# Semiparametric Information State Embedding for Policy Search under Imperfect Information

Sujay Bhatt, Weichao Mao, Alec Koppel, Tamer Başar

***Abstract*— We consider the problem of policy search in sequential decision making problems with imperfect information as encapsulated by a partially observed Markov Decision Process (POMDP) over possibly continuous state-spaces. In general, the optimal policy is history-dependent and the objective is non-convex in the policy parameters, making even stationary point policies challenging to ascertain. To address this problem class, we develop a constructive way to succinctly represent the history as an approximate *information state*, using Semiparametric Information State Embedding (SISE). SISE alternates between conditional kernel density estimation and fitting the parameters of an Echo State Networks (ESNs), a one-layer recurrent neural model. Based upon constructing SISE, we develop an actor-critic scheme for policy search over the approximate information states. Our main technical contributions are to (i) establish the convergence and generalization performance of SISE, and to (ii) derive the convergence to stationary points of our policy search scheme. Experimentally, our fusion of SISE and actor-critic yields favorable performance in practice on the canonical POMDPs of Tiger, LightDark, and a partially observed variant of CartPole.**

## I. Introduction

Partially Observed Markov Decision Processes (POMDPs) are a mathematical framework to address sequential decision making under imperfect information [1]. Consider an agent interacting with an environment that is only partially observable. The agent receives an observation $o_t \in \mathscr{O} \subseteq \mathbb{R}^{\mathrm{d}}$, which is a proxy for the underlying (hidden) state, chooses a control action $u_t \in \mathscr{U} \subseteq \mathbb{R}^{\mathrm{u}}$ at discrete time $t$, and in return, the environment reveals a reward $r_t \in \mathbb{R}$. The observation $o_t$ does not satisfy the Markov property, so an informed decision cannot be taken using only the observations. For such a non-Markovian system, it is necessary to select actions that depend on the history of observations and actions, $h_t := \{o_k, u_{k-1}, r_{k-1}\}_{0\le k\le t}$. The objective in such a sequential decision making problem is to maximize the expected discounted cumulative return, or value:

$$J_{opt} = \sup_{\pi} \ \mathbb{E}_{\pi}\Big[\sum_{j=0}^{\infty} \gamma^j \ r_j\Big]. \tag{1}$$

where $\gamma \in [0,1)$ is an economic discount factor. Let $\mathscr{F}_t := \sigma(\{o_k, u_{k-1}, r_{k-1}\}_{k\le t})$ denote the sigma algebra generated by history $h_t$. The policy of the agent is a mapping from the history to the distribution of actions, $\pi = \{\pi_k\}_{k\le t}$, such that $\pi_t : \mathscr{F}_t \to \mathbb{P}(\mathscr{A})$. Clearly, keeping track of the entire history $h_t$ for decision making becomes infeasible owing to memory and computational complexity.

In environments where the model is known and state-action spaces are finite, it is possible to maintain a distribution over the states called the *belief state*, which is sufficient for decision-making [1]. However, due to the fact that the complexity of a belief representation scales with the cardinality of the state space, it is infeasible for large or possibly continuous spaces. This motivates alternative low-dimensional representations that approximately summarize the history, and are sufficient for decision making. When such a representation of the history is (approximately) sufficient for decision making, we refer to it as an *Approximate Information State* (AIS).

A multitude of such representations are possible. One simple example of approximate information states is a fixed-memory window of the history [2], [3]. A quantitative relation between the approximation error and the window size in this method has been established very recently in [4]; and convergence of policy gradient methods using such representations is analyzed in [5]. Another example of state representation is the Predictive State Representation (PSR) [6], [7]. A PSR is a set of multi-step action-observation sequences such that knowing the results of these sequences is sufficient to predict the future observation of every possible action-observation sequence. A similar point of view aims to learn a causal state representation [8]. This method partitions histories into clusters using a metric that is related to the predicted subsequent observations, and then builds a causal graph of observations. A recent work [9] introduces a set of sufficient conditions for performance evaluation, and proves the overall performance degradation is bounded if these conditions are only approximately satisfied. Recurrent Neural Networks (RNNs) have also been widely utilized to learn approximate information states [10]–[12] due to the recent surge in using deep neural networks for reinforcement learning.

These methods demonstrate good empirical performances on various benchmark tasks, but generally do not have theoretical guarantees on the approximation error with respect to the true reward/observation likelihoods. See [13] for a detailed survey of state representation learning in control. Overall, then, there is an open question as to whether one may quantify the quality of a given construction of a distributional model for the posterior distribution over

Research supported in part by US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by Office of Naval Research (ONR) MURI Grant N00014-16-1-2710.

S. Bhatt is based in Seattle, USA. sujaybhatt.hr@gmail.com

W. Mao and T. Başar are with Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801. weichao2@illinois.edu, basar1@illinois.edu

A. Koppel is with Computational and Information Sciences Directorate, U.S. Army Research Laboratory, Adelphi, MD 20783. alec.e.koppel.civ@mail.mil

observations or rewards. This gap has made convergence of policy search in POMDPs mostly contingent on unverified conditions, to date.

**Main Contributions.** Based upon this gap, we develop joint search procedures for efficient representations of the history and policy optimization over these representations, the result of which is a conceptually grounded approach to policy search. Our approach builds upon the definition of an Approximate Information States (AIS) in [9] to contribute the following:

1) Unlike [9], [14], we provide a *constructive* algorithm (Algorithm 1) to build approximate information states by alternating between conditional kernel density estimation and fitting the parameters of an RNN, whose representational power we theoretically characterize (Theorems 2 - 4). This representation is *semi-parametric* in the sense that the conditional kernel density estimation is non-parametric, but is used as a target for fitting an RNN, which is a parametric representation.
2) We provide a policy optimization algorithm (Algorithm 2) based on the Actor-Critic methodology that operates over the approximate information states, and analyze the convergence properties of this algorithm (Theorem 1).
3) We show that the developed joint search procedure (alternate history representations & policy optimization over these representations) not only has theoretical guarantees, but empirically is competitive with [9].

## II. Approximate Information States

In this section, we begin by formalizing the notion of an Approximate Information State as any map of the history $h_t := \{o_k, u_{k-1}, r_{k-1}\}_{0\le k\le t}$ that exhibits two key properties [9]. Then, we shift to constructing Semiparametric Information State Embedding (SISE) using conditional kernel density estimates – whose bias as a function of the number of observation-action pairs processed may be rigorously characterized– and a neural network (Echo State Networks) based approximation of the low-dimensional history representation.

We proceed by more formally noting the two probabilistic attributes that must be satisfied by an AIS. The first is that it smoothly incorporates new information $(o_{k+1}, u_k, r_k)$ in a manner reminiscent of a posteriori updates, and the second is that the posterior distributions it defines over rewards or observations conditional on the history are nearly consistent.

