Learning Parameterized Prescription Policies and Disease Progression Dynamics using Markov Decision Processes

Henghui Zhu[1], Tingting Xu[1] and Ioannis Ch. Paschalidis[2]

***Abstract*— We develop an algorithm for learning physicians' prescription policies and the disease progression dynamics from Electronic Health Record (EHR) data. The prescription protocol used by physicians is viewed as a control policy which is a function of an underlying disease state in a Markov Decision Process (MDP) framework. We assume that the transition probabilities and the policy of the MDP are parameterized using some known features, such that only a small portion of them are informative. Two ℓ_1-regularized maximum likelihood estimation problems are formulated to learn the transition probabilities and the policy, respectively. A bound is established on the difference between the average reward of the estimated policy under the estimated transition dynamics and the original (unknown) policy under the true transition dynamics. Our result suggests that by using only a relatively small number of training samples, the estimate can achieve a low regret. We validate our theoretical results on a test MDP motivated by a disease treatment identification application.**

I. Introduction

Electronic health records (EHRs) are widely used in the U.S. healthcare system. They contain not only a large amount of patients' demographic information such as age and race, but also longitudinal medical history including patients' disease states and physicians' prescriptions. However, due to the limited tools and time, it is difficult for physicians to make use of such massive amounts of data to tackle medical problems. A meaningful but challenging problem is to learn the prevailing prescription policy from EHRs and to predict the disease progression of a patient under various prescriptions, which can enable physicians to identify a prescription policy that optimizes the patient's long-term health. In addition, for certain rare diseases, the amount of relevant records may be small and an algorithmic approach able to learn from limited data is desirable. Our objective is to resolve these problems by utilizing EHR data.

There are many ways to mine EHR data. This paper considers modeling the disease progression dynamics and physicians' prescription policy using an MDP model. We propose two sparse logistic regression models to obtain the best estimates of the MDP dynamics and the physician's policy from EHR data. We assume a parameterized "Boltzmann-type" policy and a parameterized "Boltzmann" conditional transition probability law in the MDP model. The parameters are obtained through maximum likelihood estimation, where we introduce an ℓ_1-norm regularization on the parameters according to the assumption that only a small portion of the features, both for the conditional transition probabilities and the policy, are informative. We prove that by using only $\Omega(\log(n)\text{poly}(1/\varepsilon))$ examples, where $\text{poly}(\cdot)$ indicates a polynomial function and n is the number of features used in modeling the conditional transition probabilities or the policy, our estimates achieve regret with order $O(\sqrt{\varepsilon})$. Consequently, the estimated policy exhibits good performance using only a relatively small number of training samples, which is particularly meaningful for some rare diseases for which there are limited available samples.

A. Related work

Mining EHR data to predict patients' future states has been studied by using a Markov chain model [1], Bayesian networks [2], and deep learning [3]. While there are many settings where this prediction could be useful, the above methods have limited power in learning prescriptions. On the other hand, although some recent works, including [4], conduct prescription analysis based on EHRs, the sequential nature of events in the EHRs is not taken into account.

Our previous works [5], [6] consider a problem of learning a policy from demonstrations where the exact MDP model is known. In this paper, we consider the substantially more general case where the MDP dynamics are unknown and need to be learned from the data.

B. Contributions

We propose two logistic regression models for estimating both the conditional transition probability of the MDP and the agent's policy. A bound on the distance between the target parameters and their estimates is derived. Having a good estimate for transition probabilities is important, since we can simulate an MDP with the estimated transition probabilities and obtain the average reward of the estimated policy. A bound on the regret is then obtained, which captures the difference between the estimated average reward and the average reward of the unknown policy in the original (unknown) MDP model. In contrast to our previous related work [5], [6], our algorithms in this paper make it possible to obtain a good average reward estimate when the MDP dynamics are not known.

* Research partially supported by the NSF under grants DMS-1664644, CNS-1645681, CCF-1527292, and IIS-1237022, by the ONR under grant MURI N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, by the Boston University Digital Health Initiative and by the Center for Information and Systems Engineering.

[1] H. Zhu and T. Xu are with the Center for Information and Systems Engineering, Boston University, {henghuiz, tingxu}@bu.edu.

[2] I. Ch. Paschalidis is with the Department of Electrical and Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, yannisp@bu.edu, http://sites.bu.edu/paschalidis/.

The rest of the paper is organized as follows. In Sec. II, we introduce our notation and the MDP model. In Sec. III, we formulate the learning problem and describe the proposed algorithm. In Sec. IV, we establish our main result on both the difference of the target parameters and their estimates and the regret of the estimated policy under the estimated MDP dynamics. In Sec. V, we conduct a simulation analysis based on a dynamic healthcare model and illustrate our theoretical results. Conclusions can be found in Sec. VI.

II. Notation and Preliminaries

We use bold letters to denote vectors and matrices; vectors are lowercase and matrices uppercase. Vectors are column vectors unless explicitly stated otherwise. Prime denotes transpose and $\|\mathbf{a}\|_p = (\sum_{i=1}^n |a_i|^p)^{1/p}$ represents the p-norm of the vector $\mathbf{a}$. We use script letters to denote sets.

