Approximate MaxEnt Inverse Optimal Control and its Application for Mental Simulation of Human Interactions (Extended Version with Proofs)

De-An Huang, Amir-massoud Farahmand, Kris M. Kitani, and J. Andrew Bagnell

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Maximum entropy inverse optimal control (MaxEnt IOC) is an effective means of discovering the underlying cost function of demonstrated human activity and can be used to predict human behavior over low-dimensional state spaces (i.e., forecasting of 2D trajectories). To enable inference in very large state spaces, we introduce an approximate MaxEnt IOC procedure to address the fundamental computational bottleneck stemming from calculating the partition function via dynamic programming. Approximate MaxEnt IOC is based on two components: approximate dynamic programming and Monte Carlo sampling. We analyze this approximation approach and provide a finite-sample error upper bound on its excess loss. We validate the proposed method in the context of analyzing dual-agent interactions from video, where we use approximate MaxEnt IOC to simulate mental images of a single agents body pose sequence (a high-dimensional image space). We experiment with sequences image data taken from RGB and RGBD data and show that it is possible to learn cost functions that lead to accurate predictions in high-dimensional problems that were previously intractable.

1 Introduction

The Maximum Entropy (MaxEnt) Inverse Optimal Control (IOC) framework is an effective approach for discovering the underlying reward model of a rational agent and enables robust sequence prediction over low-dimensional state spaces [1, 2]. The IOC framework is particularly useful in the context of understanding and modeling human activities, where the recovered reward model intuitively encodes a person's set of preferences. Furthermore, in the MaxEnt formulation of IOC, the soft-maximum value function (log-partition function) compactly describes a global distribution over every possible action sequence. The log-partition function can then be used to simulate and forecast human activities.

Of particular interest in this paper is recent work fusing computer vision and IOC to mentally (visually) simulate human activities. By integrating visual attributes of the scene as features of the reward function, it was shown that highly accurate pedestrian trajectories can be simulated in novel scenes [3]. The application of IOC to visual prediction problems, however, has been limited to

2D pedestrian trajectories since current approaches only work for problems with small state space. To extend IOC to deal with the inherent high-dimensional nature of observed human activity from image data, previous approaches [4, 5] relied on clustering techniques to quantize and reduce the size of the state space. However, coarse discretization of the state space resulted in non-smooth trajectories and inhibited the model's power to simulate the subtle qualities of activity dynamics.

At the heart of the problem of maximum entropy sequence prediction is an inference procedure which requires enumeration of all possible action sequences into the future given a set of observations. In the same way that the value function is computed for optimal control, the *log-partition function* of maximum entropy IOC can be computed using dynamic programming – differing only in the substitution of the "soft-max" operator for the "max" operator in the Bellman equations. This relationship was noted as early as [6] and formalized in [1]. While dynamic programming renders this efficient for small scale problems, more appropriate techniques are needed for dealing with problems with large state spaces.

When the state space is large, one natural approach is to use approximate dynamic programming for the approximate calculation of these functions. We draw our inspiration from value function approximation methods, which have been successful in solving high-dimensional control problems [7, 8, 9, 10], to address the high-dimensional challenges in our scenario.

The algorithmic contribution of this work is an approximate MaxEnt IOC algorithm, suitable for dealing with high-dimensional problems, that uses an Approximate Value Iteration (AVI) algorithm to compute the softmax-based value (log-partition) function. The AVI procedure uses a regression estimator at each iteration, where the choice of the estimator is not constrained. In particular, we utilize a reproducing kernel Hilbert space-based (RKHS) regularized estimator due to its flexibility and favourable properties – though the framework is more general and allows other regression estimators such as local averagers, random forests, boosting, neural networks, etc. Efficient Monte Carlo sampling then enables a dimension-independent estimate of the gradient of the reward function.

The theoretical contribution of this paper is the analysis of this approximate procedure. We provide a finite-sample upper bound guarantee on the excess loss, i.e., the loss of our approximate procedure compared to an "ideal" MaxEnt IOC procedure without any approximation in the computation of the log-partition function or the feature expectation.

2 IOC for High-Dimensional Problems

The problem of the inverse optimal control (also known as inverse reinforcement learning) is to recover an agent's (or expert's) reward function given a controller or policy (or samples from the agent's behavior) when the dynamics of the process is known.

To describe our approach to IOC, which is based on the Maximum Entropy Inverse Optimal Control of [2], we first define a parametric-reward Markov Decision Process (θ-MDP). θ-MDP is defined as a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \underline{g}, \theta)$, where $\mathcal{X}$ is a measurable state space (e.g., $\mathbb{R}^D$), $\mathcal{A}$ is a finite set of actions, $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\underline{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a mapping from state-action pairs to feature vectors of dimension d, and $\theta \in \mathbb{R}^d$ parametrizes the reward.[1] We consider θ-MDPs with finite horizon of T. For notational convenience, given a sequence $z_{1:T} = (z_1, \dots, z_T)$, we denote $\underline{f}(z_{1:T}) = \sum_{t=1}^{T} \underline{g}(z_t)$. In IOC, we assume that $\mathcal{P}$ is known (or estimated separately).

Consider a set of demonstrated trajectories $\mathcal{D}_n = \{Z_{1:T}^{(i)}\}_{i=1}^{n}$ with each trajectory $Z_{1:T} = (Z_1, \dots, Z_T) \sim \zeta$ with $Z_t = (X_t, A_t)$ and ζ being an unknown distribution over the set of

[1] $\mathcal{M}(\Omega)$ is the set of probability distributions over Ω.

trajectory. Also denote $\nu \in \mathcal{M}(\mathcal{X})$ as the distribution of X_1. We assume that this initial distribution is known. For a policy π, denote $P_\pi(Z_{1:T})$ as the distribution induced by following policy π. In the discrete state case, $P_\pi(Z_{1:T}) = \prod_{t=1}^{T-1} \mathcal{P}(X_{t+1}|X_t, A_t)\pi(A_t|X_t)$ (and similarly for continuous state spaces). Define the *causal conditioned probability* $\mathbb{P}\{A_{1:T}||X_{1:T}\} = \prod_{t=1}^{T} \mathbb{P}\{A_t|X_t\} = \prod_{t=1}^{T} \pi_t(A_t|X_t)$, which reflects the fact that future states do not influence earlier actions (compare with conditional probability $\mathbb{P}\{A_{1:T}|X_{1:T}\}$). We define the *causal entropy* H_π as $H_\pi = \mathbb{E}_{P_\pi(Z_{1:T})}[-\log \mathbb{P}\{A_{1:T}||X_{1:T}\}]$.

The primal optimization problem in Maximum Entropy Inverse Optimal Control estimator [2] is

$$\begin{aligned} &\arg\max_{\pi} H_\pi(A_{1:T}||X_{1:T}) \\ &\text{s.t.} \qquad \mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right] = \frac{1}{n}\sum_{i=1}^{n} \underline{f}\left(Z_{1:T}^{(i)}\right). \end{aligned} \tag{1}$$

The motivation behind this objective function is to find a policy π whose induced expected features, $\mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$, matches the empirical feature count of the agent, that is $\frac{1}{n}\sum_{i=1}^{n} \underline{f}(Z_{1:T}^{(i)})$, while not committing to any distribution beyond what is implied by the data. The dual of this constrained optimization problem is (Theorem 3 of [2])

$$\min_{\theta\in\mathbb{R}^d} \log \mathcal{Z}_\theta - \left\langle \theta\,,\, \frac{1}{n}\sum_{i=1}^{n} \underline{f}\left(Z_{1:T}^{(i)}\right) \right\rangle, \tag{2}$$

in which $\log \mathcal{Z}_\theta$ is the log-partition function. For notational compactness, define $\hat{b}_n, \bar{b} \in \mathbb{R}^d$ as $\hat{b}_n = \frac{1}{n}\sum_{i=1}^{n} \underline{f}(Z_{1:T}^{(i)})$ and $\bar{b} = \mathbb{E}_{Z_{1:T}\sim\zeta}\left[\underline{f}(Z_{1:T})\right]$. The vector $\bar{b}$ is the true expected feature of the agent, which is unknown.

A key observation is that one might calculate $\log \mathcal{Z}_\theta$ using a Value Iteration (VI) procedure: For any $\theta \in \mathbb{R}^d$, define $r_t(x,a) = r(x,a) = \langle \theta\,,\, \underline{g}(x,a) \rangle$, and perform the following VI procedure: Set $Q_T = r_T$, and for $t = T-1, \ldots, 1$,

$$\begin{aligned} Q_t(x,a) &= r_t(x,a) + \int \mathcal{P}(dy|x,a) V_{t+1}(y), \\ V_t(x) &= \operatorname{soft\,max}(Q_t(x,\cdot) \triangleq \log\left(\sum_{a\in\mathcal{A}} \exp(Q_t(x,a))\right). \end{aligned} \tag{3}$$

We compactly write $Q_t = r_t + \mathcal{P}^a V_{t+1}$, where $\mathcal{P}^a(\cdot|x) = \mathcal{P}(\cdot|x,a)$.

It can be shown that $\log \mathcal{Z}_\theta = \mathbb{E}_\nu[V_1(X)]$. Also the MaxEnt policy solution to (1), which is in the form of Boltzmann distribution, is $\pi_t(a|x) = \pi_{t,\theta}(a|x) = \frac{\exp(Q_t(x,a))}{\sum_{a'\in\mathcal{A}} \exp(Q_t(x,a))} = \exp(Q_t(x,a) - V_t(x))$.

Instead of (2), we aim to solve the following regularized dual objective

$$\min_{\theta\in\mathbb{R}^d} L(\theta, \hat{b}_n) \triangleq \log \mathcal{Z}_\theta - \left\langle \theta\,,\, \hat{b}_n \right\rangle + \frac{\lambda}{2}\|\theta\|_2^2, \tag{4}$$

which can be interpreted as a relaxation of the constraints in the primal as shown by [11, 12]. Adding a regularization has a Bayesian interpretation too, and corresponds to having a prior over parameters.

Algorithm 1 – Backward pass

$\mathcal{D}_m^{(t)} = \{(X_i, A_i, R_i^t, X_i')\}_{i=1}^m$, $R_i^t = \langle\, \theta\,,\, \underline{g}(X_i, A_i)\, \rangle$
$\hat{Q}_T \leftarrow 0$
for $t = T-1, \dots, 2, 1$ **do**
  $Y_i^t = R_i^t + \operatorname{soft\,max} \hat{Q}_{t+1}(X_i', \cdot)$
  $\hat{Q}_t \leftarrow \operatorname{argmin}_Q \frac{1}{m} \sum_{i=1}^m \left| Q(X_i, A_i) - Y_i^t \right|^2 + \lambda_{Q,m} \|Q\|_{\mathcal{H}}^2$
  $\hat{\pi}_t(a|x) \propto \exp(\hat{Q}(x, a))$
end for

Algorithm 2 – Forward pass

$\underline{f} \leftarrow 0$
repeat
  $\hat{X}_1 \sim \nu$
  for $t = 1, \dots T-1$ **do**
    $\hat{A}_t \sim \hat{\pi}_t(\cdot|\hat{X}_t)$, $\underline{f} \mathrel{+}= \underline{g}^t(\hat{X}_t, \hat{A}_t)$
    $\hat{X}_{t+1} \sim \mathcal{P}(\cdot|\hat{X}_t, \hat{A}_t)$
  end for
until N sample paths
$\underline{f} \leftarrow \frac{1}{N} \underline{f}$ (estimated log-partition function gradient)

It can be shown that $\nabla_\theta \log \mathcal{Z}_\theta = \mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$ with $X_1 \sim \nu$, so the gradient of the loss function, which can be used in a gradient-descent-like procedure, is

$$\nabla_\theta L(\theta, \hat{b}_n) = \mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right] - \hat{b}_n + \lambda\theta \tag{5}$$

For problems with large state space, the exact calculation of the log-partition function $\log \mathcal{Z}_\theta$ is infeasible as is the calculation of the the expected features $\mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$.

Nonetheless, one can aim to approximate the log-partition function and estimate the expected features. We use two key insights to design an algorithm that can handle large state spaces. The first is that one can approximate the VI procedure of (3) using function approximators. The Approximate Value Iteration (AVI) procedure has been successfully used and theoretically analyzed in the Approximate Dynamic Programming and RL literature [8, 9, 13].

The second insight, which is also used in some previous work such as [14], is that one can estimate an expectation by Monte Carlo sampling and the error behavior would be $O(\frac{1}{\sqrt{N}})$ (for N independent trajectories), which is a dimension-free rate. These procedures are summarized in Algorithms 1 and 2. We describe each of them in detail.

