Proximal Policy Optimization with Unbounded Score Functions for Persistent Exploration in Reinforcement Learning

Anjaly Parayil[†*], Amrit Singh Bedi[†*], Mengdi Wang[††], and Alec Koppel[†]

Abstract— **Reinforcement learning has gained attention in recent years for its ability to solve complex control tasks with costs revealed sequentially across time without a system dynamics model. We focus on the class of policy gradient methods, where one iterates in stochastic gradient ascent steps with respect to a parameterized family of policies. Central to these approaches is the Policy Gradient Theorem, which states that the gradient of the value function with respect to the policy is a product of factors: the score function and the Q function. Policy gradient method operates by performing a Monte Carlo rollout to estimate the Q function, and then evaluates the score function at the end of the trajectory, the product of which is used for stochastic ascent. Predominately in the literature, one assumes the score function is bounded in order to establish convergence, which restricts the policy parameterization as a Gaussian or Boltzmann (softmax) with bounded variance. In this work, we establish the convergence to stationarity of policy gradient method without this restriction and further establish the convergence of a projected (proximal) variant. In doing so, we permit policy parameterizations whose variance may be unbounded, which enables one to consider a class of heavy-tailed and adaptive-variance policies for reinforcement learning. We observe the improved performance in practice of these schemes, especially when myopic and farsighted decision-making are misaligned.**

I. Introduction

In reinforcement learning (RL), an autonomous agent sequentially interacts with its environment and observes rewards incrementally across time [1], and has gained attention in recent years for its successes in continuous control [2], [3], web services [4], personalized medicine [5], among other contexts. This framework, which may be mathematically defined by a Markov Decision Process (MDP) [6], is one 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 with classical optimal control is its ability to discern control policies without a system dynamics model.

Algorithms for RL may be categorized as those which operate by approximately solving Bellman's equations [7], [8] and policy gradient methods [9]. While the former may be lower variance and converge faster [10], [11], typically they require representing a Q-function for every state-action pair, which is intractable for continuous spaces, the focus of this work. For this reason, we focus on policy gradient method.

The foundation of policy search is the Policy Gradient Theorem [12], 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. Despite the maturity of policy search, several open questions regarding its limiting and finite-time behavior have only come into focus recently. This is because classically its behavior was only studied from the perspective of asymptotic stability [13], [14] (see also [15]), using tools from dynamic systems [16], [17].

More recently, the non-asymptotic performance of policy search has come to the fore. In continuous space, its finite-time performance has been linked to stochastic gradient iteration for non-convex objectives, and hence its $\mathcal{O}(1/\sqrt{k})$ rate of convergence to stationarity has been established [18], [19]. Stronger results have appeared for finite MDPs as well [20]–[22]: linear convergence to *global* optimality for softmax parameterizations. Enhancements that incorporate proximal regularization have also gained traction recently [23]–[25].

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 optimal policy may be shown to assign strictly positive likelihood to the optimal action over the entire state space [26][Lemma 9]. Under this condition, then, a version of gradient dominance [27] (known also as Polyak-Łojasiewicz inequality [28], [29]) holds [26], [30], [31], which interestingly echoes the classical notion of persistence of excitation required for accurate systems identification [32], [33].

Unfortunately, translating this condition to continuous space, the goal of this work, is somewhat elusive. That is because many common distributions in continuous space may fail to be integrable if their likelihood is lower bounded away from null over the entire state space. As a step towards satisfying this condition, we propose to study policy parameterizations defined by possibly heavy-tailed distributions [34], [35], known as Lévy Processes, which appear in fractal geometry [36], [37], finance [38], [39], pattern formation in nature [40], and networked systems [41].

Their use in non-convex optimization as a way to perturb stochastic gradient updates by α-stable Lévy noise [42], [43], inspired by earlier stochastic gradient Langevin dynamics where one instead perturbs updates using Gaussian noise [44], [45], has notably been shown to improve generalization as quantified by the tail index of the parameter estimate's limiting distribution [43], [46]. Rather than perturb

*Equal contributions.

[†]U.S. Army Research Laboratory, Adelphi, MD 20783, USA. E-mails: panjaly05@gmail.com, alec.e.koppel.civ@mail.mil, amrit0714@gmail.com

[††]Dept. of Electrical Engineering, Princeton University {mengdiw}@princeton.edu

stochastic gradient updates, we seek to directly parameterize policies using α-stable Lévy processes, motivated by the aforementioned persistent exploration conditions in finite MDPs. Doing so unfortunately invalidates the boundedness condition of the score function that is standard in the analysis of policy gradient methods in continuous spaces to date [18], [19].

Therefore, in this work, we study policy search for the setting that the score function is allowed to be unbounded but whose stochastic variance is determined by the magnitude of the population gradient (Sec. IV), as in [47]. We additionally note that the unboundedness of the score function can cause instability in the policy parameter estimates [48], which we address by introducing a proximal variant of the update [23]–[25]. Our main theoretical result is the establishment of convergence to stationarity (Theorem 1) of a variant of stochastic mirror ascent for policy optimization called Stochastic Recursive Mirror Ascent (SRMA) (Algorithm 1), which incorporates an additional recursive averaging step in its inner loop. This closes a conspicuous gap in the literature for proximal methods applied to non-convex expected value objectives *without regularization* – see [25] for a thorough study of the regularized case. Moreover, we put forth a variant of the proposed algorithm based upon gradient clipping in [23] as Algorithm 2. Experimentally, we observe that policies associated with heavy-tailed distributions obtained with these methods more effectively addresses RL problems where myopic and farsighted behavior are at odds (Sec. V).

