Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pT-Learning

Wenzhuo Zhou, Ruoqing Zhu & Annie Qu

To cite this article: Wenzhuo Zhou, Ruoqing Zhu & Annie Qu (2024) Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pT-Learning, Journal of the American Statistical Association, 119:545, 625-638, DOI: 10.1080/01621459.2022.2138760

To link to this article: https://doi.org/10.1080/01621459.2022.2138760

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 545, 625–638: Theory and Methods
https://doi.org/10.1080/01621459.2022.2138760

Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pT-Learning

Wenzhuo Zhou[a], Ruoqing Zhu[b], and Annie Qu[a]

[a]Department of Statistics, University of California Irvine, Irvine, CA; [b]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL

ABSTRACT

Recent advances in mobile health (mHealth) technology provide an effective way to monitor individuals' health statuses and deliver just-in-time personalized interventions. However, the practical use of mHealth technology raises unique challenges to existing methodologies on learning an optimal dynamic treatment regime. Many mHealth applications involve decision-making with large numbers of intervention options and under an infinite time horizon setting where the number of decision stages diverges to infinity. In addition, temporary medication shortages may cause optimal treatments to be unavailable, while it is unclear what alternatives can be used. To address these challenges, we propose a Proximal Temporal consistency Learning (pT-Learning) framework to estimate an optimal regime that is adaptively adjusted between deterministic and stochastic sparse policy models. The resulting minimax estimator avoids the *double sampling* issue in the existing algorithms. It can be further simplified and can easily incorporate off-policy data without mismatched distribution corrections. We study theoretical properties of the sparse policy and establish finite-sample bounds on the excess risk and performance error. The proposed method is provided in our `proximalDTR` package and is evaluated through extensive simulation studies and the OhioT1DM mHealth dataset. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2021
Accepted October 2022

KEYWORDS
Policy optimization; Precision medicine; Reinforcement learning; Sparse policy

1. Introduction

Mobile health (mHealth) technology has recently attracted much attention due to mobile devices such as smartphones or wearable devices for tracking physical activities and well-being. It makes real-time communications feasible between health providers and individuals (Sim 2019). In addition, the mHealth technology can also be used to collect rich longitudinal data for exploring optimal dynamic treatment regimes, which are critical in delivering long-term personalized interventions (Nahum-Shani et al. 2018). For example, the OhioT1DM mHealth study (Marling and Bunescu 2020) collects eight weeks' worth of mHealth data for type 1 diabetes patients. All patients are equipped with mobile sensor bands for continuously measuring blood glucose level, insulin dose level, heart rate, carbohydrate intake, etc. This allows us to develop tailored dynamic treatment strategies to manage patients' blood glucose levels. However, current applications of mHealth technology in clinical use encounter some unique challenges. First, using mHealth technology involves data collection and requires decision-making over a very long period. This is often referred to as the infinite horizon setting. Second, most of the mHealth applications aim to provide multi-channel interventions with a large number of treatment combinations (Yang and Van Stee 2019), or recommend continuous individualized dose levels (Marling and Bunescu 2020) for maximizing patients' clinical outcomes. Third, there is an increasing demand for implementing robust dynamic treatment regimes to meet unexpected situations such as temporary shortage of medications or budget constraints in mHealth studies (Rehg, Murphy, and Kumar 2017). It is essential to design an optimal regime that can provide an alternative optimal or near-optimal treatment option as a backup choice. However, existing statistical methodologies are not well-developed for meeting the challenges mentioned above.

Although there is a rich body of literature on estimating dynamic treatment regimes (Murphy 2003; Zhao et al. 2015; Shi et al. 2018) over a fixed period (finite horizon), only a limited number of statistical methodologies have been developed for the infinite horizon setting. Ertefaie and Strawderman (2018) proposed a variant of Greedy GQ-learning to estimate optimal regimes. Luckett et al. (2020) proposed V-learning to search for an optimal policy over a prespecified class of policies. Xu, Laber, and Staicu (2020) later extended V-learning to latent space models. Among other works, Liao, Klasnja, and Murphy (2020), Uehara, Huang, and Jiang (2020) and Shi et al. (2020) focus on a target policy or value function evaluation instead of finding an optimal policy. In the computer science field, popular approaches include learning the optimal value function first, and then recovering the corresponding optimal policy (Antos, Szepesvári, and Munos 2008; Dai et al. 2018). Other methods include the residual gradient algorithm (Baird 1995) and PCL learning (Nachum et al. 2017; Chow, Nachum, and Ghavamzadeh 2018; Nachum et al. 2018), but these algorithms encounter the *double sampling* problem (Sutton and Barto 2018). Alternatively, entropy-augmented methods (Schulman, Chen, and Abbeel 2017; Lee, Choi, and Oh 2018; Haarnoja et al. 2018) follow the principle of developing reinforcement learning algorithms with improved exploration and high robustness. However, their methods are not suitable for continuous state space or large numbers of treatment options.

CONTACT Annie Qu aqu2@uci.edu Department of Statistics, University of California Irvine, Irvine, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In this article, we propose a novel Proximal Temporal consistency Learning (pT-Learning) framework for estimating the optimal infinite horizon treatment regime. Through revisiting the standard Bellman equation from an alternative perspective, we construct a proximal counterpart simultaneously addressing the non-smoothness issues and inducing an optimal sparse policy. This distinguishes our method from commonly used approaches in the existing literature. We use the path-wise consistency property of the constructed proximal Bellman operator to incorporate off-policy data, and further propose a consistent minimax sample estimator for the optimal policy via leveraging the idea of functional space embedding.

The pT-Learning framework enjoys several unique advantages. First, it shows advantages when the number of treatment options is large and can be easily extended to a continuous treatment space. The induced optimal policy identifies treatments from a sparse subset of the treatment space, indicating that it only assigns (near-)optimal treatment options with nonzero probabilities. Also, this property adaptively adjusts the policy between stochastic and deterministic policy models through a data-driven sparsity parameter, hence, bridging the two popular models together. In addition, the induced policy is robust to unexpected situations and guarantees the recommendation of near-optimal treatment alternatives when the optimal treatment is temporarily unavailable. Second, the proposed minimax estimator can be directly optimized over the observed sample transition path without the *double sampling* required in existing algorithms. Third, pT-Learning captures the optimal policy and value function jointly over any arbitrary state-action pair. This avoids the mismatched distributions adjustments, for example, inverse propensity-score weighting in Luckett et al. (2020), for off-policy data. Fourth, our method intrinsically achieves flexibility in choosing the value function approximation class (including both linear and nonlinear function approximation) without the risk of diverging from the optimal solution. In contrast, existing methods such as Q-learning (Watkins and Dayan 1992), Greedy GQ-learning (Ertefaie and Strawderman 2018), and TD-learning (Dann, Neumann, and Peters 2014) with its stabilized variant Emphatic-TD learning (Sutton, Mahmood, and White 2016; Yu 2016; Yu, Mahmood, and Sutton 2018) may either diverge to infinity in off-policy training or only have guaranteed convergence under linear function approximation. Lastly, the proposed constrained minimax optimization problem can be reduced to an unconstrained minimization problem, which can be solved under a scalable and efficient unified *actor-critic* framework. This greatly reduces the computational cost and improves the stability of the optimal policy and value function learning.

In addition to these unique advantages, our study makes important contributions to the fundamental problem of solving the Bellman equation when function approximation is used. We provide a substantial development for addressing the decade-long *double sampling* problem in policy optimization. In addition, our approach draws a connection and provides an alternative understanding to the entropy-augmented Markov decision process problem (Schulman, Chen, and Abbeel 2017; Lee, Choi, and Oh 2018; Haarnoja et al. 2018). Our method is motivated by addressing the non-smoothness issue of the Bellman equation while inducing policy sparsity. It is fundamentally different from the principles of the existing entropy-augmented methods, which focus on improving exploration ability and algorithmic robustness. In theory, we establish the first theoretical result for the adaptivity of sparse policy distributions. Moreover, we develop finite-sample upper bounds on both the excess risk and the performance error. To the best of our knowledge, this is the first non-asymptotic result to quantify the performance error on both deterministic and stochastic policy models jointly.

2. Background and Notation

First, we introduce the background for the estimating dynamic treatment regime in infinite horizon settings, which can be modeled by the Markov decision process (MDP, Puterman 2014). The MDP is denoted as a tuple $(\mathcal{S}, \mathcal{A}, \mathbf{P}, u)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action (treatment) space, $\mathbf{P}(\cdot|s,a)$ is an unknown Markov transition kernel, and u is a unknown immediate utility function. The immediate utility at the time t is defined as $R^t = u(S^{t+1}, S^t, A^t) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. In this article, we consider a finite action space, that is, $|\mathcal{A}| < \infty$. A trajectory induced by the MDP can be written as $\mathcal{D} = \{S^1, A^1, S^2, A^2, S^3, \ldots, \ldots, S^{T+1}\}$, where $S^t \in \mathcal{S}$ is the patient's health state at t, $A^t \in \mathcal{A}$ is the action assignment at t, and T denotes the length of trajectory assumed to be nonrandom for simplicity. The observed data $\mathcal{D}_{1:n} = \{\mathcal{D}_i\}_{i=1}^n$ comprises n independent and identical distributed trajectories of $\mathcal{D}$. Here, the state evolves following the time-homogeneous Markov process. For all $t \geq 1$, $S^{t+1} \perp \left(S^1, A^1, \ldots, S^{t-1}, A^{t-1}\right) \mid \left(S^t, A^t\right)$ and $\mathbf{P}(S^{t+1} = s'|S^t = s, A^t = a) = \mathbf{P}(s'|s,a)$. A treatment regime (policy) $\pi : \mathcal{S} \to \mathcal{A}$ is a map from the state space $\mathcal{S}$ to the action space $\mathcal{A}$.

The discounted sum of utilities beyond the time t is represented as $\sum_{k=1}^{\infty} \gamma^{k-1} R^{t+k}$, where $\gamma \in (0,1)$ is called discount factor. Our goal is to find a policy π to maximize the expected discounted sum of utilities from time t until death. The infinite-horizon value function is defined as $V_t^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=1}^{\infty} \gamma^{k-1} R^{t+k} \mid S^t = s\right]$, where the expectation $\mathbb{E}_\pi$ is taken by assuming that the system follows a policy π. Accordingly, the infinite-horizon action-value function $Q_t^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=1}^{\infty} \gamma^{k-1} R^{t+k} \mid S^t = s, A^t = a\right]$ can be similarly defined as $V_t^\pi(s)$, except that taking treatment a given the state s at time t and then following π till the end. In a time-homogenous Markov process, $V_t^\pi(s)$ and $Q_t^\pi(s,a)$ would not depend on t anymore (Sutton and Barto 2018). And the optimal action-value function $Q^{\pi^*}(s,a) = \max_\pi Q^\pi(s,a)$ is *unique*, which satisfies the Bellman optimality equation (Puterman 2014): $Q^{\pi^*}(s,a) = \mathbb{E}_{S^{t+1}|s,a}[R^t + \gamma \max_{a' \in \mathcal{A}} Q^{\pi^*}\left(S^{t+1}, a'\right) \mid S^t = s, A^t = a]$, where $\mathbb{E}_{S^{t+1}|s,a}$ is a short notation for $\mathbb{E}_{S^{t+1} \sim \mathbf{P}(\cdot|s,a)}$. The policy π^* is referred to as an optimal policy, but it might not be unique. An optimal policy π^* can be obtained by taking the *greedy* action of $Q^{\pi^*}(s,a)$, such that $\pi^*(s) = \arg\max_a Q^{\pi^*}(s,a)$. Given a value function $V^\pi(s)$, the Bellman operator $\mathcal{B}$ is defined as

$$\mathcal{B}V^\pi(s) := \max_{a \in \mathcal{A}} \mathbb{E}_{S^{t+1}|s,a}\left[R^t + \gamma V^\pi(S^{t+1}) \mid S^t = s, A^t = a\right]. \tag{2.1}$$

Then $\mathcal{B}V^{\pi^*}(s) = V^{\pi^*}(s)$ for all $s \in \mathcal{S}$ where V^{π^*} is the unique fixed point of the Bellman operator $\mathcal{B}$ (Bertsekas 1997).

3. Methodology

To develop the new framework, we introduce a proximal Bellman operator and the associated sparse and adaptive optimal policy in Section 3.1. To estimate such optimal policy, we propose a minimax framework called proximal temporal consistency learning in Section 3.2. The proposed method is able to address the practical challenges in mHealth applications, and also has some theoretical guarantees, for example, the bounded performance error in Section 4.

