Abstract
This paper studies model-based bandit and reinforcement learning (RL) with nonlinear function approximations. We propose to study convergence to approximate local maxima because we show that global convergence is statistically intractable even for one-layer neural net bandit with a deterministic reward. For both nonlinear bandit and RL, the paper presents a model-based algorithm, Virtual Ascent with Online Model Learner (ViOlin), which provably converges to a local maximum with complexity bounds that only depend on the sequential Rademacher complexity of the model class. Our results imply novel sample complexity bounds on several concrete settings such as linear bandit with finite or sparse model class, and two-layer neural net bandit. A key algorithmic insight is that optimism may lead to over-exploration even for two-layer neural net model class. On the other hand, for convergence to local maxima, it suffices to maximize the virtual return if the model can also reasonably predict the gradient and Hessian of the real return.

1. Introduction
Recent progresses demonstrate many successful applications of deep reinforcement learning (RL) in robotics (Levine et al., 2016), games (Berner et al., 2019; Silver et al., 2017), computational biology (Mahmud et al., 2018), etc. However, theoretical understanding of deep RL algorithms is limited. Last few years witnessed a plethora of results on linear function approximations in RL (Zanette et al., 2020; Shariff, Szepesvári, 2020; Jin et al., 2020; Wang et al., 2019; 2020a; Du et al., 2019b; Agarwal et al., 2020a; Hao et al., 2021), but the analysis techniques appear to strongly rely on (approximate) linearity and hard to generalize to neural networks.

The goal of this paper is to theoretically analyze model-based nonlinear bandit and RL with neural net approximation, which achieves amazing sample-efficiency in practice (see e.g., (Janner et al., 2019; Clavera et al., 2019; Hafner et al., 2019a;b; Dong et al., 2020a)). We focus on the setting where the state and action spaces are continuous.

Past theoretical work on model-based RL studies families of dynamics with restricted complexity measures such as Eluder dimension (Osband, Roy, 2014), witness rank (Sun et al., 2019), the linear dimensionality (Yang, Wang, 2020), and others (Modi et al., 2020; Kakade et al., 2020; Du et al., 2021). Implications of these complexity measures have been studied, e.g., finite mixture of dynamics (Ayoub et al., 2020) and linear models (Russo, Roy, 2013) have bounded Eluder dimensions. However, it turns out that none of the complexity measures apply to the family of MDPs with even barely nonlinear dynamics, e.g., MDPs with dynamics parameterized by all one-layer neural networks with a single activation unit (and with bounded weight norms). For example, in Theorem C.2, we will prove that one-layer neural nets do not have polynomially-bounded Eluder dimension.

Given these strong impossibility results, we propose to reformulate the problem to finding an approximate local maximum policy with guarantees. This is in the same vein as the recent fruitful paradigm in non-convex optimization where researchers disentangle the problem into showing that all

1This result is also proved by the concurrent Li et al. (2021, Theorem 8) independently.
local minima are good and fast convergence to local minima (e.g., see (Ge et al., 2016; 2015; 2017; Ge, Ma, 2020; Lee et al., 2016)). In RL, local maxima can often be global as well for many cases (Agarwal et al., 2020b).2

Zero-order optimization or policy gradient algorithms can converge to local maxima and become natural potential competitors. They are widely believed to be less sample-efficient than the model-based approach because the latter can leverage the extrapolation power of the parameterized models. Theoretically, our formulation aims to characterize this phenomenon with results showing that the model-based approach’s sample complexity mostly depends (polynomially) on the complexity of the model class, whereas policy gradient algorithms’ sample complexity polynomially depend on the dimensionality of policy parameters (in RL) or actions (in bandit). Our technical goal is to answer the following question:

Can we design algorithms that converge to approximate local maxima with sample complexities that depend only and polynomially on the complexity measure of the dynamics/reward class?

We note that this question is open even if the dynamics hypothesis class is finite, and the complexity measure is the logarithm of its size. The question is also open even for nonlinear bandit problems (where dynamics class is replaced by reward function class), with which we start our research. We consider first nonlinear bandit with deterministic reward where the reward is concave in the input (action) (Amos et al., 2020) and has a sample complexity polynomially bounded by \( \frac{\Omega(s \sqrt{T})}{\epsilon} \) for a single action, \( \epsilon \)-approximate optimal action. In this case both zero-order optimization and the SquareCB algorithm in Foster and Rakhlin (2020) have sample complexity/regret that depend on the dimension of action space \( d_A \).

1. Linear bandit with finite parameter space \( \Theta \). Because \( \eta \) is concave in action \( a \), our result leads to a sample complexity \( \mathcal{O}(\log(\Theta) \log(T)) \) for finding an \( \epsilon \)-approximate optimal action. In this case both zero-order optimization and the SquareCB algorithm in Foster and Rakhlin (2020) have sample complexity/regret that depend on the dimension of action space \( d_A \).

2. Linear bandit with \( s \)-sparse or structured instance parameters. Our algorithm ViOLin achieves an \( \mathcal{O}(\text{poly}(s, 1/\epsilon)) \) sample complexity when the instance/model parameter is \( s \)-sparse and the reward is deterministic. The sample complexity of zero-order optimization depends polynomially on \( d_A \). Carpenter and Munos (2012) achieve a stronger \( \mathcal{O}(s \sqrt{T}) \) regret bound for \( s \)-sparse linear bandits with actions set \( A = S^{d-1} \). In contrast, our ViOLin algorithm applies more generally to any structured instance parameter set. Other related results either leverage the rather strong anti-concentration assumption on the action set (Wang et al., 2020b), or have implicit dimension dependency (Hao et al., 2020, Remark 4.3).

3. Two-layer neural nets bandit. Our algorithm finds an \( \epsilon \)-approximate optimal action \( \tilde{\Theta}_r \). Zero-order optimization can also find a local maximum but with \( \Omega(1) \) samples. Optimistic algorithms in this case have an exponential sample complexity (see Theorem C.3). Moreover, when the second layer of the ground-truth network contains all negative weights and the activation is convex and monotone, all local maxima are global because the reward is concave in the input (action) (Amos et al., 2017).

The results for bandit can be extended to model-based RL with deterministic nonlinear dynamics and deterministic reward.

Theorem 1.2 (Informal version of Theorem B.4). Consider RL problems with deterministic dynamics class and stochastic policy class that satisfy some Lipschitz properties. Suppose the sequential Rademacher complexity of the \( \ell_2 \) losses for learning the dynamics is bounded by \( \sqrt{\mathcal{R}(\Theta) \log(T)} \). Then, ViOLin for RL (Alg. 2) finds an \( \epsilon \)-approximate local maximum with \( \tilde{\Theta}(\Theta) \mathcal{O}(\mathcal{O}(\mathcal{R}(\Theta) \epsilon^{-8})) \) samples.

To the best of our knowledge, this is the first model-based RL algorithm with provable finite sample complexity guarantees (for local convergence) for general nonlinear dynamics. The work of (Luo et al., 2019) is the closest prior work which also shows local convergence, but its conditions likely cannot be satisfied by any parameterized models (including linear models). We also present a concrete example of RL problems with nonlinear models satisfying our Lipschitz

---

2The all-local-maxima-are-global condition only needs to hold to the ground-truth total expected reward function. This potentially can allow disentangled assumptions on the ground-truth instance and the hypothesis class.
assumptions in Example B.3 of Section B, which may also serve as a testbed for future model-based deep RL analysis. Other prior works on model-based RL do not apply to one-hidden-layer neural nets because they conclude global convergence which is not possible for one-hidden-layer neural nets in the worst case.

**Technical novelty: exploring by model-based curvature estimate.** The main challenge is that the optimism-in-face-of-uncertainty exploration principle seems to be too aggressive—for a nonlinear model class, even if the ground-truth model is linear, optimistic exploration leads to exponentially large model cumulative prediction errors and exponential sample complexity (see Theorem C.3). We use an online learning oracle to learn the dynamics, which can dramatically reduce the model prediction errors by hedging the risks. However, now, choosing actions or policies by maximizing virtual return may lack exploration. The main novelty of our algorithm is that we augment the online learning loss with gradient and Hessian estimation errors. The main insight is that, in order to ensure sufficient exploration for converging to local maxima, it suffices for the model to predict the virtual return, its gradient and Hessian reasonably accurately. We refer to the approach as “model-based curvature estimate”. Our algorithm alternates between maximizing virtual return (over action or policy) and learning the model parameters by an online learner.

Because the algorithm leverages model extrapolation, the sample complexity of model-based curvature prediction depends on the model complexity instead of action dimension in the zero-optimization approach. Foster, Rakhlin (2020) also propose algorithms that do not rely on UCB—their exploration strategy either relies on the finite discrete action space, or leverages the linear structure in the action space and has action-dimension dependency. In contrast, our algorithms’ exploration relies more on the learning of the model. Consequently, our sample complexity can be action-dimension-free.

### 2. Problem Setup and Preliminaries

In this section, we first introduce our problem setup for nonlinear bandit and reinforcement learning, and then the preliminary for online learning and sequential Rademacher complexity.

#### 2.1. Nonlinear Bandit Problem with Deterministic Reward

We consider deterministic nonlinear bandits with continuous actions. Let \( \theta \in \Theta \) be the parameter that specifies the bandit instance, \( a \in \mathbb{R}^{d_A} \) the action, and \( \eta(\theta, a) \in [0, 1] \) the reward function. Let \( \theta^* \) denote the unknown ground-truth parameter. Throughout the paper, we work under the realizability assumption that \( \theta^* \in \Theta \). A bandit algorithm aims to maximize \( \eta(\theta^*, a) \). Let \( a^* = \argmax_a \eta(\theta^*, a) \) be the optimal action (breaking tie arbitrarily). Let \( \|H\|_{sp} \) be the spectral norm of a matrix \( H \). We assume that the reward function, its gradient and Hessian matrix are Lipschitz, which are standard in the optimization literature (e.g., Johnson, Zhang (2013); Ge et al. (2015)).

**Assumption 2.1.** We assume that for all \( \theta \in \Theta \),
\[
\sup_a \|\nabla_a \eta(\theta, a)\|_2 \leq \zeta_\theta \quad \text{and} \quad \sup_a \|\nabla_a^2 \eta(\theta, a)\|_{sp} \leq \zeta_h.
\]
For every \( \theta \in \Theta \) and \( a_1, a_2 \in \mathbb{R}^{d_A} \),
\[
\|\nabla_a^2 \eta(\theta, a_1) - \nabla_a^2 \eta(\theta, a_2)\|_{sp} \leq \zeta_{3rd} \|a_1 - a_2\|_2.
\]

As a motivation to consider deterministic rewards, we prove in Theorem E.1 for a special case that no algorithm can find a local maximum in less than \( \sqrt{d_A} \) steps. The result implies that an action-dimension-free sample complexity bound is impossible with standard sub-Gaussian noise.\footnote{We rely on deterministic reward to estimate the gradient by finite difference. This method can be extended to stochastic rewards with multiple-point feedback (Liu et al., 2020).}

**Approximate local maxima.** In this paper, we aim to find a local maximum of the real reward function \( \eta(\theta^*, \cdot) \). A point \( x \) is an \((\epsilon_g, \epsilon_h)\)-approximate local maximum of a twice-differentiable function \( f(x) \) if \( \|\nabla f(x)\|_2 \leq \epsilon_g \) and \( \lambda_{max}(\nabla^2 f(x)) \leq \epsilon_h \). As argued in Sec. 1 and proved in Sec. C, because reaching a global maximum is computational and statistically intractable for nonlinear problems, we only aim to reach a local maximum.

**Sample complexity and local regret for converging to local maxima.** Let \( a_t \) be the action that the algorithm takes at time step \( t \). The sample complexity for converging to approximate local maxima is defined to be the minimal number of steps \( T \) such that there exists \( t \in [T] \) where \( a_t \) is an \((\epsilon_g, \epsilon_h)\) approximate local maximum with probability at least \( 1 - \delta \). We also define the “local regret” by comparing with an approximate local maximum. We defer the discussion to Appendix F.5.

#### 2.2. Reinforcement Learning

A finite horizon Markov decision process (MDP) with deterministic dynamics is defined by a tuple \((T, r, H, \mu_1)\). Let \( \mathcal{S} \) and \( A \) be the state and action spaces. The dynamics \( T : \mathcal{S} \times A \to \mathcal{S} \) gives next state, \( r : \mathcal{S} \times A \to [0, 1] \) is the reward function. \( H \) and \( \mu_1 \) denote the horizon and distribution of initial state respectively. Without loss of generality, we assume that the state space is disjoint for different time steps. That is, there exists disjoint sets \( S_1, \ldots, S_H \) such that \( \mathcal{S} = \bigcup_{h=1}^H S_h \), and for any \( s_h \in \mathcal{S}, a_h \in A, T(s_h, a_h) \in S_{h+1} \).

We consider parameterized policy and dynamics. Let \( \Pi = \{ \pi_\psi : \psi \in \Psi \} \) be the policy class and \( \{T_\theta : \theta \in \Theta \} \)
the dynamics class. The value function is \( V_T^\pi(s_h) \triangleq E[\sum_{i'=h}^T R(s_{i'}, a_{i'})] \), where \( a_{i'} \sim \pi(\cdot \mid s_{i'}, s_{i'+1} = T(s_{i'}, a_{i'})) \). Sharing the notation with the bandit setting, let \( \eta(\theta, \psi) = E_{s_1 \sim \pi^\theta} V_{T^\psi}(s_1) \) be the expected return of policy \( \pi^\psi \) under dynamics \( T^\theta \). Let \( \rho_{T}^\theta \) be the distribution of state action pairs when running policy \( \pi \) in dynamics \( T \). For simplicity, we do not distinguish \( \psi, \theta \) from \( \pi^\psi, T^\theta \) when the context is clear. For example, we write \( V_{\theta}^\psi \equiv V_{T^\psi}^\theta \).

2.3. Preliminary on Online Learning with Stochastic Input Components

Consider a prediction problem where we aim to learn a function that maps from \( X \) to \( Y \) parameterized by parameters in \( \Theta \). Let \( \ell((x, y) ; \theta) \) be a loss function that maps \((X \times Y) \times \Theta \to \mathbb{R}_+ \). An online learner \( \mathcal{R} \) aims to solve the prediction tasks under the presence of an adversarial nature iteratively. At time step \( t \), the following happens.

1. The learner computes a distribution \( p_t = \mathcal{R}(\{(x_i, y_i)\}^t_{i=1}) \) over the parameter space \( \Theta \).
2. The adversary selects a point \( x_t \in X \) (which may depend on \( p_t \)) and generates a sample \( \xi_t \) from some fixed distribution \( q \). Let \( x_t \triangleq (x_t, \xi_t) \), and the adversary picks a label \( y_t \in Y \).
3. The data point \((x_t, y_t)\) is revealed to the online learner.

The online learner aims to minimize the expected regret in \( T \) rounds of interactions, defined as

\[
\text{REG}^\text{OL}_T \triangleq \frac{1}{\mathcal{R}^T} \mathbb{E}_{\xi_t \sim q, \theta_t \sim p_t} \left[ \sum_{t=1}^T \ell((x_t, y_t) ; \theta) \right] - \inf_{\theta \in \Theta} \sum_{t=1}^T \ell((x_t, y_t) ; \theta) \tag{1}
\]

The difference of the formulation from the most standard online learning setup is that the \( \xi_t \) part of the input is randomized instead of adversarially chosen (and the learner knows the distribution of \( \xi_t \) before making the prediction \( p_t \)). It was introduced by Rakhlin et al. (2011), who considered a more generalized setting where the distribution \( q \) in round \( t \) can depend on \( \{x_1, \ldots, x_{t-1}\} \).

We adopt the notation from Rakhlin et al. (2011; 2015a) to define the (distribution-dependent) sequential Rademacher complexity of the loss function class \( \mathcal{L} = \{(x, y) \mapsto \ell((x, y) ; \theta) \mid \theta \in \Theta\} \). For any set \( Z \), a \( Z \)-valued tree with length \( T \) is a set of functions \( \{z_t : \{\pm 1\}^{t-1} \mapsto Z\}_{t=1}^T \). For a sequence of Rademacher random variables \( \epsilon = (\epsilon_1, \ldots, \epsilon_T) \) and for every \( 1 \leq t \leq T \), we denote \( z_t(\epsilon) \triangleq z_t(\epsilon_1, \ldots, \epsilon_{t-1}) \). For any \( Y \)-valued tree \( x \) and any \( Y \)-valued tree \( y \), we define the sequential Rademacher complexity as

\[
\mathcal{R}_T(\mathcal{L}; x, y) \triangleq \mathbb{E}_{\xi_t, \ldots, \xi_T, \epsilon} \left[ \sup_{\ell \in \mathcal{L}} \sum_{t=1}^T \epsilon_t \ell((x(\epsilon), \xi_t), y(\epsilon)) \right]. \tag{2}
\]

We also define \( \mathcal{R}_T(\mathcal{L}) = \sup_{x, y} \mathcal{R}_T(\mathcal{L}; x, y) \), where the supremum is taken over all \( X \)-valued and \( Y \)-valued trees. Rakhlit et al. (2011) proved the existence of an algorithm whose online learning regret satisfies \( \text{REG}^\text{OL}_T \leq 2\mathcal{R}_T(\mathcal{L}) \).

3. Main Results

Nonlinear bandits. Our algorithm is stated in Alg. 1 and explained in Section A. The following theorem gives the sample complexity bound.

**Theorem 3.1.** Let \( C_1 = 2 + \zeta_6^\epsilon \). Suppose the sequential Rademacher complexity of the loss function (defined in Eq. (5), Section A) is bounded by \( \sqrt{\mathcal{R}(\theta)T} \log \log(T) \). The sample complexity of Alg. 1 (for finding an \((\epsilon, 6\zeta_6^\epsilon)\)-approximate local maximum) is bounded by \( \tilde{O}(C_1^4 \mathcal{R}(\Theta) \max (\zeta_6^\epsilon \epsilon^{-8}, 2^5 \epsilon^2 \epsilon^{-6})) \).

We present a proof sketch in Section F.1. Proof of Theorem 3.1 is deferred to Section F.4. We can also boost the success probability by running Alg. 1 multiple times. In addition, we can prove that our algorithm enjoys a sublinear local regret. The theorem statement and proof is shown in Section F.5 and Section F.6 respectively.

As shown in Section 1, Theorem 3.1 implies several non-trivial results for (sparse) linear bandits and neural net bandits. We sketch these instantiations in Section A.1 and defer the proofs to Section E.7.

Model-based reinforcement learning. We can also extend the results to model-based reinforcement learning with deterministic dynamics and reward. To reason about the learning about local steps and the dynamics, we require several additional Lipschitzness assumptions. We defer the discussion to Section B.

4. Conclusion

In this paper, we design new algorithms whose sample complexity toward local maxima are bounded by the sequential Rademacher complexity of particular loss functions. By rearranging the priorities of exploration versus exploitation, our algorithms avoid over-aggressive explorations caused by the optimism in the face of uncertainty principle, and hence apply to nonlinear models and dynamics. We raise the following questions as future works.

1. Since we mainly focus on proving a regret bound that depends only on the complexity of dynamics/reward class, our convergence rate in \( \epsilon \) is likely not minimax optimal. Can our algorithms (or analysis) be modified to achieve minimax optimal regret for some of the instantiations such as sparse linear bandit and linear bandit with finite model class?
2. In the bandit setting, we focus on deterministic reward because our ViOlin algorithm relies on finite difference to estimate the gradient and Hessian of reward function. In fact, Theorem E.1 shows that action-dimension-free regret bound for linear models is impossible under standard Gaussian noise. Can we extend our algorithm to stochastic environments with additional assumptions on noises?

3. In the reinforcement learning setting, we use policy gradient lemma to upper bound the gradient/Hessian loss by the dynamics loss, which inevitable require the policies being stochastic. Despite the success of stochastic policies in deep reinforcement learning, the optimal policy may not be stochastic. Can we extend the ViOlin algorithm to reinforcement learning problems with deterministic policy hypothesis?

Acknowledgment

The authors would like to thank Yuanhao Wang, Daogao Liu, Zhizhou Ren, Jason D. Lee, Colin Wei, Akshay Krishnamurthy, Alekh Agarwal and Csaba Szepesvári for helpful discussions. TM is also partially supported by the Google Faculty Award, Lam Research, and JD.com.

References


Provable Model-based Nonlinear Bandit and Reinforcement Learning


Faury Louis, Abeille Marc, Calauzènes Clément, Fercoq Olivier. 

Filippi Sarah, Cappé Olivier, Garivier Aurélien, Szepesvári Csaba. 

Foster Dylan, Rakhlin Alexander. 


Ge Rong, Huang Furong, Jin Chi, Yuan Yang. 

Ge Rong, Jin Chi, Zheng Yi. 

Ge Rong, Lee Jason D, Ma Tengyu. 

Ge Rong, Ma Tengyu. 

Hafner Danijar, Lillicrap Timothy, Ba Jimmy, Norouzi Mohammad. 

Hafner Danijar, Lillicrap Timothy, Fischer Ian, Villegas Ruben, Ha David, Lee Honglak, Davidson James. 

Hao Botao, Lattimore Tor, Szepesvári Csaba, Wang Mengdi. 

Hao Botao, Lattimore Tor, Wang Mengdi. 

Hazan Elad, Levy Kfir, Shalev-Shwartz Shai. 

Hazan Elad, Singh Karan, Zhang Cyril. 

Hsu Daniel, Kakade Sham, Zhang Tong, others. 

Janner Michael, Fu Justin, Zhang Marvin, Levine Sergey. 

Jiang Nan, Krishnamurthy Akshay, Agarwal Alekh, Langford John, Schapire Robert E. 

Jin Chi, Allen-Zhu Zeyuan, Bubeck Sebastien, Jordan Michael I. 


Johnson Rie, Zhang Tong. 


Li Lihong. A unifying framework for computational reinforcement learning theory. 2009.


Provable Model-based Nonlinear Bandit and Reinforcement Learning


A. Model-based Algorithms for Nonlinear Bandit

In this section we study model-based algorithms for nonlinear continuous bandits problem, which is a simplification of model-based reinforcement learning. We use the notations and setup in Section 2.1.

Abstraction of analysis for model-based algorithms. Typically, a model-based algorithm explicitly maintains an estimated model \( \hat{\theta}_t \), and sometimes maintains a distribution, posterior, or confidence region of \( \hat{\theta}_t \). We will call \( \eta(\theta^*, a) \) the \textit{real reward} of action \( a \), and \( \eta(\hat{\theta}_t, a) \) the \textit{virtual reward}. Most analysis for model-based algorithms (including UCB and ours) can be abstracted as showing the following two properties:

(i) the virtual reward \( \eta(\hat{\theta}_t, a_t) \) is sufficiently high.
(ii) the virtual reward \( \eta(\hat{\theta}_t, a_t) \) is close to the real reward \( \eta(\theta^*, a_t) \) in the long run.

One can expect that a proper combination of property (i) and (ii) leads to showing the real reward \( \eta(\theta^*, a_t) \) is high in the long run. Before describing our algorithms, we start by inspecting and summarizing the pros and cons of UCB from this viewpoint.

