The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1448–1472
https://doi.org/10.1214/24-AOAS1987

NONPARAMETRIC ADDITIVE VALUE FUNCTIONS: INTERPRETABLE REINFORCEMENT LEARNING WITH AN APPLICATION TO SURGICAL RECOVERY

BY PATRICK EMEDOM-NNAMDI[1,a], TIMOTHY R. SMITH[2,d], JUKKA-PEKKA ONNELA[1,b] AND JUNWEI LU[1,c]

[1]*Department of Biostatistics, Harvard University,* [a]*patrickemedom@hsph.harvard.edu,* [b]*onnela@hsph.harvard.edu,* [c]*junweilu@hsph.harvard.edu*

[2]*Computational Neurosciences Outcomes Center, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School,* [d]*trsmith@bwh.harvard.edu*

We propose a nonparametric additive model for estimating interpretable value functions in reinforcement learning, with an application in optimizing postoperative recovery through personalized, adaptive recommendations. While reinforcement learning has achieved significant success in various domains, recent methods often rely on black-box approaches, such as neural networks, which hinder the examination of individual feature contributions to a decision-making policy. Our novel method offers a flexible technique for estimating action-value functions without explicit parametric assumptions, overcoming the limitations of the linearity assumption of classical algorithms. By incorporating local kernel regression and basis expansion, we obtain a sparse, additive representation of the action-value function, enabling local approximation and retrieval of nonlinear, independent contributions of select state features and the interactions between joint feature pairs. We validate our approach through a simulation study and apply it to spine disease recovery, uncovering recommendations aligned with clinical knowledge. This method bridges the gap between flexible machine learning techniques and the interpretability required in healthcare applications, paving the way for more personalized interventions.

1. Introduction. Optimizing postoperative recovery is crucial for improving surgical outcomes, including functional restoration and enhanced quality of life. This process is inherently complex and multifaceted, influenced by the nature of the diagnosis, the type of surgical procedure, and various patient-specific factors. Demographic characteristics, such as age, gender, and comorbidities as well as time since surgery and patient behaviors (e.g., mobility, physical activity, and sleep), significantly impact recovery trajectories (Cote et al. (2019), Panda et al. (2020a)). While common recommendations often focus on early mobilization activities, such as getting out of bed or taking light walks, developing personalized care plans presents a greater challenge. To provide tailored suggestions that account for diverse patient factors and recommend specific actions (e.g., daily step targets), clinicians require both continuous access to patient health and behavioral data and flexible algorithms capable of processing this information and providing interpretable insights.

The proliferation of smartphones and wearable devices has revolutionized our ability to collect real-time, continuous data on human behavior and health, providing valuable insights into patient recovery (Torous, Staples and Onnela (2015), Onnela (2021)). This mobile health approach, when combined with statistical machine learning techniques, offers clinicians a powerful paradigm for uncovering nuanced patterns in recovery trajectories and developing more refined, evidence-based standards of care. By leveraging high-quality, temporally-dense

Received August 2023; revised October 2024.
Key words and phrases. Reinforcement learning, interpretable value functions, nonparametric, digital health.

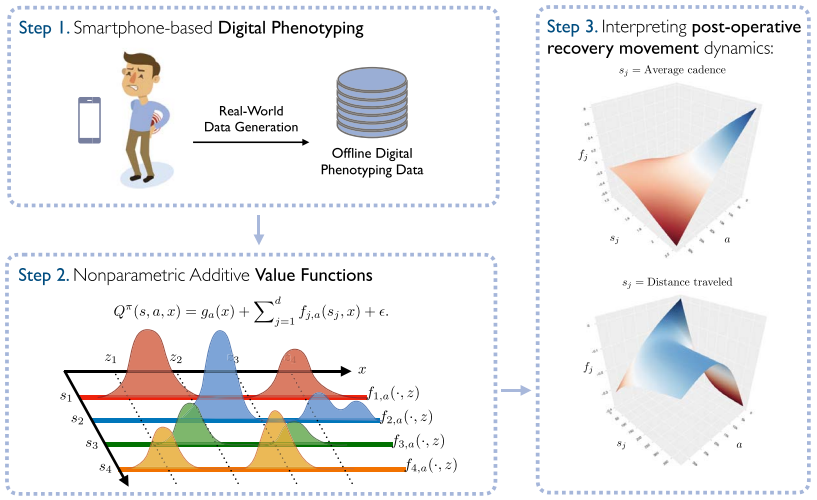


FIG. 1. *An overview of using nonparametric additive models for learning interpretable value functions. Within our setting, real-world data from subjects with select physiological disorders are collected using a smartphone-based mobile health platform. Modalities collected from subject smartphones range from raw sensor data (e.g., GPS, accelerometer, gyroscope, or magnetometer) to usage logs (e.g., anonymized communication and screen time). Relevant features* $\mathbf{s} = (s_1, \ldots, s_d)^T$ *are summarized from these modalities and are used to frame a corresponding decision-making problem (or MDP, see Section 2.1) of interest. Under the select MDP, we model the value function* $Q^\pi(\mathbf{s}, a, x)$ *as a sum of nonparametric component functions* $g_a(x)$ *and* $f_{j,a}(\mathbf{s}_j, x)$ $\forall j$. *Here we visualize the change in the shape and sparsity patterns of the component function* $f_{j,a}(\cdot, z)$ *as the candidate variable* x *changes, for example,* $x \in \{z_1, z_2, z_3, z_4\}$. *Each additive component function can be estimated and inspected using the kernel-weighted least square fixed point approximation detailed in Section 3.*

data, healthcare professionals can now enhance their understanding of postoperative recovery and pave the way for more personalized, adaptive, and effective treatment strategies (Panda et al. (2020b)).

In this paper, we introduce a novel reinforcement learning approach for estimating recovery strategies and recommendations using mobile health data (see Figure 1). Reinforcement learning is a subfield of machine learning that focuses on learning sequences of decisions that optimize long-term outcomes from experiential data (Sutton and Barto (2018)). In healthcare, reinforcement learning algorithms have been used to discover decision-making strategies for chronic disease treatments (Bothe et al. (2013), Peyser et al. (2014)), anesthesia regulation and automation (Sinzinger and Moore (2011)), chemotherapy scheduling and dosage management (Padmanabhan, Meskin and Haddad (2015), Ahn and Park (2011)), and sepsis management (Raghu et al. (2017), Peng et al. (2018)).

Implementing reinforcement learning in healthcare applications requires careful consideration of the policy estimation process and thorough examination of the learned policy's behavior prior to real-world deployment (Gottesman et al. (2019)). Typically, decision-making policies are represented as a function $\pi(\mathbf{s})$ of state features $\mathbf{s} = (s_1, \ldots, s_d)^T \in \mathbb{R}^d$, estimated using policy gradient or value-based reinforcement learning algorithms (Sutton and Barto (2018)). In value-based approaches, policies are determined by selecting the action a that maximizes the corresponding action-value function $Q^\pi(\mathbf{s}, a)$. Current methods often employ neural network function approximators, resulting in black-box algorithms that are difficult to interpret and provide minimal insight into which feature (or set of features) influenced the decision-making process (Gottesman et al. (2019)). While linear models offer interpretability, they fail to capture the complex, nonlinear relationships and interactions between clinical variables that are crucial for understanding and optimizing surgical recovery, necessitating a more flexible and interpretable approach.

To address these limitations, we propose a novel class of approximate value functions that offers a flexible, nonparametric representation of the action-value function, easily interpretable for both researchers and clinicians. Our approach allows for the inspection of a candidate variable x (e.g., time-varying/-invariant confounders or continuous-valued actions) and models action-value functions as a sum of nonparametric, additive component functions,

$$Q^{\pi}(\mathbf{s}, a, x) = g_a(x) + \sum_{j=1}^{d} f_{j,a}(\mathbf{s}_j, x) + \epsilon. \tag{1}$$

This framework enables the examination of both the marginal effect of x and its joint effect with state features $\mathbf{s}_j$ under a discretized action space. To estimate these component functions, we build upon the classical least square policy iteration (LSPI) algorithm (Lagoudakis and Parr (2004)), incorporating a kernel-hybrid approach that relaxes traditional linearity assumptions. By leveraging advances in high-dimensional, nonparametric additive regression models (Fan and Jiang (2005), Ravikumar et al. (2009), Lafferty and Wasserman (2008)), we introduce a kernel-sieve hybrid regression estimator (Lu, Kolar and Liu (2020)) to obtain a sparse additive representation of the action-value function.

To validate our methodology, we present a simulation study examining its ability to estimate nonlinear additive functions and compare its performance against modern neural network-based approaches. Furthermore, we apply our model to an ongoing mobile health study, demonstrating its capacity to learn and interpret a decision-making policy aimed at improving pain management and functional recovery in patients recovering from spine surgery through mobility management.

1.1. *Related research.* Our work contributes to the growing literature on function approximation methods for value-based reinforcement learning. Current state-of-the-art algorithms approximate action-value functions using expressive modeling architectures, such as neural networks. By combining fitted Q-iteration procedures with modern tools, such as replay buffers and target networks, these algorithms resolve the pitfalls of classical methods and solve complex, high-dimensional decision-making tasks (Riedmiller (2005), Antos, Szepesvári and Munos (2007), Van Hasselt (2010), Mnih et al. (2013, 2015)). However, the powerful flexibility of these approaches comes at the cost of interpretability inherent in algorithms such as least squares policy iteration (LSPI).

LSPI is a model free, off-policy approximate policy iteration algorithm that models the action-value function using a parametric linear approximation and finds an approximate function that best satisfies the Bellman equation (i.e., the fixed-point solution) (Lagoudakis and Parr (2004)). While LSPI provides an unbiased estimate of the action-value function, it faces significant challenges when the model is misspecified and when the dimensionality of the feature space is high (Lagoudakis and Parr (2004), Farahmand et al. (2016)). Several modifications to LSPI have been proposed in the reinforcement learning literature. In settings where the feature space is large, several approaches exist for finding sparse solutions in a linear model (Hoffman et al. (2012), Kolter and Ng (2009), Tziortziotis and Dimitrakakis (2017), Geist and Scherrer (2011)). Alternatively, Xu, Hu and Lu (2007) propose a kernel-based LSPI algorithm that operates in an infinite-dimensional Hilbert space and allows for nonlinear feature extraction by selecting appropriate kernel functions. Additionally, Howard and Nakamura (2013) propose a locally-weighted LSPI model that leverages locally-weighted to construct a nonlinear, global control policy.

To date, no approximation methods in RL have been introduced that directly allows for the nonparametric estimation concerning the additive contribution of select features and joint feature pairs. In the supervised learning literature, several approaches exist for estimating

nonparametric component functions in high-dimensional feature spaces. Several of these approaches include generalized additive models (GAM) and sparse additive models (SpAM), to which our approach draws parallels (Ravikumar et al. (2009), Hastie (2017)). To bridge these areas of research, we reformulate the policy evaluation step of the classical LSPI algorithm and propose incorporating the kernel-sieve hybrid regression estimator introduced in Lu, Kolar and Liu (2020). This approach provides a powerful function approximation technique for locally estimating action-value functions using a loss function combining both basis expansion and kernel method with a hybrid ℓ_1/ℓ_2-group Lasso penalty.

1.2. *Organization of the paper.* The remainder of this paper is organized as follows. In Section 2.2 we introduce our generalized, nonparametric model for representing action-value functions. Section 3 presents our estimation strategy for locally approximating the action-value function by combining basis expansion and kernel methods. Section 4 examines the results of a simulation study, highlighting the performance of our method in estimating sparse additive components of the action-value function. Section 5 introduces a real-world cohort of patients recovering from a neurological intervention for spine disease as a motivating case study. In Section 6 we apply our method to this mobile health study and interpret the estimated recovery strategy. Finally, Section 7 discusses the limitations of our method and proposes future extensions to address them.

2. Nonparametric additive value functions.

2.1. *Preliminaries and notation.* We consider a discrete-time, infinite horizon Markov Decision Process (MDP), defined by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma\}$, where $\mathcal{S}$ is a continuous d-dimensional state space, $\mathcal{A}$ is a set of discrete (i.e., $\mathcal{A} = \{1, \dots, k\}$) or continuous (i.e., $\mathcal{A} = \mathbb{R}$) actions, $\mathcal{P}(\mathbf{s}'|\mathbf{s}, a)$ is a next-state transition probability kernel that specifies the probability of transitioning from state $\mathbf{s} \in \mathcal{S}$ to the next state $\mathbf{s}' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, and $\gamma \in [0, 1]$ is a discount factor for weighting long-term rewards. Within this stochastic environment, the action selection strategy is determined by a deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$.

To assess the quality of a policy, the expected discounted sum of rewards, when starting at state $\mathbf{s}$ and following policy π, can be computed using the value function $V^\pi : \mathcal{S} \to \mathbb{R}$. The value function starting from state $\mathbf{s}$ is defined as

$$V^\pi(\mathbf{s}) = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i r_i \,\middle|\, \mathbf{s}_{\text{init}} = \mathbf{s}\right], \tag{2}$$

where r_i represents the reward at time i and $\mathbf{s}_{\text{init}}$ is the initial state.