II. Markov Decision Problems

In reinforcement learning (RL), an autonomous agent traversing through a state space $\mathcal{S}$ at states s, selects actions $a \in \mathcal{A}$ and transitions to another state s' according to a Markov transition density $\mathbb{P}(s'|s,a)$. Upon reaching state s', the environment reveals an instantaneous reward $r(s,a)$ which informs the merit of a given decision a starting from state s. Mathematically, this framework for interactive decision-making may be defined as a Markov Decision Process (MDP), whose components are ($\mathcal{S}$, $\mathcal{A}$, $\mathbb{P}$, r, γ). The state $\mathcal{S}$ and action space $\mathcal{A}$ may either be finite or compact real vector space such that $\mathcal{S} \subseteq \mathbb{R}^q$ and $\mathcal{A} \subseteq \mathbb{R}^p$. Moreover, γ is a discount factor that determines how much future rewards are worth relative to the next step. As is well known in MDPs [6], [49], it suffices to hypothesize the decision-maker selects actions $a_t \sim \pi(\cdot|s)$ over a time-invariant distribution $\pi(a|s) := \Pr\{a_t = a | s_t = s\}$ called a policy, which denotes the probability of action a given the agent is in state s. The goal in RL is to determine the policy that accumulates the most long-term reward on average, i.e., the value:

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_t = \pi(s_t)\right] \quad (1)$$

where s_0 denotes the initial point along a trajectory $\{s_u, a_u, r_u\}_{u=0}^{\infty}$ with short-hand notation $r_t = r(s_t, a_t)$. Here, the expectation in (1) is with respect to randomized policy $a_t \sim \pi(\cdot|s_t)$ and state transition dynamics $s_{t+1} \sim \mathbb{P}(.|s_t, a_t)$. For further reference, we further define the action-value, i.e., Q-function $Q^\pi(s,a)$ as the value conditioned on an initially selected action:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, a_t = \pi(s_t)\right]. \quad (2)$$

Our focus is on policy search over parameterized families of policies, which hypothesizes that actions are selected according to a policy $\pi_{\boldsymbol{\theta}}(\cdot|s_t)$ parameterized by vector $\boldsymbol{\theta} \in \mathbb{R}^d$. Then, we seek to estimate those parameters that maximize the cumulative return [1]:

$$\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) := V^{\pi_{\boldsymbol{\theta}}}(s_0) \quad (3)$$

where, objective is given by $J(\boldsymbol{\theta}) := V_{\pi_{\boldsymbol{\theta}}}(s_0)$. Observe that (3) is non-convex in $\boldsymbol{\theta}$, and therefore, finding the optimal policy is challenging even in the deterministic setting. However, in RL, the search procedure necessarily interacts with the transition dynamics $\mathbb{P}(s'|s,a)$ as well. Before detailing how one may implement first-order stochastic search to solve (3), we introduce several representative policy parameterizations.

A. Example Policy Parameterizations

Example 1. For continuous spaces, one of the most common policy parameterizations is the Gaussian policy

$$\pi_{\boldsymbol{\theta}}(a|s) = \mathcal{N}(a|s^\top \boldsymbol{\theta}, \sigma^2), \quad (4)$$

where the parameters $\boldsymbol{\theta}$ determine the mean (centering) of a Gaussian distribution at $s^\top \boldsymbol{\theta}$, and σ^2 is a fixed-variance hyper-parameter. One may also replace s with some feature map $\varphi(s)$ in the aforementioned inner-product to scale to higher-dimensional spaces, i.e., $\varphi : \mathcal{S} \to \mathbb{R}^d$ with $d \ll q$.

Example 2. One drawback of treating the variance (bandwidth) of the policy constant in the preceding expression is that it has a tendency to only select actions a that are near the mean $s^\top \boldsymbol{\theta}_1$. To ameliorate this issue, the variance may also be considered as a parameter of the policy:

$$\pi_{\boldsymbol{\theta}}(a|s) = \mathcal{N}(a|s^\top \boldsymbol{\theta}_1, e^{2\theta_2}), \quad (5)$$

where now the augmented parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1; \theta_2]$ determines the mean (centering) of a Gaussian distribution at $s^\top \boldsymbol{\theta}_1$ as well as the variance $\sigma^2 = e^{2\theta_2}$ [50]. Here exponentiation imposes a non-negativity domain constraint on the variance.

We next introduce a family of heavy-tailed policies motivated by power laws that arise in fractal geometry [36], finance [35], and network science [51]. Specifically, we define the family of Lévy processes called α-stable distributions.

Example 3. Symmetric α stable, $\mathcal{S}\alpha\mathcal{S}$ distributions are a generalization of a centered Gaussian distribution with $\alpha \in (0,2]$ as the tail index which determines the heaviness of the distribution's tail [52]. 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 $\sigma \in (0,\infty)$. Note that for $\alpha = 2$, it reduces to a Gaussian, and for $\alpha = 1$ we have a Cauchy distribution whose parametric form is:

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sigma\pi(1 + ((a - x_0)/\sigma)^2)}, \tag{6}$$

where, x_0 is the mode of the distribution and σ is the scaling parameter, both of which are functions of $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \theta_2]$: $x_0 = s^\top \boldsymbol{\theta}_1$, $\sigma = e^{\theta_2}$. For non-integer (fractional) value of α, the distribution does not exhibit a closed form expression, and is referred to as fractal [37]. In the financial literature, such distributions have been associated with the phenomenon of ”black swan” events [38], [39].

With potential choices of policy parameterization detailed, we take a closer look at their relative merits and drawbacks. Intuitively, policies that select actions far from a learned mean parameter over actions may be beneficial when the long-run accumulation of rewards $V^\pi(s)$ is not close to the one-step reward $r(s, a)$.

More formally, persistent exploration has been identified in *finite* (i.e., discrete) MDPs recently as a key driver for the ability to converge to the optimal policy using first-order methods [26], [30], [31]. This ability is related to the fact that under a persistent exploration condition, i.e., the initial distribution over s_0 in (1), the optimal policy may be shown to assign strictly positive likelihood to the optimal action over the entire state space [26][Lemma 9]. Under this condition, then, a version of gradient dominance (akin to strong convexity) holds (Lemma 8). Interestingly, these results echo the classical notion of persistence of excitation required for accurate systems identification [32], [33].