Pros and cons of UCB. The UCB algorithm chooses an action \( a_t \) and an estimated model \( \hat{\theta}_t \) that maximize the virtual reward \( \eta(\hat{\theta}_t, a_t) \) among those models agreeing with the observed data. The pro is that it satisfies property (i) by definition—\( \eta(\hat{\theta}_t, a_t) \) is higher than the optimal real reward \( \eta(\theta^*, a^*) \). The downside is that ensuring (ii) is challenging and often requires strong complexity measure bound such as Eluder dimension (which is not polynomial for even barely nonlinear models, as shown in Theorem C.2). The difficulty largely stems from our very limited control of \( \hat{\theta}_t \) except its consistency with the observed data. To bound the difference between the real and virtual rewards, we essentially require that any model that agrees with the past history should extrapolate to any future data accurately (as quantitatively formulated in Eluder dimension). Moreover, the difficulty of satisfying property (ii) is fundamentally caused by the over-exploration of UCB—As shown in the Theorem C.3, UCB suffers from bad sample complexity with barely nonlinear family of models.

Our key idea: natural exploration via model-based curvature estimate. We deviate from UCB by readjusting the priority of the two desiderata. We prioritize ensuring property (ii) on large model class. We leverage a strong online learning algorithm to predict \( \hat{\theta}_t \) with the objective that \( \eta(\hat{\theta}_t, a_t) \) matches \( \eta(\theta^*, a_t) \). As a result, the difference between the virtual and real reward depends on the online learnability or the sequential Rademacher complexity of the model class. Sequential Rademacher complexity turns out to be a fundamentally more relaxed complexity measure than Eluder dimension—e.g., two-layer neural networks’ sequential Rademacher complexity is polynomial in parameter norm and dimension, but the Eluder dimension is at least exponential in dimension (even with a constant parameter norm). However, an immediate consequence of using online-learned \( \hat{\theta}_t \) is that we lose optimism/exploration that ensured property (i).4

Algorithm 1 ViOlin: Virtual Ascent with Online Model Learner (for Bandit)

1: Set parameter \( \kappa_1 = 2\zeta_p \) and \( \kappa_2 = 6\log\sqrt{2}\zeta_h \). Let \( \mathcal{H}_0 = \emptyset \); choose \( a_0 \in \mathcal{A} \) arbitrarily.
2: for \( t = 1, 2, \ldots \) do
3: Run \( R \) on \( \mathcal{H}_{t-1} \) with loss function \( \ell \) (defined in equation (5)) and obtain \( p_t = R(\mathcal{H}_{t-1}) \).
4: Let \( a_t \leftarrow \arg\max_a \mathbb{E}_{x_t \sim p_t}[\eta(\theta^*, a)] \).
5: Sample \( u_t, v_t \sim \mathcal{N}(0, I_{d_A \times d_A}) \) independently.
6: Let \( \xi_t = (u_t, v_t), \bar{x}_t = (a_t, a_{t-1}) \), and \( x_t = (\bar{x}_t, \xi_t) \).
7: Compute \( y_t = [\eta(\theta^*, a_t), \eta(\theta^*, a_{t-1}), \langle \nabla_a \eta(\theta^*, a_{t-1}), u_t \rangle, \langle \nabla^2_a \eta(\theta^*, a_{t-1}) u_t, v_t \rangle] \in \mathbb{R}^4 \) by applying a finite number of actions in the real environments using equation (3) and (4).
8: Update \( \mathcal{H}_t = \mathcal{H}_{t-1} \cup \{(x_t, y_t)\} \).
9: end for

Our approach realizes property (i) in a sense that the virtual reward will improve iteratively if the real reward is not yet near a local maximum. This is much weaker than what UCB offers (i.e., that the virtual reward is higher than the optimal real reward), but suffices to show the convergence to a local maximum of the real reward function. We achieve this by demanding the estimated model \( \hat{\theta}_t \) not only to predict the real reward accurately, but also to predict the gradient \( \nabla_a \eta(\theta^*, a) \)

\footnote{More concretely, the algorithm can get stuck when (1) \( a_t \) is optimal for \( \hat{\theta}_t \), (2) \( \hat{\theta}_t \) fits actions \( a_t \) (and history) accurately, but (3) \( \hat{\theta}_t \) does not fit \( a^* \) (because online learner never sees \( a^* \)). The passivity of online learning formulation causes this issue—the online learner is only required to predict well for the point that it saw and will see, but not for those points that it never observes.}
and Hessian $\nabla^2_a \eta(\theta^*, a)$ accurately. In other words, we augment the loss function for the online learner so that the estimated model satisfies $\eta(\hat{\theta}_t, a_t) \approx \eta(\theta^*, a_t)$, $\nabla_a \eta(\hat{\theta}_t, a_t) \approx \nabla_a \eta(\theta^*, a_t)$, and $\nabla^2_a \eta(\hat{\theta}_t, a_t) \approx \nabla^2_a \eta(\theta^*, a_t)$ in the long run. This implies that when $a_t$ is not at a local maximum of the real reward function $\eta(\theta^*, \cdot)$, then it’s not at a maximum of the virtual reward $\eta(\hat{\theta}_t, \cdot)$, and hence the virtual reward will improve in the next round if we take the greedy action that maximizes it.

**Estimating projections of gradients and Hessians.** To guide the online learner to predict $\nabla_a \eta(\theta^*, a_t)$ correctly, we need a supervision for it. However, we only observe the reward $\eta(\theta^*, a_t)$. Leveraging the deterministic reward property, we use rewards at $a$ and $a + \alpha_1 u$ to estimate the projection of the gradient at a random direction $u$:

$$\langle \nabla_a \eta(\theta^*, a), u \rangle = \lim_{\alpha_1 \to 0} \frac{(\eta(\theta^*, a + \alpha_1 u) - \eta(\theta^*, a)) / \alpha_1}{\alpha_1}$$

It turns out that the number of random projections $\langle \nabla_a \eta(\theta^*, a), u \rangle$ needed for ensuring a large virtual gradient does not depend on the dimension, because we only use these projections to estimate the norm of the gradient but not necessarily the exact direction of the gradient (which may require $d$ samples.) Similarly, we can also estimate the projection of Hessian to two random directions $u, v \in \mathcal{A}$ by:

$$\langle \nabla_a^2 \eta(\theta^*, a), u, v \rangle = \lim_{\alpha_2 \to 0} \frac{(\langle \nabla_a \eta(\theta^*, a + \alpha_2 v), u \rangle - \langle \nabla_a \eta(\theta^*, a), u \rangle) / \alpha_2}{\alpha_2}$$

Algorithmically, we can choose infinitesimal $\alpha_1$ and $\alpha_2$. Note that $\alpha_1$ should be at least an order smaller than $\alpha_2$ because the limitations are taken sequentially.

We create the following prediction task for an online learner: let $\theta$ be the parameter, $x = (a, a', u, v)$ be the input, $\hat{y} = [\eta(\theta, a), \eta(\theta, a'), \langle \nabla_a \eta(\theta, a), u \rangle, \langle \nabla_a \eta(\theta, a'), u \rangle] \in \mathbb{R}^4$ be the output, and $y = [\eta(\theta^*, a), \eta(\theta^*, a'), \langle \nabla_a \eta(\theta^*, a), u \rangle, \langle \nabla_a \eta(\theta^*, a'), u \rangle] \in \mathbb{R}^4$ be the supervision, and the loss function be

$$\ell((a, a', u, v), y; \theta) \triangleq \frac{1}{2} (\|\hat{y}_1 - [y]_1\|^2 + \|\hat{y}_2 - [y]_2\|^2 + \min (\kappa_1^2, (\|\hat{y}_3 - [y]_3\|^2) + \min (\kappa_2^2, (\|\hat{y}_4 - [y]_4\|^2))$$

Here we used $[y]_i$ to denote the $i$-th coordinate of $y \in \mathbb{R}^4$ to avoid confusing with $y_t$ (the supervision at time $t$.) Our algorithm is formally stated in Alg. 1 with its sample complexity bound below.

**Theorem 3.1.** Let $C_1 = 2 + \zeta_\eta / \zeta_\Theta$. Suppose the sequential Rademacher complexity of the loss function (defined in Eq. (5), Section A) is bounded by $\sqrt{R(\Theta) T \log \log (T)}$. The sample complexity of Alg. 1 (for finding an $(\epsilon, 6\sqrt{\zeta_{\Theta \|	heta\|}})$-approximate local maximum) is bounded by $\tilde{O}(C_1^4 R(\Theta) \max (\zeta_{\Theta}^4 \epsilon^{-8}, c_2^2 \eta^{-6}))$.

We present a proof sketch in Appendix F.1. Proof of Theorem 3.1 is deferred to Appendix F.4. We can also boost the success probability by running Alg. 1 multiple times. In addition, we can prove that our algorithm enjoys a sublinear local regret. The theorem statement and proof is shown in Appendix F.5 and Appendix F.6 respectively.

**A.1. Instantiations of Theorem 3.1**

In the sequel we sketch some instantiations of our main theorem, whose proofs are deferred to Appendix F.7.

**Linear bandit with finite model class.** Consider a linear bandit problem with action set $\mathcal{A} = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$ and finite model class $\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$. Let $\eta(\theta, a) = \langle \theta, a \rangle$ be the reward. We deal with the constrained action set by using a surrogate loss $\hat{\eta}(\theta, a) = \eta(\theta, a) - \frac{1}{2} \|a\|_2^2$ and apply Theorem F.3 with reward $\hat{\eta}$. We claim that the sample complexity (of finding an $\epsilon$-suboptimal action for $\eta(\theta^*, \cdot)$) is bounded by $\tilde{O}(\log (\Theta) \epsilon^{-8})$. Note that the bound is independent of the dimension $d$. By contrast, the SquareCB algorithm (Foster, Rakhlin, 2020) depends polynomially on $d$ (see Theorem 7 of (Foster, Rakhlin, 2020)). Zero-order optimization approach (Duchi et al., 2015) in this case also gives a poly($d$) regret bound.

**Linear bandit with sparse or structured model vectors.** We consider the deterministic linear bandit setting where the model class $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq s, \|\theta\|_2 = 1\}$ consists of all $s$-sparse vectors on the unit sphere. Similarly to finite

---

3Recall that in bandit literature, action $a$ is an $\epsilon$-approximate optimal action if $\eta(\theta^*, a) \geq \eta(\theta^*, a^*) - \epsilon$. 
hypothesis case, we claim that the sample complexity of Alg. 1 is \( \tilde{O}(\log(\Theta|\epsilon^{-8}\log(d))) \). The sample complexity of our algorithm only depends on the sparsity level \( s \) (up to logarithmic factors), whereas the Eluder dimension of sparse linear hypothesis is still \( \Omega(d) \) (see Lemma F.4). Lattimore, Szepesvári (2020) show a \( \Omega(d) \) sample complexity lower bound for the sparse linear bandit problem with stochastic reward. But here we only consider a deterministic reward and continuous action.

Moreover, we can further extend the result to other linear bandit settings where \( \theta \) has an additional structure. Suppose \( \Theta = \{ \theta = \phi(z) : z \in \mathbb{R}^d \} \) for some Lipschitz function \( \phi \). Then, a similar approach gives sample complexity bound that only depends on \( s \) but not \( d \) (up to logarithmic factors). Our results can also be extended to non-linear bandits such as deterministic logistic bandits.

**Two-layer neural nets.** We consider the reward function given by two-layer neural networks with width \( m \). For matrices \( W_1 \in \mathbb{R}^{m \times d} \) and \( W_2 \in \mathbb{R}^{1 \times m} \), let \( \eta((W_1, W_2), a) = W_2\sigma(W_1a) - \frac{1}{2}\|a\|^2_2 \) for some nonlinear function \( \sigma : \mathbb{R} \to [0, 1] \) with bounded derivatives up to the third order. Recall that the \( (1, \infty) \)-norm of \( W_1^\top \) is defined by \( \max_{|i| \leq m} \sum_{j=1}^{d} |[W_1]_{i,j}| \). That is, the max 1-norm of the rows of \( W_1 \). Let the model hypothesis space be \( \Theta = \{(W_1, W_2) : \|W_1\|_{1,\infty} \leq 1, \|W_2\|_1 \leq 1 \} \) and \( \theta = (W_1, W_2) \). We claim that Alg. 1 finds an \( (\epsilon, 6\sqrt{3d}\epsilon) \)-approximate local maximum in \( O(\epsilon^{-8}\log(d)) \) steps. To the best of our knowledge, this is the first result analyzing nonlinear bandit with neural network parameterization. The result follows from analyzing the sequential Rademacher complexity for \( \eta, \langle \nabla_{\theta} \eta, u \rangle \), and \( \langle u, \nabla_{\theta}^2 \eta \cdot v \rangle \), and finally the resulting loss function \( \ell \). See Theorem F.6 in Section F.7 for details. We remark here that zero-order optimization in this case has a \( \text{poly}(d) \) sample complexity. In addition, if the second layer of the neural network \( W_2 \) contains all negative entries, and the activation function \( \sigma \) is monotone and convex, then \( \eta((W_1, W_2), a) \) is concave in the action. (This is a special case of input convex neural networks (Amos et al., 2017).) In this case, Alg. 1 finds an \( \epsilon \)-suboptimal action (see Theorem F.6).

### B. Model-based Reinforcement Learning

In this section, we extend the results in Section A to model-based reinforcement learning with deterministic dynamics and reward function.

We can always view a model-based reinforcement learning problem with parameterized dynamics and policy as a nonlinear bandit problem in the following way. The policy parameter \( \psi \) corresponds to the action \( a \) in bandit, and the dynamics parameter \( \theta \) corresponds to the model parameter \( \theta \) in bandit. The expected total return \( \eta(\theta, \psi) = \mathbb{E}_{s_1, \ldots, s_T} V_{\theta \psi}(s_1) \) is the analogue of reward function in bandit. We intend to make the same regularity assumptions on \( \eta \) as in the bandit case (that is, Assumption 2.1) with \( a \) being replaced by \( \psi \). However, when the policy is deterministic, the reward function \( \eta \) has Lipschitz constant with respect to \( \psi \) that is exponential in \( H \) (even if dynamics and policy are both deterministic with good Lipschitzness). This prohibits efficient optimization over policy parameters. Therefore we focus on stochastic policies in this section, for which we expect \( \eta \) and its derivatives to be Lipschitz with respect to \( \psi \).

Blindly treating RL as a bandit only utilizes the reward but not the state observations. In fact, one major reason why model-based methods are more sample efficient is that it supervises the learning of dynamics by state observations. To reason about the learning about local steps and dynamics, we make the following additional Lipschitzness of value functions w.r.t. the states and Lipschitzness of policies w.r.t. its parameters, beyond those assumptions for the total reward \( \eta(\theta, \psi) \) in Assumption 2.1.

**Assumption B.1.** We assume the following (analogous to Assumption 2.1) on the value function: \( \forall \psi \in \Psi, \theta \in \Theta, s, s' \in S \) we have \( |V_{\theta \psi}^\psi(s) - V_{\theta \psi}^\psi(s')| \leq L_0\|s - s'\|_2 \); \( \|\nabla_{\theta} V_{\theta \psi}^\psi(s) - \nabla_{\theta} V_{\theta \psi}^\psi(s')\|_2 \leq L_1\|s - s'\|_2 \); \( \|\nabla_{\psi}^2 V_{\theta \psi}^\psi(s) - \nabla_{\psi}^2 V_{\theta \psi}^\psi(s')\|_{sp} \leq L_2\|s - s'\|_2 \).

**Assumption B.2.** We assume the following Lipschitzness assumptions on the stochastic policies parameterization \( \pi_{\psi, \theta} \):
\[
\|E_{a \sim \pi_{\psi, \theta}(a | s)} [(\nabla_{\psi} \log \pi_{\psi, \theta}(a | s)) (\nabla_{\theta} V_{\theta \psi}^\psi(a | s))^2]\|_{sp} \leq \chi_g; \|E_{a \sim \pi_{\psi, \theta}(a | s)} [(\nabla_{\psi} \log \pi_{\psi, \theta}(a | s))^2]\|_{sp} \leq \chi_f; \|E_{a \sim \pi_{\psi, \theta}(a | s)} [(\nabla_{\psi} \log \pi_{\psi, \theta}(a | s)) (\nabla_{\theta} V_{\theta \psi}^\psi(a | s))^2]\|_{sp} \leq \chi_h.
\]

Our results depends polynomially on the parameters \( L_0, L_1, L_2, \chi_g, \chi_f, \chi_h \). To demonstrate that the Assumption B.1 and B.2 contain interesting RL problems with nonlinear models and stochastic policies, we give the following example where these parameters are all on the order of \( O(1) \).

**Example B.3.** Let state space \( S \) be the unit ball in \( \mathbb{R}^d \) and action space \( A \) be \( \mathbb{R}^d \). The (deterministic) dynamics \( T \) is given\(^6\) by

\( T_a s = \sum_{u \in S} \langle A, u^a \rangle s \).

\(^6\)Recall that the injective norm of a \( k \)-th order tensor \( A \in \mathbb{R}^{d \times k} \) is defined as \( \|A\|_{sp} = \sup_{a \in S, |a| = k} \langle A, u^a \rangle \).
We consider a family of stochastic Gaussian policies with the mean being linear in the state: $\pi(s) = N(\psi s, \sigma^2 I)$, parameterized by $\psi \in \mathbb{R}^{d \times 1}$ with $\|\psi\|_{\text{op}} \leq 1$. We consider $\sigma \in (0, 1)$ as a small constant on the order of 1.

In this setting, Assumption 2.1, B.1, and B.2 hold with all parameters $\zeta_0, \zeta_0^2, \zeta_0^3, L_0, L_1, L_2, \chi_0, \chi_f$ and $\chi_h$ bounded by $\text{poly}(\sigma, 1/\sigma, H, L_r)$.

Bounding the Lipschitz parameters is highly nontrivial and deferred to Appendix H.

We show that the difference of gradient and Hessian of the total reward can be upper-bounded by the difference of dynamics. Let $\tau_i = (s_1, a_1, \cdots, s_H, a_H)$ be a trajectory sampled from policy $\pi_{\psi_i}$ under the real dynamics $T_{\theta}$. Similarly to (Yu et al., 2020), when the value function is Lipschitz, we upper bound $\Delta_{\epsilon, 1} = |\eta(\theta, \psi_i) - \eta(\theta^*, \psi_i)|$ by the one-step model prediction errors. Similarly, we can upper bound the gradient and Hessian errors by model prediction errors. As a result, the loss function simply can be set to

$$\ell((\tau_i, \tau'_i); \theta) = \sum_{(s_h, a_h) \in \tau_i} \|T_{\theta}(s_h, a_h) - T_{\theta^*}(s_h, a_h)\|_2^2 + \sum_{(s'_h, a'_h) \in \tau'_i} \|T_{\theta}(s'_h, a'_h) - T_{\theta^*}(s'_h, a'_h)\|_2^2$$

for two trajectories $\tau, \tau'$ sampled from policy $\pi_{\psi_i}$ and $\pi_{\psi_i - 1}$, respectively. Compared to Alg. 1, the loss function is here simpler without relying on finite difference techniques to query gradients projections. Our algorithm for RL is analogous to Alg. 2 in Appendix G. Main theorem for Alg. 2 is shown below.

**Theorem B.4.** Let $c_1 = HL_0^4(4H^2\chi_h + 4H^4\chi_f + 2H^2\chi_g + 1) + HL_1^4(8H^2\chi_g + 2) + 4HL_2^2$ and $C_1 = 2 + \frac{c_1}{\sqrt{\sigma}}$. Suppose the sequential Rademacher complexity of the loss function (defined in Eq. (6)) is bounded by $\sqrt{R(\Theta)}T \text{polylog}(T)$. The sample complexity of Alg. 2 (for finding an $(\epsilon, 6\sqrt{3\text{rd}}\epsilon)$-approximate local maximum) is bounded by $O(c_1\Gamma c_1 R(\Theta) \max(\zeta_0^2, \zeta_0^3, \epsilon^{-2}))$.

**Instantiation of Theorem B.4 on Example B.3.** Applying Theorem B.4 to the Example B.3 with $\sigma = \Theta(1)$ we get the sample complexity upper bound $O(\text{poly}(\sigma, 1/\sigma, H, L_r) \log(\Theta)\epsilon^{-6})$.

**Comparison with policy gradient.** To the best of our knowledge, the best analysis for policy gradient (Williams, 1992) shows convergence to a local maximum with a sample complexity that depends polynomially on $\|\nabla_\psi \log \pi_{\psi}(a \mid s)\|_2$ (Agarwal et al., 2020b). For the instance in in Example B.3, this translates to a sample complexity guarantee on the order of $\sqrt{\sigma}/\sigma$. In contrast, our sample complexity is independent of the dimension $d$. Instead, our bound depends on the complexity of the model family $\Theta$ which could be much smaller than the ambient dimension—this demonstrates that we leverage the model extrapolation.

**C. Lower Bounds**

We prove several lower bounds to show (a) the hardness of finding global maxima, and (b) the inefficiency of using optimism in nonlinear bandit.

**Hardness of global optimality.** We show it statistically intractable to find the global optimal policy when the function class is chosen to be the neural networks with ReLU activation.

**Theorem C.1.** When the function class is chosen to be one-layer neural networks with ReLU activation, the minimax sample complexity is $\Omega(\epsilon^{-(d-2)})$.

We can also prove that the eluder dimension of the constructed reward function class is exponential.

**Theorem C.2.** The $\epsilon$-eluder dimension of one-layer neural networks is at least $\Omega(\epsilon^{-(d-1)})$.

This result is concurrently established by Li et al. (2021, Theorem 8). The proofs of both theorems are deferred to Appendices E.1 and E.2, respectively.

**Inefficiency caused by optimism in nonlinear models.** The next theorem states that the UCB algorithm that uses optimism-in-face-of-uncertainty principle can overly explore in the action space, even if the ground-truth is simple. Proof of the theorem is deferred to Appendix E.3.
Theorem C.3. Consider the case where the ground-truth reward function is linear: \( \langle \theta^*, a \rangle \) and the action set is \( a \in S^{d-1} \). If the hypothesis is chosen to be two-layer neural network with width \( d \), UCB algorithm with tightest upper confidence bound suffers exponential sample complexity.

D. Additional Related Work

There are several provable efficient algorithms without optimism for contextual bandits. Dudik et al. (2011); Agarwal et al. (2014) instantiate mirror descent for contextual bandits, and the regret depend polynomially on the number of actions. Foster, Rakhlin (2020) and Simchi-Levi, Xu (2020) exploit the exploration probability inversely depends on the empirical gap. The SquareCB algorithm (Foster, Rakhlin, 2020) also extends to infinite actions with linear structure, but the regret depends on the action dimension. Recently, Foster et al. (2020) prove an instance-dependent regret bound for contextual bandit.

Deterministic nonlinear bandits can also be formulated as zero-order optimization without noise (see Duchi et al. (2015); Liu et al. (2020) and references therein), where the reward is assumed to be any 1-Lipschitz function. In contrast, our algorithm exploits the knowledge of the reward function parameterization and achieves an action-dimension-free sample complexity. For stochastic nonlinear bandit, Filippi et al. (2010) consider generalized linear model. Valko et al. (2013); Zhou et al. (2020) focus on rewards in a RKHS or NTK. Yang et al. (2020) extends this approach to reinforcement learning. For stochastic nonlinear bandit, Filippi et al. (2010) consider generalized linear model. Valko et al. (2013); Zhou et al. (2020) focus on rewards in a RKHS or NTK. Yang et al. (2020) extends this approach to reinforcement learning.

Another line of research focuses on solving reinforcement learning by optimization on the policy space. Agarwal et al. (2020b) prove that natural policy gradient can solve tabular MDPs efficiently. Cai et al. (2020) incorporate exploration bonus in proximal policy optimization algorithm and achieves polynomial regret in linear MDP setting.

Beyond linear function approximations, there are also extensive studies on various settings that allow efficient algorithms. For example, rich observation MDPs (Krishnamurthy et al., 2016; Dann et al., 2018; Du et al., 2019a; Misra et al., 2020), state aggregation (Dong et al., 2019b; Li, 2009), Bellman rank (Jiang et al., 2016; Dong et al., 2020b) and others (Du et al., 2021; Littman et al., 2001; Munos, 2005; Wan et al., 2020; Kakade et al., 2020).