The *greedy* optimal policy $\pi^*(s) = \arg\max_a Q^{\pi^*}(s,a)$ is a deterministic policy, which means it suggests an action according to a deterministic rule without uncertainty. Several drawbacks exist for the deterministic policy class. First, the deterministic policy greedily takes an action at each decision stage, and thus fails to suggest a rule to pick up other (near-optimal) actions as back-up. This restriction leads the deterministic policy class to be nonrobust for unexpected situations whenever the true optimal action is temporarily unavailable or restricted to implement. For example, the insufficient insulin or patient's budget restricts him/her to adopt a sufficient dosage of insulin at each decision stage (Rehg, Murphy, and Kumar 2017; Marling and Bunescu 2020). Second, the deterministic policy class is suboptimal due to poor exploration of environment, which is usually caused by taking greedy actions. The insufficient exploration impairs the performance of the algorithm, especially for the MDPs with large action or state spaces (John 1994). In contrast, the stochastic policy takes actions following a probability distribution, and therefore incorporates a certain degree of randomness, and successfully encourages exploration of the dynamic environment (Singh, Jaakkola, and Jordan 1994; Schulman et al. 2015). Also, the stochastic policy could suggest other actions with specific probabilities, which makes it more flexible and robust under unexpected situations. A more detailed comparison between deterministic and stochastic policy class is provided in Section D, supplementary materials. The aforementioned advantages motivates many existing work to consider modeling a stochastic policy. In particular, one popular choice is using the Boltzmann distribution to model the optimal policy (Schulman, Chen, and Abbeel 2017; Haarnoja et al. 2018; Luckett et al. 2020). However, such policy distribution is prone to assigning nonzero probability mass to all actions, including nonoptimal and dismissible ones. This could be problematic when the cardinality of the action space is large, which may cause the policy distribution to degenerate to a uniform distribution, and potentially fails to provide a desired action recommendation.

These challenges motivate us to develop an optimal policy following a distribution whose support set is a sparse subset of the action space containing only (near-)optimal actions. We refer to this class of policies as the *sparse* policy class. Specially, the deterministic policy $\pi^*(s) = \arg\max_a Q^{\pi^*}(s,a)$ can be viewed as the most extreme case of the sparse policy and inherits some advantages of the sparse policy. However, it is greedy, resulting suboptimality issue and non-robustness to uncertainty. Therefore, we aim to design and estimate an optimal policy that enjoys a suitable and adaptive policy sparsity while inheriting the advantages of the stochastic policy model.

To start with, we first revisit the Bellman optimality equation in (2.1) from a policy-explicit view. Suppose the policy π follows a stochastic distribution, then the Bellman optimality equation can be reformulated as

$$\mathcal{B}V^{\pi^*}(s) := \max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|s),\, S^{t+1}|s,a}\Big[u(S^{t+1},s,a) + \gamma V^{\pi^*}(S^{t+1})\Big] = V^{\pi^*}(s), \tag{3.1}$$

where V^{π^*} is the unique fixed point of $\mathcal{B}$, and π^* is the maximizer of $\max_\pi \mathbb{E}_{a\sim\pi(\cdot|s),\, S^{t+1}|s,a}\Big[u(S^{t+1},s,a) + \gamma V^{\pi^*}(S^{t+1})\Big]$. We use this definition of π^* for the rest of the article. To solve the Equation (3.1), a natural idea is to jointly optimize V^{π^*} and π to minimize the discrepancy between the two sides of (3.1). However, the Equation (3.1) is nonlinear and contains a non-smooth *max* operator. When either the state or action space is large, directly solving (3.1) often results in policies that are far from the optimal solution. In addition, the discontinuity and instability caused by the *max* operator make estimation very difficult without large amounts of samples. To address these issues, we need to consider a proximal counterpart of the Bellman equation (3.1).

3.1. Proximal Bellman Operator

In this section, we propose a proximal Bellman operator that circumvents the obstacles of solving (3.1), while simultaneously inducing an adaptive and sparse optimal policy. The proposed framework bridges the gap between the deterministic and stochastic policy models and characterizes the intrinsic relationship between the two policy models.

Let $\mathcal{P}(\mathcal{A})$ be a convex probability simplex over $\mathcal{A}$, we reformulate the Bellman operator $\mathcal{B}$ under the Fenchel representation, that is,

$$\mathcal{B}V^{\pi^*}(s) = \max_{\pi\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \Big[\mathbb{E}_{S^{t+1}|s,a}\big[u(S^{t+1},s,a) + \gamma V^{\pi^*}(S^{t+1})\big]\cdot\pi(a|s) - \mu\left(\pi(a|s)\right)\Big], \tag{3.2}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product and $\mu(\cdot)$ has to be convex and continuous. We take $\mu(\cdot)\equiv 0$, which satisfies the condition without introducing additional bias. The representation (3.2) provides a basis for constructing a proximity to $\mathcal{B}V^{\pi^*}(s)$. Specifically, we consider to add a strongly convex and continuous component, so-called proximity function, $d(\pi):\mathcal{P}(\mathcal{A})\to\mathbb{R}$ to (3.2):

$$\begin{aligned}\mathcal{B}_\lambda V_\lambda^{\pi_\lambda^*}(s) &= \max_{\pi\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \Big[\mathbb{E}_{S^{t+1}|s,a}\big[u(S^{t+1},s,a) + \gamma V_\lambda^{\pi_\lambda^*}(S^{t+1})\big]\cdot\pi(a|s) + \lambda d\big(\pi(a|s)\big)\Big] \\ &= \max_{\pi\in\mathcal{P}(\mathcal{A})} \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[Q_\lambda^{\pi_\lambda^*}(s,a) + \lambda\phi\big(\pi(a|s)\big)\Big],\end{aligned} \tag{3.3}$$

where $\phi(x) = d(x)/x$ and $Q_\lambda^{\pi_\lambda^*}(s,a) := \mathbb{E}_{S^{t+1}|s,a}[u(S^{t+1},s,a) + \gamma V_\lambda^{\pi_\lambda^*}(S^{t+1})]$. Here, the proximal optimal value function $V_\lambda^{\pi_\lambda^*}$ is the unique fixed point of $\mathcal{B}_\lambda$, that is, $\mathcal{B}_\lambda V_\lambda^{\pi_\lambda^*}(s) = V_\lambda^{\pi_\lambda^*}(s)$, and

the policy π_λ^* is the maximizer of (3.3). The explicit forms of the proximal value functions are deferred in Section B, supplementary materials.

By the Fenchel transformation theorem, the proximity $\mathcal{B}_\lambda V_\lambda^{\pi_\lambda^*}(s)$ is a smooth approximation for $\mathcal{B}V^{\pi^*}(s)$ if some conditions on the proximity function $d(\cdot)$ are satisfied (Hiriart-Urruty and Lemaréchal 2012). In addition, we should notice that we are not satisfied with only achieving the smoothing purpose but aim to develop a sparse optimal policy. To achieve this goal, we define a class of proximity functions based on the κ-logarithm function (Korbel, Hanel, and Thurner 2019), that is, $d(x) = x\phi(x) = -\frac{x}{2}\log_\kappa(x)$, where $\log_\kappa(x) = \frac{1}{1-\kappa}\left(x^{1-\kappa} - 1\right)$ for $x > 0$ and $\kappa \neq 1$. In this article, we consider a special case $\kappa = 0$, and refer to the operator (3.3) as the *proximal Bellman operator*.

Accordingly, the proximal Bellman operator (3.3) has several unique properties. First, it is a valid approximation for the Bellman operator $\mathcal{B}$, where the approximation bias is bounded in Theorem S.1 provided in Section K, supplementary materials. Second, it is a smooth substitute of $\mathcal{B}$ according to the closed form of $\mathcal{B}_\lambda$ in (3.4). Third, the proximal Bellman operator $\mathcal{B}_\lambda$ induces a sparse optimal policy whose sparsity can be adjusted by the magnitude of λ.

Through verifying KKT conditions of the maximization in (3.3), the proximal Bellman operator $\mathcal{B}_\lambda$ has a closed-form equivalence (Proposition S.1 in Section K, supplementary materials):

$$\mathcal{B}_\lambda V_\lambda^{\pi_\lambda^*}(s) = \frac{\lambda}{2}\left\{1 - \sum_{a\in\mathcal{K}(s)}\left[\left(\frac{\sum_{a'\in\mathcal{K}(s)}\frac{Q_\lambda^{\pi_\lambda^*}(s,a')}{\lambda} - 1}{|\mathcal{K}(s)|}\right)^2 - \left(\frac{Q_\lambda^{\pi_\lambda^*}(s,a)}{\lambda}\right)^2\right]\right\}, \tag{3.4}$$

where $\mathcal{K}(s) = \{a_{(i)} \in \mathcal{A} : Q_\lambda^{\pi_\lambda^*}(s,a_{(i)}) > \frac{1}{i}\sum_{j=1}^{i} Q_\lambda^{\pi_\lambda^*}(s,a_{(j)}) - \frac{\lambda}{i}\}$ represents the support action set at state s. Here $a_{(i)}$ is the action with the ith largest state-action value, and $|\mathcal{K}(s)| \leq |\mathcal{A}|$ holds for any $s \in \mathcal{S}$.

The proximal Bellman operator is differentiable everywhere only except for few splitting points where the support set $\mathcal{K}(s)$ changes, and the degree of smoothness is determined by the magnitude of λ according to Proposition S.1. We visualize this property in Figure 1 under a binary action setting. As λ increases, the proximal Bellman operator $\mathcal{B}_\lambda$ becomes a more smooth approximation of $\mathcal{B}$ but the approximation bias increases accordingly, indicating that the parameter λ controls the bias and smoothness tradeoff. In addition to the smoothness property, the proximal Bellman operator $\mathcal{B}_\lambda$ induces a sparse optimal policy distribution $\pi_\lambda^*(\cdot|s)$ whose support set is a sparse subset of the action space. We illustrate this point by presenting $\pi_\lambda^*(a|s)$ in terms of the state-action value function $Q_\lambda^{\pi_\lambda^*}(s,a)$ analytically (see the proof of Proposition S.1). That is

$$\pi_\lambda^*(a|s) = \left(\frac{Q_\lambda^{\pi_\lambda^*}(s,a)}{\lambda} - \frac{\sum_{a'\in\mathcal{K}(s)}\frac{Q_\lambda^{\pi_\lambda^*}(s,a')}{\lambda} - 1}{|\mathcal{K}(s)|}\right)^+, \tag{3.5}$$

which is invariant to the location shift of the immediate utility. Given any state s, the Equation (3.5) defines a well-defined probability mass function in that $\sum_{a\in\mathcal{A}}\pi_\lambda^*(a|s) = 1$. Also, the policy π_λ^* is prone to assign a large probability to the action according to the rank of the state-action values. This ensures that the policy $\pi_\lambda^*(a|s)$ can suggest the optimal action in a given state s.

Moreover, the sparsity is shown by analyzing the support set of $\pi_\lambda^*(\cdot|s)$, that is, $\{a \in \mathcal{A} : \pi_\lambda^*(a|s) > 0\}$, which satisfies

$$\left\{a_{(i)} \in \mathcal{A} : Q_\lambda^{\pi_\lambda^*}(s,a_{(i)}) > \frac{1}{i}\sum_{j=1}^{i} Q_\lambda^{\pi_\lambda^*}(s,a_{(j)}) - \frac{\lambda}{i}\right\}. \tag{3.6}$$

The inequality (3.6) indicates that the sparsity parameter λ controls the margins between the smallest action value and the others included in the support set (3.6). In particular, the cardinality of the support set increases as λ increases. Conversely, the support set of $\pi_\lambda^*(\cdot|s)$ shrinks as λ becomes smaller. To illustrate how this mechanism works, we use a ternary action setting as an example. By the inequality (3.6), when λ is sufficiently small in that $\lambda < Q_\lambda^{\pi_\lambda^*}(s,a_{(1)}) - Q_\lambda^{\pi_\lambda^*}(s,a_{(2)})$, only $a_{(1)}$ is contained in the support set. The policy distribution becomes a deterministic rule and takes the greedy action with the largest action value $Q_\lambda^{\pi_\lambda^*}(s,a_{(1)})$. From this point of view, the Q-learning approach and its variants can be regarded as special cases in our framework. If λ increases and falls into the range $\left[Q_\lambda^{\pi_\lambda^*}(s,a_{(1)}) - Q_\lambda^{\pi_\lambda^*}(s,a_{(2)}), \sum_{i=1}^{2} Q_\lambda^{\pi_\lambda^*}(s,a_{(i)}) - 2Q_\lambda^{\pi_\lambda^*}(s,a_{(3)})\right]$, the support set

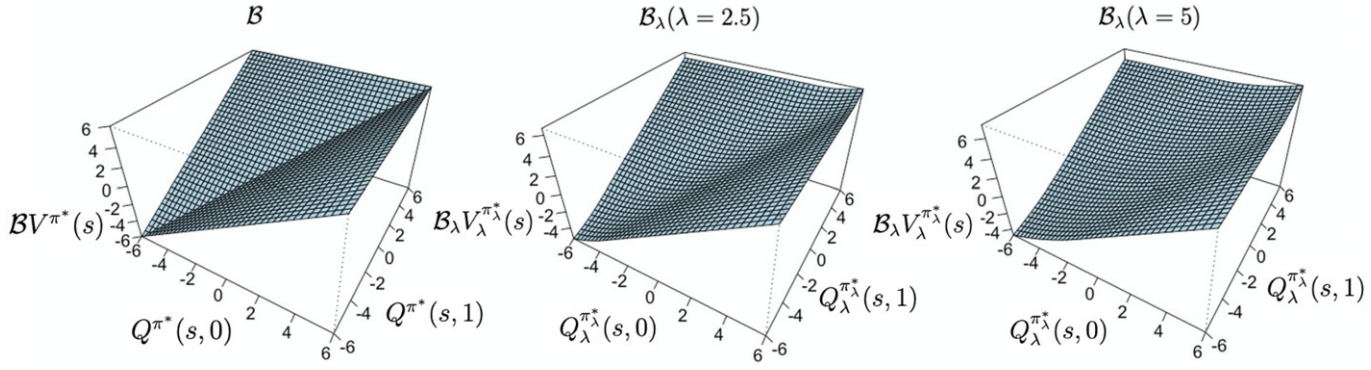

Figure 1. A comparison of the standard and proximal Bellman operator in a binary action setting.

contains two actions $a_{(1)}$ and $a_{(2)}$ with the largest state-action values, but eliminates action $a_{(3)}$. This implies that the induced policy can be adaptively adjusted between the deterministic and stochastic policy models. In Section 4, we investigate the sparsity of the optimal policy distribution in more details.