Unfortunately, translating this condition to continuous space, the setting of this work, is somewhat elusive. In particular, many common distributions in continuous space may fail to be integrable if their likelihood is lower bounded away from null over the entire state space. As a step towards satisfying this condition, in this work we study policy search under parameterizations defined by heavy-tailed distributions (Examples 2 - 3) for continuous spaces, whose likelihood approaches null as slowly as possible while still defining a valid distribution. To clarify the motivation for when near and long-term incentives may be misaligned, we introduce a representative example before continuing.

Representative Example. Consider an environment with a car trapped between two mountains of different heights in a discounted infinite-horizon setting as shown in Fig. 1. 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. We consider an incentive structure in which the amount of energy expenditure, i.e., the action squared, at each time-step is negatively penalized:

$$r(s_t, a_t) = \begin{cases} -a_t^2, & \text{for } -4.0 < s < 3.709,\ s \neq 2.667 \\ 500 - a_t^2, & \text{if } s = -4.0 \\ 10 - a_t^2, & \text{if } s = 2.667 \end{cases} \tag{7}$$

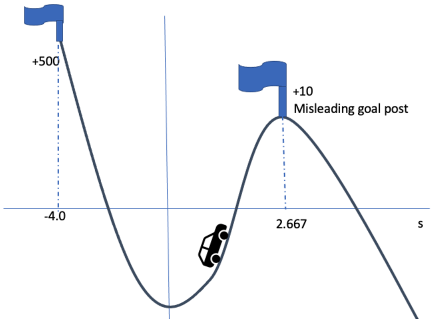


Fig. 1: Representative example for when long and short-term incentives may be misaligned in continuous space: a continuous Mountain Car-like environment with a low reward state and a bonanza atop a higher hill. Policies that do not incentivize exploration get stuck at the spurious goal.

Here $s \in [-4.0, 3.709]$ denotes the state space, and the action a_t is a one-dimensional scalar representing the speed of the vehicle $\dot{s}_t$. The environment is visualized in Fig. 1. This environment is one in which one may prioritize visiting the less rewarding state and never reach the jackpot without sufficient exploration. The potential pitfalls of this scenario is made precise experimentally in Section V.

With the motivation clarified, we shift to illuminating a technical understanding of heavy-tailed policy search. Specifically, heavy-tailed policy parameterizations, while encouraging action selection far from the mean, exhibits a downside. That is, it causes search directions to become possibly unbounded. Surmounting this issue is the focus of Section IV. In the next section, we recall policy gradient method for (3), especially in the context of Examples 2 - 3, which illuminates this boundedness issue.

III. Policy Gradient Methods

Policy gradient method is an algorithm for RL which operates by implementing approximate gradient ascent in parameter space $\mathbb{R}^d$ with respect to the value function (1). The key enabler of this method is the Policy Gradient Theorem [1], which expresses search directions in parameter space:

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \,|\, s) \cdot Q^{\pi_\theta}(s, a)\right]. \tag{8}$$

where $\rho_\theta(s, a) = \rho_{\pi_\theta}(s) \cdot \pi_\theta(a \,|\, s)$ is a probability distribution that denotes the *discounted state-action occupancy measure*, which is the product of the discounted state occupancy measure $\rho_{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}(s_k = s \,|\, s_0, \pi_\theta)$ and policy $\pi_\theta(a \,|\, s)$. In [12], both $\rho_{\pi_\theta}(s)$ and $\rho_{\boldsymbol{\theta}}(s, a)$ are established as valid probability distributions.

To compute policy search directions, then, one requires unbiased estimates of both factors in the product inside the expectation in (8). Let us focus on each factor separately, after which we assemble them into an overall procedure. The first factor is called the *score function*, which is the

Algorithm 1: Policy Gradient with Stochastic Recursive Mirror Ascent (SRMA)

```
**Initialize** : Initial parameters, $\boldsymbol{\theta}_0$, $\beta$, $\gamma$, step-size $\alpha$
**Repeat for** $k = 1, \ldots$
    Simulate trajectories, $\tau_k = (s_0, a_0, s_1, a_1, \ldots)$ by $\pi_{\boldsymbol{\theta_k}}(.|s)$
    Estimate $\hat{Q}^{\pi_{\boldsymbol{\theta}_k}}$ via Monte-Carlo rollout (9)
    Initialize $g_0 = 0$
    $\mathbf{g}_k \leftarrow \sum_{t=0}^{\tau_k} \nabla \log \pi_{\theta_k}(a_k \mid s_k) \cdot \hat{Q}^{\pi_{\theta_k}}(s_k, a_k)$
    $\hat{\mathbf{g}}_k = (1-\beta)\left(\hat{\mathbf{g}}_{k-1} - \nabla F(\boldsymbol{\theta}_{k-1}, \xi_k)\right) + \nabla F(\boldsymbol{\theta}_k, \xi_k)$
    $\boldsymbol{\theta}_{k+1} = \text{argmax}_{\boldsymbol{\theta}}\{\langle \hat{\mathbf{g}}_k, \boldsymbol{\theta}\rangle - \frac{1}{\alpha} D_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}_k)\}$
    $k \leftarrow k + 1$
**Until Convergence**
**Return:** $\boldsymbol{\theta}_k$
```

gradient of log-likelihood of selecting an action according to policy $\pi_{\boldsymbol{\theta}}$. The later factor, the Q-function [cf. (2)], may be estimated by a Monte Carlo rollout along trajectory $\{s_u, a_u\}_{u=0}^{T'}$ starting from s_0, a_0

$$\hat{Q}^{\pi_{\boldsymbol{\theta}}}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} r(s_k, a_k),$$
$$s_0 = s, a_0 = a, T' := \text{Geom}(1-\gamma^{1/2}) \tag{9}$$

