PAGE-PG: A Simple and Loopless Variance-Reduced Policy Gradient Method with Probabilistic Gradient Estimation

Matilde Gargiani [1] **Andrea Zanelli** [2] **Andrea Martinelli** [1] **Tyler H. Summers** [3] **John Lygeros** [1]

Abstract

Despite their success, policy gradient methods suffer from high variance of the gradient estimator, which can result in unsatisfactory sample complexity. Recently, numerous variance-reduced extensions of policy gradient methods with provably better sample complexity and competitive numerical performance have been proposed. After a compact survey on some of the main variance-reduced REINFORCE-type methods, we propose ProbAbilistic Gradient Estimation for Policy Gradient (PAGE-PG), a novel loopless variance-reduced policy gradient method based on a probabilistic switch between two types of update. Our method is inspired by the PAGE estimator for supervised learning and leverages importance sampling to obtain an unbiased gradient estimator. We show that PAGE-PG enjoys a $\mathcal{O}\left(\epsilon^{-3}\right)$ average sample complexity to reach an ϵ-stationary solution, which matches the sample complexity of its most competitive counterparts under the same setting. A numerical evaluation confirms the competitive performance of our method on classical control tasks.

1. Introduction

Policy gradient methods have proved to be really effective in many challenging deep reinforcement learning (RL) applications (Sutton & Barto, 2018). Their success is also due to their versatility as they are applicable to any differentiable policy parametrization, including complex neural networks, and they admit easy extensions to model-free settings and continuous state and action spaces. This class of methods has a long history in the RL literature that dates back to (Williams, 1992), but only very recent work (Agarwal et al., 2019) has characterized their theoretical properties, such as convergence to a globally optimal solution and sample and iteration complexity. Since in RL it is generally not possible to compute the exact gradient, but we rely on sample-based approximations, policy gradient methods are negatively affected by the high-variance of the gradient estimator, which slows down convergence and leads to unsatisfactory sample complexity. To reduce the variance of the gradient estimators, actor-critic methods are deployed, where not only the policy, but also the state-action value function or the advantage function are parameterized (Mnih et al., 2016). Alternatively, taking inspiration from stochastic optimization, various variance-reduced policy gradient methods have been proposed (Sidford et al., 2018; Papini et al., 2018; Xu et al., 2019; 2021; Yuan et al., 2020; Zhang et al., 2021).

In this work, we focus on variance-reduced extensions of REINFORCE-type methods, such as REINFORCE (Williams, 1992), GPOMDP (Baxter & Bartlett, 2001) and their variants with baseline (Sutton & Barto, 2018). After reviewing the principal variance-reduced extensions of REINFORCE-type methods, we introduce a novel variance-reduced policy gradient method, PAGE-PG, based on the recently proposed PAGE estimator for supervised learning (Li et al., 2021). We prove that PAGE-PG only takes $\mathcal{O}\left(\epsilon^{-3}\right)$ trajectories on average to achieve an ϵ-stationary policy, which translates into a near-optimal solution for gradient dominated objectives. This result matches the bounds on total sample complexity of the most competitive variance-reduced REINFORCE-type methods under the same setting. The key feature of our method consists in replacing the double-loop structure typical of variance-reduced methods with a probabilistic switch between two types of updates. According to recent works in supervised learning (Kovalev et al., 2020; Li et al., 2021), variance-reduced methods that do not rely on the classical double-loop structure, also called *loopless*, are easier to tune, analyze and generally lead to superior and more robust practical behavior. For policy gradient optimization, similar advantages are discussed in (Yuan et al., 2020; Huang et al., 2020), where the authors propose STORM-PG and IS-MBPG, respectively, which, to the best of our knowledge, are the only

[1] Automatic Control Laboratory at ETH Zurich [2] Institute for Dynamic Systems and Control at ETH Zurich [3] Department of Mechanical Engineering at the University of Texas at Dallas. Correspondence to: Matilde Gargiani <gmatilde@ethz.ch>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

other loopless variance-reduced REINFORCE-type policy gradient counterparts to our method. Both STORM-PG and IS-MBPG are based on the idea of incorporating momentum in the update, while our method is based on replacing the outer loop with a coin flip which triggers with a certain probability the computation of a large-batch gradient estimate. In addition, with respect to STORM-PG, our method enjoys a better theoretical rate of convergence. Our experiments show the competitive performance of PAGE-PG on classical control tasks. Finally, we describe the limitations of the considered methods, discuss promising future extensions as well as the importance of incorporating noise annealing and adaptive strategies in the PAGE-PG's update. These might favor exploration in the early stages of training and improve convergence in presence of complex non-concave landscapes (Neelakantan et al., 2015; Smith et al., 2018; Zhou et al., 2019).

Main contributions. Our main contributions are summarized below.

- We propose PAGE-PG, a novel loopless variance-reduced extension of REINFORCE-type methods based on a probabilistic update.
- We show that PAGE-PG enjoys a fast rate of convergence and achieves an ϵ-stationary policy within $\mathcal{O}\left(\epsilon^{-3}\right)$ trajectories on average. We further show that, with gradient dominated objectives, similar results are valid for near-optimal solutions.

2. Problem Setting

In this section, we describe the problem setting and briefly discuss the necessary background material on REINFORCE-type policy gradient methods.

Markov Decision Process. The RL paradigm is based on the interaction between an agent and the environment. In the standard setting, the agent observes the state of the environment and, based on that observation, plays an action according to a certain policy. As a consequence, the environment transits to a next state and a reward signal is emitted from the environment back to the agent. This process is repeated over a horizon of length $H > 0$, with $H < \infty$ in the episodic setting and $H \to \infty$ in the infinite-horizon setting. From a mathematical viewpoint, Markov Decision Processes (MDPs) are a widely utilized mathematical tool to describe RL tasks. In this work we consider discrete-time episodic MDPs $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho\}$, where $\mathcal{S}$ is the state space; $\mathcal{A}$ is the action space; P is a Markovian transition model, where $P(s' \,|\, s, a)$ defines the transition density to state s' when taking action a in state s; $r : \mathcal{S} \times \mathcal{A} \to [-U, U]$ is the reward function, where $U > 0$ is a constant; $\gamma \in (0, 1)$ is the discount factor; and ρ is the initial state distribution. The agent selects the actions according to a stochastic stationary policy π, which, given a state s, defines a density distribution over the action space $\pi(\cdot \,|\, s)$. A trajectory $\tau = \{s_h, a_h\}_{h=0}^{H-1}$ is a collection of states and actions with $s_0 \sim \rho$ and, for any time-step $h \geq 0$, $a_h \sim \pi(\cdot \,|\, s_h)$ and $s_{h+1} \sim P(\cdot \,|\, s_h, a_h)$. We denote the trajectory distribution induced by policy π_θ as $p(\tau \,|\, \theta)$.

The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ associated with a policy π and initial state s is defined as

$$V^\pi(s) := \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h r(s_h, a_h) \,|\, \pi, s_0 = s\right] ,$$

where the expectation is taken with respect to the trajectory distribution. With an overloaded notation, we denote with $V^\pi(\rho)$ the expected value under the initial state distribution ρ, i.e.,

$$V^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}\left[V^\pi(s_0)\right] . \tag{1}$$

The goal of the agent generally is to find the policy π that maximizes $V^\pi(\rho)$ in (1).

Policy Gradient. Given finite state and action spaces, the policy can be exactly coded with $|\mathcal{S}| \times |\mathcal{A}|$ parameters in the tabular setting. However, the tabular setting becomes intractable for large state and action spaces. In these scenarios, as well as in infinite countable and continuous spaces, we generally resort to parametric function approximations. In particular, instead of optimizing over the full space of stochastic stationary policies, we restrict our attention to the class of stochastic policies that is described by a finite-dimensional differentiable parametrization $\Pi_\theta = \left\{\pi_\theta \,|\, \theta \in \mathbb{R}^d\right\}$, such as a deep neural network (Levine & Koltun, 2014). The addressed problem therefore becomes

$$\max_{\theta \in \mathbb{R}^d} V^{\pi_\theta}(\rho) . \tag{2}$$

We denote with V^* the optimal value. To simplify the notation, we use $V(\theta)$ to denote $V^{\pi_\theta}(\rho)$, θ to denote π_θ and $R(\tau) = \sum_{h=0}^{H-1} \gamma^h r(s_h, a_h)$ to denote the discounted cumulative reward associated with trajectory τ. Problem (2) can be addressed via gradient ascent, which updates the parameter vector by taking fixed steps of length $\eta > 0$ along the direction of the gradient. The iterations are defined as

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta V(\theta_t) , \tag{3}$$

where the gradient is given by

$$\nabla_\theta V(\theta) = \mathbb{E}_{\tau \sim p(\cdot \,|\, \theta)}\left[\sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h \,|\, s_h) R(\tau)\right] . \tag{4}$$

In the model-free setting, we cannot compute the exact gradient as we do not have access to the MDP dynamics.