3.2. Proximal Temporal Consistency Learning

Next, we introduce the proposed proximal temporal consistency learning framework for estimating the induced sparse policy π_λ^*. Our development mainly hinges on a path-wise property of the proximal Bellman operator $\mathcal{B}_\lambda$ and the functional space embedding (Gretton et al. 2012). The proposed minimax estimator is able to address the *double sampling* issue and easily incorporate off-policy data, while intrinsically preserving the convergence in off-policy training with flexible function approximations.

First, we show that the proximal Bellman operator $\mathcal{B}_\lambda$ enjoys the temporal consistency property (Rawlik, Toussaint, and Vijayakumar 2013). Most recently, many state-of-the-art methods based on this type of property have achieved great success in real-life applications (Nachum et al. 2017; Chow, Nachum, and Ghavamzadeh 2018; Nachum et al. 2018). The temporal consistency property connects the optimal policy and value function in one equation along with any arbitrary state-action pair, which provides an elegant way to incorporate off-policy data. We present the temporal consistency property for $\mathcal{B}_\lambda$ in the following proposition.

Proposition 3.1. For any $s \in \mathcal{S}, a \in \mathcal{A}$ and $\lambda \in (0, +\infty)$, if $V_\lambda^{\pi_\lambda^*}$ is the fixed point of the proximal Bellman operator, that is, $\mathcal{B}_\lambda V_\lambda^{\pi_\lambda^*} = V_\lambda^{\pi_\lambda^*}$, and π_λ^* is the induced policy following the Equation (3.5), then $(V_\lambda^{\pi_\lambda^*}, \pi_\lambda^*)$ is a solution of the following proximal temporal consistency equation:

$$\mathbb{E}_{S^{t+1}|s,a}\big[u(S^{t+1}, s, a) + \gamma V_\lambda^{\pi_\lambda}(S^{t+1})\big] - \lambda\phi'(\pi_\lambda(a|s)) - \Psi(s) + \psi(a|s) - V_\lambda^{\pi_\lambda}(s) = 0, \tag{3.7}$$

where $\Psi(s) : \mathcal{S} \to [-\lambda/2, 0]$ and $\psi(a|s) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ are Lagrangian multipliers with $\psi(a|s) \cdot \pi_\lambda(a|s) = 0$, and $\phi'(x) = x - \frac{1}{2}$. Following, we call the discrepancy on the left-hand side of the Equation (3.7) the proximal temporal consistency error (pT-error). Here, the "temporal consistency" characterizes that the equation holds for temporal transitions.

Since the Equation (3.7) holds for any arbitrary state-action pair (s, a), the policy π_λ^* can be estimated directly over the observed transition pairs. Using this property lets us skip to adjust mismatched distributions using, for example, inverse propensity-score weighting (Luckett et al. 2020), and avoids the curse of high variance (Jiang and Li 2016) carried over from the data distribution corrections. To solve the Equation (3.7), an intuitive idea is to minimize the pT-error with the L^2 loss, that is

$$\min_{V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi} \mathbb{E}_{S^t, A^t}\Big[\big(\mathcal{T}_{\pi_\lambda}(S^t, A^t, V_\lambda^{\pi_\lambda}) - \Psi(S^t) + \psi(A^t|S^t) - V_\lambda^{\pi_\lambda}(S^t)\big)^2\Big] \tag{3.8}$$

$$\text{s.t.} \quad \pi_\lambda(a|s) \cdot \psi(a|s) = 0, \ \psi(a|s) \geq 0$$
$$\text{and} \ -\frac{\lambda}{2} \leq \Psi(s) \leq 0, \ \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}, \tag{3.9}$$

where $\mathcal{T}_{\pi_\lambda}(S^t, A^t, V_\lambda^{\pi_\lambda}) = \mathbb{E}_{S^{t+1}|S^t,A^t}\big[u(S^{t+1}, S^t, A^t) + \gamma V_\lambda^{\pi_\lambda}(S^{t+1})\big] - \lambda\phi'(\pi_\lambda(A^t|S^t))$, and $\mathbb{E}_{S^t,A^t}$ is a short notation for $\mathbb{E}_{\{S^t,A^t\}\sim d_{\pi_b}}$ where d_{π_b} is the behavior data distribution induced by the behavior policy π_b. The behavior policy π_b is the policy the decision maker follows in collecting the data.

However, directly minimizing the sample version of (3.8) is not realizable, as the conditional expectation $\mathbb{E}_{S^{t+1}|S^t,A^t}$ in $\mathcal{T}_{\pi_\lambda}(S^t, A^t, V_\lambda^{\pi_\lambda})$ is unknown. To approximate this unknown expectation, bootstrapping can be applied but produces an inconsistent and biased sample estimator, where an extra conditional variance will be involved as a bias component (Fan et al. 2020). This problem is usually referred to as the *double-sampling* issue (Baird 1995) and exists in many policy optimization approaches.

To avoid the *double-sampling* bias, we consider embedding a Lebesgue measurable function class to the averaged pT-error. In particular, we define a *critic* function $h \in \mathcal{H} = L^2(\mathcal{S} \times \mathcal{A})$ and formulate a novel embedding function $\mathcal{L}_{\text{weight}}$ as follows,

$$\mathcal{L}_{\text{weight}}(V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi, h) := \mathbb{E}_{S^t,A^t}\big[h(S^t, A^t)\big(\mathcal{T}_{\pi_\lambda}(S^t, A^t, V_\lambda^{\pi_\lambda}) - \Psi(S^t) + \psi(A^t|S^t) - V_\lambda^{\pi_\lambda}(S^t)\big)\big]. \tag{3.10}$$

The *critic* function is introduced here to fit the discrepancy of (3.7) and promotes the transition pairs with large pT-error. Unlike the naive L^2 loss (3.8), the nested conditional expectation $\mathbb{E}_{S^{t+1}|S^t,A^t}$ in $\mathcal{T}_{\pi_\lambda}(S^t, A^t, V_\lambda^{\pi_\lambda})$ is not inside the square function anymore. Intuitively, the $\mathcal{L}_{\text{weight}}$ circumvents the *double sampling* issue since the second order moment of bootstrapping samples is not involved anymore and the extra condition variance vanishes. Fundamentally, the $\mathcal{L}_{\text{weight}}$ offers compensation for insufficient sampling on $\mathbf{P}(s'|s, a)$. The marginal information of the state-action pairs $d_{\pi_b}(s, a)$ and the transition kernel $\mathbf{P}(s'|s, a)$ can be aggregated as a joint distribution $p_{\pi_b}(s', s, a)$ by the linearity of the expectation. Instead of extracting the information from the transition kernel $\mathbf{P}(s'|s, a)$, that is, approximating the conditional expectation $\mathbb{E}_{S^{t+1}|S^t,A^t}$, the $\mathcal{L}_{\text{weight}}$ can be approximated using the sample-path transition pairs drawn from the joint distribution $p_{\pi_b}(s', s, a)$. In essence, this explains why the $\mathcal{L}_{\text{weight}}$ can address the *double sampling* issue.

In the following theorem, we proceed to show $\mathcal{L}_{\text{weight}}(V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi, h)$ can identify the optimal value function and optimal policy $(V_\lambda^{\pi_\lambda^*}, \pi_\lambda^*)$.

Theorem 3.1. Suppose $\mathcal{S} \times \mathcal{A}$ is Lebesgue measurable, and for any $h(\cdot)$ in a bounded L^2 space, that is, $h(\cdot) \in \mathcal{H}_{L^2}^{\zeta} := \{h : \|h\|_{L^2} \leq \zeta\}$ for $\zeta \in (0, +\infty)$, then the loss $\mathcal{L}_{\text{weight}}(V_\lambda^{\pi_\lambda^*}, \pi_\lambda^*, \Psi, \psi, h) = 0$. Conversely, if there exists $(V_\lambda^{\widetilde{\pi}_\lambda}, \widetilde{\pi}_\lambda)$ such that $\mathcal{L}_{\text{weight}}(V_\lambda^{\widetilde{\pi}_\lambda}, \widetilde{\pi}_\lambda, \Psi, \psi, h) = 0$, then $(V_\lambda^{\widetilde{\pi}_\lambda}, \widetilde{\pi}_\lambda)$ satisfies the proximal temporal consistency Equation (3.7).

As $\mathcal{L}_{\text{weight}}(V_\lambda^{\pi_\lambda^*}, \pi_\lambda^*, \Psi, \psi, h) = 0$ holds for any $h \in \mathcal{H}_{L^2}^{\zeta}$, Theorem 3.1 leads to a minimax optimization problem with a

valid loss $\mathcal{L}^2_{\text{weight}}$,

$$\min_{V_\lambda^{\pi_\lambda},\pi_\lambda,\Psi,\psi} \max_{h\in\mathcal{H}^\zeta_{L^2}} \mathcal{L}^2_{\text{weight}}(V_\lambda^{\pi_\lambda},\pi_\lambda,\Psi,\psi,h). \tag{3.11}$$

The minimax optimization of (3.11) gives a clear direction to estimate $(V_\lambda^{\pi_\lambda^*},\pi_\lambda^*)$, but it still remains intractable. As the *critic* function $h(\cdot)$ could be any arbitrary function in $\mathcal{H}^\zeta_{L^2}$, it makes infeasible to find a proper representation for $h(\cdot)$. Therefore, we introduce a tractable framework for (3.11), where $h(\cdot)$ can be appropriately represented. Define $h^* : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be the optimal *critic* function if it satisfies

$$h^*(\cdot)\in\arg\max_{h\in\mathcal{H}^\zeta_{L^2}} \mathcal{L}^2_{\text{weight}}(V_\lambda^{\pi_\lambda^*},\pi_\lambda^*,\Psi,\psi,h). \tag{3.12}$$

In the following, we shows that $h^*(\cdot)$ lies in a class of continuous functions under some regular continuity conditions which are easily satisfied in practice.

Theorem 3.2. Let $\mathbb{C}(\mathcal{S}\times\mathcal{A})$ be all continuous functions on $\mathcal{S}\times\mathcal{A}$. For any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $s'\in\mathcal{S}$, the optimal *critic* function $h^*(s,a)$ has the following properties:

1. (*Continuity*) Suppose the utility function $u(s',s,a)$ and the transition kernel $\mathbf{P}(s'|s,a)$ are continuous over (s,a) for any s', then $h^*\in\mathcal{H}^\zeta_{L^2}\cap\mathbb{C}(\mathcal{S}\times\mathcal{A})$ and is *unique*.

2. (*Lipschitz-continuity*) Suppose $u(s',s,a)$ is uniformly M_u-Lipschitz continuous and $\mathbf{P}(s'|s,a)$ is M_p-Lipschitz continuous over (s,a) for any s', where M_u and M_p are some Lipschitz constants, then there must exist a Lipschitz constant M_{h^*} such that $h^*(s,a)$ is M_{h^*}-Lipschitz continuous over (s,a).

Theorem 3.2 states that the optimal *critic* function $h^*(s,a)$ is continuous over (s,a) if only the utility function and the transition kernel are continuous. This continuity condition widely holds for precision medicine and reinforcement learning problems. As we mentioned, the h^* could be any arbitrary function in $\mathcal{H}^\zeta_{L^2}$, which imposes exceptional difficulty in the representation of h^*. Theorem 3.2 indicates that it is sufficient to represent h in a bounded continuous function space which preserves the optimal solution of (3.11). This provides us a basis for constructing a tractable framework for solving (3.11).

We propose to represent h^* in a bounded reproducing kernel Hilbert space (RKHS) such that $\mathcal{H}^\zeta_{\mathcal{K}} := \{h : \|h\|^2_{\mathcal{H}_{\text{RKHS}}}\le\zeta\}$. When $\mathcal{H}^\zeta_{\mathcal{K}}$ is reproduced by a universal kernel, the error $err(\zeta) := \sup_{f\in\mathbb{C}(\mathcal{S}\times\mathcal{A})}\inf_{h\in\mathcal{H}^\zeta_{\mathcal{K}}}\|f-h\|_\infty$ decreases as ζ increases and vanishes to zero as ζ goes to infinity (Bach 2017). Therefore, any continuous function can be approximated by a function in $\mathcal{H}^\zeta_{\mathcal{K}}$ with arbitrarily small error. As a result, solving the inner maximization of (3.11) over $h\in\mathcal{H}^\zeta_{\mathcal{K}}$ is feasible when $h^*\in\mathcal{H}^\zeta_{L^2}\cap\mathbb{C}(\mathcal{S}\times\mathcal{A})$, which also makes the optimization become tractable.

Although the optimization (3.11) is tractable with RKHS representation, solving the minimax optimization is still a challenging task. Int the following, we transform the minimax optimization to an easier solvable minimization problem, leveraging the idea of the kernel embedding (Gretton et al. 2012).

