the ability of PG [cf. (4.2)] to maximize the value function (3.2). As $J(\boldsymbol{\theta})$ is non-convex in the policy parameter $\boldsymbol{\theta}$, the best pathwise result one may hope for is convergence to stationarity unless additional structure is present. Thus, we first study sample complexity in terms of the rate of decrease of the expected gradient norm $\mathbb{E}[\|\nabla J(\boldsymbol{\theta}_k)\|]$, which we pursue under Assumptions 4.2-4.4 regarding the integrability of the norm of the score function with respect to the policy and Hölder continuity. This generality is necessitated by heavy-tailed policy parameterizations as previously mentioned, and has not been considered in prior works such as (Sutton et al., 2000; Zhang et al., 2020c,b; Liu et al., 2020).

Assumptions 4.2-4.4 present unique confounders to the RL setting that do not manifest in vanilla stochastic programming under relaxed smoothness conditions (Shapiro et al., 2009; Nemirovski et al., 2009). Specifically, they cause integrability and smoothness complications with respect to the occupancy measure $\rho_{\boldsymbol{\theta}}(s, a)$ induced by the MDP, which upends conditions on the objective and policy gradient in existing analyses. These complications are overcome in Lemmas 5.2 - 5.3, which first require partitioning the action space into sets where the score function is and is not almost surely bounded according to an exploration tolerance parameter (Definition 5.1), which is unique to this work. Next, we make precise this discussion, establishing the convergence rate to stationarity of (4.2). Later in this section, we formalize that iterates escape narrow extrema and tend to jump towards wider peaks.

5.1 Attenuation Rate of the Expected Gradient Norm

We first focus on convergence rates to stationarity. To do so, we begin by establishing that Assumption 4.4 regarding the Hölder continuity of the score function implies approximate

Hölder continuity on the overall policy gradient. First, we partition the action space according to when the score function is almost surely bounded and where it is integrable according via a constant $\lambda > 0$ defined next.

Definition 5.1 *(***Exploration Tolerance)** *Define as $\mathcal{A}(\lambda)$ the set of subsets of action space such that*

$$\mathcal{A}(\lambda) := \left\{ \mathcal{C} \subseteq \mathcal{A} \ : \ \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \cdot \pi_{\boldsymbol{\theta}}(a|s) da \leq \lambda, \forall s, \boldsymbol{\theta} \right\}. \tag{5.1}$$

Then, λ is the exploration tolerance parameter of a policy in an MDP with unbounded score function. Intuitively, $\mathcal{A}(\lambda)$ is the collection of all region of action space $\mathcal{C} \subseteq \mathcal{A}$, such that the expectation of score function under policy $\pi_{\boldsymbol{\theta}}(a|s)$ over region $\mathcal{A}\setminus\mathcal{C}$ is upper bounded by λ. And for the region in $\mathcal{A}(\lambda)$ associated with λ, we define an upper bound for the score function as

$$B(\lambda) = \inf_{\mathcal{C}\in\mathcal{A}(\lambda)} \sup_{(s,a)\in\mathcal{S}\times\mathcal{C}} \sup_{\boldsymbol{\theta}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\|. \tag{5.2}$$

Definition 5.1 induces a tradeoff between the restriction on the range of values an action may take by a subset $\mathcal{C} \subset \mathcal{A}$ with the scale of $B(\lambda)$. Observe that for the Cauchy (Example 3), constant $B(0)$ exists and is finite. A broader characterization of λ as a function of the policy is given in Appendix E.

Lemma 5.2 *Under Assumptions 4.1 - 4.4, with λ as in Definition 5.1, the policy gradient* (4.2) *satisfies*

$$\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\| \leq M_J \left[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\beta} + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \lambda \right], \tag{5.3}$$

for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ with

$$M_J := \max\left\{ \frac{2U_R M}{(1-\gamma)^2}, \frac{M_Q B^{1/2}}{1-\gamma} + \frac{U_R B(\lambda) M_\rho}{(1-\gamma)^2}, \frac{2U_R}{(1-\gamma)^2} \right\},$$

and $0 < \beta \leq 1$. Here, U_R denotes the reward upper bound from Assumption 4.1, $M_\rho = \frac{\sqrt{B}}{1-\gamma}$, and $M_Q = \frac{\gamma U_R M_\rho}{1-\gamma}$.

See Appendix B for proof. Lemma 5.2 generalizes a comparable statement regarding the Lipschitz continuity of the score function typically imposed to establish a Lipschitz property of the policy gradient. Next, we provide an intermediate Lemma 5.3 (see Appendix C for proof) crucial to establishing the main convergence rates to stationarity of Algorithm 1.

Lemma 5.3 *Under Assumptions 4.1 - 4.4, value function $J(\boldsymbol{\theta})$ satisfies the smoothness condition*

$$|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| \leq M_J \left[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{1+\beta} + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + \lambda \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \right] \tag{5.4}$$

for all $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \mathbb{R}^d$ and M_J is as defined in Lemma (5.2) *with exploration tolerance λ as in* (5.1).

Now, we formalize the convergence rate for Algorithm 1 as Theorem 5.4.

Theorem 5.4 *Under Assumptions 4.1-4.4, with objective J bounded above by J^*, and Hölder continuity parameter β bounded by the tail-index α as $\beta \in (0, \alpha - 1]$, under constant step-size selection $\eta = 1/K^{\frac{\beta}{\beta+1}}$, the policy gradient updates of $\boldsymbol{\theta}_k$ in Algorithm 1 [cf. (4.2)] converges to stationarity:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\Big] \leq \frac{a_\beta}{K^{\frac{\beta}{1+\beta}}} + \mathcal{O}(\lambda), \qquad a_\beta = \big((L_J)^{1/(\beta+1)}\big)\,(J^* - J(\boldsymbol{\theta}_1))^{\beta/(\beta+1)} \tag{5.5}$$

with problem-dependent constant L_J defined in (D.7), and exploration tolerance λ as in Def. 5.1.

Theorem 5.4 (proof in Appendix D) establishes that the iteration complexity of Algorithm 1 is $\mathcal{O}\big(1/\zeta^{1+(1/\beta)}\big)$ when $\lambda = \mathcal{O}(\zeta)$, where ζ is the accuracy parameter. This result contrasts the standard rate of $\mathcal{O}\left(1/\zeta^2\right)$ for non-convex optimization (Bottou et al., 2018), which restricts the policy parameterization to be Gaussian (Bhatt et al., 2019; Zhang et al., 2020c; Liu et al., 2020), i.e., $\alpha = 2$. This means that heavy-tailed parameterizations result in slower convergence; however, we note that the rate of decrease of the expected gradient norm may not comprehensively encapsulate the non-convex landscape of value function. An additional subtlety is the effect of continuous action spaces, which are partitioned into sets where the score function is and is not bounded in accordance with the exploration tolerance parameter λ (Def. 5.1). In existing analyses of literature (Zhang et al., 2020c; Paternain et al., 2020; Liu et al., 2020), the effect of λ is assumed null ($\lambda = 0$), which overlooks the effect of action space coverage during policy search. Next, we establish that this perceived slower rate of heavy-tailed policies is overruled by their tendency towards local extrema with wider peaks, under a hypothesis that they admit a representation as a discretization of a Lévy Process.

5.2 Metastability and Convergence to Wide Peaks

In the previous subsection, we established that the attenuation rate of the expected gradient norm for heavy-tailed policies is actually *slower* than the rate associated smoother policies. This fact seemingly contradicts prior experimental results which demonstrate that they tend towards policies that achieve higher reward more quickly (Garg et al., 2021). The nature of this confounder has to do with the fact that expected gradient norm may only characterize how close a policy is to stationarity, but not how quickly a policy moves from one stationary point to another.

To make sense of this quandary, we turn to characterizing (i) the time that Algorithm 1 takes to escape a (possibly spurious) local extremum, and (ii) how the proportion of time spent at a local maxima depends on its width and the policy's tail index. These results hinge upon introducing into RL for the first time of *metastability* of dynamical systems under the influence of weak random perturbations (Tzen et al., 2018). Similar results have been employed for SGD in the context of training neural networks in supervised learning (Nguyen et al., 2019b; Gurbuzbalaban et al., 2020); however, it is unclear how one neural parameterization induces gradient noise whose distribution has a heavier from another. By contrast, here, this aspect is directly determined by the policy parameterization's tail index,

Algorithm	Iter. complexity	Exit time (Def. 5.5)	Trans. time (Def. 5.6)
PG	$\mathcal{O}\left(1/\zeta^2\right)$	$\mathcal{O}\left(e^{-2J(a)/\epsilon^2}\right)$	$\mathcal{O}\left(e^{2(J(\bar{\boldsymbol{\theta}}_i)-J(\bar{\boldsymbol{\theta}}_i))/\epsilon^2}\right)$
HPG	$\mathcal{O}\left(1/\zeta^{1+\frac{1}{\beta}}\right)$	$\mathcal{O}\left(\frac{\alpha}{2}\frac{a^\alpha}{\epsilon^\alpha}\right)$	$\mathcal{O}\left(1/\epsilon^\alpha\right)$

Table 1: Summary of iteration complexity, exit time, and transition time results for vanilla PG and heavy-tailed PG, with ϵ as the jump process coefficient, and ζ as accuracy parameter for $\mathbb{E}[\|\nabla J(\boldsymbol{\theta}_k)\|] \leq \zeta$. Employing a policy with a faster tail probability decay rate such as a Gaussian (larger α) may take exponential time to escape a spurious local extrema, whereas a heavy-tailed policy escapes in polynomial time, as a function of the width a of the set containing a local maxima (5.8) and its escape direction (5.10).

which *we choose* in Algorithm 1. Moreover, in the aforementioned works, the analysis is only for the scalar-dimensional case, whereas here we consider dimension $d > 1$.

We begin then by rewriting (4.2) in terms of the true policy gradient and the stochastic error $\hat{\nabla} J(\boldsymbol{\theta}_k) - \nabla J(\boldsymbol{\theta}_k)$, with the noise process hypothesized as an α-tailed distribution, given by

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) + \eta^{1/\alpha} \eta^{\frac{\alpha-1}{\alpha}} S_k \,, \tag{5.6}$$

where, $S_k \in \mathbb{R}^d$ is $\mathcal{S}\alpha\mathcal{S}$ distributed random vector. Subsequently, we impose that the score function [cf. (4.1)] is dissipative (Assumption 5.9).

Hereafter, we rewrite discrete-time process $\boldsymbol{\theta}_k$ as $\boldsymbol{\theta}^k$ with superscript to disambiguate between continuous and discrete time. (5.6) holds under a hypothesis that the stochastic errors associated with policy gradient steps are heavy-tailed, which is observed experimentally in (Garg et al., 2021). In Sec. 6, we experimentally corroborate that policies induce gradient noise with a proportionate tail index (Fig. 2). The continuous-time analogue of (5.6), i.e., $(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k)/\eta$ as $\eta \to 0$, defines Stochastic Differential Equation (SDE) driven by an $\alpha-$stable Lévy process as (Tzen et al., 2018)

$$d\boldsymbol{\theta}_t^\epsilon = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t^\epsilon) dt + \epsilon d\mathbf{L}_t^\alpha \,, \tag{5.7}$$

where, $\epsilon := \eta^{\frac{\alpha-1}{\alpha}}$ is a coefficient of the jump process (similar to diffusion coefficient in Brownian motion), and $\mathbf{L}_t^\alpha$ denotes the multi-dimensional α-stable Lévy motion in $\mathbb{R}^d$. With these details in place, we impose some additional structure (Assumption 5.7) on the non-convex landscape of the objective $J(\boldsymbol{\theta})$ in (3.1), namely, within the region of the objective's assumed finitely many local maxima, each one is separated by only a local minimum and no saddle points. With the operating hypothesis that there are finitely many extrema of the objective, denote as $\mathcal{G}_i \subset \mathbb{R}^d$ the neighborhood of the i-th local (arbitrary) maximizer $\bar{\boldsymbol{\theta}}_i$:

$$\mathcal{G}_i := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i\| < a + \xi\} \,, \qquad \partial\mathcal{G}_i := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i\| = a + \xi\} \tag{5.8}$$

where, a, $\xi > 0$ are scalar radius parameters, and $\partial\mathcal{G}_i$ denotes the boundary of this neighborhood.

Exit Time and Transition Time. We next define the metastability quantities of exit and transition time in both continuous and discrete-time, assuming that (5.7) and (5.6) are initialized at $\boldsymbol{\theta}_0 \in \mathcal{G}_i$.

Definition 5.5 *(**Exit time from** $\mathcal{G}_i$) The time required for the continuous-time process* (5.7) *and discrete-time process* (5.6)*, respectively, to exit $\mathcal{G}_i$ along standard basis vector $\mathbf{r} \in \mathbb{R}^d$ is defined by*

$$\hat{\tau}_{\xi,a}(\epsilon) \triangleq \inf\{t \geq 0 : \boldsymbol{\theta}_t^\epsilon \in \Omega_i^+(\bar{\delta})\}, \qquad \bar{\tau}_{\xi,a}(\epsilon) \triangleq \inf\{K \in \mathbb{N} : \boldsymbol{\theta}^K \in \Omega_i^+(\bar{\delta})\}. \tag{5.9}$$

Here, a and ξ denote scalar radius parameters (cf. (5.8)). For all $(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i)$ at a distance $\bar{\delta}$ from $\partial\mathcal{G}_i$, we define its distance to $\partial\mathcal{G}_i$ along standard basis vector $\mathbf{r} \in \mathbb{R}^d$, where $\mathbf{r}$ is as in (5.7) with $\mathbf{L}_t^\alpha = \mathbf{r}L_t$ in terms of the lines in $\mathbb{R}^d$ as $g_{i_{\boldsymbol{\theta}}}(t) = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i + t \cdot \mathbf{r}$ for $t \in \mathbb{R}$. Then, for all $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i\| < \bar{\delta}$, $\bar{\delta} \in (0, a + \xi)$, the distance function to the boundary along $\mathbf{r}$ is defined as

$$d^+(\boldsymbol{\theta}) := \inf\{t > 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \partial\mathcal{G}_i\}\ , \qquad d^-(\boldsymbol{\theta}) := \sup\{t < 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \partial\mathcal{G}_i\}, \tag{5.10}$$

where (5.10) define distance between any point of interest and the boundary of domain along the unit vector, $\mathbf{r}$, we have $g_{i_{\boldsymbol{\theta}}}(t) \notin \mathcal{G}_i$ for $t \notin (d^-(\boldsymbol{\theta}), d^+(\boldsymbol{\theta}))$ for all i. We say the point exits the domain $\mathcal{G}_i$ in the direction $\mathbf{r}$ when it enters the $\bar{\delta}$-tubes outside $\mathcal{G}_i$ defined by

$$\Omega_i^+(\bar{\delta}) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \| \left\langle (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i), \mathbf{r} \right\rangle \mathbf{r} - (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i)\| < \bar{\delta},\ \left\langle (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_i), \mathbf{r} \right\rangle > 0\} \cap \mathcal{G}_i^c. \tag{5.11}$$

We underscore that $\bar{\tau}(\cdot)$ represents the exit time of discrete-time process $\boldsymbol{\theta}^k$, whereas $\hat{\tau}(\cdot)$ denotes that of continuous-time stochastic process $\boldsymbol{\theta}_t^\epsilon$.

Definition 5.6 *(**Transition time from** $\mathcal{G}_i$ **to** $\mathcal{G}_j$) Under the existence of a unit vector $\mathbf{r}$ along the direction connecting the domains $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ between two distinct local maxima, we define the transition time from a neighborhood of one local maxima to another, i.e., from $\mathcal{G}_i$ to $\mathcal{G}_j$, $i \neq j$ in respective continuous-time [cf.* (5.7)*] and discrete-time* (5.6)

$$\hat{T}_i(\epsilon) = \inf\{t > 0 : \boldsymbol{\theta}_t^\epsilon \in \cup_{i \neq j} \mathcal{G}_j\}\ , \qquad \bar{T}_i(\epsilon) = \inf\{K \in \mathbb{N} : \boldsymbol{\theta}^K \in \cup_{i \neq j} \mathcal{G}_j\}\ . \tag{5.12}$$

We begin by stating a technical assumptions which are required for the theorems presented in Section 5.2. The first is regarding the non-convex landscape of J and the later is regarding the Lévy jump process in (5.7).

Assumption 5.7 *Following statements holds for function, J:*

1. *The set of local maxima of the value function J consists of r distinct points $\{m_i\} = \{J(\bar{\boldsymbol{\theta}}_i)\}$ separated by $r - 1$ local minima $\{s_i\}$.*
2. *The function J possesses the strict-saddle property, i.e., all its local maxima satisfy $\nabla^2 J(\boldsymbol{\theta}) \prec 0$ and all its other stationary points satisfy $\lambda_{\min}(\nabla^2 J(\boldsymbol{\theta})) > 0$.*
3. *The value function $J(\boldsymbol{\theta})$ satisfies the growth condition; $J'(\boldsymbol{\theta}) > |\boldsymbol{\theta}|^{1+c}$ for $c > 0$ and $|\boldsymbol{\theta}|$ sufficiently large, i.e. the function increases to infinity with infinite $\boldsymbol{\theta}$.*

Assumption 5.8
1. *$L_0^\alpha = 0$ almost surely.*
2. *For $t_0 < t_1 < \ldots < t_N$, the increments $(L_{t_i}^\alpha)$ are independent $(i = 1, \ldots, N)$.*
3. *The difference $(L_t^\alpha - L_s^\alpha)$ and L_{t-s}^α have the same distribution: $\mathcal{S}\alpha\mathcal{S}(t-s)^{1/\alpha}$ for $s < t$.*

4. *L_t^α is continuous in probability: for all $\delta > 0$ and $s \geq 0$, $\mathcal{P}(|L_t^\alpha - L_s^\alpha| > \delta) \to 0$ as $t \to s$.*

We also first present an additional condition we require on the score function.

Assumption 5.9 *For some $m > 0$ and $b \geq 0$, $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\cdot)$ is (m, b, c)-dissipative, which implies that $c_\alpha \langle \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\cdot) \rangle \geq m\|\boldsymbol{\theta}\|^{1+c} - b$, for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

We impose the following structural assumption on $\mathcal{G}_i$ [cf. (5.8)] such that desired properties for a domain perturbed by a Lévy noise in multi-dimensional space holds Imkeller et al. (2010).

Assumption 5.10 *The following assumptions hold for $\mathcal{G}_i$:*

1. *We denote by $\Omega_i := \{\boldsymbol{\theta} \in \mathbb{R}^d : \boldsymbol{\theta} = t \cdot \mathbf{r}_i$ for a $t \in \mathbb{R}\}$ the straight line in the direction of $\mathbf{r}_i$. Let $\nabla J(\cdot) : \bar{\mathcal{G}} \to \mathbb{R}^d$ and the set $\bar{\mathcal{G}} \cap \Omega$ is connected. There exists numbers $a, b > 0$ and a closed interval $I := [-b, a]$ such that for all $t \in (-b, a)$ we have: $t \cdot \mathbf{r} \in \mathcal{G}_i$. Since $\bar{\boldsymbol{\theta}}_i \in \mathcal{G}$, $\mathcal{G}_i$ is open, and $\mathcal{G}_i \cap \Omega \neq \emptyset$.*

2. *The boundary of $\mathcal{G}_i$ defined by $\partial\mathcal{G}_i$ is a $\mathcal{C}^1$-manifold so that the vector field n of the outer normals on the boundary exists. We assume $\langle \nabla J(\boldsymbol{\theta}), n(\boldsymbol{\theta}) \rangle \leq -\frac{1}{C}$, for all $\boldsymbol{\theta} \in \mathcal{G}_i$. This means that $\nabla J(\cdot)$ points into $\mathcal{G}_i$.*

3. *Local extrema, $\bar{\boldsymbol{\theta}}_i$ is an attractor of the domain, i.e. for every starting value $\boldsymbol{\theta} \in \mathcal{G}_i$, the deterministic solution vanishes asymptotically.*

4. *There exists atleast one set of domains, $\mathcal{G}_{i-1}$, $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ such that $\mathcal{G}_{i-1}$, $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ are connected, $\partial\mathcal{G}_i \cap \partial\mathcal{G}_j \neq \emptyset$, $j \in \{i, i-1\}$. We assume existence of a local minima at the intersection of $\partial\mathcal{G}_i$ and $\partial\mathcal{G}_j$.*

5. *There exists a discrete instant K such that exit time $\hat{\tau}_{\xi,a}$ [c.f. (5.9)] greater than $K\eta$, $K > 0$ and $\boldsymbol{\theta}_{\hat{\tau}}^\epsilon \in \Omega^+$ for $\hat{\tau}_{\xi,a}(\epsilon) \geq K\eta$.*

Assumption 5.7 is regarding the level sets of the value function within the vicinity of stationary points versus local extrema. Assumption 5.7.1 ensures that there is positive volume separating distinct extrema, which imposes that the value function, and hence reward, cannot be extremely similar for policies whose relative merits are different. Observe that the strict saddle property (Assumption 5.7.2) has been studied before in the context of policy gradient method, as it is a sufficient condition for the correlated negative curvature condition Zhang et al. (2020c), which holds whenever the policy parameterization is associated with a positive definite Fischer information matrix, and the reward function is strictly positive or strictly negative (Zhang et al., 2020c, Assumption 4.5). Assumption 5.7.3 is easy to satisfy for any policy that does not threshold large values of the derivative, such as the Gaussian or Cauchy – direct calculation reveals that it holds for these cases, but it does not hold for a *truncated* Gaussian.

