Provable Model-based Nonlinear Bandit and Reinforcement Learning: Shelve Optimism, Embrace Virtual Curvature

Kefan Dong [1] **Jiaqi Yang** [2] **Tengyu Ma** [1]

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 sample complexity that only depends 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.

[1]Stanford University [2]Tsinghua University. Correspondence to: Kefan Dong <kefandong@stanford.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

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.[1] (See more evidence below.)

The limited progress on neural net approximation is to some extent not surprising. Given a deterministic dynamics with *known* neural net parameters, finding the best parameterized policy still involves optimizing a complex non-concave function, which is in general computationally intractable. More fundamentally, we find that it is also *statistically intractable* for solving the one-hidden-layer neural net bandit problem (which is a strict sub-case of deep RL). In other words, it requires exponential (in the input dimension) samples to find the global maximum (see Theorems C.1). This also shows that conditions in past work that guarantee global convergence cannot apply to neural nets.

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

[1]This 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 function is given by $\eta(\theta, a)$ for action $a \in \mathcal{A}$ under instance $\theta \in \Theta$. We use sequential Rademacher complexity (Rakhlin et al., 2015a;b) to capture the complexity of the reward function η. Our main result for nonlinear bandit is stated as follows.

Theorem 1.1 (Informal version of Theorem 3.1)**.** *Suppose the sequential Rademacher complexity of a loss function class (defined later) induced by the reward function class* $\{\eta(\theta, \cdot) : \theta \in \Theta\}$ *is bounded by* $\sqrt{R(\Theta)T\mathrm{polylog}(T)}$. *Then, there exists an algorithm (ViOlin, Alg. 1) that finds an* ϵ*-approximate local maximum with* $\widetilde{\mathcal{O}}(R(\Theta)\epsilon^{-8})$ *samples.*

In contrast to zero-order optimization, which does not use the parameterization of η and has a sample complexity depending on the action dimension, our bound only depends on the complexity of the reward function class. This suggests that our algorithm exploits the extrapolation power of the reward function class. To the best of our knowledge, this is the first action-dimension-free result for both linear and nonlinear bandit problems. More concretely, we instantiate our theorem to the following settings and get new results that leverage the model complexity (more in Section A.1).

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

2. Linear bandit with s-sparse or structured instance parameters. Our algorithm ViOlin achieves an $\mathcal{O}(\mathrm{poly}(s, 1/\epsilon))$ 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 $d_{\mathcal{A}}$. Carpentier, Munos (2012) achieve a stronger $\widetilde{\mathcal{O}}(s\sqrt{T})$ regret bound for s-sparse linear bandits with actions set $\mathcal{A} = S^{d-1}$. 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 ϵ-approximate local maximum $\widetilde{\mathcal{O}}(\mathrm{poly}(1/\epsilon))$. Zero-order optimization can also find a local maximum but with $\Omega(d_{\mathcal{A}})$ 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* ℓ_2 *losses for learning the dynamics is bounded by* $\sqrt{R(\Theta)T\mathrm{polylog}(T)}$. *Then, ViOlin for RL (Alg. 2) finds an* ϵ*-approximate local maximum policy with* $\widetilde{\mathcal{O}}(R(\Theta)\epsilon^{-8})$ *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

[2]The 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 key 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, $a \in \mathbb{R}^{d_{\mathcal{A}}}$ the action, and $\eta(\theta, a) \in [0, 1]$ the reward function. Let $\theta^\star$ denote the unknown ground-truth parameter. Throughout the paper, we work under the *realizability assumption* that $\theta^\star \in \Theta$. A bandit algorithm aims to maximize $\eta(\theta^\star, a)$. Let $a^\star = \operatorname{argmax}_a \eta(\theta^\star, a)$ be the optimal action (breaking tie arbitrarily). Let $\|H\|_{\mathrm{sp}}$ 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* $\theta \in \Theta$, $\sup_a \|\nabla_a \eta(\theta, a)\|_2 \le \zeta_g$ *and* $\sup_a \left\|\nabla_a^2 \eta(\theta, a)\right\|_{\mathrm{sp}} \le \zeta_h$. *For every* $\theta \in \Theta$ *and* $a_1, a_2 \in \mathbb{R}^{d_{\mathcal{A}}}$, $\left\|\nabla_a^2 \eta(\theta, a_1) - \nabla_a^2 \eta(\theta, a_2)\right\|_{\mathrm{sp}} \le \zeta_{3\mathrm{rd}} \|a_1 - a_2\|_2$.

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 $\sqrt{d_{\mathcal{A}}}$ steps. The result implies that an action-dimension-free sample complexity bound is impossible with standard sub-Gaussian noise.[3]

Approximate local maxima. In this paper, we aim to find a local maximum of the real reward function $\eta(\theta^\star, \cdot)$. A point x is an (ϵ_g, ϵ_h)-approximate local maximum of a twice-differentiable function $f(x)$ if $\|\nabla f(x)\|_2 \le \epsilon_g$, and $\lambda_{max}(\nabla^2 f(x)) \le \epsilon_h$. 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 (ϵ_g, ϵ_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 $\langle T, r, H, \mu_1 \rangle$. Let $\mathcal{S}$ and $\mathcal{A}$ be the state and action spaces. The dynamics $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ gives next state, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. H and μ_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 $\mathcal{S}_1, \cdots, \mathcal{S}_H$ such that $\mathcal{S} = \cup_{h=1}^H \mathcal{S}_h$, and for any $s_h \in \mathcal{S}, a_h \in a$, $T(s_h, a_h) \in \mathcal{S}_{h+1}$.

We consider parameterized policy and dynamics. Let $\Pi = \{\pi_\psi : \psi \in \Psi\}$ be the policy class and $\{T_\theta : \theta \in \Theta\}$

[3] 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).

the dynamics class. The value function is $V_T^\pi(s_h) \triangleq \mathbb{E}[\sum_{h'=h}^H r(s_{h'}, a_{h'})]$, where $a_h \sim \pi(\cdot \mid s_h), s_{h+1} = T(s_h, a_h)$. Sharing the notation with the bandit setting, let $\eta(\theta, \psi) = \mathbb{E}_{s_1\sim\mu_1} V_{T_\theta}^{\pi_\psi}(s_1)$ be the expected return of policy π_ψ under dynamics T_θ. Let ρ_T^π be the distribution of state action pairs when running policy π in dynamics T. For simplicity, we do not distinguish ψ, θ from π_ψ, T_θ when the context is clear. For example, we write $V_\theta^\psi = V_{T_\theta}^{\pi_\psi}$.

2.3. Preliminary on Online Learning with Stochastic Input Components

Consider a prediction problem where we aim to learn a function that maps from $\mathcal{X}$ to $\mathcal{Y}$ parameterized by parameters in Θ. Let $\ell((x, y); \theta)$ be a loss function that maps $(\mathcal{X} \times \mathcal{Y}) \times \Theta \to \mathbb{R}_+$. 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 $p_t = \mathcal{R}(\{(x_i, y_i)\}_{i=1}^{t-1})$ over the parameter space Θ.
2. The adversary selects a point $\bar{x}_t \in \bar{\mathcal{X}}$ (which may depend on p_t) and generates a sample ξ_t from some fixed distribution q. Let $x_t \triangleq (\bar{x}_t, \xi_t)$, and the adversary picks a label $y_t \in \mathcal{Y}$.
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

$$\mathrm{REG}_T^{\mathrm{OL}} \triangleq \mathop{\mathbb{E}}_{\substack{\xi_t\sim q, \theta_t\sim p_t \\ \forall 1\le t\le T}} \left[\sum_{t=1}^T \ell((x_t, y_t); \theta_t) - \inf_{\theta\in\Theta} \sum_{t=1}^T \ell((x_t, y_t); \theta)\right]. \quad (1)$$

The difference of the formulation from the most standard online learning setup is that the ξ_t part of the input is randomized instead of adversarially chosen (and the learner knows the distribution of ξ_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 $\{x_1, \cdots, x_{t-1}\}$.

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

$$\mathfrak{R}_T(\mathcal{L}; \boldsymbol{x}, \boldsymbol{y}) \triangleq \mathbb{E}_{\xi_1,\cdots,\xi_t}\mathbb{E}_\epsilon \left[\sup_{\ell\in\mathcal{L}} \sum_{t=1}^T \epsilon_t \ell((\boldsymbol{x}(\epsilon), \xi_t), \boldsymbol{y}(\epsilon))\right]. \quad (2)$$

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

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 $C_1 = 2 + \zeta_g/\zeta_h$. Suppose the sequential Rademacher complexity of the loss function (defined in Eq. (5), Section A) is bounded by $\sqrt{R(\Theta)T\mathrm{polylog}(T)}$. The sample complexity of Alg. 1 (for finding an $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-approximate local maximum) is bounded by $\widetilde{\mathcal{O}}(C_1^4 R(\Theta) \max\left(\zeta_h^4\epsilon^{-8}, \zeta_{3\mathrm{rd}}^2\epsilon^{-6}\right))$.*

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 nontrivial results for (sparse) linear bandits and neural net bandits. We sketch these instantiations in Section A.1 and defer the proofs to Section F.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 ϵ 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

Agarwal Alekh, Hsu Daniel, Kale Satyen, Langford John, Li Lihong, Schapire Robert. Taming the monster: A fast and simple algorithm for contextual bandits // International Conference on Machine Learning. 2014. 1638–1646.

Agarwal Alekh, Jiang Nan, Kakade Sham M. Reinforcement learning: Theory and algorithms // CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep. 2019.

Agarwal Alekh, Kakade Sham, Krishnamurthy Akshay, Sun Wen. FLAMBE: Structural complexity and representation learning of low rank MDPs // arXiv preprint arXiv:2006.10814. 2020a.

Agarwal Alekh, Kakade Sham M, Lee Jason D, Mahajan Gaurav. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes // Proceedings of Thirty Third Conference on Learning Theory. 125. 09–12 Jul 2020b. 64–66. (Proceedings of Machine Learning Research).

Amos Brandon, Xu Lei, Kolter J Zico. Input convex neural networks // International Conference on Machine Learning. 2017. 146–155.

Ayoub Alex, Jia Zeyu, Szepesvari Csaba, Wang Mengdi, Yang Lin F. Model-Based Reinforcement Learning with Value-Targeted Regression // Proceedings of the 37th International Conference on Machine Learning. 2020.

Berner Christopher, Brockman Greg, Chan Brooke, Cheung Vicki, Debiak Przemyslaw, Dennison Christy, Farhi David, Fischer Quirin, Hashme Shariq, Hesse Chris, others . Dota 2 with large scale deep reinforcement learning // arXiv preprint arXiv:1912.06680. 2019.

Cai Qi, Yang Zhuoran, Jin Chi, Wang Zhaoran. Provably efficient exploration in policy optimization // International Conference on Machine Learning. 2020. 1283–1294.

Carpentier Alexandra, Munos Rémi. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit // Artificial Intelligence and Statistics. 2012. 190–198.

Clavera Ignasi, Fu Yao, Abbeel Pieter. Model-Augmented Actor-Critic: Backpropagating through Paths // International Conference on Learning Representations. 2019.

Dann Christoph, Jiang Nan, Krishnamurthy Akshay, Agarwal Alekh, Langford John, Schapire Robert E. On Oracle-Efficient PAC-RL with Rich Observations // Advances in Neural Information Processing Systems. 2018.

Dong Kefan, Luo Yuping, Yu Tianhe, Finn Chelsea, Ma Tengyu. On the expressivity of neural networks for deep reinforcement learning // International Conference on Machine Learning. 2020a. 2627–2637.

Dong Kefan, Peng Jian, Wang Yining, Zhou Yuan. Root-n-Regret for Learning in Markov Decision Processes with Function Approximation and Low Bellman Rank // Conference on Learning Theory. 2020b. 1554–1557.

Dong Shi, Ma Tengyu, Van Roy Benjamin. On the Performance of Thompson Sampling on Logistic Bandits // Conference on Learning Theory. 2019a. 1158–1160.

Dong Shi, Van Roy Benjamin. An information-theoretic analysis for thompson sampling with many actions // arXiv preprint arXiv:1805.11845. 2018.

Dong Shi, Van Roy Benjamin, Zhou Zhengyuan. Provably efficient reinforcement learning with aggregated states // arXiv preprint arXiv:1912.06366. 2019b.

Du Simon S, Kakade Sham M, Lee Jason D, Lovett Shachar, Mahajan Gaurav, Sun Wen, Wang Ruosong. Bilinear Classes: A Structural Framework for Provable Generalization in RL // arXiv preprint arXiv:2103.10897. 2021.

Du Simon S, Krishnamurthy Akshay, Jiang Nan, Agarwal Alekh, Dudik Miroslav, Langford John. Provably efficient RL with Rich Observations via Latent State Decoding // International Conference on Machine Learning. 2019a. 1665–1674.

Du Simon S, Luo Yuping, Wang Ruosong, Zhang Hanrui. Provably efficient Q-learning with function approximation via distribution shift error checking oracle // Advances in Neural Information Processing Systems. 2019b. 8058–8068.

Duchi John C, Jordan Michael I, Wainwright Martin J, Wibisono Andre. Optimal rates for zero-order convex optimization: The power of two function evaluations // IEEE Transactions on Information Theory. 2015. 61, 5. 2788–2806.

Dudik Miroslav, Hsu Daniel, Kale Satyen, Karampatziakis Nikos, Langford John, Reyzin Lev, Zhang Tong. Efficient optimal learning for contextual bandits // arXiv preprint arXiv:1106.2369. 2011.

Faury Louis, Abeille Marc, Calauzènes Clément, Fercoq Olivier. Improved optimistic algorithms for logistic bandits // International Conference on Machine Learning. 2020. 3052–3060.

Filippi Sarah, Cappé Olivier, Garivier Aurélien, Szepesvári Csaba. Parametric bandits: the Generalized Linear case // Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1. 2010. 586–594.

Foster Dylan, Rakhlin Alexander. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles // Proceedings of the 37th International Conference on Machine Learning. 119. 13–18 Jul 2020. 3199–3210. (Proceedings of Machine Learning Research).

Foster Dylan J, Rakhlin Alexander, Simchi-Levi David, Xu Yunzong. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective // arXiv preprint arXiv:2010.03104. 2020.

Ge Rong, Huang Furong, Jin Chi, Yuan Yang. Escaping from saddle points—online stochastic gradient for tensor decomposition // Conference on Learning Theory. 2015. 797–842.

Ge Rong, Jin Chi, Zheng Yi. No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis // arXiv preprint arXiv:1704.00708. 2017.

Ge Rong, Lee Jason D, Ma Tengyu. Matrix completion has no spurious local minimum // Advances in Neural Information Processing Systems. 2016. 2973–2981.

Ge Rong, Ma Tengyu. On the optimization landscape of tensor decompositions // Mathematical Programming. 2020. 1–47.

Hafner Danijar, Lillicrap Timothy, Ba Jimmy, Norouzi Mohammad. Dream to Control: Learning Behaviors by Latent Imagination // International Conference on Learning Representations. 2019a.

Hafner Danijar, Lillicrap Timothy, Fischer Ian, Villegas Ruben, Ha David, Lee Honglak, Davidson James. Learning latent dynamics for planning from pixels // International Conference on Machine Learning. 2019b. 2555–2565.

Hao Botao, Lattimore Tor, Szepesvári Csaba, Wang Mengdi. Online Sparse Reinforcement Learning // International Conference on Artificial Intelligence and Statistics. 2021. 316–324.

Hao Botao, Lattimore Tor, Wang Mengdi. High-Dimensional Sparse Linear Bandits // arXiv preprint arXiv:2011.04020. 2020.

Hazan Elad, Levy Kfir, Shalev-Shwartz Shai. Beyond convexity: Stochastic quasi-convex optimization // Advances in Neural Information Processing Systems. 2015. 1594–1602.

Hazan Elad, Singh Karan, Zhang Cyril. Efficient regret minimization in non-convex games // International Conference on Machine Learning. 2017. 1433–1441.

Hsu Daniel, Kakade Sham, Zhang Tong, others . A tail inequality for quadratic forms of subgaussian random vectors // Electronic Communications in Probability. 2012. 17.

Janner Michael, Fu Justin, Zhang Marvin, Levine Sergey. When to trust your model: Model-based policy optimization // arXiv preprint arXiv:1906.08253. 2019.

Jiang Nan, Krishnamurthy Akshay, Agarwal Alekh, Langford John, Schapire Robert E. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable // arXiv preprint arXiv:1610.09512. 2016.

Jin Chi, Allen-Zhu Zeyuan, Bubeck Sebastien, Jordan Michael I. Is Q-learning provably efficient? // Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018. 4868–4878.

Jin Chi, Yang Zhuoran, Wang Zhaoran, Jordan Michael I. Provably efficient reinforcement learning with linear function approximation // Conference on Learning Theory. 2020. 2137–2143.

Johnson Rie, Zhang Tong. Accelerating stochastic gradient descent using predictive variance reduction // Advances in Neural Information Processing Systems. 2013. 315–323.

Kakade Sham, Krishnamurthy Akshay, Lowrey Kendall, Ohnishi Motoya, Sun Wen. Information theoretic regret bounds for online nonlinear control // arXiv preprint arXiv:2006.12466. 2020.

Kakade Sham M, Kanade Varun, Shamir Ohad, Kalai Adam. Efficient learning of generalized linear and single index models with isotonic regression // Advances in Neural Information Processing Systems. 2011. 927–935.

Krishnamurthy Akshay, Agarwal Alekh, Langford John. PAC reinforcement learning with rich observations // Advances in Neural Information Processing Systems. 2016. 1840–1848.

Lattimore Tor, Szepesvári Csaba. Bandit algorithms. 2020.

Laurent Beatrice, Massart Pascal. Adaptive estimation of a quadratic functional by model selection // Annals of Statistics. 2000. 1302–1338.

Lee Jason D, Simchowitz Max, Jordan Michael I, Recht Benjamin. Gradient descent only converges to minimizers // Conference on learning theory. 2016. 1246–1257.

Levine Sergey, Finn Chelsea, Darrell Trevor, Abbeel Pieter. End-to-end training of deep visuomotor policies // The Journal of Machine Learning Research. 2016. 17, 1. 1334–1373.

Li Gene, Kamath Pritish, Foster Dylan J., Srebro Nathan. Eluder Dimension and Generalized Rank. 2021.

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

Li Lihong, Lu Yu, Zhou Dengyong. Provably optimal algorithms for generalized linear contextual bandits // International Conference on Machine Learning. 2017. 2071–2080.

Littman Michael L, Sutton Richard S, Singh Satinder P. Predictive representations of state. // NIPS. 14, 1555. 2001. 30.

Liu Sijia, Chen Pin-Yu, Kailkhura Bhavya, Zhang Gaoyuan, Hero III Alfred O, Varshney Pramod K. A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning: Principals, Recent Advances, and Applications // IEEE Signal Processing Magazine. 2020. 37, 5. 43–54.

Luo Yuping, Xu Huazhe, Li Yuanzhi, Tian Yuandong, Darrell Trevor, Ma Tengyu. Algorithmic Framework for Model-based Deep Reinforcement Learning with Theoretical Guarantees // International Conference on Learning Representations. 2019.

Mahmud Mufti, Kaiser Mohammed Shamim, Hussain Amir, Vassanelli Stefano. Applications of deep learning and reinforcement learning to biological data // IEEE transactions on neural networks and learning systems. 2018. 29, 6. 2063–2079.

Misra Dipendra, Henaff Mikael, Krishnamurthy Akshay, Langford John. Kinematic state abstraction and provably efficient rich-observation reinforcement learning // International conference on machine learning. 2020. 6961–6971.