In control problems where we are interested in improving our action selection strategy, it is useful to consider the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Given a policy π, Q^π represents the expected discounted sum of rewards after taking action a and state $\mathbf{s}$ and following policy π thereafter, that is,

$$Q^\pi(\mathbf{s}, a) = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i r_i \,\middle|\, \mathbf{s}_{\text{init}} = \mathbf{s}, a_{\text{init}} = a\right]. \tag{3}$$

Due to the Markovian property of our MDP, the action-value function (as well as the value function) is a fixed point of the Bellman operator $Q^\pi = \mathcal{T}_\pi Q^\pi$, where the operator $\mathcal{T}_\pi$ is defined as

$$(\mathcal{T}_\pi Q)(\mathbf{s}, a) = R(\mathbf{s}, a) + \gamma \int_{\mathcal{S}} Q\big(\mathbf{s}', \pi(\mathbf{s}')\big)\, d\mathcal{P}(\mathbf{s}'|\mathbf{s}, a) \tag{4}$$

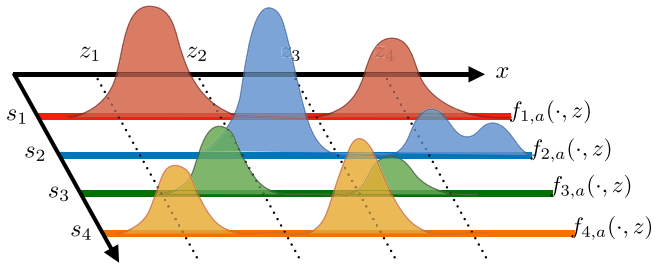


FIG. 2. *Representation of a nonparametric additive value function* $Q^\pi(\mathbf{s}, a, x)$ *with respect to the candidate variable* x *as detailed in* (5).

or, equivalently, in vector form as $\mathcal{T}_\pi Q = \mathcal{R} + \gamma \mathcal{P}^\pi Q$, where $\mathcal{R} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is a reward vector and $\mathcal{P}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is the induced transition matrix when following policy π after a next state transition according to $\mathcal{P}(\mathbf{s}'|\mathbf{s}, a)$.

For a given MDP, the optimal action-value function is defined as $Q^*(\mathbf{s}, a) = \sup_\pi Q^\pi(\mathbf{s}, a)$ for all states and actions $(\mathbf{s}, a) \in \mathcal{S} \times \mathcal{A}$. For a given action-value function Q, we define a greedy policy π as $\pi(\mathbf{s}) = \arg\max_{a \in \mathcal{A}} Q(\mathbf{s}, a)$ for all $\mathbf{s} \in \mathcal{S}$. The greedy policy with respect to the optimal action-value function Q^* is then an optimal policy, denoted as π^*. Hence, obtaining Q^* allows us to arrive at an optimal action selection strategy.

2.2. *Generalized framework.* For an arbitrary policy π, we introduce a generalized framework for modeling the action-value function Q^π as a sum of nonparametric additive component functions, as shown in Figure 2. Our approach handles both discrete (i.e., $\mathcal{A} = \{1, \ldots, k\}$) and continuous (i.e., $\mathcal{A} = \mathbb{R}$) action spaces, while allowing for the incorporation of potentially time-varying or time-invariant variables.

First, we present our generalized nonparametric framework for modeling Q^π as

$$Q^\pi(\mathbf{s}, a, x) = g_a(x) + \sum_{j=1}^{d} f_{j,a}(\mathbf{s}_j, x) + \epsilon. \tag{5}$$

Under this model we expand the input space of Q^π to include the candidate variable $x \in \mathbb{R}$, and discretize the action space such that $a \in \{1, \ldots, k\}$ if $\mathcal{A}$ is not already discrete. Accordingly, $g_a(\cdot)$ represents the additive marginal effect of x under action a, and $f_j(\cdot, \cdot)$ represents the additive joint effect of interactions between x and state features $\mathbf{s}_j$ under action a. Without making specific assumptions on the functional form of $g_a(\cdot)$ and $f_{j,a}(\cdot, \cdot)$, our model allows us to carefully examine additive nonlinear relationships that exist among relevant state features, actions, and the variable x.

Second, our choice in x allows us to explore several unique representations of the additive components in (5). For example, x can represent time-varying or time-invariant confounders (e.g., age, gender, or the number of days since a surgical event) as well as continuous-valued actions $a \in \mathbb{R}$,

$$x = \begin{cases} \mathbf{s}_0, & \text{that is, a candidate state feature or confounder,} \\ a, & \text{that is, a continuous action.} \end{cases} \tag{6}$$

EXAMPLE 1. When $x = \mathbf{s}_0$, the additive functions in (5), respectively, equate to $g_a(x) = g_a(\mathbf{s}_0)$ and $f_{j,a}(\mathbf{s}_j, x) = f_{j,a}(\mathbf{s}_j, \mathbf{s}_0)$. Furthermore, we can augment the state space S, using x to form $S_+ = \{x, s_1, \ldots, s_d\}$, and succinctly represent $Q^\pi(\mathbf{s}, a, \mathbf{s}_0)$ as $Q^\pi(\mathbf{s}_+, a)$, where $\mathbf{s}_+ \in S_+$ and

$$Q^\pi(\mathbf{s}_+, a) = g_a(\mathbf{s}_0) + \sum_{j=1}^{d} f_{j,a}(\mathbf{s}_j, \mathbf{s}_0) + \epsilon. \tag{7}$$

Thus, under the discrete action a, $g_a(\mathbf{s}_0)$ models the nonlinear marginal effect of the confounder or state feature $\mathbf{s}_0$, whereas $f_{j,a}(\mathbf{s}_j, \mathbf{s}_0)$ models the nonlinear interaction between $\mathbf{s}_0$ and state features $\mathbf{s}_j$.

EXAMPLE 2. Similarly, when $x = a$, the additive functions in (5), respectively, equate to $g_a(a) = g(a)$ and $f_{j,a}(\mathbf{s}_j, a) = f_j(\mathbf{s}_j, a)$. Under this choice of x, we avoid explicit discretization of the action space $\mathcal{A}$ and directly treat a as a continuous action. Thus, for a given state action pair, $Q^\pi(\mathbf{s}, a, a)$ reduces to $Q^\pi(\mathbf{s}, a)$, where

$$Q^\pi(\mathbf{s}, a) = g(a) + \sum_{j=1}^{d} f_j(\mathbf{s}_j, a) + \epsilon, \tag{8}$$

and the additive marginal effect of selecting a continuous action a is modeled as $g(a)$, and $f_j(\mathbf{s}_j, a)$ represents the additive effect of selecting action a under state feature value $\mathbf{s}_j$.

3. Kernel sieve hybrid–least squares policy iteration. We introduce a general approach for estimating $Q^\pi(\mathbf{s}, a)$ for both discrete and continuous action spaces. This estimation strategy provides an intuitive way to: (1) locally approximate the action-value function as an additive model with independent state features spanned by a B-spline basis expansion, (2) retrieve an estimate of the nonlinear additive components, and (3) obtain a sparse representation of the action-value function by selecting relevant regions of the domain of the component functions.

3.1. *Basis expansion.* First, we model the action-value function using a centered B-spline basis expansion of the additive component functions. Let $\{\psi_1, \ldots, \psi_m\}$ be a set of normalized B-spline basis functions. For each component function, we project $f_{j,a}$ onto the space spanned by the basis, $\mathcal{B}_m = \text{Span}(\psi_1, \ldots, \psi_m)$. Thus, $f_{j,a}(\mathbf{s}_j, x)$ can be expressed as $\sum_{\ell=1}^{m} \varphi_{j\ell}(\mathbf{s}_j)\boldsymbol{\beta}_{j\ell;a}(x)$, where $\varphi_{j\ell}$ are locally centered B-spline basis functions defined as $\varphi_{j\ell}(\mathbf{s}) = \psi_\ell(\mathbf{s}) - E[\psi_\ell(\mathbf{s}_j)]$ for the jth component function and the ℓth basis component.

As we will discuss in Section 3.2, our estimation strategy relies on performing a locally-weighted least-squares minimization of an objective criterion with respect to a fixed value of the variable x. As such, we locally express our model in (5) by: (i) setting $x = z$, where $z \in \mathcal{X}$ is some arbitrary fixed value, and by (ii) using the aforementioned centered B-spline basis expansion,

$$Q^\pi(\mathbf{s}, a, x = z) \approx \alpha_{a,z} + \sum_{j=1}^{d} \sum_{\ell=1}^{m} \varphi_{j\ell}(\mathbf{s}_j)\beta_{j\ell;a,z}. \tag{9}$$

Under this local model, $\alpha_{a,z} \in \mathbb{R}$ represents that marginal effect $g_a(x)$ when z is a fixed value of x. Accordingly, $\beta_{j\ell;a,z} \in \mathbb{R}$ is the coordinate corresponding to the ℓth B-spline basis of the jth state-feature under z.

In Examples 1 and 2, we observe two choices for representing x that highlight the generalizability of our model structure. Under theses examples, the local additive components in (9) can also be reexpressed as follows:

$$\alpha_{a,z} = \begin{cases} \alpha_{a,z} & \text{when } x = \mathbf{s}_0, \\ \alpha_z & \text{when } x = a \end{cases} \quad \text{and} \quad \beta_{j\ell;a,z} = \begin{cases} \beta_{j\ell;a,z} & \text{when } x = \mathbf{s}_0, \\ \beta_{j\ell;z} & \text{when } x = a. \end{cases} \tag{10}$$

Since the dynamics of the MDP are unknown, our estimation strategy relies on a batch dataset $\mathcal{D} = \{(\mathbf{s}^{[i]}, a^{[i]}, r^{[i]}, \mathbf{s}^{[i]\prime}, x^{[i]})\}_{i=1}^{N}$ of sampled transitions from the MDP of interest, where $\mathbf{s}^{[i]\prime} \sim P(\cdot|\mathbf{s}^{[i]}, a^{[i]})$ and $x^{[i]}$ is the associated value of the candidate variable x. When

we consider all observations in the dataset $\mathcal{D}$, we can equivalently reexpress (9) in vector form as

$$Q^{\pi}_{\boldsymbol{\beta}_+} = \tilde{\boldsymbol{\Phi}}\boldsymbol{\beta}_+, \tag{11}$$

where $\boldsymbol{\beta}_+ = (\boldsymbol{\beta}^T_{1+}, \ldots, \boldsymbol{\beta}^T_{|\mathcal{A}|+})^T \in \mathbb{R}^{(1+dm)|\mathcal{A}|}$, $\boldsymbol{\beta}_{a+} = (\alpha_{a,z}, \boldsymbol{\beta}^T_{1;a,z}, \ldots, \boldsymbol{\beta}^T_{d;a,z})^T \in \mathbb{R}^{1+dm}$, and

$$\tilde{\boldsymbol{\Phi}} = \begin{pmatrix} \phi(\mathbf{s}^{[1]}, a^{[1]})^T \\ \vdots \\ \phi(\mathbf{s}^{[N]}, a^{[N]})^T \end{pmatrix} = \begin{pmatrix} \varphi_+(\mathbf{s}^{[1]})^T \mathbb{1}(a^{[1]} = 1) & \cdots & \varphi_+(\mathbf{s}^{[1]})^T \mathbb{1}(a^{[1]} = k) \\ \vdots & & \\ \varphi_+(\mathbf{s}^{[N]})^T \mathbb{1}(a^{[N]} = 1) & \cdots & \varphi_+(\mathbf{s}^{[N]})^T \mathbb{1}(a^{[N]} = k) \end{pmatrix} \tag{12}$$

such that $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{N\times(1+dm)|\mathcal{A}|}$, $\varphi_+(\mathbf{s}) = (1 \;\; \varphi_2(\mathbf{s}_1) \;\; \cdots \;\; \varphi_d(\mathbf{s}_d))^T \in \mathbb{R}^{1+dm}$, and $\varphi_j(\mathbf{s}_j) \in \mathbb{R}^m$ is a B-spline basis component vector. Note that $E[\psi_\ell(\mathbf{s}_j)]$ is estimated as $\bar{\psi}_{j\ell} = N^{-1}\sum_{i=1}^N \psi_\ell(\mathbf{s}^{[i]}_j)$ by using a sample of N data points from the dataset $\mathcal{D}$.

3.2. *Kernel-weighted least squares fixed point approximation.* We estimate our model parameters $\boldsymbol{\beta}_+$ by minimizing a kernel-weighted version of the classical projected Bellman error (PBE).

