Provable Interactive Learning with Hindsight Instruction Feedback

Dipendra Misra[*1] **Aldo Pacchiano**[*2] **Robert Schapire**[*1]

Abstract

We study interactive learning in a setting where the agent has to generate a response (e.g., an action or trajectory) given a context and an instruction. In contrast, to typical approaches that train the system using reward or expert supervision on response, we study *learning with hindsight labeling* where a teacher provides an instruction that is most suitable for the agent's generated response. This hindsight labeling of instruction is often easier to provide than providing expert supervision of the optimal response which may require expert knowledge or can be impractical to elicit. We initiate the theoretical analysis of *interactive learning with hindsight labeling*. We first provide a lower bound showing that in general, the regret of any algorithm must scale with the size of the agent's response space. Next we study a specialized setting where the underlying instruction-response distribution can be decomposed as a low-rank matrix. We introduce an algorithm called `LORIL` for this setting, and show that it is a no-regret algorithm with the regret scaling with $\sqrt{T}$ and depends on the *intrinsic rank* but does not depend of the agent's response space. We provide experiments showing the performance of `LORIL` in practice for 2 domains.

1. Introduction

Success of a machine learning approach is intimately tied to the ease of getting training data. For example, language models (Brown et al., 2020; Achiam et al., 2023), which are one of the most successful applications of machine learning, are trained on an abundance of language data which is both easy to elicit from non-expert users and is available offline. In contrast, consider the task of a robot following instructions specified by a human user (Misra et al., 2016; Blukis et al., 2019; Myers et al., 2023). It is expensive to collect ground truth robot trajectories making standard imitation learning (IL) approaches (Pomerleau, 1991) expensive to apply, whereas reinforcement learning (RL) (Sutton & Barto, 2018) approaches suffer from high sample complexity. This makes IL and RL– the two most common ways of training agents, expensive in practice. Motivated by the limitations of IL and RL, a line of work has proposed using *hindsight labeling*, where the agent (robot in our example) generates a response (trajectory) given an instruction, and a teacher instead of providing expensive ground truth response, provides the instruction that is suitable for the agent's response (Fried et al., 2018; Nguyen et al., 2021). This reverses the labeling problem to an easier labeling problem, since instructions are typically in a format such as natural language, which can be inexpensively elicited from non-expert users in contrast to robot trajectories. While this approach has been applied empirically, a theoretical understanding remains absent. In this work, we initiate the theoretical understanding of interactive learning from hindsight instruction.

We consider the learning setup illustrated in Figure 1. In this setup, a teacher is teaching an agent to navigate in a virtual home environment. In each round, the world gives an instruction and a context to the agent. The instruction in this case is expressed in natural language. The context is an image that provides information about the environment such as the position, color, and sizes of different objects. The goal of the agent is to generate a trajectory that follows the given instruction. In the beginning, the agent lacks any language understanding and, therefore, cannot generate correct trajectories. We assume access to a teacher that can provide an instruction that best describes the agent's trajectory. This type of feedback can be viewed as a *hindsight instruction*, as it was the correct instruction in hindsight for the trajectory generated by the agent. In each round, the agent generates a trajectory and receives a hindsight instruction from the teacher. This allows the agent to learn a mapping from the instruction space to the trajectory space, which helps improve the agent's policy. We call this learning approach as **L**earning from **H**indsight **I**nstruction (`LHI`).

There are several different approaches for training a

[*]Equal contribution [1]Microsoft Research [2]Broad Institute of MIT and Harvard, Boston University.. Correspondence to: Aldo Pacchiano <pacchian@bu.edu>, Dipendra Misra <dipendra.Misra@microsoft.com>, Robert Schapire <schapire@microsoft.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

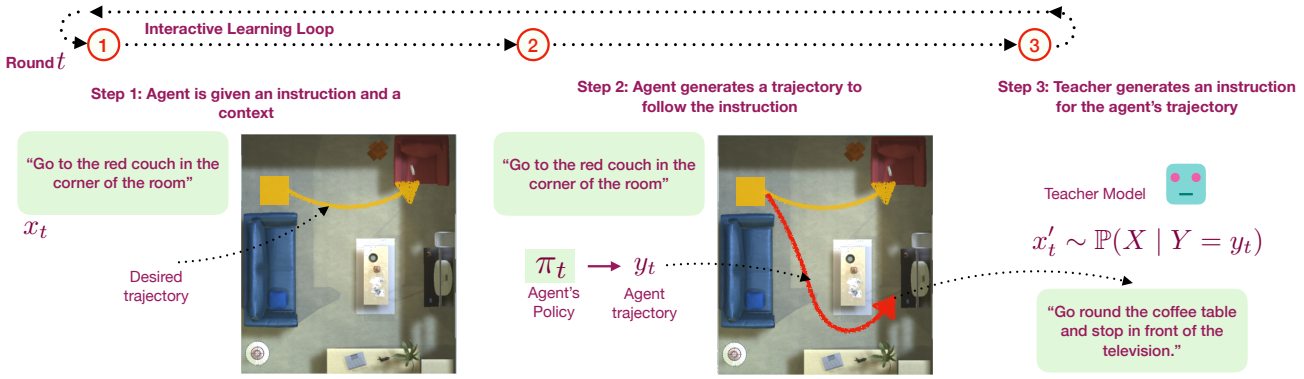

Figure 1. Shows sketch of our interactive **L**earning from **H**indsight **I**nstruction (`LHI`) setting. The agent interacts with the world iteratively. In each round (or time step), the agent is given an instruction x_t and a context s_t. In our case, the context s_t is the house layout. In response, the agent generates a trajectory (response) y_t which is then labeled by a teacher model with an instruction x'_t (*hindsight instruction*). The agent *never* receives any expert response or rewards.

decision-making agent. One of the most commonly used approaches is *imitation learning* (IL) where a teacher provides access to expert demonstrations allowing the agent to learn the right behavior (Ross et al., 2011). For the example in Figure 1, this will require the teacher to be able to understand the agent's action space and dynamics. This often requires domain expertise and can only be provided by expert teachers and may require specialized tools.[1] In contrast, a non-expert user can easily provide an instruction for the red trajectory in Figure 1.

Reinforcement learning (RL) is another widely used approach for training agents that overcomes the expense of collecting expert demonstrations by directly optimizing a reward function that is more user-friendly to provide (Sutton & Barto, 2018). However, RL approaches are less sample efficient than IL approaches making them less suited for real-world settings.In contrast, learning from hindsight instruction uses instruction feedback which is user-friendly and more natural for humans to provide. Further, instruction feedback contains significantly richer information than scalar rewards which can help in reducing the sample complexity compared to RL (Nguyen et al., 2021).

Because of its promise, learning from hindsight labeling has been explored in various applications (Andrychowicz et al., 2017; Fried et al., 2018; Nguyen et al., 2021). However, a principled understanding of this setting remains absent despite these empirical results. In particular, we focus on an interactive learning setting where a teacher trains the agent using hindsight instructions. For these settings, the natural evaluation metric is regret which penalizes the agent for failing to follow a given instruction. A key challenge in designing algorithms for this setting is that the agent has to both *exploit* to follow the given instruction, but also *explore* to improve its understanding capabilities more generally. In this work, we initiate the theoretical study of this setting. We first present a formal interactive learning setting and define a notion of regret. Motivated by natural settings where the teacher is a human user, we assume access to a black box teacher which can generate a sampled instruction given the agent's response but where the agent *does not have access* to the teacher's probability values. The agent is evaluated using a *hidden reward* given by the probability of the teacher labeling the agent's response by the original given instruction.

We first prove a lower bound for this setting showing that in the worst case, the regret bounds for any algorithm will scale polynomially with the size of the agent's response space. In many applications such as our robot navigation example, the agent's response is a trajectory, leading to an exponentially large response space. However, in practice using function approximations and featurization enables generalization in infinitely large spaces. Motivated by this we introduce a low-rank setting where where the agent has access to a feature representation of the response and context and the teacher's distribution admits a low-rank decomposition in this feature space. We introduce an algorithm `LORIL` for this setting and derive regret bounds that scale as $\sqrt{T}$ with the horizon T. Importantly, the regret *does not depend* upon the size of the agent's response or the size of instruction or context space, and instead depends on the rank of the teacher's distribution, which can be significantly smaller in practice.

We evaluate `LORIL` on two tasks. In the first setting, we use a synthetic task where low-rank assumption holds and show

[1] One such commonly used approach is motion capture where a human can record behavior that can be transferred to a humanoid agent but this requires specialized tools.

Protocol 1 Shows the protocol for our setting: Learning from Hindsight Instruction (LHI). The line in blue needs to be implemented by an algorithm implementing the protocol.

for $t = 1, 2, \cdots, T$ **do**
  World presents s_t, x_t possibly adversarially
  Agent generates a response $y_t \in \mathcal{Y}$
  Teacher describes the response $x'_t \sim \mathbb{P}(X \mid y_t, s_t)$
  Evaluate using a hidden reward $r_t = \mathbb{P}(x_t \mid y_t, s_t)$
Return $\sum_{t=1}^{T} r_t$

that LORIL achieves lower regret compared to baselines. In the second setting, we apply LORIL to a setting with natural language instruction and images and show that insights from LORIL help *even when the low-rank assumption does not hold*.

We include a discussion on the related literature after presenting our results in Section 7. The code for all experiments in the paper can be found at https://github.com/microsoft/Intrepid.

2. Preliminary and Overview

We first introduce the mathematical notations before providing an overview of our setup.

Notation. For a given $N \in \mathbb{N}$, we define $[N] = \{1, 2, \cdots, N\}$. For a given countable set $\mathcal{U}$, we denote the set of all distributions over $\mathcal{U}$ by $\Delta(\mathcal{U})$. For a given positive-definite matrix $A \in \mathbb{R}^{d \times d}$, we define the induced norm of a vector $v \in \mathbb{R}^d$ as $\|v\|_A = \sqrt{v^\top A v}$.

Interactive Learning from Hindsight Instruction. We define the space of contexts as $\mathcal{S}$, the space of instructions as $\mathcal{X}$ and the space of all possible agent response as $\mathcal{Y}$. We assume $\mathcal{S}$, $\mathcal{X}$, and $\mathcal{Y}$ to be finite for our analysis but allow them to be arbitrarily large. The finiteness is only a mild assumption in practice, as it still allows us to handle the most common data types. For example, if $\mathcal{X}$ denotes the space of all $m \times n$ RGB images with each pixel taking values in $\{0, 1, \cdots, 255\}$, then $|\mathcal{X}| = 256^{3mn}$ which is an exponentially large but finite value.

The agent interacts with the world repeatedly over T rounds. Protocol 1 shows our learning framework. In the t^{th} round, the world presents a context s_t and an instruction x_t sampled according to a fixed distribution $D_t(\cdot, \cdot \mid x_1, y_1, s_1, \cdots, x_{t-1}, s_{t-1}, y_{t-1})$ that can depend on the past history. Given the instruction and the context, the agent generates a response $y_t \in \mathcal{Y}$. Ideally, we want the agent to generate a response that fulfills the intent of the instruction. After generating the response, the agent receives a *hindsight instruction* x'_t sampled from a fixed conditional distribution model $\mathbb{P}(X \mid Y = y_t, S = s_t)$. This conditional distribution models a *teacher* that provides an instruction that is most appropriate for the agent response. In a typical setting, this teacher will be modeled using a human-in-the-loop.

The agent *does not* have access to $\mathbb{P}(X \mid Y = y_t, S = s_t)$ but can observe a sample from this distribution by generating a response and asking the teacher to label it with an instruction. This is because in a human-in-the-loop setting, we don't have access to the human teacher's distribution.