E. Missing Proofs in Section C

In this section, we prove several negative results.

E.1. Proof and Remarks of Theorem C.1

Proof. We consider the class \( \mathcal{I} = \{ I_{\theta, \varepsilon} : \| \theta \|_2 \leq 1, \varepsilon > 0 \} \) of infinite-armed bandit instances, where in the instance \( I_{\theta, \varepsilon} \), the reward of pulling action \( x \in \mathbb{B}_2^d(1) \) is deterministic and is equal to

\[
\eta(I_{\theta, \varepsilon}, x) = Ax \{ \langle x, \theta \rangle - 1 + \varepsilon, 0 \}.
\]

We prove the theorem by proving the minimax regret. The sample complexity then follows from the canonical sample complexity-regret reduction (Jin et al., 2018, Section 3.1). Let \( \mathcal{A} \) denote any algorithm. Let \( R^T_{\mathcal{A}, I} \) be the \( T \)-step regret of algorithm \( \mathcal{A} \) under instance \( I \). Then we have

\[
\inf_{\mathcal{A}} \sup_I \mathbb{E}[R^T_{\mathcal{A}, I}] \geq \Omega(T^{\frac{d}{d+2}}).
\]

Fix \( \varepsilon = c \cdot T^{1/(d-1)} \). Let \( \Theta \) be an \( \varepsilon \)-packing of the sphere \( \{ x \in \mathbb{R}^d : \| x \|_2 = 1 \} \). Then we have \( |\Theta| \geq \Omega(\varepsilon^{-(d-1)}) \). So we choose \( c > 0 \) to be a numeric constant such that \( T \leq |\Theta|/2 \). Let \( \mu \) be the distribution over \( \Theta \) such that \( \mu(\theta) = \frac{1}{|\Theta|} \) for every \( \theta \in \Theta \). Note that for any action \( a \in \mathbb{B}_2^d(1) \), there is at most one \( \theta \in \Theta \) such that \( \eta(I_{\theta, \varepsilon}, a) \neq 0 \), because \( \Theta \) is a packing. Since \( T \leq |\Theta|/2 \), there exists \( \theta^* \in \Theta \) such that \( \mu(\theta^*) \leq 1/2 \). Therefore, with probability \( 1/2 \), the algorithm \( \mathcal{A} \) would obtain reward \( r_t = \eta(I_{\theta^*, \varepsilon}, a) = 0 \) for every time step \( t = 1, \ldots, T \). Note that under instance \( I_{\theta^*, \varepsilon} \), the optimal action is to choose \( a_t \equiv \theta^* \), which would give reward \( r_t^* \equiv \varepsilon \). Therefore, with probability \( 1/2 \), we have \( \mathbb{E}[R^T_{\mathcal{A}, I_{\theta^*, \varepsilon}}] \geq \varepsilon T/2 \geq \Omega(T^{\frac{d}{d+2}}) \).

We also note that Theorem C.1 does require ReLU activation, because if the ReLU function is replaced by a strictly monotone link function with bounded derivatives (up to third order), it is the setting of deterministic generalized linear bandit problem, which does allow a global regret that depends polynomially on dimension (Filippi et al., 2010; Dong et al., 2019a; Li et al., 2020).
We first provide a proof sketch to the theorem. We consider the following reward function.

We formalize UCB algorithm under deterministic environments as follows. At every time step
With two-layer neural networks, we can relax the use of ReLU activation—Theorem C.2 holds with two-layer neural

\[ \Theta = \{ \theta \} \]

The hypothesis space is \( \eta \) be the set of parameters that is consistent with \( \Theta \) being centered at \( \theta \). Let \( \eta \) represents a linear reward. In the following we use \( \theta \) as \( \{ \theta \} \), \( (\theta, \alpha) \) = \( \langle \theta, \alpha \rangle \). Theorem C.2 holds with two-layer neural networks and leaky-ReLU activations (Xu et al., 2015) because \( O(1) \) leaky-ReLU can implement a ReLU activation. We conjecture that with more layers, the impossibility result also holds for a broader sets of activations.

E.2. Proof of Theorem C.2

**Proof.** We adopt the notations from Appendix E.1. We use \( \dim_{\mathcal{F}}(\mathcal{F}, \varepsilon) \) to denote the \( \varepsilon \)-cludder dimension of the function class \( \mathcal{F} \). Let \( \Theta \) be an \( \varepsilon \)-packing of the sphere \( \{ x \in \mathbb{R}^d : \| x \|_2 = 1 \} \). We write \( \Theta = \{ \theta_1, \ldots, \theta_n \} \). Then we have \( n \geq \Omega(\varepsilon^{-(d-1)}) \).

Next we establish that \( \dim_{\mathcal{F}}(\mathcal{F}, \varepsilon) \geq \Omega(\varepsilon^{-(d-1)}) \). For each \( i \in [n] \), we define the function \( \eta_i(a) = \eta(I_{\theta_i}, \varepsilon, a) \) \( \in \mathcal{F} \). Then for \( i \leq n-1 \), we have \( \eta_i(a_j) = \eta_{i+1}(a_j) \) for \( j \leq i \), while \( \varepsilon = \eta_i(a_i) \neq \eta_{i+1}(a_i) = 0 \). Therefore, \( \eta \) is \( \varepsilon \)-independent of its predecessors. As a result, we have \( \dim_{\mathcal{F}}(\mathcal{F}, \varepsilon) \geq n-1 \). \( \square \)

E.3. Proof of Theorem C.3

First of all, we review the UCB algorithm in deterministic environments.

We formalize UCB algorithm under deterministic environments as follows. At every time step \( t \), the algorithm maintains a upper confidence bound \( C_t : \mathcal{A} \to \mathbb{R} \). The function \( C_t \) satisfies \( \eta(\Theta^*, \alpha) \leq C_t(\alpha) \). And then the action for time step \( t \) is \( a_t \leftarrow \arg \max C_t(\alpha) \). Let \( \Theta_t \) be the set of parameters that is consistent with \( \eta(\theta_1, \alpha_1), \ldots, \eta(\theta_t, \alpha_t) \). That is, \( \Theta_t = \{ \theta \in \Theta : \eta(\theta, \alpha_t) = \eta(\theta^*, \alpha_t) \} \). In a deterministic environment, the tightest upper confidence bound is \( C_t(\alpha) = \sup_{\theta \in \Theta_t} \eta(\theta, \alpha) \).

We first provide a proof sketch to the theorem. We consider the following reward function.

\[ \eta((\theta_1^*, \theta_2^*, \alpha), a) = \frac{1}{64} \langle a, \theta_1^* \rangle + \alpha \max \left( \langle \theta_2^*, a \rangle - \frac{31}{32}, 0 \right) \]

Note that the reward function \( \eta \) can be clearly realized by a two-layer neural network with width \( 2d \). When \( \alpha = 0 \) we have \( \eta((\theta_1^*, \theta_2^*, \alpha), a) = \frac{1}{64} \langle \theta_1^*, a \rangle \), which represents a linear reward. Informally, optimism based algorithm will try to make the second term large (because optimistically the algorithm hopes \( \alpha = 1 \)), which leads to an action \( a_t \) that is suboptimal for ground-truth reward (in which case \( \alpha = 0 \)). In round \( t \), the optimism algorithm observes \( \langle \theta_2^*, a_i \rangle = 0 \), and can only eliminate an exponentially small fraction of \( \theta_2^* \) from the hypothesis. Therefore the optimism algorithm needs exponential number of steps to determine \( \alpha = 0 \) and stops exploration. Formally, the prove is given below.

**Proof.** Consider a bandit problem where \( \mathcal{A} = S^{d-1} \) and

\[ \eta((\theta_1^*, \theta_2^*, \alpha), a) = \frac{1}{64} \langle a, \theta_1^* \rangle + \alpha \max \left( \langle \theta_2^*, a \rangle - \frac{31}{32}, 0 \right) \]

The hypothesis space is \( \Theta = \{ \theta_1, \theta_2, \alpha : \| \theta_1 \|_2 \leq 1, \| \theta_2 \|_2 \leq 1, \alpha \in [0, 1] \} \). Then the reward function \( \eta \) can be clearly realized by a two-layer neural network with width \( d \). Note that when \( \alpha = 0 \) we have \( \eta((\theta_1^*, \theta_2^*, \alpha), a) = \frac{1}{64} \langle \theta_1^*, a \rangle \), which represents a linear reward. In the following we use \( \Theta^* = (\theta_1^*, \theta_2^*, 0) \) as a shorthand.

The UCB algorithm is described as follows. At every time step \( t \), the algorithm maintains a upper confidence bound \( C_t : \mathcal{A} \to \mathbb{R} \). The function \( C_t \) satisfies \( \eta(\Theta^*, \alpha) \leq C_t(\alpha) \). And then the action for time step \( t \) is \( a_t \leftarrow \arg \max C_t(\alpha) \).

Let \( \mathcal{P} = \{ p_1, p_2, \ldots, p_n \} \) be an \( \frac{1}{2} \)-packing of the sphere \( S^{d-1} \), where \( n = \Omega(2^d) \). Let \( B(p_i, \frac{1}{2}) \) be the ball with radius \( 1/4 \) centered at \( p_i \), and \( B_i = B(p_i, \frac{1}{4}) \cup S^{d-1} \). We prove the theorem by showing that the UCB algorithm will explore every packing in \( \mathcal{P} \). That is, for any \( i \in [n] \), there exists \( t \) such that \( a_t \in B_i \). Since we have \( \sup_{a \in B_j} \langle p_i, a \rangle \leq 31/32 \) for all \( j \neq i \), this over-exploration strategy leads to a sample complexity (for finding a \( (31/2048)\)-suboptimal action) at least \( \Omega(2^d) \) when \( \Theta^* = (p_i, p_i, 0) \).

Let \( \Theta_t \) be the set of parameters that is consistent with \( \eta(\Theta^*, a_1), \ldots, \eta(\Theta^*, a_t-1) \). That is, \( \Theta_t = \{ \theta \in \Theta : \eta(\theta, a_t) = \eta(\Theta^*, a_t), \forall t < t \} \). Since our environment is deterministic, a tightest upper confidence bound is \( C_t(\alpha) = \sup_{\theta \in \Theta_t} \eta(\theta, \alpha) \).
Let $A_t = \{a_1, \ldots, a_t\}$. It can be verified that for any $\theta_2 \in S^{d-1}$, $\eta((\theta_1^*, \theta_2, 1), \cdot)$ is consistent with $\eta(\theta^*, \cdot)$ on $A_{t-1}$ if $B(\theta_2, \frac{1}{4}) \cup A_{t-1} = \emptyset$. As a result, for any $\theta_2$ such that $B(\theta_2, \frac{1}{4}) \cup A_{t-1} = \emptyset$ we have

$$C_t(\theta_2) \geq \frac{1}{32} \geq \frac{1}{128} + \sup_a \eta(\theta^*, a).$$  

(8)

Next we prove that for any $i \in [n]$, there exists $t$ such that $a_i \in B(p_i, \frac{1}{2})$. Note that $\eta(\theta, \cdot)$ is $\frac{\eta}{\Theta}$ Lipschitz for every $\theta \in \Theta$. As a result, $C_t(a_\tau + \xi) \leq C_t(a_\tau) + \frac{\eta}{\Theta} \|\xi\|_2 = \eta(\theta^*, a_\tau) + \frac{\eta}{\Theta} \|\xi\|_2$ for all $\tau < t$. Consequently,

$$C_t(a_\tau + \xi) \leq \sup_a \eta(\theta^*, a) + \frac{1}{128} = \frac{3}{128} \tag{9}$$

for any $\tau < t$ and $\xi$ such that $\|\xi\|_2 \leq \frac{1}{130}$. In other words, Eq. (9) upper bounds the upper confidence bound for actions that is taken by the algorithm, and Eq. (8) lower bounds the upper confidence bound for actions that is not taken.

Now, for the sake of contradiction, assume that there exists $t \in B(\theta_2, \frac{1}{4})$ is never taken by the algorithm. By Eq. (9) we have $C_t(\theta_2) \geq \frac{1}{32}$ for all $t$. Let $H_t = \cup_{i=1}^{t-1} B(a_{\tau}, \frac{1}{2})$. By Eq. (9) we have $C_t(a) \leq \frac{3}{128}$ for all $a \in H_t$. Because $a_t \leftarrow \arg \max_a C_t(a)$ and $\max_{a \in H_t} C_t(a) < C_t(\theta_2)$ we conclude that $a_t \not\in H_t$. Therefore, $\{a_t\}$ is a $(1/130)$-packing. However, the $(1/130)$-packing of $S^{d-1}$ has a size bounded by $130^d$, which leads to contradiction.

For any $\theta_2$ there exists $t \leq 130^d$ such that $a_t \in B(\theta_2, \frac{1}{4})$. \qed

### E.4. Hardness of stochastic environments

As a motivation to consider deterministic rewards, the next theorem proves that a poly(log|\Theta|) sample complexity is impossible for finding local optimal action even under mild stochastic environment.

**Theorem E.1.** There exists an bandit problem with stochastic reward and hypothesis class with size $\log|\Theta| = \tilde{O}(1)$, such that any algorithm requires $\Omega(d)$ sample to find a $(0.1, 1)$-approximate second order stationary point with probability at least $3/4$.

A similar theorem is proved in Lattimore, Szepesvári (2020, Section 23.3) (in a somewhat different context) with minor differences in the constructed hard instances.

**Proof.** We consider a linear bandit problem with hypothesis class $\Theta = \{e_1, \ldots, e_d\}$. The action space is $S^{d-1}$. The stochastic reward function is given by $\eta(\theta, a) = \langle \theta, a \rangle + \xi$ where $\xi \sim \mathcal{N}(0, 1)$ is the noise. Define the set $A_i = \{a \in S^{d-1} : |\langle a, e_i \rangle| \geq 0.9\}$. By basic algebra we get, $A_i \cap A_j = \emptyset$ for all $i \neq j$.

The manifold gradient of $\eta(\theta, \cdot)$ on $S^{d-1}$ is

$$\nabla \eta(\theta, a) = (I - aa^T)\theta.$$

By triangular inequality we get $\|\nabla \eta(\theta, a)\|_2 \geq \|\theta\|_2 - \|a, \theta\|$. Consequently, $\|\nabla \eta(\theta, a)\|_2 \geq 0.1$ for $a \not\in A_i$. In other words, $\left(S^{d-1} \setminus A_i\right)$ does not contain any $(0.1, 1)$-approximate second order stationary point for $\eta(\theta, \cdot)$.

For a fixed algorithm, let $a_1, \ldots, a_T$ be the sequence of actions chosen by the algorithm, and $x_t = \langle \theta^*, a_t \rangle + \xi_t$. Next we prove that with $T \leq d$ steps, there exists $i \in [d]$ such that $\Pr_i \left(a_T \in A_i \right) \leq 1/2$, where $\Pr_i$ denotes the probability space generated by $\theta^* = \theta_i$. Let $\Pr_0$ be the probability space generated by $\theta^* = 0$. Let $E_{i, T}$ be the event that the algorithm outputs an action $a \in A_i$ at the last step $T$. By Pinsker inequality we get,

$$\mathbb{E}_i[E_{i, T}] \leq \mathbb{E}_0[E_{i, T}] + \sqrt{\frac{1}{2} D_{KL}(\Pr_i, \Pr_0)}. \tag{10}$$

Using the chain rule of KL-divergence and the fact that $D_{KL}(\mathcal{N}(0, 1), \mathcal{N}(a, 1)) = \frac{a^2}{2}$, we get

$$\mathbb{E}_i[E_{i, T}] \leq \mathbb{E}_0[E_{i, T}] + \sqrt{\frac{1}{4} \mathbb{E}_0 \left(\sum_{t=1}^{T} \langle a_t, \theta_i \rangle^2\right)^2}. \tag{11}$$
Consequently,
\[
\sum_{i=1}^{d} \mathbb{E}[E_{i,T}] \leq \sum_{i=1}^{d} \mathbb{E}[E_{0,T}] + \sum_{i=1}^{d} \sqrt{\frac{1}{4} \mathbb{E}_{0}\left[\sum_{t=1}^{T} \langle a_t, \theta_t \rangle^2 \right]}
\]
\[
\leq 1 + \sqrt{\frac{d}{4} \mathbb{E}_{0}\left[\sum_{t=1}^{T} \langle a_t, \theta_t \rangle^2 \right]} \leq 1 + \sqrt{\frac{dT}{4}},
\]
which means that
\[
\min_{i \in [d]} \mathbb{E}[E_{i,T}] \leq \frac{1}{d} + \sqrt{\frac{T}{4d}}.
\]

Therefore when \( T \leq d \), there exists \( i \in [d] \) such that \( \mathbb{E}[E_{i,T}] \leq \frac{3}{4}. \)

F. Missing Proofs in Section A

In this section, we show missing proofs in Section A. We also define the notion of local regret, and prove a sublinear (local) regret result.

F.1. Proof Sketch for Theorem 3.1

Proof of Theorem 3.1 consists of the following parts:

i. Because of the design of the loss function (Eq. 5), the online learner guarantees that \( \theta_t \) can estimate the reward, its gradient and hessian accurately, that is, for \( \theta_t \sim p_t, \eta(\theta^*, a_t) \approx \eta(\theta_t, a_t), \nabla_a \eta(\theta^*, a_{t-1}) \approx \nabla_a \eta(\theta_t, a_{t-1}), \) and \( \nabla_a^2 \eta(\theta^*, a_{t-1}) \approx \nabla_a^2 \eta(\theta_t, a_{t-1}). \)

ii. Because of (i), maximizing the virtual reward \( \mathbb{E}_{\theta_t} \eta(\theta_t, a) \) w.r.t \( a \) leads to improving the real reward function \( \eta(\theta^*, a) \) iteratively (in terms of finding second-order local improvement direction).

Concretely, define the errors in rewards and its derivatives: \( \Delta_{t,1} = |\eta(\theta_t, a_t) - \eta(\theta^*, a_t)|, \) \( \Delta_{t,2} = |\eta(\theta_t, a_{t-1}) - \eta(\theta^*, a_{t-1})|, \) \( \Delta_{t,3} = \|\nabla_a \eta(\theta_t, a_{t-1}) - \nabla_a \eta(\theta^*, a_{t-1})\|_2, \) and \( \Delta_{t,4} = \|\nabla_a^2 \eta(\theta_t, a_{t-1}) - \nabla_a^2 \eta(\theta^*, a_{t-1})\|_2. \) Let \( \Delta_t^2 = \sum_{t=1}^{T} \Delta_{t,i}^2 \) be the total error which measures how closeness between \( \theta_t \) and \( \theta^* \).

Assuming that \( \Delta_{t,j} \)'s are small, to show (ii), we essentially view \( a_t = \text{argmax}_{a \in A} \mathbb{E}_{\theta_t} \eta(\theta_t, a) \) as an approximate update on the real reward \( \eta(\theta^*, \cdot) \) and show it has local improvements if \( a_{t-1} \) is not a critical point of the real reward:

\[
\eta(\theta^*, a_t) \geq \Delta_t \mathbb{E}_{\theta_t} \eta(\theta_t, a_t)
\]
\[
\geq \sup_a \mathbb{E}_{\theta_t} \left[ \eta(\theta_t, a_{t-1}) + \langle a - a_{t-1}, \nabla_a \eta(\theta_t, a_{t-1}) \rangle - \frac{\zeta h}{2} \|a - a_{t-1}\|^2 \right]
\]
\[
\geq \Delta_t \sup_a \mathbb{E}_{\theta_t} \left[ \eta(\theta^*, a_{t-1}) + \langle a - a_{t-1}, \nabla_a \eta(\theta^*, a_{t-1}) \rangle - \frac{\zeta h}{2} \|a - a_{t-1}\|^2 \right]
\]
\[
\geq \eta(\theta^*, a_{t-1}) + \frac{1}{2\zeta h} \|\nabla_a \eta(\theta^*, a_{t-1})\|^2.
\]

Here in equations (15) and (17), we use the symbol \( \geq_{\Delta_t} \) to present informal inequalities that are true up to some additive errors that depend on \( \Delta_t \). This is because equation (15) holds up to errors related to \( \Delta_{t,1} = |\eta(\theta_t, a_t) - \eta(\theta^*, a_t)| \), and equation (17) holds up to errors related to \( \Delta_{t,2} = |\eta(\theta_t, a_{t-1}) - \eta(\theta^*, a_{t-1})| \) and \( \Delta_{t,3} = \|\nabla_a \eta(\theta_t, a_{t-1}) - \nabla_a \eta(\theta^*, a_{t-1})\|_2 \). Eq. (16) is a second-order Taylor expansion around the previous iteration \( a_{t-1} \) and utilizes the definition \( a_t = \text{argmax}_{a \in A} \mathbb{E}_{\theta_t} \eta(\theta_t, a) \). Eq. (18) is a standard step to show the first-order improvement of gradient descent (the so-called “descent lemma”). We also remark that \( a_t \) is the maximizer of the expected reward \( \mathbb{E}_{\theta_t} \eta(\theta_t, a) \) instead of \( \eta(\theta_t, a) \) because the adversary in online learning cannot see \( \theta_t \) when choosing adversarial point \( a_t \).

The following lemma formalizes the proof sketch above, and also extends it to considering second-order improvement. The proof can be found in Appendix F.2.
Lemma F.1. In the setting of Theorem F.3, when \( a_{t-1} \) is not an \((\epsilon, 6\sqrt{3}\sigma_3\epsilon)-\)approximate second order stationary point, we have \( \eta(\theta^*, a_1) \geq \eta(\theta^*, a_{t-1}) + \min \left( \zeta_h^{-1} \epsilon^2/4, \zeta_h^{-1/2} \epsilon \right) - C_1 \mathbb{E}_{\theta_1 \sim p_1} [\Delta_t] \).

Next, we show part (i) by linking the error \( \Delta_t \) to the loss function \( \ell \) (Eq. (5)) used by the online learner. The errors \( \Delta_{t,1}, \Delta_{t,2} \) are already part of the loss function. Let \( \hat{\Delta}_{t,3} = \langle \nabla_q \eta(\theta_t, a_{t-1}) - \nabla_q \eta(\theta^*, a_{t-1}), u_t \rangle \) and \( \hat{\Delta}_{t,4} = \langle \nabla^2_q \eta(\theta_t, a_{t-1}) - \nabla^2_q \eta(\theta^*, a_{t-1}) u_t, v_t \rangle \) be the remaining two terms (without the clipping) in the loss (Eq. (5)). Note that \( \Delta_{t,3} \) is supposed to bound \( \Delta_{t,3} \) because \( \mathbb{E}_{u_t} [\Delta_{t,3}^2] = \Delta_{t,3}^2 \). Similarly, \( \mathbb{E}_{u_t, v_t} [\Delta_{t,4}^2] = \| \nabla^2_q \eta(\theta_t, a_{t-1}) - \nabla^2_q \eta(\theta^*, a_{t-1}) \|_F^2 \geq \Delta_{t,4}^2 \). We clip \( \Delta_{t,3} \) and \( \Delta_{t,4} \) to make them uniformly bounded and improve the concentration with respect to the randomness of \( u \) and \( v \) (the clipping is conservative and is often not active). Let \( \hat{\Delta}_t^2 = \Delta_{t,1}^2 + \Delta_{t,2}^2 + \min \left( \kappa_t^2, \Delta_{t,3}^2 \right) + \min \left( \kappa_t^2, \Delta_{t,4}^2 \right) \) be the error received by the online learner at time \( t \). The argument above can be rigorously formalized into a lemma that upper bound \( \Delta_t \) by \( \Delta_t \), which will be bounded by the sequential Rademacher complexity.