Modi Aditya, Jiang Nan, Tewari Ambuj, Singh Satinder. Sample complexity of reinforcement learning using linearly combined model ensembles // International Conference on Artificial Intelligence and Statistics. 2020. 2010–2020.

Munos Rémi. Error bounds for approximate value iteration // Proceedings of the National Conference on Artificial Intelligence. 2005. 1006.

Osband Ian, Roy Benjamin Van. Model-based reinforcement learning and the eluder dimension // Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1. 2014. 1466–1474.

Rakhlin Alexander, Sridharan Karthik, Tewari Ambuj. Online learning: stochastic, constrained, and smoothed adversaries // Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011. 1764–1772.

Rakhlin Alexander, Sridharan Karthik, Tewari Ambuj. Online Learning via Sequential Complexities // Journal of Machine Learning Research. 2015a. 16, 6. 155–186.

Rakhlin Alexander, Sridharan Karthik, Tewari Ambuj. Sequential complexities and uniform martingale laws of large numbers // Probability Theory and Related Fields. 2015b. 161, 1-2. 111–153.

Russo Daniel, Roy Benjamin Van. Eluder dimension and the sample complexity of optimistic exploration // Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. 2013. 2256–2264.

Shariff Roshan, Szepesvári Csaba. Efficient planning in large MDPs with weak linear function approximation // arXiv preprint arXiv:2007.06184. 2020.

Silver David, Schrittwieser Julian, Simonyan Karen, Antonoglou Ioannis, Huang Aja, Guez Arthur, Hubert Thomas, Baker Lucas, Lai Matthew, Bolton Adrian, others . Mastering the game of Go without human knowledge // Nature. 2017. 550, 7676. 354.

Simchi-Levi David, Xu Yunzong. Bypassing the Monster: A Faster and Simpler Optimal Algorithm for Contextual Bandits under Realizability // arXiv preprint arXiv:2003.12699. 2020.

Sun Wen, Jiang Nan, Krishnamurthy Akshay, Agarwal Alekh, Langford John. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches // Conference on Learning Theory. 2019. 2898–2933.

Reinforcement learning: An introduction. // . 2011.

Valko Michal, Korda Nathan, Munos Rémi, Flaounas Ilias, Cristianini Nello. Finite-time analysis of kernelised contextual bandits // Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. 2013. 654–663.

Wang Ruosong, Salakhutdinov Ruslan, Yang Lin F. Provably Efficient Reinforcement Learning with General Value Function Approximation // Advances in Neural Information Processing Systems. 2020a.

Wang Yining, Chen Yi, Fang Ethan X, Wang Zhaoran, Li Runze. Nearly Dimension-Independent Sparse Linear Bandit over Small Action Spaces via Best Subset Selection // arXiv preprint arXiv:2009.02003. 2020b.

Wang Yining, Wang Ruosong, Du Simon S, Krishnamurthy Akshay. Optimism in Reinforcement Learning with Generalized Linear Function Approximation // arXiv preprint arXiv:1912.04136. 2019.

Williams Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine learning. 1992. 8, 3-4. 229–256.

Xu Bing, Wang Naiyan, Chen Tianqi, Li Mu. Empirical Evaluation of Rectified Activations in Convolutional Network. 2015.

Yang Lin, Wang Mengdi. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound // International Conference on Machine Learning. 2020. 10746–10756.

Yang Zhuoran, Jin Chi, Wang Zhaoran, Wang Mengdi, Jordan Michael I. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations // arXiv preprint arXiv:2011.04622. 2020.

Yu Tianhe, Thomas Garrett, Yu Lantao, Ermon Stefano, Zou James, Levine Sergey, Finn Chelsea, Ma Tengyu. MOPO: Model-based Offline Policy Optimization // arXiv preprint arXiv:2005.13239. 2020.

Zanette Andrea, Lazaric Alessandro, Kochenderfer Mykel, Brunskill Emma. Learning Near Optimal Policies with Low Inherent Bellman Error // International Conference on Machine Learning. 2020.

Zhou Dongruo, Li Lihong, Gu Quanquan. Neural contextual bandits with UCB-based exploration // International Conference on Machine Learning. 2020. 11492–11502.

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 $\eta(\theta^\star, a)$ the *real reward* of action a, and $\eta(\hat{\theta}_t, a)$ the *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 $\eta(\theta^\star, a_t)$ in the long run.

One can expect that a proper combination of property (i) and (ii) leads to showing the real reward $\eta(\theta^\star, a_t)$ 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 $\eta(\theta^\star, a^\star)$. 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 $\eta(\theta^\star, a_t)$. 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: **Vi**rtual Ascent with **On**lin**e** Model Learner (for Bandit)

1: Set parameter $\kappa_1 = 2\zeta_g$ and $\kappa_2 = 640\sqrt{2}\zeta_h$. 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 ℓ (defined in equation (5)) and obtain $p_t = \mathcal{R}(\mathcal{H}_{t-1})$.
4: Let $a_t \leftarrow \operatorname{argmax}_a \mathbb{E}_{\theta_t \sim p_t}[\eta(\theta_t, a)]$.
5: Sample $u_t, v_t \sim \mathcal{N}(0, I_{d_\mathcal{A} \times d_\mathcal{A}})$ 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 $y_t = [\eta(\theta^\star, a_t), \eta(\theta^\star, a_{t-1}), \langle \nabla_a \eta(\theta^\star, a_{t-1}), u_t\rangle, \langle \nabla_a^2 \eta(\theta^\star, a_{t-1}) u_t, v_t\rangle] \in \mathbb{R}^4$ 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 $\nabla_a \eta(\theta^\star, a)$

[4]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 $a^\star$ (because online learner never sees $a^\star$). 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 $\nabla_a^2\eta(\theta^\star, a)$ accurately. In other words, we augment the loss function for the online learner so that the estimated model satisfies $\eta(\hat{\theta}_t, a_t) \approx \eta(\theta^\star, a_t)$, $\nabla_a\eta(\hat{\theta}_t, a_t) \approx \nabla_a\eta(\theta^\star, a_t)$, and $\nabla_a^2\eta(\hat{\theta}_t, a_t) \approx \nabla_a^2\eta(\theta^\star, a_t)$ in the long run. This implies that when a_t is not at a local maximum of the real reward function $\eta(\theta^\star, \cdot)$, 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 $\nabla_a\eta(\theta^\star, a_t)$ correctly, we need a supervision for it. However, we only observe the reward $\eta(\theta^\star, a_t)$. 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:

$$\langle \nabla_a\eta(\theta^\star, a), u\rangle = \lim_{\alpha_1\to 0} \left(\eta(\theta^\star, a + \alpha_1 u) - \eta(\theta^\star, a)\right)/\alpha_1 \tag{3}$$

It turns out that the number of random projections $\langle \nabla_a\eta(\theta^\star, a), u\rangle$ 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 $u, v \in d_{\mathcal{A}}$ by:

$$\begin{aligned}\left\langle \nabla_a^2\eta(\theta^\star, a)v, u\right\rangle &= \lim_{\alpha_2\to 0} \left(\langle \nabla_a\eta(\theta^\star, a + \alpha_2 v), u\rangle - \langle \nabla_a\eta(\theta^\star, a), u\rangle\right)/\alpha_2 \\ &= \lim_{\alpha_2\to 0}\lim_{\alpha_1\to 0} \left((\eta(\theta^\star, a + \alpha_1 u + \alpha_2 v) - \eta(\theta^\star, a + \alpha_2 v)) - (\eta(\theta^\star, a + \alpha_1 u) - \eta(\theta^\star, a))\right)/(\alpha_1\alpha_2)\end{aligned} \tag{4}$$

Algorithmically, we can choose infinitesimal α_1 and α_2. Note that α_1 should be at least an order smaller than α_2 because the limitations are taken sequentially.

We create the following prediction task for an online learner: let θ be the parameter, $x = (a, a', u, v)$ be the input, $\hat{y} = [\eta(\theta, a), \eta(\theta, a'), \langle \nabla_a\eta(\theta, a'), u\rangle, \langle \nabla_a^2\eta(\theta, a')u, v\rangle] \in \mathbb{R}^4$ be the output, and $y = [\eta(\theta^\star, a), \eta(\theta^\star, a'), \langle \nabla_a\eta(\theta^\star, a'), u\rangle, \langle \nabla_a^2\eta(\theta^\star, a')u, v\rangle] \in \mathbb{R}^4$ be the supervision, and the loss function be

$$\begin{aligned}\ell(((a, a', u, v), y); \theta) \triangleq &([\hat{y}]_1 - [y]_1)^2 + ([\hat{y}]_2 - [y]_2)^2 + \min\left(\kappa_1^2, ([\hat{y}]_3 - [y]_3)^2\right) \\ &+ \min\left(\kappa_2^2, ([\hat{y}]_4 - [y]_4)^2\right)\end{aligned} \tag{5}$$

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 $C_1 = 2 + \zeta_g/\zeta_h$. Suppose the sequential Rademacher complexity of the loss function (defined in Eq. (5), Section A) is bounded by $\sqrt{R(\Theta)T\mathrm{polylog}(T)}$. The sample complexity of Alg. 1 (for finding an $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-approximate local maximum) is bounded by $\widetilde{\mathcal{O}}(C_1^4 R(\Theta)\max\left(\zeta_h^4\epsilon^{-8}, \zeta_{3\mathrm{rd}}^2\epsilon^{-6}\right))$.*

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 $\tilde{\eta}(\theta, a) \triangleq \langle \theta, a\rangle - \frac{1}{2}\|a\|_2^2$ and apply Theorem F.3 with reward $\tilde{\eta}$. We claim that the sample complexity (of finding an ϵ-suboptimal action for $\eta(\theta^\star, a)$) is bounded by $\widetilde{\mathcal{O}}(\log|\Theta|\epsilon^{-8})$.[5] 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 $\mathrm{poly}(d)$ 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

[5] Recall that in bandit literature, action a is an ϵ-approximate optimal action if $\eta(\theta^\star, a) \geq \eta(\theta^\star, a^\star) - \epsilon$.

hypothesis case, we claim that the sample complexity of Alg. 1 is $\widetilde{\mathcal{O}}(\log|\Theta|\epsilon^{-8}\mathrm{polylog}(d))$. 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 θ has an additional structure. Suppose $\Theta = \{\theta = \phi(z) : z \in \mathbb{R}^s\}$ for some Lipschitz function ϕ. Then, a simlar 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 $\eta((W_1, W_2), a) = W_2\sigma(W_1 a) - \frac{1}{2}\|a\|_2^2$ for some nonlinear link 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 $\max_{i\in[m]} \sum_{j=1}^d |[W_1]_{i,j}|$. That is, the max 1-norm of the rows of W_1. Let the model hypothesis space be $\Theta = \{(W_1, W_2) : \|W_1^\top\|_{1,\infty} \le 1, \|W_2\|_1 \le 1\}$ and $\theta \triangleq (W_1, W_2)$. We claim that Alg .1 finds an $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-approximate local maximum in $\widetilde{\mathcal{O}}\big(\epsilon^{-8}\mathrm{polylog}(d)\big)$ 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 η, $\langle\nabla_a\eta, u\rangle$, and $\langle u, \nabla_a^2\eta \cdot v\rangle$, and finally the resulting loss function ℓ. See Theorem F.6 in Section F.7 for details. We remark here that zero-order optimization in this case has a $\mathrm{poly}(d)$ sample complexity. In addition, if the second layer of the neural network W_2 contains all negative entries, and the activation function σ 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 ϵ-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 ψ corresponds to the action a in bandit, and the dynamics parameter θ corresponds to the model parameter θ in bandit. The expected total return $\eta(\theta, \psi) = \mathbb{E}_{s_1\sim\mu_1} V_{T_\theta}^{\pi_\psi}(s_1)$ is the analogue of reward function in bandit. We intend to make the same regularity assumptions on η as in the bandit case (that is, Assumption 2.1) with a being replaced by ψ. However, when the policy is deterministic, the reward function η has Lipschitz constant with respect to ψ 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 η and its derivatives to be Lipschitz with respect to ψ.

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 the dynamics, we make the following additional Lipschitzness of value functions w.r.t to the states and Lipschitzness of policies w.r.t to 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:* $\forall\psi \in \Psi, \theta \in \Theta, s, s' \in \mathcal{S}$ *we have* $|V_\theta^\psi(s) - V_\theta^\psi(s')| \le L_0\|s - s'\|_2$; $\|\nabla_\psi V_\theta^\psi(s) - \nabla_\psi V_\theta^\psi(s')\|_2 \le L_1\|s - s'\|_2$; $\|\nabla_\psi^2 V_\theta^\psi(s) - \nabla_\psi^2 V_\theta^\psi(s')\|_{\mathrm{sp}} \le L_2\|s - s'\|_2$.

Assumption B.2. *We assume the following Lipschitzness assumptions on the stochastic policies parameterization* π_ψ.[6] $\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi \log\pi_\psi(a \mid s))(\nabla_\psi \log\pi_\psi(a \mid s))^\top]\|_{\mathrm{sp}} \le \chi_g$; $\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi \log\pi_\psi(a \mid s))^{\otimes 4}]\|_{\mathrm{sp}} \le \chi_f$; $\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi^2 \log\pi_\psi(a \mid s))(\nabla_\psi^2 \log\pi_\psi(a \mid s))^\top]\|_{\mathrm{sp}} \le \chi_h$.

Our results depends polynomially on the parameters $L_0, L_1, L_2, \chi_g, \chi_f$ and χ_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* $\mathcal{S}$ *be the unit ball in* $\mathbb{R}^d$ *and action space* $\mathcal{A}$ *be* $\mathbb{R}^d$. *The (deterministic) dynamics* T *is given*

[6] Recall that the injective norm of a k-th order tensor $A \in \mathbb{R}^{d^{\otimes k}}$ is defined as $\|A^{\otimes k}\|_{\mathrm{sp}} = \sup_{u\in S^{d-1}} \langle A, u^{\otimes k}\rangle$.

by $T(s,a) = \mathrm{N}_\theta(s+a)$, where N is a nonlinear model parameterized by θ, e.g., a neural network. We assume that θ belongs to a finite hypothesis class Θ that satisfies $\|\mathrm{N}_\theta(s+a)\|_2 \le 1$ for all $\theta \in \Theta, s \in \mathcal{S}, a \in \mathcal{A}$. Assume that the reward function $r(s,a)$ is L_r-Lipschitz w.r.t ℓ_2-norm, that is, satisfying $|r(s_1,a_1) - r(s_2,a_2)| \le L_r(\|s_1 - s_2\|_2 + \|a_1 - a_2\|_2)$. We consider a family of stochastic Gaussian policies with the mean being linear in the state: $\pi_\psi(s) = \mathcal{N}(\psi s, \sigma^2 I)$, parameterized by $\psi \in \mathbb{R}^{d\times d}$ with $\|\psi\|_{\mathrm{op}} \le 1$. 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 $\zeta_g, \zeta_h, \zeta_{3\mathrm{rd}}, L_0, L_1, L_2, \chi_g, \chi_f$ and χ_h bounded by $\mathrm{poly}(\sigma, 1/\sigma, H, L_r)$.

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 $\tau_t = (s_1, a_1, \cdots, s_H, a_H)$ be a trajectory sampled from policy π_{ψ_t} under the real dynamics $T_{\theta^\star}$. Similarly to (Yu et al., 2020), when the value function is Lipschitz, we upper bound $\Delta_{t,1} = |\eta(\theta_t, \psi_t) - \eta(\theta^\star, \psi_t)|$ 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

$$\ell((\tau_t, \tau_t'); \theta) = \sum_{(s_h,a_h)\in\tau_t} \|T_\theta(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2 + \sum_{(s_h',a_h')\in\tau_t'} \|T_\theta(s_h',a_h') - T_{\theta^\star}(s_h',a_h')\|_2^2 \tag{6}$$

for two trajectories τ, τ' sampled from policy π_{ψ_t} and $\pi_{\psi_{t-1}}$ 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. 1 by using the loss function in Eq. (6). Our algorithm is presented in Alg. 2 in Appendix G. Main theorem for Alg. 2 is shown below.

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

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 $\widetilde{\mathcal{O}}\big(\mathrm{poly}(\sigma, 1/\sigma, H, L_r)\log|\Theta|\epsilon^{-8}\big)$.

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 $\sqrt{d}/\sigma$. In contrast, our sample complexity is independent of the dimension d. Instead, our bound depends on the complexity of the model family Θ 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 $\Omega(\varepsilon^{-(d-2)})$.*

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

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

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: $\langle \theta^\star, a\rangle$ 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.

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; Zanette 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 $\mathcal{I} = \{I_{\theta,\varepsilon} : \|\theta\|_2 \le 1, \varepsilon > 0\}$ 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

$$\eta(I_{\theta,\varepsilon}, x) = Ax\{\langle x, \theta\rangle - 1 + \varepsilon, 0\}. \tag{7}$$

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

$$\inf_{\mathcal{A}} \sup_{I\in\mathcal{I}} \mathbb{E}[R^T_{\mathcal{A},I}] \ge \Omega(T^{\frac{d-2}{d-1}}).$$

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

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.,

2017). In this case, our Theorem F.3 can also give polynomial global regret result: because all local maxima of the reward function is global maximum (Hazan et al., 2015; Kakade et al., 2011) and it also satisfies the strict-saddle property (Ge et al., 2015), all approximate local maxima are global. This shows that our framework does separate the intractable cases from the tractable by the notions of local and global regrets.

With two-layer neural networks, we can relax the use of ReLU activation—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 $\dim_E(\mathcal{F}, \varepsilon)$ to denote the ε-eluder dimension of the function class $\mathcal{F}$. Let Θ be an ε-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 $\dim_E(\mathcal{F}, \varepsilon) \geq \Omega(\varepsilon^{-(d-1)})$. For each $i \in [n]$, we define the function $f_i(\boldsymbol{a}) = \eta(I_{\theta_i,\varepsilon}, \boldsymbol{a}) \in \mathcal{F}$. Then for $i \leq n-1$, we have $f_i(\theta_j) = f_{i+1}(\theta_j)$ for $j \leq i-1$, while $\varepsilon = f_i(\theta_i) \neq f_{i+1}(\theta_i) = 0$. Therefore, θ_i is $\frac{\varepsilon}{2}$-independent of its predecessors. As a result, we have $\dim_E(\mathcal{F}, \varepsilon) \geq n-1$. □

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 $\eta(\theta^\star, a) \leq C_t(a)$. And then the action for time step t is $a_t \leftarrow \arg\max C_t(a)$. Let Θ_t be the set of parameters that is consistent with $\eta(\theta^\star, a_1), \cdots, \eta(\theta^\star, a_{t-1})$. That is, $\Theta_t = \{\theta \in \Theta : \eta(\theta, a_\tau) = \eta(\theta^\star, a_\tau), \forall \tau < t\}$. In a deterministic environment, the tightest upper confidence bound is $C_t(a) = \sup_{\theta \in \Theta_t} \eta(\theta, a)$.

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

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

Note that the reward function η can be clearly realized by a two-layer neural network with width $2d$. When $\alpha = 0$ we have $\eta((\theta_1^\star, \theta_2^\star, \alpha), a) = \frac{1}{64} \langle \theta_1^\star, a \rangle$, 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 $\langle \theta_2^\star, a_t \rangle = 0$, and can only eliminate an exponentially small fraction of $\theta_2^\star$ 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

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

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 η can be clearly realized by a two-layer neural network with width d. Note that when $\alpha = 0$ we have $\eta((\theta_1^\star, \theta_2^\star, \alpha), a) = \frac{1}{64} \langle \theta_1^\star, a \rangle$, which represents a linear reward. In the following we use $\boldsymbol{\theta}^\star = (\theta_1^\star, \theta_2^\star, 0)$ 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 $\eta(\boldsymbol{\theta}^\star, a) \leq C_t(a)$. And then the action for time step t is $a_t \leftarrow \operatorname{argmax} C_t(a)$.