where $\hat{Q}^{\pi_{\boldsymbol{\theta}}}(s,a)$ denotes an estimator for the Q-function assuming actions follow policy $\pi_{\boldsymbol{\theta}}$, and T' is a randomized time-horizon for the rollout length chosen according to a geometric distribution. The reason for using random rollout horizons is that the estimate in (9) may be shown to yield unbiased Q-estimates for the infinite-horizon discounted setting under consideration – see [19][Theorem 4.3]. With this in hand, assuming initialization for the Monte Carlo rollout for estimating Q at the state-action pair at the previous iteration k, i.e., $(s_0, a_0) = (s_{k-1}, a_{k-1})$, then policy gradient method operates by collecting the stochastic gradient estimate

$$\hat{\nabla} J(\boldsymbol{\theta}_k) = \nabla \log \pi_{\boldsymbol{\theta}_k}(a_k \mid s_k) \cdot \hat{Q}^{\pi_{\boldsymbol{\theta}_k}}(s_k, a_k) \tag{10}$$

and then performing the stochastic gradient ascent as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \hat{\nabla} J(\boldsymbol{\theta}_k) \tag{11}$$

where α denoted the step size. Note the need for two time-scales: index k , in (9) denotes rollout trajectory information that occurs on a faster time-scale than policy gradient updates, which are indexed by slower time-scale k.

By employing the iteration (11), one may obtain convergence to stationary points of (3) (see [19], [21]), or in some cases, convergence to global optimality [24]–[26], [30]. To do so, however, to date, most results require the score function to be deterministically bounded over the entire state space and action space [12]. Unfortunately, this restriction precludes the use of heavy-tailed distributions from Examples 2-3. In particular, the expressions for the score function associated with the adaptive-variance Gaussian takes the form

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(s,a) = \begin{bmatrix} \frac{(a - s^\top \boldsymbol{\theta}_1)s}{\sigma^2} \\ \frac{\left(a - s^\top \boldsymbol{\theta}_1\right)^2}{\sigma^2} - 1 \end{bmatrix}, \tag{12}$$

where the above gradient is evaluated with respect to $\boldsymbol{\theta}_1$ and θ_2, and that of the Cauchy distribution is given as

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(s,a) = \begin{bmatrix} \frac{2}{1 + \left(\frac{(a - s^\top \boldsymbol{\theta}_1)}{\sigma}\right)^2} \left(\frac{(a - s^\top \boldsymbol{\theta}_1)}{\sigma^2}\right) s \\ -\mathbf{1} + \frac{2}{1 + \left(\frac{(a - s^\top \boldsymbol{\theta}_1)}{\sigma}\right)^2} \left(\frac{a - s^\top \boldsymbol{\theta}_1}{\sigma}\right)^2 \end{bmatrix}. \tag{13}$$

The expression in (12) and (13) illuminate the presence of σ in the denominator. Moreover, the definition of the α stable distribution specifies $\sigma = e^{\theta_2} \in (0, \infty)$. Thus, if one holds σ as a fixed constant (as is customary in practice), it is justifiable to assume an upper bound on the norm of the gradient of score function. However, for variable σ, the domain of σ implies the score function is unbounded as σ may be arbitrary close to 0. Fixing $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ for sufficient performance is application-specific and nontrivial to discern, often resulting in trial and error procedures whose performance is difficult to quantify [50].

Therefore, policy search over the family of heavy-tailed distributions requires score functions to be unbounded. However, we note that this boundedness issue can cause numerical instabilities in the sequence of policy parameters in practice. We propose to mitigate this issue by taking inspiration from proximal policy optimization [23], which restricts movement of policy parameters approximately through the introduction of proximal regularization into the standard gradient ascent update. Doing so prevents the update from becoming too large even when the gradient is large. In the stochastic setting, such regularization as been studied as stochastic mirror ascent [53], [54]. Next, we discuss the technical limitations of existing stochastic mirror ascent approaches, which motivate a modification that uses an additional recursive averaging step. The stochastic mirror ascent update for (3) is given by

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left\{ \left\langle \mathbf{g}_k, \boldsymbol{\theta} \right\rangle - \frac{1}{\alpha} D_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}_k) \right\}, \tag{14}$$

where $\mathbf{g}_k$ denotes the unbiased stochastic estimate of the gradient $\nabla J(\boldsymbol{\theta}_k)$ and where D_ψ denotes a Bregman divergence defined with respect to the strongly convex function $\psi(\mathbf{x})$ with ζ as the strong convexity parameter. We remark here that the update in (14) boils down the the standard stochastic gradient ascent (hence policy gradient in (11)) for $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$. To analyze the update in (14), we define the Bregman gradient $\mathcal{G}^{\psi}_{\alpha, \mathbf{g}_k}(\boldsymbol{\theta}_t)$ corresponding to the stochastic estimate of the gradient $\mathbf{g}_k$ as a generalized notion of gradient [53], [54], which allows us to reformulate (14) as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_t + \alpha \mathcal{G}^{\psi}_{\alpha, \mathbf{g}_k}(\boldsymbol{\theta}_t) \tag{15}$$

Optimality Criteria. For the convergence to first-order stationary point of stochastic mirror ascent, we focus on attenuation of the norm of the generalized gradient to a small

Algorithm 2: Exploratory Policy Search

```
Initialize : Initial policy parameter, $\boldsymbol{\theta}_0$, $\epsilon$, $\gamma$
Repeat for $k = 1, \ldots$
Simulate trajectories, $\tau_k = (s_0, a_0, s_1, a_1, \ldots)$ by $\pi_{\boldsymbol{\theta_k}}(.|s)$
Estimate $\hat{Q}^{\pi_{\boldsymbol{\theta}_k}}$ via Monte-Carlo rollout (9)
for $(s_k, a_k) \in \tau_k$ do
    Obtain projected gradient using (19)
    Update policy parameter, $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha \hat{\nabla} J^{\mathrm{p}}(\boldsymbol{\theta}_k)$
    $k \leftarrow k + 1$
end for
 Until Convergence
Return: $\boldsymbol{\theta}_k$
```

