On the Global Optimality of Policy Gradient Methods in General Utility Reinforcement Learning

Anas Barakat[1], **Souradip Chakraborty**[2], **Peihong Yu**[2], **Pratap Tokekar**[2], **Amrit Singh Bedi**[3]

[1] Singapore University of Technology and Design,
[2] University of Maryland-College Park, [3] University of Central Florida

Abstract

Reinforcement learning with general utilities (RLGU) offers a unifying framework to capture several problems beyond standard expected returns, including imitation learning, pure exploration, and safe RL. Despite recent fundamental advances in the theoretical analysis of policy gradient (PG) methods for standard RL and recent efforts in RLGU, the understanding of these PG algorithms and their scope of application in RLGU still remain limited. In this work, we establish global optimality guarantees of PG methods for RLGU in which the objective is a general concave utility function of the state-action occupancy measure. In the tabular setting, we provide global optimality results using a new proof technique building on recent theoretical developments on the convergence of PG methods for standard RL using gradient domination. Our proof technique opens avenues for analyzing policy parameterizations beyond the direct policy parameterization for RLGU. In addition, we provide global optimality results for large state-action space settings beyond prior work which has mostly focused on the tabular setting. In this large scale setting, we adapt PG methods by approximating occupancy measures within a function approximation class using maximum likelihood estimation. Our sample complexity only scales with the dimension induced by our approximation class instead of the size of the state-action space.

1 Introduction

Reinforcement learning with general utilities (RLGU) has emerged as a general framework to unify a range of RL applications where the objective of the RL agent cannot be simply cast as a standard expected cumulative reward (Zhang et al., 2020). For instance, in imitation learning, the objective is to learn a policy minimizing the divergence between the induced state-action occupancy measure and expert demonstrations (Ho and Ermon, 2016). In pure exploration, the goal is to learn a policy to explore the state space in a reward-free setting by maximizing the entropy of the state occupancy measure induced by the agent's policy (Hazan et al., 2019). Other examples include risk-averse and constrained RL (Garcıa and Fernández, 2015), diverse skills discovery (Eysenbach et al., 2019), and experiment design (Mutny et al., 2023).

It is well known that the standard RL objective can be written as a linear functional of the occupancy measure. To capture all the aforementioned applications, the RLGU objective is a possibly nonlinear functional of the state action occupancy measure induced by the policy (Zhang et al., 2020). Due to non-linearity, policy gradient algorithms for solving RLGU problems face the major bottleneck of occupancy measure estimation. Prior works (Hazan et al., 2019; Zhang et al., 2020) have mostly focused on the tabular setting where the state-action occupancy measure needs to be estimated for *each* state-action pair using Monte Carlo estimation via sampling trajectories. However, this setting is restrictive for larger state and action spaces where tabular methods become intractable due to

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-2661

the curse of dimensionality. This scalability issue stands as an important challenge to overcome to establish RLGU as a general unified framework for which efficient algorithms exist to solve its larger state-action space instances.

The understanding of PG methods and their scope of application still remains limited despite recent fundamental advances in the theoretical understanding of PG methods for standard RL and recent efforts in RLGU. Specifically:

- In the tabular setting, existing global optimality results rely on hidden convexity which consists in seeing the problem as a convex problem in the occupancy measure. This approach leaves unclear and open the question of the connection of existing analysis with the recent advances in the analysis of PG methods for standard expected return RL as highlighted as future work in Zhang et al. (2020).
- Most existing results focus on the tabular setting. Beyond this restrictive setting, few recent results propose to approximate the occupancy measure using function approximation and either Mean Square Estimation (MSE) (Barakat et al., 2023) or Maximum Likelihood Estimation (MLE) (Huang and Jiang, 2024). However these works only establish first-order stationarity guarantees and hence fall short of providing global convergence guarantees for RLGU beyond standard RL. Moreover, they have several other limitations that render them either inefficient in addressing large state-action space settings due to a dependence on the state action space or suboptimal even for standard expected return objective function. We refer the reader to our related work section below and section 4.4 for a more detailed discussion.

Main contributions. In this work, we investigate the question of global optimality of PG methods for RLGU. Our contributions are summarized as follows:

- In the tabular setting, we establish a new structural property of the RLGU objective in the form of a gradient domination inequality (cf. Sec. 3). This result generalizes existing results for standard expected return objectives in RL and enables global optimality guarantees for PG methods in RLGU within the tabular setting.
- We address the scalability challenge by proposing a simple algorithm for the general and flexible RLGU framework with global optimality guarantees. In this algorithm, an actor performs policy parameter updates whereas a critic approximates the state-action occupancy measure via maximum likelihood estimation (MLE) within a function approximation class (cf. Sec. 4). We analyze the sample complexity of our algorithm under suitable assumptions. Our analysis relies on a total variation performance bound for occupancy measure approximation via MLE which scales with the dimension of the parameters of the function approximation class rather than the state-action space size. Using this result, we establish first-order stationarity and global optimality guarantees for our algorithm for nonconcave and concave utilities respectively (cf. Sec. 4.3).

Related Works. The general framework of RLGU, also known as *convex RL*, has been recently introduced in the literature (Hazan et al., 2019; Cheung, 2019; Zhang et al., 2021; Zahavy et al., 2021; Geist et al., 2022; Bai et al., 2022). Hazan et al. (2019) initially focused on the particular instance of maximum entropy exploration problem and Zhang et al. (2020) proposed a variational policy gradient method to solve the RLGU problem. Zhang et al. (2021) then introduced a simpler (variance-reduced) policy gradient method to solve the (possibly nonconcave) RLGU problem using a simpler policy gradient theorem (see also Kumar et al. (2022)). Later, Barakat et al. (2023) proposed an even simpler single-loop normalized policy gradient algorithm to solve RLGU. Zahavy et al. (2021) leveraged Fenchel duality to cast the convex RL problem into a saddle-point problem that can be solved using standard RL algorithms. In a line of works, Mutti et al. (2022b,a, 2023) formulated the convex RL problem in finite trials instead of infinite realizations and considered an objective which is any convex function of the empirical state distribution computed from a finite number of realizations. Ying et al. (2023a) introduced policy-based primal-dual methods for solving convex constrained CMDPs and Ying et al. (2023b) further addressed a multi-agent RL problem with general utilities. All the aforementioned works focus on the tabular setting. In particular, most of these works use a count-based Monte Carlo estimate of the occupancy measure that cannot scale to large state-action spaces. More recently, Huang et al. (2023) provided sample-efficient online/offline RL algorithms with density features in low-rank MDPs for occupancy estimation. Only few recent works (Barakat et al., 2023; Huang and Jiang, 2024) propose to go beyond the tabular setting, we discuss them in more details in section 4.4. See also appendix B for a more detailed related work discussion.

Notation. 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 for vectors and the spectral norm for matrices. We interchangeably denote functions $f : \mathcal{X} \to \mathbb{R}$ over a finite set $\mathcal{X}$ as vectors $f \in \mathbb{R}^{|\mathcal{X}|}$ with components $f(x)$ with a slight abuse of notations.

2 Problem Formulation

MDP with General Utility. Consider a discrete-time discounted Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}, F, \rho, \gamma)$, 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, $F : \Lambda \to \mathbb{R}$ is a general utility function defined over the space Λ of probability measures on the product state-action space $\mathcal{X} := \mathcal{S} \times \mathcal{A}$, 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 then environment transitions to a state $s_{t+1} \in \mathcal{S}$ 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 $\mathbb{E}$) for the associated expectation. We define for any policy $\pi \in \Pi$ the (normalized) state and state-action occupancy measures $d^\pi \in \Delta(\mathcal{S}), \lambda^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ respectively[1]:

$$d^\pi(s) := (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\rho,\pi}(s_t = s)\,, \quad \lambda^\pi(s,a) := d^\pi(s)\,\pi(a|s)\,. \tag{1}$$

The general utility function F assigns a real to each occupancy measure λ^π induced by a policy $\pi \in \Pi$. We note that λ^π will also be seen as a vector of the Euclidean space $\mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$. In the rest of this work, we will consider a class of policies parametrized by a vector $\theta \in \mathbb{R}^d$ for some fixed integer $d \in \mathbb{N}$. We shall denote by $\pi_\theta \in \Pi$ such a policy in this class.

Policy optimization. The goal of the RL agent is to find a policy π_θ solving the problem:

$$\max_{\theta \in \mathbb{R}^d} F(\lambda^{\pi_\theta})\,, \tag{2}$$

where λ^{π_θ} is defined in (1), F is a smooth function supposed to be upper bounded and $F^\star$ is used to denote the maximum in (2). The agent has access to trajectories of finite length H generated from the MDP under the initial distribution ρ and the policy π_θ. 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 when the state transition kernel $\mathcal{P}$ is unknown. This general utility problem was described, for instance, in Zhang et al. (2021) (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 $r \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, in which case we recover the expected return function as an objective:

$$V^{\pi_\theta}(r) := \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t)\right]\,. \tag{3}$$

Remark 1. *We prefer the terminology of 'RL with general utilities' to 'convex RL' since the objective may even be nonconvex in the occupancy measure in full generality. Although our focus in this work is on concave utilities, we provide first-order stationarity theoretical guarantees for the nonconcave case. While the convex RL literature exclusively focuses on the case of concave utilities, a lot of applications of interest do not fall under this umbrella and inherently involve nonconcave utilities. We provide several such examples in Appendix D.*

3 Global Optimality in the Tabular Setting

In this section, our main goal is to establish a structural property of the RLGU objective as a function of its policy parameters depending on the policy parameterization. Recall that even when

[1] We mostly drop the dependence on the initial distribution ρ in the notation throughout the paper, except for the statement and proof of Theorem 1.

https://doi.org/10.52202/085713-2661

the functional F is concave in the occupancy measure, the RLGU objective is in general nonconcave in the policy parameters. Recent works in the last few years (Agarwal et al., 2021; Bhandari and Russo, 2024; Mei et al., 2020) have shown that the expected return in standard RL, which is also nonconcave in the policy parameters, satisfies a gradient domination inequality. This interesting property implies that any stationary point of the objective is actually a globally optimal point. In this section we extend this result to RLGU and show that the RLGU objective function also satisfies a similar, but different, gradient domination inequality when the functional F is concave. Our results open an avenue for going beyond expected return objectives by exploiting the underlying dynamic programming structure of occupancy measures which is key to our results.

In view of our analysis, we recall how to 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:

$$[\nabla_\theta \lambda(\theta)]^T r = \nabla_\theta V^{\pi_\theta}(r) = \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) \sum_{t'=0}^{t} \nabla \log \pi_\theta(a_{t'}|s_{t'})\right], \tag{4}$$

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

$$\nabla_\theta F(\lambda(\theta)) = [\nabla_\theta \lambda(\theta)]^T \nabla_\lambda F(\lambda(\theta)) = \nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\lambda(\theta))} . \tag{5}$$

The classical policy gradient in the standard RL setting uses rewards which are obtained via interaction with the environment. In RLGU, there is no reward function but rather a *pseudoreward* $\nabla_\lambda F(\lambda(\theta))$ depending on the unknown occupancy measure induced by the policy. We assume that the gradient of F w.r.t. its variable λ is a known function to the agent. This is the case in most prior works, e.g. when F is the negative entropy function in pure exploration (Hazan et al., 2019), or a KL divergence in imitation learning (Ho and Ermon, 2016), a penalized objective in constrained RL, or other objectives in experiment design (Mutny et al., 2023). The policy gradient identity (5) shows that the gradient of the RLGU objective with respect to its policy parameters coincides with the standard expected return policy gradient evaluated at the reward function $\nabla_\lambda F(\lambda(\theta))$. This observation is essential for our development and we believe it could be of independent interest in settings involving varying rewards as a function of policy parameters. Using this key insight, the next result shows that the RLGU objective satisfies a gradient domination inequality when the policy parametrization is tabular, extending the structural result for standard expected return (Agarwal et al., 2021; Bhandari and Russo, 2021; Xiao, 2022).