**Definition 1.** *(Approximate Information State)* An Approximate Information State $\mathscr{I}_t = F(h_t) \in \mathscr{X} \subseteq \mathbb{R}^{\mathrm{a}}$ is any function of the history $h_t$ that satisfies the following properties [9]:

1) Consistency of Conditionally Expected Reward: For some $\varepsilon_r > 0$,
$$|\mathbb{E}[r_t|h_t, u_t] - \mathbb{E}[r_t|\mathscr{I}_t, u_t]| \le \varepsilon_r. \tag{2}$$
2) State-like Evolution and Near-unbiasedness of Observation Distribution:
(i) The evolution across time appends observation-action pairs: $\mathscr{I}_{t+1} = \mathscr{H}(\mathscr{I}_t, o_{t+1}, u_t, r_t)$, where $\mathscr{H}$ is a Lipschitz map in $(o, r)$.
(ii) The posterior distribution of $o_{t+1}$ conditioned on $\mathscr{I}_t$ and history $h_t$ as well as control $u_t$ are close, i.e., for some $\varepsilon_o > 0$, the conditional distributions
$$\mu(E) = \mathbb{P}(o_{t+1} \in E|h_t, u_t)\,, \upsilon(E) = \mathbb{P}(o_{t+1} \in E|\mathscr{I}_t, u_t), \tag{3}$$
are close $d(\mu, \upsilon) \le \varepsilon_o$ with respect to some metric over distributions $d(\cdot,\cdot)$.

For a given history, using the parameters of the POMDP, the belief distribution is a deterministic (Bayes' rule) mapping that leads to a unique distribution over the (underlying) states that is sufficient for decision making. Similarly, for the information state representation using an RNN/ LSTM, a given history is deterministically mapped to a unique representation sufficient for decision making. The probability measures in (3) are well defined via the Ionescu-Tulcea Theorem, see [15].

Approximate Information States are a generalization of the belief states to representations of history that satisfy belief-like (Definition 1) properties with a few key differences: (i) Information states need not lie on a compact space unlike belief states. (ii) The dimension of information states is a hyper-parameter that is tuned for low-cost decision making. For example, in the CartPole experiment (Sec.V), the state space is $\mathbb{R}^4$, making the belief space an intractable multivariate distribution.

### A. Conditional Predictive Density Estimation

In this subsection, we develop a non-parametric method of kernel density estimation to fit the conditional observation predictive distribution $\mathbb{P}(o|\mathscr{I}, a)$ and the reward distribution $\mathbb{P}(r|\mathscr{I}, u)$ using the approximate information state (AIS) $\mathscr{I}$ under the operating hypothesis that $\mathscr{I}$ is fixed. In Sec. II-B, we will shift to how to estimate $\mathscr{I}$.

To construct the aforementioned kernel density estimates, we require the ability to generate observation-reward tuples $\{o_t, r_t\}_{t\ge 0}$, which may be addressed under the hypothesis that we have access to a *generative model* or a simulator of the environment that provides observations and rewards in response to actions of the agent given AIS. Specifically, the agent is able to query an environment oracle, which provides the data $\{o_t, r_t\}_{t\ge 0}$. The generative model furnishes the data required to perform conditional predictive density estimation.

For convenience, let $f(\cdot|\mathscr{I}, u)$ denote the conditional density associated with distribution $\mathbb{P}(\cdot|\mathscr{I}, u)$. Let $N$ denote the number of unique actions and information states, i.e., a collection of tuples of the form $\{u_t, \mathscr{I}_t\}_{t\le N}$. For each action and approximate information state, $(u_t, \mathscr{I}_t)$, we generate $L$ distinct observations and rewards simulated for each action and approximate information state to form a data set of the form $\mathscr{D}_N^L := \{o_{lt}, r_{lt}\}_{t,l=1}^{N,L}$, meaning that there are $NL$ calls to the generative model (simulation oracle). Based upon this trajectory information, the sample estimates for the conditional density of observations $\hat{f}_N(o|\mathscr{I}, u)$ and rewards

$\hat{f}_N(r|\mathscr{I},u)$ take the following forms, respectively:

$$\hat{f}_N(o|\mathscr{I},u) = \frac{1}{Lb_o^d}\frac{\sum_{i=1}^N \mathbb{K}\left(\frac{\|u-u_i\|}{b_u}\right)\ \mathbb{K}\left(\frac{\|\mathscr{I}-\mathscr{I}_i\|}{b_{\mathscr{I}}}\right)\sum_{l=1}^L \mathbb{K}\left(\frac{\|o-o_{il}\|}{b_o}\right)}{\sum_{i=1}^N \mathbb{K}\left(\frac{\|u-u_i\|}{b_u}\right)\ \mathbb{K}\left(\frac{\|\mathscr{I}-\mathscr{I}_i\|}{b_{\mathscr{I}}}\right)} \quad (4)$$

$$\hat{f}_N(r|\mathscr{I},u) = \frac{1}{Lb_r}\frac{\sum_{i=1}^N \mathbb{K}\left(\frac{\|u-u_i\|}{b_a}\right)\ \mathbb{K}\left(\frac{\|\mathscr{I}-\mathscr{I}_i\|}{b_{\mathscr{I}}}\right)\sum_{l=1}^L \mathbb{K}\left(\frac{\|r-r_{il}\|}{b_r}\right)}{\sum_{i=1}^N \mathbb{K}\left(\frac{\|u-u_i\|}{b_u}\right)\ \mathbb{K}\left(\frac{\|\mathscr{I}-\mathscr{I}_i\|}{b_{\mathscr{I}}}\right)} \quad (5)$$

where the kernel $\mathbb{K}$ is any nonnegative function of its scalar argument, and $b_{\{.\}} > 0$ denotes the kernel bandwidth. The key parameters, i.e., sufficient statistics, in kernel density estimation are the data and the bandwidth parameters. The bandwidth parameters can be optimized using any of the procedures outlined in [16], [17].

### *B. Echo State Networks*

Echo State Networks (ESNs) integrate the history of observations, actions and rewards in its reservoir states, and can control the weight of the historical information [18], and are an elementary version of a recurrent neural network. The input to an ESN is observation-action-reward triple $(o_t, a_{t-1}, r_{t-1})$, and the history $h_t$ is summarized in the reservoir states $x_t \in \mathscr{X} \subseteq \mathbb{R}^{\mathrm{a}}$. The temporal evolution of such a network is governed by the following non-linear dynamical system

$$x_t = (1-\beta)x_{t-1} + \beta f_x\Big(W^x x_{t-1} + W^y \begin{bmatrix} o_t \\ u_{t-1} \\ r_{t-1} \end{bmatrix}\Big), \quad (6)$$

where $\beta \in [0,1]$ is the tunable leaking factor that controls the weight of historical information, and $W^x \in \mathbb{R}^{\mathrm{a}\times\mathrm{a}}$ and $W^y \in \mathbb{R}^{\mathrm{a}\times(\mathrm{d}+\mathrm{u}+1)}$ denote the weight matrices of the ESN. The vector function $f_x : \mathbb{R}^{\mathrm{a}} \to \mathbb{R}^{\mathrm{a}}$ is understood to act component-wise on its argument and is typically a sigmoid.
The ESN parameters that can be optimized are:

$$\delta = \{\beta, W^x, W^y\}. \quad (7)$$

The parameters $\delta$ in (6) are optimized so that the predictive reward distribution $\mathbb{P}(r_{t+1}|\mathscr{I}_t,u_t)$ and the predictive observation distribution $\mathbb{P}(o_{t+1}|\mathscr{I}_t,u_t)$ are close to the true values $\mathbb{P}(r_{t+1}|h_t,u_t)$ and $\mathbb{P}(o_{t+1}|h_t,u_t)$. This ensures that $\mathscr{I}_t$ satisfies Definition 1.