Instead, given a certain policy θ, we simulate a finite number $N > 0$ of trajectories, which are then used to approximate Equation (4) via Monte Carlo

$$\hat{\nabla}_\theta V^{\mathrm{RF}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h^i \mid s_h^i) R(\tau_i) \,, \quad (5)$$

where each trajectory $\tau_i = \left\{s_h^i, a_h^i\right\}_{h=0}^{H-1}$ is generated according to the trajectory distribution $p(\cdot \mid \theta)$. The estimator in Equation (5) is also known as the REINFORCE estimator. An alternative is given by the GPOMDP estimator

$$\hat{\nabla}_\theta V^{\mathrm{GPOMDP}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=0}^{H-1} \gamma^h r(s_h^i, a_h^i) Z_{\theta,h} \,, \quad (6)$$

where for compactness $Z_{\theta,h} = \sum_{z=0}^{h} \nabla_\theta \log \pi_\theta(a_z^i \mid s_z^i)$. Both REINFORCE and GPOMDP are unbiased estimators of the gradient, but they are not equivalent in terms of variance. Specifically for GPOMDP, by only considering the reward-to-go instead of the full reward, we are removing potentially noisy terms and therefore lowering the variance of our estimator (Zhao et al., 2011). In addition, since $\mathbb{E}\left[\nabla_\theta \log \pi_\theta(a \mid s) b(s)\right] = 0$ with $b(s)$ being a function of the state, e.g. the value function $V^\pi(s)$, both the REINFORCE and GPOMDP estimators can be used in combination with a baseline.

The discussed estimators (with or without baseline) are deployed in place of the exact gradient in Equation (3), leading to the REINFORCE and GPOMDP algorithms. These methods are reminiscent of stochastic gradient ascent (Bottou, 2012) that also relies on sample-based estimates of the true gradient.

Notation. With an overloaded notation, we use $g(\tau_i \mid \theta) = \sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h^i \mid s_h^i) R(\tau_i)$ for the REINFORCE estimator, and $g(\tau_i \mid \theta) = \sum_{h=0}^{H-1} \gamma^h r(s_h^i, a_h^i) Z_{\theta,h}$ for the GPOMDP estimator.

3. Related Work

Variance-reduction techniques have been first introduced for training supervised machine learning models, such as logistic regression, support vector machines and neural networks. Supervised learning is often recast into a finite-sum empirical risk minimization problem that in its simplest takes the form

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(f_i(\theta)) \,. \quad (7)$$

In the supervised learning scenario, n is the size of the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and ℓ is a loss function that measures the discrepancy between the model prediction $f_i(\theta) = f(x_i; \theta)$ and the true value y_i. If f is smooth and satisfies the Polyak-Łojasiewicz (PL) condition, gradient descent with an appropriate constant step-size enjoys a global linear rate of convergence (Karimi et al., 2020). Despite its fast rate, often the full gradient computation makes the iterations of gradient descent too expensive because of the large size of the training dataset. Mini-batch gradient descent replaces the full gradient with an estimate computed over a randomly sampled subset of the available samples. The method requires a decreasing step-size to control the variance and achieve convergence. As a consequence, the lower-iteration cost comes at the price of a slower sublinear convergence rate (Karimi et al., 2020). A better trade-off between computational costs and convergence rate is achieved by variance-reduced gradient methods, such as SVRG (Johnson & Zhang, 2013), Katyusha (Allen-Zhu, 2017), SARAH (Nguyen et al., 2017), STORM (Cutkosky & Orabona, 2019), L-SVRG and L-Katyusha (Kovalev et al., 2020), and PAGE (Li et al., 2021). Because of their provably superior theoretical properties and their competitive numerical performance, these methods have attracted great attention from the machine learning community in the past decade. A key structural feature of classical variance-reduced methods, such as SVRG, Katyusha and SARAH, is the double-loop structure. In the outer loop a full pass over the training data is made in order to compute the exact gradient, which is then used in the inner loop together with new stochastic gradient information to construct a variance-reduced estimator of the gradient. However, as underlined in (Zhou et al., 2019), the double-loop structure complicates the analysis and tuning, since the optimal length of the inner loop depends on the value of some structural constants that are generally unknown and often very hard to estimate. This inconvenience has fueled recent efforts from the supervised learning community to develop loopless variance-reduced gradient methods, such as L-SVRG, L-Katyusha and PAGE. In these methods the outer loop is replaced by a probabilistic switch between two types of updates: with probability p a full gradient computation is performed, while with probability $1 - p$ the previous gradient estimate is reused with a small adjustment that varies based on the method. In particular, L-SVRG and L-Katyusha, which are designed for smooth and strongly convex objective functions, recover the same fast theoretical rates of their loopy counterparts, SVRG and Katyusha, but they require less tuning and lead to superior and more robust practical behavior (Kovalev et al., 2020). Similar results also hold for PAGE (Li et al., 2021), which is designed for non-convex problems. In particular, in the non-convex finite-sum setting, PAGE achieves the optimal convergence rate and numerical evidence confirms its competitive performance.

Motivated by the great success of variance-reduction techniques in supervised learning, numerous recent works have explored their deployment in RL, and, in particular, their adaptation for policy gradient (Papini et al., 2018; Xu et al., 2021; Yuan et al., 2020; Zhang et al., 2021). As discussed in Section 2, the gradient is generally estimated based on a finite number N of observed trajectories. In online RL, the trajectories are sampled at each policy change. This is particularly costly since it requires one to simulate the system N times for every parameter update. Generally the batch-size is tuned to achieve the best trade-off between the cost of simulating the system and the variance of the estimator. This motivates the use of variance-reduced gradient methods, that use past gradients to reduce variance, leading to an improvement in terms of sample complexity. The deployment of variance-reduction techniques to solve Problem (2) is not straightforward and requires some adaptations to deal with the specific challenges of the RL setting (Papini et al., 2018). Differently from the finite-sum scenario, Problem (2) can not be recast as a finite-sum problem, unless both the state and action spaces are finite. This and the fact that the MDP dynamics are unknown prevent from the computation of the full gradient, which is generally replaced by an estimate based on a large batch-size.

An additional difficulty comes from the fact that the data distribution changes over time, since it depends on the parameter θ, which gets updated during training. This is known as *distribution shift* and requires the deployment of importance weighting in order to reuse past information without adding a bias to the gradient estimator. In particular, suppose we have two policies θ_1 and θ_2, where θ_2 is used for the interaction with the system, while we aim at obtaining an unbiased estimator of the gradient with respect to θ_1. The unbiased off-policy extension of the REINFORCE estimator (Papini et al., 2018) is obtained by replacing $g(\tau_i \,|\, \theta)$ in Equation (5) with the following quantity

$$g^{\omega_{\theta_2}}(\tau_i \,|\theta_1) = \omega(\tau_i \,|\, \theta_2, \theta_1) \sum_{h=0}^{H-1} \nabla_\theta \log \pi_{\theta_1}(a_h^i \,|\, s_h^i) R(\tau_i) \,, \tag{8}$$

where $\omega(\tau_i \,|\, \theta_2, \theta_1) = \Pi_{j=0}^{H-1} \frac{\pi_{\theta_1}(a_j^i \,|\, s_j^i)}{\pi_{\theta_2}(a_j^i \,|\, s_j^i)}$ is the importance weight for the full trajectory realization τ_i. Similarly, the off-policy extension of the GPOMDP estimator (Papini et al., 2018) is obtained by replacing $g(\tau_i \,|\, \theta)$ in Equation (6) with the following quantity

$$g^{\omega_{\theta_2}}(\tau_i \,|\theta_1) = \sum_{h=0}^{H-1} \omega_{0:h}(\tau_i \,|\, \theta_2, \theta_1) \gamma^h r(s_h^i, a_h^i) Z_{\theta_1, h} \,, \tag{9}$$

where $\omega_{0:h}(\tau_i \,|\, \theta_2, \theta_1) = \Pi_{j=0}^{h} \frac{\pi_{\theta_1}(a_j^i \,|\, s_j^i)}{\pi_{\theta_2}(a_j^i \,|\, s_j^i)}$ is the importance weight for the trajectory realization τ_i truncated at time h. Clearly, for the importance weights to be well-defined, the policy θ_2 needs to have a non-zero probability of selecting any action in every state. This assumption is implicitly required to hold where needed throughout the paper. It is easy to verify that, for both the off-policy extensions of REINFORCE and GPOMDP, $\mathbb{E}_{\tau \sim p(\cdot \,|\, \theta_2)} \left[g^{\omega_{\theta_2}}(\tau \,|\, \theta_1) \right] = \mathbb{E}_{\tau \sim p(\cdot \,|\, \theta_1)} \left[g(\tau \,|\, \theta_1) \right]$, leading to an unbiased estimator of the gradient at θ_1.

3.1. Variance-Reduced REINFORCE-type Methods

We now briefly review some of the state-of-the-art variance-reduced REINFORCE-type methods to solve Problem (2). We use $g(\tau \,|\, \theta)$ and $g^{\omega_{\theta_2}}(\tau \,|\, \theta_1)$ to refer to both the REINFORCE and GPOMDP estimators, without and with importance sampling, respectively.

