Towards Faster Global Convergence of Robust Policy Gradient Methods

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, global convergence has been achieved for non-robust MDPs with an iteration complexity of $O(\frac{1}{\epsilon})$ for finding an ϵ-optimal policy, for which PL condition derived from performance difference lemma has played a key role. This work extends performance difference lemma to s-rectangular robust MDPs from which PL condition can be derived. We further, present a simplified proof for the policy gradient convergence for non-robust case, which together with robust performance difference lemma, can lead to global convergence of robust policy gradient.

1 Introduction

In sequential decision-making problems, Markov decision processes (MDPs) provide an analytical framework for learning a policy that performs best in a fixed environment. However, given that optimal policies can be highly sensitive to the parameter values [16], the robust MDP setting alternatively seeks strategies that are robust to uncertain environments [19, 6]. It quantifies the level of uncertainty through a set determining the possible range of model perturbations. Then, a robust policy is optimal if it reaches maximal performance under the worst model parameters over the uncertainty set. Developing algorithms that efficiently solve robust MDPs is of great interest, as these can yield better generalization guarantees [33].

Without some structural assumptions on the uncertainty set, solving robust MDPs can be NP-hard [31]. Therefore, to preserve tractability, we often assume that the uncertainty set is convex and s-rectangular, i.e., it can be expressed as a Cartesian product over states [19, 6, 31, 4, 13, 28]. In that case, standard solvers for MDPs carry over to robust MDPs. Further simplification may consider (s, a)-rectangular uncertainty sets, i.e., independent uncertainty over each state-action pair, but this can lead to more conservative strategies. In fact, maintaining the problem tractable while relaxing the uncertainty set structure may be of interest when seeking less wary robust solutions [5, 15].

Policy gradient (PG) methods have been proven workhorse in reinforcement learning (RL) that is being in many variants [25, 23, 11, 9]. Recently, global convergence of PG methods have been established [1, 32, 17], crucially exploiting the PL condition [10] type property of non-robust MDPs. This PL condition is derived from performance difference lemma [8] that expresses the difference of values function w.r.t. any two policies with difference in polices, occupation measure of one policy, and Q-value of other policy. This techniques achieves a global convergence with state-of-the-art iteration complexity of $O(SA\epsilon^{-1})$ [32]. Unfortunately, these techniques can't be directly applied to robust policy gradient due to following main reasons: a) The robust MDPs can be non-smooth as compared to non-robust MDPs. b) Many structural properties doesn't carry over to robust MDPs from non-robust MDPs. In addition, the proof techniques are overly complicated. Nonetheless, there have been recent developments of robust policy gradient methods [29, 12, 27], that enjoys global convergence properties. Precisely, [27] established global convergence of robust policy gradient methods for general uncertainty set, utilizing techniques from game theory. However, the approach

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

has a much more expensive iteration complexity of $O(S^4A^2\epsilon^{-4})$. Further, [29], demonstrates a much faster global convergence with iteration complexity of $O(SA\epsilon^{-3})$ using the smoothing techniques and establishing performance difference lemma for sa-rectangular robust MDPs. However, the analysis is tailor-made for sa-rectangular R-contamination uncertainty set, which crucially relies on the simplicity of the regularizer term arising from robustness [28]. Hence, this technique can't be applied for more general uncertainty sets.

In this work, we make following contributions:

- We extend the performance difference lemma to s-rectangular robust MDPs, leading to PL type condition.
- We provide a simplified and intuitive proof for global convergence rate for non-robust MDPs. This together with PL type condition and smoothness (robust MDPs may be smooth for some special type of uncertainty sets) yields global convergence rate for robust MDPs with similar rate as non-robust MDPs.

Related Work

Table 1: Iteration Complexity for Global Convergence of Policy Gradient Methods

Robust MDPs	Complexity	Remark
Non-Robust	$O(SA\epsilon^{-1})$	[32]
(s,a) rectangular R-Contamination	$O(SA\epsilon^{-3})$	[29]
General s-rectangular	$O(SA\epsilon^{-1})$	Ours
General	$O(S^4A^2\epsilon^{-4})$	[27]

Non-Robust MDPs. Policy gradient is derived in [25] for non-robust MDPs which is widely used in practice with many variants [23, 11, 24]. Recently, there have been global convergences guarantees results [1, 3] with an iteration complexity $O(1/\epsilon)$ for finding ϵ-optimal policy [32].

(s,a)-rectangular R-Contamination Robust MDPs. The paper [29] derives policy gradient for R-rectangular robust MDPs complexity $O(S^2A\log(\frac{1}{\epsilon}))$ similar to non-robust MDPs. Further, it establishes global convergence policy gradient with an iteration complexity $O(1/\epsilon^3)$ for finding ϵ-optimal policy assuming oracle policy gradient.

General (s,a)-rectangular Robust MDPs The paper [14] establishes global convergence for robust mirror policy decent for (s,a)-rectangular robust MDPs in general with an iteration complexity $O(1/\epsilon)$ and $O(log(1/\epsilon))$ for finding ϵ-optimal policy, with two increasing-stepsize schemes. However, it assumes the oracle access to policy gradient.

General Robust MDPs The paper [27] establishes global convergence for Double-Loop Robust Policy Gradient for general robust MDPs with an iteration complexity $O(1/\epsilon^4)$ for finding ϵ-optimal policy, assuming the oracle access to policy gradient. Solving the policy gradient upto ϵ tolerance via value methods that takes (s,a)-rectangular and s-rectangular case with complexity of $O(S^4A\log(1/\epsilon))$ and $O(S^4A^3\log(1/\epsilon))$ respectively using convex optimizations tools. Our techniques are completely different than this work.