Assumption 5.8 imposes conditions on the Lévy processes that drive the heavy-tailed noise. Theoretically they are difficult to verify, but we note that they are strictly more general than standard assumptions in the ODE analysis of stochastic approximation that underlies the stability analysis of reinforcement learning – see Borkar and Meyn (2000). Moreover, we empirically verify that the noise satisfies the conditions required to be jump

process with index α in Figure 2, due to the fact that if the gradient is heavy tailed, then the noise associated with the stochastic errors is heavy-tailed.

Assumption 5.9 holds for any policy parameterization which is an increasing function of the norm. Observe that it holds for the policy in Example 2 directly when the policy parameter θ lies in compact space.

Assumption 5.10 imposes structure on the landscape of the value function. Assumption 5.10.2 imposes that the gradient is negatively correlated with the normal vector pointing away from a neighborhood of a stationary point, which usually holds. Assumption 5.10.3 ensures that the gradient is null near a local extrema, i.e., the policy gradient becomes null at a local extrema. Assumption 5.10.4 imposes that there is some intersection between the neighborhoods of extrema, which means that one locally optimal policy may have similar cumulative return to another of comparable quality. Assumption 5.10.5 imposes that the transition time between the neighborhoods of local extrema is governed by choice of learning rate up to a constant factor, which typically holds in practice.

The following theorems present the first exit time and transition time probabilities of the proposed heavy-tailed policy gradient setting, (4.2) when initialized within $\mathcal{G}_i \subset \mathbb{R}^d$ such that (5.8) holds.

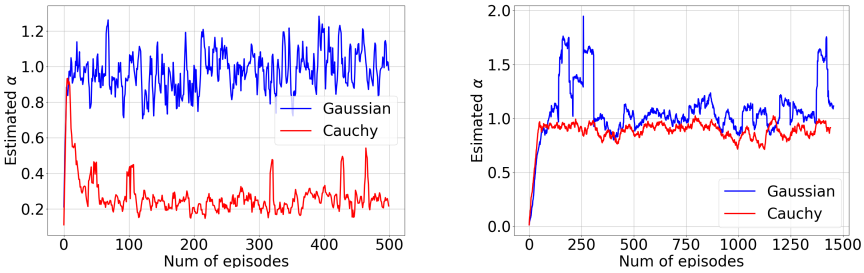
Theorem 5.11 *(Exit Time Dependence on Tail Index) Suppose Assumptions 4.1- 5.10 hold, the value function J is initialized near local maxima $\bar{\boldsymbol{\theta}}_i$, and the policy gradient update in* (4.2) *is run under a heavy-tailed policy parameterization that induces tail index α in its stochastic error* (5.6)*. Then, the likelihood of its exit time from neighborhood $\mathcal{G}_i$ [cf.* (5.8)*] of $\bar{\boldsymbol{\theta}}_i$ larger than K is upper bounded as*

$$\mathcal{P}^{\boldsymbol{\theta}_0}(\bar{\tau}_{0,a}(\epsilon) > K) \leq (2/\epsilon^{\rho\alpha})\epsilon^{\alpha}(d^+)^{-\alpha} + \mathcal{O}\big(\big(dK\eta^{2-(1/\alpha)}\big)^{\beta}\big) + \mathcal{O}\big(1 - \big(1 - C_\alpha d^{1+(\alpha/2)}\eta \exp(\alpha M_J \eta)\epsilon^{\alpha}((\xi/3))^{-\alpha}\big)^{K} + \delta\big) \tag{5.13}$$

with initialization $\boldsymbol{\theta}_0$, d^+ [cf. (5.10)*] denotes distance between $\boldsymbol{\theta}_0$ and the boundary of $\mathcal{G}_i$, $\rho \in (0,1)$ is a positive constant and $\boldsymbol{\theta} \in \mathbb{R}^d$. Moreover, the Hölder continuity constant satisfies $\beta \in (0, \alpha - 1)$, $\delta > 0$, $\xi > 0$, η is the step-size, $k_1 := 1/\eta^{\alpha-1}$, and ϵ [cf.* (5.7)*] is the jump process coefficient.*

See proof in Appendix G. Observe that as $\epsilon \to 0$ in (5.13), the right hand side of (5.13) depends on the distance of $\boldsymbol{\theta}_0$ from the boundary $\partial\mathcal{G}$ and tail-index α. Further, the dependence on d^+ (cf. (5.10)) implicitly hinges upon the width a of the neighborhood of the extrema (5.8), which noticeably decreases with heavier tails (smaller α), meaning that heavier-tailed policies increase the likelihood of escape and tend towards wider maxima. The intricacy of the expression precludes easy interpretation. Thus, consider the average exit time for the single dimensional case in Table 1, in which there exists only a single direction of exit, which coincides with (Imkeller and Pavlyukevich, 2006; Nguyen et al., 2019b). In contrast to proposed heavy-tailed setting wherein exit time is a function of width of the neighborhood, exit time for PG under, e.g., a Gaussian parameterization, depends exponentially on the value at the extrema. Next, we discuss the transition time from one extrema to another.

Theorem 5.12 *(Transition Time Dependence on Tail Index) Suppose Assumptions 4.1- 5.10 hold and the value function J is initialized near a local maxima $\bar{\boldsymbol{\theta}}_i$. Then in the limit*



(a) 1D Mario environment. (b) PMC (Sec. 3)

Figure 2: **(a)** Tail index estimation of HPG updates for a 1D Mario (Matheron et al., 2019). **(b)** Tail index estimation for Pathological Mountain car. In both, estimates are averaged over latest 50 episodes. Observe that a Cauchy policy induces a tail index lower than the Gaussian policy, and the volatility of the blue sample path stems from training being uncompleted during estimation.

$\epsilon \to 0$, *the policy gradient update in* (4.2) *under a heavy-tailed parameterization with tail index* α *associated with its induced stochastic error* (5.6)*, transitions from* $\mathcal{G}_i$ *to the boundary of* $(i+1)$*-th local maxima with probability,* $\mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1})$ *lower bounded as a function of tail index* α*:*

$$\lim_{\epsilon \to 0} \mathcal{P}^{\theta_0}(\theta^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) \geq \frac{d_{ij}^{-\alpha}}{((d_{ij}^+)^{-\alpha} + (-d_{ij}^-)^{-\alpha}} - \delta, \tag{5.14}$$

where $\delta > 0$*, escape distance from extrema are defined as* $d_{ij}^+(\boldsymbol{\theta}) := \inf\{t > 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}\}$ *and* $d_{ij}^-(\boldsymbol{\theta}) := \sup\{t < 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \Omega_i^-(\bar{\delta}) \cap \partial\mathcal{G}_{i-1}\}$*, and* $\Omega_i^+(\bar{\delta})$ *is defined before Def. 5.6.*

Similar to exit time, the transition time probability [cf. (5.14)] (proof in Appendix H) depends on the width of boundary and the tail index, which noticeably also decreases for heavier tails (smaller α), and depends on the width of the neighborhood containing a local maxima. For ease of interpretation, the single-dimensional case for both vanilla PG and HPG are given in Table 1. Transition times are asymptotically exponentially distributed in the limit of small noise and scale with $1/\epsilon^\alpha$ for HPG, whereas transition time for Brownian is exponentially distributed with ϵ^α replaced by *exponential dependence* $e^{2J(\cdot)/\epsilon^2}$ for a Gaussian policy. Thus, in the small noise limit, Brownian-motion driven PG needs exponential time to transition from one peak to another, whereas the Lévy-driven process requires polynomial time, illuminating that heavy-tailed policies quickly jump away from spurious extrema.

6. Experiments

In this section, we evaluate the proposed HPG (Algorithm 1) as compared to some common approaches for policy search. Before doing so, we demonstrate experimentally evidence that the heavy-tailed policies results in heavy tailed policy gradients. Then, we provide experiments for the Pathological Mountain Car (PMC) (Sec. 3) and 1D Mario environment

Matheron et al. (2019). For PMC, we consider an incentive structure in which the amount of energy expenditure, i.e., the action squared, at each time-step is negatively penalized and the reward structure is given by

$$r(s_t, a_t) = -a_t^2 \mathbb{1}_{\{-4.0<s<3.709, s\neq 2.667\}} + (500 - a_t^2)\mathbb{1}_{\{s=-4.0\}} + (10 - a_t^2)\mathbb{1}_{\{s=2.67\}}.$$

Here, s denotes the state space, and the action a_t is a one-dimensional scalar representing the speed of the vehicle $\dot{s}_t$. In 1D Mario environments, state $s \in [-4.0, 3.709]$ and the actions are

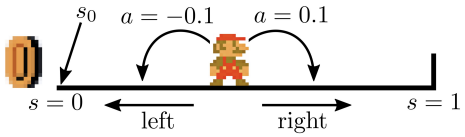


Figure 3: 1D Mario environment Matheron et al. (2019).

confined to $[-20, 20]$. On the other hand, as the name suggests, the 1D Mario environment is one-dimensional with continuous state and action spaces with incentive structure and state transition defined as $r(s_t, a_t) = \mathbb{1}_{\{s_t+a_t<0\}}$, and $s_{t+1} = \min\{1, \max\{0, s_t + a_t\}\}$ where, state, $s \in [0, 1]$ and action $a \in [-0.1, 0.1]$. Each episodes are initialized at $s_0 = 0$.

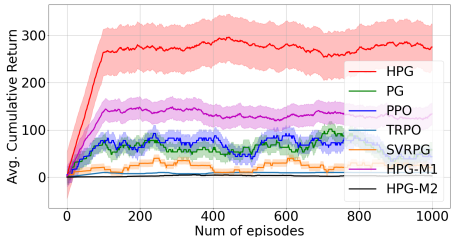
Before presenting the experiments, first in Fig. 2, we depict the estimation of tail index α (using method in Mohammadi et al. (2015)) for gradient estimates [cf. (4.2)] with a Cauchy and Gaussian policy. The lower the value of α the heavier the tail is of the policy gradient. In Fig. 2(a), we observe that the average estimate for the Gaussian policy settles to a value of one, while the corresponding value for Cauchy values settles around 0.2 for 1D Mario environment. A similar plot for PMC is in Fig. 2(b): note that the tail-index estimate of Cauchy settles around unity and the corresponding value for Gaussian exhibits volatility since the policy has yet to converge. For the tail index estimation, we utilized the logic presented in Mohammadi et al. (2015) for the α estimation reiterated here in the form of Theorem 6.1 for quick reference.

Theorem 6.1 *Mohammadi et al. (2015) Let $\{\mathbf{X}_i\}_{i=1}^K$ be the collection of random variables with $\mathbf{X}_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ and $K = K_1 \times K_2$. Define $Y_i \triangleq \sum_{i=1}^{K_1} \mathbf{X}_{j+(i-1)K_1}$ for $i \in [1, K_2]$. Then the estimator*

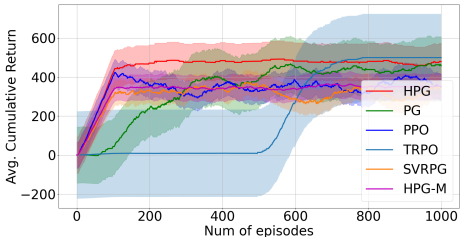
$$\frac{\hat{1}}{\alpha} \triangleq \frac{1}{\log K_1}\left(\frac{1}{K_2}\sum_{i=1}^{K_2}\log|\mathbf{Y}_i| - \frac{1}{K}\sum_{i=1}^{K}\log|\mathbf{X}_i|\right) \tag{6.1}$$

converges to $\frac{1}{\alpha}$ almost surely as $K_2 \to \infty$.

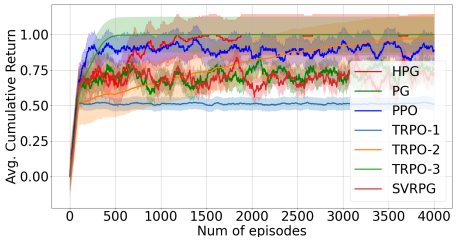
Note that $\{\mathbf{X}_i\}$ corresponds to the samples from policy gradient estimates. The aforementioned approach has been employed recently for estimating the tail-index of stochastic gradients Garg et al. (2021); Simsekli et al. (2019). Next, we present the main experiments results corroborating the findings in the main paper. Additional experiments with continuous control environments are provided in Appendix I.1.



(a) Constant scale σ PMC



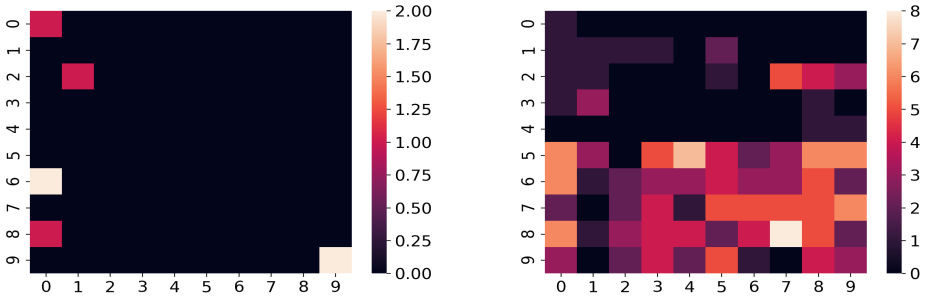
(b) Tuned scale σ PMC.



(c) Tuned scale σ Mario.

Figure 4: **(a)** We plot the average cumulative returns for PMC environment over latest 100 episodes for Gaussian and Cauchy policies with constant σ. The importance of searching over a heavy-tailed (Cauchy) distribution is clear, as the Gaussian policy converges to spurious behavior. **(b)** We plot average cumulative returns for PMC environment with variable σ over latest 100 episodes. **(c)** Average commutative return for 1D Mario with variable sigma. For TRPO, TRPO-1, 2, and 3, are respectively for trust region parameters 10^{-10}, 10^{-6}, and 10^{-5}.

Fig 4 compares the average commutative reward performance of HPG (for a Cauchy policy) to GPOMDP Baxter and Bartlett (2001) with a Gaussian policy with fixed and tuned variance parameters Papini et al. (2020) (which we abbreviate as PG), as well as Proximal Policy Optimization (PPO) Schulman et al. (2017), Trust Region Policy Optimization Schulman et al. (2015a), and Stochastic Variance Reduced PG (SVRPG) Papini et al. (2018), for constant variance as well as variable variance. In order to evaluate the Meta stable characteristics of the algorithms, we initialize each episodes at $s = 2.26$, in the neighborhood of the local minima, $s = 2.67$. Firstly in Fig. 4(a), we evaluate the performance on PMC environment when the scale of the HPG is a constant $\sigma = 3.0$ and the variance of Gaussian policy is also fixed $\sigma = 3.0$. Secondly, we present the results with variable scale σ for PMC and 1D Mario environments in Fig. 4(b)-4(c). All the experiments use a discounted factor of $\gamma = 0.97$ and we use a diminishing step-size ranging from 0.005 to 5×10^{-9}. All the simulations are performed for 1000 episodes using a batch size of $B_k = 5$ and with cumulative returns averaged over 100 episodes. For the comparison with PPO, the policy ratio for PPO is allowed to vary in the interval $[1 - \epsilon_1, 1 + \epsilon_1]$ with $\epsilon_1 = 0.2$. Note that a fined tuned value of ϵ_1 can result in a better performance as shown in Fig. 4(b). However, note that the best performance feasible for PPO is same as that of PG. The trust region parameters for TRPO,



(a) PMC: Occupancy measure, HPG

(b) PMC: Occupancy measure, PG

Figure 5: Single episode occupancy measure for PMC when initialized in the neighborhood of spurious local extrema. We plot it for a test episode after network is trained for 1000 episodes with HPG policy and PG policy. States from $[-4.0, 3.709]$ are discretized into 100 states to calculate the state visitation frequency. Desired extrema of $s = -4.0$ corresponds to $(0, 0)$ and the initial state $s = 2.26$ is $(0, 8)$. Dark color in the map corresponds to region not visited during the test episode. **(a)** Single episode occupancy measure for PMC with HPG. The importance of searching over a heavy-tailed (Cauchy) distribution is clear, as the policy takes heavier jumps and reach to desired goal faster. **(b)** Single episode occupancy measure for for PMC with PG. Overall, we may observe that a Gaussian policy results in an occupancy measure which exhibits diffuse probability across the state space, failing to concentrate around actions associated with higher reward, whereas the heavy-tailed distribution results in an occupancy measure that assigns high likelihood to a small number of extreme actions in a manner reminiscent of the "black swan" phenomenon Taleb (2007).

aka. maximum KL- divergence allowed is set to 0.001. In addition, here we also evaluate performance of the HPG against Stochastic Variance-Reduced Policy Gradient (SVRPG) (Papini et al. (2018)). The number of epochs for SVRPG is fixed to 1000 and epoch size, $m = 10$. Further for PMC environment, we have included the comparisons with HPG-M which denotes the moderate tailed policy gradient with $\alpha = 1.3$. In Fig. 4(a), HPG-M1, HPG-M2 denotes the different instance of moderate tailed policy gradient with $\sigma = 5, 3$, respectively. For the experiments in Fig 4, we use a simple network without hidden layers.

From meta stability results of Section 5.2, it is the nature of jumps initiated by heavy tailed policies which results in better meta stable characteristics and results in faster escape from spurious local extrema. In order to establish this fact, we plot the single test episode state visitation frequency (aka single episode occupancy measures) of HPG and PG once the training is done. The test episode is initialized at $s = 2.26$ (neighborhood of local extrema) and states of the environment $s \in [-4, 3.709]$ are discretized into 100 states and the heatmap of the state visitation frequency is shown in Fig. 5.

7. Conclusion

We focused on PG method in infinite-horizon RL problems. Inspired by persistent exploration that mitigates the tendency of policies to become mired at spurious behavior, we sought to

nearly satisfy it in continuous settings through heavy-tailed policies. Doing so invalidated several aspects of existing analyses, which motivated studying the sample complexity of policy search when the score function is Hölder continuous and its norm is integrable with respect to the policy, and introducing an exploration tolerance parameter to quantify the degree to which the score function may be unbounded.

Moreover, we established that heavy-tailed policies induce heavy-tailed transition dynamics, which jump away from local extrema as formally quantified by the metastability characteristics of its Lévy process representation. We discerned that policies a heavier tail induce transitions away from a local extrema more quickly than one with a lighter tail, and tend towards extrema with more volume, which we empirically associated with more stable policies for a few RL problems in practice. The characterization of jumps defined by metastability provides a lens through which approximate persistent exploration may be satisfied in continuous space.

8. Acknowledgments

Bedi would like to acknowledge the support of Northrup Grumman Seed Grant, Army Cooperative Agreement W911NF2120076, and Amazon Research Awards 2022. Mengdi Wang acknowledges support by NSF grants DMS-1953686, IIS-2107304, CMMI-1653435, and ONR grant 1006977.

References

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020a.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020b.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020c.

Dilip Arumugam and Benjamin Van Roy. Randomized value functions via posterior state-abstraction sampling. *arXiv preprint arXiv:2010.02383*, 2020.

