A cubic-regularized policy Newton algorithm for reinforcement learning

Project Report

Submitted by

Mizhaan Prajit Maniyar

NA17B014

in partial fulfilment of requirements
for the award of the dual degree of

BACHELOR OF TECHNOLOGY in
NAVAL ARCHITECTURE AND OCEAN ENGINEERING

and

MASTER OF TECHNOLOGY in
ROBOTICS

DEPARTMENT OF ENGINEERING DESIGN INDIAN
INSTITUTE OF TECHNOLOGY MADRAS CHENNAI
600 036

June, 2022

CERTIFICATE

This is to certify that the project titled **A cubic-regularized policy Newton algorithm for reinforcement learning**, submitted by **Mizhaan Prajit Maniyar (NA17B014)** to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology in Naval Architecture and Ocean Engineering and Master of Technology in Robotics**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Name of Guide:

Prashanth L. A.

Signature:

Name of student:

Mizhaan Prajit Maniyar

Signature: MPManiyar

ABSTRACT

KEYWORDS : Reinforcement Learning, Newton methods, Cubic Regularized Newton methods, Policy Gradient theorem, Policy Hessian Theorem, Stochastic optimization.

We consider the problem of control in a reinforcement learning (RL) context. Policy gradient algorithms are a popular solution approach for this problem, and are usually shown to converge to a stationary point of the value function. In this paper, we propose a policy Newton algorithm that incorporates cubic regularization. Our algorithm employs the likelihood ratio method to form estimates of the gradient and Hessian of the value function using sample trajectories. We establish convergence of our proposed algorithm to a second-order stationary point of the value function, which implies avoidance of traps in the form of saddle points. Further, we numerically validate our algorithm on a few simple benchmarks.

Contents

1 Introduction

Markov decision processes (MDPs) provide a framework for analyzing problems in sequential decision making under uncertainty. An algorithm solving the MDP finds a policy that maximizes a performance objective such as the discounted cumulative reward. A direct solution approach for MDPs requires the knowledge of the underlying transition dynamics. In practical settings, such information is seldom available, and one usually resorts to reinforcement learning (RL) algorithms that solve an MDP using sample trajectories.

Classic RL algorithms suffer from the curse of dimensionality associated with large state spaces. A popular solution approach to overcome this problem is to consider a parametric representation of policies, and search for the best policy within this class using a stochastic gradient (SG) algorithm. This constitutes the 'policy gradient' approach, and an expression for the gradient of the objective forms the basis of such algorithms.

The analysis of policy gradient algorithms usually establish convergence to stationary points in the long run, or to approximate stationary points in the non-asymptotic regime. However, such points also include traps such as local minima and saddle points. Using second-order information, one can avoid such traps, and we adopt such an approach in this paper.

Our contributions are summarized as follows: First, we propose a cubic-regularized policy Newton method for solving a finite horizon MDP. Second, we establish convergence of our proposed algorithm to an approximate second order stationary point of the objective function. To the best of our knowledge, a policy Newton algorithm incorporating cubic-regularization has not been studied earlier in the literature.

Related work. Policy gradient algorithms and their analysis has received a lot of research attention, cf. [Fazel et al., 2018, Agarwal et al., 2020, Sutton et al., 1999, Mohammadi et al., 2021, Papini et al., 2018, Vijayan and Prashanth, 2021, Zhang et al., 2020]. In [Furmston et al., 2016], the authors propose policy Newton algorithms for solving an MDP in a setting where the model (or the transition dynamics) is known. In [Shen et al., 2019], the authors propose a policy gradient algorithm that incorporates second-order information, and establish convergence to an approximate stationary point. In contrast, we show that our algorithm, which also uses second-order information, avoids traps and converges to a local maxima of the objective.

Summary of previous work. In the last report, we followed the objective formulation as given by Furmston et al. [2016]. However, in this paper we follow the formulation given by Shen et al. [2019], which avoids the calculation of of the derivative of the Q-values for the Hessian estimation as before. Furthermore, instead of using the method of PT-inverse, we use a regularization technique called cubic-regularization proposed by Nesterov and Polyak [2006] to tackle the cases where the Hessian estimate can be degenerate.

2 Problem formulation

An MDP is a tuple of the form $(\mathcal{S}, \mathcal{A}, P, R)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the underlying transition dynamics that govern the state evolution of the MDP agent and R be the transition rewards accumulated by the agent during its trajectory. Let the condition probability distribution of transition into the next state s_{h+1} given that the agent is at state $s_h \in \mathcal{S}$ and action $a_h \in \mathcal{A}$ is $P(s_{h+1}|s_h, a_h)$. The actions are chosen according to a probability distribution $\pi(a_h|s_h)$ which is conditioned over the current state. We shall call π the policy that our agent follows. We also assume that the policy is parametrized by a vector $\theta \in \mathbb{R}^d$ and use the notation π_θ as a shorthand for the distribution $\pi(a_h|s_h; \theta)$. If our agent follows a trajectory for a given time horizon H then we define the trajectory $\tau := (s_1, a_1, \ldots, s_H, a_H)$ as a collection of states and actions pairs. We first define the probability of a trajectory τ following a policy π as

$$p(\tau; \pi) := \pi(a_H|s_H) \left(\prod_{h=1}^{H-1} P(s_{h+1}|s_h, a_h)\pi(a_h|s_h) \right) \rho(s_1). \tag{1}$$

Also, note that $p(\tau; \pi_\theta) = p(\tau; \theta)$ as they are both conditioned on the same information. The discounted cumulative reward for a trajectory τ with a discount factor $\gamma < 1$ is given as

$$\mathcal{G}(\tau) := \sum_{h=1}^{H} \gamma^{h-1} r(s_h, a_h),$$

where the transition reward r is just a function of the current state and action. Our objective is to maximize the *expected* discounted cumulative reward given by

$$J(\theta) := \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[\mathcal{G}(\tau)\right] = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[\sum_{h=1}^{H} \gamma^{h-1} r(s_h, a_h) \right]. \tag{2}$$

Our aim is to find the policy that maximizes the above objective, i.e.

$$\theta^* \in \operatorname*{argmax}_{\theta \in \mathbb{R}^d} J(\theta), \tag{3}$$

where θ^* is an *optimal policy*.

3 Cubic-regularized policy Newton algorithm

A stochastic gradient algorithm to find a local optima of the objective function in the problem (3), would perform an incremental update of the policy parameter as follows:

$$\theta_{k+1} = \theta_k + \eta M(\theta_k)\nabla J(\theta_k),$$

where $\eta \in \mathbb{R}^+$ is the step size and $M(\theta)$ is a preconditioning matrix that could depend on our policy parameter θ. If J is smooth and $M(\theta)$ is positive-definite, then the policy parameter update ensures an increase in the objective viz. the total expected reward, for sufficiently small η. Note that if $M(\theta)$ is the identity matrix, then the update rule above corresponds to a gradient step, while $M(\theta) = -\nabla^2 J(\theta)^{-1}$ would result in a Newton step.

In a typical RL setting, it is not feasible to find the exact gradient or Hessian of the objective function, since the underlying transition dynamics of the environment are unknown. Instead, one has to form sample-based estimates of these quantities. Now, if we use an estimate of the Hessian in place of the preconditioning matrix we cannot assure a stable gradient ascent as our estimate may not be negative-definite as is required. This makes the classical Newton update a bad candidate for our policy search algorithm. Nesterov and Polyak [2006] motivates an algorithm called cubic-regularized Newton method in a deterministic setting which tackles these issues and more. They show that the standard Newton step ($\eta = 1$) can alternatively be presented as follows:

$$\theta_{k+1} = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \left\{ \langle \nabla J(\theta_k), \theta - \theta_k \rangle + \frac{1}{2} \left\langle \nabla^2 J(\theta_k)(\theta - \theta_k), \theta - \theta_k \right\rangle \right\}.$$

The quantity that we are optimizing in the above equation is called as the *auxiliary* function. The cubic regularized Newton step adds a cubic term to this auxiliary function in the following manner:

$$\theta_{k+1} = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \left\{ \langle \nabla J(\theta_k), \theta - \theta_k \rangle + \frac{1}{2} \left\langle \nabla^2 J(\theta_k)(\theta - \theta_k), \theta - \theta_k \right\rangle - \frac{\alpha}{6} \left\| \theta - \theta_k \right\|^3 \right\},$$

where $\alpha \in \mathbb{R}^+$ is called the regularization parameter. In this modified newton step, we can now use the estimates of the gradient and Hessian and come to a convergence bound as shown by Balasubramanian and Ghadimi [2022]. We present a similar algorithm to Balasubramanian and Ghadimi [2022] but by using the estimates derived by Shen et al. [2019] in the reinforcement learning setting.