constant ϵ as $\mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\mathbf{g}_k}(\boldsymbol{\theta})\|^2\right] \leq \epsilon$ which defines ϵ first-order stationarity. We first note that it is well established in [53] that with batch size (number of stochastic gradient samples per iteration) of $B_k = 1$, it holds that

$$\mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\mathbf{g}_k}(\boldsymbol{\theta})\|^2\right] \leq \frac{D_\psi + \frac{\sigma^2}{\zeta}\sum_{k=1}^{K}\alpha_k}{\sum_{k=1}^{K}(\zeta\alpha_k - L\alpha_k^2)} \tag{16}$$

where α_k is the step size used at each k. From the right hand side of (16), we can conclude that it is lower bounded by $\frac{\sigma^2}{\zeta^2}$, no matter what the step size is selected. In the literature, batch methods (where batch size B_k increases with k) are used to overcome this issues and obtain convergent algorithms proposed in [53], [54]. But obtaining a convergent algorithm for SMA algorithm with general non-convex *un-regularized* objective and fixed batch size per iteration remains a challenge.

In this work, we address this issue via a recursive averaging step together with a difference of two gradient evaluations at each k given as

$$\hat{\mathbf{g}}_k = (1-\beta)\left(\hat{\mathbf{g}}_{k-1} - \nabla F(\boldsymbol{\theta}_{k-1}, \xi_k)\right) + \nabla F(\boldsymbol{\theta}_k, \xi_k)\;, \tag{17}$$

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{\langle \hat{\mathbf{g}}_k, \boldsymbol{\theta}\rangle - \frac{1}{\alpha} D_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}_k)\}. \tag{18}$$

which we call Stochastic Recursive Mirror Ascent (SRMA). Note that the major difference we have as compared to the existing SMA algorithms is in the update for gradient estimation in (17) for each k. To get the gradient estimate mentioned in (17), we require 2 ($\mathcal{O}(1)$) gradient samples per iteration. This method is summarized as Algorithm 1.

In Algorithm 1, a bottleneck arises in step 7, where solving the optimization problem to obtain the next policy parameter may be demanding unless one specifies a Bregman divergence whose closed-form minimizer is evaluable. One way to resolve this, as is typical of proximal policy optimization (PPO) [23], is to introduce "gradient clipping" to promote stability. We identify this operation as a special case of Moreau-Yosida regularization [55], which may be analyzed in the framework of mirror descent based approaches and has been studied recently for the case of bounded score functions in [24], [25]. Next we present the explicit form of gradient clipping-based updates in Algorithm 2 with the understanding that it is a special case of Algorithm 1. The update direction for gradient clipping-based PPO is given as

$$\hat{\nabla} J^{\mathrm{p}}(\boldsymbol{\theta}_k) = \Pi_{\mathcal{B}}\left(\nabla \log \pi_{\boldsymbol{\theta}_k}(a \,|\, s) \cdot \hat{Q}^{\pi_{\boldsymbol{\theta}_k}}(s_k, a_k)\right) \tag{19}$$

where $\mathcal{B}$ represents a Euclidean closed ball, $\mathcal{B}(r) = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < \delta\right\}$ and $\delta > 0$ is the tuning parameter. The projection in (19) restricts the norm of the gradient by δbut otherwise maintains its directionality the direction same as the original gradient. This modified gradient of (19) yields parameter updates in the form of (15). Next we proceed to prove the convergence of the proposed iterates.

IV. Convergence Analysis

In this section, we establish that Algorithm 1 converges to stationarity in the sense of $\mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\mathbf{g}_k}(\boldsymbol{\theta})\|^2\right] \leq \epsilon$. Without loss of generality, we reformulate the problem in the syntax of minimization, that is, we consider minimizing a function $F(\boldsymbol{\theta}) := -J(\boldsymbol{\theta})$, with $J(\boldsymbol{\theta})$ as the cumulative return under policy π_θ in (3). Hence the problem we consider for the analysis is given by

$$\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) \tag{20}$$

Let $\nabla F(\boldsymbol{\theta}_k)$ denote the gradient of function $F(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_k$ and $\xi_k \supseteq \{(s_t, a_t, r_t)\}_{t=0}^{T_k'}$ as the trajectory generated at instance k to estimate the policy gradient, with $s_0 = s_{k-1}, a_0 = a_{k-1}$ as the starting point of the trajectory from step $k-1$. The associated stochastic gradient estimate is denoted as $\nabla F(\boldsymbol{\theta}_k, \xi_k) := \mathbf{g}_k$. With this modified syntax (20), we may rewrite the parameter update for θ_k as

$$\hat{\mathbf{g}}_k = (1-\beta)\left(\hat{\mathbf{g}}_{k-1} - \nabla F(\boldsymbol{\theta}_{k-1}, \xi_k)\right) + \nabla F(\boldsymbol{\theta}_k, \xi_k)\;, \tag{21}$$

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{\langle \hat{\mathbf{g}}_k, \boldsymbol{\theta}\rangle + \frac{1}{\alpha} D_\psi(\boldsymbol{\theta}, \boldsymbol{\theta}_k)\}. \tag{22}$$

We denote as $\mathcal{F}_k := \{\xi_1, \xi_2, \cdots, \xi_{k-1}\}$ the set of random quantities (trajectories) until point k, which we used to state the assumptions next.

Assumption 1. *The stochastic estimate is unbiased means* $\mathbb{E}[\nabla F(\boldsymbol{\theta}_k, \xi_k) \mid \mathcal{F}_k] = \nabla F(\boldsymbol{\theta}_k)$ *for all* k.