We consider a finite-state Markov Decision Process (MDP), denoted by a four-tuple $(\mathscr{X}, \mathscr{A}, \mathbf{P}, R)$. $\mathscr{X}$ is the MDP state-space, e.g., patients' state of disease. $\mathscr{A}$ is the set of possible actions, e.g., physicians' prescriptions for patients. For any state-action pair $(x,a) \in \mathscr{X} \times \mathscr{A}$, $\mathbf{P}(y|x,a)$ denotes the conditional transition probability from state x to state y after taking action a. The function R denotes the one-step reward function of the MDP, e.g., a utility function of a patient. Finally, a policy function μ is a function that maps each state x to a distribution of actions and $\mu(a|x)$ denotes the probability of taking action a at state x.

Particularly for MDPs with large state-action spaces, we approximate the conditional transition probabilities and policy function using a class of functions. We consider the following class of Boltzmann-type function approximations:

$$\mathbf{P}_\xi(y|x,a) = \frac{\exp\{\xi' \psi(x,a,y)\}}{\sum_{z \in \mathscr{N}_x} \exp\{\xi' \psi(x,a,z)\}}, \tag{1}$$

$$\mu_\theta(a|x) = \frac{\exp\{\theta' \phi(x,a)\}}{\sum_{b \in \mathscr{A}} \exp\{\theta' \phi(x,b)\}}, \tag{2}$$

where the functions $\psi : \mathscr{X} \times \mathscr{A} \times \mathscr{X} \to [0,1]^N$ and $\phi : \mathscr{X} \times \mathscr{A} \to [0,1]^n$ are features mapping state-action-state tuples (x,a,y) and state-action pairs (x,a), respectively. $\mathscr{N}_x$ is the set of all possible next states starting at state x and $\xi \in \mathbb{R}^N$ and $\theta \in \mathbb{R}^n$ are parameter vectors. In the interest of brevity, we will refer to transition probabilities ξ and mean transition probabilities induced by ξ. Similarly, we will say policy θ and mean the policy with parameter θ.

Consider an MDP with conditional transition probabilities $\mathbf{P}_\xi$ using policy θ. Then the state x of the MDP follows a Markov chain, and we denote its transition matrix as $\mathbf{M}_{\xi,\theta}$, where $\mathbf{M}_{\xi,\theta}(y|x) = \sum_{a \in \mathscr{A}} \mu_\theta(a|x) \mathbf{P}_\xi(y|x,a)$ for all state pairs (x,y). Since the conditional transition probability and the policy are Boltzmann-type, the Markov chain has a unique stationary distribution denoted by $\pi_{\xi,\theta}(x)$. Therefore, the state-action pair (x,a) also has a unique stationary distribution, and we denote it by $\eta_{\xi,\theta}(x,a) = \pi_{\xi,\theta}(x)\mu_\theta(a|x)$. Moreover, we can define the average reward function as $\overline{R}(\xi,\theta) = \sum_{(a,x)} \eta_{\xi,\theta}(x,a) R(x,a)$, where R is the one step reward function.

III. Problem Formulation

In many real-world settings of reinforcement learning, including ones in the healthcare domain, we have access to observations of states visited by an agent and the corresponding actions. We do not though know the dynamics of the system and are not able to simulate it. In this case, we wish to learn both the policy of the agent and the conditional transitional probabilities of the MDP from data.

In particular, consider a target conditional transition probability ξ^* and a target policy θ^*, which are not necessarily optimal. We seek to learn both these parameter vectors. Denote by $\mathscr{S} := \mathscr{S}(\xi^*, \theta^*) = \{(x_i, a_i, y_i) : i = 1, \ldots, m\}$ samples obtained by playing policy θ^* in the MDP with conditional transition probabilities ξ^*. We assume a set of state-action samples $\{(x_i, a_t) : i = 1, \ldots, m\}$ that are i.i.d. and drawn from the stationary distribution $\eta_{\xi,\theta}$. The next state y_i is chosen according to transition probability ξ^* under state x_i and action a_i. Then, the samples $\{(x_i, a_i, y_i)\}$ in $\mathscr{S}$ are i.i.d. according to the distribution $\mathscr{D} \sim \mathbf{P}_{\xi^*}(y|x,a)\eta_{\xi^*,\theta^*}(x,a)$. The goal of this paper is to learn the conditional transition probability matrix $\mathbf{P}_{\xi^*}$ and the target policy θ^* from the demonstrations in $\mathscr{S}$.

This paper assumes that the parameter vectors are sparse, which means there are only $q < N$ non-zero elements in ξ and $r < n$ non-zero elements in θ. This is because it is relatively easy to include many features, while only a few of them may play a very important role. We further assume that θ and ξ are bounded by K, elementwise.