Lemma F.2. In the setting of Theorem F.3, we have \( \mathbb{E}_{a_t \sim \mathcal{P}_t \left( \theta_{t-1} \right), \theta_t \left( \theta_{t-1} \right)} \left[ \sum_{t=1}^T \hat{\Delta}_t^2 \right] \geq \frac{1}{2} \mathbb{E}_{\theta_t \sim \mathcal{P}_t} [\sum_{t=1}^T \Delta_t^2] \).

We defer the proof to Appendix F.3. With Lemma F.1 and Lemma F.2, we can prove Theorem 3.1 by keeping track of the performance \( \eta(\theta^*, a_t) \). The full proof can be found in Appendix F.4.

F.2. Proof of Lemma F.1

Proof. We prove the lemma by showing that algorithm 1 improves reward \( \eta(\theta^*, a_t) \) in the following two cases:

1. \( \| \nabla_q \eta(\theta^*, a_{t-1}) \|_2 \geq \epsilon \), or
2. \( \| \nabla_q \eta(\theta^*, a_{t-1}) \|_2 \leq \epsilon \) and \( \lambda_{\max} \left( \nabla^2_q \eta(\theta^*, a_{t-1}) \right) \geq 6\sqrt{3}\sigma_3\epsilon \).

Case 1: For simplicity, let \( g_t = \nabla_q \eta(\theta^*, a_{t-1}) \). In this case we assume \( \| g_t \|_2 \geq \epsilon \). Define function

\[
\hat{\eta}_t(\theta, a) = \eta(\theta, a_{t-1}) + \langle a - a_{t-1}, \nabla_q \eta(\theta, a_{t-1}) \rangle - \zeta_h \| a - a_{t-1} \|^2_2 \tag{19}
\]

to be the local first order approximation of function \( \eta(\theta, a) \). By the Lipschitz assumption (namely, Assumption 2.1), we have \( \eta(\theta, a) \geq \hat{\eta}_t(\theta, a) \) for all \( \theta \in \Theta, a \in \mathcal{A} \). By the definition of \( \Delta_{t,2} \) and \( \Delta_{t,3} \), we get

\[
\hat{\eta}_t(\theta, a) \geq \hat{\eta}_t(\theta^*, a) - \Delta_{t,2} - \| a - a_{t-1} \|_2 \Delta_{t,3} \tag{20}
\]

In this case we have

\[
\eta(\theta^*, a_t) \geq \mathbb{E}_{\theta_t \sim p_t} [\eta(\theta_t, a_t) - \Delta_{t,1}] \\
\geq \sup_a \mathbb{E}_{\theta_t \sim p_t} [\eta(\theta_t, a) - \Delta_{t,1}] \\
\geq \sup_a \mathbb{E}_{\theta_t \sim p_t} [\hat{\eta}_t(\theta, a) - \Delta_{t,1}] \\
\geq \sup_a \mathbb{E}_{\theta_t \sim p_t} \left[ \eta(\theta^*, a_{t-1}) + \frac{1}{4\zeta_h} \| g_t \|^2 - \Delta_{t,1} - \| a - a_{t-1} \|_2 \Delta_{t,3} \right] \geq \mathbb{E}_{\theta_t \sim p_t} \left[ \eta(\theta^*, a_{t-1}) + \frac{\epsilon^2}{4\zeta_h} - \mathbb{E}_{\theta_t \sim p_t} \left[ \left( 2 + \frac{\epsilon}{\sqrt{\zeta_h}} \right) \Delta_t \right] \right] \tag{By Eq. (20)}
\]

\[
\geq \eta(\theta^*, a_{t-1}) + \frac{\epsilon^2}{4\zeta_h} - \mathbb{E}_{\theta_t \sim p_t} \left[ \left( 2 + \frac{\epsilon}{\sqrt{\zeta_h}} \right) \Delta_t \right] \geq \frac{\epsilon^2}{4\zeta_h} - \mathbb{E}_{\theta_t \sim p_t} \left[ \left( 2 + \frac{\epsilon}{\sqrt{\zeta_h}} \right) \Delta_t \right] \tag{By Cauchy-Schwarz}
\]

Case 2: Let \( H_t = \nabla^2_q \eta(\theta^*, a_{t-1}) \). Define \( v_t \in \text{argmax}_{v: \| v \|_2 = 1} v^\top H_t v \). In this case we have \( \| g_t \|_2 \leq \epsilon \) and

\[
v_t^\top H_t v_t \geq 6\sqrt{3}\sigma_3 \epsilon \| v_t \|^2_2 \tag{21}
\]

Define function

\[
\hat{\eta}_t(\theta, a) = \eta(\theta, a_{t-1}) + \langle a - a_{t-1}, \nabla_q \eta(\theta, a_{t-1}) \rangle
\]
As a result, we have
\[ \frac{1}{2} \left\langle \nabla^2 \eta(\theta, a) - \nabla^2 \eta(\theta, a) \right\rangle - \frac{e^{3/2}}{2} \|a-a_{t-1}\|^2 \]

(22)
to be the local second order approximation of function \( \eta(\theta, a) \). By the Lipschitz assumption (namely, Assumption 2.1), we have \( \eta(\theta, a) \geq \eta(\theta, a) \) for all \( \theta \in \Theta, a \in A \).

By Eq. (21), we can exploit the positive curvature by taking \( a' = a_{t-1} + 4 \sqrt{\frac{\zeta_{3rd}}{t}} \nu_t \). Concretely, by basic algebra we get:
\[ \hat{\eta}(\theta^*, a') \geq \eta(\theta^*, a_{t-1}) - \epsilon \|a' - a_{t-1}\|^2 + 3 \sqrt{\frac{\zeta_{3rd}}{t}} \epsilon \|a' - a_{t-1}\|^2 - \frac{e^{3/2}}{2} \|a' - a_{t-1}\|^3 \]
\[ \geq \eta(\theta^*, a_{t-1}) + 12 \frac{e^3}{\zeta_{3rd}} \Delta_{t,3} \]
Combining the two cases together, we get the desired result.

F.3. Proof of Lemma F.2

**Proof.** Define \( \mathcal{F}_t \) to be the \( \sigma \)-field generated by random variable \( u_{1:t}, v_{1:t}, \theta_{1:t} \). In the following, we use \( \mathbb{E}[\cdot | \mathcal{F}_t] \) as a shorthand for \( \mathbb{E}[\cdot] \).

Let \( g_t = \nabla \eta(\theta_t, a_{t-1}) - \nabla \eta(\theta^*, a_{t-1}) \). Note that condition on \( \theta_t \) and \( \mathcal{F}_{t-1} \), \( \langle g_t, u_t \rangle \) follows the distribution \( \mathcal{N}(0, \|g_t\|^2) \).

By Assumption 2.1, \( \|g_t\| \leq 2s_2 = \zeta_1 \). As a result,
\[ \mathbb{E}_{t-1} \left[ \min \left( \|\kappa^2_t, (g_t, u_t)^2 \right) | \theta_t \right] \geq \frac{1}{2} \mathbb{E}_{t-1} \left[ \|g_t, u_t\|^2 | \theta_t \right] = \frac{1}{2} \|g_t\|^2. \]

(25)
By the tower property of expectation we get
\[ \mathbb{E}_{t-1} \left[ \min \left( \|\Delta_{t,3} \right) \right] \geq \frac{1}{2} \mathbb{E}_{t-1} \left[ \|\kappa^2_t, \Delta_{t,3} \right] \]

(26)
Now we turn to the term \( \Delta_{D,4}^2 \). Let \( H_t = \nabla \eta(\theta_t, a_{t-1}) - \nabla \eta(\theta^*, a_{t-1}) \). Define a random variable \( x = (u_t^T H_t v_t)^2 \). Note that \( u_t, v_t \) are independent, we have
\[ \mathbb{E}_{t-1} \left[ x | \theta_t \right] = \mathbb{E}_{t-1} \left[ \|H_t v_t\|^2 | \theta_t \right] = \|H_t\|^2 \geq \|H_t\|^2_{sp}. \]

(27)
Since \( u_t, v_t \) are two Gaussian vectors, random variable \( x \) has nice concentrateability properties. Therefore we can prove that the min operator in the definition of \( \Delta_t \) does not change the expectation too much. Formally speaking, by Lemma 1.6, condition on \( \mathcal{F}_{t-1} \) and \( \theta_t \), we have \( \mathbb{E} \left[ \min \left( \kappa^2_t, x \right) \right] \geq \frac{1}{2} \min \left( \zeta^2_t, \mathbb{E}[x] \right) \), which leads to
\[ \mathbb{E}_{t-1} \left[ \min \left( \kappa^2_t, \Delta_{D,4}^2 \right) \right] \geq \frac{1}{2} \mathbb{E}_{t-1} \left[ \min \left( \zeta^2_t, \|H_t\|^2_{sp} \right) \right] \geq \frac{1}{2} \mathbb{E}_{t-1} \left[ \min \left( \zeta^2_t, \|H_t\|^2_{sp} \right) \right] = \frac{1}{2} \mathbb{E}_{t-1} \left[ \|H_t\|^2_{sp} \right]. \]

(28)
Combining Eq. (26) and Eq. (28), we get the desired inequality.
F.4. Proof of Theorem 3.1

In this section we show that Alg. 1 finds a \((\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon)\)-approximate local maximum in polynomial steps. In the following, we treat \(\zeta_g, \zeta_h, \zeta_{3rd}\) as constants.

Proof of Theorem 3.1. We prove this theorem by contradiction. Suppose \(\Pr \left[ a_{t+1} \in \mathcal{A}_{\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon} \right] \leq 0.5\) for all \(t \in [T]\), we prove that \(T \lesssim \tilde{O}(C_4^4 R(\Theta) \max (\zeta_g^4 \epsilon^{-8}, \zeta_{3rd}^2 \epsilon^{-6}))\).

Define \(\nu = \min \left( \frac{1}{4\zeta_h^2}, \frac{1}{\zeta_{3rd}^3/2} \right)\). Recall that \(C_1 = 2 + \frac{\zeta_g}{\zeta_h}\). By Lemma F.1, when \(a_t\) is not a \((\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon)\)-approximate local maximum we have

\[
\eta(\theta^*, a_{t+1}) \geq \eta(\theta^*, a_t) + \nu - C_1 E_t[\Delta_{t+1}].
\]

(29)

Similar to the proof of Theorem F.3, when \(a_t\) is a \((\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon)\)-approximate local maximum we have

\[
\eta(\theta^*, a_{t+1}) \geq \eta(\theta^*, a_t) - C_1 E_t[\Delta_{t+1}].
\]

(30)

As a result, when \(\Pr \left[ a_{t+1} \in \mathcal{A}_{\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon} \right] \leq 0.5\) we get

\[
E[\eta(\theta^*, a_{t+1})] \geq E[\eta(\theta^*, a_t)] + \frac{\nu}{2} - C_1 E[\Delta_{t+1}].
\]

(31)

Take summation of Eq. (31) over \(t \in [T]\) leads to

\[
E[\eta(\theta^*, a_T) - \eta(\theta^*, a_0)] \geq \frac{\nu T}{2} - C_1 E \left[ \sum_{t=1}^{T} \Delta_t \right].
\]

(32)

Lemma F.2 leads to

\[
E \left[ \sum_{t=1}^{T} \Delta_t \right] \leq 2TE \left[ \sum_{t=1}^{T} \Delta_t^2 \right] \leq 2T^{3/4}(R(\Theta)\text{polylog}(T))^{1/4}.
\]

(33)

Combining with Eq. (32) we have

\[
1 \geq E[\eta(\theta^*, a_T) - \eta(\theta^*, a_0)] \geq \frac{\nu T}{2} - 2C_1 T^{3/4}(R(\Theta)\text{polylog}(T))^{1/4}.
\]

(34)

As a result, we can solve an upper bound of \(T\). In particular, we get

\[
T \lesssim C_4^4 R(\Theta) \max (\zeta_h^4 \epsilon^{-8}, \zeta_{3rd}^2 \epsilon^{-6}) \text{polylog}(R(\Theta), 1/\epsilon, C_1, \zeta_h, \zeta_{3rd}).
\]

(35)

Consequently, when \(T \geq \tilde{O}(C_4^4 R(\Theta) \max (\zeta_h^4 \epsilon^{-8}, \zeta_{3rd}^2 \epsilon^{-6}))\), there exists \(t \in [T]\) such that \(\Pr \left[ a_{t+1} \in \mathcal{A}_{\epsilon, 6\sqrt{\zeta_{3rd}}\epsilon} \right] > 0.5\).

Finally, by running Alg. 1 \(\log(1/\delta)\) times, we can get a high probability guarantee. \(\square\)

F.5. Local Regret

We also define the “local regret” by comparing with an approximate local maximum. Formally speaking, let \(\mathcal{A}_{\epsilon_g, \epsilon_h}\) be the set of all \((\epsilon_g, \epsilon_h)\)-approximate local maximum of \(\eta(\theta^*, .)\). The \((\epsilon_g, \epsilon_h)\)-local regret of a sequence of actions \(a_1, \ldots, a_T\) is defined as

\[
\text{REG}_{\epsilon_g, \epsilon_h}(T) = \sum_{t=1}^{T} \left( \inf_{a \in \mathcal{A}_{\epsilon_g, \epsilon_h}} \eta(\theta^*, a) - \eta(\theta^*, a_t) \right).
\]

(36)

Our goal is to achieve a \((\epsilon_g, \epsilon_h)\)-local regret that is sublinear in \(T\) and inverse polynomial in \(\epsilon_g\) and \(\epsilon_h\). With a sublinear regret (i.e., \(\text{REG}_{\epsilon_g, \epsilon_h}(T) = o(T)\)), the average performance \(\frac{1}{T} \sum_{t=1}^{T} \eta(\theta^*, a_t)\), converges to that of an approximate local maximum of \(\eta(\theta^*, .)\). Hazan et al. (2017) also work on local regret in the setting of online non-convex games. Their local regret notation is different from ours.

The main theorem for local regret guarantee is stated below.
Theorem F.3. Let $\mathcal{R}_T$ be the sequential Rademacher complexity of the family of the losses defined in Eq. (5). Let $C_1 = 2 + \zeta_0 / \zeta_0$. Under Assumption 2.1, for any $\epsilon \leq \min \left(1, \frac{\zeta_0}{\zeta_0} \right)$, we can bound the $(\epsilon, 6\sqrt{\zeta_0 \epsilon})$-local regret of Alg. 1 from above by

$$
\mathbb{E}[\text{REG}_{\epsilon, 6\sqrt{\zeta_0 \epsilon}}(T)] \leq \left(1 + C_1 \sqrt{4T \mathcal{R}_T} \right) \max \left(4\zeta_0 \epsilon^{-2}, \sqrt{\zeta_0 \epsilon^{-3/2}} \right).
$$

Note that when the sequential Rademacher complexity $\mathcal{R}_T$ is bounded by $\tilde{O}(T \sqrt{T})$ (which is typical), we have $O(\sqrt{T \mathcal{R}_T}) = \tilde{O}(T^{3/4}) = o(T)$ regret. As a result, Alg. 1 achieves a $O(\text{poly}(1/\epsilon))$ sample complexity by the sample complexity-regret reduction (Jin et al., 2018, Section 3.1). The proof is deferred to Section F.6.

F.6. Proof of Theorem F.3

Proof. Let $\delta_t = \inf_{a \in \mathcal{A}_{(\epsilon), 6\sqrt{\zeta_0 \epsilon}}} \eta(\theta^*, a) - \eta(\theta^*, a_t)$. By the definition of regret we have $\text{REG}_{\epsilon, 6\sqrt{\zeta_0 \epsilon}}(T) = \sum_{t=1}^T \delta_t$.

Define $\nu = \min \left(\frac{1}{4\epsilon^2}, \frac{1}{\sqrt{\zeta_0}} \epsilon^{3/2} \right)$ for simplicity. Recall that $C_1 = 2 + \frac{\zeta_0}{\zeta_0}$. In the following we prove by induction that for any $t_0$,

$$
\mathbb{E}_{t_0-1} \left[ \sum_{t=t_0}^T \delta_t \right] \leq \mathbb{E}_{t_0-1} \left[ \frac{1}{\nu} \left( \delta_{t_0} + C_1 \sum_{t=t_0+1}^T \Delta_t \right) \right].
$$

(38)

For the base case where $t_0 = T$ Eq. (38) trivially holds because $\nu \leq 1$.

Now suppose Eq. (38) holds for any $t > t_0$ and consider time step $t_0$. When $a_{t_0} \notin \mathcal{A}_{(\epsilon), 6\sqrt{\zeta_0 \epsilon}}$, applying Lemma F.1 we get

$$
\eta(\theta^*, a_{t_0+1}) \geq \eta(\theta^*, a_{t_0}) + \nu - C_1 \mathbb{E}_{t_0} [\Delta_{t_0+1}].
$$

(39)

As a result,

$$
\mathbb{E}_{t_0-1} \left[ \sum_{t=t_0}^T \delta_t \right] = \mathbb{E}_{t_0-1} \left[ \delta_{t_0} + \sum_{t=t_0+1}^T \delta_t \right]
$$

$$
\leq \mathbb{E}_{t_0-1} \left[ \delta_{t_0} + \frac{1}{\nu} \left( \delta_{t_0+1} + C_1 \sum_{t=t_0+2}^T \Delta_t \right) \right] \quad \text{(By induction hypothesis)}
$$

$$
\leq \mathbb{E}_{t_0-1} \left[ \delta_{t_0} - 1 + \frac{1}{\nu} \left( \delta_{t_0} + C_1 \Delta_{t_0+1} + C_1 \sum_{t=t_0+2}^T \Delta_t \right) \right] \quad \text{(By Eq. (39))}
$$

$$
\leq \mathbb{E}_{t_0-1} \left[ \frac{1}{\nu} \left( \delta_{t_0} + C_1 \sum_{t=t_0+1}^T \Delta_t \right) \right] .
$$

On the other hand, when $a_{t_0} \in \mathcal{A}_{(\epsilon), 6\sqrt{\zeta_0 \epsilon}}$ we have

$$
\eta(\theta^*, a_{t_0+1}) \geq \mathbb{E}_{\theta_{t_0+1}} [\eta(\theta_{t_0+1}, a_{t_0+1}) - \Delta_{t_0+1, 1}] \geq \mathbb{E}_{\theta_{t_0+1}} [\eta(\theta_{t_0+1}, a_{t_0}) - \Delta_{t_0+1, 1}] \geq \mathbb{E}_{\theta_{t_0+1}} [\eta(\theta^*, a_{t_0}) - \Delta_{t_0+1, 1} - \Delta_{t_0+1, 2}] \geq \eta(\theta^*, a_{t_0}) - C_1 \mathbb{E}_{\theta_{t_0+1}} [\Delta_{t_0+1}].
$$

Consequently, by basic algebra we get $\delta_{t_0+1} \leq \delta_{t_0} + C_1 \mathbb{E}_{t_0} [\Delta_{t_0+1}]$. Note that since $a_{t_0} \in \mathcal{A}_{(\epsilon), 6\sqrt{\zeta_0 \epsilon}}$, we have $\delta_{t_0} \leq 0$. As a result,

$$
\mathbb{E}_{t_0-1} \left[ \sum_{t=t_0}^T \delta_t \right] \leq \mathbb{E}_{t_0-1} \left[ \delta_{t_0} + \frac{1}{\nu} \left( \delta_{t_0+1} + C_1 \sum_{t=t_0+2}^T \Delta_t \right) \right] \quad \text{(By induction hypothesis)}
$$

$$
\leq \mathbb{E}_{t_0-1} \left[ \frac{1}{\nu} \left( \delta_{t_0+1} + C_1 \sum_{t=t_0+2}^T \Delta_t \right) \right] \quad (\delta_{t_0} \leq 0)
$$
we have verifies Assumption 2.1. For completeness, in the following we prove that the Eluder dimension for sparse linear model is \( O(s) \) by finding a complexity bound fits perfectly with the covering number technique. That is, we can discretize the hypothesis set of linear bandit. In the following we prove that the hypothesis set has a small covering number. Note that the regularized reward \( \tilde{\ell}(\theta, a) \) is bounded \( \|\tilde{\ell}(\theta, a)\|_2 \leq \sqrt{2\log(d)} \), and therefore our local regret and the standard regret coincide (up to some conversion of the errors) when realizability holds, we have \( \inf_{\theta} \sum_{t=1}^{T} \ell((x_t, y_t); \theta) = 0 \). Therefore, by Lemma F.2 and the definition of online learning regret (see Eq. (1)) we have

\[
E[\text{REG}_{\epsilon, \delta \sqrt{\Delta T}}(T)] \leq \frac{1}{\epsilon} \left( 1 + C_1 \sqrt{2T \sum_{t=1}^{T} \Delta_t^2} \right) \leq \frac{1}{\epsilon} \left( 1 + C_1 \sqrt{4T \mathcal{R}_T} \right).
\]

(42)

**F.7. Instantiations of Theorem 3.1**

In this section we rigorously prove the instantiations discussed in Section A.

**Linear bandit with finite model class.** A full proof of this claim needs a few steps: (i) realizing that \( \eta(\theta^*, a) \) is concave in \( a \) with no bad local maxima, and therefore our local regret and standard regret coincide (up to some conversion of the errors); (ii) invoking Rakshin et al. (2015b, Lemma 3) to show that the sequential Rademacher complexity \( \mathcal{R}_T \) is bounded by \( O(\sqrt{2\log(\Theta)}/T) \), and (iii) verifying \( \tilde{\eta} \) satisfies the conditions (Assumption 2.1) on the actions that the algorithm will visit.

Recall that the linear bandit reward is given by \( \eta(\theta, a) = \langle \theta, a \rangle \), and the constrained reward is \( \tilde{\eta}(\theta, a) = \eta(\theta, a) - \frac{1}{2} \|a\|_2^2 \).

In order to deal with \( \ell_2 \) regularization which violates Assumption 2.1, we bound the set of actions Alg. 1 takes. Consider the regularized reward \( \tilde{\eta}(\theta, a) \). Recall that Alg. 1 chooses action \( a_t = \arg\max_{a \in A} E_{\theta_t \sim P_t}[\tilde{\eta}(\theta_t, a)] \). By optimality condition we have \( a_t = E_{\theta_t \sim P_t}[\tilde{\eta}] \). Consequently \( \|a_t\|_2 \leq E_{\theta_t \sim P_t}[\|\tilde{\eta}\|_2] \leq 1 \).

Because we only apply Lemma F.1 and Lemma F.2 to actions that is taken by the algorithm. Theorem F.3 holds even if Assumption 2.1 is satisfied locally for \( \|a\|_2 \leq 1 \). Since the gradient and Hessian of regularization term is \( a \) and \( I_d \) respectively, we have \( \|\nabla_a \tilde{\eta}(\theta, a)\|_2 \leq \|\nabla_a \eta(\theta, a)\|_2 + 1 \) and \( \|\nabla_a^2 \tilde{\eta}(\theta, a)\|_{sp} \leq \|\nabla_a^2 \eta(\theta, a)\|_{sp} + 1 \) when \( \|a\|_2 \leq 1 \), which verifies Assumption 2.1.

In the following we prove that any \( (\epsilon, 1) \)-approximate local maximum for \( \tilde{\eta}(\theta^*, a) \) is an \( \epsilon \)-suboptimal action for \( \eta(\theta^*, a) \).