Let $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$ be an $\frac{1}{2}$-packing of the sphere S^{d-1}, where $n = \Omega(2^d)$. Let $\mathcal{B}(p_i, \frac{1}{4})$ be the ball with radius $1/4$ centered at p_i, and $B_i = \mathcal{B}(p_i, \frac{1}{4}) \cup S^{d-1}$. 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 $\sup_{a_j \in B_j} \langle p_i, a_j \rangle \leq 31/32$ 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 $\boldsymbol{\theta}^\star = (p_i, p_i, 0)$.

Let Θ_t be the set of parameters that is consistent with $\eta(\boldsymbol{\theta}^\star, a_1), \cdots, \eta(\boldsymbol{\theta}^\star, a_{t-1})$. That is, $\Theta_t = \{\boldsymbol{\theta} \in \Theta : \eta(\boldsymbol{\theta}, a_\tau) = \eta(\boldsymbol{\theta}^\star, a_\tau), \forall \tau < t\}$. Since our environment is deterministic, a tightest upper confidence bound is $C_t(a) = \sup_{\boldsymbol{\theta} \in \Theta_t} \eta(\boldsymbol{\theta}, a)$.

Let $A_t = \{a_1, \cdots, a_t\}$. It can be verified that for any $\theta_2 \in S^{d-1}$, $\eta((\theta_1^\star, \theta_2, 1), \cdot)$ is consistent with $\eta(\boldsymbol{\theta}^\star, \cdot)$ on A_{t-1} if $\mathcal{B}(\theta_2, \frac{1}{4}) \cup A_{t-1} = \emptyset$. As a result, for any θ_2 such that $\mathcal{B}(\theta_2, \frac{1}{4}) \cup A_{t-1} = \emptyset$ we have

$$C_t(\theta_2) \geq \frac{1}{32} > \frac{1}{128} + \sup_a \eta(\boldsymbol{\theta}^\star, a). \tag{8}$$

Next we prove that for any $i \in [n]$, there exists t such that $a_t \in \mathcal{B}(p_i, \frac{1}{4})$. Note that $\eta(\boldsymbol{\theta}, \cdot)$ is $\frac{65}{64}$ Lipschitz for every $\boldsymbol{\theta} \in \Theta$. As a result, $C_t(a_\tau + \xi) \leq C_t(a_\tau) + \frac{65}{64} \|\xi\|_2 = \eta(\boldsymbol{\theta}^\star, a_\tau) + \frac{65}{64} \|\xi\|_2$ for all $\tau < t$. Consequently,

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

for any $\tau < t$ and ξ 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 actions in $\mathcal{B}(\theta_2, \frac{1}{4})$ is never taken by the algorithm. By Eq. (9) we have $C_t(\theta_2) \geq \frac{1}{32}$ for all t. Let $\mathcal{H}_t = \cup_{\tau=1}^{t-1} \mathcal{B}(a_\tau, \frac{1}{130})$. By Eq. (9) we have $C_t(a) \leq \frac{3}{128}$ for all $a \in \mathcal{H}_t$. Because $a_t \leftarrow \operatorname{argmax}_a C_t(a)$ and $\max_{a \in \mathcal{H}_t} C_t(a) < C_t(\theta_2)$, we conclude that $a_t \notin \mathcal{H}_t$. 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 θ_2 there exists $t \leq 130^d$ such that $a_t \in \mathcal{B}(\theta_2, \frac{1}{4})$. □

E.4. Hardness of stochastic environments

As a motivation to consider deterministic rewards, the next theorem proves that a $\mathrm{poly}(\log|\Theta|)$ 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* $\log|\Theta| = \widetilde{\mathcal{O}}(1)$*, 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 $\Theta = \{e_1, \cdots, e_d\}$. The action space is S^{d-1}. The stochastic reward function is given by $\eta(\theta, a) = \langle \theta, a \rangle + \xi$ where $\xi = \mathcal{N}(0, 1)$ 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

$$\operatorname{grad} \eta(\theta, a) = \left(I - aa^\top\right)\theta.$$

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

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

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

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

$$\mathbb{E}_i[E_{i,T}] \leq \mathbb{E}_0[E_{i,T}] + \sqrt{\frac{1}{4} \mathbb{E}_0 \left[\sum_{t=1}^T \langle a_t, \theta_i \rangle^2\right]}. \tag{11}$$

Consequently,

$$\sum_{i=1}^{d} \mathbb{E}_i[E_{i,T}] \leq \sum_{i=1}^{d} \mathbb{E}_0[E_{i,T}] + \sum_{i=1}^{d} \sqrt{\frac{1}{4}\mathbb{E}_0\left[\sum_{t=1}^{T} \langle a_t, \theta_i \rangle^2\right]} \tag{12}$$

$$\leq 1 + \sqrt{\frac{d}{4}\mathbb{E}_0\left[\sum_{i=1}^{d}\sum_{t=1}^{T} \langle a_t, \theta_i \rangle^2\right]} \leq 1 + \sqrt{\frac{dT}{4}}, \tag{13}$$

which means that

$$\min_{i\in[d]} \mathbb{E}_i[E_{i,T}] \leq \frac{1}{d} + \sqrt{\frac{T}{4d}}. \tag{14}$$

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

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 θ_t can estimate the reward, its gradient and hessian accurately, that is, for $\theta_t \sim p_t$, $\eta(\theta^\star, a_t) \approx \eta(\theta_t, a_t)$, $\nabla_a \eta(\theta^\star, a_{t-1}) \approx \nabla_a \eta(\theta_t, a_{t-1})$, and $\nabla_a^2 \eta(\theta^\star, a_{t-1}) \approx \nabla_a^2 \eta(\theta_t, a_{t-1})$.
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 $\eta(\theta^\star, a)$ iteratively (in terms of finding second-order local improvement direction.)

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

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

$$\eta(\theta^\star, a_t) \gtrapprox_{\Delta_t} \mathbb{E}_{\theta_t}[\eta(\theta_t, a_t)] \tag{15}$$

$$\geq \sup_a \mathbb{E}_{\theta_t}\left[\eta(\theta_t, a_{t-1}) + \langle a - a_{t-1}, \nabla_a \eta(\theta_t, a_{t-1})\rangle - \frac{\zeta_h}{2}\|a - a_{t-1}\|_2^2\right] \tag{16}$$

$$\gtrapprox_{\Delta_t} \sup_a \mathbb{E}_{\theta_t}\left[\eta(\theta^\star, a_{t-1}) + \langle a - a_{t-1}, \nabla_a \eta(\theta^\star, a_{t-1})\rangle - \frac{\zeta_h}{2}\|a - a_{t-1}\|_2^2\right] \tag{17}$$

$$\geq \eta(\theta^\star, a_{t-1}) + \frac{1}{2\zeta_h}\|\nabla_a \eta(\theta^\star, a_{t-1})\|_2^2. \tag{18}$$

Here in equations (15) and (17), we use the symbol $\gtrapprox_{\Delta_t}$ to present *informal* inequalities that are true up to some additive errors that depend on Δ_t. This is because equation (15) holds up to errors related to $\Delta_{t,1} = |\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)|$, and equation (17) holds up to errors related to $\Delta_{t,2} = |\eta(\theta_t, a_{t-1}) - \eta(\theta^\star, a_{t-1})|$ and $\Delta_{t,3} = \|\nabla_a \eta(\theta_t, a_{t-1}) - \nabla_a \eta(\theta^\star, a_{t-1})\|_2$. Eq. (16) is a second-order Taylor expansion around the previous iteration a_{t-1} and utilizes the definition $a_t = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{\theta_t}\eta(\theta_t, a)$. 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 θ_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 $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-approximate second order stationary point, we have $\eta(\theta^\star, a_t) \geq \eta(\theta^\star, a_{t-1}) + \min\left(\zeta_h^{-1}\epsilon^2/4, \zeta_{3\mathrm{rd}}^{-1/2}\epsilon^{3/2}\right) - C_1\mathbb{E}_{\theta_t\sim p_t}[\Delta_t]$.*

Next, we show part (i) by linking the error Δ_t to the loss function ℓ (Eq. (5)) used by the online learner. The errors $\Delta_{t,1}, \Delta_{t,2}$ are already part of the loss function. Let $\tilde{\Delta}_{t,3} = \langle \nabla_a\eta(\theta_t, a_{t-1}) - \nabla_a\eta(\theta^\star, a_{t-1}), u_t\rangle$ and $\tilde{\Delta}_{t,4} = \left\langle \nabla_a^2\eta(\theta_t, a_{t-1}) - \nabla_a^2\eta(\theta^\star, a_{t-1})u_t, v_t\right\rangle$ be the remaining two terms (without the clipping) in the loss (Eq. (5)). Note that $\tilde{\Delta}_{t,3}$ is supposed to bound $\Delta_{t,3}$ because $\mathbb{E}_{u_t}[\tilde{\Delta}_{t,3}^2] = \Delta_{t,3}^2$. Similarly, $\mathbb{E}_{u_t,v_t}[\tilde{\Delta}_{t,4}^2] = \|\nabla_a^2\eta(\theta_t, a_{t-1}) - \nabla_a^2\eta(\theta^\star, a_{t-1})\|_{\mathrm{F}}^2 \geq \Delta_{t,4}^2$. We clip $\tilde{\Delta}_{t,3}$ and $\tilde{\Delta}_{t,4}$ 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 $\tilde{\Delta}_t^2 = \Delta_{t,1}^2 + \Delta_{t,2}^2 + \min\left(\kappa_1^2, \tilde{\Delta}_{t,3}^2\right) + \min\left(\kappa_2^2, \tilde{\Delta}_{t,4}^2\right)$ be the error received by the online learner at time t. The argument above can be rigorously formalized into a lemma that upper bound Δ_t by $\tilde{\Delta}_t$, which will be bounded by the sequential Rademacher complexity.

Lemma F.2. *In the setting of Theorem F.3, we have $\mathbb{E}_{u_{1:T}, v_{1:T}, \theta_{1:T}}[\sum_{t=1}^T \tilde{\Delta}_t^2] \geq \frac{1}{2}\mathbb{E}_{\theta_{1:T}}[\sum_{t=1}^T \Delta_t^2]$.*

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 $\eta(\theta^\star, a_t)$. 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 $\eta(\theta^\star, a_t)$ in the following two cases:

1. $\|\nabla_a\eta(\theta^\star, a_{t-1})\|_2 \geq \epsilon$, or
2. $\|\nabla_a\eta(\theta^\star, a_{t-1})\|_2 \leq \epsilon$ and $\lambda_{\max}\left(\nabla_a^2\eta(\theta^\star, a_{t-1})\right) \geq 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}$.

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

$$\bar{\eta}_t(\theta, a) = \eta(\theta, a_{t-1}) + \langle a - a_{t-1}, \nabla_a\eta(\theta, a_{t-1})\rangle - \zeta_h\|a - a_{t-1}\|_2^2 \tag{19}$$

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

$$\bar{\eta}_t(\theta_t, a) \geq \bar{\eta}_t(\theta^\star, a) - \Delta_{t,2} - \|a - a_{t-1}\|_2\Delta_{t,3}. \tag{20}$$

In this case we have

$$\begin{aligned}
&\eta(\theta^\star, a_t) \geq \mathbb{E}_{\theta_t\sim p_t}[\eta(\theta_t, a_t) - \Delta_{t,1}] \\
&\geq \sup_a \mathbb{E}_{\theta_t\sim p_t}[\eta(\theta_t, a) - \Delta_{t,1}] && \text{(By the optimality of } a_t\text{)} \\
&\geq \sup_a \mathbb{E}_{\theta_t\sim p_t}[\bar{\eta}_t(\theta_t, a) - \Delta_{t,1}] \\
&\geq \sup_a \mathbb{E}_{\theta_t\sim p_t}[\bar{\eta}_t(\theta^\star, a) - \Delta_{t,1} - \Delta_{t,2} - \|a - a_{t-1}\|_2\Delta_{t,3}] && \text{(By Eq. (20))} \\
&\geq \mathbb{E}_{\theta_t\sim p_t}\left[\eta(\theta^\star, a_{t-1}) + \frac{1}{4\zeta_h}\|g_t\|_2^2 - \Delta_{t,1} - \Delta_{t,2} - \frac{\|g_t\|_2}{2\zeta_h}\Delta_{t,3}\right] && \text{(Take } a = a_{t-1} + \tfrac{g_t}{2\zeta_h}\text{)} \\
&\geq \eta(\theta^\star, a_{t-1}) + \frac{\epsilon^2}{4\zeta_h} - \mathbb{E}_{\theta_t\sim p_t}\left[\left(2 + \frac{\zeta_g}{\zeta_h}\right)\Delta_t\right] && \text{(By Cauchy-Schwarz)}
\end{aligned}$$

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

$$v_t^\top H_t v_t \geq 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}\,\|v_t\|_2^2\,. \tag{21}$$

Define function

$$\hat{\eta}_t(\theta, a) = \eta(\theta, a_{t-1}) + \langle a - a_{t-1}, \nabla_a\eta(\theta, a_{t-1})\rangle$$

$$+\frac{1}{2}\left\langle \nabla_a^2 \eta(\theta_t, a_{t-1})(a-a_{t-1}), a-a_{t-1}\right\rangle - \frac{\zeta_{3\mathrm{rd}}}{2}\left\|a-a_{t-1}\right\|_2^3 \tag{22}$$

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

By Eq. (21), we can exploit the positive curvature by taking $a' = a_{t-1} + 4\sqrt{\frac{\epsilon}{\zeta_{3\mathrm{rd}}}} v_t$. Concretely, by basic algebra we get:

$$\begin{aligned}\hat{\eta}_t(\theta^\star, a') &\geq \eta(\theta^\star, a_{t-1}) - \epsilon\left\|a'-a_{t-1}\right\|_2 + 3\sqrt{\zeta_{3\mathrm{rd}}\epsilon}\left\|a'-a_{t-1}\right\|_2^2 - \frac{\zeta_{3\mathrm{rd}}}{2}\left\|a'-a_{t-1}\right\|_2^3 \\ &\geq \eta(\theta^\star, a_{t-1}) + 12\sqrt{\frac{\epsilon^3}{\zeta_{3\mathrm{rd}}}}. \end{aligned} \tag{23}$$

Combining with the definition of $\Delta_{t,2}$, $\Delta_{t,3}$ and $\Delta_{t,4}$, for any $a \in \mathcal{A}$ we get

$$\hat{\eta}_t(\theta_t, a) \geq \hat{\eta}_t(\theta^\star, a) - \Delta_{t,2} - \left\|a-a_{t-1}\right\|_2 \Delta_{t,3} - \frac{1}{2}\left\|a-a_{t-1}\right\|_2^2 \Delta_{t,4}. \tag{24}$$

As a result, we have

$$\begin{aligned}
\eta(\theta^\star, a_t) &\geq \mathbb{E}_{\theta_t\sim p_t}[\eta(\theta_t, a_t) - \Delta_{t,1}] \\
&\geq \mathbb{E}_{\theta_t\sim p_t}[\eta(\theta_t, a') - \Delta_{t,1}] && \text{(By the optimality of } a_t\text{)} \\
&\geq \mathbb{E}_{\theta_t\sim p_t}\left[\hat{\eta}_t(\theta^\star, a') - \Delta_{t,1} - \Delta_{t,2} - \left\|a-a_{t-1}\right\|_2 \Delta_{t,3} - \frac{1}{2}\left\|a-a_{t-1}\right\|_2^2 \Delta_{t,4}\right] && \text{(By Eq. (24))} \\
&\geq \eta(\theta^\star, a_{t-1}) + 12\sqrt{\frac{\epsilon^3}{\zeta_{3\mathrm{rd}}}} - \mathbb{E}_{\theta_t\sim p_t}\left[\Delta_{t,1} + \Delta_{t,2} + 4\sqrt{\frac{\epsilon}{\zeta_{3\mathrm{rd}}}}\Delta_{t,3} + \frac{8\epsilon}{\zeta_{3\mathrm{rd}}}\Delta_{t,4}\right] && \text{(By Eq. (23))} \\
&\geq \eta(\theta^\star, a_{t-1}) + 12\sqrt{\frac{\epsilon^3}{\zeta_{3\mathrm{rd}}}} - \mathbb{E}_{\theta_t\sim p_t}[2\Delta_t]. && \text{(When } 16\epsilon \leq \zeta_{3\mathrm{rd}}\text{)}
\end{aligned}$$

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 σ-field generated by random variable $u_{1:t}, v_{1:t}, \theta_{1:t}$. In the following, we use $\mathbb{E}_t[\cdot]$ as a shorthand for $\mathbb{E}[\cdot \mid \mathcal{F}_t]$.

Let $g_t = \nabla_a \eta(\theta_t, a_{t-1}) - \nabla_a \eta(\theta^\star, a_{t-1})$. Note that condition on θ_t and $\mathcal{F}_{t-1}$, $\langle g_t, u_t\rangle$ follows the distribution $\mathcal{N}(0, \|g_t\|_2^2)$. By Assumption 2.1, $\|g_t\|_2 \leq 2\zeta_g = \kappa_1$. As a result,

$$\mathbb{E}_{t-1}\left[\min\left(\kappa_1^2, \langle g_t, u_t\rangle^2\right) \mid \theta_t\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\langle g_t, u_t\rangle^2 \mid \theta_t\right] = \frac{1}{2}\left\|g_t\right\|_2^2. \tag{25}$$

By the tower property of expectation we get

$$\mathbb{E}_{t-1}\left[\min\left(\kappa_1^2, \tilde{\Delta}_{t,3}^2\right)\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\Delta_{t,3}^2\right]. \tag{26}$$

Now we turn to the term $\tilde{\Delta}_{t,4}^2$. Let $H_t = \nabla_a^2 \eta(\theta_t, a_{t-1}) - \nabla_a^2 \eta(\theta^\star, a_{t-1})$. Define a random variable $x = \left(u_t^\top H_t v_t\right)^2$. Note that u_t, v_t are independent, we have

$$\mathbb{E}_{t-1}[x \mid \theta_t] = \mathbb{E}_{t-1}\left[\left\|H_t v_t\right\|_2^2 \mid \theta_t\right] = \left\|H_t\right\|_{\mathrm{F}}^2 \geq \left\|H_t\right\|_{\mathrm{sp}}^2. \tag{27}$$

Since u_t, v_t are two Gaussian vectors, random variable x has nice concentratebility properties. Therefore we can prove that the $\min$ operator in the definition of $\tilde{\Delta}_t$ does not change the expectation too much. Formally speaking, by Lemma I.6, condition on $\mathcal{F}_{t-1}$ and θ_t, we have $\mathbb{E}\left[\min\left(\kappa_2^2, x\right)\right] \geq \frac{1}{2}\min\left(\zeta_h^2, \mathbb{E}[x]\right)$, which leads to

$$\mathbb{E}_{t-1}\left[\min(\kappa_2^2, \tilde{\Delta}_{t,4}^2)\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\min(\zeta_h^2, \|H_t\|_{\mathrm{F}}^2)\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\min(\zeta_h^2, \|H_t\|_{\mathrm{sp}}^2)\right] = \frac{1}{2}\mathbb{E}_{t-1}\left[\|H_t\|_{\mathrm{sp}}^2\right]. \tag{28}$$

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 $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-approximate local maximum in polynomial steps. In the following, we treat $\zeta_g, \zeta_h, \zeta_{3\mathrm{rd}}$ as constants.