Given the parameterizations in (1) and (2), logistic regression is suitable for learning the parameters from data. We introduce an ℓ_1 constraint to induce sparse estimates. Following [7], [5], we regress the conditional transition probabilities as follows:

$$\begin{aligned} \max_{\xi \in \Re^N} \quad & \sum_{i=1}^m \log \mathbf{P}_\xi(y_i|x_i,a_i) \\ \text{s.t.} \quad & \|\xi\|_1 \le B_\xi, \end{aligned} \tag{3}$$

where B_ξ is a parameter that controls the sparsity of the parameter vector estimate. Similarly, we regress the policy of the agent using

$$\begin{aligned} \max_{\theta \in \mathscr{R}^n} \quad & \sum_{i=1}^m \log \mu_\theta(a_i|x_i) \\ \text{s.t} \quad & \|\theta\|_1 \le B_\theta, \end{aligned} \tag{4}$$

where B_θ is again a parameter affecting the sparsity of the solution. Training from data is formulated as Algorithm 1.

IV. Main Results

In this section, we establish theoretical results on the performance of our proposed algorithm. We will first obtain a bound on the target parameter vector and its estimate under Algorithm 1. Then we will establish a bound on the regret in the reward of the MDP.

We first define the Kullback-Leibler (KL) divergence between two conditional transition probability vectors and two policies. The KL divergence characterizes the difference

Algorithm 1 Training algorithm to estimate the target policy transition parameter ξ^* (or target policy θ^*) from the samples $\mathscr{S}$.

Initialization : Fix $0 < \gamma < 1$ and $C > qK$ (or $C > rK$).
Split the data set $\mathscr{S}$ into two sets $\mathscr{S}_1$ and $\mathscr{S}_2$ of size γm and $(1-\gamma)m$ respectively. $\mathscr{S}_1$ is used for training and $\mathscr{S}_2$ is used for cross-validation.
Training:
for $B = 0, 1, \ldots, C$ **do**
Solve the optimization problem (3) (or (4)) for each B on the set $\mathscr{S}_1$, and let ξ_B (or θ_B) denote the optimal solution.
end for
Validation: Among the ξ_B's (or θ_B's) from the training step, select the one with the lowest "hold-out" error on $\mathscr{S}_2$, i.e, $\hat{B} = \arg\min_{B\in\{0,1,\ldots,C\}} \varepsilon_{\mathscr{S}_2}(\xi_B)$ and set $\hat{\xi} = \xi_{\hat{B}}$ (or $\hat{B} = \arg\min_{B\in\{0,1,\ldots,C\}} \zeta_{\mathscr{S}_2}(\theta_B)$ and set $\hat{\theta} = \theta_{\hat{B}}$), where $\varepsilon_{\mathscr{S}_2}(\cdot)$ and $\zeta_{\mathscr{S}_2}(\cdot)$ denote the expected negative log-likelihood of the transition probabilities and policy function on the set $\mathscr{S}_2$, respectively.

between two distributions. For conditional transition probability vectors ξ_1 and ξ_2 at (x,a), the KL divergence between them is

$$D(\mathbf{P}_{\xi_1}(\cdot|x,a)\|\mathbf{P}_{\xi_2}(\cdot|x,a)) = \sum_y \mathbf{P}_{\xi_1}(y|x,a)\log\frac{\mathbf{P}_{\xi_1}(y|x,a)}{\mathbf{P}_{\xi_2}(y|x,a)}.$$

And for policies θ_1 and θ_2 at state x, the KL divergence between them is

$$D(\mu_{\theta_1}(\cdot|x)\|\mu_{\theta_2}(\cdot|x)) = \sum_a \mu_{\theta_1}(a|x)\log\frac{\mu_{\theta_1}(a|x)}{\mu_{\theta_2}(a|x)}.$$

Next, we need to define the average KL divergence under some distribution. Recall $\pi_{\xi,\theta}$ and $\eta_{\xi,\theta}$ denote the stationary distribution of the Markov chains for states and state-action pairs under some conditional transition probability ξ and policy θ. Then, the average KL divergence for conditional transition probabilities $D_{\xi,\theta}(\mathbf{P}_{\xi_1}\|\mathbf{P}_{\xi_2})$ and policies $D_{\xi,\theta}(\mu_{\theta_1}\|\mu_{\theta_2})$ are defined as follows:

$$D_{\xi,\theta}(\mathbf{P}_{\xi_1}\|\mathbf{P}_{\xi_2}) = \sum_{x,a}\eta_{\xi,\theta}(x,a)D\left(\mathbf{P}_{\xi_1}(\cdot|x,a)\|\mathbf{P}_{\xi_2}(\cdot|x,a)\right),$$
$$D_{\xi,\theta}(\mu_{\theta_1}\|\mu_{\theta_2}) = \sum_x \pi_{\xi,\theta}(x)D(\mu_{\theta_1}(\cdot|x)\|\mu_{\theta_2}(\cdot|x)).$$

In [6], we obtained a bound of the average KL divergence between the target policy μ_{θ^*} and its estimate $\mu_{\hat{\theta}}$ as follows.