David Avnir, Ofer Biham, Daniel Lidar, and Ofer Malcai. Is the geometry of nature fractal? *Science*, 279(5347):39–40, 1998.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

Albert-László Barabási et al. Emergence of scaling in complex networks. *Handbook of Graphs and Networks: From the Genome to the Internet. Berlin: Wiley-VCH*, 2003.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *International Conference on Machine Learning*, pages 1716–1731. PMLR, 2022.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

Richard Ernest Bellman. *Dynamic Programming*. Courier Dover, 1957. ISBN 0486428095.

Dimitir P Bertsekas and Steven Shreve. *Stochastic optimal control: the discrete-time case*. 2004.

DP Bertsekas and DA Castanon. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE transactions on automatic control*, 34(6):589–598, 1989.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Shalabh Bhatnagar, Richard Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Sujay Bhatt, Alec Koppel, and Vikram Krishnamurthy. Policy gradient using weak derivatives for reinforcement learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5531–5537. IEEE, 2019.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SICON*, 38(2):447–469, 2000.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Maurice C Bryson. Heavy-tailed distributions: properties and tests. *Technometrics*, 16(1): 61–68, 1974.

Souradip Chakraborty, Amrit Singh Bedi, Kasun Weerakoon, Prithvi Poddar, Alec Koppel, Pratap Tokekar, and Dinesh Manocha. Dealing with sparse rewards in continuous control robotics via heavy-tailed policy optimization. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 989–995. IEEE, 2023.

Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International conference on machine learning*, pages 834–843. PMLR, 2017.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Thomas L Dean and Robert Givan. Model minimization in markov decision processes. In *AAAI/IAAI*, 1997.

Adithya M Devraj and Sean P Meyn. Zap q-learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2232–2241, 2017.

Yaqi Duan, Zheng Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32, 2019.

Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for q-learning. *Journal of machine learning Research*, 5(1), 2003.

Sergio M Focardi and Frank J Fabozzi. Fat tails, scaling, and stable laws: a critical look at modeling extremal events in financial phenomena. *The Journal of Risk Finance*, 2003.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J Zico Kolter, Sivaraman Balakrishnan, Zachary C Lipton, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. *arXiv preprint arXiv:2102.10264*, 2021.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898. PMLR, 2015.

Xiaoxiao Guo, Satinder Singh, Honglak Lee, Richard L Lewis, and Xiaoshi Wang. Deep learning for real-time atari game play using offline monte-carlo tree search planning. *Advances in neural information processing systems*, 27:3338–3346, 2014.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. *arXiv preprint arXiv:2006.04740*, 2020.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

John E Hutchinson. Fractals and self similarity. *Indiana University Mathematics Journal*, 30(5):713–747, 1981.

Peter Imkeller and Ilya Pavlyukevich. First exit times of sdes driven by stable lévy processes. *Stochastic Processes and their Applications*, 116(4):611–642, 2006.

Peter Imkeller, Ilya Pavlyukevich, and Michael Stauch. First exit times of non-linear dynamical systems in $\mathbb{R}^d$ perturbed by multifractal lévy noise. *Journal of Statistical Physics*, 141(1):94–119, 2010.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 179–188. PMLR, 2015.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Taisuke Kobayashi. Student-t policy in reinforcement learning to acquire global optimum of robot control. *Applied Intelligence*, 49(12):4335–4347, 2019.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *NeurIPS*, pages 1008–1014, 2000.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for Markov decision processes. *SICON*, 38(1):94–123, 1999.

Michael R Kosorok and Erica EM Moodie. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32:10565–10576, 2019.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud. The problem with ddpg: understanding failures in deterministic environments with sparse rewards. *arXiv preprint arXiv:1911.11679*, 2019.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33, 2020a.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020b.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Mohammad Mohammadi, Adel Mohammadpour, and Hiroaki Ogata. On estimating the tail index and the spectral measure of multivariate s *alpha* s-stable distributions. *Metrika*, 78 (5):549–561, 2015.

Kumpati S Narendra and Anuradha M Annaswamy. Persistent excitation in adaptive systems. *International Journal of Control*, 45(1):127–160, 1987.

Kumpati S Narendra and Anuradha M Annaswamy. *Stable adaptive systems*. Courier Corporation, 2012.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5949–5958, 2017.

Than Huy Nguyen, Umut Simsekli, and Gaël Richard. Non-asymptotic analysis of fractional langevin monte carlo for non-convex optimization. In *International Conference on Machine Learning*, pages 4810–4819. PMLR, 2019a.

Thanh Huy Nguyen, Umut Şimşekli, Mert Gürbüzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *arXiv preprint arXiv:1906.09069*, 2019b.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4033–4041, 2016.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *ICML*, pages 4026–4035, 2018.

Matteo Papini, Andrea Battistello, and Marcello Restelli. Balancing learning speed and stability in policy gradient via adaptive exploration. In *International Conference on Artificial Intelligence and Statistics*, pages 1188–1199. PMLR, 2020.

Santiago Paternain, Juan Bazerque, Austin Small, and Alejandro Ribeiro. Stochastic policy gradient ascent in reproducing kernel hilbert spaces. *IEEE Transactions on Automatic Control*, 2020.

Robin Pemantle et al. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Martin L Puterman. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Daniel Russo. Approximation benefits of policy gradient methods with aggregated states. *arXiv preprint arXiv:2007.11684*, 2020.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1): 1–96, 2018.

Yagiz Savas, Melkior Ornik, Murat Cubuktepe, and Ufuk Topcu. Entropy maximization for constrained markov decision processes. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 911–918. IEEE, 2018.

Michael Scheutzow. A stochastic gronwall lemma. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 16(02):1350019, 2013.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc., 2020a.

Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pages 8970–8980. PMLR, 2020b.

Satinder Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems 7*, 7:361, 1995.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS*, pages 1057–1063, 2000.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement Learning: An Introduction*. 2 edition, 2017.

Nassim Nicholas Taleb. *The black swan: The impact of the highly improbable*, volume 2. Random house, 2007.

John B Taylor and John C Williams. A black swan in the money market. *American Economic Journal: Macroeconomics*, 1(1):58–83, 2009.

H Thanh, S Simsekli, M Gurbuzbalaban, and G RICHARD. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *Advances in neural information processing systems*, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1):59–94, 1996.

Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. *Proceedings of Machine Learning Research vol*, 75:1–19, 2018.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Longjie Xie, Xicheng Zhang, et al. Ergodicity of stochastic differential equations with jumps and singular coefficients. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 175–229. Institut Henri Poincaré, 2020.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proceedings of the 34nd International Conference on Neural Information Processing Systems*, 2020a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. *arXiv preprint arXiv:2010.11364*, 2020b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020c.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. 2018.

Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2810–2818, 2019.

Appendix

Appendix A. Technical Details of Policy Search

A.1 Finiteness of Integral in (4.1)

Lemma A.1 *The integral in* (4.1) *is finite.*

Proof Consider the integral for the policy gradient from (4.1), we obtain

$$I := \frac{1}{1-\gamma}\int_{\mathcal{S}\times\mathcal{A}} \rho_{\pi_{\boldsymbol{\theta}}}(s)\cdot\pi_{\boldsymbol{\theta}}(a\,|\,s)\cdot\|\nabla\log[\pi_{\boldsymbol{\theta}}(a\,|\,s)]\|\cdot|Q_{\pi_{\boldsymbol{\theta}}}\cdot(s,a)|dsda$$

From Assumption (4.1) and the definition of Q function, we note that $Q_{\pi_{\boldsymbol{\theta}}}(s,a)\leq\frac{U_R}{1-\gamma}$ for any $\boldsymbol{\theta}, s$ and a. Therefore we could upper bound the above integral as follows

$$\begin{aligned} I &\leq \frac{U_R}{(1-\gamma)^2}\int_{\mathcal{S}\times\mathcal{A}} \rho_{\pi_{\boldsymbol{\theta}}}(s)\cdot\|\nabla\log\pi_{\boldsymbol{\theta}}(a\,|\,s)\|\cdot\pi_{\boldsymbol{\theta}}(a\,|\,s)dsda \\ &\leq \frac{U_R\sqrt{B}}{(1-\gamma)^2}\int_{s\in\mathcal{S}} \rho_{\pi_{\boldsymbol{\theta}}}(s)\cdot ds \qquad \text{(A.1)} \end{aligned}$$

where we obtain (A.1) from Assumption 4.2 after applying the Jensen's inequality which implies that

$$\int_{\mathcal{A}}\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a\,|\,s)\|^x\cdot\pi_{\boldsymbol{\theta}}(a\,|\,s)\cdot da\leq B^{\frac{x}{2}}$$

for $x\in(1,2]$. Since $\rho_{\pi_{\boldsymbol{\theta}}}(a)$ is the occupancy measure distribution for states s, we can obtain the following upper bound for the integral as $I\leq\frac{U_R\sqrt{B}}{(1-\gamma)^2}$. ■

A.2 Proof of Lemma 4.3

Proof Let us start by considering the stochastic gradient $\hat{\nabla}J(\boldsymbol{\theta})$ as defined in (4.2)

$$\mathbb{E}[\hat{\nabla}J(\boldsymbol{\theta})\,|\,\boldsymbol{\theta}] = \mathbb{E}\Big\{\sum_{t=0}^{T}\gamma^{t/2}\cdot R(s_t,a_t)\cdot\Big(\sum_{\tau=0}^{t}\nabla\log\pi_{\boldsymbol{\theta}_k}(a_\tau\,|\,s_\tau)\Big)\Big\}. \qquad \text{(A.2)}$$

In order to relate the above expression to the true gradient, we introduce the infinite sum via the identity function notation $\mathbb{1}_{T\geq t}$ and modify (A.2) as

$$\mathbb{E}[\hat{\nabla}J(\boldsymbol{\theta})\,|\,\boldsymbol{\theta}] = \mathbb{E}\Big\{\lim_{N\to\infty}\sum_{t=0}^{N}\mathbb{1}_{T\geq t}\cdot\gamma^{t/2}\cdot R(s_t,a_t)\cdot\Big(\sum_{\tau=0}^{t}\nabla\log\pi_{\boldsymbol{\theta}_k}(a_\tau\,|\,s_\tau)\Big)\Big\}. \qquad \text{(A.3)}$$

To interchange the limit and expectation in the above expression via Dominated Convergence Theorem in (A.3), we need first to ensure that the individual terms are dominated by an integrable function. To do so, we consider the term inside the expectation in (A.3) as

$$\begin{aligned}&\left\| \sum_{t=0}^{N} \mathbb{1}_{T\geq t} \cdot \gamma^{t/2} \cdot R(s_t, a_t) \cdot \left(\sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right) \right\| \\ &\qquad \leq U_R \sum_{t=0}^{N} \mathbb{1}_{T\geq t} \cdot \gamma^{t/2} \cdot \left(\sum_{\tau=0}^{t} \|\nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau)\| \right), \end{aligned} \tag{A.4}$$

which follows from the bound $R(s_t, a_t) \leq U_R$. The sum on the right-hand side of (A.4) is integrable with respect to the occupancy measure via Assumption 4.2. Therefore, we may apply Dominated Convergence Theorem in (A.3) in order to exchange expectation and limit as follows

$$\begin{aligned}\mathbb{E}[\hat{\nabla} J(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}] &= \lim_{N\to\infty} \sum_{t=0}^{N} \mathbb{E}\left[\mathbb{E}_T\left[\mathbb{1}_{T\geq t}\right] \cdot \gamma^{t/2} \cdot R(s_t, a_t) \cdot \left(\sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right) \right] \\ &= \lim_{N\to\infty} \sum_{t=0}^{N} \mathbb{E}\left[\mathbb{P}[T \geq t] \cdot \gamma^{t/2} \cdot R(s_t, a_t) \cdot \left(\sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right) \right]. \end{aligned} \tag{A.5}$$

where (A.5) holds since $\mathbb{E}_T[\mathbb{1}_{T\geq t}] = \mathbb{P}[T \geq t]$. Next, we note that since $T \sim \text{Geom}(1-\gamma^{1/2})$ which implies that $\mathbb{P}[T \geq t] = \gamma^{t/2}$, hence we can write

$$\mathbb{E}[\hat{\nabla} J(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \cdot \left(\sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right) \right]. \tag{A.6}$$

After rearranging the order of summation in the above expression, we could write

$$\begin{aligned}\mathbb{E}[\hat{\nabla} J(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}] &= \mathbb{E}\left[\sum_{\tau=0}^{\infty} \sum_{t=\tau}^{\infty} \gamma^t \cdot R(s_t, a_t) \cdot \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right] \\ &= \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \cdot \sum_{t=\tau}^{\infty} \gamma^{t-\tau} \cdot R(s_t, a_t) \cdot \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right] \\ &= \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \cdot Q_{\pi_{\boldsymbol{\theta}}}(s_\tau, a_\tau) \cdot \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \,|\, s_\tau) \right] \\ &= \nabla J(\boldsymbol{\theta}). \end{aligned} \tag{A.7}$$

which is as stated in Lemma 4.3. ■

Appendix B. Proof of Lemma 5.2

Before providing the proof for the statement of Lemma 5.2, we discuss an intermediate Lemma B.1 provided next.

Lemma B.1 *For any given* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathbb{R}^d$*, it holds that*

1) the occupancy measure is Lipschitz continuous, which implies that

$$||\rho_{\boldsymbol{\theta}}(s,a) - \rho_{\boldsymbol{\theta}'}(s,a)||_1 \leq M_\rho \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \tag{B.1}$$

2) and the $Q_{\pi_{\boldsymbol{\theta}}}$ *is also Lipschitz continuous which satisfies*

$$||Q_{\pi_{\boldsymbol{\theta}}}(s,a) - Q_{\pi_{\boldsymbol{\theta}'}}(s,a)||_1 \leq M_Q \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \tag{B.2}$$

where $M_\rho = \frac{\sqrt{B}}{1-\gamma}$ *and* $M_Q = \frac{\gamma U_R M_\rho}{1-\gamma}$.

Proof Proof of statement (1). In order to bound the term $||\rho_{\boldsymbol{\theta}}(s,a) - \rho_{\boldsymbol{\theta}'}(s,a)||_1$, let us define a function $d(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as follows

$$\begin{aligned} d(\boldsymbol{\theta}, \boldsymbol{\theta}') =& \|\rho_{\boldsymbol{\theta}}(s,a) - \rho_{\boldsymbol{\theta}'}(s,a)\|_1 && \text{(B.3)} \\ =& \int_{\mathcal{S}\times\mathcal{A}} |\rho_{\boldsymbol{\theta}}(s,a) - \rho_{\boldsymbol{\theta}'}(s,a)| \cdot ds\ da. && \text{(B.4)} \end{aligned}$$

Next, we evaluate the gradient of $d(\boldsymbol{\theta}, \boldsymbol{\theta}')$ with respect to $\boldsymbol{\theta}$ and tale norm as

$$\begin{aligned} \|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta}, \boldsymbol{\theta}')\| \leq& \left\| \int_{\mathcal{S}\times\mathcal{A}} \text{sign}[\rho_{\boldsymbol{\theta}}(s,a) - \rho_{\boldsymbol{\theta}'}(s,a)] \cdot \nabla_{\boldsymbol{\theta}} \rho_{\boldsymbol{\theta}}(s,a) \cdot ds\ da \right\| \\ \leq& \int_{\mathcal{S}\times\mathcal{A}} \|\nabla_{\boldsymbol{\theta}} \rho_{\boldsymbol{\theta}}(s,a)\| ds\ da, && \text{(B.5)} \end{aligned}$$

where $\text{sign}[x] = +1$ if $x \geq 0$ and $\text{sign}[x] = -1$ if $x < 0$. We recall the definition of the occupancy measure and write

$$\rho_{\boldsymbol{\theta}}(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p(s_t = s, a_t = a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0)) \tag{B.6}$$

where $p(s_t = s, a_t = a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0))$ is the probability of visiting state s and a at the t^{th} instant and $\xi(s_0)$ denotes the initial state distribution. We write the explicit form of $p(s_t = s, a_t = a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0))$ as

$$\begin{aligned} p(s_t{=}s, a_t{=}a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0)) = \int_{\mathcal{S}\times\mathcal{A}} & \xi(s_0) \cdot \pi_{\boldsymbol{\theta}}(a_0|s_0) p(s_1|s_0,a_0) \times \\ & \times \pi_{\boldsymbol{\theta}}(a_1|s_1) p(s_2|s_1,a_1) \times \\ & \quad \vdots \\ & \times \pi_{\boldsymbol{\theta}}(a_{t-1}|s_{t-1}) p(s_t|s_{t-1},a_{t-1}) \times \\ & \times \pi_{\boldsymbol{\theta}}(a_t{=}a|s_t{=}s) \cdot d\mathbf{s}_{t-1} d\mathbf{a}_{t-1} \end{aligned} \tag{B.7}$$

where $d\mathbf{s}_{t-1} = ds_0 ds_1 \cdots ds_{t-1}$, and $d\mathbf{a}_{t-1} = da_0 da_1 \cdots da_{t-1}$ denotes the integration of the state action pairs so far till t. Let us collect the state action pair trajectory till t as $\mathcal{T}_t := \{(s_0,a_0),(s_1,a_1),\cdots,(s_t,a_t)\}$ and write

$$p_{\boldsymbol{\theta}}(\mathcal{T}_t) = \xi(s_0) \cdot \pi_{\boldsymbol{\theta}}(a_0|s_0) p(s_1|s_0,a_0) \times \cdots \times p(s_t|s_{t-1},a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a_t|s_t). \tag{B.8}$$

Using the notation in (B.8), we could rewrite (B.7) as

$$p(s_t{=}s, a_t{=}a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0)) = \int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s_t|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a_t{=}a|s_t{=}s) \cdot d\mathcal{T}_{t-1}. \quad \text{(B.9)}$$

Calculating the gradient on both sides of (B.6), we get

$$\nabla_{\boldsymbol{\theta}} \rho_{\boldsymbol{\theta}}(s, a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \nabla_{\boldsymbol{\theta}} p(s_t = s, a_t = a \mid \pi_{\boldsymbol{\theta}}, \xi(s_0)). \quad \text{(B.10)}$$

Using the definition (B.9), we could write the gradient in (B.10) as follows

$$\nabla_{\boldsymbol{\theta}} \rho_{\boldsymbol{\theta}}(s, a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Big[\int_{\mathcal{T}_{t-1}} \nabla_{\boldsymbol{\theta}} \big\{ p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \big\} \cdot d\mathcal{T}_{t-1} \Big], \quad \text{(B.11)}$$

where by default, for the term $t = 0$, we let $p_{\boldsymbol{\theta}}(\mathcal{T}_{-1}) \equiv 1$, and $p(s|s_{t-1}, a_{t-1}) = \xi(s)$. Now we shift to upper bound the right hand side of (B.5) using the simplified definition in (B.11). Let us rewrite the inequality in (B.5) as

$$\|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta}, \boldsymbol{\theta}')\| \leq \int_{\mathcal{S}\times\mathcal{A}} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Big\| \int_{\mathcal{T}_{t-1}} \nabla_{\boldsymbol{\theta}} \big\{ p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \big\} \cdot d\mathcal{T}_{t-1} \Big\| \cdot ds\ da, \quad \text{(B.12)}$$

where note that the integration over $\mathcal{S} \times \mathcal{A}$ *is now outside the norm as correctly pointed out by the reviewer*. For this correct version of (B.12) the argument from the reviewer will no longer hold and will not result in any confusion. Note that $\nabla_{\boldsymbol{\theta}} \log \big\{ p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \big\} = \sum_{i=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(a_i|s_i)$, with $(s_t, a_t) = (s, a)$, then we have

$$\begin{aligned}
&\Big\| \int_{\mathcal{T}_{t-1}} \nabla_{\boldsymbol{\theta}} \big\{ p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \big\} d\mathcal{T}_{t-1} \Big\| \qquad \text{(B.12)} \\
&= \Big\| \int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot \sum_{i=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(a_i|s_i) d\mathcal{T}_{t-1} \Big\| \\
&\leq \int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot \ \Big\| \sum_{i=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(a_i|s_i) \Big\| d\mathcal{T}_{t-1} \\
&\leq \int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot (t+1)\sqrt{B} d\mathcal{T}_{t-1}
\end{aligned}$$

