Approximate MaxEnt Inverse Optimal Control

De-An Huang
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Amir-massoud Farahmand*
Mitsubishi Electric Research Laboratories
Cambridge, MA, USA

Kris M. Kitani
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

J. Andrew Bagnell
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Maximum entropy inverse optimal control (MaxEnt IOC) is an effective means of discovering the underlying cost function of demonstrated agent's activity. To enable inference in 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. This approach has a finite-sample error upper bound guarantee 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 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]

Keywords: Inverse Optimal Control, Inverse Reinforcement Learning, Maximum Entropy Principle, Human Activity Recognition, Finite-Sample Error Bound

Acknowledgements

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.

*This work was done while the author was at the Robotics Institute, Carnegie Mellon University.
[1]This work is a summary of Huang et al. [1].

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 [2, 3]. For instance, Kitani et al. [4] used MaxEnt IOC in the context of understanding and modeling human activities, where the recovered reward model encodes a person's set of preferences. They considered the computer vision problem of mentally (visually) simulate human activities. By integrating visual attributes of the scene as features of the reward function, they showed that highly accurate pedestrian trajectories could be simulated in novel scenes.

Most current approaches of MaxEnt IOC, however, are limited to problems with small state space. So for example, its application to visual prediction problems has been limited to 2D pedestrian trajectories. To extend MaxEnt IOC to deal with the inherent high-dimensional nature of observed human activity from image data, previous approaches [5, 6] 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. To address this problem, we recently introduced an *approximate MaxEnt IOC* algorithm that is suitable for dealing with problems with large or high-dimensional state space [1]. This paper summarizes the algorithm, the theoretical results, and the application of the framework in the context of analyzing dual-agent interactions from video.

At the heart of the problem of maximum entropy MaxEnt IOC and sequence prediction is an inference problem of computing the *log-partition function* that 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 [7] and formalized in [2]. While dynamic programming renders this efficient for small scale problems, more appropriate techniques are needed for dealing with problems with large state space.

When the state space is large, one natural approach is to use approximate dynamic programming for the approximate calculation of these functions. The approximate MaxEnt IOC algorithm of this paper in fact uses Approximate Value Iteration (AVI) to compute the softmax-based value (log-partition) function. The AVI procedure uses a regression estimator at each iteration. The choice of regression estimator is flexible and one can choose to work with local averagers, random forests, boosting, deep neural networks, etc. In this work, we particularly utilize a reproducing kernel Hilbert space-based (RKHS) regularized estimator due to its flexibility and favourable properties. We also briefly mention the theoretical properties of this algorithm and 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, which is also known as inverse reinforcement learning, is to recover an agent's (or expert's) reward function based on its policy (or samples from the agent's behavior) when the dynamics of the process is known. Our approach to IOC is based on the Maximum Entropy Inverse Optimal Control of [3]. Let us 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 state space, $\mathcal{A}$ 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.[2] In our approach, the state space $\mathcal{X}$ can be large, e.g., $\mathbb{R}^D$. We consider θ-MDPs with finite horizon of T. Given a sequence $z_{1:T} = (z_1, \ldots, z_T)$, 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 $Z_{1:T} = (Z_1, \ldots, Z_T) \sim \zeta$ with $Z_t = (X_t, A_t)$ and ζ being a distribution over the set of trajectory. Also denote $\nu \in \mathcal{M}(\mathcal{X})$ as the distribution of X_1. We only assume that the initial distribution ν is known, but the joint distribution ζ is not. 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 [3] is

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

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

Algorithm 1 – Backward pass

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

Algorithm 2 – Forward pass

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

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 [3])

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

in which $\log \mathcal{Z}_\theta$ is the log-partition function (Theorem 3 of [3]). 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 $r_t(x, a) = r(x, a) = \left\langle \theta \,,\, \underline{g}(x, a) \right\rangle$, and perform the following VI procedure: Set $Q_T = r_T$, and for $t = T-1, \ldots, 1$,

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

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 $\log \mathcal{Z}_\theta = \mathbb{E}_\nu\left[V_1(X)\right]$. 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 the original dual objective (2), we aim to solve the 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 [8]. 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 [9]. The second insight, which is also used in some previous work such as [10], 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.[3] Given these samples, one can estimate Q_t with $\hat{Q}_t$ by

[3] In general, the distribution η used for the regression estimator is different from ζ. Furthermore, for simplicity of presentation we consider 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 $\mathcal{D}_n$.

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 $\hat{V}_{t+1} = \log\left(\sum_{a\in\mathcal{A}} \exp(\hat{Q}_t(x,a))\right)$. 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.

The choice of regression estimator is flexible in approximate MaxEnt IOC algorithm. It is desirable to use a powerful estimator that makes $\varepsilon_{\text{reg}}(t)$ as small as possible. One such a choice is the family of regularized least-squares estimators, which is also used in the Regularized Fitted Q-Iteration algorithm for control [11]: $\hat{Q}_t \leftarrow \text{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}} \frac{1}{m}\sum_{i=1}^m \left|Q(X_i, A_i) - \left(R_i + \hat{V}_{t+1}(X'_i)\right)\right|^2 + \lambda_{Q,m} J(Q)$. 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. One particular choice is $\mathcal{F}^{|\mathcal{A}|}$ being a reproducing kernel Hilbert space (RKHS) and J being its corresponding norm, i.e., $J(Q) = \|Q\|^2_{\mathcal{H}}$. The AVI procedure with the RKHS-based formulation is summarized in Algorithm 1.