In LPSI we observe that a simple procedure for estimating a linear action-value function is to force the approximate function to be a fixed point under the projected Bellman operator (i.e., $\Pi\mathcal{T}_\pi Q_{\boldsymbol{\beta}_+} \approx Q_{\boldsymbol{\beta}_+}$). For this condition to hold, the fixed point of the Bellman operator $\mathcal{T}_\pi$ must lie in the space of approximate value functions spanned by the basis functions over all possible state action pairs, $\mathcal{C}(\boldsymbol{\Phi})$. By construction it is known that $Q_{\boldsymbol{\beta}_+} = \boldsymbol{\Phi}\boldsymbol{\beta}_+ \in \mathcal{C}(\boldsymbol{\Phi})$. However, since there is no guarantee that $\mathcal{T}_\pi Q_{\boldsymbol{\beta}_+}$ (i.e., the result of the Bellman operator) is in $\mathcal{C}(\boldsymbol{\Phi})$, it first must be projected onto $\mathcal{C}(\boldsymbol{\Phi})$, using the projection operator Π, such that $\Pi\mathcal{T}_\pi Q^{\pi}_{\boldsymbol{\beta}_+} = \boldsymbol{\Phi}\boldsymbol{u}^*$ where $\boldsymbol{u}^*$ is the solution to the following least-squares problem:

$$\boldsymbol{u}^* = \operatorname*{argmin}_{Q\in\mathcal{C}(\boldsymbol{\Phi})} \| Q_{\boldsymbol{\beta}_+} - \mathcal{T}_\pi Q_{\boldsymbol{\beta}_+} \|^2_2 = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^k} \| \boldsymbol{\Phi}u - \mathcal{T}_\pi \boldsymbol{\Phi}\boldsymbol{\beta}_+ \|^2_2. \tag{13}$$

Empirically, $\boldsymbol{u}^*$ can be estimated using a sample-based feature design matrix $\tilde{\boldsymbol{\Phi}}$ constructed from a dataset of N transitions $\mathcal{D}$,

$$\boldsymbol{u}^* = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^k} \| \tilde{\boldsymbol{\Phi}}\boldsymbol{u} - \hat{\mathcal{T}}_\pi \tilde{\boldsymbol{\Phi}}\boldsymbol{\beta}_+ \|^2_2 \tag{14}$$

$$= \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^k} \sum_{i=1}^N \big(\phi(\mathbf{s}^{[i]}, a^{[i]})^T \boldsymbol{u} - \big[r^{[i]} + \gamma\phi(\mathbf{s}^{[i]\prime}, \pi(\mathbf{s}^{[i]\prime}))^T \boldsymbol{\beta}_+\big]\big)^2, \tag{15}$$

where $\hat{\mathcal{T}}_\pi$ is the empirical Bellman operator $(\hat{\mathcal{T}}_\pi Q_\beta)(\mathbf{s}, a) = r(\mathbf{s}, a) + \gamma Q_\beta(\mathbf{s}', \pi(\mathbf{s}))$ defined using a single transition $\{\mathbf{s}, a, r, \mathbf{s}'\}$ from $\mathcal{D}$.

Rather than performing the projection step according to an ℓ_2-norm, we propose using a *kernel-weighted* norm with weights that are centered at a fixed value z that lies within the domain of the candidate variable x. Let $K : \mathcal{X} \to \mathbb{R}$ be a symmetric kernel function with bounded support. We denote $K_h(\cdot) = h^{-1}K(\cdot/h)$, where $h > 0$ is the bandwidth. The solution $\boldsymbol{u}^*_z$ to the kernel-weighted projection step is estimated as follows:

$$\boldsymbol{u}^*_z = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^k} \sum_{i=1}^N K_h(x^{[i]} - z)\big(\phi(\mathbf{s}^{[i]}, a^{[i]})^T \boldsymbol{u} - \big[r^{[i]} + \gamma\phi(\mathbf{s}^{[i]\prime}, \pi(\mathbf{s}^{[i]\prime}))^T \boldsymbol{\beta}_+\big]\big)^2 \tag{16}$$

$$= \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^k} (\tilde{\boldsymbol{\Phi}}\boldsymbol{u} - \hat{\mathcal{T}}_\pi \tilde{\boldsymbol{\Phi}}\boldsymbol{\beta}_+)^T \boldsymbol{W}_z (\tilde{\boldsymbol{\Phi}}\boldsymbol{u} - \hat{\mathcal{T}}_\pi \tilde{\boldsymbol{\Phi}}\boldsymbol{\beta}_+) = (\tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \tilde{\boldsymbol{\Phi}})^{-1} \tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \hat{\mathcal{T}}_\pi \tilde{\boldsymbol{\Phi}}\boldsymbol{\beta}_+, \tag{17}$$

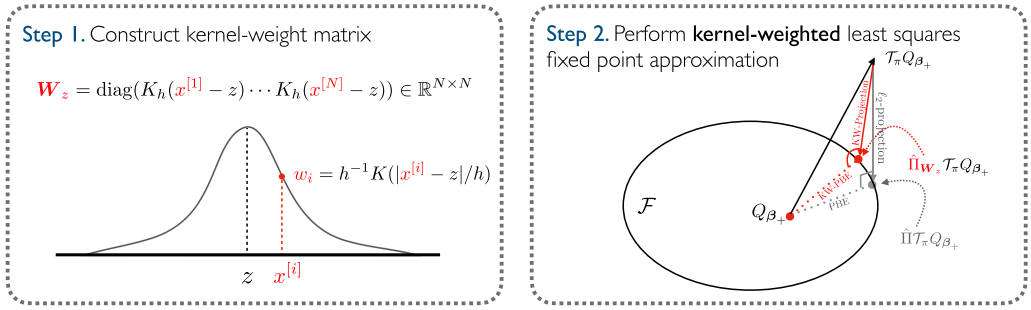


FIG. 3. *A step-by-step illustration of kernel-weighted least squares fixed point approximation. First, using observations gathered in the batch dataset $\mathcal{D}$, we construct a diagonal kernel-weight matrix $\boldsymbol{W}_z$, where each diagonal weight is a function of the distance between the observed candidate variable $x^{[i]}$ and the fixed value z. Second, let $\mathcal{F}$ be $\mathcal{C}(\boldsymbol{\Phi})$, that is, the space of approximate value functions. Since applying the Bellman operator $\mathcal{T}_\pi$ to an arbitrary value function $Q_{\boldsymbol{\beta}_+}$ can push the resulting quantity $\mathcal{T}_\pi Q^\pi_{\boldsymbol{\beta}_+}$ out of the space $\mathcal{F}$, we perform a projection step using the constructed kernel-weighted matrix $\boldsymbol{W}_z$ (detailed in Section 3.2). This approach differs from classical least squares fixed point approximation (shown in grey), where an ℓ_2-projection operator $\hat{\Pi}$ is used. Lastly, we find $\hat{\boldsymbol{\beta}}_+$ that minimizes the ℓ_2-norm between $Q_{\boldsymbol{\beta}_+}$ and $\hat{\Pi}_{\boldsymbol{W}_z}\mathcal{T}_\pi Q^\pi_{\boldsymbol{\beta}_+}$, that is, the kernel-weighted projected bellman error.*

where $\boldsymbol{W}_z = \text{diag}(K_h(x^{[1]} - z) \cdots K_h(x^{[N]} - z)) \in \mathbb{R}^{N \times N}$ is a diagonal kernel-weight matrix. Under this weighted norm, transitions with a candidate variable $x^{[i]}$ that are local to z contribute more to the overall fit of the least squares minimization. Accordingly, the empirical kernel-weighted projection operator is $\hat{\Pi}_{\boldsymbol{W}_z} = \tilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \tilde{\boldsymbol{\Phi}})^{-1} \tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z$.

Using the projection operator $\hat{\Pi}_{\boldsymbol{W}_z}$, we can now directly find $\boldsymbol{\beta_+}$ that minimizes the kernel-weighted empirical PBE represented as

$$\mathcal{E}_\mathcal{D} = \| Q^\pi_{\boldsymbol{\beta}_+} - \hat{\Pi}_{\boldsymbol{W}_z} \mathcal{T}_\pi Q^\pi_{\boldsymbol{\beta}_+} \|_2^2 = \| \tilde{\boldsymbol{\Phi}} \boldsymbol{\beta}_+ - \tilde{\boldsymbol{\Phi}} \underbrace{(\tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \tilde{\boldsymbol{\Phi}})^{-1} \tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \mathcal{T}_\pi \tilde{\boldsymbol{\Phi}} \boldsymbol{\beta}_+}_{\tilde{g}(\boldsymbol{\beta}_+)} \|_2^2. \tag{18}$$

Since $\tilde{\boldsymbol{\Phi}} \tilde{g}(\beta) \in \mathcal{C}(\tilde{\boldsymbol{\Phi}})$, minimizing this objective function is equivalent to solving for $\boldsymbol{\beta}_+$ in $\boldsymbol{\Phi}\boldsymbol{\beta}_+ = \boldsymbol{\Phi}\tilde{g}(\boldsymbol{\beta}_+)$, which can be simplified as

$$\underbrace{\tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z (\tilde{\boldsymbol{\Phi}} - \gamma \tilde{\boldsymbol{\Phi}}')}_{\boldsymbol{A}_z} \boldsymbol{\beta}_+ = \underbrace{\tilde{\boldsymbol{\Phi}}^T \boldsymbol{W}_z \tilde{R}}_{\boldsymbol{b}_z}, \tag{19}$$

where $\tilde{\boldsymbol{\Phi}}' = (\phi(\mathbf{s}^{[1]\prime}, \pi(\mathbf{s}^{[1]\prime}))^T \quad \cdots \quad \phi(\mathbf{s}^{[N]\prime}, \pi(\mathbf{s}^{[N]\prime}))^T)^T$. Thus, the solution to minimizing the kernel-weighted empirical PBE can be obtained analytically as $\hat{\boldsymbol{\beta}}_+ = \boldsymbol{A}_z^{-1} \boldsymbol{b}_z$. This procedure is summarized in Figure 3.

3.3. *Componentwise regularization via group Lasso.* Since we are interested in obtaining a sparse representation of the elements in $\boldsymbol{\beta}_+$, we apply a penalty to an estimating equation $\mathcal{L}_z(\boldsymbol{\beta}_+)$ of (19). Since the components of our basis functions are grouped by features, we incorporate a group Lasso penalty that performs group-level variable selection by jointly constraining all coefficients that belong to a given feature. Consequently, the primary objective function for our estimator is

$$\mathcal{L}_z(\boldsymbol{\beta}_+) + \lambda \mathcal{R}(\boldsymbol{\beta}_+) = -\frac{1}{2} \boldsymbol{\beta}_+^T \tilde{\boldsymbol{A}}_z \boldsymbol{\beta}_+ - \boldsymbol{\beta}_+^T \tilde{\boldsymbol{b}}_z + \lambda \sum_a^{|\mathcal{A}|} \Big(\sqrt{m} \cdot |\alpha_a| + \sum_{j \geq 2} \| \boldsymbol{\beta}_{j;a} \|_2 \Big), \tag{20}$$

where λ is a regularization parameter. Note that the group Lasso penalty $\mathcal{R}(\boldsymbol{\beta}_+)$ includes a $\sqrt{m}$ factor, which is used to appropriately scale the strength of the regularization term λ

Algorithm 1: KSH-LSTDQ (via Randomized Coordinate Descent for Group Lasso)

Input: z (Fixed value), $\boldsymbol{\beta}_+^{(0)}$ (Initial weights), $K(\cdot)$ (Kernel function), $0 \leq \gamma < 1$ (Discount factor), μ (Step size), ϵ (Stopping Criteria), λ (Regularization Parameter), π (Current Policy)

Data: Dataset of transitions $\mathcal{D} = \{(\mathbf{s}^{[i]}, a^{[i]}, r^{[i]}, \mathbf{s}^{[i]\prime}, x^{[i]})\}_{i=1}^N$

Initialization: Construct $\boldsymbol{W}_z$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}'$

while $\|\boldsymbol{\beta}_+^{(t+1)} - \boldsymbol{\beta}_+^{(t)}\| \geq \epsilon$ **do**

  Select $j \in [d]$ with probability $1/d$

  for $a \in \mathcal{A}$ **do**

    Update $\boldsymbol{\beta}_{j,a}^{(t+1)} \leftarrow \mathcal{U}_{\lambda_j/\mu}(\boldsymbol{\beta}_{j;a}^{(t)} - \mu\, \boldsymbol{\Phi}_{\cdot j;a}^T \boldsymbol{W}_z((\boldsymbol{\Phi} - \gamma\boldsymbol{\Phi}')\boldsymbol{\beta}_+^{(t)} - \tilde{R}))$

  end

end

applied to $|\alpha_a|$ and to that of the coefficients of the B-splines basis functions. This ensures that the grouped coefficients get evenly penalized.

To estimate $\boldsymbol{\beta}_+$ under the objective function (20), we use the randomized coordinate descent method for composite functions proposed in Richtárik and Takáč (2014). Under this procedure we: (1) randomly select a coordinate j from $\{1, \ldots, d\}$ under a fixed action a and (2) update the current estimate of $\beta_{j;a}^{(t)}$. We then repeat steps (1) and (2) until convergence to $\hat{\boldsymbol{\beta}}_+$. Each update in (2) can be written in closed form as

$$U(\boldsymbol{\beta}_{j;a}^{(t)}) = \mathcal{U}_{\lambda_j/\mu}(\boldsymbol{\beta}_{j;a}^{(t)} - \mu \nabla_{j;a} \mathcal{L}_z(\boldsymbol{\beta}_+^{(t)})) \tag{21}$$

$$= \mathcal{U}_{\lambda_j/\mu}(\boldsymbol{\beta}_{j;a}^{(t)} - \mu \boldsymbol{\Phi}_{\cdot j;a}^T \boldsymbol{W}_z((\boldsymbol{\Phi} - \gamma\boldsymbol{\Phi}')\boldsymbol{\beta}_+^{(t)} - \tilde{R})), \tag{22}$$

where $\mathcal{U}_\lambda$ is a soft-thresholding operator defined as $\mathcal{U}_\lambda(\mathbf{v}) = (\mathbf{v}/\|\mathbf{v}\|_2) \cdot \max\{0, \|\mathbf{v}\|_2 - \lambda\}$, μ is the step size, and λ_j are regularization parameters. Note that $\lambda_1 = \lambda\sqrt{m}$ and $\lambda_j = \lambda\ \forall j \in [d]$. Details of the estimation procedure of KSH-LSTDQ are provided in Algorithm 1.