Substitute the above inequality to B.12 gives us

$$\|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta}, \boldsymbol{\theta}')\| \leq \int_{\mathcal{S}\times\mathcal{A}} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot \int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1}) \cdot p(s|s_{t-1}, a_{t-1}) \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot (t+1)\sqrt{B} \cdot d\mathcal{T}_{t-1} \cdot ds\ da,$$

We note that we can take $(1-\gamma) \sum_{t=0}^{\infty} \gamma^t$ outside the integration $\int_{\mathcal{S}\times\mathcal{A}}$, hence we get

$\|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta}, \boldsymbol{\theta}')\|$

$$\leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \int_{\mathcal{S}\times\mathcal{A}}\int_{\mathcal{T}_{t-1}} p_{\boldsymbol{\theta}}(\mathcal{T}_{t-1})\cdot p(s|s_{t-1},a_{t-1})\cdot \pi_{\boldsymbol{\theta}}(a|s)\cdot (t+1)\sqrt{B}\cdot d\mathcal{T}_{t-1}\cdot ds\ da.$$

After adjusting the limits of integration and from the definition of $p_{\boldsymbol{\theta}}(\mathcal{T}_t) = \xi(s_0)\cdot \pi_{\boldsymbol{\theta}}(a_0|s_0)p(s_1|s_0,a_0)\times\cdots\times p(s_t|s_{t-1},a_{t-1})\cdot\pi_{\boldsymbol{\theta}}(a_t|s_t)$ from (B.8) in the main paper, we can write

$$\|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta},\boldsymbol{\theta}')\| \leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\int_{\mathcal{T}_t}\big((t+1)\sqrt{B}\big)\cdot p_{\boldsymbol{\theta}}(\mathcal{T}_t)\cdot d\mathcal{T}_t.$$

From Assumption 2, it holds that $\int_{\mathcal{A}}\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\|\cdot\pi_{\boldsymbol{\theta}}(a|s)da \leq \sqrt{B}$. We can write the above inequality as

$$\|\nabla_{\boldsymbol{\theta}} d(\boldsymbol{\theta},\boldsymbol{\theta}')\| \leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t(t+1)\sqrt{B} = \frac{\sqrt{B}}{1-\gamma} := M_\rho, \tag{B.15}$$

where we have used the fact that arithmetic–geometric sequence $\sum_{t=0}^{\infty}(t+1)\gamma^t = \frac{\gamma}{(1-\gamma)^2}$. Since the gradient of function $d(\boldsymbol{\theta},\boldsymbol{\theta}')$ with respect to $\boldsymbol{\theta}$ is bounded, it implies that $d(\boldsymbol{\theta},\boldsymbol{\theta}')$ is Lipschitz with respect to $\boldsymbol{\theta}$, which further implies that

$$|d(\boldsymbol{\theta},\boldsymbol{\theta}') - d(\boldsymbol{\theta}',\boldsymbol{\theta}')| \leq M_\rho\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|, \tag{B.16}$$

for all $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^d$. Next, substituting the definition $d(\boldsymbol{\theta},\boldsymbol{\theta}') = \|\rho_{\boldsymbol{\theta}}(s,a)-\rho_{\boldsymbol{\theta}'}(s,a)\|_1$ into (B.16) and noting that $d(\boldsymbol{\theta}',\boldsymbol{\theta}')=0$, we get

$$\|\rho_{\boldsymbol{\theta}}(s,a)-\rho_{\boldsymbol{\theta}'}(s,a)\|_1 \leq M_\rho\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|, \tag{B.13}$$

which is as stated in Lemma B.1(1).

Proof of statement (2). Let us define the occupancy measure,

$$\mu^{sa}_{\boldsymbol{\theta}}(s',a') = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\cdot p(s_t=s',a_t=a'\mid \pi_{\boldsymbol{\theta}}, s_0\sim p(\cdot|s,a)) \tag{B.14}$$

where the initial state $s_0\sim p(\cdot|s,a)$. With the help of occupancy measure $\mu^{sa}_{\boldsymbol{\theta}}(s',a')$, we could write the $Q_{\pi_{\boldsymbol{\theta}}}$ function for state action pair (s,a) as

$$Q_{\pi_{\boldsymbol{\theta}}}(s,a) = R(s,a) + \frac{\gamma}{1-\gamma}\int R(s',a')\cdot\mu^{sa}_{\boldsymbol{\theta}}(s',a')\cdot ds'da'. \tag{B.15}$$

Using the definition in (B.15), we can write

$$|Q_{\pi_{\boldsymbol{\theta}}}(s,a) - Q_{\pi_{\boldsymbol{\theta}'}}(s,a)| \leq \frac{\gamma U_R}{1-\gamma}\|\mu^{sa}_{\boldsymbol{\theta}}(s',a') - \mu^{sa}_{\boldsymbol{\theta}'}(s',a')\|_1. \tag{B.16}$$

Utilizing the upper bound in (B.13), we can write Using the definition in (B.15), we can write

$$|Q_{\pi_{\boldsymbol{\theta}}}(s,a) - Q_{\pi_{\boldsymbol{\theta}'}}(s,a)| \leq \frac{\gamma U_R M_\rho}{1-\gamma}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|. \tag{B.17}$$

as stated in Lemma B.1(2).

■

Now we shift focus to prove the statement of Lemma 5.2 as follows.

Proof [Proof of Lemma 5.2] To obtain a bound on the gradient norm difference for J, we start by considering the term $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\|$ and expanding it using the definition in (4.1), we get

$$\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\| \leq \frac{1}{1-\gamma} \Bigg\| \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) \nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da - \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}_2}}(s,a) \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s) \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da \Bigg\|. \tag{B.18}$$

Add and subtract the terms $Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s) \rho_{\pi_{\boldsymbol{\theta}_1}}(s) \pi_{\boldsymbol{\theta}_1}(a|s)$ and $Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)$ $\rho_{\pi_{\boldsymbol{\theta}_2}}(s) \pi_{\boldsymbol{\theta}_1}(a|s)$ inside the first integral of (B.18) and use triangle inequality to obtain

$$\begin{aligned} &\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\| \\ &\leq \underbrace{\frac{1}{1-\gamma} \Bigg\| \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) \Big(\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s) \Big) \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da \Bigg\|}_{\mathbf{I}_1} \\ &\quad + \underbrace{\frac{1}{1-\gamma} \Bigg\| \int_{\mathcal{S}\times\mathcal{A}} \Big(Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) - Q_{\pi_{\boldsymbol{\theta}_2}}(s,a) \Big) \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s) \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da \Bigg\|}_{\mathbf{I}_2} \\ &\quad + \underbrace{\frac{1}{1-\gamma} \Bigg\| \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s) \Big(\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) \Big) ds\, da \Bigg\|}_{\mathbf{I}_3}. \end{aligned} \tag{B.19}$$

Hence, we could write the equation in (B.19) as follows using the triangle inequality:

$$\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\| \leq \mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3. \tag{B.20}$$

Now we will bound each of the above terms separately. Let us start with $\mathbf{I}_1$ and take the norm inside the integral to get

$$\mathbf{I}_1 \leq \frac{1}{1-\gamma} \int_{\mathcal{S}\times\mathcal{A}} |Q_{\pi_{\boldsymbol{\theta}_1}}(s,a)| \, \|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \, \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da, \tag{B.21}$$

From Assumption 4.1, we have $|Q_{\pi_{\boldsymbol{\theta}_1}}(s,a)| \leq \frac{U_R}{(1-\gamma)}$ which may be substituted into the right-hand side of (B.21) as follows

$$\mathbf{I}_1 \leq \frac{U_R}{(1-\gamma)^2} \int_{\mathcal{S}\times\mathcal{A}} \|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da. \tag{B.22}$$

From Assumption (4.4) regarding the Hölder continuity of the score function, we have

$$\|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \leq M \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\beta}, \tag{B.23}$$

where $\beta \in (0,1]$. This expression (B.23) may be substituted into (B.22) to write

$$\mathbf{I}_1 \leq \frac{U_R M}{(1-\gamma)^2}\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\beta} \int_{\mathcal{S}\times\mathcal{A}} \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a)\cdot \, ds\, da\right). \tag{B.24}$$

The above integral is a valid probability measure which further implies that it integrates to unit. Therefore, we have that

$$\mathbf{I}_1 \leq \frac{U_R M \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\beta}}{(1-\gamma)^2}. \tag{B.25}$$

Now let us consider the expression associated with $\mathbf{I}_2$ in (B.19) and take the norm inside the integral , we write

$$\mathbf{I}_2 \leq \frac{1}{1-\gamma}\int_{\mathcal{S}\times\mathcal{A}} \left|Q_{\pi_{\boldsymbol{\theta}_1}}(s,a) - Q_{\pi_{\boldsymbol{\theta}_2}}(s,a)\right| \cdot \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da. \tag{B.26}$$

Note that Q function is Lipschitz as given in (B.2), hence we can upper bound (B.26) as follows

$$\begin{aligned}
\mathbf{I}_2 \leq & \frac{M_Q}{1-\gamma}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \int_{\mathcal{S}\times\mathcal{A}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da & \text{(B.27)}\\
\leq & \frac{M_Q B^{1/2}}{1-\gamma}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|. & \text{(B.28)}
\end{aligned}$$

Now, we are only left to bound $\mathbf{I}_3$ in (B.20). Let us rewrite $\mathbf{I}_3$ as follows

$$\begin{aligned}
\mathbf{I}_3 = & \frac{1}{1-\gamma}\left\| \int_{\mathcal{S}\times\mathcal{A}} Q_{\pi_{\boldsymbol{\theta}_1}}(s,a)\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\left(\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)\right) ds\, da\right\| \\
\leq & \frac{1}{1-\gamma}\int_{\mathcal{S}\times\mathcal{A}} |Q_{\pi_{\boldsymbol{\theta}_1}}(s,a)| \cdot \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot |\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)| ds\, da. & \text{(B.29)}
\end{aligned}$$

Using the bound $|Q_{\pi_{\boldsymbol{\theta}_1}}(s,a)| \leq \frac{U_R}{(1-\gamma)}$, we can write

$$\mathbf{I}_3 \leq \frac{U_R}{(1-\gamma)^2}\int_{\mathcal{S}\times\mathcal{A}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot |\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)| ds\, da. \tag{B.30}$$

Next, we need to bound the right-hand side of (B.30). According to definitions provided in (5.1)-(5.2), for any s, there exists a set $\mathcal{C} \in \mathcal{A}(\lambda)$ s.t.

$$\int_{\mathcal{S}}\int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \cdot \rho_{\boldsymbol{\theta}}(s,a) \cdot ds da \leq \lambda, \tag{B.31}$$

and $\sup_{a\in\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \leq B(\lambda)$. We proceed to upper bound the right-hand side of (B.30) by splitting this integral over the action space into two parts and employing the quantities in Definition 5.1, specifically, (5.1), as

$$\mathbf{I}_3 \leq \frac{U_R}{(1-\gamma)^2}\int_{\mathcal{S}}\int_{\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot |\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)| ds\, da.$$

$$+ \frac{U_R}{(1-\gamma)^2} \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot |\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)| ds\, da. \tag{B.32}$$

From (5.2) and the definition of $\|\cdot\|_1$, we can write

$$\begin{aligned}\mathbf{I}_3 \leq & \frac{U_R B(\lambda)}{(1-\gamma)^2} \|\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)\|_1 \\ & + \frac{U_R}{(1-\gamma)^2} \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot |\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) - \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)| ds\, da. \end{aligned} \tag{B.33}$$

From the Lipschitz continuity of the occupancy measure in (B.1) and triangle inequality, we may write

$$\begin{aligned}\mathbf{I}_3 \leq & \frac{U_R B(\lambda) M_\rho}{(1-\gamma)^2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\ & + \frac{U_R}{(1-\gamma)^2} \underbrace{\int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot \left(\rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) + \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a)\right) ds\, da}_{\mathcal{Z}}. \end{aligned} \tag{B.34}$$

Let us focus on the second term $\mathcal{Z}$ of the right-hand side of (B.34). We expand its expression using (5.1) as

$$\begin{aligned}\mathcal{Z} = & \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da \\ & + \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da. \end{aligned} \tag{B.35}$$

Next, from the Hölder continuity of the score function, by adding and subtracting $\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s)$ inside the norm for the first term, we may write

$$\begin{aligned}\mathcal{Z} \leq & M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^\beta + \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s)\| \cdot \rho_{\pi_{\boldsymbol{\theta}_1}}(s,a) ds\, da \\ & + \int_{\mathcal{S}} \int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \cdot \rho_{\pi_{\boldsymbol{\theta}_2}}(s,a) ds\, da \\ \leq & M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^\beta + 2\lambda. \end{aligned} \tag{B.36}$$

Substituting (B.36) into the right hand side of (B.34), we get

$$\mathbf{I}_3 \leq \frac{U_R M}{(1-\gamma)^2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^\beta + \frac{U_R B(\lambda) M_\rho}{(1-\gamma)^2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \frac{2U_R \lambda}{(1-\gamma)^2}. \tag{B.37}$$

Next, substituting the upper bounds for $\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$ into the right hand side of (B.20), we get

$$\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_2)\| \leq M_J [\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^\beta + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \lambda]$$

where M_J is defined as

$$M_J := \max\left\{ \frac{2U_R M}{(1-\gamma)^2}, \frac{M_Q B^{1/2}}{1-\gamma} + \frac{U_R B(\lambda) M_\rho}{(1-\gamma)^2}, \frac{2U_R}{(1-\gamma)^2} \right\}. \tag{B.38}$$

which concludes the proof of Lemma 5.2. ■

Appendix C. Proof of Lemma 5.3

Proof The proof technique is motivated from the analysis in Nguyen et al. (2019a). Consider a curve $g(t) \triangleq J(\boldsymbol{\theta}_2 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2))$. Then $g'(t) = \langle \nabla J(\boldsymbol{\theta}_2 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle$. The integral of $g'(t)$ from $t = 0$ to $t = 1$ can be expressed as

$$\int_{t=0}^{t=1} g'(t)dt = g(1) - g(0) = J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2).$$

by the Fundamental theorem of calculus. Now subtracting $\langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle$ on the both sides of the above expression, we get

$$|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| = \left| \int_{t=0}^{t=1} g'(t)dt - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \right| \tag{C.1}$$

Using the expression for $g'(t)$ above expression takes the form

$$\begin{aligned} &\big|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \big| \\ &\quad = \left| \int_{t=0}^{t=1} \langle \nabla J(\boldsymbol{\theta}_2 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \, dt - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \right| \\ &\quad = \left| \int_{t=0}^{t=1} \langle \nabla J(\boldsymbol{\theta}_2 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)) - \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \, dt \right|. \end{aligned} \tag{C.2}$$

Using Cauchy-Schwartz inequality for inner product on the right-hand side of the previous expression

$$|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| \leq \int_{t=0}^{t=1} \|\nabla J(\boldsymbol{\theta}_2 + t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)) - \nabla J(\boldsymbol{\theta}_2)\| \, \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \, dt. \tag{C.3}$$

Apply Lemma 5.2 to the first factor of the integrand on the right-hand side, which simplifies the preceding expression to

$$\begin{aligned} &\big|J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - \langle \nabla J(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \big| \\ &\quad \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \int_{t=0}^{t=1} M_J \left[\|t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^{\beta} + \|t(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\| + \lambda \right] dt \\ &\quad \leq M_J \left[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{1+\beta} + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + \lambda \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \right], \end{aligned} \tag{C.4}$$

which is as stated in (5.4). ■

Appendix D. Proof of Theorem 5.4

Proof We begin by unraveling the statement of Lemma (5.3) to write an approximate ascent relationship on the objective $J(\boldsymbol{\theta})$ as:

$$J(\boldsymbol{\theta}_{k+1}) \geq J(\boldsymbol{\theta}_k) + \langle \nabla J(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle$$

$$- M_J \left[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^{1+\beta} + \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2 + \lambda \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\| \right]. \tag{D.1}$$

Substitute the expression for the policy gradient (4.2) in place of $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$ into (D.1) as

$$\begin{aligned} J(\boldsymbol{\theta}_{k+1}) \geq & J(\boldsymbol{\theta}_k) + \eta \left\langle \nabla J(\boldsymbol{\theta}_k), \hat{\nabla} J(\boldsymbol{\theta}_k) \right\rangle \\ & - M_J \left[\|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\|^{1+\beta} + \|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\|^2 + \lambda \|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\| \right]. \end{aligned} \tag{D.2}$$

For $c = 1, 1+\beta, 2$, using Assumption 4.1 along with the Jensen's inequality indicates that

$$\begin{aligned} \|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\|^c = & \left\| \eta \sum_{t=0}^{T_k} \gamma^{t/2} \cdot R(s_t, a_t) \cdot \left(\sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau) \right) \right\|^c \\ \leq & \eta^c U_R^c \cdot \left(\sum_{t=0}^{T_k} \gamma^{t/2} \right)^c \cdot \left\| \sum_{t=0}^{T_k} \frac{\gamma^{t/2}}{\sum_{t=0}^{T_k} \gamma^{t/2}} \sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau) \right\|^c \\ \leq & \eta^c U_R^c \cdot \left(\sum_{t=0}^{T_k} \gamma^{t/2} \right)^{c-1} \cdot \sum_{t=0}^{T_k} \gamma^{t/2} \left\| \sum_{\tau=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau) \right\|^c. \end{aligned} \tag{D.3}$$

Applying the Jensen's inequality again, we can write

$$\begin{aligned} \|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\|^c \leq & \eta^c U_R^c \cdot \left(\sum_{t=0}^{T_k} \gamma^{t/2} \right)^{c-1} \cdot \sum_{t=0}^{T_k} \gamma^{t/2} (t+1)^{c-1} \sum_{\tau=0}^{t} \|\nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau)\|^c \\ \leq & \eta^c U_R^c \cdot T_k \cdot \sum_{t=0}^{T_k} \gamma^{t/2} (t+1)^{c-1} \sum_{\tau=0}^{t} \|\nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau)\|^c. \end{aligned} \tag{D.4}$$

Taking the expectation on both sides in (D.4), we have

$$\begin{aligned} \mathbb{E}[\|\eta \hat{\nabla} J(\boldsymbol{\theta}_k)\|^c] \leq & \eta^c U_R^c \cdot \mathbb{E}\left[T_k \cdot \sum_{t=0}^{T_k} \gamma^{t/2} (t+1)^{c-1} \sum_{\tau=0}^{t} \|\nabla \log \pi_{\boldsymbol{\theta}_k}(a_\tau \mid s_\tau)\|^c \right] \\ \leq & \eta^c U_R^c \cdot \sum_{T=0}^{+\infty} (1 - \gamma^{1/2}) \gamma^{T/2} \cdot T \cdot \sum_{t=0}^{T} \gamma^{t/2} (t+1)^c B^{c/2} \\ \leq & \eta^c U_R^c B^{c/2} \cdot \sum_{T=0}^{+\infty} (1 - \gamma^{1/2}) \gamma^{T/2} \cdot (T+1) \cdot \sum_{t=0}^{+\infty} \gamma^{t/2} (t+1)(t+2) \\ = & \frac{2 \eta^c U_R^c B^{c/2}}{(1 - \gamma^{1/2})^4}. \end{aligned} \tag{D.5}$$

Taking expectation on the both sides of (D.2) conditioning on $\boldsymbol{\theta}_k$, denoted as $\mathbb{E}_k$ and utilizing the bound in (D.5), gives

$$\mathbb{E}_k \left[J(\boldsymbol{\theta}_{k+1}) \right] \geq J(\boldsymbol{\theta}_k) + \eta \|\nabla J(\boldsymbol{\theta}_k)\|^2 - \frac{2 M_J}{(1 - \gamma^{1/2})^4} \left(\eta^2 U_R^2 B + \eta^{1+\beta} U_R^{1+\beta} B^{\frac{1+\beta}{2}} + \eta \lambda U_R B^{1/2} \right) \tag{D.6}$$

After rearranging the term and defining,

$$L_J := \frac{2M_J}{(1-\gamma^{1/2})^4} \cdot \max\left\{U_R^2 B, U_R^{1+\beta} B^{\frac{1+\beta}{2}}, U_R B^{1/2}\right\}, \tag{D.7}$$

we can write (D.6) as

$$\begin{aligned}\mathbb{E}_k\left[J(\boldsymbol{\theta}_{k+1})\right] \geq & J(\boldsymbol{\theta}_k) + \eta\|\nabla J(\boldsymbol{\theta}_k)\|^2 - L_J\left(\eta^{1+\beta} + \eta^2 + \eta\lambda\right) &\text{(D.8)}\\ \geq & J(\boldsymbol{\theta}_k) + \eta\|\nabla J(\boldsymbol{\theta}_k)\|^2 - 2L_J\eta^{1+\beta} - L_J\eta\lambda. &\text{(D.9)}\end{aligned}$$

Let J^* be the optimal function value, then it holds that $J(\boldsymbol{\theta}_{k+1}) \leq J^*$. Calculating the total expectation in (D.6) and taking sum from $k = 0, \ldots, K-1$, we get

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right] \leq \frac{J^* - J(\boldsymbol{\theta}_0)}{\eta} + 2KL_J\eta^{\beta} + KL_J\lambda. \tag{D.10}$$

Divide both sides by K, we get

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right] \leq \frac{J^* - J(\boldsymbol{\theta}_0)}{\eta K} + 2L_J\eta^{\beta} + L_J\lambda. \tag{D.11}$$

Now we specify the step-size as a constant $\eta = c_\beta / K^{\frac{1}{1+\beta}}$ with $c_\beta = \left(\frac{1}{2\beta L_J}\left(J^* - J(\boldsymbol{\theta}_0)\right)\right)^{1/(1+\beta)}$. Doing so permits us to rewrite (D.11) as follows

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right] \leq \frac{a_\beta}{K^{\frac{\beta}{1+\beta}}} + \mathcal{O}(\lambda), \tag{D.12}$$

where we define the constant $a_\beta = (2L_J)^{1/(\beta+1)}\left(J^* - J(\boldsymbol{\theta}_0)\right)^{\beta/(\beta+1)}$, as stated in (5.4). Observe that in existing analyses in the literatureZhang et al. (2020c), the $\mathcal{O}(\lambda)$ term is assumed to be null. Thus, standard rates are recovered as a special case. ■

Appendix E. Instantiations of λ in Example 1-3

In this section, we discuss about the parameter λ in detail. Note that the specific value of λ would depend upon the policy class being considered. Therefore, we derive the values of λ for Example 2-3 (Example 1 is special case of Example 2 for $\alpha = 2$). For the sake of analysis in this section, we assume that θ belongs to some compact set Θ.