Stochastic Varaice-Reduced Policy Gradient (SVRPG), first proposed in (Papini et al., 2018) and then further analyzed in (Xu et al., 2019), adapts the stochastic variance-reduced gradient method for finite-sum problems (Johnson & Zhang, 2013) to deal with the RL challenges as discussed above. The method is characterized by a double loop structure, where the outer iterations are called epochs. At the s-th epoch, a snapshot of the current iterate θ_0^s is taken. Then, $N >> 1$ trajectories $\{\tau_i\}_{i=1}^N$ are collected based on the current policy and used to compute the gradient estimator $v_0^s = \frac{1}{N} \sum_{i=1}^N g(\tau_i \,|\, \theta_0^s)$. For every epoch, m iterations in the inner loop are performed. At the t-th iteration of the inner loop with $t = 0, \ldots, m-1$, the parameter vector is updated by

$$\theta_{t+1}^s = \theta_t^s + \eta v_t^s \,, \tag{10}$$

where $\eta > 0$. Then $B << N$ trajectories $\{\tau_j\}_{j=1}^B$ are collected according to the current policy θ_{t+1}^s and an estimate of the gradient at θ_{t+1}^s is produced

$$v_{t+1}^s = \frac{1}{B} \sum_{j=1}^B g(\tau_j \,|\, \theta_{t+1}^s) + v_0^s - \frac{1}{B} \sum_{j=1}^B g^{\omega_{\theta_{t+1}^s}}(\tau_j \,|\, \theta_0^s) \,.$$

After m iterations in the inner loop, the snapshot is refreshed by setting $\theta_0^{s+1} = \theta_m^s$, and the process is repeated for a fixed number of iterations. See Algorithm 1 in Section A of the Appendix.

Stochastic Recursive Variance-Reduced Policy Gradient (SRVRPG) (Xu et al., 2021) is inspired from the SARAH method for supervised learning (Nguyen et al., 2017). Differently from SVRPG, SRVRPG incorporates in the update the concept of momentum, which helps convergence by dampening the oscillations typical of first-order methods. In particular, the estimate produced in the inner iterations for all $t = 0, \ldots, m-1$ is

$$v_{t+1}^s = \frac{1}{B} \sum_{i=1}^B g(\tau_j \,|\, \theta_{t+1}^s) + v_t^s - \frac{1}{B} \sum_{i=1}^B g^{\omega_{\theta_{t+1}^s}}(\tau_j \,|\, \theta_t^s) \,, \tag{11}$$

where $\{\tau_i\}_{i=1}^B$ are generated according to policy θ_{t+1}^s and v_0^s is the large batch-size estimate computed at the s-th epoch. See Algorithm 2 in Section A of the Appendix.

Stochastic Recursive Momentum Policy Gradient (STORM-PG) (Yuan et al., 2020) blends the key components of STORM (Cutkosky & Orabona, 2019), a state-of-the-art variance-reduced gradient estimator for finite-sum problems, with policy gradient algorithms. A major drawback of SVRPG and SRVRPG is the restarting mechanism, namely, the alternation between large and small batches of sampled trajectories which ensures control of the variance. As discussed for the finite-sum scenario, the double-loop structure complicates the theoretical analysis and the tuning procedure. STORM-PG circumvents the issue by deploying an exponential moving averaging mechanism that exponentially discounts the accumulated variance. The method only requires one to collect a large batch of trajectories at the first iteration and then relies on small batch updates. Specifically, STORM-PG starts by collecting $N >> 1$ trajectory samples $\{\tau_i\}_{i=1}^N$ according to an initial policy θ_0. Those samples are deployed to calculate an initial gradient estimate $v_0 = \frac{1}{N}\sum_{i=1}^N g(\tau_i \,|\, \theta_0)$, which is used in place of the gradient to update the parameter vector as in Equation (3). Then T iterations are performed where at the t-th iteration the parameter vector is updated as described in Equation (3), but replacing the gradient with the following estimate

$$v_t = \frac{1}{B}\sum_{i=1}^{B} g(\tau_i \,|\, \theta_t) + (1-\alpha)\Bigg[v_{t-1} - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1})\Bigg] , \tag{12}$$

where $\alpha \in (0, 1]$ and $\{\tau_i\}_{i=1}^B$ are generated with policy θ_t. Notice that if $\alpha = 1$, we recover the REINFORCE method, while if $\alpha = 0$ we recover the SRVRPG update. See Algorithm 3 in Section A of the Appendix.

IS-MBPG's update is identical to (12) but α is adjusted at every iteration (Huang et al., 2020).

4. PAGE-PG

PAGE (Li et al., 2021) is a novel variance-reduced stochastic gradient estimator for Problem (7), where f is differentiable but possibly non-convex. Let $\mathcal{B}_t$ be a set of randomly selected indices without replacement from $\{1, \ldots, n\}$ and $|\mathcal{B}_t| = B << n$, where the subscript t refers to the iteration. The PAGE estimator is based on a small adjustment to the mini-batch gradient estimator. Specifically, it is initialized to the full gradient $g_0 = \nabla_\theta f(\theta_0)$ at θ_0. For the subsequent iterations, the PAGE estimator is defined as follows

$$g_t = \begin{cases} \frac{1}{n}\sum\limits_{i=1}^{n} \nabla f_i(\theta_t) & \text{prob. } p_t \\ \frac{1}{B}\sum\limits_{i\in\mathcal{B}_t} \nabla f_i(\theta_t) + g_{t-1} - \frac{1}{B}\sum\limits_{i\in\mathcal{B}_t} \nabla f_i(\theta_{t-1}) & \text{prob. } 1-p_t. \end{cases} \tag{13}$$

This unbiased gradient estimator is used to update the parameter vector in a gradient-descent fashion

$$\theta_{t+1} = \theta_t - \eta g_t \, ,$$

where $\eta > 0$ is a fixed step-size. Therefore, PAGE is based on switching with probability p_t between gradient descent and a mini-batch version of SARAH (Nguyen et al., 2017).

As suggested by its name, PAGE-PG is designed by blending the key ideas of PAGE with policy gradient methods. As discussed in Section 3, we can not simply use the PAGE estimator for policy gradient in its original formulation but some adjustments are required. In particular, we substitute the exact gradient computations with an estimate based on a large batch-size $N >> B$. To deal with the distribution shift, we deploy importance weighting in a similar fashion as the variance-reduced policy gradient methods discussed in Section 3. PAGE-PG works by initially sampling N trajectories with the initial policy θ_0 and using those samples to build a solid gradient estimate $v_0 = \frac{1}{N}\sum_{i=1}^N g(\tau_i \,|\, \theta_0)$. For any $t > 0$, PAGE-PG deploys the following estimate

$$v_t = \begin{cases} \frac{1}{N}\sum\limits_{i=1}^{N} g(\tau_i \,|\, \theta_t) & \text{prob. } p_t \\ \frac{1}{B}\sum\limits_{i=1}^{B} g(\tau_i \,|\, \theta_t) + v_{t-1} - \frac{1}{B}\sum\limits_{i=1}^{B} g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1}) & \text{prob. } 1-p_t, \end{cases} \tag{14}$$

where τ_i is drawn according to policy θ_t for any i. The parameter vector is updated according to Equation (3), where v_t is deployed in place of the gradient. Notice that for $p_t = p = 1$ we recover the REINFORCE/GPOMDP method with batch-size N. See Algorithm 4 in Section A of the Appendix for a pseudo-code description. As it appears in Equation (14), the double loop-structure that characterizes SVRPG and SRVRPG is replaced by a probabilistic switch between two estimators. In particular, the probability of switching p_t plays an analogous role to the hyperparameter dictating the length of the inner loop in SVRPG and SRVRPG as it determines the frequency with which the gradient estimate based on a large batch is updated. Consequently, a smaller probability of switching tends to generate more noisy gradient estimates and viceversa. As discussed in Section 6, this could be exploited in some iteration-varying strategy that regulates the level of exploration by adjusting the hyperparameter p_t on the fly. One possibility could be to start with a small value of p_t and then gradually increase it as the training progresses. This can potentially prevent the iterates from getting stuck in some bad local maximizer by promoting exploration in the early stages of training, while producing more stable gradient estimates towards the end of training.

4.1. Theoretical Analysis

For the convergence analysis, we focus on the GPOMDP estimator, since it is generally preferred over the REINFORCE one because of its better performance. Therefore in this section we use $g(\tau_i \,|\, \theta) = \sum_{h=0}^{H-1} \gamma^h r(s_h^i, a_h^i) Z_{\theta,h}$ and $g^{\omega_{\theta_2}}(\tau_i \,|\theta_1) = \sum_{h=0}^{H-1} \omega_{0:h}(\tau_i \,|\, \theta_2, \theta_1)\gamma^h r(s_h^i, a_h^i) Z_{\theta_1,h}$. We also consider a constant probability of switching $p_t = p$ in order to simplify the analysis. We refer to Section C in the Appendix for the proofs and to Section B in the Appendix for the technical lemmas. After discussing the fundamental assumptions, we focus on studying the sample complexity of PAGE-PG to reach an ϵ-stationary solution. We further show that, when the objective is gradient-dominated, since ϵ-stationarity translates into near-optimality, the derived results are also valid for near-optimal solutions.

The theoretical analysis of variance-reduced policy gradient methods generally focuses on deriving, under certain assumptions, an upper bound on the number of sampled trajectories that are needed to achieve an ϵ-stationary solution.

Definition 4.1 (ϵ-stationary solution)**.** Let $\epsilon > 0$. $\theta \in \mathbb{R}^d$ is an ϵ-stationary solution if and only if $\|\nabla_\theta V(\theta)\| \le \epsilon$.

Based on Definition 4.1, a stochastic policy gradient based algorithm reaches an ϵ-stationary solution if and only if $\mathbb{E}\left[\|\nabla_\theta V(\theta_{\text{out}})\|^2\right] \le \epsilon^2$, where θ_{out} is the output of the algorithm after T iterations and the expected value is taken with respect to all the sources of randomness involved in the process. Our analysis is based on the following assumptions.

Assumption 4.2 (Bounded log-policy gradient norm)**.** For any $a \in \mathcal{A}$ and $s \in \mathcal{S}$ there exists a constant $G > 0$ such that $\|\nabla_\theta \log \pi_\theta(a \,|\, s)\| \le G$ for all $\theta \in \mathbb{R}^d$.