Note that \( \nabla_a \tilde{\eta}(\theta, a) = \theta - a \). Therefore, for any \( a \in A(\epsilon, 1) \) we have \( \|\theta^* - a\|_2 \leq \epsilon \). Applying Lemma I.11 we have

\[
(1 - \langle \theta^*, a \rangle)^2 \leq \|\theta^* - a\|_2^2 \leq \epsilon^2.
\]

(43)

Combining with the fact that \( \|a_t\|_2 \leq 1 \) for any \( t \in [T] \), we prove the claim.

**Linear bandit with sparse or structured model vectors.** In this case, the reduction is exactly the same as that in linear bandit. In the following we prove that the sparse linear hypothesis has a small covering number. Note that the \( \log(|\Theta|) \) sample complexity bound fits perfectly with the covering number technique. That is, we can discretize the hypothesis set \( \Theta \) by finding a \( 1/poly(d, 1/\epsilon) \)-covering of the loss function \( \mathcal{L} = \{\ell(\cdot, \theta) : \theta \in \Theta\} \). And then the sample complexity of our algorithm depends polynomially on the log-covering number. Since the log-covering number of the set of \( s \)-sparse vectors is bounded by \( O(s \log(dT)) \), we get the desired result.

For completeness, in the following we prove that the Eluder dimension for sparse linear model is \( \Omega(d) \).
Lemma F.4. Let $e_1, \cdots, e_d$ be the basis vectors and $f_i(a) = \langle e_i, a \rangle$. Specifically, define $f_0(a) = 0$. Define the function class $\mathcal{F} = \{f_i : 0 \leq i \leq d\}$. The Eluder dimension of $\mathcal{F}$ is at least $d$.

Proof. In order to prove the lower bound for Eluder dimension, we only need to find a sequence $a_1, \cdots, a_d$ such that $a_i$ is independent with its predecessors. In the sequel we consider the action sequence $a_1 = e_1, a_2 = e_2, \cdots, a_d = e_d$.

Now we prove that for any $i \in [d]$, $a_i$ is independent with $a_j$ where $j < i$. Indeed, consider functions $f_i$ and $f_0$. By definition we have $f_i(a_j) = f_0(a_j), \forall j < i$. However, $f_i(a_i) = 1 \neq 0 = f_0(a_i)$.

Deterministic logistic bandits. For deterministic logistic bandits, the reward function is given by $\eta(\theta, a) = (1 + e^{-\langle \theta, a \rangle})^{-1}$. The model class is $\Theta \subseteq S^{d-1}$ and the action space is $\mathcal{A} = S^{d-1}$. Similarly, we run Alg. 1 on an unbounded action space with regularized loss $\tilde{\eta}(\theta, a) = \eta(\theta, a) - \frac{1}{2} \|a\|_2^2$ where $c = c(e + 1)^{-2}$ is a constant. The optimal action in this case is $a^* = \theta^*$. Note that the loss function is not concave, but it satisfies that all local maxima are global. As a result, we claim that our algorithm finds an $\epsilon$-suboptimal action for $\eta(\theta^*, a)$. First of all, we bound the set of actions Alg. 1 can choose. By basic algebra we get

\[
\nabla_a \tilde{\eta}(\theta, a) = \frac{\exp(-\theta^\top a)}{(1 + \exp(-\theta^\top a))^2} \theta - ca.
\]

As a result, we have $a^* = \theta^*$. Recall that $a_t = \arg\max_{a \in \mathcal{A}} E_{\theta_t} [\tilde{\eta}(\theta_t, a)]$. By optimality condition we get

\[
ca = E_{\theta_t} \left[ \frac{\exp(-\theta_t^\top a)}{(1 + \exp(-\theta_t^\top a))^2} \theta_t \right].
\]

Multiply $a^\top$ to both hand side we get

\[
c \|a\|_2^2 = E_{\theta_t} \left[ \frac{\exp(-\theta_t^\top a)}{(1 + \exp(-\theta_t^\top a))^2} \theta_t^\top a \right].
\]

Define $f(x) = \frac{x \exp(-x)}{c(1 + \exp(-x))^2}$. Eq. (46) implies

\[
\|a\|_2^2 = E_{\theta_t} [f(\theta_t^\top a)] \leq \sup_{x \in [-\|a\|_2, \|a\|_2]} f(x).
\]

Solving Eq. (47) we get $\|a\|_2 \leq 1$.

Now translate an $(\epsilon, 1)$-approximate local maximum for $\tilde{\eta}(\theta^*, a)$ to an $O(\epsilon)$-approximate optimal action for $\eta(\theta^*, a)$. Note that $\tilde{\eta}(\theta^*, \cdot)$ is $(1/20)$-strongly concave. As a result, for any $\epsilon \in \mathbb{R}_{(\epsilon, 1)}$ we get

\[
\tilde{\eta}(\theta^*, a^*) - \tilde{\eta}(\theta^*, a) \lesssim \| \nabla_a \tilde{\eta}(\theta^*, a) \|_2^2 \lesssim \epsilon^2.
\]

Define $r(x) \triangleq (1 + \exp(-x))$ for shorthand. By Taylor expansion, for any $x \in \mathbb{R}$ there exists $\xi \in \mathbb{R}$ such that $r(x) = r(1) + (x - 1)r'(x) + (x - 1)^2 r''(\xi)$. As a result,

\[
\hat{\eta}(\theta^*, a^*) - \hat{\eta}(\theta^*, a) = r(1) - \frac{c}{2} - r'(\langle \theta^*, a \rangle) + \frac{c}{2} \|a\|_2^2
\]

\[
= (1 - \langle \theta^*, a \rangle) r'(1) - (1 - \langle \theta^*, a \rangle)^2 r''(\xi) - \frac{c}{2} + \frac{c}{2} \|a\|_2^2
\]

\[
= \frac{c}{2} - c \langle \theta^*, a \rangle + \frac{c}{2} \|a\|_2^2 - (1 - \langle \theta^*, a \rangle)^2 r''(\xi)
\]

\[
\geq \left( \frac{c}{2} - r''(\xi) \right)(1 - \langle \theta^*, a \rangle)^2
\]

(Recall that $r'(1) = c$.)
we can bound the sequential Rademacher complexity of Assumption F.5.

Theorem F.6. Let \( \{W_1, W_2\} \) be the parameter hypothesis. Under the setting of Theorem 3.1 with Assumption F.5, Alg. 1 finds an \( \epsilon, 6\sqrt{\log d}\)-approximate local maximum in \( \tilde{O}(\epsilon^{-8}\log d) \) steps. In addition, if the neural network is input concave, Alg. 1 finds an \( \epsilon \)-suboptimal action in \( \tilde{O}(\epsilon^{-4}\log d) \) steps.

Proof. We prove the theorem by first bounding the sequential Rademacher complexity of the loss function, and then applying Theorem F.3. Let \( \theta = (W_1, W_2) \). Recall that \( u \circ v \) denotes the element-wise product. By basic algebra we get,

\[
\langle \nabla_a \eta(\theta, a), u \rangle = W_2(\sigma'(W_1 a) \circ W_1 u),
\]

\[
u^\top \nabla^2_a \eta(\theta, a) v = W_2(\sigma''(W_1 a) \circ W_1 u \circ W_1 v).
\]

First of all, we verify that the regularized reward \( \tilde{\eta}(\theta, a) \) satisfies Assumption 2.1. Indeed we have

\[
\|\nabla_a \eta(\theta, a)\|_2 = \sup_{u \in \mathbb{S}^{d-1}} \langle \nabla_a \eta(\theta, a), u \rangle \leq 1,
\]

\[
\|\nabla^2_a \eta(\theta, a)\|_{sp} = \sup_{u, v \in \mathbb{S}^{d-1}} \nu^\top \nabla^2_a \eta(\theta, a) v \leq 1,
\]

\[
\|\nabla^2_a \eta(\theta, a_1) - \nabla^2_a \eta(\theta, a_2)\|_{sp} = \sup_{u, v \in \mathbb{S}^{d-1}} \| W_2((\sigma''(W_1 a_1) - \sigma''(W_1 a_2)) \circ W_1 u \circ W_1 v) \| \leq |a_1 - a_2|_2.
\]

Observe that \( |\eta(\theta, a)| \leq |a|_\infty \), we have \( \tilde{\eta}(\theta, a) \) satisfies Assumption 2.1. As a result, action \( a \) taken by Alg. 1 satisfies \( \|a\|_2 \leq 2 \) for all \( t \). Since the gradient and Hessian of regularizaton term \( a \) and \( I_d \) respectively, we have \( \|\nabla_a \tilde{\eta}(\theta, a)\|_2 \leq 2 \) and \( \|\nabla^2_a \tilde{\eta}(\theta, a)\|_{sp} \leq 2 \). It follows that Assumption 2.1 holds with constant Lipschitzness for actions \( a \) such that \( |a| \leq 1 \).

In the following we bound the sequential Rademacher complexity of the loss function. By Rakhlın et al. (2015a, Proposition 15), we can bound the sequential Rademacher complexity of \( \Delta^2_{1,1} \) and \( \Delta^2_{1,2} \) by \( \tilde{O}(\sqrt{T \log d}) \). Next we turn to higher order terms.

First of all, because the \( (1, \infty) \) norm of \( W_1 \) is bounded, we have \( \|W_1 u\|_\infty \leq \|u\|_\infty \). It follows from the upper bound of \( \sigma'(x) \) that \( \|\sigma'(W_1 a) \circ W_1 u\|_\infty \leq \|u\|_\infty \). Therefore we get

\[
\langle \nabla_a \eta(\theta, a), u \rangle \leq \|W_2\|_1 \|\sigma'(W_1 a) \circ W_1 u\|_\infty \leq \|u\|_\infty,
\]

\[
u^\top \nabla^2_a \eta(\theta, a) v \leq \|u\|_\infty \|v\|_\infty.
\]

Let \( B = (1 + \|u\|_\infty)(1 + \|v\|_\infty) \) for shorthand. We consider the error term \( \hat{\Delta}^2_{1,3} = (\langle \nabla_a \eta(\theta, a), u \rangle - [y_{\text{true}}]_3^2)^2 \). Let \( G_1 \) be the function class \( \{\langle \nabla_a \eta(\theta, a), u \rangle - [y_{\text{true}}]_3^2 : \theta \in \Theta\} \), and \( G_2 \) = \( \{\nabla_a \eta(\theta, a), u : \theta \in \Theta\} \). Applying Rakhlín et al. (2015a, Lemma 4) we get

\[ \mathcal{R}_T(G_1) \leq B \log^{3/2}(T^2)\mathcal{R}_T(G_2). \]
Define $G_3 = \{\sigma'(w_1^\top a) \cdot w_1^\top u : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$. In the following we show that $\mathcal{R}_T(G_2) \lesssim \mathcal{R}_T(G_3)$. For any sequence $u_1, \ldots, u_T$ and $A$-valued tree $a$, we have

$$\mathcal{R}_T(G_2) = \mathbb{E}_\epsilon \left[ \sup_{\|W_2\|_2 \leq 1} \sum_{i=1}^T \epsilon_i \left( \sum_{j=1}^w [W_2]_{ji} g_j(a_i(\epsilon)) \right) \right] \leq \mathbb{E}_\epsilon \left[ \sup_{\|W_2\|_2 \leq 1} \|W_2\|_1 \sup_{j \in [w]} \sum_{i=1}^T \epsilon_i g_j(a_i(\epsilon)) \right] \leq \mathbb{E}_\epsilon \left[ \sup_{g \in G_3} \left\| \sum_{i=1}^T \epsilon_i (g_j(a_i(\epsilon))) \right\| \right] \tag{55}$$

Since we have $0 \in \mathcal{G}_3$ by taking $w_1 = 0$, by symmetry we have

$$\mathbb{E}_\epsilon \left[ \sup_{g \in G_3} \left\| \sum_{i=1}^T \epsilon_i (g_j(a_i(\epsilon))) \right\| \right] \leq 2\mathbb{E}_\epsilon \left[ \sup_{g \in G_3} \sum_{i=1}^T \epsilon_i (g_j(a_i(\epsilon))) \right] = 2\mathcal{R}_T(G_3). \tag{56}$$

Now we bound $\mathcal{R}_T(G_3)$ by applying the composition lemma of sequential Rademacher complexity (namely Rakhlin et al. (2015a, Lemma 4)). First of all we define a relaxed function hypothesis $G_4 = \{\sigma'(w_1^\top a) \cdot w_1^\top u : w_1, w_1' \in \mathbb{R}^d, \|w_1\|_1 \leq 1, \|w_1'\|_1 \leq 1\}$. Since $G_3 \subset G_4$ we have $\mathcal{R}_T(G_3) \leq \mathcal{R}_T(G_4)$. Note that we have $|\sigma'(w_1^\top a)| \leq 1$ and $w_1^\top u \leq \|u\|_{\infty}$. Let $\phi(x, y) = xy$, which is $(3c)$-Lipschitz for $|x|, |y| \leq c$. Define $G_5 = \{\sigma'(w_1^\top a) : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$ and $G_6 = \{w_1^\top u : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$. Rakhlin et al. (2015a, Lemma 4) gives $\mathcal{R}_T(G_4) \lesssim B \log^{3/2}(T^2) \left( \mathcal{R}_T(G_5) + \mathcal{R}_T(G_6) \right)$. Note that $G_5$ is a generalized linear hypothesis and $G_6$ is linear, we have $\mathcal{R}_T(G_5) \lesssim B \log^{3/2}(T^2) \sqrt{T \log(d)}$ and $\mathcal{R}_T(G_6) \lesssim B \sqrt{T \log(d)}$.

In summary, we get $\mathcal{R}_T(G_1) = O\left(\text{poly}(B)\log^{3/2}(d, T) \sqrt{T}\right)$. Since the input $u_t \sim \mathcal{N}(0, I_{d \times d})$, we have $B \lesssim \log(dT)$ with probability $1/T$. As a result, the distribution dependent Rademacher complexity of $\hat{\Delta}_{\ell,3}^2$ in this case is bounded by $O\left(\text{polylog}(d, T) \sqrt{T}\right)$.

Similarly, we can bound the sequential Rademacher complexity of the Hessian term $\hat{\Delta}_{\ell,4}^2$ by $O\left(\text{polylog}(d, T) \sqrt{T}\right)$ by applying composition lemma with Lipschitz function $\phi(x, y, z) = xyz$ with bounded $|x|, |y|, |z|$. By Rakhlin et al. (2015a, Lemma 4), composing with the min operator only introduces $\text{polylog}(T)$ terms in the sequential Rademacher complexity. As a result, the sequential Rademacher complexity of the loss function can be bounded by

$$\mathcal{R}_T = O\left(\text{polylog}(d, T) \sqrt{T}\right).$$

Applying Theorem 3.1, the sample complexity of Alg. 1 is bounded by $\tilde{O}(\epsilon^{-8} \text{polylog}(d))$.

When the neural network is input concave (see (Amos et al., 2017)), the regularized reward $\hat{\eta}(\theta, a)$ is $\Omega(1)$-strongly concave. As a result, for any $a \in \mathcal{A}_{\epsilon, 1}$ we have $\hat{\eta}(\theta^*, a) \geq \hat{\eta}(\theta^*, a^*) - O(\epsilon^2)$. Hence, Alg. 1 finds an $\epsilon$-suboptimal action for regularized loss in $\tilde{O}(\epsilon^{-4} \text{polylog}(d))$ steps.

G. Missing Proofs in Section B

First of all, we present our algorithm in Alg. 2.

The approximate local maximum is defined in the same as in the bandit setting, except that the gradient and Hessian matrix are taken w.r.t. to the policy parameter space $\psi$. We also assume realizability ($\theta^* \in \Theta$) and the Lipschitz assumptions as in Assumption 2.1 (with action $a$ replaced by policy parameter $\psi$).

In the following we present the proof sketch for Theorem B.4. Compare to the bandit case, we only need to prove an analog of Lemma F.2, which means that we need to upper-bound the error term $\Delta_t$ by the difference of dynamics, as discussed
Algorithm 2 Virtual Ascent with Online Model Learner (ViOlin for RL)

1: Let $\mathcal{H}_0 = \emptyset$; choose $a_0 \in \mathcal{A}$ arbitrarily.
2: for $t = 1, 2, \cdots$ do
3: Run $\mathcal{R}$ on $\mathcal{H}_{t-1}$ with loss function $\ell$ (defined in Eq. (6)) and obtain $p_t = \mathcal{A}(\mathcal{H}_{t-1})$.
4: $\psi_t \leftarrow \arg\max_{\psi} \mathbb{E}_{\theta_t \sim p_t} [\eta(\theta_t, \psi)]$;
5: Sample one trajectory $\tau_t$ from policy $\pi_{\psi_t}$, and one trajectory $\tau'_t$ from policy $\pi_{\psi_{t-1}}$.
6: Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{\tau, \tau'\}$.
7: end for

Before. Formally speaking, let $\tau = (s_1, a_1, \cdots, s_H, a_H)$ be a trajectory sampled from policy $\pi_{\psi}$ under the ground-truth dynamics $T_{\theta^*}$. By telescoping lemma (Lemma I.16) we get

$$V_\theta^\psi(s_1) - V_\theta^\psi(s_1) = \mathbb{E}_{T \sim \rho_{\theta^*}} \left[ \sum_{h=1}^{H} (V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^*}(s_h, a_h))) \right].$$

(59)

Lipschitz assumption (Assumption B.1) yields,

$$|V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^*}(s_h, a_h))| \leq L_0 \|T_\theta(s_h, a_h) - T_{\theta^*}(s_h, a_h)\|_2.$$  

(60)

Combining Eq. (59) and Eq. (60) and apply Cauchy-Schwartz inequality gives an upper bound for $[\Delta_i]^2$ and $[\Delta_i^2]$. As for the gradient term, we will take gradient w.r.t. $\psi$ to both sides of Eq. (59). The gradient inside expectation can be dealt with easily. And the gradient w.r.t. the distribution $\rho_{\theta^*}^\psi$, can be computed by policy gradient lemma (Lemma I.17). As a result we get

$$\nabla_{\psi} V_\theta^\psi(s_1) - \nabla_{\psi} V_{\theta^*}^\psi(s_1)$$

$$= \mathbb{E}_{T \sim \rho_{\theta^*}} \left[ \sum_{h=1}^{H} \nabla_{\psi} \log \pi_{\psi}(a_h \mid s_h) \left( \sum_{h=1}^{H} (V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^*}(s_h, a_h))) \right) \right]$$

$$+ \mathbb{E}_{T \sim \rho_{\theta^*}} \left[ \sum_{h=1}^{H} \nabla_{\psi} V_\theta^\psi(T_\theta(s_h, a_h)) - \nabla_{\psi} V_{\theta^*}^\psi(T_{\theta^*}(s_h, a_h)) \right].$$

(61)

The first term can be bounded by vector-form Cauchy-Schwartz and Assumption B.2, and the second term is bounded by Assumption B.1. Similarly, this approach can be extended to second order term. As a result, we have the following lemma.

**Lemma G.1.** Under the setting of Theorem B.4, we have

$$c_1 \mathbb{E}_{\tau_t, \tau'_t, \theta^*} \left[ \Delta_t^2 \right] \geq \mathbb{E}_{\theta^*} \left[ \Delta_t^2 \right].$$

(62)

Proof of Lemma G.1 is shown in Appendix G.1. Proof of Theorem B.4 is exactly the same as that of Theorem 3.1 except for replacing Lemma F.2 with Lemma G.1.

**G.1. Proof of Lemma G.1**

**Proof.** The lemma is proven by combining standard telescoping lemma and policy gradient lemma. Specifically, let $\rho_{\theta^*}^\psi$ be the distribution of trajectories generated by policy $\pi$ and dynamics $T$. By telescoping lemma (Lemma I.16) we have,

$$V_{\theta^*}^\psi(s_1) - V_{\theta^*}^\psi(s_1) = \mathbb{E}_{T \sim \rho_{\theta^*}} \left[ \sum_{h=1}^{H} (V_{\theta^*}^\psi(T_{\theta^*}(s_h, a_h)) - V_{\theta^*}^\psi(T_{\theta^*}(s_h, a_h))) \right].$$

(63)

By the Lipschitz assumption (Assumption B.1),

$$|V_{\theta^*}^\psi(T_{\theta^*}(s_h, a_h)) - V_{\theta^*}^\psi(T_{\theta^*}(s_h, a_h))| \leq L_0 \|T_{\theta^*}(s_h, a_h) - T_{\theta^*}(s_h, a_h)\|_2.$$  

(64)
Applying policy gradient lemma (namely, Lemma I.17) to RHS of Eq. (70) we get,

\[
\Delta^2_{t,1} = \left( V^\psi_{\theta_t}(s_0) - V^\psi_{\theta^*}(s_0) \right)^2 \leq H L_0^2 \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \sum_{h=1}^{H} \|T_{\theta_t}(s_h, a_h) - T_{\theta^*}(s_h, a_h)\|_2^2 \right]. \tag{65}
\]

Similarly we get,

\[
\Delta^2_{t,2} = \left( V^\psi_{\theta_t}(s_0) - V^\psi_{\theta^*}(s_0) \right)^2 \leq H L_0^2 \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \sum_{h=1}^{H} \|T_{\theta_t}(s_h, a_h) - T_{\theta^*}(s_h, a_h)\|_2^2 \right]. \tag{66}
\]

Now we turn to higher order terms. First of all, by Hölder inequality and Assumption B.2, we can prove the following:

- \( \left\| \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} \nabla \log \pi_{\psi}(a_h | s_h) \right) \left( \sum_{h=1}^{H} \nabla \log \pi_{\psi}(a_h | s_h) \right)^\top \right] \right\|_\infty \leq H^2 \chi_g, \forall \psi \in \Psi \);

- \( \left\| \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} \nabla \log \pi_{\psi}(a_h | s_h) \right)^\otimes 4 \right] \right\|_\infty \leq H^4 \chi_f, \forall \psi \in \Psi \);

- \( \left\| \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} \nabla^2 \log \pi_{\psi}(a | s) \right) \left( \sum_{h=1}^{H} \nabla^2 \log \pi_{\psi}(a | s) \right)^\top \right] \right\|_\infty \leq H^2 \chi_h, \forall \psi \in \Psi \).

Indeed, consider the first statement. Define \( g_h = \nabla \log \pi_{\psi}(a_h | s_h) \) for shorthand. Then we have

\[
\left\| \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} g_h \right) \left( \sum_{h=1}^{H} g_h \right)^\top \right] \right\|_\infty = \sup_{u \in \mathbb{S}^{d-1}} \left\| u^\top \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} g_h \right) \left( \sum_{h=1}^{H} g_h \right)^\top \right] u \right\|
\]

\[
\leq \sup_{u \in \mathbb{S}^{d-1}} \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} g_h \right)^2 \right] \leq \sup_{u \in \mathbb{S}^{d-1}} \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ H \sum_{h=1}^{H} (u, g_h)^2 \right] \]

\[
\leq \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ H \sum_{h=1}^{H} \left\| g_h \right\|_2^2 \right] \leq H^2 \chi_g. \tag{67}
\]

Similarly we can get the second and third statement.