Assumption 1 (Concavity)**.** *The utility function $F : \Lambda \to \mathbb{R}$ is concave.*

Theorem 1. ***(RLGU objective gradient domination)*** *Let Assumption 1 hold. Consider a direct policy parametrization ($\pi_\theta(a|s) = \theta_{s,a}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$). Then for every $\theta \in \mathbb{R}^d$,*

$$F(\lambda(\theta^*)) - F(\lambda(\theta)) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star(\nabla_\lambda F(\lambda(\theta)))}}{\mu} \right\|_\infty \max_{\bar{\pi} \in \Pi} \langle \bar{\pi} - \pi_\theta, \nabla_\theta F(\lambda(\theta)) \rangle , \tag{6}$$

where $d = |\mathcal{S}| \cdot |\mathcal{A}|$, $\pi^\star(r) \in \arg\max_{\pi \in \Pi} V^\pi(r)$ for any $r \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, θ^* is an optimal policy parameter and μ is any state distribution s.t. $\mu(s) > 0$ for all states $s \in \mathcal{S}$.

A few comments are in order regarding this result:

- The gradient domination inequality depends on a distribution mismatch coefficient which itself depends on the pseudo-reward function $\nabla_\lambda F(\lambda(\theta))$. Theorem 1 recovers the standard gradient domination result in linear RL, as in Lemma 4.1 of Agarwal et al. (2021), when the pseudo-reward is constant and equal to the true reward.
- The mismatch coefficient is finite for any θ, provided the reference distribution μ has full support.
- If one seeks a uniform upper bound over all θ, the coefficient can be large. However, this worst-case bound is still finite, and can be upper-bounded by the state space size if μ is uniform. It can also be upper bound by its maximum over the reward functions which are bounded by the range of pseudo-rewards (see e.g. Assumption 3).

Remark 2. *Recall that we assume throughout the paper that the state and action spaces are finite as mentioned in section 2. In particular, Theorem 1 holds under this setting. While we mention the applicability of our algorithm to continuous state spaces in practice (in remark 3 below), the current analysis does not readily extend to that setting. Ensuring boundedness of the distribution mismatch coefficient can be challenging beyond our finite setting, as mentioned in e.g. Koren et al. (2025) showing that policy gradient methods can then converge to locally (in contrast to globally) optimal policies (in classical linear RL).*

Proof. We introduce a few useful notations for the state action value functions induced by any fixed reward functions $r \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ and any $\theta \in \mathbb{R}^d$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$Q^{\pi_\theta}_{s,a}(r) := \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t,a_t) | s_0 = s, a_0 = a\right], V^{\pi_\theta}_s(r) := \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q^{\pi_\theta}_{s,a}(r)\,. \tag{7}$$

Proof. First, using the gradient domination result (Agarwal et al., 2021, Lemma 4) for standard expected return and any fixed reward function r, we obtain for every tabular policy π,

$$V^{\pi^\star(r)}(r) - V^{\pi}(r) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star(r)}}{\mu} \right\|_\infty \max_{\bar{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \langle \bar{\pi} - \pi, \nabla_\pi V^\pi(r) \rangle\,. \tag{8}$$

Plugging in $r = r_\theta := \nabla_\lambda F(\lambda(\theta))$ and using (5), we get

$$V^{\pi^\star(r_\theta)}(r_\theta) - V^{\pi}(r_\theta) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star(r_\theta)}}{\mu} \right\|_\infty \max_{\bar{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \langle \bar{\pi} - \pi, \nabla_\theta F(\lambda(\theta)) \rangle\,. \tag{9}$$

As $\pi^\star(r) \in \arg\max_\pi V^\pi(r)$, we have $V^{\pi^\star(r)}(r) \geq V^\pi(r)$ for every policy π . Using this inequality with $\pi = \pi^\star \in \arg\max_\pi F(\lambda^\pi)$ gives

$$V^{\pi^\star(r_\theta)}(r_\theta) - V^{\pi}(r_\theta) \geq V^{\pi^\star}(r_\theta) - V^{\pi_\theta}(r_\theta) = \langle r_\theta, \lambda^{\pi^\star} - \lambda^{\pi_\theta} \rangle \geq F(\lambda(\theta^\star)) - F(\lambda(\theta))\,, \tag{10}$$

where the identity stems from the definition of a value function using occupancy measures and the last inequality follows from using concavity of F w.r.t. its occupancy measure argument (Assumption 1). Combining (9) and (10) concludes the proof. □

Note that in the last step, we have only used concavity at a given point $\lambda^{\pi^\star}$ which means that only a weaker version of Assumption 1 is needed. This means that we can potentially go beyond standard objectives in RLGU. We leave such investigations for future work.

Gradient domination inequalities can be readily used to show $\mathcal{O}(1/k)$ iteration complexity results for policy gradient methods using similar techniques to Xiao (2022) for instance (see also more recently Kumar et al. (2024) for actor-critic methods). Zhang et al. (2021) previously provided global optimality results for a PG algorithm using hidden convexity. However their technique has several limitations: (a) it does not connect to structural properties of standard expected returns and standard policy gradients. In particular they do not show gradient domination for RLGU objectives as developed above and this is explicitly mentioned in their work; and (b) it requires a restrictive assumption (see (5) below) that is not easily verifiable beyond the tabular policy. We believe that our proof technique can be extended to the case of the softmax policy building on the results of Mei et al. (2020) (Lemma 8 specifically). We leave this question for future work.

4 Global Optimality Beyond the Tabular Setting

In this section, we propose a policy gradient algorithm to solve the policy optimization problem (2) with general utilities for larger state-action spaces. We start by elaborating on the challenges faced to solve such a large-scale problem. Section 4.1 mainly contains known material from the recent literature (Zhang et al., 2021), we report it here separately from the problem formulation in section 2 to motivate our algorithmic design. The rest of the section presents our algorithmic contributions.

https://doi.org/10.52202/085713-2661

4.1 Challenges for Large-scale RLGU

One of the main challenges in solving the general utility problem (2) via a policy gradient algorithm based on (5) is to estimate the unknown state-action occupancy measure $\lambda(\theta)$ in large scale settings involving huge state and action spaces. This problem is arguably more delicate than that of estimating action-value functions in cumulative expected reward RL problems. First, while action-value functions satisfy a *forward* Bellman equation, occupancy measures satisfy a *backward* Bellman flow equation. This fundamental difference makes it hard to design stochastic algorithms minimizing mean-square Bellman errors as it is customary in algorithms using function approximation to solve standard RL problems (see end of appendix B for further explanations). Second and foremost, while prior work has used Monte Carlo estimates for this quantity, such count-based estimates are not tractable beyond small tabular settings. Indeed, for very large state-action spaces, it is not tractable to compute and store a table of count-based estimates of the true occupancy measure containing all the values for all the state-action pairs. In the next section, we propose an approach to tackle this issue.

Stochastic Policy Gradient. In view of performing a stochastic policy gradient algorithm, we would like to estimate the policy gradient $\nabla_\theta F(\lambda(\theta))$ in (5). We can use the standard reinforce estimator suggested by Eq. (4). 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:

$$g(\tau, \theta, r) = \sum_{t=0}^{H-1} \left(\sum_{h=t}^{H-1} \gamma^h r(s_h, a_h) \right) \nabla \log \pi_\theta(a_t|s_t) \,. \tag{11}$$

Given (5), we also need to estimate the state-action occupancy measure $\lambda(\theta)$ (when F is nonlinear)[2]. Prior work has exclusively focused on the tabular setting using a Monte-Carlo estimate of this occupancy measure $\lambda^{\pi_\theta} = \lambda(\theta)$ (see (1)) truncated at the horizon H by $\lambda(\tau) = \sum_{h=0}^{H-1} \gamma^h \delta_{s_h,a_h}$ 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, and $\tau = \{(s_h, a_h)\}_{0\le h\le H-1}$ is a trajectory of length H generated by the MDP controlled by the policy π_θ .

Remark 3. ***(Extension to continuous state-action spaces)*** *Our algorithm can be used in the continuous (compact) state-action space setting since it only relies on using policy gradients and MLE which are both scalable. We stick to the discrete state action space notation for simplicity.*

4.2 Occupancy Measure Estimation

In this section, we address the challenge of occupancy measure estimation in large state action spaces. Given a policy π_θ, our goal is to estimate the unknown occupancy measure d^{π_θ} induced by this policy using state samples obtained from executing the policy. Since the normalized occupancy measure is a probability distribution, we propose to perform maximum likelihood estimation. Before presenting this procedure, we elaborate on the motivation behind approximating the occupancy measure by a parametrized distribution in a given function class of neural networks for example.

Motivation. Besides the practical motivation of using distribution approximation to scale to larger state-action space settings, we provide some theoretical motivation. Recall that action-value functions are linear in the feature map for linear (or low-rank) MDPs for solving standard cumulative sum RL problems (see Proposition 2.3 in Jin et al. (2020)). Similarly, it turns out that state-occupancy measures are linear (or affine in the discounted setting) in density features in low-rank MDPs. We refer the reader to Appendix C for a proof of this statement (see also Lemma 16, 17 in Huang et al. (2023)). Therefore, in this case, it is natural to approximate occupancy measures via linear function approximation using some density features. More generally, for an arbitrary MDP, we propose to approximate the (normalized) state occupancy measure d^{π_θ} induced by a policy π_θ directly by a probability distribution in a certain parametric class of probability distributions:

$$\Lambda := \{ p_\omega \in \Delta(\mathcal{S}) \,|\, \omega \in \Omega \subseteq \mathbb{R}^m \} \,, \tag{12}$$

where for instance $m \ll |\mathcal{S}|$. An example of such a parametrization for a given $\omega \in \mathbb{R}^m$ is the softmax σ_ω defined over the state space by $\sigma_\omega(s) := \exp(\psi_\omega(s))/Z(\omega)$, where $Z(\omega) :=$

[2]In the cumulative reward setting, the utility F is linear w.r.t. λ and $\nabla_\lambda F(\lambda(\theta))$ is independent of $\lambda(\theta)$.

$\sum_{s' \in \mathcal{S}} \exp(\psi_\omega(s'))$ and where $\psi_\omega : \mathcal{S} \to \mathbb{R}$ is a given mapping which can be a neural network in practice. For continuous state spaces, practitioners can consider for instance Gaussian mixture models with means and covariance matrices encoded by trainable neural networks.

Maximum Likelihood Estimation (MLE). For simplicity, we suppose we have access to i.i.d. state samples following the distribution d^{π_θ} throughout our exposition. We refer the reader to Appendix E.1 for a discussion about how to sample such states. Given the parametric distribution class Λ defined in (12) and a data set $\mathcal{D} := \{s_i\}_{i=1,\cdots,n} \in \mathcal{S}^n$ of n i.i.d. state samples following the distribution d^{π_θ} induced by the current policy π_θ, we construct the standard MLE

$$\hat{d}^{\pi_\theta} := p_{\omega^*}\,, \qquad \omega^* \in \arg\max_{\omega \in \Omega} \frac{1}{n} \sum_{i=1}^{n} \log p_\omega(s_i)\,. \tag{13}$$

An estimator of the state-action occupancy measure λ^{π_θ} is then given by $\hat{\lambda}^{\pi_\theta}(s,a) = \hat{d}^{\pi_\theta}(s)\,\pi_\theta(a|s)$ for any $s \in \mathcal{A}, a \in \mathcal{A}$ (see (1)). Using MLE is important for our scalability goal. Barakat et al. (2023) recently proposed a different procedure based on mean square error estimation. Please see appendix B for a detailed comparison with this work highlighting the merits of our approach. In practice, a neural network learns the parameters of a chosen parametrized distribution class for approximating the true occupancy measure by maximizing the log-likelihood loss (13) over the samples generated (see appendix E.1 for sampling).