The teacher model $\mathbb{P}(X \mid Y, S = s_t)$ is in practice highly stochastic since there can be many possible ways to describe instructions for a given response. Further, the space of all possible responses and instructions can be impractically large, necessitating the use of function approximation.

Computing Regret. Given a state s_t and context x_t, the ideal response should maximize the probability of the teacher labeling the response with the right instruction x_t, i.e., the agent should play $y = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(x_t \mid y, s_t)$. We can, therefore, view $\mathbb{P}(x_t \mid y, s_t)$ as a *latent reward* for generating response y. This leads to a natural notion of regret given by:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(\max_{y \in \mathcal{Y}} \mathbb{P}(x_t \mid s_t, y) - \mathbb{P}(x_t \mid s_t, y_t) \right), \tag{1}$$

where s_t, x_t are the context and instruction in round t and y_t is the response generated by the agent.

There can be alternative ways to define regret in Equation 1. Log-probabilities $\log \mathbb{P}(X \mid S, Y)$ may appear more natural to use instead of probabilities, however, the former is unbounded which makes it ill-suited for defining reward. For example, if in the first round, the agent generates a response y_1 for which $\mathbb{P}(x_1 \mid y_1, s_1) = 0$, then the agent's regret is unbounded irrespective of the agent's performance in later rounds. Another choice for reward is the likelihood of the response $\mathbb{P}(y \mid x_t, s_t)$. However, this requires assuming a prior distribution over y which can be hard to realize.

[DM: Discuss relation to LLMs]

3. Lower Bound in the General Case

We first prove that it is impossible to design an algorithm for Protocol 1 with a regret bound that doesn't scale polynomially in the size of plausible responses $|\mathcal{Y}|$.

We introduce the concept of 'stochastic worlds' to prove our lower bound. A stochastic world W consists of a set of instructions, contexts marginal distribution $\mathbb{P}_W(X, S)$ and a conditional distribution of instructions given responses and contexts $\mathbb{P}_W(X|Y, S)$.

When at time t an agent $\mathbb{A}$ interacts via Protocol 1 with a stochastic world W, the world produces an instruction x_t and context s_t sampled from a *time-independent* distribution $\mathbb{P}_W(X, S)$. We use the notation $\mathbb{P}_{W,\mathbb{A}}$ and $\mathbb{E}_{W,\mathbb{A}}$ to denote the measure and expectations over trajectories $(s_1, x_1, y_1, x'_1, \cdots, s_T, x_T, y_T, x'_T)$ resulting from the interaction between $\mathbb{A}$ and world W. We show that for any $K \in \mathbb{N}$, and any algorithm $\mathbb{A}$, there is a stochastic world where the regret of algorithm $\mathbb{A}$ satisfies $\text{Reg}(T) \geq \Omega(\sqrt{KT})$ when $\mathbb{A}$ interacts with W through Protocol 1.

To prove our main result we exhibit a family of stochastic worlds $\{W_i\}$, such that world W_i is defined by context space $S = \{s_o\}$ instruction space $\mathcal{X} = \{A, B\}$, response set $\mathcal{Y} = [K]$ marginal $\mathbb{P}_{W_i}(X, S) = \text{Uniform}((A, s_o), (B, s_o))$ for all $i \in [K]$, and conditional

$$\mathbb{P}_{W_i}(X|y_j) = 1/2 + \sqrt{K/T} \cdot \mathbf{1}(j = i) \cdot (1 - 2 \cdot \mathbf{1}(X = B)).$$

The context distribution is a delta mass around context s_o. In world W_i the optimal response for instruction $X = A$ equals y_i, and the optimal response for instruction $X = B$ is any y_j for $j \neq i$. Any suboptimal decision, regardless of the instruction incurs in regret of order $\sqrt{K/T}$. Our main result is the following.

Theorem 1. *Let $T \geq 256 \log(2e)$ and $K \geq 8e$. For any algorithm, there is at least one stochastic world $W_{\hat{i}}$ such that Reg$(T) \geq \frac{\sqrt{KT}}{8}$ such that with probability at least $1/4e$.*

The proof can be found in Appendix A. Lemma 1 implies the expected regret lower bound,

Corollary 2. *If the conditions of Lemma 1 hold then for any algorithm there exists at least one stochastic world $W_{\hat{i}}$ such that $\overline{Reg}(T) \geq \Omega(\sqrt{KT})$. Where,*

$$\overline{Reg}(T) = \mathbb{E}_{W_{\hat{i}},\mathbb{A}} \left[\sum_{t=1}^{T} \max_{y \in Y} \mathbb{P}(x_t|y, s_o) - \mathbb{P}(x_t|y_t, s_o) \right].$$

The proof of this result can also be found in Appendix A. Theorem 1 shows that for tractable hindsight learning it is necessary to impose structural assumptions on the conditional probabilities $\mathbb{P}(X|Y, S)$. We explore one such assumption in the next sections.

4. Provable Learning in Low-Rank Setting

The analysis in Section 3 shows that the regret scales as $\Omega(\sqrt{|\mathcal{Y}|})$ which makes this an intractable setting when $\mathcal{Y}$ is extremely large. For settings with typically large input or output spaces, it is natural in practice to use function approximation. For example, a trajectory can be encoded using a neural network to a representation that contains the relevant information. In statistical learning theory, significant progress has been made in the study of learning with function approximation (Misra et al., 2020; Sekhari et al., 2021; Foster et al., 2021). In particular, problems with low-rank structures (Agarwal et al., 2020; Jin et al., 2020) have received significant attention due to their abilities to model commonly occurring settings and the success of corresponding algorithms in real-world problems even where the low-rank assumption is violated (Henaff et al., 2022). Motivated by this, we introduce and study a setting where the teacher model $\mathbb{P}(X \mid Y, S)$ admits a low-rank decomposition.

Low-Rank Teacher Model. We consider a specialization of our general setup where the teacher model follows a low-rank decomposition. Formally, we assume that there exists $f^\star : \mathcal{X} \to \mathbb{R}^d$ and $g^\star : \mathcal{Y} \to \mathbb{R}^d$ such that

$$\forall s \in \mathcal{S}, x \in \mathcal{X}, y \in \mathcal{Y}, \qquad \mathbb{P}(x \mid y, s) = f^\star(x)^\top g^\star(y, s),$$

where d is the *intrinsic dimension* of the problem which is much smaller than the size of $\mathcal{S}, \mathcal{X}$, and $\mathcal{Y}$ which can all be infinitely large. We assume that the agent has knowledge of $g^\star$ but does not know $f^\star$.

We assume access to a model class $\mathcal{F}$ to learn $f^\star$. Our goal is to get regret guarantees that do not scale with $|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{S}|$ and instead only depend on the intrinsic dimension d of the problem and the statistical complexity of $\mathcal{F}$.

`LORIL` Algorithm. We present the "Learning in **LO**w-**R**ank models from **I**nstruction **L**abels" algorithm (`LORIL`): for low-rank teacher models in Algorithm 1. The algorithm assumes access to the embedding function $g^\star$ for encoding the agent's response. In practice, such a function can be available either using a pre-trained representation model or by using a self-supervised learning objective such as autoencoding. We discuss some implementation choices later in the experiment section.

`LORIL` implements Protocol 1. In the t^{th} round, the algorithm first computes a maximum likelihood estimation $\hat{f}_t$ of $f^\star$ using the historical data (line 3). We use this to define a policy π_t to generate a response y_t. `LORIL` is based on the principle of optimism under uncertainty. As per this principle, we first compute an appropriate uncertainty measure $b_t(y)$ for a response $y \in \mathcal{Y}$ such that we know with high probability that the true value of a response, i.e., $f^\star(x_t)^\top g^\star(y, s_t)$ lies in $\left[\hat{f}_t(x_t)^\top g^\star(y, s_t) - b_t(y, s_t), \hat{f}_t(x_t)^\top g^\star(y, s_t) + b_t(y, s_t)\right]$ with high probability. As $\hat{f}_t(x_t)^\top g^\star(y, s_t)$ is the current estimate of the value of a response y in the t^{th} round, we can view $b(y, s_t)$ as defining a confidence interval for a given response y and context s_t. Second, we take the action that has the maximum possible value in the confidence

interval, namely:

$$y_t = \arg\max_{y \in \mathcal{Y}} \quad \Big(\underbrace{\hat{f}_t(x_t)^\top g^\star(y, s_t)}_{\text{estimated model value}} + \underbrace{b_t(y, s_t)}_{\text{bonus}} \Big). \tag{2}$$

For low-rank models, we will show that $b_t(y, s_t)$ can be expressed as $\mathcal{O}(\|g^\star(y, s_t)\|_{\widehat{\Sigma}_t^{-1}})$ where $\widehat{\Sigma}_t = \lambda_t \mathbb{I} + \sum_{l=1}^{t-1} g^\star(y_l, s_l) g^\star(y_l, , s_l)^\top$ is a positive definite matrix capturing information about historical data. This can be viewed as a positive definite matrix $\lambda_t \mathbb{I}$ were $\lambda_t > 0$ and a sum of rank one positive semi-definite matrices $g^\star(y_l, s_l) g^\star(y_l, s_l)^\top$ which form a covariance matrix $\sum_{l=1}^{t-1} g^\star(y_l, s_l) g^\star(y_l, s_l)^\top$. The quantity $b_t(y, s_t)$ can be viewed as a bonus in Equation 2 and is known as *elliptic bonus* in the literature and is a frequently appearing quantity in the study of linear models (Abbasi-Yadkori et al., 2011) and low-rank models (Agarwal et al., 2020).[2]

The agent computes the optimistic response $y_t = \pi_t(x_t)$ (line 6) and plays it. In response, the teacher provides a description x'_t (line 7) which is added along with the agent response to the historical data.

Note that the agent never has direct access to $f^\star$ or the true model $\mathbb{P}(X \mid Y, S)$, but only has access through feedback generated by the teacher model and through its knowledge of $g^\star$ and $\mathcal{F}$. Further, the agent is also unaware of the true horizon length T which is often unknown in practice.

Computational Efficiency. The computation of the covariance matrix can be performed easily as can the computation of bonus b_t for a given response. The inverse of the covariance matrix can be computed efficiently in a numerically stable way using the Sherman–Morrison formula (Sherman, 1949). The two main computationally expensive steps in LORIL are maximum-likelihood estimation and solving the optimization in line 5-6. The maximum likelihood estimation is routinely computed for complex function classes such as deep neural networks in practice. However, in this case, the main challenge is in defining a function class $\mathcal{F}$ such that $f(\cdot)^\top g^\star(y, s)$ is a well-defined distribution. This question has been addressed for low-rank models (Zhang et al., 2022) and we expect the same tools to also help here. The optimization in line 5-6 can be trivially solved when $\mathcal{Y}$ is small enough to be enumerated. When $\mathcal{Y}$ is exponentially large, this step can be challenging. One strategy can be to use a proposal distribution $q(y \mid x_t, s_t)$ to generate a set of K responses, and then find the response with maximum objective value in Equation 4. The proposal distribution can be trained by performing MLE on historic data and modeling y autoregressively. However, we leave a computational study with exponentially large $\mathcal{Y}$ for future work. [DM: Discuss a specific algorithm specially in RL framework]

[2]The word elliptic comes from the fact that for a positive definite matrix Σ, the set $\{y \mid y^\top \Sigma^{-1} y \leq 1\}$ denotes an ellipsoid centered at 0.