2 Preliminaries

Notation: We denote the cardinal of an arbitrary finite set $\mathcal{Z}$ by $|\mathcal{Z}|$. Given two real functions $\mathbf{a},\mathbf{b}:\mathcal{Z}\to\mathbb{R}$, their inner product is $\langle\mathbf{a},\mathbf{b}\rangle_{\mathcal{Z}} := \sum_{z\in\mathcal{Z}}\mathbf{a}(z)\,\mathbf{b}(z)$, which induces the L_2-norm $\|\mathbf{a}\|_2 := \sqrt{\langle\mathbf{a},\mathbf{a}\rangle_{\mathcal{Z}}}$. More generally, for any $p\in[1,\infty]$, the L_p-norm of $\mathbf{a}$ is $\|\mathbf{a}\|_p := (\sum_{z\in\mathcal{Z}}|\mathbf{a}(z)|^p)^{\frac{1}{p}}$. Its conjugate norm satisfies $\|\mathbf{a}\|_q = \max_{\|\mathbf{b}\|\le 1}\langle\mathbf{a},\mathbf{b}\rangle$, where q is the conjugate value of p, that is , $\frac{1}{q} = 1-\frac{1}{p}$. The probability simplex over $\mathcal{Z}$ is denoted by $\Delta_{\mathcal{Z}} := \{\mathbf{a}:\mathcal{Z}\to\mathbb{R}_+ \,|\, \sum_{z\in\mathcal{Z}}\mathbf{a}(z)=1\}$ and $\mathbf{0}$ (resp. $\mathbf{1}$) is the vector of all zeros (resp. all ones) with appropriate dimensions. Finally, $I_{\mathcal{Z}}$ designates the identity matrix in $\mathbb{R}^{\mathcal{Z}\times\mathcal{Z}}$.

2.1 Markov Decision Processes

A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \gamma, \mu, P, R)$ such that $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces respectively, $\gamma \in [0, 1)$ is a discount factor and $\mu \in \Delta_{\mathcal{S}}$ the initial state distribution. Denoting $\mathcal{X} := \mathcal{S} \times \mathcal{A}$, the couple (P, R) corresponds to the MDP model with $P : \mathcal{X} \to \Delta_{\mathcal{S}}$ being a transition kernel and $R : \mathcal{X} \to \mathbb{R}$ a reward function. A policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ maps each state to a probability distribution over $\mathcal{A}$, and we denote by Π the set of such functions. For any policy $\pi \in \Pi$, $R^\pi \in \mathbb{R}^{\mathcal{S}}$ is the expected immediate reward defined as $R^\pi(s) := \langle \pi_s, R(s, \cdot)\rangle_{\mathcal{A}}, \quad \forall s \in \mathcal{S}$, where π_s is a shorthand for $\pi(\cdot|s)$. We similarly define the stochastic matrix induced by π as $P^\pi(s'|s) := \langle \pi_s, P(s'|s, \cdot)\rangle_{\mathcal{A}}, \quad \forall s, s' \in \mathcal{S}$. Our goal is to maximize the discounted return over the set of policies Π:

$$\rho^\pi_{(P,R)} := \mathbf{E}\left[\sum_{n=0}^{\infty} \gamma^n R(s_t, a_t) \mid \pi, P, s_0 \sim \mu\right]. \tag{1}$$

The above return can be rewritten as [20]

$$\rho^\pi_{(P,R)} = \langle \mu, v^\pi_{(P,R)}\rangle = \langle R, d^\pi_{P,\mu}\rangle,$$

where $v^\pi_{(P,R)}$ is value function defined as

$$v^\pi_{(P,R)}(s) := \mathbf{E}\left[\sum_{n=0}^{\infty} \gamma^n R(s_n, a_n) \mid \pi, P, s_0 = s\right]$$

and $d^\pi_{P,k} \in \mathbb{R}^{\mathcal{S}}$ is occupation measure defined as

$$d^\pi_{P,k} := k^T (I - \gamma P^\pi)^{-1}, \qquad \forall k \in \mathbb{R}^{\mathcal{S}}.$$

Remark 1. *Generally we take $k \in \Delta^{\mathcal{S}}$, we extend this definition for later use in the robust MDPs.*

We can obtain the optimal policy $\pi^*_{(P,R)}$, which is a maximizer of (1), via a policy gradient method. The policy gradient is given by [26]

$$\nabla \rho^\pi_{(P,R)} = \sum_{s,a} d^\pi_{P,\mu}(s) Q^\pi_{(P,R)}(s, a) \nabla \pi(a|s),$$

where the Q-value is defined as

$$Q^\pi_{(P,R)}(s, a) := R(s, a) + \gamma \sum_{s',a} P(s'|s, a) v^\pi_{(P,R)}(s').$$

The value function $v^\pi_{(P,R)}$ can be computed via value iteration. Given a policy $\pi \in \Pi$, the evaluation Bellman operator is given by

$$T^\pi_{(P,R)} v = R^\pi + \gamma P^\pi v, \quad \forall v \in \mathbb{R}^{\mathcal{S}},$$

which is a contraction whose unique fixed point is $v^\pi_{(P,R)}$.

2.2 Robust Markov Decision Processes

Generally, the system's dynamics may be unknown or partially known. Thus, if the agent does not account for model uncertainties during training, its performance can significantly drop after deployment while testing [16]. The robust MDP framework addresses this issue by assuming that $(P, r) \in \mathcal{U}$ where $\mathcal{U}$ is an compact uncertainty set, and by aiming to maximize return under the worst-case model. As standard in the robust RL literature, we assume $\mathcal{U}$ to be compact and convex, so that a worst-case model exists and can be computed in polynomial time [30].

The *robust* performance of a policy $\pi \in \Pi$ is defined as

$$\rho^\pi_{\mathcal{U}} := \min_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)},$$

and the robust optimal return

$$\rho^*_{\mathcal{U}} := \max_{\pi\in\Pi} \rho^\pi_{\mathcal{U}}$$

is attained at $\pi^*_{\mathcal{U}} \in \arg\max_{\pi\in\Pi} \rho^\pi_{\mathcal{U}}$ [18, 7, 31]. Let $(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})$ be worst values associated with policy π, defined as

$$(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}}) :\in \operatorname*{arg\,inf}_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)}.$$

The robust value function, robust Q-value and robust occupation measure, can be defined as

$$v^\pi_{\mathcal{U}} = v^\pi_{(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})}, \qquad Q^\pi_{\mathcal{U}} = Q^\pi_{(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})}, \quad \text{and} \quad d^\pi_{\mathcal{U}} = d^\pi_{(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})},$$

for all policy π [12].

However, solving robust MDPs with general convex uncertainty sets is known to be strongly NP-hard, and optimal policies can be stochastic as well as history-dependent [30]. Moreover, no meaningful robust Bellman operators exist for general uncertainty sets that can give rise to value iteration methods.