**Assumption 1.** The conditional kernel densities $f(o|u,\mathscr{I})$ and $f(r|u,\mathscr{I})$ are bounded and non-zero with probability 1:

$$0 < \xi_0 \le |f(o|u,\mathscr{I})| \le 1, \text{ and } 0 < \xi_r \le |f(r|u,\mathscr{I})| \le 1. \quad (8)$$

Assumption 1 is satisfied by any kernel whose operator norm decays sufficiently quickly, i.e., the associated kernel matrices exhibit eigenvalue decay, as with Gaussian and polynomial kernels.

**Remark**: *With a slight abuse of notation, $\mathscr{I}(\delta)$ is used to denote that the reservoir state/AIS propagated using the echo state parameters $\delta$, and data hence collected is denoted as $\mathscr{D}_N^L(\delta_\tau)$. Note that $\mathscr{I}(\delta^*)$ is the AIS, but we use reservoir state/AIS interchangeably to simplify exposition.*

**Algorithm 1** **S**emiparametric **I**nformation **S**tate **E**mbedding (SISE)

**Input:** Number of evaluative actions $N$, the number of observations (and rewards) per action $L$, initial parameters of Echo State Network $\delta_0$, threshold parameter $\vartheta_I > 0$. Information state simulation parameters $\widehat{\mu}, \widehat{\sigma}$.
  **While** $|\delta_{\tau+1} - \delta_\tau| > \vartheta_I$ **do**:
    Generate AIS $\mathscr{I}_0 \in \mathbb{R}^{\mathrm{x}}$ from $\mathscr{N}(\widehat{\mu}, \widehat{\sigma})$. *// This is any arbitrary distribution.*
    Collect $\mathscr{D}_N^L(\delta_\tau) = \{(o_{n,l}, r_{n,l})\}_{l=1,n=1}^{L,N}$ using $NL$ calls to the simulator.
    Using $\mathscr{D}_N^L(\delta_\tau)$, compute $\hat{f}_N(o \mid \mathscr{I}(\delta_\tau), u)$ and $\hat{f}_N(r \mid \mathscr{I}(\delta_\tau), u)$ for $\delta_\tau$ [cf. Sec.II-A]
    Using $\hat{f}_N(o \mid \mathscr{I}(\delta_\tau), u)$ and $\hat{f}_N(r \mid \mathscr{I}(\delta_\tau), u)$, solve (12) to obtain $\delta_{\tau+1}$, where
      $\delta_{\tau+1} = \mathrm{argmax}_\delta \mathscr{L}(\delta)$ [cf. (11)].
    $\tau = \tau + 1$.
  **End**
**Output:** the converged echo state parameters $\delta_{\tau^*}$, where $|\delta_{\tau^*+1} - \delta_{\tau^*}| \le \vartheta_I$.

Subsequently, we reinterpret $\mathscr{I}_t = \mathscr{I}_t(\delta)$ as being parameterized by $\delta$, the echo-state parameters (7), which we seek to fit in a maximum likelihood fashion to the density estimates defined in Sec.II-A. Specifically, we consider the following bounded loss functions $\mathscr{L}_r \in [0,1]$ and $\mathscr{L}_o \in [0,1]$ that respectively quantify the closeness to the reward and observation densities:

$$\mathscr{L}_r(\delta) = \frac{-1}{\xi_r \cdot NL}\sum_{m=1,l=1}^{N,L} \log(\hat{f}_N(r_{ml}|\mathscr{I}_m(\delta), u_m)), \quad (9)$$

$$\mathscr{L}_o(\delta) = \frac{-1}{\xi_o \cdot NL}\sum_{m=1,l=1}^{N,L} \log(\hat{f}_N(o_{ml}|\mathscr{I}_m(\delta), u_m)). \quad (10)$$

Thus, (9) approximates the KL-divergence between empirical and the true density $f(r|u,h)$. Moreover $r_m, o_m$ denote the observed reward and observation from the simulator at time $m$. We propose the composite loss function

$$\mathscr{L}(\delta) = \frac{1}{2}(\mathscr{L}_o(\delta) + \mathscr{L}_r(\delta)) \quad (11)$$

that accounts for both the fitness of reward and observation distributions. We then seek to minimize it over the empirical data set $\mathscr{D}_N^L$,

$$\delta^* = \mathrm{argmax}_\delta \mathscr{L}(\delta)\,, \quad (12)$$

which yields parameters that minimize the KL-divergence between the approximate and true predictive distributions. Reservoir states associated with the ESN propagated across time using $\delta^*$ in (12) define the AIS we consider.

### C. Alternating Minimization for Approximate Information States

In this section, we assemble the pieces of the previous subsections into an alternating minimization procedure to compute the parameters of the Echo State Network that map any history into approximate information states. This mapping facilitates (approximately) optimal sequential decision making using a policy parametrization to be subsequently defined.

The resulting procedure which we call **S**emiparametric **I**nformation **S**tate **E**mbedding (SISE), is summarized as Algorithm 1. The algorithm operates by alternating conditional kernel density estimates over a batch of $N \times L$ samples – $L$ observation-reward pairs $(r_{il}, o_{il})_{l=1}^{L}$ are simulated upon for a fixed action $u_i$ and AIS $\mathscr{I}_i(\delta)$, where the action is selected using any arbitrary distribution over $\mathscr{A}$, and AIS is propagated using fixed ESN parameters $\delta$ starting from $\mathscr{I}_0$. We assume that the sampled observations are also identically distributed. This can be achieved, for example, for each $\mathscr{I}_n(\delta)$, by repeatedly applying randomly chosen $u_n$ (from arbitrary distribution over $\mathscr{A}$) for duration $L$ to get a sequence of observations and rewards. Over this fixed batch of data $\mathscr{D}_N^L(\delta) = (r_{il}, o_{il})_{i=1,l=1}^{N,L}$, we alternate between solving for the conditional kernel density and the ESN parameters (12) that define the AIS. This process is repeated for steps $\tau = 1, 2, \ldots$ until the termination criterion $|\delta_{\tau+1} - \delta_\tau| \leq \vartheta_I$ is satisfied. The output is an AIS $\mathscr{I}(\delta)$ which may be rigorously shown to satisfy the properties of Definition 1, as we detail in subsequent sections. Before doing so, we expand upon how the AIS may be employed for policy search.