We now start with the following assumptions that are standard on the regularity of the MDP and the smoothness of our parameterized policy π_θ.

(A1) (Bounded rewards)**.** *The absolute value of the reward function of the MDP is bounded, i.e.*

$$|r(s,a)| \leq R, \qquad \forall (s,a) \in (\mathcal{S} \times \mathcal{A}).$$

(A2) (Parametrization regularity)**.** *For any choice of the parameters θ and any state-action pair (s,a), we have*

$$\|\nabla \log \pi(a|s;\theta)\| \leq G \quad \textit{and} \quad \left\|\nabla^2 \log \pi(a|s;\theta)\right\| \leq L.$$

(A3) (Lipschitz Hessian)**.** *For any set of parameters (θ_1, θ_2) and any state-action pair (s,a), we have*

$$\left\|\nabla^2 \log \pi(a|s;\theta_1) - \nabla^2 \log \pi(a|s;\theta_2)\right\| \leq C \left\|\theta_1 - \theta_2\right\|.$$

Note that (A1) and (A2) are standard in the literature of policy gradient and actor-critic algorithms as shown in Shen et al. [2019]. Furthermore, (A3) is also standard in second order policy search algorithms like Zhang et al. [2020]. Also, from the above assumptions, it can be shown that (see Lemmas 1 and 2 in Appendix 5.1)

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq G_{\mathcal{H}} \|\theta_1 - \theta_2\|, \text{ and } \left\|\nabla^2 J(\theta_1) - \nabla^2 J(\theta_2)\right\| \leq L_{\mathcal{H}} \|\theta_1 - \theta_2\|, \text{ where}$$
$$G_{\mathcal{H}} := \frac{HG^2R + LR}{(1-\gamma)^2} \text{ and } L_{\mathcal{H}} := \frac{H^2G^3R + 3HGLR + CR}{(1-\gamma)^2}. \tag{4}$$

The above result tells us that the objective is smooth and thus its gradient and Hessian are well defined. We now present the policy gradient and Hessian theorem.

Theorem 1 (Policy gradient and Hessian theorem)**.** *Let $\Psi_i(\tau) := \sum_{h=i}^{H} \gamma^{h-1} r(s_h, a_h)$, and $\Phi(\theta;\tau) = \sum_{i=1}^{H} \Psi_i(\tau) \log \pi(a_i|s_i;\theta)$. Then, under the assumptions (A1), (A2), (A3) the gradient $\nabla J(\theta)$ and the Hessian $\nabla^2 J(\theta)$ of the objective* (2) *are given by*

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[\nabla \Phi(\theta;\tau)\right],$$
$$\nabla^2 J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[\nabla \Phi(\theta;\tau) \nabla^\top \log p(\tau;\theta) + \nabla^2 \Phi(\theta;\tau)\right].$$

Proof. Refer to Section 5.1. □

Note that the above theorem gives us a way to formulate unbiased estimates which would be the terms under the expectation denoted as

$$g(\theta;\tau) := \nabla \Phi(\theta;\tau), \quad \mathcal{H}(\theta;\tau) := \nabla \Phi(\theta;\tau) \nabla^\top \log p(\tau;\theta) + \nabla^2 \Phi(\theta;\tau).$$

The above estimates are calculated by the information obtained from a given trajectory τ and policy parameter

θ. We simulate multiple trajectories and calculate these estimates for each of them and then take its average as the final estimates for our algorithm 1.

```
Algorithm 1: Cubic-regularized policy Newton
Input  : Initial parameter $\theta_0 \in \mathbb{R}^d$, a non-negative sequence $\{\alpha_k\}$, positive integer sequences
         $\{m_k\}$ and $\{b_k\}$, and an iteration limit $N \geq 1$.
for $k = 1, \ldots, N$ do
    /* Monte Carlo simulation                                              */
    Simulate $\max\{m_k, b_k\}$ number of trajectories according to $\theta_{k-1}$, randomly pick $m_k$ trajectories
      for set $\mathcal{T}_m$ and $b_k$ trajectories for set $\mathcal{T}_b$;
    /* Gradient and Hessian estimation                                     */
        $$\bar{g}_k = \frac{1}{m_k} \sum_{\tau \in \mathcal{T}_m} \sum_{h=1}^{H} \Psi_h(\tau) \nabla \log \pi(a_h | s_h; \theta_{k-1}),$$
        $$\bar{\mathcal{H}}_k = \frac{1}{b_k} \sum_{\tau \in \mathcal{T}_b} \left( \sum_{h=1}^{H} \Psi_h(\tau) \nabla \log \pi(a_h | s_h; \theta_{k-1}) \sum_{h'=1}^{H} \nabla^\top \log \pi(a_{h'} | s_{h'}; \theta_{k-1}) \right)$$
        $$+ \frac{1}{b_k} \sum_{\tau \in \mathcal{T}_b} \sum_{h=1}^{H} \Psi_h(\tau) \nabla^2 \log \pi(a_h | s_h; \theta_{k-1})$$
    where the state-action pairs $(s_h, a_h) \in \tau$ belong to the respective trajectories;
    /* Policy update (cubic regularized Newton step)                       */
    Compute
        $$\theta_k = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \left\{ \tilde{J}^k(\theta) \equiv \tilde{J}(\theta, \theta_{k-1}, \bar{\mathcal{H}}_k, \bar{g}_k, \alpha_k) \right\},$$
      where
        $$\tilde{J}(x, y, \mathcal{H}, g, \alpha) = \langle g, x - y \rangle + \frac{1}{2} \langle \mathcal{H}(x - y), x - y \rangle - \frac{\alpha}{6} \|x - y\|^3 . \qquad (5)$$
end for
Output : Policy $\theta_N$
```

4 Main results

In this section, we mathematically define a first and second order stationary point in optimization and then prove that our proposed algorithm 1 achieves the same.

Definition 1 (ϵ-First-order stationary point)**.** *Assume that a solution $\bar{x} \in \mathcal{X}$ as output of an algorithm and a target accuracy $\epsilon > 0$ are given, and the function to be maximized f is non-convex. Then $\bar{x}$ is called an ϵ first-order stationary point (ϵ-FOSP), if*

$$\mathbb{E}\left[\|\nabla f(\bar{x})\|\right] \leq \epsilon,$$

The above definition states that a first-order stationary point is a point where the gradient is zero if $\epsilon = 0$. This could potentially be a saddle point. In order to verify whether it is an optima, we need information with regards to the curvature. This motivates us to our second definition:

Definition 2 (ϵ-Second-order stationary point)**.** *Assume that a solution $\bar{x} \in \mathcal{X}$ as output of an algorithm and a target accuracy $\epsilon > 0$ are given, and the function to be maximized f is non-convex. Then for some $\rho > 0$, $\bar{x}$ is called an ϵ second-order stationary point (ϵ-SOSP) if*

$$\max\left\{\sqrt{\mathbb{E}\left[\|\nabla f(\bar{x})\|\right]}, \frac{1}{\sqrt{\rho}}\mathbb{E}\left[\lambda_{\max}\left(\nabla^2 f(\bar{x})\right)\right]\right\} \leq \sqrt{\epsilon},$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalue of a given matrix, respectively.

In the above definition, if $\epsilon = 0$ then $\bar{x}$ is a second-order stationary point. Therefore, a second order stationary point is where the gradient is zero and the Hessian is negative semi-definite. Such definitions are standard in many second order optimization literature, cf. [Balasubramanian and Ghadimi, 2022, Tripuraneni et al., 2018]. It is now evident that such an algorithm that outputs an ϵ-SOSP avoids saddle points.

We now state the result that establishes convergence of Algorithm 1 to an ϵ-SOSP of the objective (2).

Theorem 2. *Let $\{\theta_1, \ldots, \theta_N\}$ be computed by Algorithm 1 with the following parameters:*

$$\alpha_k = 3L_{\mathcal{H}}, N = \frac{12\sqrt{L_{\mathcal{H}}}\left(J^* - J(\theta_0)\right)}{\epsilon^{\frac{3}{2}}}, m_k = \frac{25G_g^2}{4\epsilon^2}, b_k = \frac{36\sqrt[3]{30(1+2\log 2d)}d^{\frac{2}{3}}G_{\mathcal{H}}^2}{\epsilon}. \quad (6)$$

Let θ_R be picked uniformly at random from $\{\theta_1, \ldots, \theta_N\}$. Then, we have

$$5\sqrt{\epsilon} \geq \max\left\{\sqrt{\mathbb{E}\left[\|\nabla J(\theta_R)\|\right]}, \frac{5}{6\sqrt{L_{\mathcal{H}}}}\mathbb{E}\left[\lambda_{\max}\left(\nabla^2 J(\theta_R)\right)\right]\right\}, \quad (7)$$

where $G_{\mathcal{H}}$ and $L_{\mathcal{H}}$ are defined as in (4).

