Reinforcement Learning with General Utilities: Simpler Variance Reduction and Large State-Action Space

Anas Barakat [1] **Ilyas Fatkhullin** [1] **Niao He** [1]

Abstract

We consider the reinforcement learning (RL) problem with general utilities which consists in maximizing a function of the state-action occupancy measure. Beyond the standard cumulative reward RL setting, this problem includes as particular cases constrained RL, pure exploration and learning from demonstrations among others. For this problem, we propose a simpler single-loop parameter-free normalized policy gradient algorithm. Implementing a recursive momentum variance reduction mechanism, our algorithm achieves $\tilde{\mathcal{O}}(\epsilon^{-3})$ and $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexities for ϵ-first-order stationarity and ϵ-global optimality respectively, under adequate assumptions. We further address the setting of large finite state action spaces via linear function approximation of the occupancy measure and show a $\tilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity for a simple policy gradient method with a linear regression subroutine.

1. Introduction

While the classical Reinforcement Learning (RL) problem consists in learning a policy maximizing the expected cumulative sum of rewards through interaction with an environment, several other problems of practical interest are concerned with objectives involving more general utilities. Examples of such problems include pure exploration in RL via maximizing the entropy of the state visitation distribution (see for e.g., Hazan et al. (2019); Mutti et al. (2022a)), imitation learning via minimizing an f-divergence between state-action occupancy measures of an agent and an expert (Ghasemipour et al., 2020), risk-sensitive (Zhang et al., 2021a) or risk-averse RL maximizing, for instance, the Conditional Value-at-Risk (Garcıa & Fernández, 2015), constrained RL (see for e.g., Altman (1999); Borkar (2005); Bhatnagar & Lakshmanan (2012); Miryoosefi et al. (2019); Efroni et al. (2020)), experiment design (Mutny et al., 2023) and diverse skill discovery (Eysenbach et al., 2019) among others. We refer the interested reader to Table 1 in Zahavy et al. (2021); Mutti et al. (2022b), Zhang et al. (2020) and references therein for further examples and a more comprehensive description of such problems.

Recently, Zhang et al. (2020; 2021b) proposed a unified formulation encapsulating all the aforementioned problems as a maximization of a functional (which may not be concave) over the set of state-action occupancy measures. Interestingly, this formulation generalizes standard RL which corresponds to maximizing a linear functional of the state-action occupancy measure (Puterman, 2014). Subsuming the standard RL problem, the case where the objective functional is convex (concave for maximization) in the occupancy measure is known as Convex RL (Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022; Mutti et al., 2022b).

Unlike the standard RL problem which enjoys a nice additive structure, the more general nonlinear functional alters the additive structure of the problem, invalidates the classical Bellman equations as a consequence and hence hinders the standard use of the dynamical programming machinery (see for e.g., Bertsekas (2019); Sutton & Barto (2018)). While value-based methods are not meaningful anymore in this general (nonlinear) utilities setting, Zhang et al. (2020; 2021b) proposed a direct policy search method to solve the RL problem with general utilities. This class of methods directly updates a parametrized policy along the gradient direction of the objective function. More precisely, Zhang et al. (2021b) propose a double-loop Policy Gradient (PG) method called TSIVR-PG implementing a variance reduction mechanism requiring two large batches and checkpoints. Similarly to existing variance-reduced PG methods in the standard RL setting, the algorithm makes use of importance sampling (IS) weights to account for the distribution shift inherent to the RL setting. Interestingly, while most existing variance-reduced PG methods make an unrealistic and unverifiable assumption which guarantees that the IS weights variance is bounded at each iteration of the algorithm, Zhang et al. (2021b) alleviate this issue by using a gradient truncation mechanism. Such a strategy consisting in performing a

[1]Department of Computer Science, ETH Zurich, Switzerland. Correspondence to: A.B. <anas.barakat@inf.ethz.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

truncated gradient step can be formulated as solving a trust-region subproblem at each iteration, which is reminiscent of trust-region based algorithms such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017). In particular, implementing TSIVR-PG requires tuning a gradient truncation radius depending on problem parameters while also choosing adequate large batches. Besides these algorithmic considerations, a major limitation of recent prior work (Zhang et al., 2021b; 2020; Kumar et al., 2022) is the need to estimate the unknown occupancy measure at each state-action pair. In several problems of practical scale, the number of states and/or actions is prohibitively large and renders tabular methods intractable. For instance, the size of a state space grows exponentially with the number of state variables. This is commonly known as the curse of dimensionality.

In this paper, we consider the RL problem with general utilities. Our contributions are as follows:

- We propose a novel single-loop normalized PG algorithm called N-VR-PG using only a single trajectory per iteration. In particular, our algorithm does not require the knowledge of problem specific parameters, large batches nor checkpoints unlike TSIVR-PG in Zhang et al. (2021b). Instead of gradient truncation, we propose to use a normalized update rule for which no additional gradient truncation hyperparameter is needed. At the heart of our algorithm design is a recursive double variance reduction mechanism implemented with momentum for both the stochastic policy gradient and the occupancy measure estimator (in the tabular setting), akin to STORM (Cutkosky & Orabona, 2019) in stochastic optimization.

- We show that using a normalized gradient update guarantees bounded IS weights for the softmax parametrization. Unlike in most prior works focusing on the particular case of the standard RL setting, variance of IS weights is automatically bounded and no further assumption is needed. We further demonstrate that IS weights can also be similarly controlled when using a gaussian policy for continuous state-action spaces under mild assumptions.

- In the general utilities setting with finite state-action spaces and softmax policy, we show that our algorithm requires $\tilde{\mathcal{O}}(\varepsilon^{-3})$ samples to reach an ε-stationary point of the objective function and $\tilde{\mathcal{O}}(\varepsilon^{-2})$ samples to reach an ε-globally optimal policy by exploiting the hidden concavity of the problem when the utility function is concave and the policy is overparametrized. In the standard RL setting, we further show that such sample complexity results also hold for continuous state-action spaces when using the gaussian policy under adequate assumptions.

- Beyond the tabular setting, we consider the case of large finite state and action spaces which has not been previously addressed in this general setting to the best of our knowledge. We consider approximating the unknown state-action occupancy measure itself by a linear combination of pre-selected basis functions via a least-mean-squares solver. This linear function approximation procedure combined with a stochastic policy gradient method results in an algorithm for solving the RL problem with general nonlinear utilities for large state and action spaces. Specifically, we show that our PG method requires $\tilde{\mathcal{O}}(\varepsilon^{-4})$ samples to guarantee an ε-first-order stationary point of the objective function up to an error floor due to function approximation.

Related works. We briefly discuss standard RL before closely related works for RL with general utility.

Variance-reduced PG for standard RL. In the last few years, there has been a vast array of work around variance-reduced PG methods for solving the standard RL problem with a cumulative sum of rewards to reduce the high variance of the stochastic policy gradients (see for e.g., Papini et al. (2018); Xu et al. (2020a); Pham et al. (2020); Gargiani et al. (2022)). Yuan et al. (2020); Huang et al. (2020) proposed momentum-based policy gradient methods. All the aforementioned works use IS and make an unverifiable assumption stipulating that the IS weights variance is bounded. To relax this unrealistic assumption, Zhang et al. (2021b) provide a gradient truncation mechanism complementing IS for the specific case of the softmax parameterization whereas Shen et al. (2019); Salehkaleybar et al. (2022) incorporate second-order information for which IS is not needed. Even in the special case of standard cumulative reward, our algorithm differs from prior work in that it combines the following features: it is single-loop, runs with a single trajectory per iteration and uses a normalized update rule to control the IS weights without further assumption. In particular, our algorithm does not make use of second order information and thus our analysis does not require second-order smoothness conditions. Typically, variance-reduced PG methods guarantee a $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity to reach a first-order stationary policy, improving over its $\tilde{\mathcal{O}}(\varepsilon^{-4})$ counterpart for vanilla PG. Subsequently to the recent work of Agarwal et al. (2021) which provided global optimality guarantees for PG methods despite the non-concavity of the problem, several works (Liu et al., 2020; Zhang et al., 2021b; Ding et al., 2021; 2022; Yuan et al., 2022; Masiha et al., 2022; Yuan et al., 2023) established global optimality guarantees for stochastic PG methods with or without variance reduction under policy parametrization. The best known sample complexity to reach an ϵ-globally optimal policy is $\tilde{\mathcal{O}}(\epsilon^{-2})$ and was achieved via policy mirror descent without parametrization (Lan, 2022; Xiao, 2022), with log-linear policies recently (Yuan et al., 2023) and via variance-reduced PG for softmax parametrization by exploiting hidden convex-

ity (Zhang et al., 2021b). Very recently, Fatkhullin et al. (2023) obtained a $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for Fisher-non-degenerate parametrized policies.

RL with General Utility. There is a huge literature addressing control problems with nonstandard utilities that we cannot hope to give justice to. Let us mention though some early examples in Operations Research such as inventory problems with constraints on the probability of shortage (Derman & Klein, 1965) and variance-penalized MDPs (Filar et al., 1989; Kallenberg, 1994) where the problem is formulated as a nonlinear program in the space of state-action frequencies. In the rest of this section, we briefly discuss the most relevant research to the present paper. Zhang et al. (2020) study the policy optimization problem where the objective function is a concave function of the state-action occupancy measure to include several known problems such as constrained MDPs, exploration and learning from demonstrations. To solve this problem for which dynamic programming cannot be employed, Zhang et al. (2020) investigate policy search methods and first define a variational policy gradient for RL with general utilities as the solution to a stochastic saddle point problem. Exploiting the hidden convexity structure of the problem, they further show global optimality guarantees when having access to exact policy gradients. However, the procedure to estimate even a single policy gradient via the proposed primal-dual stochastic approximation method from sample paths turns out to be complex. Leveraging the formulation of the RL problem as a stochastic composite optimization problem, Zhang et al. (2021b) later proposed a (variance-reduced) stochastic PG approach for solving general utility RL ensuring a $\tilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity to find an ϵ-stationary policy under smoothness of the utility function and the policy parametrization and a $\tilde{\mathcal{O}}(\epsilon^{-2})$ global optimality sample complexity for a concave utility with an overparametrized policy. When the utility is concave as a function of the occupancy measure, the corresponding RL problem is known as Convex RL or Convex MDPs. Using Fenchel duality, Zahavy et al. (2021) casted the convex MDP problem as a min-max game between a policy player and a cost player producing rewards that the policy player must maximize. An insightful consequence of this viewpoint is that any algorithm solving the standard RL problem can be used for solving the more general convex MDP problem. In the present paper, we adopt the direct policy search approach with policy parametrization proposed in Zhang et al. (2021b) instead of the dual viewpoint. Geist et al. (2022) show that Convex RL is a subclass of Mean-Field games. Zhang et al. (2022) consider a decentralized version of the problem with general utilities with a network of agents.

2. Preliminaries

Notations. For a given finite set $\mathcal{X}$, we use the notation $|\mathcal{X}|$ for its cardinality and $\Delta(\mathcal{X})$ for the space of probability distributions over $\mathcal{X}$. We equip any Euclidean space with its standard inner product denoted by $\langle \cdot, \cdot \rangle$. The notation $\|\cdot\|$ refers to both the standard 2-norm and the spectral norm for vectors and matrices respectively.

Markov Decision Process with General Utility. Consider a discrete-time discounted Markov Decision Process (MDP) with a general utility function $\mathbb{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, F, \rho, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability kernel, $F : \mathcal{M}(\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$ is a general utility function defined over the space of measures $\mathcal{M}(\mathcal{S} \times \mathcal{A})$ on the product space $\mathcal{S} \times \mathcal{A}$, ρ is the initial state distribution and $\gamma \in (0, 1)$ is the discount factor. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state $s \in \mathcal{S}$ to a distribution $\pi(\cdot|s)$ over the action space $\mathcal{A}$. The set of all stationary policies is denoted by Π. At each time step $t \in \mathbb{N}$ in a state $s_t \in \mathcal{S}$, the RL agent chooses an action $a_t \in \mathcal{A}$ with probability $\pi(a_t|s_t)$ and the environment transitions to a state s_{t+1} with probability $\mathcal{P}(s_{t+1}|s_t, a_t)$. We denote by $\mathbb{P}_{\rho,\pi}$ the probability distribution of the Markov chain $(s_t, a_t)_{t\in\mathbb{N}}$ induced by the policy π with initial state distribution ρ. We use the notation $\mathbb{E}_{\rho,\pi}$ (or often simply $\mathbb{E}$ instead) for the associated expectation. We define for any policy $\pi \in \Pi$ the state-action occupancy measure $\lambda^\pi \in \mathcal{M}(\mathcal{S} \times \mathcal{A})$ as:

$$\lambda^\pi(s, a) \overset{\text{def}}{=} \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\rho,\pi}(s_t = s, a_t = a)\,. \tag{1}$$

We denote by Λ the set of such occupancy measures, i.e., $\Lambda \overset{\text{def}}{=} \{\lambda^\pi : \pi \in \Pi\}$. Then, the general utility function F assigns a real to each occupancy measure λ^π induced by a policy $\pi \in \Pi$. A state-action occupancy measure λ^π will also be seen as a vector of the Euclidean space $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Policy parametrization. In this paper, we will consider the common softmax policy parametrization defined for every $\theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$ by:

$$\pi_\theta(a|s) = \frac{\exp(\psi(s, a; \theta))}{\sum_{a' \in \mathcal{A}} \exp(\psi(s, a'; \theta))}\,, \tag{2}$$

where $\psi : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$ is a smooth function. The softmax parametrization will be important for controlling IS weights for variance reduction. However, some of our results will not require this specific parameterization and we will explicitly indicate it when appropriate.

Problem formulation. The goal of the RL agent is to find a policy π_θ (determined by the vector θ) solving the problem:

$$\max_{\theta \in \mathbb{R}^d} F(\lambda^{\pi_\theta})\,, \tag{3}$$

where F is a smooth function supposed to be upper bounded and $F^\star$ is used in the remainder of this paper to denote the maximum in (3). The agent has only access to (a) trajectories of finite length H generated from the MDP under

the initial distribution ρ and the policy π_θ and (b) the gradient of the utility function F with respect to (w.r.t.) its variable λ. In particular, provided a time horizon H and a policy π_θ with $\theta \in \mathbb{R}^d$, the learning agent can simulate a trajectory $\tau = (s_0, a_0, \cdots, s_{H-1}, a_{H-1})$ from the MDP whereas the state transition kernel $\mathcal{P}$ is unknown. This general utility problem was described, for instance, in Zhang et al. (2021b) (see also Kumar et al. (2022)). Recall that the standard RL problem corresponds to the particular case where the general utility function is a linear function, i.e., $F(\lambda^{\pi_\theta}) = \langle r, \lambda^{\pi_\theta} \rangle$ for some vector $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ in which case we recover the expected return function as an objective:

$$V^{\pi_\theta}(r) \stackrel{\text{def}}{=} \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t)\right] . \tag{4}$$

In the standard RL case, we shall use the notation $J(\theta) \stackrel{\text{def}}{=} V^{\pi_\theta}(r)$ where r is the corresponding reward function.

Policy Gradient for General Utilities. Following the exposition in (Zhang et al., 2021b) (see also more recently (Kumar et al., 2022)), we derive the policy gradient for the general utility objective. For convenience, we use the notation $\lambda(\theta)$ for λ^{π_θ}. Since the cumulative reward can be rewritten more compactly $V^{\pi_\theta}(r) = \langle \lambda^{\pi_\theta}, r \rangle$, it follows from the policy gradient theorem that:

$$\begin{aligned}[\nabla_\theta \lambda(\theta)]^T r &= \nabla_\theta V^{\pi_\theta}(r) \\ &= \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) \sum_{t'=0}^{t} \nabla \log \pi_\theta(a_{t'}|s_{t'})\right] , \end{aligned} \tag{5}$$

where $\nabla_\theta \lambda(\theta)$ is the Jacobian matrix of the vector mapping $\lambda(\theta)$. Using the chain rule, we have

$$\begin{aligned}\nabla_\theta F(\lambda(\theta)) &= [\nabla_\theta \lambda(\theta)]^T \nabla_\lambda F(\lambda(\theta)) \\ &= \nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\lambda(\theta))} . \end{aligned} \tag{6}$$

Stochastic Policy Gradient. In light of (6), in order to estimate the policy gradient $\nabla_\theta F(\lambda(\theta))$ for general utilities, we can use the standard reinforce estimator suggested by Eq. (5) but we also need to estimate the state-action occupancy measure $\lambda(\theta)$ (when F is nonlinear)[1]. Define for every reward function r (which is also seen as a vector in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$), every $\theta \in \mathbb{R}^d$ and every H-length trajectory τ simulated from the MDP with policy π_θ and initial distribution ρ the (truncated) policy gradient estimate:

$$g(\tau, \theta, r) = \sum_{t=0}^{H-1} \left(\sum_{h=t}^{H-1} \gamma^h r(s_h, a_h)\right) \nabla \log \pi_\theta(a_t|s_t) . \tag{7}$$

[1] In the cumulative reward setting, notice that the general utility function F is linear and $\nabla_\lambda F(\lambda(\theta))$ is independent of $\lambda(\theta)$.

We also define an estimator for the state-action occupancy measure $\lambda^{\pi_\theta} = \lambda(\theta)$ (see (1)) truncated at the horizon H by:

$$\lambda(\tau) = \sum_{h=0}^{H-1} \gamma^h \delta_{s_h,a_h} , \tag{8}$$

where for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\delta_{s,a} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is a vector of the canonical basis of $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, i.e., the vector whose only non-zero entry is the (s, a)-th entry which is equal to 1.

Importance Sampling. Given a trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots, s_{H-1}, a_{H-1})$ of length H generated under the initial distribution ρ and the policy π_θ for some $\theta \in \mathbb{R}^d$, we define for every $\theta' \in \mathbb{R}^d$ the IS weight:

$$w(\tau|\theta', \theta) \stackrel{\text{def}}{=} \prod_{h=0}^{H-1} \frac{\pi_{\theta'}(a_h|s_h)}{\pi_\theta(a_h|s_h)} . \tag{9}$$

Since the problem is nonstationary in the sense that updating the parameter θ shifts the distribution over trajectories, it follows that for any $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $\mathbb{E}_{\rho,\pi_\theta}[g(\tau, \theta, r) - g(\tau, \theta', r)] \neq \nabla_\theta V^{\pi_\theta}(r) - \nabla_\theta V^{\pi_{\theta'}}(r)$. Using the IS weights, we correct this bias to obtain

$$\begin{aligned}&\mathbb{E}_{\rho,\pi_\theta}[g(\tau, \theta, r) - w(\tau|\theta', \theta) g(\tau, \theta', r)] \\ &\qquad = \nabla_\theta V^{\pi_\theta}(r) - \nabla_\theta V^{\pi_{\theta'}}(r) .\end{aligned}$$

The use of IS weights is standard in variance-reduced PG.

3. Normalized Variance-Reduced Policy Gradient Algorithm

In this section, we present our N-VR-PG algorithm (see Algorithm 1) to solve the RL problem with general utilities. This algorithm has two main distinctive features compared to vanilla PG and existing algorithms (Zhang et al., 2021b): (i) recursive variance reduction: instead of using the stochastic PG and occupancy measure estimators respectively reported in (7) and (8), we use recursive variance-reduced estimators for both the PG and the state-action occupancy measure akin to STORM in stochastic optimization (Cutkosky & Orabona, 2019). This leads to a simple single-loop algorithm using a single trajectory per iteration and for which no checkpoints nor any second order information are needed; (ii) normalized PG update rule: normalization will be crucial to control the IS weights used in the estimators. We elaborate more on the motivation for using it in Section 4.1.

Remark 3.1. In Algorithm 1, note that $g(\tau_t, \theta_t, r_{t-1})$ and $g(\tau_t, \theta_{t-1}, r_{t-2})$ are used in v_t instead of $g(\tau_t, \theta_t, r_t)$ and $g(\tau_t, \theta_{t-1}, r_{t-1})$ respectively to address measurability and independence issues in the analysis.

Remark 3.2 (Standard RL). In the cumulative reward setting, estimating the occupancy measure is not needed. Hence, Algorithm 1 simplifies (see Algorithm 4 in Appendix A).

Algorithm 1 N-VR-PG (General Utilities)

Input: θ_0, T, H, $\{\eta_t\}_{t\geq 0}$, $\{\alpha_t\}_{t\geq 0}$.
Sample τ_0 of length H from $\mathbb{M}$ and π_{θ_0}
$\lambda_0 = \lambda(\tau_0, \theta_0); r_0 = \nabla_\lambda F(\lambda_0); r_{-1} = r_0$
$d_0 = g(\tau_0, \theta_0, r_0)$
$\theta_1 = \theta_0 + \alpha_0 \frac{d_0}{\|d_0\|}$
for $t = 1, \ldots, T-1$ **do**
  Sample τ_t of length H from MDP $\mathbb{M}$ and π_{θ_t}
  $u_t = \lambda(\tau_t)(1 - w(\tau_t|\theta_{t-1}, \theta_t))$
  $\lambda_t = \eta_t \lambda(\tau_t) + (1-\eta_t)(\lambda_{t-1} + u_t)$
  $r_t = \nabla_\lambda F(\lambda_t)$
  $v_t = g(\tau_t, \theta_t, r_{t-1}) - w(\tau_t|\theta_{t-1}, \theta_t) g(\tau_t, \theta_{t-1}, r_{t-2})$
  $d_t = \eta_t g(\tau_t, \theta_t, r_{t-1}) + (1-\eta_t)(d_{t-1} + v_t)$
  $\theta_{t+1} = \theta_t + \alpha_t \frac{d_t}{\|d_t\|}$
end for

4. Convergence Analysis of N-VR-PG

We first introduce our assumptions regarding the regularity of the policy parametrization and the utility function F.

Assumption 4.1. In the softmax parametrization (2), the map $\psi(s, a; \cdot)$ is twice continuously differentiable and there exist $l_\psi, L_\psi > 0$ s.t. (i) $\max_{s\in\mathcal{S}, a\in\mathcal{A}} \sup_\theta \|\nabla \psi(s, a; \theta)\| \leq l_\psi$ and (ii) $\max_{s\in\mathcal{S}, a\in\mathcal{A}} \sup_\theta \|\nabla^2 \psi(s, a; \theta)\| \leq L_\psi$.

Assumption 4.2. There exist constants $l_\lambda, L_\lambda, L_{\lambda,\infty} > 0$ s.t. for all $\lambda, \lambda' \in \Lambda$, $\|\nabla_\lambda F(\lambda)\|_\infty \leq l_\lambda$ and

$$\|\nabla_\lambda F(\lambda) - \nabla_\lambda F(\lambda')\|_\infty \leq L_\lambda \|\lambda - \lambda'\|_2 ,$$
$$\|\nabla_\lambda F(\lambda) - \nabla_\lambda F(\lambda')\|_\infty \leq L_{\lambda,\infty} \|\lambda - \lambda'\|_1 .$$

Assumptions 4.1 and 4.2 were previously considered in Zhang et al. (2021b; 2020) and guarantee together that the objective function $\theta \mapsto F(\lambda^{\pi_\theta})$ is smooth. Assumption 4.2 is automatically satisfied for the cumulative reward setting (i.e., F linear) if the reward function is bounded.

4.1. Normalization ensures boundedness of IS weights

Most prior works suppose that the variance of the IS weights is bounded. Such assumption cannot be verified. In this section we provide an alternative algorithmic way based on the softmax policy to control the IS weights without the aforementioned assumption. Since our algorithm only uses IS weights for two consecutive iterates, our key observation is that a normalized gradient update rule automatically guarantees bounded IS weights. In particular, compared to Zhang et al. (2021b), we do not use a gradient truncation mechanism which requires an additional truncation hyperparameter depending on the problem parameters and dictates a non-standard stationarity measure (see Remark 4.6). This simple algorithmic modification requires several adjustments in the convergence analysis (see Appendix E and F). We formalize the result in the following lemma.

Lemma 4.3. *Let Assumption 4.1 hold true. Suppose that the sequence (θ_t) is updated via $\theta_{t+1} = \theta_t + \alpha_t \frac{d_t}{\|d_t\|}$ where $d_t \in \mathbb{R}^d$ is any non-zero update direction and α_t is a positive stepsize. Then, for every integer t and any trajectory τ of length H, we have $w(\tau|\theta_t, \theta_{t+1}) \leq \exp\{2H l_\psi \alpha_t\}$. If, in addition, $H = \mathcal{O}(\frac{\log T}{1-\gamma})$ and $\alpha_t = \alpha = T^{-\frac{2}{3}}$, then there exists a constant $W > 0$ s.t. $w(\tau|\theta_t, \theta_{t+1}) \leq W$. Moreover, we have $\mathrm{Var}\left[w(\tau_{t+1}|\theta_t, \theta_{t+1})\right] \leq C_w \alpha^2$ where τ_{t+1} is a trajectory of length H sampled from $\pi_{\theta_{t+1}}$ and $C_w \overset{def}{=} H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$.*

In this lemma, the variance of the IS weights decreases over time at a rate controlled by α^2 and this result will be crucial for our convergence analysis of N-VR-PG. We show in Lemma E.19 in the Appendix that such a result also holds for Gaussian policies for continuous state action spaces.

4.2. First-order stationarity

In this section, we show that N-VR-PG requires $\tilde{\mathcal{O}}(\varepsilon^{-3})$ samples to reach an ε-first-order stationary (FOS) point of the objective function for RL with general utilities.[2]

Theorem 4.4. *Let Assumptions 4.1 and 4.2 hold. Let $\alpha_0 > 0$ and let $T \geq 1$ be an integer. Set $\alpha_t = \frac{\alpha_0}{T^{2/3}}, \eta_t = \left(\frac{2}{t+1}\right)^{2/3}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Then, $\mathbb{E}\left[\left\|\nabla_\theta F(\lambda(\bar{\theta}_T))\right\|\right] \leq \mathcal{O}\left(\frac{1+(1-\gamma)^3\Delta\alpha_0^{-1}+(1-\gamma)^{-1}\alpha_0}{(1-\gamma)^3 T^{1/3}}\right)$, where $\Delta \overset{def}{=} F^\star - \mathbb{E}[F(\lambda(\theta_1))]$ and $\bar{\theta}_T$ is sampled uniformly at random from $\{\theta_1, \cdots, \theta_T\}$ of Algorithm 1.*

Remark 4.5. In terms of dependence on $(1-\gamma)^{-1}$, we significantly improve over the result of Zhang et al. (2021b) which does not make it explicit. We defer a detailed comparison regarding this dependence to Appendix B.

Remark 4.6. Unlike Zhang et al. (2021b) which utilizes a gradient truncation radius, our sample complexity does not depend on the inverse of this gradient truncation hyperparameter which might be small. Indeed, to translate their guarantee from the non-standard gradient mapping dictated by gradient truncation to the standard stationarity measure (used in our result), one has to incur an additional multiplicative constant δ^{-1} where δ is the gradient truncation radius (see Lemma 5.4 in (Zhang et al., 2021b)).

Recalling the notation $J(\theta) = V^{\pi_\theta}(r)$ (see (4)) for the standard RL setting, we can state the following corollary.

Corollary 4.7. *Under the setting of Theorem 4.4, if we set $\alpha_0 = 1-\gamma$, then $\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] \leq \mathcal{O}\left((1-\gamma)^{-2} T^{-1/3}\right)$.*

The next result addresses the case of continuous state-action spaces in the standard RL setting using a Gaussian policy.