Theorem 3.3. Suppose $h\in\mathcal{H}^\zeta_{\mathcal{K}}$ is reproduced by a universal kernel $K(\cdot,\cdot)$ in (3.11) , then the minimax optimization (3.11) can be decoupled to a single-stage minimization problem wherein

$$\begin{aligned}\min_{V_\lambda^{\pi_\lambda},\pi_\lambda,\Psi,\psi}\ \mathcal{L}_U := \mathbb{E}_{S^t,\widetilde{S}^t,A^t,\widetilde{A}^t,S^{t+1},\widetilde{S}^{t+1}}\Big[\Big(\widetilde{\mathcal{T}}_{\pi_\lambda}(S^t,A^t,V_\lambda^{\pi_\lambda})\\ -\Psi(S^t)+\psi(A^t|S^t)-V_\lambda^{\pi_\lambda}(S^t)\Big)\cdot\zeta K\big(\{S^t,A^t\},\{\widetilde{S}^t,\widetilde{A}^t\}\big)\\ \Big(\widetilde{\mathcal{T}}_{\pi_\lambda}(\widetilde{S}^t,\widetilde{A}^t,V_\lambda^{\pi_\lambda})-\Psi(\widetilde{S}^t)+\psi(\widetilde{A}^t|\widetilde{S}^t)-V_\lambda^{\pi_\lambda}(\widetilde{S}^t)\Big)\Big],\end{aligned} \tag{3.13}$$

where $\widetilde{\mathcal{T}}_{\pi_\lambda}(S^t,A^t,V_\lambda^{\pi_\lambda}) = u(S^{t+1},S^t,A^t)+\gamma V_\lambda^{\pi_\lambda}(S^{t+1})-\lambda\phi'(\pi_\lambda(A^t|S^t))$ and $(\widetilde{S}^t,\widetilde{A}^t,\widetilde{S}^{t+1})$ is an independent copy of the transition pair (S^t,A^t,S^{t+1}).

Moreover, if the universal kernel $K(\cdot,\cdot)$ is strictly positive definite (Stewart 1976), the loss $\mathcal{L}_U$ is nonnegative definite. This implies that $\mathcal{L}_U=0$ when $(V_\lambda^{\pi_\lambda},\pi_\lambda)=(V_\lambda^{\pi_\lambda^*},\pi_\lambda^*)$, and thus the loss $\mathcal{L}_U$ is a valid loss can identify $(V_\lambda^{\pi_\lambda^*},\pi_\lambda^*)$.

Given the observed data $\mathcal{D}_{1:n}$ with the length of trajectory T, to minimize $\mathcal{L}_U$ under the constraints in (3.9), we propose a trajectory-based U-statistic estimator to capture the within-trajectory loss. Subsequently, the total loss $\mathcal{L}_U$ can be aggregated as the empirical mean of n iid within-trajectory loss. That is,

$$\begin{aligned}\min_{V_\lambda^{\pi_\lambda},\pi_\lambda,\Psi,\psi}\ \widehat{\mathcal{L}_U} = \frac{1}{n}\sum_{i=1}^n\frac{2\zeta}{T(T-1)}\sum_{1\le j\ne k\le T}\Big[\big(\widetilde{\mathcal{T}}_{\pi_\lambda}(S_i^j,A_i^j,V_\lambda^{\pi_\lambda})\\ -\Psi(S_i^j)+\psi(A_i^j|S_i^j)-V_\lambda^{\pi_\lambda}(S_i^j)\big)\cdot K\big(\{S_i^j,A_i^j\},\{S_i^k,A_i^k\}\big)\\ \big(\widetilde{\mathcal{T}}_{\pi_\lambda}(S_i^k,A_i^k,V_\lambda^{\pi_\lambda})-\Psi(S_i^k)+\psi(A_i^k|S_i^k)-V_\lambda^{\pi_\lambda}(S_i^t)\big)\Big].\\ \text{s.t.}\quad \pi_\lambda(a|s)\cdot\psi(a|s)=0,\ \psi(a|s)\ge 0\ \text{ and }\ -\frac{\lambda}{2}\le\Psi(s)\le 0,\\ \text{for all } s\in\mathcal{S}\text{ and } a\in\mathcal{A}.\end{aligned} \tag{3.14}$$

Unlike the inconsistent sample estimator in (3.8), the proposed sample estimator $\widehat{\mathcal{L}_U}$ is consistent. The consistency is shown in Theorem 4.2 in Section 4, through examining the tail behavior of $\widehat{\mathcal{L}_U}$. In addition, the gradient of the proposed loss $\widehat{\mathcal{L}_U}$ can be approximated by the sampled transitions and optimized using any gradient descent type of algorithm. Hence, the proposed method achieves great flexibility in the function approximations of $(V_\lambda^{\pi_\lambda^*},\pi_\lambda^*)$, allowing both linear and nonlinear approximations without the risk of divergence from the optimal solution. In comparison, popular methods, including the Q-learning (Watkins and Dayan 1992), TD-learning (Dann, Neumann, and Peters 2014) and Greedy GQ-learning (Maei et al. 2010; Ertefaie and Strawderman 2018), either diverge to infinity in off-policy training or have guaranteed convergence only in using linear function approximations. Also, we should note that the pT-Learning framework connects to the importance weighted variants of off-policy TD-learning algorithms when the *critic* function $h(\cdot)$ is restricted into a bounded linear function space. We provide a discussion on this connection in Section E, supplementary materials.

4. Theory

In this section, we establish the theoretical properties of the pT-Learning framework. Our study mainly focuses on three major parts. In the first part, we formally study the sparsity of the policy distribution π_λ^*. Theorem 4.1 shows that the cardinality of the nonzero probability set for π_λ^* is well-controlled by the magnitude of λ, which implies that π_λ^* is adaptively adjusted between the deterministic and stochastic policy models. We believe this is the first theoretical result on investigating the sparsity of policy distributions. In the second part, we establish a convergence rate of the empirical risk $\widehat{\mathcal{L}_U}$ toward the true risk $\mathcal{L}_U$. In particular, Theorem 4.2 provides a sharp exponential concentration bound, indicating that $\widehat{\mathcal{L}_U}$ is a consistent estimate for $\mathcal{L}_U$ and the *double sample* issue is solved. In the last part, we measure the excess risk of $\widehat{\mathcal{L}_U}$ which is applicable for both the parametric and nonparametric function spaces in Theorem 4.3. Furthermore, Theorem 4.4 provides a finite-sample upper bound on the performance of the estimated optimal value function. The bound indicates both the excess risk and the approximation bias of the proximal Bellman operator will affect the performance error. To the best of our knowledge, this is the first nonasymptotic result to quantify the performance error on deterministic and stochastic policy models jointly. In the following, we present our major assumptions and theorems. The proofs and additional theoretical results and assumptions are deferred to supplementary materials.

First, we define a notation called Δ_π-sparsity to characterize the cardinality of nonzero probability actions for some policy π.

Definition 4.1. Denote the maximum cardinality of the support set for a given policy distribution $\pi(\cdot|s)$ over $s \in \mathcal{S}$ as Δ_π, that is,

$$\Delta_\pi := \max_{s\in\mathcal{S}} |\{a \in \mathcal{A} : \pi(a|s) \neq 0\}|,$$

where $|\cdot|$ is a cardinality operator and $\Delta_\pi \leq |\mathcal{A}|$. Then, we call the policy distribution $\pi(\cdot|s)$ a Δ_π-degree sparse policy distribution.

Before we characterize the sparsity of the proposed sparse policy model, we first present a boundedness assumption.

Assumption 1. The immediate utility function $u(S^{t+1}, S^t, A^t)$ is uniformly bounded by $R_{\max}$, that is, $\|u\|_\infty \leq R_{\max} < \infty$.

Assumption 1 is a regular assumption to impose a boundedness condition on the MDP (Antos, Szepesvári, and Munos 2008; Liao, Klasnja, and Murphy 2020).

Theorem 4.1. Under Assumption 1 and Assumptions S.1–S.2 provided in Section I, supplementary materials, for all $s \in \mathcal{S}$, the sparse policy distribution $\pi_\lambda^*(\cdot|s)$ in (3.5) has the following properties: the policy π_λ^* is $\Delta_{\pi_\lambda^*}$-degree sparse where $\Delta_{\pi_\lambda^*} \to 1$ as $\lambda \to 0$, and $\Delta_{\pi_\lambda} \to |\mathcal{A}|$ as $\lambda \to \infty$. In particular, for any $\epsilon > 0$, there exists a positive number λ_0 such that for all $\lambda \geq \lambda_0$, the sparse policy distribution $\pi_\lambda^*(\cdot|s)$ approaches a discrete uniform distribution $\mathcal{U}$ with a probability mass function $|\mathcal{A}|^{-1}$, that is, $|\pi_\lambda^*(a|s) - |\mathcal{A}|^{-1}| < \epsilon$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Theorem 4.1 shows that the degree of sparseness, or the cardinality of the nonzero probability actions set, of the sparse policy $\pi_\lambda^*(\cdot|s)$ can be controlled by λ. In the most extreme case, when $\lambda \to 0$, the policy π_λ^* becomes a deterministic policy. Alternatively, as $\lambda \to +\infty$, the induced policy π_λ^* degenerates to a uniform distribution that assigns equal probabilities to all action arms.

Before we present the rest of the theoretical results, we need to introduce the stationarity and dependency of the stochastic process $\{(S^t, A^t)\}_{t\geq 1}$. For each single patient trajectory, it is easily observed that the sequence $\{(S^t, A^t)\}_{t\geq 1}$ is a stationary Markov chain. In terms of sample dependency, Farahmand et al. (2016) and Liao, Klasnja, and Murphy (2020) assume that within-trajectory samples are independent in order to reduce technical difficulties in the theoretical developments. However, this assumption might be too restrictive and often violated in practice. In our work, we consider the sample dependency and establish more rigorous theoretical results under the notion of the mixing process (Kosorok 2008). Specifically, for any stationary sequence of dependent random variables $\{S^t, A^t\}_{t\geq 1}$, let $\mathcal{F}_b^c$ be the σ-field generated by $\{S^b, A^b\}, \ldots, \{S^c, A^c\}$ and define $\beta(k) = \mathbb{E}\big[\sup_{m\geq 1}\{|P(B \mid \mathcal{F}_1^m) - P(B)| : B \in \mathcal{F}_{m+k}^\infty\}\big]$, and we say that the process $\{S^t, A^t\}_{t\geq 1}$ is β-mixing if $\beta(k) \to 0$ as $k \to \infty$. In the following, we assume a mixing condition to quantify the within-trajectory sample dependency.

Assumption 2. For a strictly stationary sequence $\{(S^t, A^t)\}_{t\geq 1}$, there exists a constant $\delta_1 > 1$ such that the β-mixing coefficient corresponding to $\{(S^t, A^t)\}_{t\geq 1}$ satisfies $\beta(k) \lesssim \exp(-\delta_1 k)$ for $k \geq 1$.

Assumption 2 implies an exponential decay rate of the mixing process, which is typically for deriving a polynomial decay rate of the estimation error (Kosorok 2008).

In Theorem 4.2, we study the tail behaviors of the empirical loss $\widehat{\mathcal{L}_U}$, and establish an exponential concentration inequality for $\widehat{\mathcal{L}_U}$ under an exponential rate of the β-mixing condition.

Theorem 4.2. For any sparsity parameter $\lambda < \infty$ and $\epsilon > 0$, under Assumptions 1–2 and S.1–S.2 provided in Section I, supplementary materials, we have ϵ-divergence of $|\widehat{\mathcal{L}_U} - \mathcal{L}_U|$ bounded in probability for sufficiently large T, that is

$$\mathbb{P}\left(|\widehat{\mathcal{L}_U} - \mathcal{L}_U| > \epsilon\right) \leq 2\Bigg[\exp\left(-\frac{c_1\epsilon^2 T/4 - c_0 c_1 \epsilon U_{\max}^2\sqrt{T}}{U_{\max}^2(\epsilon/2 - c_0 U_{\max}^2/\sqrt{T})\log T \log\log 4T + U_{\max}^4}\right) + \exp\left(-\frac{n\epsilon^2}{2U_{\max}^4}\right)\Bigg],$$

where c_0, c_1 are some constants depending on δ_1; and $U_{\max} = \frac{6R_{\max}+(5-4\gamma)\lambda}{2(1-\gamma)}$.

Theorem 4.2 shows that $\widehat{\mathcal{L}_U}$ is a consistent estimator for the loss $\mathcal{L}_U$ with a high probability at an exponential rate. The concentration bound is sharper than the bound established in Borisov and Volodko (2009). Here, we require an exponential decaying mixing rate, which is standard in the literature of deriving concentration inequalities for weakly dependent data (Borisov and Volodko 2009; Merlevède, Peligrad, and Rio

2011). It should be possible to relax this mixing condition to a polynomial-decay rate of β-mixing with imposing an additional exponential α-mixing condition. However, that is out of the scope of this article.

We denote $\mathcal{L}_U^* = \inf_{\{V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi\}} \mathcal{L}_U(V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi)$ as the minimal risk, also called Bayes risk (Bartlett, Jordan, and McAuliffe 2006). Also, we define the empirical risk minimizer of $\widehat{\mathcal{L}_U}$ as

$$(\widehat{V}_\lambda^{\theta_1}, \widehat{\pi}_\lambda^{\theta_2}, \widehat{\Psi}^{\xi}, \widehat{\psi}^{\theta_3}) = \arg\min_{(V_\lambda^{\theta_1}, \pi_\lambda^{\theta_2}, \Psi^{\xi}, \psi^{\theta_3}) \in \Theta_1 \times \Theta_2 \times \Xi \times \Theta_3} \widehat{\mathcal{L}_U}(V_\lambda^{\theta_1}, \pi_\lambda^{\theta_2}, \Psi^{\xi}, \psi^{\theta_3}),$$

where Θ_1, Θ_2, Ξ, and Θ_3 are function spaces corresponding to $V_\lambda^{\theta_1}, \pi_\lambda^{\theta_2}, \Psi^{\xi}$, and ψ^{θ_3}, respectively.