Proof. Refer to Section 5.2. □

Remark 1. *As a consequence of Theorem 2, to obtain an ϵ second-order stationary point of the problem, the total number of trajectories required to compute the gradient and the Hessian are bounded by $O\left(\frac{1}{\epsilon^{\frac{7}{2}}}\right)$ and $O\left(\frac{d^{\frac{2}{3}}}{\epsilon^{\frac{5}{2}}}\right)$, respectively. This is of a higher order in contrast to the HAPG algorithm proposed by Shen et al. [2019], which requires $O\left(\frac{1}{\epsilon^{3}}\right)$ number of trajectories. However, the total number of time steps that it required for our algorithm to converge is of $O\left(\frac{1}{\epsilon^{1.5}}\right)$ versus the $O\left(\frac{1}{\epsilon^{2}}\right)$ from HAPG. Furthermore, our algorithm ensures a convergence to an ϵ-SOSP thereby avoiding saddle points, while HAPG is shown to converge to a first-order stationary point which could potentially include traps.*

5 Convergence analysis

5.1 Proof of Theorem 1

Proof. We can re-write the objective function (2) as follows:

$$J(\theta) := \mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\mathcal{G}(\tau)\right] = \mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\sum_{h=1}^{H}\gamma^{h-1}r(s_h,a_h)\right] = \sum_{h=1}^{H}\mathbb{E}_{\tau_h\sim p(\tau_h;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\right].$$

The equality above uses the fact that trajectories are independent. last two lines are equal as the term inside the expectation is independent of future events, i.e. the trajectory $(s_{1:h},a_{1:h})$ does not depend on the trajectory $(s_{h+1:H},a_{h+1:H})$. Replacing the expectation by the integral over all trajectories

$$J(\theta) = \sum_{h=1}^{H}\int_{\tau_h}\gamma^{h-1}r(s_h,a_h)p(\tau_h;\theta)\,d\tau_h.$$

Taking the derivative of the above equation

$$\nabla J(\theta) = \sum_{h=1}^{H}\int_{\tau_h}\gamma^{h-1}r(s_h,a_h)\nabla p(\tau_h;\theta)\,d\tau_h.$$

using the log trick where $\nabla p(\tau_h;\theta) = p(\tau_h;\theta)\nabla\log p(\tau_h;\theta)$, we obtain

$$\begin{aligned}\nabla J(\theta) &= \sum_{h=1}^{H}\int_{\tau_h}\gamma^{h-1}r(s_h,a_h)\nabla\log p(\tau_h;\theta)\,p(\tau_h;\theta)\,d\tau_h\\ &= \sum_{h=1}^{H}\mathbb{E}_{\tau_h\sim p(\tau_h;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\nabla\log p(\tau_h;\theta)\right].\end{aligned}$$

From (1), we can show that $\nabla\log p(\tau;\theta) = \sum_{h=1}^{H}\nabla\log\pi(a_h|s_h;\theta)$, and thus

$$\begin{aligned}\nabla J(\theta) &= \sum_{h=1}^{H}\mathbb{E}_{\tau_h\sim p(\tau_h;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\sum_{i=1}^{h}\nabla\log\pi(a_i|s_i;\theta)\right]\\ &= \sum_{h=1}^{H}\sum_{i=1}^{h}\mathbb{E}_{\tau_h\sim p(\tau_h;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\nabla\log\pi(a_i|s_i;\theta)\right]\\ &= \sum_{h=1}^{H}\sum_{i=1}^{h}\mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\nabla\log\pi(a_i|s_i;\theta)\right].\end{aligned}$$

where in the last equality we use that $\gamma^{h-1}r(s_h,a_h)\nabla\log\pi(a_i|s_i;\theta)$ with $i\le h$ is independent of the randomness after a_h. Exchanging the order of summation

$$\nabla J(\theta) = \sum_{i=1}^{H}\sum_{h=i}^{H}\mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\gamma^{h-1}r(s_h,a_h)\nabla\log\pi(a_i|s_i;\theta)\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{H} \mathbb{E}_{\tau\sim p(\tau;\theta)} \left[\left(\sum_{h=i}^{H} \gamma^{h-1} r(s_h, a_h) \right) \nabla \log \pi(a_i|s_i;\theta) \right] \\
&= \sum_{i=1}^{H} \mathbb{E}_{\tau\sim p(\tau;\theta)} \left[\Psi_i(\tau) \nabla \log \pi(a_i|s_i;\theta) \right].
\end{aligned}
$$

This concludes the proof of the first claim.

For the second claim, notice that

$$
\begin{aligned}
\nabla^2 J(\theta) &= \nabla \left(\int_\tau \nabla\Phi(\theta;\tau) p(\tau;\theta)\, d\tau \right) \\
&= \int_\tau \nabla\Phi(\theta;\tau)\nabla^\top p(\tau;\theta) + \nabla^2\Phi(\theta;\tau)p(\tau;\theta)\, d\tau \\
&= \int_\tau \left(\nabla\Phi(\theta;\tau)\nabla^\top \log p(\tau;\theta) + \nabla^2\Phi(\theta;\tau) \right) p(\tau;\theta)\, d\tau \\
&= \mathbb{E}_{\tau\sim p(\tau;\theta)} \left[\nabla\Phi(\theta;\tau)\nabla^\top \log p(\tau;\theta) + \nabla^2\Phi(\theta;\tau) \right].
\end{aligned}
$$

Hence, we conclude the proof. □

5.2 Proof of Theorem 2

The proof proceeds through a sequence of lemmas.

Lemma 1. *Under Assumptions (A1) and (A2), we have for any parameter θ and trajectory τ*

$$
\begin{aligned}
&\|\nabla\Phi(\theta;\tau)\| \le \frac{GR}{(1-\gamma)^2}, \quad \|\nabla^2\Phi(\theta;\tau)\| \le \frac{LR}{(1-\gamma)^2} \\
&\|g(\theta;\tau)\| \le G_g, \quad \text{and} \quad \|\mathcal{H}(\theta;\tau)\| \le G_{\mathcal{H}},
\end{aligned}
$$

where $G_g = \frac{GR}{(1-\gamma)^2}$, and $G_{\mathcal{H}} := \frac{HG^2R+LR}{(1-\gamma)^2}$.

Proof. Using the definition of $\Phi(\theta;\tau)$, we have

$$
\|\nabla\Phi(\theta;\tau)\| = \left\| \sum_{i=1}^{H} \Psi_i(\tau)\nabla\log\pi(a_i|s_i;\theta) \right\| \le \sum_{i=1}^{H} |\Psi_i(\tau)| \cdot \|\nabla\log\pi(a_i|s_i;\theta)\| \le G\sum_{i=1}^{H} |\Psi_i(\tau)|.
$$

We can establish a bound on $|\Psi_i(\tau)|$ as follows:

$$
|\Psi_i(\tau)| = |\sum_{h=i}^{H} \gamma^{h-1} r(s_h, a_h)| \le R\sum_{h=i}^{H}\gamma^{h-1} \le \frac{R\gamma^{i-1}}{1-\gamma}.
$$

Therefore,

$$\|\nabla \Phi(\theta;\tau)\| \le G \sum_{i=1}^{H} \frac{R\gamma^{i-1}}{1-\gamma} \le \frac{GR}{1-\gamma} \sum_{i=1}^{H} \gamma^{i-1} \le \frac{GR}{(1-\gamma)^2}.$$

Similarly,

$$\begin{aligned}\left\|\nabla^2 \Phi(\theta;\tau)\right\| &= \left\| \sum_{i=1}^{H} \Psi_i(\tau) \nabla^2 \log \pi(a_i|s_i;\theta) \right\| \\ &\le \sum_{i=1}^{H} |\Psi_i(\tau)| \cdot \left\|\nabla^2 \log \pi(a_i|s_i;\theta)\right\| \\ &\le L \sum_{i=1}^{H} |\Psi_i(\tau)| \le \frac{LR}{(1-\gamma)^2}.\end{aligned}$$

It is now easy to show that the gradient estimate $g(\theta;\tau)$ is bounded as follows:

$$\|g(\theta;\tau)\| = \|\nabla \Phi(\theta;\tau)\| \le \frac{GR}{(1-\gamma)^2} := G_g.$$