Proposed Algorithm. Based on sections 4.1 and 4.2, we propose a stochastic policy gradient algorithm which consists of two main steps: *(i)* Compute an approximation of the unknown state-action occupancy measure $\lambda^{\pi_\theta} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ for a fixed parameter $\theta \in \mathbb{R}^d$ with MLE using collected state samples (see (13)); *(ii)* Perform stochastic policy gradient ascent using the stochastic policy gradient defined in (11) using the estimated occupancy measure computed in step (i). The resulting algorithm is Algorithm 1 which is model-free as we do not estimate the transition kernel.

Algorithm 1 PG for RLGU with Occupancy Measure Approximation (PG-OMA)

Input: $\theta_0 \in \mathbb{R}^d, T, N \geq 1, \alpha > 0, H$.
for $t = 0, \ldots, T-1$ **do**
  `//Occupancy approximation for pseudo-reward learning`
  Compute the MLE estimator $\hat{\lambda}_t = \hat{d}^{\pi_{\theta_t}} \cdot \pi_{\theta_t}$ using policy π_{θ_t} (see (13)).
  $\hat{r}_t = \nabla_\lambda F(\hat{\lambda}_t)$
  `//Policy parameter update`
  Sample a batch of N independent trajectories $(\tau_t^{(i)})_{1\leq i\leq N}$ of length H using π_{θ_t} .
  $\theta_{t+1} = \theta_t + \frac{\alpha}{N} \sum_{i=1}^{N} g(\tau_t^{(i)}, \theta_t, \hat{r}_t)$ (see (11))
end for
Return: θ_T

Remark 4. *When running Algorithm 1, note that the vector $\hat{\lambda}_t \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ (and hence the vector r_t) is not computed for all state-action pairs. 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 (11).*

Our occupancy measure estimation step can be seen as a critic for pseudo-reward learning. Notice though that this critic is not approximating a value function like in standard RL but rather the occupancy measure which is a distribution.

4.3 Global Optimality for Policy Gradient with Occupancy Measure Approximation

Statistical Complexity of Occupancy Estimation. In this section, we suppose we are given a data set of i.i.d. state-action pair samples following the (normalized) occupancy measure λ^π induced by a fixed given policy π . As previously explained, we approximate λ^π by a function (or parametrized density) in the function class Λ defined in (12). We make the following assumption to control the complexity of our function approximation class.

https://doi.org/10.52202/085713-2661

Assumption 2 (Function approximation class regularity)**.** *The following holds true:*

(i) (parameter compactness) The set Ω is compact, we denote by $B_\omega := \max_{\omega\in\Omega} \|\omega\|_\infty$;

(ii) (realizability) The (normalized) occupancy measure to be estimated satisfies: $\lambda^\pi \in \Lambda$;

(iii) (Lipschitzness) $\forall \omega, \bar{\omega} \in \Omega, \forall x \in \mathcal{X}, \exists L(x) \in \mathbb{R}$ s.t. $|p_\omega(x) - p_{\bar{\omega}}(x)| \leq L(x)\|\bar{\omega} - \omega\|_\infty$ with $B_L := \int_{\mathcal{X}} L(x)dx < +\infty$.

Assumption 2 is satisfied for instance for the class of generalized linear models, i.e. $\Lambda := \{p_\omega(x) = g(\omega^T \phi(x)), \forall x \in \mathcal{X} : p_\omega \in \Delta(\mathcal{X}), \omega \in \Omega\}$ where $g : \mathbb{R} \to [0, 1]$ is an increasing Lipschitz continuous function and $\phi : \mathcal{X} \to \mathbb{R}^d$ is a given feature map s.t. $\int \|\phi(x)\|_1 dx \leq B_L$ for some $B_L > 0$. Notice that features can be normalized appropriately to satisfy the assumption. A similar assumption has been made in the case of linear MDPs in Huang et al. (2023) (Assumption 1). The realizability assumption holds in the case of low-rank MDPs since state occupancy measures are linear in density features in low-rank MDPs (see discussion in section 4.2 and Appendix C). This realizability assumption can be relaxed at the price of incurring an error due to function approximation that cannot vanish if the true occupancy measures do not belong to our function approximation class.

We now state our sample complexity result for occupancy measure estimation via MLE in view of our PG sample complexity analysis. This result relies on arguments developed in the statistics literature Van de Geer (2000); Zhang (2006). These techniques were adapted to the RL setting for low-rank MDPs in e.g. Agarwal et al. (2020). Our proof builds on Huang et al. (2023) which we slightly adapt for our purpose (see Appendix E.2).

Proposition 1. *Let Assumption 2 hold true. Then for any $\delta > 0$, the MLE $\hat{\lambda}^{\pi_\theta}$ defined using (13) satisfies with probability at least $1 - \delta$, $\|\hat{\lambda}^{\pi_\theta} - \lambda^{\pi_\theta}\|_1 \leq 6\sqrt{\frac{12\, m \log\left(\frac{2\lceil B_\omega B_L n\rceil}{\delta}\right)}{n}}$.*

The above result translates into a sample complexity of $\tilde{\mathcal{O}}(m\,\varepsilon^{-2})$ to guarantee an ε-approximation of the true occupancy measure (in the l_1-norm distance) using samples. We highlight that our sample complexity only depends on the dimension m of the parameter space and does not scale with the size of the state-action space. Hence the MLE procedure we use is the key ingredient to scale our algorithm to large state-action spaces. To the best of our knowledge, existing algorithms for solving the RLGU problem (with nonlinear utility) are limited to the restrictive tabular setting.

Global convergence guarantees. Now we establish sample complexity guarantees for Algorithm 1. We start by introducing the assumptions required for our results and discuss their relevance.

Assumption 3 (Policy parametrization)**.** *The following holds for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. For every $\theta \in \mathbb{R}^d$, $\pi_\theta(a|s) > 0$. Moreover, the function $\theta \mapsto \pi_\theta(a|s)$ is continuously differentiable and the score function $\theta \mapsto \nabla \log \pi_\theta(a|s)$ is bounded by some positive constant B.*

This standard assumption is satisfied for instance by the common softmax policy parametrization defined for every $\theta \in \mathbb{R}^d$, $(s, a) \in \mathcal{S}\times\mathcal{A}$ by $\pi_\theta(a|s) = \frac{\exp(\psi(s,a;\theta))}{\sum_{a'\in\mathcal{A}} \exp(\psi(s,a';\theta))}$, where $\psi : \mathcal{S}\times\mathcal{A}\times\mathbb{R}^d \to \mathbb{R}$ is a smooth function such that the map $\psi(s, a; \cdot)$ is twice continuously differentiable for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ and for which 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) $\max_{s\in\mathcal{S},a\in\mathcal{A}} \sup_\theta \|\nabla^2\psi(s, a; \theta)\| \leq L_\psi$.

We now make a smoothness assumption on the utility function which is standard in the RLGU literature (Hazan et al., 2019; Zhang et al., 2020, 2021; Barakat et al., 2023; Ying et al., 2023a). This assumption captures most of the problems of interest in RLGU including pure exploration (using the smoothed entropy), learning from demonstrations (using the smoothed KL) as well as standard linear RL and CMDPs. Other entropic measures or l^2 (quadratic) losses are also possible. For instance, the smoothed entropy defined as $H_\sigma(x) = -x \log(x + \sigma)$ (for $\sigma > 0$) is $1/(2\sigma)$-smooth w.r.t. the infinity norm and has been used in RLGU, see e.g. Lemma 4.3 in Hazan et al. (2019).

Assumption 4 (General utility smoothness)**.** *There exist constants $l_\lambda, L_\lambda > 0$ s.t. for all $\lambda_1, \lambda_2 \in \Lambda$, $\|\nabla_\lambda F(\lambda_1)\|_2 \leq l_\lambda$ and $\|\nabla_\lambda F(\lambda_1) - \nabla_\lambda F(\lambda_2)\|_2 \leq L_\lambda \|\lambda_1 - \lambda_2\|_2$.*

Under Assumptions 3 and 4, the function $\theta \mapsto F(\lambda^{\pi_\theta})$ is L_θ-smooth (see Lemma 3 for the expression). Using this property, the next result shows that our algorithm enjoys a first-order stationary guarantee in terms of the non-convex general utility objective.

Theorem 2. ***(Nonconcave general utility)*** *Let Assumptions 3, 4 hold. Then the iterates generated by Algorithm 1 with step sizes $\alpha_t \leq 1/(2L_\theta)$ and $T \geq 1$ iterations satisfy:*

$$\mathbb{E}[\|\nabla_\theta F(\lambda^{\pi_{\theta_\tau}})\|^2] \leq \frac{16(F^\star - \mathbb{E}[F(\lambda^{\pi_{\theta_1}})])}{\alpha T} + \frac{C_1}{N} + C_2 \mathbb{E}[\|\hat{\lambda}_\tau - \lambda^{\pi_{\theta_\tau}}\|_2^2], \tag{14}$$

where τ is a uniform random variable over $\{1, \cdots, T\}$ and expectation is w.r.t. all randomness (in (θ_t) and τ).

The above upper bound shows a decomposition of the first order stationarity error into three terms: the first two are the typical errors incurred by PG methods whereas the third one is due to occupancy measure approximation. In particular, choosing the number of iterations T, the batch size N (of sampled trajectories) appropriately and the number n of samples used in MLE for occupancy measure approximation, we obtain the following sample complexity result.

Corollary 1. *Let Assumptions 2, 3, 4 hold. Setting the number of iterations to $T = \mathcal{O}(\epsilon^{-1})$, the batch size for PG to $N = \mathcal{O}(\epsilon^{-1})$, the horizon to $H = \mathcal{O}(\frac{1}{1-\gamma}\log(\frac{1}{\epsilon}))$ and the number of samples for occupancy measure MLE to $n = \mathcal{O}(m\epsilon^{-1})$ for some precision $\epsilon > 0$ in Theorem 2, it holds that $\mathbb{E}[\|\nabla_\theta F(\lambda^{\pi_{\theta_\tau}})\|^2] \leq \epsilon$. The total sample complexity is $T(N+n)H = \tilde{\mathcal{O}}(m\epsilon^{-2})$.*[a]

[a] The notation $\mathcal{O}(\cdot)$ hides polynomial and logarithmic dependence on problem parameters independent of the desired accuracy ϵ and the dimension m, $\tilde{\mathcal{O}}(\cdot)$ hides in addition logarithmic dependence on ϵ.

In several applications in RLGU, the utility function F is concave w.r.t. its occupancy measure variable. We now turn to proving global performance bounds in this setting.

Notice that the general utility objective is in general nonconcave w.r.t. the policy parameter θ. Despite this non-concavity, we can exploit the so-called *hidden convexity* (concavity in our setting) of the problem Zhang et al. (2021). We require an additional regularity assumption on the policy parametrization which has been previously made in Zhang et al. (2021); Ying et al. (2023a); Barakat et al. (2023). This is a local assumption relating parameterized policies and their corresponding occupancy measures, maintaining the hidden convexity structure. While this assumption holds for a tabular policy parametrization, it is delicate to relax it further, see e.g. Appendix C in Barakat et al. (2023) for a discussion.

Assumption 5 (Policy overparametrization)**.** *For the softmax policy defined above, 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 $\bar{\eta} > 0$ s.t. for every positive real $\eta \leq \bar{\eta}$, $(1-\eta)\lambda(\theta) + \eta\lambda(\theta^*) \in \mathcal{V}_{\lambda(\theta)}$ where π_{θ^*} is an optimal policy.*

The following result makes use of the concavity of the utility function F to obtain a global optimality guarantee for the iterates of our algorithm under the assumption that the occupancy measures induced by the policies encountered during the run of the algorithm are uniformly well-approximated.