This algorithm allows us to retrieve an estimate of $\boldsymbol{\beta}_+$ relative to the fixed value z and, accordingly, approximate the additive functions in (5) as

$$\hat{g}_a(z) = \hat{\alpha}_{a,z} \quad \text{and} \quad \hat{f}_{j,a}(\mathbf{s}_j, z) = \sum_{k=1}^m \varphi_{jk}(\mathbf{s}_j)\hat{\beta}_{jk;a,z} \quad \forall j \geq 2. \tag{23}$$

To retrieve a nonlinear, smooth estimate of $g_a(\cdot)$ and $f_{j,a}(\mathbf{s}_j, \cdot)$, respectively, we compute the estimators $\hat{\alpha}_{a,z}$ and $\hat{f}_{j,a}(\mathbf{s}_j, z)$ for each value of z contained within the set $\mathcal{Z} = \{z_1, \ldots, z_M\}$ that densely covers the domain of x. This procedure amounts to running Algorithm 1 M times (i.e., once for each element in $\mathcal{Z}$).

3.4. *Approximate policy iteration.* The aforementioned estimation strategy is a policy evaluation method for obtaining an approximate representation of the action-value function Q^π under a fixed policy π. By using policy iteration, we can construct a procedure for estimating Q^* under an improved, or potentially optimal policy π^* (Howard (1960), Bertsekas (2011)).

To perform policy iteration, we begin with an arbitrary policy π_0 or the behavioral policy π_b used to generate $\mathcal{D}$. At each iteration t, we evaluate the current policy π_t by estimating Q^{π_t} according to (9) over a grid of local points $\mathcal{Z} = \{z_1, \ldots, z_M\}$. The policy improvement step follows by using the recently approximated action-value function Q^{π_t} to generate the new

Algorithm 2: KSH-LSPI

Data: $\mathcal{D} = \{(\mathbf{s}^{[i]}, a^{[i]}, r^{[i]}, \mathbf{s}^{[i]\prime}, x^{[i]})\}_{i=1}^{N}$

Input: $\mathcal{Z}$ (Array of fixed values), $\{\boldsymbol{\beta}_{+(i)}^{\text{init}}\}_{i=1}^{|\mathcal{Z}|}$ (Initial weights), $K(\cdot)$ (Kernel function), $0 \leq \gamma < 1$ (Discount factor), μ (Step size), ϵ (Stopping Criteria), λ (Regularization Parameter)

Initialization: Construct $\boldsymbol{\Phi}$

$\mathcal{B}^{(t)} = [\boldsymbol{\beta}_{+(i)}^{\text{init}}]_{i=1}^{|\mathcal{Z}|}$ (Matrix of model weights)

while $\|\mathcal{B}^{(t+1)} - \mathcal{B}^{(t)}\|_F \leq \epsilon$ **do**

  Construct $\boldsymbol{\Phi}'$ w.r.t. a greedy policy $\pi(s) = \arg\max_a \hat{Q}^{\pi}(s, a)$ using Section 3.4.

  for $i \in \{1 \ldots, |\mathcal{Z}|\}$ **do**

    $\mathcal{B}_{i\bullet}^{(t+1)} \leftarrow$ KSH-LSTDQ($\boldsymbol{\Phi}$, $\boldsymbol{\Phi}'$, $\mathcal{B}_{i\bullet}^{(t)}$, $z_i, \ldots$)

  end

end

greedy policy π_{t+1}. Since Q^{π_t} is represented using a grid of m local models (each computed with respect to each fixed value of x), the action selection strategy and representation of π_{t+1} is closely determined by our choice of x. This process, as detailed in Algorithm 2, repeats until convergence.

EXAMPLE 3. When $x = \mathbf{s}_0$, we represent the greedy policy $\pi(\mathbf{s})$ using the local model whose value of z is closest to $\mathbf{s}_0$. In other words, let $\mathcal{Z} = \{z_1, \ldots, z_M\}$ and $\mathcal{B} \in \mathbb{R}^{M \times (1+(d-1)m)|\mathcal{A}|}$ be a matrix of weights, where each row i corresponds to a set of model weights estimated under the value z_i. The greedy policy $\pi(\mathbf{s})$ is defined as $\pi(\mathbf{s}) = \arg\max_a \boldsymbol{\phi}(\mathbf{s}, a)^T \mathcal{B}_{i^*\bullet}$, where $i^* = \arg\min_{i \in \{1\ldots,M\}} |\mathbf{s}_0 - z_i|$.

EXAMPLE 4. When $x = a$, the greedy policy is represented as the fixed value z of local model that maximizes its associated action-value function. Let $\mathcal{Z} = \{z_1, \ldots, z_M\}$ and $\mathcal{B} \in \mathbb{R}^{M \times (1+(d-1)m)}$ be a matrix of weights, where each row i corresponds to a set of model weights estimated under the value z_i. The greedy policy $\pi(\mathbf{s})$ is defined as $\pi(\mathbf{s}) = z_{i^*}$, where $i^* = \arg\max_i \boldsymbol{\phi}(\mathbf{s}, \cdot)^T \mathcal{B}_{i\bullet}$.

4. Simulation study. In this section we perform a simulation study to examine the key properties of the KSH-LSTDQ and KSH-LSPI algorithms. We highlight the performance of the KSH-LSTDQ algorithm in estimating nonlinear marginal additive functions $g_a(x)$, examine its sensitivity to different environmental and model configurations, and compare the performance of the KSH-LSPI algorithm against several neural network-based approaches.

4.1. *Estimating marginal components.* We consider a multidimensional, continuous state MDP with binary actions and an additive reward function. For each sampled trajectory, the initial state vector $\mathbf{s}^{(0)} \in \mathbb{R}^d$ is sampled uniformly from $[0, 1]^d$. At each time step t, we randomly sample an action $a^{(t)} \in \{0, 1\}$ with probability $\frac{1}{2}$. Accordingly, the state transition function is defined as

$$\mathbf{s}^{(t+1)} = (-1)^a \big(\sin(\mathbf{x} + a\mathbf{u}) + 0.1\big(\mathbf{s}^{(t)}\big)^2\big) + \boldsymbol{\epsilon}, \tag{24}$$

where $\mathbf{x} \sim \text{Unif}(0, 2)^d$, $\mathbf{u} \sim \text{Unif}(0, 1)^d$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. This transition function allows us to explore complex, nonlinear dynamics into the MDP. We construct a reward function

$$r(\mathbf{s}, a) = u_1(\mathbf{s}_1, a) + u_2(\mathbf{s}_2, a) \tag{25}$$

with reward components that are reliant only on the state features $\mathbf{s}_1$ and $\mathbf{s}_2$, where

$$u_1(\mathbf{s}_1, a) = \big(5\mathbf{s}_1^2 + 5\big)\mathbb{1}(a = 1) - \big(2\mathbf{s}_1^3 - 5\big)\mathbb{1}(a = 0), \tag{26}$$

$$u_2(\mathbf{s}_2, a) = \big(5\sin\big(\mathbf{s}_2^2\big) + 5\big)\mathbb{1}(a = 1) + (4\mathbf{s}_2 - 5)\mathbb{1}(a = 0), \tag{27}$$

and $g_j(\mathbf{s}_j, a) = 0\ \forall j \geq 3$. For an arbitrary policy π, the construction of this reward function induces a corresponding action-value function that is additive with respect each nonzero reward component, specifically,

$$\begin{aligned} Q^\pi(\mathbf{s}, a) &= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i r\big(\mathbf{s}^{(i)}, a^{(i)}\big) \,\middle|\, \mathbf{s}^{(0)} = \mathbf{s}, a^{(0)} = a\right] \\ &= \sum_{j=1}^{2} \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i u_j\big(\mathbf{s}_j^{(i)}, a^{(i)}\big) \,\middle|\, s^{(0)} = \mathbf{s}, a^{(0)} = a\right] \\ &= U_1(\mathbf{s}_1, a) + U_2(\mathbf{s}_2, a). \end{aligned}$$

Using n trajectories sampled from this MDP (represented as a batch dataset $\mathcal{D}$), we evaluate the behavioral policy (i.e., $\pi(s^{[i]}) = a^{[i]}$) and retrieve the marginal component function $g_a(x)$ of the following nonparametric additive model:

$$Q^\pi(\mathbf{s}, a) = g_a(\mathbf{s}_i) + \sum_{j \in [d]/i} f_{j,a}(\mathbf{s}_j, \mathbf{s}_i) + \epsilon, \tag{28}$$

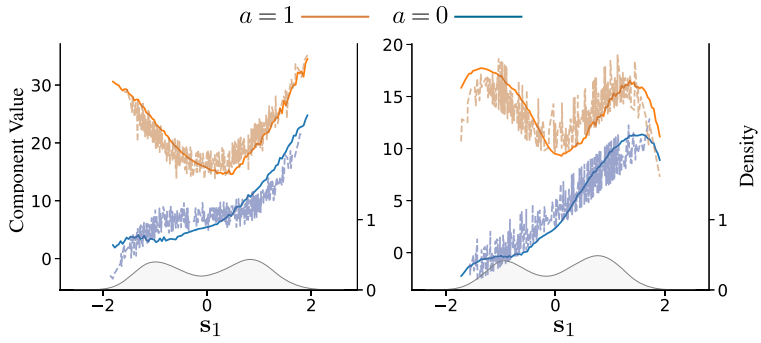
where $x = \mathbf{s}_i$ and $i \in \{1, 2\}$. For each component function, we explored using a B-spline basis (with hyperparameters including bandwidth h, number of basis functions m, and the degree of each piecewise polynomial) or a trigonometric basis function (hyperparameters include the number of basis functions m). We select the model's hyperparameters (including the regularization penalty λ and step size μ) using five-fold cross-validation. For each combination of hyperparameters, we compute the average out-of-sample Bellman loss across the five folds and select hyperparameters θ^* that minimize the average loss. To measure the performance of our model against a target, we utilize Monte Carlo (MC) sampling on the MDP to retrieve a direct estimate of the $Q^\pi(s, a)$, evaluated as

$$\hat{Q}^\pi_{\mathrm{MC}}(\mathbf{s}, a) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\ell} \gamma^j r_{ij}, \tag{29}$$

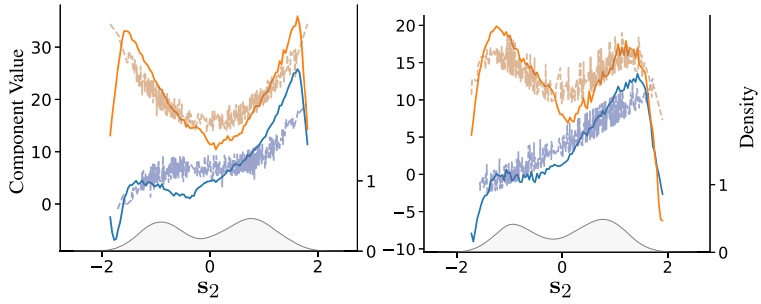
where ℓ is the length of each trajectory and n is the number of sampled trajectories. Since our action-value function is additive, we can similarly construct MC estimates for the component functions $U_1(\mathbf{s}_1, a)$ and $U_2(\mathbf{s}_2, a)$. Lastly, using a prespecified grid of points $\mathcal{Z} = \{z_1, \ldots, z_M\}$, we repeat Algorithm 1 M times to obtain smooth estimates of $g_a(\mathbf{s}_i)$, as described in Section 3.3.

Figures 4 and 5 present the nonlinear marginal component functions of the estimated model, as described in equation (28), and an ablation study over MDP and model hyperparameters, respectively. The dataset $\mathcal{D}$ comprised $n = 100$ sampled trajectories, each with a length of $\ell = 10$. MC estimates were obtained by sampling 100 trajectories of length ℓ.

Figure 4 compares the MC estimate of $u_i(\mathbf{s}_i, a)$ with the estimated component function $\hat{g_a}(\mathbf{s}_i)$ derived from the KSH-LSTDQ estimator, using bandwidths of $h = 0.01$ and $h = 0.1$. In this scenario we set the state space dimensionality to $d = 10$ and the discount factor to $\gamma = 0.5$. As the kernel function's bandwidth increases, the model generates a smoother component function, attributable to larger weights being assigned to observations further from each local point z. Notably, in sparsely sampled regions of the domain, the component function values tend toward zero, particularly with smaller bandwidths. This effect is partially driven by the

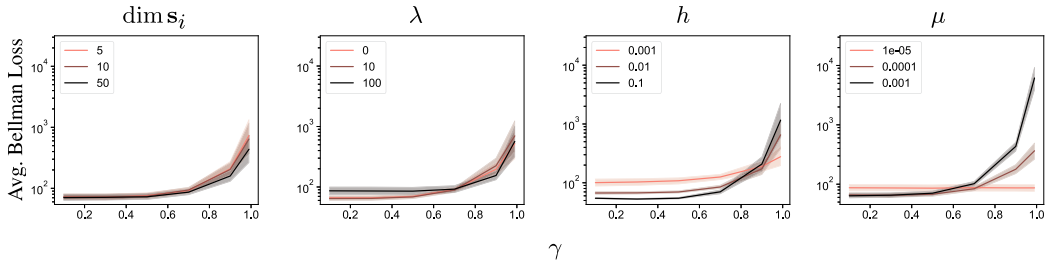


(a) Marginal effect $\hat{g}_a(\mathbf{s}_1)$ vs. Monte Carlo estimate of $U_1(\mathbf{s}_1, a)$.