Assumption 4.3 (Smoothness)**.** π_θ is twice differentiable and for any $a \in \mathcal{A}$ and $s \in \mathcal{S}$ there exists a constant $M > 0$ such that $\|\nabla_\theta^2 \log \pi_\theta(a \,|\, s)\| \le M$ for all $\theta \in \mathbb{R}^d$.

Assumption 4.4 (Finite variance)**.** There exists a constant $\sigma > 0$ such that $\text{Var}(g(\tau \,|\, \theta)) \le \sigma^2$ for all $\theta \in \mathbb{R}^d$.

Assumption 4.5 (Finite importance weight variance)**.** For any policy pair $\theta_a, \theta_b \in \mathbb{R}^d$ and with $\tau \sim p(\cdot \,|\, \theta_b)$, the importance weight $\omega(\tau \,|\, \theta_b, \theta_a) = \frac{p(\tau \,|\, \theta_a)}{p(\tau \,|\, \theta_b)}$ is well-defined. In addition, there exists a constant $W > 0$ such that $\text{Var}(\omega(\tau \,|\theta_b, \theta_a\,)) \le W$.

The same set of assumptions is considered in (Papini et al., 2018; Xu et al., 2019; 2021; Yuan et al., 2020). By analyzing PAGE-PG in the same setting as its counterparts, we are able to compare them from a theoretical viewpoint.

Remark 4.6. While the non-convex optimization community agrees on the rationality of Assumptions 4.2-4.4, in (Zhang et al., 2021) the authors argue that Assumption 4.5 on the boundedness of the importance weight variance is uncheckable and very stringent. In a more limited setting (finite MDPs only) than the one considered in this work, they are able to remove such assumption via the introduction of a gradient-truncation mechanism that provably controls the variance of the importance weights in off-policy sampling. This approach is for now outside the scope of this work, but can be addressed in future work by adopting a trust region policy optimisation perspective. In practice, to ensure that Assumption 4.5 is met, one can resort to small step-sizes so that $p(\tau \,|\, \theta_b) \approx p(\tau \,|\, \theta_a)$ and the weight is bounded. This, however, comes at the cost of slower convergence, as also confirmed by our numerical experiments in Section 5.

For completeness, we report the following fundamental proposition from (Xu et al., 2021), which is used consistently in our proofs.

Proposition 4.7. *Let τ_i be a realization of $\tau \sim p(\cdot \,|\, \theta_1)$. Under Assumptions 4.2-4.3:*

1. $\|g(\tau_i \,|\, \theta_1) - g(\tau_i \,|\, \theta_2)\| \le L\|\theta_1 - \theta_2\|$ *for all* $\theta_1, \theta_2 \in \mathbb{R}^d$*, where* $L := MU/(1-\gamma)^2 + 2G^2U/(1-\gamma)^3$*,*
2. $V(\theta)$ *is* L*-smooth and twice differentiable, i.e.* $\|\nabla_\theta^2 V(\theta)\| \le L$*.*
3. $\|g(\tau_i \,|\, \theta)\| \le C_g$ *for all* $\theta \in \mathbb{R}^d$ *and* $C_g := GU/(1-\gamma)^2$*.*

Theorem 4.8. *Suppose that Assumptions 4.2-4.5 hold and select $\eta > 0$, $p \in (0, 1]$ and $B \in \mathbb{N}$ such that $\eta^2 \le \min\left\{p/(1-p) \cdot B/2C, 1/4L^2\right\}$. The average expected squared gradient norm after T iterations of PAGE-PG satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_\theta V(\theta_t)\|^2\right] \le \frac{2\,(V^* - V(\theta_0))}{\eta T} + \frac{\sigma^2}{N} + \frac{\sigma^2}{pNT}\,.$$

See Lemma B.1 in Section B of the Appendix for the definition of C. Theorem 4.8 states that under a proper choice of step-size, batch-size and probability of switching, the average expected squared gradient norm of the performance function after T iterations of PAGE-PG is in the order of $\mathcal{O}\left(\frac{1}{T} + \frac{1}{NT} + \frac{1}{N}\right)$. The first term $\mathcal{O}\left(\frac{1}{T}\right)$ characterizes the convergence of PAGE-PG, while the second and third terms come from the variance of the gradient estimator computed at the iterations with large batches. Our convergence rate improves over the rate $\mathcal{O}\left(\frac{1}{T} + \frac{1}{B} + \frac{1}{N}\right)$ of SVRPG (Papini et al., 2018) and over the rate $\mathcal{O}\left(\frac{1}{T} + \frac{1}{B} + \frac{1}{TN}\right)$ of STORM-PG (Yuan et al., 2020), by avoiding the dependency on the small batch-size B. Compared to the rate of SRVR-PG $\mathcal{O}\left(\frac{1}{T} + \frac{1}{N}\right)$, our analysis leads to an extra $\mathcal{O}\left(\frac{1}{TN}\right)$ term which arises from the variance of the first gradient estimator. By selecting $p = \frac{B}{N}$ and $B = \mathcal{O}\left(1\right)$, we recover the rate $\mathcal{O}\left(\frac{1}{T} + \frac{1}{N}\right)$.

Corollary 4.9. *Under the conditions of Theorem 4.8, set $\eta = \sqrt{B}/\sqrt{2CN}$, $p = 1/N$, $N = \mathcal{O}\left(\epsilon^{-2}\right)$ and*

$B = \mathcal{O}(1)$. *Then* $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla_\theta V(\theta_t)\right\|^2\right] \leq \epsilon^2$ *within* $\mathcal{O}\left(\epsilon^{-3}\right)$ *trajectories on average with* $\epsilon \to 0$.

Table 1. Sample complexities of comparable algorithms for finding an ϵ-stationary solution.

Method	Sample-Complexity	No-Restart
REINFORCE	$\mathcal{O}\left(\epsilon^{-4}\right)$	–
GPOMDP	$\mathcal{O}\left(\epsilon^{-4}\right)$	–
SVRPG	$\mathcal{O}\left(\epsilon^{-10/3}\right)$	✗
SRVRPG	$\mathcal{O}\left(\epsilon^{-3}\right)$	✗
STORM-PG	$\mathcal{O}\left(\epsilon^{-3}\right)$	✓
PAGE-PG	$\mathcal{O}\left(\epsilon^{-3}\right)$	✓

Under the considered assumptions, REINFORCE-type methods, including REINFORCE, GPOMDP as well as their variants with baselines, need $\mathcal{O}(\epsilon^{-4})$ samples to achieve an ϵ-stationary solution. By incorporating stochastic variance reduction techniques the complexity can be reduced. In particular, SVRPG achieves an $\mathcal{O}(\epsilon^{-10/3})$ sample complexity (Xu et al., 2019) while the more sophisticated SRVRPG (Xu et al., 2021) and STORM-PG (Yuan et al., 2020) achieve an $\mathcal{O}(\epsilon^{-3})$ sample complexity. According to Corollary 4.9, PAGE-PG needs on average $\mathcal{O}\left(\epsilon^{-3}\right)$ trajectories to achieve an ϵ-stationary solution, which makes it competitive from a theoretical viewpoint with its state-of-the-art counterparts. The discussed results on sample complexity are summarized in Table 1.

Recent works (Agarwal et al., 2019; Bhandari & Russo, 2021) have shown that, despite its non-concavity, with certain policy parametrizations, such as direct and softmax, the objective of Problem (2) is gradient dominated. We complete our theoretical analysis by extending the results of Theorem 4.8 to gradient-dominated objectives.

Assumption 4.10 (Gradient dominancy)**.** There exists a constant $\lambda > 0$ such that $V^* - V(\theta) \leq \lambda\left\|\nabla_\theta V(\theta)\right\|^2$ for all $\theta \in \mathbb{R}^d$.

Corollary 4.11. *Consider the same setting as in Theorem 4.8 where also Assumption 4.10 holds. Then,*

$$V^* - \max_{t \leq T} \mathbb{E}\left[V(\theta_t)\right] \leq \frac{2\lambda\left(V^* - V(\theta_0)\right)}{\eta T} + \frac{\sigma^2\lambda}{N} + \frac{\sigma^2\lambda}{pNT}.$$

The gradient dominancy condition implies that any stationary policy is also globally optimal. Consequently, as formalized in Corollary 4.11, the results from Theorem 4.8 are valid for near-optimal policies, with the only difference that in this case the upper bound on the suboptimality is also proportional to the gradient dominancy constant.

5. Numerical Evaluation

In this section we numerically evaluate the performance of the discussed variance-reduced policy gradient methods on two state-of-the-art model-free reinforcement learning tasks from OpenAI Gym (Brockman et al., 2016). In order to conduct the numerical evaluation of the discussed methods, we implemented them, along with GPOMDP, in a Pytorch-based toolbox. In addition, the toolbox interfaces OpenAI Gym (Brockman et al., 2016) allowing the user to easily train RL agents on different environments with the discussed methods. Finally, Pytorch (Paszke et al., 2019) provides the possibility of speeding up the computation via the deployment of graphical processing units (GPUs). The toolbox is publicly available at `https://gitlab.ethz.ch/gmatilde/vr_reinforce`.

5.1. Benchmarks

For the empirical evaluation of the discussed methods we consider the Acrobot and the Cartpole environments from OpenAI Gym.