To perform AVI, we use samples in the form of $\mathcal{D}_m^{(t)} = \{(X_i, A_i, R_i, X_i'\}_{i=1}^m$ with $X_i \sim \eta \in \mathcal{M}(\mathcal{X})$, $A_i \sim \pi_b(X_i)$, $R_i \sim \mathcal{R}(\cdot|X_i)$, and $X_i' \sim \mathcal{P}(\cdot|X_i, A_i)$. Here π_b is a behavior policy.[2] Given these samples, one can estimate Q_t with $\hat{Q}_t$ by solving a regression problem in which the input variables are $Z_i = (X_i, A_i)$ and the target values are $R_i + \hat{V}_{t+1}(X_i')$, and $\hat{V}_{t+1} =$

[2]In general, the distribution η used for the regression estimator is different from ζ. Furthermore, for simplicity of presentation and analysis, we assume that η is fixed for all time steps, but this is not necessary. In practice one might choose to use $\mathcal{D}_m^{(t)} = \mathcal{D}_n^{(t)}$ extracted from the demonstrated trajectories $\mathcal{D}_n$.

$\log\left(\sum_{a\in\mathcal{A}}\exp(\hat{Q}_t(x,a))\right)$. That is,

$$\hat{Q}_t \leftarrow \text{Regress}\left(\left\{\left((X_i,A_i), R_i + \hat{V}_{t+1}(X'_i)\right)\right\}_{i=1}^{m}\right).$$

Let us define $\tilde{Q}_t = r_t + \mathcal{P}^a\hat{V}_{t+1}$ and note that $\mathbb{E}\left[R_i + \hat{V}_{t+1}(X'_i)|(X_i,A_i)\right] = \tilde{Q}_t(X_i,A_i)$, i.e., $\tilde{Q}_t$ is the target regression function. We will shortly see that the quality of approximation, which is quantified by $\varepsilon_{\text{reg}}(t) \triangleq \|\hat{Q}_t - \tilde{Q}_t\|_2$, affects the excess error of approximate MaxEnt IOC procedure. One way to improve this error is by using powerful regression estimator such as the regularized least-squares estimators, similar to Regularized Fitted Q-Iteration [15]:

$$\hat{Q}_t \leftarrow \underset{Q\in\mathcal{F}^{|\mathcal{A}|}}{\text{argmin}}\frac{1}{m}\sum_{i=1}^{m}\left|Q(X_i,A_i) - \left(R_i + \hat{V}_{t+1}(X'_i)\right)\right|^2 + \lambda_{Q,m}J(Q).$$

Here $\mathcal{F}^{|\mathcal{A}|}$ is the set of action-value functions, $J(Q)$ is the regularization functional, which allows us to control the complexity, and $\lambda_{Q,m} > 0$ is the regularization coefficient. The regularizer $J(Q)$ measures the complexity of function Q. Different choices of $\mathcal{F}^{|\mathcal{A}|}$ and J lead to different notions of complexity, e.g., various definitions of smoothness, sparsity in a dictionary, etc. For example, $\mathcal{F}^{|\mathcal{A}|}$ could be a reproducing kernel Hilbert space (RKHS) and J its corresponding norm, i.e., $J(Q) = \|Q\|^2_{\mathcal{H}}$. The AVI procedure with the RKHS-based formulation is summarized in Algorithm 1. Note that one may use any other regression method in this algorithm, and the theory would still hold.

To estimate $\mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$ we may use Monte Carlo sampling: Draw a sample state from the initial distribution ν and then follow the sequence of policies π_t and count the features along the trajectory. Repeat this procedure N times (Algorithm 2). Because of the approximation of AVI, we do not have Q_t and consequently π_t, so we use $\hat{Q}_t$ and its corresponding Boltzmann policy $\hat{\pi}_t$. Therefore, instead of finding $\hat{\theta}_n$ minimizing the loss, i.e., $\nabla_\theta L(\hat{\theta}_n,\hat{b}_n) = 0$, we find a $\tilde{\theta}_n$ that makes the following "distorted" gradient of loss zero:

$$\nabla_\theta\tilde{L}(\theta,\hat{b}_n) = \frac{1}{N}\sum_{i=1}^{N}\underline{f}\left(\hat{Z}^{(i)}_{1:T}\right) - \hat{b}_n + \lambda\theta, \tag{6}$$

where $\hat{Z}^{(i)}_{1:T} \sim P_{\hat{\pi}}(Z_{1:T})$. This causes some error in the estimation of $\mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$. Also note that we do not have the true expected feature $\bar{b}$, but only $\hat{b}_n$. We would like to compare the loss of our procedure, that is $L(\tilde{\theta}_n,\hat{b}_n)$, compared to the best possible loss assuming that the log-partition function could be solved exactly, the expectation was calculated exactly, and the true expected feature vector was available, i.e., $\min_{\theta\in\mathbb{R}^d} L(\theta,\bar{b})$. Appendix A is devoted to the analysis of these sources of error in the quality of the obtained solution. Here we only report the main result.

Before presenting the result, we require a few more definitions. For $\theta, b \in \mathbb{R}^d$, define $L(\theta,b) = \log\mathcal{Z}_\theta - \langle\,\theta\,,\,b\,\rangle + \frac{\lambda}{2}\|\theta\|_2^2$. Let $\theta^* \leftarrow \text{argmin}_{\theta\in\mathbb{R}^d} L(\theta,\bar{b})$ and $\tilde{\theta}_n$ be the solution of $\nabla_\theta\tilde{L}(\tilde{\theta}_n,\hat{b}_n) = 0$. We use $\|\underline{g}(z)\|_p$ $(1 \le p \le \infty)$ to denote the usual vector space l_p-norm and we define $\|\underline{g}\|_{p,\infty} = \sup_z\|\underline{g}(z)\|_p$. We also define the following concentrability coefficients, similar to [16, 17, 18].

Definition 1 (Concentrability Coefficient of the Future-State Distribution)**.** *Given $\mu_1,\mu_2 \in \mathcal{M}(\mathcal{X})$, $k \ge 0$, and an arbitrary sequence of policies $(\pi_i)_{i=1}^k$, let $\mu_1\mathcal{P}^{\pi_1}\ldots P^{\pi_k} \in \mathcal{M}(\mathcal{X})$ denote the future-state distribution obtained when the first state is distributed according to μ_1 and then we follow the sequence of policies $(\pi_i)_{i=1}^k$. Define*

$$C_{\mu_1,\mu_2}(k) \triangleq \sup_{\pi_1,\ldots,\pi_k}\left\|\frac{\mathrm{d}\left(\mu_1\mathcal{P}^{\pi_1}\ldots P^{\pi_k}\right)}{\mathrm{d}\mu_2}\right\|_\infty.$$

If $\mu_1 \mathcal{P}^{\pi_1} \dots P^{\pi_k}$ is not absolutely continuous w.r.t. μ_2,[3] *we set $C_{\mu_1,\mu_2} = \infty$.*

Theorem 1. *Fix $\delta > 0$. Suppose that the excess error of the regression estimate at each time step $t = 1, \dots, T-1$ is upper bounded by $\varepsilon_{reg}(t) \geq \|\hat{Q}_t - \tilde{Q}_t\|_{2,2(\eta)}$. Choose an arbitrary $\mu \in \mathcal{M}(\mathcal{X})$. Define*

$$\varepsilon^2 \triangleq \left\|\underline{g}\right\|_{1,\infty}^2 (T+1) \left[\frac{|\mathcal{A}|^2}{4} \sum_{t=1}^{T-1} (T+1-t)^3 C_{\nu,\mu}^2(t-1) \sum_{k=0}^{T-t} C_{\mu,\eta}(k) \varepsilon_{reg}^2(t+k) + 4T \left(\frac{8\ln(2/\delta)}{N} + \frac{1}{N} \right) \right].$$

The excess loss is then upper bounded by

$$L(\tilde{\theta}_n, \bar{b}) - L(\theta^*, \bar{b}) \leq \frac{16 \left\|\underline{g}\right\|_{2,\infty}^2 T \left(\frac{16\ln(2/\delta)}{n} + \frac{2}{n} \right)}{\lambda} + \frac{2\sqrt{2} \left\|\underline{g}\right\|_{2,\infty} \sqrt{T} \left(\sqrt{\frac{8\ln(2/\delta)}{n}} + \frac{1}{\sqrt{n}} \right) \varepsilon}{\lambda} + \frac{\varepsilon^2}{2\lambda},$$

with probability at least $1 - \delta$.

Notice the effect of the number of demonstrated trajectories n and the value of ε on the excess loss $L(\tilde{\theta}_n, \bar{b}) - \min_\theta L(\theta, \bar{b})$. By increasing n, the first two terms in the upper bound decreases with a dominantly $O(\frac{\varepsilon}{\lambda\sqrt{n}})$ behavior. The value of ε depends on several factors including the regression errors $\varepsilon_{\text{reg}}(t)$, the number of Monte Carlo trajectories N used in the Forward pass, and the behavior of MDP characterized by the concentrability coefficients.

The regression error depends on the regression estimator we use, the number of samples m, and the intrinsic difficulty of the regression problem characterized by its smoothness, sparsity, etc. For instance, if the input space $\mathcal{X}$ is D-dimensional and the regression function is k-times smooth, i.e., it belongs to the Sobolev space $\mathbb{W}^k(\mathbb{R}^D)$, the error ε_{reg} of the optimal estimator has $O(m^{-\frac{k}{2k+D}})$ behavior. The regularized least-squares estimators can achieve optimal error rate for a large class of problems including Sobolev spaces and many RKHSs. More examples of these standard results in the statistical learning theory are reported by [19, 20]. We would like to emphasize that the analysis here is not for a specific regression estimator and one may use decision trees, random forest, deep neural networks, etc. for the task of regression.

3 Mental Simulation of Human Interactions

We validate our approach in the context of analyzing dual-agent interactions from video, in which the actions of one person are used to predict the actions of another [4]. The key idea is that dual-agent interactions can be modelled as an optimal control problem, where the actions of the initiating agent induces a cost topology over the space of reactive poses – a space in which the reactive agent plans an optimal pose trajectory. Therefore, IOC can be applied to recover this underlying reactive cost function, which allows us to simulate mental images of the reactive body pose.

A visualization of the setting is shown in Figure 1. As shown in the figure, the ground truth sequence contains both the true reaction sequence $q_{1:T} = (q_1, \dots, q_T)$ on the left hand side (LHS) and the pose sequence of the initiating agent (obervation) $o_{1:T} = (o_1, \dots, o_T)$ on the right hand side (RHS). At training time, n demonstrated interaction pairs $\{q_{1:T}^{(i)}\}_{i=1}^n$ and $\{o_{1:T}^{(i)}\}_{i=1}^n$ are provided to learn the reward model of human interaction. At test time, only the initiating actions on the

[3] For two measures μ_1 and μ_2 on the same measurable space, we say that μ_1 is absolutely continuous with respect to μ_2 (or μ_2 dominates μ_1) and denote $\mu_1 \ll \mu_2$ iff $\mu_2(A) = 0 \Rightarrow \nu_1(A) = 0$.

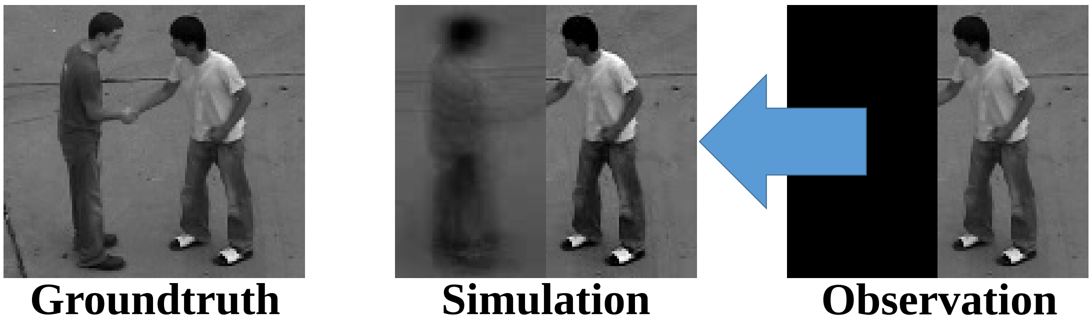


Figure 1: Examples of ground truth, partial observation, and visual simulation over occluded regions.

RHS $o_{1:T}$ are observed, and we perform inference over the previously learned reactive model to obtain an optimal reaction sequence $x_{1:T}$.

We follow [4] and model dual-agent interaction as a MDP in the following way. We use a high-dimensional HOG [21] feature of an image patch around a person as our state (pose) representation (Figure 2). The HOG feature is weighted by the probability of the foreground to filter out the background. This results in a continuous vector of 819 dimensions (64×112 bounding box). The actions are defined as the transition between states (poses), which are *deterministic* because we assume humans have perfect control over their body and one action will deterministically bring the pose to the next state.