[2] All the proofs of our results are provided in the Appendix.

Notably, we rely on similar considerations as for the softmax policy to control the variance of IS weights. We defer a precise statement of this result to Appendix E.4.

Theorem 4.8 (informal)**.** *Using the Gaussian policy under some regularity conditions,* N-VR-PG *(see Algorithm 4) requires $\tilde{\mathcal{O}}(\varepsilon^{-3})$ to reach an ε-first-order stationary point of the expected return J.*

4.3. Global optimality

In this section, we show that N-VR-PG only requires $\tilde{\mathcal{O}}(\varepsilon^{-2})$ samples to reach an ε-globally optimal policy under a concave reparametrization of the RL problem with concave utilities and an additional overparametrization assumption. Our results and assumptions match the recent results in Zhang et al. (2021b) for finite state-action spaces.

Assumption 4.9. The utility function F is concave.

Assumption 4.10. For the softmax policy parametrization in (2), the following three requirements hold: (i) For any $\theta \in \mathbb{R}^d$, there exist relative neighborhoods $\mathcal{U}_\theta \subset \mathbb{R}^d$ and $\mathcal{V}_{\lambda(\theta)} \subset \Lambda$ respectively containing θ and $\lambda(\theta)$ s.t. the restriction $\lambda|_{\mathcal{U}_\theta}$ forms a bijection between $\mathcal{U}_\theta$ and $\mathcal{V}_{\lambda(\theta)}$; (ii) There exists $l > 0$ s.t. for every $\theta \in \mathbb{R}^d$, the inverse $(\lambda|_{\mathcal{U}_\theta})^{-1}$ is l-Lipschitz continuous; (iii) There exists $\bar{\epsilon} > 0$ s.t. for every positive real $\epsilon \leq \bar{\epsilon}$, $(1-\epsilon)\lambda(\theta) + \epsilon\lambda(\theta^*) \in \mathcal{V}_{\lambda(\theta)}$ where π_{θ^*} is the optimal policy.

For the tabular softmax parametrization (i.e., $\psi(s,a;\theta) = \theta_{s,a}, d = |\mathcal{S}||\mathcal{A}|$), a continuous local inverse can be defined whereas computing the Lipschitz constant l is more involved as reported in Zhang et al. (2021b) (see Appendix C for a discussion of Assumption 4.10). Relaxing this strong assumption is left for future work.

Remark 4.11. Compared to Assumption 5.11 in Zhang et al. (2021b), Assumption 4.10 is quasi-identical with the slight difference that it does not depend on the gradient truncation hyperparameter δ used in Zhang et al. (2021b).

Our global optimality convergence result is as follows.

Theorem 4.12. *Let Assumptions 4.1, 4.2 and 4.9 hold. Additionally, let Assumption 4.10 be satisfied with $\bar{\epsilon} \geq \frac{\alpha_0(1-\gamma)}{2\ell_\theta(T+1)^a}$ for some integer $T \geq 1$ and reals $\alpha_0 > 0$, $a \in (0,1)$. Set $\alpha_t = \frac{\alpha_0}{(T+1)^a}$, $\eta_t = \frac{2}{t+1}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Then the output θ_T of* N-VR-PG *(see Algorithm 1) satisfies*

$$F^\star - \mathbb{E}\left[F(\lambda(\theta_T))\right] \leq \mathcal{O}\left(\frac{\alpha_0^2}{(1-\gamma)^3(T+1)^{2a-\frac{3}{2}}}\right),$$

Thus, setting $\alpha_0 = (1-\gamma)^{3/2}$, the sample complexity to achieve $F^* - \mathbb{E}\left[F(\lambda(\theta_T))\right] \leq \varepsilon$ is $\mathcal{O}\left(\varepsilon^{\frac{-2}{4a-3}}\right)$.

Corollary 4.13. *In the setting of Theorem 4.12,* N-VR-PG *(see Algorithm 4) requires $\tilde{\mathcal{O}}\left(\varepsilon^{\frac{-2}{2a-1}}\right)$ samples to achieve $J^\star - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon$ where $J^\star$ is the optimal expected return.*

Remark 4.14. We refer the reader to Appendix F.2 for a precise statement of Corollary 4.13. If we know problem parameters and choose time varying step-sizes $\alpha_t = \frac{\alpha_0}{t}$, then we can obtain exactly $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity.

We can state a similar global optimality result to Corollary 4.13 for continuous state-action spaces (see Appendix F.3).

5. Large State-Action Space Setting

An important limitation of Algorithm 1 and the prior work (Zhang et al., 2021b) is the need to estimate the occupancy measure for each state-action pair in the case of general nonlinear utilities. This procedure is intractable if the state and/or action spaces are prohibitively large and finite or even worse infinite/continuous. In the case of infinite or continuous state-action spaces, the occupancy measure λ^{π_θ} induced by a policy π_θ cannot be represented by a vector in finite dimensions. Thus, the derivative of the utility function F w.r.t. its variable λ is not well defined in the chain rule in (6) for the policy gradient. Therefore, more adequate notions of derivative for optimization on the space of measures are probably needed and this would require different methodological and algorithmic tools which go beyond the scope of this work. In this paper, we propose to do a first step by considering the setting of large *finite* state and action spaces which is already of practical interest.

5.1. PG for RL with General Utilities via linear function approximation of the occupancy measure

Similarly to the classical linear function approximation of the (action-)value function in standard RL, we propose to approximate the (truncated) state-action occupancy measure by a linear combination of pre-selected basis functions in order to break the so-called curse of dimensionality. Our exposition is similar in spirit to the compatible function approximation framework (Sutton et al., 1999) which was recently extended in Agarwal et al. (2021) (see also Yuan et al. (2023) for a recent example). However, we are not concerned here by the approximation of the action-value function nor are we considering the NPG (or Q-NPG) method but we are rather interested in approximating the discounted occupancy measure. Recall that we are considering the more general problem of RL with general utilities. Beyond this connection with existing work, we shall precise that our approach mostly shares the use of standard least squares regression for estimating an unknown function which is the state-action occupancy measure in our case.

Let m be a positive integer and let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ be

a feature map. We shall approximate the truncated[3] state-action occupancy measure for a given policy π_θ ($\theta \in \mathbb{R}^d$ fixed) by a linear combination of feature vectors from the feature map, i.e., for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\lambda_H^{\pi_\theta}(s,a) \approx \langle \phi(s,a), \omega_\theta \rangle \,, \tag{10}$$

for some $\omega_\theta \in \mathbb{R}^m$ that we shall compute. Typically, the dimension m is much smaller than $|\mathcal{S}| \times |\mathcal{A}|$. The feature map summarizes the most important characteristics of state-action pairs. Typically, this map is designed based on experience and domain-specific knowledge or intuition regarding the MDP. Standard examples of basis functions for the feature map include radial basis functions, wavelet networks or polynomials. Nevertheless, designing such a feature map is an important practical question that is often problem-specific and we will not address it in this work.

In order to compute such a vector ω_θ, we will use linear regression. Accordingly, we define the expected regression loss measuring the estimation quality of any parameter ω for every $\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^m$ by:

$$L_\theta(\omega) \overset{\text{def}}{=} \mathbb{E}_{s\sim\rho, a\sim\mathcal{U}(\mathcal{A})}[(\lambda_H^{\pi_\theta}(s,a) - \langle \phi(s,a), \omega \rangle)^2] \,, \tag{11}$$

where ρ is the initial distribution in the MDP and $\mathcal{U}(\mathcal{A})$ is the uniform distribution over the action space $\mathcal{A}$.[4] In practice, we cannot minimize L_θ exactly since this would require having access to the true state-action occupancy measure and averaging over all state-action pairs $s \sim \rho, a \sim \mathcal{U}(\mathcal{A})$. Therefore, we compute an approximate solution $\hat{\omega}_\theta \approx \arg\min_\omega L_\theta(\omega)$. For this procedure, we need: (i) unbiased estimates of the true truncated state-action occupancy measure $\lambda_H^{\pi_\theta}(s,a)$ (or the non-truncated one $\lambda^{\pi_\theta}(s,a)$) for $s \sim \rho, a \sim \mathcal{U}(\mathcal{A})$ and (ii) a regression solver based on samples to minimize L_θ as defined in (11). As for item (i), we use a Monte-Carlo estimate $\hat{\lambda}_H^{\pi_\theta}(s,a)$ of the truncated occupancy measure computed from a single rollout (see Algorithm 5 for details).[5] An unbiased stochastic gradient of the function L_θ in (11) is then given by

$$\hat{\nabla}_\omega L_\theta(\omega) \overset{\text{def}}{=} 2(\langle \phi(s,a), \omega \rangle - \hat{\lambda}_H^{\pi_\theta}(s,a))\, \phi(s,a) \,. \tag{12}$$

We can then solve the regression problem consisting in minimizing L_θ in (11) via the averaged SGD algorithm (see Algorithm 2) as proposed in Bach & Moulines (2013).

Using this procedure, we propose a simple stochastic PG algorithm for solving the RL problem with general utilities

[3]We could use the non-truncated occupancy measure (see Appendix G). For simplicity of exposition, we use the truncated version, the difference between both quantities is of the order of γ^H.

[4]Other exploratory sampling distributions for s and a can be considered, we choose ρ and $\mathcal{U}(\mathcal{A})$ for simplicity.

[5]We can also compute an unbiased estimator of the true occupancy measure $\lambda^{\pi_\theta}(s,a)$ via a standard procedure with a random horizon H following a geometric distribution (see Algorithm 6).

Algorithm 2 (averaged) SGD for Occupancy Measure Estimation via Linear Function Approximation

Input: $\omega_0 \in \mathbb{R}^m, K \geq 1, \beta > 0, \rho, \pi_\theta$.
for $k = 0, \dots, K-1$ **do**
  Sample $s \sim \rho; a \sim \mathcal{U}(\mathcal{A})$
  Compute an estimator $\hat{\lambda}_H^{\pi_\theta}(s,a)$ via Algorithm 5
  $\hat{\nabla}_\omega L_\theta(\omega_k) \overset{\text{def}}{=} 2(\langle \phi(s,a), \omega_k \rangle - \hat{\lambda}_H^{\pi_\theta}(s,a))\, \phi(s,a)$
  $\omega_{k+1} = \omega_k - \beta\, \hat{\nabla}_\omega L_\theta(\omega_k)$
end for
Return: $\hat{\omega}_\theta = \frac{1}{K} \sum_{k=1}^K \omega_k$

for large state action spaces. Since this large-scale setting has not been priorly addressed for general utilities to the best of our knowledge, we focus on a simpler PG algorithm without the variance reduction and normalization features of our algorithm in Section 3. Incorporating variance reduction to occupancy measure estimates seems more involved with our linear regression procedure for function approximation. We leave it for future work to design a method with improved sample complexity using variance reduction.

Algorithm 3 Stochastic PG for RL with General Utilities via Linear Function Approximation of the Occupancy Measure

Input: $\theta_0 \in \mathbb{R}^d, T, N \geq 1, \alpha > 0, K \geq 1, \beta > 0, H$.
Run Algorithm 2 with policy π_{θ_0} and define from its output $\hat{\lambda}_0(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_0} \rangle$.
$r_{-1} = \nabla_\lambda F(\hat{\lambda}_0)$
for $t = 0, \dots, T-1$ **do**
  Run Algorithm 2 with policy π_{θ_t} and define from its output $\hat{\lambda}_t(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_t} \rangle$.
  $r_t = \nabla_\lambda F(\hat{\lambda}_t)$
  Sample a batch of N independent trajectories $(\tau_t^{(i)})_{1\leq i\leq N}$ of length H from $\mathbb{M}$ and π_{θ_t}
  $\theta_{t+1} = \theta_t + \frac{\alpha}{N} \sum_{i=1}^N g(\tau_t^{(i)}, \theta_t, r_{t-1})$
end for
Return: θ_T

Remark 5.1. When running Algorithm 3, notice that the vector $\hat{\lambda}_t \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ (and hence the vector r_t) does not need to be computed for all state-action pairs as this would be unrealistic and even impossible in the large state-action setting we are considering. Indeed, at each iteration, one does only need to compute $(r_t(s_h^{(t)}, a_h^{(t)}))_{0\leq h\leq H-1}$ where $\tau_t = (s_h^{(t)}, a_h^{(t)})_{0\leq h\leq H-1}$ to obtain the stochastic policy gradient $g(\tau_t, \theta_t, r_{t-1})$ as defined in (7).

5.2. Convergence and sample complexity analysis

In this section, we provide a convergence analysis of Algorithm 3. For every integer t, let $\omega_*(\theta_t) \in \arg\min_\omega L_{\theta_t}(\omega)$. We decompose the regression loss into the statistical error measuring the accuracy of our approximate solution and

the approximation error measuring the distance between the true occupancy measure and its best linear approximation using the feature map ϕ:

$$L_{\theta_t}(\hat{\omega}_t) = \underbrace{L_{\theta_t}(\hat{\omega}_t) - L_{\theta_t}(\omega_*(\theta_t))}_{\text{statistical error}} + \underbrace{L_{\theta_t}(\omega_*(\theta_t))}_{\text{approximation error}} ,$$

where we use the shorthand notation $\hat{\omega}_t = \hat{\omega}_{\theta_t}$ and $\hat{\omega}_{\theta_t}$ is the output of Algorithm 2 after K iterations. We assume that both the statistical and approximation errors are uniformly bounded along the iterates of our algorithm. Such assumptions have been considered for instance in a different context in the compatible function approximation framework (see Assumptions 6.1.1 and Corollary 21 in Agarwal et al. (2021), also Assumptions 1 and 5 in Yuan et al. (2023)).

Assumption 5.2 (Bounded statistical error)**.** There exists $\epsilon_{\text{stat}} > 0$ s.t. for all iterations $t \geq 0$ of Algorithm 3, we have $\mathbb{E}[L_{\theta_t}(\hat{\omega}_{\theta_t}) - L_{\theta_t}(\omega_*(\theta_t))] \leq \epsilon_{\text{stat}}$.

We will see in the next section that we can guarantee $\epsilon_{\text{stat}} = \mathcal{O}(1/K)$ where K is the number of iterations of SGD (Algorithm 2) to find the approximate solution $\hat{\omega}_t$ at each iteration t of Algorithm 3.

Assumption 5.3 (Bounded approximation error)**.** There exists $\epsilon_{\text{approx}} > 0$ s.t. for all iterations $t \geq 0$ of Algorithm 3, we have $\mathbb{E}[L_{\theta_t}(\omega_*(\theta_t))] \leq \epsilon_{\text{approx}}$.

This error is due to function approximation and depends on the expressiveness of the approximating function class. The true state-action occupancy measure to be estimated may not lie in the function approximation class under consideration.

Theorem 5.4. *Let Assumptions 4.1, 4.2, 5.2 and 5.3 hold true. In addition, suppose that there exists $\rho_{\min} > 0$ s.t. $\rho(s) \geq \rho_{\min}$ for all $s \in \mathcal{S}$. Let $T \geq 1$ be an integer and let (θ_t) be the sequence generated by Algorithm 3 with a positive step size $\alpha = \mathcal{O}(1)$ and batch size $N \geq 1$. Then,*

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{1}{N}\right) + \mathcal{O}(\gamma^{2H}) + \mathcal{O}(\epsilon_{stat} + \epsilon_{approx}) , \quad (13)$$

where $\bar{\theta}_T \in \{\theta_1, \cdots, \theta_T\}$ uniformly at random.

A few comments are in order regarding Theorem 5.4 : (1) The specific structure of the softmax parametrization is not needed for Theorem 5.4. Indeed, this softmax parametrization is only useful to control IS weights used for variance reduction in Algorithm 1. Assumption 4.1 can be replaced by any smooth policy parametrization satisfying the same standard conditions with $\nabla \log \pi_\theta$ instead of ψ ; (2) If the true (truncated) occupancy measure does not lie in the class of linear functions described, a positive function approximation error ϵ_{approx} is incurred due to the bias induced by the limited expressiveness of the linear function approximation. A possible natural alternative is to consider richer classes such as neural networks to approximate the state-action occupancy measure and reduce the approximation bias. In this more involved case, the expected least squares (or other metrics) regression loss would likely become nonconvex and introduce further complications in our analysis. Such an extension would require other technical tools that are beyond the scope of the present paper and we leave it for future work.

In order to establish the total sample complexity of our algorithm, we need to compute the number of samples needed in the occupancy measure estimation subroutine of Algorithm 2. To do so, we now specify the number of SGD iterations required in Algorithm 2 to approximately solve our regression problem. In particular, we will show that we can achieve $\epsilon_{\text{stat}} = \mathcal{O}(1/K)$ where K is the number of iterations of the SGD subroutine using Theorem 1 in Bach & Moulines (2013). Before stating our result, we make an additional standard assumption on the feature map ϕ .

Assumption 5.5. The feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ satisfies: (i) There exists $B > 0$ s.t. for all $s \in \mathcal{S}, a \in \mathcal{A}$, $\|\phi(s,a)\| \leq B$ and (ii) There exists $\mu > 0$ s.t. $\mathbb{E}_{s\sim\rho, a\sim\mathcal{U}(\mathcal{A})}[\phi(s,a)\phi(s,a)^T] \succcurlyeq \mu I_m$ where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix.

Assumption 5.5 guarantees that the covariance matrix of the feature map is invertible. Similar standard assumptions have been commonly considered for linear function approximation settings (Tsitsiklis & Van Roy, 1997).

We are now ready to state a corollary of Theorem 5.4 establishing the total sample complexity of Algorithm 3 to achieve an ϵ-stationary point of the objective function.

Corollary 5.6. *Let Assumptions 4.1, 4.2, 5.3 and 5.5 hold in the setting of Theorem 5.4 where we run the SGD subroutine of Algorithm 2 with step size $\beta = 1/8B^2$ and $\omega_0 = 0$ for K iterations at each timestep t of Algorithm 3. Then, for every $\epsilon > 0$, setting $T = \mathcal{O}(\epsilon^{-2})$, $N = \mathcal{O}(\epsilon^{-2})$, $K = \mathcal{O}(\epsilon^{-2})$ and $H = \mathcal{O}(\log(\frac{1}{\epsilon}))$ guarantees that $\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|] \leq \mathcal{O}(\epsilon) + \mathcal{O}(\sqrt{\epsilon_{approx}})$ where $\bar{\theta}_T \in \{\theta_1, \cdots, \theta_T\}$ uniformly at random. The total sample complexity to reach an ϵ-stationary point (up to the $\mathcal{O}(\sqrt{\epsilon_{approx}})$ error floor) is given by $T \times (K + M) \times H = \tilde{\mathcal{O}}(\epsilon^{-4})$.*

In terms of the target accuracy ϵ, this result matches the optimal sample complexity to obtain an ϵ-FOSP for nonconvex smooth stochastic optimization via SGD (without variance reduction) up to a log factor.

6. Numerical Simulations

In this section, we present two simple numerical experiments to illustrate the performance of our algorithm compared to prior work and complement our theoretical contri-

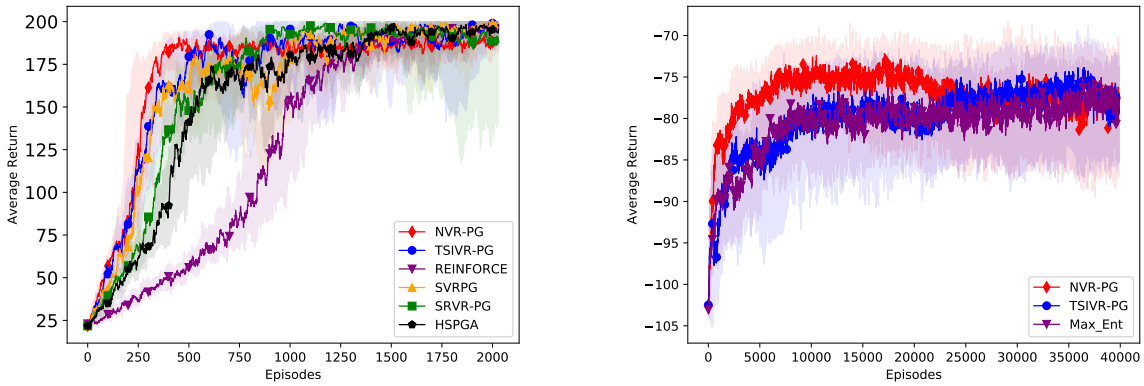

Figure 1. (right) Nonlinear objective maximization in the FrozenLake environment and (left) Standard RL in the CartPole environment. In both cases, the performance curves represent the median return over 20 runs of the algorithms (with 20 seeds) and the shaded colored areas are computed with the 1/4 and 3/4 quantiles of the outcomes.

butions. Our implementation is based on the code provided in Zhang et al. (2021b).[6] Our goal is to show that our algorithm can be competitive compared to existing algorithms while gaining simplicity. We leave further experimental investigations in larger scale problems for future work.

(a) Nonlinear objective function maximization. We consider a general utility RL problem where the objective function $F : \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}$ is a nonlinear function of the occupancy measure defined for every $\lambda \in \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}$ by:

$$F(\lambda) \stackrel{\text{def}}{=} \sum_{s\in\mathcal{S}} \log\left(\sum_{a\in\mathcal{A}} \lambda_{s,a} + \sigma\right),$$

where σ is a small constant which we set to $\sigma = 0.125$. We test our algorithm in the FrozenLake8x8 benchmark environment available in OpenAI gym (Brockman et al., 2016). The result of the experiment is illustrated in Figure 1 (right). The performance curves show that our NVR-PG algorithm shows a relatively faster convergence compared to the TSIVR-PG algorithm (Zhang et al., 2021b) and the MaxEnt algorithm which is specific to the maximum entropy exploration problem (by Hazan et al. (2019)) while the final performances are comparable (see also the overlapping shaded areas). We refer the reader to Section 6.3 in Zhang et al. (2021b) for further details regarding our setting.

(b) Standard RL. While the focus of our work is on the general utility case beyond the standard RL setting, we also perform simulations for the particular case where the objective is a linear function of the state action occupancy measure (i.e., the standard cumulative reward setting) in the CartPole benchmark environment (Brockman et al., 2016).

[6]Available in OpenReview (https://openreview.net/forum?id=Re_VXFOyyO).

Figure 1 (left) shows that our algorithm is competitive with TSIVR-PG (actually even slightly faster, see between 250-500 episodes and see also the shaded areas) and all other algorithms which are not designed for the general utility case (REINFORCE (Williams, 1992), SVRPG (Xu et al., 2020b), SRVR-PG (Xu et al., 2020a), HSPGA (Pham et al., 2020)) while gaining simplicity compared to existing variance-reduced methods. Indeed, our algorithm is single-loop and does not require two distinct batch sizes and checkpoints nor does it require bounded importance sampling weights. Hyperparameters of the algorithms are tuned.

7. Perspectives

Compared to the standard RL setting, the general utilities setting is much less studied. A better understanding of the hidden convexity structure of the problem and its interplay with general policy parametrization would be interesting to derive global optimality guarantees under milder assumptions which would accommodate more practical and expressive policy parametrizations such as neural networks. Regarding the case of large state action spaces, future avenues of research include designing more efficient procedures and guarantees for approximating and estimating the occupancy measure to better address the curse of dimensionality as well as investigating the dual point of view for designing more efficient algorithms. Addressing the case of continuous state-action spaces is also an interesting research direction.

Acknowledgements

This work was supported by ETH AI Center doctoral fellowship, ETH Foundations of Data Science (ETH-FDS), and ETH Research Grant funded via ETH Zurich Foundation.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Altman, E. Constrained markov decision processes. 1999.

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Bedi, A. S., Parayil, A., Zhang, J., Wang, M., and Koppel, A. On the sample complexity and metastability of heavy-tailed policy search in continuous control. *arXiv preprint arXiv:2106.08414*, 2021.

Bedi, A. S., Chakraborty, S., Parayil, A., Sadler, B. M., Tokekar, P., and Koppel, A. On the hidden biases of policy mirror ascent in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1716–1731. PMLR, 17–23 Jul 2022.

Bertsekas, D. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

Bhatnagar, S. and Lakshmanan, K. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

Borkar, V. S. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54 (3):207–213, 2005.

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

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Derman, C. and Klein, M. Some remarks on finite horizon markovian decision models. *Operations research*, 13(2): 272–278, 1965.

Ding, Y., Zhang, J., and Lavaei, J. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Ding, Y., Zhang, J., and Lavaei, J. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 1910–1934. PMLR, 2022.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.

Fatkhullin, I., Etesami, J., He, N., and Kiyavash, N. Sharp analysis of stochastic optimization under global Kurdyka-łojasiewicz inequality. *Advances in Neural Information Processing Systems*, 2022.

Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. *To appear at the International Conference on Machine Learning, arXiv preprint arXiv:2302.01734*, 2023.

Filar, J. A., Kallenberg, L. C., and Lee, H.-M. Variance-penalized markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

Gadat, S., Panloup, F., and Saadane, S. Stochastic Heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Gargiani, M., Zanelli, A., Martinelli, A., Summers, T., and Lygeros, J. PAGE-PG: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7223–7240. PMLR, 2022.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 489–497, 2022.

Ghasemipour, S. K. S., Zemel, R., and Gu, S. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pp. 1259–1277. PMLR, 2020.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019.

Huang, F., Gao, S., Pei, J., and Huang, H. Momentum-based policy gradient methods. In *International conference on machine learning*, pp. 4422–4433. PMLR, 2020.

Kallenberg, L. C. Survey of linear programming for standard and nonstandard markovian control problems. part i: Theory. *Zeitschrift für Operations Research*, 40(1):1–42, 1994.

Kumar, N., Wang, K., Levy, K., and Mannor, S. Policy gradient for reinforcement learning with general utilities. *arXiv preprint arXiv:2210.00991*, 2022.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pp. 1–48, 2022.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Masiha, S., Salehkaleybar, S., He, N., Kiyavash, N., and Thiran, P. Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions. In *Advances in Neural Information Processing Systems*, 2022.

Miryoosefi, S., Brantley, K., Daume III, H., Dudik, M., and Schapire, R. E. Reinforcement learning with convex constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Mutny, M., Janik, T., and Krause, A. Active exploration via experiment design in markov chains. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7349–7374. PMLR, 25–27 Apr 2023.

Mutti, M., De Santi, R., and Restelli, M. The importance of non-markovianity in maximum state entropy exploration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16223–16239. PMLR, 17–23 Jul 2022a.

Mutti, M., Santi, R. D., Bartolomeis, P. D., and Restelli, M. Challenging common assumptions in convex reinforcement learning. *Advances in Neural Information Processing Systems*, 2022b.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Pham, N., Nguyen, L., Phan, D., Nguyen, P. H., Dijk, M., and Tran-Dinh, Q. A hybrid stochastic policy gradient algorithm for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 374–385. PMLR, 2020.

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

Salehkaleybar, S., Khorasani, S., Kiyavash, N., He, N., and Thiran, P. Adaptive momentum-based policy gradient with second-order information. *arXiv preprint arXiv:2205.08253*, 2022.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. Hessian aided policy gradient. In *International conference on machine learning*, pp. 5729–5738. PMLR, 2019.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Tsitsiklis, J. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

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

Xiao, L. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282): 1–36, 2022.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020a.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pp. 541–551. PMLR, 2020b.