2.2.1 Robust Gradient Method

To optimize the robust return, we can rely on the projected gradient ascent rule:

$$\pi_{k+1} :=\mathbf{proj}_\Pi(\pi_k + \eta\nabla_\pi \rho^{\pi_k}_{\mathcal{U}}), \tag{2}$$

where η is the learning rate and $\mathbf{proj}_\Pi$ denotes the orthogonal projection on Π. The gradient $\nabla_\pi \rho^{\pi_k}_{\mathcal{U}}$ of the robust return may not exist due to non-differentiability. However, sub-differential can be defined as

$$\partial_\pi \rho^\pi_{\mathcal{U}} := \nabla_\pi \rho^\pi_{(P,R)} \Big|_{(P,R)=(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})} . \tag{3}$$

Given the oracle access to sub-gradient $\partial\rho^\pi_{\mathcal{U}}$, the projected gradient ascent

$$\pi_{k+1} :=\mathbf{proj}_\Pi(\pi_k + \eta\partial_\pi \rho^{\pi_k}_{\mathcal{U}}), \tag{4}$$

converges to an ϵ-optimal policy $\pi^*_{\mathcal{U}}$, with iteration complexity $O(S^4A^2\epsilon^{-4})$, under similar conditions to the non-robust setting [27]. Unfortunately, this approach is generally not applicable since the computation of the gradient of the robust policy $\nabla\rho^\pi_{\mathcal{U}}$ is generally intractable; and the latter is due to the NP Hardness of robust MDPs for general convex uncertainty sets.

2.2.2 Rectangular Uncertainty set

To overcome intractability, the uncertainty set has to be both convex and $\mathtt{s}$-rectangula,r which allows the optimal policies to be stationary, even though stochastic [31]. Uncertainty set $\mathcal{U}^{\mathtt{s}}$ is called $\mathtt{s}$-rectangular if it can decomposed over states, that is

$$\mathcal{U}^{\mathtt{s}} =\big(\times_{s\in\mathcal{S}}\mathcal{P}_s \big) \times \big(\times_{s\in\mathcal{S}}\mathcal{R}_s \big) .$$

Further simplification may take $\mathtt{sa}$-rectangular uncertainty sets, namely [6, 19]

$$\mathcal{U}^{\mathtt{sa}} =\big(\times_{(s,a)\in\mathcal{X}}\mathcal{P}_{s,a} \big) \times \big(\times_{(s,a)\in\mathcal{X}}\mathcal{R}_{s,a} \big) .$$

$\mathtt{sa}$-rectangular robust MDPs are much more conservative than $\mathtt{s}$-rectangular robust MDPs, and admit deterministic optimal robust policy [7, 18, 31]. Under this rectangularity assumption, the robust value function $v^\pi_{\mathcal{U}}$ is well defined as

$$v^\pi_{\mathcal{U}} := \min_{(P,R)\in\mathcal{U}} v^\pi_{(P,R)},$$

which is unique fixed point of the γ-contractive robust Bellman operator $T^\pi_{\mathcal{U}}$ well defined as

$$T^\pi_{\mathcal{U}} v := \min_{(P,R)\in\mathcal{U}} T^\pi_{(P,R)} v, \qquad \forall v \in \mathbb{R}^{\mathcal{S}},$$

and it also allows the optimal robust value function $v^*_{\mathcal{U}}$ to be well defined as

$$v^*_{\mathcal{U}} := \max_\pi v^\pi_{\mathcal{U}},$$

which is the unique fixed point of the γ-contractive optimal robust Bellman operator $T^*_{\mathcal{U}}$ well defined as

$$T^*_{\mathcal{U}} v := \max_\pi T^\pi_{\mathcal{U}} v, \qquad \forall v \in \mathbb{R}^{\mathcal{S}}$$

[31]. Since it is a contraction map, robust value iteration, $v_{n+1} := T^*_{\mathcal{U}} v_n$, converges to the optimal robust value function $v^*_{\mathcal{U}}$ linearly, making robust value iteration an attractive approach. Once the robust value function is obtained, the optimal robust policy can be computed as

$$\pi^*_{\mathcal{U}} \in \arg\max_\pi T^\pi_{\mathcal{U}} v^*_{\mathcal{U}}$$

[31]. However, the evaluation of each Bellman operator can still be prohibitive for practical use.

3 Main

In this section, we outline the simplified proof techniques used to proof global convergence rate of non-robust policy gradient that is presented in [27]. This can used to prove global convergence rate for robust MDPs too.

Assumption 1. *We assume the set of policies* Π *and set of uncertainty set* $\mathcal{U}$ *are convex and compact.*

The above assumption is very mild that is satisfied in most of the settings.

The technique has two main parts which when combined with a cohesive bond, yields into the desired result. We begin with the first assumption below.

Assumption 2. *The function* $\rho_{\mathcal{U}}^{\pi}$ *is differentiable.*

The assumption doesn't hold for general uncertainty set, however we believe it holds for many useful uncertainty sets. Observe that we do not require the function $\rho_{\mathcal{U}}^{\pi}$ to be convex.

Proposition 1. *The derivative of robust return* $\nabla\rho_{\mathcal{U}}^{\pi}$ *is Lipschitz in policy with Lipschitz constant* L *.*

Proof. Recall, the set of policies Π is convex, and the derivative $\nabla\rho_{\mathcal{U}}^{\pi}$ is assumed to exist, hence the following exists $L = \max_{\pi\in\Pi}\|\nabla\rho_{\mathcal{U}}^{\pi}\|$. Moreover, L is the desired Lipschitz constant[22]. □

We consider the learning rule

$$\pi_{k+1} = \pi_k + \eta\nabla\rho_{\mathcal{U}}^{\pi_k}.$$

Let T (resp. G) be the next step gradient step operator (resp. the effective gradient step operator) defined as

$$T(\pi) := \mathbf{proj}_{\Pi}(\pi + \frac{1}{L}\nabla\rho_{\mathcal{U}}^{\pi}), \tag{5}$$

$$G(\pi) := L(T(\pi) - \pi) \tag{6}$$

Lemma below states that the assumption 2 ensures the a minimum fixed improvement on the gradient ascent given the right step size. Further, the improvement is lower bounded by the norm of the 'effective gradient' times some constant.