Theorem 3. ***(Concave general utility)*** *Let Assumptions 3 to 5 hold. Assume further that there exists $\epsilon_{MLE} > 0$ s.t. $\mathbb{E}[\|\hat{\lambda}_t - \lambda(\theta_t)\|_2^2] \leq \epsilon_{MLE}$ uniformly over $T \geq 1$ iterations of Algorithm 1 with step sizes $\alpha_t \leq 1/(2L_\theta)$. Then the iterate output θ_T of Algorithm 1 satisfies for any $\eta < \bar{\eta}$,*

$$\mathbb{E}[F^\star - F(\lambda(\theta_T))] \leq (1-\eta)^T \delta_0 + C_3 \frac{\eta}{\alpha} + C_4 \frac{\alpha}{\eta}\left(\frac{1}{N} + \epsilon_{MLE}\right), \tag{15}$$

for some positive constants C_3, C_4 explicit in Appendix E.4, (53) and $\delta_0 := \mathbb{E}[F^\star - F(\lambda(\theta_0))]$.

The above bound shows how the global optimality function value gap depends on the estimation error ϵ_{MLE} of occupancy measures. In the next result, we use Proposition 1 to reduce the estimation

https://doi.org/10.52202/085713-2661

error. Since occupancy measures are supposed to be realizable, we can approximate them arbitrarily well using enough samples (see Proposition 1). Indeed by picking the number of samples $n = \mathcal{O}(m/\epsilon^2)$, the error ϵ_{MLE} is smaller than the desired function value gap accuracy ϵ.

We obtain the following sample complexity by specifying the step size and number of iterations of our algorithm as well as large enough batch size and number of samples for MLE using Proposition 1.

Corollary 2. *Let Assumptions 2 to 5 hold. For any given precision $\epsilon > 0$, set $T = \frac{1}{\eta}\log(\frac{\delta_0}{\epsilon}), \alpha = \mathcal{O}(\epsilon), \eta = \mathcal{O}(\epsilon^2), N = \mathcal{O}(\epsilon^{-2})$, $H = \mathcal{O}(\frac{1}{1-\gamma}\log(\frac{1}{\epsilon}))$ and $n = \mathcal{O}(m\epsilon^{-2})$, then the total sample complexity to obtain $\mathbb{E}[F^\star - F(\lambda(\theta_t))] \leq \epsilon$ is $T(N+n)H = \tilde{\mathcal{O}}(m\epsilon^{-4})$.*

4.4 Novelty and Comparison to Prior Work

In this section, we discuss the two most relevant works (Barakat et al., 2023; Huang and Jiang, 2024).

Comparison to Huang and Jiang (2024). They focus primarily on the finite-horizon, expected return setting, with a brief extension to general utilities leaving global optimality for future work. In contrast, our work directly targets infinite-horizon discounted RLGU. Furthermore, we establish a *last-iterate global* convergence guarantee with a rate for RLGU, improving upon their best-iterate rate for expected returns. Their analysis does not immediately extend to RLGU. Finally, although we both use MLE occupancy measure estimation, their method requires estimating both occupancy and log-gradient occupancy via recursive regression, introducing additional estimation error. Focusing on the online setting, our method only requires MLE occupancy estimation.

Comparison to Barakat et al. (2023, section 5). Before commenting on the limitations in the MSE approach of Barakat et al. (2023) for scaling to large spaces, we highlight first two preliminary points: (a) Global convergence. In contrast to our work which establishes global convergence guarantees (see Theorem 3, Corollary 2), Barakat et al. (2023) only provide a first-order stationarity guarantee; (b) Technical analysis. Our occupancy measure MLE estimation combined with our PG algorithm requires a different analysis even for our first-order stationarity guarantee. We exploit hidden convexity (similarly to Zhang et al. (2021)) to obtain global optimality and we isolate and propagate errors induced by occupancy measure approximation in the PG method. This leads to a function value gap recursion with errors satisfied by the optimality function value gap. See appendix E. Furthermore, Barakat et al. (2023) propose to approximate the occupancy measure using a specific mean square error loss estimation procedure whereas we use an MLE procedure. This difference is important given the main goal and motivation of scaling to larger state action spaces. We argue that *their* MSE formulation has important limitations for occupancy measure approximation in terms of scalability and other fundamental aspects (see appendix B for a detailed discussion).

5 Conclusion

In this paper, we have investigated the question of global optimality of PG algorithms for RLGU beyond the standard expected return RL setting in both the tabular and large state action space settings. Promising future directions include extensions to more general policy parameterizations, continuous state–action spaces, and average reward settings building on recent analysis of policy gradient methods (Kumar et al., 2025). We hope this work will stimulate further research in view of designing efficient and scalable algorithms for solving real-world problems.

Acknowledgments and Disclosure of Funding

We thank anonymous reviewers for their useful comments.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*.

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

Bai, Q., Agarwal, M., and Aggarwal, V. (2022). Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm. *Journal of Artificial Intelligence Research*.

Barakat, A., Fatkhullin, I., and He, N. (2023). Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*.

Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*.

Bhandari, J. and Russo, D. (2024). Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927.

Cheung, W. C. (2019). Regret minimization for reinforcement learning with vectorial feedback and complex objectives. *Advances in Neural Information Processing Systems*, 32.

De Santi, R., Prajapat, M., and Krause, A. (2024). Global reinforcement learning : Beyond linear and convex rewards via submodular semi-gradient methods. In *International Conference on Machine Learning*.

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

Fatkhullin, I., He, N., and Hu, Y. (2023). Stochastic optimization under hidden convexity. *arXiv preprint arXiv:2401.00108*.

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

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*.

Hallak, A. and Mannor, S. (2017). Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*.

Huang, A., Chen, J., and Jiang, N. (2023). Reinforcement learning in low-rank MDPs with density features. In *International Conference on Machine Learning*.

Huang, A. and Jiang, N. (2024). Occupancy-based policy gradient: Estimation, convergence, and optimality. In *Advances in Neural Information Processing Systems*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

Koren, U., Kumar, N., Gadot, U., Ramponi, G., Levy, K. Y., and Mannor, S. (2025). Policy gradient with tree search: Avoiding local optimas through lookahead. *arXiv preprint arXiv:2506.07054*.

Kumar, N., Agrawal, P., Ramponi, G., Levy, K. Y., and Mannor, S. (2024). On the convergence of single-timescale actor-critic. *arXiv preprint arXiv:2410.08868*.

Kumar, N., Murthy, Y., Shufaro, I., Levy, K. Y., Srikant, R., and Mannor, S. (2025). Global convergence of policy gradient in average reward MDPs. In *International Conference on Learning Representations*.

 https://doi.org/10.52202/085713-2661

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

Lin, K., Jie, C., and Marcus, S. I. (2018). Probabilistically distorted risk-sensitive infinite-horizon dynamic programming. *Automatica*, 97:1–6.

Lin, K. and Marcus, S. I. (2013). Dynamic programming with non-convex risk-sensitive measures. In *2013 American Control Conference*, pages 6778–6783. IEEE.

Liu, J., Li, W., Lin, D., Wei, K., and Zhang, Z. (2025). On the convergence of projected policy gradient for any constant step sizes. *Journal of Machine Learning Research*, 26(193):1–35.

Liu, J., Li, W., and Wei, K. (2024). Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*.

Mutny, M., Janik, T., and Krause, A. (2023). Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*.

Mutti, M., De Santi, R., De Bartolomeis, P., and Restelli, M. (2023). Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42.

Mutti, M., De Santi, R., and Restelli, M. (2022a). The importance of non-markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*.

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

Prajapat, M., Mutny, M., Zeilinger, M., and Krause, A. (2024). Submodular reinforcement learning. In *International Conference on Learning Representations*.

Prashanth, L., Jie, C., Fu, M., Marcus, S., and Szepesvári, C. (2016). Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*.

Van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

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

Ying, D., Guo, M. A., Ding, Y., Lavaei, J., and Shen, Z.-J. (2023a). Policy-based primal-dual methods for convex constrained markov decision processes. In *AAAI Conference on Artificial Intelligence*.

Ying, D., Zhang, Y., Ding, Y., Koppel, A., and Lavaei, J. (2023b). Scalable primal-dual actor-critic method for safe multi-agent RL with general utilities. In *Advances in Neural Information Processing Systems*.

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

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. (2021). Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*.

Zhang, J., Ni, C., Szepesvari, C., and Wang, M. (2021). On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*.

Zhang, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180 – 2210.

A Appendix

Contents

1 Introduction **1**

2 Problem Formulation **3**

3 Global Optimality in the Tabular Setting **3**

4 Global Optimality Beyond the Tabular Setting **5**

4.1 Challenges for Large-scale RLGU 6

4.2 Occupancy Measure Estimation 6

4.3 Global Optimality for Policy Gradient with Occupancy Measure Approximation . . 7

4.4 Novelty and Comparison to Prior Work 10

5 Conclusion **10**

A Appendix **13**

B Extended Related Work Discussion **14**

C Occupancy Measures in Low-Rank MDPs **19**

D Examples of Nonconcave RLGU Problems **19**

E Proofs for Section 4.3 **20**

E.1 State Sampling for MLE 20

E.2 Proof of Proposition 1 20

E.3 Proof of Theorem 2 22

E.4 Proof of Theorem 3 24

E.5 Useful technical result 25

F Future Work **25**

 https://doi.org/10.52202/085713-2661

B Extended Related Work Discussion

Table 1: Comparison to closest related works about RLGU.

Reference	First-order stationarity rate[1]	Global optimality rate[2]	Beyond tabular[3]	No state space size dependence[4]
Hazan et al. (2019)	✘	$\tilde{\mathcal{O}}(\epsilon^{-3})^{\&}$	✘	✘
Zhang et al. (2020)	$\tilde{\mathcal{O}}(\epsilon^{-2})^{*}$	$\tilde{\mathcal{O}}(\epsilon^{-1})^{*}$	✘	✘
Zhang et al. (2021)	$\tilde{\mathcal{O}}(\epsilon^{-3})^{\#}$	$\tilde{\mathcal{O}}(\epsilon^{-2})^{\#}$	✘	✘
Zahavy et al. (2021)	✘	$\tilde{\mathcal{O}}(\epsilon^{-3})^{\&}$	✘	✘
Barakat et al. (2023) (sec. 4)	$\tilde{\mathcal{O}}(\epsilon^{-3})^{\#}$	$\tilde{\mathcal{O}}(\epsilon^{-2})^{\#}$	✘	✘
Barakat et al. (2023) (sec. 5)	$\tilde{\mathcal{O}}(\epsilon^{-4})$	✘	✓	✘
Mutti et al. (2023)$^{+}$	✘	$\tilde{\mathcal{O}}(\epsilon^{-2})^{\&}$	✓	✘
This paper	$\tilde{\mathcal{O}}(m\epsilon^{-4})^{\S}$	$\tilde{\mathcal{O}}(m\epsilon^{-4})^{\S}$	✓	✓

$\tilde{\mathcal{O}}$ hides logarithmic factors in the accuracy ϵ, mainly due to the horizon length in the infinite horizon discounted reward setting.

[1] refers to the number of samples (or number of iterations in the deterministic case when specified) to achieve a given first-order stationarity ϵ, i.e. $\mathbb{E}[\|\nabla_\theta F(\lambda(\bar{\theta}_T))\|] \leq \epsilon$ where $\bar{\theta}_T$ is sampled uniformly at random from the iterates of the algorithm $\{\theta_1, \cdots, \theta_T\}$ until timestep T.

[2] refers to the number of samples (or number of iterations in the deterministic case when specified) to achieve global optimality under convexity of the general utility function F w.r.t. its occupancy measure variable, i.e. $F^* - F(\lambda(\theta_T)) \leq \epsilon$ where F^* is the maximum utility achieved for an optimal policy and θ_T is the last iterate of the algorithm generated after T steps.