Proof of Theorem 3.1. We prove this theorem by contradiction. Suppose $\Pr\left[a_{t+1} \in \mathfrak{A}_{\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}\right] \leq 0.5$ for all $t \in [T]$, we prove that $T \lesssim \widetilde{\mathcal{O}}(C_1^4 R(\Theta) \max\left(\zeta_h^4 \epsilon^{-8}, \zeta_{3\mathrm{rd}}^2 \epsilon^{-6}\right))$.

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

$$\eta(\theta^\star, a_{t+1}) \geq \eta(\theta^\star, a_t) + \upsilon - C_1 \mathbb{E}_t[\Delta_{t+1}]. \tag{29}$$

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

$$\eta(\theta^\star, a_{t+1}) \geq \eta(\theta^\star, a_t) - C_1 \mathbb{E}_t[\Delta_{t+1}]. \tag{30}$$

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

$$\mathbb{E}[\eta(\theta^\star, a_{t+1})] \geq \mathbb{E}[\eta(\theta^\star, a_t)] + \frac{\upsilon}{2} - C_1 \mathbb{E}[\Delta_{t+1}]. \tag{31}$$

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

$$\mathbb{E}[\eta(\theta^\star, a_T) - \eta(\theta^\star, a_0)] \geq \frac{\upsilon T}{2} - C_1 \mathbb{E}\left[\sum_{t=1}^T \Delta_t\right]. \tag{32}$$

Lemma F.2 leads to

$$\mathbb{E}\left[\sum_{t=1}^T \Delta_t\right] \leq \sqrt{2T\mathbb{E}\left[\sum_{t=1}^T \tilde{\Delta}_t^2\right]} \leq 2T^{3/4}(R(\Theta)\mathrm{polylog}(T))^{1/4}. \tag{33}$$

Combining with Eq. (32) we have

$$1 \geq \mathbb{E}[\eta(\theta^\star, a_T) - \eta(\theta^\star, a_0)] \geq \frac{\upsilon T}{2} - 2C_1 T^{3/4}(R(\Theta)\mathrm{polylog}(T))^{1/4}. \tag{34}$$

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

$$T \lesssim C_1^4 R(\Theta) \max\left(\zeta_h^4 \epsilon^{-8}, \zeta_{3\mathrm{rd}}^2 \epsilon^{-6}\right) \mathrm{polylog}(R(\Theta), 1/\epsilon, C_1, \zeta_h, \zeta_{3\mathrm{rd}}). \tag{35}$$

Consequently, when $T \geq \widetilde{\mathcal{O}}(C_1^4 R(\Theta) \max\left(\zeta_h^4 \epsilon^{-8}, \zeta_{3\mathrm{rd}}^2 \epsilon^{-6}\right))$, there exists $t \in [T]$ such that $\Pr\left[a_{t+1} \in \mathfrak{A}_{\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}\right] > 0.5$. Finally, by running running Alg. 1 $\log(1/\delta)$ times, we can get a high probability guarantee. □

F.5. Local Regret

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

$$\mathrm{REG}_{\epsilon_g, \epsilon_h}(T) = \sum_{t=1}^T (\inf_{a \in \mathfrak{A}_{\epsilon_g, \epsilon_h}} \eta(\theta^\star, a) - \eta(\theta^\star, a_t)). \tag{36}$$

Our goal is to achieve a (ϵ_g, ϵ_h)-local regret that is sublinear in T and inverse polynomial in ϵ_g and ϵ_h. With a sublinear regret (i.e., $\mathrm{REG}_{\epsilon_g, \epsilon_h}(T) = o(T)$), the average performance $\frac{1}{T}\sum_{t=1}^T \eta(\theta^\star, a_t)$, converges to that of an approximate local maximum of $\eta(\theta^\star, \cdot)$. 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 $\mathfrak{R}_T$ be the sequential Rademacher complexity of the family of the losses defined in Eq. (5). Let $C_1 = 2 + \zeta_g/\zeta_h$. Under Assumption 2.1, for any $\epsilon \leq \min(1, \zeta_{3\mathrm{rd}}/16)$, we can bound the $(\epsilon, 6\sqrt{\zeta_{3\mathrm{rd}}\epsilon})$-local regret of Alg. 1 from above by*

$$\mathbb{E}\left[\mathrm{REG}_{\epsilon,6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}(T)\right] \leq \left(1 + C_1\sqrt{4T\mathfrak{R}_T}\right)\max\left(4\zeta_h\epsilon^{-2}, \sqrt{\zeta_{3\mathrm{rd}}}\epsilon^{-3/2}\right). \tag{37}$$

Note that when the sequential Rademacher complexity $\mathfrak{R}_T$ is bounded by $\widetilde{\mathcal{O}}(R\sqrt{T})$ (which is typical), we have $\mathcal{O}(\sqrt{T\mathfrak{R}_T}) = \widetilde{\mathcal{O}}(T^{3/4}) = o(T)$ regret. As a result, Alg. 1 achieves a $\mathcal{O}(\mathrm{poly}(1/\epsilon))$ 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 $\delta_t = \inf_{a\in\mathfrak{A}_{(\epsilon),6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}} \eta(\theta^\star, a) - \eta(\theta^\star, a_t)$. By the definition of regret we have $\mathrm{REG}_{\epsilon,6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}(T) = \sum_{t=1}^T \delta_t$. Define $\upsilon = \min\left(\frac{1}{4\zeta_h}\epsilon^2, \frac{1}{\zeta_{3\mathrm{rd}}^{1/2}}\epsilon^{3/2}\right)$ for simplicity. Recall that $C_1 = 2 + \frac{\zeta_g}{\zeta_h}$. In the following we prove by induction that for any t_0,

$$\mathbb{E}_{t_0-1}\left[\sum_{t=t_0}^T \delta_t\right] \leq \mathbb{E}_{t_0-1}\left[\frac{1}{\upsilon}\left(\delta_{t_0} + C_1\sum_{t=t_0+1}^T \Delta_t\right)\right]. \tag{38}$$

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

Now suppose Eq. (38) holds for any $t > t_0$ and consider time step t_0. When $a_{t_0} \notin \mathfrak{A}_{(\epsilon),6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}$, applying Lemma F.1 we get $\eta(\theta^\star, a_{t_0+1}) \geq \eta(\theta^\star, a_{t_0}) + \upsilon - C_1\mathbb{E}_{t_0}[\Delta_{t_0+1}]$. By basic algebra we get,

$$\delta_{t_0+1} \leq \delta_{t_0} - \upsilon + C_1\mathbb{E}_{t_0}[\Delta_{t_0+1}]. \tag{39}$$

As a result,

$$\begin{aligned}
\mathbb{E}_{t_0-1}\left[\sum_{t=t_0}^T \delta_t\right] &= \mathbb{E}_{t_0-1}\left[\delta_{t_0} + \sum_{t=t_0+1}^T \delta_t\right] \\
&\leq \mathbb{E}_{t_0-1}\left[\delta_{t_0} + \frac{1}{\upsilon}\left(\delta_{t_0+1} + C_1\sum_{t=t_0+2}^T \Delta_t\right)\right] && \text{(By induction hypothesis)} \\
&\leq \mathbb{E}_{t_0-1}\left[\delta_{t_0} - 1 + \frac{1}{\upsilon}\left(\delta_{t_0} + C_1\Delta_{t_0+1} + C_1\sum_{t=t_0+2}^T \Delta_t\right)\right] && \text{(By Eq. (39))} \\
&\leq \mathbb{E}_{t_0-1}\left[\frac{1}{\upsilon}\left(\delta_{t_0} + C_1\sum_{t=t_0+1}^T \Delta_t\right)\right].
\end{aligned}$$

On the other hand, when $a_{t_0} \in \mathfrak{A}_{(\epsilon),6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}$ we have

$$\begin{aligned}
&\eta(\theta^\star, a_{t_0+1}) \geq \mathbb{E}_{\theta_{t_0+1}}[\eta(\theta_{t_0+1}, a_{t_0+1}) - \Delta_{t_0+1,1}] \\
\geq &\mathbb{E}_{\theta_{t_0+1}}[\eta(\theta_{t_0+1}, a_{t_0}) - \Delta_{t_0+1,1}] && \text{(By the optimality of } a_{t_0+1}) \\
\geq &\mathbb{E}_{\theta_{t_0+1}}[\eta(\theta^\star, a_{t_0}) - \Delta_{t_0+1,1} - \Delta_{t_0+1,2}] \geq \eta(\theta^\star, a_{t_0}) - C_1\mathbb{E}_{\theta_{t_0+1}}[\Delta_{t_0+1}].
\end{aligned}$$

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 $a_{t_0} \in \mathfrak{A}_{(\epsilon),6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}$, we have $\delta_{t_0} \leq 0$. As a result,

$$\begin{aligned}
\mathbb{E}_{t_0-1}\left[\sum_{t=t_0}^T \delta_t\right] &\leq \mathbb{E}_{t_0-1}\left[\delta_{t_0} + \frac{1}{\upsilon}\left(\delta_{t_0+1} + C_1\sum_{t=t_0+2}^T \Delta_t\right)\right] && \text{(By induction hypothesis)} \\
&\leq \mathbb{E}_{t_0-1}\left[\frac{1}{\upsilon}\left(\delta_{t_0+1} + C_1\sum_{t=t_0+2}^T \Delta_t\right)\right] && (\delta_{t_0} \leq 0)
\end{aligned}$$

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

Combining the two cases together we prove Eq. (38). It follows that

$$\mathbb{E}\left[\mathrm{REG}_{\epsilon,6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}(T)\right] = \mathbb{E}\left[\sum_{t=0}^{T}\delta_t\right] \leq \mathbb{E}\left[\frac{1}{\upsilon}\left(\delta_0 + C_1\sum_{t=1}^{T}\Delta_t\right)\right] \tag{40}$$

$$\leq \frac{1}{\upsilon}\left(1 + C_1\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\Delta_t^2}\right]\right) \leq \frac{1}{\upsilon}\left(1 + C_1\sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t^2\right]}\right). \tag{41}$$

Note that 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

$$\mathbb{E}\left[\mathrm{REG}_{\epsilon,6\sqrt{\zeta_{3\mathrm{rd}}\epsilon}}(T)\right] \leq \frac{1}{\upsilon}\left(1 + C_1\sqrt{2T\mathbb{E}\left[\sum_{t=1}^{T}\tilde{\Delta}_t^2\right]}\right) \leq \frac{1}{\upsilon}\left(1 + C_1\sqrt{4T\mathfrak{R}_T}\right). \tag{42}$$

□

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 $\eta(\theta^\star, a)$ is concave in a with no bad local maxima, and therefore our local regret and the standard regret coincide (up to some conversion of the errors); (ii) invoking Rakhlin et al. (2015b, Lemma 3) to show that the sequential Rademacher complexity $\mathfrak{R}_T$ is bounded by $\mathcal{O}(\sqrt{(2\log|\Theta|)/T})$, 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 ℓ_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 $a_t = \mathrm{argmax}_{a\in\mathcal{A}}\mathbb{E}_{\theta_t\sim p_t}[\tilde{\eta}(\theta_t, a)]$. By optimality condition we have $a_t = \mathbb{E}_{\theta_t\sim p_t}[\theta_t]$. Consequently $\|a_t\|_2 \leq \mathbb{E}_{\theta_t\sim p_t}[\|\theta_t\|_2] \leq 1$.

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 $\|a\|_2 \lesssim 1$. Since the gradient and Hessian of regularization term is a and I_d respectively, we have $\|\nabla_a\tilde{\eta}(\theta, a)\|_2 \lesssim \|\nabla_a\eta(\theta, a)\|_2 + 1$ and $\left\|\nabla_a^2\tilde{\eta}(\theta, a)\right\|_{\mathrm{sp}} \lesssim \left\|\nabla_a^2\eta(\theta, a)\right\|_{\mathrm{sp}} + 1$ when $\|a\|_2 \lesssim 1$, which verifies Assumption 2.1.

In the following we prove that any $(\epsilon, 1)$-approximate local maximum for $\tilde{\eta}(\theta^\star, a)$ is an ϵ-suboptimal action for $\eta(\theta^\star, a)$. Note that $\nabla_a\tilde{\eta}(\theta, a) = \theta - a$. Therefore, for any $a \in \mathfrak{A}_{(\epsilon,1)}$ we have $\|\theta^\star - a\|_2 \leq \epsilon$. Applying Lemma I.11 we have

$$(1 - \langle\theta^\star, a\rangle)^2 \leq \|\theta^\star - a\|_2^2 \leq \epsilon^2. \tag{43}$$

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 Θ by finding a $1/\mathrm{poly}(d, 1/\epsilon)$-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 $\mathcal{O}(s\log(dT))$, 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 $\mathcal{F} = \{f_i : 0 \le i \le d\}$. 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 $\tilde{\eta}(\theta, a) = \eta(\theta, a) - \frac{c}{2}\|a\|_2^2$ where $c = e(e+1)^{-2}$ is a constant. The optimal action in this case is $a^\star = \theta^\star$. 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 ϵ-suboptimal in $\widetilde{\mathcal{O}}(\log|\Theta|\epsilon^{-8})$ steps. Compared with algorithms that specially designed for logistic bandits (Faury et al., 2020; Dong, Van Roy, 2018; Filippi et al., 2010; Li et al., 2017), our regret bound obtained by reduction is not optimal.

In the following we prove that any $(\epsilon, 1)$-approximate local maximum for $\tilde{\eta}(\theta^\star, a)$ is an $\mathcal{O}(\epsilon)$-suboptimal action for $\eta(\theta^\star, a)$. First of all, we bound the set of actions Alg. 1 can choose. By basic algebra we get

$$\nabla_a \tilde{\eta}(\theta, a) = \frac{\exp(-\theta^\top a)}{(1+\exp(-\theta^\top a))^2}\theta - ca. \tag{44}$$

As a result, we have $a^\star = \theta^\star$. Recall that $a_t = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{\theta_t\sim p_t}[\tilde{\eta}(\theta_t, a)]$. By optimality condition we get

$$ca = \mathbb{E}_{\theta_t}\left[\frac{\exp(-\theta_t^\top a)}{(1+\exp(-\theta_t^\top a))^2}\theta_t\right]. \tag{45}$$

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

$$c\|a\|_2^2 = \mathbb{E}_{\theta_t}\left[\frac{\exp(-\theta_t^\top a)}{(1+\exp(-\theta_t^\top a))^2}\theta_t^\top a\right]. \tag{46}$$

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

$$\|a\|_2^2 = \mathbb{E}_{\theta_t}[f(\theta_t^\top a)] \le \sup_{x\in[-\|a\|_2, \|a\|_2]} f(x). \tag{47}$$

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

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

$$\tilde{\eta}(\theta^\star, a^\star) - \tilde{\eta}(\theta^\star, a) \lesssim \|\nabla_a \tilde{\eta}(\theta^\star, a)\|_2^2 \lesssim \epsilon^2. \tag{48}$$

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,

$$\begin{aligned}
&\tilde{\eta}(\theta^\star, a^\star) - \tilde{\eta}(\theta^\star, a) = r(1) - \frac{c}{2} - r(\langle\theta^\star, a\rangle) + \frac{c}{2}\|a\|_2^2 \\
&= (1-\langle\theta^\star, a\rangle)r'(1) - (1-\langle\theta^\star, a\rangle)^2 r''(\xi) - \frac{c}{2} + \frac{c}{2}\|a\|_2^2 \\
&= \frac{c}{2} - c\langle\theta^\star, a\rangle + \frac{c}{2}\|a\|_2^2 - (1-\langle\theta^\star, a\rangle)^2 r''(\xi) && \text{(Recall that } r'(1) = c.\text{)} \\
&= \frac{c}{2}\|\theta^\star - a\|_2^2 - (1-\langle\theta^\star, a\rangle)^2 r''(\xi) \\
&\ge \left(\frac{c}{2} - r''(\xi)\right)(1-\langle\theta^\star, a\rangle)^2 && \text{(By Lemma I.11)}
\end{aligned}$$

$$\geq (1-\langle\theta^\star, a\rangle)^2/50 \gtrsim (r(1)-r(\langle\theta^\star,a\rangle))^2 \qquad \text{(The reward function is Lipschitz.)}$$
$$\gtrsim (\eta(\theta^\star,a^\star)-\eta(\theta^\star,a))^2.$$

Consequently, $\eta(\theta^\star, a) \geq \eta(\theta^\star, a^\star) - \mathcal{O}(\epsilon)$.

Two-layer neural network. Recall that a two-layer neural network is defined by $\eta((W_1, W_2), a) = W_2\sigma(W_1 a)$,where σ is the activation function. For a matrix $W_1 \in \mathbb{R}^{m\times d}$, the $(1,\infty)$-norm is defined by $\max_{i\in[m]}\sum_{j=1}^d |[W_1]_{i,j}|$. We make the following assumptions regarding the activation function.

Assumption F.5. *For any $x, y \in \mathbb{R}$, the activation function $\sigma(\cdot)$ satisfies*

$$\sup_x |\sigma(x)| \leq 1, \quad \sup_x |\sigma'(x)| \leq 1, \quad \sup_x |\sigma''(x)| \leq 1, \tag{49}$$
$$|\sigma''(x) - \sigma''(y)| \leq |x-y|. \tag{50}$$

The following theorem summarized our result in this setting.

Theorem F.6. *Let $\Theta = \{(W_1, W_2) : \|W_2\|_1 \leq 1, \|W_1\|_{1,\infty} \leq 1\}$ be the parameter hypothesis. Under the setting of Theorem 3.1 with Assumption F.5, Alg. 1 finds an $(\epsilon, 6\sqrt{\zeta_{\mathrm{3rd}}\epsilon})$-approximate local maximum in $\widetilde{\mathcal{O}}\big(\epsilon^{-8}\mathrm{polylog}(d)\big)$ steps. In addition, if the neural network is input concave, Alg. 1 finds an ϵ-suboptimal action in $\widetilde{\mathcal{O}}\big(\epsilon^{-4}\mathrm{polylog}(d)\big)$ 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 $u \odot v$ denotes the element-wise product. By basic algebra we get,

$$\langle \nabla_a \eta(\theta, a), u\rangle = W_2(\sigma'(W_1 a) \odot W_1 u), \tag{51}$$
$$u^\top \nabla_a^2 \eta(\theta, a) v = W_2(\sigma''(W_1 a) \odot W_1 u \odot W_1 v). \tag{52}$$

First of all, we verify that the regularized reward $\tilde{\eta}(\theta, a) \triangleq \eta(\theta, a) - \frac{1}{2}\|a\|_2^2$ satisfies Assumption 2.1. Indeed we have

$$\|\nabla_a \eta(\theta,a)\|_2 = \sup_{u\in S^{d-1}} \langle \nabla_a\eta(\theta,a), u\rangle \leq 1,$$
$$\left\|\nabla_a^2\eta(\theta,a)\right\|_{\mathrm{sp}} = \sup_{u,v\in S^{d-1}} u^\top \nabla_a^2 \eta(\theta,a) v \leq 1,$$
$$\left\|\nabla_a^2\eta(\theta,a_1) - \nabla_a^2\eta(\theta,a_2)\right\|_{\mathrm{sp}} = \sup_{u,v\in S^{d-1}} W_2((\sigma''(W_1a_1) - \sigma''(W_1a_2)) \odot W_1 u \odot W_1 v)$$
$$\leq \|a_1 - a_2\|_2 .$$

Observe that $|\eta(\theta, a)| \leq \|a\|_\infty$, we have $\tilde{\eta}(\theta, a) < 0$ when $\|a\|_2 > 2$. As a result, action a_t taken by Alg. 1 satisfies $\|a_t\|_2 \leq 2$ for all t. Since the gradient and Hessian of regularization term is a and I_d respectively, we have $\|\nabla_a\tilde{\eta}(\theta,a)\|_2 \lesssim \|\nabla_a\eta(\theta,a)\|_2 + 1$ and $\left\|\nabla_a^2\tilde{\eta}(\theta,a)\right\|_{\mathrm{sp}} \lesssim \left\|\nabla_a^2\eta(\theta,a)\right\|_{\mathrm{sp}} + 1$. It follows that Assumption 2.1 holds with constant Lipschitzness for actions a such that $\|a\| \lesssim 1$.