Yuan, H., Lian, X., Liu, J., and Zhou, Y. Stochastic Recursive Momentum for Policy Gradient Methods, 2020.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3332–3380. PMLR, 2022.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations*, 2023.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Zhang, J., Bedi, A. S., Wang, M., and Koppel, A. Cautious reinforcement learning via distributional risk in the dual domain. *IEEE Journal on Selected Areas in Information Theory*, 2(2):611–626, 2021a. doi: 10.1109/JSAIT.2021.3081108.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021b.

Zhang, J., Bedi, A. S., Wang, M., and Koppel, A. Multi-agent reinforcement learning with general utilities via decentralized shadow reward actor-critic. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 9031–9039, Jun. 2022.

Contents

Appendix

A. N-VR-PG algorithm for standard RL setting with cumulative reward

In this section, we report the special case of Algorithm 3 for the standard cumulative sum of rewards setting.

Algorithm 4 N-VR-PG (Standard Cumulative Reward)

Input: $\theta_0, T, H, \{\eta_t\}_{t\geq 0}, \{\alpha_t\}_{t\geq 0}$.
Sample τ_0 of length H from $\mathbb{M}$
$d_0 = g(\tau_0, \theta_0)$
$\theta_1 = \theta_0 + \alpha_0 \frac{d_0}{\|d_0\|}$
for $t = 1, \ldots, T-1$ **do**
  Sample τ_t of length H from $\mathbb{M}$ and π_{θ_t}
  $v_t = g(\tau_t, \theta_t) - w(\tau_t|\theta_{t-1}, \theta_t) g(\tau_t, \theta_{t-1})$
  $d_t = \eta_t g(\tau_t, \theta_t) + (1-\eta_t)(d_{t-1} + v_t)$
  $\theta_{t+1} = \theta_t + \alpha_t \frac{d_t}{\|d_t\|}$
end for

Remark A.1. If the direction d_t in Algorithms 1 and 4 is null, then we formally take $\theta_{t+1} = \theta_t$. Note that $d_t \neq 0$ with probability 1 in general. Observe for instance that with $\eta_t = 1$, $d_t = 0$ means that the stochastic policy gradient is equal to zero which means we are already at a first-order stationary point in expectation.

B. Dependence on $(1-\gamma)^{-1}$

In this section, we discuss the dependence of our convergence guarantee in Theorem 4.4 on $(1-\gamma)^{-1}$ where γ is the discount factor of the MDP. The dependence on $1-\gamma$ of our proposed method is $\tilde{\mathcal{O}}((1-\gamma)^{-6}\varepsilon^{-3})$ compared to $\tilde{\mathcal{O}}((1-\gamma)^{-25}\varepsilon^{-3})$ for TSIVR-PG of (Zhang et al., 2021b).

Dependence on $(1-\gamma)^{-1}$ of our N-VR-PG algorithm. It follows from Theorem 4.4 by setting $\alpha_0 = (1-\gamma)^2$ that we need $\tilde{\mathcal{O}}((1-\gamma)^{-6}\varepsilon^{-3})$ samples to reach an ε-stationary point of the utility function (i.e., $\mathbb{E}\|\nabla_\theta F(\lambda(\theta_{out}))\| \leq \varepsilon$).

Derivation of explicit dependence on $(1-\gamma)^{-1}$ in Theorem 5.9 of (Zhang et al., 2021b). Although the dependence is not made explicit in the aforementioned work, we can use their intermediate results in the proofs in order to derive it. We use their notations in the following.

From the last two lines of page 26 (in the proof of Theorem 5.9), we can infer that $\mathbb{E}\left[\|\mathcal{G}(\theta_{out})\|\right] \leq \mathcal{O}(C_3\varepsilon)$, where $C_3 = \mathcal{O}(H(1-\gamma)^{-7})$ which is defined in the statement of Lemma F.2 on page 23. Here in $\mathcal{O}$ we only hide the dependence on the smoothness constants and other numerical constants. The statement of Theorem 5.9 guarantees that in order to achieve this, we need $T(mB+N)H = \mathcal{O}(H\varepsilon^{-3})$ number of samples. By setting $\varepsilon_1 = C_3\varepsilon$, this translates to $\mathcal{O}(HC_3^3\varepsilon_1^{-3}) = \mathcal{O}(H^4(1-\gamma)^{-21}\varepsilon_1^{-3}) = \tilde{\mathcal{O}}((1-\gamma)^{-25}\varepsilon_1^{-3})$ samples to achieve $\mathbb{E}\left[\|\mathcal{G}(\theta_{out})\|\right] \leq \varepsilon_1$, where in the last step we used the expression of $H = \frac{2}{1-\gamma}\log(1/\varepsilon)$. Moreover, if we translate this guarantee to a more standard stationarity measure $\mathbb{E}\left[\|\nabla_\theta F(\lambda(\theta))\|\right]$, the dependence on $(1-\gamma)^{-1}$ may further degrade for TSIVR-PG, see Lemma 5.4 in (Zhang et al., 2021b) where they establish $\mathbb{E}[\|\nabla_\theta F(\lambda(\theta))\|] = \mathcal{O}(\delta^{-1}\mathbb{E}\left[\|\mathcal{G}(\theta)\|\right])$ and $\delta = \mathcal{O}(H^{-1}) = \tilde{\mathcal{O}}(1-\gamma)$ is the truncation parameter. Indeed, their convergence result is stated in terms of the gradient mapping (because their algorithm uses a truncation mechanism with hyperparameter δ) and then translated to the standard first-order stationarity measure we use in

this work.

C. Further discussion of Assumption 4.10

A few comments are in order regarding Assumption 4.10:

1. In Assumption 4.10, the uniformity of the Lipschitz constant l (independent of θ) and $\bar{\epsilon}$ is important. Without this requirement, for instance, for item (iii), the existence of $\bar{\epsilon}_\theta > 0$ (depending on θ) with the desired property for every $\theta \in \mathbb{R}^d$ is always guaranteed since $\mathcal{V}_{\lambda(\theta)}$ is an open set.

2. As it was recently reported in Zhang et al. (2020), the direct parametrization satisfies the bijection Assumption 4.10. We refer the reader to Appendix H in (Zhang et al., 2020) for a complete proof of this fact. Notice also that Assumption 4.10 which was first introduced in (Zhang et al., 2021b) is a local version of Assumption 1 in (Zhang et al., 2020) and is hence less restrictive.

3. As for the softmax parametrization, verifying Assumption 4.10 is more challenging as we explain in the main part of the present paper. It is a delicate and interesting question to investigate whether Assumption 4.10 accommodates complex policy parametrizations such as practical neural networks or if it can be relaxed to do so.

D. Proof of Lemma 4.3 in Section 4.1

The proof of this lemma is simple and combines an elementary technical lemma from Zhang et al. (2021b) (Lemma 5.6) upper bounding the IS weights for the special case of the softmax parametrization with Lemma B.1 in Xu et al. (2020a) which provides a bound on the variance of IS weights which are bounded as a function of the squared euclidean distance between two policy parameters.

Proof. Using the softmax parametrization (2) with Assumption 4.1 satisfied, Lemma 5.6 in Zhang et al. (2021b) stipulates that for every $\theta, \theta' \in \mathbb{R}^d$ and every truncated trajectory $\tau = (s_t, a_t)_{0 \leq t \leq H-1}$ of length H, the IS weights defined in (9) satisfy:

$$w(\tau|\theta', \theta) \leq \exp\{2Hl_\psi \|\theta - \theta'\|\}\,. \tag{14}$$

It suffices to observe that $\|\theta_{t+1} - \theta_t\| = \alpha_t$ with the normalized update rule to prove that for every integer t and any trajectory τ of length H, we have $w(\tau|\theta_t, \theta_{t+1}) \leq \exp\{2Hl_\psi \alpha_t\}$. This proves the first part of the result.

Combining this first part with Lemma B.1 in Xu et al. (2020a), we obtain the desired fine-grained control of the IS weights variance. Specifically, if τ_{t+1} is a trajectory of length H generated following the initial distribution ρ and policy $\pi_{\theta_{t+1}}$, then

$$\mathbb{E}[w(\tau_{t+1}|\theta_t, \theta_{t+1})] = 1\,, \tag{15}$$

$$\mathrm{Var}\,[w(\tau_{t+1}|\theta_t, \theta_{t+1})] \leq C_w \alpha^2\,, \tag{16}$$

where $C_w \overset{\text{def}}{=} H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$. □

E. Proofs for Section 4.2: First-order stationarity

From the technical point of view, our proofs for the general utility setting depart from the proof techniques of Zhang et al. (2021b) in several ways although we share several common points. First, normalized policy gradient requires an adequate ascent-like lemma which is different from the analysis of gradient truncation mechanism. Second, our algorithm uses a different variance reduction scheme which does not require checkpoints and consists of a single loop. This requires careful changes in the proof. Compared to standard STORM variance reduction proofs in stochastic optimization (Cutkosky & Orabona, 2019), our setting involves two estimators which are intertwined, namely the state-action occupancy measure estimate and the stochastic policy gradient. This makes the analysis more complex and we refer the reader to the decomposition in (24) and the subsequent lemmas to observe this. In contrast, STORM only involves the stochastic gradient and uses a different update rule. More broadly, we believe the techniques we use here could also be useful for stochastic composite optimization.

E.1. Proof sketch

For the convenience of the reader, we highlight the main steps of the proof in this subsection before diving into the full detailed proof. The main steps consist in:

(a) showing an ascent-like lemma on the general utility function (see Lemma E.1). Notice that our algorithm is not a standard policy gradient algorithm but features normalization which requires a particular treatment;

(b) controlling the variance error due to two coupled stochastic estimates: the stochastic estimates of the state-action occupancy measures (for distinct policy parameters) and the stochastic policy gradients. Controlling these coupled estimates in our single-loop batch-free algorithm constitutes one of the main challenges of the proofs. More precisely, we use the following steps:

(i) We decompose the overall stochastic policy gradients errors in estimating the true policy gradients in (24) into two errors (see (25)): the error due to the state action occupancy measure estimation (which provides an estimate of the reward sequence) and the error due to the policy gradient given an estimate of the reward sequence.

(ii) We control each one of the aforementioned errors by establishing recursions in Lemma E.5 and Lemma E.8 respectively. Then, we solve the resulting recursions in the second parts of the lemmas.

(iii) We sum up each one of the expected errors over time in Lemma E.6 and Lemma E.6 and we obtain an estimation of the overall error in Lemma E.10 by combining both errors using Lemma E.4.

Further technical steps needed are described in the full proof (see for e.g. Lemma E.7).

(c) incorporating the estimates obtained in the second step to the descent lemma and telescoping the obtained inequality to derive our final convergence guarantee (see the proof of Theorem E.11 for the concluding steps).

E.2. Proof of Theorem 4.4 (General utilities setting)

In this section, we provide a proof for the case of general utilities. Notice that the case of cumulative rewards is a particular case. The first lemma is a an ascent-like lemma which follows from smoothness of the objective function. Before stating the lemma, we define the error sequence (e_t) for every integer t as follows:

$$e_t \stackrel{\text{def}}{=} d_t - \nabla_\theta F(\lambda_H(\theta_t))\,. \tag{17}$$

Lemma E.1. *Let Assumptions 4.1 and 4.2 hold true. Then, the sequence (θ_t) generated by Algorithm 1 and the sequence (e_t) satisfy for every integer $t \geq 0$,*

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_t)) + \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2\,. \tag{18}$$

Proof. Since the objective function $\theta \mapsto F(\lambda(\theta))$ is L_θ-smooth by Lemma H.1, we obtain the following by using the update rule of the sequence (θ_t):

$$\begin{aligned}
F(\lambda(\theta_{t+1})) &\geq F(\lambda(\theta_t)) + \langle \nabla_\theta F(\lambda(\theta_t)), \theta_{t+1} - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= F(\lambda(\theta_t)) + \alpha_t\langle \nabla_\theta F(\lambda(\theta_t)), \frac{d_t}{\|d_t\|}\rangle - \frac{L_\theta}{2}\alpha_t^2 \\
&= F(\lambda(\theta_t)) + \alpha_t\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle + \alpha_t\langle \nabla_\theta F(\lambda(\theta_t)) - \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle - \frac{L_\theta}{2}\alpha_t^2 \\
&\geq F(\lambda(\theta_t)) + \alpha_t\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle - \alpha_t\|\nabla_\theta F(\lambda(\theta_t)) - \nabla_\theta F(\lambda_H(\theta_t))\| - \frac{L_\theta}{2}\alpha_t^2 \\
&\stackrel{(a)}{\geq} F(\lambda(\theta_t)) + \alpha_t\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle - \alpha_t D_\lambda\gamma^H - \frac{L_\theta}{2}\alpha_t^2\,,
\end{aligned} \tag{19}$$

where (a) follows from Lemma H.2-(i).

Then we control the scalar product term. We distinguish two different cases:

Case 1: $\|e_t\| \leq \frac{1}{2}\|\nabla_\theta F(\lambda_H(\theta_t))\|$. In this case, we have

$$\begin{aligned}\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle &= \frac{1}{\|d_t\|}(\|\nabla_\theta F(\lambda_H(\theta_t))\|^2 + \langle \nabla_\theta F(\lambda_H(\theta_t)), e_t\rangle) \\ &\geq \frac{1}{\|d_t\|}(\|\nabla_\theta F(\lambda_H(\theta_t))\|^2 - \|\nabla_\theta F(\lambda_H(\theta_t))\| \cdot \|e_t\|) \\ &= \frac{1}{\|d_t\|}\|\nabla_\theta F(\lambda_H(\theta_t))\| \cdot (\|\nabla_\theta F(\lambda_H(\theta_t))\| - \|e_t\|) \\ &\geq \frac{1}{3}\|\nabla_\theta F(\lambda_H(\theta_t))\|,\end{aligned}$$

where the last inequality follows from observing that $\|d_t\| \leq \|e_t\| + \|\nabla_\theta F(\lambda_H(\theta_t))\| \leq \frac{3}{2}\|\nabla_\theta F(\lambda_H(\theta_t))\|$.

Case 2: $\|e_t\| \geq \frac{1}{2}\|\nabla_\theta F(\lambda_H(\theta_t))\|$. In this case, we simply have

$$\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle \geq -\|\nabla_\theta F(\lambda_H(\theta_t))\| \geq -2\|e_t\|.$$

Combining both cases, we obtain:

$$\begin{aligned}\alpha_t\langle \nabla_\theta F(\lambda_H(\theta_t)), \frac{d_t}{\|d_t\|}\rangle &\geq \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda_H(\theta_t))\| - 2\alpha_t\|e_t\| \\ &\geq \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - \frac{\alpha_t}{3}D_\lambda\gamma^H - 2\alpha_t\|e_t\|. \end{aligned} \tag{20}$$

Combining (19) and (20), we get

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_t)) + \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2,$$

which completes the proof. □

We now proceed with some preliminary results in order to control the error term $\|e_t\|$ in expectation.

The next lemma appeared in Prop. E.1 (Zhang et al., 2021b).

Remark E.2. With a slight abuse of notation, $\tau \sim \pi_\theta$ means that the trajectory τ (of length H) is sampled from the MDP controlled by the policy π_θ. We adopt this notation to highlight the dependence on the parametrized policy π_θ, the MDP being fixed in the problem formulation.

Lemma E.3. *For any reward vector $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have*

$$\mathbb{E}_{\tau\sim\pi_\theta}[\lambda(\tau)] = \lambda_H(\theta), \tag{21}$$

$$\mathbb{E}_{\tau\sim\pi_\theta}[g(\tau,\theta,r)] = [\nabla_\theta\lambda_H(\theta)]^T r. \tag{22}$$

In particular, under Assumption 4.2,

$$\mathbb{E}_{\tau\sim p(\cdot|\pi_\theta)}[g(\tau,\theta,\nabla_\lambda F(\lambda_H(\theta)))] = [\nabla_\theta\lambda_H(\theta)]^T\nabla_\lambda F(\lambda_H(\theta)) = \nabla_\theta F(\lambda_H(\theta)). \tag{23}$$

Proof. The proof follows from the definitions of the estimators $\lambda(\tau)$ and $g(\tau,\theta,r)$ in Eqs. (8)-(7) and the definition of the truncated state-action occupancy measure $\lambda_H(\theta)$ (see Eq.(1)) as well as the policy gradient theorem in Eq. (5). □

In view of controlling the error sequence (e_t), we first observe the following decomposition:

$$\begin{aligned}e_t &= d_t - \nabla_\theta F(\lambda_H(\theta_t)) \\ &= d_t - [\nabla_\theta\lambda_H(\theta_t)]^T r_{t-1} + [\nabla_\theta\lambda_H(\theta_t)]^T(r_{t-1} - \nabla_\lambda F(\lambda_H(\theta_t))) \\ &= d_t - [\nabla_\theta\lambda_H(\theta_t)]^T r_{t-1} + [\nabla_\theta\lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t))). \end{aligned} \tag{24}$$

Given the previous decomposition, we define two useful additional notations:

$$\hat{e}_t \overset{\text{def}}{=} d_t - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1}\,, \tag{25}$$

$$\tilde{e}_t \overset{\text{def}}{=} \lambda_t - \lambda_H(\theta_t)\,. \tag{26}$$

Using these notations, we establish the following result relating the error $\|e_t\|^2$ to the errors $\|\hat{e}_t\|^2$ and $\|\tilde{e}_t\|^2$ in expectation.

Lemma E.4. *Let Assumptions 4.1 and 4.2 hold true. Then we have for every integer $t \geq 1$,*

$$\mathbb{E}[\|e_t\|] \leq \mathbb{E}[\|\hat{e}_t\|] + C_1\,\mathbb{E}[\|\tilde{e}_{t-1}\|] + C_2\,\alpha_{t-1}\,, \tag{27}$$

where $C_1 \overset{def}{=} \frac{2L_\lambda^2 l_\psi}{(1-\gamma)^2}$ and $C_2 \overset{def}{=} \frac{2L_\lambda L_{\lambda,\infty} l_\psi}{(1-\gamma)^2}$.

Proof. It follows from the decomposition in (24) and the definitions (25)-(26) that

$$\mathbb{E}[\|e_t\|] \leq \mathbb{E}[\|\hat{e}_t\|] + \mathbb{E}[\|[\nabla_\theta \lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t)))\|]\,. \tag{28}$$

We now control the second term in the above inequality. The following step is similar to the treatment in (Zhang et al., 2021b)(Eq.(18)). Indeed, the policy gradient theorem (see (5)) yields

$$[\nabla_\theta \lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t))) = \mathbb{E}\left[\sum_{t'=0}^{H-1} \gamma^{t'} [\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t))]_{s_{t'},a_{t'}} \cdot \left(\sum_{h=0}^{t'} \nabla_\theta \log \pi_\theta(a_h, s_h)\right)\right]$$

As a consequence,

$$\|[\nabla_\theta \lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t)))\| \leq \mathbb{E}\left[\sum_{t'=0}^{H-1} \gamma^{t'} \|\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t))\|_\infty \left\|\sum_{h=0}^{t'} \nabla_\theta \log \pi_\theta(a_h, s_h)\right\|\right]\,. \tag{29}$$

Then, using Assumption 4.2, we have

$$\begin{aligned}
\|\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t))\|_\infty &\leq \|\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_{t-1}))\|_\infty + \|\nabla_\lambda F(\lambda_H(\theta_{t-1})) - \nabla_\lambda F(\lambda_H(\theta_t))\|_\infty \\
&\leq L_\lambda \|\lambda_{t-1} - \lambda_H(\theta_{t-1})\| + L_{\lambda,\infty} \|\lambda_H(\theta_{t-1}) - \lambda_H(\theta_t)\|_1 \\
&\leq L_\lambda \|\tilde{e}_{t-1}\| + L_{\lambda,\infty} \|\theta_t - \theta_{t-1}\|\,,
\end{aligned} \tag{30}$$

where the last inequality follows from Lemma H.1-(ii). Plugging this inequality in (29) and using Lemma H.1-(i) yields

$$\begin{aligned}
\|[\nabla_\theta \lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t)))\| &\leq \mathbb{E}\left[\sum_{t'=0}^{H-1} 2(t'+1) l_\psi \gamma^{t'} (L_\lambda \|\tilde{e}_{t-1}\| + L_{\lambda,\infty} \|\theta_t - \theta_{t-1}\|)\right] \\
&= \left(\sum_{t'=0}^{H-1} 2(t'+1) l_\psi \gamma^{t'} L_\lambda\right) (L_\lambda \mathbb{E}[\|\tilde{e}_{t-1}\|] + L_{\lambda,\infty} \alpha_{t-1}) \\
&\leq \frac{2L_\lambda l_\psi}{(1-\gamma)^2} (L_\lambda \mathbb{E}[\|\tilde{e}_{t-1}\|] + L_{\lambda,\infty} \alpha_{t-1})\,.
\end{aligned} \tag{31}$$

Hence, taking the total expectation, we obtain

$$\mathbb{E}[\|[\nabla_\theta \lambda_H(\theta_t)]^T(\nabla_\lambda F(\lambda_{t-1}) - \nabla_\lambda F(\lambda_H(\theta_t)))\|] \leq \frac{2L_\lambda^2 l_\psi}{(1-\gamma)^2} \mathbb{E}[\|\tilde{e}_{t-1}\|] + \frac{2L_\lambda L_{\lambda,\infty} l_\psi}{(1-\gamma)^2} \alpha_{t-1}\,. \tag{32}$$

Combining (28) and (32) yields the desired inequality. □

We now control each one of the errors in the right-hand side of the previous lemma in what follows. We start with the error $\tilde{e}_t$ (see (26)) induced by the (truncated) state-action occupancy measure estimation.

Lemma E.5. *Let Assumption 4.1 hold. Then, for every integer $t \geq 1$, if $\eta_t \in [0,1]$ we have*

$$\mathbb{E}[\|\tilde{e}_t\|^2] \leq (1-\eta_t)\mathbb{E}[\|\tilde{e}_{t-1}\|^2] + \frac{2C_w}{(1-\gamma)^2}\alpha_{t-1}^2 + \frac{2}{(1-\gamma)^2}\eta_t^2\,, \tag{33}$$

where we recall that $\tilde{e}_t = \lambda_t - \lambda_H(\theta_t)$ and $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$ as defined in Lemma 4.3. Moreover,

(i) if $\eta_t = \frac{2}{t+1}$, then for all integers $t \geq 1$, we have

$$\mathbb{E}[\|\tilde{e}_t\|] \leq \frac{4}{(1-\gamma)}\eta_t \cdot t^{1/2} + \frac{2C_w^{1/2}}{(1-\gamma)}\alpha_{t-1} \cdot t^{1/2}. \tag{34}$$

(ii) if $\eta_t = \left(\frac{2}{t+1}\right)^q$ and $\alpha_t = \alpha\left(\frac{2}{t+1}\right)^p$ for some reals $\alpha > 0$, $q \in (0,1)$, $p \geq 0$ and all integers $t \geq 1$, then we have

$$\mathbb{E}\left[\|\tilde{e}_t\|^2\right] \quad \leq \quad \frac{(2C+1)\eta_{t+1}}{(1-\gamma)^2} + \frac{2CC_w}{(1-\gamma)^2}\alpha_t^2\eta_{t+1}^{-1}\,, \tag{35}$$

where $C > 0$ is an absolute numerical constant.

Proof. We start with the proof of (33). Using the update rule of the sequence (λ_t) in Algorithm 1, we first derive a recursion on the error sequence $\tilde{e}_t$ from the following decomposition:

$$\begin{aligned}
\tilde{e}_t &= \lambda_t - \lambda_H(\theta_t) \\
&= \eta_t\lambda(\tau_t) + (1-\eta_t)(\lambda_{t-1} + u_t) - \lambda_H(\theta_t) \\
&= (1-\eta_t)\tilde{e}_{t-1} + (1-\eta_t)(\lambda_H(\theta_{t-1}) + u_t) + \eta_t(\lambda(\tau_t) - \lambda_H(\theta_t)) - (1-\eta_t)\lambda_H(\theta_t) \\
&= (1-\eta_t)\tilde{e}_{t-1} + (1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\,,
\end{aligned}$$

where

$$\tilde{y}_t \overset{\text{def}}{=} \lambda(\tau_t) - \lambda_H(\theta_t)\,, \tag{36}$$

$$\tilde{z}_t \overset{\text{def}}{=} u_t - (\lambda_H(\theta_t) - \lambda_H(\theta_{t-1})) = \lambda(\tau_t)(1 - w(\tau_t|\theta_{t-1},\theta_t)) - (\lambda_H(\theta_t) - \lambda_H(\theta_{t-1}))\,. \tag{37}$$

Using these notations, we have

$$\mathbb{E}[\|\tilde{e}_t\|^2] = (1-\eta_t)^2\mathbb{E}[\|\tilde{e}_{t-1}\|^2] + \mathbb{E}[\|(1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\|^2] + \mathbb{E}[\langle(1-\eta_t)\tilde{e}_{t-1}, (1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\rangle]\,. \tag{38}$$

Then, we notice that the scalar product term is equal to zero. We consider for this the filtration $(\mathcal{F}_t)$ of σ-algebras defined s.t. for every integer t, $\mathcal{F}_t \overset{\text{def}}{=} \sigma(\theta_k, \tau_k : k \leq t)$ where τ_t is a (random) trajectory of length H generated following the policy π_{θ_t}. This σ-algebra represents the history of all the random variables until time t. As a consequence, the random variable $\tilde{e}_t$ being $\mathcal{F}_{t-1}$-measurable, it follows from the tower property of the conditional expectation that

$$\begin{aligned}
\mathbb{E}[\langle(1-\eta_t)\tilde{e}_{t-1}, (1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\rangle] &= \mathbb{E}[\mathbb{E}[\langle(1-\eta_t)\tilde{e}_{t-1}, (1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\rangle|\mathcal{F}_{t-1}]] \\
&= \mathbb{E}[\langle(1-\eta_t)\tilde{e}_{t-1}, \mathbb{E}[(1-\eta_t)\tilde{z}_t + \eta_t\tilde{y}_t\rangle|\mathcal{F}_{t-1}]] \\
&= 0\,,
\end{aligned} \tag{39}$$

where the last step stems from the fact that $\mathbb{E}[\tilde{z}_t|\mathcal{F}_{t-1}] = \mathbb{E}[\tilde{y}_t|\mathcal{F}_{t-1}] = 0$, recall for this that $\tau_t \sim \pi_{\theta_t}$ for every t and see the definitions (36) and (37).