Algorithm 1 LORIL($g^\star$, $\mathcal{F}$): Learning in **LO**w-**R**ank models from **I**nstruction **L**abels

Require: Response embedding function $g^\star : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}^d$
Require: Model class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}^d\}$
1: Define $\lambda_t = \frac{1}{t}$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Compute MLE estimator $\widehat{f}_t$ using $\{x'_\ell, y_\ell, s_\ell\}_{\ell=1}^{t-1}$.

$$\hat{f}_t = \arg\max_{f \in \mathcal{F}} \sum_{\ell=1}^{t-1} \ln \hat{f}_t(x'_\ell)^\top g^\star(y_\ell, s_\ell)$$

4: Define empirical covariance matrix

$$\widehat{\Sigma}_t = \lambda_t I + \sum_{\ell=1}^{t-1} g^\star(y_\ell, s_\ell) g^\star(y_\ell, s_\ell)^\top \tag{3}$$

5: Define policy for this round

$$\pi_t : x, s \mapsto \operatorname*{argmax}_{y \in \mathcal{Y}} \Big(\hat{f}_t(x)^\top g^\star(y, s) + b_t(y, s) \Big) \tag{4}$$

where

$$b_t(y, s) = C' \left(\sqrt{\log(t|\mathcal{F}|/\delta)} + \sqrt{\lambda_t} B \right) \|g^\star(y, s)\|_{\widehat{\Sigma}_t^{-1}}$$

and C' is determined by Equation 7 in Lemma 1 .
6: Agent generates the response $y_t = \pi_t(x_t, s_t)$
7: Teacher describes the response $x'_t \sim \mathbb{P}(X \mid y_t, s_t)$
8: Evaluate using a *hidden reward* $r_t = \mathbb{P}(x_t \mid y_t, s_t)$
return $\sum_{t=1}^{T} r_t$

5. Theoretical Analysis

Our main result is to show Algorithm 1 satisfies a sublinear regret bound in the realizable setting,

Assumption 1. *[Realizability] The teacher model* $\mathbb{P}(x|y, s) = f^\star(x)^\top g^\star(y, s)$ *satisfies* $f^\star \in \mathcal{F}$ *for a* known *model class* $\mathcal{F}$. *Moreover, all teacher models parametrized by feature maps* $f \in \mathcal{F}$ *are valid distributions, i.e.,* $f^\top(X)^\top g^\star(y, s) \in \Delta(\mathcal{X})$, *for any* $y \in \mathcal{Y}$ *and* $s \in \mathcal{S}$.

We also assume the feature maps in $\mathcal{F}$ are bounded.

Assumption 2. *There exists a constant* $B > 0$ *such that for any* $f \in \mathcal{F}$ *we have* $\sup_{x \in \mathcal{X}} \|f(x)\| \leq B$ *and* $\sup_{y \in \mathcal{Y}, s \in \mathcal{S}} \|g^\star(y, s)\| \leq B$.

Our main theoretical result is a high probability upper bound for the regret of Algorithm 1.

Theorem 3. *[Regret bound of* LORIL*] When Assumption 1 and Assumption 2 hold and $\lambda_t = 1/t$ the regret of Algorithm 1 satisfies*

$$Reg(T) = \mathcal{O}\Big(B\sqrt{Td\log(1+TB)} + \sqrt{Td\log(T|\mathcal{F}|/\delta)\log(1+TB)}\Big)$$

with probability at least $1-3\delta$ for all $T \in \mathbb{N}$.

The analysis of Algorithm 1 in Theorem 3 is based on the principle of optimism. For any $(s,x,y) \in \mathcal{S}\times\mathcal{X}\times\mathcal{Y}$ thequantity $\hat{f}_t(x)^\top g^\star(y,s) + b_t(y,s)$ can be understood as an estimator for the value of $\mathbb{P}(x|y,s)$ where the corrective bonus $b_t(y,s)$ takes into account the accuracy of the empirical estimator $\widehat{\mathbb{P}}_t(x|y,s) = \hat{f}_t(x)^\top g^\star(y,s)$. By defininingthe bonus function $b_t : \mathcal{Y} \to \mathbb{R}$ as an appropriately scaled multiple of $\|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}}$ it overestimates $\mathbb{P}(x|y,s)$. That is, for all $s \in \mathcal{S}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $t \in \mathbb{N}$,

$$\mathbb{P}(x|y,s) \le \widehat{\mathbb{P}}_t(x|y,s) + b_t(y,s) \tag{5}$$

with probability at least $1-2\delta$. When Equation 5 is satisfied the policy definition of line 5 immediately implies that

$$\max_{y\in\mathcal{Y}} \mathbb{P}(x|y,s_t) \le \max_{y\in\mathcal{Y}} \widehat{\mathbb{P}}_t(x|y,s_t) + b_t(y,s_t) = \widehat{\mathbb{P}}_t(x|y_t,s_t) + b_t(y_t,s_t) \tag{6}$$

To prove Equation 5 we develop the following supporting result,

Lemma 1. *When Assumption 1 and Assumption 2 hold, then with probability at least $1-2\delta$ we have:*

$$\sup_{x\in\mathcal{X}} \left\|f^\star(x) - \widehat{f}_t(x)\right\|_{\widehat{\Sigma}_t} \le C'\left(\sqrt{\log(t|\mathcal{F}|/\delta)} + \sqrt{\lambda_t}B\right) \tag{7}$$

for all $t \in \mathbb{N}$ simultaneously.

This result provides a bound for the maximum error in the estimation of $f^\star(x)$ as measured by the data norm $\|\cdot\|_{\widehat{\Sigma}_t}$. It will prove crucial in bounding the error of the empirical models $\widehat{\mathbb{P}}_t(x|y,s)$. The detailed version of this result can its proof can be found in Lemma 2 in Appendix B.

We denote as $\mathcal{E}$ the event that Equation 7 holds. In this case, we can upper bound the prediction error of the empirical model $\widehat{\mathbb{P}}_t(x|y,s) = \hat{f}_t(x)^\top g^\star(y,s)$ for all $(s,x,y) \in \mathcal{S}\times\mathcal{X}\times\mathcal{Y}$.

$$\begin{aligned}\left|\widehat{\mathbb{P}}_t(x|y,s) - \mathbb{P}(x|y,s)\right| &= \left|\left(f^\star(x) - \hat{f}_t(x)\right)^\top g^\star(y,s)\right| \\ &\overset{(i)}{\le} \left\|f^\star(x) - \hat{f}_t(x)\right\|_{\widehat{\Sigma}_t} \|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}} \\ &\overset{(ii)}{\le} b_t(y,s)\end{aligned} \tag{8}$$

where (i) holds because for all $v, w \in \mathbb{R}^d$ and invertible $\Sigma \in \mathbb{R}^{d\times d}$, the Cauchy-Schwartz inequality implies $\langle v, w\rangle = \langle \Sigma^{1/2}v, \Sigma^{-1/2}w\rangle \le \|v\|_\Sigma \|w\|_{\Sigma^{-1}}$ and (ii) by upper bounding $\left\|f^\star(x) - \hat{f}_t(x)\right\|_{\widehat{\Sigma}_t}$ using the RHS of Equation 7.

These are the necessary ingredients to finalize our sketch of Theorem 3. When $\mathcal{E}$ holds the following inequalities are satisfied,

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^T \left(\mathbb{P}(x_t \mid \pi^\star(x_t), s_t) - \mathbb{P}(x_t \mid \pi_t(x_t), s_t)\right) \\ &\overset{(a)}{\le} \sum_{t=1}^T \hat{f}_t(x)^\top g^\star(y_t,s_t) - f^\star(x_t)^\top g^\star(y_t,s_t) + b_t(y_t,s_t) \\ &\overset{(b)}{\le} \sum_{t=1}^T 2b_t(y_t,s_t)\end{aligned}$$

where (a) is a consequence of Optimism (Equation 6). And inequality (b) of the prediction error bound from Equation 8. What we have managed to achieve at this point is to upper bound the regret by a sum of estimation errors along the features of the responses played by the algorithm at each time-step. Finally, substituting the definition of the bonus terms and invoking a standard sum of inverse norms bound from the linear bandits literature (see for example Proposition 3 in (Pacchiano et al., 2021) and Proposition 7 in Appendix C)

$$\begin{aligned}\sum_{t=1}^T b_t(y_t) &\le \mathcal{O}\left(\sqrt{\log(T|\mathcal{F}|/\delta)}\sum_{t=1}^T \|g^\star(y_t,s_t)\|_{\widehat{\Sigma}_t^{-1}}\right) \\ &\le \mathcal{O}\left(\sqrt{Td\log(T|\mathcal{F}|/\delta)\log(1+TB)}\right)\end{aligned}$$

This finalizes the proof sketch of Theorem 3. These results rely on the assumption that $g^\star$ is known. Removing that assumption yields a substantially harder problem as it makes it more difficult to leverage the linearity structure. Although this scenario can be dealt with by deriving algorithms and bounds depending on statistical capacity measures such as the eluder dimension (Russo & Van Roy, 2013) for the combined $f(x)^\top g(y,s)$ model class we leave the derivation of a sharper analysis of this setting for future work.

6. Empirical Study

We evaluate LORIL in two settings. The first is a synthetic task that satisfies the low-rank teacher setting and all our assumptions and is designed to provide a proof of concept of LORIL. The second setting is a grounded setting with real images, natural language instructions, and where the teacher model is not low-rank. Our goal with these experiments is not to present challenging settings for exploration, but to show how various components of LORIL can be implemented empirically. Our second experiment also evaluates whether insights from LORIL carry over to more realistic settings even where our assumptions are violated.

6.1. Evaluating on a Synthetic Task

Environment. For a given intrinsic dimension d, instruction size $|\mathcal{X}|$ and response size $|\mathcal{Y}|$, we randomly initialize two matrices $F \in \mathbb{R}^{|\mathcal{X}| \times d}$ and $G \in \mathbb{R}^{d \times |\mathcal{Y}|}$. We ignore the context in this setting by defining $\mathcal{S}$ as a singleton $\{s_0\}$. We define $\mathcal{X} = [|\mathcal{X}|]$ and $\mathcal{Y} = [|\mathcal{Y}|]$ and so an instruction x and a response y are positive integers. We define these matrices by first initializing them with values sampled iid with standard Gaussian distribution. We then take their exponent and divide by a temperature coefficient τ. We then normalize F and G row-wise such that $FG \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$ is a stochastic matrix whose columns sum to 1. For a given $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we view the matrix entry $(FG)_{xy}$ as denoting the value of the teacher distribution $\mathbb{P}(X = x \mid Y = y, S = s_0)$. We can view the x^{th} row of F and the y^{th} column of G as denoting $f^\star(x)$ and $g^\star(y)$ respectively.

Baselines. We evaluate the following baselines. *Random*: the agent takes uniformly random actions. *ϵ-Greedy*: the agent performs maximum-likelihood estimation on the historic data to learn an estimate $\hat{f}_t$ similar to LORIL; however, unlike LORIL, the exploration is not performed using elliptic bonus but using ϵ-greedy, where with ϵ probability a random action is taken and with the remaining probability, we take the greedy action $\arg\max_{y \in \mathcal{Y}} \hat{f}_t(x_t)^\top g^\star(y)$. *Greedy*: This is same as ϵ-Greedy with $\epsilon = 0$ and only exploits based on historic data. We tune the hyperparameters λ and C' for LORIL and ϵ for ϵ-greedy using grid search.

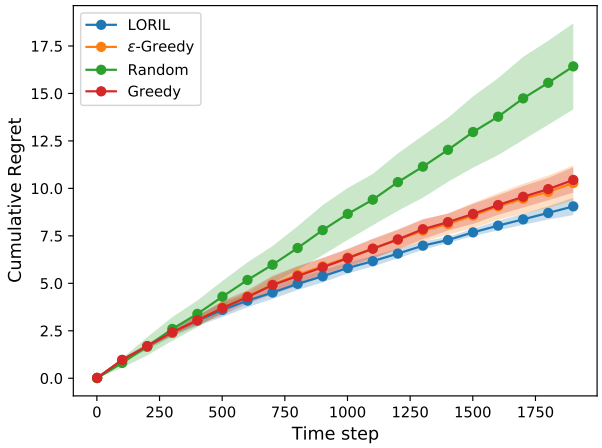

Figure 2. Comparison of LORIL against baselines on the controlled task. We run each baseline 3 times and report the average. The shaded areas show the standard deviation.