Theorem 1 ([6]): Let $\varepsilon > 0$ and $\delta > 0$. In order to guarantee that with probability at least $1-\delta$, $\hat{\theta}$ produced by Algorithm 1 performs as well as θ^*, i.e.,

$$D_{\xi^*,\theta^*}(\mu_{\theta^*}\|\mu_{\hat{\theta}}) \le \varepsilon, \tag{5}$$

it suffices that

$$m = \Omega\Big((\log n)\cdot \mathrm{poly}(r, K, C, H, \log(1/\delta), 1/\varepsilon)\Big),$$

where H is the maximum number of actions per state for the MDP, and $\mathrm{poly}(x)$ denotes a polynomial in elements of x. Specifically, in terms of only H, $m = \Omega(H^3)$.

Similarly, we bound the average KL divergence between the target conditional transition probability $\mathbf{P}_{\xi^*}$ and its estimate $\mathbf{P}_{\hat{\xi}}$ in the following corollary. The proof is similar to the one in Theorem 1 and hence omitted.

Corollary 2: Let $\varepsilon > 0$ and $\delta > 0$. In order to guarantee that with probability at least $1-\delta$, $\hat{\xi}$ produced by Algorithm 1 performs as well as ξ^*, i.e.,

$$D_{\xi^*,\theta^*}(\mathbf{P}_{\xi^*}\|\mathbf{P}_{\hat{\xi}}) \le \varepsilon, \tag{6}$$

it suffices that

$$m = \Omega\Big((\log N)\cdot \mathrm{poly}(q, K, C, M, \log(1/\delta), 1/\varepsilon)\Big),$$

where M is defined as $M = \max_x |\mathscr{N}_x|$. Specifically, in terms of only M, $m = \Omega(M^3)$.

Next, we develop a bound on regret of the MDP. In particular, the regret $\mathrm{Reg}(\mathscr{S})$ is defined as

$$\mathrm{Reg}(\mathscr{S}) = \overline{R}(\xi^*, \theta^*) - \overline{R}(\hat{\xi}, \hat{\theta}),$$

where $\hat{\xi}$ and $\hat{\theta}$ are the estimated conditional transition probability and policy parameters from the samples $\mathscr{S}$, respectively.

In our previous work [6], we have established a bound on the regret when the transition probability of the MDP is known. As we argued, in the healthcare application we focus on, it is difficult to know what is the exact disease progression model. Also, for most systems, it is expensive, even impossible, to perform simulations. For example, if we wish to know what is the average reward of a policy estimate in a disease progression model, we can not just let patients follow this policy and observe their status over a long period of time. Alternatively, if we have an estimate of the transition probability of the MDP, we can easily simulate the MDP with the transition probability estimate and find the reward of the estimated policy. Then a natural question is, what is the difference between the average reward in the simulated MDP with estimated transition probabilities and policy and the average reward in the actual MDP under the true target policy. To arrive at an answer, we need some definitions and lemmas.

Definition 1 ([5]): The fundamental matrix of a Markov chain with transition probability matrix $\mathbf{M}_{\xi,\theta}$ induced by conditional transition probability parameter ξ and policy θ is

$$\mathbf{Z}_{\xi,\theta} = (\mathbf{A}_{\xi,\theta} + \mathbf{e}\pi'_{\xi,\theta})^{-1},$$

where $\mathbf{e}$ denotes the vector of all 1's, $\mathbf{A}_{\xi,\theta} = \mathbf{I} - \mathbf{M}_{\xi,\theta}$ and $\pi_{\xi,\theta}$ denotes the stationary distribution associated with $\mathbf{M}_{\xi,\theta}$.

Definition 2 ([5]): The group inverse of a square matrix $\mathbf{A}$ denoted as $\mathbf{A}^{\#}$ is the unique matrix satisfying

$$\mathbf{A}\mathbf{A}^{\#}\mathbf{A} = \mathbf{A},\ \mathbf{A}^{\#}\mathbf{A}\mathbf{A}^{\#} = \mathbf{A},\ \mathbf{A}\mathbf{A}^{\#} = \mathbf{A}^{\#}\mathbf{A}.$$

Definition 3 ([5]): Consider a matrix $\mathbf{B}$ with equal row sums. Its ergodic coefficient is defined as

$$\tau(\mathbf{B}) = \sup_{\mathbf{v}'\mathbf{e}=0;\|\mathbf{v}\|_1=1} \|\mathbf{v}'\mathbf{B}\|_1 = \frac{1}{2}\max_{i,j}\sum_s |b_{is} - b_{js}|. \tag{7}$$

Lemma 3: ([8, Lemma 11.6.1]) Given any two probability vectors $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^n$, we have

$$D(\mathbf{p}_1 \| \mathbf{p}_2) \geq \frac{1}{2\ln 2} \|\mathbf{p}_1 - \mathbf{p}_2\|_1^2. \tag{8}$$

Lemma 4 ([5]): For the stochastic matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, assume π_1 and π_2 are their unique stationary distributions, respectively. Let $\mathbf{E} = \mathbf{P}_1 - \mathbf{P}_2$. Then,

$$\|\pi_1 - \pi_2\|_1 \leq \kappa \|\pi_1' \mathbf{E}\|_1, \tag{9}$$

where κ is a constant that can take one of the following values: $\kappa = \|\mathbf{Z}_2\|_1$, or $\kappa = \|\mathbf{A}_2^{\#}\|_1$, or $\kappa = 1/(1-\tau(\mathbf{P}_2))$, or $\kappa = \tau(\mathbf{Z}_2) = \tau(\mathbf{A}_2^{\#})$.