In the following we bound the sequential Rademacher complexity of the loss function. By Rakhlin et al. (2015a, Proposition 15), we can bound the sequential Rademacher complexity of $\Delta_{t,1}^2$ and $\Delta_{t,2}^2$ by $\widetilde{\mathcal{O}}\big(\sqrt{T\log d}\big)$. 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 $\|\sigma'(W_1 a) \odot W_1 u\|_\infty \leq \|u\|_\infty$. Therefore we get

$$\langle \nabla_a \eta(\theta,a), u\rangle \leq \|W_2\|_1 \|\sigma'(W_1 a) \odot W_1 u\|_\infty \leq \|u\|_\infty . \tag{53}$$

Similarly, we get

$$u^\top \nabla_a^2 \eta(\theta,a) v \leq \|u\|_\infty \|v\|_\infty . \tag{54}$$

Let $B = (1+\|u\|_\infty)(1+\|v\|_\infty)$ for shorthand. We consider the error term $\tilde{\Delta}_{t,3}^2 = (\langle\nabla_a\eta(\theta,a),u\rangle - [y_t]_3)^2$. Let $\mathcal{G}_1$ be the function class $\{(\langle\nabla_a\eta(\theta,a),u\rangle - [y_t]_3)^2 : \theta\in\Theta\}$, and $\mathcal{G}_2 = \{\langle\nabla_a\eta(\theta,a),u\rangle : \theta\in\Theta\}$. Applying Rakhlin et al. (2015a, Lemma 4) we get

$$\mathfrak{R}_T(\mathcal{G}_1) \lesssim B\log^{3/2}(T^2)\mathfrak{R}_T(\mathcal{G}_2).$$

Define $\mathcal{G}_3 = \{\sigma'(w_1^\top a) \cdot w_1^\top u : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$. In the following we show that $\mathfrak{R}_T(\mathcal{G}_2) \lesssim \mathfrak{R}_T(\mathcal{G}_3)$. For any sequence $u_1, \cdots, u_T$ and $\mathcal{A}$-valued tree $\boldsymbol{a}$, we have

$$\mathfrak{R}_T(\mathcal{G}_2) = \mathbb{E}_\epsilon \left[\sup_{\substack{W_2 : \|W_2\|_1 \leq 1 \\ g_1, \cdots, g_w \in \mathcal{G}_3}} \sum_{t=1}^T \epsilon_t \left(\sum_{j=1}^w [W_2]_j g_j(\boldsymbol{a}_t(\epsilon)) \right) \right] \tag{55}$$

$$\leq \mathbb{E}_\epsilon \left[\sup_{\substack{W_2 : \|W_2\|_1 \leq 1 \\ g_1, \cdots, g_w \in \mathcal{G}_3}} \|W_2\|_1 \sup_{j \in [w]} \left| \sum_{t=1}^T \epsilon_t g_j(\boldsymbol{a}_t(\epsilon)) \right| \right] \tag{56}$$

$$\leq \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}_3} \left| \sum_{t=1}^T \epsilon_t (g_j(\boldsymbol{a}_t(\epsilon))) \right| \right]. \tag{57}$$

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

$$\mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}_3} \left| \sum_{t=1}^T \epsilon_t (g_j(\boldsymbol{a}_t(\epsilon))) \right| \right] \leq 2\mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}_3} \sum_{t=1}^T \epsilon_t (g_j(\boldsymbol{a}_t(\epsilon))) \right] = 2\mathfrak{R}_T(\mathcal{G}_3). \tag{58}$$

Now we bound $\mathfrak{R}_T(\mathcal{G}_3)$ 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 $\mathcal{G}_4 = \{\sigma'((w_1')^\top a) \cdot w_1^\top u : w_1, w_1' \in \mathbb{R}^d, \|w_1\|_1 \leq 1, \|w_1'\|_1 \leq 1\}$. Since $\mathcal{G}_3 \subset \mathcal{G}_4$ we have $\mathfrak{R}_T(\mathcal{G}_3) \leq \mathfrak{R}_T(\mathcal{G}_4)$. Note that we have $\left|\sigma'(w_1^\top a)\right| \leq 1$ and $w_1^\top u \leq \|u\|_\infty$. Let $\phi(x, y) = xy$, which is $(3c)$-Lipschitz for $|x|, |y| \leq c$. Define $\mathcal{G}_5 = \{\sigma'(w_1^\top a) : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$ and $\mathcal{G}_6 = \{w_1^\top u : w_1 \in \mathbb{R}^d, \|w_1\|_1 \leq 1\}$. Rakhlin et al. (2015a, Lemma 4) gives $\mathfrak{R}_T(\mathcal{G}_4) \lesssim B \log^{3/2}(T^2)(\mathfrak{R}_T(\mathcal{G}_5) + \mathfrak{R}_T(\mathcal{G}_6))$. Note that $\mathcal{G}_5$ is a generalized linear hypothesis and $\mathcal{G}_6$ is linear, we have $\mathfrak{R}_T(\mathcal{G}_5) \lesssim B \log^{3/2}(T^2)\sqrt{T \log(d)}$ and $\mathfrak{R}_T(\mathcal{G}_6) \lesssim B\sqrt{T \log(d)}$.

In summary, we get $\mathfrak{R}_T(\mathcal{G}_1) = \mathcal{O}\Big(\text{poly}(B)\text{polylog}(d, T)\sqrt{T}\Big)$. 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 $\tilde{\Delta}_{t,3}^2$ in this case is bounded by $\mathcal{O}\Big(\text{polylog}(d, T)\sqrt{T}\Big)$.

Similarly, we can bound the sequential Rademacher complexity of the Hessian term $\tilde{\Delta}_{t,4}^2$ by $\mathcal{O}\Big(\text{polylog}(d, T)\sqrt{T}\Big)$ 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 $\text{poly}(\log(T))$ terms in the sequential Rademacher complexity. As a result, the sequential Rademacher complexity of the loss function can be bounded by

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

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

When the neural network is input concave (see (Amos et al., 2017)), the regularized reward $\tilde{\eta}(\theta, a)$ is $\Omega(1)$-strongly concave. As a result, for any $a \in \mathfrak{A}_{\epsilon,1}$ we have $\tilde{\eta}(\theta^\star, a) \geq \tilde{\eta}(\theta^\star, a^\star) - \mathcal{O}(\epsilon^2)$. Hence, Alg. 1 finds an ϵ-suboptimal action for regularized loss in $\widetilde{\mathcal{O}}\big(\epsilon^{-4}\text{polylog}(d)\big)$ 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 ψ. We also assume realizability ($\theta^\star \in \Theta$) and the Lipschitz assumptions as in Assumption 2.1 (with action a replaced by policy parameter ψ).

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 Δ_t by the difference of dynamics, as discussed

Algorithm 2 Virtual Ascent with **Onlin**e Model Learner (ViOlin for RL)

Let $\mathcal{H}_0 = \emptyset$; choose $a_0 \in \mathcal{A}$ arbitrarily.
for $t = 1, 2, \cdots$ **do**
  Run $\mathcal{R}$ on $\mathcal{H}_{t-1}$ with loss function ℓ (defined in Eq. (6)) and obtain $p_t = \mathcal{A}(\mathcal{H}_{t-1})$.
  $\psi_t \leftarrow \operatorname{argmax}_\psi \mathbb{E}_{\theta_t \sim p_t}[\eta(\theta_t, \psi)]$;
  Sample one trajectory τ_t from policy π_{ψ_t}, and one trajectory τ_t' from policy $\pi_{\psi_{t-1}}$.
  Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(\tau, \tau')\}$
end for

before. Formally speaking, let $\tau_t = (s_1, a_1, \cdots, s_H, a_H)$ be a trajectory sampled from policy π_{ψ_t} under the ground-truth dynamics $T_{\theta^\star}$. By telescope lemma (Lemma I.16) we get

$$V_\theta^\psi(s_1) - V_{\theta^\star}^\psi(s_1) = \mathbb{E}_{\tau \sim \rho_{\theta^\star}^\psi}\left[\sum_{h=1}^H \left(V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h, a_h))\right)\right]. \tag{59}$$

Lipschitz assumption (Assumption B.1) yields,

$$\left|V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h, a_h))\right| \le L_0 \left\|T_\theta(s_h, a_h) - T_{\theta^\star}(s_h, a_h))\right\|_2. \tag{60}$$

Combining Eq. (59) and Eq. (60) and apply Cauchy-Schwartz inequality gives an upper bound for $[\Delta_t]_1^2$ and $[\Delta_t]_2^2$. As for the gradient term, we will take gradient w.r.t. ψ to both sides of Eq. (59). The gradient inside expectation can be dealt with easily. And the gradient w.r.t. the distribution $\rho_{\theta^\star}^\psi$ can be computed by policy gradient lemma (Lemma I.17). As a result we get

$$\begin{aligned}
&\nabla_\psi V_\theta^\psi(s_1) - \nabla_\psi V_{\theta^\star}^\psi(s_1) \\
&= \mathbb{E}_{\tau \sim \rho_{\theta^\star}^\psi}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)\left(\sum_{h=1}^H \left(V_\theta^\psi(T_\theta(s_h, a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h, a_h))\right)\right)\right] \\
&\quad + \mathbb{E}_{\tau \sim \rho_{\theta^\star}^\psi}\left[\sum_{h=1}^H \left(\nabla_\psi V_\theta^\psi(T_\theta(s_h, a_h)) - \nabla_\psi V_\theta^\psi(T_{\theta^\star}(s_h, a_h))\right)\right].
\end{aligned} \tag{61}$$

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*

$$c_1 \mathbb{E}_{\tau_{1:t}, \tau_{1:t}', \theta_{1:t}}\left[\bar{\Delta}_t^2\right] \ge \mathbb{E}_{\theta_{1:t}}\left[\Delta_t^2\right]. \tag{62}$$

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 ρ_T^π be the distribution of trajectories generated by policy π and dynamics T. By telescoping lemma (Lemma I.16) we have,

$$V_{\theta_t}^{\psi_t}(s_1) - V_{\theta^\star}^{\psi_t}(s_1) = \mathbb{E}_{\tau \sim \rho_{\theta^\star}^{\psi_t}}\left[\sum_{h=1}^H \left(V_{\theta_t}^{\psi_t}(T_{\theta_t}(s_h, a_h)) - V_{\theta_t}^{\psi_t}(T_{\theta^\star}(s_h, a_h))\right)\right]. \tag{63}$$

By the Lipschitz assumption (Assumption B.1),

$$\left|V_{\theta_t}^{\psi_t}(T_{\theta_t}(s_h, a_h)) - V_{\theta_t}^{\psi_t}(T_{\theta^\star}(s_h, a_h))\right| \le L_0 \left\|T_{\theta_t}(s_h, a_h) - T_{\theta^\star}(s_h, a_h))\right\|_2. \tag{64}$$

Consequently

$$\Delta_{t,1}^2 = \left(V_{\theta_t}^{\psi_t}(s_0) - V_{\theta^\star}^{\psi_t}(s_0)\right)^2 \leq HL_0^2 \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi_t}}\left[\sum_{h=1}^H \|T_{\theta_t}(s_h,a_h) - T_{\theta^\star}(s_h,a_h))\|_2^2\right]. \tag{65}$$

Similarly we get,

$$\Delta_{t,2}^2 = \left(V_{\theta_t}^{\psi_{t-1}}(s_0) - V_{\theta^\star}^{\psi_{t-1}}(s_0)\right)^2 \leq HL_0^2 \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi_{t-1}}}\left[\sum_{h=1}^H \|T_{\theta_t}(s_h,a_h) - T_{\theta^\star}(s_h,a_h))\|_2^2\right]. \tag{66}$$

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

- $\left\|\mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)^\top\right]\right\|_{\text{sp}} \leq H^2\chi_g, \forall \psi \in \Psi;$
- $\left\|\mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)^{\otimes 4}\right]\right\|_{\text{sp}} \leq H^4\chi_f, \forall \psi \in \Psi;$
- $\left\|\mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h)\right)\left(\sum_{h=1}^H \nabla_\psi^2 \log \pi_\psi(a \mid s)\right)^\top\right]\right\|_{\text{sp}} \leq H^2\chi_h, \forall \psi \in \Psi.$

Indeed, consider the first statement. Define $g_h = \nabla_\psi \log \pi_\psi(a_h \mid s_h)$ for shorthand. Then we have

$$\left\|\mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H g_h\right)\left(\sum_{h=1}^H g_h\right)^\top\right]\right\|_{\text{sp}} = \sup_{u\in S^{d-1}} u^\top \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H g_h\right)\left(\sum_{h=1}^H g_h\right)^\top\right] u \tag{67}$$

$$= \sup_{u\in S^{d-1}} \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left\langle u, \left(\sum_{h=1}^H g_h\right)\right\rangle^2\right] \leq \sup_{u\in S^{d-1}} \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[H\sum_{h=1}^H \langle u, g_h\rangle^2\right] \tag{68}$$

$$\leq \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[H\sum_{h=1}^H \sup_{u\in S^{d-1}} \langle u, g_h\rangle^2\right] = \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[H\sum_{h=1}^H \left\|gg^\top\right\|_{\text{sp}}\right] \leq H^2\chi_g. \tag{69}$$

Similarly we can get the second and third statement.

For any fixed ψ and θ we have

$$V_\theta^\psi(s_1) - V_{\theta^\star}^\psi(s_1) = \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\sum_{h=1}^H \left(V_\theta^\psi(T_\theta(s_h,a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h,a_h))\right)\right]. \tag{70}$$

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

$$\begin{aligned}
&\nabla_\psi V_\theta^\psi(s_1) - \nabla_\psi V_{\theta^\star}^\psi(s_1) \\
&= \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)\left(\sum_{h=1}^H \left(V_\theta^\psi(T_\theta(s_h,a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h,a_h))\right)\right)\right] \\
&\quad + \mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi}}\left[\sum_{h=1}^H \left(\nabla_\psi V_\theta^\psi(T_\theta(s_h,a_h)) - \nabla_\psi V_\theta^\psi(T_{\theta^\star}(s_h,a_h))\right)\right].
\end{aligned} \tag{71}$$

Define the following shorthand:

$$G_\theta^\psi(s,a) = V_\theta^\psi(T_\theta(s,a)) - V_\theta^\psi(T_{\theta^\star}(s,a)), \tag{72}$$

$$f = \sum_{h=1}^{H} \nabla_\psi \log \pi_\psi(a_h \mid s_h). \tag{73}$$

In the following we also omit the subscription in $\mathbb{E}_{\tau\sim\rho^{\psi}_{\theta^\star}}$ when the context is clear. It followed by Eq. (71) that

$$\begin{aligned}
&\left\|\nabla_\psi V^\psi_\theta(s_1) - \nabla_\psi V^\psi_{\theta^\star}(s_1)\right\|_2^2 \\
&\le 2\left\|\mathbb{E}\left[f\left(\sum_{h=1}^{H} G^\psi_\theta(s_h,a_h)\right)\right]\right\|_2^2 + 2\left\|\mathbb{E}\left[\sum_{h=1}^{H}\nabla_\psi G^\psi_\theta(s_h,a_h)\right]\right\|_2^2 \\
&\le 2\left\|\mathbb{E}\left[ff^\top\right]\right\|_{\mathrm{sp}}\mathbb{E}\left[\left(\sum_{h=1}^{H} G^\psi_\theta(s_h,a_h)\right)^2\right] + 2\left\|\mathbb{E}\left[\sum_{h=1}^{H}\nabla_\psi G^\psi_\theta(s_h,a_h)\right]\right\|_2^2 && \text{(By Lemma I.7)} \\
&\le 2H\left\|\mathbb{E}\left[ff^\top\right]\right\|_{\mathrm{sp}}\mathbb{E}\left[\sum_{h=1}^{H} G^\psi_\theta(s_h,a_h)^2\right] + 2H\mathbb{E}\left[\sum_{h=1}^{H}\left\|\nabla_\psi G^\psi_\theta(s_h,a_h)\right\|_2^2\right].
\end{aligned}$$

Now, plugin $\psi = \psi_{t-1}, \theta = \theta_t$ and apply Assumption B.1 we get

$$\begin{aligned}
\Delta^2_{t,3} &= \left\|\nabla_\psi V^{\psi_{t-1}}_{\theta_t}(s_1) - \nabla_\psi V^{\psi_{t-1}}_{\theta^\star}(s_1)\right\|_2^2 \\
&\le (2HL_1^2 + 2H^3\chi_g L_0^2)\mathbb{E}_{\tau\sim\rho^{\psi_{t-1}}_{\theta^\star}}\left[\sum_{h=1}^{H}\|T_{\theta_t}(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2\right].
\end{aligned}$$

For any fixed ψ, θ, define the following shorthand:

$$g = \sum_{h=1}^{H}\Big(V^\psi_\theta(T_\theta(s_h,a_h)) - V^\psi_\theta(T_{\theta^\star}(s_h,a_h))\Big). \tag{74}$$

Apply policy gradient lemma again to RHS of Eq. (71) we get

$$\begin{aligned}
&\nabla^2_\psi V^\psi_\theta(s_1) - \nabla^2_\psi V^\psi_{\theta^\star}(s_1) \\
&= \mathbb{E}\left[(\nabla_\psi g)f^\top\right] + \mathbb{E}\left[f(\nabla_\psi g)^\top\right] + \mathbb{E}\left[\nabla^2_\psi g\right] + \mathbb{E}\left[g\left(\sum_{h=1}^{H}\nabla^2_\psi \log\pi_\psi(a_h\mid s_h)\right)\right] + \mathbb{E}\left[g\left(ff^\top\right)\right].
\end{aligned}$$

As a result of Lemma I.8 and Lemma I.9 that,

$$\begin{aligned}
&\left\|\nabla^2_\psi V^\psi_\theta(s_1) - \nabla^2_\psi V^\psi_{\theta^\star}(s_1)\right\|_{\mathrm{sp}}^2 \\
&= 4\left\|\mathbb{E}\left[(\nabla_\psi g)f^\top\right] + \mathbb{E}\left[f(\nabla_\psi g)^\top\right]\right\|_{\mathrm{sp}}^2 + 4\left\|\mathbb{E}\left[\nabla^2_\psi g\right]\right\|_{\mathrm{sp}}^2 + 4\left\|\mathbb{E}\left[g\left(ff^\top\right)\right]\right\|_{\mathrm{sp}}^2 \\
&\quad + 4\left\|\mathbb{E}\left[g\left(\sum_{h=1}^{H}\nabla^2_\psi \log\pi_\psi(a_h\mid s_h)\right)\right]\right\|_{\mathrm{sp}}^2 \\
&\le 8\sup_{u,v\in S^{d-1}}\mathbb{E}[\langle\nabla_\psi g, u\rangle\langle f, v\rangle]^2 + 4\mathbb{E}\left[\left\|\nabla^2_\psi g\right\|_{\mathrm{sp}}^2\right] + 4\mathbb{E}\left[g^2\right]\left\|\mathbb{E}\left[f^{\otimes 4}\right]\right\|_{\mathrm{sp}} \\
&\quad + 4\mathbb{E}\left[g^2\right]\left\|\mathbb{E}\left[\left(\sum_{h=1}^{H}\nabla^2_\psi \log\pi_\psi(a_h\mid s_h)\right)\left(\sum_{h=1}^{H}\nabla^2_\psi \log\pi_\psi(a_h\mid s_h)\right)^\top\right]\right\|_{\mathrm{sp}}.
\end{aligned} \tag{75}$$