(1) We start with the Example 3, for which we note that the score function is absolutely bounded over the full action space $\mathcal{A}$. For this case, then, $\lambda = 0$ and $B(\lambda)$ exists and is finite.

(2) For the moderate tail case (Example 2), note that the policy distribution is given by

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sigma\mathcal{A}_\alpha}\exp\left\{-\frac{\|a - \phi(s)^T\boldsymbol{\theta}\|^\alpha}{\sigma^\alpha}\right\}. \tag{E.1}$$

Therefore the score function could be written as

$$\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\| = \sigma^{-\alpha}\|a - \phi(s)^T\boldsymbol{\theta}\|^{\alpha-1}\|\phi(s)\| \leq \sigma^{-\alpha}\mathcal{D}_\phi\|a - \phi(s)^T\boldsymbol{\theta}\|^{\alpha-1} \tag{E.2}$$

as long as the feature map is bounded as $\|\phi(s)\| \leq \mathcal{D}_\phi$. Suppose θ belongs to some compact set Θ. Let us construct the set $\mathcal{C}$ as

$$\mathcal{C} := \{a \in \mathcal{A} : \exists \boldsymbol{\theta} \in \Theta \; s.t. \; |a - \phi(s)^T\boldsymbol{\theta}| \leq R\}$$

where R is a finite positive constant. Now, let us look at the following integral

$$\begin{aligned}
&\int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot da \\
&\quad \leq \sigma^{-\alpha}\mathcal{D}_\phi \int_{\mathcal{A}\setminus\mathcal{C}} \|a - \phi(s)^T\boldsymbol{\theta}\|^{\alpha-1} \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot da
\end{aligned} \tag{E.3}$$

which follows from the upper bound in (E.2). From the definition in (E.1), we can write

$$\begin{aligned}
&\int_{\mathcal{A}\setminus\mathcal{C}} \|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \cdot \pi_{\boldsymbol{\theta}}(a|s) \cdot da \\
&\leq \frac{\mathcal{D}_\phi}{\sigma^{1+\alpha}\mathcal{A}_\alpha} \int_{\mathcal{A}\setminus\mathcal{C}} \|a - \phi(s)^T\boldsymbol{\theta}\|^{\alpha-1} \cdot \exp\left\{-\frac{\|a-\phi(s)^T\boldsymbol{\theta}\|^\alpha}{2\sigma^\alpha}\right\} \cdot \exp\left\{-\frac{\|a - \phi(s)^T\boldsymbol{\theta}\|^\alpha}{2\sigma^\alpha}\right\} da \\
&\leq \frac{\mathcal{D}_\phi}{\sigma^{1+\alpha}\mathcal{A}_\alpha} \int_{\mathcal{A}\setminus\mathcal{C}} \|a - \phi(s)^T\boldsymbol{\theta}\|^{\alpha-1} \cdot \exp\left\{-\frac{\|a - \phi(s)^T\boldsymbol{\theta}\|^\alpha}{2\sigma^\alpha}\right\} \cdot da \cdot \exp\left\{-\frac{R^\alpha}{2\sigma^\alpha}\right\} \\
&\leq \frac{\mathcal{D}_\phi}{\sigma^{1+\alpha}\mathcal{A}_\alpha} \cdot \sigma^\alpha \cdot B_\alpha \cdot \exp\left\{-\frac{R^\alpha}{2\sigma^\alpha}\right\} \\
&\leq \frac{\mathcal{D}_\phi}{\sigma\mathcal{A}_\alpha} \cdot B_\alpha \cdot \exp\left\{-\frac{R^\alpha}{2\sigma^\alpha}\right\},
\end{aligned} \tag{E.4}$$

where $B_\alpha := \int |a|^{\alpha-1} \exp\{-\frac{|a|^\alpha}{2}\} < \infty$. The above equation will be less than λ if we have

$$\frac{\mathcal{D}_\phi}{\sigma\mathcal{A}_\alpha} \cdot B_\alpha \cdot \exp\left\{-\frac{R^\alpha}{2\sigma^\alpha}\right\} \leq \lambda, \tag{E.5}$$

which implies that

$$\left(\frac{R}{\sigma}\right)^\alpha \geq 2\log\left(\frac{\mathcal{D}_\phi B_\alpha}{\sigma\mathcal{A}_\alpha\lambda}\right). \tag{E.6}$$

The above expression provides the bound for $B(\lambda)$ as

$$\begin{aligned}
B(\lambda) &\leq \max_{a\in\mathcal{C}} \max_{s\in\mathcal{S}} \max_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\| \leq \max_{a\in\mathcal{C}} \max_{s\in\mathcal{S}} \max_{\boldsymbol{\theta}\in\Theta} \frac{\mathcal{D}_\phi}{\sigma} \left(\frac{\|a - \phi(s)^T\boldsymbol{\theta}\|}{\sigma}\right)^{\alpha-1} \\
&\leq \frac{\mathcal{D}_\phi}{\sigma} \left(\frac{\mathcal{D}_\Theta\mathcal{D}_\phi}{\sigma} + 2\log\left(\frac{\mathcal{D}_\phi B_\alpha}{\sigma\mathcal{A}_\alpha\lambda}\right)\right)^{\frac{\alpha-1}{\alpha}} \\
&= \mathcal{O}\left(\log\frac{1}{\lambda}\right)^{\frac{\alpha-1}{\alpha}},
\end{aligned} \tag{E.7}$$

where $\mathcal{D}_\Theta := \max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ is the diameter of Θ. Note that the bound in (E.7) is small even for a very small value of λ. This permits us to relax the standard assumption of absolutely bounded score function for continuous action spaces.

Appendix F. Exit Time Analysis

We first present the technical preliminaries required for the analysis in this section as follows.

F.1 Technical Preliminaries

Technical Results for Univariate Case $d = 1$. We continue then with the formal definition of first exit time for a SDE in continuous one dimensional case for simplicity. Consider a neighborhood $B_i := [-b, a]$ around the i-th local extrema $\bar{\boldsymbol{\theta}}_i$ in a single dimension. Let the process is initialized at $\boldsymbol{\theta}_0$ which is inside B_i. We are interested in the first exit time from B_i starting from a point $\boldsymbol{\theta}_0 \in B_i$. The first exit time from $B_i := [-b, a]$ for a process defined by continuous SDE is defined as

$$\hat{\tau}(\epsilon) = \inf\{t \geq 0 : \boldsymbol{\theta}_t^\epsilon \notin [-b, a]\},$$

as random perturbation in the SDE, $\epsilon \to 0$, [cf. (5.7)]. We denote the first exit time for continuous SDE using $\hat{\tau}(\cdot)$. Under the Assumption 5.7 (3.), we invoke results from Imkeller and Pavlyukevich (2006) for the first exit times of continuous SDEs (5.7) in the univariate case $d = 1$.

Further, we impose that the multi-dimensional Lévy motion $\mathbf{L}_t^\alpha$ in (5.7) admits a representation as a $\mathbf{L}_t^\alpha = \mathbf{r}L_t$ with $\mathbf{r} \in \mathbb{R}^d$ as a standard basis vector in d-dimensions, which determines the direction of the jump process, and L_t is a scalar α-stable Lévy motion. This restriction is needed in order to tractably study the transient behavior of (5.7) in terms of its exit time from regions of attraction Imkeller et al. (2010), specifically, in applying Lemma F.10, as well as characterizing the proportion of time jumping between its limit points, to be discussed next. We note that such analyses for general d-dimensional Lévy motion is an open problem in stochastic processes.

Theorem F.1 *Imkeller and Pavlyukevich (2006) Consider the SDE* (5.7)*, in the univariate case $d = 1$ ($\theta \leftarrow \boldsymbol{\theta}$) and assume that it has a unique strong solution. Assume further that the there exists an objective J with a global maximum at zero, satisfying the conditions $J'(\theta)\theta < 0$ for every $\theta \in \mathbb{R}$, $J(0) = 0$, $J'(\theta) = 0$ if and only if $\theta = 0$ and $J''(0) < 0$. Then, there exist positive constants ϵ_0, γ, δ, and $C > 0$ such that for $0 < \epsilon \leq \epsilon_0$, the following holds in the limit of small ϵ:*

$$\exp\left(-u\epsilon^\alpha\left(\frac{1}{a^\alpha}\right)\frac{(1+C\epsilon^\delta)}{\alpha}\right)(1-C\epsilon^\delta) \leq \mathcal{P}(\hat{\tau}(\epsilon) > u) \leq \exp\left(-u\epsilon^\alpha\left(\frac{1}{a^\alpha}\right)\frac{(1+C\epsilon^\delta)}{\alpha}\right)(1+C\epsilon^\delta) \tag{F.1}$$

uniformly for all $\theta \leq a - \epsilon^\gamma$ and $u \geq 0$. Consequently

$$\mathbb{E}\left[\hat{\tau}_a(\epsilon)\right] = \frac{\alpha a^\alpha}{\epsilon^\alpha}(1 + \mathcal{O}(\epsilon^\delta)) \tag{F.2}$$

uniformly for all $\theta \leq a - \epsilon^\gamma$.

Theorem F.2 *Imkeller and Pavlyukevich (2006), Consider the SDE* (5.7)*, in dimension $d = 1$ and assume that it has a unique strong solution. Assume further that there exists an*

objective J with a global maximum at zero, satisfying the conditions $J'(\theta)\theta < 0$ for every $\theta \in \mathbb{R}$, $J(0) = 0$, $J'(\theta) = 0$ if and only if $\theta = 0$ and $J^{''}(0) < 0$, the following results hold in the limit if small ϵ:

1. *First exit time is exponentially large in ϵ^{-2}. Assume for definiteness $J(a) > J(-b)$. Then for any $\delta > 0$, $\theta \in B_i$.*

$$\mathcal{P}_{\boldsymbol{\theta}}(\exp\left(-2J(a) - \delta\right)/\epsilon^2 < \hat{\tau}(\epsilon) < \exp\left(-2J(a) + \delta\right)/\epsilon^2) \to 1 \; as \, \epsilon \to 0 \tag{F.3}$$

   Moreover, $\epsilon^2 \log \mathbb{E}_{\boldsymbol{\theta}}[\hat{\tau}(\epsilon)] \to 2J(a)$.

2. *The mean of first exit time is given by*

$$\mathbb{E}_\theta(\hat{\tau}(\epsilon) \approx \frac{\epsilon\sqrt{\pi}}{J'(a)\sqrt{J''(0)}} \exp\left(2J(a)/\epsilon^2\right) \tag{F.4}$$

3. *Normalized first exit time is exponentially distributed: for $u \geq 0$*

$$\mathcal{P}_{\boldsymbol{\theta}}\left(\frac{\hat{\tau}(\epsilon)}{\mathbb{E}_\theta(\hat{\tau}(\epsilon)} > u\right) \to \exp\left(-u\right) \, as \, \epsilon \to 0 \tag{F.5}$$

   uniformly in θ on compact subsets of $(-b, a)$.

Note that the above results hold for continuous SDEs in one dimensional space.
Next we present extend the exit time results from a domain $\mathcal{G}_i \subset \mathbb{R}^d$ around i-th local maxima of $J(\cdot)$, $\bar{\boldsymbol{\theta}}_i$ Imkeller et al. (2010) with an assumption that the system is perturbed by a single one-dimensional Lévy process with α-stable component.
Multi-Dimensional Case $d > 1$. Before proceeding to the the statement of results and proofs, we define assumptions and terminologies associated with the multi-dimensional space, i.e., subsequently $\boldsymbol{\theta} \in \mathbb{R}^d$. The reason for separately stating the scalar case and the multi-dimensional case, is that results and conditions for the scalar-dimensional case are invoked in generalizing to the multi-dimensional case, specifically, in Lemmas F.8 and F.9. For simplicity, we assume the tail-index (α) of perturbations are identical in all the directions. The dynamical system of (5.7) when perturbed by single dimensional Lévy process is given by

$$\boldsymbol{\theta}_t^\epsilon(\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 + \int_0^t b(\boldsymbol{\theta}_s^\epsilon(\boldsymbol{\theta}_0))ds + \epsilon \mathbf{r} L_t^i, \epsilon > 0, \boldsymbol{\theta} \in \mathcal{G}, t \geq 0 \tag{F.6}$$

where, $\mathbf{r} \in \mathbb{R}^d$ is the unit vector.

We define the inner parts of $\mathcal{G}_i$ by $\mathcal{G}_{i_{\bar{\delta}}} := \{z \in \mathcal{G}_i : \text{dist}(z, \partial\mathcal{G}_i) \geq \bar{\delta}\}$. Therefore, the following holds: Sets $\mathcal{G}_{\bar{\delta}}$ are positively invariant for all $\bar{\delta} \in (0, a + \xi)$ (cf. (5.8)), in the sense that the deterministic solutions starting in $\mathcal{G}_{\bar{\delta}}$ do not leave this set for all times $t \geq 0$. We have $\Omega^-(\bar{\delta}) \cap \mathcal{G}_{i_{\bar{\delta}}}^c(\bar{\delta}) = \emptyset$ and $\Omega^+(\bar{\delta}) \cap \mathcal{G}_{i_\delta}^c(\bar{\delta}) = \emptyset$. The preceding statements follows from Imkeller et al. (2010).

Next we state exit time results from the domain, $\mathcal{G}_i$ for a system defined by (F.6).

Theorem F.3 *Expressions 3.4, 3.8 Imkeller et al. (2010) For $\bar{\delta} \in (0, \bar{\delta}_0)$ and initial state $\boldsymbol{\theta}_0 \in \mathcal{G}_i$, $\boldsymbol{\theta}_t^\epsilon$ following* (F.6) *exits from the domain $\mathcal{G}_i$ in a little tube in the direction of α, Furthermore, for every $\bar{\delta} \in (0, \bar{\delta}_0)$ the probability to exit in direction of perturbation $+\mathbf{r}$ is given by*

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}_{\hat{\tau}}^\epsilon \in \Omega_i^+(\delta)\right) = \frac{2}{\epsilon^{\rho\alpha}} \epsilon^\alpha (d^+)^{-\alpha}, \tag{F.7}$$

where $(a, b) > 0$, $\rho \in (0, 1)$, $\min(a, b) > \epsilon^{1-\rho}$.

Furthermore, for every $\bar{\delta} \in (0, \bar{\delta}_0)$ the probabilities to exit in direction $\pm\mathbf{r}$ are given by

$$\lim_{\epsilon \to 0} \mathcal{P}\left(\boldsymbol{\theta}_{\hat{\tau}}^\epsilon \in \Omega_i^+(\bar{\delta})\right) = \frac{p^+}{p_s} \tag{F.8}$$

$$\lim_{\epsilon \to 0} \mathcal{P}\left(\boldsymbol{\theta}_{\hat{\tau}}^\epsilon \in \Omega_i^-(\bar{\delta})\right) = \frac{p^-}{p_s}, \tag{F.9}$$

for all $\boldsymbol{\theta}_t \in \mathcal{G}$ and

$$p_s := ((d^+)^{-\alpha} + (-d^-)^{-\alpha}) \tag{F.10}$$

$$p^+ := (d^+)^{-\alpha} \tag{F.11}$$

$$p^- := (-d^-)^{-\alpha}, \tag{F.12}$$

d^+ and d^- define distance from boundary of interest $\partial\mathcal{G}_i$ along $\pm\mathbf{r}$.

The above expression is obtained by using a single $\mathbf{r}$ in (3.4) of Imkeller et al. (2010). It is to be noted that for a general process perturbed by finitely many single dimensional Lévy processes with different tail indices, the exit time depends on the smallest tail-index, and the system exits from the domain in the direction of the process with smallest α_i.

In this section we derive the first exit time behavior for the proposed heavy tailed setting of Algorithm 1. We proceed by defining some key quantities of interest and lemmas used in the proof of Theorem 5.11. The first is the Itô formula for stochastic differential equations, and then we present the Bellman-Gronwall inequality.

Definition F.4 *(Itô formula)Xie et al. (2020)* Let N be a Poisson random measure with intensity measure $dt\nu(dz)$, where ν is a Lévy measure on $\mathbb{R}^d$, ie., $\int_{\mathbb{R}^d}(|z|^2 \wedge 1)\nu(dz) < +\infty$, $\nu(\{0\}) = 0$. The compensated Poisson random measure $\widetilde{N}$ is defined as $\widetilde{N}(dt, dz) := N(dt, dz) - dt\nu(dz)$. Consider the following SDE in $\mathbb{R}^d$ with jumps:

$$dX_t = b_t(X_t)dt + \int_{|z|<R} g_t(X_{t-}, Z)\widetilde{N}(dt, dz) + \int_{|z|\geq R} g_t(X_{t-}, Z)N(dt, dz),$$

where $R > 0$ is a fixed constant. Suppose $g(\mathbf{x}) \in \mathcal{C}^2(\mathbb{R})$ is a twice continuously differentiable function (in particular all second-partial-derivatives are continuous functions). Suppose $Y_t = g(\mathbf{X}_t)$ is again an Itô process, then we have

$$dY_t = [\mathcal{L}_1^{b_t} g + \mathcal{L}_\nu^g h](\mathbf{X}_t)dt + dM(t),$$

where M_t is local martingale, $\mathcal{L}_1^{b_t}$ is the first order differential operator associated with drift (b_t), and $\mathcal{L}_\nu^g h$ is the non-local operator associated with jump coefficient $g()$ such that:

$$\mathcal{L}_\nu^g u(x) := \int_{|z|<R} \left[u(x + g_t(x,z)) - u(x) - g_t(x,z) \cdot \nabla u(x)\right] \nu(dz)$$
$$\int_{|z|\geq R} \left[u(x + g_t(x,z)) - u(x)\right] \nu(dz)$$

Definition F.5 *(Bellman-Gronwall inequality)* Assume $\phi : [0,T] \to \mathbb{R}$ is a bounded nonnegative measurable function, $C : [0,T] \to \mathbb{R}$ is a nonnegative integrable function and $B \geq 0$ is a constant with the property that

$$\phi(t) \leq B + \int_0^t C(\tau)\phi(\tau)d\tau \forall t \in [0,T]. \tag{F.13}$$

Then

$$\phi(t) \leq B \exp\left(\int_{t=0}^T C(\tau)d\tau\right) \forall t \in [0,T]. \tag{F.14}$$

Next we provide a lemma regarding the difference between a Lévy process with two different tail indices α and $k\eta$.