For any fixed \( \psi \) and \( \theta \) we have

\[
V^\psi_{\theta}(s_1) - V^\psi_{\theta^*}(s_1) = \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \sum_{h=1}^{H} \left( V^\psi_{\theta}(T_{\theta}(s_h, a_h)) - V^\psi_{\theta^*}(T_{\theta^*}(s_h, a_h)) \right) \right]. \tag{70}
\]

Applying policy gradient lemma (namely, Lemma I.17) to RHS of Eq. (70) we get,

\[
\nabla \psi V^\psi_{\theta}(s_1) - \nabla \psi V^\psi_{\theta^*}(s_1) = \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \left( \sum_{h=1}^{H} \nabla \log \pi_{\psi}(a_h | s_h) \right) \left( \sum_{h=1}^{H} \left( V^\psi_{\theta}(T_{\theta}(s_h, a_h)) - V^\psi_{\theta^*}(T_{\theta^*}(s_h, a_h)) \right) \right) \right]
\]

\[
+ \mathbb{E}_{\tau \sim \rho^\psi_{\theta^*}} \left[ \sum_{h=1}^{H} \left( \nabla \psi V^\psi_{\theta}(T_{\theta}(s_h, a_h)) - \nabla \psi V^\psi_{\theta^*}(T_{\theta^*}(s_h, a_h)) \right) \right]. \tag{71}
\]

Define the following shorthand:

\[
G^\psi_{\theta}(s, a) = V^\psi_{\theta}(T_{\theta}(s, a)) - V^\psi_{\theta^*}(T_{\theta^*}(s, a)), \tag{72}
\]
\[ f = \sum_{h=1}^{H} \nabla_{\psi} \log \pi_{\psi}(a_h \mid s_h). \] (73)

In the following we also omit the subscription in \( \mathbb{E}_{\tau \sim \rho_{\theta, \psi}^*} \) when the context is clear. It followed by Eq. (71) that
\[
\left\| \nabla_{\psi} V_{\theta}^\psi(s_1) - \nabla_{\psi} V_{\theta^*}^\psi(s_1) \right\|_2^2 \\
\leq 2 \left\| \mathbb{E} \left[ f \left( \sum_{h=1}^{H} G_{\theta}^\psi (s_h, a_h) \right) \right] \right\|_2^2 + 2 \left\| \mathbb{E} \left[ \sum_{h=1}^{H} \nabla_{\psi} G_{\theta}^\psi (s_h, a_h) \right] \right\|_2^2 \\
\leq 2 \left\| \mathbb{E} [f f^\top] \right\|_{sp} \mathbb{E} \left[ \left( \sum_{h=1}^{H} G_{\theta}^\psi (s_h, a_h) \right)^2 \right] + 2 \left\| \mathbb{E} \left[ \sum_{h=1}^{H} \nabla_{\psi} G_{\theta}^\psi (s_h, a_h) \right] \right\|_2^2 \quad \text{(By Lemma 1.7)}
\]
\[
\leq 2H \left\| \mathbb{E} [f f^\top] \right\|_{sp} \mathbb{E} \left[ \sum_{h=1}^{H} \left| G_{\theta}^\psi (s_h, a_h) \right|^2 \right] + 2H \mathbb{E} \left[ \sum_{h=1}^{H} \left\| \nabla_{\psi} G_{\theta}^\psi (s_h, a_h) \right\|_2^2 \right].
\]

Now, plugin \( \psi = \psi_{t-1}, \theta = \theta_t \) and apply Assumption B.1 we get
\[
\Delta_{t,3}^2 = \left\| \nabla_{\psi} V_{\theta_t}^{\psi_{t-1}}(s_1) - \nabla_{\psi} V_{\theta^*}^{\psi_{t-1}}(s_1) \right\|_2^2 \\
\leq (2HL_t^2 + 2H^3 \chi_t L_0^2) \mathbb{E}_{\tau \sim \rho_{\psi_{t-1}}^*} \left[ \sum_{h=1}^{H} \left| T_{\theta_t}(s_h, a_h) - T_{\theta^*}(s_h, a_h) \right| \right].
\]

For any fixed \( \psi, \theta \), define the following shorthand:
\[
g = \sum_{h=1}^{H} \left( V_{\theta}^{\psi}(T_{\theta}(s_h, a_h)) - V_{\theta}^{\psi}(T_{\theta^*}(s_h, a_h)) \right). \] (74)

Apply policy gradient lemma again to RHS of Eq. (71) we get
\[
\nabla_{\psi}^2 V_{\theta}^{\psi}(s_1) - \nabla_{\psi}^2 V_{\theta^*}^{\psi}(s_1) \\
= \mathbb{E}[(\nabla_{\psi} g)^\top] + \mathbb{E}[f(\nabla_{\psi} g)^\top] + \mathbb{E}[\nabla_{\psi}^2 g] + \mathbb{E} \left[ g \left( \sum_{h=1}^{H} \nabla_{\psi}^2 \log \pi_{\psi}(a_h \mid s_h) \right) \right] + \mathbb{E}[g f f^\top].
\]

As a result of Lemma 1.8 and Lemma 1.9 that,
\[
\left\| \nabla_{\psi}^2 V_{\theta}^{\psi}(s_1) - \nabla_{\psi}^2 V_{\theta^*}^{\psi}(s_1) \right\|_{sp}^2 \\
= 4 \left\| \mathbb{E}[(\nabla_{\psi} g)^\top] + \mathbb{E}[f(\nabla_{\psi} g)^\top] \right\|_{sp}^2 + 4 \left\| \mathbb{E}[\nabla_{\psi}^2 g] \right\|_{sp}^2 + 4 \left\| \mathbb{E}[g f f^\top] \right\|_{sp}^2 \\
+ 4 \left\| \mathbb{E} \left[ g \left( \sum_{h=1}^{H} \nabla_{\psi}^2 \log \pi_{\psi}(a_h \mid s_h) \right) \right] \right\|_{sp}^2 \\
\leq 8 \sup_{u,v \in S_0} \mathbb{E}[(\nabla_{\psi} g,u)(f,v)^2] + 4 \mathbb{E}[\left\| \nabla_{\psi}^2 g \right\|_{sp}^2] + 4 \mathbb{E}[g^2] \left\| \mathbb{E}[f f^\top] \right\|_{sp}^2 \\
+ 4 \mathbb{E}[g^2] \left\| \left( \sum_{h=1}^{H} \nabla_{\psi}^2 \log \pi_{\psi}(a_h \mid s_h) \right) \right\|_{sp}^2 \left( \sum_{h=1}^{H} \nabla_{\psi}^2 \log \pi_{\psi}(a_h \mid s_h) \right)^\top. \] (75)

Note that by Hölder's inequality,
\[
\sup_{u,v \in S_0} \mathbb{E}[(\nabla_{\psi} g,u)(f,v)^2] \leq \sup_{u,v \in S_0} \mathbb{E}[\nabla_{\psi} g,u]^2 \mathbb{E}[f,v]^2 \leq \mathbb{E}[\left\| \nabla_{\psi} g \right\|_{2}^2] \mathbb{E}[f f^\top]_{sp}.
\]
By Assumption B.1 we get,

\[
E[g^2] = E \left[ \left( \sum_{h=1}^{H} \left( V^\psi_{\theta'}(T\theta(s_h, a_h)) - V^\psi_{\theta'}(T\theta^\star(s_h, a_h)) \right) \right)^2 \right] 
\leq H E \left[ \sum_{h=1}^{H} \left( V^\psi_{\theta'}(T\theta(s_h, a_h)) - V^\psi_{\theta'}(T\theta^\star(s_h, a_h)) \right)^2 \right] 
\leq H L^2_0 \sum_{h=1}^{H} \|T\theta(s_h, a_h) - T\theta^\star(s_h, a_h)\|^2_2 .
\] (76)

Similarly, we have

\[
E[\|\nabla_{\psi} g\|^2_2] \leq H L^2_1 \sum_{h=1}^{H} \|T\theta(s_h, a_h) - T\theta^\star(s_h, a_h)\|^2_2 ,
\] (79)

\[
E[\|\nabla_{\psi} g\|^2_{sp}] \leq H L^2_2 \sum_{h=1}^{H} \|T\theta(s_h, a_h) - T\theta^\star(s_h, a_h)\|^2_2 .
\] (80)

Combining with Eq. (75) we get,

\[
\Delta^2_{i,4} = \|\nabla_{\psi} V^\psi_{\theta_i}(s_1) - \nabla_{\psi} V^\psi_{\theta^\star}(s_1)\|^2_{sp} 
\leq \left( 8 H^3 L^2_1 X^g + 4 H L^2_2 + 4 L^2_0 \right) \|T\theta(s_h, a_h) - T\theta^\star(s_h, a_h)\|^2_2 .
\] (77)

By noting that \(\Delta^2 = \sum_{i=1}^{4} \Delta^2_{i,4}\), we get the desired upper bound.

**H. Analysis of Example B.3**

Recall that our RL instance is given as follows:

\[
T(s, a) = N\theta(s + a),
\] (81)

\[
\pi_{\psi}(s) = N(\psi s, \sigma^2 I).
\] (82)

And the assumptions are listed below.

- Lipschitzness of reward function: \(|r(s_1, a_1) - r(s_2, a_2)| \leq L_r(\|s_1 - s_2\|_2 + \|a_1 - a_2\|_2)\).
- Bounded Parameter: we assume \(\|\psi\|_{op} \leq O(1)\).

In the sequel we verify the assumptions of Theorem B.4.

**H.1. Verifying Assumption B.2.**

**Verifying item 1.** Recall that \(\psi \in \mathbb{R}^{d \times d}\). By algebraic manipulation, for all \(s, a\) we get,

\[
\nabla_{\psi} \log \pi_{\psi}(a | s) = \frac{1}{\sigma^2} \text{vec}(a - \psi s) \otimes s
\] (83)

where \(\text{vec}(x)\) denotes the vectorization of tensor \(x\). Define random variable \(u = a - \psi s\). By the definition of policy \(\pi_{\psi}(s)\) we have \(u \sim N(0, \sigma^2 I)\). As a result,

\[
\|E_{u \sim N(\cdot | s)} [(\nabla_{\psi} \log \pi_{\psi}(a | s)) (\nabla_{\psi} \log \pi_{\psi}(a | s))^T]\|_{sp}
\] (84)
Verifying item 2. Similarly, using the equation where \( \mathbb{E}_{x \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} [x^4] = 3 \sigma^4 \) we have

\[
\mathbb{E}_{a \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} [(\nabla \log \pi_{\psi}(a \mid s))(\nabla \log \pi_{\psi}(a \mid s))^T]_{\mathbf{x}} \leq \frac{1}{\sigma^2} \Delta \chi_g.
\]

Verifying item 3. Since \( \nabla_{\psi}^2 \log \pi_{\psi}(a \mid s) \) is PSD, we have

\[
\mathbb{E}_{a \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} [(\nabla_{\psi}^2 \log \pi_{\psi}(a \mid s))(\nabla_{\psi}^2 \log \pi_{\psi}(a \mid s))^T]_{\mathbf{x}} \leq \frac{3}{\sigma^4} \Delta \chi_f.
\]

By algebraic manipulation, for all \( s, a \in \mathbb{R}^d \) and \( v \in \mathbb{R}^{d \times d} \) we have

\[
v^T (\nabla_{\psi}^2 \log \pi_{\psi}(a \mid s)) v = - \sum_{i=1}^{d} \left( \sum_{j=1}^{d} v_{i,j} s_j \right)^2.
\]

Consequently,

\[
(v^T (\nabla_{\psi}^2 \log \pi_{\psi}(a \mid s)) v)^4 \leq \|s\|_2^4 \|v\|_2^4 \leq \frac{1}{\sigma^8} \Delta \chi_h.
\]


Verifying item 1. We verify Assumption 2.1 by applying policy gradient lemma. Recall that

\[
\eta(\theta, \psi) = \mathbb{E}_{\tau \sim \rho_{\psi}^{H}} \left[ \sum_{h=1}^{H} r(s_h, a_h) \right].
\]

By policy gradient lemma (Lemma I.17) we have

\[
\nabla_{\psi} \eta(\theta, \psi) = \mathbb{E}_{\tau \sim \rho_{\psi}^{H}} \left[ \left( \sum_{h=1}^{H} \nabla_{\psi} \log \pi_{\psi}(a_h \mid s_h) \right) \left( \sum_{h=1}^{H} r(s_h, a_h) \right) \right].
\]
By Eq. (83), condition on $s_h$ we get
\[ \nabla_\psi \log \pi_\psi(a_h \mid s_h) = \frac{1}{\sigma^2} \text{vec}(u \otimes s_h) \] (98)
where $u = a_h - \psi s_h \sim \mathcal{N}(0, \sigma^2 I)$. Define the shorthand $g = \sum_{h=1}^{H} r(h, a_h)$. Note that by Hölder inequality,
\[
\| \mathbb{E}[\langle \nabla_{\psi} \log \pi_\psi(a_h \mid s_h)g \rangle] \|_2^2 = \sup_{\psi \in \mathbb{R}^{d \times d}, \|\psi\|_2 = 1} \mathbb{E}[\langle \nabla_{\psi} \log \pi_\psi(a_h \mid s_h), v \rangle] g^2 \leq \sup_{\psi \in \mathbb{R}^{d \times d}, \|\psi\|_2 = 1} \mathbb{E}[\langle \nabla_{\psi} \log \pi_\psi(a_h \mid s_h), v \rangle] \mathbb{E}[g^2].
\] (99)

Since $v \in \mathbb{R}^{d \times d}$, if we view $v$ as a $d \times d$ matrix then $\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle = \frac{1}{\sigma^2} \langle vs_h, u \rangle$. Because $u$ is an isotropic Gaussian random vector, $\langle vs_h, u \rangle \sim \mathcal{N}(0, \sigma^2 \|vs_h\|_2^2)$. Consequently,
\[
\mathbb{E}[\langle \nabla_{\psi} \log \pi_\psi(a_h \mid s_h), v \rangle] \leq \frac{1}{\sigma^2} \|vs_h\|_2^2 \leq \frac{1}{\sigma^2} \|v\|_F^2 \|s_h\|_2^2 \leq \frac{1}{\sigma^2}. \] (101)

It follows that $\|\mathbb{E}[\langle \nabla_{\psi} \log \pi_\psi(a_h \mid s_h)g \rangle] \|_2^2 \leq \frac{H^2}{\sigma^2}$. By triangular inequality and Eq. (97) we get
\[ \|\nabla_\psi \eta(\theta, \psi)\|_2 \leq H^2 / \sigma. \] (102)

**Verifying item 2.** Define the shorthand $f = \sum_{h=1}^{H} \nabla_\psi \log \pi_\psi(a_h | s_h)$. Use policy gradient lemma on Eq. (97) again we get, for any $v, w \in \mathbb{R}^{d \times d}$,
\[
v^T \mathbb{E}_{\rho \sim \rho_\psi} \left[ \left( f, v \right) \left( f, w \right) g + \sum_{h=1}^{H} v^T \nabla^2_\psi \log \pi_\psi(a_h \mid s_h)w \right] \] (103)

For the first term inside the expectation, we bound it by using Hölder inequality twice. Specifically, for any $h, h' \in [H]$ we have
\[
\mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle \langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w \rangle g] \leq \mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle \right]^{1/4} \mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w \rangle \right]^{1/4} \mathbb{E}[g]^{1/2}. \] (104)

Similarly, $\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle \sim \frac{1}{\sigma^2} \mathcal{N}(0, \sigma^2 \|vs_h\|_2^2)$ and $\langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w \rangle \sim \frac{1}{\sigma^2} \mathcal{N}(0, \sigma^2 \|ws_{h'}\|_2^2)$. As a result,
\[
\mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle \langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w \rangle g] \leq \frac{1}{\sigma^2} \|v\|_2 \|s_h\|_2 \|w\|_2 \|s_{h'}\|_2 \|H \leq \frac{3H^3}{\sigma^2}. \] (106)

Therefore the first term of Eq. (103) can be bounded by $\frac{3H^3}{\sigma^2}$. Now we bound the second term of Eq. (103). By algebraic manipulation we have
\[
v^T \nabla^2_\psi \log \pi_\psi(a_h \mid s_h)w = -\frac{1}{\sigma^2} \langle ws_h, vs_h \rangle. \] (108)

Consequently,
\[
\mathbb{E} \left[ \sum_{h=1}^{H} v^T \nabla^2_\psi \log \pi_\psi(a_h \mid s_h)w \right] g \leq \frac{H^2}{\sigma^2} \|w\|_2 \|v\|_2 \leq \frac{H^2}{\sigma^2}. \] (109)

In summary, we have $\|\nabla^2_\psi \eta(\theta, \psi)\|_{op} \leq \frac{4H^3}{\sigma^2}$. 
Verifying item 3. Now we turn to the last item in Assumption 2.1. First of all, following Eq. (108), we have \( \nabla_\psi^3 \log \pi_\psi(a_h \mid s_h) = 0 \). As a result, applying policy gradient lemma to Eq. (103) again we get
\[
\langle \nabla_\psi^3 \eta(\theta, \psi), v \otimes w \otimes x \rangle = \mathbb{E}_{\tau \sim \rho_\psi^3} \left[ \langle f, v \rangle \langle f, w \rangle g \right]
\]
\[
+ \mathbb{E}_{\tau \sim \rho_\psi^3} \left[ \langle f, x \rangle \left( \sum_{h=1}^{H} v^T \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)w \right) g \right]
\]
\[
+ \mathbb{E}_{\tau \sim \rho_\psi^3} \left[ \langle f, v \rangle \left( \sum_{h=1}^{H} x^T \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)w \right) g \right]
\]
\[
+ \mathbb{E}_{\tau \sim \rho_\psi^3} \left[ \langle f, w \rangle \left( \sum_{h=1}^{H} v^T \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)x \right) g \right].
\]

Following the same argument, by Hölder inequality, for any \( h_1, h_2, h_3 \in [H] \) we have
\[
\mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle \langle \nabla_\psi \log \pi_\psi(a_{h'}, s_{h'})w \rangle \langle \nabla_\psi \log \pi_\psi(a_{h''}, s_{h''})x \rangle g \right]
\]
\[
\leq \mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v \rangle^6 \right]^{1/6} \mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_{h'}, s_{h'})w \rangle^6 \right]^{1/6} \mathbb{E} \left[ \langle \nabla_\psi \log \pi_\psi(a_{h''}, s_{h''})x \rangle^6 \right]^{1/6} H
\]
\[
\leq \frac{15H^4}{\sigma^3}.
\]

On the other hand,
\[
\mathbb{E} \left[ \langle f, x \rangle \left( \sum_{h=1}^{H} v^T \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)w \right) g \right]
\]
\[
\leq \mathbb{E} \left[ \langle f, x \rangle^2 \right]^{1/2} \mathbb{E} \left[ \left( \sum_{h=1}^{H} v^T \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)w \right) g \right]^{1/2}
\]
\[
\leq \frac{H^4}{\sigma^3}.
\]

By symmetry, Eq. (110) can be upper bounded by
\[
\langle \nabla_\psi^3 \eta(\theta, \psi), v \otimes w \otimes x \rangle \leq \frac{7H^4}{\sigma^3}.
\]


Verifying item 1. We verify Assumption B.1 by coupling argument. First of all, consider the Lipschitzness of value function. By Bellman equation we have
\[
V_\theta^\psi(s) = \mathbb{E}_{a \sim \pi_\psi} \left[ r(s, a) + V_\theta^\psi(T(s, a)) \right]
\]
\[
= \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)} \left[ r(s, \psi s + u) + V_\theta^\psi(N_\theta(s + \psi s + u)) \right].
\]

Define \( B = 1 + \| \psi \|_{\text{op}} \) for shorthand. For two states \( s_1, s_2 \in S \), by the Lipschitz assumption on reward function we have
\[
|r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u)| \leq L_r B \| s_1 - s_2 \|_2.
\]

Then consider the second term in Eq. (116). Since we have \( |V_\theta^\psi| \leq H \) and
\[
\text{TV}(\mathcal{N}(s_1 + \psi s_1, \sigma^2 I), \mathcal{N}(s_2 + \psi s_2, \sigma^2 I)) \leq \frac{1}{2\sigma} \| s_1 + \psi s_1 - s_2 - \psi s_2 \|_2 \leq \frac{B \| s_1 - s_2 \|_2}{2\sigma},
\]
it follows that
\[
\left| \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)} \left[ V_\theta^\psi(N_\theta(s_1 + \psi s_1 + u)) \right] - \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)} \left[ V_\theta^\psi(N_\theta(s_2 + \psi s_2 + u)) \right] \right| \leq \frac{HB}{2\sigma} \| s_1 - s_2 \|_2.
\]
As a result, item 1 of Assumption B.1 holds as follows
\[ |V_\theta^v(s_1) - V_\theta^v(s_2)| \leq \left( \frac{HB}{2\sigma} + L_r B \right) \|s_1 - s_2\|_2. \] (118)

**Verifying item 2.** Now we turn to verifying the Lipschitzness of gradient term. Recall that by policy gradient lemma we have for every \( v \in \mathbb{R}^{d \times d}, \)
\[ \left\langle \nabla_v V_\theta^v(s), v \right\rangle = E_{a \sim \pi_v(s)} \left[ \left\langle \nabla_v V_\theta^v(N_\theta(s + a)), v \right\rangle \right] \] (119)
\[ + E_{a \sim \pi_v(s)} \left[ \nabla_v \log \pi_v(a | s, v) \left( r(s, a) + V_\theta^v(N_\theta(s + a)) \right) \right] \] (120)
\[ = E_{a \sim N(0, \sigma^2 I)} \left[ \left\langle \nabla_v V_\theta^v(N_\theta(s + \psi s + u)), v \right\rangle \right] \] (121)
\[ + E_{a \sim N(0, \sigma^2 I)} \left[ \nabla_v \log \pi_v(\psi s + u | s, v) \left( r(s, \psi s + u) + V_\theta^v(N_\theta(s + \psi s + u)) \right) \right]. \] (122)

Because for any two vectors \( g_1, g_2 \in \mathbb{R}^{d \times d} \|g_1 - g_2\|_2 = \sup_{v \in \mathbb{R}^{d \times d}} \left\langle g_1 - g_2, v \right\rangle, \) Lipschitzness of Eq. (119) for every \( v \in \mathbb{R}^{d \times d}, \|v\|_2 = 1 \) implies Lipschitzness of \( \nabla_v V_\theta^v(s). \)

By the boundness of \( \left\| \nabla_v V_\theta^v(s) \right\|_2 \) (specifically, item 1 of Assumption 2.1), we have
\[ \left| E_{u \sim N(0, \sigma^2 I)} \left[ \left\langle \nabla_v V_\theta^v(N_\theta(s_1 + \psi s_1 + u)), v \right\rangle \right] - E_{u \sim N(0, \sigma^2 I)} \left[ \left\langle \nabla_v V_\theta^v(N_\theta(s_2 + \psi s_2 + u)), v \right\rangle \right] \right| \] \[ \leq \frac{H^2}{\sigma} TV(N(\psi s_1, \sigma^2 I), N(\psi s_2, \sigma^2 I)) \leq \frac{H^2 B}{2\sigma} \|s_1 - s_2\|_2. \] (123)

For the reward term in Eq. (123), recalling \( v \in \mathbb{R}^{d \times d} \) we have
\[ E_{a \sim \pi_v(s)} \left[ (\nabla \log \pi_v(\psi s + u | s, v), r(s, \psi s + u)) \right] = E_{a \sim N(0, \sigma^2 I)} \left[ (v, r) \right]. \]

Note that
\[ E_{u \sim N(0, \sigma^2 I)} [(v_1, u) r(s_1, \psi s_1 + u)] - E_{u \sim N(0, \sigma^2 I)} [(v_2, u) r(s_2, \psi s_2 + u)] \] \[ = E_{u \sim N(0, \sigma^2 I)} [(v_1, u) r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u)] \] \[ + E_{u \sim N(0, \sigma^2 I)} [(v_1 - v_2, u) r(s_2, \psi s_2 + u)]. \] (124)