It follows from (38) and (39) that

$$\mathbb{E}[\|\tilde{e}_t\|^2] \leq (1-\eta_t)^2\mathbb{E}[\|\tilde{e}_{t-1}\|^2] + 2(1-\eta_t)^2\mathbb{E}[\|\tilde{z}_t\|^2] + 2\eta_t^2\mathbb{E}[\|\tilde{y}_t\|^2]\,. \tag{40}$$

Then, we upperbound each one of the last two terms in (40). As for the first term, since $\mathbb{E}[\tilde{z}_t] = 0$, we have the following

$$\mathbb{E}[\|\tilde{z}_t\|^2] \leq \mathbb{E}[\|u_t\|^2] = \mathbb{E}[\|\lambda(\tau_t)(1 - w(\tau_t|\theta_{t-1}, \theta_t))\|^2] = \mathbb{E}[(1 - w(\tau_t|\theta_{t-1}, \theta_t))^2 \|\lambda(\tau_t)\|^2]\,. \tag{41}$$

Given the definition of $\lambda(\tau_t)$ in (8), we first observe that with probability one,

$$\|\lambda(\tau_t)\| \leq \sum_{t=0}^{H-1} \gamma^t \|\delta_{s_t,a_t}\| = \sum_{t=0}^{H-1} \gamma^t \leq \frac{1}{1-\gamma}\,. \tag{42}$$

Using Lemma 4.3 together with the previous bound, we get

$$\mathbb{E}[(1-w(\tau_t|\theta_{t-1}, \theta_t))^2 \|\lambda(\tau_t)\|^2] \leq \frac{1}{(1-\gamma)^2}\mathbb{E}[(1-w(\tau_t|\theta_{t-1}, \theta_t))^2] = \frac{1}{(1-\gamma)^2} \operatorname{Var}\left[w(\tau_t|\theta_{t-1}, \theta_t)\right] \leq \frac{C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,. \tag{43}$$

We deduce from (41) and (43) together that

$$\mathbb{E}[\|\tilde{z}_t\|^2] \leq \frac{C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,. \tag{44}$$

Regarding the last term in (40), since $\mathbb{E}[\tilde{y}_t] = 0$, we observe that

$$\mathbb{E}[\|\tilde{y}_t\|^2] \leq \mathbb{E}[\|\lambda(\tau_t)\|^2] \leq \frac{1}{(1-\gamma)^2}\,, \tag{45}$$

where the last inequality stems from (42). Incorporating (44) and (45) into (40) leads to the following inequality

$$\mathbb{E}[\|\tilde{e}_t\|^2] \leq (1-\eta_t)^2\mathbb{E}[\|\tilde{e}_{t-1}\|^2] + \frac{2C_w}{(1-\gamma)^2}(1-\eta_t)^2\alpha_{t-1}^2 + \frac{2}{(1-\gamma)^2}\eta_t^2\,, \tag{46}$$

which concludes the proof of the first since $\eta_t \in [0, 1]$.

Proof of (34): In order to derive (34), we apply Lemma H.4 with $\eta_t = \frac{2}{t+1}$, $\beta_t = \frac{2}{(1-\gamma)^2}\eta_t^2 + \frac{2C_w}{(1-\gamma)^2}\alpha_{t-1}^2$. Using $\mathbb{E}[\|\tilde{e}_0\|^2] \leq \mathbb{E}[\|\lambda(\tau_t)\|^2] \leq \frac{1}{(1-\gamma)^2}$, we derive

$$\begin{aligned}
\mathbb{E}[\|\tilde{e}_t\|] &\leq \left(\mathbb{E}[\|\tilde{e}_t\|^2]\right)^{\frac{1}{2}} \leq \left(\frac{4}{(1-\gamma)^2(t+1)^2} + \frac{2}{(1-\gamma)^2}\eta_t^2 \cdot t + \frac{2C_w}{(1-\gamma)^2}\alpha_{t-1}^2 \cdot t\right)^{1/2} \\
&\leq \frac{2}{(1-\gamma)(t+1)} + \frac{2}{(1-\gamma)}\eta_t \cdot t^{1/2} + \frac{2C_w^{1/2}}{(1-\gamma)}\alpha_{t-1} \cdot t^{1/2} \\
&\leq \frac{4}{(1-\gamma)}\eta_t \cdot t^{1/2} + \frac{2C_w^{1/2}}{(1-\gamma)}\alpha_{t-1} \cdot t^{1/2}\,.
\end{aligned} \tag{47}$$

Proof of (35): Let $\eta_t = \left(\frac{2}{t+1}\right)^q$ for some $q \in (0, 1)$. In order to derive (35), we unroll the recursion (33) from $t = 1$ to $t = t'$ where $t' \leq T - 1$. Denoting $\beta_t = \frac{2}{(1-\gamma)^2}\eta_t^2 + \frac{2C_w}{(1-\gamma)^2}\alpha_{t-1}^2$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{e}_{t'}\|^2\right] &\leq \prod_{\tau=1}^{t'}(1-\eta_\tau)\mathbb{E}\left[\|\tilde{e}_0\|^2\right] + \sum_{t=1}^{t'}\beta_t \prod_{\tau=t+1}^{t'}(1-\eta_\tau) \\
&\leq \frac{\eta_{t'+1}}{(1-\gamma)^2} + C\beta_{t'+1}\eta_{t'+1}^{-1}\,,
\end{aligned} \tag{48}$$

where we used $\mathbb{E}\left[\|\tilde{e}_0\|^2\right] \leq \mathbb{E}\left[\|\lambda(\tau_t)\|^2\right] \leq \frac{1}{(1-\gamma)^2}$ and the results of Lemmas H.5-H.6 with $C > 1$ being a numerical constant.

□

Lemma E.6. *Let Assumption 4.1 hold. Let $\alpha_0 > 0$ and consider an integer $T \geq 1$. Set $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$ and $\alpha_t = \frac{\alpha_0}{T^{2/3}}$ for every nonzero integer $t \leq T$. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\tilde{e}_t\|] \leq \frac{C\left(1+C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}}, \tag{49}$$

where $C > 0$ is an absolute numerical constant.

Proof. Summing up inequality (35) from Lemma E.6 from $t = 1$ to $t = T$ and choosing $\alpha_t = \alpha = \frac{\alpha_0}{T^{2/3}}$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\tilde{e}_t\|\right] &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\tilde{e}_t\|^2\right]\right)^{1/2} \\
&\leq \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\tilde{e}_t\|^2\right]\right)^{1/2} \\
&\leq \left(\frac{1}{T}\sum_{t=1}^{T}\frac{\eta_{t+1}}{(1-\gamma)^2} + C\frac{2}{(1-\gamma)^2}\eta_{t+1} + \frac{2C_w}{(1-\gamma)^2}\alpha^2\eta_{t+1}^{-1}\right)^{1/2} \\
&\overset{(i)}{\leq} \left(\frac{3\eta_{T-1}}{(1-\gamma)^2} + \frac{6C\eta_{T-1}}{(1-\gamma)^2} + \frac{2CC_w}{(1-\gamma)^2}\frac{\alpha^2}{\eta_{T+1}}\right)^{1/2} \\
&\leq \left(\frac{9C\eta_{T-1}}{(1-\gamma)^2} + \frac{2CC_w}{(1-\gamma)^2}\frac{\alpha^2}{\eta_{T+1}}\right)^{1/2} \\
&\leq \left(\frac{18C}{(1-\gamma)^2T^{2/3}} + \frac{3CC_w}{(1-\gamma)^2}\frac{\alpha_0^2}{T^{2/3}}\right)^{1/2} \\
&\leq \frac{12C^{1/2}\left(1+C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}}\,.
\end{aligned} \tag{50}$$

where (i) follows from observing that

$$\sum_{t=1}^{T}\eta_{t+1} \leq 2^{\frac{2}{3}}\sum_{t=1}^{T}\frac{1}{(t+2)^{\frac{2}{3}}} \leq 2^{\frac{2}{3}}\int_{t=1}^{T}\frac{1}{(t+1)^{\frac{2}{3}}} \leq 3\cdot 2^{\frac{2}{3}}T^{\frac{1}{3}} = 3\,T\frac{2^{\frac{2}{3}}}{T^{\frac{2}{3}}} = 3\,T\,\eta_{T-1}. \tag{51}$$

□

In view of controlling the error $\hat{e}_t$, we first state a technical lemma that will be useful. This result controls the expected squared difference between two consecutive estimates of the (truncated) state-action occupancy measure.

Lemma E.7. *Suppose Assumption 4.1 holds. Then for all integers $t \geq 1$,*

$$\mathbb{E}[\|\lambda_{t-1} - \lambda_t\|^2] \leq \frac{3\eta_t^2}{(1-\eta_t)^2}\mathbb{E}[\|\tilde{e}_t\|^2] + \frac{3\eta_t^2}{(1-\eta_t)^2(1-\gamma)^2} + \frac{3C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,, \tag{52}$$

where $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$ as defined in Lemma 4.3. Moreover, if in addition $\eta_t = \left(\frac{2}{t+1}\right)^q$ for some $q \in [0,1)$ then for every integer $t \geq 1$,

$$\mathbb{E}[\|\lambda_{t-1} - \lambda_t\|^2] \leq \frac{12C\eta_t^2}{(1-\gamma)^2} + \frac{6C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,, \tag{53}$$

where $C > 0$ is a numerical constant.

Proof. Using the update rule of the truncated occupancy measure estimate sequence (λ_t), we have

$$\begin{aligned}\lambda_{t-1} - \lambda_t &= \lambda_{t-1} - [\eta_t \lambda(\tau_t) + (1-\eta_t)(\lambda_{t-1} + u_t)] \\ &= \eta_t(\lambda_{t-1} - \lambda(\tau_t)) - (1-\eta_t)u_t \\ &= \eta_t(\lambda_{t-1} - \lambda_t) + \eta_t(\lambda_t - \lambda(\tau_t)) - (1-\eta_t)u_t\,.\end{aligned}$$

As a consequence, we have

$$\begin{aligned}\lambda_{t-1} - \lambda_t &= \frac{\eta_t}{1-\eta_t}(\lambda_t - \lambda(\tau_t)) - u_t \\ &= \frac{\eta_t}{1-\eta_t}\tilde{e}_t + \frac{\eta_t}{1-\eta_t}(\lambda_H(\theta_t) - \lambda(\tau_t)) - u_t\,. \end{aligned} \tag{54}$$

Taking expectation of the square of the previous identity, we obtain the following bound:

$$\mathbb{E}[\|\lambda_{t-1} - \lambda_t\|^2] \leq \frac{3\eta_t^2}{(1-\eta_t)^2}\mathbb{E}[\|\tilde{e}_t\|^2] + \frac{3\eta_t^2}{(1-\eta_t)^2}\mathbb{E}[\|\lambda(\tau_t) - \lambda_H(\theta_t)\|^2] + 3\mathbb{E}[\|u_t\|^2]\,. \tag{55}$$

Recall now that $\mathbb{E}[\|u_t\|^2] \leq \frac{C_w}{(1-\gamma)^2}\alpha_{t-1}^2$ and $\mathbb{E}[\|\lambda(\tau_t) - \lambda_H(\theta_t)\|^2] \leq \frac{1}{(1-\gamma)^2}$ from (41)-(44) and (45) respectively. Incorporating these bounds into (55) yields:

$$\mathbb{E}[\|\lambda_{t-1} - \lambda_t\|^2] \leq \frac{3\eta_t^2}{(1-\eta_t)^2}\mathbb{E}[\|\tilde{e}_t\|^2] + \frac{3\eta_t^2}{(1-\eta_t)^2(1-\gamma)^2} + \frac{3C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,.$$

This completes the proof of (52). We now set $\eta_t = \left(\frac{2}{t+1}\right)^q$ for some $q \in (0,1)$. By (76) in the proof of Lemma E.5, we have

$$\mathbb{E}\left[\|\tilde{e}_t\|^2\right] \quad \leq \quad \frac{\eta_t}{(1-\gamma)^2} + C\beta_t\eta_t^{-1}\,, \tag{56}$$

where $\beta_t = \frac{2}{(1-\gamma)^2}\eta_t^2 + \frac{2C_w}{(1-\gamma)^2}\alpha_{t-1}^2$, and $C > 0$ is a numerical constant. Thus,

$$\begin{aligned}\mathbb{E}[\|\lambda_{t-1} - \lambda_t\|^2] &\leq \frac{3\eta_t^2}{(1-\eta_t)^2}\left(\frac{\eta_t}{(1-\gamma)^2} + C\beta_t\eta_t^{-1}\right) + \frac{3\eta_t^2}{(1-\eta_t)^2(1-\gamma)^2} + \frac{3C_w}{(1-\gamma)^2}\alpha_{t-1}^2 \\ &\leq \frac{12C\eta_t^2}{(1-\gamma)^2} + \frac{6C_w}{(1-\gamma)^2}\alpha_{t-1}^2\,.\end{aligned}$$

□

We are now ready to prove a recursive upper bound on the error sequence $(\hat{e}_t)$ defined in (25). Notice that this result is of the same flavor as Lemma E.5 which we already proved. In particular, the result illustrates a variance reduction effect stemming from the variance reduction updates used for both the stochastic policy gradients and the state-action occupancy measure estimates.

Lemma E.8. *Suppose Assumptions 4.1 and 4.2 hold. Then, for every integer $t \geq 2$,*

$$\mathbb{E}[\|\hat{e}_t\|^2] \leq (1-\eta_t)^2\mathbb{E}[\|\hat{e}_{t-1}\|^2] + C_3\eta_{t-1}^2 + C_4\alpha_{t-2}^2, \tag{57}$$

where $C_3 \overset{def}{=} \frac{288Cl_\psi^2L_\lambda^2}{(1-\gamma)^6} + \frac{32l_\lambda^2l_\psi^2}{(1-\gamma)^4}$, $C_4 \overset{def}{=} \frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_wl_\psi^2]}{(1-\gamma)^4} + \frac{144C_wl_\psi^2L_\lambda^2}{(1-\gamma)^6}$, and $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$ as defined in Lemma 4.3. Moreover,

(i) if $\eta_t = \frac{2}{t+1}$, then for all integers $t \geq 1$, we have

$$\mathbb{E}[\|\hat{e}_t\|] \quad \leq \quad \frac{2\hat{E}}{t+1} + 2C_3^{1/2}\eta_t \cdot t^{1/2} + C_4^{1/2}\alpha_{t-2} \cdot t^{1/2}\,. \tag{58}$$

(ii) if $\eta_t = \left(\frac{2}{t+1}\right)^q$ *and* $\alpha_t = \alpha\left(\frac{2}{t+1}\right)^p$ *for some reals* $\alpha > 0$, $q \in (0,1)$, $p \geq 0$ *and all integers* $t \geq 1$, *then we have*

$$\mathbb{E}\left[\|\hat{e}_t\|^2\right] \quad \leq \quad \hat{E}^2\eta_{t+1} + 2CC_3\eta_{t+1} + CC_4\alpha_{t-1}^2\eta_{t+1}^{-1}, \tag{59}$$

where $\hat{E} = \frac{4l_\lambda l_\psi}{(1-\gamma)^2}$.

Proof. The first step of the proof consists in decomposing the error $\hat{e}_t$ in a suitable way using the update rule of the sequence (d_t) so that for every integer $t \geq 2$,

$$\begin{aligned}
\hat{e}_t &= d_t - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} \\
&= (1-\eta_t)(d_{t-1} + v_t) + \eta_t g(\tau_t, \theta_t, r_{t-1}) - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} \\
&= (1-\eta_t)(\hat{e}_{t-1} + [\nabla_\theta \lambda_H(\theta_{t-1})]^T r_{t-2} + v_t) + \eta_t(g(\tau_t, \theta_t, r_{t-1}) - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1}) - (1-\eta_t)[\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} \\
&= (1-\eta_t)\hat{e}_{t-1} + (1-\eta_t)\hat{z}_t + \eta_t \hat{y}_t \,,
\end{aligned}$$

where

$$\hat{y}_t \stackrel{\text{def}}{=} g(\tau_t, \theta_t, r_{t-1}) - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} \,, \tag{60}$$

$$\hat{z}_t \stackrel{\text{def}}{=} v_t - ([\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} - [\nabla_\theta \lambda_H(\theta_{t-1})]^T r_{t-2}) \tag{61}$$

$$= g(\tau_t, \theta_t, r_{t-1}) - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1} + [\nabla_\theta \lambda_H(\theta_{t-1})]^T r_{t-2} - w(\tau_t|\theta_{t-1}, \theta_t) g(\tau_t, \theta_{t-1}, r_{t-2}) \,. \tag{62}$$

Then we use similar derivations to (38) and (39). We consider again the same filtration $(\mathcal{F}_t)$ of σ-algebras where $\mathcal{F}_t$ represents the randomness until time t (including time t and random trajectories τ_t of length H generated by following the policy π_{θ_t}). Therefore, we have (see Lemma E.3)

$$\mathbb{E}[\hat{y}_t|\mathcal{F}_{t-1}] = 0 \,; \quad \mathbb{E}[\hat{z}_t|\mathcal{F}_{t-1}] = 0 \,. \tag{63}$$

Note here as a comment that the reason why we used r_{t-1} instead of r_t in Algorithm 1 (for the sequence (v_t)) becomes clearer here in the previous identities: r_{t-1} is $\mathcal{F}_{t-1}$- measurable unlike r_t and this allows to obtain a null conditional expectation avoiding in particular dependency issues between r_t and θ_t.

Employing (63) and using the same derivations as in (39) leads to

$$\begin{aligned}
\mathbb{E}[\|\hat{e}_t\|^2] &= \mathbb{E}[\|(1-\eta_t)\hat{e}_{t-1} + (1-\eta_t)\hat{z}_t + \eta_t\hat{y}_t\|^2] \\
&= (1-\eta_t)^2\mathbb{E}[\|\hat{e}_{t-1}\|^2] + \mathbb{E}[\|(1-\eta_t)\hat{z}_t + \eta_t\hat{y}_t\|^2] \\
&\leq (1-\eta_t)^2\mathbb{E}[\|\hat{e}_{t-1}\|^2] + 2(1-\eta_t)^2\mathbb{E}[\|\hat{z}_t\|^2] + 2\eta_t^2\mathbb{E}[\|\hat{y}_t\|^2] \,.
\end{aligned} \tag{64}$$

We now derive a bound for the term $\mathbb{E}[\|\hat{z}_t\|^2]$. Observe for this that

$$\begin{aligned}
\mathbb{E}[\|\hat{z}_t\|^2] &\leq \mathbb{E}[\|g(\tau_t, \theta_t, r_{t-1}) - w(\tau_t|\theta_{t-1}, \theta_t) g(\tau_t, \theta_{t-1}, r_{t-2})\|^2] \\
&= \mathbb{E}[\|g(\tau_t, \theta_t, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-1}) + g(\tau_t, \theta_{t-1}, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-2}) \\
&\qquad + g(\tau_t, \theta_{t-1}, r_{t-2})(1 - w(\tau_t|\theta_{t-1}, \theta_t))\|^2] \\
&\leq 3\mathbb{E}[\|g(\tau_t, \theta_t, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-1})\|^2] + 3\mathbb{E}[\|g(\tau_t, \theta_{t-1}, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-2})\|^2] \\
&+ 3\mathbb{E}[(1 - w(\tau_t|\theta_{t-1}, \theta_t))^2\|g(\tau_t, \theta_{t-1}, r_{t-2})\|^2] \,.
\end{aligned} \tag{65}$$

Each term in this inequality is upper bounded separately in what follows.

Term 1: $\mathbb{E}[\|g(\tau_t, \theta_t, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-1})\|^2]$**.** Using Lemma H.3-(ii), we obtain

$$\begin{aligned}
\|g(\tau_t, \theta_t, r_{t-1}) - g(\tau_t, \theta_{t-1}, r_{t-1})\| &\leq \frac{2(l_\psi^2 + L_\psi)}{(1-\gamma)^2}\|r_{t-1}\|_\infty \cdot \|\theta_t - \theta_{t-1}\| \\
&= \frac{2(l_\psi^2 + L_\psi)}{(1-\gamma)^2}\|\nabla_\lambda F(\lambda_{t-1})\|_\infty \cdot \alpha_{t-1} \\
&\leq \frac{2(l_\psi^2 + L_\psi)l_\lambda}{(1-\gamma)^2}\alpha_{t-1} \,,
\end{aligned} \tag{66}$$

where the last inequality stems from Assumption 4.2. We deduce from this the bound for the first term:

$$\mathbb{E}[\|g(\tau_t,\theta_t,r_{t-1})-g(\tau_t,\theta_{t-1},r_{t-1})\|^2] \leq \frac{4(l_\psi^2+L_\psi)^2 l_\lambda^2}{(1-\gamma)^4}\alpha_{t-1}^2\,. \tag{67}$$

Term 2: $\mathbb{E}[\|g(\tau_t,\theta_{t-1},r_{t-1})-g(\tau_t,\theta_{t-1},r_{t-2})\|^2]$**.** Together with Assumption 4.2, Lemma H.3-(i) yields

$$\begin{aligned}\|g(\tau_t,\theta_{t-1},r_{t-1})-g(\tau_t,\theta_{t-1},r_{t-2})\| &\leq \frac{2l_\psi}{(1-\gamma)^2}\|r_{t-1}-r_{t-2}\|_\infty \\ &= \frac{2l_\psi}{(1-\gamma)^2}\|\nabla_\lambda F(\lambda_{t-1})-\nabla_\lambda F(\lambda_{t-2})\|_\infty \\ &\leq \frac{2l_\psi L_\lambda}{(1-\gamma)^2}\|\lambda_{t-1}-\lambda_{t-2}\|\,.\end{aligned} \tag{68}$$

Invoking Lemma E.7, we obtain from (68)

$$\begin{aligned}\mathbb{E}[\|g(\tau_t,\theta_{t-1},r_{t-1})-g(\tau_t,\theta_{t-1},r_{t-2})\|^2] &\leq \frac{4l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\mathbb{E}[\|\lambda_{t-1}-\lambda_{t-2}\|^2] \\ &\leq \frac{4l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\left(\frac{12C\eta_{t-1}^2}{(1-\gamma)^2}+\frac{6C_w}{(1-\gamma)^2}\alpha_{t-2}^2\right) \\ &\leq \frac{48l_\psi^2 L_\lambda^2}{(1-\gamma)^6}\left(C\eta_{t-1}^2+C_w\alpha_{t-2}^2\right)\,.\end{aligned} \tag{69}$$

Term 3: $\mathbb{E}[(1-w(\tau_t|\theta_{t-1},\theta_t))^2\|g(\tau_t,\theta_{t-1},r_{t-2})\|^2]$**.**

First, observe that

$$\begin{aligned}\|g(\tau_t,\theta_{t-1},r_{t-2})\| &\overset{(a)}{=} \left\|\sum_{t=0}^{H-1}\left(\sum_{h=t}^{H-1}\gamma^h r_{t-2}(s_h,a_h)\right)\nabla\log\pi_\theta(a_t|s_t)\right\| \\ &\leq \sum_{t=0}^{H-1}\sum_{h=t}^{H-1}\gamma^h\|r_{t-2}\|_\infty\cdot\|\nabla\log\pi_\theta(a_t|s_t)\| \\ &\overset{(b)}{\leq} l_\lambda\sum_{t=0}^{H-1}\sum_{h=t}^{H-1}\gamma^h\|\nabla\log\pi_\theta(a_t|s_t)\| \\ &\overset{(c)}{\leq} 2l_\lambda l_\psi\sum_{t=0}^{H-1}\sum_{h=t}^{H-1}\gamma^h \\ &= 2l_\lambda l_\psi\sum_{h=0}^{H-1}\sum_{t=0}^{h}\gamma^h \\ &\leq 2l_\lambda l_\psi\sum_{h=0}^{H-1}(h+1)\gamma^h \\ &\leq \frac{2l_\lambda l_\psi}{(1-\gamma)^2}\,,\end{aligned} \tag{70}$$

where (a) follows from the expression of the stochastic policy gradient (7), (b) stems from Assumption 4.2 and (c) is a consequence of Lemma H.1-(i).

Using Lemma 4.3 together with the previous bound yields

$$\begin{aligned}\mathbb{E}[(1-w(\tau_t|\theta_{t-1},\theta_t))^2\|g(\tau_t,\theta_{t-1},r_{t-2})\|^2] &\leq \frac{4l_\lambda^2 l_\psi^2}{(1-\gamma)^4}\mathbb{E}[(1-w(\tau_t|\theta_{t-1},\theta_t))^2] \\ &= \frac{4l_\lambda^2 l_\psi^2}{(1-\gamma)^4}\operatorname{Var}\left[w(\tau_t|\theta_{t-1},\theta_t)\right] \\ &\leq \frac{4l_\lambda^2 l_\psi^2 C_w}{(1-\gamma)^4}\alpha_{t-1}^2\,. \end{aligned} \tag{71}$$

Collecting (67), (69) and (71) in (65), we obtain

$$\begin{aligned}\mathbb{E}[\|\hat{z}_t\|^2] &\leq \frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_w l_\psi^2]}{(1-\gamma)^4}\alpha_{t-1}^2 + \frac{144 l_\psi^2 L_\lambda^2}{(1-\gamma)^6}\left(C\eta_{t-1}^2 + C_w\alpha_{t-2}^2\right) \\ &\leq \frac{144 C l_\psi^2 L_\lambda^2}{(1-\gamma)^6}\eta_{t-1}^2 + \left(\frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_w l_\psi^2]}{(1-\gamma)^4} + \frac{144 C_w l_\psi^2 L_\lambda^2}{(1-\gamma)^6}\right)\alpha_{t-2}^2\,. \end{aligned} \tag{72}$$

We now bound the term $\mathbb{E}[\|\hat{y}_t\|^2]$ in (64). First, recall from (60) that $\hat{y}_t = g(\tau_t,\theta_t,r_{t-1}) - [\nabla_\theta \lambda_H(\theta_t)]^T r_{t-1}$. Then, with probability one,

$$\begin{aligned}\|\hat{y}_t\| &\leq \|g(\tau_t,\theta_t,r_{t-1})\| + \|[\nabla_\theta\lambda_H(\theta_t)]^T r_{t-1}\| \\ &\overset{(a)}{\leq} \frac{2l_\lambda l_\psi}{(1-\gamma)^2} + \|[\nabla_\theta\lambda_H(\theta_t)]^T r_{t-1}\| \\ &\overset{(b)}{\leq} \frac{4l_\lambda l_\psi}{(1-\gamma)^2}\,, \end{aligned} \tag{73}$$

where (a) stems from the same bound as in (70) and (b) also follows from a similar bound to (70). Indeed, notice using (5) that

$$\begin{aligned}\|[\nabla_\theta\lambda_H(\theta_t)]^T r_{t-1}\| &= \left\|\mathbb{E}\left[\sum_{t'=0}^{H-1}\gamma^{t'} r_{t-1}(s_t',a_t')\left(\sum_{h=0}^{t'}\nabla\log\pi_\theta(a_h|s_h)\right)\right]\right\| \\ &\leq \mathbb{E}\left[\sum_{t'=0}^{H-1}\gamma^{t'}\|r_{t-1}\|_\infty\sum_{h=0}^{t'}\|\nabla\log\pi_\theta(a_h|s_h)\|\right] \\ &\overset{(a)}{\leq} 2l_\lambda l_\psi\sum_{t'=0}^{H-1}(t'+1)\gamma^{t'} \\ &\leq \frac{2l_\lambda l_\psi}{(1-\gamma)^2}\,, \end{aligned} \tag{74}$$

where again (a) stems from Assumption 4.2 and Lemma H.1-(i).

We conclude from (64), (72) and (73) that

$$\mathbb{E}[\|\hat{e}_t\|^2] \leq (1-\eta_t)^2\mathbb{E}[\|\hat{e}_{t-1}\|^2] + \left(\frac{288 C l_\psi^2 L_\lambda^2}{(1-\gamma)^6} + \frac{32 l_\lambda^2 l_\psi^2}{(1-\gamma)^4}\right)\eta_{t-1}^2 + \left(\frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_w l_\psi^2]}{(1-\gamma)^4} + \frac{144 C_w l_\psi^2 L_\lambda^2}{(1-\gamma)^6}\right)\alpha_{t-2}^2\,,$$

where $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$ as defined in Lemma 4.3.