## III. Policy Search using Approximate Information States

With our construction of the approximate information state (Def. 1) $\mathscr{I}$ crystallized, we now shift to how this scheme may be incorporated into making policy search efficient in terms of the parameter dimension. That is, whereas in a general POMDP, the policy must be defined over the sigma algebra $\mathscr{F}_t$ (of dimensionality $d_t$), instead we may now define it only over the approximate information state $\mathscr{I} \in \mathscr{X} \subset \mathbb{R}^a$ which is of fixed dimension $a$ (see the beginning of Sec. II).

More specifically, we consider policy parameterizations of the form $\pi : \mathscr{X} \to \mathbb{P}(\mathscr{U})$, which map the approximate information state $\mathscr{I}$ to control decisions selected as $u_t \sim \pi(\cdot|\mathscr{I}_t)$. We subsequently parameterize the policy by a *fixed-dimensional* parameter vector $\theta \in \mathbb{R}^p$. So, a policy $\pi_\theta : \mathscr{X} \to \mathbb{P}(\mathscr{U})$ is a parametric function of the approximate information state $\mathscr{I}$. An example parametrization is $\pi_\theta(\cdot|\mathscr{I}) = \mathscr{N}(\theta^T \mathscr{I}; \sigma)$, where $\mathscr{N}$ is a Gaussian distribution with fixed known variance $\sigma^2$. Here the dimension $a$ of information state equals the dimension $p$ of the policy parameters.

This parameterization allows us to define policy search schemes typical of (fully observable) Markov Decision Processes (MDPs), but whose state space is the Approximate Information State. Before doing so, we make precise the implications of the AIS being "sufficient for decision-making."

**Algorithm 2** Actor-Critic using Approximate Information States

**Input:** Initialization parameters $\lambda_0 \in \mathbb{R}^v$ and $\theta_0 \in \mathbb{R}^p$, $\mathscr{I}_0 \in \mathbb{R}^a$, and step-sizes $\{\alpha_t, \beta_t\}_{t \geq 0}$.
**For** $t = 1, 2, \cdots$ **do**:
  Propagate $\mathscr{I}_t(\delta^*)$ using $\delta^*$ obtained from Algorithm 1, take action $u \sim \pi_{\theta_t}(\cdot|\mathscr{I}_t)$.
  Compute reward $R(\mathscr{I}_t, u_t)$ and the next state $\mathscr{I}_{t+1}(\delta^*)$ using (6).
  Compute Temporal Difference (TD)-error

$$\Delta_t = R(\mathscr{I}_t, u_t) + \gamma \cdot V_{\lambda_t}(\mathscr{I}_{t+1}) - V_{\lambda_t}(\mathscr{I}_t) \tag{13}$$

  Update the parameters as:

$$\theta_{t+1} = \theta_t + \alpha_t \cdot \Delta_t \cdot \nabla_\theta \log \pi_{\theta_t}(u_t|\mathscr{I}_t). \tag{14}$$

$$\lambda_{t+1} = \lambda_t + \beta_t \cdot \Delta_t \cdot \nabla_\lambda V_{\lambda_t}(\mathscr{I}_t). \tag{15}$$

**End**
**Output:** Sequences $\{\theta_t\}_{t \geq 0}$ and $\{\lambda_t\}_{t \geq 0}$.

As formalized in [9, Theorem 3], given an AIS that satisfies Def. 1, one may write the Bellman optimality equation as follows. Let $\mathscr{J}(\cdot) : \mathscr{X} \to \mathbb{R}$ denote the value function on the AIS space, i.e.

$$\mathscr{J}(\mathscr{I}) = \sup_{u \in \mathscr{U}} \Big\{ R(\mathscr{I}, u) + \gamma \int_{o \in \mathscr{O}, r \in \mathbb{R}} \mathbb{P}(o, r|\mathscr{I}, u)\, \mathscr{J}(\mathscr{H}(\mathscr{I}, u, o, r)) \Big\}, \tag{16}$$

where $R(\mathscr{I}, u) = \mathbb{E}[r|\mathscr{I}, u]$, and $\mathscr{H}(\mathscr{I}_t, o_{t+1}, u_t, r_t) = \mathscr{I}_{t+1}$.

Here $r$ denotes the reward output of the simulator and $\mathscr{H}(\cdot)$ is the next AIS mapping which is described by the non-linear dynamical system update of ESN in (6). It was shown in [9, Theorem 5] that $\mathscr{J}$ in (16) is close to $J_{opt}$ in (1). This implies that, once Echo State Network parameters are tuned, one can invoke policy search algorithms for reinforcement learning problems to compute the optimal policy of the imperfect information problem.

Next, we will include the convergence result for Actor-Critic on continuous state-action spaces. As before, let $\{\pi_\theta : \theta \in \mathbb{R}^p\}$ denote the family of parameterized policies, and let $V_\lambda : \lambda \in \mathbb{R}^v$ denote the parametrized family of value functions. Define

$$g(\lambda, \theta) := \mathbb{E}\Big\{ \Big[ R(\mathscr{I}, u) + \gamma \cdot V_\lambda(\mathscr{I}') - V_\lambda(\mathscr{I}) \Big] \nabla_\lambda V_\lambda(\mathscr{I}) \Big\}, \tag{17}$$

where $\mathbb{E} := \mathbb{E}_{\mathscr{I} \sim d_{\pi_\theta}(\mathscr{I}),\ u \sim \pi_\theta(\cdot|\mathscr{I})}$, $\mathscr{I}'$ denotes the next state, and $d_{\pi_\theta}(\mathscr{I}) := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(\mathscr{I}_t = \mathscr{I}|\pi_\theta)$ is the information state-occupancy measure of policy $\pi_\theta$. Here $V_\lambda(\mathscr{I})$ denotes the critic approximated using non-linear function approximation, and $\pi_\theta(\cdot|\mathscr{I})$ denotes the actor policy.

**Assumption 2.**

- For any $\mathscr{I} \in \mathbb{R}^a$ and $\lambda \in \mathbb{R}^v$, let $|V_\lambda(\mathscr{I})| < \infty$, $\|V_\lambda(\mathscr{I})\|_2 < \infty$, and $\nabla_\lambda V_\lambda(\mathscr{I})$ is Lipschitz continuous.

Also, $\|\nabla_\theta \log \pi(a|\mathscr{I})\|_2 < \infty$ for all $(u,\mathscr{I}) \in \mathscr{A} \times \mathbb{R}^v$ and $\theta \in \mathbb{R}^p$.

- For function $g(\lambda,\theta)$ in (17), for each $\theta$, the ODE $\dot{\lambda}(t) = g(\lambda(t),\theta)$ has a local asymptotically stable equilibrium in $K(\theta)$,

$$K(\theta) := \arg\min_{\lambda} \|V_\lambda - \mathscr{T}^{\pi_\theta} V_\lambda\|^2_{d_{\pi_\theta}(\mathscr{I})}, \qquad (18)$$

where $\mathscr{T}^{\pi_\theta}$ is the Bellman operator.