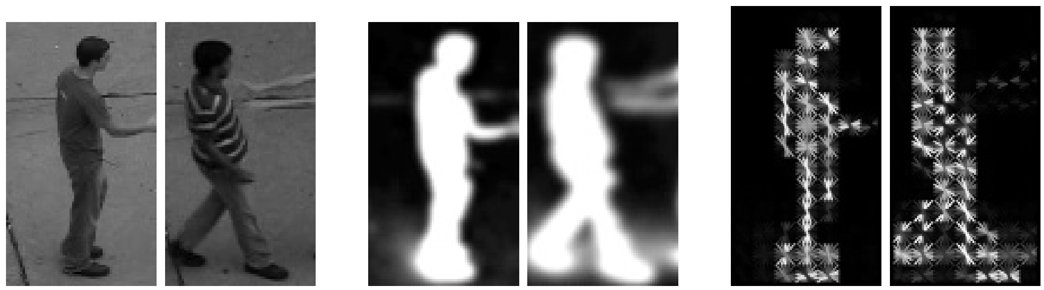
The features define the expressiveness of our cost function and are crucial to our method in modeling the dynamics of human interaction. We assume that the pose sequence $o_{1:T}$ of the initiating agent is observable on the RHS. For each frame t, we compute different features $\underline{g}_t(x, a) = (g_t^1, \ldots, g_t^d)$ from the sequence $o_{1:T}$. We modified the discrete features in [4] to adapt them to our approximate MaxEnt IOC for continuous state space.

Cooccurrence. Given a pose o_t on the RHS, we want to know how often a reactive pose x_t occurs on the LHS. This can be captured by the cooccurrence probability of poses on both LHS and RHS. We use kernel density estimation (Gaussian kernel with bandwidth 0.5) to approximate the cooccurrence probability $P_{co}(x, o)$ of LHS pose x and RHS pose o. Given a RHS pose o, we use the conditional probability $P_{co}(x|o)$ as our cooccurrence feature $g_t^1(x, a)$.

Transition. We want to know what actions will occur at a pose x, which model the probable transitions between consecutive states. Therefore, the second feature is the transition probability $g_t^2(x, a) = P_{tr}(x_a|x)$, where x_a is the state we will get to by performing action a at state x. Again, we use kernel density estimation to approximate $P_{tr}(x_a|x)$.

Smoothness. In addition to transition statistics from the training data, it is unlikely that the centroid position of human will change drastically between 2 frames and actions that induce high centroid velocity should be penalized. Therefore, we use the *smoothness* feature as $g_t^3(x, a) = 1 - \sigma(|v(x, a)|)$, where $\sigma(\cdot)$ is the sigmoid function, and $v(x, a)$ is the centroid velocity of action a at state x. These two features are independent of time step t.

Symmetry. In addition to the magnitude of centroid velocity, the relative velocity of the interacting agents is informative fir the current interaction. For example, in the hugging activity, the agents are approaching each other and will have a negative relative sign of centroid velocity. Therefore, we define two relative velocity features *attraction* and *repulsion* based on its sign. The



(a) Input image (b) Foreground map (c) HOG

Figure 2: 819-d HOG features (c) are weighted by the foreground map (b) and extracted from the original image (a).

feature *attraction* $g_t^4(x, a) = 1$ if and only if the interacting agents are moving in a symmetric way. We also define a complementary feature *repulsion* $g_t^5(x, a)$, which captures the case when the agents repel each other.

4 Experiments

Given two people interacting, we observe only the actions of the initiator on the right hand side (RHS) and attempt to simulate the reaction on the left hand side (LHS). Since the ground truth distribution over all possible reaction sequences is not available, we measure how well the learned policy is able to describe the single ground truth pose sequence. For evaluation, we used videos from three datasets, *UT-interaction 1*, *UT-interaction 2* [22], and *SBU Kinect Interaction Dataset* [23] where the UTI datasets consist of only RGB videos, and SBU dataset consists of RGB-D (color plus depth) human interaction videos. In each interaction video, we occlude the ground truth reaction $q_{1:T} = (q_1, \ldots, q_T)$ on the LHS, observe $o_{1:T} = (o_1, \ldots, o_T)$ the action of the initiating agent on the RHS, and attempt to visually simulate $q_{1:T}$.

4.1 Metrics and Baselines

We compare the ground truth sequence with the learned policy using two metrics. The first one is probabilistic, which measures the probability of performing the ground truth reaction under the learned policy. A higher probability means the learned policy is more consistent with the ground truth reaction. We use the Negative Log-Likelihood (NLL):

$$-\log P(q_{1:T}|o_{1:T}) = -\sum_t \log P(q_t|q_{t-1}, o_{1:T}), \tag{7}$$

as our metric. In a MDP, $P(q_t|q_{t-1}, o_{1:T}) = \pi_{t-1}(a_{t-1}|q_{t-1})$, where the action a_{t-1} brings q_{t-1} to q_t. The second metric is deterministic, which directly measures the physical HOG distance (or joint distances for the skeleton video) of the ground truth reaction $q_{1:T}$ and the reaction simulated by the learned policy. The deterministic metric is the average image feature distance (AFD)

$$\frac{1}{T}\sum_t ||q_t - x_t||^2 \tag{8}$$

Table 1: AFD and NLL per activity category for UTI

AFD/NLL	NN	HMM	DMDP	KRL	Ours
shake	4.57/447	5.99/285	4.33/766	5.26/467	**4.08**/**213**
hug	4.78/507	8.89/339	**3.40**/690	4.11/475	3.53/**239**
kick	6.29/283	6.03/**184**	5.34/476	5.94/286	**3.92**/197
point	3.38/399	6.16/**321**	3.20/714	3.66/382	**3.06**/391
punch	3.81/246	5.85/193	4.06/396	4.71/254	**3.44**/**145**
push	4.21/315	7.73/214	**3.75**/446	4.67/324	3.94/**145**

where x_t is the resulting reaction pose at frame t.

For model evaluation, we select four baselines to compare with the proposed method. The first baseline is the per frame nearest neighbor (NN) [24], which only uses the *co-occurrence* feature at each frame *independently* and does not take into account the temporal context of states. For each observation o_t, we find the corresponding nearest LHS state with the highest cooccurrenceas $x_t^{NN} = \arg\max_x \hat{P}_{co}(x|o_t)$.

The second baseline is the hidden Markov model (HMM) [25], which has been widely used to recover hidden time sequences given the observation. This fits our goal of simulating the hidden reactions given the observed actions of the initiating agent. HMM is defined by the transition probabilities $P(x_t|x_{t-1})$ and emission probabilities $P(o_t|x_t)$, which are equivalent to our *transition* and *cooccurrence* features. The weights for these two features are always the same in HMM, while our algorithm learns the optimal feature weights θ. We use the forward-backward algorithm to compute the likelihood. The optimal state sequence $x_{1:T}^{HMM}$ is computed by the Viterbi algorithm.

For the third baseline, we quantize the continuous state space into K discrete state by k-means clustering and apply the discrete Markov decision process (DMDP) inference used in [3]. The likelihood for MDP is computed by the stepwise product of the policy executions defined in (7).

The forth baseline is the kernel-based reinforcement learning (KRL) [26] value function approximation presented in [4], which applies kernel regression on a value function learned by MaxEnt IOC to get a continuous value function over the whole state space. For a fair comparison for value function approximation we do not implement the extended mean-shift inference proposed in [4].

4.2 Performance on 819-D HOG Space

We first evaluate our method on *UT-interaction 1*, and *UT-interaction 2* [22] datasets. The UTI datasets consists of RGB videos only, and some examples have been shown in Figure 1. The UTI datsets consist of 6 actions: hand shaking, hugging, kicking, pointing, punching, pushing. Each action has a total of 10 sequences for both datasets. We use 10-fold evaluation as in [27]. We empirically set $K = 100$ for k-means and Gaussian kernel with bandwidth 0.5 for kernel density estimation. For the regression estimator in Backward pass (Algorithm 1), we use RKHS-based regularized least-squares estimator with a Gaussian kernel (equivalent to estimating the mean function of a Gaussian process with a Gaussian covariance kernel). We set $\lambda = \lambda_{Q,n} = 0.05$ as regularization coefficients. The average NLL and image feature distance per activity for each baseline is shown in Table 1. To evaluate the accuracy of our Monte Carlo (MC) sampling algorithm, we compare with the Forward pass in [3] using our learned policy $\hat{\pi}$ ("Exact" in Table 1 and 2). Empirical results verify that our MC sampling strategy ($N = 500$) is able to achieve comparable performance. All optimal control based methods (DMDP and proposed) outperform the other two baselines in terms of image feature distance. Although the MDP is able to achieve a

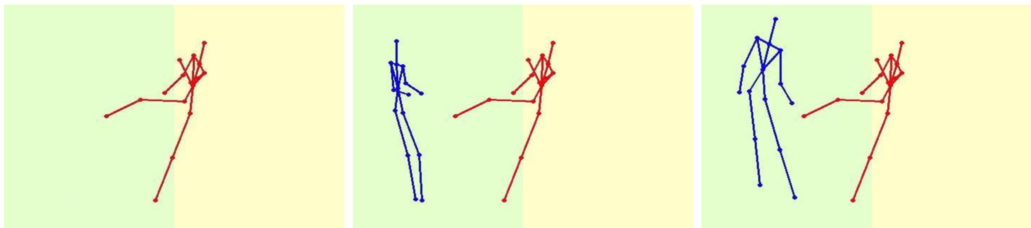


Figure 3: Observation, our simulation result, and the ground truth skeleton images of Kinect in SBU dataset.

Table 2: AFD and NLL per activity category for SBU dataset

AFD/NLL	NN	HMM	DMDP	KRL	Ours (Exact)
kick	0.81/93	0.92/92	**0.65**/88	0.92/75	0.67/**58**
push	0.51/125	0.60/127	**0.45**/119	0.61/99	0.48/**78**
shake	0.48/149	1.41/151	0.42/145	0.54/121	**0.42**/**109**
hug	0.61/137	0.67/137	0.48/132	0.81/113	**0.47**/**96**
exchange	0.63/108	3.84/112	**0.53**/104	0.74/88	0.54/**72**
punch	0.56/98	0.66/99	**0.48**/93	0.66/78	0.52/**67**

lower image feature distance than NN and HMM, the NLL is worse without proper regularization. Furthermore, the proposed approximate MaxEnt IOC consistently outperforms the KRL value function approximation. Our method directly performs IOC on the continuous state space rather than interpolating value function of discretized state space.

4.3 Performance on 45-D Human Joint Space

We extend our framework to 3D human joint space. We evaluate our method on *SBU Kinect Interaction Dataset* [23], in which interactions performed by two people are captured by a RGB-D sensor and tracked skeleton positions at each frame are provided. In this case, the state space becomes a 15×3 (joint number times x, y, z) dimensional continuous vector. The SBU dataset consist of 8 actions: approaching, departing, kicking,pushing, shaking hands, hugging, exchanging object, punching. The first two actions (approaching & departing) are excluded from our experiments because the action of the initiating agent is to stand still and provides no information for forecasting. 7 participants performed activities in the dataset and results in 21 video sets, where each set contains videos of a pair of different people performing all interactions. We use 7-fold evaluation, in which videos of one participants are held out for one fold. The average NLL and AFD per activity are shown in Table 2. Again, the proposed model performs best. We note that in this lower-dimensional problem, the quantized model (DMDP) is able to achieve comparable performance.

5 Discussion

Our experiments demonstrate that it is possible to accurately mentally simulate (extrapolate images of body pose) using the IOC framework. The results are indicative of two important application domains that are enabled by this new framework: (1) anomaly detection and (2) reasoning about activities under heavy occlusion. Since the IOC framework can be used to simulate "typical" or expected sequential visualizations of human activity, they can be compared to observed activity to detect anomalous behavior. The same framework can be used to extrapolate a sequence of human poses even when a person might be fully occluded by exiting the field of the view of the camera or stand behind an obstruction.

The task of learning the underlying reward function of a Markov decision process from observed behavior has been studied as an inverse optimal control problem [1], also called inverse reinforcement learning [28] or structural estimation [6]. In many approaches, parameters of the reward function are learned in an iterative procedure with repeated calls to a forward control or inference problem [28, 29, 1], though one may estimate the value function directly [30] or solve a single large quadratic program [29]. The work of [30], however, is developed for linearly-solvable MDPs, and more general MDPs should first be approximately embedded in the class of linearly-solvable MDPs. In addition, the rewards of linearly-solvable MDPs are assumed to be independent of actions.We follow Ziebart *et al.* [1, 2], who formalized MaxEnt IOC, showing that the soft-maximum value function can be efficiently computed with dynamic programming for problems with finite state spaces.