[3] means that the large scale state action space is discussed and addressed, i.e., the work is not restricted to the tabular setting in which occupancy measures are estimated using a simple Monte Carlo (count-based) estimator for each state $s \in \mathcal{S}$. For a more extended discussion regarding this point and comparison to prior work, please see the rest of this section below.

[4] means that the performance bounds provided for first-order stationarity or global optimality do not depend on the state space size.

[&] These results do not hold for the last iterate like for the other results but rather for a mixture of policies in Hazan et al. (2019) (Theorem 4.4), an averaged occupancy measure over the iterates in Zahavy et al. (2021) (Lemma 2) and an average regret guarantee leading to a statistical (rather than computational) complexity in Mutti et al. (2023) (Theorem 5).

[*] This is for the deterministic setting only, i.e. only reporting the number of iterations required. The rate is further improved to be linear under strong convexity of the general utility function. Other results provided report sample complexities.

[#] These results make use of variance reduction in the tabular setting to obtain improved sample complexities compared to vanilla PG algorithms.

[+] This result considers a different (single trial) problem formulation compared to ours (and other works in the literature), see detailed discussion below for a comparison.

[§] m refers to the dimension of the function approximation class parameter for occupancy measure approximation, see eq. (12) and section 4.3. It should be noted here that we suppose access to a maximizer of the log-likelihood (13) (which requires some computational complexity that we do not discuss here), this is common in sample complexity analysis. Note also that all the other results suffer from a dependence on the size of the state space (explicit or hidden in the statements).

Comparison to Barakat et al. (2023). The work of Barakat et al. (2023) is mostly focused on the tabular setting (secs. 1 to 4). Section 5 therein is the only relevant section to our work which focuses on the large state action space setting. We list here several fundamental differences with our work and crucial improvements in terms of scalability:

(a) **MSE vs MLE.** The aforementioned work we compare to here uses a mean squared error estimator (MSE) whereas we use a maximum likelihood estimator (MLE), this difference turns out to be crucial for scalability. This is because mean square error estimation for occupancy measure estimation fails to scale to large state action spaces. To see this, consider an even simpler setting: suppose we have an unknown distribution $p^\star$ over a space $\mathcal{X}$ and i.i.d. samples $X_i \sim p^\star$ with $i = 1, \cdots, n$. MLE provides a TV bound $\|p - p^\star\|_1 \leq \epsilon$ where the accuracy ϵ is some $|\mathcal{X}|$-independent quantity that only depends on the sample size and complexity of the hypothesis class. In stark contrast, mean square regression would lead to

$\mathbb{E}_{x\sim p^\star}[(p(x)-p^\star(x))^2] \leq \epsilon$. By the Cauchy-Schwartz inequality (which is tight if the error $p(x)-p^\star(x)$ is relatively uniform over the space), we obtain $\mathbb{E}_{x\sim p^\star}[|p(x)-p^\star(x)|] \leq \sqrt{\epsilon}$. While this bound is close to the TV error bound above, it has an extra $p^\star(x)$ which implies an extra $|\mathcal{X}|$ dependence compared to the MLE approach if $p^\star$ is close to uniform. This is fundamentally not scalable. Note that MLE works even for densities over continuous spaces as it is already extensively used in the statistics literature. Please see also below (in the same section) for an extended discussion regarding MLE vs MSE;

(b) **Dependence on the state space size.** Their results do not make the dependence on the state space explicit and do not show an (exclusive) dependence on the dimension d of the state action feature map. It is required in their Theorem 5.4 that $\rho(s) \geq \rho_{\min}$. Notice that if ρ covers the whole state space like in the uniform distribution case, then $1/\rho_{\min}$ scales as the state space size. The dependence on this quantity is not made explicit in Theorem 5.4. After a close investigation of their proof, one can spot the dependence on $1/\rho_{\min}$ (which scales with S) in their constants (see e.g. in the constant $\tilde{C}_2$ in eq. (130) p. 41 in the detailed version of the theorem, see also eq. (139) p. 42 and eq. (143) p. 43 for more details).

(c) **Global convergence.** In contrast to our work (see our theorem 2 and corollary 2), they only provide a first-order stationarity guarantee and they do not provide global convergence guarantees;

(d) **Technical analysis.** From the technical viewpoint, our occupancy measure MLE estimation procedure combined with our PG algorithm requires a different theoretical analysis even for our first order stationarity guarantee. Please see appendix E below;

(e) **Experiments.** They do not provide any simulations testing their algorithm in section 5 for large state action spaces, Fig. 1 therein is only for the tabular setting.

More about MSE vs MLE. It is known that MSE is equivalent to MLE when the errors in a linear regression problem follow a normal distribution. However, as first preliminary comments regarding the comparison to the approach in Barakat et al. (2023), we additionally note that: (a) they only discuss the finite state action space setting for which this connection to MLE is not relevant and (b) there is no discussion nor any assumption about normality of the errors or any extension to the continuous state action space setting, we also observe that the occupancy measure values are bounded between 0 and $1/(1-\gamma)$ (or 0 and 1 for the normalized occupancies) which is a finite support that cannot be the support of a Gaussian distribution.

Beyond these first comments, let us now elaborate in more details on their approach and its potential regarding scalability to provide further clarifications. Our goal is to learn the (normalized) state occupancy measure d^{π_θ} induced by a given policy π_θ which is a probability distribution. In the discrete setting, this boils down to estimate $d^{\pi_\theta}(s)$ for every $s \in \mathcal{S}$. Note first that this quantity can be extremely small for very large state space settings which are the focus of our work, making the probabilities hard to model especially when using a regression approach.

The approach adopted in Barakat et al. (2023) consists in seeing this estimation problem as a regression problem. In more details, since the whole distribution needs to be estimated, they propose to consider an expected mean square error over the state space (rather than solving $|\mathcal{S}|$ regression problems - one for each $\lambda^{\pi_\theta}(s)$ - which is not affordable given the scalability objective). Hence the mean square loss they define is an expected error over a state distribution ρ to obtain an aggregated objective. This is less usual and specific to our occupancy measure estimation problem (this aggregation is not the mean over observations). This introduces a scalability issue as we recall that we would like to estimate $d^{\pi_\theta}(s)$ for every $s \in \mathcal{S}$, so the aggregated MSE objective considered there (see eq. (11) p. 7 therein) introduces a discrepancy w.r.t. the initial objective of estimating the whole distribution.

We do not exclude that a mean square error approach under suitable statistical model assumptions might address the occupancy measure estimation problem in a scalable way for large state action spaces for the continuous setting. However, this is not addressed in Barakat et al. (2023), their regression approach needs to be amended to address issues we mentioned above to be applicable and relevant to occupancy measure estimation and we are not sure that can be even achieved to tackle the problem for both discrete and continuous settings as we do.

Illustrative example for the limitations of MSE vs MLE for probability distribution estimation. We provide a simple illustrative example. Consider a simple case where the distribution $p^*(x)$ to be estimated is uniform ($p^*(x) = 1/|\mathcal{X}|$ where $|\mathcal{X}|$ is the size of the state space). The estimated

 https://doi.org/10.52202/085713-2661

distribution $p(x) = 2/|\mathcal{X}|$ on one half of the space and 0 on the other-half i.e this distribution is non-uniform, assigning a higher probability to events in one part of the space and zero probability to events in the other part. The expected loss thus incurred in this scenario using regression (namely $\sum_{x\in\mathcal{X}} p^*(x)(p(x) - p^*(x))^2$) scales as $O(1/|\mathcal{X}|^2)$ after a simple computation. This means that with large cardinality of the space, it becomes impossible to detect the difference between the two models even with infinite data when doing regression, whereas MLE does not suffer from this issue.

The primary difference between regression and MLE is that MLE results in a useful TV error bound (see Zhang (2006) and Huang et al. (2023) (Lemma 12) which we make use of in our analysis) i.e $\|p - p^*\|_1 \leq \epsilon$, where ϵ is independent of the cardinality of the space $|\mathcal{X}|$ and depends only on the sample size and complexity of the hypothesis class. In contrast, in the case of regression (MSE) where the expected loss is optimized, we get

$$\mathbb{E}_{x\sim p^*}\|p - p^*\|^2 \leq \epsilon, \mathbb{E}_{x\sim p^*}\|p - p^*\| \leq \sqrt{\epsilon}, \tag{16}$$

where the second inequality stems from an application of the Cauchy-Schwartz inequality. Note that we can write the left-hand side of the above last inequality as $\sum_{x\in\mathcal{X}} |p(x) - p^*(x)| \cdot p^*(x) \leq \epsilon$, which would eventually lead to the total variation norm upper-bounded by $\sqrt{\epsilon} \times |\mathcal{X}|$, assuming $p^*(x) = 1/|\mathcal{X}|$ to be uniform for illustration, thus incurring a large error while estimating the distribution.

Comments about limitations of their formulation.

The expected loss is over the initial state distribution (defining the MDP). Take the extreme case where we initialize at a single state (note that this is also realistic, e.g. a robot starting at a given deterministic state). Then this means that the expected loss boils down to estimating the occupancy measure only at that state. However we need to estimate it as accurately as possible for all states and there is no reason why the occupancy measure should be supported by the same set of states as the initial distribution (which we should have freedom about). Note also that the occupancy measure itself depends on the initial distribution. In principle, the distribution used for defining the expected loss (over state action pairs) should be different from the initial state distribution defining the MDP.

Coverage and scalability problem. Now you may argue that it is enough to take an initial distribution (or just ρ distribution for the expected loss if one assumes it is unrelated to the initial distribution) that just needs to cover the support of the occupancy measure we want to estimate. Note that the occupancy measure is unknown and we want to estimate it so we have a priori no clue about its support. One might then think about just taking the uniform distribution as an initial distribution to be sure to cover the whole state space equally. This choice is problematic for several reasons: (a) First, this introduces a bias: Why would we need to estimate the occupancy measure equally well in all the states if the occupancy measure is concentrated on a specific set of states which is not necessarily the entire state space? (b) Second and most importantly, if we make such a choice, we have now $\rho_{\min} = 1/|\mathcal{S}|$ (say in the discrete state space setting) and the first order stationarity bound in Barakat et al. (2023) scales with $1/\rho_{\min}$, this makes the result not scalable to large state spaces. You might argue that we do not need the uniform distribution but just to take a distribution covering the entire state space (not necessarily equally well), i.e. which has a support equal to the entire state space. Then again, this introduces a bias as the loss minimization might focus on states which are irrelevant to the occupancy measure we want to estimate. Furthermore, note that MSE might not be the best metric for comparing distributions because it focuses on pointwise differences (in our case states or state-action pairs). In our setting, how the weights of the MSE loss are chosen for fitting our probability distribution of interest is important.

Shortcomings of using MSE compared to MLE. There are a number of shortcomings of using MSE compared to MLE for fitting a probability distribution in general, we summarize them here:

Consistency and efficiency. MLE maximizes the likelihood function, ensuring the estimator is consistent (i.e. converges to the true parameter value as the sample size increases) and asymptotically efficient (i.e. achieves the lowest possible variance among unbiased estimators for large samples). In contrast, as MSE minimizes mean squared errors, it does not guarantee properties like consistency or efficiency unless under specific assumptions on the data such as normality of the errors in which case both coincide. Note that the approach in Barakat et al. (2023) does not fit a Gaussian distribution to the normalized occupancy measure. We use favorable statistical properties of MLE (see Proposition 1).

Sensitivity to scaling. MLE operates on probabilities and likelihoods, these are normalized and scale-invariant. This makes MLE suitable for probability distribution fitting in general. MSE is rather better used for pointwise estimation in statistics (which is also indirectly used for distribution estimation via estimating parameters such as Gaussian means), MSE depends on the scale of the observations which might make it less robust in some settings.