Note that by Hölder's inequality,

$$\sup_{u,v\in S^{d-1}}\mathbb{E}[\langle\nabla_\psi g, u\rangle\langle f, v\rangle]^2 \le \sup_{u,v\in S^{d-1}}\mathbb{E}\left[\langle\nabla_\psi g, u\rangle^2\right]\mathbb{E}\left[\langle f, v\rangle^2\right] \le \mathbb{E}\left[\|\nabla_\psi g\|_2^2\right]\left\|\mathbb{E}\left[ff^\top\right]\right\|_{\mathrm{sp}}.$$

By Assumption B.1 we get,

$$\mathbb{E}[g^2] = \mathbb{E}\left[\left(\sum_{h=1}^{H}\left(V_\theta^\psi(T_\theta(s_h,a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h,a_h))\right)\right)^2\right] \tag{76}$$

$$\leq H\mathbb{E}\left[\sum_{h=1}^{H}\left(V_\theta^\psi(T_\theta(s_h,a_h)) - V_\theta^\psi(T_{\theta^\star}(s_h,a_h))\right)^2\right] \tag{77}$$

$$\leq HL_0^2\mathbb{E}\left[\sum_{h=1}^{H}\|T_\theta(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2\right]. \tag{78}$$

Similarly, we have

$$\mathbb{E}[\|\nabla_\psi g\|_2^2] \leq HL_1^2\mathbb{E}\left[\sum_{h=1}^{H}\|T_\theta(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2\right], \tag{79}$$

$$\mathbb{E}[\left\|\nabla_\psi^2 g\right\|_{\mathrm{sp}}^2] \leq HL_2^2\mathbb{E}\left[\sum_{h=1}^{H}\|T_\theta(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2\right]. \tag{80}$$

Combining with Eq. (75) we get,

$$\Delta_{t,4}^2 = \left\|\nabla_\psi^2 V_{\theta_t}^{\psi_t}(s_1) - \nabla_\psi V_{\theta^\star}^{\psi_t}(s_1)\right\|_{\mathrm{sp}}^2$$
$$\leq \left(8H^3L_1^2\chi_g + 4HL_2^2 + 4L_0^2(H^3\chi_h + H^5\chi_f)\right)\mathbb{E}_{\tau\sim\rho_{\theta^\star}^{\psi_{t-1}}}\left[\sum_{h=1}^{H}\|T_{\theta_t}(s_h,a_h) - T_{\theta^\star}(s_h,a_h)\|_2^2\right].$$

By noting that $\Delta_t^2 = \sum_{i=1}^4 \Delta_{i,t}^2$, we get the desired upper bound. □

H. Analysis of Example B.3

Recall that our RL instance is given as follows:

$$T(s,a) = \mathrm{N}_\theta(s+a), \tag{81}$$

$$\pi_\psi(s) = \mathcal{N}(\psi s, \sigma^2 I). \tag{82}$$

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 $\|\psi\|_{\mathrm{op}} \leq \mathcal{O}(1)$.

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,

$$\nabla_\psi \log \pi_\psi(a \mid s) = \frac{1}{\sigma^2}\mathrm{vec}((a - \psi s)\otimes s) \tag{83}$$

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

$$\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi \log \pi_\psi(a \mid s))(\nabla_\psi \log \pi_\psi(a \mid s))^\top]\|_{\mathrm{sp}} \tag{84}$$

$$=\frac{1}{\sigma^4}\sup_{v\in S^{d\times d-1}}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[\langle v,\mathrm{vec}(u\otimes s)\rangle^2\Big]. \tag{85}$$

Note that $\langle v,\mathrm{vec}(u\otimes s)\rangle=\sum_{1\le i,j\le d}[u]_i[s]_j[v]_{i,j}$. Because u is isotropic, $[u]_i$ are independent random variables where $[u]_i\sim\mathcal{N}(0,\sigma^2)$. Therefore $\langle v,\mathrm{vec}(u\otimes s)\rangle\sim\mathcal{N}\left(0,\sigma^2\sum_{i=1}^d\left(\sum_{j=1}^d s_jv_{i,j}\right)^2\right)$. Combining with Eq. (85) we get,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[\langle v,\mathrm{vec}(u\otimes s)\rangle^2\Big]=\sigma^2\sum_{i=1}^d\left(\sum_{j=1}^d s_jv_{i,j}\right)^2 \tag{86}$$

$$\le\sigma^2\sum_{i=1}^d\left(\sum_{j=1}^d s_j^2\right)\left(\sum_{j=1}^d v_{i,j}^2\right)\le\|s\|_2^2\,\|v\|_2^2\,. \tag{87}$$

Consequently we have

$$\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi\log\pi_\psi(a\mid s))(\nabla_\psi\log\pi_\psi(a\mid s))^\top]\|_{\mathrm{sp}}\le\frac{1}{\sigma^2}\triangleq\chi_g. \tag{88}$$

Verifying item 2. Similarly, using the equation where $\mathbb{E}_{x\sim\mathcal{N}(0,\sigma^2)}[x^4]=3\sigma^4$ we have

$$\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi\log\pi_\psi(a\mid s))^{\otimes 4}]\|_{\mathrm{sp}}=\frac{1}{\sigma^8}\sup_{v\in S^{d\times d-1}}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[\langle v,\mathrm{vec}(u\otimes s)\rangle^4\Big]. \tag{89}$$

$$\le\frac{3}{\sigma^8}\left(\sigma^2\sum_{i=1}^d\left(\sum_{j=1}^d s_jv_{i,j}\right)^2\right)^2\le\frac{3}{\sigma^4}\|s\|_2^4\,\|v\|_2^4\le\frac{3}{\sigma^4}\triangleq\chi_f. \tag{90}$$

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

$$\|\mathbb{E}_{a\sim\pi_\psi(\cdot|s)}[(\nabla_\psi^2\log\pi_\psi(a\mid s))(\nabla_\psi^2\log\pi_\psi(a\mid s))^\top]\|_{\mathrm{sp}} \tag{91}$$

$$=\sup_v\mathbb{E}\big[v^\top(\nabla_\psi^2\log\pi_\psi(a\mid s))(\nabla_\psi^2\log\pi_\psi(a\mid s))^\top v\big] \tag{92}$$

$$=\sup_v\mathbb{E}\Big[\big\|(\nabla_\psi^2\log\pi_\psi(a\mid s))^\top v\big\|_2^2\Big]=\sup_v\mathbb{E}\Big[\big(v^\top(\nabla_\psi^2\log\pi_\psi(a\mid s))v\big)^4\Big]. \tag{93}$$

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

$$v^\top(\nabla_\psi^2\log\pi_\psi(a\mid s))v=-\sum_{i=1}^d\left(\sum_{j=1}^d v_{i,j}s_j\right)^2. \tag{94}$$

Consequently,

$$\big(v^\top(\nabla_\psi^2\log\pi_\psi(a\mid s))v\big)^4\le\|s\|_2^4\,\|v\|_2^4\le 1\triangleq\chi_h. \tag{95}$$

H.2. Verifying Assumption 2.1.

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

$$\eta(\theta,\psi)=\mathbb{E}_{\tau\sim\rho_\theta^\psi}\left[\sum_{h=1}^H r(s_h,a_h)\right]. \tag{96}$$

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

$$\nabla_\psi\eta(\theta,\psi)=\mathbb{E}_{\tau\sim\rho_\theta^\psi}\left[\left(\sum_{h=1}^H\nabla_\psi\log\pi_\psi(a_h\mid s_h)\right)\left(\sum_{h=1}^H r(s_h,a_h)\right)\right]. \tag{97}$$

By Eq. (83), condition on s_h we get

$$\nabla_\psi \log \pi_\psi(a_h \mid s_h) = \frac{1}{\sigma^2}\mathrm{vec}(u \otimes s_h) \tag{98}$$

where $u = a_h - \psi s_h \sim \mathcal{N}(0, \sigma^2 I)$. Define the shorthand $g = \sum_{h=1}^H r(s_h, a_h)$. Note that by Hölder inequality,

$$\|\mathbb{E}[\nabla_\psi \log \pi_\psi(a_h \mid s_h) g]\|_2^2 = \sup_{v \in \mathbb{R}^{d\times d}, \|v\|_2 = 1} \mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle g]^2 \tag{99}$$

$$\leq \sup_{v \in \mathbb{R}^{d\times d}, \|v\|_2 = 1} \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle^2\Big]\mathbb{E}\big[g^2\big]. \tag{100}$$

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

$$\mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle^2\Big] = \frac{1}{\sigma^2}\|v s_h\|_2^2 \leq \frac{1}{\sigma^2}\|v\|_{\mathrm{F}}^2 \|s_h\|_2^2 \leq \frac{1}{\sigma^2}. \tag{101}$$

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

$$\|\nabla_\psi \eta(\theta, \psi)\|_2 \leq H^2/\sigma. \tag{102}$$

Verifying item 2. Define the shorthand $f = \sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)$. Use policy gradient lemma on Eq. (97) again we get, for any $v, w \in \mathbb{R}^{d\times d}$,

$$v^\top \nabla_\psi^2 \eta(\theta, \psi) w = \mathbb{E}_{\tau \sim \rho_\theta^\psi}\left[\langle f, v\rangle \langle f, w\rangle g + \left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h) w\right) g\right]. \tag{103}$$

For the first term inside the expectation, we bound it by using Hölder inequality twice. Specifically, for any $h, h' \in [H]$ we have

$$\mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle \langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w\rangle g] \tag{104}$$

$$\leq \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle^4\Big]^{1/4} \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w\rangle^4\Big]^{1/4} \mathbb{E}\big[g^2\big]^{1/2}. \tag{105}$$

Similarly, $\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle \sim \frac{1}{\sigma^2}\mathcal{N}(0, \sigma^2 \|v s_h\|_2^2)$ and $\langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w\rangle \sim \frac{1}{\sigma^2}\mathcal{N}(0, \sigma^2 \|w s_{h'}\|_2^2)$. As a result,

$$\mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h \mid s_h), v\rangle \langle \nabla_\psi \log \pi_\psi(a_{h'} \mid s_{h'}), w\rangle g] \tag{106}$$

$$\leq 3\frac{1}{\sigma^2}\|v\|_2 \|s_h\|_2 \|w\|_2 \|s_{h'}\|_2 H \leq \frac{3H}{\sigma^2}. \tag{107}$$

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

$$v^\top \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h) w = -\frac{1}{\sigma^2}\langle w s_h, v s_h\rangle. \tag{108}$$

Consequently,

$$\mathbb{E}\left[\left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h \mid s_h) w\right) g\right] \leq \frac{H^2}{\sigma^2}\|w\|_2 \|v\|_2 \leq \frac{H^2}{\sigma^2}. \tag{109}$$

In summary, we have $\left\|\nabla_\psi^2 \eta(\theta, \psi)\right\|_{\mathrm{op}} \leq \frac{4H^3}{\sigma^2}$.

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

$$\begin{aligned}
\left\langle \nabla_\psi^3 \eta(\theta,\psi), v\otimes w\otimes x\right\rangle =& \mathbb{E}_{\tau\sim\rho_\theta^\psi}[\langle f,v\rangle \langle f,w\rangle \langle f,w\rangle g] && (110)\\
&+\mathbb{E}_{\tau\sim\rho_\theta^\psi}\left[\langle f,x\rangle \left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h\mid s_h) w\right) g\right] && (111)\\
&+\mathbb{E}_{\tau\sim\rho_\theta^\psi}\left[\langle f,v\rangle \left(\sum_{h=1}^H x^\top \nabla_\psi^2 \log \pi_\psi(a_h\mid s_h) w\right) g\right] && (112)\\
&+\mathbb{E}_{\tau\sim\rho_\theta^\psi}\left[\langle f,w\rangle \left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h\mid s_h) x\right) g\right]. && (113)
\end{aligned}$$

Following the same argument, by Hölder inequality, for any $h_1, h_2, h_3 \in [H]$ we have

$$\begin{aligned}
&\mathbb{E}[\langle \nabla_\psi \log \pi_\psi(a_h\mid s_h), v\rangle \langle \nabla_\psi \log \pi_\psi(a_{h'}\mid s_{h'}), w\rangle \langle \nabla_\psi \log \pi_\psi(a_{h'}\mid s_{h'}), x\rangle g]\\
\le& \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_h\mid s_h), v\rangle^6\Big]^{1/6} \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_{h'}\mid s_{h'}), w\rangle^6\Big]^{1/6} \mathbb{E}\Big[\langle \nabla_\psi \log \pi_\psi(a_{h'}\mid s_{h'}), x\rangle^6\Big]^{1/6} H\\
\le& \frac{\sqrt{15}H}{\sigma^3}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\langle f,x\rangle \left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h\mid s_h) w\right) g\right]\\
\le& \mathbb{E}\Big[\langle f,x\rangle^2\Big]^{1/2} \mathbb{E}\left[\left(\left(\sum_{h=1}^H v^\top \nabla_\psi^2 \log \pi_\psi(a_h\mid s_h) w\right) g\right)^2\right]^{1/2}\\
\le& \frac{H^3}{\sigma^3}.
\end{aligned}$$

By symmetricity, Eq. (110) can be upper bounded by

$$\left\langle \nabla_\psi^3 \eta(\theta,\psi), v\otimes w\otimes x\right\rangle \le \frac{7H^4}{\sigma^3}. \tag{114}$$

H.3. Verifying Assumption B.1.

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

$$\begin{aligned}
V_\theta^\psi(s) =& \mathbb{E}_{a\sim\pi_\psi(s)}\Big[r(s,a) + V_\theta^\psi(T(s,a))\Big] && (115)\\
=& \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[r(s,\psi s+u) + V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u))\Big]. && (116)
\end{aligned}$$

Define $B = 1 + \|\psi\|_{\mathrm{op}}$ for shorthand. For two states $s_1, s_2 \in \mathcal{S}$, by the Lipschitz assumption on reward function we have

$$|r(s_1,\psi s_1+u) - r(s_2,\psi s_2+u)| \le L_r B \|s_1 - s_2\|_2. \tag{117}$$

Then consider the second term in Eq. (116). Since we have $|V_\theta^\pi| \le H$ and

$$\mathrm{TV}\big(\mathcal{N}(s_1+\psi s_1,\sigma^2 I), \mathcal{N}(s_2+\psi s_2,\sigma^2 I)\big) \le \frac{1}{2\sigma}\|s_1+\psi s_1 - s_2 - \psi s_2\|_2 \le \frac{B\|s_1-s_2\|_2}{2\sigma},$$

it follows that

$$\left|\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[V_\theta^\psi(\mathrm{N}_\theta(s_1+\psi s_1+u))\Big] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\Big[V_\theta^\psi(\mathrm{N}_\theta(s_2+\psi s_2+u))\Big]\right| \le \frac{HB}{2\sigma}\|s_1-s_2\|_2.$$

As a result, item 1 of Assumption B.1 holds as follows

$$\left|V_\theta^\psi(s_1) - V_\theta^\psi(s_2)\right| \leq \left(\frac{HB}{2\sigma} + L_r B\right) \|s_1 - s_2\|_2 . \tag{118}$$

Verifying item 2. Now we turn to verifying the Lipschitzness of gradient term. Recall that by policy gradient lemma we have for every $v \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}
&\left\langle \nabla_\psi V_\theta^\psi(s), v \right\rangle && (119)\\
&= \mathbb{E}_{a\sim\pi_\psi(s)}\left[\left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s+a)), v\right\rangle\right] && (120)\\
&\quad + \mathbb{E}_{a\sim\pi_\psi(s)}\left[\langle \nabla_\psi \log \pi_\psi(a \mid s), v\rangle \left(r(s,a) + V_\theta^\psi(\mathrm{N}_\theta(s+a))\right)\right] && (121)\\
&= \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u)), v\right\rangle\right] && (122)\\
&\quad + \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle \nabla_\psi \log \pi_\psi(\psi s+u \mid s), v\rangle \left(r(s,\psi s+u) + V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u))\right)\right]. && (123)
\end{aligned}$$

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

By the boundness of $\left\|\nabla_\psi V_\theta^\psi(s)\right\|_2$ (specifically, item 1 of Assumption 2.1), we have

$$\begin{aligned}
&\left|\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s_1+\psi s_1+u)), v\right\rangle\right] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s_2+\psi s_2+u)), v\right\rangle\right]\right|\\
&\leq \frac{H^2}{\sigma}\mathrm{TV}\big(\mathcal{N}(s_1+\psi s_1, \sigma^2 I), \mathcal{N}(s_2+\psi s_2, \sigma^2 I)\big) \leq \frac{H^2}{\sigma}\frac{B}{2\sigma}\|s_1 - s_2\|_2 .
\end{aligned}$$

For the reward term in Eq. (123), recalling $v \in \mathbb{R}^{d\times d}$ we have

$$\mathbb{E}_{a\sim\pi_\psi(s)}[\langle \nabla_\psi \log \pi_\psi(\psi s+u \mid s), v\rangle\, r(s,\psi s+u)] = \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs, u\rangle\, r(s,\psi s+u)].$$

Note that

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle\, r(s_1,\psi s_1+u)] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_2, u\rangle\, r(s_2,\psi s_2+u)] && (124)\\
&= \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle\, (r(s_1,\psi s_1+u) - r(s_2,\psi s_2+u))] && (125)\\
&\quad + \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1, u\rangle - \langle vs_2, u\rangle) r(s_2,\psi s_2+u)]. && (126)
\end{aligned}$$

Note that u is isotropic. Applying Lemma I.10 we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1, u\rangle - \langle vs_2, u\rangle) r(s_2,\psi s_2+u)] \leq \sigma\|vs_1 - vs_2\|_2 \leq \sigma\|s_1 - s_2\|_2 . \tag{127}$$

We can also bound the term in Eq. (125) by

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle\, (r(s_1,\psi s_1+u) - r(s_2,\psi s_2+u))] && (128)\\
&\leq \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle vs_1, u\rangle^2\right]^{1/2} \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[(r(s_1,\psi s_1+u) - r(s_2,\psi s_2+u))^2\right]^{1/2} && (129)\\
&\leq \sigma\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[L_r^2 B^2 \|s_1 - s_2\|_2^2\right]^{1/2} \leq \sigma L_r B\|s_1 - s_2\|_2 . && (130)
\end{aligned}$$

Now we deal with the last term in Eq. (123). Let $f(s,u) = V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u))$ for shorthand. Similarly we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle \nabla_\psi \log \pi_\psi(\psi s+u \mid s), v\rangle\, V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u))\right] = \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs, u\rangle\, f(s,u)]. \tag{131}$$

By the same telescope sum we get,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle\, f(s_1,u)] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_2, u\rangle\, f(s_2,u)] \tag{132}$$

$$= \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle (f(s_2,u) - f(s_2,u))] \tag{133}$$
$$+ \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1, u\rangle - \langle vs_2, u\rangle) f(s_2,u)]. \tag{134}$$

Applying Lemma I.10 we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1, u\rangle - \langle vs_2, u\rangle) f(s_2,u)] \le \sigma H \left\| vs_1 - vs_2 \right\|_2 \le \sigma H \left\| s_1 - s_2 \right\|_2. \tag{135}$$

Applying Lemma I.12 we have

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

In summary, we have

$$\left\| \nabla_\psi V_\theta^\psi(s_1) - \nabla_\psi V_\theta^\psi(s_2) \right\|_2 \le \mathrm{poly}(H, B, \sigma, 1/\sigma, L_r) \left\| s_1 - s_2 \right\|_2. \tag{137}$$