- Step-sizes for the actor $\sum \alpha_t = \infty$ and $\sum \alpha_t^2 < \infty$, critic $\sum \beta_t = \infty$ and $\sum \beta_t^2 < \infty$, and time-scale $\lim_{t\to\infty} \frac{\alpha_t}{\beta_t} = 0$.

**Theorem 1** (Actor-Critic)**.** *Let the ESN parameters be obtained as in* (12)*. Suppose Assumption 2 holds. We have, using Algorithm 2, $\{\lambda_t,\theta_t\} \to \{\lambda^*,\theta^*\}$ w.p.1, where $\lambda^* \in K(\theta^*)$ in (18).*

Here $\mathscr{J}_\theta(\mathscr{I};\lambda^*) := V_{\lambda^*}(\mathscr{I})$ denotes the value function on the AIS space with policy parameters $\theta$, and $\theta^* \in \left\{\theta : \nabla_\theta \mathscr{J}_\theta(\mathscr{I};\lambda^*) = 0\right\}$ denotes the policy parameters corresponding to stationary points of the value function. Theorem 1 essentially says that on the faster time-scale the parameters of the value function converge to the stable equilibrium points of the related ODE, while the policy parameters converge to stationary points on the slower time-scale. The proof uses arguments in [9] to map the POMDP to (approximate) Information State MDP, and then perform two-time scale convergence analysis of Actor-Critic for continuous MDP similar to [19]. We further note that non-asymptotic analysis of Algorithm 2 using either lock in probability [20], [21] or supermartingales [22] remains an open problem due to the interaction between the estimation steps involving the approximate information state, the actor, and the critic.

## IV. Convergence Analysis

In this section, we rigorously establish the performance of Algorithm 1 as satisfying Definition 1. The analysis proceeds in two phases: first we establish the attenuation of the bias of the conditional kernel density estimates as a function of sample size $N$. Then, based upon this sub-sampling error, we are able to characterize the bias defined by the echo-state parameters generated by Algorithm 1 with respect to the true predictive distributions of the observation and reward.

To proceed with the consistency of the conditional density in (4), we fix the metric over distributions employed to quantify their distance from the true densities as the $L_1-$ distributional difference defined as follows

$$\int\int |\hat{f}_N(o|\mathscr{I},u) - f(o|\mathscr{I},u)| do \mathbb{P}(du \cdot d\mathscr{I}) \qquad (19)$$

For simplicity of notation, denote the input feature vector by $\Phi = [\mathscr{I},u] \in \mathscr{X} \times \mathscr{U}$.

**Remark**: *Below we will describe the approximation for the observation predictive distribution. The methodology carries over for the reward distribution.*

Due to space limitations, the proofs are made available in an extended technical report associated with this work [23]. Next, we specify the technical conditions under which the AIS is consistent.

**Assumption 3.** For $C_1 > 0$, $\rho_1 \in (0,1]$, and $\Phi_{(\cdot)} \in \mathscr{X} \times \mathscr{U}$, the predictive observation density $f(o|\Phi)$ is such that

$$\int_{\mathbb{R}} |f(o|\Phi_1) - f(o|\Phi_2)| \le C_1 \|\Phi_1 - \Phi_2\|^{\rho_1} \qquad (20)$$

**Assumption 4.** The observation predictive densities are such that the joint is a product of the marginals, i.e., $f(\varpi|\Phi) = \prod_{i=1}^d f(\varpi_i|\Phi)$, where $\varpi = [\varpi_1,\varpi_2,\cdots,\varpi_d] \in \mathbb{R}^d$. The marginal predictive observation densities $f(\cdot|\Phi)$ are Hölder-continuous with exponent $\rho_2 \in (0,1]$, i.e,

$$|f(\varpi_1|\Phi) - f(\varpi_2|\Phi)| \le C_2 \cdot |\varpi_1 - \varpi_2|^{\rho_2} \qquad (21)$$

for $C_2 > 0$, $\varpi_{(\cdot)} \in \mathbb{R}$, and $\Phi \in \mathscr{X} \times \mathscr{U}$.

**Assumption 5.**

- The kernel $\mathbb{K}$ has finite integral and moments over its scalar domain $v \in \mathbb{R}$. $\int_\infty^\infty \mathbb{K}^2(v)\ dv < \infty$ and $\int_{\mathbb{R}} \mathbb{K}(v) \cdot |v|^{\rho_2} < \infty$ where $\rho_2$ is the parameter in (21).
- Moreover, its point-wise evaluations are finite $\sup_{v\in\mathbb{R}} |\mathbb{K}(v)| < \infty$ .

These assumptions are standard in establishing the consistency of kernel density estimates in the multi-variate input case [24]. Assumption 3 imposes conditions on the underlying ground truth density we are estimating when conditioned on different points. Assumption 4 imposes continuity conditions on the ground-truth density with respect to its input argument, meaning that the likelihood may not vary arbitrarily for nearby points. The condition on the joint distribution factorizing into marginals is standard for vector-valued kernel density estimation. A conditional density satisfying Assumptions 3-5 is called *regular*.

Under these conditions, we have the following posterior contraction rate for conditional kernel density estimate associated with the observations over hidden states.

**Theorem 2.** *Suppose the conditional density $f(o|\Phi)$ is regular. For all $N \in \mathbb{N}$, there exists a constant $C > 0$ such that the following is true for every $\Phi \in \mathscr{X} \times \mathscr{U}$:*

$$\mathbb{E}\Big(\int\int |\hat{f}_N(o|\Phi) - f(o|\Phi)| do \mathbb{P}(d\Phi)\Big) \le C \cdot N^{\frac{-\rho_1}{\rho_1+(u+a)}} \qquad (22)$$

*where $\rho_1 \in (0,1]$ is such that*

$$\int_{\mathbb{R}} |f(o|\Phi_1) - f(o|\Phi_2)| \le C_1 \|\Phi_1 - \Phi_2\|^{\rho_1},$$

*for some $C_1 > 0, \Phi_1, \Phi_2 \in \mathscr{X} \times \mathscr{U}$.*

Theorem 2 provides the consistency and rate of convergence of the conditional density estimate, when the error of the density estimate is measured by the $L_1$-error, as a function of the sample size. The proof uses analysis similar to [24].

Next we shift to analyzing the approximate information state which is fit to the target of the conditional kernel density estimates.

**Theorem 3.** *Let $\delta \in \Delta$ denote the parameters of the Echo State Network, where $\Delta$ is the compact parameter space.*

*Let $\tau^*$ denote the termination time of Algorithm 1 for a given threshold parameter $\vartheta_I$. Let $\mathscr{L}$ denote the loss function in (11). As the threshold $\vartheta_I \to 0$, the following holds for the alternating minimization in Algorithm 1:*

$$\delta_{\tau^*} = argmax_{\delta} \mathscr{L}(\delta) = \delta^*. \tag{23}$$

Theorem 3 provides a consistency result for Algorithm 1 using Echo State Networks. The proof uses convergence analysis of alternating minimization procedures similar to [25]. Theorem 2 and Theorem 3 are necessary to characterize the errors using the densities conditioned on the approximate information states, which is formalized next. We continue by first defining the minimal model fitness of the AIS as follows.