Lemma F.6 *Thanh et al. (2019) For any $u > 0$, $\eta > 0$ and $K \in \mathbb{N}$, there exist a constant C_α such that:*

$$\max_{k\in 0,\ldots,K-1} \mathcal{P}[\sup_{t\in[k\eta,(k+1)\eta]} \|L^\alpha(t) - L^{k\eta}(t)\| \geq u] \leq C_\alpha d^{1+\frac{\alpha}{2}} \eta u^{-\alpha} \tag{F.15}$$

and

$$\mathcal{P}[\max_{k\in 0,\ldots,K-1} \sup_{t\in[k\eta,(k+1)\eta]} \|L^\alpha(t) - L^{k\eta}(t)\| \geq u] \leq 1 - (1 - C_\alpha d^{1+\frac{\alpha}{2}} \eta u^{-\alpha})^K \tag{F.16}$$

Next we state a stochastic variant of Gronwall's inequality which is also used in the proof.

Lemma F.7 *Stochastic Gronwall's inequality Scheutzow (2013)* Let Z and H be nonnegative, adapted processes with continuous path and assume that ψ is nonnegative and progressively measurable. Let M be a continuous local martingale starting at 0. If

$$Z(t) \leq \int_0^t \psi(s)Z(s) + M(t) + H(t)$$

holds for all $t \geq 0$, then for $p \in (0,1)$ and μ and $\nu > 1$ such that $\frac{1}{\mu} + \frac{1}{\nu} = 1$ and $p\nu < 1$, we have

$$\mathbb{E}\left(\sup_{0\leq s\leq t} Z^p(s)\right) \leq (c_{p\nu} + 1)^{1/\nu} \left(\mathbb{E}\exp\left(p\mu \int_0^t \psi(s)ds\right)\right)^{1/\mu} (\mathbb{E}(H^*(t))^{p\nu})^{1/\nu},$$

where, a real valued process $Y^*(t) := \sup_{0\leq s\leq t} Y(s)$.

Next we provide a bound on the second moment of parameter vector $\boldsymbol{\theta}$ when integrated with respect to Lévy measure ν.

Lemma F.8 *Thanh et al. (2019) Let ν be the Lévy measure of a d-dimensional Lévy process L_α whose components are independent scalar symmetric α-stable Lévy processes $L_1, \ldots, L_d$. Then there exists a constant $C > 0$ such that the following inequality holds with $k_1 \geq 1$ and $2 > \alpha > 1$:*

$$\frac{1}{k^{2/\alpha}} \int_{\|\boldsymbol{\theta}\|<1} \|\boldsymbol{\theta}\|^2 \nu(d\boldsymbol{\theta}) + \frac{1}{2k_1^{1/\alpha}} \int_{\|\boldsymbol{\theta}\|\geq 1} \|\boldsymbol{\theta}\| \nu(d\boldsymbol{\theta}) \leq C \frac{d}{k_1^{1/\alpha}}$$

Lemma F.9 *Given the Assumptions 4.1- 5.8 and the proposed Heavy tailed setting of* (5.6)*, for $\lambda \in (0, 1)$, $\boldsymbol{\theta} \in \mathbb{R}^d$, there exists constants C_1 and C_2 depending on λ, dissipativity constants (m, b), and $\boldsymbol{\theta}_0$, the following holds on the expected value of $\boldsymbol{\theta}$ for all $t > 0$ such that*

$$\mathbb{E}\left(\sup_{s\in[0,t]} (\|\boldsymbol{\theta}_s\|)^\lambda\right) \leq C_1 \left(1 + C_2 \left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right) t\right)^\lambda, \; k_1 \geq 1, \; 1 < \alpha < 2, \tag{F.17}$$

where (m, b) are the dissipativity constants from Assumption 5.9 and k_1 is a function of stepsize η, $k_1 := 1/\eta^{\alpha-1}$.

Proof Here we derive an upper bound on $\mathbb{E}\left(\sup_{s\in[0,t]}(\|\boldsymbol{\theta}_s\|)^\lambda\right)$ for the heavy-tailed policy gradient setting of (5.6) perturbed by a single dimensional Lévy process in the direction of unit vector $\mathbf{r}$. We build up on the existing results from Thanh et al. (2019); Xie et al. (2020) for Lévy process and SDE. Using Itô formula, the whole expression is divided into two terms. We simplify the second term using properties of a Lévy process from Thanh et al. (2019); Xie et al. (2020). Further we build upon the heavy-tailed setting and its properties such as dissipativity of score function and Hólder continuity to simplify the first term. Further the combined expression is simplified using direct application of stochastic Gronwall's inequality. We start with the continuous equivalent of (5.6) defined by $d\boldsymbol{\theta}_t = b(\boldsymbol{\theta}_{t-})dt + k_1^{-1/\alpha} dL_t^\alpha$ (t_- denote left limit of the process) and $k_1 := 1/\eta^{\alpha-1}$ (k_1 is always greater than 1 as $\eta^{\alpha-1} < 1$), we use $F = -J$ for simplicity, $b(\cdot) := -\nabla F(\cdot)$ using (5.7). Note that this representation is equivalent to (5.7) and helps to invoke some of the existing results in the analysis. Here the direction of perturbation r is absorbed into L_t^α without the loss of generality (cf. (F.6)). In order to upper bound $\mathbb{E}\left(\sup_{s\in[0,t]}(\|\boldsymbol{\theta}_s\|)^\lambda\right)$, we start by defining a function, $g_1(\boldsymbol{\theta}) \triangleq (1 + \|\boldsymbol{\theta}\|^2)^{1/2}$. Using direct application of Itô's formula from Definition F.4 with jump coefficient g being $k_1^{-1/\alpha}$, we can write Itô formula for (5.6) as follows

$$\begin{aligned} dg_1(\boldsymbol{\theta}_t) = & \underbrace{\langle b(\boldsymbol{\theta}_t), \nabla g(\boldsymbol{\theta}_t))\rangle}_{T_1} dt \\ & + \underbrace{\int_{\mathbb{R}^d} \left(g_1(\boldsymbol{\theta}_t + k_1^{-1/\alpha}\boldsymbol{\theta}) - g_1(\boldsymbol{\theta}_t) - \mathbb{I}_{\|\boldsymbol{\theta}\|<1} \left\langle k_1^{-1/\alpha}, \nabla g_1(\boldsymbol{\theta}_t)\right\rangle)\nu(d\boldsymbol{\theta})\right)}_{T_2} dt + dM(t), \end{aligned} \tag{F.18}$$

here, $M(t)$ is defined as local martingale and ν be the Lévy measure of a d-dimensional Lévy process L^α . We have $\partial_i g_1(\boldsymbol{\theta}) = \boldsymbol{\theta}_i(1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2$.

Next we unfold the expressions for T_1 and T_2 using the expression for the policy gradient in (4.2).

Expression for T_1: Expression for T_1 along with the policy gradient of (4.2) takes the form

$$T_1 = \langle -\nabla F(\boldsymbol{\theta}), \nabla g_1(\boldsymbol{\theta})\rangle = \left\langle -\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{\boldsymbol{\theta}}}\left[\nabla \log \pi_{\boldsymbol{\theta}}(s,a) Q_{\pi_{\boldsymbol{\theta}}}(s,a)\right], \boldsymbol{\theta}\right\rangle (1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.19}$$

Now using Assumption 4.1 to upper bound $\|\hat{Q}_{\pi_{\boldsymbol{\theta}}}(s,a)\|$, expression reduces to

$$T_1 \leq \frac{U_R}{1-\gamma}\left\langle -\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{\boldsymbol{\theta}}}\left[\nabla \log \pi_{\boldsymbol{\theta}}(s,a)\right], \boldsymbol{\theta}\right\rangle (1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.20}$$

$$\leq \frac{U_R}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{\boldsymbol{\theta}}}\underbrace{\left[\left\langle -\frac{1}{1-\gamma}\nabla \log \pi_{\boldsymbol{\theta}}(s,a), \boldsymbol{\theta}\right\rangle\right]}_{T_3}(1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.21}$$

The expression, T_3 is simplified using (m,b,β)-dissipative assumption of Assumption 5.9 ($T_3 = -\langle \nabla \log \pi_{\boldsymbol{\theta}}(s,a), \boldsymbol{\theta}\rangle \leq -m\|\boldsymbol{\theta}\|^{1+\beta} + b$). Unfolding the expectation operator defined with respect to the occupancy measure of the MDP under policy π_θ allows us to write

$$T_1 = \langle -\nabla F(\boldsymbol{\theta}), \nabla g_1(\boldsymbol{\theta})\rangle \leq \frac{U_R}{(1-\gamma)^2}\left(\int_{\mathcal{S}\times\mathcal{A}}\left(-m\left\|\boldsymbol{\theta}\right\|^{1+\beta}+b\right)(1-\gamma)\rho_{\pi_{\boldsymbol{\theta}}}(s)\cdot\pi_{\boldsymbol{\theta}}(a|s)\,ds\,da\right) \times (1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.22}$$

The integral $\left(\int_{\mathcal{S}\times\mathcal{A}}(1-\gamma)\rho_{\pi_{\boldsymbol{\theta}}}(s)\cdot\pi_{\boldsymbol{\theta}}(a|s)\,ds\,da\right)$ from the preceding expression is a valid probability measure and hence integrates to unit. Therefore, we can simplify the right-hand side as

$$T_1 \leq \frac{U_R}{(1-\gamma)^2}\underbrace{\left(-m\left\|\boldsymbol{\theta}\right\|^{1+\beta}+b\right)}_{T_4}(1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.23}$$

Adding and subtracting m inside T_4 gives

$$T_1 \leq \frac{U_R}{(1-\gamma)^2}\left(-m\underbrace{\left(\|\boldsymbol{\theta}\|^{1+\beta}+1\right)}_{T_5}+m+b\right)(1+\|\boldsymbol{\theta}\|^2)^{-1/2}/2 \tag{F.24}$$

In order to simplify the term T_5, we evaluate $(1+\|\boldsymbol{\theta}\|^2)^{\beta_1}$.

It turns out that the application of Bernoulli's inequality is advantageous when we split the evaluation of T_5 into two cases, namely, $\|\boldsymbol{\theta}\|^2 < 1$ and $\|\boldsymbol{\theta}\|^2 > 1$ and relate the resultant inequality to $g_1(\boldsymbol{\theta})$. Consider the case where $\|\boldsymbol{\theta}\|^2 < 1$, using Bernouli's inequality for $0 < \beta_1 = (1+\beta)/2 < 1$,

$$(1+\|\boldsymbol{\theta}\|^2)^{\beta_1} \leq 1+\beta_1\|\boldsymbol{\theta}\|^2 \tag{F.25}$$

As $0 < \beta_1 < 1$, we get

$$(1 + \|\boldsymbol{\theta}\|^2)^{\beta_1} \leq 1 + \beta_1 \|\boldsymbol{\theta}\|^2 \leq 1 + \|\boldsymbol{\theta}\|^2$$

Similarly, as $0 < \|\boldsymbol{\theta}\|^2 < 1$, $\|\boldsymbol{\theta}\|^2 < \|\boldsymbol{\theta}\|^{2\beta_1}$ and we have

$$(1 + \|\boldsymbol{\theta}\|^2)^{\beta_1} \leq 1 + \|\boldsymbol{\theta}\|^2 \leq 1 + \|\boldsymbol{\theta}\|^{2\beta_1}$$

Next, consider the case, $\|\boldsymbol{\theta}\|^2 > 1$, therefore $1/\|\boldsymbol{\theta}\|^2 < 1$. Further following the same argument from previous case with Bernouli's inequality

$$\left(1 + \frac{1}{\|\boldsymbol{\theta}\|^2}\right)^{\beta_1} \leq 1 + \frac{\beta_1}{\|\boldsymbol{\theta}\|^2} \leq 1 + \frac{1}{\|\boldsymbol{\theta}\|^2} \leq 1 + \frac{1}{\|\boldsymbol{\theta}\|^{2\beta_1}} \tag{F.26}$$

Multiplying both side of the above inequality by $\|\boldsymbol{\theta}\|^{2\beta_1}$

$$\left(\|\boldsymbol{\theta}\|^2 + 1\right)^{\beta_1} \leq \|\boldsymbol{\theta}\|^{2\beta_1} + 1 \tag{F.27}$$

Therefore, using $\beta_1 = (1 + \beta)/2$, we have

$$\left(\|\boldsymbol{\theta}\|^2 + 1\right)^{(1+\beta)/2} \leq \|\boldsymbol{\theta}\|^{1+\beta} + 1 \tag{F.28}$$

Now we can write the above inequality for all values of $\|\boldsymbol{\theta}\|$ as

$$-m(\|\boldsymbol{\theta}\|^2 + 1)^{(1+\beta)/2} \geq -m(\|\boldsymbol{\theta}\|^{1+\beta} + 1) \tag{F.29}$$

Now we substitute this inequality (F.29) back into the expression for the dissipativity-based upper-bound on the policy gradient in (F.24) to write

$$\begin{aligned}
\langle -\nabla F(\boldsymbol{\theta}), \nabla g_1(\boldsymbol{\theta}) \rangle &\leq \frac{U_R}{(1-\gamma)^2} \left(-m(\|\boldsymbol{\theta}\|^2 + 1)^{(1+\beta)/2} + m + b\right) (1 + \|\boldsymbol{\theta}\|^2)^{-1/2}/2 && \text{(F.30)} \\
&= \frac{U_R}{2(1-\gamma)^2} \left(-m(\|\boldsymbol{\theta}\|^2 + 1)^{\beta/2} + (m+b) \underbrace{(1 + \|\boldsymbol{\theta}\|^2)^{-1/2}}_{\in(0,1]}\right) && \text{(F.31)} \\
&\leq \frac{U_R}{2(1-\gamma)^2} \left(-m(\|\boldsymbol{\theta}\|^2 + 1)^{\beta/2} + m + b\right) && \text{(F.32)} \\
&\leq \frac{U_R}{2(1-\gamma)^2} \left(-m g_1(\boldsymbol{\theta})^{\beta} + m + b\right). && \text{(F.33)}
\end{aligned}$$

The last inequality is obtained by plugging in the expression for $g_1(\boldsymbol{\theta}) = (1 + \|\boldsymbol{\theta}\|^2)^{1/2}$.

Expression for T_2: Similarly, we analyze the second term, T_2. Note that T_2 does not depend on the heavy-tailed gradients and simplification is based on the standard results on properties of Lévy process from Xie et al. (2020). We get

$$T_2 = \int_{\mathbb{R}^d} \left(g_1(\boldsymbol{\theta}_t + k^{-1/\alpha}\boldsymbol{\theta}) - g_1(\boldsymbol{\theta}_t) - \mathbb{I}_{\|\boldsymbol{\theta}\|<1} \left\langle k^{-1/\alpha}, \nabla g_1(\boldsymbol{\theta}_t) \right\rangle) \nu(d\boldsymbol{\theta})\right) dt \tag{F.34}$$

$$\leq \frac{1}{2k_1^{2/\alpha}} \int_{\|\boldsymbol{\theta}\|<1} \|\boldsymbol{\theta}\|^2 \nu(d\boldsymbol{\theta}) + \frac{1}{2k_1^{1/\alpha}} \int_{\|\boldsymbol{\theta}\|\geq 1} \|\boldsymbol{\theta}\| \nu(d\boldsymbol{\theta}) \leq C \frac{d}{k_1^{1/\alpha}} \tag{F.35}$$

Note that the above expression is a standard result for Lévy process (Xie et al. (2020), Expression 7.6). Using Lemma F.8 on the right-hand-side inequality, we get

$$T_2 \leq \frac{1}{2k_1^{2/\alpha}} \int_{\|\boldsymbol{\theta}\|<1} \|\boldsymbol{\theta}\|^2 \nu(d\boldsymbol{\theta}) + \frac{1}{2k_1^{1/\alpha}} \int_{\|\boldsymbol{\theta}\|\geq 1} \|\boldsymbol{\theta}\| \nu(d\boldsymbol{\theta}) \leq C \frac{d}{k_1^{1/\alpha}} \tag{F.36}$$

Using (F.33) and (F.36) in (F.18) and integrating the expression from 0 to t gives

$$g_1(\boldsymbol{\theta}_t) - g_1(\boldsymbol{\theta}_0) \leq \int_0^t \left(\frac{U_R}{(1-\gamma)^2} (-m g_1(\boldsymbol{\theta}_t)^\beta + m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) ds + M(t) \tag{F.37}$$

$$\leq \int_0^t \left(\frac{U_R}{(1-\gamma)^2} (m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) ds + M(t) \tag{F.38}$$

The above expression can be simplified using Gronwall's inequality for stochastic equations (Lemma 3.8 of Scheutzow (2013), Lemma F.7 in the Appendix) to upper bound $\mathbb{E}\left(\sup_{s\in[0,t]} g_1(\boldsymbol{\theta}_s)\right)$ in the given interval. Now following the similar argument of Thanh et al. (2019), an upper bound on expression on $\mathbb{E}\left(\sup_{s\in[0,t]} g_1(\boldsymbol{\theta}_s)\right)$ also upper bounds $\mathbb{E}\left(\sup_{s\in[0,t]} \|\boldsymbol{\theta}_s\|\right)$ as $\|\boldsymbol{\theta}\|$ is always less than $\|1 + \boldsymbol{\theta}\|$. Therefore, let us first proceed with Gronwall's inequality and obtain an upper bound on $\mathbb{E}\left(\sup_{s\in[0,t]} g_1(\boldsymbol{\theta}_s)\right)$ and further relate it to $\mathbb{E}\left(\sup_{s\in[0,t]} \|\boldsymbol{\theta}_s\|\right)$. Upon comparing the above expression with Stochastic Gronwall's inequality of Lemma F.7 (Theorem 4, Scheutzow (2013)) for nonnegative adapted processes Z and H such that

$$Z(t) \leq \int_0^t \psi(s) Z(s) + M(t) + H(t)$$

we have $g_1(\cdot)$ equivalent to $Z(\cdot)$, $H(\cdot) := \int_0^t \left(\frac{U_R}{(1-\gamma)^2} (m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) ds$, and $\sup_{s\in[0,\,t]} H^*(s) = \left(\frac{U_R}{(1-\gamma)^2} (m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) t$, $p = \lambda$. As (F.37) holds for all $t \geq 0$, using Lemma F.7 we have

$$\mathbb{E}\left(\sup_{s\in[0,t]} g_1(\boldsymbol{\theta}_s)^\lambda \right) \leq (c_{p\nu} + 1)^{1/\nu} \left(\mathbb{E} g_1(\boldsymbol{\theta}_0) + \left(\frac{U_R}{(1-\gamma)^2} (m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) t \right)^\nu \tag{F.39}$$

where, $\nu > 0$, $c_{p\nu} := (4 \wedge \frac{1}{p\nu}) \frac{\sin \pi\nu}{\pi p\nu}$ (Proposition 1, Scheutzow (2013)). Now for $\lambda\nu < 1$, $p \in (0, 1)$ from Lemma F.7, we get

$$\mathbb{E}\left(\sup_{s\in[0,t]} g_1(\boldsymbol{\theta}_s)^\lambda \right) \leq c_\lambda \left(\mathbb{E} g_1(\boldsymbol{\theta}_0) + \left(\frac{U_R}{(1-\gamma)^2} (m + b)/2 + C \frac{d}{k_1^{1/\alpha}} \right) t \right)^\lambda \tag{F.40}$$

where, $c_\lambda := (c_{p\nu}+1)^\lambda$. As we have $g_1(\boldsymbol{\theta}) \geq \|\boldsymbol{\theta}\|$, the above inequality is lower bounded by $\mathbb{E}\left(\sup_{s\in[0,t]} \|\boldsymbol{\theta}\|)^\lambda\right)$ and we get

$$\mathbb{E}\left(\sup_{s\in[0,t]} \|\boldsymbol{\theta}\|)^\lambda\right) \leq C_1\left(1 + C_2\left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right)t\right)^\lambda, \tag{F.41}$$

where C_1 and C_2 are positive constants depending on c_λ and $\mathbb{E}g_1(\boldsymbol{\theta}_0)$. ■

The following lemma upper bounds the error between (5.6) and its continuous time equivalent SDE (5.7) for $t \in [k\eta, (k+1)\eta]$ and derives its probabilistic interpretation. Note that this result is useful in translating exit time results for the proposed setting as defined in Theorem 5.11.