Next, we show that the Hessian estimate $\mathcal{H}(\theta;\tau)$ is bounded. Notice that

$$\begin{aligned}\|\mathcal{H}(\theta;\tau)\| &= \left\|\nabla \Phi(\theta;\tau) \nabla^{\top} \log p(\tau;\theta) + \nabla^2 \Phi(\theta;\tau)\right\| \\ &\le \|\nabla \Phi(\theta;\tau)\| \cdot \|\nabla \log p(\tau;\theta)\| + \left\|\nabla^2 \Phi(\theta;\tau)\right\| \\ &\le \frac{GR}{(1-\gamma)^2} \|\nabla \log p(\tau;\theta)\| + \frac{LR}{(1-\gamma)^2}.\end{aligned}$$

Using the relation $\nabla \log p(\tau;\theta) = \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h;\theta)$, we have

$$\|\nabla \log p(\tau;\theta)\| \le \sum_{h=1}^{H} \|\nabla \log \pi(a_h|s_h;\theta)\| \le HG.$$

Therefore, we obtain

$$\|\mathcal{H}(\theta;\tau)\| \le \frac{HG^2R + LR}{(1-\gamma)^2} = G_{\mathcal{H}}.$$

We now conclude the proof. □

From the above lemma, one can easily interpret that the objective function, i.e. $J(\theta)$ and its gradient, i.e. $\nabla J(\theta)$ are Lipschitz continuous. We now need to show that the Hessian of the objective, i.e. $\nabla^2 J(\theta)$ is Lipschitz.

Lemma 2. *Under Assumptions (A1), (A2) and (A3), we have for any* (θ_1, θ_2)*,*

$$\left\|\nabla^2 J(\theta_1) - \nabla^2 J(\theta_2)\right\| \leq L_{\mathcal{H}} \left\|\theta_1 - \theta_2\right\|, \textit{ where} \tag{8}$$
$$L_{\mathcal{H}} := \frac{H^2 G^3 R + 3HGLR + CR}{(1-\gamma)^2}.$$

Proof. We begin with the integral expression for the Hessian of our objective, i.e.

$$\nabla^2 J(\theta) = \int_\tau \nabla\Phi(\theta;\tau)\nabla^\top p(\tau;\theta) + p(\tau;\theta)\nabla^2\Phi(\theta;\tau)\, d\tau.$$

Therefore, we have

$$\begin{aligned}
\nabla^2 J(\theta_1) - \nabla^2 J(\theta_2) &= \int_\tau \nabla\Phi(\theta_1;\tau)\nabla^\top p(\tau;\theta_1) + p(\tau;\theta_1)\nabla^2\Phi(\theta_1;\tau)\, d\tau \\
&- \int_\tau \nabla\Phi(\theta_2;\tau)\nabla^\top p(\tau;\theta_2) + p(\tau;\theta_2)\nabla^2\Phi(\theta_2;\tau)\, d\tau \\
&= \int_\tau \left(\nabla\Phi(\theta_1;\tau)\nabla^\top p(\tau;\theta_1) - \nabla\Phi(\theta_2;\tau)\nabla^\top p(\tau;\theta_2)\right)\, d\tau \\
&- \int_\tau \left(p(\tau;\theta_1)\nabla^2\Phi(\theta_1;\tau) - p(\tau;\theta_2)\nabla^2\Phi(\theta_2;\tau)\right)\, d\tau \\
\text{Hence, } \left\|\nabla^2 J(\theta_1) - \nabla^2 J(\theta_2)\right\| &\leq \int_\tau \left\|\nabla\Phi(\theta_1;\tau)\nabla^\top p(\tau;\theta_1) - \nabla\Phi(\theta_2;\tau)\nabla^\top p(\tau;\theta_2)\right\|\, d\tau \\
&- \int_\tau \left\|p(\tau;\theta_1)\nabla^2\Phi(\theta_1;\tau) - p(\tau;\theta_2)\nabla^2\Phi(\theta_2;\tau)\right\|\, d\tau. \qquad (9)
\end{aligned}$$

For ease of notation, let $\Phi_j := \Phi(\theta_j;\tau)$ and $p_j := p(\tau;\theta_i)$. Considering the first integrand in (9)

$$\begin{aligned}
\left\|\nabla\Phi_1\nabla^\top p_1 - \nabla\Phi_2\nabla^\top p_2\right\| &\leq \left\|\nabla\Phi_1\nabla^\top p_1 - \nabla\Phi_1\nabla^\top p_2\right\| + \left\|\nabla\Phi_1\nabla^\top p_2 - \nabla\Phi_2\nabla^\top p_2\right\| \\
&\leq \left\|\nabla\Phi_1\right\|\left\|\nabla p_1 - \nabla p_2\right\| + \left\|\nabla p_2\right\|\left\|\nabla\Phi_1 - \nabla\Phi_2\right\|. \qquad (10)
\end{aligned}$$

Using mean value theorem for vector-valued functions, we have

$$\nabla p_1 - \nabla p_2 = \nabla^2 p_{h_1}(\theta_1 - \theta_2),$$

where $p_{h_1} = p(\tau;\theta_{h_1})$, and $\theta_{h_1} = (1-h_1)\theta_1 + h_1\theta_2$ for some $h_1 \in [0,1]$. Therefore,

$$\begin{aligned}
\left\|\nabla p_1 - \nabla p_2\right\| &\leq \left\|\nabla^2 p_{h_1}\right\|\left\|\theta_1 - \theta_2\right\| \\
&= \left\|\nabla(p_{h_1}\nabla\log p_{h_1})\right\|\left\|\theta_1 - \theta_2\right\| \\
&= \left\|\nabla p_{h_1}\nabla^\top \log p_{h_1} + p_{h_1}\nabla^2\log p_{h_1})\right\|\left\|\theta_1 - \theta_2\right\|
\end{aligned}$$

$$\begin{aligned} &\leq p_{h_1}\left(\left\|\nabla \log p_{h_1}\right\|^2+\left\|\nabla^2 \log p_{h_1}\right\|\right)\left\|\theta_1-\theta_2\right\| \\ &\leq p_{h_1}\left(H^2 G^2+H L\right)\left\|\theta_1-\theta_2\right\| . \end{aligned} \tag{11}$$

Similarly,

$$\left\|\nabla \Phi_1-\nabla \Phi_2\right\| \leq\left\|\nabla^2 \Phi_{h_2}\right\|\left\|\theta_1-\theta_2\right\|, \tag{12}$$

where $\Phi_{h_2}=\Phi(\theta_{h_2};\tau)$, and $\theta_{h_2}=(1-h_2)\theta_1+h_2\theta_2$ for some $h_2\in[0,1]$. Plugging (11) and (12) in (10), we have

$$\left\|\nabla \Phi_1 \nabla^{\top} p_1-\nabla \Phi_2 \nabla^{\top} p_2\right\| \leq p_{h_1}\left(H^2 G^2+H L\right)\left\|\nabla \Phi_1\right\|\left\|\theta_1-\theta_2\right\|+p_2\left\|\nabla \log p_2\right\|\left\|\nabla^2 \Phi_{h_2}\right\|\left\|\theta_1-\theta_2\right\|$$

$$\leq p_{h_1} \frac{H^2 G^3 R+H G L R}{(1-\gamma)^2}\left\|\theta_1-\theta_2\right\|+p_2 \frac{H G L R}{(1-\gamma)^2}\left\|\theta_1-\theta_2\right\| .$$

We used the results from Lemma 1 in the second inequality. Integrating on both sides of the above inequality, we obtain

$$\int_\tau\left\|\nabla \Phi_1 \nabla^{\top} p_1-\nabla \Phi_2 \nabla^{\top} p_2\right\| d \tau \leq \frac{H^2 G^3 R+2 H G L R}{(1-\gamma)^2}\left\|\theta_1-\theta_2\right\| .$$

In the above inequality, we used the fact that $\int_\tau p(\tau;\theta)\,d\tau=1$ for all θ. Moving on to the second integrand in (9)

$$\begin{aligned} \left\|p_1 \nabla^2 \Phi_1-p_2 \nabla^2 \Phi_2\right\| &\leq\left\|p_1 \nabla^2 \Phi_1-p_1 \nabla^2 \Phi_2\right\|+\left\|p_1 \nabla^2 \Phi_2-p_2 \nabla^2 \Phi_2\right\|, \\ &\leq p_1\left\|\nabla^2 \Phi_1-\nabla^2 \Phi_2\right\|+\left\|\nabla^2 \Phi_2\right\|\left|p_1-p_2\right| . \end{aligned} \tag{13}$$