**Definition 2** (Minimum Risk)**.** *Let $\mathscr{L} \in \mathbb{L}$, where $\mathbb{L}$ denotes the class of loss functions. With $\mathbb{E}(\cdot)$ denoting the expectation w.r.t data distribution, we define minimum risk of* (11) *as*

$$\mathtt{R}^*(\mathbb{L}) = \inf_{\mathscr{L} \in \mathbb{L}} \mathbb{E}(\mathscr{L}). \tag{24}$$

**Theorem 4.** *Suppose the conditional density $f(o|\Phi)$ is regular. There is an $\varepsilon_o(N, \delta^*)$ such that the following holds:*

$$d_{TV}(f(o|\mathscr{I}, u), f(o|h, u)) \leq \varepsilon_o(N, \delta^*). \tag{25}$$

*Letting $g = \min\{\frac{1}{\sqrt{2}}, \frac{\rho_1}{\rho_1 + (a+u)}\}$, w.prob at least $1-\gamma$,*

$$\varepsilon_o(N, \delta^*) = \sqrt{\mathtt{R}^*(\mathbb{L})} + O(\frac{1}{N^g}) \,, \tag{26}$$

*that is, $\lim_{N\to\infty} \varepsilon_o(N, \delta^*) = \sqrt{\mathtt{R}^*(\mathbb{L})}$.*

Theorem 4 provides an explicit characterization of the error in approximating the history using approximate information states. The proofs of all results are made available in an extended technical report associated with this work [23].

## V. Numerical Experiments

In this section, we evaluate the empirical performances of algorithms developed for solving POMDP problems, which constitute an important class of sequential decision making problems under imperfect information. We compare our performance with two baselines: the AIS algorithm proposed in [9] and the vanilla Actor-Critic algorithm. AIS uses a Gated Recurrent Unit network [26] (or GRU, a variant of recurrent neural networks) to learn an approximate information state, and then utilizes the REINFORCE algorithm [27] for policy search. The two-part neural network is trained in an end-to-end fashion, which is generally believed to yield better empirical performance but on the other hand prohibits any theoretical analysis. AIS can be considered as a state-of-the-art algorithm for policy search in POMDPs; however, the RNN employed there is heuristically assumed to satisfy Definition 1 whereas Algorithm 1 rigorously does so as formalized in preceding sections. As for the vanilla Actor-Critic baseline, we simply treat the observation of the agent as the state of the system, without devoting any additional effort to take care of the partial observability. This scheme inevitably leads to sub-optimal performances, but can still be used as a meaningful comparison baseline.

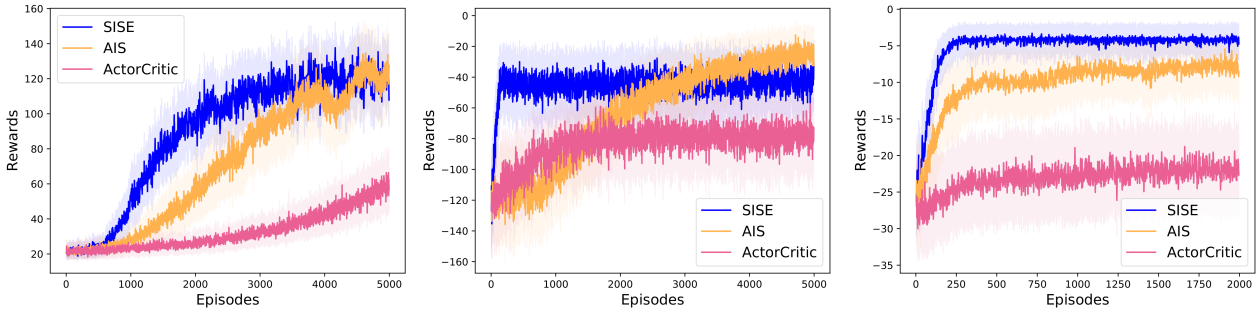
We evaluate the performances of the algorithms on three POMDP tasks: (i) CartPole [28] with superficial disturbance in the observations, (ii) Tiger [29], and (iii) LightDark [30], which are next described in detail.

**Task Descriptions.** In the original CartPole task, a pole is attached to a cart through a joint. The task is to apply an appropriate horizontal force to the cart to prevent the pole from falling. The observation in this task is 4-dimensional, and each of the 4 dimensions represents, respectively, the cart position, the cart velocity, the pole angle, and the pole angular velocity. In our experiments, we modify this task into a partially observable one by manually removing the cart velocity from the observations, leading to a 3-dimensional observation space. A unit reward is given for every time step that the pole remains standing (up to 15 degrees from being vertical).

In Tiger, the agent stands in front of two closed doors and decides which door to open. Behind one door is a tiger with a negative reward and behind the other is a treasure. Instead of opening a door immediately, the agent can also choose to listen for the tiger noises and locate the tiger, but listening is neither accurate nor free (a small negative reward is given). This task is inherently partially observable because the agent never has a direct observation of the environment state (i.e., the location of the tiger).

In LightDark, the agent has to locate itself in the plane along its way to reach a certain destination. The agent has a noisy sensor whose accuracy depends on the brightness of its current location. The illumination condition varies spatially in the plane, and the agent needs to move to a brighter place first to be able to better localize itself. A unit negative reward is given for every time step that the agent remains away from the destination by a certain radius. In our simulations, we use a one dimensional variant of the original LightDark task. Again, this task is inherently partially observable because the agent never obtains an accurate estimation of the state (its location).

**Parameter Selection.** For SISE, Algorithm 1 is implemented using an Echo State Network (ESN) with a hidden layer of size 16. We set the leaking factor in (6) to be $\beta = 0.2$, and the number of calls to the simulator as $N = 3000$ and $L = 200$. We use the Smooth L1-Norm as the kernel function $\mathbb{K}$. The learned information states are then fed into an Actor-Critic algorithm for policy search. The actor is instantiated using a two layer fully connected neural network with a hidden layer of 16 neurons. The output size of the actor network is equal to the dimension of the action space. This output can be interpreted as the mean vector of a multivariate normal distribution, where the action to be taken is sampled from this distribution with an identity covariance matrix. The value network is a two layer fully connected network with a hidden size of 16 and an output size of 1. We iterate between estimating the information states (Algorithm 1) and policy search (Algorithm 2) for 5 times to ensure that the policy used for sample collection is consistent with our behavioral policy. We use the ADAM optimizer [31] with a learning rate 0.05 for the ESN training, ADAM (0.005) for




(a) CartPole (b) Tiger (c) LightDark

Fig. 1. Rewards of the three algorithms on (a) CartPole, (b) Tiger, and (c) LightDark, respectively. Shaded areas denote the standard deviations of rewards. Note that SISE attains competitive performance with the benchmarks as the number of episodes increases, and in some cases outperforms them, as quantified by the accumulated reward.

the actor network, and ADAM (0.05) for the value network. All activations used in our simulations, including $f_x$ in (6), are Rectified Linear Unit (ReLU) [32] functions. The AIS algorithm is implemented in the same way as defined in [9], with one difference that we replace the categorical sampling of actions with a multivariate normal sampling to handle the continuous action space. The vanilla Actor-Critic algorithm is implemented similarly as the policy search part in SISE, except that it uses the present observation as the input to the policy network instead of a learned information state.