Lemma 1. ***[Sufficient Increase Lemma]*** *Gradient ascent ensures the monotone improvement in the robust return. Precisely,*

$$\rho_{\mathcal{U}}^{\pi_{k+1}} - \rho_{\mathcal{U}}^{\pi_k} \geq \frac{1}{2L}\|G(\pi_k)\|^2 = \frac{L}{2}\|\pi_{k+1} - \pi_k\|^2, \qquad \forall k.$$

Proof. Proved in the appendix. It just uses convexity of the projection set Π, differentiability and smoothness of the objective function. □

Note that the above lemma is enough to ensure iterates $\{\rho_{\mathcal{U}}^{\pi_k}\}$ converge to some some value $\hat{\rho}$, as the iterates forms monotonically increasing sequence. However, it doesn't imply the $\hat{\rho}$ is global maxima or local maxima for that matter. This just implies, the iterates $\rho_{\mathcal{U}}^{\pi_k}$ keeps on increasing until the gradient $G(\pi_k)$ doesn't diminish to zero.

Hence, for the global optimality, we need second part, to ensure that the norm of the gradient vanishes only when the sub-optimality vanishes. In order to do so, we first extend performance difference lemma to $\mathtt{s}$-rectangular case, as stated below.

Lemma 2 (Robust Performance Difference). *For* $\mathtt{s}$*-rectangular uncertainty set* $\mathcal{U}$*, and for any policies* $\pi_1, \pi_2 \in \Pi$*, the difference in robust value is bounded as*

$$\rho_{\mathcal{U}}^{\pi_2} - \rho_{\mathcal{U}}^{\pi_1} \leq \sum_{(s,a)\in\mathcal{X}} d_{P_{\mathcal{U}}^{\pi_1}}^{\pi_2}(s)\ \big(\ \pi_2(a|s) - \pi_1(a|s)\ \big)\ Q_{\mathcal{U}}^{\pi_1}(s,a).$$

Proof. Let two arbitrary policies $\pi_1, \pi_2 \in \Pi$. Denote by $(P_1, r_1) := (P_{\mathcal{U}}^{\pi_1}, r_{\mathcal{U}}^{\pi_1}), (P_2, r_2) := (P_{\mathcal{U}}^{\pi_2}, r_{\mathcal{U}}^{\pi_2})$ their respective worst kernels (the worst values exist, because $\mathcal{U}$ is assumed to be

compact). We now proceed similarly as in [14, 1]. Since $v^{\pi_2}_{(P_2,r_2)}$ is the unique fixed point of the robust Bellman operator, it holds that

$$v^{\pi_2}_{(P_2,r_2)} = \min_{(P,r)\in\mathcal{U}} T^{\pi_2}_{(P,r)} v^{\pi_2}_{\mathcal{U}} \leq T^{\pi_2}_{(P_1,r_1)} v^{\pi_2}_{\mathcal{U}}.$$

On the other hand, by definition of (P_1, r_1), we have $v^{\pi_1}_{\mathcal{U}} = T^{\pi_1}_{(P_1,r_1)} v^{\pi_1}_{\mathcal{U}}$. Therefore,

$$\begin{aligned} v^{\pi_2}_{\mathcal{U}} - v^{\pi_1}_{\mathcal{U}} &\leq T^{\pi_2}_{(P_1,r_1)} v^{\pi_2}_{\mathcal{U}} - T^{\pi_1}_{(P_1,r_1)} v^{\pi_1}_{\mathcal{U}} \\ &= r^{\pi_2}_1 + \gamma P^{\pi_2}_1 v^{\pi_2}_{\mathcal{U}} - r^{\pi_1}_1 - \gamma P^{\pi_1}_1 v^{\pi_1}_{\mathcal{U}} \\ &= r^{\pi_2}_1 - r^{\pi_1}_1 + \gamma \underbrace{(P^{\pi_2}_1 v^{\pi_2}_{\mathcal{U}} - P^{\pi_1}_1 v^{\pi_1}_{\mathcal{U}})}_{(1)}. \end{aligned}$$

Using the identity $a_2 b_2 - a_1 b_1 = a_2(b_2 - b_1) + (a_2 - a_1) b_1$ on expression (1), it results that:

$$\begin{aligned} v^{\pi_2}_{\mathcal{U}} - v^{\pi_1}_{\mathcal{U}} \leq & r^{\pi_2}_1 - r^{\pi_1}_1 + \gamma P^{\pi_2}_1 \left(v^{\pi_2}_{\mathcal{U}} - v^{\pi_1}_{\mathcal{U}} \right) + \gamma \left(P^{\pi_2}_1 - P^{\pi_1}_1 \right) v^{\pi_1}_{\mathcal{U}} \\ = & r^{\pi_2 - \pi_1}_1 + \gamma \left(P^{\pi_2 - \pi_1}_1 \right) v^{\pi_1}_{\mathcal{U}} + \gamma P^{\pi_2}_1 \left(v^{\pi_2}_{\mathcal{U}} - v^{\pi_1}_{\mathcal{U}} \right) \\ = & (\pi_2 - \pi_1)[r_1 + \gamma P_1 v^{\pi_1}_{\mathcal{U}}] + \gamma P^{\pi_2}_1 \left(v^{\pi_2}_{\mathcal{U}} - v^{\pi_1}_{\mathcal{U}} \right) \end{aligned}$$

This implies,

$$\begin{aligned} & v^{\pi_2}_{\mathcal{U}}(s) - v^{\pi_1}_{\mathcal{U}}(s) \leq \sum_{s',a'} d^{\pi_2}_{P^{\pi_1}_{\mathcal{U}},s}(s')[R^{\pi_1}_{\mathcal{U}}(s',a') + \gamma \sum_{s"} P^{\pi_1}_{\mathcal{U}}(s"|s',a') v^{\pi_1}_{\mathcal{U}}(s")](\pi_2(a'|s') - \pi_1(a'|s')) \\ \implies & \rho^{\pi_2}_{\mathcal{U}} - \rho^{\pi_1}_{\mathcal{U}} \leq \sum_{s,a} d^{\pi_2}_{P^{\pi_1}_{\mathcal{U}}}(s)[R^{\pi_1}_{\mathcal{U}}(s,a) + \gamma \sum_{s'} P^{\pi_1}_{\mathcal{U}}(s'|s,a) v^{\pi_1}_{\mathcal{U}}(s')][\pi_2(a|s) - \pi_1(a|s)]. \end{aligned}$$

□

The performance difference lemma bounds the the difference in the robust return of two policies by using only robust Q-value of one policy. This can be used to bound the sub-optimality of a policy with in terms of robust Q-value of the policy which in turn can be related to its policy gradient. This notion is formalized next.