Using mean value theorem,

$$\begin{aligned} \left|p_1-p_2\right| &\leq\left\|\nabla p_{h_3}\right\|\left\|\theta_1-\theta_2\right\| \\ &\leq p_{h_3}\left\|\nabla \log p_{h_3}\right\|\left\|\theta_1-\theta_2\right\| \\ &\leq p_{h_3}(H G)\left\|\theta_1-\theta_2\right\| . \end{aligned}$$

Considering the first term in (13),

$$\left\|\nabla^2 \Phi_1-\nabla^2 \Phi_2\right\|=\left\|\sum_{i=1}^H \Psi_i(\tau)\left(\nabla^2 \log \pi\left(a_i \mid s_i ; \theta_1\right)-\nabla^2 \log \pi\left(a_i \mid s_i ; \theta_2\right)\right)\right\|$$

$$
\begin{aligned}
&\le \sum_{i=1}^{H} |\Psi_i(\tau)| \left\| \nabla^2 \log \pi(a_i|s_i;\theta_1) - \nabla^2 \log \pi(a_i|s_i;\theta_2) \right\| \\
&\le C \left\| \theta_1 - \theta_2 \right\| \sum_{i=1}^{H} \frac{R\gamma^{i-1}}{1-\gamma} \le \frac{CR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\|.
\end{aligned}
$$

Where we used (A3) in the second last inequality above. Plugging the above results in (13), we have

$$
\left\| p_1 \nabla^2 \Phi_1 - p_2 \nabla^2 \Phi_2 \right\| \le p_1 \frac{CR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\| + p_{h_3} \frac{HGLR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\|.
$$

Integrating on both sides, we obtain

$$
\int_{\tau} \left\| p_1 \nabla^2 \Phi_1 - p_2 \nabla^2 \Phi_2 \right\| \, d\tau \le \frac{HGLR + CR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\|.
$$

Therefore, substituting in the original equation

$$
\begin{aligned}
\left\| \nabla^2 J(\theta_1) - \nabla^2 J(\theta_2) \right\| &\le \frac{H^2 G^3 R + 2HGLR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\| + \frac{HGLR + CR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\| \\
&= \frac{H^2 G^3 R + 3HGLR + CR}{(1-\gamma)^2} \left\| \theta_1 - \theta_2 \right\|.
\end{aligned}
$$

Hence, we conclude the proof. □

Remark 2. *From* (8)*, it can be easily seen that*

$$
\begin{aligned}
&\left\| \nabla J(\theta_1) - \nabla J(\theta_2) - \nabla^2 J(\theta_2)(\theta_1 - \theta_2) \right\| \le \frac{L_{\mathcal{H}}}{2} \left\| \theta_1 - \theta_2 \right\|^2, \\
&|J(\theta_1) - J(\theta_2) - \langle \nabla J(\theta_2), \theta_1 - \theta_2 \rangle - \frac{1}{2} \left\langle \theta_1 - \theta_2, \nabla^2 J(\theta_2)(\theta_1 - \theta_2) \right\rangle | \le \frac{L_{\mathcal{H}}}{6} \left\| \theta_1 - \theta_2 \right\|^3.
\end{aligned}
$$

Lemma 3. *Let* $\bar{\theta} = \operatorname{argmax}_{x \in \mathbb{R}^d} \tilde{J}(x, \theta, \mathcal{H}, g, \alpha)$*. Then, we have*

$$
\begin{aligned}
&g + \mathcal{H}(\bar{\theta} - \theta) - \frac{\alpha}{2} \left\| \bar{\theta} - \theta \right\| (\bar{\theta} - \theta) = 0, \\
&\mathcal{H} - \frac{\alpha}{2} \left\| \bar{\theta} - \theta \right\| I_d \preceq 0.
\end{aligned}
$$

Proof. See Lemma 4.3 from Balasubramanian and Ghadimi [2022]. □

We now derive the second and third order error bounds on our Hessian estimate.

Lemma 4. *Let $\bar{g}_k$ and $\bar{\mathcal{H}}_k$ be computed by 1, and assume $b_k \geq 4(1 + 2\log 2d)$. Then we have*

$$\mathbb{E}\left[\|\bar{g}_k - \nabla J(\theta_{k-1})\|^2\right] \leq \frac{G_g^2}{m_k}, \qquad \mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|^3\right] \leq \frac{4\sqrt{15(1+2\log 2d)}dG_{\mathcal{H}}^3}{b_k^{\frac{3}{2}}} \tag{14}$$

Proof. Using the fact that the estimate $\bar{g}_k$ is unbiased, we have

$$\mathbb{E}\left[\|\bar{g}_k - \nabla J(\theta_{k-1})\|^2\right] \leq \mathbb{E}\left[\|\bar{g}_k\|^2\right] \leq \frac{1}{m_k^2}\mathbb{E}\left[\sum_{\tau \in \mathcal{T}_m} \|g(\theta_{k-1};\tau)\|^2\right] \leq \frac{G_g^2}{m_k}.$$

This establishes the first bound in (14).

Now we turn to proving the second bound in (14). By Theorem 1 in Tropp [2016], we have

$$\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|^2\right] \leq \frac{2C(d)}{b_k^2}\left(\left\|\sum_{\tau\in\mathcal{T}_b}\mathbb{E}\left[\Delta_{k,\tau}^2\right]\right\| + C(d)\mathbb{E}\left[\max_i \|\Delta_{k,\tau}\|^2\right]\right),$$

where $\Delta_{k,\tau} = \mathcal{H}(\theta_{k-1};\tau) - \nabla^2 J(\theta_{k-1})$ and $C(d) = 4(1 + 2\log 2d)$. Hence, we can see that

$$\mathbb{E}\left[\|\Delta_{k,\tau}\|^2\right] \leq \mathbb{E}\left[\|\mathcal{H}(\theta_{k-1};\tau)\|^2\right] \leq G_{\mathcal{H}}^2,$$

which together with the above inequality and the fact that

$$\left\|\sum_{\tau\in\mathcal{T}_b}\mathbb{E}\left[\Delta_{k,\tau}^2\right]\right\| \leq \sum_{\tau\in\mathcal{T}_b}\left\|\mathbb{E}\left[\Delta_{k,\tau}^2\right]\right\| \leq \sum_{\tau\in\mathcal{T}_b}\mathbb{E}\left[\|\Delta_{k,\tau}\|^2\right],$$

implies

$$\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|^2\right] \leq \frac{2C(d)}{b_k^2}\left(b_k G_{\mathcal{H}}^2 + C(d)G_{\mathcal{H}}^2\right) \leq \frac{4C(d)}{b_k}G_{\mathcal{H}}^2,$$

where in the last inequality we use the assumption that $b_k \geq C(d)$. Using Holder's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|^3\right] &\leq \mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\| \cdot \left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|_F^2\right] \\ &\leq \left(\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|^2\right] \cdot \mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|_F^4\right]\right)^{\frac{1}{2}}.\end{aligned} \tag{15}$$

Note that $\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1}) = \frac{1}{b_k}\sum_{\tau\in\mathcal{T}_b}\Delta_{k,\tau}$, therefore we have

$$\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|_F^4\right] = \mathbb{E}\left[\left\|\frac{1}{b_k}\sum_{\tau\in\mathcal{T}_b}\Delta_{k,\tau}\right\|_F^4\right] = \frac{1}{b_k^4}\mathbb{E}\left[\left\|\sum_{\tau\in\mathcal{T}_b}\Delta_{k,\tau}\right\|_F^4\right],$$

which together with Rosenthal's inequality (see Lemma 10 in Appendix B) implies

$$\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|_F^4\right] \leq \frac{3\mathbb{E}\left[\|\Delta_{k,\tau}\|_F^4\right]}{b_k^2}.$$

Using the fact that $\|\cdot\|_F \leq \sqrt{d}\,\|\cdot\|$ and the inequality from Lemma 9 in Appendix B, we have

$$\mathbb{E}\left[\left\|\bar{\mathcal{H}}_k - \nabla^2 J(\theta_{k-1})\right\|_F^4\right] \leq \frac{3d^2\mathbb{E}\left[\|\Delta_{k,\tau}\|^4\right]}{b_k^2} \leq \frac{15d^2\mathbb{E}\left[\|\mathcal{H}(\theta_{k-1};\tau_i)\|^4\right]}{b_k^2} \leq \frac{15d^2G_{\mathcal{H}}^4}{b_k^2},$$

which when combined in (15) leads to the second bound in (14). □

We next state a result that will be used in a subsequent lemma.