Verifying item 3. Lastly, we verify the Lipschitzness of Hessian term. Applying policy gradient lemma to Eq. (119) again we have

$$w^\top \nabla_\psi^2 V_\theta^\psi(s) v \tag{138}$$
$$= \mathbb{E}_{a\sim\pi_\psi(s)}\left[w^\top \nabla_\psi^2 V_\theta^\psi(\mathrm{N}_\theta(s+a)) v \right] \tag{139}$$
$$+ \mathbb{E}_{a\sim\pi_\psi(s)}\left[\left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s+a)), v \right\rangle \langle \nabla_\psi \log \pi_\psi(a \mid s), w\rangle \right] \tag{140}$$
$$+ \mathbb{E}_{a\sim\pi_\psi(s)}\left[\langle \nabla_\psi \log \pi_\psi(a \mid s), v\rangle \left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s+a)), w \right\rangle \right] \tag{141}$$
$$+ \mathbb{E}_{a\sim\pi_\psi(s)}\left[\langle \nabla_\psi \log \pi_\psi(a \mid s), v\rangle \langle \nabla_\psi \log \pi_\psi(a \mid s), w\rangle \left(r(s,a) + V_\theta^\psi(\mathrm{N}_\theta(s+a)) \right) \right]. \tag{142}$$

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 $\left\| \nabla_\psi^2 V_\theta^\psi(\mathrm{N}_\theta(s+a)) \right\|_{\mathrm{op}}$ (specifically, item 2 of Assumption 2.1) we have

$$\left| \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[w^\top \nabla_\psi^2 V_\theta^\psi(\mathrm{N}_\theta(s_1 + \psi s_1 + u)) v \right] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[w^\top \nabla_\psi^2 V_\theta^\psi(\mathrm{N}_\theta(s_1 + \psi s_1 + u)) v \right] \right|$$
$$\le \frac{4H^3}{\sigma^2} \mathrm{TV}\big(\mathcal{N}(s_1 + \psi s_1, \sigma^2 I), \mathcal{N}(s_2 + \psi s_2, \sigma^2 I)\big) \le \frac{3H^3}{\sigma^2} \frac{B}{2\sigma} \left\| s_1 - s_2 \right\|_2.$$

For the terms in Eq. (140), let $f(s,u) = \left\langle \nabla_\psi V_\theta^\psi(\mathrm{N}_\theta(s + \psi s + a)), v \right\rangle$. Repeat the same argument when verifying item 2 again, we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle ws_1, u\rangle (f(s_2,u) - f(s_2,u))] \le 6B \frac{H^2}{\sigma} \left\| s_1 - s_2 \right\|_2 \left(1 + \frac{1}{\sigma}\right). \tag{143}$$

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

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle \langle ws_1, u\rangle r(s_1, \psi s_1 + u)] - \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_2, u\rangle \langle ws_2, u\rangle r(s_2, \psi s_2 + u)] \tag{144}$$
$$= \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1, u\rangle \langle ws_1, u\rangle (r(s_1, \psi s_1 + u) - r(s_2, \psi s_2 + u))] \tag{145}$$
$$+ \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1, u\rangle \langle ws_1, u\rangle - \langle vs_2, u\rangle \langle ws_2, u\rangle) r(s_2, \psi s_2 + u)]. \tag{146}$$

Note that u is isotropic. Applying Lemma I.15 we have

$$\begin{aligned}&\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1,u\rangle\langle ws_1,u\rangle-\langle vs_2,u\rangle\langle ws_2,u\rangle)r(s_2,\psi s_2+u)]\\ \le&\sqrt{3}\sigma^2(\|vs_1-vs_2\|_2+\|ws_1-ws_2\|_2)\le 2\sqrt{3}\sigma^2\|s_1-s_2\|_2,\end{aligned}$$

We can also bound the term in Eq. (145) by

$$\begin{aligned}&\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1,u\rangle\langle ws_1,u\rangle(r(s_1,\psi s_1+u)-r(s_2,\psi s_2+u))]\\ \le&\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle vs_1,u\rangle^4\right]^{1/4}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle ws_1,u\rangle^4\right]^{1/4}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[(r(s_1,\psi s_1+u)-r(s_2,\psi s_2+u))^2\right]^{1/2}\\ \le&\sqrt{3}\sigma^2\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[L_r^2B^2\|s_1-s_2\|_2^2\right]^{1/2}\le\sqrt{3}\sigma^2L_rB\|s_1-s_2\|_2.\end{aligned}$$

Now we deal with the last term in Eq. (142). Let $f(s,u)=V_\theta^\psi(\mathrm{N}_\theta(s+\psi s+u))$ for shorthand. Similarly we have

$$\mathbb{E}_{a\sim\pi_\psi(s)}\left[\langle\nabla_\psi\log\pi_\psi(a\mid s),v\rangle\langle\nabla_\psi\log\pi_\psi(a\mid s),w\rangle V_\theta^\psi(\mathrm{N}_\theta(s+a))\right] \tag{147}$$

$$=\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs,u\rangle\langle ws,u\rangle f(s,u)]. \tag{148}$$

By the same telescope sum we get,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1,u\rangle\langle ws_1,u\rangle f(s_1,u)]-\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_2,u\rangle\langle ws_2,u\rangle f(s_2,u)] \tag{149}$$

$$=\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1,u\rangle\langle ws_1,u\rangle(f(s_1,u)-f(s_2,u))] \tag{150}$$

$$+\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1,u\rangle\langle ws_1,u\rangle-\langle vs_2,u\rangle\langle ws_2,u\rangle)f(s_2,u)]. \tag{151}$$

Applying Lemma I.15 we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle vs_1,u\rangle\langle ws_1,u\rangle-\langle vs_2,u\rangle\langle ws_2,u\rangle)f(s_2,u)] \tag{152}$$

$$\le\sqrt{3}\sigma^2H(\|vs_1-vs_2\|_2+\|ws_1-ws_2\|_2)\le 2\sqrt{3}\sigma^2H\|s_1-s_2\|_2. \tag{153}$$

Applying Lemma I.13 we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle vs_1,u\rangle\langle ws_1,u\rangle(f(s_2,u)-f(s_2,u))]\le\mathrm{poly}(H,\sigma,1/\sigma)B\|s_1-s_2\|_2. \tag{154}$$

In summary, we have

$$\left\|\nabla_\psi^2V_\theta^\psi(s_1)-\nabla_\psi^2V_\theta^\psi(s_2)\right\|_{\mathrm{op}}\le\mathrm{poly}(H,B,\sigma,1/\sigma,L_r)\|s_1-s_2\|_2. \tag{155}$$

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*

$$\forall t>0,\Pr\left[\|Av\|_2^2\ge\mathrm{Tr}(\Sigma)+2\sqrt{\mathrm{Tr}(\Sigma^2)t}+2\|\Sigma\|_{\mathrm{op}}t\right]\le\exp(-t). \tag{156}$$

Corollary I.2. *Under the same settings of Lemma I.1,*

$$\forall t>1,\Pr\left[\|Av\|_2^2\ge\|A\|_{\mathrm{F}}^2+4\|A\|_{\mathrm{F}}^2t\right]\le\exp(-t). \tag{157}$$

Proof. Let λ_i be the i-th eigenvalue of Σ. By the definition of Σ we have $\lambda_i \geq 0$. Then we have

$$\begin{aligned}
\mathrm{Tr}(\Sigma) &= \sum_{i=1}^{n} \lambda_i = \|A\|_{\mathrm{F}}^2 \,, \\
\mathrm{Tr}(\Sigma^2) &= \sum_{i=1}^{n} \lambda_i^2 \leq \left(\sum_{i=1}^{n} \lambda_i\right)^2 = \|A\|_{\mathrm{F}}^4 \,, \\
\|\Sigma\|_{\mathrm{op}} &= Ax_{i\in[n]}\lambda_i \leq \sum_{i=1}^{n} \lambda_i = \|A\|_{\mathrm{F}}^2 \,.
\end{aligned}$$

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*

$$\forall t \geq 1, \Pr\left[(u^\top H v)^2 \geq t \|H\|_{\mathrm{F}}^2\right] \leq 3\exp(-\sqrt{t}/4). \tag{158}$$

Proof. Condition on v, $u^\top H v$ is a Gaussian random variable with mean zero and variance $\|Hv\|_2^2$. Therefore we have,

$$\forall v, \Pr\left[\left(u^\top H v\right)^2 \geq \sqrt{t} \|Hv\|_2^2\right] \leq \exp(-\sqrt{t}/2). \tag{159}$$

By Corollary I.2 and basic algebra we get,

$$\Pr\left[\|Hv\|_2^2 \geq \sqrt{t} \|H\|_{\mathrm{F}}^2\right] \leq 2\exp(-\sqrt{t}/4). \tag{160}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{I}\left[(u^\top H v)^2 \geq t \|H\|_{\mathrm{F}}^2\right]\Big] \\
\leq\ &\mathbb{E}\Big[\mathbb{I}\left[(u^\top H v)^2 \geq \sqrt{t} \|Hv\|_2^2 \text{ or } \|Hv\|_2^2 \geq \sqrt{t} \|H\|_{\mathrm{F}}^2\right]\Big] \\
\leq\ &\mathbb{E}\Big[\mathbb{I}\left[(u^\top H v)^2 \geq \sqrt{t} \|Hv\|_2^2\right] \mid v\Big] + \mathbb{E}\Big[\mathbb{I}\left[\|Hv\|_2^2 \geq \sqrt{t} \|H\|_{\mathrm{F}}^2\right]\Big] \\
\leq\ &3\exp(-\sqrt{t}/4). \qquad \text{(Combining Eq. (159) and Eq. (160))}
\end{aligned}$$

□

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 $a = (a_1, \cdots, a_n)$ be non-negative coefficient. Let*

$$\|a\|_2^2 = \sum_{i=1}^{n} a_i^2.$$

Then for any positive t,

$$\Pr\left(\sum_{i=1}^{n} a_i y_i^2 \leq \sum_{i=1}^{n} a_i - 2\|a\|_2 \sqrt{t}\right) \leq \exp(-t). \tag{161}$$

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*

$$\Pr\left[(u^\top H v)^2 \geq \frac{1}{8} \|H\|_{\mathrm{F}}^2\right] \geq \frac{1}{64}. \tag{162}$$

Proof. Since u, v are independent, by the isotropy of Guassian vectors we can assume that $H = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$. 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,

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

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

$$\begin{aligned}
&\Pr\left[\|Hv\|_2^2 \geq \frac{1}{2}\|H\|_{\mathrm{F}}^2\right] \\
\geq\ &\Pr\left[\|Hv\|_2^2 \geq \|H\|_{\mathrm{F}}^2 - \frac{1}{2}\sqrt{\sum_{i=1}^n \lambda_i^4}\right] \\
=\ &\Pr\left[\sum_{i=1}^n \lambda_i^2 v_i^2 \geq \sum_{i=1}^n \lambda_i^2 - \frac{1}{2}\sqrt{\sum_{i=1}^n \lambda_i^4}\right] && \text{(By definition)} \\
\geq\ &1 - \exp(-1/16) \geq \frac{1}{32}. && (164)
\end{aligned}$$

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

$$\Pr\left[(u^\top H v)^2 \geq \frac{1}{8}\|H\|_{\mathrm{F}}^2\right] \geq \Pr\left[(u^\top H v)^2 \geq \frac{1}{4}\|Hv\|_2^2, \|Hv\|_{\mathrm{F}}^2 \geq \frac{1}{2}\|H\|_{\mathrm{F}}^2\right] \geq \frac{1}{64}.$$

□

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*

$$\min\left(c_1^2, \|H\|_{\mathrm{F}}^2\right) \leq 2\mathbb{E}\Big[\min\left(\kappa_2^2, \left(u^\top H v\right)^2\right)\Big]. \tag{165}$$

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

Case 1: $\|H\|_{\mathrm{F}} \leq \kappa_2/40$. In this case we exploit the tail bound of random variable x. Specifically,

$$\begin{aligned}
&\mathbb{E}\Big[\left(u^\top H v\right)^2\Big] - \mathbb{E}\Big[\min\left(\kappa_2^2, \left(u^\top H v\right)^2\right)\Big] \\
=\ &\int_{\kappa_2^2}^{\infty} \Pr\left[x \geq t\right] dt \\
\leq\ &3\int_{\kappa_2^2}^{\infty} \exp\left(-\frac{1}{4}\sqrt{\frac{t}{\|H\|_{\mathrm{F}}^2}}\right) dt && \text{(By Lemma I.3)} \\
=\ &24\exp\left(-\frac{\kappa_2}{4\|H\|_{\mathrm{F}}}\right)\|H\|_{\mathrm{F}}\left(\kappa_2 + 4\|H\|_{\mathrm{F}}\right) \\
\leq\ &48\exp\left(-\frac{\kappa_2}{4\|H\|_{\mathrm{F}}}\right)\|H\|_{\mathrm{F}}\,\kappa_2 && (4\|H\|_{\mathrm{F}} \leq \kappa_2 \text{ in this case}) \\
\leq\ &48 \cdot \frac{4\|H\|_{\mathrm{F}}}{384\kappa_2}\|H\|_{\mathrm{F}}\,\kappa_2 && (\exp(-x) \leq \tfrac{1}{384x} \text{ when } x \geq 10) \\
\leq\ &\frac{\|H\|_{\mathrm{F}}^2}{2}.
\end{aligned}$$

As a result,

$$\mathbb{E}\Big[\min\left(\kappa_2^2, \left(u^\top H v\right)^2\right)\Big] \geq \mathbb{E}\Big[\left(u^\top H v\right)^2\Big] - \frac{\|H\|_{\mathrm{F}}^2}{2} = \frac{\|H\|_{\mathrm{F}}^2}{2}. \tag{166}$$

Case 2: $\|H\|_{\mathrm{F}} > \kappa_2/40$. In this case, we exploit the anti-concentration result of random variable x. Note that by the choice of κ_2, we have

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

As a result,

$$\begin{aligned}
&\mathbb{E}\Big[\min\Big(\kappa_2^2, \big(u^\top H v\big)^2\Big)\Big] \\
&\geq 64c_1^2 \Pr\Big[\min\Big(\kappa_2^2, \big(u^\top H v\big)^2\Big) \geq 64c_1^2\Big] \\
&\geq 64c_1^2 \Pr\Big[\big(u^\top H v\big)^2 \geq 64c_1^2\Big] && \text{(By definition of } \kappa_2\text{)} \\
&\geq 64c_1^2 \Pr\Big[\big(u^\top H v\big)^2 \geq \frac{1}{8}\|H\|_{\mathrm{F}}^2\Big] \\
&\geq c_1^2. && \text{(By Lemma I.5)}
\end{aligned}$$

Therefore, in both cases we get

$$\mathbb{E}\Big[\min\Big(\kappa_2^2, \big(u^\top H v\big)^2\Big)\Big] \geq \frac{1}{2}\min\Big(c_1^2, \|H\|_{\mathrm{F}}^2\Big), \tag{167}$$

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*

$$\|\mathbb{E}[rx]\|_2^2 \leq \big\|\mathbb{E}\big[xx^\top\big]\big\|_{\mathrm{op}} \mathbb{E}\big[r^2\big]. \tag{168}$$

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,

$$\begin{aligned}
\|\mathbb{E}[rx]\|_2^2 &= \sup_{u \in S^{d-1}} \langle u, \mathbb{E}[rx]\rangle^2 = \sup_{u \in S^{d-1}} \mathbb{E}[r\langle u, x\rangle]^2 \\
&\leq \sup_{u \in S^{d-1}} \mathbb{E}\Big[\langle u, x\rangle^2\Big]\mathbb{E}\big[r^2\big] && \text{(Hölder Ineqaulity)} \\
&= \big\|\mathbb{E}\big[xx^\top\big]\big\|_{\mathrm{op}} \mathbb{E}\big[r^2\big].
\end{aligned}$$

□

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

$$\|\mathbb{E}[rH]\|_{\mathrm{sp}}^2 \leq \big\|\mathbb{E}\big[HH^\top\big]\big\|_{\mathrm{sp}} \mathbb{E}\big[r^2\big]. \tag{169}$$

Proof. Note that for any matrix $\boldsymbol{G} \in \mathbb{R}^d$, $\|H\|_{\mathrm{sp}}^2 = \sup_{u,v \in S^{d-1}} \big(u^\top \boldsymbol{G} v\big)^2$. As a result,

$$\begin{aligned}
\|\mathbb{E}[rH]\|_2^2 &= \sup_{u,v \in S^{d-1}} \big(u^\top \mathbb{E}[rH] v\big)^2 = \sup_{u,v \in S^{d-1}} \mathbb{E}\big[r\big(u^\top H v\big)\big]^2 \\
&\leq \sup_{u,v \in S^{d-1}} \mathbb{E}\Big[\big(u^\top H v\big)^2\Big]\mathbb{E}\big[r^2\big] && \text{(Hölder Ineqaulity)} \\
&= \sup_{u,v \in S^{d-1}} \mathbb{E}\big[u^\top H v v^\top H^\top u\big]\mathbb{E}\big[r^2\big] \\
&\leq \sup_{u \in S^{d-1}} \mathbb{E}\big[u^\top H H^\top u\big]\mathbb{E}\big[r^2\big] \\
&= \big\|\mathbb{E}\big[HH^\top\big]\big\|_{\mathrm{sp}} \mathbb{E}\big[r^2\big].
\end{aligned}$$

□

Lemma I.9. *For a random matrix $x \in \mathbb{R}^d$ and a positive random variable r, we have*

$$\left\|\mathbb{E}\left[rxx^\top\right]\right\|_{\text{sp}}^2 \leq \left\|\mathbb{E}[x^{\otimes 4}]\right\|_{\text{sp}} \mathbb{E}\left[r^2\right]. \tag{170}$$

Proof. Since r is non-negative, we have $\mathbb{E}\left[rxx^\top\right] \succeq 0$. As a result,

$$\left\|\mathbb{E}\left[rxx^\top\right]\right\|_{\text{sp}} = \sup_{u \in S^{d-1}} u^\top \mathbb{E}\left[rxx^\top\right] u.$$

It follows that

$$\begin{aligned}
&\left\|\mathbb{E}\left[rxx^\top\right]\right\|_{\text{sp}}^2 = \sup_{u \in S^{d-1}} \left(u^\top \mathbb{E}\left[rxx^\top\right] u\right)^2 = \sup_{u \in S^{d-1}} \mathbb{E}\left[r \left\langle u, x\right\rangle^2\right]^2 \\
\leq &\sup_{u \in S^{d-1}} \mathbb{E}\left[\left\langle u, x\right\rangle^4\right] \mathbb{E}\left[r^2\right] && \text{(Hölder Inequality)} \\
= &\sup_{u \in S^{d-1}} \left\langle u^{\otimes 4}, \mathbb{E}[x^{\otimes 4}]\right\rangle \mathbb{E}\left[r^2\right] \\
= &\left\|\mathbb{E}\left[x^{\otimes 4}\right]\right\|_{\text{sp}} \mathbb{E}\left[r^2\right].
\end{aligned}$$

□

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 $g \in \mathbb{R}^d$, $\langle g, u\rangle \sim \mathcal{N}(0, \|g\|_2^2)$ when $u \sim \mathcal{N}(0, I)$,

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*

$$\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[(\langle p, u\rangle - \langle q, u\rangle) f(u)] \leq \sigma B \left\|p - q\right\|_2. \tag{171}$$