The next step consists in bounding the sub-optimality of any policy according to the gradient, as we do next. Leveraging Lemma 2, the domination lemma below upper bounds the performance gap by variational gradients.

Lemma 3 (**PL condition / Gradient Domination lemma**). *For any policy $\pi \in \Pi$, its sub-optimality is bounded by its policy gradient as*

$$\rho^*_{\mathcal{U}} - \rho^{\pi}_{\mathcal{U}} \leq C_{\text{PL}} \max_{\pi' \in \Pi} \langle \pi' - \pi, \nabla \rho^{\pi}_{\mathcal{U}} \rangle,$$

where $C_{\text{PL}} := \max_{(\pi,s)\in\Pi\times\mathcal{S}} \frac{d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s)}{d^{\pi}_{\mathcal{U}}(s)}$.

Proof. From performance difference lemma, we have

$$\rho^*_{\mathcal{U}} - \rho^{\pi}_{\mathcal{U}} \leq \sum_s d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s) \sum_a \left(\pi^*_{\mathcal{U}}(a|s) - \pi(a|s) \right) Q^{\pi}_{\mathcal{U}}(s,a) \tag{7}$$

$$\leq \max_{\pi'} \sum_s d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s) \underbrace{\sum_a \left(\pi'(a|s) - \pi(a|s) \right) Q^{\pi}_{\mathcal{U}}(s,a)}_{\geq 0} \tag{8}$$

$$= \max_{\pi'} \sum_s \frac{d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s)}{d^{\pi}_{\mathcal{U}}(s)} d^{\pi}_{\mathcal{U}}(s) \underbrace{\sum_a \left(\pi'(a|s) - \pi(a|s) \right) Q^{\pi}_{\mathcal{U}}(s,a)}_{\geq 0} \tag{9}$$

$$\leq \left(\max_s \frac{d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s)}{d^{\pi}_{\mathcal{U}}(s)} \right) \max_{\pi'} \sum_s d^{\pi}_{\mathcal{U}}(s) \sum_a \left(\pi'(a|s) - \pi(a|s) \right) Q^{\pi}_{\mathcal{U}}(s,a) \tag{10}$$

$$= \left(\max_s \frac{d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s)}{d^{\pi}_{\mathcal{U}}(s)} \right) \max_{\pi'} \left\langle \pi' - \pi, \nabla \rho^{\pi}_{\mathcal{U}} \right\rangle . \tag{11}$$

$$\tag{12}$$

It holds for all $(P, R) \in \arg\min_{P,R} \rho^{\pi}_{\mathcal{U}}$, hence it also holds for minimum sub-gradient. □

Note that the constant C_{PL} is bounded constant for $\mu \succ 0$, same as non-robust counterpart [1]. The distributional mismatch constant C_{PL}, we obtain here is very similar to the one prevailing in non-robust MDPs, where it is also known as the Polyak-Łojasiewicz constant [1]. However, our mismatch constant is calculated w.r.t. the worst transition kernel whereas in the non-robust case, it corresponds to the nominal.

For the sake of intuition, assume the domain Π is a unit ball around π, then right hand sides becomes $C_{PL}\|\partial \rho^{\pi}_{\mathcal{U}}\|$. This intuitively shows why the above assumption is called PL condition.

Now, we have both the parts: One that lower bounds the gradient and the other that upper bounds it. However, they are not exactly in very compatible forms, hence we require the result below that acts a cohesive bond between the two.

Lemma 4. *(Cohesive Bond) For all $\pi \in \Pi$, we have*

$$\max_{\pi' \in \Pi} \langle \nabla \rho^{\pi_{k+1}}_{\mathcal{U}}, \pi' - \pi_{k+1} \rangle \leq 2\|G(\pi_k)\| \mathbf{diam}(\Pi),$$

where $\mathbf{diam}(\Pi) := \max_{\pi, \pi' \in \Pi} \|\pi - \pi'\|$ is the diameter of Π.

Proof. Proved in the appendix, however it is also a consequence of the second projection theorem [2][Thm. 9.8]. □

Now equating the effective gradient in Lemma 3 and Lemma 1 using the Lemma 4 as intermediary, we get the sub-optimality recursion below.

Theorem 1. *Take $\eta = \frac{1}{L}$ as a learning rate. Then, the scaled sub-optimality $a_k = \frac{\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}}}{8LC^2_{PL}\mathbf{diam}(\Pi)^2}$ follows the recursion*

$$a^2_{k+1} + a_{k+1} - a_k \leq 0.$$

Proof. From the PL condition proved in Lemma 3, we have

$$\rho^*_{\mathcal{U}} - \rho^{\pi_{k+1}}_{\mathcal{U}} \leq C_{PL} \max_{\pi'} \langle \pi' - \pi_{k+1}, \nabla \rho^{\pi_{k+1}}_{\mathcal{U}} \rangle \tag{13}$$

$$\leq 2\|G(\pi_k)\| \mathbf{diam}(\Pi), \qquad \text{(from Lemma 4)} \tag{14}$$

$$\leq C_{\text{PL}} \cdot 2\sqrt{2L(\rho^{\pi_{k+1}}_{\pi} - \rho^{\pi_k}_{\mathcal{U}})} \cdot \mathbf{diam}(\Pi), \qquad \text{(from Lemma 1)} \tag{15}$$

Squaring both sides and adding subtracting ρ^* in RHS, we get

$$\left(\rho^*_{\mathcal{U}} - \rho^{\pi_{k+1}}_{\mathcal{U}} \right)^2 \leq 8C^2_{\text{PL}} L \mathbf{diam}(\Pi)^2 \left((\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}}) + (\rho^*_{\mathcal{U}} - \rho^{\pi_{k+1}}_{\pi}) \right)$$

Setting $a_k := \frac{\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}}}{8LC^2_{\text{PL}}\mathbf{diam}(\Pi)^2}$, the sequence $(a_k)_{k\in\mathbb{N}}$ satisfies the recursion $a^2_{k+1} \leq a_k - a_{k+1}$. $\square$

The sub-optimality recursion derived in the theorem above, illustrates how the sub-optimality at time $k+1$ depends at the sub-optimality at time k. Moreover, the sub-optimality recursion has the quadratic form and $a_k \geq 0$, hence its solution is given as

$$a_{k+1} \leq \sqrt{\frac{1}{4} + a_k} - \frac{1}{2}.$$

As a sanity check, we observe that $\sqrt{\frac{1}{4} + a} - \frac{1}{2} \leq a$ for all $a \geq 0$, implying that $(a_k)_{k\in\mathbb{N}}$ is monotonically decreasing. Further, 0 is the only non-negative fixed point of the $\sqrt{\frac{1}{4} + a} - \frac{1}{2} = a$ implying that $(a_k)_{k\in\mathbb{N}}$ monotonically decreases to 0.