Proof of (58): In order to derive (58), we apply Lemma H.4 with $\eta_t = \frac{2}{t+1}$, $\beta_t = C_3\eta_{t-1}^2 + C_4\alpha_{t-2}^2$. Using $\mathbb{E}[\|\hat{e}_0\|^2] \leq \hat{E}^2$, we derive

$$\begin{aligned}\mathbb{E}[\|\hat{e}_t\|] &\leq \left(\mathbb{E}[\|\hat{e}_t\|^2]\right)^{\frac{1}{2}} \leq \left(\frac{4\hat{E}^2}{(t+1)^2} + 4C_3\eta_t^2 \cdot t + C_4\alpha_{t-2}^2 \cdot t\right)^{1/2} \\ &\leq \frac{2\hat{E}}{t+1} + 2C_3^{1/2}\eta_t \cdot t^{1/2} + C_4^{1/2}\alpha_{t-2} \cdot t^{1/2}\,. \end{aligned} \tag{75}$$

Proof of (59): Let $\eta_t = \left(\frac{2}{t+1}\right)^q$ for some $q \in (0,1)$. In order to derive (59), we unroll the recursion from $t = 1$ to $t = t' \leq T$. Denoting $\beta_t = C_3\eta_{t-1}^2 + C_4\alpha_{t-2}^2$, we derive

$$\begin{aligned}\mathbb{E}\left[\|\hat{e}_{t'}\|^2\right] &\leq \left(\prod_{\tau=1}^{t'}(1-\eta_\tau)\right)\mathbb{E}\left[\|\hat{e}_0\|^2\right] + \sum_{t=1}^{t'}\beta_t\prod_{\tau=t+1}^{t'}(1-\eta_\tau) \\ &\leq \hat{E}^2\eta_{t'+1} + C\beta_{t'+1}\eta_{t'+1}^{-1} \\ &\leq \hat{E}^2\eta_{t'+1} + CC_3\eta_{t'}^2\eta_{t'+1}^{-1} + CC_4\alpha_{t'-1}^2\eta_{t'+1}^{-1} \\ &\leq \hat{E}^2\eta_{t'+1} + 2CC_3\eta_{t'+1} + CC_4\alpha_{t'-1}^2\eta_{t'+1}^{-1}\cdot\,, \end{aligned} \tag{76}$$

where we used the results of Lemmas H.5-H.6 and $\mathbb{E}\left[\|\hat{e}_0\|^2\right] \leq \hat{E}^2$ with $\hat{E} = \frac{4l_\lambda l_\psi}{(1-\gamma)^2}$, which can be derived similarly to (73).

□

In the next lemma, we derive an estimate of the average expected error $\mathbb{E}[\|\hat{e}_t\|]$ from the recursion we have just established in Lemma E.8.

Lemma E.9. *Suppose Assumptions 4.1 and 4.2 hold. Let $T \geq 1$ be an integer, let $\alpha_0 > 0$ and set $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$, $\alpha_t = \frac{\alpha_0}{T^{2/3}}$ for every integer t. Then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{e}_t\|\right] \leq \frac{C\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}}, \tag{77}$$

where $\hat{E} = \frac{4l_\lambda l_\psi}{(1-\gamma)^2}$, $C_3 = \frac{288Cl_\psi^2L_\lambda^2}{(1-\gamma)^6} + \frac{32l_\lambda^2l_\psi^2}{(1-\gamma)^4}$, $C_4 = \frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_wl_\psi^2]}{(1-\gamma)^4} + \frac{144C_wl_\psi^2L_\lambda^2}{(1-\gamma)^6}$, $C_w = H((8H+2)l_\psi^2+2L_\psi)(W+1)$ as defined in Lemma 4.3, and $C > 1$ is a numerical constant.

Proof. Summing up inequality (59) from Lemma E.8 from $t = 1$ to $t = T$ and choosing $\alpha_t = \alpha = \frac{\alpha_0}{T^{2/3}}$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{e}_t\|\right] &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\hat{e}_t\|^2\right]\right)^{1/2} \\ &\leq \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{e}_t\|^2\right]\right)^{1/2} \\ &\leq \left(\frac{1}{T}\sum_{t=1}^{T}\hat{E}^2\eta_{t+1} + 2CC_3\eta_{t+1} + CC_4\alpha^2\eta_{t+1}^{-1}\right)^{1/2} \\ &\overset{(i)}{\leq} \left(3(\hat{E}^2 + 2CC_3)\eta_{T-1} + CC_3\alpha^2\eta_{T+1}^{-1}\right)^{1/2} \\ &\leq \left(\frac{12(\hat{E}^2+CC_3)}{T^{2/3}} + \frac{2CC_4\alpha_0^2}{T^{2/3}}\right)^{1/2} \\ &\leq \frac{4C^{1/2}\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}}, \end{aligned}$$

where (i) holds by (51). □

We are now ready to state the main lemma controlling the error sequence (e_t) in expectation as defined in (17).

Lemma E.10. *Suppose Assumptions 4.1 and 4.2 hold. Let $T \geq 1$ be an integer, let $\alpha_0 > 0$ and set $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$, $\alpha_t = \frac{\alpha_0}{T^{2/3}}$ for every integer t. Then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|e_t\|] \leq \frac{C\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}} + \frac{CC_1\left(1 + C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}} + \frac{C_2\alpha_0}{T^{2/3}}, \tag{78}$$

where $C_1 = \frac{2L_\lambda^2 l_\psi}{(1-\gamma)^2}$ and $C_2 = \frac{2L_\lambda L_{\lambda,\infty} l_\psi}{(1-\gamma)^2}$, $C_3 = \frac{288 C l_\psi^2 L_\lambda^2}{(1-\gamma)^6} + \frac{32 l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$, $C_4 = \frac{12 l_\lambda^2[(l_\psi^2 + L_\psi)^2 + C_w l_\psi^2]}{(1-\gamma)^4} + \frac{144 C_w l_\psi^2 L_\lambda^2}{(1-\gamma)^6}$, C is a numerical constant and $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$ as defined in Lemma 4.3.

Proof. By Lemma E.5 and E.9 we have the bounds

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{e}_t\|\right] \leq \frac{C\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}}, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\tilde{e}_t\|] \leq \frac{C\left(1 + C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}}. \tag{79}$$

Summing up the result of Lemma E.4 from $t = 1$ to $t = T$ and using the above bounds, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|e_t\|] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\hat{e}_t\|] + \frac{C_1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\tilde{e}_{t-1}\|] + \frac{C_2}{T}\sum_{t=1}^{T}\alpha_{t-1} \\ &\leq \frac{C\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}} + \frac{CC_1\left(1 + C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}} + \frac{C_2\alpha_0}{T^{2/3}}.\end{aligned} \tag{80}$$

□

End of the proof of Theorem 4.4. We conclude the proof of Theorem 4.4 which is first recalled in the following for the convenience of the reader.

Theorem E.11. *Let Assumptions 4.1 and 4.2 hold. Let $\alpha_0 > 0$ and consider an integer $T \geq 1$. Set $\alpha_t = \frac{\alpha_0}{T^{2/3}}$, $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Let $\bar{\theta}_T$ be sampled from the iterates $\{\theta_1, \cdots, \theta_T\}$ of Algorithm 1 uniformly at random. Then, we have*

$$\mathbb{E}\left[\left\|\nabla_\theta F(\lambda(\bar{\theta}_T))\right\|\right] \leq \mathcal{O}\left(\frac{1 + (1-\gamma)^3\Delta\alpha_0^{-1} + (1-\gamma)^{-1}\alpha_0}{(1-\gamma)^3 T^{1/3}}\right). \tag{81}$$

If moreover $\alpha_0 = (1-\gamma)^2\sqrt{\Delta}$, then $\mathbb{E}\left[\left\|\nabla_\theta F(\lambda(\bar{\theta}_T))\right\|\right] \leq \mathcal{O}\left(\frac{1+(1-\gamma)\sqrt{\Delta}}{(1-\gamma)^3 T^{1/3}}\right)$.

Proof. By Lemma E.1, we have for every integer t,

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_t)) + \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2. \tag{82}$$

Setting constant step-size $\alpha_t = \alpha = \frac{\alpha_0}{T^{2/3}}$, taking expectation, telescoping and rearranging, we get

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_\theta F(\lambda(\theta_t))\|\right] &\leq \frac{3(F^\star - F(\lambda(\theta_1)))}{\alpha T} + \frac{6}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|\right] + \frac{3L_\theta\alpha}{2} + 4D_g\gamma^H \\ &\leq \frac{3(F^\star - F(\lambda(\theta_1)))}{\alpha_0 T^{1/3}} + \frac{6C\left(\hat{E} + C_3^{1/2} + C_4^{1/2}\alpha_0\right)}{T^{1/3}} + \frac{6CC_1\left(1 + C_w^{1/2}\alpha_0\right)}{(1-\gamma)}\frac{1}{T^{1/3}} \\ &\quad + \frac{C_2\alpha_0}{T^{2/3}} + \frac{3L_\theta\alpha_0}{2T^{2/3}} + 4D_g\gamma^H \\ &= \mathcal{O}\left(\frac{(1-\gamma)^{-3} + \Delta\alpha_0^{-1} + \alpha_0(1-\gamma)^{-4}}{T^{1/3}}\right), \end{aligned} \tag{83}$$

where we set $H = (1-\gamma)^{-1}\log(T)$. Setting $\alpha_0 = (1-\gamma)^2\sqrt{\Delta}$, we obtain the desired result. □

E.3. Proof of Corollary 4.7 (Cumulative reward setting)

For this particular case, we redefine the error sequence (e_t) by overloading the notation since it plays a similar role. Define the error sequence (e_t) for every integer t as follows:

$$e_t \stackrel{\text{def}}{=} d_t - \nabla J_H(\theta_t)\,, \tag{84}$$

where the truncated cumulative reward $J_H(\theta)$ is defined as follows for any policy parameter $\theta \in \mathbb{R}^d$:

$$J_H(\theta) = \mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t r(s_t, a_t)\right]. \tag{85}$$

We start by stating a complete version of Corollary 4.7 which we shall prove in this section.

Corollary E.12 (FOS convergence of N-VR-PG). *Let Assumptions 4.1 and 4.2 hold. Let $\alpha_0 > 0$ and let T be an integer larger than 1. Set $\alpha_t = \frac{\alpha_0}{T^{2/3}}$, $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Let $\bar{\theta}_T$ be sampled from the iterates $\{\theta_1, \cdots, \theta_T\}$ of N-VR-PG (Algorithm 4) uniformly at random. Then we have*

$$\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] \leq \mathcal{O}\left(\frac{J^* - J(\theta_1)}{\alpha_0 T^{1/3}} + \frac{V + (L_g + GC_w^{1/2})\alpha_0}{T^{1/3}}\right), \tag{86}$$

where V, L_g, G, and C_w are defined in Lemma 4.3, H.3, and E.13. Moreover, if we set $\alpha_0 = 1-\gamma$, then

$$\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2 T^{1/3}}\right).$$

The proof of this result follows the same lines as the proof of Theorem 4.4 which addresses the more general setting of general utilities. In the special case of cumulative rewards, recall that the estimation of the state-action occupancy measure is not required. In the following, we provide for clarity the intermediate results required to prove our result, mirroring the proof of the more general result of Theorem 4.4.

In order to derive the improved dependence on the $(1-\gamma)$ factor in the final rate compared to Theorem 4.4, we will apply the following results from Yuan et al. (2022, Lemma 4.2, (68) and Lemma 4.4, (19)) and (Xu et al., 2020a, Proposition 4.2 (1) and (3)), which offer a tighter dependence on $1-\gamma$ in the standard cumulative reward setting. We use the notation $g(\tau, \theta)$ instead of $g(\tau, \theta, r)$ in this simpler standard RL setting.

Lemma E.13. *Let Assumption 4.1 hold true and let $\tau = \{s_0, a_0, \cdots, s_{H-1}, a_{H-1}\}$ be an arbitrary trajectory of length H. Then the following statements hold:*

(i) The objective function $\theta \mapsto J(\theta)$ is L_θ-smooth with $L_\theta \stackrel{def}{=} \frac{2\|r\|_\infty(L_\psi + 3l_\psi^2)}{(1-\gamma)^2}$.

(ii) *For all* $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|g(\tau,\theta_1) - g(\tau,\theta_2)\| \leq L_g\|\theta_1 - \theta_2\|$ *where* $L_g \overset{def}{=} \frac{2(l_\psi^2 + L_\psi)\|r\|_\infty}{(1-\gamma)^2}$,

(iii) *For all* $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g(\tau,\theta) - \nabla_\theta J_H(\theta)\|^2\right] \leq V^2$ *where* $V \overset{def}{=} \frac{2l_\psi\|r\|_\infty}{(1-\gamma)^{3/2}}$.

(iv) *For all* $\theta \in \mathbb{R}^d$, $\|g(\tau,\theta)\| \leq G$ *where* $G \overset{def}{=} \frac{2l_\lambda l_\psi}{(1-\gamma)^2}$.

We start with the next lemma which corresponds exactly to Lemma E.1.

Lemma E.14. *Let Assumption 4.1 hold true. Then, the sequence* (θ_t) *generated by Algorithm 4 and the sequence* (e_t) *defined in* (84) *satisfy for every integer* $t \geq 0$,

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha_t}{3}\|\nabla J(\theta_t)\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_g\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2\,. \tag{87}$$

Proof. The proof is identical to the proof of Lemma E.1 upon noticing that Assumption 4.2 is not needed here since smoothness of the objective function J is a standard result in the RL literature following from Assumption 4.1 (see for e.g., Lemma 4.4 in Yuan et al. (2022)). □

Then next lemma is similar to Lemma E.8 and controls the error e_t in Lemma E.14. We provide here a complete statement and proof of this result for clarity and completeness since the corresponding lemma in the more general case is more involved. Indeed, the latter result involves an additional error due to the occupancy measure estimation which is not required in our present setting.

Lemma E.15. *Under Assumption 4.1, we have for every integer* $t \geq 1$

$$\mathbb{E}[\|e_t\|^2] \leq (1-\eta_t)^2\mathbb{E}[\|e_{t-1}\|^2] + 2V^2\eta_t^2 + 4(L_g^2 + G^2C_w)(1-\eta_t)^2\alpha_{t-1}^2\,. \tag{88}$$

where V, G, L_g *are constants defined in Lemma H.3 and E.13. If in addition*

(i) $\eta_t = \frac{2}{t+1}$, *then for all* $t \geq 1$, *we have*

$$\mathbb{E}[\|e_t\|] \leq 4V\eta_t \cdot t^{1/2} + 2(L_g + GC_w^{1/2})\alpha_{t-1} \cdot t^{1/2}\,. \tag{89}$$

(ii) $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$, *then for all integers* $T \geq 1$, *if* $\alpha_t = \frac{\alpha_0}{T^{2/3}}$ *for some* $\alpha_0 > 0$, *we have*

$$\sum_{t=1}^{T}\mathbb{E}[\|e_t\|] \leq \frac{C\left(V + \left(L_g + GC_w^{1/2}\right)\alpha_0\right)}{T^{1/3}}\,, \tag{90}$$

where $C > 0$ *is an absolute numerical constant.*

Proof. Using the update rule of the sequence (d_t) and recalling the definition of the error $e_t = d_t - \nabla J_H(\theta_t)$, we have

$$\begin{aligned}
e_t &= d_t - \nabla J_H(\theta_t)\\
&= (1-\eta_t)(d_{t-1} + v_t) + \eta_t g(\tau_t,\theta_t) - \nabla J_H(\theta_t)\\
&= (1-\eta_t)(e_{t-1} + \nabla J_H(\theta_{t-1}) + v_t) + \eta_t g(\tau_t,\theta_t) - \nabla J_H(\theta_t)\\
&= (1-\eta_t)e_{t-1} + \eta_t(g(\tau_t,\theta_t) - \nabla J_H(\theta_t)) + (1-\eta_t)(v_t - (\nabla J_H(\theta_t) - \nabla J_H(\theta_{t-1})))\,.
\end{aligned}$$

Introducing additional notation for convenience:

$$y_t \overset{\text{def}}{=} g(\tau_t,\theta_t) - \nabla J_H(\theta_t)\,, \tag{91}$$

$$z_t \overset{\text{def}}{=} v_t - (\nabla J_H(\theta_t) - \nabla J_H(\theta_{t-1}))\,, \tag{92}$$

we obtain the following useful decomposition:

$$e_t = (1-\eta_t)e_{t-1} + \eta_t y_t + (1-\eta_t)z_t\,. \tag{93}$$

Defining $\mathcal{F}_t$ as the σ-algebra generated by all the random variables until time t, we observe that $\mathbb{E}[y_t|\mathcal{F}_{t-1}] = \mathbb{E}[z_t|\mathcal{F}_{t-1}] = 0$. As a consequence, we have

$$\begin{aligned}\mathbb{E}[\|e_t\|^2] &= (1-\eta_t)^2\mathbb{E}[\|e_{t-1}\|^2] + \mathbb{E}[\|\eta_t y_t + (1-\eta_t)z_t\|^2] \\ &\leq (1-\eta_t)^2\mathbb{E}[\|e_{t-1}\|^2] + 2\eta_t^2\mathbb{E}[\|y_t\|^2] + 2(1-\eta_t)^2\mathbb{E}[\|z_t\|^2]\,.\end{aligned} \tag{94}$$

We now control each one of the last two terms in the previous inequality. For the first term, we have by Lemma E.13-(iii)

$$\mathbb{E}[\|y_t\|^2] \leq V^2\,. \tag{95}$$

Concerning the second remaining term, using Lemma H.3-(ii) and Lemma E.13-(iv), we write

$$\begin{aligned}\mathbb{E}[\|z_t\|^2] &\leq \mathbb{E}[\|v_t\|^2] \\ &= \mathbb{E}[\|g(\tau_t,\theta_t) - w(\tau_t|\theta_{t-1},\theta_t)g(\tau_t,\theta_{t-1})\|^2] \\ &= \mathbb{E}[\|g(\tau_t,\theta_t) - g(\tau_t,\theta_{t-1}) + g(\tau_t,\theta_{t-1})(1-w(\tau_t|\theta_{t-1},\theta_t))\|^2] \\ &\leq 2L_g^2\mathbb{E}[\|\theta_t-\theta_{t-1}\|^2] + 2G^2\mathbb{E}[(1-w(\tau_t|\theta_{t-1},\theta_t))^2] \\ &= 2L_g^2\mathbb{E}[\|\theta_t-\theta_{t-1}\|^2] + 2G^2\,\mathrm{Var}(w(\tau_t|\theta_{t-1},\theta_t)) \\ &\leq 2(L_g^2 + G^2C_w)\mathbb{E}[\|\theta_t-\theta_{t-1}\|^2] \\ &= 2(L_g^2 + G^2C_w)\alpha_{t-1}^2\,.\end{aligned} \tag{96}$$

Combining (94) with (95) and (96) concludes the first part of the proof.

Applying Lemma H.4 with $\eta_t = \frac{2}{t+1}$, $\beta_t = 2V^2\eta_t^2 + 4(L_g^2 + G^2C_w)(1-\eta_t)^2\alpha_{t-1}^2$ and using $\mathbb{E}[\|e_0\|^2] \leq V^2$, we get

$$\begin{aligned}\mathbb{E}[\|e_t\|] &\leq \left(\mathbb{E}[\|e_t\|^2]\right)^{\frac{1}{2}} \leq \left(\frac{4V^2}{(t+1)^2} + 2V^2\eta_t^2\cdot t + 4(L_g^2+G^2C_w)\alpha_{t-1}^2\cdot t\right)^{1/2} \\ &\leq \frac{2V}{(t+1)} + 2V\eta_t\cdot t^{1/2} + 2(L_g + GC_w^{1/2})\alpha_{t-1}\cdot t^{1/2} \\ &\leq 4V\eta_t\cdot t^{1/2} + 2(L_g + GC_w^{1/2})\alpha_{t-1}\cdot t^{1/2}\,.\end{aligned} \tag{97}$$

In order to derive (90), we unroll the recursion (88) from $t=1$ to $t=t'$, where $t' \leq T$. Denoting $\beta_t = 2V^2\eta_t^2 + 4(L_g^2 + G^2C_w)(1-\eta_t)^2\alpha_{t-1}^2$, we have

$$\begin{aligned}\mathbb{E}\left[\|e_{t'}\|^2\right] &\leq \prod_{\tau=1}^{t'}(1-\eta_\tau)\mathbb{E}\left[\|e_0\|^2\right] + \sum_{t=0}^{t'}\beta_t\prod_{\tau=t+1}^{t'}(1-\eta_\tau) \\ &\leq V^2\eta_{t'+1} + C\beta_{t+1}\eta_{t'+1}^{-1}\,,\end{aligned} \tag{98}$$

where we used $\mathbb{E}\left[\|e_0\|^2\right] \leq V^2$ and the result of Lemma H.6 with $C>0$ being a numerical constant. Finally, summing up

the above inequality from $t'=1$ to $t'=T$ and choosing $\alpha_t=\alpha=\frac{\alpha_0}{T^{2/3}}$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|\right] &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|e_t\|^2\right]\right)^{1/2}\\
&\leq \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|^2\right]\right)^{1/2}\\
&\leq \left(\frac{1}{T}\sum_{t=1}^{T}V^2\eta_{T+1}+2CV^2\eta_{t+1}+4C(L_g^2+G^2C_w)\alpha^2\eta_{t+1}^{-1}\right)^{1/2}\\
&\overset{(i)}{\leq} \left(3V^2\eta_{T-1}+6CV^2\eta_{T-1}+4C(L_g^2+G^2C_w)\frac{\alpha^2}{\eta_{T+1}}\right)^{1/2}\\
&\leq \left(9CV^2\eta_T+6C(L_g^2+G^2C_w)\frac{\alpha_0^2}{\eta_T}\frac{1}{T^{4/3}}\right)^{1/2}\\
&\leq \frac{4C^{1/2}\left(V+\left(L_g+GC_w^{1/2}\right)\alpha_0\right)}{T^{1/3}}.
\end{aligned}$$

where in (i) we used (51). □

End of Proof of Corollary 4.7. The last steps of the proof are standard. Taking expectation on both sides of the result of Lemma E.14, telescoping and rearranging, we have for every integer $T\geq 1$ and constant step-size $\alpha_t=\frac{\alpha_0}{T^{2/3}}$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|\right] &\leq \frac{3(J^*-J(\theta_1))}{\alpha_0 T}T^{2/3}+\frac{6}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|\right]+\frac{3L_\theta\alpha_0}{T^{2/3}}+4D_g\gamma^H\\
&\leq \frac{3(J^*-J(\theta_1))}{\alpha_0 T^{1/3}}+\frac{6C\left(V+\left(L_g+GC_w^{1/2}\right)\alpha_0\right)}{T^{1/3}}+\frac{3L_\theta\alpha_0}{T^{2/3}}+4D_g\gamma^H \qquad (99)
\end{aligned}$$

with $C>0$ being a numerical constant, where we applied Lemma E.15 to bound $\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|\right]$. Then choosing H large enough, we have after T iterations

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|\right] \leq \mathcal{O}\left(\frac{J^*-J(\theta_1)}{\alpha_0 T^{1/3}}+\frac{V+(L_g+GC_w^{1/2})\alpha_0}{T^{1/3}}\right),$$

which concludes the first part of the corollary.

As for the second part of the statement, we know from Lemma H.3 that

$$J^*-J(\theta_1)=\mathcal{O}\left(\frac{1}{1-\gamma}\right),\quad V=\mathcal{O}\left(\frac{1}{(1-\gamma)^{3/2}}\right),\quad G=\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right),$$
$$L_g=\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right),\quad C_w^{1/2}=\mathcal{O}\left(\frac{1}{1-\gamma}\right).$$

If we set $\alpha_0=1-\gamma$, we derive the desired bound.

E.4. Proof of Theorem 4.8 (Cumulative reward setting for continuous state-action space and Gaussian policy)

In this section, we consider continuous state and action spaces where $\mathcal{S}=\mathbb{R}^p$ and $\mathcal{A}=\mathbb{R}^q$ for two positive integers $p,q\geq 1$. Our focus is on the popular class of Gaussian policies which are common to handle the case of continuous state action spaces in practice.

Let $\sigma > 0$. Define for every $\theta \in \mathbb{R}^d$ a map $\mu_\theta : \mathcal{S} \to \mathbb{R}^q$. Then, we define the Gaussian policy π_θ for each parameter $\theta \in \mathbb{R}^d$ and each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ as follows:

$$\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|\mu_\theta(s) - a\|^2}{2\sigma^2}\right). \tag{100}$$

Let us mention that μ_θ is a parametrization of the Gaussian mean which can be a neural network in practice. The standard deviation σ can be fixed or parametrized as well in practice. We consider a fixed standard deviation for the purpose of our discussion.

Remark E.16. Note that one can consider even more general parametrizations such as the exponential family or symmetric α-stable policies which include the Gaussian policy as a particular case. We refer the interested reader to the nice exposition in Bedi et al. (2021) for a discussion around such heavy-tailed policy parametrizations (see also Bedi et al. (2022)).

We make the following standard smoothness assumption on our Gaussian policy parametrization.

Assumption E.17. In the Gaussian parametrization (100), the map $\theta \mapsto \mu_\theta(s)$ is continuously differentiable for every $s \in \mathcal{S}$, l_μ-Lipschitz continuous (uniformly in $s \in \mathcal{S}$) and there exist $M_g > 0, M_h > 0$ s.t. for every $\theta \in \mathbb{R}^d, (s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\nabla \log \pi_\theta(a|s)\| \leq M_g$, $\|\nabla^2_\theta \log \pi_\theta(a|s)\| \leq M_h$.

Notice that conditions on the map $\theta \mapsto \mu_\theta(s)$ and its higher-order derivatives can be enforced for every $s \in \mathcal{S}$ so that the desired regularity conditions on the policy parametrization in Assumption E.17 are satisfied upon considering a set of actions lying in a compact set. Consider for instance the simpler case where $q = 1$ and the mean of the policy is parametrized with a linear function, i.e., $\mu_\theta(s) = \phi(s)^T\theta$ for some feature map $\phi : \mathcal{S} \to \mathbb{R}^d$. Then, the boundedness of $\|\nabla^2_\theta \log \pi_\theta(a|s)\|$ is automatically satisfied since $\nabla^2_\theta \log \pi_\theta(a|s)$ is the matrix $-\frac{1}{\sigma^2}\phi(s)\phi(s)^T$ which is independent from the parameter θ . As for the first condition, it is satisfied if the feature map ϕ as well as $\phi_\theta(s)$ are bounded over the state space and the policy parameter space while the action set is also bounded. Notice though that Assumption E.17 can be relaxed to hold in expectation (over state-action pairs) in order to include an even larger class of policies (Yuan et al., 2022). In this work, we do not pursue such relaxations and assume the standard bound for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ for simplicity (Xu et al., 2020a; Liu et al., 2020).

Similarly to the softmax parametrization setting with Lemma E.13, under Assumption E.17, one can show smoothness of the expected return function J and derive useful bounds for the norm and the variance of stochastic gradients, see (Yuan et al., 2022, Lemma 4.2, (68) and Lemma 4.4, (19)), and (Xu et al., 2020a, Proposition 4.2 (1) and (3)).