Several approaches for inference and learning in high-dimensional problems have been proposed. Computational efficiency is straightforward for linear dynamical systems with quadratic costs [31]. [32] leverage a related local quadratic approximation of the log-partition function for the forward problem. [33] learn local reward functions by considering a local linear-quadratic model. [14] show that in the special case of continuous paths in $\mathbb{R}^D$ and the reward function of a high-dimensional problem possessing low-dimensional structure, a globally optimal solution can be attained. In contrast with these methods, our framework considers a **global** approximation and global reward learning not limited to continuous paths in $\mathbb{R}^D$ (admitting, e.g., discrete variables or stochastic dynamics) nor a low-dimensional reward constraint, and comes with finite-sample complexity guarantees.

Our formulation focuses on the prediction of decision, but similar model can also arise from information-theoretical constraints on decision making [34, 35, 36]. In this context, Monte Carlo sampling has been utilized in [37] to approximate the path integral computation, and function approximation of the desirability function has also been explored in [38, 39]. The contribution of our work, however, lies in the combined application of these approaches to the context of learning a predictive model based on inverse reinforcement learning. Furthermore, we analyze this procedure and provide a finite-sample upper bound guarantee on the excess loss.

A Proofs

To prove the main theoretical result of this paper, Theorem 1, we have to develop several intermediate results. First, we present a general high-probability upper bound on the l_2-norm of the empirical average from the true expectation (Section A.1), which will be used in later analyses. Afterwards, we analyze the Backward Pass (Section A.2) and Forward Pass (Section A.3). Finally, we analyze the statistical properties of the MaxEnt IOC and provide a high probability upper bound on the excess error (Section A.4).

Let us first define some notations. The transition probability kernel of the MDP is $\mathcal{P}: \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$. Given a stochastic policy π, we define $\mathcal{P}^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ as $\mathcal{P}^\pi(\cdot|x) \triangleq \sum_{a \in \mathcal{A}} \pi(a|x)\mathcal{P}(\cdot|x,a)$. Sometimes we use $\mathcal{Z} = \mathcal{X} \times \mathcal{A}$ and $z = (x,a)$ as a shorthand. This should be clear from the context.

We use $\nu, \mu, \eta, \rho, \xi, \ldots$ to denote a probability distribution defined on $\mathcal{X}$. Given a probability distribution $\nu \in \mathcal{M}(\mathcal{X})$ and a probability transition kernel $\mathcal{P}^\pi$, we define the next-state probability distribution as $(\nu\mathcal{P}^\pi)(\cdot) = \int \nu(\mathrm{d}x)\mathcal{P}^\pi(\cdot|x)$.

Given $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the Boltzmann policy $\pi : \mathcal{X} \to \mathcal{M}(\mathcal{A})$ is defined as

$$\pi(a|x) = \frac{e^{Q(x,a)}}{\sum_{a' \in \mathcal{A}} e^{Q(x,a')}}.$$

Define measurable functions $\underline{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ with the identification $\underline{g}(z) = (g_1(z), \ldots, g_d(z))$ and $\underline{f} : (\mathcal{X} \times \mathcal{A})^T \to \mathbb{R}$ defined as $\underline{f}(z_{1:T}) = \sum_{t=1}^T \underline{g}(z_t)$. Also recall that we use $\left\|\underline{g}(z)\right\|_p$ $(1 \le p \le \infty)$ to denote the usual vector space l_p-norm and we define $\left\|\underline{g}\right\|_{p,\infty} = \sup_z \left\|\underline{g}(z)\right\|_p$.

Given two policies π_1, π_2, the point-wise l_1 and l_2-distances between them are defined by the usual vector norms on $\mathbb{R}^{|\mathcal{A}|}$, that is, $\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_1 = \sum_{a \in \mathcal{A}} |\pi_1(a|x) - \pi_2(a|x)|$ and $\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_2 = \sqrt{\sum_{a \in \mathcal{A}} |\pi_1(a|x) - \pi_2(a|x)|^2}$. We use similarly defined definitions for $\|Q(x,\cdot)\|_1$ and $\|Q(x,\cdot)\|_2$. Given a distribution $\mu \in \mathcal{M}(\mathcal{X})$, we define $\|\pi_1 - \pi_2\|_{1,1(\mu)} = \|\pi_1 - \pi_2\|_{1,\mu} = \int \mathrm{d}\mu(x) \sum_{a \in \mathcal{A}} |\pi_1(a|x) - \pi_2(a|x)|$, and $\|Q_1 - Q_2\|_{2,2(\mu)} = \|Q_1 - Q_2\|_\mu = \int \mathrm{d}\mu(x) \|Q_1(x,\cdot) - Q_2(x,\cdot)\|_2^2$. For two distributions $\rho_1, \rho_2 \in \mathcal{M}(\mathcal{X})$, we denote $\|\rho_1 - \rho_2\|_1 = \int |\rho_1(\mathrm{d}x) - \rho_2(\mathrm{d}x)|$.

A.1 Deviation of the l_2-Norm of the Empirical Average from the Expectation

Consider a fixed multivariate function $\Psi : \mathcal{Z} \to \mathbb{R}^d$ $(d \ge 1)$ with $\Psi(\cdot) = (\psi_1(\cdot), \ldots, \psi_d(\cdot))$. Suppose that we are given a set of n independent and identically distributed (i.i.d.) samples $\{Z_i\}_{i=1}^n$ drawn from distribution $\rho \in \mathcal{M}(\mathcal{Z})$. We would like to compare the l_2-norm between $\mathbb{E}\left[\Psi(Z)\right]$ (with $Z \sim \rho$) and $\frac{1}{n}\sum_{i=1}^n \Psi(Z_i)$. The following lemma provides such a guarantee. We will use this lemma in our further analysis.

Lemma 2. *Assume that $\|\Psi(z)\|_2 \le B$ for all $z \in \mathcal{Z}$ almost surely. For any $\delta > 0$, we have*

$$\left\| \mathbb{E}\left[\Psi(Z)\right] - \frac{1}{n}\sum_{i=1}^n \Psi(Z_i) \right\|_2 \le 2B\left[\sqrt{\frac{8\ln(1/\delta)}{n}} + \sqrt{\frac{1}{n}}\right],$$

with probability at least $1 - \delta$.

Proof. Since $\|\Psi(z)\|_2 = \sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \langle\, h\,,\, \Psi(z)\,\rangle$, we have

$$
\begin{aligned}
&\mathbb{P}\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \Psi(Z_i) - \mathbb{E}\left[\Psi(Z)\right] \right\|_2 > t \right\} = \\
&\mathbb{P}\left\{ \sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \left\langle\, h\,,\, \frac{1}{n}\sum_{i=1}^{n} \Psi(Z_i) - \mathbb{E}\left[\Psi(Z)\right] \,\right\rangle \right| > t \right\} = \\
&\mathbb{P}\left\{ \sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} \langle\, h\,,\, \Psi(Z_i)\,\rangle - \mathbb{E}\left[\langle\, h\,,\, \Psi(Z)\,\rangle\right] \right| > t \right\} = \\
&\mathbb{P}\left\{ \sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} f_h(Z_i) \right| > t \right\},
\end{aligned}
$$

in which we used the linearity of the inner product and summation to exchange their ordering in the second equality and defined and substituted $f_h(z) \triangleq \langle\, h\,,\, \Psi(z)\,\rangle - \mathbb{E}\left[\langle\, h\,,\, \Psi(Z)\,\rangle\right]$ in the last equality.

Note that $\mathbb{E}\left[f_h(Z_i)\right] = 0$. Also because $|\langle\, h\,,\, \Psi(z)\,\rangle| \leq \|\Psi(z)\|_2 \|h\|_2 \leq B\|h\|_2$ by assumption, we have $\sup_{\|h\|_2\leq 1} \sup_{z\in\mathcal{Z}} |\langle\, h\,,\, \Psi(z)\,\rangle| \leq B$. So $\sup_{z\in\mathcal{Z}} |f_h(z)| \leq 2B$ for all h that has an l_2-norm equal or smaller than 1. By Theorem 14.2 of Bühlmann and van de Geer [40] (a concentration of measure inequality), we get that for any fixed $\delta > 0$, we have

$$
\left\| \frac{1}{n}\sum_{i=1}^{n} \Psi(Z_i) - \mathbb{E}\left[\Psi(Z)\right] \right\|_2 \leq \mathbb{E}\left[\sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} f_h(Z_i) \right| \right] + 2B\sqrt{\frac{8\ln(1/\delta)}{n}}, \tag{9}
$$

with probability at least $1-\delta$.

Let $\varepsilon_1, \dots, \varepsilon_n$ be a Rademacher sequence (i.e., ε_i are i.i.d. random variables taking values in $\{-1, +1\}$ with equal probability) independent of Z_is. We use a symmetrization theorem (Theorem 14.3 of Bühlmann and van de Geer [40]) to get

$$
\mathbb{E}\left[\sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} f_h(Z_i) \right| \right] \leq 2\mathbb{E}\left[\sup_{h\in\mathbb{R}^d, \|h\|_2\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \langle\, h\,,\, \Psi(Z_i)\,\rangle \right| \right]. \tag{10}
$$

The expectation on the RHS is the Rademacher complexity of a linear function space with an l_2-constraint on the weights, i.e.,

$$
\mathcal{H}(1) = \left\{\, z \mapsto \langle\, w\,,\, \Psi(z)\,\rangle \,:\, w \in \mathbb{R}^d, \|w\|_2 \leq 1, \|\Psi(z)\|_2 \leq B \,\right\}.
$$

By Theorem 3 of Kakade et al. [41], we get that the Rademacher complexity $R_n(\mathcal{H}(1))$ is upper bounded by $\frac{B}{\sqrt{n}}$. This upper bound along (9) and (10) conclude the proof. $\square$

Remark 1. One can see this lemma as a vector-valued version of Hoeffding's inequality. Notice that the result is independent of the dimension d of the vector space.

A.2 Analysis of the Backward Pass

We provide an error propagation result for the Approximate Value Iteration (AVI) procedure of the Backward pass. The difference with results such as Munos [17], Farahmand et al. [18] is that instead of the Bellman optimality operator commonly used in Approximate Dynamic Programming (ADP) and RL, we use a Bellman operator that uses softmax (3).

In AVI, we are facing the error propagation phenomenon: Consider time $t = T$. The regression estimation error leads to having an estimate $\hat{Q}_T$ instead of $Q_T = r_T$. This in turn leads to some error in $\hat{V}_T$, which is used to estimate $\hat{Q}_{T-1}$. The estimation of $\hat{Q}_{T-1}$ has not only the usual regression estimation error, but also the error caused by having $\hat{V}_T$ instead of V_T. The same happens for the estimates in earlier iterations. The following theorem analyzes the error propagation in AVI with softmax. Before stating the theorem, recall that $\tilde{Q}_t = r_t + \mathcal{P}^a \hat{V}_{t+1}$ and $\mathbb{E}\left[R_i + \hat{V}_{t+1}(X_i')|(X_i, A_i)\right] = \tilde{Q}_t(X_i, A_i)$.

Theorem 3. *Assume that $C_{\mu,\eta}(k) < \infty$ for $k = 0, \dots, T-t$. We have*

$$\left\|Q_t - \hat{Q}_t\right\|_{2,2(\mu)}^2 \leq |\mathcal{A}|(T+1-t)\sum_{k=0}^{T-t} C_{\mu,\eta}(k)\left\|\hat{Q}_{t+k} - \tilde{Q}_{t+k}\right\|_{2,2(\eta)}^2 .$$

Proof. We have

$$\varepsilon_t \triangleq Q_t - \hat{Q}_t = Q_t - \tilde{Q}_t + \underbrace{\tilde{Q}_t - \hat{Q}_t}_{\triangleq \delta_t} = \mathcal{P}^a\left[V_{t+1} - \hat{V}_{t+1}\right] + \delta_t. \tag{11}$$

Define functions $\bar{\varepsilon}, \bar{\delta} : \mathcal{X} \to \mathbb{R}$ by $\bar{\varepsilon}(x) = \max_a |\varepsilon(x,a)|$ and $\bar{\delta}(x) = \max_a |\delta(x,a)|$.