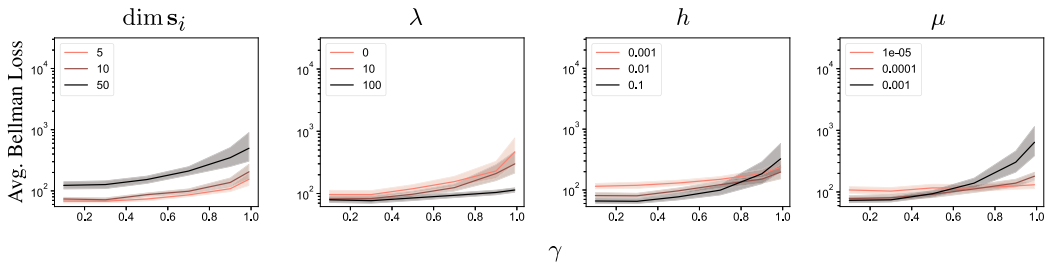


(b) Marginal effect $\hat{g}_a(\mathbf{s}_2)$ vs. Monte-Carlo estimate of $U_2(\mathbf{s}_2, a)$.

FIG. 4. *A comparison of estimated marginal component functions* $\hat{g}_a(\mathbf{s}_i)$ *and MC-estimates of* $u_i(\mathbf{s}_i, a)$, *as described in Section 4.1. For each action the solid lines represent the estimate of the marginal component function,* $g_a(\mathbf{s}_i)$, *of* $Q^\pi(s, a)$ *as modeled in equation (28) under a bandwidth of* $h = 0.01$ (*left*) *and* $h = 0.1$ (*right*) *under* $\gamma = 0.5$, *while the dashed line represents the Monte Carlo estimate of* $U_i(\mathbf{s}_i, a)$. *The observed distribution of the state feature* $\mathbf{s}_i$ *is displayed using the density in grey.*



(a) Simulation results using a B-spline basis expansion.



(b) Simulation results using a trigonometric polynomial basis expansion.

FIG. 5. *Results of the simulation study described in Section 4. Average Bellman loss (y-axis) vs. discount factor* γ (*y-axis*) *stratified by the dimensionality of* $\mathbf{s}_i$, *regularization penalty* λ, *bandwidth* h, *and step size* μ. *For each parameter setting, the solid lines represent the average Bellman loss as a function of* γ, *while the different line colors denote variations in* $\dim \mathbf{s}_i$, λ, h, *and* μ.

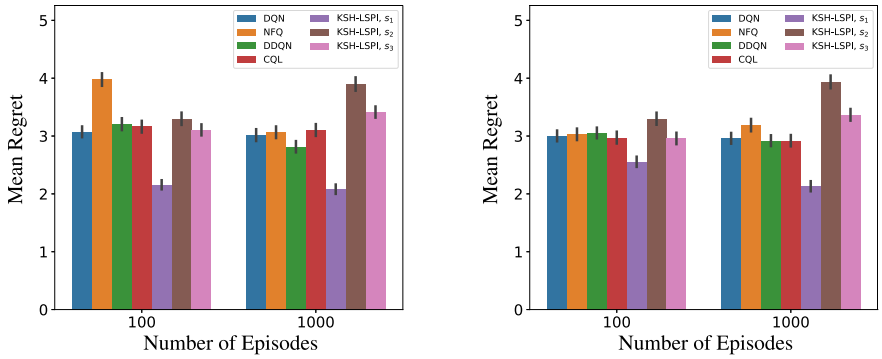
group Lasso penalty, which shrinks estimates within these sparse regions toward 0. While our model generally captures the underlying function's shape in nonsparse domain regions, our estimates may exhibit slight bias for complex functions.

Figure 5 illustrates results from an ablation study, depicting the average Bellman loss estimated across all folds on the y-axis against six different discount factors ($\gamma = 0.1, \ldots, 0.99$) on the x-axis. The results are further stratified by state space dimensionality $\dim \mathbf{s}_i$, regularization parameter λ, bandwidth h, step size μ, and choice of basis expansion method.

Generally, we observe, that as the discount factor γ increases, the average Bellman loss decreases. This can be attributed to increased instability in model estimation, as future next-state transitions and actions (i.e., $\gamma \boldsymbol{\Phi}'$) are weighted more heavily, potentially leading to a singular matrix $\boldsymbol{A}_z$ in equation (19). The results also indicate that the effect of increasing γ can be mitigated by increasing the strength of the regularization parameter (especially in trigonometric polynomial basis functions) and by using smaller step sizes. Furthermore, we observe minimal sensitivity to changes in the state dimension for B-spline basis functions, compared to trigonometric basis functions, where the average Bellman loss decreases as the state dimensionality grows. Lastly, we generally observe that as the bandwidth increases, the average Bellman loss across γ decreases.

4.2. *Comparison to neural approaches*. We evaluate the performance of the KSH-LSPI algorithm against several widely-used neural network-based approaches, specifically: neural fitted Q-iteration (NFQ), deep Q-network (DQN), double deep Q-network (DDQN), and conservative Q-learning (CQL) (Mnih et al. (2013), Van Hasselt (2010), Riedmiller (2005), Kumar et al. (2020)). Each model is trained using a batch dataset of experiences, which is gathered from a random policy interacting in an MDP with correlated state features and an additive reward function. Similar to equation (25), the reward function is dependent on the first two state features $\{\mathbf{s}_1, \mathbf{s}_2\}$ and the selected action a. Appendix A.1 in the Supplementary Material provides a detailed description of the MDP and the data generation process (Emedom-Nnamdi et al. (2025)). In each experiment we adjust the dimensionality of the MDP's state space and the number of episodes used to generate the batch dataset. For the KSH-LSPI algorithm, we fit a separate model, where the candidate feature x is represented as one of the first three state features $\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$. Here each state feature, denoted as $\mathbf{s}_i$, contributes to the marginal component, $g_a(\mathbf{s}_i)$, as illustrated in equation (28). We perform policy iteration in accordance with Example 3, setting the maximum number of allowed policy iterations to three. Detailed specifications and architectures of both the KSH-LSPI models and the neural network-based approaches can be found in Appendix A.2 in the Supplementary Material (Emedom-Nnamdi et al. (2025)).

The estimated policies for each approach were evaluated within the MDP used to generate the training batch data set. Specifically, each policy was rolled out for 10 time steps (i.e., an episode), 1000 times. We performed a regret analysis, where, at the end of each episode, the difference between the optimal reward at each time step and the reward obtained by the current policy was calculated. The average of these differences over all episodes was then computed to obtain the estimated mean regret for each experiment. Figure 6 presents results from the regret analysis, where the dimensionality of the state space and the number of episodes used to generate the batch dataset were varied. Within each experiment we observe that the KSH-LSPI model under a candidate feature of $\mathbf{s}_2$ and $\mathbf{s}_3$, performs similarly to neural network-based approaches, when the number of episodes is 100, and worse when the number of episodes used in the generated batch data set increases to 1000. Conversely, when $\mathbf{s}_1$ (i.e., the feature that accounts for the most variation within the observed rewards) is set as the candidate feature, the KSH-LSPI model outperforms the neural-network based models and is further improved as the number of episodes increases. These results highlight a key sensitivity



(a) Dimensionality of $\mathbf{s}_i$ is 5. (b) Dimensionality of $\mathbf{s}_i$ is 50.

FIG. 6. *Regret analysis comparing the performance of KSH-LSPI models, where the candidate feature x is independently represented using state features $\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ and neural network-based approaches, as described in Section 4.2. Within each subfigure the dimensionality of the state space and the number of episodes used to generate the batch dataset are varied.*

in the KSH-LSPI model; that is, the appropriate selection of the candidate feature x largely influences model performance.

Lastly, the marginal performance differences among the neural network approaches may be attributable to the simulation environment. For example, in more complex, high-dimensional environments with offline data, we would expect approaches such as CQL to provide more significant benefits.

5. Motivating case study. Postoperative recovery is defined as the period of functional improvement that occurs from the end of surgery and hospital discharge to the instance in which normal function has been restored (Bowyer and Royse (2016)). Depending on the type of surgery administered, this period of functional recovery can vary drastically and be accompanied by mild to severe complications. For patients who received corrective surgery for spine disease, postoperative recovery is impacted by the complexity of the diagnosis and surgical procedure received. Additional barriers to recovery for spine disease patients include stress, pain, cognitive dysfunction, and potential postoperative complications (Wainwright, Immins and Middleton (2016)).

To improve postoperative recovery and care for spine patients, physicians employ a multi-pronged approach focusing on protocols that accelerate functional recovery, decrease postoperative complications, and improve subjective patient experience (Elsarrag et al. (2019)). As part of this effort, patient mobilization and consistent pain management are heavily suggested (Burgess and Wainwright (2019)). To advance these efforts, physicians require objective measurements of a patient's functional capacity and pain over the course of their recovery (Cote et al. (2019), Panda et al. (2020a), Karas et al. (2020), Boaro, Reeder and Siddi (2021)). With respect to spine patients, such measurements can provide a formal understanding and quantification of mobilization activities that expedite overall patient recovery and minimize the risk of complications.

We consider $n = 67$ neurosurgical spine patients with a median age of 57 years (IQR: 48–65.5) who were enrolled between June 2016 and March 2020 as part of a mobile health study at Brigham and Women's Hospital. Each patient underwent a neurosurgical intervention in relation to their spine disease. For data collection, patients installed the Beiwe application on their smartphones. Beiwe is a high-throughput research platform that was developed by the Onnela lab at Harvard T.H. Chan School of Public Health for smartphone-based digital phenotyping on iOS and Android devices. Passive features collected on Beiwe include

TABLE 1
Subset of GPS and accelerometer-based summary statistics. Definitions can be found on the Forest GitHub repository (www.github.com/onnela-lab/forest)

Distance Traveled (km)	Radius of Gyration (km)	Average flight duration (km)
Time Spent at Home (hours)	Maximum Diameter (km)	Fraction of the day spent stationary
Max. Distance from Home (km)	Num. Significant Places Visited	Time Spent Walking
Average flight length (km)	Number of Steps	Average Cadence

GPS and accelerometer data in their raw unprocessed form, Bluetooth and WiFi logs, and anonymized phone call and text message logs. Samples were collected from the GPS data stream for one minute every five minutes and from the accelerometer data stream for 10 seconds every 10 seconds. Using raw data sampled from GPS and accelerometer sensors, a set of behavioral characteristics is calculated regarding patient mobility at the daily level (Liu and Onnela (2021)). A subset of these features are represented in Table 1. For active data collection, patients were electronically surveyed once daily at 5 p.m. Eastern Standard Time to evaluate their current pain level. The prompt of the micro-survey was "Please rate your pain over the last 24 hours on a scale from 0 to 10, where 0 is no pain at all and 10 is the worst pain imaginable."

In conjunction with the daily self-reported micro-surveys, these constructed features allow researchers to objectively identify postoperative trends in mobility and pain as it relates to overall functional recovery, as shown in Figures 7 and 8 (Boaro, Reeder and Siddi (2021), Cote et al. (2019)). To this end, we leverage reinforcement learning to estimate and interpret mobility-based action-value functions that provide recommendations concerning questions such as "What level of mobilization is advisable after surgery?" and "How should these levels be adjusted given a patient's current condition?" The overall goal of these recommendations is to manage a patient's overall pain level and promote improved recovery. Furthermore, by utilizing an interpretable representation of the estimated action-value function, we seek to identify clinical and behavioral features that are important to consider for decision-making.