Model and Optimization. We model $f \in \mathcal{F}$ as $f(x) = \left(\frac{\exp(\theta_{xi})}{\sum_{x' \in \mathcal{X}} \exp(\theta_{x'i})}\right)_{i=1}^{d}$, where $(\theta_{xi})_{x \in \mathcal{X}, i \in [d]}$ are the parameters that we train. For any $f \in \mathcal{F}$, we can verify that $f(x)^\top g^\star(y)$ is a valid conditional distribution over x given y. We perform maximum likelihood estimation using Adam optimization.

Results. Figure 2 shows cumulative regret over time steps for LORIL and baselines. We ran each experiment 3 times with different seeds. We select hyperparameters for each algorithm based on the mean final regret. We can see that LORIL performs better than all baselines achieving the best regret which is 12.3% smaller than than the next best baseline. Improvements over the greedy baseline show that exploration helps, whereas improvements over ϵ-greedy show that using elliptic bonus for exploration provides better regret bounds.

6.2. Evaluation on an Image Selection Task

We evaluate LORIL on an image classification task where the true model does not admit a low-rank decomposition. In reinforcement learning, it has been found that using the elliptic bonus for exploration is helpful in real-world settings where low-rank assumption doesn't hold (Henaff et al., 2022). Our goal in this subsection is to test if a similar result holds for our setting.

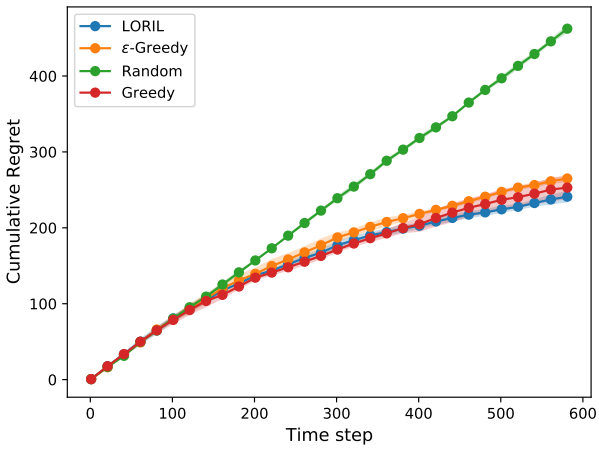

Figure 3. Results on the image classification task. We run each baseline 5 times and report the average performance. The shaded areas show the standard deviation.
[DM: add a figure of the model]

Environment. The instruction space $\mathcal{X}$ is in natural language where for a given $x \in \mathcal{X}$, we denote the i^{th} token by x_i. The agent has an action space with $|\mathcal{Y}| = K$ actions. In each round, the world assigns an image of an object to each action, and the agent is given a natural language instruction describing the image that the agent should select. The agent has a non-trivial context $s \in \mathcal{S}$ that contains the identity of

the object assigned to each action in the given round. No two actions have the same image but the image associated with each action can change across rounds. We sample an object by picking an object of a given type and a given color from a set of types and colors. The instruction space $\mathcal{X}$ is in natural language as in our motivating example in Figure 1. We use a set of templates to generate instructions for describing a given image using the object type and color.

Model. We model a function class $\mathcal{H} = \{h : \mathcal{Y} \times \mathcal{S} \to \Delta(\mathcal{X})\}$ using a deep neural network. Given an action y and context s, we encode the image associated with the action with an encoding $g^\star(y, s) \in \mathbb{R}^d$. We model $g^\star$ as a 3-layer convolutional neural network with LeakyReLu activation. In this setting, we don't assume that the environment provides the $g^\star$. Instead, we train $g^\star$ using an autoencoder objective and a set of offline images sampled from the environment. Alternatively, we could have used a pre-trained image representation model such as ResNet (He et al., 2016) or CLIP (Radford et al., 2021).

We use the encoding $g^\star(y, s)$ to generate a distribution over texts $x = (x_1, \cdots, x_n)$ using a two-layer Gated Recurrent Unit (GRU). Specifically, we apply a fully connected layer to $g^\star(y, s)$ to reshape it to an appropriate size and use it to initialize the hidden state of the GRU for all layers.

Results. Figure 3 shows the results. Similar to our previous experiment, `LORIL` performs better than baselines, achieving 4.8% less regret than the next best baseline, even though the setting does not admit a low-rank structure. We also note that these tasks were not designed to present a challenging scenario for exploration, and consequently the gains relative to baselines are smaller.

7. Related Work

Provably-efficient Interactive Learning. The ubiquitous nature of interactive learning has resulted in significant attention devoted to its theoretical understanding. Protocol 1 superficially resembles a contextual bandit problem but stands in contrast with the scenario where the learner receives a (possibly noisy) reward signal after taking an action in a given context. The key difference is that while in the contextual bandit setting the feedback equals an unbiased sample of the reward corresponding to the arm and context, in our setting the feedback is produced from a conditional distribution of instructions that does not immediately relate to the reward. Thus, it is not possible to immediately adapt a contextual bandit algorithm to provide regret bounds for Protocol 1. There is a vast literature dedicated to developing sublinear regret algorithms for contextual bandit problems. Early efforts to incorporate contextual information into bandit problems led to the development of algorithms such as LinUCB (Chu et al., 2011), OFUL (Abbasi-Yadkori et al., 2011), and Linear Thompson Sampling (Agrawal & Goyal, 2013), for the setting when there is a linear relationship between the context and the reward. A long line of work has also focused on studying guarantees for imitation learning (Ross et al., 2011; Rashidinejad et al., 2021), and policy optimization (Kearns & Singh, 2002; Auer & Ortner, 2006; Azar et al., 2017; Foster et al., 2021). More recently, there has been a focus on developing statistically efficient RL algorithms with function approximation (Misra et al., 2020; Jin et al., 2021; Foster et al., 2021). Our work focuses on provable learning similar to these methods and uses similar tools for analysis, but focuses on a novel learning setting with a different type of feedback than IL and RL.

Low-rank Interactive Learning. Low-rank models have been studied in bandit settings (Abbasi-Yadkori et al., 2011), contextual bandit settings (Chu et al., 2011), and in more general multi-step reinforcement learning (Jin et al., 2020; Agarwal et al., 2020). One of the appeals of low-rank models is that they can generalize tabular MDPs and provide a way to study function approximation settings which is standard in empirical studies. This is also our motivation for studying low-rank models. Further, low-rank models are one of the most expressive settings for which both statistically and computationally efficient algorithms exist.

Learning using Hindsight Feedback. Several different works have found it advantageous to use hindsight feedback to convert a failed example into a positive example by relabeling it with a different goal (or in our case instruction) (Andrychowicz et al., 2017; Li et al., 2020; Nair et al., 2018). These approaches typically solve goal-conditioned RL where a failed trajectory is labeled with its final state as the goal. However, these approaches focus on empirical performance and do not provide regret bounds.[DM: cite Aviv's paper]

Instruction Following. The task of developing agents that can follow natural language instructions has received significant attention since the early days of AI (Winograd, 1972). Several approaches have developed methods that train these systems using imitation learning (Mei et al., 2016) and reinforcement learning (Misra et al., 2018; Hill et al., 2021). Training agents with hindsight instruction labeling has been previously explored for instruction following in (Nguyen et al., 2021; Fried et al., 2018). The main focus of these results is on empirical performance and they either provide no theoretical analysis, or in the case of (Nguyen et al., 2021) only provide asymptotic analysis. In contrast, we provide the first finite-sample regret bounds for learning from instruction labeling.

8. Conclusion

In this work we define a *formal* interactive learning setup for hindsight instruction and initiate its theoretical understanding. Among other things we present a lower bound indicating that hindisght instruction learning in the general case can be statistically intractable, thus implying the necessity of imposing structural conditions for statistically efficient learning with hindsight feedback. We present an algorithm LORIL that has no-regret when the underlying teacher distribution has low-rank. The regret of LORIL scales $\widetilde{\mathcal{O}}\left(\sqrt{T}\right)$ with the horizon and only depends on the rank of the distribution and does not depend on the size of the agent's response space or instruction space. We finalize our work with an experimental demonstration of LORIL in a variety of synthetic and grounded scenarios. This work represents a first exploration of the hindsight instruction setup and therefore many exciting research directions remain open. Chief among them is to design provably efficient algorithms for hindsight instruction under less restrictive function approximation assumptions and that are also computationally efficient. We foresee that the algorithmic framework introduced by the Decision Estimation Coefficient literature (Foster et al., 2023; 2021) can serve as the basis of the development of algorithms for hindsight instruction that are both computationally tractable and statistically efficient and that can lead to practical impact in scenarios such as training language models and robotics.

Impact Statement

This paper presents an interactive learning algorithm that learns from hindsight instructions. The main contributions of this paper are theoretical, however, algorithmic principles from our work can be useful in various empirical studies. A key application can be in training embodied systems using feedback provided by a human or a language model. An important thing to keep in mind is ensuring safety and privacy of any human in the loop during the training process. Further, if a language model is used to generate hindsight instructions, then care must be taken to ensure that hallucinations and implicit bias in the model does not lead to undesired behavior in the embodied agent or robot. While these questions are important, they are orthogonal to the focus of our study.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Blukis, V., Terme, Y., Niklasson, E., Knepper, R. A., and Artzi, Y. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Foster, D. J., Golowich, N., and Han, Y. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., and Darrell, T. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henaff, M., Raileanu, R., Jiang, M., and Rocktäschel, T. Exploration via elliptical episodic bonuses. *Advances in Neural Information Processing Systems*, 35:37631–37646, 2022.

Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., and Clark, S. Grounded language learning fast and slow. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=wpSWuz_hyqA.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Li, A., Pinto, L., and Abbeel, P. Generalized hindsight for reinforcement learning. *Advances in neural information processing systems*, 33:7754–7767, 2020.

Mei, H., Bansal, M., and Walter, M. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., and Artzi, Y. Mapping instructions to actions in 3D environments with visual goal prediction. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Misra, D. K., Sung, J., Lee, K., and Saxena, A. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, 2016.

Myers, V., He, A. W., Fang, K., Walke, H. R., Hansen-Estruch, P., Cheng, C.-A., Jalobeanu, M., Kolobov, A., Dragan, A., and Levine, S. Goal representations for instruction following: A semi-supervised language interface to control. In *Conference on Robot Learning*, pp. 3894–3908. PMLR, 2023.

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6292–6299. IEEE, 2018.

Nguyen, K. X., Misra, D., Schapire, R., Dudík, M., and Shafto, P. Interactive learning from activity description. In *International Conference on Machine Learning*, pp. 8096–8108. PMLR, 2021.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2827–2835. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/pacchiano21a.html.

Pomerleau, D. A. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.,

et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Sekhari, A., Dann, C., Mohri, M., Mansour, Y., and Sridharan, K. Agnostic reinforcement learning with low-rank mdps and rich observations. *Advances in Neural Information Processing Systems*, 34:19033–19045, 2021.

Sherman, J. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of mathematical statistics*, 20(4):621, 1949.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Winograd, T. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022.