We now have all the ingredients for our main result.

Theorem 5: Given $\varepsilon > 0$ and $\delta > 0$, suppose $m = \Omega((\log(\max(N,n))) \cdot \mathrm{poly}(r, K, C, M, \log(1/\delta), 1/\varepsilon, H)$ i.i.d. samples are used by Algorithm 1 to estimate $\hat{\xi}$ and $\hat{\theta}$. Then, with probability of at least $1-\delta$, we have

$$|\overline{R}(\xi^*, \theta^*) - \overline{R}(\hat{\xi}, \hat{\theta})| \leq 2\sqrt{\ln 2 \varepsilon} R_{\max}(1 + 2\kappa),$$

where $R_{\max} = \max_{(x,a) \in \mathscr{X} \times \mathscr{A}} |R(x,a)|$, κ is a constant that depends on the conditional transition probability $\hat{\xi}$ and policy $\hat{\theta}$, and κ can be any of the following: $\kappa = \|\mathbf{Z}_2\|_1$, or $\kappa = \|\mathbf{A}_2^{\#}\|_1$, or $\kappa = 1/(1-\tau(\mathbf{P}_2))$, or $\kappa = \tau(\mathbf{Z}_2) = \tau(\mathbf{A}_2^{\#})$.

Proof: We will follow the line of development in the proof to Theorem III.3 in [5]. First, we bound the regret as the sum of two parts:

$$\begin{aligned}
&\mathrm{Reg}(\mathscr{S}) = \overline{R}(\xi^*, \theta^*) - \overline{R}(\hat{\xi}, \hat{\theta}) \\
=&\sum_x \sum_a [\eta_{\xi^*,\theta^*}(x,a) - \eta_{\hat{\xi},\hat{\theta}}(x,a)] R(x,a) \\
=&\sum_x \pi_{\xi^*,\theta^*}(x) \sum_a \mu_{\theta^*}(a|x) R(x,a) \\
&- \sum_x \pi_{\hat{\xi},\hat{\theta}}(x) \sum_a \mu_{\hat{\theta}}(a|x) R(x,a) \\
=&\sum_x \pi_{\xi^*,\theta^*}(x) \sum_a [\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)] R(x,a) \\
&- \sum_x [\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)] \sum_a \mu_{\hat{\theta}}(a|x) R(x,a) \\
\leq&\left| \sum_x \pi_{\xi^*,\theta^*}(x) \sum_a [\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)] R(x,a) \right| \\
+&\left| \sum_x [\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)] \sum_a \mu_{\hat{\theta}}(a|x) R(x,a) \right|. \qquad (10)
\end{aligned}$$

Note that the first absolute sum has terms $\sum_a [\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)]$ for all x that are related to the estimation error in fitting the policy policy $\hat{\theta}$ to θ^*. The second part has terms $\sum_x |\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)|$ that are related to the perturbation of the stationary distribution of the Markov chain by applying the fitted policy $\hat{\theta}$. In the following, we bound each term separately. We begin with the first term:

$$\begin{aligned}
&\left| \sum_x \pi_{\xi^*,\theta^*}(x) \sum_a [\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)] R(x,a) \right| \\
\leq&\sum_x \pi_{\xi^*,\theta^*}(x) \sum_a |(\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x))| \cdot |R(x,a)| \\
\leq&R_{\max} \sum_x \pi_{\xi^*,\theta^*}(x) \sum_a |(\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x))| \\
=&R_{\max} \sum_x \pi_{\xi^*,\theta^*}(x) \|(\mu_{\theta^*}(\cdot|x) - \mu_{\hat{\theta}}(\cdot|x))\|_1. \qquad (11)
\end{aligned}$$

The bound in (11) is related to the difference in the log-loss of the policies θ^* and $\hat{\theta}$.

By using Lemma 3, we obtain

$$\begin{aligned}
&\left| \sum_x \pi_{\xi^*,\theta^*}(x) \sum_a [\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)] R(x,a) \right| \qquad (12) \\
&\leq R_{\max} \sum_x \pi_{\xi^*,\theta^*}(x) \sqrt{2\ln 2 D\left(\mu_{\theta^*}(\cdot|x) \| \mu_{\hat{\theta}}(\cdot|x)\right)} \\
&\leq \sqrt{2\ln 2} R_{\max} \cdot \\
&\quad \sqrt{\sum_x \pi_{\xi^*,\theta^*}(x) D\left(\mu_{\theta^*}(\cdot|x) \| \mu_{\hat{\theta}}(\cdot|x)\right)} \\
&= \sqrt{2\ln 2} R_{\max} \sqrt{D_{\xi^*,\theta^*}(\mu_{\theta^*} \| \mu_{\hat{\theta}})} \\
&\leq \sqrt{2\ln 2 \varepsilon} R_{\max}. \qquad (13)
\end{aligned}$$

In the first inequality, we applied Lemma 3 by setting $\mathbf{p}_1 = \mu_{\theta^*}(\cdot|x)$ and $\mathbf{p}_2 = \mu_{\hat{\theta}}(\cdot|x)$ for each x. In the second inequality, we applied Jensen's inequality. The final inequality is obtained by Theorem 1.