In Theorem 4.3, we establish the excess risk bound and convergence rate of $\mathcal{L}_U(\widehat{V}_\lambda^{\theta_1}, \widehat{\pi}_\lambda^{\theta_2}, \widehat{\Psi}^{\xi}, \widehat{\psi}^{\theta_3}) - \mathcal{L}_U^*$. The following assumption on the function space capacity is needed in order to develop the theoretical results.

Assumption 3. There exists a constant $\mathbf{C} > 0$, and $q \in (0, 2)$ such that for any $\varepsilon > 0$, $0 < \lambda \le \lambda_{\max} < \infty$, $R_{\max} < \infty$, the following condition on metric entropy is satisfied,

$$\log\Big(\max\big\{\mathcal{N}(\varepsilon, \Theta_1, \|\cdot\|_\infty), \mathcal{N}(\varepsilon, \Theta_2, \|\cdot\|_\infty), \mathcal{N}(\varepsilon, \Xi, \|\cdot\|_\infty), \mathcal{N}(\varepsilon, \Theta_3, \|\cdot\|_\infty)\big\}\Big) \le \mathbf{C}\left(\frac{\max\{\frac{4R_{\max}+(2-\gamma)\lambda_{\max}}{2-2\gamma}, 1\}}{\varepsilon}\right)^q,$$

where $\|\cdot\|_\infty$ denotes the supreme norm and $(V_\lambda^{\theta_1}, \pi_\lambda^{\theta_2}, \Psi^{\xi}, \psi^{\theta_3}) \in \Theta_1 \times \Theta_2 \times \Xi \times \Theta_3$.

Assumption 3 characterizes the complexity of the function spaces. In general, it is more difficult to estimate the functions as ϵ decreases. This assumption is satisfied in the function spaces such as the RKHS and Sobolev space (Geer and van de Geer 2000; Steinwart and Christmann 2008). Moreover, we compare this assumption with the assumption 3 in (Antos, Szepesvári, and Munos 2008) which considers parametric function spaces with finite effective dimension $D_{\mathcal{F}}$. Their assumption is equivalent to assuming $\log \mathcal{N}\left(\varepsilon, \Theta_1, \{\mathcal{D}_i\}_{i=1}^n\right) \le D_{\mathcal{F}} \log(1/\varepsilon)$, which is only able to account for the capacity of the finite dimension space. In contrast, our assumption is more general and allows our theorems can be applied to both finite and infinite dimension space.

Theorem 4.3. Under Assumptions 1–3 and S.1–S.4 provided in Section I, supplementary materials, for any $\delta \in (0, 1]$, $q \in (0, 2)$ and $\lambda \in (0, \lambda_{\max})$, the excess risk $\mathcal{E}(\mathcal{L}_U) := \mathcal{L}_U(\widehat{V}_\lambda^{\theta_1}, \widehat{\pi}_\lambda^{\theta_2}, \widehat{\Psi}^{\xi}, \widehat{\psi}^{\theta_3}) - \mathcal{L}_U^*$ is upper bounded with probability at least $1 - \delta$ for a sufficiently large T, that is

$$\mathcal{E}(\mathcal{L}_U) \le c_2 n^{-1} + c_3\sqrt{\log\left(\frac{2}{\delta}\right)} n^{-\frac{1}{2}} + J_1(\delta)\left(\frac{\log\log\left(2^{-\frac{\delta_1}{1+\delta_1}} c_4^{\frac{\delta_1}{1+\delta_1}} T^{\frac{\delta_1}{1+\delta_1}}\right)}{T^{\frac{\delta_1}{1+\delta_1}}}\right)^{1/2} + J_2(\delta)\left(T^{\frac{1-\delta_1}{(\delta_1+1)(2+q)}}\right) + J_3(\delta)\left(\frac{\log\left(c_5 T^{\frac{2}{(1+\delta_1)}}\right)}{T^{\frac{2}{2+q}}}\right)^{1/2}, \tag{4.1}$$

where the terms $\{J_1(\delta), J_2(\delta), J_3(\delta)\}$ are functions of $\delta, \delta_1, \mathbf{C}, c_4, q, U_{\max}, \lambda_{\max}, R_{\max}, \gamma$, and ζ; additionally, the constants terms $\{c_2, c_3, c_4, c_5\}$ depend on $\delta, \delta_1, \mathbf{C}, U_{\max}, \lambda_{\max}, R_{\max}$, and γ.

If the constant and logarithmic terms are omitted, the excess risk can be simplified and achieves the rate

$$\mathcal{E}(\mathcal{L}_U) \asymp \mathcal{O}\left(n^{-\frac{1}{2}} \vee \left(T^{\frac{\delta_1-1}{\delta_1+1}}\right)^{-\frac{1}{2+q}} \vee \left(\frac{T^{\frac{\delta_1}{1+\delta_1}}}{\log\log(T^{\frac{\delta_1}{1+\delta_1}})}\right)^{-\frac{1}{2}}\right).$$

The first two terms in (4.1) control the statistical error from the empirical estimation over n iid trajectories. The third and fourth terms bound the trajectory-based stochastic error from the variability inherent in weakly dependent within-trajectory samples. It observes that the third and fourth terms depends on the complexity of the function spaces $\Theta_1 \times \Theta_2 \times \Xi \times \Theta_3$. In addition, the last term is a remainder term due to using the *block devices* in dealing with the sample dependency. In general, this remainder term converges to zero much faster than other main terms, especially when the sample dependency is weak in that δ_1 is large. Hence, the remainder term does not affect the established convergence rate.

Theorem 4.3 provides a more powerful result than just a consistency of estimation as it establishes a finite-sample upper bound guarantee for the risk. In particular, if the process $\{(S^t, A^t)\}_{t\ge1}$ forgets its past history sufficiently fast, that is, $\delta_1 \to \infty$, then $T^{(\delta_1-1)/(\delta_1+1)}$ and $T^{\delta_1/(1+\delta_1)}$ both converge to T. Since the term $\log\log(T^{\delta_1/(1+\delta_1)})$ is a negligible to $\mathcal{O}(T^{\delta_1/(1+\delta_1)})$, we may achieve the optimal sample complexity upper bound $\mathcal{O}(n^{-1/(2+q)})$ if T is the same order of n. In particular, when the state space $\mathcal{S}$ is an open Euclidean ball in $\mathbb{R}^d$, and for a second-order Sobolev space $\mathbb{W}^{j,2}$ with $j > d/2$, one can choose $q = d/j$ to obtain the risk upper bound of the rate $\mathcal{O}(n^{-d/2(j+d)})$ if T is the same order of n. Besides establishing the risk bound in infinite dimensional function spaces, Theorem 4.3 can be also applied for the finite dimensional function spaces. That is, Theorem 4.3 recovers the best upper error bound $n^{-1/2}$ when $q \to 0$.

In the following, we provide a finite sample upper bound on the performance error of the estimated optimal value function. Let $K(\cdot,\cdot)$ be a continuous integral strictly positive definite kernel on the compact metric space $\mathcal{S} \times \mathcal{A}$, and there exists an orthonormal basis $e_1, e_2, \ldots$, for $L^2(\mathcal{S} \times \mathcal{A})$. Also, we let $\kappa_1, \kappa_2, \ldots$, be corresponding eigenvalues such that $K(\{s,a\}, \{\widetilde{s}, \widetilde{a}\}) = \sum_{j=1}^\infty \kappa_j e_j(\{s,a\}) e_j(\{\widetilde{s}, \widetilde{a}\})$, where $\{s,a\}, \{\widetilde{s}, \widetilde{a}\} \in \mathcal{S} \times \mathcal{A}$. The L^2 norm that $\sqrt{\int f^2(s) \mathrm{d} d_{\pi_b}(s)}$ for the observed data distribution d_{π_b} over $\mathcal{S}$ is denoted by $\|f\|_{L^2}$.

Theorem 4.4. Under Assumptions 1–3 and S.1–S.4 provided in Section I, supplementary materials, for $\lambda \in (0, \lambda_{\max})$ and the finite minimum eigenvalue $\kappa_{\min} = \min\{\kappa_1, \kappa_2, \ldots,\} < \infty$, the performance error between the estimated optimal value function $\widehat{V}_\lambda^{\theta_1}$ with respect to the proximal Bellman operator $\mathcal{B}_\lambda$

and the optimal value function V^{π^*} with respect to the standard Bellman operator $\mathcal{B}$ is upper bounded by

$$\|\widehat{V}_\lambda^{\theta_1} - V^{\pi^*}\|_{L^2}^2 \le \frac{c_6\mathcal{E}(\mathcal{L}_U)}{\kappa_{\min}(1-\gamma)^2} + \frac{c_7\lambda^2}{(1-\gamma)^2}|\mathcal{A}| + \frac{c_8\lambda^2|\mathcal{A}| + \lambda^2 c_9}{(1-\gamma)^2|\mathcal{A}|},$$

where c_6, c_7, c_8, c_9 are some terms of $\delta_1, \mathbf{C}, c_2, c_3, c_4, c_5, J_1(\delta)$, $J_2(\delta), J_3(\delta), q, U_{\max}, \zeta, \lambda_{\max}, R_{\max}$, and γ in which $c_7 > c_8$.

The above bound provides an insight regarding the performance error of the proposed method. Specifically, under the regularity conditions, the L^2 distance between the estimated optimal proximal value function $\widehat{V}_\lambda^{\theta_1}$ and the optimal value function V^{π^*} is upper bounded, and this gap diminishes with the growth of sample size n and time length T, and with the decay of the smoothing parameter λ. Note that the last two terms is approaching to zero as λ is sufficiently small. This implies that only a small smoothing bias is involved in the finite sample bound of $\|\widehat{V}_\lambda^{\theta_1} - V^{\pi^*}\|_{L_2}^2$. Moreover, combined with the sample bound for an excess risk $\mathcal{E}(\mathcal{L}_U)$ in Theorem 4.3, Theorem 4.4 indicates that the behavior of the upper bound as a function of samples $\mathcal{O}(n^{-1/(2+q)})$ is the best if T is the same order of n. In general, a large λ increases the smoothing error but decreases the approximation error as the solution function space is better behaved due to stronger smoothness. However, the approximation error is not reflected in our sample bound because we make the zero approximation error assumption, that is, Assumption S.4 in supplementary materials.

5. Implementation and Algorithm

For optimizing (3.14), the functions $(V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi)$ are required to be parameterized by a class of models for practical implementations. One may parameterize $(V_\lambda^{\pi_\lambda}, \pi_\lambda, \Psi, \psi)$ by $\{V_\lambda^{\pi_\lambda}(\cdot;\beta), \pi_\lambda(\cdot;\theta), \Psi(\cdot;\omega), \psi(\cdot;\xi)\}$, where $(\theta, \beta, \omega, \xi)$ are associated parameters. Note that the policy (actor) $\pi_\lambda(\cdot;\theta)$ and the value function (critic) $V_\lambda^{\pi_\lambda}(\cdot;\beta)$ are separately represented by the two sets of parameters (θ, β). This parameterization representation is aligned with the *actor-critic* paradigm in estimating the policy, which requires assistance and evaluation from the critic (Mnih et al. 2016). In our algorithm development, we introduce a unified *actor-critic* paradigm where the actor itself can play a critic role such that the actor can be self-supervised during the training process. We allow both $V_\lambda^{\pi_\lambda}$ and π_λ to be represented in terms of $Q_\lambda^{\pi_\lambda}$ based on the connections in (3.3) and (3.5), respectively. In other words, it is sufficient to only parameterize $Q_\lambda^{\pi_\lambda}$ for $V_\lambda^{\pi_\lambda}$ and π_λ. Specifically, if we parameterize $Q_\lambda^{\pi_\lambda}$ by $Q_\lambda^{\pi_\lambda}(\cdot;\theta)$, then $\{V_\lambda^{\pi_\lambda}, \pi_\lambda\}$ can be parameterized by the same set of parameter θ, that is, $\{V_\lambda^{\pi_\lambda}(\cdot;\theta), \pi_\lambda(\cdot;\theta)\}$. One advantage of the new diagram is to reduce the parameter space, form two sets of parameters to one set, and hence relaxing the computational intensity. Another key advantage is that the new diagram only need to track the target policy $\pi_\lambda(\cdot;\theta)$ instead of tracking the nonstationary target $V_\lambda^{\pi_\lambda}(\cdot;\theta)$ which usually results in divergence issues.