Lemma F.10 *Given the proposed heavy-tailed setting of* (4.2) *is initialized at* $\boldsymbol{\theta}_0$ *with step-size* $\eta : (\exp(M_J\eta)\eta(B+M_J) \leq \xi/3)$, $\xi > 0$, *there exist a set of positive constants* C_α, C_1, *and* C_2 *such that the following holds*

$$\begin{aligned}
&\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t - \boldsymbol{\theta}_{k\eta}\| \geq \xi\right) \\
&\leq \exp(M_J\eta)M_J\eta\frac{C_1\left(1 + C_2\left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right)K\eta\right)^\beta}{\xi/3} \\
&\quad + 1 - \left(1 - C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp(\alpha M_J\eta)\epsilon^\alpha\left(\frac{\xi}{3}\right)^{-\alpha}\right)^K.
\end{aligned} \tag{F.42}$$

Proof In order to upper bound the error between the discrete and continuous equivalents, we start by analyzing the difference in dynamics for $t \in [k\eta, (k+1)\eta]$ and $t = k\eta$. Using the continuous equivalent (F.6) and integrating the expression over 0 to t gives

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \int_0^t \nabla J(\boldsymbol{\theta}_t^\epsilon(u))du + \epsilon r L^\alpha(t) \tag{F.43}$$

For $t = k\eta$, ie., the beginning of the interval considered, we have

$$\boldsymbol{\theta}_{k\eta} = \boldsymbol{\theta}_0 + \int_0^{k\eta} \nabla J(\boldsymbol{\theta}_t^\epsilon(u))du + \epsilon r L^\alpha(k\eta) \tag{F.44}$$

Using the above two expressions, $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\|$ takes the form

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \leq \int_{k\eta}^t \|\nabla J(\boldsymbol{\theta}_u)\|du + \epsilon\|r\|\|L^\alpha(t) - L^\alpha(k\eta)\| \tag{F.45}$$

Here $\boldsymbol{\theta}_u$ denotes $\boldsymbol{\theta}_t^\epsilon(u)$, r is the unit vector defining the direction of perturbation. We are interested in the dynamics when $t \in [k\eta, (k+1)\eta)$. Adding and subtracting $\nabla J(\boldsymbol{\theta}_{k\eta})$ inside the integral of the above expression yields

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \leq \int_{k\eta}^t \|(\nabla J(\boldsymbol{\theta}_u) - \nabla J(\boldsymbol{\theta}_{k\eta})\|du + \eta\|\nabla J(\boldsymbol{\theta}_{k\eta})\| + \epsilon\|L^\alpha(t) - L^\alpha(k\eta)\| \tag{F.46}$$

Next, Apply Lemma 5.2 and the Hölder continuity of the policy gradients inside the integral to write:

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \leq \int_{k\eta}^{t} M_J \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\|^{\beta} du + \eta \|\nabla J(\boldsymbol{\theta}_{k\eta})\| + \epsilon \|L^{\alpha}(t) - L^{\alpha}(k\eta)\| \tag{F.47}$$

By employing Hölder continuity and Cauchy Schwartz inequality for the second term, we obtain

$$\|\nabla J(\boldsymbol{\theta}_{k\eta})\| - \|\nabla J(\boldsymbol{\theta}_0)\| \leq \|\nabla J(\boldsymbol{\theta}_{k\eta}) - \nabla J(\boldsymbol{\theta}_0)\| \leq M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} \tag{F.48}$$

Using $\|\nabla J(0)\| \leq B$ in the above expression, we get

$$\|\nabla J(\boldsymbol{\theta}_{k\eta})\| \leq M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B \tag{F.49}$$

Substituting (F.49) for the second term of (F.47) results in

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \leq \int_{k\eta}^{t} M_J \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\|^{\beta} du + \eta \left(M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B \right) + \epsilon \|L^{\alpha}(t) - L^{\alpha}(k\eta)\| \tag{F.50}$$

For $\beta < 1$, $\|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\|^{\beta} < \|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\| + 1$. This fact allows us to rewrite the above expression as

$$\begin{aligned}
\|\theta_t - \boldsymbol{\theta}_{k\eta}\| &\leq \int_{k\eta}^{t} M_J (\|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\|) du + \eta \left(M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B + M_J \right) + \epsilon \|L^{\alpha}(t) - L^{\alpha}(k\eta)\| \\
&\leq \int_{k\eta}^{t} M_J (\|\boldsymbol{\theta}_u - \boldsymbol{\theta}_{k\eta}\|) du + \eta \left(M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B + M_J \right) + \epsilon \sup_{t \in [k\eta, (k+1)\eta]} \|L^{\alpha}(t) - L^{\alpha}(k\eta)\|.
\end{aligned}$$

Now this is in the exact form of Gronwall's inequality of Definition F.5 with $\phi(\cdot)$ being $\|\theta_t - \boldsymbol{\theta}_{k\eta}\|$ and $B := \eta \left(M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B + M_J \right) + \epsilon \sup_{t \in [k\eta, (k+1)\eta]} \|L^{\alpha}(t) - L^{\alpha}(k\eta)\|$ and $C(\cdot) := M_J$, therefore direct application of the inequality yields

$$\sup_{t \in [k\eta, (k+1)\eta]} \|\theta_t - \boldsymbol{\theta}_{k\eta}\| \leq \exp(M_J \eta) \left(\left(\eta \left(M_J \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B + M_J \right) \right) + \epsilon \sup_{t \in [k\eta, (k+1)\eta]} \|L^{\alpha}(t) - L^{\alpha}(k\eta)\| \right). \tag{F.51}$$

Evaluating the maximum of the above expression in the interval $k \in [0,\, K-1]$

$$\begin{aligned}
\max_{0 \leq k \leq K-1} \sup_{t \in [k\eta, (k+1)\eta]} \|\theta_t - \boldsymbol{\theta}_{k\eta}\| \leq \exp(M_J \eta) \Bigg(\Bigg(\eta \left(M_J \max_{0 \leq k \leq K-1} \|\boldsymbol{\theta}_{k\eta}\|^{\beta} + B + M_J \right) \Bigg) \\
+ \epsilon \max_{0 \leq k \leq K-1} \sup_{t \in [k\eta, (k+1)\eta]} \|L^{\alpha}(t) - L^{\alpha}(k\eta)\| \Bigg).
\end{aligned} \tag{F.52}$$

As we are interested in evaluating the probability of $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \in B^C$ (cf. (G.8)), consider the cases where error between continuous and discrete process exceeds $\|\xi\|$, ie., $\max_{0 \leq k \leq K-1} \sup_{t \in [k\eta, (k+1)\eta]} \|\theta_t - \boldsymbol{\theta}_{k\eta}\| \geq \xi$,

$$\mathcal{P}^{\boldsymbol{\theta}_0} \left(\max_{0 \leq k \leq K-1} \sup_{t \in [k\eta, (k+1)\eta]} \|\theta_t - \boldsymbol{\theta}_{k\eta}\| \geq \xi \right). \tag{F.53}$$

Now with the assumption that each term on the right hand side of (F.52) contributes equally, we have

$$\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t-\boldsymbol{\theta}_{k\eta}\|\leq\left(\exp\left(M_J\eta\right)\eta M_J\max_{0\leq k\leq K-1}\|\boldsymbol{\theta}_{k\eta}\|^\beta\geq\frac{\xi}{3}\right)+\left(\exp\left(M_J\eta\right)\eta(B+M_J)\geq\frac{\xi}{3}\right)+\left(\exp\left(M_J\eta\right)\left(\epsilon\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|L^\alpha(t)-L^\alpha(k\eta)\|\right)\geq\frac{\xi}{3}\right). \tag{F.54}$$

Next we evaluate the probability that the above expression holds given the process is initialized at $\boldsymbol{\theta}_0$

$$\begin{aligned}&\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t-\boldsymbol{\theta}_{k\eta}\|\geq\xi\right)\\&\leq\mathcal{P}^{\boldsymbol{\theta}_0}\left(\exp\left(M_J\eta\right)\eta M_J\max_{0\leq k\leq K-1}\|\boldsymbol{\theta}_{k\eta}\|^\beta\geq\xi/3\right)+\mathcal{P}^{\boldsymbol{\theta}_0}\left(\exp\left(M_J\eta\right)\eta(B+M_J)\geq\xi/3\right)\\&\quad+\epsilon\exp\left(M_J\eta\right)\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|L^\alpha(t)-L^\alpha(k\eta)\|\geq\xi/3\right).\end{aligned} \tag{F.55}$$

From Markov's inequality, we can write

$$\mathcal{P}(X\geq u)\leq\mathbb{E}\left[X\right]/u. \tag{F.56}$$

Using Markov's inequality for first term on the right hand side of (F.55) gives

$$\begin{aligned}&\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t-\boldsymbol{\theta}_{k\eta}\|\geq\xi\right)\leq\exp\left(M_J\eta\right)M_J\eta\frac{\mathbb{E}\left[\max_{0\leq k\leq K-1}\|\boldsymbol{\theta}_{k\eta}\|^\beta\right]}{\xi/3}\\&+\mathcal{P}^{\boldsymbol{\theta}_0}\left(\exp\left(M_J\eta\right)\eta(B+M_J)\geq\xi/3\right)+\\&\underbrace{\mathcal{P}^{\boldsymbol{\theta}_0}\left(\exp\left(M_J\eta\right)\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|L^\alpha(t)-L^\alpha(k\eta)\|\geq\epsilon^{-1}\xi/3\right)}_{T_6}.\end{aligned} \tag{F.57}$$

Using the choice of η such that $(\exp\left(M_J\eta\right)\eta(B+M_J))$ is always less than $\xi/3$, second term of the right hand side inequality equals to zero. Using properties of Lévy process from Lemma F.6 (Lemma 3, Thanh et al. (2019)), the last term of the above inequality simplifies to

$$T_6\leq1-\left(1-C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M_J\eta\right)\epsilon^\alpha\left(\frac{\xi}{3}\right)^{-\alpha}\right)^K. \tag{F.58}$$

Using the above expression in (F.57)

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\leq k\leq K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t-\boldsymbol{\theta}_{k\eta}\|\geq\xi\right)\leq\exp\left(M_J\eta\right)M_J\eta\frac{\mathbb{E}\left[\max_{0\leq k\leq K-1}\|\boldsymbol{\theta}_{k\eta}\|^\beta\right]}{\xi/3}$$

$$+1-\left(1-C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M_J\eta\right)\epsilon^\alpha\left(\frac{\xi}{3}\right)^{-\alpha}\right)^K. \tag{F.59}$$

Now the expression for $\mathbb{E}\left(\max_{0\le k\le K-1}\|\boldsymbol{\theta}_{k\eta}\|\right)^\beta)$ can be obtained by the direct substitution of inequality from Lemma F.9. Therefore, we get

$$\begin{aligned}
&\mathcal{P}^{\boldsymbol{\theta}_0}\left(\max_{0\le k\le K-1}\sup_{t\in[k\eta,(k+1)\eta]}\|\theta_t-\boldsymbol{\theta}_{k\eta}\|\ge\xi\right)\\
&\le\exp\left(M_J\eta\right)M_J\eta\frac{C_1\left(1+C_2\left(\frac{U_R}{(1-\gamma)^2}(m+b)/2+C\frac{d}{k_1^{1/\alpha}}\right)K\eta\right)^\beta}{\xi/3}\\
&\quad+1-\left(1-C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M_J\eta\right)\epsilon^\alpha\left(\frac{\xi}{3}\right)^{-\alpha}\right)^K,
\end{aligned} \tag{F.60}$$

where $\beta\in(0,1)$, U_R is the upper bound on the reward function from Assumption 4.1, constants C_1 and C_2 depends on β, dissipativity constants (m, b), and tail index α, $k_1:=1/\eta^{\alpha-1}$. ■

Appendix G. Proof of Theorem 5.11: Exit time Analysis for Heavy-tailed Policy Search

Next, we present results on the first exit time for the proposed heavy-tailed policy gradient setting.

Proof

We start along the lines of Simsekli et al. (2019); Thanh et al. (2019); Tzen et al. (2018) and relate the proposed setting of (5.6) to the continuous SDE of (5.7). We define the set of K points, $k=1\ldots,K$ obtained from (5.6) which are at a maximum distance of a from the local maxima of interest, $\bar{\boldsymbol{\theta}}_i$ is defined as

$$A\triangleq\{(\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^k):\max_{k\le K}\|\boldsymbol{\theta}^k-\bar{\boldsymbol{\theta}}_i\|\le a\} \tag{G.1}$$

Next we define a set N_a as the neighborhood in Euclidean distance centered at $\bar{\boldsymbol{\theta}}_i$:

$$N_a\triangleq\{\boldsymbol{\theta}\in\mathbb{R}^d:\|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}_i\|\le a\}. \tag{G.2}$$

As we initialized both discrete and continuous process at $\boldsymbol{\theta}_0$, processes defined by (5.6) and (5.7) are considered close enough if both of their exit times from N_a are close. For the moment, we assume step-size $\eta_k=\eta$ is constant, and consider the linearly interpolated version of (5.6) given by

$$d\hat{\boldsymbol{\theta}}_t=b(\hat{\boldsymbol{\theta}}_t)dt+\epsilon d\mathbf{L}_\alpha^t, \tag{G.3}$$

Note that the unit vector direction of perturbation is absorbed into $d\mathbf{L}_{\alpha}^{t}$ (cf. (F.6)). Here, $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\theta}}_t\}_{t\geq 0}$ denotes the whole process with drift term, b defined as

$$b(\hat{\boldsymbol{\theta}}_t) \triangleq \sum_{k=0}^{\infty} \nabla J(\hat{\boldsymbol{\theta}}_{k\eta}) \mathbb{I}_{[k\eta,(k+1)\eta)}(t). \tag{G.4}$$

where $\mathbb{I}$ denotes the indicator function such that $\mathbb{I} = 1$ if $t \in [k\eta,\, (k+1)\eta)$. The above expression for the drift (deterministic) component of the continuous-time process can also be expressed as

$$b(\hat{\boldsymbol{\theta}}_t) \triangleq \frac{1}{1-\gamma} \sum_{k=0}^{\infty} \mathbb{E}_{s,a\sim\rho_{\hat{\boldsymbol{\theta}}_{k\eta}}} [\nabla \log \pi_{\hat{\boldsymbol{\theta}}_{k\eta}} Q_{\pi_{\hat{\boldsymbol{\theta}}_{k\eta}}}(s,a)] \mathbb{I}_{[k\eta,(k+1)\eta)}(t).$$

First exit time of a continuous SDE (5.7) is summarized in Theorem F.3 Imkeller et al. (2010). To invoke this result, we first more rigorously establish the connection between the exit times of the discrete [cf. (5.6)] and continuous-time [cf. (5.7)] processes. Then, we may invoke this result to obtain the desired statement associated with the heavy-tailed policy search scheme in discrete time given in (5.6).

For the time being, we assume the distance (divergence) between the underlying distributions of processes (5.6) and (5.7) sampled at discrete instants, η, $k\eta$, ..., $K\eta$ is bounded by δ. Since (G.3) is the interpolated version of (5.6), this sampling specification also implies that there exists an optimal transport plan or optimal coupling between $\{\boldsymbol{\theta}_s\}_{s\in[0,\ldots,K]}$ and $\{\hat{\boldsymbol{\theta}}_s\}_{s\in[0,\ldots,K]}$. Therefore, using optimal coupling argument, data processing inequality for the relative entropy, and Pinsker's inequality, there exists a coupling between the random variables of both the processes such that coupling M between $\{\boldsymbol{\theta}_s\}_{s\in[0,K\eta]}$ and $\{\hat{\boldsymbol{\theta}}_s\}_{s\in[0,K\eta]}$ satisfies Tzen et al. (2018)

$$M(\{\boldsymbol{\theta}_s\}_{s\in[0,K\eta]} \neq \{\hat{\boldsymbol{\theta}}_s\}_{s\in[0,K\eta]}) \leq \delta. \tag{G.5}$$

Now, probability that the interpolated process and the continuous-time process are not equivalent is upper bounded by δ, i.e.,

$$\mathcal{P}\left(\{\boldsymbol{\theta}_s\}_{s\in[0,K\eta]} \neq \{\hat{\boldsymbol{\theta}}_s\}_{s\in[0,K\eta]}\right) \leq \delta \tag{G.6}$$

Typically, this upper-bound depends on the choice of algorithm step-size, η (See Assumption 6, Simsekli et al. (2019)), and depends on the KL-divergence between the underlying distributions.

Next we relate the expressions for the exit time of continuous and discrete-time processes, (5.6) and (5.7), respectively. Let $\{\hat{\boldsymbol{\theta}}_s \in A, \text{for}\, k = 1, \ldots,\, K\}$. As $\hat{\boldsymbol{\theta}}_s$ is the linearly interpolated version of the discrete process defined by (5.7) we get $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K \in A$. The statement $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K \in A$ implies exit time $\bar{\tau}$ of the discrete process from a domain characterized by a is greater than K, $\mathcal{P}(\bar{\tau}_{0,a}(\epsilon) > K)$. Therefore, using (G.6), the following events can happen with a probability dependent on δ if we have $\bar{\tau}_{0,a}(\epsilon) > K$

$$\boldsymbol{\theta}_t \begin{cases} \in A & \text{for } t = \eta, \ldots, K\eta \\ \notin A: & \text{with probability } < \delta \end{cases}$$

This fact allows us to write the probability of $\bar{\tau}_{0,a}(\epsilon) > K$ for some constant K as follows:

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\bar{\tau}_{0,a}(\epsilon) > K) \leq \mathcal{P}^{\boldsymbol{\theta}_0}\left((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A\right)\right) + \delta \tag{G.7}$$

Note that $((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A)$ implies maximum distance from local minimum is a.

Now we have an expression connecting the probability for continuous time processes, $\boldsymbol{\theta}_t$ at $t = k\eta$, $k = 1, \ldots, K$ being inside A to the exit time of discrete equivalents given both of them are initialized with same value. However, the identity of continuous random variables $\boldsymbol{\theta}_t$ between $[k\eta, (k+1)\eta]$ are still unknown. To understand their behavior within this range, we study their behavior to sets N_a [cf. (G.2)]. Once we augment that into (G.7), we are in the position to derive for exit time results for discrete process in terms of its continuous equivalent. These steps are formalized next.

Upper bound on $\mathcal{P}^{\boldsymbol{\theta}_0}\left((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A\right)$: In order to analyze $\boldsymbol{\theta}_t$ for $t \in [k\eta, (k+1)\eta]$ and to link it with N_a, we follow the approach used in Thanh et al. (2019) and impose an upper bound on $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\|$, $\forall t \in [k\eta, (k+1)\eta]$. Let us assume $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\|$ is bounded by ξ for the time between $k\eta$ and $(k+1)\eta$. We define a set B such that

$$B \triangleq \{\max_{0 \leq k \leq K-1} \sup_{t \in [k\eta, (k+1)\eta]} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\| \leq \xi\} \tag{G.8}$$

For a clear understanding of the process, we illustrate the previously defined sets in Fig. 6

Let A defines a hyper sphere with K points $\boldsymbol{\theta}^k \in \mathbb{R}^d$ from discrete process (5.6). The radius of A is defined by a and $\bar{\boldsymbol{\theta}}_i$ be its center. Now we have another hyper sphere (shown in gray), B of radius $a + \xi$ such that B defines the event $\boldsymbol{\theta}_t$, $t \in [k\eta, (k+1)\eta]$ such that the maximum error between $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^k\| \leq \xi$. Now, the event $[(\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A] \cap B)$ ensures that $\boldsymbol{\theta}_t$ is close to N_a for $t = \eta, \ldots, K\eta$. Now let us relate $\mathcal{P}^{\boldsymbol{\theta}_0}\left((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A\right)$ from (G.6) to exit time of the continuous process.