We next bound the second term in (10) using techniques from perturbation analysis of eigenvalues of a matrix:

$$\begin{aligned}
&\left| \sum_x (\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)) \sum_a \mu_{\hat{\theta}}(a|x) R(x,a) \right| \\
&\leq \sum_x |\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)| \sum_a |\mu_{\hat{\theta}}(a|x) R(x,a)| \\
&\leq R_{\max} \sum_x |\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)| \sum_a \mu_{\hat{\theta}}(a|x) \\
&= R_{\max} \sum_x |\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)| \qquad (14) \\
&\leq R_{\max} \kappa \|\pi'_{\xi^*,\theta^*}(\mathbf{M}_{\xi^*,\theta^*} - \mathbf{M}_{\hat{\xi},\hat{\theta}})\|_1. \qquad (15)
\end{aligned}$$

Here, (14) follows by noting that $\sum_a \mu_{\hat{\theta}}(a|x) = 1$ for all x. (15) can be obtained using Lemma 4.

We now apply the definition of the conditional transition probability $\mathbf{P}_\xi$ associated with policy θ to the last equality. Then,

$$\begin{aligned}
&\|\pi'_{\xi^*,\theta^*}(\mathbf{M}_{\xi^*,\theta^*} - \mathbf{M}_{\hat{\xi},\hat{\theta}})\|_1 \\
=&\sum_y \Big| \sum_x \pi_{\xi^*,\theta^*}(x) \\
&\quad \sum_a [\mathbf{P}_{\xi^*}(y|x,a) \mu_{\theta^*}(a|x) - \mathbf{P}_{\hat{\xi}}(y|x,a) \mu_{\hat{\theta}}(a|x)] \Big| \\
\leq&\sum_x \pi_{\xi^*,\theta^*}(x) \sum_y \sum_a \left| \mathbf{P}_{\hat{\xi}}(y|x,a) \left[\mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x)\right] \right| \\
+&\sum_x \sum_a \pi_{\xi^*,\theta^*}(x) \mu_{\theta^*}(a|x) \sum_y \left| \mathbf{P}_{\hat{\xi}}(y|x,a) - \mathbf{P}_{\xi^*}(y|x,a) \right| \\
\leq&\sum_x \pi_{\xi^*,\theta^*}(x) \sum_a \left| \mu_{\theta^*}(a|x) - \mu_{\hat{\theta}}(a|x) \right| \\
+&\sum_{x,a} \pi_{\xi^*,\theta^*}(x,a) \sum_y \left| \mathbf{P}_{\hat{\xi}}(y|x,a) - \mathbf{P}_{\xi^*}(y|x,a) \right|, \qquad (16)
\end{aligned}$$

where (16) follows by noting that $\sum_y \mathbf{P}(y|x,a) = 1$ for all (x,a). Now, using the steps (12) - (13), Theorem 1 and

Corollary 2, we can bound (11) as

$$\left|\sum_x (\pi_{\hat{\xi},\hat{\theta}}(x) - \pi_{\xi^*,\theta^*}(x)) \sum_a \mu_{\hat{\theta}}(a|x) R(x,a)\right| \leq 2\sqrt{2\varepsilon \ln 2}\kappa R_{\max}. \quad (17)$$

Finally, combining (13) and (17) and applying the bound in Theorem 1, the result in Theorem 5 follows. ∎

We note that the constant κ in Theorem 5 is referred to as condition number. The regret is thus governed by the condition number of the estimated policy under the estimated transition dynamics; the smaller the condition number, the smaller is the regret.

V. A Healthcare Dynamics and Policy Learning Example

A. Background Settings

In this section, we propose an MDP model to simulate the drug effects on patients with some chronic disease. The state of the MDP is denoted as $\mathbf{x} = (x_1, x_2)$, where $x_1, x_2 \in \{0, \ldots, 10\}$. Here x_1 represents the severity of the disease and x_2 represents the severity of comorbidities or complications the patient may be facing. The actions of the MDP are related to the drugs prescribed to the patients. Suppose there are two types of drugs that focus on different diagnostic features, i.e., a type-1 drug mainly relieves the disease symptoms, while a type-2 drug works on the comorbidities or complications. Therefore, the action set can be represented as $\mathbf{a} = (a_1, a_2)$, where $a_i \in \{0,1\}$ indicates whether the patient takes the type-i drug or not.