Lemma E.18. *Let Assumption E.17 hold true and let $\tau = \{s_0, a_0, \cdots, s_{H-1}, a_{H-1}\}$ be an arbitrary trajectory of length H. Then the following statements hold:*

(i) The objective function $\theta \mapsto J(\theta)$ is L_θ-smooth with $L_\theta \overset{def}{=} \frac{\|r\|_\infty(M_g^2+M_h)}{(1-\gamma)^2}$.

(ii) For all $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|g(\tau, \theta_1) - g(\tau, \theta_2)\| \leq L_g\|\theta_1 - \theta_2\|$ with $L_g \overset{def}{=} \frac{2M_g^2\|r\|_\infty}{(1-\gamma)^3} + \frac{M_h\|r\|_\infty}{(1-\gamma)^2}$,

(iii) For all $\theta \in \mathbb{R}^d$, $\mathbb{E}\left[\|g(\tau, \theta) - \nabla_\theta J_H(\theta)\|^2\right] \leq V^2$ with $V \overset{def}{=} \frac{M_g\|r\|_\infty}{(1-\gamma)^{3/2}}$.

(iv) For all $\theta \in \mathbb{R}^d$, $\|g(\tau, \theta)\| \leq G$ with $G \overset{def}{=} \frac{M_g\|r\|_\infty}{(1-\gamma)^2}$.

Given a trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots, s_{H-1}, a_{H-1})$ of length H generated under the initial distribution ρ and the Gaussian policy π_θ as defined in (100) for some $\theta \in \mathbb{R}^d$, recall the definition of the IS weight for every $\theta' \in \mathbb{R}^d$:

$$w(\tau|\theta', \theta) \overset{\text{def}}{=} \prod_{h=0}^{H-1} \frac{\pi_{\theta'}(a_h|s_h)}{\pi_\theta(a_h|s_h)}. \tag{101}$$

Lemma E.19. *Let $H \geq 1$ be an integer and let Assumption E.17 be satisfied. Suppose that the sequence (θ_t) is updated via $\theta_{t+1} = \theta_t + \alpha_t\frac{d_t}{\|d_t\|}$ where $d_t \in \mathbb{R}^d$ is any nonzero update direction and α_t is a positive stepsize. If τ_{t+1} is a (random) trajectory of length H generated following the initial distribution ρ and the Gaussian policy $\pi_{\theta_{t+1}}$ as defined in (100), then*

$$\mathbb{E}[w(\tau_{t+1}|\theta_t, \theta_{t+1})] = 1\,, \tag{102}$$

$$\text{Var}\left[w(\tau_{t+1}|\theta_t, \theta_{t+1})\right] \leq C_w\alpha_t^2\,, \tag{103}$$

where the IS weight $w(\tau_{t+1}|\theta_t,\theta_{t+1})$ is as defined in (101) and $C_w \overset{def}{=} (2H^2M_g + HM_h)(W+1)$.

Proof. The first identity follows from the definitions of the expectation and the IS weight. We now prove the second identity. For any $\theta \in \mathbb{R}^d$, let $p(\cdot|\pi_\theta)$ denote the probability distribution induced by the policy π_θ over the space of random trajectories of length H initialized with the state distribution ρ. The probability density is then given by

$$p(\tau|\pi_\theta) = \rho(s_0)\,\pi_\theta(a_0|s_0)\prod_{t=1}^{H-1}\mathcal{P}(s_t|s_{t-1},a_{t-1})\,\pi_\theta(a_t|s_t)\,, \tag{104}$$

where $\tau = (s_0, a_0, \cdots, s_{H-1}, a_{H-1})$.

We use the shorthand notations $\theta_1 = \theta_t, \theta_2 = \theta_{t+1}$ and $\tau = \tau_{t+1}$ for the rest of this proof. Then, we have

$$\begin{aligned}\mathbb{E}[w(\tau|\theta_1,\theta_2)^2] &= \int \frac{p(\tau|\pi_{\theta_1})^2}{p(\tau|\pi_{\theta_2})}d\tau \\ &= \int \rho(s_0)\,\pi_\theta(a_0|s_0)\prod_{t=1}^{H-1}\mathcal{P}(s_t|s_{t-1},a_{t-1})\,\frac{\pi_{\theta_1}(a_t|s_t)^2}{\pi_{\theta_2}(a_t|s_t)}d\tau\,.\end{aligned} \tag{105}$$

We bound the above integral starting from the integral of the last term of the product[7] which writes as follows:

$$\int \mathcal{P}(s_{H-1}|s_{H-2},a_{H-2})\int \frac{\pi_{\theta_1}(a_{H-1}|s_{H-1})^2}{\pi_{\theta_2}(a_{H-1}|s_{H-1})}da_{H-1}ds_{H-1}\,. \tag{106}$$

We shall now compute the integral w.r.t. a_{H-1}. In dimension 1 for the action variable a_{H-1} (similar derivations hold for higher dimensions), we have for every s,

$$\begin{aligned}&\int_{-\infty}^{+\infty}\frac{\pi_{\theta_1}(x\mid s)^2}{\pi_{\theta_2}(x\mid s)}dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left[-\frac{2\left(x-\mu_{\theta_1}(s)\right)^2-\left(x-\mu_{\theta_2}(s)\right)^2}{2\sigma^2}\right]dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{2\mu_{\theta_1}(s)^2-\mu_{\theta_2}(s)^2}{2\sigma^2}\right]\int_{-\infty}^{+\infty}\exp\left[-\frac{x^2-2\left(2\mu_{\theta_1}(s)-\mu_{\theta_2}(s)\right)x}{2\sigma^2}\right]dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{2\mu_{\theta_1}(s)^2-\mu_{\theta_2}(s)^2-\left(2\mu_{\theta_1}(s)-\mu_{\theta_2}(s)\right)^2}{2\sigma^2}\right]\int_{-\infty}^{+\infty}\exp\left[-\frac{\left(x-\left(2\mu_{\theta_1}(s)-\mu_{\theta_2}(s)\right)\right)^2}{2\sigma^2}\right]dx \\ &= \exp\left[-\frac{2\mu_{\theta_1}(s)^2-\mu_{\theta_2}(s)^2-\left(2\mu_{\theta_1}(s)-\mu_{\theta_2}(s)\right)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{\left(\mu_{\theta_2}(s)-\mu_{\theta_1}(s)\right)^2\Big]}{\sigma^2}\right]\end{aligned} \tag{107}$$

As a consequence, we obtain

$$\begin{aligned}&\int \mathcal{P}(s_{H-1}|s_{H-2},a_{H-2})\int \frac{\pi_{\theta_1}(a_{H-1}|s_{H-1})^2}{\pi_{\theta_2}(a_{H-1}|s_{H-1})}da_{H-1}ds_{H-1} \\ &\le \int \mathcal{P}(s_{H-1}|s_{H-2},a_{H-2})\exp\left(\frac{\left\|\mu_{\theta_1}(s_{H-1})-\mu_{\theta_2}(s_{H-1})\right\|^2}{\sigma^2}\right)ds_{H-1} \\ &\overset{(i)}{\le} \int \mathcal{P}(s_{H-1}|s_{H-2},a_{H-2})\exp\left(\frac{l_\mu^2\left\|\theta_1-\theta_2\right\|^2}{\sigma^2}\right)ds_{H-1} \\ &\overset{(ii)}{=} \exp\left(\frac{l_\mu^2\alpha_t^2}{\sigma^2}\right)\,,\end{aligned} \tag{108}$$

[7] Notice that the integrand is nonnegative and we can integrate in any order by Tonelli's theorem.

where (i) follows from the l_μ-Lipschitzness of the parametrized mean in Assumption E.17 and (ii) utilizes the normalized update rule as well as the fact that $\mathcal{P}(\cdot|s_{H-2}, a_{H-2})$ is a transition probability kernel. Using a similar reasoning to bound the different integrals like in (106) backward from $H-1$ to 0 successively, we obtain the following bound on the second moment of IS weights in (105):

$$\mathbb{E}[w(\tau|\theta_1, \theta_2)^2] \leq \exp\left(\frac{H l_\mu^2 \alpha_t^2}{\sigma^2}\right). \tag{109}$$

Therefore, similarly to the argument in Lemma 4.3, we obtain:

$$\mathrm{Var}(w(\tau|\theta_1, \theta_2)) \leq W, \tag{110}$$

where $W = \mathcal{O}(1)$ is a numerical constant, which can be ensured, for example, by setting the step-sizes as $\alpha_t = \alpha = T^{-2/3}$. As a consequence, we can apply Lemma B.1 in (Xu et al., 2020a) and derive the bound on the variance of IS weights:

$$\mathrm{Var}\left[w(\tau_{t+1}|\theta_t, \theta_{t+1})\right] \leq C_w \|\theta_{t+1} - \theta_t\|^2 = C_w \alpha_t^2,$$

where the last step follows by the update rule $\theta_{t+1} = \theta_t + \alpha \frac{d_t}{\|d_t\|}$. This concludes the proof. □

Given the above results, we immediately obtain convergence of Algorithm 4 for Gaussian policy parametrization.

Corollary E.20 (Stationary convergence of N-VR-PG). *Let Assumption E.17 hold. Let $\alpha_0 > 0$ and let T be an integer larger than 1. Set $\alpha_t = \frac{\alpha_0}{T^{2/3}}$, $\eta_t = \left(\frac{2}{t+1}\right)^{2/3}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Let $\bar{\theta}_T$ be sampled from the iterates $\{\theta_1, \cdots, \theta_T\}$ of N-VR-PG (Algorithm 4) uniformly at random. Then we have*

$$\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] \leq \mathcal{O}\left(\frac{J^* - J(\theta_1)}{\alpha_0 T^{1/3}} + \frac{V + (L_g + G C_w^{1/2})\alpha_0}{T^{1/3}}\right), \tag{111}$$

where V, L_g, G, and C_w are defined in Lemma E.18 and E.19. Moreover, if we set $\alpha_0 = 1-\gamma$, then

$$\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2 T^{1/3}}\right).$$

Proof. Given the results of Lemma E.18 and E.19, the proof of this statement follows immediately from the result of Corollary E.20. We notice that in order to specify the dependence on $1-\gamma$, we invoke Lemma E.18, which is analogous to the corresponding Lemma E.13 for softmax policy parameterization. The only difference in terms of the dependence on $1-\gamma$ is in the bound for L_g. However, this fact does not affect the final dependence on $1-\gamma$ since it is dominated by other terms in (111). □

F. Proofs for Section 4.3: Global optimality convergence

F.1. Proof of Theorem 4.12 (General utilities setting)

In this section, to prove our global convergence result under an additional concave reparametrization assumption, we refine the result of Lemma E.1. The proof is similar to the proof of Lemma 5.12 in Zhang et al. (2021b). Nevertheless, we would like to mention that it deviates from the latter in that our algorithm is significantly different and its normalized nature requires a significantly different treatment. In particular, the reader can appreciate from the statement of the result that the error term $\|e_t\|$ to the gradient estimation is not squared unlike in (Zhang et al., 2021b) and controlling its magnitude required different proof techniques given the different recursive loopless variance reduction mechanism that we consider.

Lemma F.1. *Let Assumptions 4.1, 4.2 and Assumption 4.9 hold. Additionally, let Assumption 4.10 be satisfied with some positive $\bar{\epsilon}$. Then, the sequence (θ_t) generated by Algorithm 1 and the sequence (e_t) satisfy for every positive real $\epsilon \leq \min\left\{\bar{\epsilon}, \frac{\alpha_t(1-\gamma)}{2\ell_\theta}\right\}$ and every integer t,*

$$F(\lambda(\theta^*)) - F(\lambda(\theta_{t+1})) \leq (1-\epsilon)(F(\lambda(\theta^*)) - F(\lambda(\theta_t))) + 2\alpha_t\|e_t\| + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 + \frac{4}{3}\alpha_t D_\lambda \gamma^H + \frac{L_\theta}{2}\alpha_t^2. \tag{112}$$

Proof. Lemma E.1 provides the following inequality:

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_t)) + \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2\,. \tag{113}$$

Now, for any $\epsilon < \bar{\epsilon}$, the concavity reparametrization assumption implies that $(1-\epsilon)\lambda(\theta_t) + \epsilon\lambda(\theta^*) \in \mathcal{V}_{\lambda(\theta_t)}$ and therefore we have

$$\theta_\epsilon \overset{\text{def}}{=} (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}((1-\epsilon)\lambda(\theta_t) + \epsilon\lambda(\theta^*)) \in \mathcal{U}_{\theta_t}\,. \tag{114}$$

It also follows from the smoothness of the objective function $\theta \mapsto F(\lambda(\theta))$ that

$$F(\lambda(\theta_t)) \geq F(\lambda(\theta_\epsilon)) - \langle\nabla_\theta F(\lambda(\theta_t)), \theta_\epsilon - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_\epsilon - \theta_t\|^2\,. \tag{115}$$

Combining (113) and (115) yields

$$\begin{aligned} F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_\epsilon)) - \langle\nabla_\theta F(\lambda(\theta_t)), \theta_\epsilon - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_\epsilon - \theta_t\|^2 \\ + \frac{\alpha_t}{3}\|\nabla_\theta F(\lambda(\theta_t))\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2\,. \end{aligned} \tag{116}$$

Then, we notice that:

(i) By assumption, the mapping $\lambda \circ (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}$ coincides with the identity mapping on the set $\mathcal{U}_{\theta_t}$. Hence, given the definition of θ_ϵ in (114), we have

$$\begin{aligned} F(\lambda(\theta_\epsilon)) &= F((1-\epsilon)\lambda(\theta_t) + \epsilon\lambda(\theta^*)) \\ &\geq (1-\epsilon)F(\lambda(\theta_t)) + \epsilon F(\lambda(\theta^*))\,, \end{aligned} \tag{117}$$

where the last step follows from the concavity of the function F.

(ii) Again since the mapping $\lambda \circ (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}$ coincides with the identity mapping on the set $\mathcal{U}_{\theta_t}$ and using the (uniform) lipschitzness of the inverse mapping $(\lambda|_{\mathcal{U}_{\theta_t}})^{-1}$, we have

$$\begin{aligned} \|\theta_\epsilon - \theta_t\| &= \|(\lambda|_{\mathcal{U}_{\theta_t}})^{-1}((1-\epsilon)\lambda(\theta_t) + \epsilon\lambda(\theta^*)) - (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}(\lambda(\theta_t))\| \\ &\leq l_\theta\epsilon\|\lambda(\theta_t) - \lambda(\theta^*)\| \\ &\leq \frac{2l_\theta\epsilon}{(1-\gamma)}\,. \end{aligned} \tag{118}$$

(iii) Using the Cauchy-Schwarz inequality together with the inequality established in the previous item gives

$$\begin{aligned} |\langle\nabla_\theta F(\lambda(\theta_t)), \theta_\epsilon - \theta_t\rangle| &\leq \|\nabla_\theta F(\lambda(\theta_t))\| \cdot \|\theta_\epsilon - \theta_t\| \\ &\leq \frac{2l_\theta\epsilon}{1-\gamma}\|\nabla_\theta F(\lambda(\theta_t))\|\,. \end{aligned} \tag{119}$$

Substituting the inequalities (117), (118) and (119) into (116) leads to

$$\begin{aligned} F(\lambda(\theta_{t+1})) \quad &\geq \quad (1-\epsilon)F(\lambda(\theta_t)) + \epsilon F(\lambda(\theta^*)) + \left(\frac{\alpha_t}{3} - \frac{2l_\theta\epsilon}{1-\gamma}\right)\|\nabla_\theta F(\lambda(\theta_t))\| - \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 - 2\alpha_t\|e_t\| \\ &\qquad -\frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2 \\ &\geq \quad (1-\epsilon)F(\lambda(\theta_t)) + \epsilon F(\lambda(\theta^*)) - \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 - 2\alpha_t\|e_t\| - \frac{4}{3}D_\lambda\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2\,, \end{aligned} \tag{120}$$

where the last step follows from the condition $\epsilon \leq \frac{\alpha_t(1-\gamma)}{2\ell_\theta}$.

Finally, substracting $F(\lambda(\theta^*))$ from both sides and rearranging the terms gives the desired result:

$$F(\lambda(\theta^*)) - F(\lambda(\theta_{t+1})) \leq (1-\epsilon)(F(\lambda(\theta^*)) - F(\lambda(\theta_t))) + 2\alpha_t\|e_t\| + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 + \frac{4}{3}\alpha_t D_\lambda \gamma^H + \frac{L_\theta}{2}\alpha_t^2 . \quad (121)$$

□

Theorem F.2 (Global convergence of N-VR-PG for general utilities). *Let Assumptions 4.1 and 4.9 hold. Additionally, let Assumption 4.10 be satisfied with $\bar{\epsilon} \geq \frac{\alpha_0(1-\gamma)}{2\ell_\theta(T+1)^a}$ for some integer $T \geq 1$ and reals $\alpha_0 > 0$, $a \in (0,1)$. Set $\alpha_t = \frac{\alpha_0}{(T+1)^a}$, $\eta_t = \frac{2}{t+1}$ for every integer t and $H = (1-\gamma)^{-1}\log(T+1)$. Then the output θ_T of N-VR-PG (see Algorithm 1) satisfies*

$$F(\lambda(\theta^*)) - \mathbb{E}\left[F(\lambda(\theta_T))\right] \leq \mathcal{O}\left(\frac{\alpha_0^2}{(1-\gamma)^3(T+1)^{2a-\frac{3}{2}}}\right),$$

where $F(\lambda(\theta^*))$ is the optimal utility value. Therefore, the sample complexity to achieve $F(\lambda(\theta^*)) - \mathbb{E}\left[F(\lambda(\theta_T))\right] \leq \varepsilon$ is $\mathcal{O}\left(\varepsilon^{\frac{-2}{4a-3}}\right)$.

Proof. Define $\delta_t \overset{\text{def}}{=} \mathbb{E}\left[F(\lambda(\theta^*)) - F(\lambda(\theta_t))\right]$. Applying expectation to the result of Lemma F.1, we have for $\epsilon \leq \min\left\{\bar{\epsilon}, \frac{\alpha_t(1-\gamma)}{2\ell_\theta}\right\}$,

$$\begin{aligned}
\delta_{t+1} &\leq (1-\epsilon)\delta_t + 2\alpha_t\mathbb{E}\left[\|e_t\|\right] + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 + \frac{4}{3}\alpha_t D_\lambda\gamma^H + \frac{L_\theta}{2}\alpha_t^2 \\
&\leq (1-\epsilon)\delta_t + 2\alpha_t\mathbb{E}\left[\|\hat{e}_t\|\right] + 2C_1\alpha_t\mathbb{E}\left[\|\tilde{e}_t\|\right] + 2C_2\alpha_t\alpha_{t-1} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 + \frac{4}{3}\alpha_t D_\lambda\gamma^H + \frac{L_\theta}{2}\alpha_t^2 \quad (122)
\end{aligned}$$

where in the last step we apply Lemma E.4 with $C_1 \overset{\text{def}}{=} \frac{2L_\lambda^2 l_\psi}{(1-\gamma)^2}$ and $C_2 \overset{\text{def}}{=} \frac{2L_\lambda L_{\lambda,\infty} l_\psi}{(1-\gamma)^2}$.

By Lemma E.5 (Equation (34)), for $\eta_t = \frac{2}{t+1}$, we have

$$\mathbb{E}[\|\tilde{e}_t\|] \leq \frac{4}{(1-\gamma)}\eta_t \cdot t^{1/2} + \frac{2C_w^{1/2}}{(1-\gamma)}\alpha_{t-1}\cdot t^{1/2}. \quad (123)$$

With the same η_t as above, by Lemma E.8 (Equation (58)), we have

$$\mathbb{E}[\|\hat{e}_t\|] \leq \frac{2\hat{E}}{t+1} + 2C_3^{1/2}\eta_t\cdot t^{1/2} + C_4^{1/2}\alpha_{t-2}\cdot t^{1/2}, \quad (124)$$

where $C_3 \overset{\text{def}}{=} \frac{288Cl_\psi^2L_\lambda^2}{(1-\gamma)^6} + \frac{32l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$, $C_4 \overset{\text{def}}{=} \frac{12l_\lambda^2[(l_\psi^2+L_\psi)^2+C_w l_\psi^2]}{(1-\gamma)^4} + \frac{144C_w l_\psi^2 L_\lambda^2}{(1-\gamma)^6}$, and $C_w = H((8H+2)l_\psi^2 + 2L_\psi)(W+1)$.

Unrolling (122) from $t = T-1$ to $t = 0$, using (123) and (124) and setting $\alpha_t = \alpha$, we have

$$\begin{aligned}
\delta_T &\leq (1-\epsilon)^T\delta_0 + 2\alpha\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|\hat{e}_t\|\right] + C_1\mathbb{E}\left[\|\tilde{e}_t\|\right]\right) + 2C_2\alpha^2 T + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon}D_\lambda\gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq (1-\epsilon)^T\delta_0 + 2\alpha\sum_{t=0}^{T-1}\left(\frac{2\hat{E}}{t+1} + 2C_3^{1/2}\eta_t\cdot t^{1/2} + C_4^{1/2}\alpha\cdot t^{1/2}\right) + 2C_2\alpha^2 T \\
&\quad + 2C_2\alpha\sum_{t=0}^{T-1}\left(\frac{4}{(1-\gamma)}\eta_t\cdot t^{1/2} + \frac{2C_w^{1/2}}{(1-\gamma)}\alpha\cdot t^{1/2}\right) + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon}D_\lambda\gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq (1-\epsilon)^T\delta_0 + 4\alpha\hat{E}\log(T) + 8\alpha C_3^{1/2}(T+1)^{1/2} + 2C_4^{1/2}\alpha^2\cdot(T+1)^{3/2} + 2C_2\alpha^2 T \\
&\quad + \frac{16C_2\alpha}{(1-\gamma)}(T+1)^{1/2} + \frac{4C_2C_w^{1/2}}{(1-\gamma)}\alpha^2(T+1)^{3/2} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon}D_\lambda\gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon}.
\end{aligned}$$

Notice that $(1-\epsilon)^T \leq \exp\left(T\log(1-\epsilon)\right) \leq \exp\left(-\epsilon T\right)$. Finally setting $\alpha = \frac{\alpha_0}{(T+1)^a}$, for $0 < a < 1$ and $\epsilon = \min\left\{\bar{\epsilon}, \frac{\alpha(1-\gamma)}{2\ell_\theta}\right\} = \frac{\alpha(1-\gamma)}{2\ell_\theta}$, we obtain

$$\begin{aligned}
\delta_T &\leq \exp\left(-\frac{\alpha_0(1-\gamma)}{2\ell_\theta}T^{1-a}\right) + \frac{4\hat{E}\log(T)\alpha_0}{(T+1)^a} + \left(8C_3^{1/2} + \frac{16C_2}{(1-\gamma)}\right)\frac{\alpha_0}{(T+1)^{a-1/2}} + \frac{2C_2\alpha_0^2}{(T+1)^{2a-1}} \\
&\quad + \left(2C_4^{\frac{1}{2}} + \frac{4C_2C_w^{1/2}}{(1-\gamma)}\right)\frac{\alpha_0^2}{(T+1)^{2a-\frac{3}{2}}} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon}D_\lambda\gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq \exp\left(-\frac{\alpha_0(1-\gamma)}{2\ell_\theta}T^{1-a}\right) + \frac{4\hat{E}\log(T)\alpha_0}{(T+1)^a} + \left(8C_3^{1/2} + \frac{16C_2}{(1-\gamma)}\right)\frac{\alpha_0}{(T+1)^{a-1/2}} + \frac{2C_2\alpha_0^2}{(T+1)^{2a-1}} \\
&\quad + \left(2C_4^{\frac{1}{2}} + \frac{4C_2C_w^{1/2}}{(1-\gamma)}\right)\frac{\alpha_0^2}{(T+1)^{2a-\frac{3}{2}}} + \frac{3L_\theta l_\theta\alpha_0}{(1-\gamma)(T+1)^a} + \frac{8\ell_\theta D_\lambda}{3(1-\gamma)}\gamma^H \\
&\leq \mathcal{O}\left(\frac{1}{(1-\gamma)^3(T+1)^{2a-\frac{3}{2}}}\right),
\end{aligned}$$

where the last step follows by setting $H = (1-\gamma)^{-1}\log(T)$ and noticing that $2a - \frac{3}{2} < a - \frac{1}{2}$ for $a \in (0,1)$, $C_4 = \mathcal{O}\left((1-\gamma)^{-6}\right)$, $C_w = \mathcal{O}\left((1-\gamma)^{-2}\right)$, $C_2 = \mathcal{O}\left((1-\gamma)^{-2}\right)$. □

F.2. Proof of Corollary 4.13 (Cumulative reward setting)

We first recall that similarly to Section E.3, for cumulative reward setting, we redefine the error sequence (e_t) as

$$e_t = d_t - \nabla J_H(\theta_t),$$

where the truncated cumulative reward $J_H(\theta)$ is defined as

$$J_H(\theta) = \mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t r(s_t, a_t)\right].$$

Now we state a complete version of Corollary 4.13, which we shall prove in this section.

Corollary F.3 (Global convergence of N-VR-PG). *Let Assumptions 4.1 and 4.9 hold. Additionally, let Assumption 4.10 be satisfied with $\bar{\epsilon} \geq \frac{\alpha_0(1-\gamma)}{2\ell_\theta(T+1)^a}$ for some integer $T \geq 1$ and reals $\alpha_0 > 0$, $a \in (0,1)$. Set $\alpha_t = \frac{\alpha_0}{(T+1)^a}$, $\eta_t = \frac{2}{t+2}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Then the output θ_T of N-VR-PG (see Algorithm 4) satisfies*

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{\alpha_0 V}{(T+1)^{a-\frac{1}{2}}}\right),$$

where J^* is the optimal expected return and V is defined in Lemma E.13. Therefore, the sample complexity to achieve $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon$ is $\mathcal{O}\left(\varepsilon^{\frac{-2}{2a-1}}\right)$.

Remark F.4. If we are allowed to select α_0 based on the problem parameters (only the bound on $(1-\gamma)$ is actually needed here), then the dependence on $(1-\gamma)^{-1}$ in the above theorem can be made arbitrary small.