6. Application to surgical recovery. Using data from the spine disease cohort in Section 5, we use nonparametric additive models to estimate action-value functions associated with:

1. A *behavioral policy* that aims to mimic decisions commonly taken by patients, and

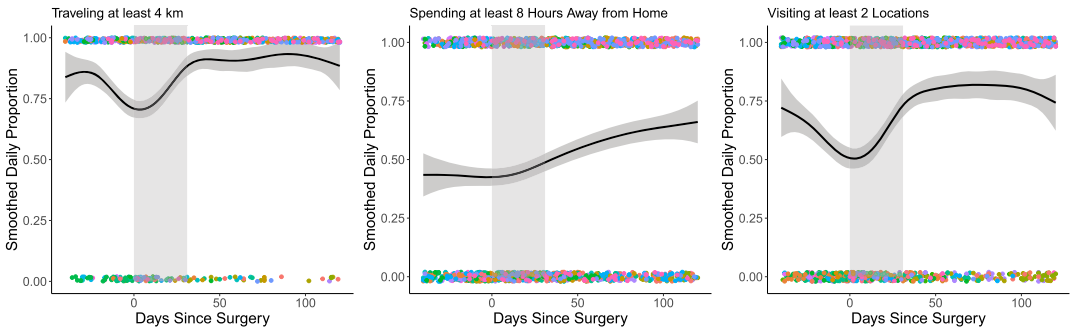


FIG. 7. *Smoothed mobility proportions (with standard errors represented in grey) for spine disease cohort centered on day of surgery. The lighter shaded area corresponds to the first 30 postoperative days. Receiving a neurosurgical intervention is followed by a period of decreased mobility, where patients tend to travel less and stay at home over a longer duration. Individual-level differences in recovery are driven by factors such as the type of surgery received, the specific diagnosis of spine disease, and patient demographics.*

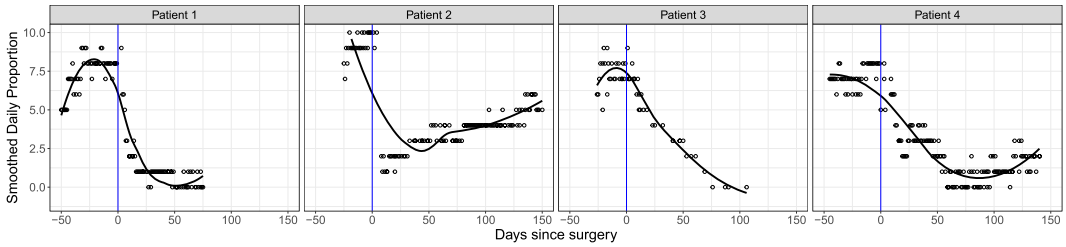


FIG. 8. *Pre- and postoperative pain responses with time centered on the day of surgery* (*i.e.*, *blue line*) *with a fitted local regression* (*i.e.*, *black line*) *for a random selection of patients. While surgery corresponds to a sharp decline in self-reported pain, we observe a heterogeneous recovery experience among these four patients.*

2. An *improved policy* retrieved from performing approximate policy iteration on the estimated behavioral policy.

In both cases the estimated decision-making policy aims to suggest the daily number of steps necessary to reduce long-term ($\gamma \gg 0$) postoperative pain response. We explore both discrete and continuous action spaces and provide a practical interpretation of the additive functional components, as presented in equations (7) and (8), respectively.

6.1. *Data preprocessing*. We consider the recovery period of $n = 67$ neurosurgical spine disease patients with a mean postoperative follow-up of 87 days (SD = 51.21 days). Baseline clinical information on this study cohort can be found in Table 2. Behavioral features, derived from raw GPS and accelerometer data, were summarized on a daily time scale to

TABLE 2
Participant demographic information and mobile health data for spine disease cohort. Summaries are computed according to the first 60 *postoperative days since surgery* (*including the day of surgery*)

Variable	*n* (%) or Median (25th–75th)
Demographic Data	
Age	57.0 (48.0–65.5)
Female gender	34 (50.7)
Site of surgery	
Cervical	19 (28.4)
Lumbar	27 (40.3)
Thoracic	2 (3.0)
Multiple	18 (26.9)
Data Collection	
GPS days of follow-up	61 (49–61)
Accelerometer days of follow-up	61 (50.5–61)
Daily pain survey response rate	59.4 (42.4–76.9)
Digital Phenotypes	
Number of places visited	3 (2–5)
Time spent at home (hours)	18.3 (12.9–21.9)
Distance traveled (km)	32.3 (10.8–62.3)
Maximum distance from home (km)	10.6 (4.5–25.5)
Radius of gyration (km)	1.50 (0.18-5.01)
Time spent not moving	21.2 (20.2–22.2)
Average cadence	1.64 (1.55–1.74)
Number of steps	948.6 (356.9–2005)

closely monitor each patients' clinical recovery and/or progression after surgery. These features include passively sampled summary statistics that uniquely describe a patient's daily mobility and activity levels.

We construct a simple MDP where each time step t corresponds to a day since surgery. The state space, $S \in \mathbb{R}^d$, is a multidimensional, continuous state vector that consists of relevant behavioral features and patient-specific demographic information (i.e., age and days since surgery). In total, $d = 9$ features were used in this analysis.[1] The action space $\mathcal{A} \in \mathbb{R}$ represents the number of steps taken per day. For the discrete action model, the action space, $\mathcal{A} \in \{0, 1\}$, is binarized such that 0 represents moving less than the subject-level preoperative median number of steps taken per day and 1 represents moving above this threshold. The rewards, $r \in \mathbb{R}$, are chosen to be the negative value of the self-reported pain score, where each score is taken from a numerical rating scale between 0 (i.e., no pain) and 10 (i.e., worst pain imaginable). Lastly, we consider a discount factor γ of 0.5 to examine estimated policies that aim to reduce long-term pain response.

Under this MDP we consider up to the first 60 days since surgery for each patient. Patients with a postoperative follow-up period of less than five days were excluded. Entries with missing values in either the behavioral features or the daily self-reported pain scores were removed. The batch dataset $\mathcal{D}$ with $N = 1409$ daily transitions was constructed using data collected from the study cohort and represented using the MDP. All state features were normalized to a [0, 1] range for model fitting.

6.2. *Model fitting*. To estimate the action-value function associated with the behavioral policy π_b, we use Algorithm 1 and construct $\mathbf{\Phi}'$ using the observed next-state action contained within each patient-level trajectory in $\mathcal{D}$. That is, $\mathbf{\Phi}'_{i\cdot} = \phi(\mathbf{s}'^{[i]}, a^{[i+1]})$ for the ith observed transition. Accordingly, we estimate the action-value function for the improved policy π^* using approximate policy iteration (as detailed Algorithm 2) on the action-value function associated with the behavioral policy. For both discrete and continuous action versions of the general model (5), we use a Gaussian kernel $K(u) = e^{-\frac{1}{2}u^2}$ and a grid of evenly spaced points $\mathcal{Z}$ within a [0, 1] range for discretization. In the discrete action model, we estimate the marginal effect $g_a(x)$ and the additive joint effects $f_{j,a}(\mathbf{s}_j, x)$ for $j \geq 2$ for each candidate state feature x in a set $\mathcal{H}$, where $x \neq \mathbf{s}_j$.

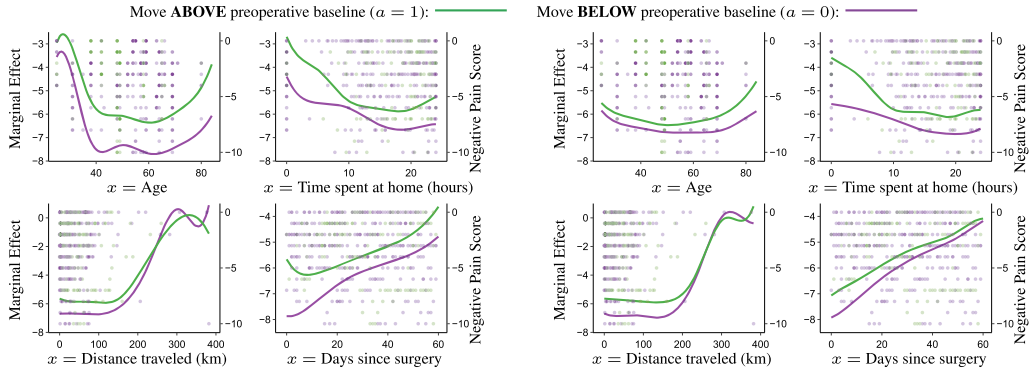
To select the hyperparameters of the KSH-LSTDQ estimator (i.e., the degree of the B-spline functions, the number of basis functions, the bandwidth, and the regularization penalty), we partitioned the dataset $\mathcal{D}$ into training and validation sets according to a 80%–20% patient-level split and performed a grid search. Using these partitions, we retrieved a set of hyperparameters that minimized the validation mean squared error between the estimated action-value under the behavioral policy when $\gamma = 0$ and the true immediate rewards,

$$\text{MSE}(\mathcal{D}_{\text{Val}}) = \frac{1}{|\mathcal{D}_{\text{Val}}|} \sum_{(\mathbf{s},a,r)\sim\mathcal{D}_{\text{Val}}} \big(\hat{Q}^{\pi_b}(\mathbf{s}, a) - r\big)^2.$$

Accordingly, these hyperparameters were used to retrieve the KSH-LSTDQ estimators for MDPs where γ is set to 0.5. The set of hyperparameters used for each estimated model is shown in Appendix A.2 in the Supplementary Material (Emedom-Nnamdi et al. (2025)).

6.3. *Results and interpretations*. We visualize and interpret the estimated additive component functions of the action-value functions associated with the behavioral and improved policies.

[1]These features include: Age, number of days since surgery, time spent at home (hours), distance traveled (km), maximum distance from home (km), radius of gyration (km), average cadence, and time spent not moving (hours).



(a) **Behavioral policy** (b) **Policy Iteration**

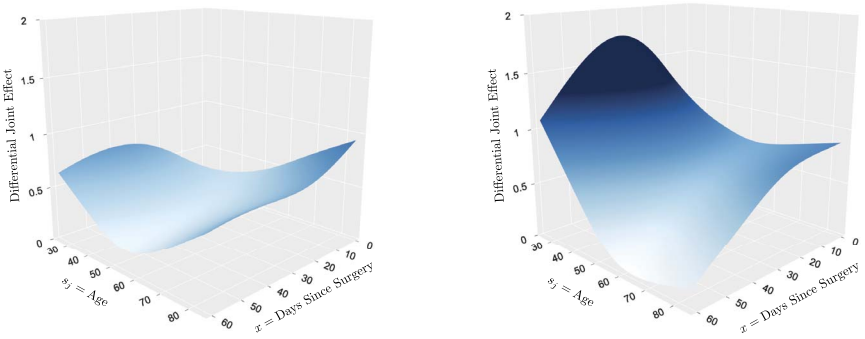
FIG. 9. *A comparison of the marginal component function* $\hat{g}_a(x)$ *of* $Q^\pi(\mathbf{s}, a, x)$ *estimated under the behavioral policy* $\pi = \pi_b$ *vs. the improved policy* $\pi = \pi^*$. *Each subfigure is associated with a separate nonparametric additive model of* $Q^\pi(\mathbf{s}, a, x)$, *where the state feature representing* x *is changed. For each action the solid lines represent the estimate of the marginal component function* $\hat{g}_a(x)$ *over the range of observed values of* x, *whereas the points represent the value of the associated observed rewards* (*i.e.*, *negative pain score*) *over* x.

6.3.1. *Discrete action model.* In the discrete action model, we estimate a nonparametric action-value function for each candidate state feature in the set $\mathcal{H}$, which we represent as x. Here $\mathcal{H}$ consists of the following features: Age, number of days since surgery, time spent at home (hours), and distance traveled (km).

Marginal state feature effects. In Figure 9 we examine the marginal effect $\hat{g}_a(x)$ of the estimated action-value function for each candidate state feature $x \in \mathcal{H}$ under the behavioral policy π_b and the improved policy π^* constructed using policy iteration. Specifically, $\hat{g}_a(x)$ estimates the marginal change in *long-term* negative pain response for each state feature x under action a. Across each subfigure in Figure 9(a), moving above a patient's preoperative baseline number of steps ($a = 1$) within a given day is associated with a higher immediate negative pain response in comparison to the converse action ($a = 0$), regardless of the value of x. This observation is in line with clinical research that suggests movement at or above a patient's preoperative baseline is associated with improved postoperative functional recovery (Duc et al. (2013), Ozkara et al. (2015), Cote et al. (2019)).

Within each select action, the marginal effect shows a nonlinear change in long-term pain response. When $x = age$, Figure 9(a) suggests a marginal increase in long-term negative pain response for younger and older ages. This observation supports clinical studies suggesting the existence of age-related pain sensitivity that peaks during mid-life (Yezierski (2012)). Additionally, when $x = time\ spent\ at\ home$, our model suggests that spending more time at home is associated with a nonlinear decrease in negative pain response. Furthermore, when $x = distance\ traveled$, we observe that, regardless of the selected action, traveling less than 150 km is associated with a constant effect on negative pain response, whereas traveling beyond 150 km within a given day is associated with an increasing effect. We note that this association is possibly due to survivorship bias present in the model estimates, where a few patients report minimal pain during periods of excessive travel.

Lastly, when $x = days\ since\ surgery$, our model examines the impact of mobilization as the number of days since surgery increases. We note the difference in the marginal effect among both actions is maximized for days closest to the onset of surgery, suggesting that increased mobilization during early periods of recovery may be associated with decreased pain response. This finding supports current clinical practice that suggests early mobilization enhances surgical recovery, a cornerstone of postoperative pain management (Wainwright, Immins and Middleton (2016), Burgess and Wainwright (2019)).