We assume the following: $\mathbf{P}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)$ depends only on state difference $\mathbf{z}_t = \mathbf{x}_{t+1} - \mathbf{x}_t$ and action $\mathbf{a}_t$. The state can only transit to its neighboring states, specifically, $||\mathbf{z}_t||_1 \leq 1$ for all $\mathbf{x}_t$ and $\mathbf{a}_t$. The conditional state transition dynamics under all actions are described in Table I. We also assume a bouncing boundary condition for this MDP.

TABLE I
State transition probabilities conditioned on actions.

$\mathbf{z}_t =$	$(0,0)$	$(0,1)$	$(0,-1)$	$(-1,0)$	$(1,0)$
$\mathbf{a}_t = (0,0)$	0.7	0.1	0.05	0.05	0.1
$\mathbf{a}_t = (1,0)$	0.4	0.1	0.05	0.35	0.1
$\mathbf{a}_t = (0,1)$	0.4	0.1	0.35	0.05	0.1
$\mathbf{a}_t = (1,1)$	0.2	0.1	0.3	0.3	0.1

When the patient takes no drug, i.e., $\mathbf{a}_t = (0,0)$, both x_1 and x_2 will tend to remain the same or increase. When the patient takes type-1 drug, i.e., $\mathbf{a}_t = (1,0)$, the drug relieves the disease symptoms. When the patient takes type-2 drug, i.e., $\mathbf{a}_t = (0,1)$, the drug relieves the comorbidities or complications. Finally, when both drugs are taken together, i.e., $\mathbf{a}_t = (1,1)$, the drugs' effects are less powerful due to their interactions.

Suppose that the patient collects the corresponding immediate reward when he (or she) enters a certain state. Physicians (experts) adopt some policy to maximize the long-term average reward of patients. We are interested in learning their policy and the transition dynamics from observed state-action-state tuples. We associate a reward R_0 to the best state $(0,0)$, and assume the associated reward "spreads"according to a Gaussian distribution. Specifically, the immediate reward for state $\mathbf{x}$ is

$$R(\mathbf{x}, \mathbf{a}) = R(\mathbf{x}) = R_0 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\|\mathbf{x}\|^2}{2\sigma^2}},$$

for all $\mathbf{a}$, where R_0 specifies the amplitude of reward and σ adjusts the discounting rate of reward as the disease state gets worse. Such a continuous reward function is adjustable to achieve desirable behavior [9]. We set $R_0 = 30$ and $\sigma = 5$ in our experiment.

The key step of our learning method is to design appropriate features for the regression. Inspired by the reproducing kernel Hilbert space method [10], we select some representative state based on which the features are defined. Regarding the transition dynamics, we assume

$$\mathbf{P}_\xi(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t) = \frac{\exp\{\xi' \psi(\mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{a}_t)\}}{\sum_{\mathbf{x}\in\mathscr{X}} \exp\{\xi' \psi(\mathbf{x} - \mathbf{x}_t, \mathbf{a}_t)\}},$$

and define the representative states belonging to $\{(i,j) : i,j \in \{-1,0,1\}\}$, where the corresponding features are:

$$\psi_{ij}(\mathbf{z}, \mathbf{a}) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{\frac{-\|\mathbf{z}-(i,j)\|^2}{2}}, & \text{if } \mathbf{a} = \mathbf{a}_t, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{z} = \mathbf{x}_{t+1} - \mathbf{x}_t$.

Similarly, assuming that the expert policy has the form

$$\mu_\theta(\mathbf{a}|\mathbf{x}) = \frac{\exp\{\theta' \phi(\mathbf{x}, \mathbf{a})\}}{\sum_{b\in\mathscr{A}} \exp\{\theta' \phi(\mathbf{x}, b)\}},$$

we define the representative states belonging to $\{(i,j) : i,j \in \{0,2,4,6,8,10\}\}$, where the corresponding features are:

$$\phi_{ij}(\mathbf{x}, \mathbf{a}) = \sum_{\mathbf{y}} \mathbf{P}_\xi(\mathbf{y}|\mathbf{x}, \mathbf{a}) f_{ij}(\mathbf{y}),$$

in which $\mathbf{y}$ is a potential next state from $\mathbf{x}$, and

$$f_{ij}(\mathbf{y}) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\|(i,j)-\mathbf{y}\|^2}{2}}.$$

B. Simulation and Learning Performance

For the MDP specified above, we use value iteration to find the optimal policy as the target policy, and generate the state-action samples based on this policy. Given the state-action samples, we have two objectives: one is to estimate and evaluate the conditional transition probability under actions $\mathbf{P}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)$, and the other is to learn and evaluate the physician's (expert) policy $\mu(\mathbf{a}_t|\mathbf{x}_t)$. We compare the average rewards with three policy estimates as follows:

1) ℓ_1-regularized policy: The policy is trained by the ℓ_1-regularized logistic regression using Algorithm 1.
2) Unregularized policy: The policy is trained by solving the logistic regression problems using Algorithm 1 but without the ℓ_1-regularization.
3) Greedy policy: the policy where a patient takes the drug which maximizes the expected immediate next step reward.