Let us provide an upper bound on $V_{t+1} - \hat{V}_{t+1}$. Observe that

$$\begin{aligned} V_{t+1}(x') - \hat{V}_{t+1}(x') &= \log\left(\frac{\sum_{a'} \exp(Q_{t+1}(x',a'))}{\sum_{a'} \exp(\hat{Q}_{t+1}(x',a'))}\right) \\ &= \log\left(\frac{\sum_{a'} \exp(\hat{Q}_{t+1}(x',a') \cdot \exp(\varepsilon_{t+1}(x',a'))}{\sum_{a'} \exp(\hat{Q}_{t+1}(x',a')}\right). \end{aligned}$$

To simplify the notation, define vectors $\underline{w}, \underline{\varepsilon} \in \mathbb{R}^{|\mathcal{A}|}$ with the identification $w_a = \exp(\hat{Q}_{t+1}(x',a))$ and $\varepsilon_a = \varepsilon_{t+1}(x',a)$ for all $a = 1, \dots, |\mathcal{A}|$.[4] Define function $f : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$

$$f(\underline{\varepsilon}) = \log\left(\frac{\sum_a w_a \exp(\varepsilon_a)}{\sum_a w_a}\right),$$

and observe that

$$\frac{\partial f}{\partial \varepsilon_a} = \frac{w_a \exp(\varepsilon_a)}{\sum_{a'} w_a' \exp(w_a')}. \tag{12}$$

By Taylor's and mean value theorems, we have $f(\underline{\varepsilon}) = f(0) + \langle \nabla f(\xi), \xi \rangle$ for some $\xi = c\underline{\varepsilon}$ with $c \in (0,1)$. As $f(0) = 0$, we only need to upper bound $\langle \nabla f(\xi), \xi \rangle$. We use the l_1/l_∞ decomposition:

$$\langle \nabla f(\xi), \xi \rangle \leq \sup_\xi \|\nabla f(\xi)\|_1 \|\underline{\varepsilon}\|_\infty .$$

By (12), we see that $\|\nabla f(\xi)\|_1 = 1$, so

$$|f(\underline{\varepsilon})| \leq \|\underline{\varepsilon}\|_\infty = \bar{\varepsilon}_{t+1}(x'). \tag{13}$$

[4] Note that this is a slight abuse of notation as $a \in \mathcal{A}$ is not necessarily an integer between 1 to $|\mathcal{A}|$. Since $\mathcal{A}$ is finite, however, we may always define such a correspondence.

Because $V_{t+1}(x') - \hat{V}_{t+1}(x') = f(\varepsilon_{t+1}(x', \cdot))$, we write (11) as

$$\varepsilon_t(x,a) = \delta_t(x,a) + \int \mathcal{P}(\mathrm{d}x'|x,a) f(\varepsilon_{t+1}(x',\cdot)),$$

thus by (13) and the Jensen's inequality, we have

$$|\varepsilon_t(x,a)| \le |\delta_t(x,a)| + \int \mathcal{P}(\mathrm{d}x'|x,a)\bar{\varepsilon}_{t+1}(x').$$

Taking $\max_a$ from both sides, we get $\bar{\varepsilon}_t(x) \le \bar{\delta}_t(x) + \max_a \int \mathcal{P}(\mathrm{d}x'|x,a)\bar{\varepsilon}_{t+1}(x') = \bar{\delta}_t(x) + \sup_{\pi_t} \left(\mathcal{P}^{\pi_t}\bar{\varepsilon}_{t+1}\right)(x)$. Thus,

$$\begin{aligned}
\bar{\varepsilon}_t &\le \bar{\delta}_t + \sup_{\pi_t} \mathcal{P}^{\pi_t}\bar{\varepsilon}_{t+1} \le \bar{\delta}_t + \sup_{\pi_t} \mathcal{P}^{\pi_t}\left[\bar{\delta}_{t+1} + \sup_{\pi_{t+1}} \mathcal{P}^{\pi_{t+1}}\bar{\varepsilon}_{t+2}\right] \le \cdots \\
&\le \bar{\delta}_t + \sup_{\pi_t \dots \pi_{T-1}} \sum_{k=t}^{T-1} \mathcal{P}^{\pi_t} \dots \mathcal{P}^{\pi_k}\bar{\delta}_{k+1}.
\end{aligned} \tag{14}$$

To obtain the result of theorem, we would like to calculate $\mu\bar{\varepsilon}_t^2$. First, note that we can change the order of supremum and integration. To see this, consider a measurable function g and let $\tilde{\pi}$ be the policy that achieves $\sup_\pi \mathcal{P}^\pi g$, i.e., $\mathcal{P}^{\tilde{\pi}} g = \sup_\pi \mathcal{P}^\pi g$. Also let $\bar{\pi}$ be the policy that achieves $\sup_\pi \mu\mathcal{P}^\pi g$, i.e., $\mu\mathcal{P}^{\bar{\pi}} g = \sup_\pi \mu\mathcal{P}^\pi g$. We have

$$\begin{aligned}
\int \mathrm{d}\mu(x) \int \mathcal{P}(\mathrm{d}x'|x,\bar{\pi}(x))g(x') &\overset{(i)}{\ge} \int \mathrm{d}\mu(x) \int \mathcal{P}(\mathrm{d}x'|x,\tilde{\pi}(x))g(x') \\
&\overset{(ii)}{\ge} \int \mathrm{d}\mu(x) \int \mathcal{P}(\mathrm{d}x'|x,\bar{\pi}(x))g(x'),
\end{aligned}$$

where the inequality (i) is because of the optimizer property of $\bar{\pi}$ and the inequality (ii) is because of the optimizer property of $\tilde{\pi}$. This proves that $\mu \sup_\pi \mathcal{P}^\pi g = \sup_\pi \mu\mathcal{P}^\pi g$. This and (14) show that

$$\begin{aligned}
\mu\bar{\varepsilon}_t^{\,2} &\le \sup_{\pi_t \dots \pi_{T-1}} \mu \left|\bar{\delta}_t + \sum_{k=t}^{T-1} \mathcal{P}^{\pi_t} \dots \mathcal{P}^{\pi_k}\bar{\delta}_{k+1}\right|^2 \\
&\overset{\text{(i)}}{\le} (T+1-t) \sup_{\pi_t \dots \pi_{T-1}} \left[\mu\bar{\delta}_t^2 + \sum_{k=t}^{T-1} \mu\left|\mathcal{P}^{\pi_t} \dots \mathcal{P}^{\pi_k}\bar{\delta}_{k+1}\right|^2\right] \\
&\overset{\text{(ii)}}{\le} (T+1-t) \sup_{\pi_t \dots \pi_{T-1}} \left[\mu\bar{\delta}_t^2 + \sum_{k=t}^{T-1} \mu\mathcal{P}^{\pi_t} \dots \mathcal{P}^{\pi_k}\bar{\delta}_{k+1}^2\right] \\
&\le (T+1-t) \left[\mu\bar{\delta}_t^2 + \sum_{k=t}^{T-1} \sup_{\pi_t \dots \pi_k} \mu\mathcal{P}^{\pi_t} \dots \mathcal{P}^{\pi_k}\bar{\delta}_{k+1}^2\right] \\
&\overset{\text{(iii)}}{\le} (T+1-t) \left[C_{\mu,\eta}(0)\, \eta\bar{\delta}_t^2 + \sum_{k=t}^{T-1} C_{\mu,\eta}(k-t+1)\, \eta\bar{\delta}_{k+1}^2\right].
\end{aligned}$$

We used the Cauchy-Schwarz's inequality in (i), then the Jensen's inequality in (ii), and finally performed a change of measure argument and used the concentrability coefficient in (iii).

The final step is to relate $\mu\bar{\varepsilon}_t{}^2$ and $\eta\bar{\delta}^2_{k+1}$ to $\|\bar{\varepsilon}_t\|_{2,2(\mu)}$ and $\|\bar{\delta}_{k+1}\|_{2,2(\eta)}$. This can be done by noticing that $\eta\bar{\delta}^2_{k+1} = \int \mathrm{d}\eta(x) \max_a |\delta_{k+1}(x,a)|^2 \le \int \mathrm{d}\eta(x) \sum_a |\delta_{k+1}(x,a)|^2 = \|\delta_{k+1}\|^2_{2,2(\eta)}$ and $\|\varepsilon_t\|^2_{2,2(\mu)} = \int \mathrm{d}\mu(x) \sum_a |\varepsilon_t(x,a)|^2 \le |\mathcal{A}| \int \mathrm{d}\mu(x) \max_a |\varepsilon_t(x,a)|^2$. □

Remark 2. If the regression errors at all iterations is in the order of $\varepsilon_{\mathrm{reg}}$ (i.e., $\|\hat{Q}_{t+k} - \tilde{Q}_{t+k}\|_{2,2(\eta)} \approx \varepsilon_{\mathrm{reg}}$), we have $|\mathcal{A}|(T+1-t)\varepsilon^2_{\mathrm{reg}} \sum_{k=0}^{T-t} C_{\mu,\eta}(k)$ behavior.

Remark 3. It is well-known that AVI that uses Bellman optimality operator may not converge [42]. The same can happen to Backward pass too. One possible reason is that the regression error $\|\hat{Q}_t - \tilde{Q}_t\|_{2,2(\eta)}$ at some iteration t would be large. This happens when one cannot find a good approximation $\hat{Q}_t$ to its target function $\tilde{Q}_t = r_t + \mathcal{P}^a \hat{V}_{t+1}$. One possible cause is that we have function approximation error (bias), i.e., the function space $\mathcal{F}^{|\mathcal{A}|}$ used in the regression estimation is not rich enough and $\tilde{Q}_t \notin \mathcal{F}^{|\mathcal{A}|}$. Refer to the discussion of Inherent Bellman Error by Munos and Szepesvári [13] for more detail. On the other hand, we might have a large estimation error (variance), which is caused by either not having enough data samples or the function space being too rich and complex. Obviously, there is a tradeoff between estimation error and function approximation error. By choosing a powerful regression estimator that automatically balances these two sources of errors, for example through a model selection procedure, we can make sure that the regression errors are small. The question of model selection in the reinforcement learning context is discussed in detail by Farahmand and Szepesvári [43].

The other possible source of having a large error is that the concentrability coefficients $C_{\mu,\eta}(k)$ become very large. This might happen, for example, when the distribution η does not have a support in a particular region of the state space but the agent that has an initial distribution μ can go to that region. In this case, even if all $\|\hat{Q}_{t+k} - \tilde{Q}_{t+k}\|_{2,2(\eta)}$ are very small, one cannot guarantee that $\|Q_t - \hat{Q}_t\|_{2,2(\mu)}$ is small too. Refer to [17, 18, 13] for more discussion on the role of concentrability coefficients.

Remark 4. An interesting observation is that for π being the Boltzmann policy corresponding to Q,

$$\sum_{a\in\mathcal{A}} \pi(a|x)Q(x,a) = \frac{\sum_{a\in\mathcal{A}} \exp(Q(x,a))Q(x,a)}{\sum_{a'\in\mathcal{A}} \exp(Q(x,a'))} \neq \log\left(\sum_{a\in\mathcal{A}} \exp(Q_t(x,a))\right) = V(x).$$

Therefore, V should not be interpreted as the expected value of Q weighted according to policy π. So a rollout-based estimate, which follows π_t (from $t=1$ to $t=T$) and adds the rewards r_ts collected on the trajectory, does not provide an unbiased estimate of $Q_1(x,\cdot)$.

A.3 Analysis of the Forward Pass

The goal of the Forward pass is to provide an estimate of the gradient of the log-partition function $\nabla_\theta \log \mathcal{Z}_\theta = \mathbb{E}_{P_\pi(Z_{1:T})}\left[\underline{f}(Z_{1:T})\right]$. Two sources of errors affect this estimate: approximation and estimation errors. The approximation error is caused by the errors in the estimation of $(Q_t)_{t=1}^T$ by $(\hat{Q}_t)_{t=1}^T$ in the Backward pass. The estimation error, on the other hand, is caused by using Monte Carlo sampling to estimate the expectation. In this section, we analyze the effect of these errors in the calculation of the gradient of the log-partition function. The main result of this section is Theorem 11.

The setup of the Forward pass is as follows. Assume that we are given two sequences of action-value functions, $(Q)_{t=1}^T$ and $(\hat{Q})_{t=1}^T$. The former sequence is the true action-value functions, i.e., if the value iteration was done exactly in the Backward pass, while the latter sequence is for when we have approximation error in the Backward pass, which is the case in Approximate MaxEnt IOC. These sequences define corresponding Boltzmann policy sequences $(\pi_t)_{t=1}^T$ and $(\hat{\pi}_t)_{t=1}^T$.

Given an initial distribution $\nu \in \mathcal{M}(\mathcal{X})$, define two sequences $\rho_1 = \nu$, $\rho_2 = \nu\mathcal{P}^{\pi_1}$, $\rho_3 = \nu\mathcal{P}^{\pi_1}\mathcal{P}^{\pi_2}$, ..., $\rho_T = \nu\mathcal{P}^{\pi_1}\cdots\mathcal{P}^{\pi_{T-1}}$ and $\hat{\rho}_1 = \nu$, $\hat{\rho}_2 = \nu\mathcal{P}^{\hat{\pi}_1}$, $\hat{\rho}_2 = \nu\mathcal{P}^{\hat{\pi}_1}\mathcal{P}^{\hat{\pi}_2}$, ..., $\hat{\rho}_2 = \nu\mathcal{P}^{\hat{\pi}_1}\cdots\mathcal{P}^{\hat{\pi}_{T-1}}$. We denote $\underline{\rho} = (\rho_1, \dots, \rho_T)$ and $\underline{\hat{\rho}} = (\hat{\rho}_1, \dots, \hat{\rho}_T)$.