For $(\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A$, we can have two possibilities: $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{k\eta}\|$, $t \in [k\eta, (k+1)\eta]$ can be either inside or outside B. Therefore,

$$\mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A) \leq \mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A \cap B) + \mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in B^c) \tag{G.9}$$

Note that $\mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A \cap B) \implies \boldsymbol{\theta}_t \in N_a$, implies exit time of the corresponding continuous process is greater than $K\eta$. Therefore, first term on the right hand side of the above expression is equivalent to probability that the exit time of the continuous process from a domain characterized by $a + \xi$ is greater than $K\eta$. And we get,

$$\mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in A) \leq \mathcal{P}^{\boldsymbol{\theta}_0}(\tau_{\xi,a}(\epsilon) \geq K\eta) + \mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in B^c) \tag{G.10}$$

Using the above expression in (G.7)

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\bar{\tau}_{0,a}(\epsilon) > K\right) \leq \mathcal{P}^{\boldsymbol{\theta}_0}(\tau_{\xi,a}(\epsilon) \geq K\eta) + \mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in B^c) + \delta \tag{G.11}$$

Now the second term on the right hand side of the expression defines the probability that $\{\boldsymbol{\theta}_{k\eta}\}_{k=1,\ldots,K}$ are not confined in the set B, $\mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in B^c)$. Using upper bound on $\mathcal{P}^{\boldsymbol{\theta}_0}((\boldsymbol{\theta}_\eta, \ldots, \boldsymbol{\theta}_{K\eta}) \in B^c)$ from Lemma F.10 in the above inequality, we get

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\bar{\tau}_{0,a}(\epsilon) > K\right) \leq \mathcal{P}(\hat{\tau}_{\xi,a}(\epsilon) \geq K\eta)$$

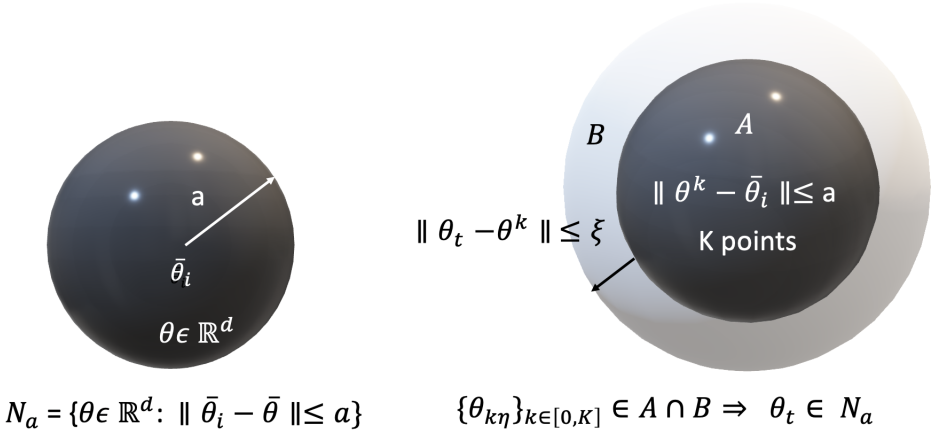


Figure 6: Sketch of intuition for the analysis of the exit time behavior of the discrete time process. We study the behavior of discrete and continuous process initialized at same point. On the left we depict N_a, the neighborhood of the local extrema for a generic element $\boldsymbol{\theta} \in \mathbb{R}^d$ with respect to which we study the exit time behavior. In order to do that we first analyze the behavior of the continuous-time process at discrete time indices which is piecewise constant within intervals $t = \eta, \cdots, K\eta$ within a neighborhood of local maximum $\bar{\boldsymbol{\theta}}_i$ defined in Assumption 5.10(3.). This interpolation process bridges the gap between the continuous-time process and the discrete time process. On the right we depict an equivalent neighborhood for the discrete-time process. We show that the probabilistic behavior of the continuous and discrete neighborhoods are comparable up to constant factors that depend on algorithm step-size and interpolation parameters (cf. (G.7)). Characterizing this probabilistic behavior does not bound the behavior of the continuous process between the instants $t \in [k\eta, (k+1)\eta]$. To do so, we introduce event B and study the behavior of continuous process between $[k\eta, (k+1)\eta]$ (Lemma F.10). Now relating events $A \cap B$ and B^c, we obtain an upper bound on the exit time of discrete process.

$$+ \exp(M_J \eta) M_J \eta \frac{C_1 \left(1 + C_2 \left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right) K\eta\right)^{\beta}}{\xi/3}$$
$$+ 1 - \left(1 - C_\alpha d^{1+\frac{\alpha}{2}} \eta \exp(\alpha M \eta) \epsilon^\alpha \left(\frac{\xi}{3}\right)^{-\alpha}\right)^K + \delta. \tag{G.12}$$

From here, we invoke the results for exit time in multi-dimensional space Imkeller et al. (2010) formalized in Theorem F.3. To do so, let $\boldsymbol{\theta}_0 \in \mathcal{G}_i$ ($\mathcal{G}_i$ be a domain containing the i-th local minima $\bar{\boldsymbol{\theta}}_i$ such that $\mathcal{G}_i \subset \mathbb{R}^d$) and the process escapes to the delta-tube, $\Omega_i^+(\bar{\delta})$ (cf. (5.11)) from $\mathcal{G}_i$ using jumps initiated by Lévy process of tail index, α (cf. (5.11)).

Therefore, the first term of the inequality (G.12) corresponds to the time at which $\boldsymbol{\theta}$ exits $\mathcal{G}_i$, i.e., the time at which $\boldsymbol{\theta} \in \Omega_i^+(\bar{\delta})$. Note that the probability defined in the above expression is in terms of the time at which the process exits a given domain, i.e. the probability that the exit time is greater than $K\eta$. Using Assumption 5.10, there exists a K such that $\hat{\tau}_{\xi,a}$ greater than $K\eta$, $K > 0$ and $\boldsymbol{\theta}_{\hat{\tau}}^{\epsilon} \in \Omega_i^+$ for $\hat{\tau}_{\xi,a}(\epsilon) \geq K\eta$ (meaning, probability

that the exit time is greater than $K\eta$ is proportional to the probability of the process entering the desired delta tube), specifically, that $\mathcal{P}(\hat{\tau}_{\xi,a}(\epsilon) \geq K\eta) \propto \mathcal{P}^{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}^{\epsilon}_{\hat{\tau}} \in \Omega^{+}_{i}(\bar{\delta})\right)$.

Therefore, substituting $\mathcal{P}(\hat{\tau}_{\xi,a}(\epsilon) \geq K\eta)$ with $\mathcal{P}^{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}^{\epsilon}_{\hat{\tau}} \in \Omega^{+}_{i}(\bar{\delta})\right)$ in (G.12) allows us to write

$$\begin{aligned}\mathcal{P}^{\boldsymbol{\theta}_0}\left(\bar{\tau}_{0,a}(\epsilon) > K\right) \leq & \mathcal{P}^{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}^{\epsilon}_{\hat{\tau}} \in \Omega^{+}_{i}(\bar{\delta})\right) \\ & + \exp\left(M_J\eta\right) M_J\eta \frac{C_1\left(1 + C_2\left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right)K\eta\right)^{\beta}}{\xi/3} \\ & + 1 - \left(1 - C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M\eta\right)\epsilon^{\alpha}\left(\frac{\xi}{3}\right)^{-\alpha}\right)^{K} + \delta \qquad \text{(G.13)}\end{aligned}$$

where, $\mathcal{P}^{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}^{\epsilon}_{\hat{\tau}} \in \Omega^{+}_{i}\right)$ indicates the first exit time for a continuous SDE from $\mathcal{G}_i$. Note that $\Omega^{+}_{i}(\cdot)$ indicates delta-tube outside the boundary of $\mathcal{G}_i$ with perturbations along the positive basis vector. Now direct application of Theorem F.3 for the first term on the right-hand-side of the inequality, we get

$$\begin{aligned}\mathcal{P}^{\boldsymbol{\theta}_0}\left(\bar{\tau}_{0,a}(\epsilon) > K\right) \leq & \frac{2}{\epsilon^{\rho\alpha}}\epsilon^{\alpha}(d^{+})^{-\alpha} \\ & + \exp\left(M_J\eta\right) M_J\eta \frac{C_1\left(1 + C_2\left(\frac{U_R}{(1-\gamma)^2}(m+b)/2 + C\frac{d}{k_1^{1/\alpha}}\right)K\eta\right)^{\beta}}{\xi/3} \\ & + 1 - \left(1 - C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M_J\eta\right)\epsilon^{\alpha}\left(\frac{\xi}{3}\right)^{-\alpha}\right)^{K} + \delta \qquad \text{(G.14)} \\ \leq & \frac{2}{\epsilon^{\rho\alpha}}\epsilon^{\alpha}(d^{+})^{-\alpha} + \mathcal{O}\left(\frac{d}{k_1^{1/\alpha}}K\eta\right)^{\beta} \\ & + \mathcal{O}\left(1 - \left(1 - C_\alpha d^{1+\frac{\alpha}{2}}\eta\exp\left(\alpha M_J\eta\right)\epsilon^{\alpha}\left(\frac{\xi}{3}\right)^{-\alpha}\right)^{K} + \delta\right) \qquad \text{(G.15)}\end{aligned}$$

where, d^{+} denotes distance function to the boundary along the positive r unit vector (cf. (5.10)), ρ is a positive constant such that $\rho \in (0,1)$, $a + \xi > \epsilon^{1-\rho}$, d is the dimension of $\boldsymbol{\theta}$, Hölder continuity constant, $\beta \in (0,1)$, $\delta > 0$, $C_\alpha > 0$, $\xi > 0$, η is the step-size, U_R, γ are the parameters of proposed RL setting, positive constants, C, $C_1\,C_2$, are functions of dissipativity constants of score function. Note that first term of the above inequality denotes the probability that the stochastic process exits from a given domain $\mathcal{G}_i$ when perturbed by a jump process along unit vector $\mathbf{r}$ with tail index α. Observe that the exit time is only a function of distance between the point at which the process is initialized and the boundary of the domain around local minima (extrema), but does not depend on the function values (height of the local minima (extrema)). ■

Appendix H. Proof for Theorem 5.12: Transition time for the Proposed Heavy-tailed setting

In this section, we derive transition time results for the proposed heavy-tailed setting from one local maxima to another as defined in Theorem 5.11. Suppose the domain $\mathcal{G}_i$ satisfies the Assumption 5.10 defined in Section 5.2. Further there exists a unit vector, $+\mathbf{r}$ in the direction connecting the domains $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ and we define the distance function to the the intersection of boundaries $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$ by:

$$d_{ij}^{+}(\boldsymbol{\theta}) := \inf\{t > 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \partial\mathcal{G}_i \cap \partial\mathcal{G}_{i+1}\}, \tag{H.1}$$

$$d_{ij}^{-}(\boldsymbol{\theta}) := \sup\{t < 0 : g_{i_{\boldsymbol{\theta}}}(t) \in \partial\mathcal{G}_i \cap \partial\mathcal{G}_{i-1}\}. \tag{H.2}$$

Let $\boldsymbol{\theta}_0 \in \mathcal{G}_i$. As ϵ is a coefficient that multiplies the noise process in (5.7), it is clear that the error between both the processes are lower bounded by ϵ, ie. $\left(\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{k\eta}\| > \epsilon\right)$. Using Markov's inequality, the above statement can be expressed in the language of probability as

$$\mathcal{P}^{\boldsymbol{\theta}_0}\left(\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{k\eta}\| > \epsilon\right) \leq \frac{\mathbb{E}_{\boldsymbol{\theta}_0}\left[\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{k\eta}\|\right]}{\epsilon} \leq \frac{\delta}{2}. \tag{H.3}$$

Let us say the probability of above event given $\boldsymbol{\theta}^0 = \boldsymbol{\theta}_0$ is $\delta/2$. Given the continuous time instant $t = k\eta$, there exists a unit vector, $\mathbf{r}$ in the direction connecting the domains $\mathcal{G}_i$ and $\mathcal{G}_{i+1}$. and let $\boldsymbol{\theta}_{k\eta} \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$, $\forall \bar{\delta} \in (0, \bar{\delta}_0)$, the following events can happen with probabilities dependent on δ:

$$\boldsymbol{\theta}^k \begin{cases} \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1} & \text{with } \mathcal{P}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) \\ \notin \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1} : \|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{k\eta}\| > \epsilon & \text{with probability } < \delta/2 \\ \notin \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1} : \max\|\boldsymbol{\theta}^k - N_a\| < \epsilon & \text{with probability } < \delta/2 \end{cases}$$

The above case-by-case study illuminates that either $\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$ or $\boldsymbol{\theta}^k \notin \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$ (for which the probability adds up to δ). Note that $\Omega_i^+(\bar{\delta})$ denotes $\bar{\delta}$- tubes outside of $\mathcal{G}_i$ (cf. (5.11)). This fact allows us to write the probability that $\boldsymbol{\theta}_{k\eta} \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$ as follows:

$$\mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_{k\eta} \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) \leq \mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) + \delta \tag{H.4}$$

Evaluating the above inequality in the limit of $\epsilon \to 0$ yields

$$\lim_{\epsilon\to 0} \mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_{k\eta} \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) \leq \lim_{\epsilon\to 0} \mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) + \delta \tag{H.5}$$

Observe that the probability that $\boldsymbol{\theta}^k \notin \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$ is δ. Note that the value of δ can be obtained analyzing KL-divergence between underlying distirbutions of (5.6) and (G.3), the interpolated version of (5.6). The consequence of the aforementioned analysis is an upper bound on the step-size, η (See Assumption 6, Simsekli et al. (2019)). Now the expression on the left-hand side of the above inequality is the probability that the process reaches to the intersection of $\bar{\delta}$ tube and $\mathcal{G}_{i+1}$ in the limit $\epsilon \to 0$.

Using Theorem F.3, for every $\bar{\delta} \in (0, a + \xi)$ the probabilities to exit in direction $\pm r$ are given by

$$\lim_{\epsilon\to 0} \mathcal{P}\left(\Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}\right) = \frac{p_i^+}{p_s} \tag{H.6}$$

for all $\boldsymbol{\theta}_t \in \mathcal{G}_i$ and

$$p_s := ((d_{ij}^+)^{-\alpha} + (-d_{ij}^-)^{-\alpha}) \tag{H.7}$$

$$p_i^+ := (d_{ij}^+)^{-\alpha} \tag{H.8}$$

$$p_i^- := (-d_{ij}^-)^{-\alpha}, \tag{H.9}$$

where d_{ij}^+ and d_{ij}^- define distance from $\boldsymbol{\theta}_0$ to $\Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}$ along $\mathbf{r}$. Now using the above expression in (H.5) yields lower bound on the transition probability from $\mathcal{G}_i$ to $\mathcal{G}_{i+1}$, $\mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1})$ in the limit $\epsilon \to 0$

$$\lim_{\epsilon \to 0} \mathcal{P}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}^k \in \Omega_i^+(\bar{\delta}) \cap \partial\mathcal{G}_{i+1}) \geq \frac{d_{ij}^{-\alpha}}{((d_{ij}^+)^{-\alpha} + (-d_{ij}^-)^{-\alpha}} - \delta. \tag{H.10}$$

Appendix I. Supplemental Experiments

I.1 Additional Experimental Instantiations

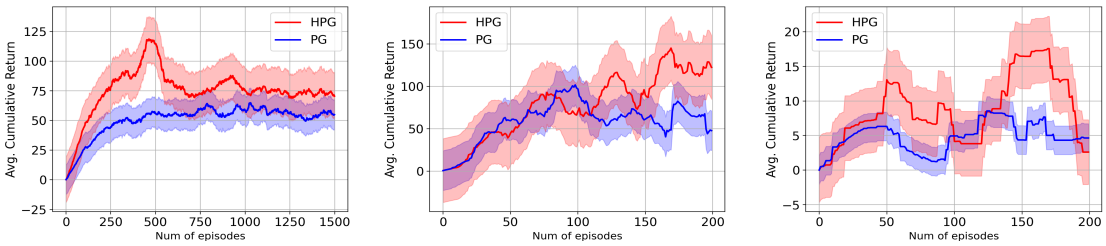
In this section, we evaluate the performance of the proposed algorithm on more complex environments from OpenAI gym and Roboschool, namely Inverted Pendulum from Roboschool and sparse versions of Roboschool REACHER-V2 & HALFCHEETAH-V2.

I.1.1 Inverted Pendulum

Next, we evaluate the performance of HPG on Inverted Pendulum-v1 from Roboschool in Fig. 7(a). The objective is to keep the pendulum balanced while keeping the cart upon which it is mounted away from the borders. Environment consists of a observation space of dimension 5 and an action space of dimension 1. Here we use a neural network with single hidden layer consisting of 64 neurons. All the other parameters are same as given in the previous experiments.

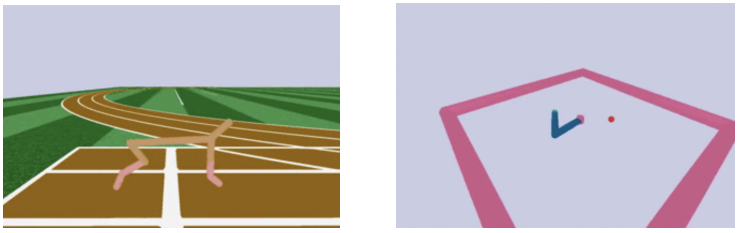
I.1.2 Sparse Variant of HalfCheetah

The environment (shown in Fig. 8(a)) consists of planar biped robot in a continuous environment with $\mathcal{S} \in \mathbb{R}^{26}$ and $\mathcal{A} \in \mathbb{R}^6$. The task runs on the Roboschool simulator. In the task, the agent is a two-dimensional cheetah controlled by seven actuators. The state representation corresponds to the position and velocity of the cheetah, current angle and angular velocity of its joints as well as the torque on the actuators. The reward function in this environment is a mixture between the amount of contact the limbs have on the floor and the displacement of the agent from the starting position. The reward structure is made sparse by removing the control penalty to each actions. In addition, a step reward of 2 is provided when the x-component of speed of the Cheetah is greater than 2 as in Matheron et al. (2019). We plot the initial average cumulative returns for sparse Chettah over latest 25 episodes in Fig. 7(b). Observe that HPG, by virtue of its extreme action selection according to a heavy-tailed distribution, causes more jumps in the state space, which results in better exploration of the environment. The result is faster policy improvement than vanilla PG.



(a) Inverted Pendulum (b) Sparse Half Chettah (c) Sparse Reacher

Figure 7: **(a)** Average cumulative returns for Inverted Pendulum from Roboschool over latest 100 episodes. HPG shows better ability to explore over light tailed policies. **(b)** We plot the initial average cumulative returns for sparse Chettah over latest 25 episodes. Observe that HPG better explores the environment and shows improved performance over PG. **(c)** We plot the initial average cumulative returns for sparse Reacher over latest 25 episodes for HPG which shows better ability to explore over PG. For clarity of only distinguishing the performance differences associated with policy parameterization, we omit the additional comparators of Figure 4.



(a) HalfCheetah environment. (b) Reacher environment.

Figure 8: Additional environments

I.1.3 Sparse Variant of Roboschool Reacher

The environment (shown in Fig. 8(b)) consists of an robot arm in a 2D space which tries to reach a target. Reacher is a continuous environment with $\mathcal{S} \in \mathbb{R}^9$ and $\mathcal{A} \in \mathbb{R}^2$. A sparse variant of REACHER-V2 is created by imposing a step reward of 1 when the distance between the arm and the target is less than 0.06, and 0 otherwise Matheron et al. (2019). The target is fixed to a position of $[0.19, 0]$ and removed distance to the target from the observations. In addition, the reward structure is sparsified by removing the control penalty. We implement the proposed algorithm using policies with $\alpha = 1, 2$. Here we choose the mode of the policy: $\mathcal{S} \to \mathcal{A}$ to be a neural network that has two hidden layers. Each hidden layer consists of 10 neurons with tanh being the activation function. Other parameters of the Algorithm are same as the previous experiments. Fig. 7(c) shows that HPG, by virtue of its extreme action selection according to a heavy-tailed distribution, causes more jumps in the state space, which results in better exploration of the environment. The result is faster policy improvement than vanilla PG.