Acrobot. The Acrobot system comprises two joints and two links, where the joint between the two links is actuated. Initially, the links are hanging downwards, and the goal is to swing the end of the lower link up to a given height. A reward of -1 is emitted every time the goal is not achieved. As soon as the target height is reached or 500 time-steps are elapsed, the episode ends. The state space is continuous with dimension 6. The action space is discrete and 3 possible actions can be selected: apply a positive torque, apply a negative torque, do nothing. To model the policy, we use a neural softmax parametrization. In particular, we deploy a neural network with two hidden layers, width 32 for both layers and Tanh as activation function.

Cartpole. The Cartpole system is a classical control environment that comprises a pole attached by an un-actuated joint to a cart that moves along a frictionless track. The pendulum starts upright, and the goal is to prevent it from falling over. A reward of $+1$ is provided for every time-step that the pole remains within 15 degrees from the upright position. The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from its initial position. The state space is continuous with dimension 4. The action space is discrete with 2 available actions: apply a force of $+1$ or -1 to the cart. As for the Acrobot, to model the policy we use a neural softmax parametrization. In particular, we deploy a neural network with two hidden layers, width 32 for both layers and Tanh as activation function. The maximum episode length is set to 200.

Unfortunately the hyperparameter settings of Theorem 4.8 and Corollary 4.9, as well as those indicated by the theoretical analysis of the other considered methods, are functions of problem-dependent constants which are typically unknown and/or very expensive to estimate. Therefore for our benchmarks we set $N = 100$, $B = 5$ and $m = 10$ and $\gamma = 0.9999$ and we rely on grid-search for tuning the step-size and the other hyperparameters. See Section D in the Appendix for more details on the choice of the hyperparameters. Notice in addition that, since the hyperparameter configurations in Theorem 4.8 and Corollary 4.9 are sufficient but not necessary requirements, there could be potentially different configurations that lead to similar results. For each algorithm, we run the experiment 5 times with random initialization of the environments. The curves (solid-lines) are obtained by taking the mean over the independent runs and the shaded areas represent the $\pm\sigma$ standard deviations.

The experiments in Figures 1 and 2 show that, given enough episodes, all of the algorithms are able to solve the tasks, achieving near-optimal returns. For the Acrobot environment in Figure 1, the SRVRPG and GPOMDP algorithms take the biggest number of episodes to find an optimal policy, while STORM-PG and SVRPG are the fastest in terms of number of episodes. This might be due to the step-size, which, for certain methods, needs to be set to particularly small values to enforce finite importance weight variance and ensure convergence. For the Cartpole environment in Figure 2, as expected, the GPOMDP algorithm takes the longest to find the optimal policy, followed in order by SVRPG, SRVRPG, STORM-PG and PAGE-PG. Notice that these empirical observations corroborate the theoretical findings on the sample complexity. Finally, our benchmarks demonstrate the competitive performance of PAGE-PG with respect to its counterparts.

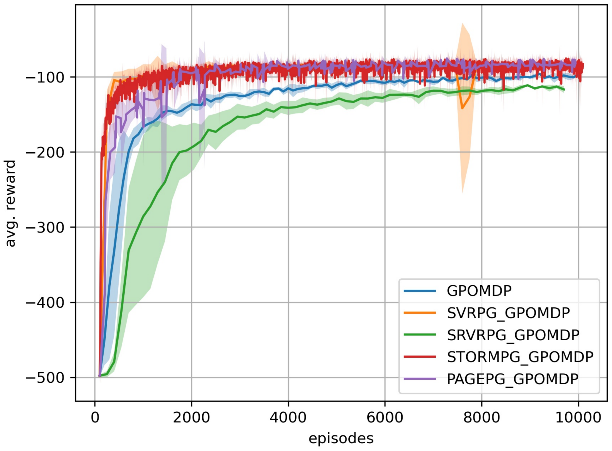

Figure 1. Average reward versus number of episodes for GPOMDP (blue), SVRPG (orange), SRVRPG (green), STORM-PG (red) and PAGE-PG (light purple) on the Acrobot environment. The solid line represents the mean and the shaded areas are calculated as the $\pm\sigma$ of the outcomes over 5 independent runs.

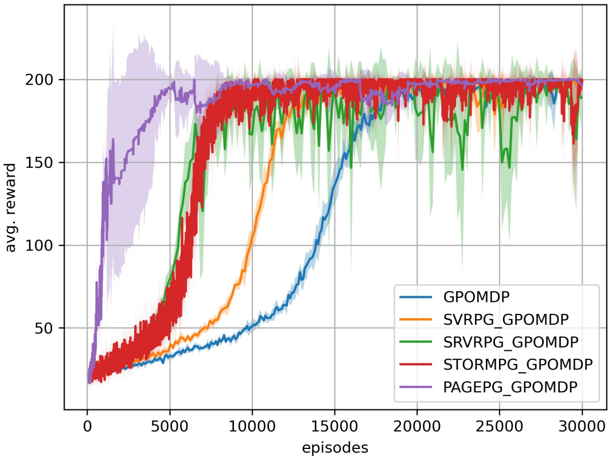

Figure 2. Average reward versus number of episodes for GPOMDP (blue), SVRPG (orange), SRVRPG (green), STORM-PG (red) and PAGE-PG (light purple) on the Cartpole environment. The solid line represents the mean and the shaded areas are calculated as the $\pm\sigma$ of the outcomes over 5 independent runs.

6. Conclusions, Limitations & Future Works

After a brief survey on the main variance reduced policy gradient methods based on REINFORCE-type algorithms, we formulate a novel variance-reduced extension, PAGE-PG, inspired from the PAGE gradient estimator for optimization of non-convex finite-sum problems. To the best of the authors' knowledge, our method is the first variance-reduced policy gradient method that replaces the outer loop with a probabilistic switch. This key feature of PAGE-PG facilitates the theoretical analysis while preserving a fast theoretical rate and a low sample complexity. In addition, our numerical evaluation shows that PAGE-PG has a competitive performance with respect to its counterparts.

Our benchmarks and theoretical results on the sample complexity confirm that variance-reduced techniques successfully manage to reduce the sample complexity of REINFORCE-type algorithms, speeding up the convergence in terms of number of sampled trajectories. At the same time, it is possible to identify the following limitations:
Unrealistic and uncheckable assumption on importance weight variance. As underlined in Remark 4.6, all the discussed variance-reduced policy gradient methods heavily rely on the stringent and uncheckable assumption that the importance weights have bounded variance for every iteration of the algorithms (Assumption 4.5). To enforce indirectly this assumption, very small values of the step-size are needed, resulting in a dramatic slow-down of the convergence rate. A more efficient alternative could be the deployment of a gradient-truncation strategy, as proposed in (Zhang et al., 2021) for the case of finite MDPs. This modification, which corresponds to the solution of a trust-region

subproblem, is simple and efficient since it does not involve significant extra computational costs but, at the same time, requires to migrate from vanilla REINFORCE-type methods to trust-region based algorithms such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017) for comparisons.

Extreme sensitivity to hyperparameters. Our benchmarks suggest an extreme sensitivity to the hyperparameters, especially the choice of the step-size. Time-consuming and resource-expensive tuning procedures are required to select a proper configuration of hyperparameters. To alleviate this issue, the update direction should be computed also taking into account second-order information, as it is done in (Huang et al., 2020) for HA-MBPG. Second-order methods are notably more robust against the step-size selection than first-order methods, since their update includes information on the local curvature (Agarwal et al., 2019; Gargiani et al., 2020).

Noise annealing strategies. Empirical evidence suggests that, in the presence of complex non-concave landscapes, exploration in the form of noise injection is of critical importance in the early stages of training to prevent convergence to spurious local maximizers (Chung et al., 2021). Entropy regularization is often used to improve exploration, since it indirectly injects noise in the training process by favoring the selection of more stochastic policies (Ahmed et al., 2019). Unfortunately, by adding a regularizer to Problem (2) we are effectively changing the optimal policy. An alternative approach would be increasing the batch-size during training. In this perspective, a promising heuristic to further improve the convergence of PAGE-PG could consist in gradually increasing the probability of switching p_t. Finally, as also pointed out in (Papini et al., 2018), since the variance of the updates depends on the snapshot policy as well as on the sampled trajectories, it is realistic to imagine that predefined schemes for the probability of switching are not going to perform as well as adaptive ones, which adjust the value of p_t based on some measure of the variance.

We leave for future development the aforementioned extensions, which we believe would counteract the current limitations of the analyzed methods.

Acknowledgements

This work has been supported by the European Research Council (ERC) under the H2020 Advanced Grant no. 787845 (OCAL). The authors thank Niao He and Goran Banjac for the insightful discussions.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Gaurav, M. On the theory of policy gradient methods: optimality, approximation, and distribution shift. *arXiv:1908.00261*, 2019.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 151–160. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ahmed19a.html.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pp. 1200–1205, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055448. URL https://doi.org/10.1145/3055399.3055448.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, nov 2001. ISSN 1076-9757.

Bhandari, J. and Russo, D. On the linear convergence of policy gradient methods for finite MDPs. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2386–2394. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bhandari21a.html.

Bottou, L. Stochastic gradient descent tricks. 2012. URL https://cilvr.cs.nyu.edu/diglib/lsml/bottou-sgd-tricks-2012.pdf.

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

Chung, W., Thomas, V., Machado, M. C., and Roux, N. L. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1999–2009. PMLR, 2021.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.

neurips.cc/paper/2019/file/b8002139cdde66b87638f7f91d169d96-Paper.pdf.