Now, we investigate the convergence rate for a_k. Observe that if $a_0, a_k \gg 1$, then $a_{k+1} \approx \sqrt{a_k}$ and $a_k \approx (a_0)^{\frac{1}{2^k}}$. That is, the convergence rate is super-exponential! Yet, in most cases, $8LC^2_{\text{PL}}\mathbf{diam}(\Pi)^2 \gg 1$ and $\rho^*_{\mathcal{U}} - \rho^{\pi_0}_{\mathcal{U}} = 8LC^2_{\text{PL}}\mathbf{diam}(\Pi)^2 a_0 = O(1)$ is bounded so we are more interested in the case where $a_0 \ll 1$. In fact, in an MDP with a reward smaller than 1, we do have $\rho^{\pi}_{\mathcal{U}} = O(1)$.

In this regime, the sub-optimality recursion $a_{k+1} - a_k \leq -a^2_{k+1}$ suggests the ordinary differential equation $\frac{da}{dk} \leq -a^2$ whose solution is $a(k) \leq \frac{1}{k + \frac{1}{a(0)}} \leq \frac{1}{k}$. This intuitively indicates an $O(\frac{1}{\epsilon})$ iteration complexity for achieving an ϵ-optimal solution, which we state below formally.

Corollary 1 (**Global optimality**). *For all iterations $k \geq 1$, it holds that:*

$$\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}} \leq \max\left(\frac{8LC^2_{\text{PL}}\mathbf{diam}(\Pi)^2}{k}, 2^{-\frac{k}{2}}\right)(\rho^*_{\mathcal{U}} - \rho^{\pi_0}_{\mathcal{U}}).$$

Proof. The sub-optimality recursion yields the desired result which follows directly from [32]. $\square$

4 Non-Differentiable Case

This section is devoted to studying the convergence without the differentiability, as in many case, differentiability assumption 2 may not satisfy. Non-differentiability restricts us from using 'sufficient increase lemma'. This lemma was crucially used in the section above that guarantees some improvement proportional to the 'effective gradient'.

We our study below indicates that it is possible to use 'PL-condition' to ensure improvement while gradient ascent, however it accounts for slower convergence rate of $O(\epsilon^{-2})$ compared to $O(\epsilon^{-1})$ in differentiable case.

We easily get the non-differential version of PL condition, as stated in the result below.

Lemma 5 (**PL condition**). *For any policy $\pi \in \Pi$, it holds that*

$$\rho^*_{\mathcal{U}} - \rho^{\pi}_{\mathcal{U}} \leq C_{\text{PL}} \min_{\partial} \max_{\pi' \in \Pi} \langle \pi' - \pi, \partial\rho^{\pi}_{\mathcal{U}} \rangle,$$

where $C_{\text{PL}} := \max_{(\pi,s)\in\Pi\times\mathcal{S}} \frac{d^{\pi^*_{\mathcal{U}}}_{P^{\pi}_{\mathcal{U}}}(s)}{d^{\pi}_{\mathcal{U}}(s)}$.

The above condition, immediately implies that the robust return has no saddle points, that is, global maxima is the only point where sub-differential is zero.

Theorem 2. *(No Saddle Points) The robust MDPs have no saddle points, that is, zero sub-gradient implies global optima. In other words,*

$$\partial\rho^{\pi}_{\mathcal{U}} = 0 \implies \rho^{\pi}_{\mathcal{U}} = \rho^*_{\mathcal{U}}.$$

Proof. From the above PL condition, we have

$$\rho^*_{\mathcal{U}} - \rho^{\pi}_{\mathcal{U}} \leq C_{\text{PL}} \min_{\partial} \max_{\pi' \in \Pi} \langle \pi' - \pi, \partial\rho^{\pi}_{\mathcal{U}} \rangle.$$

If we have $\partial\rho^{\pi}_{\mathcal{U}} = 0$, that implies $\rho^*_{\mathcal{U}} - \rho^{\pi}_{\mathcal{U}} \leq 0$, which in turn, implies $\rho^*_{\mathcal{U}} = \rho^{\pi}_{\mathcal{U}}$. $\square$

In the next subsection, we show how the above PL condition can be used for convergence in one-dimensional case. Extending this proof technique to a full fledged multi-dimensional case, remains for the future work.

4.1 Illustrative case of Single Dimension

This section is devoted for showcasing the potential of this technique in one-dimensional case, without projection. That is, we consider the update rule:

$$x_{k+1} = x_k + \eta_t \partial f(x_k).$$

For $x \in X \subset \mathbb{R}, f : X \to \mathbb{R}$ and with domain set X being convex and compact, lets assume equivalent PL condition as

$$f(x^*) - f(x) \leq c \min_{\partial} \max_{x'} \langle x' - x, \partial f(x) \rangle \leq c \min_{\partial} diam(X) |\partial f(x)|.$$

Now, WLOG lets assume $\partial f(x_0) \geq 0$. Further, note that robust return $\rho^{\pi}_{\mathcal{U}}$ is Lipschitz (proved in appendix). So, we also assume the function $f(x)$ is Lipschitz with Lipschitz constant L.

Proposition 2. *Assuming* $\min_{\partial} \partial f(x_i) > 0$ *for all* $0 \leq i \leq k$*, then either* $\min_{\partial} \partial f(x_{k+1}) > 0$ *or* $f(x_k) \leq \eta_k$.