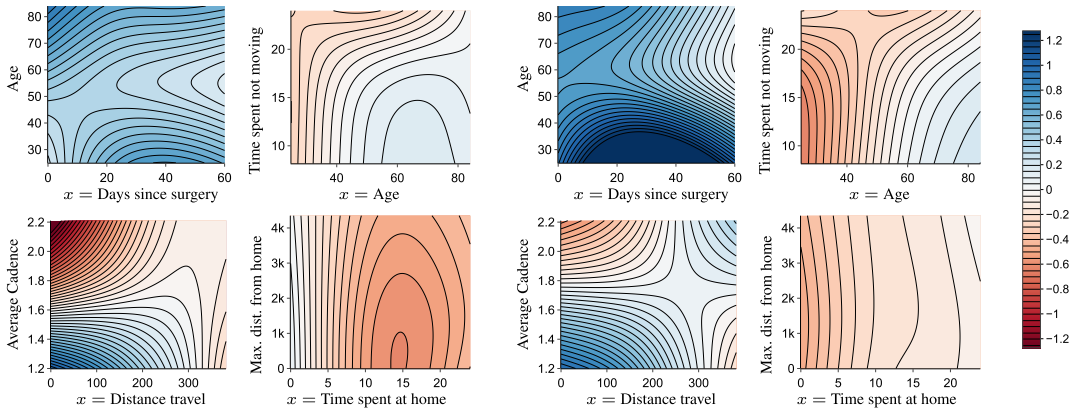
(a) **Behavioral policy** (b) **Policy Iteration**

FIG. 10. *Surface plots representing the difference between the joint component functions* $\hat{f}_{j,1}(\mathbf{s}_j, x)$ *and* $\hat{f}_{j,0}(\mathbf{s}_j, x)$ *(i.e., the differential benefit of selecting action* $a = 1$ *over* $a = 0$*) of* $Q^{\pi}(\mathbf{s}, a, x)$ *estimated under the behavioral policy* $\pi = \pi_b$ *vs. the improved policy* $\pi = \pi^*$.

The differences between the marginal effects associated with the behavioral and improved policies, as shown in Figure 9(b), are subtle. While the underlying trends and ordering of actions are relatively consistent, the estimated effect sizes appear to be smaller for select candidate state features under the improved policy (e.g., $x =$ age, time spent at home, or days since surgery) compared to those of the behavioral policy.

Joint effects betwen state features. In Figures 10 and 11, we examine the joint effect $f_{j,a}(x, \mathbf{s}_j)$ of the estimated action-value function between select state features $\mathbf{s}_j$ (i.e., age, time spent not moving, average cadence, and maximum distance from home) and candidate state features $x \in \mathcal{H}$. The value of each joint feature pair corresponds to a nonlinear effect on an estimated smooth surface representing the additive, long-term change in negative pain response. We specifically examine the benefit of selecting a given action over its converse by visualizing the difference between the joint effects under both actions, that is, $f_{j,1} - f_{j,0}$. Differences greater than zero indicate an additive preference for action $a = 1$, over the converse $a = 0$.

When examining the joint effect between $x =$ *days since surgery* and $\mathbf{s}_j =$ *age*, we observe that regardless of the value of each corresponding feature, moving more than the preopera-



(a) **Behavioral policy** (b) **Policy Iteration**

FIG. 11. *Contour plots representing the differential benefit of selecting action* $a = 1$ *over* $a = 0$ *with respect to joint effects* $\hat{f}_{j,a}(\mathbf{s}_j, x)$ *under* $Q^{\pi}(\mathbf{s}, a, x)$ *estimated under the behavioral policy* $\pi = \pi_b$ *vs. the improved policy* $\pi = \pi^*$. *Each subfigure is associated with a separate nonparametric additive model of* $Q^{\pi}(\mathbf{s}, a, x)$*, where the state feature representing* x *is changed.*

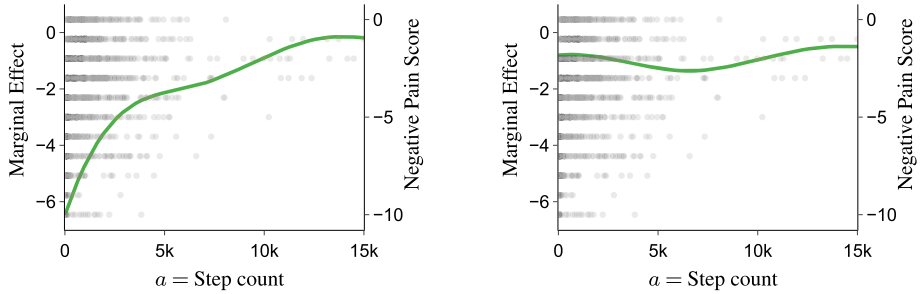
tive baseline is associated with an increase in negative pain response throughout the domain of the joint component function. Interestingly, this association is more pronounced among younger patients under the improved policy. This observation is consistent with clinical research suggesting a relationship between increased postoperative movement and improved rehabilitation and its potential modification by factors such as age (Ozkara et al. (2015), Duc et al. (2013), Jaensson, Dahlberg and Nilsson (2019)).

When $x = \textit{distance traveled}$ and $\mathbf{s}_j = \textit{average cadence}$, maintaining a slower average walking cadence over longer distances seems to be associated with the suggested action of moving beyond the preoperative baseline step count. This association is relatively consistent across both the behavioral and improved policies. However, under the improved policy, a positive association is also noticed with faster walking cadences near the upper boundary of total distance traveled. In general, the differential relationship between $x = \textit{distance traveled}$ and $\mathbf{s}_j = \textit{average cadence}$ could be indicative of the shift from automaticity to executive control of locomotion, as seen in rehabilitation literature (Clark (2015)). This shift may occur as distances increase or as walking becomes more challenging (e.g., due to physical exertion, elevated pain response, or injury), requiring individuals to expend more cognitive effort (i.e., executive control) to manage their gait.

6.3.2. *Continuous action model.* We estimate a nonparametric action-value function for continuous actions, specifically the number of steps taken, under both behavioral and improved policies.

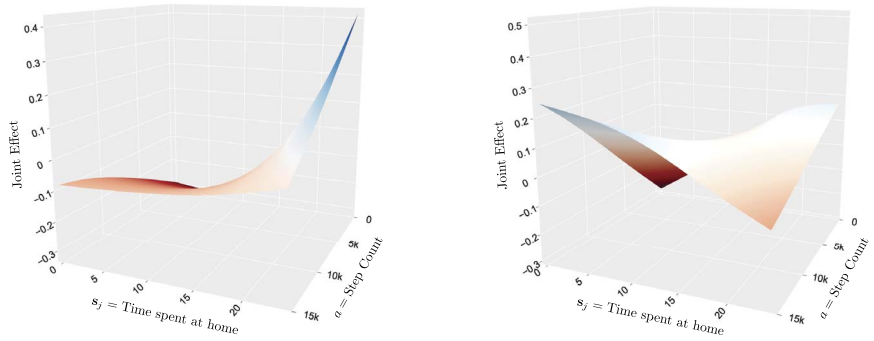
Marginal effects. In Figure 12 we examine the marginal effect $\hat{g}(a)$ of the estimated action-value function. Specifically, $\hat{g}(a)$ estimates the marginal change in *long-term* negative pain response for a select value of a, or number of steps taken. As with the discrete action model, we observe a positive association between the number of steps taken and the long-term negative pain response in the continuous action model, especially under the behavioral policy. For the behavioral policy (as shown in Figure 12(a)), we observe that the marginal effect is log-shaped and increases with number of steps taken. For the improved policy (Figure 12(b)), the marginal effect remains relatively constant across the observed number of steps.

Joint effects between state features and actions. In Figures 13 and 14, we examine the joint effects $\hat{f}_j(\mathbf{s}_j, a)$ between select state features $\mathbf{s}_j$ and the continuous action a of the estimated action-value functions. When $a = \textit{step count}$ and $\mathbf{s} = \textit{time spent at home}$, we observed that increased time spent at home beyond 15 hours is associated with a positive increase in negative pain response across observed values of step count under the behavioral policy. This trend changes under the improved policy, where the joint effect is maximized when step count



(a) **Behavioral policy** (b) **Policy Iteration**

FIG. 12. *The marginal component function $\hat{g}(a)$ of $Q^{\pi}(\mathbf{s}, a)$ estimated under the behavioral policy $\pi = \pi_b$ vs. the improved policy $\pi = \pi^*$. The solid lines represent the estimate of the marginal component function $\hat{g}(a)$ over the observed values a, whereas the points represent the value of the associated observed reward (i.e., negative pain score).*



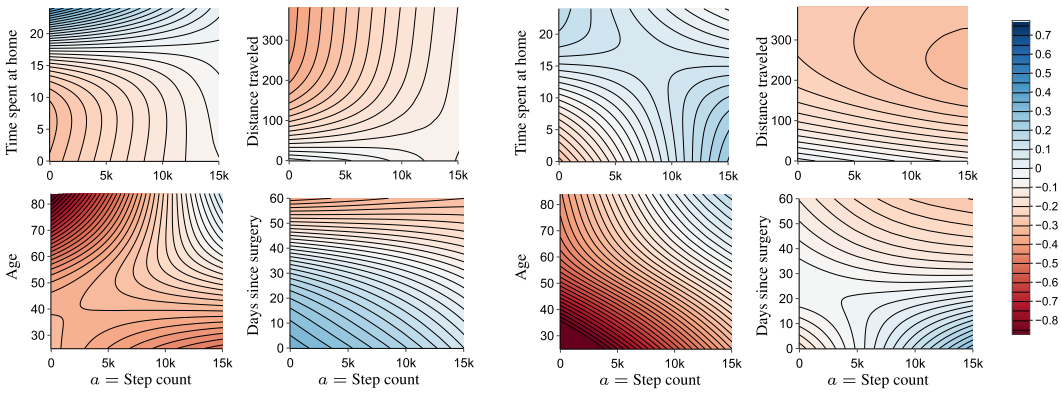
(a) **Behavioral policy** (b) **Policy Iteration**

FIG. 13. *Surface plots representing the joint component functions* $\hat{f}_j(\mathbf{s}_j, a)$ *of* $Q^{\pi}(\mathbf{s}, a)$ *estimated under the behavioral policy* $\pi = \pi_b$ *vs. the improved policy* $\pi = \pi^*$.

increases and time spent at home decreases and when step count is minimized and time spent at home increases. This suggests the joint effect of the improved policy may prioritize two distinct recovery strategies: patients can either maintain higher activity levels with less time at home or focus on rest with reduced activity.

Similar to the discrete action model, we observe a differential change in the joint effect associated with age. Under the improved and behavioral policy, an increase in step count across age is associated with an relative increase in the negative pain response.

7. Discussion. In this study we introduce a flexible, nonparametric additive representation of action-value functions. Our approach, KSH-LSPI, distinguishes itself from LSPI by avoiding parametric assumptions regarding the action-value function's form beyond additivity. It achieves this by incorporating ideas directly from local kernel regression and spline methods. This estimation approach affords KSH-LSPI the ability to capture the nonlinear additive contribution of each state action feature represented in the model. Furthermore, by introducing a group Lasso penalty to our primary objective function, we perform componentwise variable selection and retrieve a parsimonious representation of the action-value function. As demonstrated in our simulation, this approach can achieve competitive performance



(a) **Behavioral policy** (b) **Policy Iteration**

FIG. 14. *Contour plots representing the joint component functions* $\hat{f}_j(\mathbf{s}_j, a)$ *of* $Q^{\pi}(\mathbf{s}, a)$ *estimated under the behavioral policy* $\pi = \pi_b$ *vs. the improved policy* $\pi = \pi^*$. *Each subfigure is associated with the same nonparametric additive model but represents a different state feature* $\mathbf{s}_j$ *on the y-axis.*

against modern neural network methods, particularly in scenarios with limited training data, while maintaining interpretability.

In the simulation study, we evaluated the performance of the proposed estimator and examined its sensitivity to changes in its hyperparameters. Future work aims to delve deeper, examining the estimator's finite sample properties both theoretically and through further simulations. The application of the proposed method to the spine disease dataset also provides new insights into mobilization behaviors that support postoperative pain management and reaffirmed several well-studied clinical findings. In future applications to spine disease recovery, we hope to extend the model by including categorical features, such as gender, race, and diagnosis, as well as additional clinical features such as medication use.

Our empirical analyses also revealed several methodological limitations. The method exhibits sensitivity to the discount factor, where higher values lead to estimation instability due to increased weighting of future states. The performance of the method depends on the selection of candidate features, where suboptimal choices can result in performance inferior to neural network approaches, an effect that becomes more pronounced as the training data increases. In regions of sparse data, component function values trend toward zero, an effect exacerbated by the group Lasso penalty and smaller bandwidth selections. Although the method effectively captures general functional shapes in well-sampled regions, it exhibits bias when estimating more complex underlying functions. Additionally, we found notable dependencies on the selected basis function, where trigonometric basis functions display greater sensitivity to state dimensions, compared to B-spline bases, and require stronger regularization for higher discount factors.

Furthermore, our application is not without limitations. The results of our model require careful interpretation and should not be deemed significant without comprehensive uncertainty quantification. In future adaptations of the KSH-LSPI model, we hope to formalize our uncertainty concerning our model estimates by incorporating a form of interval estimation. In the offline reinforcement learning setting, uncertainty-based approaches have shown promise in offline RL by prioritizing risk adverse policies when performing policy improvement (Sonabend-W et al. (2020), O'Donoghue et al. (2017), Ghavamzadeh et al. (2016)). This naturally brings light to a limitation concerning our method's approach for policy improvement. After evaluating the current policy using KSH-LSTDQ, our policy improvement step greedily selects actions that maximize the estimated action-value function. Unfortunately, function approximation methods in offline reinforcement learning are prone to providing overly optimistic values for state action pairs that are unobserved in the training data. Hence, safe policy improvement steps, within the actor-critic framework, that regularize the learned policy toward the behavioral policy is encouraged in offline reinforcement learning, especially in healthcare applications (Wang et al. (2020)). Another potential remedy would be to initialize our algorithm using an initial policy that closely reflects behaviors that would be suggested by a clinical expert. Initialization using physician-guided policies helps prevent the algorithm from becoming overly optimistic by selecting best actions that physicians themselves may select (Gottesman et al. (2019)).