Gargiani, M., Zanelli, A., Diehl, M., and Hutter, F. On the promise of the stochastic generalized Gauss-Newton method for training DNNs. *arXiv preprint arXiv:2006.02409v4*, 2020.

Huang, F., Gao, S., Pei, J., and Huang, H. Momentum-based policy gradient methods. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4422–4433. PMLR, 13–18 Jul 2020.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. *arXiv preprint arXiv:1608.04636v4*, 2020.

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In Kontorovich, A. and Neu, G. (eds.), *Proceedings of Machine Learning Research*, volume 117 of *31st International Conference on Algorithmic Learning Theory*, pp. 1–17, 2020.

Levine, S. and Koltun, V. Learning complex neural network policies with trajectory optimization. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 829–837, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/levine14.html.

Li, Z., Bao, H., Zhang, X., and Richtarik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. *Proceeding of the 38-th International Conference on Machine Learning*, 2021.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/mniha16.html.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/nguyen17b.html.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4026–4035. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/papini18a.html.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347v2*, 2017.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. SODA '18, pp. 770–787, USA, 2018.

Society for Industrial and Applied Mathematics. ISBN 9781611975031.

Smith, S. L., Kindermans, P., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=B1Yy1BxCZ.

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

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv:1905.12615v1*, 2019.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv:1909.08610v3*, 2021.

Yuan, H., Lian, X., Liu, J., and Zhou, Y. Stochastic recursive momentum for policy gradient methods. *arXiv:2003.04302v1*, 2020.

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., and Wang, M. On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv:2102.08607*, 2021.

Zhao, T., Hachiya, H., Niu, G., and Sugiyama, M. Analysis and improvement of policy gradient estimation. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/85d8ce590ad8981ca2c8286f79f59954-Paper.pdf.

Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E., and Zhao, T. Towards understanding the importance of noise in training neural networks. *arXiv preprint arXiv:1909.03172*, 2019.

A. Algorithmic Description of Variance-Reduced REINFORCE-type Methods

Algorithm 1 SVRPG

```
  Input: initial parameter $\theta_0$, large batch-size $N$, small batch-size $B$, step-size $\eta > 0$, inner-loop length $m \in \mathbb{N}$, number of epochs $S \in \mathbb{N}$
  initialize $\theta_m^0 = \theta_0$
  for $s = 1$ to $S$ do
    set $\theta_0^s = \theta_m^{s-1}$
    collect $N$ trajectories with policy $\theta_0^s$
    $v_0^s = \frac{1}{N} \sum_{i=1}^{N} g(\tau_i \,|\, \theta_0^s)$
    for $t = 0$ to $m - 1$ do
      $\theta_{t+1}^s = \theta_t^s + \eta v_t^s$
      sample $B$ trajectories with policy $\theta_{t+1}^s$
      $v_{t+1}^s = \frac{1}{B} \sum_{i=1}^{B} g(\tau_i \,|\, \theta_{t+1}^s) + v_0^s - \frac{1}{B} \sum_{i=1}^{B} g^{\omega_{\theta^s_{t+1}}}(\tau_i \,|\, \theta_0^s)$
    end for
  end for
  Output: $\theta_{\text{out}}$ chosen uniformly at random from $\{\theta_t^s\}_{t=1,\, s=1}^{m,\, S}$
```

Algorithm 2 SRVRPG

```
  Input: initial parameter $\theta_0$, large batch-size $N$, small batch-size $B$, step-size $\eta > 0$, inner-loop length $m \in \mathbb{N}$, number of epochs $S \in \mathbb{N}$
  initialize $\theta_m^0 = \theta_0$
  for $s = 1$ to $S$ do
    set $\theta_0^s = \theta_m^{s-1}$
    collect $N$ trajectories with policy $\theta_0^s$
    $v_0^s = \frac{1}{N} \sum_{i=1}^{N} g(\tau_i \,|\, \theta_0^s)$
    for $t = 0$ to $m - 1$ do
      $\theta_{t+1}^s = \theta_t^s + \eta v_t^s$
      sample $B$ trajectories with policy $\theta_{t+1}^s$
      $v_{t+1}^s = \frac{1}{B} \sum_{i=1}^{B} g(\tau_i \,|\, \theta_{t+1}^s) + v_t^s - \frac{1}{B} \sum_{i=1}^{B} g^{\omega_{\theta^s_{t+1}}}(\tau_i \,|\, \theta_t^s)$
    end for
  end for
  Output: $\theta_{\text{out}}$ chosen uniformly at random from $\{\theta_t^s\}_{t=1,\, s=1}^{m,\, S}$
```

Algorithm 3 STORM-PG

```
  Input: initial parameter $\theta_0$, large batch-size $N$, small batch-size $B$, step-size $\eta > 0$, momentum parameter $\alpha \in (0, 1]$
  collect $N$ trajectories with policy $\theta_0$
  $v_0 = \frac{1}{N} \sum_{i=1}^{N} g(\tau_i \,|\, \theta_0)$
  for $t = 0$ to $T - 1$ do
    $\theta_{t+1} = \theta_t + \eta v_t$
    collect $B$ trajectories with policy $\theta_{t+1}$
    $v_{t+1} = \frac{1}{B} g(\tau_i \,|\, \theta_{t+1}) + (1 - \alpha) \left[ v_t - \frac{1}{B} \sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i \,|\, \theta_t) \right]$
  end for
  Output: $\theta_{\text{out}}$ chosen uniformly at random from $\{\theta_t\}_{t=1}^{T}$
```

Algorithm 4 PAGE-PG

```
**Input:** initial parameter $\theta_0$, large batch-size $N$, small batch-size $B$, step-size $\eta > 0$, probability $p \in (0, 1]$
collect $N$ trajectories with policy $\theta_0$
$v_0 = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_0)$
**for** $t = 0$ **to** $T - 1$ **do**
    $\theta_{t+1} = \theta_t + \eta v_t$
    $v_{t+1} = \begin{cases} \frac{1}{N}\sum\limits_{i=1}^{N} g(\tau_i \,|\, \theta_t) & \text{prob. } p \\ \frac{1}{B}\sum\limits_{i=1}^{B} g(\tau_i \,|\, \theta_t) + v_{t-1} - \frac{1}{B}\sum\limits_{i=1}^{B} g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1}) & \text{prob. } 1-p \end{cases}$
**end for**
**Output:** $\theta_{\text{out}}$ chosen uniformly at random from $\{\theta_t\}_{t=1}^{T}$
```

B. Technical Lemmas

Lemma B.1. *Let θ_t and θ_{t+1} denote two consecutive iterates of PAGE-PG and let $g(\tau \,|\, \theta_{t+1})$ and $g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)$ denote the on-policy and off-policy GPOMDP estimates computed at the iterates θ_{t+1} and θ_t respectively, and where $\tau \sim p(\cdot \,|\, \theta_{t+1})$. Then,*

$$\mathbb{E}\left[\left\|g(\tau \,|\, \theta_{t+1}) - g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)\right\|^2\right] \le C \cdot \mathbb{E}\left[\left\|\theta_{t+1} - \theta_t\right\|^2\right] ,$$

where $C := 2(L^2 + C_\omega)$, $L := MU/(1-\gamma)^2 + 2G^2U/(1-\gamma)^3$ and $C_\omega := 24UG^2(2G^2 + M)(W+1)\gamma/(1-\gamma)^5$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\|g(\tau \,|\, \theta_{t+1}) - g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)\|^2\right] &= \mathbb{E}\left[\|g(\tau \,|\, \theta_{t+1}) - g(\tau \,|\, \theta_t) + g(\tau \,|\, \theta_t) - g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)\|^2\right] \\
&\overset{(a)}{\le} 2\mathbb{E}\left[\|g(\tau \,|\, \theta_{t+1}) - g(\tau \,|\, \theta_t)\|^2\right] + 2\mathbb{E}\left[\|g(\tau \,|\, \theta_t) - g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)\|^2\right] \\
&\overset{(b)}{\le} 2L^2\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] + 2\mathbb{E}\left[\|g(\tau \,|\, \theta_t) - g^{\omega_{\theta_{t+1}}}(\tau \,|\, \theta_t)\|^2\right] \\
&\overset{(c)}{\le} 2L^2\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] + 2C_\omega\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] \\
&= 2(L^2 + C_\omega)\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] ,
\end{aligned}$$

where Inequality (a) follows from the fact that, given any arbitrary triplet of vectors (x, y, z), then $\|x - z + z - y\|^2 = \|x-z\|^2 + \|z-y\|^2 + 2\langle x-z, z-y\rangle = \|x-z\|^2 + \|z-y\|^2 + \|x-z\|^2 + \|z-y\|^2 - \|x-2z+y\|^2 \le 2\|x-z\|^2 + 2\|z-y\|^2$. Inequality (b) is derived by considering the first point of Proposition 4.7 (see Proposition 4.2 in (Xu et al., 2021) for a detailed proof). Inequality (c) is obtained by considering Inequality (B.9) in (Xu et al., 2021). □

Lemma B.2. *Let v_t and θ_{t+1} denote the gradient estimate and the iterate generated by PAGE-PG at iteration $t + 1$, respectively. Under Assumptions 4.2-4.5, the estimation error at iteration $t + 1$ can be bounded as follows*

$$\mathbb{E}\left[\|v_{t+1} - \nabla_\theta V(\theta_{t+1})\|^2\right] \le (1-p)\mathbb{E}\left[\left\|v_t - \nabla_\theta V(\theta_t)\right\|^2\right] + \frac{\eta^2(1-p)C}{B}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{p\sigma^2}{N} ,$$

where the expectation is taken with respect to all the sources of randomness up to iteration $t + 1$.