Robustness to outliers. As MLE models probabilities directly, it might be more robust to outliers depending on the distribution. MSE can be more sensitive to outliers and extreme values as it relies on squared errors which amplify such outliers. In our setting, this is also relevant as we are interested in estimating occupancy measures on large state spaces, this induces small probability values (even extremely small for some of them) and squaring differences makes it worse, see a discussion in appendix B.

Satisfying distribution constraints. MLE naturally adapts to the distribution's shape and constraints to satisfy them. In contrast, MSE can lead to estimates that violate distribution constraints such as probability normalization or predicting a negative variance for a normal distribution. Therefore, post-processing might be required to ensure these are satisfied.

Comparison to Theorem 5, section 3 in Mutti et al. (2023). We enumerate the differences between our results and settings in the following:

1. **Problem formulation.** As mentioned in the short related work section in the main part, Mutti et al. (2023) consider a finite trial version of the convex RL problem which has its own merits (for settings where the objective itself only cares about the performance on the finite number of realizations the agent can have access to instead of an expected objective which can be interpreted as an infinite realization access setting, see discussion therein) but this formulation is different from ours. Both coincide when the number of trials they consider goes to infinity. Although the problem formulations are different, let us comment further on some additional differences in our results.

2. **Assumptions.** They assume linear realizability of the utility function F with known feature vectors (Assumption 4, p. 17 therein). Our setting differs for two reasons: (1) We do not approximate the utility function itself but rather the occupancy measure and (2) we train a neural network to learn an occupancy measure approximation by maximizing a log-likelihood loss. In our case, our analog (similar but different in formulation and nature) assumption would be our function approximation class regularity assumption (Assumption 2). We do not suppose access to feature vectors which are given. Nevertheless, we do suppose that we can solve the log-likelihood optimization problem to optimality (which is approximated in practice and widely used among practitioners).

3. **Algorithm.** The algorithm they use is model-based, they repeatedly solve a regression problem to approximate the utility function F using samples and use optimism for ensuring sufficient exploration. Our policy gradient algorithm is model-free and we rather rely on MLE for approximating occupancy measures rather than regression.

4. **Analysis.** Under concavity of the utility function, we provide a last iterate global optimality guarantee whereas Mutti et al. (2023) establish an average regret guarantee which is different in nature. Their proof relies on a reduction to an online learning once-per-episode framework. Our proof ideas are different: We combine a gradient optimization analysis exploiting hidden convexity with a statistical complexity analysis for occupancy measure estimation. Overall, our results combine optimization and statistical guarantees whereas their results focus purely on the statistical complexity (as their problem is computationally hard).

Comparison to Mutti et al. (2023). Let us list first a few advantages/differences w.r.t. the aforementioned work:

Advantages of our analysis/approach.

No planning oracle access. We do not suppose access to a planning oracle able to solve any convex MDP efficiently. This is precisely the point of our optimization guarantees for our PG algorithm which updates policies incrementally. Nevertheless, we point out here that Mutti et al. (2023) address a slightly different (finite-trial) convex RL problem which is computationally intractable. Our problem coincides with their infinite trial variant of the problem.

 https://doi.org/10.52202/085713-2661

No access required to feature vectors for function approximation. We learn occupancy measure approximations rather than supposing access to a set of features to approximate utilities (i.e. $F(\lambda)$ in our notations), i.e. we do not suppose access to a set of basis feature vectors for our approximation.

Model-free and no dependence on state action space size. We do not estimate the transition kernel, our algorithm is model-free as we only require access to sampled trajectories. In particular, we do not require to go through the entire state action space to estimate each entry of the transition kernel. Therefore, our performance bounds do not have dependence on the state action space sizes as their regret bound.

Policy parameterization. We consider policy parameterization instead of tabular policies which results in a practical algorithm. We do require a strong assumption though (Assumption 5) to obtain our global optimality result.

Advantages of our model-free PG algorithm. We inherit the usual advantages of model-free vs model-based algorithms: (a) Model-free are often simpler to implement because they do not require learning or using a model of the environment. Our algorithm directly focuses on learning a policy (even if we do also approximate the occupancy measures but not the transitions themselves like in model-based approaches); (b) Robustness: Inaccuracies in environment modeling propagate to policy optimization and can significantly degrade performance. Model-free methods directly learn policies from interaction with the environment; (c) PG methods are particularly suitable for complex and high-dimensional environments settings. Of course, model-based methods may also have advantages over model-free ones.

About hardness of occupancy measure estimation. We comment here on one of the challenges discussed in the main part as for estimating the occupancy measure. An occupancy measure induced by a policy π satisfies the identity $\lambda^\pi(s,a) = \mu_0(s,a) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \mathcal{P}(s|s',a')\pi(a|s')\lambda^\pi(s',a')$ where μ_0 is the initial state action distribution. Notice that the sum is not over the next action s in transition kernel P but rather the 'backward' state actions (s',a'). In contrast, an action value function in standard RL rather satisfies a 'forward' Bellman equation. In contrast to the standard Bellman equation which can be written using an expectation and leads to a sampled version of the Bellman fixed point equation, the equation satisfied by the occupancy measure cannot be written under an expectation form and does not naturally lead to any stochastic algorithm. This issue is recognized in the literature in Huang et al. (2023) (see also Hallak and Mannor (2017)).

More about assumptions. Besides the points above, we provide a few comments regarding assumptions:

1. Some of our assumptions are quite similar. For instance, Mutti et al. (2023) assume linear realizability of the utility function with known feature vectors (Assumption 4, p. 17 therein) and then assume access to a regression problem solver with cross-entropy loss to approximate the utility function. We rather have a similar but different function approximation class regularity assumption (Assumption 1) and we suppose access to an optimizer which solves our log-likelihood loss maximization problem to approximate occupancy measures (see Eq. (8)). We both assume concavity of the utility function.
2. We require smoothness assumptions on the utility function (Assumption 3) whereas Mutti et al. (2023) only require Lipschitzness of the same function (Assumption 1 therein). This is because smoothness is important for deriving optimization guarantees as we make use of gradient information whereas Lipschitzness is enough for developing their statistical analysis.
3. Mutti et al. (2023) assume access to an optimal planner (as mentioned above), we do not need such a requirement as we provide optimization guarantees using our PG algorithm.
4. We need policy parametrization assumptions as previously discussed, Mutti et al. (2023) do not consider policy parametrization.

Recently, Prajapat et al. (2024) and De Santi et al. (2024) proposed to go beyond linear and convex rewards by considering rewards which are defined globally over trajectories instead of locally over states.

For more recent work concerning the analysis of policy gradient methods in classical linear RL, we refer the reader to Liu et al. (2024, 2025).

C Occupancy Measures in Low-Rank MDPs

In this section, we show that occupancy measures have a linear structure in the so-called density features in low-rank MDPs. We provide a proof for completeness. Similar results were established in Lemma 16, 17 in Huang et al. (2023) for the finite-horizon setting. Throughout this section, we use the same notations as in the main part of this paper.

Definition C.1 (Low-rank MDPs)**.** *An MDP is said to be low-rank with dimension $d \geq 1$ if there exists a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and there exist d unknown measures $(\mu_1, \cdots, \mu_d)$ over the state space $\mathcal{S}$ such that for every states $(s, s') \in \mathcal{S}$ and every action $a \in \mathcal{A}$ it holds that*

$$P(s'|s, a) = \langle \phi(s, a), \mu(s') \rangle \,, \tag{17}$$

with $\|\phi\|_\infty \leq 1$ without loss of generality.

Before stating the result, recall that for any policy $\pi \in \Pi$, a state-occupancy measure is defined for every state $s \in \mathcal{S}$ as follows:

$$d^\pi(s) := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\rho,\pi}(s_t = s) \,. \tag{18}$$

Lemma 1. *Consider a low-rank infinite horizon discounted MDP. Then, for any policy $\pi \in \Pi$, there exists a vector $\omega_\pi \in \mathbb{R}^d$ such that the state-action occupancy measure d^π induced by the policy π satisfies for any state $s \in \mathcal{S}$,*

$$d^\pi(s) = \rho(s) + \langle \omega_\pi, \mu(s) \rangle \,, \tag{19}$$

where we use the notation $\mu(s) := (\mu_1(s), \cdots, \mu_d(s))^T$.

Proof. Let $\pi \in \Pi$. It follows from the definition of the state-occupancy measure d^π induced by the policy π that it satisfies the following (backward) Bellman flow equation for every state $s \in \mathcal{S}$:

$$d^\pi(s) = \rho(s) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} P(s|s', a') \pi(a'|s') d^\pi(s') \,. \tag{20}$$

Using the definition of a low-rank MDP and (17) in particular, we obtain:

$$\begin{aligned} d^\pi(s) &= \rho_0(s) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \langle \phi(s', a'), \mu(s) \rangle \pi(a'|s') d^\pi(s') && (21) \\ &= \rho_0(s) + \left\langle \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \phi(s', a') \pi(a'|s') d^\pi(s'),\, \mu(s) \right\rangle && (22) \\ &= \rho_0(s) + \langle \omega_\pi, \phi(s) \rangle \,, && (23) \end{aligned}$$

where we define $\omega_\pi := \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \phi(s', a') \pi(a'|s') d^\pi(s')$. □

D Examples of Nonconcave RLGU Problems

Nonconvexity is ubiquitous in real-world applications and we provide below a few examples where it naturally arises beyond the standard convex RL examples in the literature. First of all, we would like to mention risk-sensitive RL with non-convex risk measures inspired by Cumulative Prospect Theory (CPT) (with S-shaped utility curves). Nonconvex criteria are important for modeling human decisions. See e.g. (Lin and Marcus, 2013; Lin et al., 2018) for a discussion about their relevance and importance. See also Remark 1 and figure 2 p. 3 in Prashanth et al. (2016).

Applications include for instance:

- **Robotics control:** in control tasks, it is common to deal with nonconvex objectives such as minimizing energy consumption while achieving a task or maximizing the success rate of a manipulation task.
- **Portfolio Management:** Utility functions in finance may be non-convex due to risk measures or transaction costs for example.

https://doi.org/10.52202/085713-2661

- **Traffic Control:** RL can be used to optimize traffic flow and minimize congestion. The utility function may involve non-convex terms such as travel time, queue lengths, and safety constraints.
- **Supply Chain Management:** RL can be applied to inventory control, pricing, and logistics optimization. The utility function may include non-convex components such as demand forecasting, supply chain disruptions, and dynamic pricing.

We leave the experimental investigation of those applications for future work. We hope our work will foster more research in this direction.

E Proofs for Section 4.3

E.1 State Sampling for MLE

In this section, we briefly discuss how to sample states following the (normalized) state occupancy d^{π_θ} for a given policy π_θ . In particular, these states are used for the MLE procedure described in section 4.2. The idea consists in sampling states following the transition kernel $\mathcal{P}$ and the policy π_θ for a random horizon following a geometric distribution of parameter γ where γ is the discount factor, starting from a state drawn from the initial distribution. The detailed sampling procedure is described in Algorithm 2, borrowed and adapted from Yuan et al. (2023) (Algorithm 3 p. 22) which provides a clear presentation of the idea as well as a simple supporting proof (see Lemma 4 p. 23 therein). This procedure has been commonly used in the literature, see e.g. Algorithm 1 p. 30 and Algorithm 3 p. 34 in Agarwal et al. (2021).