Lemma 4. *Given two action-value functions Q_1 and Q_2, and their corresponding Boltzmann policies π_1 and π_2, for any $x \in \mathcal{X}$ we have*

$$\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_2 \leq \frac{1}{2}\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_2 ,$$

and

$$\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_1 \leq \begin{cases} \frac{1}{2}\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_1 \\ \frac{\sqrt{|\mathcal{A}|}}{2}\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_2 \end{cases} .$$

We actually only use the upper bound on $\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_1$ in this paper, but we report the l_2 result as well as it might be useful in other contexts.

To prove this lemma, we require two intermediate results: a multivariate form of mean value theorem and Gershgorin Circle theorem. These are not new results, but for the sake of completeness, we report them here.

Lemma 5. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be a continuously differentiable function and $J : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ be its Jacobian matrix, that is $J_{ij} = \frac{\partial f_i(x)}{\partial x_j}$. We then have for any $x, \Delta x \in \mathbb{R}^m$,*

$$\|f(x+\Delta x) - f(x)\|_2 \leq \sup_{x'}\|J(x')\|_2\|\Delta x\|_2 ,$$
$$\|f(x+\Delta x) - f(x)\|_1 \leq \sup_{x'}\|J(x')\|_1\|\Delta x\|_1 .$$

A matrix l_1 and l_2-norms in this lemma are vector-induced norms on $\mathbb{R}^m$, and have the property that for an $m\times m$ matrix A, $\|A\|_2 = \sigma_{\max}(A)$ and $\|A\|_1 = \max_j \sum_i |A_{ij}|$.

Proof of Lemma 5. Consider a continuously differentiable function $g : \mathbb{R} \to \mathbb{R}$. By the fundamental theorem of calculus, $g(1) - g(0) = \int_0^1 g'(t)\mathrm{d}t$. For each component f_i of f, define $g_i(u) = f_i(x + u\Delta x)$, so $f_i(x+\Delta x) - f_i(x) = g_i(1) - g_i(0) = \int_0^1 g_i'(t)\mathrm{d}t = \int_0^1 \left[\sum_{j=1}^d \frac{\partial f_i}{\partial x_j}(x + t\Delta x).\Delta x_j\right]\mathrm{d}t$. For the vector-valued function f, we get $f(x+\Delta x) - f(x) = \int_0^1 J(x+t\Delta x)\Delta x\,\mathrm{d}t$, therefore,

$$\begin{aligned}\|f(x+\Delta x) - f(x)\|_2 = \left\|\int_0^1 J(x+t\Delta x)\Delta x \mathrm{d}t\right\|_2 &\leq \int_0^1 \|J(x+t\Delta x)\|_2\|\Delta x\|_2\,\mathrm{d}t \\ &\leq \sup_{x'}\|J(x')\|_2\|\Delta x\|_2\int_0^1 \mathrm{d}t.\end{aligned}$$

The l_1-norm result is obtained using the l_1-norm instead of the l_2-norm in the last step. □

Lemma 6 (Gershgorin Circle Theorem – Appendix 7 of Lax [44])**.** *Let A be an $m\times m$ complex-valued matrix. Let $r_i = \sum_{j\neq i}|A_{ij}|$. Define D_i to be the circular disc consisting of all complex numbers z satisfying $|z - A_{ii}| \leq r_i$ $(i = 1, \dots, m)$. Every eigenvalue of A is contained in one of the discs D_i.*

Equipped with these results, we are ready to prove Lemma 4.

Proof of Lemma 4. We only focus on a single state x. To simplify the notation, we use $u \in \mathbb{R}^{|\mathcal{A}|}$ to refer to $Q(x, \cdot)$. Set $p_k = p_k(u) = \pi(a_k|x) = \frac{\exp(u_k)}{\sum_i \exp(u_i)}$ for all $k = 1, \dots, |\mathcal{A}|$. We have

$$\frac{\partial p_k}{\partial u_i}(u) = \begin{cases} p_k(1 - p_k) & i = k, \\ -p_i p_k & i \neq k. \end{cases}$$

By Lemma 5, we have

$$\|p(u) - p(u_0)\|_2 \leq \sup_{u'} \|J(u')\|_2 \|\Delta u\|_2 . \tag{15}$$

For the l_2-induced norm $\|J\|_2$, we have $\|J\|_2 = \sigma_{\max}(J) = \sqrt{\lambda_{\max}(J^\top J)} = \lambda_{\max}(J)$, where the last equality is because J is symmetric and $\lambda_{\max}(J^\top J) = \lambda^2_{\max}(J)$.

To find $\lambda_{\max}(J)$, we use the Gershgorin Circle theorem (Lemma 6). Using the notation of that lemma, $r_i = p_i \sum_{j \neq i} p_j = p_i(1 - p_i)$. The centre of each circle D_i is $p_i(1 - p_i)$. Because J is symmetric, its eigenvalues are all real, so the maximum value that an eigenvalue in D_i may take on the real line is $2p_i(1 - p_i)$. So

$$\lambda_{\max} \leq \max_i 2p_i(u)(1 - p_i(u)) \leq \frac{1}{2}.$$

By setting $u_0 = Q_1(x, \cdot)$ and $u = Q_2(x, \cdot)$, alongside the above upper bound on $\lambda_{\max}$ and (15), we get that $\|\pi_2(\cdot|x) - \pi_1(\cdot|x)\|_1 = \|p(u) - p(u_0)\|_1 \leq \frac{1}{2} \|Q_1(x, \cdot) - Q_2(x, \cdot)\|_2$. We also have $\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_1 \leq \sqrt{|\mathcal{A}|} \|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_2$.

Finally, to related $\|\pi_1(\cdot|x) - \pi_2(\cdot|x)\|_1$ to $\|Q_1(x, \cdot) - Q_2(x, \cdot)\|_1$, we use the other part of Lemma 5 to get $\|p(u) - p(u_0)\|_1 \leq \sup_{u'} \|J(u')\|_1 \|\Delta u\|_1$. We then have $\|J\|_1 = \max_j \sum_i |J_{ij}| = \max_j \{p_j(1 - p_j) + p_j \sum_{i \neq j} p_j\} = 2 \max_j p_j(1 - p_j) \leq \frac{1}{2}$, which leads to the desired result. □

The next lemma upper bounds the difference in the next-state distributions induced by a mismatch in the initial distributions and policies followed by two agents.

Lemma 7. *Consider densities ρ_1 and ρ_2 over $\mathcal{X}$, policies π_1 and π_2, and their corresponding transition probability kernels $\mathcal{P}^{\pi_1}$ and $\mathcal{P}^{\pi_2}$. We have*

$$\|\rho_1 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_2}\|_1 \leq \|\rho_1 - \rho_2\|_1 + \|\pi_1 - \pi_2\|_{1,1(\rho_2)} .$$

Proof. By the triangle inequality, we have $\|\rho_1 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_2}\|_1 \leq \|\rho_1 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_1}\|_1 + \|\rho_2 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_2}\|_1$. We upper bound each term. For the first one, we have

$$\begin{aligned}
\|\rho_1 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_1}\|_1 &= \int_y \left| \int_x (\rho_1(\mathrm{d}x) - \rho_2(\mathrm{d}x)) \mathcal{P}^{\pi_1}(\mathrm{d}y|x) \right| \\
&\leq \int_y \int_x |\rho_1(\mathrm{d}x) - \rho_2(\mathrm{d}x)| \, \mathcal{P}^{\pi_1}(\mathrm{d}y|x) \\
&= \int_x |\rho_1(\mathrm{d}x) - \rho_2(\mathrm{d}x)| \underbrace{\int_y \mathcal{P}^{\pi_1}(\mathrm{d}y|x)}_{=1} = \int_x |\rho_1(\mathrm{d}x) - \rho_2(\mathrm{d}x)| \\
&= \|\rho_1 - \rho_2\|_1 .
\end{aligned} \tag{16}$$

Here we used the Jensen's inequality. Similarly, for the second term, we have

$$
\begin{aligned}
\|\rho_2 \mathcal{P}^{\pi_1} - \rho_2 \mathcal{P}^{\pi_2}\|_1 &= \int_y \left| \int_x \rho_2(\mathrm{d}x) \left(\mathcal{P}^{\pi_1}(\mathrm{d}y|x) - \mathcal{P}^{\pi_2}(\mathrm{d}y|x)\right) \right| \\
&\leq \int_x \rho_2(\mathrm{d}x) \int_y \left|\mathcal{P}^{\pi_1}(\mathrm{d}y|x) - \mathcal{P}^{\pi_2}(\mathrm{d}y|x)\right| . \\
&= \int_x \rho_2(\mathrm{d}x) \int_y \left| \sum_a \mathcal{P}(\mathrm{d}y|x,a)(\pi_1(a|x) - \pi_2(a|x)) \right| \\
&\leq \int_x \rho_2(\mathrm{d}x) \int_y \sum_a \mathcal{P}(\mathrm{d}y|x,a)|\pi_1(a|x) - \pi_2(a|x)| \\
&\leq \int_x \rho_2(\mathrm{d}x) \sum_a |\pi_1(a|x) - \pi_2(a|x)| \underbrace{\int_y \mathcal{P}(\mathrm{d}y|x,a)}_{=1} \\
&= \|\pi_1 - \pi_2\|_{1,1(\rho_2)} .
\end{aligned}
$$

□

The next lemma provides an upper bound on the distribution mismatch between ρ_{k+1} and $\hat{\rho}_{k+1}$.

Lemma 8. *Let $\rho_1 = \hat{\rho}_1 = \nu$ and $\rho_{k+1} = \nu \mathcal{P}^{\pi_1} \ldots P^{\pi_k}$ and $\hat{\rho}_{k+1} = \nu \mathcal{P}^{\hat{\pi}_1} \ldots P^{\hat{\pi}_k}$. Assume that $C_{\nu,\mu}(i) < \infty$ for $i = 0, \ldots, k-1$. We then have*

$$
\|\rho_{k+1} - \hat{\rho}_{k+1}\|_1 \leq \sum_{i=1}^{k} \|\pi_i - \hat{\pi}_i\|_{1,1(\rho_i)} \leq \sum_{i=1}^{k} C_{\nu,\mu}(i-1) \|\pi_i - \hat{\pi}_i\|_{1,1(\mu)} .
$$

Proof. First, we apply Lemma 7 recursively:

$$
\begin{aligned}
\|\rho_{k+1} - \hat{\rho}_{k+1}\|_1 &\leq \|\pi_k - \hat{\pi}_k\|_{1,1(\rho_k)} + \|\rho_k - \hat{\rho}_k\|_1 \\
&\leq \|\pi_k - \hat{\pi}_k\|_{1,1(\rho_k)} + \|\pi_{k-1} - \hat{\pi}_{k-1}\|_{1,1(\rho_{k-1})} + \|\rho_{k-1} - \hat{\rho}_{k-1}\|_1 \leq \cdots \\
&\leq \sum_{i=1}^{k} \|\pi_i - \hat{\pi}_i\|_{1,1(\rho_i)} .
\end{aligned}
$$

Afterward, we do a change of measure argument:

$$
\begin{aligned}
\int \mathrm{d}\rho_i(x) \|\pi_i(\cdot|x) - \hat{\pi}_i(\cdot|x)\|_1 &= \int \frac{\mathrm{d}\rho_i(x)}{\mathrm{d}\mu} \mathrm{d}\mu(x) \|\pi_i(\cdot|x) - \hat{\pi}_i(\cdot|x)\|_1 \\
&\leq C_{\nu,\mu}(i-1) \|\pi_i - \hat{\pi}_i\|_{1,1(\mu)} .
\end{aligned}
$$

□

We are now ready to state the following result, which shows the effect of $\|\pi_k - \hat{\pi}_k\|_1$ on the approximation error $\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat{\rho}}}\left[\underline{f}(Z_{1:T})\right]\|$ caused by the distribution mismatch.

Lemma 9. *Assume that* $C_{\nu,\mu}(t) < \infty$ *for* $t = 0, \dots, T-1$ *and* $\left\|\underline{g}\right\|_{1,\infty} < \infty$. *We then have*

$$\begin{aligned}\left\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(Z_{1:T})\right]\right\|_2 &\le \left\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(Z_{1:T})\right]\right\|_1 \\ &\le \left\|\underline{g}\right\|_{1,\infty} \sum_{t=1}^{T-1} (T-t) \left\|\pi_t - \hat\pi_t\right\|_{1,1(\rho_t)} \\ &\le \left\|\underline{g}\right\|_{1,\infty} \sum_{t=1}^{T-1} (T-t) C_{\nu,\mu}(t-1) \left\|\pi_t - \hat\pi_t\right\|_{1,1(\mu)} .\end{aligned}$$

The assumption on the concentrability coefficients is only required for the last inequality.