Proof. If $\min_{\partial} \partial f(x_{k+1}) \leq 0$ and we already have $\min_{\partial} \partial f(x_k) > 0$, this implies the existence of a local minima between x_k and x_{k+1}. But result above have proved that there is no-saddle point. Hence, a global optima $f(x*) = \max_{x \in X} f(x)$ must have been achieved between x_k and x_{k+1}. Further, since the function f is Lipschitz, and $|x_{k+1} - x^*| \leq |x_{k+1} - x_k| \leq \eta_k$ This implies $f^* - f(x_{k+1}) \leq L\eta_k$. □

Now, assume $\min_{\partial} \partial f(x_i) > 0$ for all $0 \leq i \leq k+1$, and from standard calculus [21], we have

$$\begin{aligned}
f(x_{k+1}) - f(x_k) = \int_{x_k}^{x_{k+1}} \partial f(x) dx &\leq (x_{k+1} - x_k) \min_{x \in [x_k, x_{k+1}]} \min_{\partial} |\partial f(x)|, \\
&\geq (x_{k+1} - x_k) \min_{x \in [x_k, x_{k+1}]} \frac{f(x^*) - f(x)}{c \cdot diam(X)}, \qquad \text{(from PL condition)}, \\
&\geq (x_{k+1} - x_k) \frac{f(x^*) - f(x_{k+1})}{c \cdot diam(X)}, \qquad \text{(f is increasing in } [x_0, x_{k+1}]), \\
&\geq \eta_k \frac{f(x^*) - f(x_{k+1})}{c \cdot diam(X)}, \qquad \text{(learning rule)},
\end{aligned}$$

As proved in appendix, the above recursion implies

$$f(x^*) - f(x_{k+1}) \leq c \cdot diam(X) \frac{f(x^*) - f(x_0)}{\sum_{l=0}^{k} \eta_l}.$$

Choosing the learning rate $\eta_k = \frac{1}{\sqrt{k}+1}$, implies the $O(\epsilon^{-2})$ convergence. Extending this proof technique for general multi-dimension case, is our work in progress.

5 Discussion

We established global convergence for s-rectangular robust MDPs with iteration complexity of $O(SA\epsilon^{-1})$, which is much faster than existing complexity of of $O(S^4A^2\epsilon^{-4})$ [27], given the robust return is differentible w.r.t. policy. Moreover, our proof trivially yields a simpler and more intuitive proof non-robust MDPs by taking single environment uncertainty set. Further, we tried to alleviate the differentiablity condition on the robust return, yielding iteration complexity of $O(SA\epsilon^{-2})$, by using PL condition for sufficient increase lemma. We showed this can be done for one-dimension, which is motivating, and extending this to a general case, is our work in progress.

References

[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

[2] Amir Beck. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

[3] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods, 2019.

[4] Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization, 2021.

[5] Vineet Goyal and Julien Grand-Clément. Robust markov decision process: Beyond rectangularity, 2018.

[6] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, May 2005.

[7] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[8] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*. Citeseer, 2002.

[9] Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[10] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, 2020.

[11] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[12] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes, 2023.

[13] Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization, 2022.

[14] Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust markov decision process, 2022.

[15] Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Math. Oper. Res.*, 41(4):1484–1509, nov 2016.

[16] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[17] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods, 2022.

[18] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[19] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Oper. Res.*, 53:780–798, 2005.

[20] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[21] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

[22] Walter Rudin. Real and complex analysis. 1987. *Cited on*, 156, 1987.

[23] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015.

[24] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[25] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[26] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pages 1057–1063. Citeseer, 1999.

[27] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. On the convergence of policy gradient in robust mdps, 2022.

[28] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty, 2021.

[29] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning, 2022.

[30] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[31] Berç Rustem Wolfram Wiesemann, Daniel Kuhn. Robust markov decision processes. *Mathematics of Operations Research 38(1):153-183*, 2012.

[32] Lin Xiao. On the convergence rates of policy gradient methods, 2022.

[33] Huan Xu and Shie Mannor. Robustness and generalization, 2010.

A Robust MDPs: Helper results

Proposition 3. *The robust return $\rho^{\pi}_{\mathcal{U}}$ is Lipschitz in policy with some Lipschitz constant L .*

Proof. We know that $\rho^{\pi}_{(P,R)}$ is Lipschitz in π [1], and let its Lipschitz constant be $L_{(P,R)}$. Let $L = \max_{(P,R)\in\mathcal{U}} L_{(P,R)}$ be maximum of all Lipschitz constant. Compactness of $\mathcal{U}$ ensures L exist and it is finite. WLOG assume $\rho^{\pi_1}_{\mathcal{U}} \geq \rho^{\pi_2}_{\mathcal{U}}$, then

$$\begin{aligned}
\rho^{\pi_1}_{\mathcal{U}} - \rho^{\pi_2}_{\mathcal{U}} &= \min_{(P,R)} \rho^{\pi_1}_{(P,R)} - \min_{(P,R)} \rho^{\pi_2}_{(P,R)}, \\
&= \min_{(P,R)} \rho^{\pi_1}_{(P,R)} - \rho^{\pi_2}_{(P^{\pi_2}_{\mathcal{U}},R^{\pi_2}_{\mathcal{U}})}, \\
&\leq \rho^{\pi_1}_{(P^{\pi_2}_{\mathcal{U}},R^{\pi_2}_{\mathcal{U}})} - \rho^{\pi_2}_{(P^{\pi_2}_{\mathcal{U}},R^{\pi_2}_{\mathcal{U}})}, \\
&\leq L_{(P^{\pi_2}_{\mathcal{U}},R^{\pi_2}_{\mathcal{U}})} \|\pi_1 - \pi_2\|, \\
&\leq L\|\pi_1 - \pi_2\|.
\end{aligned}$$

□

Proposition 4. *The recursion*

$$f(x_{k+1}) - f(x_k) \geq c\eta_k(f(x^*) - f(x_{k+1}))$$

implies $f(x^*) - f(x_{k+1}) \leq \frac{f(x^*)-f(x_0)}{c\sum_{l=0}^{k}\eta_l}$.