Lemma 5. *If for any two matrices A and B, and a scalar c, we have*

$$A \preceq B + cI,$$

where I is the identity matrix of the appropriate dimensions, then the following holds

$$c \geq \lambda_{max}(A) - \|B\|$$

Proof. See Appendix A. □

Lemma 6. *Let $\{\theta_k\}$ be computed by Algorithm 1. Then, we have*

$$\begin{aligned}&\sqrt{\mathbb{E}\left[\|\theta_k - \theta_{k-1}\|^2\right]} && (16)\\ &\geq \max\left\{\sqrt{\frac{\mathbb{E}\left[\|\nabla J(\theta_k)\|\right] - \delta_k^g - \delta_k^{\mathcal{H}}}{L_{\mathcal{H}} + \alpha_K}}, \frac{2}{\alpha_k + 2L_{\mathcal{H}}}\left[\mathbb{E}\left[\lambda_{\max}\left(\nabla^2 J(\theta_k)\right)\right] - \sqrt{2(\alpha_k + L_{\mathcal{H}})\delta_k^{\mathcal{H}}}\right]\right\},\end{aligned}$$

where $\delta_k^g, \delta_k^{\mathcal{H}} > 0$ are chosen such that

$$\mathbb{E}\left[\|\nabla J(\theta_{k-1}) - \bar{g}_k\|^2\right] \leq (\delta_k^g)^2, \qquad \mathbb{E}\left[\left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|^3\right] \leq \left(2(L_{\mathcal{H}} + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}. \quad (17)$$

Proof. Firstly, note that δ_k^g and $\delta_k^{\mathcal{H}}$ are inversely proportional to $\sqrt{m_k}$ and b_k, respectively and are therefore well-defined. Now, by the equality condition in Lemma 3, we have

$$\|\nabla J(\theta_k)\| \leq \left\|\nabla J(\theta_k) - \nabla J(\theta_{k-1}) - \nabla^2 J(\theta_{k-1})(\theta_k - \theta_{k-1})\right\| + \|\nabla J(\theta_{k-1}) - \bar{g}_k\|$$

$$
\begin{aligned}
&+ \left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \cdot \|\theta_k - \theta_{k-1}\| + \frac{\alpha_k}{2} \|\theta_k - \theta_{k-1}\|^2 \\
&\leq \frac{(L_{\mathcal{H}} + \alpha_k)}{2} \|\theta_k - \theta_{k-1}\|^2 + \|\nabla J(\theta_{k-1}) - \bar{g}_k\| + \left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \cdot \|\theta_k - \theta_{k-1}\| \\
&\leq (L_{\mathcal{H}} + \alpha_k) \|\theta_k - \theta_{k-1}\|^2 + \|\nabla J(\theta_{k-1}) - \bar{g}_k\| + \frac{\left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|^2}{2(L_{\mathcal{H}} + \alpha_k)},
\end{aligned}
$$

where in the inequality comes from Young's inequality. We now take expectation on both sides to obtain

$$
\frac{\left(\mathbb{E}\left[\|\nabla J(\theta_k)\| - \delta_k^g - \delta_k^{\mathcal{H}}\right]\right)}{L_{\mathcal{H}} + \alpha_k} \leq \mathbb{E}\left[\|\theta_k - \theta_{k-1}\|^2\right]. \tag{18}
$$

By the inequality in Lemma 3, and the smoothness result in Lemma 2, we have

$$
\begin{aligned}
\nabla^2 J(\theta_k) &\preceq \nabla^2 J(\theta_{k-1}) + L_{\mathcal{H}} \|\theta_k - \theta_{k-1}\| I_d = \nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k + \bar{\mathcal{H}}_k + L_{\mathcal{H}} \|\theta_k - \theta_{k-1}\| I_d \\
&\preceq \nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k + \frac{(\alpha_k + 2L_{\mathcal{H}}) \|\theta_k - \theta_{k-1}\|}{2} I_d,
\end{aligned}
$$

which together with Lemma 5, implies

$$
\frac{(\alpha_k + 2L_{\mathcal{H}}) \|\theta_k - \theta_{k-1}\|}{2} \geq \lambda_{\max}\left(\nabla^2 J(\theta_k)\right) - \left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\|.
$$

Taking expectation on both side, and using the definition $\delta_k^{\mathcal{H}}$ in (17), we have

$$
\begin{aligned}
\sqrt{\mathbb{E}\left[\|\theta_k - \theta_{k-1}\|^2\right]} &\geq \mathbb{E}\left[\|\theta_k - \theta_{k-1}\|\right] \\
&\geq \frac{2}{\alpha_k + 2L_{\mathcal{H}}} \left[\mathbb{E}\left[\lambda_{\max}\left(\nabla^2 J(\theta_k)\right)\right] - \sqrt{2(\alpha_k + L_{\mathcal{H}}))\delta_k^{\mathcal{H}}}\right].
\end{aligned}
$$

combining the above inequality with (18), we obtain (16). □

Lemma 7. *Let $\{\theta_k\}$ be computed by Algorithm [1] for a given iteration limit $N \geq 1$, we have*

$$\mathbb{E}\left[\|\theta_R - \theta_{R-1}\|^3\right] \tag{19}$$
$$\leq \frac{36}{\sum_{k=1}^N \alpha_k}\left[J^* - J(\theta_0) + \sum_{k=1}^N \frac{4\,(\delta_k^g)^{\frac{3}{2}}}{\sqrt{3\alpha_k}} + \sum_{k=1}^N \left(\frac{18\sqrt[4]{2}}{\alpha_k}\right)^2 \left((L_{\mathcal{H}} + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}\right],$$

where R is a random variable whose probability distribution $P_R(\cdot)$ is supported on $\{1, \ldots, N\}$ and given by

$$P_R(R = k) = \frac{\alpha_k}{\sum_{k=1}^N \alpha_k} \qquad k = 1, \ldots, N, \tag{20}$$

and $\delta_k^g, \delta_k^{\mathcal{H}} > 0$ are defined as before in (17).

Proof. We can see that by Lemma 2, (5) and the fact that $\alpha_k \geq L_{\mathcal{H}}$, we have

$$J(\theta_k) \geq J(\theta_{k-1}) + \tilde{J}^k(\theta_k) - \|\nabla J(\theta_{k-1} - \bar{g}_k)\| \, \|\theta_k - \theta_{k-1}\| - \frac{1}{2}\left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \|\theta_k - \theta_{k-1}\|^2. \tag{21}$$

Moreover, by Lemma 3, we have

$$\tilde{J}^k(\theta_k) = -\frac{1}{2}\left\langle \bar{\mathcal{H}}_k(\theta_k - \theta_{k-1}), (\theta_k - \theta_{k-1})\right\rangle + \frac{\alpha_k}{3}\|\theta_k - \theta_{k-1}\|^3 \geq \frac{\alpha_k}{12}\|\theta_k - \theta_{k-1}\|^3. \tag{22}$$

Combining (21) and (22), we obtain

$$\begin{aligned}
\frac{\alpha_k}{12}\|\theta_k - \theta_{k-1}\|^3 &\leq J(\theta_k) - J(\theta_{k-1}) + \|\nabla J(\theta_{k-1} - \bar{g}_k)\| \, \|\theta_k - \theta_{k-1}\| \\
&\quad + \frac{1}{2}\left\|\nabla^2 J(\theta_{k-1}) - \bar{\mathcal{H}}_k\right\| \|\theta_k - \theta_{k-1}\|^2 \\
&\leq J(\theta_k) - J(\theta_{k-1}) + \frac{4}{\sqrt{3\alpha_k}}\|\nabla J(\theta_{k-1} - \bar{g}_k)\|^{\frac{3}{2}} \\
&\quad + \left(\frac{9\sqrt{2}}{\alpha_k}\right)^2 \left\|\nabla^2 J(\theta_{k-1} - \bar{\mathcal{H}}_k)\right\|^3 + \frac{\alpha_k}{18}\|\theta_k - \theta_{k-1}\|^3,
\end{aligned} \tag{23}$$

where the last inequality follows from the fact $ab \leq \frac{a^p}{\lambda^p p} - \frac{\lambda^q b^q}{q}$ for p, q satisfying $\frac{1}{p} + \frac{1}{q} = 1$. We now take expectation on both sides of (23) and use (17) to obtain

$$\frac{\alpha_k}{36}\mathbb{E}\left[\|\theta_k - \theta_{k-1}\|^3\right] \leq J(\theta_k) - J(\theta_{k-1}) + \frac{4\,(\delta_k^g)^{\frac{3}{2}}}{\sqrt{3\alpha_k}} + \left(\frac{18\sqrt[4]{2}}{\alpha_k}\right)^2 \left((L_{\mathcal{H}} + \alpha_k)\delta_k^{\mathcal{H}}\right)^{\frac{3}{2}}.$$

Summing over $k = 1, \ldots, N$, dividing both sides by $\sum_{k=1}^N \alpha_k$ and noting (20), we obtain the bound in(19). □

And now we finally state the proof of Theorem 2.