The push for interpretability in machine learning models, especially within healthcare contexts, is driven by a need for transparency in decision-making processes. Compared to the powerful, but often less interpretable neural network methodologies, nonparametric additive models for value-functions offers a representation where decision-making policies can be understood and scrutinized. Such interpretability is essential for potential clinical applications, given the need for clinicians to trust and validate the recommendations derived from these decision-making models. In conclusion, the KSH-LSPI model, although it has areas that require further refinement, provides a promising framework that aligns with the demand for both efficacy and transparency.

Acknowledgments. The authors would like to thank the anonymous referees, the Associate Editor, and the Editor for their constructive feedback and input.

Funding. Junwei Lu is supported by NSF Grant DMS-2434664, William F. Milton Fund, NIH/NCI R35CA220523, and NIH/NINDS R01NS098023.

SUPPLEMENTARY MATERIAL

Supplement A—Appendices (DOI: 10.1214/24-AOAS1987SUPPA; .pdf). We provide appendices containing additional information regarding the simulation study and the hyperparameter selection for the application to surgical recovery.

Supplement B—Python code (DOI: 10.1214/24-AOAS1987SUPPB; .zip). We provide a python implementation of the KSH-LSPI algorithm to reproduce the simulations and analysis performed in the paper. This code is provided as a supplement and is also available at https://github.com/patricknnamdi/ksh-lspi.

REFERENCES

AHN, I. and PARK, J. (2011). Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Biosystems* **106** 121–129. https://doi.org/10.1016/J.BIOSYSTEMS.2011.07.005

ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2007). Fitted Q-iteration in continuous action-space MDPs. *Adv. Neural Inf. Process. Syst.* **20**.

BERTSEKAS, D. P. (2011). Approximate policy iteration: A survey and some new methods. *J. Control Theory Appl.* **9** 310–335. MR2833999 https://doi.org/10.1007/s11768-011-1005-3

BOARO, A., REEDER, H. T. and SIDDI, F. (2021). Smartphone GPS signatures of patients undergoing spine surgery correlate with mobility and current gold standard outcome measures Digital Phenotyping in Neurosurgery View project Intracranial aneurysm View project. *J. Neurosurg., Spine* **35** 796–806. https://doi.org/10.3171/2021.2.SPINE202181

BOTHE, M. K., DICKENS, L., REICHEL, K., TELLMANN, A., ELLGER, B., WESTPHAL, M. and FAISAL, A. A. (2013). The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev. Med. Devices* **10** 661–673. https://doi.org/10.1586/17434440.2013.827515

BOWYER, A. J. and ROYSE, C. F. (2016). Postoperative recovery and outcomes—What are we measuring and for whom? *Anaesthesia* **71** 72–77. https://doi.org/10.1111/anae.13312

BURGESS, L. C. and WAINWRIGHT, T. W. (2019). What is the evidence for early mobilisation in elective spine surgery? A narrative review. *Healthcare* **7** 92. https://doi.org/10.3390/healthcare7030092

CLARK, D. J. (2015). Automaticity of walking: Functional significance, mechanisms, measurement and rehabilitation strategies. *Front. Human Neurosci.* **9** 246. https://doi.org/10.3389/fnhum.2015.00246

COTE, D. J., BARNETT, I., ONNELA, J.-P. and SMITH, T. R. (2019). Digital phenotyping in patients with spine disease: A novel approach to quantifying mobility and quality of life. *World Neurosurg.* **126** e241–e249. https://doi.org/10.1016/j.wneu.2019.01.297

DUC, C., SALVIA, P., LUBANSU, A., FEIPEL, V. and AMINIAN, K. (2013). Objective evaluation of cervical spine mobility after surgery during free-living activity. *Clin. Biomech.* **28** 364–369. https://doi.org/10.1016/j.clinbiomech.2013.03.006

ELSARRAG, M., SOLDOZY, S., PATEL, P., NORAT, P., SOKOLOWSKI, J. D., PARK, M. S., TVRDIK, P. and KALANI, M. Y. S. (2019). Enhanced recovery after spine surgery: A systematic review. *Neurosurg. Focus* **46** 1–8. https://doi.org/10.3171/2019.1.FOCUS18700

EMEDOM-NNAMDI, P., SMITH, T. R., ONNELA, J.-P. and LU, J. (2025). Supplement to "Nonparametric additive value functions: Interpretable reinforcement learning with an application to surgical recovery." https://doi.org/10.1214/24-AOAS1987SUPPA, https://doi.org/10.1214/24-AOAS1987SUPPB

FAN, J. and JIANG, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100** 890–907. https://doi.org/10.1198/016214504000001439

FARAHMAND, A., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.* **17** Paper No. 139, 66 pp. MR3555030

GEIST, M. and SCHERRER, B. (2011). l1-penalized projected Bellman residual. Technical Report No. 11.

GHAVAMZADEH, M., MANNOR, S., PINEAU, J. and TAMAR, A. (2016). Bayesian reinforcement learning: A survey. *Found. Trends Mach. Learn.* **8** 359–483. https://doi.org/10.1561/2200000049

GOTTESMAN, O., JOHANSSON, F., KOMOROWSKI, M., FAISAL, A., SONTAG, D., DOSHI-VELEZ, F. and CELI, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nat. Med.* **25** 16–18. https://doi.org/10.1038/s41591-018-0310-5

HASTIE, T. J. (2017). Generalized additive models. In *Statistical Models in S* 249–307. https://doi.org/10.1201/9780203738535-7

HOFFMAN, M. W., LAZARIC, A., GHAVAMZADEH, M. and MUNOS, R. (2012). Regularized least squares temporal difference learning with nested ℓ_2 and ℓ_1 penalization. In *Recent Advances in Reinforcement Learning. EWRL* 2011. *Lecture Notes in Computer Science* **7188**. Springer, Berlin. https://doi.org/10.1007/978-3-642-29946-9_13

HOWARD, M. and NAKAMURA, Y. (2013). Locally weighted least squares policy iteration for model-free learning in uncertain environments. In *IEEE International Conference on Intelligent Robots and Systems* 1223–1229. https://doi.org/10.1109/IROS.2013.6696506

HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York. MR0118514

JAENSSON, M., DAHLBERG, K. and NILSSON, U. (2019). Factors influencing day surgery patients' quality of postoperative recovery and satisfaction with recovery: A narrative review. *Perioper. Med.* **8** 3. https://doi.org/10.1186/s13741-019-0115-1

KARAS, M., MARINSEK, N., GOLDHAHN, J., FOSCHINI, L., RAMIREZ, E. and CLAY, I. (2020). Predicting subjective recovery from lower limb surgery using consumer wearables. *Digit. Biomark.* **4** 73–86. https://doi.org/10.1159/000511531

KOLTER, J. Z. and NG, A. Y. (2009). Regularization and feature selection in least-squares temporal difference learning. In *ACM International Conference Proceeding Series* **382**. https://doi.org/10.1145/1553374.1553442

KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative Q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020-December**. https://doi.org/10.48550/arxiv.2006.04779

LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963 https://doi.org/10.1214/009053607000000811

LAGOUDAKIS, M. G. and PARR, R. (2004). Least-squares policy iteration. *J. Mach. Learn. Res.* **4** 1107–1149. MR2125347 https://doi.org/10.1162/1532443041827907

LIU, G. and ONNELA, J.-P. (2021). Bidirectional imputation of spatial GPS trajectories with missingness using sparse online Gaussian process. *J. Amer. Med. Inform. Assoc.* **28** 1777–1784. https://doi.org/10.1093/jamia/ocab069

LU, J., KOLAR, M. and LIU, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *J. Amer. Statist. Assoc.* **115** 2084–2099. MR4189778 https://doi.org/10.1080/01621459.2019.1689984

MNIH, V., KAVUKCUOGLU, K., SILVER, D., GRAVES, A., ANTONOGLOU, I., WIERSTRA, D. and RIEDMILLER, M. (2013). Playing Atari with deep reinforcement learning. Preprint. Available at arXiv:1312.5602.

MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533. https://doi.org/10.1038/nature14236

O'DONOGHUE, B., OSBAND, I., MUNOS, R. and MNIH, V. (2017). The uncertainty Bellman equation and exploration. In 35*th International Conference on Machine Learning*, *ICML* 2018 **9** 6154–6173.

ONNELA, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* **46** 45–54. https://doi.org/10.1038/s41386-020-0771-3

OZKARA, G. O., OZGEN, M., OZKARA, E., ARMAGAN, O., ARSLANTAS, A. and ATASOY, M. A. (2015). Effectiveness of physical therapy and rehabilitation programs starting immediately after lumbar disc surgery. *Turk. Neurosurg.* **25** 372–379. https://doi.org/10.5137/1019-5149.JTN.8440-13.0

PADMANABHAN, R., MESKIN, N. and HADDAD, W. M. (2015). Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed. Signal Process. Control* **22** 54–64. https://doi.org/10.1016/J.BSPC.2015.05.013

PANDA, N., SOLSKY, I., HAWRUSIK, B., LIU, G., REEDER, H., LIPSITZ, S., DESAI, E. V., LOWERY, K. W., MILLER, K. et al. (2020a). Smartphone global positioning system (GPS) data enhances recovery assessment after breast cancer surgery. *Ann. Surg. Oncol.* **28** 985–994. https://doi.org/10.1245/s10434-020-09004-5

PANDA, N., SOLSKY, I., HUANG, E. J., LIPSITZ, S., PRADARELLI, J. C., DELISLE, M., CUSACK, J. C., GADD, M. A., LUBITZ, C. C. et al. (2020b). Using smartphones to capture novel recovery metrics after cancer surgery. *JAMA Surg.* **155** 123–129. https://doi.org/10.1001/jamasurg.2019.4702

PENG, X., DING, Y., WIHL, D., GOTTESMAN, O., KOMOROWSKI, M., LEHMAN, L.-W. H., ROSS, A., FAISAL, A. and DOSHI-VELEZ, F. (2018). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. *AMIA Annual Symp. Proc.* **2018** 887–896.

PEYSER, T., DASSAU, E., BRETON, M. and SKYLER, J. S. (2014). The artificial pancreas: Current status and future prospects in the management of diabetes. *Ann. N.Y. Acad. Sci.* **1311** 102–123. https://doi.org/10.1111/nyas.12431

RAGHU, A., KOMOROWSKI, M., CELI, L. A., SZOLOVITS, P. and GHASSEMI, M. (2017). Continuous state-space models for optimal sepsis treatment—A deep reinforcement learning approach 1–17.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

RICHTÁRIK, P. and TAKÁČ, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.* **144** 1–38. MR3179953 https://doi.org/10.1007/s10107-012-0614-z

RIEDMILLER, M. (2005). Neural fitted Q iteration—First experiences with a data efficient neural reinforcement learning method. In *Machine Learning*: *ECML* 2005. *Lecture Notes in Computer Science* **3720**. Springer, Berlin. https://doi.org/10.1007/11564096_32

SINZINGER, E. D. and MOORE, B. (2011). Sedation of simulated Icu patients using reinforcement learning based control. *Int. J. Artif. Intell. Tools* **14** 137–156. https://doi.org/10.1142/S021821300500203X

SONABEND-W, A., LU, J., CELI, L. A., CAI, T. and SZOLOVITS, P. (2020). Expert-supervised reinforcement learning for offline policy learning and evaluation. *Adv. Neural Inf. Process. Syst.* **2020-December**.

SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

TOROUS, J., STAPLES, P. and ONNELA, J.-P. (2015). Realizing the potential of mobile mental health: New methods for new data in psychiatry. *Curr. Psychiatry Rep.* **17** 602. https://doi.org/10.1007/s11920-015-0602-0

TZIORTZIOTIS, N. and DIMITRAKAKIS, C. (2017). Bayesian inference for least squares temporal difference regularization. Technical Report.

VAN HASSELT, H. (2010). Double Q-learning. *Adv. Neural Inf. Process. Syst.* **23**.

WAINWRIGHT, T. W., IMMINS, T. and MIDDLETON, R. G. (2016). Enhanced recovery after surgery (ERAS) and its applicability for major spine surgery. *Bailliere's Best Pract. Res., Clin. Anaesthesiol.* **30** 91–102. https://doi.org/10.1016/j.bpa.2015.11.001

WANG, Z., NOVIKOV, A., ZOLNA, K., SPRINGENBERG, J. T., REED, S., SHAHRIARI, B., SIEGEL, N., MEREL, J., GULCEHRE, C. et al. (2020). Critic regularized regression. *Adv. Neural Inf. Process. Syst.* **2020-December**.

XU, X., HU, D. and LU, X. (2007). Kernel-based least squares policy iteration for reinforcement learning. *IEEE Trans. Neural Netw.* **18** 973–992. https://doi.org/10.1109/TNN.2007.899161

YEZIERSKI, R. P. (2012). The effects of age on pain sensitivity: Preclinical studies. *Pain Med.* **13 Suppl 2** 27–36. https://doi.org/10.1111/j.1526-4637.2011.01311.x