Assumption 2. *The variance of the stochastic gradient satisfies the growth condition* $\mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}_k, \xi_k) - \nabla F(\boldsymbol{\theta}_k)\|^2\right] \leq m_0 + m_1\|\nabla F(\boldsymbol{\theta}_k)\|^2$ *for all* k *where* m_0, m_1 *are scalars.*

Assumption 3. *The original gradient* $\nabla F(\boldsymbol{\theta}_k)$ *and the stochastic Bergman gradient* $\mathcal{G}^{\psi}_{\alpha,\hat{\mathbf{g}}_k}(\boldsymbol{\theta}_k)$ *satisfy the error bound condition*

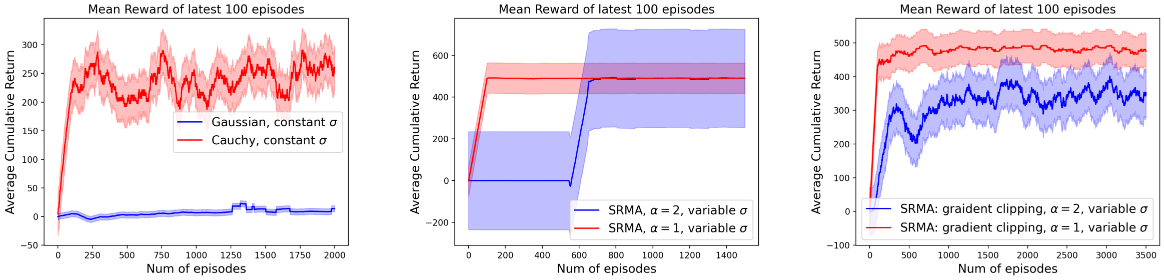
$$\mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}_k) - \mathcal{G}^{\psi}_{\alpha,\hat{\mathbf{g}}_k}(\boldsymbol{\theta}_k)\|^2\right] \leq m_2 + m_3\mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\hat{\mathbf{g}}_k}(\boldsymbol{\theta}_k)\|^2\right] \tag{23}$$

where m_2, m_3 *are scalars.*

Assumption 4. *The objective function* $F(\cdot)$ *is is* L*-smooth.*

Assumption 5. *The instantaneous objective function gradient* $\nabla F(\boldsymbol{\theta}_k, \xi_k)$ *is is* L_1- *Lipschitz which implies that*

$$\|\nabla F(\boldsymbol{\theta}_1, \xi_k) - \nabla F(\boldsymbol{\theta}_2, \xi_k)\| \leq L_1\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|. \tag{24}$$



(a) (b) (c)

Fig. 2: **(a)** Policy gradient with constant variance for the Mountain Car-like environment with misaligned long and short-term incentives in continuous space in Fig. 1. We visualize the average cumulative returns over latest 100 episodes for Gaussian and Cauchy policies with constant sigma. The importance of searching over a heavy-tailed (Cauchy) distribution is clear, as the Gaussian policy converges to spurious behavior.

(b) Stochastic Recursive Mirror Ascent (SRMA) applied to Fig. 1, where we visualize average cumulative returns over latest 100 episodes for policies with variable sigma using Algorithm 1. Variable σ provides favorable exploratory behavior, ensuring one reaches the bonanza of 500 reward. Moreover, interpolation [cf. (17)]/proximal steps [cf. (18)] preclude instability in the sample path of policy updates. Heavy-tailed distribution reaches the bonanza in a fraction of the required episodes of the Gaussian.

(c) Stochastic Recursive Mirror Ascent (SRMA) as instantiated with gradient clipping (Algorithm 2), with episodes initialized at false goal post: Average cumulative returns over latest 100 episodes for policies with variable sigma where each episode is initialized at the false goal point. Cauchy policy rapidly escapes the spurious stationary point towards the large goal at the top of the mountain.

Assumptions 1-2 are standard in the optimization literature for the unbounded gradient settings [18], [47]. The condition in Assumption 3 assumes a bound between the original gradient estimate and the generalized gradient evaluated at the current biased estimate of the gradient $\mathbf{g}_k$. Note that for the case when $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ and we utilized the stochastic unbiased gradient at each k, Assumption 3 would boils down to the statement of Assumption 2. Assumption 4-5 are related to the smoothness of the objective function F and the stochastic estimate of the objective function. Before proceeding with the main result of this work, we recall two important properties of the generalized gradient from [56, Lemma 1]:

$$\langle \mathbf{g}_k, \mathcal{G}^{\psi}_{\alpha,\mathbf{g}_k}(\boldsymbol{\theta}_t)\rangle \geq \zeta \|\mathcal{G}^{\psi}_{\alpha,\mathbf{g}_k}(\boldsymbol{\theta}_t)\|^2 \tag{25}$$

$$\|\mathcal{G}^{\psi}_{\alpha,\mathbf{g}_1}(\boldsymbol{\theta}) - \mathcal{G}^{\psi}_{\alpha,\mathbf{g}_2}(\boldsymbol{\theta})\| \leq \frac{1}{\zeta}\|\mathbf{g}_1 - \mathbf{g}_2\|. \tag{26}$$

The inequalities (25) and (26) will be used in the analysis. Next, we present an intermediate lemma which bounds the stochastic errors associated with gradient estimation $\varepsilon_k := \mathbb{E}\left[\|\mathbf{w}_k\|^2\right]$ where $\mathbf{w}_k = \hat{\mathbf{g}}_k - \nabla F(\boldsymbol{\theta}_k)$.

Lemma 1. *Let $\varepsilon_k := \mathbb{E}\left[\|\mathbf{w}_k\|^2\right]$, For all $k \geq 1$, it holds that*

$$\begin{aligned}\varepsilon_k \leq &(1-\beta)^2\varepsilon_{k-1} + 2\alpha^2 L\mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\hat{\mathbf{g}}_k}(\boldsymbol{\theta}_k)\|^2\right] + 2m_0\beta^2 \\ &+ 2m_1\beta^2\|\nabla F(\boldsymbol{\theta}_k)\|^2.\end{aligned} \tag{27}$$

The proof of Lemma 1 is provided in Appendix VII of the supplementary material [57]. The result in Lemma 1 bounds the per step expected value of the the norm of error for each k. Next we present the main theorem of this paper.