A. Lower Bound Proofs

Theorem 1. *Let $T \geq 256\log(2e)$ and $K \geq 8e$. For any algorithm, there is at least one stochastic world $W_{\hat{i}}$ such that Reg$(T) \geq \frac{\sqrt{KT}}{8}$ such that with probability at least $1/4e$.*

Proof. Throughout the proof we'll use the notation $\epsilon = \sqrt{K/T}$ so that problem W_i has distribution $\mathbb{P}(X, S) = \text{Uniform}((A, s_o), (B, s_o))$ and

$$\mathbb{P}_{W_i}(X|y_j) = 1/2 + \epsilon \cdot \mathbf{1}(j = i) \cdot (1 - 2 \cdot \mathbf{1}(X = B)).$$

Let's start by defining the empty problem $W_\emptyset$ as

$$\mathbb{P}(X, S) = \text{Uniform}((A, s_o), (B, s_o))$$
$$\mathbb{P}(A|y_i) = \mathbb{P}(B|y_i) = 1/2, \quad \forall i$$

Let's consider an arbitrary algorithm for $\mathbb{A}$ for learning with hindsight labeling and consider its interaction with problem W. We'll use the notation $\mathbb{P}_{W,\mathbb{A}}$ and $\mathbb{E}_{W,\mathbb{A}}$ to denote the measure and expectations induced by problem W and algorithm $\mathbb{A}$.

First, we consider algorithm $\mathbb{A}$'s interaction with problem $W_\emptyset$.

Let's consider an arbitrary algorithm for $\mathbb{A}$ for learning with hindsight labeling and its interactions with $W_\emptyset$.

Let $n_i(T) = \sum_{t=1}^{T} \mathbf{1}(y_t = i)$ denote the (random) number of times the learner selected $y_t = i$ from time 1 to T.

We'll use the notation $\ell_1, \cdots, \ell_K$ to denote the ordering of indices in $[K]$ such that,

$$\mathbb{E}_{\mathbb{P}_{W_\emptyset,\mathbb{A}}}[n_{\ell_1}(T)] \leq \cdots \leq \mathbb{E}_{\mathbb{P}_{W_\emptyset,\mathbb{A}}}[n_{\ell_K}(T)].$$

Notice that for all $j \leq \lfloor K/2 \rfloor$,

$$\mathbb{E}_{\mathbb{P}_{W_\emptyset,\mathbb{A}}}[n_{\ell_1}(T)] \leq \frac{2T}{K}, \forall \ell \leq \lfloor K/2 \rfloor \tag{9}$$

Let's start by noting that for any W_i, the KL distance between $\mathbb{P}_{W_\emptyset,\mathbb{A}}$ and $\mathbb{P}_{W_i,\mathbb{A}}$ for $i \in \{\ell_j\}_{j\in[\lfloor K/2 \rfloor]}$ satisfies the bound

$$\begin{aligned}
\text{KL}\left(\mathbb{P}_{W_\emptyset,\mathbb{A}} \parallel \mathbb{P}_{W_i,\mathbb{A}}\right) &= \mathbb{E}_{W_\emptyset,\mathbb{A}}\left[\sum_{t=1}^{T} \text{KL}(\mathbb{P}_{W_\emptyset}(X'|Y_t) \parallel \mathbb{P}_{W_i}(X'|Y_t))\right] \\
&= \mathbb{E}_{W_\emptyset,\mathbb{A}}\left[\sum_{t=1}^{T} \text{KL}(\mathbb{P}_{W_\emptyset}(X'|Y_t = i) \parallel \mathbb{P}_{W_i}(X'|Y_t = i))\mathbf{1}(Y_t = i)\right] \\
&= \mathbb{E}_{W_\emptyset,\mathbb{A}}[n_i(T)]\,\text{KL}\left(\text{Ber}(1/2) \parallel \text{Ber}(1/2 - \epsilon)\right) \\
&\overset{(i)}{\leq} \mathcal{O}\left(\frac{T\epsilon^2}{K}\right)
\end{aligned}$$

Where (i) is implied by inequality 9 for all ℓ_j with $j \in [\lfloor K/2 \rfloor]$.

Define $\mathcal{E}_i = \left\{\sum_{t=1}^{T} \mathbf{1}(Y_t = i, X_t = A) + \mathbf{1}(Y_t \neq i, X_t = B) \geq \frac{7T}{8}\right\}$. When interacting with world W_i the event $\mathcal{E}_i$ corresponds to the event where $\mathbb{A}$ makes the right decisions in at least $3T/4$ time-steps.

The complement event $\mathcal{E}_i^c = \{\sum_{t=1}^{T} \mathbf{1}(Y_t = i, X_t = A) + \mathbf{1}(Y_t \neq i, X_t = B) < \frac{7T}{8}\}$ corresponds to the event where $\mathbb{A}$ makes the correct decisions in at most $7T/8$ time-steps.

By the Huber-Bretagnolle inequality, all $i \in \{\ell_j\}_{j\in[\lfloor K/2 \rfloor]}$ satisfy

$$\begin{aligned}
\mathbb{P}_{W_\emptyset,\mathbb{A}}(\mathcal{E}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) &\geq \exp\left(-\text{KL}\left(\mathbb{P}_{W_\emptyset,\mathbb{A}} \parallel \mathbb{P}_{W_i,\mathbb{A}}\right)\right) \\
&\geq \mathcal{O}\left(-T\epsilon^2/K\right).
\end{aligned}$$

since $\epsilon = \sqrt{K/T}$, we have

$$\mathbb{P}_{W_\emptyset,\mathbb{A}}(\mathcal{E}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \geq 1/e \text{ for all } i \in \{\ell_j\}_{j\in[\lfloor K/2\rfloor]}.$$

Let $n_X(T) = \sum_{t=1}^T \mathbf{1}(X_t = X)$ denote the (random) number of times the learner selected $X_t = X$ for $X \in \{A, B\}$ from time 1 to T.

We now define $\mathcal{E}_{\text{good}} = \{n_B(T) \leq \frac{5T}{8}\}$. If $T \geq 256\log(2e)$ Proposition 4 applied with $\alpha = 1/8$ and $\delta = 1/2e$ implies that,

$$\mathbb{P}_{W_\emptyset}(\mathcal{E}_{\text{good}}) \geq 1 - 1/2e$$

Define $\widetilde{\mathcal{E}}_i = \mathcal{E}_i \cap \mathcal{E}_{\text{good}}$. For all $i \in \{\ell_j\}_{j\in[\lfloor K/2\rfloor]}$,

$$\begin{aligned}
1/e \leq \mathbb{P}_{W_\emptyset,\mathbb{A}}(\mathcal{E}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) &= \mathbb{P}_{W_\emptyset,\mathbb{A}}(\mathcal{E}_i \cap \mathcal{E}_{\text{good}}^c) + \mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \\
&\leq \mathbb{P}_{W_\emptyset,\mathbb{A}}(\mathcal{E}_{\text{good}}^c) + \mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \\
&\leq 1/2e + \mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c)
\end{aligned}$$

And therefore $\mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \geq 1/2e$. Thus,

$$\sum_{i\in\{\ell_j\}_{j\in[\lfloor K/2\rfloor]}} \mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) + \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \geq \lfloor K/2\rfloor/2e \tag{10}$$

Notice that when $\widetilde{\mathcal{E}}_i$ it follows that $\sum_{t=1}^T \mathbf{1}(Y_t = i, X_t = A) + \frac{5T}{8} \geq \sum_{t=1}^T \mathbf{1}(Y_t = i, X_t = A) + \mathbf{1}(Y_t \neq i, X_t = B) \geq \frac{7T}{8}$ and therefore $\sum_{t=1}^T \mathbf{1}(Y_t = i, X_t = A) \geq \frac{T}{4}$. This in turn implies that when $\widetilde{\mathcal{E}}_i$ holds then $\sum_{t=1}^T \mathbf{1}(Y_t \neq i, X_t = A) + \mathbf{1}(X_t = B) \leq \frac{3T}{4}$.

Notice that $\widetilde{\mathcal{E}}_i \cap \widetilde{\mathcal{E}}_j = \emptyset$ for all $i \neq j$. This is because for all $j \neq i$, when $\widetilde{\mathcal{E}}_i$ holds,

$$\sum_{t=1}^T \mathbf{1}(Y_t = j, X_t = A) + \mathbf{1}(Y_t \neq j, X_t = B) \leq \sum_{t=1}^T \mathbf{1}(Y_t \neq i, X_t = A) + \mathbf{1}(X_t = B) \leq \frac{3T}{4} < \frac{7T}{8}$$

Since $\widetilde{\mathcal{E}}_i \cap \widetilde{\mathcal{E}}_j = \emptyset$ the sum $\sum_{i\in\{\ell_j\}_{j\in[\lfloor K/2\rfloor]}} \mathbb{P}_{W_\emptyset,\mathbb{A}}(\widetilde{\mathcal{E}}_i) \leq 1$. Thus Equation 10 implies,

$$\sum_{i\in\{\ell_j\}_{j\in[\lfloor K/2\rfloor]}} \mathbb{P}_{W_i,\mathbb{A}}(\mathcal{E}_i^c) \geq \lfloor K/2\rfloor/2e - 1$$

And therefore there is an index $\hat{i} \in \{\ell_j\}_{j\in[\lfloor K/2\rfloor]}$ such that,

$$\mathbb{P}_{W_{\hat{i}},\mathbb{A}}(\mathcal{E}_{\hat{i}}^c) \geq \frac{1}{2e} - \frac{1}{\lfloor K/2\rfloor} \overset{(i)}{\geq} \frac{1}{4e}.$$

where inequality (i) holds because $K \geq 8e$.

when $\mathcal{E}_{\hat{i}}^c = \{\sum_{t=1}^T \mathbf{1}(Y_t = i, X_t = A) + \mathbf{1}(Y_t \neq i, X_t = B) < \frac{7T}{8}\}$ holds,

$$\sum_{t=1}^T \mathbf{1}(Y_t \neq i, X_t = A) + \mathbf{1}(Y_t = i, X_t = B) \geq \frac{T}{8} \tag{11}$$

also holds. When $\mathcal{E}_{\hat{i}}^c$ is satisfied and therefore Equation 11 is satisfied, the regret can be lower bounded by $\epsilon T/8 = \frac{\sqrt{KT}}{8} = \Omega(\sqrt{KT})$ since $\epsilon = \sqrt{K/T}$. We conclude that,

$$\text{Reg}(T) \geq \frac{\sqrt{KT}}{8}$$

with probability $\mathbb{P}_{W_{\hat{i}},\mathbb{A}}(\mathcal{E}_{\hat{i}}^c) \geq \frac{1}{4e}$. □

Corollary 2. *If the conditions of Lemma 1 hold then for any algorithm there exists at least one stochastic world $W_{\hat{i}}$ such that $\overline{Reg}(T) \geq \Omega(\sqrt{KT})$. Where,*

$$\overline{Reg}(T) = \mathbb{E}_{W_{\hat{i}},\mathbb{A}}\left[\sum_{t=1}^{T} \max_{y\in Y} \mathbb{P}(x_t|y,s_o) - \mathbb{P}(x_t|y_t,s_o)\right].$$

Proof. Lemma 1 implies there exists at least one problem $W_{\hat{i}}$ such that $\text{Reg}(T) \geq \frac{\sqrt{KT}}{8}$ with probability at least $1/4e$. Let's call this event $\mathcal{E}$. Therefore since $\sum_{t=1}^{T} \max_{y\in Y} \mathbb{P}(X_t|y,s_o) - \mathbb{P}(X_t|Y_t,s_o) \geq 0$ with probability one,

$$\overline{\text{Reg}}(T) = \mathbb{E}_{W_{\hat{i}},\mathbb{A}}\left[\sum_{t=1}^{T} \max_{y\in Y} \mathbb{P}(X_t|y,s_o) - \mathbb{P}(X_t|Y_t,s_o)\right] \geq \mathbb{P}_{W_{\hat{i}},\mathbb{A}}(\mathcal{E}) \cdot \frac{\sqrt{KT}}{8} \geq \Omega(\sqrt{KT}).$$

□

Proposition 4. *Let $\delta \in (0,1)$, $\alpha \in (0,1/2)$ and $\{X_i\}_{i=1}^{T}$ be T i.i.d. random variables sampled from $\text{Ber}(1/2)$. It $T \geq \frac{4}{\alpha^2}\log(1/\delta)$ then $\sum_{i=1}^{T} X_i \leq (\frac{1}{2}+\alpha)T$ with probability at least $1-\delta$. Similarly if $T \geq \frac{4}{\alpha^2}\log(1/\delta)$ then $\sum_{i=1}^{T} X_i \geq (\frac{1}{2}-\alpha)T$ with probability at least $1-\delta$.*

Proof. Let $\widehat{S} = \sum_{i=1}^{T} X_i$ be the sum of the outcomes $\{X_i\}_{i\in[T]}$. Hoeffding inequality implies

$$\widehat{S} - T/2 \leq 2\sqrt{T\log(1/\delta)}$$

with probability at least $1-\delta$. Thus, as long as $2\sqrt{T\log(1/\delta)} \leq \alpha T$, (i.e. $\frac{4}{\alpha^2}\log(1/\delta) \leq T$) we have $\widehat{S} \leq (\frac{1}{2}+\alpha)T$.