The simulation results are presented in Figure 1. The length of each episode is set to be 200 in CartPole, 20 in Tiger, and 40 in LightDark, as per the individual structure of each task. All results are averaged over 40 runs. We can see that SISE nearly matches the performance of AIS on the tasks of CartPole and Tiger, and outperforms AIS on LightDark. We believe that the performance difference in LightDark is due to the fact that LightDark is a continuous-action task, but AIS was not originally designed for continuous action spaces. Compared to the other two tasks, an additional challenge in LightDark is to explore the continuous action space more efficiently, which SISE handles well because it is designed to proactively collect samples in the first phase (Algorithm 1). The vanilla Actor-Critic algorithm achieves low rewards in all the three tasks. This is an expected result because it is not explicitly designed to handle partial observability. This result also illustrates the point that simply treating the partial observations as information states is generally not enough in partially observable environments, and a certain type of memory or abstraction of the history is necessary.

We would also like to remark that our approach permits larger learning rates than AIS, as can be seen from Figure 1 that SISE converges faster on all the three tasks. With a larger learning rate, a learning algorithm converges faster, but this may cause instability in training. By contrast, smaller learning rates yield improved limiting performance. Through exhaustive experimentation, we manually-tuned the learning rate of each approach for its individualized optimal performance. SISE allows for larger learning rates than AIS because upon beginning policy search (Algorithm 2), the approximate information state (Algorithm 1) has already converged for SISE, which allows training the policies more aggressively. By contrast, since the information states and policies are estimated simultaneously in the AIS approach, errors in information states may detrimentally influence policy search unless the step-size is sufficiently small.

## VI. Conclusion

We considered the problem of policy search in continuous state-action spaces under imperfect information. As maintaining belief states is computationally intractable, we provided an alternate approach, using approximate information states, to sequentially adapt representations of history that are learnable, tractable and sufficient for decision making. This representation is semi-parametric in the sense that the conditional kernel density estimation is non-parametric, but is used as a target for fitting a recurrent neural network (RNN), which is a parametric representation. For this approximate information state MDP, we provided an Actor-Critic algorithm and analyzed the convergence properties. We also showed that the developed joint search procedure (alternate history representations & policy optimization over these representations) not only has theoretical guarantees, but empirically is competitive with the state-of-the-art.

It is of interest to consider the non-asymptotic analysis of alternating minimization based algorithms developed in this paper. Also, we considered the convergence of non-convex policy search procedures to stationary points only. It is of interest to establish the global convergence of joint search procedures.

## References

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[2] C. C. White and W. T. Scherer, "Finite-memory suboptimal design for partially observed Markov decision processes," *Operations Research*, vol. 42, no. 3, pp. 439–455, 1994.

[3] C. Amato, D. S. Bernstein, and S. Zilberstein, "Optimizing fixed-size stochastic controllers for pomdps and decentralized pomdps," *Autonomous Agents and Multi-Agent Systems*, vol. 21, no. 3, pp. 293–320, 2010.

[4] A. D. Kara and S. Yuksel, "Near optimality of finite memory feedback policies in partially observed Markov decision processes," *arXiv preprint arXiv:2010.07452*, 2020.

[5] K. Azizzadenesheli, Y. Yue, and A. Anandkumar, "Policy gradient in partially observable environments: Approximation and convergence." *arXiv preprint arXiv:1810.07900*, 2020.

[6] M. L. Littman and R. S. Sutton, "Predictive representations of state," in *Proc. NeurIPS*, 2002, pp. 1555–1561.

[7] N. Jiang, A. Kulesza, and S. Singh, "Improving predictive state representations via gradient descent," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 1709–1715.

[8] A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello, "Learning causal state representations of partially observable environments," *arXiv preprint arXiv:1906.10437*, 2019.

[9] J. Subramanian and A. Mahajan, "Approximate information state for partially observed systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 1629–1636.

[10] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," *arXiv preprint arXiv:1507.06527*, 2015.

[11] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson, "Deep variational reinforcement learning for pomdps," *arXiv preprint arXiv:1806.02426*, 2018.

[12] A. Baisero and C. Amato, "Learning internal state models in partially observable environments," in *Reinforcement Learning under Partial Observability, NeurIPS Workshop*, 2018.

[13] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview," *Neural Networks*, vol. 108, pp. 379–392, 2018.

[14] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *arXiv preprint arXiv:2010.08843*, 2020.

[15] O. Hernández-Lerma and J. B. Lasserre, *Further topics on discrete-time Markov control processes*. Springer Science & Business Media, 2012, vol. 42.

[16] D. M. Bashtannyk and R. J. Hyndman, "Bandwidth selection for kernel conditional density estimation," *Computational Statistics & Data Analysis*, vol. 36, no. 3, pp. 279–298, 2001.

[17] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer Science & Business Media, 2012.

[18] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 659–686.

[19] Z. Yang, Z. Fu, K. Zhang, and Z. Wang, "Convergent reinforcement learning with function approximation: A bilevel optimization perspective," 2018.

[20] G. Dalal, B. Szorenyi, and G. Thoppe, "A tale of two-timescale reinforcement learning with the tightest finite-time bound," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3701–3708.

[21] T. Xu, Z. Wang, and Y. Liang, "Improving sample complexity bounds for (natural) actor-critic algorithms," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[22] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *arXiv preprint arXiv:1910.08412*, 2019.

[23] S. Bhatt, W. Mao, A. Koppel, and T. Başar, "Semiparametric information state embedding for policy search under imperfect information," *U.S. Army Research Laboratory/University of Illinois Technical Report*, 2020. [Online]. Available: https://koppel.netlify.app/assets/papers/2021_sise_report.pdf

[24] A.-K. Bott and M. Kohler, "Nonparametric estimation of a conditional density," *Annals of the Institute of Statistical Mathematics*, vol. 69, no. 1, pp. 189–214, 2017.

[25] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2049–2073, 2005.

[26] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[27] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[28] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 5, pp. 834–846, 1983.

[29] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman, "Acting optimally in partially observable stochastic domains," in *AAAI*, vol. 94, 1994, pp. 1023–1028.

[30] R. Platt Jr, R. Tedrake, L. Kaelbling, and T. Lozano-Perez, "Belief space planning assuming maximum likelihood observations," 2010.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[33] B. Hajek and M. Raginsky, *Statistical Learning Theory*. University of Illinois at Urbana-Champaign (course notes), 2018.