Proof. We expand $\underline{f}(z_{1:T})$ as $\sum_{t=1}^{T} \underline{g}(z_t)$ and use the Jensen's inequality to get

$$\begin{aligned}\left\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(Z_{1:T})\right]\right\|_1 &= \sum_{j=1}^{d} \left| \sum_{t=1}^{T} \int g_j(x_t) \left(\rho_t(\mathrm{d}x_t) - \hat\rho_t(\mathrm{d}x_t)\right)\right| \\ &\le \sum_{j=1}^{d} \sum_{t=1}^{T} \int |g_j(x_t)| \left|\rho_t(\mathrm{d}x_t) - \hat\rho_t(\mathrm{d}x_t)\right| \\ &= \sum_{t=1}^{T} \int \sum_{j=1}^{d} |g_j(x_t)| \left|\rho_t(\mathrm{d}x_t) - \hat\rho_t(\mathrm{d}x_t)\right| \\ &\le \sup_z \left\|\underline{g}(z)\right\|_1 \sum_{t=1}^{T} \left\|\rho_t - \hat\rho_t\right\|_1 .\end{aligned}$$

This inequality alongside Lemma 8 show that

$$\begin{aligned}\left\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(Z_{1:T})\right]\right\|_1 &\le \left\|\underline{g}\right\|_{1,\infty} \sum_{t=1}^{T} \sum_{i=1}^{t-1} \left\|\pi_i - \hat\pi_i\right\|_{1,1(\rho_i)} \\ &= \left\|\underline{g}\right\|_{1,\infty} \sum_{t=1}^{T-1} (T-t) \left\|\pi_t - \hat\pi_t\right\|_{1,1(\rho_i)} \\ &\le \left\|\underline{g}\right\|_{1,\infty} \sum_{t=1}^{T-1} (T-t) C_{\nu,\mu}(t-1) \left\|\pi_t - \hat\pi_t\right\|_{1,1(\mu)} ,\end{aligned}$$

where in the last inequality we used a change of measure argument and noted that $\rho_i = \nu \mathcal{P}^{\pi_1} \dots P^{\pi_{i-1}}$.

Finally for any finite-dimensional vector v, we have $\|v\|_2 \le \|v\|_1$, so $\|\mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(Z_{1:T})\right]\|_2$ is upper bounded by the same quantity too. $\square$

Let us turn to studying the estimation error caused by the Monte Carlo procedure in the Forward pass. The setup is as follows: We have N independent sample trajectories $\hat Z_{1:T}^{(i)} = (\hat Z_1^{(i)}, \dots, \hat Z_T^{(i)})$ $(i = 1, \dots, N)$ that are generated by sampling from the initial distribution $\nu \in \mathcal{M}(\mathcal{X})$ and then following the policy sequence $(\hat\pi_t)_{t=1}^{T-1}$. The underlying distribution of these samples are $\underline{\hat\rho} = (\hat\rho_1, \dots, \hat\rho_T)$ with $\hat\rho_1 = \nu$, $\hat\rho_2 = \nu \mathcal{P}^{\hat\pi_1}$, $\hat\rho_3 = \nu \mathcal{P}^{\hat\pi_1} \mathcal{P}^{\hat\pi_2}$, etc. The following result shows how far the vector of empirical averages $\frac{1}{N} \sum_{i=1}^{N} \underline{f}(\hat Z_{1:T}^{(i)})$ deviates from the true expectation $\mathbb{E}_{\underline{\hat\rho}}\left[\underline{f}(\hat Z_{1:T})\right]$ in the l_2-norm.

Lemma 10. *For any fixed $\delta > 0$, we have*

$$\left\| \mathbb{E}_{\underline{\hat{\rho}}}\left[\underline{f}(\hat{Z}_{1:T})\right] - \frac{1}{N}\sum_{i=1}^{N} \underline{f}(\hat{Z}_{1:T}^{(i)}) \right\|_2 \leq 2\sqrt{T}\,\|\underline{g}\|_{2,\infty}\left[\sqrt{\frac{8\ln(1/\delta)}{N}} + \frac{1}{\sqrt{N}}\right],$$

with probability at least $1-\delta$.

Proof. Define $x = (z_1, \dots, z_t)$ and set $X_i = (\hat{Z}_1^{(i)}, \dots, \hat{Z}_T^{(i)})$ (for $i = 1, \dots, N$). The samples X_i are i.i.d. random vectors drawn from $\underline{\hat{\rho}}$. For any z,

$$\|\underline{f}(z)\|_2^2 = \sum_{j=1}^{d}\left|\sum_{t=1}^{T} g_j(z_t)\right|^2 \leq \sum_{j=1}^{d}\sum_{t=1}^{T} g_j^2(z_t) = \sum_{t=1}^{T}\sum_{j=1}^{d} g_j^2(z_t) \leq T\,\|\underline{g}\|_{2,\infty}.$$

Apply Lemma 2 with $B = T\,\|\underline{g}\|_{2,\infty}$ to get the desired result. □

Equipped with Lemmas 4, 9 and 10, we now state the main result of this section.

Theorem 11. *(Assumptions of Part I) Given two policy sequences $(\pi_t)_{t=1}^{T-1}$ and $(\hat{\pi}_t)_{t=1}^{T-1}$, let $\underline{\rho}, \underline{\hat{\rho}} \in \mathcal{M}\left((\mathcal{X}\times\mathcal{A})^T\right)$ be defined as described earlier. Suppose that $\hat{Z}_{1:T}^{(i)} = (\hat{Z}_1^{(i)}, \dots, \hat{Z}_T^{(i)})$ $(i = 1, \dots, N)$ are sampled trajectories from $\underline{\hat{\rho}}$. Assume that $C_{\nu,\mu}(t) < \infty$ for $t = 0, \dots, T-1$ and $\|\underline{g}\|_{1,\infty} < \infty$. (Assumptions of Part II) Furthermore, suppose that policies π_t and $\hat{\pi}_t$ are Boltzmann policies corresponding to Q_t and $\hat{Q}_t$ $(t = 1, \dots, T-1)$, i.e., $\pi_t(a|x) = \frac{e^{Q_t(x,a)}}{\sum_{a'\in\mathcal{A}} e^{Q_t(x,a')}}$ (and the same relation between $\hat{\pi}_t$ and $\hat{Q}_t$). For any fixed $\delta > 0$, it holds that*

$$\left\| \mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \frac{1}{N}\sum_{i=1}^{N} \underline{f}(\hat{Z}_{1:T}^{(i)}) \right\|_2^2 \overset{\text{Part I}}{\leq}$$
$$\|\underline{g}\|_{1,\infty}^2 (T+1)\left[\sum_{t=1}^{T-1}(T-t)^2 C_{\nu,\mu}^2(t-1)\,\|\pi_t - \hat{\pi}_t\|_{1,1(\mu)}^2 + 4T\left(\frac{8\ln(1/\delta)}{N} + \frac{1}{N}\right)\right] \overset{\text{Part II}}{\leq}$$
$$\|\underline{g}\|_{1,\infty}^2 (T+1)\left[\frac{|\mathcal{A}|}{4}\sum_{t=1}^{T-1}(T-t)^2 C_{\nu,\mu}^2(t-1)\left\|Q_t - \hat{Q}_t\right\|_{2,2(\mu)}^2 + 4T\left(\frac{8\ln(1/\delta)}{N} + \frac{1}{N}\right)\right],$$

with probability at least $1-\delta$.

Proof. Fix $\delta > 0$ and evoke Lemmas 9 and 10, and notice that $\|\underline{g}\|_{2,\infty} \leq \|\underline{g}\|_{1,\infty}$ to get

$$\left\| \mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \frac{1}{N}\sum_{i=1}^{N} \underline{f}(\hat{Z}_{1:T}^{(i)}) \right\|_2 \leq$$
$$\|\underline{g}\|_{1,\infty}\left[\sum_{t=1}^{T-1}(T-t) C_{\nu,\mu}(t-1)\,\|\pi_t - \hat{\pi}_t\|_{1,1(\mu)} + 2\sqrt{T}\left(\sqrt{\frac{8\ln(1/\delta)}{N}} + \frac{1}{\sqrt{N}}\right)\right],$$

with probability at least $1-\delta$. By the Cauchy-Schwarz's inequality, we have $(\sum_{i=1}^{T}|a_i|)^2 \leq$

$T\sum_{i=1}^{T} a_i^2$, so with the same probability,

$$\left\| \mathbb{E}_{\underline{\rho}}\left[\underline{f}(Z_{1:T})\right] - \frac{1}{N}\sum_{i=1}^{N} \underline{f}(\hat{Z}_{1:T}^{(i)}) \right\|_2^2 \leq$$
$$\left\|\underline{g}\right\|_{1,\infty}^2 (T+1)\left[\sum_{t=1}^{T-1}(T-t)^2 C_{\nu,\mu}^2(t-1)\left\|\pi_t - \hat{\pi}_t\right\|_{1,1(\mu)}^2 + 4T\left(\frac{8\ln(1/\delta)}{N} + \frac{1}{N}\right)\right].$$

Finally, we apply Lemma 4 to upper bound $\|\pi_t - \hat{\pi}_t\|_{1,1(\mu)}^2 \leq \int \mathrm{d}\mu(x)\, \|\pi_t(\cdot|x) - \hat{\pi}_t(\cdot|x)\|_1^2$ by $\int \mathrm{d}\mu(x) \frac{|\mathcal{A}|}{4} \|Q_t(x,\cdot) - \hat{Q}_t(x,\cdot)\|_2^2 = \frac{|\mathcal{A}|}{4}\|Q_t - \hat{Q}_t\|_{2,2(\mu)}^2$ for all $t = 1, \ldots, T-1$. $\square$

A.4 Analysis of the Regularized MaxEnt IOC

Recall from Section 2 that we had a set of demonstrated trajectories $\mathcal{D}_n = \{Z_{1:T}^{(i)}\}_{i=1}^n$ with each trajectory $Z_{1:T} = (Z_1, \ldots, Z_T) \sim \zeta$ with $Z_t = (X_t, A_t)$. We defined $\hat{b}_n, \bar{b} \in \mathbb{R}^d$ as $\hat{b}_n = \frac{1}{n}\sum_{i=1}^n \underline{f}(Z_{1:T}^{(i)})$ and $\bar{b} = \mathbb{E}_{Z_{1:T}\sim\zeta}\left[\underline{f}(Z_{1:T})\right]$. For any $\theta, b \in \mathbb{R}^d$, we define the loss function as

$$L(\theta, b) = \log \mathcal{Z}_\theta - \langle\, \theta\, ,\, b\, \rangle + \frac{\lambda}{2}\|\theta\|_2^2.$$

Approximate MaxEnt IOC finds $\tilde{\theta}_n$ that makes the following "distorted" gradient of loss zero (cf. (6)):

$$\nabla_\theta \tilde{L}(\theta, \hat{b}_n) = \frac{1}{N}\sum_{i=1}^{N} \underline{f}\left(\hat{Z}_{1:T}^{(i)}\right) - \hat{b}_n + \lambda\theta, \quad \tilde{Z}_{1:T}^{(i)} \sim P_{\hat{\pi}}(Z_{1:T}).$$

We let $\theta^* \leftarrow \operatorname{argmin}_{\theta\in\mathbb{R}^d} L(\theta, \bar{b})$ (the ideal minimizer), $\hat{\theta}_n \leftarrow \operatorname{argmin}_{\theta\in\mathbb{R}^d} L(\theta, \hat{b}_n)$ (the minimizer with empirical average $\hat{b}_n$, which comes from the true underlying distribution), and $\tilde{\theta}_n$ be the solution of $\nabla_\theta \tilde{L}(\tilde{\theta}_n, \hat{b}_n) = 0$ (the minimizer with empirical average based on distorted distribution).

The error in the gradient estimation leads to an error in the empirical loss. The following lemma relates these quantities.