The reverse inequality can be derived using the same argument applied to the inequality $T/2 - \widehat{S} \leq 2\sqrt{T\log(1/\delta)}$ with probability at least $1-\delta$.

□

B. Regret Bounds for LORIL in low-rank distribution setting

In this section, we provide a regret bound for LORIL. We first enumerate the assumptions.

Assumption 1. *[Realizability] The teacher model $\mathbb{P}(x|y,s) = f^\star(x)^\top g^\star(y,s)$ satisfies $f^\star \in \mathcal{F}$ for a* known *model class $\mathcal{F}$. Moreover, all teacher models parametrized by feature maps $f \in \mathcal{F}$ are valid distributions, i.e., $f^\top(X)^\top g^\star(y,s) \in \Delta(\mathcal{X})$, for any $y \in \mathcal{Y}$ and $s \in \mathcal{S}$.*

Assumption 2. *There exists a constant $B > 0$ such that for any $f \in \mathcal{F}$ we have $\sup_{x\in\mathcal{X}} \|f(x)\| \leq B$ and $\sup_{y\in\mathcal{Y},s\in\mathcal{S}} \|g^\star(y,s)\| \leq B$.*

Let us consider the t^{th} round. The empirical covariance matrix is given by $\widehat{\Sigma}_t$ where

$$\widehat{\Sigma}_t = \sum_{l=1}^{t-1} g^*(y_l,s_l)g^\star(y_l,s_l)^\top + \lambda_t\mathbb{I}.$$

for a regularizer value $\lambda_t > 0$. It is easy to verify that $\widehat{\Sigma}_t$ is a positive definite matrix since $\lambda_t\mathbb{I}$ is a positive definite matrix and $C_t = \sum_{l=1}^{t-1} g^*(y_l,s_l)g^\star(y_l,s_l)^\top$ is a positive semidefinite matrix as it is a symmetric matrix and for any $v \in \mathbb{R}^d$ we have $v^\top C_t v = \sum_{l=1}^{t-1} v^\top g^\star(y_l,s_l)g^\star(y_l,s_l)^\top v = \sum_{l=1}^{t-1} \left\|g^\star(y_l,s_l)^\top v\right\|_2^2 \geq 0$. As $\widehat{\Sigma}_t$ is positive definite its inverse $\widehat{\Sigma}_1^{-1}$ and exists and is also a positive definite matrix.[3] Finally, it can be shown that one can also define the square root of the matrix $\Sigma_t^{1/2}$ and that of its inverse $\Sigma_t^{-1/2}$ and these are symmetric and positive definite as well.

[3]This is trivial to show as $v^\top\Sigma^{-1}v = (v^\top\Sigma^{-1})\Sigma(\Sigma^{-1}v) > 0$ if $\Sigma^{-1}v \neq 0$ as Σ is positive definite, further $\Sigma^{-1}v = 0 \Leftrightarrow v = 0$. Therefore, if $v \neq 0$, then $v^\top\Sigma^{-1}v > 0$ and vice versa.

Let $\widehat{\mathbb{P}}_t(x \mid y, s) = \hat{f}_t(x)^\top g^\star(y, s)$ be the model estimated in round t by maximum likelihood estimation. Given any $s, x, y \in \mathcal{S} \times \mathcal{X} \times \mathcal{Y}$, the following important inequality holds,

$$\begin{aligned}
\left|\mathbb{P}(x|y,s) - \widehat{\mathbb{P}}_t(x|y,s)\right| &= \left|\left(f^\star(x) - \hat{f}_t(x)\right)^\top g^\star(y,s)\right|, \\
&= \left|\left(\Sigma_t^{1/2}\left(f^\star(x) - \hat{f}_t(x)\right)\right)^\top \left(\Sigma_t^{-1/2} g^\star(y,s)\right)\right|, \\
&\le \sqrt{\left(f^\star(x) - \hat{f}_t(x)\right)^\top \Sigma_t^{1/2} \cdot \Sigma_t^{1/2}\left(f^\star(x) - \hat{f}_t(x)\right)} \cdot \sqrt{g^\star(y,s)^\top \Sigma_t^{-1/2} \cdot \Sigma_t^{-1/2} g^\star(y,s)}, \\
&= \left\|f^\star(x) - \hat{f}_t(x)\right\|_{\Sigma_t} \cdot \|g^\star(y,s)\|_{\Sigma_t^{-1}}, \qquad (12)
\end{aligned}$$

where the second last step uses Cauchy-Schwarz inequality. This inequality allows us to bound the error in the estimated model for a given y in terms of the error based on historical data given by $\left\|f^\star(x) - \hat{f}_t(x)\right\|_{\Sigma_t}$ and the novelty of the given input $\|g^\star(y,s)\|_{\Sigma_t^{-1}}$. We want to bound the two terms in RHS of Equation 12.

First we'll prove the following result,

Lemma 2. *When Assumption 1 and Assumption 2, then with probability at least $1 - 2\delta$ we have:*

$$\sup_{x \in \mathcal{X}} \left\|f^\star(x) - \widehat{f}_t(x)\right\|_{\widehat{\Sigma}_t} \le (32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B \qquad (13)$$

for all $t \in \mathbb{N}$ simultaneously.

Proof.

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \|f^*(x) - \widehat{f}_t(x)\|^2_{\widehat{\Sigma}_t} &= \sup_{x \in \mathcal{X}} \left(\sum_{\ell=1}^{t-1} (f^\star(x) - \hat{f}_t(x))^\top g^\star(y_\ell, s_\ell) \cdot g^\star(y_\ell, s_\ell)^\top (f^\star(x) - \hat{f}_t(x)) + \lambda_t \left\|f^\star(x) - \hat{f}_t(x)\right\|_2^2\right) \\
&= \sup_{x \in \mathcal{X}} \left(\sum_{\ell=1}^{t-1} \left(f^\star(x)^\top g^\star(y_\ell, s_\ell) - \hat{f}_t(x)^\top g^\star(y_\ell, s_\ell)\right)^2 + \lambda_t \left\|f^\star(x) - \hat{f}_t(x)\right\|_2^2\right) \\
&\le \underbrace{\sum_{\ell=1}^{t-1} \sup_{x \in \mathcal{X}} \left(f^\star(x)^\top g^\star(y_\ell, s_\ell) - \hat{f}_t(x)^\top g^\star(y_\ell, s_\ell)\right)^2}_{:= U_t \text{ first term}} + \lambda_t \underbrace{\sup_{x \in \mathcal{X}} \left\|f^\star(x) - \hat{f}_t(x)\right\|_2^2}_{:= V_t \text{ second term}}
\end{aligned}$$

We first bound the second term V_t as:

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \left\|f^\star(x) - \hat{f}_t(x)\right\|_2^2 &= \sup_{x \in \mathcal{X}} \left(\|f^\star(x)\|_2^2 + \left\|\hat{f}_t(x)\right\|_2^2 - 2 f^\star(x)^\top \hat{f}_t(x)\right), \\
&\le \sup_{x \in \mathcal{X}} \|f^\star(x)\|_2^2 + \sup_{x \in \mathcal{X}} \left\|\hat{f}_t(x)\right\|_2^2 + 2 \sup_{x \in \mathcal{X}} \left|f^\star(x)^\top \hat{f}_t(x)\right|, \\
&\le \sup_{x \in \mathcal{X}} \|f^\star(x)\|_2^2 + \sup_{x \in \mathcal{X}} \left\|\hat{f}_t(x)\right\|_2^2 + 2 \sup_{x \in \mathcal{X}} \|f^\star(x)\|_2 \left\|\hat{f}_t(x)\right\|_2, \\
&\le 4B^2,
\end{aligned}$$

where we use the assumption that $\sup_{x \in \mathcal{X}} \|f(x)\|_2 \le B$ for any $f \in \mathcal{F}$ and that $f^\star, \hat{f}_t \in \mathcal{F}$ and the second last step uses Cauchy-Schwarz inequality.

We now bound the first term U_t. For a given $f \in \mathcal{F}$ we define $\Delta^f(x, y_\ell, s_\ell) = \left(f^\star(x)^\top g^\star(y_\ell, s_\ell) - f(x)^\top g^\star(y_\ell, s_\ell)\right)^2$ and $X_\ell^f = \sup_{x \in \mathcal{X}} \Delta^f(x, y_\ell, , s_\ell)$, which allows us to write $U_t = \sum_{\ell=1}^{t-1} X_\ell^{\hat{f}_t}$.

We fix $f \in \mathcal{F}$. We have $X_\ell^f \ge 0$ by definition and as $f^\star(x)^\top g^\star(y_\ell, s_\ell), f(x)^\top g^\star(y_\ell, s_\ell) \in [0, 1]$, we also have

$$X_\ell^f = \sup_{x \in \mathcal{X}} \Delta^f(x, y_\ell, s_\ell) = \sup_{x \in \mathcal{X}} \left(f^\star(x)^\top g^\star(y_\ell, s_\ell) - f(x)^\top g^\star(y_\ell, s_\ell)\right)^2 \le 1, \qquad (14)$$

Let $\mathbb{P}_\ell \in \Delta(\mathcal{Y})$ be the marginal distribution over y_ℓ conditioned on $\{x_{\ell'}, s_{\ell'}, y_{\ell'}, x'_{\ell'}\}_{\ell'=1}^{\ell-1} \cup \{x_\ell, s_\ell\}$ then

$$\begin{aligned}
\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left(X_\ell^f\right)^2\right] &= \mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\sup_{x\in\mathcal{X}}\left(f^\star(x)^\top g^\star(y,s_\ell) - f(x)^\top g^\star(y,s_\ell)\right)^4\right] \\
&\le 16\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left\|f^\star(\cdot)^\top g^\star(y,s_\ell) - f(\cdot)^\top g^\star(y,s_\ell)\right\|_{\mathrm{TV}}^4\right], \\
&\le 16\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left\|f^\star(\cdot)^\top g^\star(y,s_\ell) - f(\cdot)^\top g^\star(y,s_\ell)\right\|_{\mathrm{TV}}^2\right],
\end{aligned}$$

where we use the fact that $\|\cdot\|_\infty \le \|\cdot\|_1$, that TV distance between two distributions is equal to half of 1-norm distance between them, and that $\|\cdot\|_{\mathrm{TV}} \le 1$. Thus, using the anytime Freedman's inequality (see Lemma 6) and union bound over all $f \in \mathcal{F}$, we get:

$$\begin{aligned}
\sum_{\ell=1}^{t-1} X_\ell^f &\le \eta\sum_{\ell=1}^{t-1}\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left(X_\ell^f\right)^2\right] + \frac{C\log(t|\mathcal{F}|/\delta)}{\eta} \\
&\le 16\eta\sum_{\ell=1}^{t-1}\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left\|f^\star(\cdot)^\top g^\star(y,s_\ell) - f(\cdot)^\top g^\star(y,s_\ell)\right\|_{\mathrm{TV}}^2\right] + \frac{C\log(t|\mathcal{F}|/\delta)}{\eta},
\end{aligned}$$

simultaneously for all $t \in \mathbb{N}$ and all $f \in \mathcal{F}$ with probability at least $1-\delta$.