For any state-action pair (s, a), we follow the new unified *actor-critic* framework to approximate $Q_\lambda^{\pi_\lambda}(s, a)$ using basis function approximations. The state-action function $Q_\lambda^{\pi_\lambda}(s, a; \theta)$ is represented by a linear combination of basis functions $\theta^{\mathrm{T}}\varphi(s, a)$, where θ is a p_Q-dimensional weight vector and $\varphi(s, a)$ is an column vector of nonlinear basis functions computed at (s, a). For $m = p_Q/|\mathcal{A}|$, the vector $\varphi(s, a)$ sets the basis function value $\varphi(s) = \big(\varphi_1(s), \varphi_2(s), \ldots, \varphi_m(s)\big)$ in the corresponding slot for a specific action a, while the values of basis function for the rest of the actions are set to be zero. That is

$$\varphi(s, a_1) = \begin{bmatrix} \varphi_1(s) \\ \cdots \\ \varphi_m(s) \\ \mathbf{0}_{m(|\mathcal{A}|-1)\times 1} \end{bmatrix}, \ldots,$$

$$\varphi(s, a_k) = \begin{bmatrix} \mathbf{0}_{m(k-1)\times 1} \\ \varphi_1(s) \\ \cdots \\ \varphi_m(s) \\ \mathbf{0}_{m(|\mathcal{A}|-k)\times 1} \end{bmatrix}, \ldots,$$

$$\varphi\big(s, a_{|\mathcal{A}|}\big) = \begin{bmatrix} \mathbf{0}_{m(|\mathcal{A}|-1)\times 1} \\ \varphi_1(s) \\ \cdots \\ \varphi_m(s) \end{bmatrix},$$

where $\mathbf{0}$ is a zero vector. Similarly, we model the two Lagrangian functions as $\Psi(s;\omega) = \omega^{\mathrm{T}}\varphi(s)$ and $\psi(a|s;\xi) = \xi^{\mathrm{T}}\varphi(s, a)$. The flexibility of these working models can be achieved by selecting different $\varphi(\cdot)$, such as B-splines, radial basis and Fourier basis functions. Also note that $Q_\lambda^{\pi_\lambda}(s, a; \theta)$ can be parameterized by a nonlinear approximation architecture and the optimization convergence is guaranteed in the pT-learning framework.

In the following, we reformulate the empirical risk minimization (3.14) to a nonlinear programming problem for which the objective is converted to a quadratic term with a nonlinear equality constraint and two nonlinear inequality constraint,

$$\begin{aligned} \min_{\theta,\omega,\xi} \widehat{\mathcal{L}_U}(\theta, \omega, \xi) &= \frac{\zeta}{n}\sum_{i=1}^{n}\big(D_i(\theta) - P_i(\theta) - Z_i(\omega) \\ &\quad + W_i(\xi)\big)^{\mathrm{T}}\Omega_i\big(D_i(\theta) - P_i(\theta) - Z_i(\omega) + W_i(\xi)\big) \\ &\text{s.t. } \pi_\lambda(a|s;\theta)\cdot\psi(a|s;\xi) = 0,\ \psi(a|s;\xi) \ge 0 \\ &\quad \text{and } -\frac{\lambda}{2} \le \Psi(s;\omega) \le 0, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}, \end{aligned} \tag{5.1}$$

where $\Omega_i \in \mathbb{R}^{T\times T}$ is a weight matrix with $\{\Omega_i\}_{jk} = \big\{2\big(K(\{S_i^j, A_i^j\}, \{S_i^k, A_i^k\}) - \mathbb{1}(j = k)\big)/T(T-1)\big\}_{jk}$, and $D_i(\theta), P_i(\theta), Z_i(\omega), W_i(\xi)$ are provided in Section A, Supplementary Materials.

To solve this quadratic programming problem with nonlinearity constraints, one may consider applying sequential quadratic programming (Fletcher 2010) to optimize the objective function with linearized constraints. However, this requires several derivatives which are required to be solved analytically before iteration. Therefore, it can be quite cumbersome in practical implementations. Another approach is to apply exterior penalty methods (Boyd, Boyd, and Vandenberghe 2004), which convert a constrained problem to a series of unconstrained optimization problems. However, the size of constraints is $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$, and thus the computation is intensive if either $|\mathcal{S}|$ or $|\mathcal{A}|$ is large.

Alternatively, we propose a computationally more efficient algorithm, solving an unconstrained optimization via imposing certain restrictions on the representation of $\{Z_i(\omega), W_i(\xi)\}$ such that the two Lagrangian functions satisfy the constraints automatically. Although this re-parameterization may sacrifice certain model flexibility, it gains computational advantages as it solves a simpler quadratic optimization problem. More specifically, we parameterize $Z_i(\omega)$ by flipping a sigmoid function: $\Psi(s;\omega) = -(\lambda/2)\cdot\left(1+\exp(-k_0(\omega^{\mathrm{T}}s-b_0))\right)^{-1} \in [-\lambda/2, 0]$, where b_0 is the sigmoid's midpoint and k_0 is the logistic growth rate. Obviously, the first inequality constraint $-\lambda/2 \le \Psi(s;\omega) \le 0$ is automatically satisfied under this parameterization. Observe that the $\pi_\lambda(a|s;\theta) > 0$ when $a \in \mathcal{K}(s)$, and $\pi_\lambda(a|s;\theta) = 0$ otherwise. Therefore, we can further reduce our parameter space via modeling $\psi(a|s;\theta)$ as a function of θ instead of ξ. That is, $\psi(a|s;\theta) = \left(\frac{\sum_{a\in\mathcal{K}(s)}\theta^{\mathrm{T}}\varphi(s,a)-\lambda}{\lambda|\mathcal{K}(s)|} - \frac{\theta^{\mathrm{T}}\varphi(s,a)}{\lambda}\right)^{+}$, where the constraints $\pi_\lambda(a|s;\theta)\cdot\psi(a|s;\theta)=0$ and $\psi(a|s;\theta)\ge 0$ are always satisfied, and $W_i(\cdot)$ in (5.1) are changed to $W_i(\theta)$. The parameter ξ is replaced by θ, and hence the computational complexity is further reduced. The representation of $\widehat{\mathcal{L}_U}(\theta,\omega,\xi)$ under $\Psi(s;\omega)$ and $\psi(a|s;\theta)$, that is, $\widehat{\mathcal{L}_U}(\theta,\omega)$, and its derivatives are provided in Section A, Supplementary Materials.

Algorithm 1 pT-Learning with Stochastic Gradient Descent

1: **Input** observed data $\mathcal{D}_{1:n}$ as the transition pairs format $\{(S_i^t, A_i^t, R_i^t, S_i^{t+1}) : t = 1,\dots,T\}_{i=1}^n$.
2: **Initialize** the primary and auxiliary parameters $(\theta,\omega) = (\theta_0,\omega_0)$, the mini-batch size n_0, the learning rates $\alpha_\theta = \alpha_\theta^0, \alpha_\omega = \alpha_\omega^0$, the scale parameter $\zeta = \zeta_0$, the factors $(\kappa_e = 1, \kappa_\pi = 1, \kappa_\alpha = 1)$, the sparsity parameter $\lambda = \lambda_0$, the bandwidth $bw = bw_0$, and the stopping criterion ε.
3: **For** $k = 1$ to $k = max.iter$
4: Randomly sample a mini-batch $\{(S_i^t, A_i^t, R_i^t, S_i^{t+1}) : t = 1,\dots,T\}_{i=1}^{n_0}$.
5: Compute the gradient w.r.t. θ as $\bar{\Delta}_\theta = \zeta_0\mathbb{P}_{n_0}[D_i(\theta) - P_i(\theta) + W_i(\theta) - Z_i(\omega)]^{\mathrm{T}}\Omega_i[\kappa_e\nabla_\theta D_i(\theta) + \kappa_\pi\nabla_\theta P_i(\theta) + \kappa_\alpha\nabla_\theta W_i(\theta)]$.
6: Compute the gradient w.r.t. ω, $\bar{\Delta}_\omega = \zeta_0\mathbb{P}_{n_0}[D_i(\theta) - P_i(\theta) + W_i(\theta) - Z_i(\omega)]^{\mathrm{T}}\Omega_i\nabla_\omega Z_i(\omega)$.
7: Decay the learning rate $\alpha_\theta^k = \mathcal{O}(k^{-1/2})$, $\alpha_\omega^k = \mathcal{O}(k^{-1})$.
8: Update the parameters of interest as $\theta^k \leftarrow \theta^{k-1} - \alpha_\theta^k\bar{\Delta}_\theta$, $\omega^k \leftarrow \omega^{k-1} - \alpha_\omega^k\bar{\Delta}_\omega$.
9: Stop if $\|\theta^k - \theta^{k-1}\| \le \varepsilon$.
10: **Return** $\widehat{\theta} = \theta^k$.

The minimization of $\widehat{\mathcal{L}_U}(\theta,\omega)$ requires $\mathcal{O}(nqT^2)$ time complexity in calculating the exact gradients. Here, we implement a stochastic gradient descent (SGD) algorithm, where the training is on mini-batch datasets and is faster than a vanilla gradient descent or BFGS algorithm. We summarize details of the proposed algorithm in Algorithm 1, where the derivations of the gradients are provided in Section A, Supplementary Materials. In addition, the time complexity of calculating the gradient requires $\mathcal{O}(|\mathcal{A}|\log(|\mathcal{A}|)Tqn)$ by a naive sorting algorithm, and we can further improve the time complexity to $\mathcal{O}(|\mathcal{A}|Tqn)$ by using the bucket-sorting algorithm (Blum et al. 1973).

6. Simulation Studies

We conduct two numerical experiments to evaluate the finite sample performance of the proposed method. In the first experiment, we consider a binary treatment setting following a benchmark generative model (Luckett et al. 2020; Liao, Klasnja, and Murphy 2020). In the second experiment, we mimic an mHealth cohort study aiming to deliver personalized interventions with 12 choices for managing an individual's glucose level. In both experiments, we compare our approach to state-of-the-art methods including linear, polynomial and Gaussian V-learning (Luckett et al. 2020), and Greedy GQ-learning (Ertefaie and Strawderman 2018). The proposed method is available in our `proximalDTR` R package.

In the first example, we let the current state $S_i^t = \left(S_{i,1}^t, S_{i,2}^t\right)^T$ be a two-dimensional vector and the current action $A_i^t \in \{1, 0\}$. The next state S_i^{t+1} is generated according to $S_{i,1}^{t+1} = (3/4)(2A_i^t - 1)S_{i,1}^t + (1/4)S_{i,1}^t S_{i,2}^t + \epsilon_{i,1}^t$ and $S_{i,2}^{t+1} = (3/4)(1 - 2A_i^t)S_{i,2}^t + (1/4)S_{i,1}^t S_{i,2}^t + \epsilon_{i,2}^t$, where the random noises $\epsilon_{i,1}^{t+1}$ and $\epsilon_{i,2}^{t+1}$ follow an independent Gaussian distribution $N(0, 0.5^2)$. Note that assigning $A_i^t = 1$ imposes a positive effect on $S_{i,1}^{t+1}$ but has a negative effect on $S_{i,2}^{t+1}$. We define a nonlinear utility function $u(S_i^{t+1}, S_i^t, A_i^t)$ such that

$$R_i^t = u(S_i^{t+1}, S_i^t, A_i^t) = (1/4)(S_{i,1}^{t+1})^3 + 2S_{i,1}^{t+1} + (1/2)(S_{i,2}^{t+1})^3 + S_{i,1}^{t+1} + (1/4)(2A_i^t - 1).$$

The initial state S_i^1 follows a Gaussian distribution $N(0, I_{2\times 2})$ and A_i^t is randomly assigned with an equal probability for each treatment arm as in micro-randomized trials.

We consider different scenarios in which the number of patients $n = \{25, 50, 100\}$ and the follow-up time length $T = \{24, 36, 48\}$. The discount factor γ is set to be 0.9, focusing on long-term benefits. To specify the basis function $\varphi(s)$ mentioned in Section 5, we consider a cubic spline containing six knots located in equal space of interval [0, 1], and then apply it to the state variables normalized between 0 and 1.

The objective function $\widehat{\mathcal{L}_U}(\theta,\omega)$ may not be convex with respect to both θ and ω. Therefore, we adopt a multiple initialization method to determine an appropriate initial point. Specifically, we choose an initial point with the smallest objective value among 50 randomly generated initial points. In our two numerical experiments, we consider a fixed $\lambda = 0.1$. Alternatively, we can also use a k-fold cross-validation procedure to select sparsity parameter λ. We choose an optimal λ which maximizes the lower bound of the empirical discounted sum of utilities, that is, $\widehat{\lambda} = \arg\max_\lambda \frac{1}{k}\sum_{r=1}^k \mathbb{P}_{n(r)}\widehat{V_\lambda^{\pi_\lambda}}^{(r)}(S^1) - \frac{\lambda\phi(0)}{1-\gamma}$, where $\widehat{V_\lambda^{\pi_\lambda}}^{(r)}$ is the proximal value function estimator on the rth training set, and $\mathbb{P}_{n(r)}$ is the empirical measure on the initial state S_1 for the rth validation set. Note that the cross-validation procedure may help to select a better λ but it is computationally expensive. The cross-validation results, the tuning parameters set-up, and sensitivity analyses on different choices of λ for model performance are provided in Section C, Supplementary Materials. To evaluate the model performance, we use the mean utility under the estimated policy as an evaluation criterion. Specifically, after obtaining the estimated policy, we simulate

100 independent individual patients following this estimated policy over 100 stages, and calculate the mean utility. Therefore, a larger value of the mean utility indicates a better policy. In Table 1, we report the average and standard deviations of the mean utilities over 50 simulations.

Table 1 shows that the proposed method outperforms the competing methods in all scenarios. When $n = 100$ and $T = 24$, compared to the baseline, that is, the observed mean utility, the proposed method improves 0.4249 while the polynomial V-learning and Greedy GQ-learning only improve 0.2510 and 0.0834, respectively. This is mainly because the proposed method does not impose restrictions on the class of policies as in V-learning. Also, the induced policy can be automatically adjusted between deterministic and stochastic policy models. In addition, the proposed method uses the smoothed proximal Bellman operator, and thus avoids discontinuity and instability. Moreover, the proposed method has advantages in off-policy training due to using the proximal temporal consistency property. In contrast, Greedy GQ-learning tends to solve the hard Bellman optimality equation, which could be difficult without large amounts of samples. In terms of computations, pT-Learning also achieves great gains over V-learning. The actual computation time tables are provided in Section C of the supplementary materials.