Proof. By Hölder inequality we have

$$\begin{aligned}
&\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}[(\langle p, u\rangle - \langle q, u\rangle) f(u)] && (172) \\
\leq &\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}\left[\left(\langle p, u\rangle - \langle q, u\rangle\right)^2\right]^{1/2} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}\left(f(u)^2\right)^{1/2}. && (173)
\end{aligned}$$

Note that u is isotropic. As a result $\langle p - q, u\rangle \sim \mathcal{N}(0, \sigma^2 \|p - q\|_2^2)$. It follows that

$$\begin{aligned}
&\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}\left[\left(\langle p, u\rangle - \langle q, u\rangle\right)^2\right]^{1/2} \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)}\left(f(u)^2\right)^{1/2} && (174) \\
\leq &\sigma \left\|p - q\right\|_2 B. && (175)
\end{aligned}$$

□

Lemma I.11. *For two vectors $x, y \in \mathbb{R}^d$, if $\|x\|_2 = 1$ we have*

$$\|x - y\|_2^2 \geq (1 - \langle x, y\rangle)^2. \tag{176}$$

Proof. By basic algebra we get

$$\begin{aligned}
&\|x - y\|_2^2 = \|x - \langle x, y\rangle x + \langle x, y\rangle x - y\|_2^2 && (177) \\
= &\|x - \langle x, y\rangle x\|_2^2 + \|\langle x, y\rangle x - y\|_2^2 - 2(1 - \langle x, y\rangle) \langle x, \langle x, y\rangle x - y\rangle && (178) \\
= &\|x - \langle x, y\rangle x\|_2^2 + \|\langle x, y\rangle x - y\|_2^2 && (179) \\
= &(1 - \langle x, y\rangle)^2 + \|\langle x, y\rangle x - y\|_2^2 \geq (1 - \langle x, y\rangle)^2. && (180)
\end{aligned}$$

□

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*

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

Proof. The lemma is proved by coupling argument. With out loss of generality, we assume that $x_1 = Ce_1$ and $x_2 = -Ce_1$ 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,

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

As a result,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[\langle p, u\rangle (f(x_1+u) - f(x_2+u))] \tag{183}$$

$$= \int_{y\in\mathbb{R}^d} F(y - x_1) \langle p, y - x_1\rangle f(y)\mathrm{d}y - F(y - x_2) \langle p, y - x_2\rangle f(y)\mathrm{d}y. \tag{184}$$

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

$$\int_{y\in\mathbb{R}^d} F(y - x_1) \langle p, y - x_1\rangle f(y)\mathrm{d}y - F(y - x_2) \langle p, y - x_2\rangle f(y)\mathrm{d}y \tag{185}$$

$$\le \int_{y\in\mathbb{R}^d} G(y)|\langle p, y - x_1\rangle - \langle p, y - x_2\rangle| f(y)\mathrm{d}y \tag{186}$$

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

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

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

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

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

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 $[y]_1 \le 0$. As a result,

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

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

$$\le \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1))\mathbb{E}_{[y]_{-1}}[(|\langle [p]_{-1}, [y - x_1]_{-1}\rangle + [p]_1[y - x_1]_1|) f(y)]. \tag{193}$$

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

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

$$\le \mathbb{E}_{[y]_{-1}}\left[\langle [p]_{-1}, [y - x_1]_{-1}\rangle^2\right]^{1/2} \mathbb{E}_{[y]_{-1}}\left[f(y)^2\right]^{1/2} \tag{195}$$

$$\leq B\sigma \left\|[p]_{-1}\right\|_2 \leq B\sigma \left\|p\right\|_2 . \tag{196}$$

On the other hand, we have

$$\int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1)) \leq \mathrm{TV}\big(\mathcal{N}(-C, \sigma^2), \mathcal{N}(C, \sigma^2)\big) \leq \frac{1}{\sigma} \left\|x_1 - x_2\right\|_2 . \tag{197}$$

It follows that,

$$\int_{y:[y]_1>0} (F(y - x_1) - G(y)) |\langle p, y - x_1\rangle| f(y) \mathrm{d}y \tag{198}$$

$$\leq B \left\|p\right\|_2 \left\|x_1 - x_2\right\|_2 + B \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1)) |[p]_1 [y - x_1]_1|. \tag{199}$$

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 $\left\|p\right\|_2 \left(\frac{3\|x_1 - x_2\|_2}{\sigma} \left\|x_1\right\|_2 + 4 \left\|x_1 - x_2\right\|_2\right)$. Therefore, we have

$$\int_{y:[y]_1>0} (F(y - x_1) - G(y)) |\langle p, y - x_1\rangle| f(y) \mathrm{d}y \leq 5B \left\|p\right\|_2 \left\|x_1 - x_2\right\|_2 + \left\|p\right\|_2 \frac{3B \left\|x_1 - x_2\right\|_2}{\sigma}. \tag{200}$$

□

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

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)} [\langle p, u\rangle \langle q, u\rangle (f(x_1 + u) - f(x_2 + u))] \leq \mathrm{poly}(B, \sigma, 1/\sigma, \left\|p\right\|_2, \left\|q\right\|_2) \left\|x_1 - x_2\right\|_2 . \tag{201}$$

Proof. Proof of this lemma is similar to that of Lemma I.13. With out loss of generality, we assume that $x_1 = Ce_1$ and $x_2 = -Ce_1$ where e_1 is the first basis vector. That is, $\left\|x_1 - x_2\right\|_2 = 2C$. 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,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)} [\langle p, u\rangle \langle q, u\rangle f(x_1 + u)] = \int_{y\in\mathbb{R}^d} F(y - x_1) \langle p, y - x_1\rangle \langle q, y - x_1\rangle f(y) \mathrm{d}y. \tag{202}$$

As a result,

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)} [\langle p, u\rangle \langle q, u\rangle (f(x_1 + u) - f(x_2 + u))] \tag{203}$$

$$= \int_{y\in\mathbb{R}^d} F(y - x_1) \langle p, y - x_1\rangle \langle q, y - x_1\rangle f(y) \mathrm{d}y - F(y - x_2) \langle p, y - x_2\rangle \langle q, y - x_2\rangle f(y) \mathrm{d}y. \tag{204}$$

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

$$\int_{y\in\mathbb{R}^d} F(y - x_1) \langle p, y - x_1\rangle \langle q, y - x_1\rangle f(y) \mathrm{d}y - F(y - x_2) \langle p, y - x_2\rangle \langle q, y - x_2\rangle f(y) \mathrm{d}y \tag{205}$$

$$\leq \int_{y\in\mathbb{R}^d} G(y) |\langle p, y - x_1\rangle \langle q, y - x_1\rangle - \langle p, y - x_2\rangle \langle q, y - x_2\rangle| f(y) \mathrm{d}y \tag{206}$$

$$+ \int_{y\in\mathbb{R}^d} (F(y - x_1) - G(y)) |\langle p, y - x_1\rangle \langle q, y - x_1\rangle| f(y) \mathrm{d}y \tag{207}$$

$$+ \int_{y\in\mathbb{R}^d} (F(y - x_2) - G(y)) |\langle p, y - x_2\rangle \langle q, y - x_2\rangle| f(y) \mathrm{d}y. \tag{208}$$

By basic algebra we have

$$\int_{y\in\mathbb{R}^d} G(y) |\langle p, y - x_1\rangle \langle q, y - x_1\rangle - \langle p, y - x_2\rangle \langle q, y - x_2\rangle| f(y) \mathrm{d}y \tag{209}$$

$$\leq \int_{y\in\mathbb{R}^d} G(y)|\langle p, x_2 - x_1\rangle \langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{210}$$

$$+ \int_{y\in\mathbb{R}^d} G(y)|\langle p, y - x_2\rangle \langle q, x_2 - x_1\rangle| f(y)\mathrm{d}y. \tag{211}$$

Continue with the first term we get

$$\int_{y\in\mathbb{R}^d} G(y)|\langle p, x_2 - x_1\rangle \langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{212}$$

$$\leq \|x_2 - x_1\|_2 \|p\|_2 \int_{y\in\mathbb{R}^d} G(y)|\langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{213}$$

$$\leq \|x_2 - x_1\|_2 \|p\|_2 \int_{y\in\mathbb{R}^d} F(y - x_1)|\langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{214}$$

$$= \|x_2 - x_1\|_2 \|p\|_2 \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[|\langle q, u\rangle| f(u + x_1)] \tag{215}$$

$$\leq \|x_2 - x_1\|_2 \|p\|_2 \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle q, u\rangle^2\right]^{1/2} \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[f(u + x_1)^2\right]^{1/2} \tag{216}$$

$$\leq \sigma B \|x_2 - x_1\|_2 \|p\|_2 \|q\|_2 . \tag{217}$$

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,

$$\int_{y\in\mathbb{R}^d} G(y)|\langle p, y - x_1\rangle \langle q, y - x_1\rangle - \langle p, y - x_2\rangle \langle q, y - x_2\rangle| f(y)\mathrm{d}y \leq 2\sigma B \|x_2 - x_1\|_2 \|p\|_2 \|q\|_2 . \tag{218}$$

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 $I = |[p]_1[y - x_1]_1|, J = |\langle [p]_{-1}, [y - x_1]_{-1}\rangle|, C = |[q]_1[y - x_1]_1|, D = |\langle [q]_{-1}, [y - x_1]_{-1}\rangle|$. When condition on $[y]_1$, A, C are constants. As a result,

$$\int_{y:[y]_1>0} (F(y - x_1) - G(y))|\langle p, y - x_1\rangle \langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{219}$$

$$= \int_{y:[y]_1>0} (F(y - x_1) - F(y - x_2))|\langle p, y - x_1\rangle \langle q, y - x_1\rangle| f(y)\mathrm{d}y \tag{220}$$

$$\leq \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1))\mathbb{E}_{[y]_{-1}}[|(I + J)(C + D)| f(y)] \tag{221}$$

$$\leq \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1)) IC\mathbb{E}_{[y]_{-1}}[f(y)] \tag{222}$$

$$+ \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1)) I\mathbb{E}_{[y]_{-1}}[Df(y)] \tag{223}$$

$$+ \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1)) C\mathbb{E}_{[y]_{-1}}[Jf(y)] \tag{224}$$

$$+ \int_{[y]_1>0} \mathrm{d}[y]_1 (P([y]_1 - [x_1]_1) - P([y_1] - [x_2]_1))\mathbb{E}_{[y]_{-1}}[JDf(y)]. \tag{225}$$

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

$$\mathbb{E}_{[y]_{-1}}[Df(y)] \leq \mathbb{E}_{[y]_{-1}}\left[D^2\right]^{1/2} \mathbb{E}_{[y]_{-1}}\left[f(y)^2\right]^{1/2} \leq B\sigma \|q\|_2 \tag{226}$$

$$\mathbb{E}_{[y]_{-1}}[Jf(y)] \leq \mathbb{E}_{[y]_{-1}}\left[J^2\right]^{1/2} \mathbb{E}_{[y]_{-1}}\left[f(y)^2\right]^{1/2} \leq B\sigma \|p\|_2 \tag{227}$$

$$\mathbb{E}_{[y]_{-1}}[JDf(y)] \leq \mathbb{E}_{[y]_{-1}}\left[J^4\right]^{1/4}\mathbb{E}_{[y]_{-1}}\left[D^4\right]^{1/4}\mathbb{E}_{[y]_{-1}}\left[f(y)^2\right]^{1/2} \leq \sqrt{3}B\sigma^2 \|p\|_2 \|q\|_2 . \tag{228}$$

Invoking Lemma I.14 we get

$$\int_{[y]_1>0} \mathrm{d}[y]_1(P([y]_1-[x_1]_1)-P([y_1]-[x_2]_1)) \leq \frac{1}{\sigma}\|x_1-x_2\|_2 , \tag{229}$$

$$\int_{[y]_1>0} \mathrm{d}[y]_1(P([y]_1-[x_1]_1)-P([y_1]-[x_2]_1))I \leq \mathrm{poly}(B,\sigma,1/\sigma)\|p\|_2 \|x_1-x_2\|_2 , \tag{230}$$

$$\int_{[y]_1>0} \mathrm{d}[y]_1(P([y]_1-[x_1]_1)-P([y_1]-[x_2]_1))C \leq \mathrm{poly}(B,\sigma,1/\sigma)\|q\|_2 \|x_1-x_2\|_2 , \tag{231}$$

$$\int_{[y]_1>0} \mathrm{d}[y]_1(P([y]_1-[x_1]_1)-P([y_1]-[x_2]_1))IC \leq \mathrm{poly}(B,\sigma,1/\sigma)\|p\|_2 \|q\|_2 \|x_1-x_2\|_2 . \tag{232}$$

As a result, we get

$$\int_{y:[y]_1>0} (F(y-x_1)-G(y))|\langle p, y-x_1\rangle \langle q, y-x_1\rangle| f(y)\mathrm{d}y \leq \mathrm{poly}(B,\sigma,1/\sigma,\|p\|_2,\|q\|_2)\|x_1-x_2\|_2 .$$

□

Lemma I.14. *Let $P(x)$ be the density function of $\mathcal{N}(0,\sigma^2)$. Given a scalar $x \geq 0$. we have*

$$\int_{y>0}(P(y-x)-P(y+x))\mathrm{d}y \leq \frac{x}{\sigma}, \tag{233}$$

$$\int_{y>0}(P(y-x)-P(y+x))|y-x|\mathrm{d}y \leq \frac{3x^2}{\sigma}+4x, \tag{234}$$

$$\int_{y>0}(P(y-x)-P(y+x))|y-x|^2\mathrm{d}y \leq \frac{x^2}{\sigma}+4\sigma x. \tag{235}$$

Proof. Note that $\mathrm{TV}\big(\mathcal{N}(-x,\sigma^2),\mathcal{N}(x,\sigma^2)\big) \leq \frac{x}{\sigma}$. Consequently,

$$\int_{y>0}(P(y-x)-P(y+x))\mathrm{d}y \leq \frac{x}{\sigma}. \tag{236}$$

Using the same TV-distance bound, we get

$$\int_{y>0}(P(y-x)-P(y+x))|y-x|\mathrm{d}y \tag{237}$$

$$\leq \int_{y>0}(P(y-x)-P(y+x))((y-x)+2x)\mathrm{d}y \tag{238}$$

$$\leq \int_{y>2x}(P(y-x)-P(y+x))(y-x)\mathrm{d}y+\frac{3x^2}{\sigma}. \tag{239}$$

Recall $P(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$. By algebraic manipulation we have

$$\int_{y>2x}\left(\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)-\exp\left(-\frac{(y+x)^2}{2\sigma^2}\right)\right)(y-x)\mathrm{d}y \tag{240}$$

$$\leq \int_{y>2x}\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\left(1-\exp\left(-\frac{4xy}{2\sigma^2}\right)\right)(y-x)\mathrm{d}y \tag{241}$$

$$\leq \int_{y>2x}\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\frac{4xy}{2\sigma^2}(y-x)\mathrm{d}y \tag{242}$$

$$\leq \int_{y>2x} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\frac{4x}{\sigma^2}(y-x)^2\mathrm{d}y \tag{243}$$
$$\leq 4x. \tag{244}$$

Now we turn to the third inequality. Because $|y-x| \leq x$ for $y \in [0, 2x]$, using the TV-distance bound we get

$$\int_{y>0}(P(y-x)-P(y+x))|y-x|^2\mathrm{d}y \leq \int_{y>2x}(P(y-x)-P(y+x))(y-x)^2\mathrm{d}y + \frac{x^2}{\sigma}. \tag{245}$$

By algebraic manipulation we have

$$\int_{y>2x}\left(\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)-\exp\left(-\frac{(y+x)^2}{2\sigma^2}\right)\right)(y-x)^2\mathrm{d}y \tag{246}$$
$$\leq \int_{y>2x}\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\left(1-\exp\left(-\frac{4xy}{2\sigma^2}\right)\right)(y-x)^2\mathrm{d}y \tag{247}$$
$$\leq \int_{y>2x}\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\frac{4xy}{2\sigma^2}(y-x)^2\mathrm{d}y \tag{248}$$
$$\leq \int_{y>2x}\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)\frac{4x}{\sigma^2}(y-x)^3\mathrm{d}y \tag{249}$$
$$\leq 4\sigma x. \tag{250}$$

□

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*

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle v,u\rangle - \langle q,u\rangle\langle w,u\rangle)f(u)] \leq \sqrt{3}\sigma^2 B(\|p-q\|_2 + \|v-w\|_2). \tag{251}$$

Proof. First of all, by telescope sum we get

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle v,u\rangle - \langle q,u\rangle\langle w,u\rangle)f(u)] \tag{252}$$
$$= \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle v,u\rangle - \langle p,u\rangle\langle w,u\rangle)f(u)] \tag{253}$$
$$+ \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle w,u\rangle - \langle q,u\rangle\langle w,u\rangle)f(u)]. \tag{254}$$

By Hölder inequality we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle v,u\rangle - \langle p,u\rangle\langle w,u\rangle)f(u)] \tag{255}$$
$$\leq \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle p,u\rangle^4\right]^{1/4}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[\langle v-w,u\rangle^4\right]^{1/4}\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}\left[f(u)^2\right]^{1/2} \tag{256}$$
$$\leq \sqrt{3}\sigma^2\|p\|_2\|v-w\|_2 B. \tag{257}$$

Similarly we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2 I)}[(\langle p,u\rangle\langle w,u\rangle - \langle q,u\rangle\langle w,u\rangle)f(u)] \leq \sqrt{3}\sigma^2\|p-q\|_2\|w\|_2 B. \tag{258}$$

□

I.2. Helper Lemmas on Reinforcement Learning

Lemma I.16 (Telescoping or Simulation Lemma, see Luo et al. (2019); Agarwal et al. (2019))**.** *For any policy π and deterministic dynamical model $T, \hat{T}$, we have*

$$V^\pi_{\hat{T}}(s_1) - V^\pi_T(s_1) = \mathbb{E}_{\tau\sim\rho^\pi_T}\left[\sum_{h=1}^H\left(V^\pi_{\hat{T}}(\hat{T}(s_h,a_h)) - V^\pi_{\hat{T}}(T(s_h,a_h))\right)\right]. \tag{259}$$

Lemma I.17 (Policy Gradient Lemma, see Sutton, Barto (2011))**.** *For any policy π_ψ, deterministic dynamical model T and reward function $r(s_h, a_h)$, we have*

$$\nabla_\psi V_T^{\pi_\psi} = \mathbb{E}_{\tau\sim\rho_T^{\pi_\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right)\left(\sum_{h=1}^H r(s_h, a_h)\right)\right] \tag{260}$$

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. ψ in both sides, we have

$$\begin{aligned}
\nabla_\psi V_T^{\pi_\psi} &= \nabla_\psi \int_\tau \Pr[\tau] \sum_{h=1}^H r(s_h, a_h)\, d\tau \\
&= \int_\tau (\nabla_\psi \Pr[\tau]) \sum_{h=1}^H r(s_h, a_h)\, d\tau \\
&= \int_\tau \Pr[\tau](\nabla_\psi \log \Pr[\tau]) \sum_{h=1}^H r(s_h, a_h)\, d\tau \\
&= \int_\tau \Pr[\tau]\left(\nabla_\psi \log \prod_{h=1}^H \pi_\psi(a_h \mid s_h)\right) \sum_{h=1}^H r(s_h, a_h)\, d\tau \\
&= \int_\tau \Pr[\tau]\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right) \sum_{h=1}^H r(s_h, a_h)\, d\tau \\
&= \mathbb{E}_{\tau\sim\rho_T^{\pi_\psi}}\left[\left(\sum_{h=1}^H \nabla_\psi \log \pi_\psi(a_h \mid s_h)\right) \sum_{h=1}^H r(s_h, a_h)\right]
\end{aligned}$$

□