Theorem 1. *Under Assumption 1-5, under step-size selections $\beta = C_1\alpha$ with $C_1 \geq 0$ and $\alpha \leq \min\left\{\frac{\zeta L'}{10}, \frac{\zeta}{8\widetilde{m}_3 C_1^2}\right\}$, in order to achieve*

$$\min_{1\leq k\leq K} \mathbb{E}\left[\|\mathcal{G}^{\psi}_{\alpha,\hat{\mathbf{g}}_k}(\boldsymbol{\theta}_k)\|^2\right] \leq \epsilon \tag{28}$$

with $\epsilon \leq \min\left\{\frac{\zeta L'}{10}, \frac{\zeta}{8\widetilde{m}_3 C_1^2}\right\}$, the iterates in Algorithm 1 requires at least $K \geq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations with $\mathcal{O}(1)$ stochastic gradients samples (calls to a simulation oracle) at each k.

We note that a related but simpler specification of step-size α also permitted: $\alpha = \frac{\alpha_0}{\sqrt{K}}$ in terms of final iteration index K with $\alpha_0 = \min\left\{\frac{\zeta L'}{10}, \frac{\zeta}{8\widetilde{m}_3 C_1^2}\right\}$.

The proof of Theorem 1 is provided in Appendix VIII of supplementary material [57]. Note that the use of recursive update for the stochastic gradient estimate in (17) permits us to achieve the $\mathcal{O}(\frac{1}{\epsilon^2})$ oracle complexity with a constant batch size of gradients $\mathcal{O}(1)$ per iteration. In related work on stochastic mirror descent based algorithms for the non-convex objective, an increasing batch size is mandatory to converge [54], [56]. We summarize the results in Table I. These results permit one to consider policy optimization over classes of heavy-tailed policies whose score functions may be unbounded, provided they satisfy the error bound conditions in Assumptions 2 - 3. We evaluate the experimental utility of these approaches next.

Refs.	Samples per Iteration	No. of Iterations
[53]	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$
[54]	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(1/\epsilon^2)$
This work	$\mathcal{O}(1)$	$\mathcal{O}(1/\epsilon^2)$

TABLE I: Summary of related results.

V. Simulations

This section validates the efficacy of proposed policy parameterization using the representative example introduced in Section II. At each episode, the position of the car is initialized within a small neighborhood $[1.15, 2.0]$. State, x is constrained to an interval $[-4.0,\ 3.709]$ and action a_k lies in $[-20,\ 20]$. The discounted factor γ is 0.97 and we use a diminishing step-size ranging from 0.005 to 5×10^{-9}. We compare performance of the policy parameterizations from Examples 1-3. First, the setting is validated using SαS distributions of constant sigma and is further extended to variable scale for a persistent exploration. Adaptive variance and associated unboundedness are taken care of using the proposed SRMA of Algorithm 1 and gradient clipping introduced in Algorithm 2.

A. $S\alpha S$ distributions with constant scale

We evaluate the representative example setting using policy parameterization for $S\alpha S$ stable distributions with α values of 1 and 2 for constant scale. In other words, this corresponds to Gaussian and Cauchy policy of (4) and (6) with a constant σ ($\sigma = 1.0$). Figure 2(a) shows the cumulative return averaged over latest 100 episodes. Heavy tailed Cauchy distribution shows better performance as compared to standard Gaussian. However, as evident from the figure, both policies exhibit high variance. Though the heavy-tailed distribution of $\alpha = 2$ achieves better performance, it is interesting to note that averaged cumulative return is less than 500, hint at convergence to misleading goals at times. The result also hints at the necessity of persistent exploration with variable scale for $S\alpha S$ distributions as discussed in Section II.

B. $S\alpha S$ distributions with SRMA

Next we consider policy parameterizations of Example 3 with variable scale. The unbounded score function and related divergence issues in the implementation are taken care of using the proposed Policy Gradient with SRMA of Algorithm 1. SαS distributions of variable scale provide better exploratory behavior, and both the policies result in convergence to the desired goal. Each episode is initialized within a small neighborhood $[1.15, 2.0]$ as in the previous case studies. We use a β value of 0.9. Fig. 2(b) also shows faster convergence of heavy-tailed distributions in addition to the lesser variance in average cumulative returns. Besides, the heavy tail of the Cauchy policy converges faster to the long-term incentive.

C. $S\alpha S$ stable distributions of variable scale: Policy Gradient with Gradient Clipping

Numerical experiments from the previous case studies show better convergence with proposed policy parameterizations of Example 3. Next, we evaluate the performance of $S\alpha S$ parameterizations using the projected policy gradients of Algorithm 2, a special case of mirror ascent with lesser computational requirements as discussed in Section III. To further illustrate the degree of exploratory behavior of the proposed distributions with tail index α, we initiate each episode with the false goal point. Policies with tail index $\alpha = 1$ and 2 manage to escape the short-term incentive with time as evident from the average cumulative return of Fig 2(b). Besides, results indicate the relatively lesser exit time of heavier tailed policy ($\alpha = 1$) as shown in Fig 2(c). It is to be noted that the absence of recursive mirror ascent introduces variance in policy updates in contrast to Fig 2(b).

VI. Conclusion

In this work, we focused on policy gradient method for solving RL problems associated with infinite-horizon discounted returns. In some problems, the one-step reward may be very far from the value of a given state, which can cause policies to become mired at spurious behavior. Inspired by the persistent exploration conditions that preclude this behavior in finite MDPs, we proposed to study ways to incorporate it in continuous settings through policies parameterized as heavy-tailed distributions. This parameterization introduced numerical challenges, namely, unbounded score functions and potentially volatile changes in policy gradients.

To address these issues, we explicitly analyzed the behavior of policy gradient when the score function may be unbounded, as well as introduced proximal variants to ensure stable updates. To assure stability of the resulting algorithm, we further introduced a gradient interpolation step in order to mitigate the mini-batch growth conditions that exist for stochastic mirror descent. The convergence the resulting iterative schemes was established under novel error bound conditions for the generalized Bregman gradient. Moreover, experimentally, we observed favorable performance of the proposed approach for escaping spurious stationary points. Future work includes a rigorous study of the generalization behavior of the resulting class of heavy-tailed policies.