In particular by setting $f = \widehat{f}_t$ this implies,

$$U_t = \sum_{\ell=1}^{t-1} X_\ell^{\widehat{f}_t} \le 16\eta\sum_{\ell=1}^{t-1}\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left\|f^\star(\cdot)^\top g^\star(y,s_\ell) - \widehat{f}_t(\cdot)^\top g^*(y,s_\ell)\right\|_{\mathrm{TV}}^2\right] + \frac{C\log(t|\mathcal{F}|/\delta)}{\eta} \tag{15}$$

with probability at least $1-\delta$. Finally, Proposition 6 implies that with probability at least $1-\delta$,

$$\sum_{\ell=1}^{t-1}\mathbb{E}_{y\sim\mathbb{P}_\ell}\left[\left\|f^\star(\cdot)^\top g^\star(y,s_\ell) - \hat{f}_t(\cdot)^\top g^\star(y,s_\ell)\right\|_{\mathrm{TV}}^2\right] \le 2\log(t|\mathcal{F}|/\delta) \tag{16}$$

for all $t \in \mathbb{N}$. Therefore a union bound implies that with probability at least $1-2\delta$, combining the bounds of Equation 15 and Equation 16,

$$U_t = \sum_{\ell=1}^{t-1} X_\ell^{\widehat{f}_t} \le 32\eta\log(t|\mathcal{F}|/\delta) + \frac{C\log(t|\mathcal{F}|/\delta)}{\eta}$$

simultaneously for all $t \in \mathbb{N}$. Optimizing for η gives us $\eta = \sqrt{\frac{C}{32}}$, which gives us $U_t \le \sqrt{32C}\log(t|\mathcal{F}|/\delta)$.

Plugging together the upper bounds for U_t and V_t we get:

$$\sup_{x\in\mathcal{X}}\left\|f^\star(x) - \hat{f}_t(x)\right\|_{\widehat{\Sigma}_t}^2 \le U_t + \lambda_t V_t \le \sqrt{32C}\log(t|\mathcal{F}|/\delta) + 4\lambda_t B^2,$$

with probability $1-2\delta$ for all $t \in \mathbb{N}$. This gives us the desired bound as $\sup_{x\in\mathcal{X}}\left\|f^\star(x) - \hat{f}_t(x)\right\|_{\widehat{\Sigma}_t} \le \sqrt{\sqrt{32C}\log(t|\mathcal{F}|/\delta) + 4\lambda_t B^2} \le (32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B$ where we use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$. □

From now on we'll define as $\mathcal{E}$ the event that Equation 13 holds for all $t \in \mathbb{N}$. As established in Lemma 2, we have $\mathbb{P}(\mathcal{E}) \ge 1-2\delta$. From Lemma 2 we can also establish the following corollary.

Corollary 5. *If event $\mathcal{E}$ holds then we have:*

$$\left|\left(f^\star(x) - \hat{f}_t(x)\right)^\top g^\star(y,s)\right| \le \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B\right)\|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}}$$

for all $s \in \mathcal{S}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and all $t \in \mathbb{N}$.

Proof. Let the event $\mathcal{E}$ hold. Then the following sequence of inequalities holds,

$$\begin{aligned}\left|\left(f^\star(x)-\hat{f}_t(x)\right)^\top g^\star(y,s)\right| &\overset{(i)}{\leq} \left\|f^\star(x)-\hat{f}_t(x)\right\|_{\widehat{\Sigma}_t} \|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}}\,, \\ &\overset{(ii)}{\leq} \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)}+2\sqrt{\lambda_t}B\right)\|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}}\,,\end{aligned}$$

where (i) holds by Inequality 12 and (ii) holds because of Lemma 2 and because we are assuming $\mathcal{E}$ holds. □

B.1. Proving Optimism for LORIL

We next show that LORIL derives a useful optimism inequality that allows us to upper bound the true model probabilities $\mathbb{P}(x \mid y,s) = f^\star(x)^\top g^\star(y,s)$ using the estimated model probabilities $\hat{f}_t(x)^\top g^\star(y,s)$.

Lemma 3. *If event $\mathcal{E}$ holds, then we have:*

$$f^\star(x)^\top g^\star(\pi^\star(x,s),s) \leq \hat{f}_t(x)^\top g^\star(\pi_t(x,s),s) + \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)}+2\sqrt{\lambda_t}B\right)\|g^\star(\pi_t(x,s),s)\|_{\widehat{\Sigma}_t^{-1}}\,, \quad (17)$$

for all $x \in \mathcal{X}$ and $t \in \mathbb{N}$, where π_t is the policy in t^{th} round of LORIL *and $\pi^\star : x,s \mapsto \arg\max_{y\in\mathcal{Y}} \mathbb{P}(x \mid y,s)$ is the optimal policy.*

Proof. If event $\mathcal{E}$ holds then for any $x \in \mathcal{S}, x \in \mathcal{X}, y \in \mathcal{Y}$, and $t \in \mathbb{N}$ we have directly from Corollary 5 the following:

$$f^\star(x)^\top g^\star(y,s) \leq \hat{f}_t(x)^\top g^\star(y,s) + \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)}+2\sqrt{\lambda_t}B\right)\|g^\star(y,s)\|_{\widehat{\Sigma}_t^{-1}}\,.$$

For any $s \in \mathcal{S}$, $x \in \mathcal{X}$ and $t \in \mathbb{N}$, this implies:

$$\begin{aligned}f^\star(x)^\top g^\star(\pi^\star(x,s),s) &\leq \hat{f}_t(x)^\top g^\star(\pi^\star(x,s),s) + \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)}+2\sqrt{\lambda_t}B\right)\|g^\star(\pi^\star(x,s),s)\|_{\widehat{\Sigma}_t^{-1}} \\ &\leq \hat{f}_t(x)^\top g^\star(\pi_t(x),s) + \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)}+2\sqrt{\lambda_t}B\right)\|g^\star(\pi_t(x),s)\|_{\widehat{\Sigma}_t^{-1}}\,,\end{aligned}$$

where the second inequality holds because of the definition of $\pi_t(x,s)$. □

B.2. Regret Upper Bound

We are ready to upper bound the regret of LORIL(Algorithm 1). Recall that we define regret for a given run of LORIL as,

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(\mathbb{P}(x_t \mid \pi^\star(x_t,s_t),s_t) - \mathbb{P}(x_t \mid \pi_t(x_t,s_t),s_t)\right).$$

The main result is the following theorem,

Theorem 3. *[Regret bound of* LORIL*] When Assumption 1 and Assumption 2 hold and $\lambda_t = 1/t$ the regret of Algorithm 1 satisfies*

$$\begin{aligned}\mathit{Reg}(T) =& \mathcal{O}\Big(B\sqrt{Td\log(1+TB)}+ \\ &\sqrt{Td\log(T|\mathcal{F}|/\delta)\log(1+TB)}\Big)\end{aligned}$$

with probability at least $1-3\delta$ for all $T \in \mathbb{N}$.

Proof. Fix $T \in \mathbb{N}$. We assume event $\mathcal{E}$ holds, then we can bound the regret as:

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \left(\mathbb{P}(x_t \mid \pi^\star(x_t, s_t), s_t) - \mathbb{P}(x_t \mid y_t, s_t)\right), \\
&= \sum_{t=1}^{T} \left(f^\star(x_t)^\top g^\star(\pi^\star(x_t), s_t) - f^\star(x_t)^\top g^\star(\pi_t(x_t), s_t)\right), \\
&\overset{(i)}{\leq} \sum_{t=1}^{T} \Big(\hat{f}_t(x_t)^\top g^\star(\pi_t(x_t), s_t) \\
&\qquad + \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B\right) \|g^\star(\pi_t(x_t), s_t)\|_{\widehat{\Sigma}_t^{-1}} - f^\star(x_t)^\top g^\star(\pi_t(x_t), s_t)\Big),
\end{aligned}$$

where we use Lemma 3 in the last step. We also have:

$$\hat{f}_t(x)^\top g^\star(\pi_t(x_t), s_t) - f^\star(x_t)^\top g^\star(\pi_t(x_t), s_t) \leq \left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B\right) \|g^\star(\pi_t(x_t), s_t)\|_{\widehat{\Sigma}_t^{-1}}$$

due to Corollary 5. Combining these two results we get:

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T} 2\left((32C)^{1/4}\sqrt{\log(t|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_t}B\right) \|g^\star(\pi_t(x_t), s_t)\|_{\widehat{\Sigma}_t^{-1}} \tag{18}$$

$$\leq \sup_{t' \in [T]} \left(2\left((32C)^{1/4}\sqrt{\log(t'|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_{t'}}B\right)\right) \cdot \sum_{t=1}^{T} \|g^\star(\pi_t(x_t), s_t)\|_{\widehat{\Sigma}_t^{-1}} \tag{19}$$

Let $\widetilde{\Sigma}_t = \sum_{\ell=1}^{t-1} g^\star(y_\ell, s_\ell)\left(g^\star(y_\ell, s_\ell)\right)^\top + \lambda_T \mathbb{I}$ for all $t \in [T]$. Further, let $D_t = \sum_{\ell=1}^{t-1} g^\star(y_\ell, s_\ell)\left(g^\star(y_\ell, s_\ell)\right)^\top$, which gives us $\widetilde{\Sigma}_t = D_t + \lambda_T \mathbb{I}$ and $\widehat{\Sigma}_t = D_t + \lambda_t \mathbb{I}$. This implies that for any $v \in \mathbb{R}^d$ we have

$$\|v\|_{\widetilde{\Sigma}_t} = \sqrt{v^\top (D_t + \lambda_T \mathbb{I})v} = \sqrt{v^\top D_t v + \lambda_T \|v\|_2^2} \leq \sqrt{v^\top D_t v + \lambda_t \|v\|_2^2} = \sqrt{v^\top (D_t + \lambda_t \mathbb{I})v} = \|v\|_{\widehat{\Sigma}_t}.$$

As $\widehat{\Sigma}_t \succeq \widetilde{\Sigma}_t \succ 0$ and $\widehat{\Sigma}_t = \widetilde{\Sigma}_t + (\lambda_t - \lambda_T)\mathbb{I}$ with $\lambda_t - \lambda_T > 0$ it follows that by Proposition 8, $\|v\|_{\widehat{\Sigma}_t^{-1}} \leq \|v\|_{\widetilde{\Sigma}_t^{-1}}$. This implies

$$\sum_{t=1}^{T} \|g^\star(\pi_t(x_t), s_t)\|_{\widehat{\Sigma}_t^{-1}} \leq \sum_{t=1}^{T} \|g^\star(\pi_t(x_t), s_t)\|_{\widetilde{\Sigma}_t^{-1}}$$