Note that \( u \) is isotropic. Applying Lemma I.10 we have
\[ E_{u \sim N(0, \sigma^2 I)} [(v_1, u) - (v_2, u) r(s_2, \psi s_2 + u)] \leq \sigma \|v_1 - v_2\|_2 \leq \sigma \|s_1 - s_2\|_2. \] (125)

We can also bound the term in Eq. (125) by
\[ E_{u \sim N(0, \sigma^2 I)} [(v_1, u) r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u)] \] \[ \leq E_{u \sim N(0, \sigma^2 I)} \left[ (v_1, u)^2 \right]^{1/2} E_{u \sim N(0, \sigma^2 I)} \left[ (r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u))^2 \right]^{1/2} \] \[ \leq \sigma E_{u \sim N(0, \sigma^2 I)} \left[ L^2 B^2 \|s_1 - s_2\|_2^2 \right]^{1/2} \leq \sigma L_r B \|s_1 - s_2\|_2. \] (126)

Now we deal with the last term in Eq. (123). Let \( f(s, u) = V_\theta^v(N_\theta(s + \psi s + u)) \) for shorthand. Similarly we have
\[ E_{u \sim N(0, \sigma^2 I)} \left[ (\nabla \log \pi_v(\psi s + u | s, v), V_\theta^v(N_\theta(s + \psi s + u)) \right] = E_{u \sim N(0, \sigma^2 I)} [(v, u) f(s, u)]. \] (127)

By the same telescope sum we get,
\[ E_{u \sim N(0, \sigma^2 I)} [(v_1, u) f(s_1, u)] - E_{u \sim N(0, \sigma^2 I)} [(v_2, u) f(s_2, u)] \] (128)
we have
\[ = \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_1, u \rangle (f(s_1, u) - f(s_2, u))]
+ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_2, u \rangle f(s_2, u)]. \] (133)

Applying Lemma I.10 we have
\[ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_2, u \rangle f(s_2, u)] \leq \sigma H \|vs_1 - vs_2\|_2 \leq \sigma H \|s_1 - s_2\|_2. \] (135)

Applying Lemma I.12 we have
\[ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_1, u \rangle (f(s_2, u) - f(s_2, u))] \leq 6BH \|s_1 - s_2\|_2 \left(1 + \frac{1}{\sigma}\right). \] (136)

In summary, we have
\[ \|\nabla_\psi V_\theta^\psi(s_1) - \nabla_\psi V_\theta^\psi(s_2)\|_2 \leq \text{poly}(H, B, \sigma, 1/\sigma, L_r) \|s_1 - s_2\|_2. \] (137)

**Verifying item 3.** Lastly, we verify the Lipschitzness of Hessian term. Applying policy gradient lemma to Eq. (119) again we have
\[ w^T \nabla^2 V_\theta^\psi(s)v = \mathbb{E}_{\alpha \sim \pi_\psi(s)}[w^T \nabla^2 V_\theta^\psi(\eta(s + \alpha))v] + \mathbb{E}_{\alpha \sim \pi_\psi(s)}[\langle \nabla_\psi V_\theta^\psi(\eta(s + \alpha)), v \rangle \langle \nabla_\psi \log \pi_\psi(a \mid s), w \rangle] + \mathbb{E}_{\alpha \sim \pi_\psi(s)}[\langle \nabla_\psi \log \pi_\psi(a \mid s), v \rangle \langle \nabla_\psi V_\theta^\psi(\eta(s + \alpha)), w \rangle] + \mathbb{E}_{\alpha \sim \pi_\psi(s)}[\langle \nabla_\psi \log \pi_\psi(a \mid s), v \rangle \langle \nabla_\psi \log \pi_\psi(a \mid s), w \rangle \langle r(s, a) + V_\theta^\psi(\eta(s + \alpha)) \rangle]. \] (138)

Recall that \( a \sim \psi s + \mathcal{N}(0, \sigma^2 I). \) In the sequel, we bound the Lipschitzness of above four terms separately.

By the upper bound of \( \|\nabla^2 V_\theta^\psi(\eta(s + \alpha))\|_{\text{op}} \) (specifically, item 2 of Assumption 2.1) we have
\[ \left| \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[w^T \nabla^2 V_\theta^\psi(\eta(s + \psi s + u))v] - \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[w^T \nabla^2 V_\theta^\psi(\eta(s + \psi s + u))v - \|w^T \nabla^2 V_\theta^\psi(\eta(s + \psi s + u))v\|_{\text{op}}] \right| \leq 4H^3 \frac{1}{\sigma^2} \text{TV}(\mathcal{N}(s_1 + \psi s_1, \sigma^2 I), \mathcal{N}(s_2 + \psi s_2, \sigma^2 I)) \leq 3H^3 \frac{B}{2\sigma} \|s_1 - s_2\|_2. \]

For the terms in Eq. (140), let \( f(s, u) = \langle \nabla_\psi V_\theta^\psi(\eta(s + \psi s + a), v) \rangle. \) Repeat the same argument when verifying item 2 again, we have
\[ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle w_s, u \rangle (f(s_2, u) - f(s_2, u))] \leq 6B \frac{H^2}{\sigma} \|s_1 - s_2\|_2 \left(1 + \frac{1}{\sigma}\right). \] (143)

Similarly, term in Eq. (141) also has the same Lipschitz constant.

Finally, we bound the term in Eq. (142). For the reward term in Eq. (142), recalling \( v, w \in \mathbb{R}^{d \times d} \) we have
\[ \mathbb{E}_{a \sim \pi_\psi(s)}[\langle \nabla_\psi \log \pi_\psi(a \mid s), v \rangle \langle \nabla_\psi \log \pi_\psi(a \mid s), w \rangle r(s, a)] = \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs, u \rangle \langle ws, u \rangle r(s, \psi s + u)]. \]

Note that
\[ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_1, u \rangle \langle ws_1, u \rangle r(s_1, \psi s_1 + u)] - \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_2, u \rangle \langle ws_2, u \rangle r(s_2, \psi s_2 + u)] = \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_1, u \rangle \langle ws_1, u \rangle (r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u))] \]
\[ + \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle vs_1, u \rangle \langle ws_1, u \rangle - \langle vs_2, u \rangle \langle ws_2, u \rangle] r(s_2, \psi s_2 + u). \] (144)
Note that $u$ is isotropic. Applying Lemma I.15 we have
\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle u s_1, u \rangle \langle w s_1, u \rangle - \langle u s_2, u \rangle \langle w s_2, u \rangle] r(s_2, \psi s_2 + u) \\
\leq \sqrt{3} \sigma^2 (\|u s_1 - u s_2\|_2 + \|w s_1 - w s_2\|_2) \leq 2 \sqrt{3} \sigma^2 \|s_1 - s_2\|_2.
\]

We can also bound the term in Eq. (145) by
\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle u s_1, u \rangle (r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u))]
\leq \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle u s_1, u \rangle]^{1/4} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle w s_1, u \rangle]^{1/4} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u)]^{1/2}
\leq \sqrt{3} \sigma^2 \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[L^2 B^2 \|s_1 - s_2\|_2]^{1/2} \leq \sqrt{3} \sigma^2 L B \|s_1 - s_2\|_2.
\]

Now we deal with the last term in Eq. (142). Let $f(s, u) = V^\psi_\theta(N\theta(s + \psi s + u))$ for shorthand. Similarly we have
\[
\mathbb{E}_{u \sim \pi_\psi(s)}[\langle \n abla \psi, \log \pi_\psi(a | s), v \rangle] \langle \n abla \psi, \log \pi_\psi(a | s), u \rangle V^\psi_\theta(N\theta(s + a))
\leq \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s, u \rangle \langle w s, u \rangle] f(s, u).
\]

By the same telescope sum we get,
\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_1, u \rangle \langle w s_1, u \rangle f(s_1, u)] - \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_2, u \rangle \langle w s_2, u \rangle f(s_2, u)]
\leq \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_1, u \rangle \langle w s_1, u \rangle f(s_1, u) - f(s_2, u)]
+ \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_1, u \rangle \langle w s_1, u \rangle - \langle v s_2, u \rangle \langle w s_2, u \rangle] f(s_2, u).
\]

Applying Lemma I.15 we have
\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_1, u \rangle \langle w s_1, u \rangle - \langle v s_2, u \rangle \langle w s_2, u \rangle] f(s_2, u)
\leq \sqrt{3} \sigma^2 H (\|v s_1 - v s_2\|_2 + \|w s_1 - w s_2\|_2) \leq 2 \sqrt{3} \sigma^2 H \|s_1 - s_2\|_2.
\]

Applying Lemma I.13 we have
\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v s_1, u \rangle \langle w s_1, u \rangle (f(s_2, u) - f(s_2, u)) \leq \text{poly}(H, \sigma, 1/\sigma) B \|s_1 - s_2\|_2.
\]

In summary, we have
\[
\left\| \n abla \psi V^\psi_\theta(s_1) - \n abla \psi V^\psi_\theta(s_2) \right\|_{\text{op}} \leq \text{poly}(H, B, \sigma, 1/\sigma, L_r) \|s_1 - s_2\|_2.
\]

I. Helper Lemmas

In this section, we list helper lemmas that are used in previous sections.

I.1. Helper Lemmas on Probability Analysis

The following lemma provides a concentration inequality on the norm of linear transformation of a Gaussian vector, which is used to prove Lemma I.3.

Lemma I.1 (Theorem 1 of Hsu et al. (2012)). For $v \sim \mathcal{N}(0, I)$ be a $n$ dimensional Gaussian vector, and $A \in \mathbb{R}^{n \times n}$. Let $\Sigma = A^\top A$, then
\[
\forall t > 0, \text{Pr} \left[ \|Av\|_2^2 \geq \text{Tr}(\Sigma) + 2 \sqrt{\text{Tr}(\Sigma^2)} t + 2 \|\Sigma\|_{\text{op}} t \right] \leq \exp(-t).
\]

Corollary I.2. Under the same settings of Lemma I.1,
\[
\forall t > 1, \text{Pr} \left[ \|Av\|_2^2 \geq \|A\|_F^2 + 4 \|A\|_F^2 t \right] \leq \exp(-t).
\]
Proof. Let $\lambda_i$ be the $i$-th eigenvalue of $\Sigma$. By the definition of $\Sigma$ we have $\lambda_i \geq 0$. Then we have
\[
\text{Tr}(\Sigma) = \sum_{i=1}^{n} \lambda_i = \|A\|_{F}^{2},
\]
\[
\text{Tr}(\Sigma^2) = \sum_{i=1}^{n} \lambda_i^2 \leq \left( \sum_{i=1}^{n} \lambda_i \right)^2 = \|A\|_{F}^{4},
\]
\[
\|\Sigma\|_{op} = \max_{\|x\|_{\infty} = 1} \|Ax\|_{2} \leq \sum_{i=1}^{n} \lambda_i = \|A\|_{F}^{2}.
\]
Plug in Eq. (156), we get the desired equation.

Next lemma proves a concentration inequality on which Lemma F.2 relies.

Lemma I.3. Given a symmetric matrix $H$, let $u, v \sim \mathcal{N}(0, I)$ be two independent random vectors, we have
\[
\forall t \geq 1, \Pr[(u^\top Hv)^2 \geq t \|H\|_{F}^{2}] \leq 3 \exp(-\sqrt{t}/4).
\]
Proof. Condition on $v$, $u^\top Hv$ is a Gaussian random variable with mean zero and variance $\|Hv\|_{2}^{2}$. Therefore we have,
\[
\forall v, \Pr[(u^\top Hv)^2 \geq \sqrt{t} \|Hv\|_{2}^{2}] \leq \exp(-\sqrt{t}/2).
\]
By Corollary I.2 and basic algebra we get,
\[
\Pr[\|Hv\|_{2}^{2} \geq \sqrt{t} \|H\|_{F}^{2}] \leq 2 \exp(-\sqrt{t}/4).
\]
Consequently,
\[
\mathbb{E}[\mathbb{I}[(u^\top Hv)^2 \geq t \|H\|_{F}^{2}]]
\leq \mathbb{E}[\mathbb{I}[(u^\top Hv)^2 \geq \sqrt{t} \|Hv\|_{2}^{2} \text{ or } \|Hv\|_{2}^{2} \geq \sqrt{t} \|H\|_{F}^{2}]]
\leq \mathbb{E}[\mathbb{I}[(u^\top Hv)^2 \geq \sqrt{t} \|Hv\|_{2}^{2}] | v] + \mathbb{E}[\mathbb{I}[(Hv)^{2} \geq \sqrt{t} \|H\|_{F}^{2}]]
\leq 3 \exp(-\sqrt{t}/4).
\]
(Combining Eq. (159) and Eq. (160))

The next two lemmas are dedicated to prove anti-concentration inequalities that is used in Lemma F.2.

Lemma I.4 (Lemma 1 of Laurent, Massart (2000)). Let $(y_1, \cdots, y_n)$ be i.i.d. $\mathcal{N}(0, 1)$ Gaussian variables. Let $\alpha = (a_1, \cdots, a_n)$ be non-negative coefficient. Let
\[
\|\alpha\|_{2}^{2} = \sum_{i=1}^{n} a_i^2.
\]
Then for any positive $t$,
\[
\Pr\left(\sum_{i=1}^{n} a_i y_i^2 \leq \sum_{i=1}^{n} a_i - 2 \|\alpha\|_{2} \sqrt{t}\right) \leq \exp(-t).
\]

Lemma I.5. Given a symmetric matrix $H \in \mathbb{R}^{n \times n}$, let $u, v \sim \mathcal{N}(0, I)$ be two independent random vectors. Then
\[
\Pr[(u^\top Hv)^2 \geq \frac{1}{8} \|H\|_{F}^{2}] \geq \frac{1}{64}.
\]
Proof. Since \( u, v \) are independent, by the isotropy of Gaussian vectors we can assume that \( H = \text{diag}(\lambda_1, \cdots, \lambda_n) \). Note that condition on \( v, u^\top H v \) is a Gaussian random variable with mean zero and variance \( \|Hv\|_2^2 \). As a result,

\[
\forall v, \Pr\left( (u^\top H v)^2 \geq \frac{1}{4} \|Hv\|_2^2 \mid v \right) \geq \frac{1}{2}. \tag{163}
\]

On the other hand, \( \|Hv\|_2^2 = \sum_{i=1}^n \lambda_i^2 v_i^2 \). Invoking Lemma I.4 we have

\[
\Pr \left[ \|Hv\|_2^2 \geq \frac{1}{2} \|H\|_F^2 \right] \geq \frac{1}{2} \sqrt{\sum_{i=1}^n \lambda_i^4} = \frac{1}{2} \sqrt{\sum_{i=1}^n \lambda_i^2} \geq \frac{1}{2} \sqrt{\sum_{i=1}^n \lambda_i^4} \tag{164}
\]

Combining Eq. (163) and Eq. (164) we get,

\[
\Pr \left[ (u^\top H v)^2 \geq \frac{1}{8} \|H\|_F^2 \right] \geq \Pr \left[ (u^\top H v)^2 \geq \frac{1}{4} \|Hv\|_2^2 \mid \|Hv\|_2^2 \geq \frac{1}{2} \|H\|_F^2 \right] \geq \frac{1}{64}. \tag{165}
\]

The following lemma justifies the cap in the loss function.

**Lemma I.6.** Given a symmetric matrix \( H \), let \( u, v \sim \mathcal{N}(0, I) \) be two independent random vectors. Let \( \kappa_2, c_1 \in \mathbb{R}_+ \) be two numbers satisfying \( \kappa_2 \geq 640 \sqrt{2} c_1 \), then

\[
\min \left( c_1^2, \|H\|_F^2 \right) \leq 2 \mathbb{E} \left[ \min \left( \kappa_2^2, (u^\top H v)^2 \right) \right]. \tag{166}
\]

Proof. Let \( x = (u^\top H v)^2 \) for simplicity. Consider the following two cases:

**Case 1:** \( \|H\|_F \leq \kappa_2/40 \). In this case we exploit the tail bound of random variable \( x \). Specifically,

\[
\mathbb{E} \left[ (u^\top H v)^2 \right] - \mathbb{E} \left[ \min \left( \kappa_2^2, (u^\top H v)^2 \right) \right] \\
= \int_{\kappa_2^2}^\infty \Pr \{ x \geq t \} dt \\
\leq 3 \int_{\kappa_2^2}^\infty \exp \left( -\frac{1}{4} \sqrt{\frac{t}{\|H\|_F^2}} \right) dt \tag{By Lemma I.3} \\
= 24 \exp \left( -\frac{\kappa_2^2}{4 \|H\|_F^2} \right) \|H\|_F \left( \kappa_2 + 4 \|H\|_F \right) \leq 48 \exp \left( -\frac{\kappa_2^2}{4 \|H\|_F^2} \right) \|H\|_F \kappa_2 \leq 48 \cdot \frac{4 \|H\|_F^2}{384 \kappa_2} \|H\|_F \kappa_2 \leq \frac{\|H\|_F^2}{2}.
\]

As a result,

\[
\mathbb{E} \left[ \min \left( \kappa_2^2, (u^\top H v)^2 \right) \right] \geq \mathbb{E} \left[ (u^\top H v)^2 \right] - \frac{\|H\|_F^2}{2} = \frac{\|H\|_F^2}{2}. \tag{166}
\]
**Case 2:** $\|H\|_F > \kappa_2/40$. In this case, we exploit the anti-concentration result of random variable $x$. Note that by the choice of $\kappa_2$, we have

$$\|H\|_F > \kappa_2/40 \implies \frac{1}{8} \|H\|_F^2 \geq 64c_1^2.$$ 

As a result,

$$\mathbb{E}\left[ \min\left( \kappa_2^2, (u^\top Hv)^2 \right) \right] \geq 64c_1^2 \Pr\left[ \min\left( \kappa_2^2, (u^\top Hv)^2 \right) \geq 64c_1^2 \right] \geq 64c_1^2 \Pr\left[ (u^\top Hv)^2 \geq \frac{1}{8} \|H\|_F^2 \right] \geq c_1^2.$$ 

(By definition of $\kappa_2$)

Therefore, in both cases we get

$$\mathbb{E}\left[ \min\left( \kappa_2^2, (u^\top Hv)^2 \right) \right] \geq \frac{1}{2} \min\left( c_1^2, \|H\|_F^2 \right), \quad (167)$$

which proofs Eq. (165).

Following lemmas are analogs to Cauchy-Schwartz inequality (in vector/matrix forms), which are used to prove Lemma G.1 for reinforcement learning case.

**Lemma I.7.** For a random vector $x \in \mathbb{R}^d$ and random variable $r$, we have

$$\|\mathbb{E}[rx]\|_2^2 \leq \|\mathbb{E}[xx^\top]\|_{\text{op}} \mathbb{E}[r^2]. \quad (168)$$

**Proof.** Note that for any vector $g \in \mathbb{R}^d$, $\|g\|_2^2 = \sup_{u \in S^{d-1}} \langle u, g \rangle^2$. As a result,

$$\|\mathbb{E}[rx]\|_2^2 = \sup_{u \in S^{d-1}} \langle u, \mathbb{E}[rx] \rangle^2 = \sup_{u \in S^{d-1}} \mathbb{E}[ r \langle u, x \rangle ]^2 \leq \sup_{u \in S^{d-1}} \mathbb{E}\left[ \|u, x\|^2 \right] \mathbb{E}[r^2] \quad \text{(Hölder Inequality)}$$

$$= \|\mathbb{E}[xx^\top]\|_{\text{op}} \mathbb{E}[r^2].$$

**Lemma I.8.** For a symmetric random matrix $H \in \mathbb{R}^{d \times d}$ and random variable $r$, we have

$$\|\mathbb{E}[rH]\|_{\text{sp}}^2 \leq \|\mathbb{E}[HH^\top]\|_{\text{sp}} \mathbb{E}[r^2]. \quad (169)$$

**Proof.** Note that for any matrix $G \in \mathbb{R}^d$, $\|G\|_{\text{sp}}^2 = \sup_{u, v \in S^{d-1}} \langle u^\top Gv \rangle^2$. As a result,

$$\|\mathbb{E}[rH]\|_2^2 = \sup_{u, v \in S^{d-1}} \langle u^\top \mathbb{E}[rH]v \rangle^2 = \sup_{u, v \in S^{d-1}} \mathbb{E}[r(u^\top Hv)^2] \leq \sup_{u, v \in S^{d-1}} \mathbb{E}\left[ \|u^\top Hv\|^2 \right] \mathbb{E}[r^2] \quad \text{(Hölder Inequality)}$$

$$= \sup_{u, v \in S^{d-1}} \mathbb{E}\left[ uu^\top Hv^\top Hu \right] \mathbb{E}[r^2] \leq \sup_{u \in S^{d-1}} \mathbb{E}\left[ uu^\top HH^\top u \right] \mathbb{E}[r^2] \leq \|\mathbb{E}[HH^\top]\|_{\text{sp}} \mathbb{E}[r^2].$$
Lemma I.9. For a random matrix \( x \in \mathbb{R}^d \) and a positive random variable \( r \), we have
\[
\| \mathbb{E}[rxx^\top] \|_{sp}^2 \leq \| \mathbb{E}[x^\otimes 4] \|_{sp} \mathbb{E}[r^2].
\] (170)

Proof. Since \( r \) is non-negative, we have \( \mathbb{E}[rxx^\top] \geq 0 \). As a result,
\[
\| \mathbb{E}[rxx^\top] \|_{sp} = \sup_{u \in S^{d-1}} u^\top \mathbb{E}[rxx^\top] u.
\]
It follows that
\[
\| \mathbb{E}[rxx^\top] \|_{sp}^2 = \sup_{u \in S^{d-1}} \left( u^\top \mathbb{E}[rxx^\top] u \right)^2 = \sup_{u \in S^{d-1}} \mathbb{E} \left[ r \langle u, x \rangle^2 \right]^2
\]
\[
\leq \sup_{u \in S^{d-1}} \mathbb{E} \left[ (u^\otimes 4, \mathbb{E}[x^\otimes 4]) \mathbb{E}[r^2] \right]
\]
\[
= \| \mathbb{E}[x^\otimes 4] \|_{sp} \mathbb{E}[r^2].
\]

\[ \square \]

Following lemmas exploit the isotropism of Gaussian vectors, and are used to verify the Lipschitzness assumption of Example B.3. In fact, we heavily rely on the fact that, for a fixed vector \( g \in \mathbb{R}^d \), \( \langle g, u \rangle \sim N(0, \| g \|_2^2) \) when \( u \sim N(0, I) \).