Lemma 12. *Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be as defined above. Assume that $\|\nabla_\theta \tilde{L}(\tilde{\theta}_n, \hat{b}_n) - \nabla_\theta L(\tilde{\theta}_n, \hat{b}_n)\| \leq \varepsilon$. We then have*

$$L(\tilde{\theta}_n, \hat{b}_n) \leq L(\hat{\theta}_n, \hat{b}_n) + \frac{\varepsilon^2}{2\lambda}.$$

Proof. First note that $h(\theta) \triangleq L(\theta, \hat{b}_n)$ is λ-strongly convex in θ as $-\left\langle\, \theta\, ,\, \hat{b}_n\, \right\rangle$ is linear, the Hessian of $\log \mathcal{Z}_\theta$ is the covariance matrix of $\bar{f}(Z_{1:T})$, which is positive semi-definite, so $\log \mathcal{Z}_\theta$ is convex, and $\frac{\lambda}{2}\|\theta\|_2^2$ is λ-strongly convex. Thus for any $\theta, \theta' \in \mathbb{R}^d$, we have

$$\begin{aligned}
h(\theta') &= h(\theta) + \nabla_\theta^\top h(\theta)(\theta' - \theta) + \frac{1}{2}(\theta' - \theta)^\top \nabla_\theta^2 h(\theta'')(\theta' - \theta) \\
&\geq h(\theta) + \nabla_\theta^\top h(\theta)(\theta' - \theta) + \frac{\lambda}{2}\|\theta' - \theta\|_2^2 \\
&\geq h(\theta) + \nabla_\theta^\top h(\theta)(\theta_* - \theta) + \frac{\lambda}{2}\|\theta_* - \theta\|_2^2 \\
&= h(\theta) - \frac{1}{2\lambda}\|\nabla_\theta h(\theta)\|_2^2,
\end{aligned}$$

where $\theta'' = \alpha\theta + (1-\alpha)\theta'$ with some $\alpha \in (0,1)$, and where $\theta_* = \theta - \frac{\nabla_\theta h(\theta)}{\lambda}$ minimizes the RHS (Section 9.1.2 of Boyd and Vandenberghe [45]). We set $\theta = \tilde{\theta}_n$ and $\theta' = \hat{\theta}_n$ to get

$$L(\tilde{\theta}_n, \hat{b}_n) \leq L(\hat{\theta}_n, \hat{b}_n) + \frac{1}{2\lambda} \left\| \nabla_\theta L(\tilde{\theta}_n, \hat{b}_n) \right\|_2^2$$

Recalling the assumption $\|\nabla_\theta \tilde{L}(\tilde{\theta}_n, \hat{b}_n) - \nabla_\theta L(\tilde{\theta}_n, \hat{b}_n)\| \leq \varepsilon$ and noticing that $\nabla_\theta \tilde{L}(\tilde{\theta}_n, \hat{b}_n) = 0$ lead to the desired result. $\square$

We are ready to state the main result of this section and the paper.

Theorem 13. *Let $\hat{\theta}_n$, $\tilde{\theta}_n$, and θ^* be as defined above. Assume that $\|\nabla_\theta \tilde{L}(\tilde{\theta}_n, \hat{b}_n) - \nabla_\theta L(\tilde{\theta}_n, \hat{b}_n)\| \leq \varepsilon$. Fix $\delta_1 > 0$. The excess loss is upper bounded by*

$$L(\tilde{\theta}_n, \bar{b}) - L(\theta^*, \bar{b}) \leq \frac{16 \left\|\underline{g}\right\|_{2,\infty}^2 T \left(\frac{16\ln(1/\delta_1)}{n} + \frac{2}{n} \right)}{\lambda} + \frac{2\sqrt{2}\left\|\underline{g}\right\|_{2,\infty} \sqrt{T} \left(\sqrt{\frac{8\ln(1/\delta_1)}{n}} + \frac{1}{\sqrt{n}} \right) \varepsilon}{\lambda} + \frac{\varepsilon^2}{2\lambda},$$

with probability at least $1-\delta_1$. Furthermore, suppose that the excess error of the regression estimate at each time step $t = 1, \ldots, T-1$ is upper bounded by $\varepsilon_{reg}(t) \geq \|\hat{Q}_t - \tilde{Q}_t\|_{2,2(\eta)}$. Choose an arbitrary $\mu \in \mathcal{M}(\mathcal{X})$. For any fixed $\delta_2 > 0$, ε can then be upper bounded as

$$\varepsilon^2 \leq \left\|\underline{g}\right\|_{1,\infty}^2 (T+1) \left[\frac{|\mathcal{A}|^2}{4} \sum_{t=1}^{T-1} (T+1-t)^3 C_{\nu,\mu}^2(t-1) \sum_{k=0}^{T-t} C_{\mu,\eta}(k) \varepsilon_{reg}^2(t+k) + 4T \left(\frac{8\ln(1/\delta_2)}{N} + \frac{1}{N} \right) \right],$$

with probability at least $1-\delta_2$.

Proof. We have

$$\begin{aligned}
e(\tilde{\theta}_n) \triangleq L(\tilde{\theta}_n, \bar{b}) - L(\theta^*, \bar{b}) &= L(\tilde{\theta}_n, \bar{b}) - L(\tilde{\theta}_n, \hat{b}_n) + L(\tilde{\theta}_n, \hat{b}_n) - L(\hat{\theta}_n, \hat{b}_n) + \\
&\qquad \underbrace{L(\hat{\theta}_n, \hat{b}_n) - L(\theta^*, \hat{b}_n)}_{\leq 0} + L(\theta^*, \hat{b}_n) - L(\theta^*, \bar{b}) \\
&\leq \left\langle \tilde{\theta}_n \, , \, \hat{b}_n - \bar{b} \right\rangle + \frac{\varepsilon^2}{2\lambda} + \left\langle \theta^* \, , \, \bar{b} - \hat{b}_n \right\rangle \\
&\leq \left\langle \tilde{\theta}_n - \theta^* \, , \, \hat{b}_n - \bar{b} \right\rangle + \frac{\varepsilon^2}{2\lambda} \\
&\leq \left\| \tilde{\theta}_n - \theta^* \right\|_2 \left\| \hat{b}_n - \bar{b} \right\|_2 + \frac{\varepsilon^2}{2\lambda},
\end{aligned} \tag{17}$$

where we used the optimizer property of $\hat{\theta}_n$ to get $L(\hat{\theta}_n, \hat{b}_n) - L(\theta^*, \hat{b}_n) \leq 0$ and we evoked Lemma 12 to upper bound $L(\tilde{\theta}_n, \hat{b}_n) - L(\hat{\theta}_n, \hat{b}_n)$. This decomposition is similar to what is used by Altun and Smola [12].

We upper bound $\|\tilde{\theta}_n - \theta^*\|_2$ by benefitting from the λ-strong convexity of $L(\theta, b)$ w.r.t. θ. Similar to the proof of Lemma 12, we get that for any θ' and θ_0, we have $L(\theta', b) - L(\theta_0, b) \geq \langle \nabla_\theta L(\theta_0) \, , \, \theta' - \theta_0 \rangle + \frac{\lambda}{2} \|\theta' - \theta_0\|_2^2$. In particular, if $\theta_0 \leftarrow \operatorname{argmin}_\theta L(\theta, b)$, we have $\nabla_\theta L(\theta_0) = 0$, so $L(\theta', b) - L(\theta_0, b) \geq \frac{\lambda}{2} \|\theta' - \theta_0\|_2^2$. Choose $b = \bar{b}$, $\theta_0 = \theta^*$, and $\theta' = \tilde{\theta}_n$, along (17) to get

$$\left\| \tilde{\theta}_n - \theta^* \right\|_2^2 \leq \frac{2e(\tilde{\theta}_n)}{\lambda} \leq \frac{2}{\lambda} \left[\underbrace{\left\| \tilde{\theta}_n - \theta^* \right\|_2 \left\| \hat{b}_n - \bar{b} \right\|_2}_{=(a)} + \underbrace{\frac{\varepsilon^2}{2\lambda}}_{=(b)} \right].$$

Two cases might happen:
Case 1: If $(a) \geq (b)$, we have $\|\tilde{\theta}_n - \theta^*\|_2^2 \leq \frac{2\times 2}{\lambda}\|\tilde{\theta}_n - \theta^*\|_2 \|\hat{b}_n - \bar{b}\|_2$, so $\|\tilde{\theta}_n - \theta^*\|_2 \leq \frac{4\|\hat{b}_n - \bar{b}\|_2}{\lambda}$.
Case 2: If $(a) < (b)$, we have $\|\tilde{\theta}_n - \theta^*\|_2 \leq \frac{\sqrt{2}\varepsilon}{\lambda}$.

From these two cases, we have

$$\left\|\tilde{\theta}_n - \theta^*\right\|_2 \leq \frac{4\left\|\hat{b}_n - \bar{b}\right\|_2 + \sqrt{2}\varepsilon}{\lambda} \tag{18}$$

From Lemma 10 (with appropriate modification of changing the sampling distribution to ζ instead of $\hat{\rho}$ in that result), we get that for any fixed $\delta_1 > 0$,

$$\left\|\hat{b}_n - \bar{b}\right\|_2 \leq 2\left\|\underline{g}\right\|_{2,\infty}\sqrt{T}\left[\sqrt{\frac{8\ln(1/\delta_1)}{n}} + \frac{1}{\sqrt{n}}\right],$$

with probability at least $1 - \delta_1$. This upper bound alongside (17) and (18) prove the first part. The second part is the direct result of Theorems 3 and 11 with some minor simplifications. □

Theorem 1 is essentially a summarized version of this theorem.

Acknowledgments

This research was sponsored in part by the Army Research Laboratory (W911NF-10-2-0061), the National Science Foundation (Purposeful Prediction: Co-robot Interaction via Understanding Intent and Goals), and the Natural Sciences and Engineering Research Council of Canada.

References

[1] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008. 1, 2, 11

[2] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. The principle of maximum causal entropy for estimating interacting processes. *IEEE Trans. on Information Theory*, 59(4): 1966–1980, April 2013. ISSN 0018-9448. 1, 2, 3, 11

[3] Kris M. Kitani, Brian D. Ziebart, J. Andrew Bagnell, and Martial Hebert. Activity forecasting. In *ECCV*, 2012. 1, 9

[4] De-An Huang and Kris M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*, 2014. 2, 6, 7, 9

[5] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014. 2

[6] John Rust. Structural estimation of Markov decision processes. *Handbook of econometrics*, 4 (4), 1994. 2, 11

[7] Gerald Tesauro. TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6:215–219, 1994. 2

[8] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *JMLR*, 6:503–556, 2005. 2, 4

[9] Martin Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In *ECML*, pages 317–328, 2005. 2, 4

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. 2

[11] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *COLT*, volume 3120, pages 472–486. 2004. 3

[12] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006. 3, 23

[13] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *JMLR*, 9: 815–857, 2008. 4, 16

[14] Paul Vernaza and J Andrew Bagnell. Efficient high dimensional maximum entropy modeling via symmetric partition functions. In *NIPS*, pages 584–592, 2012. 4, 11

[15] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian Decision Problems. In *ACC*, pages 725–730, June 2009. 5

[16] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, pages 267–274, 2002. 5

[17] Rémi Munos. Performance bounds in L_p norm for approximate value iteration. *SIAM Journal on Control and Optimization*, pages 541–561, 2007. 5, 13, 16

[18] Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *NIPS*. 2010. 5, 13, 16

[19] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. 2002. 6

[20] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008. 6

[21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 7

[22] Michael S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010. 8, 9

[23] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPRW*, 2012. 8, 10

[24] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13(1):21–27, 1967. 9

[25] Lawrence Rabiner and Biing-Hwang Juang. An introduction to hidden Markov models. *ASSP Magazine*, 1986. 9

[26] Dirk Ormoneit and [Sacute]aunak Sen. Kernel based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002. 9

[27] Yu Cao, Daniel Paul Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven J. Dickinson, Jeffrey Mark Siskind, and Song Wang. Recognize human activities from partially observed videos. In *CVPR*, 2013. 9

[28] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004. 11

[29] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *ICML*, 2006. 11

[30] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable mdps. In *ICML*, 2010. 11

[31] Brian D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, 2010. 11

[32] Anca Dragan and Siddhartha Srinivasa. Formalizing assistive teleoperation. In *RSS*, 2012. 11

[33] Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. In *ICML*, 2012. 11

[34] Arnab Nilim and Laurent E Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *NIPS*, 2003. 11

[35] E. Todorov. Linearly-solvable Markov decision problems. In *NIPS*. 2006. 11

[36] E. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *IEEE CDC*, 2012. 11

[37] Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11), 2005. 11

[38] Emanuel Todorov. Eigenfunction approximation methods for linearly-solvable optimal control problems. In *ADPRL*, pages 161–168. IEEE, 2009. 11

[39] Mingyuan Zhong and Emanuel Todorov. Moving least-squares approximations for linearly-solvable stochastic optimal control problems. *Journal of Control Theory and Applications*, 9 (3):451–463, 2011. 11

[40] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011. 13

[41] Sham Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NIPS - 21)*, 2009. 13

[42] John N. Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94, 1996. 16

[43] Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning Journal*, 85(3):299–332, 2011. 16

[44] Peter D. Lax. *Linear Algebra and Its Applications*. Wiley, 2nd edition, 2007. 17

[45] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004. 23