Proof. Let $\mathcal{F}_t$ denote the information up to iteration t. From the law of iterated expectations, we know that

$\mathbb{E}\left[\|v_{t+1}-\nabla_\theta V(\theta_{t+1})\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|v_{t+1}-\nabla_\theta V(\theta_{t+1})\|^2 \,\middle|\, \mathcal{F}_t\right]\right]$. We start by analysing the inner expectation

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{t+1}-\nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] = \\
&= (1-p)\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) + v_t - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\;\;+\; p\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&= (1-p)\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) + v_t - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) - \nabla_\theta V(\theta_t) + \nabla_\theta V(\theta_t) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\;\;+ p\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&= (1-p)\mathbb{E}\left[\left\|v_t - \nabla_\theta V(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + 2(1-p)\mathbb{E}\left[\left\langle v_t - \nabla_\theta V(\theta_t),\, \frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) \,\middle|\, \mathcal{F}_t\right] - \nabla_\theta V(\theta_{t+1})\right\rangle \,\middle|\, \mathcal{F}_t\right] \\
&\;\;+\; 2(1-p)\left\langle v_t - \nabla_\theta V(\theta_t),\, -\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) \,\middle|\, \mathcal{F}_t\right] + \nabla_\theta V(\theta_t)\right\rangle \\
&\;\;+\; (1-p)\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) + \nabla_\theta V(\theta_t) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\;\;+\; p\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\overset{(a)}{=} (1-p)\mathbb{E}\left[\left\|v_t - \nabla_\theta V(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + 2(1-p)\left\langle v_t - \nabla_\theta V(\theta_t),\, -\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) \,\middle|\, \mathcal{F}_t\right] + \nabla_\theta V(\theta_t)\right\rangle \\
&\;\;+\; (1-p)\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) + \nabla_\theta V(\theta_t) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\;\;+\; p\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\overset{(b)}{\leq} (1-p)\mathbb{E}\left[\left\|v_t - \nabla_\theta V(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + 2(1-p)\left\langle v_t - \nabla_\theta V(\theta_t),\, -\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) \,\middle|\, \mathcal{F}_t\right] + \nabla_\theta V(\theta_t)\right\rangle \\
&\;\;+\; (1-p)\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i\,|\,\theta_{t+1}) - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + p\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
&\overset{(c)}{\leq} (1-p)\mathbb{E}\left[\left\|v_t - \nabla_\theta V(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + 2(1-p)\left\langle v_t - \nabla_\theta V(\theta_t),\, -\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t) \,\middle|\, \mathcal{F}_t\right] + \nabla_\theta V(\theta_t)\right\rangle \\
&\;\;+\; \frac{(1-p)}{B^2}\mathbb{E}\left[\sum_{i=1}^{B}\left\|g(\tau_i\,|\,\theta_{t+1}) - g^{\omega_{\theta_{t+1}}}(\tau_i\,|\,\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] + \frac{p}{N^2}\,\mathbb{E}\left[\sum_{i=1}^{N}\left\|g(\tau_i\,|\,\theta_{t+1}) - \nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right],
\end{aligned}
$$

where Equality (a) follows from the fact that $\mathbb{E}_{\tau\sim p(\cdot\,|\,\theta)}\left[g(\tau\,|\,\theta)\right] = \nabla_\theta V(\theta)$. Inequality (b) is obtained considering that, for any random vector X, the variance can be bounded as follows $\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$ (see Lemma B.5 in (Papini et al., 2018) for a detailed proof) and Inequality (c) is obtained by exploiting the triangle inequality. By combining the derived upper bound with the results from Lemma B.1 and exploiting Assumption 4.4, we derive the following bound

$$
\begin{aligned}
\mathbb{E}\left[\left\|v_{t+1}-\nabla_\theta V(\theta_{t+1})\right\|^2 \,\middle|\, \mathcal{F}_t\right] \leq & (1-p)\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\
& + \ 2(1-p)\left\langle v_t-\nabla_\theta V(\theta_t),\, -\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i \,|\, \theta_t)\right] + \nabla_\theta V(\theta_t)\right\rangle \\
& + \ \frac{(1-p)C}{B}\mathbb{E}\left[\left\|\theta_{t+1}-\theta_t\right\|^2 \,\middle|\, \mathcal{F}_t\right] + \frac{p\sigma^2}{N}\,.
\end{aligned}
$$

By considering the full expectation on the derived results, we finally obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|v_{t+1}-\nabla_\theta V(\theta_{t+1})\right\|^2\right] \leq \ & (1-p)\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] \\
& + \ 2(1-p)\mathbb{E}\left[\left\langle v_t-\nabla_\theta V(\theta_t),\, -\frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i \,|\, \theta_t) + \nabla_\theta V(\theta_t)\right\rangle\right] \\
& + \ \frac{(1-p)C}{B}\mathbb{E}\left[\left\|\theta_{t+1}-\theta_t\right\|^2\right] + \frac{p\sigma^2}{N} \\
= \ & (1-p)\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] \\
& + \ 2(1-p)\mathbb{E}\left[\mathbb{E}\left[\left\langle v_t-\nabla_\theta V(\theta_t),\, -\frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_{t+1}}}(\tau_i \,|\, \theta_t) + \nabla_\theta V(\theta_t)\right\rangle \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\
& + \ \frac{(1-p)C}{B}\mathbb{E}\left[\left\|\theta_{t+1}-\theta_t\right\|^2\right] + \frac{p\sigma^2}{N} \\
= \ & (1-p)\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] + \frac{(1-p)C}{B}\mathbb{E}\left[\left\|\theta_{t+1}-\theta_t\right\|^2\right] + \frac{p\sigma^2}{N} \\
= \ & (1-p)\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] + \frac{\eta^2(1-p)C}{B}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{p\sigma^2}{N}
\end{aligned}
$$

where the second last equality is derived considering that $\mathbb{E}_{\tau\sim p(\cdot\,|\,\theta_2)}\left[g^{\omega_{\theta_2}}(\tau\,|\,\theta_1)\right] = \nabla_\theta V(\theta_1)$ for any $\theta_1,\, \theta_2$, and the last equality is obtained by considering that $\theta_{t+1} = \theta_t + \eta v_t$. □

Lemma B.3. *Let Assumptions 4.2-4.5 hold and let v_t and θ_{t+1} denote the gradient estimator and the iterate generated by PAGE-PG at iteration $t+1$, respectively. The accumulated sum of the expected estimation error satisfies the following inequality*

$$
\sum_{t=0}^{T-1}\mathbb{E}\left[\|v_t-\nabla_\theta V(\theta_t)\|^2\right] \leq \frac{\eta^2(1-p)C}{pB}\sum_{t=0}^{T-1}\mathbb{E}\left[\|v_t\|^2\right] + \frac{T\sigma^2}{N} + \frac{\sigma^2}{pN}\,.
$$

Proof. Recall that $p \in (0,1]$ is the probability that regulates the probabilistic switching. Then,

$$
\begin{aligned}
p\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] - (1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] - (1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] \\
&\quad + \mathbb{E}\left[\left\|v_0-\nabla_\theta V(\theta_0)\right\|^2\right] - \mathbb{E}\left[\left\|v_T-\nabla_\theta V(\theta_T)\right\|^2\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] - (1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right] + \mathbb{E}\left[\left\|v_0-\nabla_\theta V(\theta_0)\right\|^2\right]\,.
\end{aligned} \tag{15}
$$

Summing up over T iterations the result from Lemma B.2, we obtain the following inequality

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right]\leq(1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right]+\frac{\eta^2(1-p)C}{B}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right]+\frac{Tp\sigma^2}{N}\,. \tag{16}$$

By combining Inequality (16) with Inequality (15), we obtain the final result

$$\begin{aligned}p\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right]&\leq(1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right]+\frac{\eta^2(1-p)C}{B}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right]+\frac{Tp\sigma^2}{N}\\&\quad-(1-p)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t-\nabla_\theta V(\theta_t)\right\|^2\right]+\mathbb{E}\left[\left\|v_0-\nabla_\theta V(\theta_0)\right\|^2\right]\\&=\frac{\eta^2(1-p)C}{B}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right]+\frac{Tp\sigma^2}{N}+\mathbb{E}\left[\left\|v_0-\nabla_\theta V(\theta_0)\right\|^2\right]\\&\leq\frac{\eta^2(1-p)C}{B}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right]+\frac{Tp\sigma^2}{N}+\frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\left\|g(\tau_i\,|\,\theta_0)-\nabla_\theta V(\theta_0)\right\|^2\right]\\&\leq\frac{\eta^2(1-p)C}{B}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right]+\frac{Tp\sigma^2}{N}+\frac{\sigma^2}{N}\,,\end{aligned} \tag{17}$$

where the second last inequality is derived by deploying the definition of v_0 and the triangle inequality, while the last inequality is obtained by considering Assumption 4.4.