[34] S. Kakade and A. Tewari, *Learning Theory*. University of Illinois at Urbana-Champaign (course notes), 2008. [Online]. Available: https://ttic.uchicago.edu/~tewari/lectures/lecture15.pdf

[35] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

[36] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song, "Sbeed: Convergent reinforcement learning with nonlinear function approximation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1125–1134.

SUPPLEMENTARY MATERIAL FOR
"SEMIPARAMETRIC INFORMATION STATE EMBEDDING
FOR POLICY SEARCH UNDER IMPERFECT INFORMATION"

## VII. DEFINITIONS

**Definition 3** (Minimum Risk)**.** *Let $\mathscr{L} \in \mathbb{L}$, where $\mathbb{L}$ denotes the class of loss functions. With $\mathbb{E}(\cdot)$ denoting the expectation w.r.t data distribution, we define minimum risk of* (11) *as*

$$\mathtt{R}^*(\mathbb{L}) = \inf_{\mathscr{L}\in\mathbb{L}} \mathbb{E}(\mathscr{L}). \tag{27}$$

**Definition 4** (Rademacher Average)**.** *Let $Q \subset \mathbb{R}^N$ with $Q$ bounded. The Rademacher Average of $\mathscr{Q}$ denoted by $\mathscr{R}_N(\mathscr{Q})$ is given as*

$$\mathscr{R}_N(\mathscr{Q}) = \mathbb{E}\Big(\sup_{q\in\mathscr{Q}} \Big|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i q_i\Big|\Big), \tag{28}$$

*where $\varepsilon_i = \{-1,+1\}$ are discrete uniform random variables.*

## VIII. PROOFS OF MAIN RESULTS

*Theorem 4.* We first establish the following result that is required in the proof of the theorem.

**Lemma 1.** *With probability atleast $1-\gamma$, the empirical risk $\mathbb{E}(\mathscr{L}(\delta^*))$ can be bounded as*

$$\mathbb{E}(\mathscr{L}(\delta^*)) = \mathtt{R}^*(\mathbb{L}) + O(\frac{1}{\sqrt{N}}), \tag{29}$$

*where $\mathtt{R}^*(\mathbb{L})$ denotes the minimum attainable risk within the (loss) function class $\mathbb{L}$.*

*Lemma 1.* By [33, Corollary 6.1], we have with probability at least $1-\gamma$:

$$\mathbb{E}(\mathscr{L}(\delta^*)) \leq \mathtt{R}^*(\mathbb{L}) + 4\mathbb{E}(\mathscr{R}_N(\mathbb{L}(\delta^*))) + \sqrt{\frac{2\log(\frac{1}{\gamma})}{N}} \tag{30}$$

By Dudley's Theorem [34], [35], we have

$$\mathbb{E}(\mathscr{R}_N(\mathbb{L}(\delta^*))) \leq 12\sqrt{\frac{\pi}{2}}\cdot\sqrt{\frac{\mathrm{d}}{N}}, \tag{31}$$

where d denotes the (linear algebraic[1]) dimension of the observation space $\mathscr{O} \subseteq \mathbb{R}^{\mathrm{d}}$. The result follows. □

For the proof of Theorem 4, we have the following:

$$d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq d_{TV}(\hat{f}_N(o|\mathscr{I},u), f(u|\mathscr{I}),u) + d_{TV}(\hat{f}_N(o|\mathscr{I},u), f(u|h,u)) \tag{32}$$

Using Pinsker's Inequality for the last term:

$$d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq d_{TV}(\hat{f}_N(o|\mathscr{I},u), f(o|\mathscr{I},u)) + \sqrt{\frac{1}{2}D_{KL}(f(o|h,u), \hat{f}_N(o|\mathscr{I},u))} \tag{33}$$

Maximizing over $\delta$, from (11),

$$d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq d_{TV}(\hat{f}_N(o|\mathscr{I},u), f(o|\mathscr{I},u)) + \sqrt{\frac{\xi_o}{2}\max_{\delta}\ \mathscr{L}_o(\delta)} \tag{34}$$

$$\Rightarrow \text{ for some } \xi, \quad d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq d_{TV}(\hat{f}_N(o|\mathscr{I},u), f(o|\mathscr{I},u)) + \sqrt{\frac{\xi}{2}\max_{\delta}\ \mathscr{L}(\delta)}. \tag{35}$$

Using definition of total variation distance, we have

$$d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq \frac{1}{2}\int\int |\hat{f}_N(o|\mathscr{I},u) - f(o|\mathscr{I},u)| do\mathbb{P}(du\cdot d\mathscr{I}) + \sqrt{\frac{\xi}{2}\mathscr{L}(\delta^*)} \tag{36}$$

Considering a distribution over data, and using Fubini's theorem we have

$$d_{TV}(f(o|\mathscr{I},u), f(o|h,u)) \leq \frac{1}{2}\mathbb{E}\Big(\int\int |\hat{f}_N(o|\mathscr{I},u) - f(o|\mathscr{I},u)| do\mathbb{P}(du\cdot d\mathscr{I})\Big) + \mathbb{E}\Big(\sqrt{\frac{1}{2}\mathscr{L}(\delta^*)}\Big) \tag{37}$$

[1]There are $d$ linearly independent vectors in $\mathbb{R}^d$

By Jensen's inequality,

$$d_{TV}(f(o|\mathscr{I},u),f(o|a,h)) \leq \frac{1}{2}\mathbb{E}\Big(\int\int |\hat{f}_N(o|\mathscr{I},u) - f(o|\mathscr{I},u)|do\mathbb{P}(du\cdot d\mathscr{I})\Big) + \sqrt{\frac{\xi}{2}\mathbb{E}\Big(\mathscr{L}(\delta^*)\Big)} \tag{38}$$

$$d_{TV}(f(o|\mathscr{I},u),f(o|h,u)) \leq \frac{C}{2}\cdot N^{\frac{-\rho_1}{\rho_1+(\mathrm{u}+\mathrm{a})}} + \sqrt{\frac{\xi}{2}\mathbb{E}\Big(\mathscr{L}(\delta^*)\Big)} \tag{39}$$

Choosing $\varepsilon_o(N,\delta^*) = \frac{C}{2}\cdot N^{\frac{-\rho_1}{\rho_1+(\mathrm{u}+\mathrm{a})}} + \sqrt{\frac{\xi}{2}\mathbb{E}\Big(\mathscr{L}(\delta^*)\Big)}$, and using Lemma 1 the result follows. □

*Theorem 1.* The proof follows from arguments similar to [19, Theorem 4.4] and [36, Theorem 5], and is omitted for brevity. □

*Theorem 2.* The proof follows from arguments similar to Theorem 2 and Corollary 2 in [24], and is omitted for brevity. □

*Theorem 3.* The proof follows using similar arguments as in [25, Theorem 3], and is omitted for brevity. □