Proof. First, note that by (6), Lemma 14, we can ensure that (17) is satisfied by $\delta_k^g = 2\epsilon/5$ and $\delta_k^{\mathcal{H}} = \epsilon/144$. Moreover, by Lemma 7, we have

$$\mathbb{E}\left[\|\theta_R - \theta_{R-1}\|^3\right] \leq \frac{1}{L_{\mathcal{H}}^{\frac{3}{2}}}\left[\frac{12\sqrt{L_{\mathcal{H}}}(J^* - J(\theta_0))}{N} + 6.88\epsilon^{\frac{3}{2}}\right] \leq \frac{8\epsilon^{\frac{3}{2}}}{L_{\mathcal{H}}^{\frac{3}{2}}},$$

where we choose N according to (6). Therefore, θ_R is an 4ϵ stationary point of the problem. Furthermore, from Lemma 6, one can verify that

$$\sqrt{\mathbb{E}\left[\|\nabla J(\theta_k)\|\right]} \leq \sqrt{\left(16 + \frac{2}{5} + \frac{1}{144}\right)\epsilon} \leq 5\sqrt{\epsilon},$$

and

$$\frac{\mathbb{E}\left[\lambda_{\max}\left(\nabla^2 J(\theta_k)\right)\right]}{\sqrt{L_{\mathcal{H}}}} \leq \left(5 + \frac{\sqrt{2}}{6}\right)\sqrt{\epsilon} \leq 6\sqrt{\epsilon},$$

from which we can obtain (7). Finally, note that the total number of required samples to obtain such a solution is bounded by

$$\sum_{k=1}^{N} m_k = O\left(\frac{1}{\epsilon^{\frac{7}{2}}}\right), \qquad \sum_{k=1}^{N} b_k = O\left(\frac{d^{\frac{2}{3}}}{\epsilon^{\frac{5}{2}}}\right).$$

□

6 Conclusions and future work

In this thesis, we have presented a novel algorithm 1. As far as we know, there has been no policy based reinforcement algorithm that has shown convergence to an ϵ-SOSP. Related works like [Shen et al., 2019, Huang et al., 2020] incorporate second order information via a Hessian-vector product but their algorithm does not mention a convergence to an ϵ-SOSP.

Future work. However, our algorithm is not yet ready to be applied yet as we yet have to device an optimization sub-algorithm to compute the maximum of the convex auxiliary function. Work in this regard has been done by Tripuraneni et al. [2018] and the algorithm would be complete if we extend our analysis and follow up with relevant simulation experiments.

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, Feb 2022. ISSN 1615-3383.

M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

T. Furmston, G. Lever, and D. Barber. Approximate Newton Methods for Policy Search in Markov Decision Processes. *Journal of Machine Learning Research*, 17(226):1–51, 2016.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4422–4433. PMLR, 13–18 Jul 2020.

H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović. On the linear convergence of random search for discrete-time lqr. *IEEE Control Systems Letters*, 5(3):989–994, 2021.

Yurii Nesterov and Boris Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108:177–205, 08 2006.

M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035. PMLR, 10–15 Jul 2018.

Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR, 2019.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pages 1057–1063, 1999.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Joel A. Tropp. The Expected Norm of a Sum of Independent Random Matrices: An Elementary Approach. Springer International Publishing, 2016.

N. Vijayan and L. A. Prashanth. Smoothed functional-based gradient algorithms for off-policy reinforcement learning, 2021.

K. Zhang, A. Koppel, H. Zhu, and T. Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control. Optim.*, 58(6):3586–3612, 2020.

A Proof of Lemma 5

We state and prove a useful result that will imply bound in Lemma 5.

Lemma 8. *For any square matrix $A \in \mathbb{R}^{d\times d}$ and for all vectors $v \in \mathbb{R}^d$, the following holds*

$$v^\top Av \leq \lambda \left\|v\right\|^2, \tag{24}$$

for some $\lambda \in \mathbb{R}$, if and only if

$$\lambda_{\max}(A) \leq \lambda. \tag{25}$$

Proof. We shall first prove the forward argument which is quite trivial. We define the map

$$\lambda_v(A) := \frac{v^\top Av}{\left\|v\right\|^2}$$

Now, given (24), we shall find a v_* such that $Av_* = \lambda_{max}(A)v_*$, i.e. v_* is the eigenvector associated with the largest eigenvalue of A. As $v_* \in \mathcal{C}(A) \subseteq \mathbb{R}^d$, the following should hold

$$\lambda_{v_*}(A) \leq \lambda.$$

But,

$$\lambda_{v_*}(A) = \frac{v_*^\top Av_*}{\left\|v_*\right\|^2} = \lambda_{\max}(A)\frac{v_*^\top v_*}{\left\|v_*\right\|^2} = \lambda_{\max}(A) \leq \lambda.$$

Now to check if the converse holds, we start by arguing that $\lambda_v(A) \leq \lambda_{\max}(A)$ for all v and A. We argue that $\lambda_v(A) \in [\lambda_{\min}(A), \lambda_{\max}(A)]$, as it has the form

$$\lambda_v(A) = \frac{\sum_{i=1}^r \lambda_i a_i^2}{\sum_{i=1}^r a_i^2},$$

where r is the rank of A and a_i are the coefficients of v. Hence, $\lambda_v(A)$ can be thought of as a weighted average of all eigenvalues of A. Therefore, given (25), we have for all $v \in \mathbb{R}^d$,

$$\lambda_v \leq \lambda,$$

which satisfies (24). □

Proof. (Lemma 5) For all $v \in \mathbb{R}^d$, we have

$$v^\top Av \leq v^\top Bv + c\left\|v\right\|^2$$

$$\begin{aligned}&\leq \|B\| \cdot \|v\|^2 + c\|v\|^2 \\ &= (\|B\| + c)\|v\|^2 ,\end{aligned}$$

where in the second line, we used the Cauchy-Schwartz inequality. Now by using, Lemma 8 in the last line, we obtain

$$\lambda_{max}(A) \leq \|B\| + c, \quad \text{implying} \quad c \geq \lambda_{max}(A) - \|B\| .$$

□

B A few probabilistic inequalities

We state and prove two probabilistic inequalites, which are used in the proof of Theorem 2. In particular, the result below as well as Rosenthal's inequality (stated in Lemma 10) are used in the proof of Lemma 4

Lemma 9. *Let $Z \in \mathbb{R}^{d\times d}$ be a random matrix. Then, we have*

$$\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^4\right] \leq 5\mathbb{E}\left[\|Z\|^4\right].$$

Proof. We can re-write the expectation as

$$\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^4\right] = \mathrm{Var}\left(\|Z - \mathbb{E}[Z]\|^2\right) + \left(\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^2\right]\right)^2 .$$

Consider the first term

$$\begin{aligned}\mathrm{Var}\left(\|Z - \mathbb{E}[Z]\|^2\right) &= \mathrm{Var}\left(\|Z\|^2 + \|\mathbb{E}[Z]\|^2 - 2\langle Z, \mathbb{E}[Z]\rangle\right) \\ &= \mathrm{Var}\left(\|Z\|^2 - 2\langle Z, \mathbb{E}[Z]\rangle\right) \qquad \left(\because \mathrm{Var}\left(\|\mathbb{E}[Z]\|^2\right) = 0\right) \\ &\leq \mathrm{Var}\left(\|Z\|^2\right) + 4\mathrm{Var}\left(\langle Z, \mathbb{E}[Z]\rangle\right) + 4\sqrt{\mathrm{Var}\left(\|Z\|^2\right)}\sqrt{\mathrm{Var}\left(\langle Z, \mathbb{E}[Z]\rangle\right)}.\end{aligned}$$

Now for the second term

$$\begin{aligned}\left(\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^2\right]\right)^2 &= \left(\mathbb{E}\left[\|Z\|^2\right] - \|\mathbb{E}[Z]\|^2\right)^2 \\ &= \left(\mathbb{E}\left[\|Z\|^2\right]\right)^2 + \|\mathbb{E}[Z]\|^4 - 2\mathbb{E}\left[\|Z\|^2\right]\|\mathbb{E}[Z]\|^2 .\end{aligned}$$