Finally, Proposition 7 implies

$$\sum_{t=1}^{T} \|g^\star(y_t, s_t)\|_{\widetilde{\Sigma}_t^{-1}} \leq \sqrt{2Td\log\left(1 + \frac{TB^2}{\lambda_T}\right)},$$

where we use $\sup_{y \in \mathcal{Y}, s \in \mathcal{S}} \|g^\star(y, s)\|_2 \leq B$. Combining these we get:

$$\mathrm{Reg}(T) \leq \sup_{t' \in [T]} \left(2\left((32C)^{1/4}\sqrt{\log(t'|\mathcal{F}|/\delta)} + 2\sqrt{\lambda_{t'}}B\right)\right) \sqrt{2Td\log\left(1 + \frac{TB^2}{\lambda_T}\right)}$$

Using $\lambda_t = 1/t \leq 1$ we get:

$$\begin{aligned}
\mathrm{Reg}(T) &\leq \left(2\left((32C)^{1/4}\sqrt{\log(T|\mathcal{F}|/\delta)} + 2B\right)\right)\sqrt{2Td\log\left(1 + T^2B^2\right)} \\
&\leq 8\left((2C)^{1/4}\sqrt{\log(T|\mathcal{F}|/\delta)} + B\right)\sqrt{Td\log\left(1 + TB\right)},
\end{aligned}$$

where we use $\log(1 + T^2B^2) \leq 2\log(1 + TB)$. This gives us

$$\mathrm{Reg}(T) = \mathcal{O}\left(B\sqrt{Td\log(1 + TB)} + \sqrt{Td\log(T|\mathcal{F}|/\delta)\log(1 + TB)}\right).$$

□

C. Useful Lemmas

Lemma 4 (Hoeffding Inequality). *Let $\{Y_\ell\}_{\ell=1}^{\infty}$ be a martingale difference sequence such that Y_ℓ is $Y_\ell \in [a_\ell, b_\ell]$ almost surely for some constants a_ℓ, b_ℓ almost surely for all $\ell = 1, \cdots, t$. then*

$$\sum_{\ell=1}^{t} Y_\ell \leq 2\sqrt{\sum_{\ell=1}^{t}(b_\ell - a_\ell)^2 \ln\left(\frac{1}{\widetilde{\delta}}\right)}.$$

With probability at least $1 - \widetilde{\delta}$.

See for example Corollary 2.20 from (Wainwright, 2019).

Our results relies on the following variant of Bernstein inequality for martingales, or Freedman's inequality (Freedman, 1975), as stated in e.g., (Agarwal et al., 2014; Beygelzimer et al., 2011).

Lemma 5 (Simplified Freedman's inequality). *Let $X_1, ..., X_T$ be a bounded martingale difference sequence with $|X_\ell| \leq R$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^{T} X_\ell \leq \eta \sum_{\ell=1}^{T} \mathbb{E}_\ell[X_\ell^2] + \frac{\log(1/\delta')}{\eta}. \tag{20}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation[4] induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Lemma 6 (Anytime Freedman). *Let $\{X_t\}_{t=1}^{\infty}$ be a bounded martingale difference sequence with $|X_t| \leq R$ for all $t \in \mathbb{N}$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, there exists a universal constant $C > 0$ such that for all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^{t} X_\ell \leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell[X_\ell^2] + \frac{C\log(t/\delta')}{\eta}. \tag{21}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Proof. This result follows from Lemma 5. Fix a time-index t and define $\delta_t = \frac{\delta'}{12t^2}$. Lemma 5 implies that with probability at least $1 - \delta_t$,

$$\sum_{\ell=1}^{t} X_\ell \leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{\log(1/\delta_t)}{\eta}.$$

A union bound implies that with probability at least $1 - \sum_{\ell=1}^{t} \delta_t \geq 1 - \delta'$,

$$\begin{aligned}\sum_{\ell=1}^{t} X_\ell &\leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{\log(12t^2/\delta')}{\eta} \\ &\overset{(i)}{\leq} \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{C\log(t/\delta')}{\eta}.\end{aligned}$$

holds for all $t \in \mathbb{N}$. Inequality (i) holds because $\log(12t^2/\delta') = \mathcal{O}\left(\log(t\delta')\right)$.

□

Adapted from Theorem 21 from (Agarwal et al., 2020). See also (Geer, 2000).

[4] We will use this notation to denote conditional expectations throughout this work.

Proposition 6 (MLE Bound). *For any fixed $\delta \in (0,1)$,*

$$\sum_{\ell=1}^{t-1} \mathbb{E}_{y\sim\pi_h}\left[\left\|\hat{f}_t(\cdot)^\top g^\star(y,s_\ell) - f^\star(\cdot)^\top g^\star(y,s_\ell)\right\|_{TV}^2\right] \le 2\ln(t|\mathcal{F}|/\delta)$$

for all $t \in \mathbb{N}$ simultaneously with probability at least $1-\delta$ where $\mathbb{P}_\ell \in \Delta(\mathcal{Y})$ is the distribution over y_ℓ conditioned on $\{x_{\ell'}, s_{\ell'}, y_{\ell'}, x'_{\ell'}\}_{\ell'=1}^{\ell-1} \cup \{x_\ell, s_\ell\}$.

Proof. This result is an immediate consequence of Theorem 21 from (Agarwal et al., 2020). We convert this result into an anytime statement by invoking this result repeatedly with probability values $\delta_t = \frac{\delta}{3*t^2}$ and then applying a union bound. □

Proposition 7 (Proposition 3 from (Pacchiano et al., 2021)). *For any sequence of vectors $v_1, \cdots, v_t \subset \mathbb{R}^d$ satisfying $\|v_\ell\| \le L$ for all $\ell \in \mathbb{N}$, let Σ_t be its corresponding Gram matrix $\Sigma_t = \lambda\mathbb{I} + \sum_{\ell=1}^{t-1} v_\ell v_\ell^\top$. Then for all $t \in \mathbb{N}$, we have*

$$\sum_{\ell=1}^{T} \|v_\ell\|_{\Sigma_\ell^{-1}} \le \sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)}$$

Proposition 8. *Let $A \succ 0$ be a $d\times d$ positive definite matrix. And let $\lambda' \ge 0$. If $v \in \mathbb{R}^d$,*

$$\|v\|_{A+\lambda'\mathbb{I}} \ge \|v\|_A$$

and

$$\|v\|_{(A+\lambda'\mathbb{I})^{-1}} \le \|v\|_{A^{-1}}$$

Proof. Let $v_1, \cdots, v_d$ be an orthonormal basis of eigenvectors of A. The eigenvalues of A are positive because the matrix is assumed to be positive definite. Call $\mu_i > 0$ to the eigenvalue associated with eigenvector v_i. Elementary linear algebra shows that,

$$(A+\lambda'\mathbb{I})v_i = Av_i + \lambda' v_i = (\mu_i + \lambda')v_i$$

thus showing that v_i are also an orthonormal basis of eigenvectors for $A+\lambda'\mathbb{I}$ and have eigenvalues $\mu_i + \lambda'$. Thus,

$$\|v\|_A^2 = v^\top(A)\,v = \sum_{i=1}^d \mu_i\left(\langle v, v_i\rangle\right)^2 \le \sum_{i=1}^d (\mu_i+\lambda')\cdot\left(\langle v, v_i\rangle\right)^2 = v^\top\left(A+\lambda'\mathbb{I}\right)) = \|v\|_{A+\lambda'\mathbb{I}}^2$$

Notice that $A^{-1} = \sum_{i=1}^d \frac{1}{\mu_i} v_i v_i^\top$ and $(A+\lambda'\mathbb{I})^{-1} = \sum_{i=1}^d \frac{1}{\mu_i+\lambda'} v_i v_i^\top$. And therefore,

$$\|v\|_{A^{-1}}^2 = v^\top A^{-1} v = \sum_{i=1}^d \frac{1}{\mu_i}\left(\langle v, v_i\rangle\right)^2 \ge \sum_{i=1}^d \frac{1}{\mu_i+\lambda'}\left(\langle v, v_i\rangle\right)^2 = v^\top\left(A+\lambda'\mathbb{I}\right)^{-1} v = \|v\|_{(A+\lambda'\mathbb{I})^{-1}}^2$$

The result follows. □

D. Experimental Details

We provide additional details of our experiments in this section.

D.1. Additional Details for the First Experiment on Synthetic Task

We use the almost same implementation of `LORIL` as listed in Algorithm 1. The only change we make is that we use a simplified policy given by:

$$\pi_t(x_t) = \arg\max_{y\in\mathcal{Y}} \hat{f}_t(x_t)^\top g^\star(y,s_t) + k\,\|g^\star(y,s_t)\|_{\widehat{\Sigma}_t^{-1}}, \tag{22}$$

where k is a single hyperparameter. We also use $\lambda_t = \lambda$ which is a hyperparameter to be tuned. We use the hyperparameter values in Table 1.

Hyperparameter	Values
ϵ	grid search in $[0.05, 0.1, 0.2, 0.3]$
λ	grid search in $[0.05, 0.1, 1.0]$
k	grid search in $[0.1, 1.0, 10.0]$
optimization	Adam
learning rate	0.001
temperature used in defining F and G	0.75
$\lvert\mathcal{X}\rvert$	2000
d	10
$\lvert\mathcal{Y}\rvert$	10

Table 1. Grid search for hyperparameter for the first experiment.

Hyperparameter	Values
ϵ	grid search in $[0.05, 0.1, 0.2, 0.3]$
λ	grid search in $[0.05, 0.1, 1.0]$
k	grid search in $[0.1, 1.0, 10.0]$
optimization	Adam
learning rate	0.001
vocabulary size	34
word embedding dimension	10
GRU hidden dimension	10
dimension of $g^\star$ encoding	16
number of layers in GRU	2
possible templates	10
object types	[“square”, “rectangle”, “triangle”, “circle”]
object color	[“red”, “blue”, “green”, “yellow”, “black”, “grey”, “black”, “cyan”, “orange”]
$\lvert\mathcal{Y}\rvert$	5

Table 2. Grid search for hyperparameter for the second experiment.

D.2. Additional Details for the Second Experiment on the Image Selection Task

We use the hyperparameter values shown in Table 2. The list of templates is given below where {object1} and {color1} are variables that are replaced by the object type and its color in a given image.

1. “You are seeing a {object1} of color {color1}”
2. “The image contains a {object1} of color {color1}”
3. “There is a {color1} colored object of type {object1}”
4. “A {color1} {object1}”
5. “The object is a {object1} and its color is {color1}.”
6. “The image has a single {color1} colored {object1}.”
7. “You are seeing a {color1} colored {object1}.”
8. “There is a {color1} colored object.”
9. “You are seeing a {object1} in the image.”
10. “There is a {color1} colored object1.”

Autoencoder. We use a 3-layer autoencoder with leaky relu activations. The first two layers apply a convolution with 16 8×8 kernels with stride 4. The last layer applies 15 4×4 kernels with stride 2. The input image is an RGB image of size $200 \times 200 \times 3$. After applying the CNN encoder, we get a feature of size $16 \times 4 \times 4 = 64$. We flatten this feature and apply a fully connected layer to map it to another vector of size 64. We reshape it and pass it through a 3-layer deconvolutional network with leaky relu activation, to predict an image of the same size as the input. The first layer of the decoder applies 16 4×4 convtranspose2d of stride 2 and output padding 1. The second layer applies 16 8×8 convtranspose2d of stride 4 and output padding 1. Finally, the last layer applies 16 8×8 convtranspose2d of stride 4 with no output padding. We train the autoencoder with squared loss using Adam optimization. We apply gradient clip to clip gradient above a clipping value of 2.5. Finally, we model $g^\star(y, s)$ by first applying the encoder to generate a feature map of $16 \times 4 \times 4$, and then summing over the first dimension and flattening the remaining tensor into a 16-dimensional vector.

Compute. We use A2600 for all experiments. The entire set of experiments took 3 hours to finish. We used PyTorch to implement the code.