References

[1] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement Learning: An Introduction*, 2nd ed., 2017.

[2] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” *arXiv preprint arXiv:1506.02438*, 2015.

[3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in *International Conference on Learning Representations*, 2016.

[4] L. Zou, L. Xia, Z. Ding, J. Song, W. Liu, and D. 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*, 2019, pp. 2810–2818.

[5] M. R. Kosorok and E. E. Moodie, *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.

[6] M. L. Puterman, *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[7] R. E. Bellman, *Dynamic Programming*. Courier Dover, 1957.

[8] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[9] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[10] E. Even-Dar, Y. Mansour, and P. Bartlett, "Learning rates for q-learning." *Journal of machine learning Research*, vol. 5, no. 1, 2003.

[11] A. M. Devraj and S. P. Meyn, "Zap q-learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2232–2241.

[12] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *NeurIPS*, 2000, pp. 1057–1063.

[13] V. R. Konda and V. S. Borkar, "Actor-critic–type learning algorithms for Markov decision processes," *SICON*, vol. 38, no. 1, pp. 94–123, 1999.

[14] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *NeurIPS*, 2000, pp. 1008–1014.

[15] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor-critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[16] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[17] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[18] S. Bhatt, A. Koppel, and V. Krishnamurthy, "Policy gradient using weak derivatives for reinforcement learning," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 5531–5537.

[19] K. Zhang, A. Koppel, H. Zhu, and T. Basar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.

[20] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.

[21] J. Zhang, J. Kim, B. O'Donoghue, and S. Boyd, "Sample efficient reinforcement learning with reinforce," *arXiv preprint arXiv:2010.11364*, 2020.

[22] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," in *Conference on Learning Theory*. PMLR, 2020, pp. 64–66.

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

[24] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh, "Mirror descent policy optimization," *arXiv preprint arXiv:2005.09814*, 2020.

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

[26] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6820–6829.

[27] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Annals of Operations Research*, vol. 46, no. 1, pp. 157–178, 1993.

[28] S. Lojasiewicz, "A topological property of real analytic subsets," *Coll. du CNRS, Les équations aux dérivées partielles*, vol. 117, pp. 87–89, 1963.

[29] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[30] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *arXiv preprint arXiv:1908.00261*, 2019.

[31] J. Mei, C. Xiao, B. Dai, L. Li, C. Szepesvári, and D. Schuurmans, "Escaping the gravitational pull of softmax," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[32] K. S. Narendra and A. M. Annaswamy, "Persistent excitation in adaptive systems," *International Journal of Control*, vol. 45, no. 1, pp. 127–160, 1987.

[33] ——, *Stable adaptive systems*. Courier Corporation, 2012.

[34] M. C. Bryson, "Heavy-tailed distributions: properties and tests," *Technometrics*, vol. 16, no. 1, pp. 61–68, 1974.

[35] S. M. Focardi and F. J. Fabozzi, "Fat tails, scaling, and stable laws: a critical look at modeling extremal events in financial phenomena," *The Journal of Risk Finance*, 2003.

[36] J. E. Hutchinson, "Fractals and self similarity," *Indiana University Mathematics Journal*, vol. 30, no. 5, pp. 713–747, 1981.

[37] B. B. Mandelbrot and B. B. Mandelbrot, *The fractal geometry of nature*. WH freeman New York, 1982, vol. 1.

[38] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random house, 2007, vol. 2.

[39] J. B. Taylor and J. C. Williams, "A black swan in the money market," *American Economic Journal: Macroeconomics*, vol. 1, no. 1, pp. 58–83, 2009.

[40] D. Avnir, O. Biham, D. Lidar, and O. Malcai, "Is the geometry of nature fractal?" *Science*, vol. 279, no. 5347, pp. 39–40, 1998.

[41] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[42] M. Gurbuzbalaban, U. Simsekli, and L. Zhu, "The heavy-tail phenomenon in sgd," *arXiv preprint arXiv:2006.04740*, 2020.

[43] U. Simsekli, L. Zhu, Y. W. Teh, and M. Gurbuzbalaban, "Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8970–8980.

[44] R. Pemantle *et al.*, "Nonconvergence to unstable points in urn models and stochastic approximations," *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.

[45] S. B. Gelfand and S. K. Mitter, "Recursive stochastic algorithms for global optimization in rˆd," *SIAM Journal on Control and Optimization*, vol. 29, no. 5, pp. 999–1018, 1991.

[46] U. Simsekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu, "Hausdorff dimension, heavy tails, and generalization in neural networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 5138–5151.

[47] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[48] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[49] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: the discrete-time case*, 2004.

[50] M. Papini, A. Battistello, and M. Restelli, "Balancing learning speed and stability in policy gradient via adaptive exploration," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1188–1199.

[51] A.-L. Barabási *et al.*, "Emergence of scaling in complex networks," *Handbook of Graphs and Networks: From the Genome to the Internet. Berlin: Wiley-VCH*, 2003.

[52] T. H. Nguyen, U. $\mathcal{S}$im$\int$ekli, M. Gürbüzbalaban, and G. Richard, "First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise," *arXiv preprint arXiv:1906.09069*, 2019.

[53] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.

[54] L. Yang, G. Zheng, H. Zhang, Y. Zhang, Q. Zheng, J. Wen, and G. Pan, "Policy optimization with stochastic mirror descent," *arXiv preprint arXiv:1906.10462*, 2019.

[55] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM journal on control and optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[56] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIOPT*, vol. 23, no. 4, pp. 2341–2368, 2013.

[57] A. Parayil, A. S. Bedi, M. Wang, and A. Koppel, "Supplementary material for proximal policy optimization with unbounded score functions for persistent exploration in reinforcement learning," Mar. 2021. [Online]. Available: https://tinyurl.com/mert4uup

[58] A. Cutkosky and F. Orabona, "Momentum-based variance reduction in non-convex sgd," *arXiv preprint arXiv:1905.10018*, 2019.