Simplifying the terms under the root

$$\begin{aligned}
\sqrt{\mathrm{Var}\left(\|Z\|^2\right)} &= \sqrt{\mathbb{E}\left[\|Z\|^4\right] - \left(\mathbb{E}\left[\|Z\|^2\right]\right)^2} \\
&= \sqrt{\mathbb{E}\left[\|Z\|^4\right]}\sqrt{1 - \frac{\left(\mathbb{E}\left[\|Z\|^2\right]\right)^2}{\mathbb{E}\left[\|Z\|^4\right]}} \\
&\leq \sqrt{\mathbb{E}\left[\|Z\|^4\right]}\left(1 - \frac{\left(\mathbb{E}\left[\|Z\|^2\right]\right)^2}{2\mathbb{E}\left[\|Z\|^4\right]}\right)
\end{aligned}$$

where in the last inequality we used the fact that $\sqrt{1-x} \leq 1 - \frac{x}{2}$.

$$\begin{aligned}
\mathrm{Var}\left(\langle Z, \mathbb{E}\left[Z\right]\rangle\right) &= \mathbb{E}\left[\langle Z, \mathbb{E}\left[Z\right]\rangle^2\right] - \left(\mathbb{E}\left[\langle Z, \mathbb{E}\left[Z\right]\rangle\right]\right)^2 \\
&\leq \mathbb{E}\left[\|Z\|^2 \|\mathbb{E}\left[Z\right]\|^2\right] - \|\mathbb{E}\left[Z\right]\|^4 \\
&\leq \mathbb{E}\left[\|Z\|^2\right] \|\mathbb{E}\left[Z\right]\|^2 .
\end{aligned}$$

Putting these results together

$$\begin{aligned}
\mathbb{E}\left[\|Z - \mathbb{E}\left[Z\right]\|^4\right] &\leq \mathrm{Var}\left(\|Z\|^2\right) + \left(\mathbb{E}\left[\|Z\|^2\right]\right)^2 \\
&\quad + 4\mathrm{Var}\left(\langle Z, \mathbb{E}\left[Z\right]\rangle\right) + \|\mathbb{E}\left[Z\right]\|^4 - 2\mathbb{E}\left[\|Z\|^2\right] \|\mathbb{E}\left[Z\right]\|^2 \\
&\quad + 4\sqrt{\mathrm{Var}\left(\|Z\|^2\right)}\sqrt{\mathrm{Var}\left(\langle Z, \mathbb{E}\left[Z\right]\rangle\right)} \\
&\leq \mathbb{E}\left[\|Z\|^4\right] + 2\mathbb{E}\left[\|Z\|^2\right] \|\mathbb{E}\left[Z\right]\|^2 - 3\|\mathbb{E}\left[Z\right]\|^4 \\
&\quad + 4\sqrt{\mathbb{E}\left[\|Z\|^4\right]}\left(1 - \frac{\left(\mathbb{E}\left[\|Z\|^2\right]\right)^2}{2\mathbb{E}\left[\|Z\|^4\right]}\right)\sqrt{\mathbb{E}\left[\|Z\|^2\right] \|\mathbb{E}\left[Z\right]\|^2}.
\end{aligned}$$

Note that by Jensen's inequality, we have $\mathbb{E}\left[\|Z\|^2\right] \|\mathbb{E}\left[Z\right]\|^2 \leq \left(\mathbb{E}\left[\|Z\|^2\right]\right)^2 \leq \mathbb{E}\left[\|Z\|^4\right]$. Substituting these results above and further simplification, we have

$$\mathbb{E}\left[\|Z - \mathbb{E}\left[Z\right]\|^4\right] \leq 5\mathbb{E}\left[\|Z\|^4\right] - 3\|\mathbb{E}\left[Z\right]\|^4 \leq 5\mathbb{E}\left[\|Z\|^4\right].$$

□

Lemma 10 (Rosenthal's inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be a sequence of random $d \times d$ square matrices with $\mathbb{E}[X_i] = 0$ for all i. Then the following inequality holds*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|_F^4\right] \leq 3n^2 \mathbb{E}\left[\|X_i\|_F^4\right],$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Proof. We start by using the definition of variance as follows

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|_F^4\right] = \operatorname{Var}\left(\left\|\sum_{i=1}^{n} X_i\right\|_F^2\right) + \left(\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|_F^2\right]\right)^2. \tag{26}$$

Consider the first term in (26)

$$\begin{aligned}
\operatorname{Var}\left(\left\|\sum_{i=1}^{n} X_i\right\|_F^2\right) &= \operatorname{Var}\left(\operatorname{Tr}\left(\sum_i X_i^\top \sum_j X_j\right)\right) \\
&= \operatorname{Var}\left(\operatorname{Tr}\left(\sum_i X_i^\top X_i + 2\sum_{i<j} X_i^\top X_j\right)\right) \\
&= \operatorname{Var}\left(\sum_i \|X_i\|_F^2 + 2\sum_{i<j} \operatorname{Tr}\left(X_i^\top X_j\right)\right) \\
&= \sum_i \operatorname{Var}\left(\|X_i\|_F^2\right) + 4\sum_{i<j} \operatorname{Var}\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right).
\end{aligned}$$

Expanding the terms under summation

$$\begin{aligned}
\sum_i \operatorname{Var}\left(\|X_i\|_F^2\right) &= \sum_i \mathbb{E}\left[\|X_i\|_F^4\right] - \sum_i \left(\mathbb{E}\left[\|X_i\|_F^2\right]\right)^2, \\
\sum_{i<j} \operatorname{Var}\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right) &= \sum_{i<j} \mathbb{E}\left[\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right)^2\right] - \sum_{i<j} \left(\mathbb{E}\left[\operatorname{Tr}\left(X_i^\top X_j\right)\right]\right)^2 \\
&= \sum_{i<j} \mathbb{E}\left[\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right)^2\right]. \qquad (\because \mathbb{E}\left[\operatorname{Tr}\left(X_i^\top X_j\right)\right] = 0 \text{ for } i \neq j)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Var}\left(\left\|\sum_{i=1}^{n} X_i\right\|_F^2\right) &= \sum_i \mathbb{E}\left[\|X_i\|_F^4\right] + 4\sum_{i<j} \mathbb{E}\left[\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right)^2\right] - \sum_i \left(\mathbb{E}\left[\|X_i\|_F^2\right]\right)^2 \\
&= \sum_i \mathbb{E}\left[\|X_i\|_F^4\right] + 2\sum_{i \neq j} \mathbb{E}\left[\left(\operatorname{Tr}\left(X_i^\top X_j\right)\right)^2\right] - \sum_i \left(\mathbb{E}\left[\|X_i\|_F^2\right]\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\sum_i\sum_j \mathbb{E}\left[\left(\mathrm{Tr}\left(X_i^\top X_j\right)\right)^2\right] \leq 2\sum_i\sum_j \mathbb{E}\left[\left(\mathrm{Tr}\left(X_i^\top X_i\right)\right)^2\right] \\
&= 2n^2\mathbb{E}\left[\|X_i\|_F^4\right].
\end{aligned}
$$

In the second last line we used the property of inner products, i.e. $\langle X_i, X_j\rangle \leq \langle X_i, X_i\rangle = \|X_i\|^2$ for all (i, j) pairs. Now taking the second term in (26)

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|_F^2\right] &= \mathbb{E}\left[\mathrm{Tr}\left(\sum_i X_i^\top \sum_j X_j\right)\right] \\
&= \mathrm{Tr}\left(\sum_i\sum_j \mathbb{E}\left[X_i^\top X_j\right]\right) \\
&= \mathrm{Tr}\left(\sum_i \mathbb{E}\left[X_i^\top X_i\right]\right) = \sum_i \mathbb{E}\left[\|X_i\|_F^2\right] = n\mathbb{E}\left[\|X_i\|_F^2\right].
\end{aligned}
$$

Plugging these results in (26)

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|_F^4\right] &\leq 2n^2\mathbb{E}\left[\|X_i\|_F^4\right] + n^2\left(\mathbb{E}\left[\|X_i\|_F^2\right]\right)^2 \\
&\leq 2n^2\mathbb{E}\left[\|X_i\|_F^4\right] + n^2\mathbb{E}\left[\|X_i\|_F^4\right] \qquad \text{(Jensen's inequality.)} \\
&= 3n^2\mathbb{E}\left[\|X_i\|_F^4\right].
\end{aligned}
$$

□