Proof. By Lemma E.15, we have the control of the variance sequence for $\eta_t = \frac{2}{t+2}$ as

$$\mathbb{E}\left[\|e_t\|\right] \leq 4V\eta_t \cdot t^{1/2} + 2(L_g + GC_w^{1/2})\alpha_{t-1} \cdot t^{1/2}. \tag{125}$$

Define $\delta_t \overset{\text{def}}{=} \mathbb{E}\left[J(\theta^*) - J(\theta_t)\right]$, where in the cumulative reward case $F(\lambda(\theta)) = J(\theta)$. Let $\alpha_t = \alpha$ for all $t = 0, \ldots, T-1$. Then applying full expectation to the result of Lemma F.1, we have for $\epsilon \leq \min\left\{\bar{\epsilon}, \frac{\alpha(1-\gamma)}{2\ell_\theta}\right\}$

$$\delta_{t+1} \leq (1-\epsilon)\delta_t + 2\alpha\mathbb{E}\left[\|e_t\|\right] + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon^2 + \frac{4}{3}\alpha D_\lambda\gamma^H + \frac{L_\theta}{2}\alpha^2.$$

Unrolling the recursion from $t = 0$ to $t = T - 1$, we have

$$\begin{aligned}
\delta_T &\leq (1-\epsilon)^T \delta_0 + 2\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\|e_t\|\right] + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon} D_\lambda \gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq (1-\epsilon)^T \delta_0 + 8V\alpha \sum_{t=0}^{T-1} \eta_t \cdot t^{1/2} + 4(L_g + GC_w^{1/2})\alpha^2 T^{1/2} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon} D_\lambda \gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq (1-\epsilon)^T \delta_0 + 8V\alpha (T+1)^{1/2} + 4(L_g + GC_w^{1/2})\alpha^2 T^{1/2} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon} D_\lambda \gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon}\,.
\end{aligned}$$

Notice that $(1-\epsilon)^T \leq \exp\left(T\log(1-\epsilon)\right) \leq \exp\left(-\epsilon T\right)$. Finally setting $\alpha = \frac{\alpha_0}{(T+1)^a}$, for $0 < a < 1$ and $\epsilon = \min\left\{\bar{\epsilon}, \frac{\alpha(1-\gamma)}{2\ell_\theta}\right\} = \frac{\alpha(1-\gamma)}{2\ell_\theta}$, we obtain

$$\begin{aligned}
\delta_T &\leq \exp\left(-\frac{\alpha_0(1-\gamma)}{2\ell_0}T^{1-a}\right) + \frac{8\alpha_0 V}{(T+1)^{a-\frac{1}{2}}} + \frac{4\alpha_0^2(L_g + GC_w^{1/2})}{T^{2a-\frac{1}{2}}} + \frac{4L_\theta l_\theta^2}{(1-\gamma)^2}\epsilon + \frac{4}{3}\frac{\alpha}{\epsilon} D_\lambda \gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq \exp\left(-\frac{\alpha_0(1-\gamma)}{2\ell_0}T^{1-a}\right) + \frac{8\alpha_0 V}{(T+1)^{a-\frac{1}{2}}} + \frac{4\alpha_0^2(L_g + GC_w^{1/2})}{T^{2a-\frac{1}{2}}} + \frac{2L_\theta l_\theta \alpha_0}{(1-\gamma)(T+1)^a} + \frac{4}{3}\frac{\alpha}{\epsilon} D_\lambda \gamma^H + \frac{L_\theta}{2}\frac{\alpha^2}{\epsilon} \\
&\leq \exp\left(-\frac{\alpha_0(1-\gamma)}{2\ell_0}T^{1-a}\right) + \frac{8\alpha_0 V}{(T+1)^{a-\frac{1}{2}}} + \frac{4\alpha_0^2(L_g + GC_w^{1/2})}{T^{2a-\frac{1}{2}}} + \frac{3L_\theta l_\theta \alpha_0}{(1-\gamma)(T+1)^a} + \frac{8\ell_\theta D_\lambda}{3(1-\gamma)}\gamma^H \\
&\leq \mathcal{O}\left(\frac{1}{(T+1)^{a-\frac{1}{2}}}\right),
\end{aligned}$$

where the last step follows by setting $H = (1-\gamma)^{-1}\log(T)$. □

F.3. Global optimality in the cumulative reward setting for continuous state-action space and Gaussian policy

We first present our set of assumptions to derive global convergence results under the Gaussian policy parameterization. We start by assuming that our Gaussian policy parametrization is Fisher-non-degenerate, meaning that the Fisher information matrix induced by the policy parametrization is (uniformly) positive definite. This assumption is standard in the literature (Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022; Masiha et al., 2022; Fatkhullin et al., 2022). We remark that Fatkhullin et al. (2023) recently obtained a $\mathcal{O}(\varepsilon^{-2})$ sample complexity under similar assumptions using a similar proof technique. The key difference between our N-VR-PG method and their (N)-HARPG algorithm is that our algorithm does not require the use of second-order information. The bound of IS weights is automatically ensured by the normalization step of the algorithm and the specific structure of the Gaussian policy parametrization (Lemma E.19).

Assumption F.5. There exists $\mu_F > 0$ such that for every $\theta \in \mathbb{R}^d$, the Fisher information matrix satisfies

$$F_\rho(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[\nabla \log \pi_\theta(a|s) \nabla \log \pi_\theta(a|s)^\top] \succeq \mu_F I_d\,,$$

where $d_\rho^{\pi_\theta}(\cdot) \stackrel{\text{def}}{=} (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}_{\rho,\pi_\theta}(s_t \in \cdot)$ is the discounted state visitation measure.

For Gaussian policies with fixed covariance matrix and linear mean parametrization $\mu_\theta(s) = \phi(s)^\top\theta$, the Fisher information matrix can be written explicitly. Namely, we have $F_\rho(\theta) = \sigma^{-2}\phi(s)\phi(s)^\top$ for every $s \in \mathcal{S}$. Therefore, the above assumption is satisfied if we assume that the feature map $\phi(s)$ has full-row-rank.

Now we introduce an assumption which characterizes the expressivity of our policy parameterization class via the framework of compatible function approximation (Sutton et al., 1999; Agarwal et al., 2021). In order to state this assumption, we first define the advantage function. Define for every policy π the state-action value function $Q^\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ for every $s \in \mathcal{S}, a \in \mathcal{A}$ as:

$$Q^\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 = a, a_0 = a\right].$$

Under the same policy π, the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$ and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are defined for every $s \in \mathcal{S}, a \in \mathcal{A}$ as follows:

$$V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]\,,$$
$$A^\pi(s,a) \stackrel{\text{def}}{=} Q^\pi(s,a) - V^\pi(s)\,.$$

Now we are ready to state the *compatible function approximation error* assumption.

Assumption F.6. There exists $\varepsilon_{\text{bias}} \geq 0$ s.t. for every $\theta \in \mathbb{R}^d$, the transfer error satisfies:

$$\mathbb{E}[(A^{\pi_\theta}(s,a) - (1-\gamma)w^*(\theta)^\top \nabla \log \pi_\theta(a|s))^2] \leq \varepsilon_{\text{bias}}\,,$$

where A^{π_θ} is the advantage function, $w^*(\theta) \stackrel{\text{def}}{=} F_\rho(\theta)^\dagger \nabla J(\theta)$ where $F_\rho(\theta)^\dagger$ is the pseudo-inverse of the matrix $F_\rho(\theta)$ and expectation is taken over $s \sim d_\rho^{\pi^*}$, $a \sim \pi^*(\cdot|s)$ where π^* is an optimal policy (maximizing $J(\pi)$).

The above assumption requires that the policy parametrization π_θ should be able to approximate the advantage function A^{π_θ} by the score function $\nabla \log \pi_\theta$. Naturally $\varepsilon_{\text{bias}}$ is necessarily positive for a parameterization π_θ that does not cover the set of all stochastic policies and $\varepsilon_{\text{bias}}$ is small for a rich neural policy (Wang et al., 2020). We note that this is a common assumption which was used for instance in (Agarwal et al., 2021; Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022).

Equipped with Assumptions E.17, F.5, F.6, and following the derivations of Ding et al. (2022), we obtain a relaxed weak gradient dominance inequality.

Lemma F.7 (Relaxed weak gradient domination, (Ding et al., 2022))**.** *Let Assumptions E.17, F.5 and F.6 hold. Then*

$$\forall\, \theta \in \mathbb{R}^d, \quad \varepsilon' + \|\nabla J(\theta)\| \geq \sqrt{2\mu}\,(J^* - J(\theta))\,, \tag{126}$$

where J^* is the optimal expected return, $\varepsilon' = \frac{\mu_F\sqrt{\varepsilon_{bias}}}{M_g(1-\gamma)}$ and $\mu = \frac{\mu_F^2}{2M_g^2}$.

Corollary F.8 (Global convergence of N-VR-PG**).** *Let Assumptions E.17, F.5 and F.6 hold. Set $\alpha_t = \frac{3}{\sqrt{2\mu}(T+1)^a}$, for some $0 < a < 1$, $\eta_t = \frac{2}{t+1}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Then the output θ_T of* N-VR-PG *(see Algorithm 4) satisfies*

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^{3/2}(T+1)^{a-\frac{1}{2}}}\right) + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma},$$

where J^* is the optimal expected return. Therefore, the sample complexity to achieve $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma}$ is $\mathcal{O}\left(\varepsilon^{\frac{-2}{2a-1}}\right)$.

Proof. As in the case of softmax parametrization, given the result of Lemma E.18, and following the steps in the proof of Lemma E.15, we can derive the control of the variance sequence for $\eta_t = \frac{2}{t+1}$ as

$$\mathbb{E}\left[\|e_t\|\right] \leq 4V\eta_t \cdot t^{1/2} + 2(L_g + GC_w^{1/2})\alpha_{t-1} \cdot t^{1/2}. \tag{127}$$

Similarly to Lemma E.1, we can obtain

$$\begin{aligned} J(\theta_{t+1}) &\geq J(\theta_t) + \frac{\alpha_t}{3}\|\nabla J(\theta_t)\| - 2\alpha_t\|e_t\| - \frac{4}{3}D_g\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2 \\ &\geq J(\theta_t) + \frac{\alpha_t\sqrt{2\mu}}{3}\left(J^* - J(\theta_t)\right) - 2\alpha_t\|e_t\| - \frac{4}{3}D_g\gamma^H\alpha_t - \frac{L_\theta}{2}\alpha_t^2 - \frac{\varepsilon'\alpha_t}{3}, \end{aligned} \tag{128}$$

where in the last step we applied the relaxed weak gradient dominance condition (Lemma F.7). Now we define $\delta_t \stackrel{\text{def}}{=} \mathbb{E}\left[J(\theta^*) - J(\theta_t)\right]$. Let $\alpha_t = \alpha$ for all $t = 0, \ldots, T$. Then applying full expectation to the result of Lemma F.1, we have

$$\delta_{t+1} \leq \left(1 - \frac{\alpha\sqrt{2\mu}}{3}\right)\delta_t + 2\alpha\mathbb{E}\left[\|e_t\|\right] + \frac{4}{3}D_g\gamma^H\alpha + \frac{L_\theta}{2}\alpha^2 + \frac{\varepsilon'\alpha}{3}.$$

Unrolling the recursion from $t = 0$ to $t = T - 1$, we have

$$\begin{aligned}
\delta_T &\leq \left(1 - \frac{\alpha\sqrt{2\mu}}{3}\right)^T \delta_0 + 2\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\|e_t\|\right] + \frac{4}{\sqrt{2\mu}} D_g \gamma^H + \frac{3}{\sqrt{2\mu}} \frac{L_\theta}{2} \alpha + \frac{\varepsilon'}{\sqrt{2\mu}} \\
&\leq \left(1 - \frac{\alpha\sqrt{2\mu}}{3}\right)^T \delta_0 + 8V\alpha \sum_{t=0}^{T-1} \eta_t \cdot t^{1/2} + 4(L_g + GC_w^{1/2})\alpha^2 T^{1/2} + \frac{4}{\sqrt{2\mu}} D_g \gamma^H + \frac{3}{\sqrt{2\mu}} \frac{L_\theta}{2} \alpha + \frac{\varepsilon'}{\sqrt{2\mu}} \\
&\leq \left(1 - \frac{\alpha\sqrt{2\mu}}{3}\right)^T \delta_0 + 8V\alpha (T+1)^{1/2} + 4(L_g + GC_w^{1/2})\alpha^2 T^{1/2} + \frac{4}{\sqrt{2\mu}} D_g \gamma^H + \frac{3}{\sqrt{2\mu}} \frac{L_\theta}{2} \alpha + \frac{\varepsilon'}{\sqrt{2\mu}} .
\end{aligned}$$

Finally, setting $\alpha = \frac{3}{\sqrt{2\mu}(T+1)^a}$, for $0 < a < 1$ and noticing that $(1 - (T+1)^{-a})^T \leq \exp\left(T \log(1 - (T+1)^{-a})\right) \leq \exp\left(-T^{1-a}\right)$, we obtain

$$\begin{aligned}
\delta_T &\leq \exp\left(-T^{1-a}\right) \delta_0 + \frac{24V}{\sqrt{2\mu}(T+1)^{a-\frac{1}{2}}} + \frac{18(L_g + GC_w^{1/2})}{\mu \cdot T^{2a-\frac{1}{2}}} + \frac{4}{\sqrt{2\mu}} D_g \gamma^H + \frac{9L_\theta}{8\mu} \frac{1}{(T+1)^a} + \frac{\varepsilon'}{\sqrt{2\mu}} \\
&\leq \mathcal{O}\left(\frac{V}{\sqrt{\mu}(T+1)^{a-\frac{1}{2}}}\right) + \frac{\varepsilon'}{\sqrt{2\mu}},
\end{aligned}$$

where the last step follows by setting $H = (1-\gamma)^{-1} \log(T)$. It only remains to notice from Lemma E.18 and F.7 that

$$\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{\text{bias}}}}{M_g(1-\gamma)} = \mathcal{O}\left(\frac{1}{1-\gamma}\right), \qquad V = \frac{M_g \|r\|_\infty}{(1-\gamma)^{3/2}} = \mathcal{O}\left(\frac{1}{(1-\gamma)^{3/2}}\right).$$

□

G. Proofs for Section 5: Large state-action space setting

G.1. Unbiased estimates of the occupancy measure at state-action pairs

Notation. For a given set A, the indicator function $\mathbb{1}_A$ is equal to one on the set A and zero otherwise.

In this section, we provide two different estimators: the first one is a Monte-Carlo estimate of the truncated occupancy measure whereas the second one is an unbiased estimate of the true occupancy measure. Notice that we can also slightly modify the second estimator to a obtain a minibatch estimator via sampling (independently) similarly N different state-action pairs $(s_H^{(i)}, a_H^{(i)})_{0 \leq i \leq N}$ via the same sampling procedure as in Algorithm 6 and averaging out the outputs, i.e., considering the following estimator:

$$\hat{\lambda}^{\pi_\theta}(s,a) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{s_H^{(i)} = s,\, a_H^{(i)} = a\}} . \tag{129}$$

Algorithm 5 Monte-Carlo estimate of the truncated state-action occupancy measure for (s,a): $\lambda_H^{\pi_\theta}(s,a)$

Input: Initial state distribution ρ, state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, policy π_θ, discount factor $\gamma \in [0,1)$, truncation horizon H.
Sample a trajectory $\tau = (s_t, a_t)_{0 \leq t \leq H-1}$ from the MDP controlled by policy π_θ
$\hat{\lambda}_H^{\pi_\theta}(s,a) = \sum_{t=0}^{H} \gamma^t \mathbb{1}_{\{s_t = s,\, a_t = a\}}$
Return: $\hat{\lambda}_H^{\pi_\theta}(s,a)$.

Algorithm 6 Unbiased estimator of the state-action occupancy measure for (s,a): $\lambda^{\pi_\theta}(s,a)$

Input: Initial state distribution ρ, state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, policy π_θ, discount factor $\gamma \in [0,1)$, $h=0$.
$s_0 \sim \rho, a_0 \sim \pi_\theta(\cdot|s_0)$
Draw H from the geometric distribution Geom$(1-\gamma)$
for $h = 0, \ldots, H-1$ **do**
  $s_{h+1} \sim P(\cdot|s_h, a_h)$; $a_{h+1} \sim \pi_\theta(\cdot|s_h)$
end for
$\hat{\lambda}^{\pi_\theta}(s,a) = \mathbb{1}_{\{s_H=s,\, a_H=a\}}$
Return: $\hat{\lambda}^{\pi_\theta}(s,a)$.

G.2. Proof of Theorem 5.4: Convergence analysis under bounded statistical and approximation errors

We first state a more detailed version of Theorem 5.4.

Theorem G.1. *Let Assumptions 4.1, 4.2, 5.2 and 5.3 hold true. In addition, suppose that there exists $\rho_{\min} > 0$ s.t. the initial distribution ρ satisfies $\rho(s) \geq \rho_{\min}$ for all $s \in \mathcal{S}$. Let $T \geq 1$ be an integer and let (θ_t) be the sequence generated by Algorithm 3 with a positive step size $\alpha \leq \min(1/\sqrt{5\tilde{C}_1}, 1/2L_\theta)$ (see $\tilde{C}_1$ below) and batch size $N \geq 1$. Then, we have*

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \frac{16(F^* - \mathbb{E}[F(\lambda(\theta_1))]) + \alpha\tilde{C}_4}{\alpha T} + \frac{\tilde{C}_3}{N} + 2D_\lambda^2\gamma^{2H} + \tilde{C}_2(\epsilon_{stat} + \epsilon_{approx})\,, \tag{130}$$

where $\bar{\theta}_T$ be a random iterate drawn uniformly at random from $\{\theta_1, \cdots, \theta_T\}$, $\tilde{C}_1 \overset{def}{=} \frac{48l_\psi^3 L_{\lambda,\infty}^2}{(1-\gamma)^6}, \tilde{C}_2 \overset{def}{=} \frac{48l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\frac{|\mathcal{A}|}{\rho_{\min}}$, $\tilde{C}_3 \overset{def}{=} \frac{24l_\lambda^2 l_\psi^2}{(1-\gamma)^4}, \tilde{C}_4 \overset{def}{=} \frac{8l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$ and D_λ is defined in Lemma H.2.

Proof. We introduce the shorthand notation $u_t \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^N g(\tau_t^{(i)}, \theta_t, r_{t-1})$ for this proof. The smoothness of the objective function $\theta \mapsto F(\lambda(\theta))$ (see Lemma H.1) together with the update rule of the sequence (θ_t) yields

$$\begin{aligned}
F(\lambda(\theta_{t+1})) &\geq F(\lambda(\theta_t)) + \langle \nabla_\theta F(\lambda(\theta_t)), \theta_{t+1} - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= F(\lambda(\theta_t)) + \alpha\langle \nabla_\theta F(\lambda(\theta_t)), u_t\rangle - \frac{L_\theta\alpha^2}{2}\|u_t\|^2 \\
&= F(\lambda(\theta_t)) + \alpha\langle \nabla_\theta F(\lambda(\theta_t)) - u_t, u_t\rangle + \alpha\left(1 - \frac{L_\theta\alpha}{2}\right)\|u_t\|^2 \\
&\geq F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2 - \frac{\alpha}{2}\|u_t\|^2 + \alpha\left(1 - \frac{L_\theta\alpha}{2}\right)\|u_t\|^2 \\
&= F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2 + \frac{\alpha}{2}(1 - L_\theta\alpha)\|u_t\|^2 \\
&\overset{(i)}{\geq} F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2 + \frac{\alpha}{4}\|u_t\|^2 \\
&= F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2 + \frac{\alpha}{8}\|u_t\|^2 + \frac{\alpha}{8}\|u_t\|^2 \\
&\overset{(ii)}{\geq} F(\lambda(\theta_t)) + \frac{\alpha}{16}\|\nabla_\theta F(\lambda(\theta_t))\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2 + \frac{\alpha}{8}\|u_t\|^2\,,
\end{aligned} \tag{131}$$

where (i) follows from the condition $\alpha \leq 1/2L_\theta$ and (ii) from $\frac{1}{2}\|\nabla_\theta F(\lambda(\theta_t))\|^2 \leq \|u_t\|^2 + \|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2$.

We now control the last error term in the above inequality in expectation. Observe first that

$$\begin{aligned}
\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2] &\leq 2\,\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \nabla_\theta F(\lambda_H(\theta_t))\|^2] + 2\mathbb{E}[\|\nabla_\theta F(\lambda_H(\theta_t)) - u_t\|^2] \\
&\leq 2D_\lambda^2\gamma^{2H} + 2\mathbb{E}[\|\nabla_\theta F(\lambda_H(\theta_t)) - u_t\|^2]\,,
\end{aligned} \tag{132}$$

where the last inequality stems from Lemma H.2. Now, it remains to control $\mathbb{E}[\|\nabla_\theta F(\lambda_H(\theta_t)) - u_t\|^2]$. Using the

notation $r_t \overset{\text{def}}{=} \nabla_\lambda F(\lambda_H(\theta_t))$, we have the following decomposition:

$$\nabla_\theta F(\lambda_H(\theta_t)) - u_t = \nabla_\theta F(\lambda_H(\theta_t)) - [\nabla_\theta \lambda(\theta_t)]^T r^*_{t-1} + [\nabla_\theta \lambda(\theta_t)]^T r^*_{t-1} - [\nabla_\theta \lambda(\theta_t)]^T r_{t-1} + [\nabla_\theta \lambda(\theta_t)]^T r_{t-1} - u_t \,. \tag{133}$$

Then it follows that

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta F(\lambda_H(\theta_t)) - u_t\|^2] \leq 3\mathbb{E}[\|\nabla_\theta F(\lambda_H(\theta_t)) - [\nabla_\theta \lambda(\theta_t)]^T r^*_{t-1}\|^2] + 3\mathbb{E}[\|[\nabla_\theta \lambda(\theta_t)]^T (r^*_{t-1} - r_{t-1})\|^2] \\ + 3\mathbb{E}[\|[\nabla_\theta \lambda(\theta_t)]^T r_{t-1} - u_t\|^2] \,. \end{aligned} \tag{134}$$

We control each term in the above decomposition separately in what follows.

Term 1 in (134): For this term, we have the following series of inequalities

$$\begin{aligned}\|\nabla_\theta F(\lambda_H(\theta_t)) - [\nabla_\theta \lambda(\theta_t)]^T r^*_{t-1}\|^2 &= \|[\nabla_\theta \lambda(\theta_t)]^T (r^*_t - r^*_{t-1})\|^2 \\ &\overset{(a)}{\leq} \frac{4l_\psi^2}{(1-\gamma)^4} \|r^*_{t-1} - r^*_t\|^2_\infty \\ &\overset{(b)}{\leq} \frac{4l_\psi^2 L^2_{\lambda,\infty}}{(1-\gamma)^4} \|\lambda_H(\theta_{t-1}) - \lambda_H(\theta_t)\|_1^2 \\ &\overset{(c)}{\leq} \frac{8l_\psi^3 L^2_{\lambda,\infty}}{(1-\gamma)^6} \|\theta_t - \theta_{t-1}\|^2 \\ &\overset{(d)}{=} \frac{8l_\psi^3 L^2_{\lambda,\infty}}{(1-\gamma)^6} \|u_{t-1}\|^2 \cdot \alpha^2 \,, \end{aligned} \tag{135}$$

where (a) follows from similar derivations to (29)-(31) using (5), (b) stems from Assumption 4.2, (c) is an immediate consequence of Lemma H.1-(ii) and (d) uses the update rule of Algorithm 3.

Term 2 in (134): For this term, we start with the following inequalities:

$$\begin{aligned}\mathbb{E}[\|[\nabla_\theta \lambda(\theta_t)]^T (r^*_{t-1} - r_{t-1})\|^2] &\overset{(i)}{\leq} \frac{4l_\psi^2}{(1-\gamma)^4} \mathbb{E}[\|r_{t-1} - r^*_{t-1}\|^2_\infty] \\ &\overset{(ii)}{\leq} \frac{4l_\psi^2 L^2_\lambda}{(1-\gamma)^4} \mathbb{E}[\|\hat{\lambda}_{t-1} - \lambda_H(\theta_{t-1})\|^2] \,, \end{aligned} \tag{136}$$

where (i) follows from similar derivations to (29)-(31) using (5) and (ii) follows from Assumption 4.2. Then we decompose and upper bound the above error as follows:

$$\begin{aligned}\mathbb{E}[\|\hat{\lambda}_{t-1} - \lambda_H(\theta_{t-1})\|^2] &= \mathbb{E}[\|\langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_{t-1}} \rangle - \lambda_H(\theta_{t-1})\|^2] \\ &= \mathbb{E}[\|\langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle + \langle \phi(\cdot,\cdot), \omega_*(\theta_{t-1}) \rangle - \lambda_H(\theta_{t-1})\|^2] \\ &\leq 2\,\mathbb{E}[\|\langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle\|^2] + 2\,\mathbb{E}[\|\langle \phi(\cdot,\cdot), \omega_*(\theta_{t-1}) \rangle - \lambda_H(\theta_{t-1})\|^2] \,. \end{aligned} \tag{137}$$

Our task now is to upper bound each one of the above errors, the first one being related to the statistical error whereas the second one relates to the approximation error. Recall the definition of the regression loss function for every $\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^m$,

$$L_\theta(\omega) = \mathbb{E}_{s\sim\rho, a\sim\mathcal{U}(\mathcal{A})}[(\lambda_H^{\pi_\theta}(s,a) - \langle \phi(s,a), \omega \rangle)^2] \,, \tag{138}$$

where $\mathcal{U}(\mathcal{A})$ is the uniform distribution over the action space $\mathcal{A}$.

(a) Bounding term 1 in (137) by the statistical error. First, observe for this term that

$$\begin{aligned}\mathbb{E}[\|\langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle\|^2] &= \mathbb{E}\left[\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \langle \phi(s,a), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle^2\right] \\ &\leq \frac{|\mathcal{A}|}{\rho_{\min}} \mathbb{E}\left[\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \frac{\rho(s)}{|\mathcal{A}|} \langle \phi(s,a), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle^2\right] \\ &= \frac{|\mathcal{A}|}{\rho_{\min}} \mathbb{E}\left[\mathbb{E}_{s\sim\rho, a\sim\mathcal{U}(\mathcal{A})}[\langle \phi(s,a), \hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1}) \rangle^2]\right] \,. \end{aligned} \tag{139}$$

Then, we have for all $\omega \in \mathbb{R}^m$,

$$
\begin{aligned}
&L_{\theta_{t-1}}(\omega) - L_{\theta_{t-1}}(\omega_*(\theta_{t-1})) \\
&= \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[(\langle\phi(s,a),\omega\rangle - \lambda^{\pi_{\theta_{t-1}}}(s,a))^2] - L_{\theta_{t-1}}(\omega_*(\theta_{t-1})) \\
&= \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[(\langle\phi(s,a),\omega-\omega_*(\theta_{t-1})\rangle + \langle\phi(s,a),\omega_*(\theta_{t-1})\rangle - \lambda^{\pi_{\theta_{t-1}}}(s,a))^2] - L_{\theta_{t-1}}(\omega_*(\theta_{t-1})) \\
&= \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[\langle\phi(s,a),\omega-\omega_*(\theta_{t-1})\rangle^2] + 2\langle\omega-\omega_*(\theta_{t-1}), \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[(\langle\phi(s,a),\omega_*(\theta_{t-1})\rangle - \lambda^{\pi_{\theta_{t-1}}}(s,a))\phi(s,a)]\rangle \\
&= \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[\langle\phi(s,a),\omega-\omega_*(\theta_{t-1})\rangle^2] + \langle\omega-\omega_*(\theta_{t-1}), \nabla_\omega L_{\theta_{t-1}}(\omega_*(\theta_{t-1}))\rangle \\
&\geq \mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[\langle\phi(s,a),\omega-\omega_*(\theta_{t-1})\rangle^2]\,, \qquad (140)
\end{aligned}
$$

where the last inequality stems from the first-order optimality condition for $\omega_*(\theta_{t-1}) \in \arg\min_\omega L_{\theta_{t-1}}(\omega)$, which gives the inequality $\langle\omega-\omega_*(\theta_{t-1}), \nabla_\omega L_{\theta_{t-1}}(\omega_*(\theta_{t-1}))\rangle \geq 0$ for every $\omega \in \mathbb{R}^m$.