Lemma I.10. For two vectors \( p, q \in \mathbb{R}^d \) and a bounded function \( f : \mathbb{R}^d \to [-B, B] \), we have
\[
\mathbb{E}_{u \sim N(0, \sigma^2 I)} \left[ (\langle p, u \rangle - \langle q, u \rangle) f(u) \right] \leq \sigma B \| p - q \|_2.
\] (171)

Proof. By Hölder inequality we have
\[
\mathbb{E}_{u \sim N(0, \sigma^2 I)} \left[ (\langle p, u \rangle - \langle q, u \rangle) f(u) \right]
\]
\[
\leq \mathbb{E}_{u \sim N(0, \sigma^2 I)} \left[ \langle p, u \rangle - \langle q, u \rangle \right]^{1/2} \mathbb{E}_{u \sim N(0, \sigma^2 I)} \left( f(u)^2 \right)^{1/2}.
\] (172)
(173)

Note that \( u \) is isotropic. As a result \( \langle p - q, u \rangle \sim N(0, \sigma^2 \| p - q \|_2^2) \). It follows that
\[
\mathbb{E}_{u \sim N(0, \sigma^2 I)} \left[ \langle p, u \rangle - \langle q, u \rangle \right]^{1/2} \mathbb{E}_{u \sim N(0, \sigma^2 I)} \left( f(u)^2 \right)^{1/2}
\]
\[
\leq \sigma \| p - q \|_2 B.
\] (174)
(175)

\[ \square \]

Lemma I.11. For two vectors \( x, y \in \mathbb{R}^d \), if \( \| x \|_2 = 1 \) we have
\[
\| x - y \|_2^2 \geq (1 - \langle x, y \rangle)^2.
\] (176)

Proof. By basic algebra we get
\[
\| x - y \|_2^2 = \| x - \langle x, y \rangle x + \langle x, y \rangle x - y \|_2^2
\]
\[
= \| x - \langle x, y \rangle x \|_2^2 + \| \langle x, y \rangle x - y \|_2^2 - 2(1 - \langle x, y \rangle \langle x, y \rangle x - y)\langle x, y \rangle x - y)
\]
\[
= \| x - \langle x, y \rangle x \|_2^2 + \| \langle x, y \rangle x - y \|_2^2
\]
\[
= (1 - \langle x, y \rangle)^2 + \| \langle x, y \rangle x - y \|_2^2 \geq (1 - \langle x, y \rangle)^2.
\] (177)
(178)
(179)
(180)

\[ \square \]
Lemma I.12. For vectors \( p, x_1, x_2 \in \mathbb{R}^d \) and a bounded function \( f : \mathbb{R}^d \to [0, B] \), we have

\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[(p, u) (f(x_1 + u) - f(x_2 + u))] \leq B \|p\|_2 \|x_1 - x_2\|_2 \left( 6 + \frac{3 \|x_1\|_2 + \|x_2\|_2}{\sigma} \right).
\] (181)

**Proof.** The lemma is proved by coupling argument. With out loss of generality, we assume that \( x_1 = C e_1 \) and \( x_2 = -C e_1 \) where \( e_1 \) is the first basis vector. That is, \( \|x_1 - x_2\|_2 = 2C \). For a vector \( x \in \mathbb{R}^d \), let \( F(x) \) be the density of distribution \( \mathcal{N}(0, \sigma^2 I) \) at \( x \). Then we have,

\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[(p, u) f(x_1 + u)] = \int_{y \in \mathbb{R}^d} F(y - x_1) \langle p, y - x_1 \rangle f(y)dy.
\] (182)

As a result,

\[
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[(p, u) (f(x_1 + u) - f(x_2 + u))] = \int_{y \in \mathbb{R}^d} F(y - x_1) \langle p, y - x_1 \rangle f(y)dy - \int_{y \in \mathbb{R}^d} F(y - x_2) \langle p, y - x_2 \rangle f(y)dy.
\] (183)

Define \( G(y) = \min(F(y - x_1), F(y - x_2)) \). It follows that,

\[
\int_{y \in \mathbb{R}^d} F(y - x_1) \langle p, y - x_1 \rangle f(y)dy - \int_{y \in \mathbb{R}^d} F(y - x_2) \langle p, y - x_2 \rangle f(y)dy \\
\leq \int_{y \in \mathbb{R}^d} G(y) \langle p, y - x_1 \rangle - \langle p, y - x_2 \rangle |f(y)dy| \tag{186}
\]

\[
+ \int_{y \in \mathbb{R}^d} (F(y - x_1) - G(y)) \langle p, y - x_1 \rangle |f(y)dy| \tag{187}
\]

\[
+ \int_{y \in \mathbb{R}^d} (F(y - x_2) - G(y)) \langle p, y - x_2 \rangle |f(y)dy|. \tag{188}
\]

The term in Eq. (186) can be bounded by

\[
\int_{y \in \mathbb{R}^d} G(y) \langle p, y - x_1 \rangle - \langle p, y - x_2 \rangle |f(y)dy| \tag{189}
\]

\[
\leq \int_{y \in \mathbb{R}^d} G(y)dy \sup_{y \in \mathbb{R}^d} |\langle p, x_2 - x_1 \rangle f(y)| \leq \|x_2 - x_1\|_2 \|p\|_2 B. \tag{190}
\]

Note that the terms in Eq. (187) and Eq. (188) are symmetric. Therefore in the following we only prove an upper bound for Eq. (187). In the following, we use the notation \( [y]_{-1} \) to denote the \((d - 1)\)-dimensional vector generated by removing the first coordinate of \( y \). Let \( P(x) \) be the density of distribution \( \mathcal{N}(0, \sigma^2) \) at point \( x \in \mathbb{R} \). By the symmetricity of Gaussian distribution, \( F(y) = P([y]_1) F([y]_{-1}) \).

By definition, \( F(y - x_1) - G(y) = 0 \) for \( y \) such that \( [y]_1 \leq 0 \). As a result,

\[
\int_{y: [y]_1 > 0} (F(y - x_1) - G(y)) \langle p, y - x_1 \rangle |f(y)dy| \tag{191}
\]

\[
= \int_{y: [y]_1 > 0} (F(y - x_1) - F(y - x_2)) \langle p, y - x_1 \rangle |f(y)dy| \tag{192}
\]

\[
\leq \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) \mathbb{E}_{[y]_{-1}}[|\langle [p]_{-1}, [y - x_1]_{-1} \rangle + [p]_1 [y - x_1]_1| f(y)]. \tag{193}
\]

Note that conditioned on \( [y]_1, [y - x_1]_{-1} \sim \mathcal{N}(0, \sigma^2 I) \). Consequently,

\[
\mathbb{E}_{[y]_{-1}}[|\langle [p]_{-1}, [y - x_1]_{-1} \rangle| f(y)] \leq \mathbb{E}_{[y]_{-1}}[\langle [p]_{-1}, [y - x_1]_{-1} \rangle^2]^{1/2} \mathbb{E}_{[y]_{-1}}[f(y)^2]^{1/2} \tag{194}
\]

\[
\leq \mathbb{E}_{[y]_{-1}}[\langle [p]_{-1}, [y - x_1]_{-1} \rangle^2]^{1/2} \mathbb{E}_{[y]_{-1}}[f(y)^2]^{1/2}. \tag{195}
\]
As a result,

\[ \text{Lemma I.13.} \]

Note that the second term in Eq. (199) involves only one dimensional Gaussian distribution. Invoking Lemma I.14, the second term can be bounded by \( \|p\|_2 \), \( \|q\|_2 \), and \( \|x\|_2 \) as follows:

\[ \int_{|y|, |z| > 0} (F(y - x) - G(y)) (p, y - x) f(y) dy \leq B \|p\|_2 \|x - x_2\|_2 + B \int_{|y|, |z| > 0} d[y]_1 (P([y], [z])) (p, y - x_1) \|p\|_2 \|y - x_1\|_2. \]

It follows that,

\[ \int_{|y|, |z| > 0} (F(y - x) - G(y)) (p, y - x) f(y) dy \leq B \|p\|_2 \|x - x_2\|_2 + B \int_{|y|, |z| > 0} d[y]_1 (P([y], [z])) (p, y - x_1) \|p\|_2 \|y - x_1\|_2. \]

Note that the second term in Eq. (199) involves only one dimensional Gaussian distribution. Invoking Lemma I.14, the second term can be bounded by \( \|p\|_2 \left( \frac{3|x - x_2|}{r} \|x_1\|_2 + 4 \|x_1 - x_2\|_2 \right) \). Therefore, we have

\[ \int_{|y|, |z| > 0} (F(y - x) - G(y)) (p, y - x) f(y) dy \leq B \|p\|_2 \|x - x_2\|_2 + \|p\|_2 \frac{3B \|x - x_2\|_2}{r}. \]

\[ \square \]

**Lemma I.13.** For vectors \( p, q, x_1, x_2 \in \mathbb{R}^d \) with \( \|x_1\|_2 \leq 1 \leq \|x_2\|_2 \leq 1 \) and a bounded function \( f : \mathbb{R}^d \rightarrow [0, B] \), we have

\[ \mathbb{E}_{u \sim N(0, \sigma^2 I)} [(p, u) (q, u) (f(x_1 + u) - f(x_2 + u))] \leq \text{poly}(B, \sigma, 1/\sigma, \|p\|_2, \|q\|_2) \|x_1 - x_2\|_2. \]

**Proof.** Proof of this lemma is similar to that of Lemma I.13. With out loss of generality, we assume that \( x_1 = C e_1 \) and \( x_2 = -C e_1 \) where \( e_1 \) is the first basis vector. That is, \( \|x_1 - x_2\|_2 = 2C \). For a vector \( x \in \mathbb{R}^d \), let \( F(x) \) be the density of distribution \( N(0, \sigma^2 I) \) at \( x \). Then we have,

\[ \mathbb{E}_{u \sim N(0, \sigma^2 I)} [(p, u) (q, u) f(x_1 + u)] = \int_{y \in \mathbb{R}^d} F(y - x_1) (p, y - x_1) (q, y - x_1) f(y) dy. \]

As a result,

\[ \mathbb{E}_{u \sim N(0, \sigma^2 I)} [(p, u) (q, u) (f(x_1 + u) - f(x_2 + u))] \]

\[ = \int_{y \in \mathbb{R}^d} F(y - x_1) (p, y - x_1) (q, y - x_1) f(y) dy - F(y - x_2) (p, y - x_2) (q, y - x_2) f(y) dy. \]

Define \( G(y) = \min(F(y - x_1), F(y - x_2)) \). It follows that,

\[ \int_{y \in \mathbb{R}^d} F(y - x_1) (p, y - x_1) (q, y - x_1) f(y) dy - F(y - x_2) (p, y - x_2) (q, y - x_2) f(y) dy \]

\[ \leq \int_{y \in \mathbb{R}^d} G(y) (p, y - x_1) (q, y - x_1) - (p, y - x_2) (q, y - x_2) f(y) dy \]

\[ + \int_{y \in \mathbb{R}^d} (F(y - x_1) - G(y)) (p, y - x_1) (q, y - x_1) f(y) dy \]

\[ + \int_{y \in \mathbb{R}^d} (F(y - x_2) - G(y)) (p, y - x_2) (q, y - x_2) f(y) dy. \]

By basic algebra we have

\[ \int_{y \in \mathbb{R}^d} G(y) (p, y - x_1) (q, y - x_1) - (p, y - x_2) (q, y - x_2) f(y) dy \]
\begin{align}
\leq \int_{y \in \mathbb{R}^d} G(y) \langle p, x_2 - x_1 \rangle \langle q, y - x_1 \rangle |f(y)| dy \\
+ \int_{y \in \mathbb{R}^d} G(y) \langle p, y - x_2 \rangle \langle q, x_2 - x_1 \rangle |f(y)| dy.
\end{align}

Continue with the first term we get
\begin{align}
\int_{y \in \mathbb{R}^d} G(y) \langle p, x_2 - x_1 \rangle \langle q, y - x_1 \rangle |f(y)| dy \\
\leq \|x_2 - x_1\|_2 \|p\|_2 \int_{y \in \mathbb{R}^d} G(y) \langle q, y - x_1 \rangle |f(y)| dy \\
\leq \|x_2 - x_1\|_2 \|p\|_2 \int_{y \in \mathbb{R}^d} F(y - x_1) \langle q, y - x_1 \rangle |f(y)| dy \\
= \|x_2 - x_1\|_2 \|p\|_2 E_{u \sim N(0, \sigma^2 I)} \left[ |\langle q, u \rangle| f(u + x_1) \right] \\
\leq \|x_2 - x_1\|_2 \|p\|_2 E_{u \sim N(0, \sigma^2 I)} \left[ (q, u)^2 \right]^{1/2} E_{u \sim N(0, \sigma^2 I)} \left[ f(u + x_1)^2 \right]^{1/2} \\
\leq \sigma B \|x_2 - x_1\|_2 \|p\|_2 \|q\|_2.
\end{align}

For the same reason, the second term in Eq. (211) is also bounded by $\sigma B \|x_2 - x_1\|_2 \|p\|_2 \|q\|_2$. As a result,
\begin{align}
\int_{y \in \mathbb{R}^d} G(y) \langle p, y - x_1 \rangle \langle q, y - x_1 \rangle - \langle p, y - x_2 \rangle \langle q, y - x_2 \rangle |f(y)| dy \\
\leq 2\sigma B \|x_2 - x_1\|_2 \|p\|_2 \|q\|_2.
\end{align}

Now we turn to the term in Eq. (207). Note that the terms in Eq. (207) and Eq. (208) are symmetric. Therefore in the following we only prove an upper bound for Eq. (207). In the following, we use the notation $[y]_{-1}$ to denote the $(d-1)$-dimensional vector generated by removing the first coordinate of $y$. Let $P(x)$ be the density of distribution $\mathcal{N}(0, \sigma^2)$ at point $x \in \mathbb{R}$. By the symmetricity of Gaussian distribution, $F(y) = P([y]_1) F([y]_{-1})$.

By definition, $F(y - x_1) - G(y) = 0$ for $y$ such that $[y]_1 \leq 0$. Define the shorthand $I = |[p]_1| |y - x_1|_1|, J = |([p]_{-1}, y - x_1|_{-1}), C = |[q]_1| |y - x_1|_1|, D = |([q]_{-1}, y - x_1|_{-1}).$ When condition on $[y]_1$, $A, C$ are constants. As a result,
\begin{align}
\int_{[y]_1 > 0} (F(y - x_1) - G(y)) \langle p, y - x_1 \rangle \langle q, y - x_1 \rangle |f(y)| dy \\
= \int_{[y]_1 > 0} (F(y - x_1) - F(y - x_2)) \langle p, y - x_1 \rangle \langle q, y - x_1 \rangle |f(y)| dy \\
\leq \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) E_{[y]_{-1}} \left[ |(I + J)(C + D)| f(y) \right] \\
\leq \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) I C E_{[y]_{-1}} [f(y)] \\
+ \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) I E_{[y]_{-1}} [D f(y)] \\
+ \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) C E_{[y]_{-1}} [J f(y)] \\
+ \int_{[y]_1 > 0} d[y]_1 (P([y]_1 - [x_1]_1) - P([y]_1 - [x_2]_1)) E_{[y]_{-1}} [J D f(y)].
\end{align}

Note that conditioned on $[y]_1$, $[y - x_1]_{-1} \sim \mathcal{N}(0, \sigma^2 I)$. Consequently,
\begin{align}
E_{[y]_{-1}} [D f(y)] \leq E_{[y]_{-1}} [D^2]^{1/2} E_{[y]_{-1}} [f(y)^2]^{1/2} \leq B \sigma \|q\|_2 \\
E_{[y]_{-1}} [J f(y)] \leq E_{[y]_{-1}} [J^2]^{1/2} E_{[y]_{-1}} [f(y)^2]^{1/2} \leq B \sigma \|p\|_2
\end{align}
As a result, we get
\[
\mathbb{E}_{[y]_1} [JDf(y)] \leq \mathbb{E}_{[y]_1} \left[ J^4 \right]^{1/4} \mathbb{E}_{[y]_1} \left[ D^4 \right]^{1/4} \mathbb{E}_{[y]_1} [f(y)^2]^{1/2} \leq \sqrt{3}B\sigma^2 \|p\|_2 \|q\|_2. \tag{228}
\]

Invoking Lemma I.14 we get
\[
\int_{[y]_1,0} d[y]_1 (P([y]_1 - [x]_1) - P([y]_1 - [x]_1)) \leq \frac{1}{\sigma} \|x_1 - x_2\|_2, \tag{229}
\]
\[
\int_{[y]_1,0} d[y]_1 (P([y]_1 - [x]_1) - P([y]_1 - [x]_1))I \leq \text{poly}(B, \sigma, 1/\sigma) \|p\|_2 \|x_1 - x_2\|_2, \tag{230}
\]
\[
\int_{[y]_1,0} d[y]_1 (P([y]_1 - [x]_1) - P([y]_1 - [x]_1))C \leq \text{poly}(B, \sigma, 1/\sigma) \|q\|_2 \|x_1 - x_2\|_2, \tag{231}
\]
\[
\int_{[y]_1,0} d[y]_1 (P([y]_1 - [x]_1) - P([y]_1 - [x]_1))IC \leq \text{poly}(B, \sigma, 1/\sigma) \|p\|_2 \|q\|_2 \|x_1 - x_2\|_2. \tag{232}
\]

As a result, we get
\[
\int_{y,0} (F(y - x_1) - G(y)) \langle p, y - x_1 \rangle \langle q, y - x_1 \rangle |f(y)|dy \leq \text{poly}(B, \sigma, 1/\sigma, \|p\|_2, \|q\|_2) \|x_1 - x_2\|_2. \tag{236}
\]

**Lemma I.14.** Let \(P(x)\) be the density function of \(N(0, \sigma^2)\). Given a scalar \(x \geq 0\) we have
\[
\int_{y,0} (P(y - x) - P(y + x)) dy \leq \frac{x}{\sigma}, \tag{233}
\]
\[
\int_{y,0} (P(y - x) - P(y + x)) |y - x| dy \leq \frac{3x^2}{\sigma} + 4x, \tag{234}
\]
\[
\int_{y,0} (P(y - x) - P(y + x)) |y - x|^2 dy \leq \frac{x^2}{\sigma} + 4\sigma x. \tag{235}
\]

**Proof.** Note that \(\text{TV}(N(-x, \sigma^2), N(x, \sigma^2)) \leq \frac{x}{\sigma}\). Consequently,
\[
\int_{y,0} (P(y - x) - P(y + x)) dy \leq \frac{x}{\sigma}. \tag{236}
\]

Using the same TV-distance bound, we get
\[
\int_{y,0} (P(y - x) - P(y + x)) |y - x| dy \tag{237}
\]
\[
\leq \int_{y,0} (P(y - x) - P(y + x)) ((y - x) + 2x) dy \tag{238}
\]
\[
\leq \int_{y > 2x} (P(y - x) - P(y + x)) (y - x) dy + \frac{3x^2}{\sigma}. \tag{239}
\]

Recall \(P(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2\sigma^2} \right)\). By algebraic manipulation we have
\[
\int_{y > 2x} \left( \exp \left( -\frac{(y - x)^2}{2\sigma^2} \right) - \exp \left( -\frac{(y + x)^2}{2\sigma^2} \right) \right) (y - x) dy \tag{240}
\]
\[
\leq \int_{y > 2x} \exp \left( -\frac{(y - x)^2}{2\sigma^2} \right) \left( 1 - \exp \left( -\frac{4xy}{2\sigma^2} \right) \right) (y - x) dy \tag{241}
\]
\[
\leq \int_{y > 2x} \exp \left( -\frac{(y - x)^2}{2\sigma^2} \right) \frac{4xy}{2\sigma^2} (y - x) dy \tag{242}
\]
Now we turn to the third inequality. Because $|y - x| \leq x$ for $y \in [0, 2x]$, using the TV-distance bound we get
\begin{align*}
\int_{y>0} (P(y - x) - P(y + x))|y - x|^2 dy &\leq \int_{y>2x} (P(y - x) - P(y + x))(y - x)^2 dy + \frac{x^2}{\sigma}.
\end{align*}
(245)

By algebraic manipulation we have
\begin{align*}
\int_{y>2x} \left( \exp\left(\frac{-(y - x)^2}{2\sigma^2}\right) - \exp\left(\frac{-(y + x)^2}{2\sigma^2}\right) \right) (y - x)^2 dy &\leq \int_{y>2x} \left( \exp\left(\frac{-(y - x)^2}{2\sigma^2}\right) - \frac{4xy}{\sigma^2} \right) (y - x)^2 dy \\
&\leq \int_{y>2x} \exp\left(\frac{-(y - x)^2}{2\sigma^2}\right) \frac{4xy}{\sigma^2} (y - x)^2 dy \\
&\leq \int_{y>2x} \exp\left(\frac{-(y - x)^2}{2\sigma^2}\right) \frac{4x}{\sigma^2} (y - x)^3 dy \\
&\leq 4\sigma x.
\end{align*}
(250)

\textbf{Lemma I.15.} For four vectors $p, q, v, w \in \mathbb{R}^d$ with unit norm and a bounded function $f : \mathbb{R}^d \to [-B, B]$, we have
\begin{align*}
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle v, u \rangle - \langle q, u \rangle \langle w, u \rangle ] f(u) &\leq \sqrt{3}\sigma^2 B(\|p - q\|_2 + \|v - w\|_2).
\end{align*}
(251)

\textbf{Proof.} First of all, by telescope sum we get
\begin{align*}
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle v, u \rangle - \langle q, u \rangle \langle w, u \rangle ] f(u) &\leq \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle v, u \rangle - \langle p, u \rangle \langle w, u \rangle ] f(u) + \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle w, u \rangle - \langle q, u \rangle \langle w, u \rangle ] f(u).
\end{align*}
(253)

By Hölder inequality we have
\begin{align*}
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle v, u \rangle - \langle p, u \rangle \langle w, u \rangle ] f(u) &\leq \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle^4]^{1/4} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle v - w, u \rangle^4]^{1/4} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[f(u)^2]^{1/2} \\
&\leq \sqrt{3}\sigma^2 \|p - q\|_2 \|v - w\|_2 B.
\end{align*}
(257)

Similarly we have
\begin{align*}
\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[\langle p, u \rangle \langle w, u \rangle - \langle q, u \rangle \langle w, u \rangle ] f(u) &\leq \sqrt{3}\sigma^2 \|p - q\|_2 \|w\|_2 B.
\end{align*}
(258)

I.2. Helper Lemmas on Reinforcement Learning

\textbf{Lemma I.16} (Telescoping or Simulation Lemma, see Luo et al. (2019); Agarwal et al. (2019)). For any policy $\pi$ and deterministic dynamical model $T, \hat{T}$, we have
\begin{align*}
V_T^\pi(s_1) - V_T^\pi(s_1) = \mathbb{E}_{T \sim \rho_T} \left[ \sum_{h=1}^H \left( V_{T}^\pi(\hat{T}(s_h, a_h)) - V_{T}^\pi(T(s_h, a_h)) \right) \right].
\end{align*}
(259)
Lemma I.17 (Policy Gradient Lemma, see Sutton, Barto (2011)). For any policy $\pi_\psi$, deterministic dynamical model $T$ and reward function $r(s_h, a_h)$, we have

$$
\nabla_\psi V_{T}^{\pi_\psi} = \mathbb{E}_{\tau \sim \rho_{T}} \left[ \left( \sum_{h=1}^{H} \nabla_\psi \log \pi_\psi(a_h | s_h) \right) \left( \sum_{h=1}^{H} r(s_h, a_h) \right) \right] 
$$

(260)

**Proof.** Note that

$$
V_{T}^{\pi_\psi} = \int_{\tau} \Pr[\tau] \sum_{h=1}^{H} r(s_h, a_h) \ d\tau.
$$

Take gradient w.r.t. $\psi$ in both sides, we have

\[
\nabla_\psi V_{T}^{\pi_\psi} = \nabla_\psi \int_{\tau} \Pr[\tau] \sum_{h=1}^{H} r(s_h, a_h) \ d\tau \\
= \int_{\tau} (\nabla_\psi \Pr[\tau]) \sum_{h=1}^{H} r(s_h, a_h) \ d\tau \\
= \int_{\tau} \Pr[\tau] (\nabla_\psi \log \Pr[\tau]) \sum_{h=1}^{H} r(s_h, a_h) \ d\tau \\
= \int_{\tau} \Pr[\tau] \left( \sum_{h=1}^{H} \nabla_\psi \log \prod_{h=1}^{H} \pi_\psi(a_h | s_h) \right) \sum_{h=1}^{H} r(s_h, a_h) \ d\tau \\
= \mathbb{E}_{\tau \sim \rho_{T}} \left[ \left( \sum_{h=1}^{H} \nabla_\psi \log \pi_\psi(a_h | s_h) \right) \sum_{h=1}^{H} r(s_h, a_h) \right]
\]

$\square$