In the second example, we simulate cohorts of patients with Type 1 diabetes to mimic the mobile health study (Maahs et al. 2012). The study aims to achieve long-term glycemic control and maintain the rest of the health index in the desired ranges. Specifically, our hypothetical mHealth study targets to searching the optimal policy for suggesting multi-channel interventions; the insulin injection (IN), physical activity (PA), and dietary intake (DI), to control the blood glucose (BG) level in the desired range while maintaining healthy levels of Adiponectin (AD) and blood pressure (BP) (Fidler, Elmelund Christensen, and Gillard 2011). Hence, the immediate utility R_i^t is defined by a weighted summation of patients' health status as following,

$$\begin{aligned} R_i^t = {} & \alpha_1 \mathbb{1}(70 \le \mathrm{BG}_i^{t+1} \le 120) \\ & + \alpha_2 \mathbb{1}(\mathrm{BG}_i^{t+1} < 70 \text{ or } \mathrm{BG}_i^{t+1} > 150) \\ & + \alpha_3 \mathbb{1}(120 \le \mathrm{BG}_i^{t+1} \le 150) \\ & + \alpha_4 \mathbb{1}(5 \le \mathrm{AD}_i^{t+1} \le 23) \\ & + \alpha_5 \mathbb{1}(\mathrm{AD}_i^{t+1} < 5 \text{ or } \mathrm{AD}_i^{t+1} > 23) \\ & + \alpha_6 \mathbb{1}(66 \le \mathrm{BP}_i^{t+1} \le 80) \\ & + \alpha_7 \mathbb{1}(\mathrm{BP}_i^{t+1} < 66 \text{ or } \mathrm{BP}_i^{t+1} > 80), \end{aligned}$$

where $(\alpha_1, \ldots, \alpha_7) = (3, -3, -1, 2, -1, 2, -1)$ are weights reflecting the clinical consequences.

The patients are assigned treatment from a combination of the insulin injections (Yes/No), physical activity (No/Moderate/Strong) and dietary intake (Yes/No). Note that there are total 12 different treatment choices, that is, $\mathcal{A} = \{1, \ldots, 12\}$. The details of the 12 different treatment combinations are enumerated in Section C, supplementary materials. At each stage, the treatment $A_i^t \in \mathcal{A}$ is randomly assigned with equal probability for each treatment arm. The patient's state $(\mathrm{BG}_i^t, \mathrm{AD}_i^t, \mathrm{BP}_i^t)$ evolves according to the given dynamic model: $\mathrm{BG}_i^{t+1} = \gamma_{11}\mu_{\mathrm{BG}} + \gamma_{12}\mathrm{BG}_i^t - \gamma_{13}\mathrm{AD}_i^t + \sum_{a\in\mathcal{A}} \mu_{1a}\mathbb{1}(A_i^t = a) + \epsilon_{1i}^{t+1}$; $\mathrm{AD}_i^{t+1} = \gamma_{21}\mu_{\mathrm{AD}} + \gamma_{22}\mathrm{AD}_i^t + \gamma_{23}\mathrm{BG}_i^t + \sum_{a\in\mathcal{A}} \mu_{2a}\mathbb{1}(A_i^t = a) + \epsilon_{2i}^{t+1}$; $\mathrm{BP}_i^{t+1} = \gamma_{31}\mu_{\mathrm{BP}} + \gamma_{32}\mathrm{BP}_i^t + \sum_{a\in\mathcal{A}} \mu_{3a}\mathbb{1}(A_i^t = a) + \epsilon_{3i}^{t+1}$, where $\epsilon_{1i}, \epsilon_{2i}$ and ϵ_{3i} are individual-level Gaussian random noises. The values of coefficients in the generative model are provided in Section C, supplementary materials.

The cohort data is generated under different scenarios with a follow-up time length $T = \{24, 36\}$ and sample size $n = \{25, 75, 100\}$. The discount factor γ is set to be 0.9. We use the same evaluation criterion as in Simulation Example 1 except that the testing set is simulated over 36 stages. The results are summarized in Table 2. The pT-Learning approach achieves the best performance. For example, when $n = 100$ and $T = 36$, pT-Learning achieves 56.7% and 83.8% improvements compared to Gaussian V-learning and Greedy GQ-learning, respectively. The improvement of pT-Learning compared to V-learning is due to the sparse property of pT-Learning. For a better illustration, we visualize the sparsity and compare the estimated policy

Table 1. Example 1: the average and standard deviation (in parentheses) of the mean utilities under the estimated optimal policy based on 50 simulation runs.

n	T	Proposed	Greedy-GQL	Linear VL	Poly VL	Gauss VL	Observed
25	24	0.3827(0.121)	0.0787(0.175)	0.2561(0.011)	0.2564(0.011)	0.2561(0.011)	0.0033
	36	0.4153(0.065)	0.0716(0.234)	0.2560(0.013)	0.2561(0.011)	0.2558(0.014)	0.0025
	48	0.4001(0.078)	0.0840(0.213)	0.2578(0.012)	0.2578(0.014)	0.2575(0.012)	0.0033
50	24	0.4154(0.080)	0.0844(0.209)	0.2564(0.012)	0.2569(0.012)	0.2564(0.012)	0.0092
	36	0.4029(0.093)	0.0829(0.235)	0.2570(0.010)	0.2574(0.012)	0.2569(0.011)	0.0080
	48	0.3888(0.084)	0.0836(0.236)	0.2564(0.012)	0.2570(0.012)	0.2530(0.026)	0.0036
100	24	0.4290(0.059)	0.0875(0.240)	0.2529(0.016)	0.2551(0.011)	0.2547(0.011)	0.0041
	36	0.4050(0.062)	0.0780(0.266)	0.2538(0.010)	0.2542(0.010)	0.2535(0.010)	0.0043
	48	0.4112(0.052)	0.0945(0.254)	0.2548(0.011)	0.2552(0.015)	0.2543(0.017)	0.0090

Table 2. Example 2 (mHealth study): the average and standard deviation (in parentheses) of the mean utilities under the estimated optimal policy based on 50 simulation runs.

n	T	Proposed	Greedy-GQL	Linear VL	Poly VL	Gauss VL	Observed
25	24	2.489(0.631)	1.561(1.034)	1.743(0.748)	1.555(0.652)	1.787(0.784)	1.481
	36	2.878(0.568)	1.500(1.244)	1.694(0.702)	1.725(0.705)	1.770(0.674)	1.473
75	24	2.685(0.268)	1.588(1.310)	1.667(0.482)	1.724(0.862)	1.885(0.898)	1.464
	36	3.247(0.510)	1.772(1.277)	1.713(0.423)	2.162(0.954)	1.918(0.894)	1.478
100	24	2.785(0.401)	1.720(1.232)	1.561(0.317)	1.849(0.904)	1.981(0.573)	1.471
	36	3.409(0.429)	1.854(1.453)	1.768(0.622)	2.105(0.912)	2.177(1.086)	1.473

distribution of pT-Learning and V-learning in Section C of the supplementary materials. On the other hand, pT-Learning is better than Greedy GQ-learning mainly because Greedy GQ-learning requires modeling the entire data-generating process, and hence produces a large over-estimation error. Worst of all, the Greedy GQ-learning approach is designed to take the greedy action based on the fitted model. This again magnifies the over-estimation error and ultimately leads to a huge sub-optimality. In contrast, pT-Learning only requires modeling the optimal policy and optimal value function. Additionally, pT-Learning does not take a greedy action, but follows an adaptive and sparse stochastic policy model.

7. Application to Ohio Type 1 Diabetes Data

We apply the proposed method to two cohorts of individuals from the OhioT1DM dataset (Marling and Bunescu 2020), which was used to study the long-term blood glucose management via the just-in-time interventions on the type 1 diabetes patients. Each cohort contains six individuals of age ranging within 20–60. All patients were on insulin pump therapy with continuous glucose monitoring (CGM) sensors, and the life-event data is collected via a mobile phone app. The physiological data of the first cohorts was collected by *Basis* sensor bands, and the second cohort used *Empatica* sensor bands. Specifically, the dataset includes CGM blood glucose levels measured per 5 min, insulin dose levels delivered to the patient, meal intakes, and corresponding carbohydrate estimates. Note that the heart rate measured per 5 min is available only for patients using the *Basis* sensor band (the first cohort), while the magnitude of acceleration aggregated per minute is only available with the *Empatica* sensor band (the second cohort). The data also include other features such as the self-reported times of work, sleep and stress, etc. Based on the preliminary investigation, each patient has distinct blood glucose dynamics. Therefore, we follow the pre-processing strategy used in Zhu, Lu, and Song (2020), and treat the data of each patient as a single dataset. Then we estimate the optimal policy by treating each day as an independent sample. For the first cohort, we consider a binary intervention setting, that is, whether or not to provide the insulin injection; and for the second cohort, we study the individualized dose-finding problem by discretizing the continuous dose level into 14 disjoint intervals for intervention options. We compare the performance of pT-Learning with the same competing methods as in Section 6.

We summarize the collected measurements over 60-min intervals such that the length of each trajectory is $T = 24$. After removing missing samples and outliers, each dataset contains $n = 15$ trajectories on average. For the first six patients who wore the *Basis* sensor band, the patient's states at each stage include the average blood glucose levels $S^t_{i,1}$, the average heart rate $S^t_{i,2}$ and the total carbohydrates $S^t_{i,3}$ intake from time $t-1$ and t. For others equipped with the *Empatica* sensor band, the states are the same as the first six patients, except that the average heart rate is substituted by the average magnitude of acceleration. Here, the utility is defined as the average of the index of glycemic control (Rodbard 2009) between time $t-1$ and t, measuring the health status of the patient's glucose level. That is $R^t_i = -\frac{\mathbb{1}(S^t_{i,1}>140)|S^t_{i,1}-140|^{1.35}+\mathbb{1}(S^t_{i,1}<80)(S^t_{i,1}-80)^2}{30}$, where R^t_i is nonpositive and a larger value is preferred. Our goal is to maximize the expected discounted sum of utilities $\mathbb{E}_\pi \sum_{t\geq 1}\gamma^{t-1}R^t$. In the first cohort study, the treatment is binary, that is, $A^t_i \in \{0,1\}$. In the second study, we provide the optimal insulin dose suggestion via the uniform discretization of the continuous dose level, that is, $A^t_i \in \{0 = A_{(1)} < \cdots < A_{(14)} = \max(A)\}$.

Since the data-generating process is unknown, it is hard to use the mean utility under the estimated policy as the evaluation criterion as in Section 6. Instead, we follow Luckett et al. (2020) to use the Monte Carlo approximation of the expected discounted sum of utilities for evaluating the model performance, that is, $\mathbb{P}_n\widehat{V^\pi}(S^1_i)$, where S^1_i is the initial state for the ith trajectory. In the competing methods, the quantity $\widehat{V^\pi}(\cdot)$ represents the estimated value function. In our method, we consider the lower bound of our estimated value function $\widehat{V^{\pi_\lambda}_\lambda}(\cdot)$, that is, $\mathbb{P}_n\widehat{V^{\pi_\lambda}_\lambda}(S^1_i) - (1-\gamma)^{-1}\widehat{\lambda}\phi(0)$, to mitigate the effects of the sparsity parameter $\widehat{\lambda}$ on the utilities. This quantity can be also interpreted as the worst-case of the discounted sum of utilities produced by pT-Learning. The discounted sum of observed utilities, that is, $\mathbb{P}_n\sum_{t\geq 1}\gamma^{t-1}R^t_i$, is used as the baseline.

In our experiments, we choose two discount factors, $\gamma = 0.9$ and $\gamma = 0.8$. For the $\gamma = 0.9$ setting, the boxplots of the relative improvements of utilities are provided in Figures 2 and 3 for

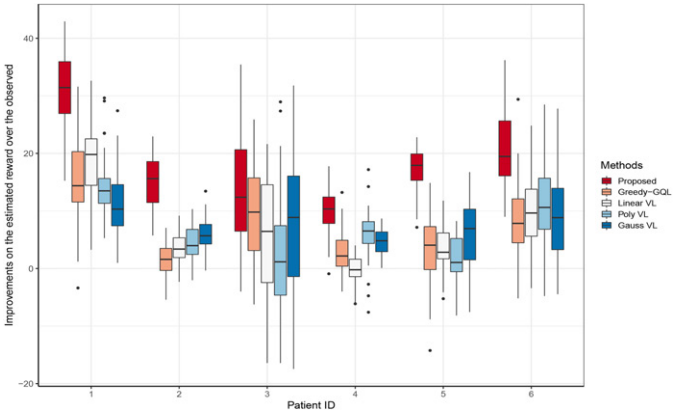

Figure 2. The first cohort patients: boxplots of the improvements on the discounted sum of utilities under estimated policy over 50 simulation runs, with $|\mathcal{A}| = 2$ and $\gamma = 0.9$.