Combining (139) with (140) and using Assumption 5.2, we obtain

$$
\mathbb{E}[\|\langle\phi(\cdot,\cdot),\hat{\omega}_{\theta_{t-1}} - \omega_*(\theta_{t-1})\rangle\|^2] \leq \frac{|\mathcal{A}|}{\rho_{\min}}\mathbb{E}[L_{\theta_{t-1}}(\hat{\omega}_{\theta_{t-1}}) - L_{\theta_{t-1}}(\omega_*(\theta_{t-1}))] \leq \frac{|\mathcal{A}|}{\rho_{\min}}\epsilon_{\text{stat}}\,. \qquad (141)
$$

(b) Bounding term 2 in (137) by the approximation error. Similar derivations as for the previous term yield

$$
\begin{aligned}
\mathbb{E}[\|\langle\phi(\cdot,\cdot),\omega_*(\theta_{t-1})\rangle - \lambda_H(\theta_{t-1})\|^2] &= \mathbb{E}\left[\sum_{s\in\mathcal{S},a\in\mathcal{A}}(\langle\phi(s,a),\omega_*(\theta_{t-1})\rangle - \lambda_H^{\pi_{\theta_{t-1}}}(s,a))^2\right] \\
&\leq \frac{|\mathcal{A}|}{\rho_{\min}}\mathbb{E}\left[\sum_{s\in\mathcal{S},a\in\mathcal{A}}\frac{\rho(s)}{|\mathcal{A}|}(\langle\phi(s,a),\omega_*(\theta_{t-1})\rangle - \lambda_H^{\pi_{\theta_{t-1}}}(s,a))^2\right] \\
&= \frac{|\mathcal{A}|}{\rho_{\min}}\mathbb{E}\left[\mathbb{E}_{s\sim\rho,a\sim\mathcal{U}(\mathcal{A})}[(\langle\phi(s,a),\omega_*(\theta_{t-1})\rangle - \lambda_H^{\pi_{\theta_{t-1}}}(s,a))^2]\right] \\
&= \frac{|\mathcal{A}|}{\rho_{\min}}\mathbb{E}[L_{\theta_{t-1}}(\omega_*(\theta_{t-1}))] \\
&\leq \frac{|\mathcal{A}|}{\rho_{\min}}\epsilon_{\text{approx}}\,. \qquad (142)
\end{aligned}
$$

Combining (136), (137), (141) and (142) yields

$$
\mathbb{E}[\|[\nabla_\theta\lambda(\theta_t)]^T(r^*_{t-1} - r_{t-1})\|^2] \leq \frac{8l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\frac{|\mathcal{A}|}{\rho_{\min}}(\epsilon_{\text{stat}} + \epsilon_{\text{approx}})\,. \qquad (143)
$$

Term 3 in (134): For this last term, we have

$$
\begin{aligned}
\mathbb{E}[\|[\nabla_\theta\lambda(\theta_t)]^T r_{t-1} - u_t\|^2] &\overset{(a)}{=} \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N([\nabla_\theta\lambda(\theta_t)]^T r_{t-1} - g(\tau_t^{(i)},\theta_t,r_{t-1}))\right\|^2\right] \\
&\overset{(b)}{=} \frac{1}{N}\mathbb{E}[\|g(\tau_t^{(i)},\theta_t,r_{t-1}) - [\nabla_\theta\lambda(\theta_t)]^T r_{t-1}\|^2] \\
&\overset{(c)}{\leq} \frac{1}{N}\mathbb{E}[\|g(\tau_t^{(i)},\theta_t,r_{t-1})\|^2] \\
&\overset{(d)}{\leq} \frac{4l_\lambda^2 l_\psi^2}{N(1-\gamma)^4}\,, \qquad (144)
\end{aligned}
$$

where (a) stems from the definition of u_t, (b) follows from using Lemma E.3 and recalling that the trajectories $(\tau_t^{(i)})_{1\leq i\leq N}$ are independently drawn in Algorithm 3, (c) is due to the inequality $\mathrm{Var}(X) \leq \mathbb{E}[\|X\|^2]$ for any random vector $X \in \mathbb{R}^d$ and (d) uses a similar bound to (70).

Combining (134), (135), (143) and (144) together with (132) gives

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2] \leq \tilde{C}_1\alpha^2\|u_{t-1}\|^2 + \tilde{C}_2(\epsilon_{\text{stat}} + \epsilon_{\text{approx}}) + \frac{\tilde{C}_3}{N} + 2D_\lambda^2\gamma^{2H}\,, \tag{145}$$

where $\tilde{C}_1 = \frac{48l_\psi^3 L_{\lambda,\infty}^2}{(1-\gamma)^6}, \tilde{C}_2 = \frac{48l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\frac{|\mathcal{A}|}{\rho_{\min}}$ and $\tilde{C}_3 = \frac{24l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$.

Rearranging (131), dividing by $\frac{\alpha}{16}$ and taking full expectation yields

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t))\|^2] \leq \frac{16}{\alpha}(F(\lambda(\theta_{t+1})) - F(\lambda(\theta_t))) - 2\mathbb{E}[\|u_t\|^2] + 10\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - u_t\|^2]\,. \tag{146}$$

Plugging (145) into (146), summing the resulting inequality for $t = 1, \cdots, T$ and dividing by T gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t))\|^2] \leq \frac{16}{\alpha T}\mathbb{E}[F(\lambda(\theta_{T+1})) - F(\lambda(\theta_1))] + \frac{1}{T}\sum_{t=1}^{T}\left(10\tilde{C}_1\alpha^2\mathbb{E}[\|u_{t-1}\|^2] - 2\mathbb{E}[\|u_t\|^2]\right) + \tilde{C}_2(\epsilon_{\text{stat}} + \epsilon_{\text{approx}}) + \frac{\tilde{C}_3}{N} + 2D_\lambda^2\gamma^{2H} \tag{147}$$

Then, we upper bound the remaining sum in the right-hand side of (147) as follows:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(10\tilde{C}_1\alpha^2\mathbb{E}[\|u_{t-1}\|^2] - 2\mathbb{E}[\|u_t\|^2]\right) &= \frac{1}{T}\sum_{t=1}^{T}(10\tilde{C}_1\alpha^2 - 2)\mathbb{E}[\|u_{t-1}\|^2] + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\|u_{t-1}\|^2 - \|u_t\|^2] \\ &\overset{(a)}{\leq} \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\|u_{t-1}\|^2 - \|u_t\|^2] \\ &\overset{(b)}{\leq} \frac{2\mathbb{E}[\|u_0\|^2]}{T} \\ &\overset{(c)}{\leq} \frac{\tilde{C}_4}{T}\,,\end{aligned} \tag{148}$$

where $\tilde{C}_4 = \frac{8l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$, (a) stems from the condition $\alpha \leq \frac{1}{\sqrt{5\tilde{C}_1}}$, (b) follows telescoping the sum and upper bounding the remaining resulting negative term by zero and (c) is a consequence of a similar bound to (70).

Finally, we obtain

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \frac{16(F^* - \mathbb{E}[F(\lambda(\theta_1))]) + \alpha\tilde{C}_4}{\alpha T} + \frac{\tilde{C}_3}{N} + 2D_\lambda^2\gamma^{2H} + \tilde{C}_2(\epsilon_{\text{stat}} + \epsilon_{\text{approx}})\,, \tag{149}$$

where $\bar{\theta}_T$ be a random iterate drawn uniformly at random from $\{\theta_1, \cdots, \theta_T\}$. This concludes the proof. □

G.3. Proof of Corollary 5.6: Sample complexity analysis

In order to establish the total sample complexity of our algorithm, we shall use Theorem 1 in Bach & Moulines (2013) for the least-mean-square algorithm corresponding to SGD for least-squares regression to explicit the number of samples needed in the occupancy measure estimation subroutine of Algorithm 2. In other words, our objective here is to precise the number of iterations of SGD needed to approximately solve our regression problem. In particular, we will show that we can achieve $\epsilon_{\text{stat}} = \mathcal{O}(1/K)$ where K is the number of iterations of the SGD subroutine. We first report Theorem 1 from Bach & Moulines (2013) before applying it to our specific case.

Theorem G.2 (Theorem 1, (Bach & Moulines, 2013)). *Let $\mathcal{H}$ be an m-dimensional Euclidean space with $m \geq 1$. Let $(x_n, z_n) \in \mathcal{H} \times \mathcal{H}$ be independent and identically distributed observations. Assume the following:*

(i) The expectations $\mathbb{E}[\|x_n\|^2]$ and $\mathbb{E}[\|z_n\|^2]$ are finite; the covariance matrix $\mathbb{E}[x_n x_n^T]$ is invertible.

(ii) The global minimum of $f(\omega) = \frac{1}{2}\mathbb{E}[\langle \omega, x_n\rangle^2 - 2\langle \omega, z_n\rangle]$ *is attained at a certain* $\omega_* \in \mathcal{H}$. *Denoting by* $\xi_n \overset{def}{=} z_n - \langle \omega_*, x_n\rangle x_n$ *the residual, assume that* $\mathbb{E}[\xi_n] = 0$.

(iii) There exist $R > 0, \sigma > 0$ *s.t.* $\mathbb{E}[\xi_n \xi_n^T] \preccurlyeq \sigma^2 \mathbb{E}[x_n x_n^T]$ *and* $\mathbb{E}[\|x_n\|^2 x_n x_n^T] \preccurlyeq R^2 \mathbb{E}[x_n x_n^T]$, *where for two matrices* $A, B \in \mathbb{R}^{m\times m}$, $A \preccurlyeq B$ *if and only if* $B - A$ *is positive semi-definite.*

Consider the Stochastic Gradient Descent (SGD) recursion started at $\omega_0 \in \mathcal{H}$ *and defined for every integer* $n \geq 1$ *as*

$$\omega_n = \omega_{n-1} - \beta' \left(\langle \omega_{n-1}, x_n\rangle x_n - z_n\right), \tag{150}$$

where $\beta' > 0$. *Then for a constant step size* $\beta' = \frac{1}{4R^2}$ *the averaged iterate* $\bar{\omega}_n \overset{def}{=} \frac{1}{n+1}\sum_{k=0}^n \omega_k$ *satisfies*

$$\mathbb{E}[f(\bar{\omega}_n) - f(\omega_*)] \leq \frac{2}{n}(\sigma\sqrt{m} + R\|\omega_0 - \omega_*\|)^2 . \tag{151}$$

Proof of Corollary 5.6. It follows from Theorem 5.4 that

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \frac{16(F^* - \mathbb{E}[F(\lambda(\theta_1))]) + \tilde{C}_4}{\alpha T} + \frac{\tilde{C}_3}{N} + 2D_\lambda^2 \gamma^{2H} + \tilde{C}_2(\epsilon_{\text{stat}} + \epsilon_{\text{approx}}) , \tag{152}$$

where $\bar{\theta}_T$ is a random iterate drawn uniformly at random from $\{\theta_1, \cdots, \theta_T\}$. We now upper bound the statistical error ϵ_{stat} as a function of the number K of SGD iterations (see Algorithm 2) by applying Theorem G.2. Let $\omega_* \in \arg\min_\omega L_\theta(\omega)$ where $\theta \in \mathbb{R}^d$ is fixed (at each iteration of Algorithm 3). To do so, we successively verify each assumption of the latter theorem in the Euclidean space $\mathbb{R}^m$. Recall from (12) that the stochastic gradient of the loss function $L_\theta(\omega)$ is given for every $\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^m$ by

$$\hat{\nabla}_\omega L_\theta(\omega) \overset{\text{def}}{=} 2(\langle \phi(s,a), \omega\rangle - \hat{\lambda}_H^{\pi_\theta}(s,a))\, \phi(s,a) . \tag{153}$$

Note here that we consider the unbiased estimator $\hat{\lambda}_H^{\pi_\theta}(s,a)$ of the truncated state-action occupancy measure as computed in Algorithm 5.

Remark G.3. One could also consider the unbiased estimator $\hat{\lambda}_H^{\pi_\theta}(s,a)$ of the true state-action occupancy measure (without truncation) using Algorithm 6 and slightly modify the definition of the expected loss with $\hat{\lambda}^{\pi_\theta}(s,a)$ instead of $\hat{\lambda}_H^{\pi_\theta}(s,a)$. The latter procedure would lead to the same result since the truncation error can be made as small as desired via setting the horizon large enough, the error being of the order of γ^H .

Take $x_n = \phi(s,a) \in \mathbb{R}^m, z_n = \hat{\lambda}_H^{\pi_\theta}(s,a)\phi(s,a) \in \mathbb{R}^m$. The observations (x_n, z_n) are indeed independent and identically distributed (for each state-action pair (s,a) sample).

(i) Given Assumption 5.5, we have $\mathbb{E}[\|x_n\|^2] = \mathbb{E}[\|\phi(s,a)\|^2] \leq B^2$. Similarly we have$\mathbb{E}[\|z_n\|^2] \leq B^2/(1-\gamma)^2$. Moreover, the covariance matrix $\mathbb{E}[\phi(s,a)\phi(s,a)^T]$ has full rank by Assumption 5.5.

(ii) Take $f = L_\theta$. Define the residual $\xi \overset{\text{def}}{=} (\hat{\lambda}_H^{\pi_\theta}(s,a) - \langle \omega_*, \phi(s,a)\rangle)\phi(s,a)$. Then we conclude the verification of the second item by observing that

$$\mathbb{E}[\xi] = \mathbb{E}\left[\frac{1}{2}\hat{\nabla}_\omega L_\theta(\omega_*)\right] = \frac{1}{2}\nabla_\omega L_\theta(\omega_*) = 0 , \tag{154}$$

where the last identity stems from the definition of the optimal solution ω_* .

(iii) As for this last item, recall again that $\|\phi(s,a)\| \leq B$ which immediately implies that $\mathbb{E}[\|x_n\|^2 x_n x_n^T] \preccurlyeq R^2\mathbb{E}[x_n x_n^T]$ with $R = B$. It remains to show that the covariance matrix of ξ satisfies $\mathbb{E}[\xi\xi^T] \preccurlyeq \sigma^2\mathbb{E}[\phi(s,a)\phi(s,a)^T]$ for some positive constant σ that we will now determine. First, we write

$$\begin{aligned}\mathbb{E}[\xi\xi^T] &= \mathbb{E}[(\hat{\lambda}_H^{\pi_\theta}(s,a) - \langle \omega_*, \phi(s,a)\rangle)^2 \phi(s,a)\phi(s,a)^T] \\ &= \mathbb{E}\left[\mathbb{E}[(\hat{\lambda}_H^{\pi_\theta}(s,a) - \langle \omega_*, \phi(s,a)\rangle)^2 | s,a]\phi(s,a)\phi(s,a)^T\right] ,\end{aligned} \tag{155}$$

where the conditional expectation $\mathbb{E}[\cdot|s,a]$ is w.r.t. randomness induced by sampling the state-action pair (s,a). Then, we have for every $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\mathbb{E}[(\hat{\lambda}_H^{\pi_\theta}(s,a) - \langle \omega_*, \phi(s,a)\rangle)^2|s,a] = \mathbb{E}[(\hat{\lambda}_H^{\pi_\theta}(s,a))^2 - 2\hat{\lambda}_H^{\pi_\theta}(s,a)\langle \omega_*, \phi(s,a)\rangle + \langle \omega_*, \phi(s,a)\rangle^2|s,a]\,. \tag{156}$$

We know that $|\hat{\lambda}_H^{\pi_\theta}(s,a)| \leq \frac{1}{1-\gamma}$. It remains to bound $\|\omega_*\|$ to be able to upper bound the quantity of (156). Recall for this that $\nabla_\omega L_\theta(\omega_*) = 0$, i.e., $\mathbb{E}[(\langle \phi(s,a), \omega_*\rangle - \lambda_H^{\pi_\theta}(s,a))\phi(s,a)] = 0$, which can be rewritten as follows:

$$\mathbb{E}[\lambda_H^{\pi_\theta}(s,a)\phi(s,a)] = \mathbb{E}[\phi(s,a)\phi(s,a)^T]\,\omega_*\,.$$

Therefore, we obtain by invoking Assumption 5.5 that $\omega_* = \mathbb{E}[\phi(s,a)\phi(s,a)^T]^{-1}\mathbb{E}[\lambda_H^{\pi_\theta}(s,a)\phi(s,a)]$ and hence

$$\|\omega_*\| \leq \frac{B}{\mu(1-\gamma)}\,. \tag{157}$$

Using this inequality, it follows from (156) that:

$$\begin{aligned}
\mathbb{E}[(\hat{\lambda}_H^{\pi_\theta}(s,a) - \langle \omega_*, \phi(s,a)\rangle)^2|s,a] &\leq \frac{1}{(1-\gamma)^2} + \frac{2B}{1-\gamma}\|\omega_*\| + B^2\|\omega_*\|^2 \\
&\leq \frac{1}{(1-\gamma)^2}\left(1 + \frac{2B^2}{\mu} + \frac{B^4}{\mu^2}\right) \\
&= \frac{1}{(1-\gamma)^2}\left(1 + \frac{B^2}{\mu}\right)^2\,.
\end{aligned} \tag{158}$$

Hence, the missing part of item (iii) is satisfied with $\sigma = \frac{1}{1-\gamma}(1 + \frac{B^2}{\mu})$.

We conclude the proof by using the result of Theorem G.2 with $\beta' = 2\beta = \frac{1}{4B^2}$ and $\omega_0 = 0$ to obtain after K iterations of the SGD subroutine (see Algorithm 2)

$$\begin{aligned}
\mathbb{E}[L_\theta(\bar{\omega}_K) - L_\theta(\omega_*)] &\leq \frac{4}{K}(\sigma\sqrt{m} + R\|\omega_*\|)^2 \\
&= \frac{4}{(1-\gamma)^2 K}\left(\frac{B^2}{\mu}(1+\sqrt{m}) + \sqrt{m}\right)^2\,,
\end{aligned} \tag{159}$$

where $\bar{\omega}_K$ is the output of Algorithm 2. As a consequence, we have

$$\epsilon_{\text{stat}} \leq \frac{4}{(1-\gamma)^2 K}\left(\frac{B^2}{\mu}(1+\sqrt{m}) + \sqrt{m}\right)^2\,. \tag{160}$$

Plugging this inequality into (152) leads to

$$\begin{aligned}
\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \frac{16(F^* - \mathbb{E}[F(\lambda(\theta_1))]) + \tilde{C}_4}{\alpha T} + \frac{\tilde{C}_3}{N} + 2D_\lambda^2\gamma^{2H} \\
+ \frac{4\tilde{C}_2}{(1-\gamma)^2 K}\left(\frac{B^2}{\mu}(1+\sqrt{m}) + \sqrt{m}\right)^2 + \tilde{C}_2\epsilon_{\text{approx}}\,,
\end{aligned} \tag{161}$$

We set the number of iterations T, the batch size N, the number of iterations K in the subroutine of Algorithm 2 and the horizon H to guarantee that $\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|^2] \leq \mathcal{O}(\epsilon^2) + \mathcal{O}(\epsilon_{\text{approx}})$ where the expectation is taken over both the randomness inherent to the sequence produced by the algorithm together with the uniform sampling defining $\bar{\theta}_T$. Given (161), choosing $T = \mathcal{O}(\epsilon^{-2})$, $N = \mathcal{O}(\epsilon^{-2})$, $K = \mathcal{O}(\epsilon^{-2})$ and $H = \mathcal{O}(\log(\frac{1}{\epsilon}))$ concludes the proof. In particular, the total sample complexity to solve the RL problem with general utilities with occupancy measure approximation in order to achieve an ϵ-approximate stationary point of the objective function (up to the $\mathcal{O}(\sqrt{\epsilon_{\text{approx}}})$ error floor) is given by $T \times (K+N) \times H = \tilde{\mathcal{O}}(\epsilon^{-4})$, where $\tilde{O}$ hides a logarithmic factor in ϵ.

H. Useful technical lemma

In this section, we gather a few technical results that are useful throughout the proofs of our results.

H.1. Smoothness, Lipschitzness and truncation error technical lemmas

The following result from (Zhang et al., 2021b)(Lemma 5.3) ensures in particular that the objective function $\theta \mapsto F(\lambda^{\pi_\theta})$ is smooth which is used to derive an ascent-like lemma in our convergence analysis.

Lemma H.1. *Let Assumptions 4.1 and 4.2 hold. Then, the following statements hold:*

(i) $\forall \theta \in \mathbb{R}^d, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \|\nabla \log \pi_\theta(a|s)\| \leq 2l_\psi$, $\|\nabla^2_\theta \log \pi_\theta(a|s)\| \leq 2(L_\psi + l_\psi^2)$, *and* $\|\nabla_\theta F(\lambda(\theta))\| \leq \frac{2l_\psi l_\lambda}{(1-\gamma)^2}$.

(ii) $\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|\lambda^{\pi_{\theta_1}} - \lambda^{\pi_{\theta_2}}\|_1 \leq \frac{2l_\psi}{(1-\gamma)^2}\|\theta_1 - \theta_2\|$ *and* $\|\lambda_H(\theta_1) - \lambda_H(\theta_2)\|_1 \leq \frac{2l_\psi}{(1-\gamma)^2}\|\theta_1 - \theta_2\|$.

(iii) The objective function $\theta \mapsto F(\lambda^{\pi_\theta})$ *is* L_θ*-smooth with* $L_\theta = \frac{4L_{\lambda,\infty} l_\psi^2}{(1-\gamma)^4} + \frac{8l_\psi^2 l_\lambda}{(1-\gamma)^3} + \frac{2l_\lambda(L_\psi + l_\psi^2)}{(1-\gamma)^2}$.

Proof. See Lemma 5.3 in (Zhang et al., 2021b). The second part of item (ii) was not reported in the aforementioned reference but the proof follows the same lines upon replacing the infinite horizon but the finite one H for the truncated state-action occupancy measures. □

The next lemma controls the truncation error due to truncating simulated trajectories to the horizon H in our infinite horizon setting. Notably, this error vanishes geometrically fast with the horizon H .

Lemma H.2. *Let Assumptions 4.1 and 4.2 be satisfied. Then, we have for any* $H \geq 1$ *and every* $\theta \in \mathbb{R}^d$*:*

(i) $\|\nabla_\theta F(\lambda_H(\theta)) - \nabla_\theta F(\lambda(\theta))\| \leq D_\lambda \gamma^H$ *where* $D_\lambda^2 = \frac{8l_\psi^2 L_\lambda^2}{(1-\gamma)^6} + 16 l_\psi^2 l_\lambda^2 \left(\frac{(H+1)^2}{(1-\gamma)^2} + \frac{1}{(1-\gamma)^4}\right)$.

(ii) $\|\nabla J_H(\theta) - \nabla J(\theta)\| \leq D_g \gamma^H$ *with* $D_g \overset{def}{=} \frac{2l_\psi \|r\|_\infty}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + H}$ *and* r *is the fixed reward function in the cumulative reward setting* .

Proof. See Lemma E.3 in (Zhang et al., 2021b) for the first item. The second item is standard and follows directly from using the policy gradient expression together with Lemma H.1-(i). □

The following result whose proof follows immediately from (7) and Assumption 4.1 (see Lemma E.2, (Zhang et al., 2021b)) establishes the Lipschitz continuity of the policy gradient estimator w.r.t. the policy parameter and the reward variable.

Lemma H.3. *Let Assumption 4.1 hold true and let* $\tau = \{s_0, a_0, \cdots, s_{H-1}, a_{H-1}\}$ *be an arbitrary trajectory of length* H*. Then the following statements hold:*

(i) $\forall \theta \in \mathbb{R}^d, \forall r_1, r_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \|g(\tau,\theta,r_1) - g(\tau,\theta,r_2)\| \leq \frac{2l_\psi}{(1-\gamma)^2}\|r_1 - r_2\|_\infty$,

(ii) $\forall \theta_1, \theta_2 \in \mathbb{R}^d, \forall r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \|g(\tau,\theta_1,r) - g(\tau,\theta_2,r)\| \leq L_g\|\theta_1 - \theta_2\|$ *where* $L_g \overset{def}{=} \frac{2(l_\psi^2 + L_\psi)\|r\|_\infty}{(1-\gamma)^2}$,

H.2. Technical lemma for solving a recursion

The next lemma is useful for solving recursions appearing in our analysis to derive convergence rates.

Lemma H.4. *Let* τ *be a positive integer and let* $\{r_t\}_{t\geq 1}$ *be a non-negative sequence satisfying for every integer* $t \geq 1$

$$r_t \leq (1-\eta_t) r_{t-1} + \beta_t,$$

where $\{\beta_t\}_{t\geq 1}$ *is a non-negative sequence. Then for* $\eta_t = \frac{2}{t+\tau}$ *we have for every integer* $T \geq 1$

$$r_T \leq \frac{\tau^2 r_0}{(T+\tau)^2} + \frac{\sum_{t=1}^T \beta_t (t+\tau)^2}{(T+\tau)^2}.$$

Proof. Notice that $1-\eta_t = \frac{t+\tau-2}{t+\tau}$. Then for all $t \geq 1$

$$r_t \leq \frac{t+\tau-2}{t+\tau} r_{t-1} + \beta_t.$$

Multiplying both sides by $(t+\tau)^2$, we get

$$\begin{aligned}(t+\tau)^2 r_t &\leq (t+\tau-2)(t+\tau) r_{t-1} + \beta_t (t+\tau)^2 \\ &\leq (t+\tau-1)^2 r_{t-1} + \beta_t (t+\tau)^2.\end{aligned}$$

By summing this inequality from $t=1$ to T, we obtain

$$(T+\tau)^2 r_T \leq \tau^2 r_0 + \sum_{t=1}^{T} \beta_t (t+\tau)^2.$$

□

H.3. Technical lemma for decreasing stepsizes

Lemma H.5. *Let $q \in [0,1]$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$ for every integer t. Then for every integer t and any integer $T \geq 1$ we have*

$$\eta_t (1-\eta_{t+1}) \leq \eta_{t+1}, \tag{162}$$

$$\prod_{t=0}^{T-1} (1-\eta_{t+1}) \leq \eta_T \,.$$

Proof. For every integer t we have

$$1-\eta_{t+1} = 1 - \left(\frac{2}{t+3}\right)^q \leq 1 - \frac{1}{t+3} = \frac{t+2}{t+3} \leq \frac{\eta_{t+1}}{\eta_t}.$$

Using the above result, we can write

$$\prod_{t=0}^{T-1} (1-\eta_{t+1}) \leq \prod_{t=0}^{T-1} \frac{\eta_{t+1}}{\eta_t} = \frac{\eta_T}{\eta_0} = \eta_T.$$

□

Lemma H.6. *Let $q \in [0,1)$, $p \geq 0$, $\beta_0 > 0$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$, $\beta_t = \beta_0 \left(\frac{2}{t+2}\right)^p$ for every integer t. Then for any integers t and $T \geq 1$, it holds*

$$\sum_{t=0}^{T-1} \beta_t \prod_{\tau=t+1}^{T-1} (1-\eta_\tau) \leq C \beta_T \eta_T^{-1},$$

where $C > 1$ is an absolute constant depending on p and q.

Proof. See, for instance, (Gadat et al., 2018, Proposition B.1) or (Fatkhullin et al., 2023, Lemma 15). □