Algorithm 2 Sampler for $s \sim d_\rho^{\pi_\theta}$

Input: Initial state distribution ρ, policy π_θ, discount factor $\gamma \in [0, 1)$
Initialize $s_0 \sim \rho, a_0 \sim \pi_\theta(\cdot|s_0)$, time step $h, t = 0$, variable $X = 1$
while $X = 1$ **do**
  With probability γ**:**
  Sample $s_{h+1} \sim \mathcal{P}(\cdot \mid s_h, a_h)$
  Sample $a_{h+1} \sim \pi_\theta(\cdot|s_{h+1})$
  $h \leftarrow h + 1$
  EndWith
  Otherwise with probability $1 - \gamma$**:**
  $X = 0$ `(Accept` s_h`)`
  EndOtherwise
end while
Return: s_h

E.2 Proof of Proposition 1

Proposition 1 and its proof are largely based on the work of Huang et al. (2023): We follow and reproduce their proof strategy here. Since the latter paper deals with a more complex setting that does not exactly fit our current focus, we provide a proof for clarity and completeness.

We start by defining the concept of l_1 optimistic cover. This cover will be immediately useful to quantify the complexity of our (possibly infinite) approximating function class Γ defined in (12).

In the following, we denote by $\{\mathcal{X} \to \mathbb{R}\}$ the set of functions defined on $\mathcal{X}$ with values in $\mathbb{R}$.

Definition E.1 (Definition 3 in Huang et al. (2023))**.** *For a given function class $\Lambda \subseteq \Delta(\mathcal{X})$, the function class $\bar{\Lambda} \subseteq (\mathcal{X} \to \mathbb{R})$ is said to be an l_1 optimistic cover of Λ with scale $\kappa > 0$ if:*

$$\forall \lambda \in \Lambda, \quad \exists\, \bar{\lambda} \in \bar{\Lambda} \quad \textit{s.t.} \quad \|\lambda - \bar{\lambda}\|_1 \leq \kappa, \quad \textit{and} \quad \lambda(x) \leq \bar{\lambda}(x), \forall x \in \mathcal{X}\,. \tag{24}$$

Remark 5. *Notice that $\bar{\Lambda}$ does not need to be a set containing only probability distributions if Λ is a set of probability distributions, namely the set of (normalized) occupancy measures as we will be considering in the rest of this section.*

We now provide a general statistical guarantee for the maximum likelihood estimator (MLE) defined in (13) supposing we have access to an optimistic cover of the space of distributions used for computing the MLE estimator.

Proposition 2 (Lemma 12 in Huang et al. (2023))**.** *Let $\mathcal{D} := \{x_i\}_{i=1}^n$ be a dataset of state-action pairs drawn i.i.d from some fixed probability distribution $\lambda^* \in \Delta(\mathcal{X})$. Let $\Lambda \subseteq \Delta(\mathcal{X})$ be a function class such that:*

(i) (realizability) $\lambda^* \in \Lambda$,

(ii) (probability distribution class) $\forall \lambda \in \Lambda, \lambda \in \Delta(\mathcal{X})$,

(iii) (covering) Λ has a finite l_1-optimistic cover $\bar{\Lambda} \subseteq \{\mathcal{X} \to \mathbb{R}_{\geq 0}\}$ with scale κ (see Definition E.1).

Then, for any $\delta > 0$, we have with probability at least $1 - \delta$,

$$\|\hat{\lambda} - \lambda^*\|_1 \leq \kappa + \sqrt{\frac{12 \log\left(\frac{|\bar{\Lambda}|}{\delta}\right)}{n} + 6\kappa}\,, \tag{25}$$

where $\hat{\lambda}$ is the MLE estimator defined in (13) computed using the dataset $\mathcal{D}$ and $|\bar{\Lambda}|$ is the cardinality of the finite cover $\bar{\Lambda}$.

In view of using Proposition 2, the next lemma constructs an l_1 optimistic cover for the function approximation class Λ used to computed the MLE. For the reader's convenience, we recall that

$$\Lambda := \{p_\omega : \omega \in \Omega \subseteq \mathbb{R}^d, p_\omega \in \Delta(\mathcal{X})\}\,. \tag{26}$$

Lemma 2. *Let Assumption 2 hold. Then there exists a finite l_1-optimistic cover $\bar{\Lambda} \subseteq \{\mathcal{X} \to \mathbb{R}_{\geq 0}\}$ of the function class Λ with scale $\kappa > 0$ and size at most $2\lceil \frac{B_\omega B_L}{\kappa} \rceil^m$ where m is the dimension of the parameter space $\Omega \subseteq \mathbb{R}^m$.*

Proof. The proof follows the same lines as the proof of Lemma 22 p. 41 in Huang et al. (2023). Let $\lambda \in \Lambda$, i.e., $\lambda = p_\omega$ for some $\omega \in \Omega$. Let $\kappa' > 0$. Define the set $\mathcal{B}(\omega, \kappa') := \kappa' \lfloor \frac{\omega}{\kappa'} \rfloor + [0, \kappa']^m$ which is a cubic κ'-neighborhood of the point $\omega \in \Omega$. Now define the function f_ω for every $x \in \mathcal{X}$ as follows:

$$f_\omega(x) := \max_{\bar{\omega} \in \mathcal{B}(\omega, \kappa')} p_{\bar{\omega}}(x)\,. \tag{27}$$

By construction, we immediately have $f_\omega(x) \geq p_\omega(x) \geq 0$. Note that f_ω might not be a probability distribution though. Then using Assumption 2 we also have

$$\begin{aligned}
\|f_\omega - p_\omega\|_1 &= \int |f_\omega(x) - p_\omega(x)| dx \\
&= \int |\max_{\bar{\omega} \in \mathcal{B}(\omega, \kappa')} p_{\bar{\omega}}(x) - p_\omega(x)| dx \\
&\leq \int \max_{\bar{\omega} \in \mathcal{B}(\omega, \kappa')} |p_{\bar{\omega}}(x) - p_\omega(x)| dx \leq \int \max_{\bar{\omega} \in \mathcal{B}(\omega, \kappa')} |L(x)| \cdot \|\bar{\omega} - \omega\|_\infty dx \leq B_L \kappa'\,.
\end{aligned} \tag{28}$$

To conclude, we observe that there are at most $2\lceil \frac{B_\omega}{\kappa'} \rceil^m$ unique functions in the l_1-optimistic cover $\bar{\Lambda}$ of Λ which is of scale $B_L \kappa'$. Setting $\kappa' = \frac{\kappa}{B_L}$ concludes the proof. □

End of Proof of Proposition 1. We conclude the proof by using Proposition 2 together with Lemma 2 above, choosing a scale $\kappa = \frac{1}{n}$ where n is the number of samples used for computing the MLE and plugging $|\bar{\Lambda}| \leq 2\lceil B_\omega B_L n \rceil^m$. We obtain after simple upper-bounding inequalities,

$$\|\hat{d}^{\pi_\theta} - d^{\pi_\theta}\|_1 \leq 6\sqrt{\frac{12\, m \log\left(\frac{2\lceil B_\omega B_L n \rceil}{\delta}\right)}{n}}\,. \tag{29}$$

https://doi.org/10.52202/085713-2661

E.3 Proof of Theorem 2

The proof follows similar lines to the proof of Theorem 5.4 in Barakat et al. (2023). However, our occupancy measure estimation procedure is different in the present case. We provide a full proof here for completeness.

We introduce the shorthand notation $\bar{g}_t := \frac{1}{N}\sum_{i=1}^{N} g(\tau_t^{(i)}, \theta_t, r_t)$ for this proof. Using the smoothness of the objective function $\theta \mapsto F(\lambda(\theta))$ (see Lemma 3 in Appendix E.5) and the update rule of the sequence (θ_t), we have

$$
\begin{aligned}
F(\lambda(\theta_{t+1})) &\geq F(\lambda(\theta_t)) + \langle \nabla_\theta F(\lambda(\theta_t)), \theta_{t+1} - \theta_t \rangle - \frac{L_\theta}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= F(\lambda(\theta_t)) + \alpha \langle \nabla_\theta F(\lambda(\theta_t)), \bar{g}_t \rangle - \frac{L_\theta \alpha^2}{2}\|\bar{g}_t\|^2 \\
&= F(\lambda(\theta_t)) + \alpha \langle \nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t, \bar{g}_t \rangle + \alpha\left(1 - \frac{L_\theta \alpha}{2}\right)\|\bar{g}_t\|^2 \\
&\geq F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 - \frac{\alpha}{2}\|\bar{g}_t\|^2 + \alpha\left(1 - \frac{L_\theta \alpha}{2}\right)\|\bar{g}_t\|^2 \\
&= F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{2}(1 - L_\theta \alpha)\|\bar{g}_t\|^2 \\
&\overset{(i)}{\geq} F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{4}\|\bar{g}_t\|^2 \\
&= F(\lambda(\theta_t)) - \frac{\alpha}{2}\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2 \\
&\overset{(ii)}{\geq} F(\lambda(\theta_t)) + \frac{\alpha}{16}\|\nabla_\theta F(\lambda(\theta_t))\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2 \,, \qquad (30)
\end{aligned}
$$

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

We now control the last error term in the above inequality in expectation. Recalling that $\nabla_\theta F(\lambda(\theta)) = \nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\lambda(\theta))}$ for any $\theta \in \mathbb{R}^d$, we have

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2] = \mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(r)_{r=\nabla_\lambda F(\nabla(\theta_t))} - \bar{g}_t\|^2] \\
&\leq 2\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\lambda(\theta_t))} - \nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\hat{\lambda}_t)}\|^2] + 2\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\hat{\lambda}_t)} - \bar{g}_t\|^2] \,. \qquad (31)
\end{aligned}
$$

Now, we upper bound each one of the two terms above separately. For convenience, we introduce the notations $r_t := \nabla_\lambda F(\lambda(\theta_t))$ and $\hat{r}_t := \nabla_\lambda F(\hat{\lambda}_t)$.

Upper bound of the term $\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(r_t) - \nabla_\theta V^{\pi_\theta}(\hat{r}_t)\|^2]$ in (31). Using the policy gradient theorem (see (4)) yields

$$
\nabla_\theta V^{\pi_\theta}(r_t) - \nabla_\theta V^{\pi_\theta}(\hat{r}_t) = \mathbb{E}\left[\sum_{t'=0}^{H-1} \gamma^{t'}[\nabla_\lambda F(\lambda(\theta_t))) - \nabla_\lambda F(\hat{\lambda}_t)]_{s_{t'}, a_{t'}} \cdot \left(\sum_{h=0}^{t'} \nabla_\theta \log \pi_\theta(a_h, s_h)\right)\right] . \qquad (32)
$$

Notice that the above expectation is only taken w.r.t. the state action pairs in the random trajectory of length H . Taking the norm, we obtain

$$\begin{aligned}\|\nabla_\theta V^{\pi_\theta}(r_t) - \nabla_\theta V^{\pi_\theta}(\hat{r}_t)\|_2 &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t'=0}^{H-1} \gamma^{t'} \|\nabla_\lambda F(\lambda(\theta_t))) - \nabla_\lambda F(\hat{\lambda}_t)\|_\infty \left\|\sum_{h=0}^{t'} \nabla_\theta \log \pi_\theta(a_h, s_h)\right\|_2\right] \\ &\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t'=0}^{H-1} 2l_\psi(t'+1)\gamma^{t'} \|\nabla_\lambda F(\lambda(\theta_t))) - \nabla_\lambda F(\hat{\lambda}_t)\|_\infty\right] \\ &\overset{(c)}{\leq} \mathbb{E}\left[\sum_{t'=0}^{H-1} 2l_\psi L_\lambda(t'+1)\gamma^{t'} \|\lambda(\theta_t) - \hat{\lambda}_t\|_2\right] \\ &\overset{(d)}{\leq} \frac{2l_\psi L_\lambda}{(1-\gamma)^2}\|\lambda(\theta_t) - \hat{\lambda}_t\|_2\,, \end{aligned} \tag{33}$$

where (a) follows from using the triangle inequality together with the definition of the sup norm, (b) uses Lemma 3 (i) in Appendix E.5, (c) is a consequence of Assumption 4 together with the fact that $\|x\|_\infty \leq \|x\|_2$ for any $x \in \mathbb{R}^d$, and (d) stems from the upper bound $\sum_{t'=0}^{H-1}(t'+1)\gamma^{t'} \leq \sum_{t'=0}^{\infty}(t'+1)\gamma^{t'} = \frac{1}{(1-\gamma)^2}$. Hence we have shown that

$$\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(r_t) - \nabla_\theta V^{\pi_\theta}(\hat{r}_t)\|_2^2] \leq \frac{4l_\psi^2 L_\lambda^2}{(1-\gamma)^4}\mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2]\,. \tag{34}$$