Proof. Taking $a_k = f(x^*) - f(x_k)$, the above recursion implies,

$$\begin{aligned}
& a_k - a_{k+1} \geq c\eta_k a_{k+1} \\
\implies & \frac{a_k}{a_{k+1}} \geq c\eta_k + 1 \\
\implies & \Pi_{k=0}^n \frac{a_k}{a_{k+1}} \geq \Pi_{k=0}^n (c\eta_k + 1) \geq c\sum_{k=0}^n \eta_k \\
\implies & \frac{a_0}{a_{n+1}} \geq c\sum_{k=0}^n \eta_k. \\
\implies & \frac{a_{n+1}}{a_0} \leq \frac{1}{c\sum_{k=0}^n \eta_k}.
\end{aligned}$$

□

B Optimization: Global Convergence Rate of the Gradient Projection Method

Let a compact convex $C \subset \mathbb{R}^n$ and $f \in C_L^{1,1}(C)$. We aim to maximize f, i.e., find: $f^* := \max_{x \in C} f(x)$. Applying [2][Lemma 4.22], for all $x, y \in C$, it holds that:

$$f(y) \geq f(x) + \langle \nabla f(y), y - x \rangle - \frac{L}{2}\|y - x\|^2. \tag{16}$$

Let T (resp. G) be the next step gradient step operator (resp. the effective gradient step operator) defined as

$$T(x) := \mathbf{proj}_C(x + \frac{1}{L}\nabla f(x)), \tag{17}$$

$$G(x) := L(T(x) - x) \tag{18}$$

By construction of the projection operator, if $T(x) \in C$, then $T(x) = x + \frac{1}{L}\nabla f(x)$ so that $L(T(x) - x) = \nabla f(x) = G(x)$.

Proposition 5. *For all x, y in a convex set C, it holds that:*

$$\langle \nabla f(x) - G(x), T(x) - y \rangle \geq 0.$$

Proof. By the second projection theorem [2][Thm. 9.8], for any $x, y \in C$, we have

$$\begin{aligned}
& \langle x + \frac{1}{L}\nabla f(x) - T(x), T(x) - y \rangle \geq 0 \\
\iff & \langle \frac{1}{L}\nabla f(x) + x - T(x), T(x) - y \rangle \geq 0 \\
\iff & \frac{1}{L}\langle \nabla f(x) - G(x), T(x) - y \rangle \geq 0 \qquad [x - T(x) = -\frac{1}{L}G(x)] \\
\iff & \frac{1}{L}\langle \nabla f(x) - G(x), T(x) - y \rangle \geq 0 \\
\iff & \langle \nabla f(x) - G(x), T(x) - y \rangle \geq 0.
\end{aligned}$$

□

Similarly to the sufficient decrease lemma for constrained problems [2][Lemma 9.11], the above proposition enables us to establish a sufficient increase lemma for our constrained maximization.

Lemma (Sufficient Increase Lemma). *For all $f \in C_L^{1,1}(C)$, it holds that*

$$f(T(x)) - f(x) \geq \frac{1}{2L}\|G(x)\|^2, \qquad \forall x \in C.$$

Proof. Since $f \in C_L^{1,1}(C)$, Eq. (16) holds so that

$$\begin{aligned} f(T(x)) &\geq f(x) + \langle \nabla f(x), T(x) - x \rangle - \frac{L}{2}\|T(x) - x\|^2 \\ &= f(x) + \langle \nabla f(x), T(x) - x \rangle - \frac{1}{2L}\|G(x)\|^2. \qquad \text{[By Eq. (18)]} \end{aligned} \tag{19}$$

Set $x = y$ and apply Prop. 5. This yields $\langle \nabla f(x), T(x) - x \rangle \geq \langle G(x), T(x) - x \rangle$, which we incorporate into Eq. (19) to obtain:

$$\begin{aligned} f(T(x)) &\geq f(x) + \langle G(x), T(x) - x \rangle - \frac{1}{2L}\|G(x)\|^2 \\ &= f(x) + \langle G(x), \frac{1}{L}G(x) \rangle - \frac{1}{2L}\|G(x)\|^2 \qquad \text{[By Eq. (18)]} \\ &= f(x) + \frac{1}{2L}\|G(x)\|^2. \end{aligned}$$

This ends the proof. □

The above result does not require f to be concave. In the unconstrained case, we have $G = \nabla f$ in the above result. The intuition behind the above increase is: At a given point the function behaves like a linear function where the change in gradient is slow due to L-smoothness. The above result ensures, the iterates $\{x_k\}$ increases the function value by atleast $\frac{1}{2L}\|G(x_k)\|^2$, but this only guarantee local convergence.

We establish the following result which is also a consequence of the second projection theorem [2][Thm. 9.8].

Lemma 6. *For all $x \in C$, we have*

$$\max_{y \in C} \langle \nabla f(T(x)), y - T(x) \rangle \leq 2\|G(x)\| \mathbf{diam}(C),$$

where $\mathbf{diam}(C) := \max_{x,y \in C} \|x - y\|$ is the diameter of C.

Proof. For all $x, y \in C$, we have:

$$\begin{aligned} \langle \nabla f(T(x)), y - T(x) \rangle &= \langle \nabla f(T(x)) - \nabla f(x) + \nabla f(x), y - T(x) \rangle && \text{[Subtract \& add } \nabla f(x)\text{]} \\ &= \langle \nabla f(T(x)) - \nabla f(x), y - T(x) \rangle + \langle \nabla f(x), y - T(x) \rangle && \text{[Linearity of scalar product]} \\ &\leq \|\nabla f(T(x)) - \nabla f(x)\| \|y - T(x)\| + \langle \nabla f(x), y - T(x) \rangle && \text{[Cauchy-Schwartz inequality]} \\ &\leq L\|T(x) - x\| \|y - T(x)\| + \langle \nabla f(x), y - T(x) \rangle && [f \in C_L^{1,1}(C)] \\ &= \|G(x)\| \|y - T(x)\| + \langle \nabla f(x), y - T(x) \rangle && \text{[By definition of } G \text{ - Eq. (18)]} \\ &\leq \|G(x)\| \|y - T(x)\| + \langle \nabla G(x), y - T(x) \rangle && \text{[Prop. 5]} \\ &\leq 2\|G(x)\| \|y - T(x)\| && \text{[Cauchy-Schwartz inequality]} \\ &\leq 2\|G(x)\| \mathbf{diam}(C). && \text{[By construction, } T(x) \in C\text{]} \end{aligned}$$

Since the resulting bound is for arbitrary $y \in C$, we can set the maximum over $y \in C$ to conclude that

$$\max_{y \in C} \langle \nabla f(T(x)), y - T(x) \rangle \leq 2\|G(x)\| \mathbf{diam}(C).$$

□