To estimate $\mathbb{E}_{P_\pi(Z_{1:T})}\left[f(Z_{1:T})\right]$ 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 as well as the error caused by the Monte Carlo sampling, the solution $\tilde{\theta}_n$ of the approximate MaxEnt IOC procedure would have an error. We compare its loss to the ideal, but unavailable, case when the log-partition function could be solved exactly, the expectation was calculated exactly, and the true expected feature vector was available, i.e., $\min_{\theta\in\mathbb{R}^d} L(\theta, \bar{b})$. Huang et al. [1] provides a finite-sample error upper bound guarantee that compares 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., $\min_{\theta\in\mathbb{R}^d} L(\theta, \bar{b})$. Under certain reasonable simplifying assumptions, the result is that for any $\delta > 0$, it holds that

$$L(\tilde{\theta}_n, \bar{b}) - \min_{\theta\in\mathbb{R}^d} L(\theta, \bar{b}) \le c\frac{T^3\sqrt{T\ln(1/\delta)}}{\lambda\sqrt{n}}\,\varepsilon_{\text{reg}},$$

with probability at least $1-\delta$.[4] Here ε_{reg} is an upper bound on the sequence $\left(\varepsilon_{\text{reg}}(t)\right)_{t=1}^{T-1}$. The value of $c > 0$ depends on the MDP and distributions ν and η through concentrability coefficients [12, 13, 14], and the number of actions. The regression error ε_{reg} 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.

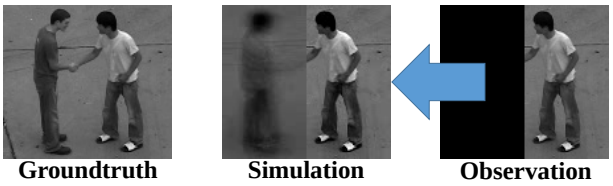
3 Mental Simulation of Human Interactions

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 [5]. 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 1a. As shown in the figure, the ground truth sequence contains both the true reaction sequence $q_{1:T} = (q_1, \ldots, q_T)$ on the left hand side (LHS) and the pose sequence of the initiating agent (obervation) $o_{1:T} = (o_1, \ldots, o_T)$ on the right hand side (RHS). At training time, n demonstrated interaction pairs $\{q^{(i)}_{1:T}\}_{i=1}^n$ and $\{o^{(i)}_{1:T}\}_{i=1}^n$ are provided to learn the reward model of human interaction. At test time, only the initiating actions on the 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 [5] and model dual-agent interaction as an MDP in the following way. We use an 819-dimensional HOG feature [15] of an image patch around a person as our state (pose) representation. 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.

Given two people interacting, we observe only the actions of the initiator on the RHS and attempt to simulate the reaction on the LHS. For evaluation, we used videos from *UT-interaction 1*, *UT-interaction 2* datasets [16]. The UTI datasets consist of RGB videos and has six actions: hand shaking, hugging, kicking, pointing, punching, pushing. In each interaction

[4] The simplifications are that $N \ge nT$, the number of samples m used in the regression estimation is in the same order as n, and the regression problem is not trivial, so ε_{reg} does not go to zero faster than $1/\sqrt{m}$. These are all reasonable assumptions.



(a)

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**

(b)

Figure 1: (a) Examples of ground truth, partial observation, and visual simulation over occluded regions. (b) AFD and NLL per activity category for UTI

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 $q_{1:T}$. Our baselines for comparisons are per frame nearest neighbor (NN), a hidden Markov model (HMM), discretized MDP-based (DMDP) MaxEnt IOC formulation [4], and the smoothing kernel-based RL (KRL) approach to MaxEnt IOC [5]. We compare the ground truth sequence with the learned policy using two metrics of Negative Log-Likelihood (NLL) $-\log P(q_{1:T}|o_{1:T}) = -\sum_t \log P(q_t|q_{t-1}, o_{1:T})$ and the average image feature distance (AFD) $\frac{1}{T}\sum_t ||q_t - x_t||^2$, where x_t is the resulting reaction pose at frame t.

The average NLL and image feature distance per activity for each baseline is shown in Figure 1b. To evaluate the accuracy of our Monte Carlo (MC) sampling algorithm, we compare with the Forward pass in [4] using our learned policy $\hat{\pi}$. Empirical results verify that our MC sampling strategy ($N = 500$) is able to achieve comparable performance. All optimal control based methods (DMDP and the proposed method) outperform the other two baselines in terms of image feature distance. Although the DMDP is able to achieve a lower than NN and HMM image feature distance, its NLL is worse then theirs. 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. For more details, additional experiments, and comparison with other IOC and IRL algorithms refer to Huang et al. [1].

References

[1] De-An Huang, Amir-massoud Farahmand, Kris M Kitani, and J. Andrew Bagnell. Approximate MaxEnt inverse optimal control and its application for mental simulation of human interactions. In *AAAI*, January 2015.

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

[3] 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.

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

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

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

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

[8] 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.

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

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

[11] 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.

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

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

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

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

[16] 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.