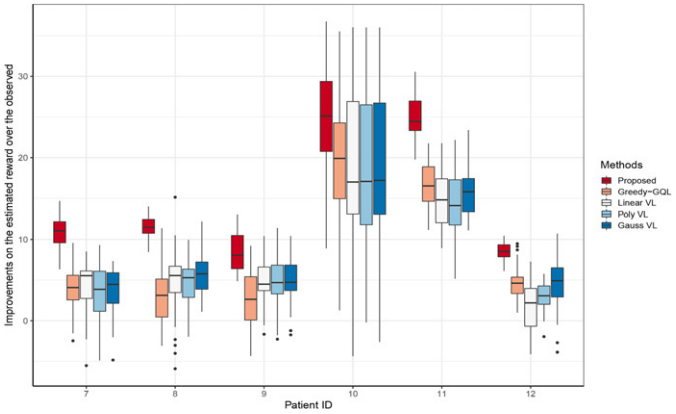

Figure 3. The second cohort patients: boxplots of the improvements on the discounted sum of utilities under estimated policy over 50 simulation runs, with $|\mathcal{A}| = 14$ and $\gamma = 0.9$.

the first and second cohort data, respectively. Additional results are presented in Section C, supplementary materials. Figure 2 indicates that pT-Learning achieves better improvements over competing methods across all patients. For example, for the Patient 1, the proposed method has 98.0% and 51.2% improvement rates compared to Greedy GQ-learning and the linear V-learning, respectively. This is mainly because pT-Learning does not impose restrictions on the class of policies and thus is more flexible. The standard deviation of the proposed method is the smallest among all approaches, reflecting that pT-Learning is relatively stable due to the smoothness property of the proximal Bellman operator and the implemented unified *actor-critic* framework. For the second cohort study with a large cardinality treatment space, pT-Learning substantially outperforms the competing methods in Figure 3. For the seventh patient, the improvement of pT-Learning to Greedy GQ-learning and Gaussian V-learning attains 175.4% and 187.0%, respectively. This result shows that pT-Learning achieves high efficiency in the continuous treatment space, and our sparse policy estimation has a clear benefit with large numbers of treatment options.

8. Discussion

In this article, we propose a novel proximal temporal consistency learning framework for estimating the optimal dynamic treatment regime in infinite time horizon settings. The constructed proximal Bellman operator directly leads to a smoothed Bellman optimality equation, while simultaneously inducing a sparse optimal policy. The proposed minimax policy estimator resolves the double sampling issue and can be easily optimized by a scalable and efficient SGD algorithm.

Several improvements and extensions are worth exploring in the future. First, we may extend our algorithm to deal with strong temporal dependency. The idea of the experience-replay (Mnih et al. 2015) might be useful. It is shown that the gradients calculated by the experience-replay algorithm are $1 - \mathcal{O}(n^{-1})$-nearly independent. Second, it is interesting to extend pT-Learning to the long-term average reward setting (Murphy et al. 2016). Third, developing statistical inference methods for quantifying uncertainty of the policy and value function is also important.

Under the nonstationary learning setting where the environment varies over time, a relatively high exploration is typically preferred, as it may lead to a better policy estimation in the long run. Therefore, the V-learning method potentially gains more benefits than pT-Learning in policy estimations due to a higher exploration rate. To improve the exploration ability of pT-Learning, one may consider to adopt the ε-greedy strategy (Sutton and Barto 1998). Other future directions include developing a rigorous extension to continuous action space with theoretical justifications, and constructing a state-varying $\lambda(s)$ which results in a group-wise smoothness and sparsity. The detailed discussion of the aforementioned extensions is provided in supplementary materials.

Supplementary Materials

The supplementary materials provide all technical proofs of main theorems as well as the additional discussions and numerical experiments, and theoretical results.

Acknowledgments

The authors thank the Editor, Associate Editor and the anonymous reviewers for their insightful suggestions and helpful feedback which improved the article significantly.

Disclosure Statement

The authors declare no financial or nonfinancial interest that has arisen from the direct applications of this research.

Funding

This work is supported by NSF grants DMS 2210640, DMS 1952406 and DMS 2210657.

References

Antos, A., Szepesvári, C., and Munos, R. (2008), "Learning Near-Optimal Policies with Bellman-Residual Minimization based Fitted Policy Iteration and a Single Sample Path," *Machine Learning*, 71, 89–129. [625,631,632]

Bach, F. (2017), "Breaking the Curse of Dimensionality with Convex Neural Networks," *The Journal of Machine Learning Research*, 18, 629–681. [630]

Baird, L. (1995), "Residual Algorithms: Reinforcement Learning with Function Approximation," in *Machine Learning Proceedings 1995*, pp. 30–37, Elsevier. [625,629]

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006), "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, 101, 138–156. [632]

Bertsekas, D. P. (1997), "Nonlinear Programming," *Journal of the Operational Research Society*, 48, 334–334. [626]

Blum, M., Floyd, R. W., Pratt, V. R., Rivest, R. L., and Tarjan, R. E. (1973), "Time Bounds for Selection," *Journal of Computer and System Sciences*, 7, 448–461. [634]

Borisov, I., and Volodko, N. (2009), "Exponential Inequalities for the Distributions of Canonical U-and V-Statistics of Dependent Observations," *Siberian Advances in Mathematics*, 19, 1–12. [631]

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press. [633]

Chow, Y., Nachum, O., and Ghavamzadeh, M. (2018), "Path Consistency Learning in Tsallis Entropy Regularized MDPs," in *International Conference on Machine Learning*, pp. 979–988. [625,629]

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. (2018), "SBEED: Convergent Reinforcement Learning with Nonlinear Function Approximation," in *International Conference on Machine Learning*, pp. 1125–1134, PMLR. [625]

Dann, C., Neumann, G., and Peters, J. (2014), "Policy Evaluation with Temporal Differences: A Survey and Comparison," *Journal of Machine Learning Research*, 15, 809–883. [626,630]

Ertefaie, A., and Strawderman, R. L. (2018), "Constructing Dynamic Treatment Regimes over Indefinite Time Horizons," *Biometrika*, 105, 963–977. [625,626,630,634]

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020), "A Theoretical Analysis of Deep Q-Learning," in *Learning for Dynamics and Control*, pp. 486–489, PMLR. [629]

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016), "Regularized Policy Iteration with Nonparametric Function Spaces," *The Journal of Machine Learning Research*, 17, 4809–4874. [631]

Fidler, C., Elmelund Christensen, T., and Gillard, S. (2011), "Hypoglycemia: An Overview of Fear of Hypoglycemia, Quality-of-Life, and Impact on Costs," *Journal of Medical Economics*, 14, 646–655. [635]

Fletcher, R. (2010), "The Sequential Quadratic Programming Method," in *Nonlinear Optimization*, eds. G. Di Pillo, and F. SchoenSpringer, pp. 165–214, Berlin: Springer. [633]

Geer, S. A., and van de Geer, S. (2000), *Empirical Processes in M-estimation* (Vol. 6), Cambridge: Cambridge University Press. [632]

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A Kernel Two-Sample Test," *The Journal of Machine Learning Research*, 13, 723–773. [629,630]

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018), "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *International Conference on Machine Learning*, pp. 1861–1870, PMLR. [625,626,627]

Hiriart-Urruty, J.-B., and Lemaréchal, C. (2012), *Fundamentals of Convex Analysis*, Berlin: Springer. [628]

Jiang, N., and Li, L. (2016), "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning," in *International Conference on Machine Learning*, pp. 652–661, PMLR. [629]

John, G. H. (1994), "When the Best Move isn't Optimal: Q-learning with Exploration," in *AAAI-94*. [627]

Korbel, J., Hanel, R., and Thurner, S. (2019), "Information Geometric Duality of ϕ-Deformed Exponential Families," *Entropy*, 21, 112–125. [628]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer. [631]

Lee, K., Choi, S., and Oh, S. (2018), "Sparse Markov Decision Processes with Causal Sparse Tsallis Entropy Regularization for Reinforcement Learning," *IEEE Robotics and Automation Letters*, 3, 1466–1473. [625,626]

Liao, P., Klasnja, P., and Murphy, S. (2020), "Off-policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health," *Journal of the American Statistical Association*, 116, 382–391. [625,631,634]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020), "Estimating Dynamic Treatment Regimes in Mobile Health using v-learning," *Journal of the American Statistical Association*, 115, 692–706. [625,626,627,629,634,636]

Maahs, D. M., Mayer-Davis, E., Bishop, F. K., Wang, L., Mangan, M., and McMurray, R. G. (2012), "Outpatient Assessment of Determinants of Glucose Excursions in Adolescents with Type 1 Diabetes: Proof of Concept," *Diabetes Technology & Therapeutics*, 14, 658–664. [635]

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010), "Toward Off-policy Learning Control with Function Approximation," in *27th ICML*, pp. 719–726. [630]

Marling, C., and Bunescu, R. (2020), "The ohioT1DM Dataset for Blood Glucose Level Prediction: Update 2020," *KHD@ IJCAI*. [625,627,636]

Merlevède, F., Peligrad, M., and Rio, E. (2011), "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields*, 151, 435–474. [632]

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016), "Asynchronous Methods for Deep Reinforcement Learning," in *International Conference on Machine Learning*, pp. 1928–1937. [633]

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015), "Human-Level Control through Deep Reinforcement Learning," *Nature*, 518, 529–533. [637]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [625]

Murphy, S. A., Deng, Y., Laber, E. B., Maei, H. R., Sutton, R. S., and Witkiewitz, K. (2016), "A Batch, Off-policy, Actor-Critic Algorithm for Optimizing the Average Reward." arXiv preprint arXiv:1607.05047. [637]

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017), "Bridging the Gap between Value and Policy based Reinforcement Learning," in *Advances in Neural Information Processing Systems*, pp. 2775–2785. [625,629]

——— (2018), "Trust-PCL: An Off-Policy Trust Region Method for Continuous Control," in *International Conference on Learning Representations*. [625,629]

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2018), "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support," *Annals of Behavioral Medicine*, 52, 446–462. [625]

Puterman, M. L. (2014), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Hoboken, NJ: Wiley. [626]

Rawlik, K., Toussaint, M., and Vijayakumar, S. (2013), "On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference," in *Twenty-third IJCAI*. [629]

Rehg, J. M., Murphy, S. A., and Kumar, S. (2017), *Mobile Health*, Cham: Springer. [625,627]

Rodbard, D. (2009), "Interpretation of Continuous Glucose Monitoring Data: Glycemic Variability and Quality of Glycemic Control," *Diabetes Technology & Therapeutics*, 11, S–55. [636]

Schulman, J., Chen, X., and Abbeel, P. (2017), "Equivalence between Policy Gradients and Soft q-learning." arXiv preprint arXiv:1704.06440. [625,626,627]

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015), "Trust Region Policy Optimization," in *International Conference on Machine Learning*, pp. 1889–1897. [627]

Shi, C., Fan, A., Song, R., and Lu, W. (2018), "High-Dimensional A-learning for Optimal Dynamic Treatment Regimes," *Annals of Statistics*, 46, 925–967. [625]

Shi, C., Zhang, S., Lu, W., and Song, R. (2020), "Statistical Inference of the Value Function for Reinforcement Learning in Infinite Horizon Settings." arXiv preprint arXiv:2001.04515. [625]

Sim, I. (2019), "Mobile Devices and Health," *New England Journal of Medicine*, 381, 956–968. [625]

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994), "Learning without State-Estimation in Partially Observable Markovian Decision Processes," in *Machine Learning 1994*, Elsevier, pp. 284–292. [627]

Steinwart, I., and Christmann, A. (2008), *Support Vector Machines*, New York: Springer. [632]

Stewart, J. (1976), "Positive Definite Functions and Generalizations, an Historical Survey," *The Rocky Mountain Journal of Mathematics*, 6, 409–434. [630]

Sutton, R. S., and Barto, A. G. (1998), *Introduction to Reinforcement Learning* (Vol. 135), Cambridge, MA: MIT Press. [637]

——— (2018), *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press. [625,626]

Sutton, R. S., Mahmood, A. R., and White, M. (2016), "An Emphatic Approach to the Problem of Off-Policy Temporal-Difference Learning," *The Journal of Machine Learning Research*, 17, 2603–2631. [626]

Uehara, M., Huang, J., and Jiang, N. (2020), "Minimax Weight and q-function Learning for Off-Policy Evaluation," in *International Conference on Machine Learning*, pp. 9659–9668, PMLR. [625]

Watkins, C. J., and Dayan, P. (1992), "Q-learning," *Machine Learning*, 8, 279–292. [626,630]

Xu, Z., Laber, E., and Staicu, A.-M. (2020), "Latent-State Models for Precision Medicine." arXiv preprint arXiv:2005.13001. [625]

Yang, Q., and Van Stee, S. K. (2019), "The Comparative Effectiveness of Mobile Phone Interventions in Improving Health Outcomes: Meta-Analytic Review," *JMIR mHealth and uHealth*, 7, e11244. [625]

Yu, H. (2016), "Weak Convergence Properties of Constrained Emphatic Temporal-Difference Learning with Constant and Slowly Diminishing Stepsize," *The Journal of Machine Learning Research*, 17, 7745–7802. [626]

Yu, H., Mahmood, A. R., and Sutton, R. S. (2018), "On Generalized Bellman Equations and Temporal-Difference Learning," *The Journal of Machine Learning Research*, 19, 1864–1912. [626]

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [625]

Zhu, L., Lu, W., and Song, R. (2020), "Causal Effect Estimation and Optimal Dose Suggestions in Mobile Health," in *International Conference on Machine Learning*, pp. 11588–11598, PMLR. [636]