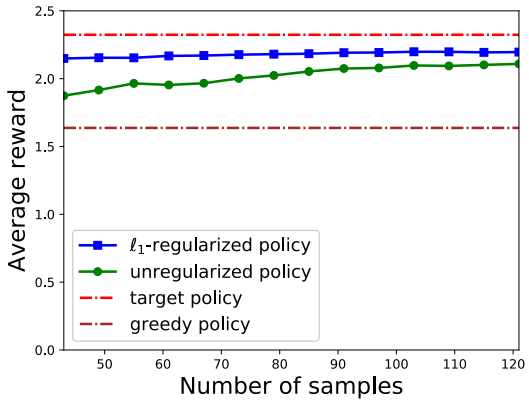

Fig. 1. The average reward of the MDP under different policy estimates for different sample sizes.

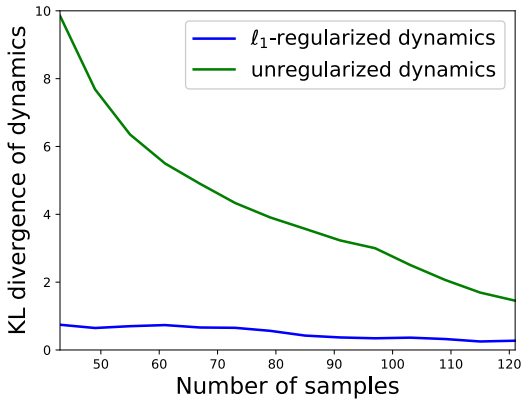

Fig. 2. Average KL divergence of the different conditional transition probability estimates for different sample sizes.

To investigate the performance of our proposed algorithm under different sample sizes, we implement Algorithm 1 for finding the conditional transition probability and policy estimates using different sample sizes. For a fixed sample size, we sample data and perform regression for 100 times and evaluate its average performance. The average reward of different policy estimates is shown in Fig. 1.

We observe that the average reward of the ℓ_1-regularized policy is close to the one of the target policy, which is consistent with the result in Theorem 5. Additionally, the ℓ_1-regularized policy outperforms both the greedy policy and the unregularized policy, especially when the sample size is small. This indicates the importance of the ℓ_1-regularization.

Also, we compare two different conditional transitional probability estimates as follows:

1) ℓ_1-regularized policy: The conditional transition probability is trained by the ℓ_1-regularized logistic regression using Algorithm 1.
2) Unregularized policy: The conditional transition probability is trained by solving the logistic regression problems using Algorithm 1 but without the ℓ_1-regularization.

Fig. 2 shows the KL-divergence of the target and estimated conditional transition probability $D_{\xi^*,\theta^*}(\mathbf{P}_{\xi^*}\|\mathbf{P}_{\hat{\xi}}) = \varepsilon(\hat{\xi}) - \varepsilon(\xi^*)$ as a function of the sample size. We observe that the regularized conditional transition probability estimate is more accurate than its unregularized one, especially when the sample size is small. Also, we can see in the figure that an accurate estimate is achieved by the regularized algorithm with only very few samples, which is consistent with Corollary 2. Note that our algorithm works well even when the target conditional transition probability and policy do not even follow the parameterized form in (1) and (2) respectively.

VI. Conclusion

This paper considers a problem of learning the dynamics and policy of an MDP based on demonstrations. We implement a sparse logistic regression algorithm to estimate the parameterized conditional transitional probabilities and policy. Theoretical results are established for obtaining a bound on the difference in target parameters and their estimates. In addition, we derive a bound on the regret of the policy estimates.

The proposed algorithms can be applied in many real world MDP estimation problems, such as mining Electronic Health Record data. Our algorithms are shown to achieve satisfactory performance in the simulated disease progression models. The learned conditional transition probabilities and the prescription policy are useful for analysis of chronic disease progression and drug effects.

References

[1] Y.-Y. Liu, H. Ishikawa, M. Chen, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden markov model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 444–451.

[2] J. Weiss, S. Natarajan, and D. Page, "Multiplicative forests for continuous-time processes," in *Advances in Neural Information Processing Systems*, 2012, pp. 458–466.

[3] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[4] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized diabetes management using electronic medical records," *Diabetes care*, vol. 40, no. 2, pp. 210–217, 2017.

[5] M. K. Hanawal, H. Liu, H. Zhu, and I. C. Paschalidis, "Learning parameterized policies for Markov decision processes through demonstrations," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE, 2016, pp. 7087–7092.

[6] ——, "Learning policies for Markov decision processes from data," *arXiv preprint arXiv:1701.05954*, 2017.

[7] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proceedings of the International Conference on Machine Learning (ICML)*, June 2004.

[8] T. A. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[9] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Reward function and initial values: better choices for accelerated goal-directed reinforcement learning," in *International Conference on Artificial Neural Networks*. Springer, 2006, pp. 840–849.

[10] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, pp. 1171–1220, 2008.