□

C. Proof of the Main Theoretical Results

Proof of Theorem 4.8. In the considered setting we know from Proposition 4.7 that $V(\theta)$ is L-smooth, where $L:=MU/(1-\gamma)^2+2G^2U/(1-\gamma)^3$. Consequently, we can write the following lower bound on $V(\theta_{t+1})$

$$\begin{aligned}V(\theta_{t+1})&\geq V(\theta_t)+\left\langle\nabla V(\theta_t),\theta_{t+1}-\theta_t\right\rangle-\frac{L}{2}\|\theta_{t+1}-\theta_t\|^2\\&=V(\theta_t)+\eta\left\langle\nabla V(\theta_t),v_t\right\rangle-\frac{\eta^2L}{2}\left\|v_t\right\|^2\\&\overset{(a)}{=}V(\theta_t)+\frac{\eta}{2}\left\|\nabla V(\theta_t)\right\|^2+\frac{\eta}{2}\left\|v_t\right\|^2-\frac{\eta}{2}\left\|v_t-V(\theta_t)\right\|^2-\frac{\eta^2L}{2}\left\|v_t\right\|^2\\&=V(\theta_t)+\frac{\eta}{2}\left\|\nabla V(\theta_t)\right\|^2+\frac{\eta}{2}\left(1-\eta L\right)\left\|v_t\right\|^2-\frac{\eta}{2}\left\|v_t-V(\theta_t)\right\|^2,\end{aligned}$$

where Equality (a) is derived by considering that $\langle x,y\rangle=\frac{\|x\|^2}{2}+\frac{\|y\|^2}{2}-\frac{\|y-x\|^2}{2}$. Since by design choice $\eta\leq1/(2L)$, then

$$V(\theta_{t+1})\geq V(\theta_t)+\frac{\eta}{2}\left\|\nabla V(\theta_t)\right\|^2+\frac{\eta}{4}\left\|v_t\right\|^2-\frac{\eta}{2}\left\|v_t-V(\theta_t)\right\|^2.$$

By rearranging some terms and changing the direction of the inequality, we obtain

$$V(\theta_t)-V(\theta_{t+1})\leq-\frac{\eta}{2}\left\|\nabla V(\theta_t)\right\|^2-\frac{\eta}{4}\left\|v_t\right\|^2+\frac{\eta}{2}\left\|v_t-V(\theta_t)\right\|^2. \tag{18}$$

Summing up over the first T iterations, we obtain

$$\sum_{t=0}^{T-1}\left(V(\theta_t)-V(\theta_{t+1})\right)=V(\theta_0)-V(\theta_T)\leq-\frac{\eta}{2}\sum_{t=0}^{T-1}\left\|\nabla V(\theta_t)\right\|^2-\frac{\eta}{4}\sum_{t=0}^{T-1}\left\|v_t\right\|^2+\frac{\eta}{2}\sum_{t=0}^{T-1}\left\|v_t-V(\theta_t)\right\|^2. \tag{19}$$

By taking the expectation on both sides of Equation (19) and considering the fact that $V^* \geq V(\theta)$ for all $\theta \in \mathbb{R}^d$, we get

$$\begin{aligned}
V(\theta_0) - V^* &\leq V(\theta_0) - \mathbb{E}\left[V(\theta_T)\right] \\
&\leq -\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right] - \frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t - V(\theta_t)\right\|^2\right] \\
&\overset{(a)}{\leq} -\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right] - \frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{\eta^3(1-p)C}{2pB}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{\eta T\sigma^2}{2N} + \frac{\eta\sigma^2}{2pN} \\
&= -\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right] - \frac{\eta}{2}\left(\frac{1}{2} - \frac{\eta^2(1-p)C}{pB}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|v_t\right\|^2\right] + \frac{\eta T\sigma^2}{2N} + \frac{\eta\sigma^2}{2pN} \\
&\overset{(b)}{\leq} -\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right] + \frac{\eta T\sigma^2}{2N} + \frac{\eta\sigma^2}{2pN}\,,
\end{aligned} \tag{20}$$

where Inequality (a) follows from Lemma B.3 and Inequality (b) follows from the fact that $\frac{\eta^2(1-p)}{pB} \leq 1/(2C)$. Finally, rearranging the terms and multiplying both sides by $\frac{2}{\eta T}$, we obtain the final result

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right] \leq \frac{2\left(V^* - V(\theta_0)\right)}{\eta T} + \frac{\sigma^2}{N} + \frac{\sigma^2}{pNT}\,.$$

□

Proof of Corollary 4.9. Let $\Delta = V^* - V(\theta_0)$. We set $p = \frac{1}{N}$ and $\eta = \frac{\sqrt{B}}{\sqrt{2CN}}$. Notice that with this choice of parameters, we verify the constraints on the parameter selection, i.e. $\eta^2 \leq \min\left\{\frac{Bp}{2C(1-p)}\,, \frac{1}{4L^2}\right\}$. In particular, with this choice for the probability of switching and the step-size, we get that $\eta^2 = \frac{B}{2CN} = \frac{Bp}{2C} \leq \frac{Bp}{2C(1-p)}$. In addition, since $\frac{B}{N} \leq 1$ and $C = 2(C_\omega + L^2)$, then $\eta^2 \leq \frac{1}{4L^2}$.

We now want to derive some values of T and N such that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla_\theta V(\theta_T)\right\|^2\right] \leq \epsilon^2\,. \tag{21}$$

For that, we utilize the results of Theorem 4.8 as follows

$$\frac{2\Delta}{\eta T} + \frac{\sigma^2}{N} + \frac{\sigma^2}{pNT} \leq \epsilon^2\,. \tag{22}$$

We decide to partition ϵ^2 equally among the three terms in (22)

$$\frac{2\Delta}{\eta T} \leq \frac{\epsilon^2}{3}\,, \quad \frac{\sigma^2}{N} \leq \frac{\epsilon^2}{3}\,, \quad \frac{\sigma^2}{pNT} \leq \frac{\epsilon^2}{3}\,, \tag{23}$$

from which we infer that, with these specific choices, Inequality (21) is verified for all values of N and T such that

$$N \geq \frac{3\sigma^2}{\epsilon^2}\,, \qquad\qquad T \geq \max\left\{\frac{6\Delta}{\epsilon^2}\frac{\sqrt{2CN}}{\sqrt{B}}\,, \frac{3\sigma^2}{\epsilon^2}\right\}\,. \tag{24}$$

For the sake of compactness, we define $K_1 = 6\Delta\sqrt{2C}$ and $K_2 = 3\sigma^2$. The average number of trajectories over T iterations is given by the following expression

$$pTN + (1-p)TB = T\left(pN + (1-p)B\right)\,. \tag{25}$$

We want to study the average sample complexity of PAGE-PG to reach an ϵ-stationary solution when $\epsilon \to 0$. To do that, we set $T = \frac{K_1}{\epsilon^2}\frac{\sqrt{N}}{\sqrt{B}} + \frac{K_2}{\epsilon^2}$, such that the constraints on T from Equation (24) are verified. Finally, by plugging this value for the number of iterations and the choice of p in Equation (25), we get

$$\begin{aligned} T\left(pN + (1-p)B\right) &= T\left(1 + B - \frac{B}{N}\right) \\ &= \left(\frac{K_1}{\epsilon^2}\frac{\sqrt{N}}{\sqrt{B}} + \frac{K_2}{\epsilon^2}\right)\left(1 + B - \frac{B}{N}\right) . \end{aligned} \tag{26}$$

By setting the batch-size parameters to $B = \mathcal{O}\left(1\right)$ and $N = \mathcal{O}\left(\epsilon^{-2}\right)$, we finally get

$$\begin{aligned} \left(\frac{K_1}{\epsilon^2}\frac{\sqrt{N}}{\sqrt{B}} + \frac{K_2}{\epsilon^2}\right)\left(1 + B - \frac{B}{N}\right) &= \mathcal{O}\left(\epsilon^{-3}\right)\mathcal{O}\left(1\right) \\ &= \mathcal{O}\left(\epsilon^{-3}\right) . \end{aligned} \tag{27}$$

□

Proof of Corollary 4.11. We combine the gradient dominancy condition and the results from Theorem 4.8 as follows

$$\begin{aligned} V^* - \max_{t \leq T} \mathbb{E}\left[V(\theta_t)\right] &\leq V^* - \mathbb{E}\left[V(\theta_a)\right] \\ &\leq \lambda \mathbb{E}\left[\left\|\nabla_\theta V(\theta_a)\right\|^2\right] \\ &\leq \frac{\lambda}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla_\theta V(\theta_t)\right\|^2\right] \\ &\leq \frac{2\lambda\left(V^* - V(\theta_0)\right)}{\eta T} + \frac{\sigma^2\lambda}{N} + \frac{\sigma^2\lambda}{pNT} , \end{aligned} \tag{28}$$

where $a := \arg\min_{t \leq T} \mathbb{E}\left[\left\|\nabla V(\theta_t)\right\|^2\right]$.

□

D. Additional Details on the Hyperparameters

Table 2. Hyperparameter setting for the Acrobot benchmark.

METHOD	η	α	p_t
GPOMDP	10^{-4}	–	–
SVRPG	10^{-5}	–	–
SRVRPG	$6 \cdot 10^{-6}$	–	–
STORM-PG	10^{-4}	0.9	–
PAGE-PG	$5 \cdot 10^{-6}$	–	0.01, 0.4

Table 3. Hyperparameter setting for the Cartpole benchmark.

METHOD	η	α	p_t
GPOMDP	10^{-4}	–	–
SVRPG	10^{-4}	–	–
SRVRPG	10^{-6}	–	–
STORM-PG	$4 \cdot 10^{-5}$	0.99	–
PAGE-PG	$5 \cdot 10^{-5}$	–	0.8