Upper bound of the term $\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(\hat{r}_t) - \bar{g}_t\|^2]$ in (31). Recalling the definition of $\bar{g}_t$, we have

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta V^{\pi_\theta}(\hat{r}_t) - \bar{g}_t\|^2] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\nabla_\theta V^{\pi_\theta}(\hat{r}_t) - g(\tau_t^{(i)}, \theta_t, \hat{r}_t))\right\|^2\right] \\ &\overset{(a)}{=} \frac{1}{N}\mathbb{E}[\|g(\tau_t^{(i)}, \theta_t, \hat{r}_t) - \nabla_\theta V^{\pi_\theta}(\hat{r}_t)\|^2] \\ &\overset{(b)}{\leq} \frac{1}{N}\mathbb{E}[\|g(\tau_t^{(i)}, \theta_t, \hat{r}_t)\|^2] \\ &\overset{(c)}{\leq} \frac{4l_\lambda^2 l_\psi^2}{(1-\gamma)^4 N}\,, \end{aligned} \tag{35}$$

where (a) follows from the fact that the expectation of $g(\tau_t^{(i)}, \theta_t, \hat{r}_t)$ w.r.t. the random trajectory $\tau_t^{(i)}$ conditioned on θ_t and $\hat{r}_t$ is precisely given by $\nabla_\theta V^{\pi_\theta}(\hat{r}_t)$ by the policy gradient theorem (see (4)), notice also that all the N trajectories are drawn i.i.d. As for (b), use the fact that the variance of a random variable is upper bounded by its second moment. Finally (c) stems from using the expression of $g(\tau_t^{(i)}, \theta_t, \hat{r}_t)$ in (11) together with Assumptions 3, 4 and Lemma 3 (i) in Appendix E.5. The proof of this last point follows similar lines to (33).

Combining both the previous upper bounds we have now established above, we obtain

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2] \leq \frac{\tilde{C}_1}{N} + \tilde{C}_2 \cdot \mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2]\,, \tag{36}$$

where $\tilde{C}_1 := \frac{8l_\lambda^2 l_\psi^2}{(1-\gamma)^4}$ and $\tilde{C}_2 := \frac{8l_\psi^2 L_\lambda^2}{(1-\gamma)^4}$.

End of Proof of Theorem 2. We are now ready to conclude the proof of our result. Going back to (30), rearranging the terms and taking expectation, we obtain

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t))\|^2] \leq \frac{16}{\alpha}\mathbb{E}[F(\lambda(\theta_{t+1})) - F(\lambda(\theta_t))] + 10\,\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2]. \tag{37}$$

Plugging the bound (36) into the previous inequality, we obtain

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t))\|^2] \leq \frac{16}{\alpha}\mathbb{E}[F(\lambda(\theta_{t+1})) - F(\lambda(\theta_t))] + \frac{10\tilde{C}_1}{N} + 10\tilde{C}_2 \cdot \mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2]\,, \tag{38}$$

https://doi.org/10.52202/085713-2661

Summing the previous inequality for $t = 1$ to T, telescoping the right hand side and using the upper bound $F^\star$ on the objective function leads to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t))\|^2] \leq \frac{16(F^\star - \mathbb{E}[F(\lambda(\theta_1))])}{\alpha T} + \frac{10\tilde{C}_1}{N} + \frac{10\tilde{C}_2}{T}\sum_{t=1}^{T}\mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2]\,. \quad (39)$$

Setting $C_1 := 10\tilde{C}_1$ and $C_2 := \tilde{C}_2$ gives the desired result.

E.4 Proof of Theorem 3

The proof of this result borrows some ideas from Zhang et al. (2021) and Barakat et al. (2023). However the algorithm we are analyzing is different and the proof deviates from the aforementioned results accordingly.

Remark 6. *A different technical analysis can be found in Fatkhullin et al. (2023) by considering a particular case of their theorem 5 dealing with stochastic optimization under hidden convexity. However, their general setting is not focused on our specific RLGU setting using policy parametrization and specifying the assumptions needed as a consequence. More importantly, we are considering a context in which unknown occupancy measures are approximated via function approximation using relevant collected state samples and our theorem accounts for the induced error. In contrast, Fatkhullin et al. (2023) assume access to an unbiased estimate of the gradient of the utility function which is not readily available in our RLGU setting since occupancy measures are unknown and estimated via function approximation with a supporting sample complexity guarantee. Besides these differences, we conduct a different analysis which is rather inspired by the proofs in Zhang et al. (2021) and Barakat et al. (2023) as previously mentioned.*

It follows from smoothness of the objective function $\theta \mapsto F(\lambda(\theta))$ (see (30)) that for every iteration t,

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_t)) + \frac{\alpha}{16}\|\nabla_\theta F(\lambda(\theta_t))\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2\,. \quad (40)$$

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

$$\theta_\eta := (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}((1-\eta)\lambda(\theta_t) + \eta\lambda(\theta^*)) \in \mathcal{U}_{\theta_t}\,. \quad (41)$$

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

$$F(\lambda(\theta_t)) \geq F(\lambda(\theta_\eta)) - \langle \nabla_\theta F(\lambda(\theta_t)), \theta_\eta - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_\eta - \theta_t\|^2\,. \quad (42)$$

Combining (40) and (42), we obtain

$$\begin{aligned} F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_\eta)) - \langle \nabla_\theta F(\lambda(\theta_t)), \theta_\eta - \theta_t\rangle - \frac{L_\theta}{2}\|\theta_\eta - \theta_t\|^2 \\ + \frac{\alpha}{16}\|\nabla_\theta F(\lambda(\theta_t))\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2\,. \quad (43) \end{aligned}$$

Now, pick $a \leq \frac{1}{16}$, using Young's inequality gives

$$\langle \nabla_\theta F(\lambda(\theta_t)), \theta_\eta - \theta_t\rangle \leq a\alpha\|\nabla_\theta F(\lambda(\theta_t))\|^2 + \frac{1}{a\alpha}\|\theta_\eta - \theta_t\|^2\,. \quad (44)$$

Plugging this inequality into (43) yields

$$\begin{aligned} F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_\eta)) + (\frac{\alpha}{16} - a\alpha)\|\nabla_\theta F(\lambda(\theta_t))\|^2 + \frac{\alpha}{8}\|\bar{g}_t\|^2 \\ - \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\|\theta_\eta - \theta_t\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2\,. \quad (45) \end{aligned}$$

Therefore, since $a \leq \frac{1}{16}$, we obtain

$$F(\lambda(\theta_{t+1})) \geq F(\lambda(\theta_\eta)) - \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\|\theta_\eta - \theta_t\|^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2\,. \quad (46)$$

Using the definition of θ_η and the concavity of F (Assumption 1), we now control each one of the terms $F(\lambda(\theta_\eta))$ and $\|\theta_\eta - \theta_t\|^2$.

(i) By concavity of F (Assumption 1) and using the definition of θ_η, we have

$$F(\lambda(\theta_\eta)) = F((1-\eta)\lambda(\theta_t) + \eta\lambda(\theta^*)) \geq (1-\eta)F(\lambda(\theta_t)) + \eta F(\lambda(\theta^*))\,. \tag{47}$$

(ii) Using the uniform Lipschitzness of the inverse mapping $(\lambda|_{\mathcal{U}_{\theta_t}})^{-1}$ (see Assumption 5), we have

$$\begin{aligned}\|\theta_\eta - \theta_t\|^2 &= \|(\lambda|_{\mathcal{U}_{\theta_t}})^{-1}((1-\eta)\lambda(\theta_t) + \eta\lambda(\theta^*)) - (\lambda|_{\mathcal{U}_{\theta_t}})^{-1}(\lambda(\theta_t))\|^2 \\ &\leq l_\theta^2\eta^2\|\lambda(\theta_t) - \lambda(\theta^*)\|^2 \\ &\leq \frac{4l_\theta^2\eta^2}{(1-\gamma)^2}\,.\end{aligned} \tag{48}$$

Injecting (47) and (48) into (46) yields

$$F(\lambda(\theta_{t+1})) \geq (1-\eta)F(\lambda(\theta_t)) + \eta F(\lambda(\theta^*)) - \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\frac{4l_\theta^2}{(1-\gamma)^2}\eta^2 - \frac{5}{8}\alpha\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2\,. \tag{49}$$

Rearranging the above inequality, adding F^* to both sides, taking expectation and using the notation $\delta_t := \mathbb{E}[F^* - F(\lambda(\theta_t))]$, we obtain

$$\delta_{t+1} \leq (1-\eta)\delta_t + \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\frac{4l_\theta^2}{(1-\gamma)^2}\eta^2 + \frac{5}{8}\alpha\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2]\,. \tag{50}$$

Recall then from (36) that

$$\mathbb{E}[\|\nabla_\theta F(\lambda(\theta_t)) - \bar{g}_t\|^2] \leq \frac{\tilde{C}_1}{N} + \tilde{C}_2 \cdot \mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2]\,. \tag{51}$$

Since $\mathbb{E}[\|\lambda(\theta_t) - \hat{\lambda}_t\|_2^2] \leq \epsilon_{\text{MLE}}$ uniformly over the iterations, we get by combining (50) and (51) that

$$\delta_{t+1} \leq (1-\eta)\delta_t + \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\frac{4l_\theta^2}{(1-\gamma)^2}\eta^2 + \frac{5}{8}\alpha\left(\frac{\tilde{C}_1}{N} + \tilde{C}_2\epsilon_{\text{MLE}}\right)\,. \tag{52}$$

Finally, unrolling this recursion gives

$$\delta_T \leq (1-\eta)^T\delta_0 + \left(\frac{L_\theta}{2} + \frac{1}{a\alpha}\right)\frac{4l_\theta^2}{(1-\gamma)^2}\eta + \frac{5}{8}\frac{\alpha}{\eta}\left(\frac{\tilde{C}_1}{N} + \tilde{C}_2\epsilon_{\text{MLE}}\right)\,. \tag{53}$$

E.5 Useful technical result

Lemma 3 (Lemma 5.3, Zhang et al. (2021))**.** *Let Assumptions 3 and 4 hold. Then, the following statements hold:*

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

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

F Future Work

We comment here on a few future directions of improvement:

- In our PG algorithm, the estimations of the state occupancy measure need to be relearned for each policy parameter θ_t. We believe a regularized policy optimization approach could lead to a more efficient procedure. Indeed, by enforcing policy parameters to be not too far from each other, it would allow to reuse estimations of the occupancy measure from previous iterations to obtain better and more reliable estimations.

https://doi.org/10.52202/085713-2661

- The state-occupancy measure can be very complicated and hence difficult to estimate, especially in complex high-dimensional state settings. The use of massively overparametrized neural networks for occupancy measure approximation might therefore be of much help in such complex settings as practice shows that overparametrized neural networks do perform well in general. Establishing theoretical guarantees in this regime is certainly an interesting question to extend our work.
- It would definitely be interesting to conduct experiments in very large scale environments such as DMLab or Atari. Our work makes progress towards solving larger scale real-world RLGU problems and offers a promising approach supported by theoretical guarantees.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide clear statements in line with the abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of some of our assumptions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

https://doi.org/10.52202/085713-2661

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide complete proofs in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

https://doi.org/10.52202/085713-2661

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

 https://doi.org/10.52202/085713-2661

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

https://doi.org/10.52202/085713-2661