A Relative Ignorability Framework for Decision-Relevant Observability in Control Theory and Reinforcement Learning

MaryLena Bleile**, Minh-Nhat Phung† and Minh-Binh Tran†*

**Sanofi Pharmaceuticals, New York, NY 10018, USA
Email: MaryLena.Bleile@sanofi.com
Permanent Email: marylenableile@gmail.com

†Department of Mathematics, Texas A&M University,
College Station, TX 77843, USA
Emails: pmnt1114@tamu.edu & minhbinh@tamu.edu

August 5, 2025

Abstract

Sequential decision-making systems routinely operate with missing or incomplete data. Classical reinforcement learning theory, which is commonly used to solve sequential decision problems, assumes Markovian observability, which may not hold under partial observability. Causal inference paradigms formalise ignorability of missingness. We show these views can be unified and generalized in order to guarantee Q-learning convergence even when the Markov property fails. To do so, we introduce the concept of *relative ignorability*. Relative ignorability is a graphical-causal criterion which refines the requirements for accurate decision-making based on incomplete data. Theoretical results and simulations both reveal that non-markovian stochastic processes whose missingness is relatively ignorable with respect to causal estimands can still be optimized using standard Reinforcement Learning algorithms. These results expand the theoretical foundations of safe, data-efficient AI to real-world environments where complete information is unattainable.

1 Introduction

Reinforcement learning theory traditionally assumes that agents have complete access to state information [15, 1]. However, real-world applications usually involve missing or unobserved state components, even if these are not explicitly acknowledged or modelled. Consider, for example, the clinical AI agent developed by Komorowski, et al [7], which was shown to provide treatment suggestions for sepsis care. The AI clinician makes decisions based on 48 clinical features, aiming to minimize overall and 90-day hospital mortality rates. The input feature set, which included demographic data, vital signs, lab values, and medication history, notably does not include medical insurance status, which has been shown to influence sepsis outcome even after controlling for the hospital treatments received [9, 13]. In the absence of complete information, therefore, the standard convergence theorems do not apply.
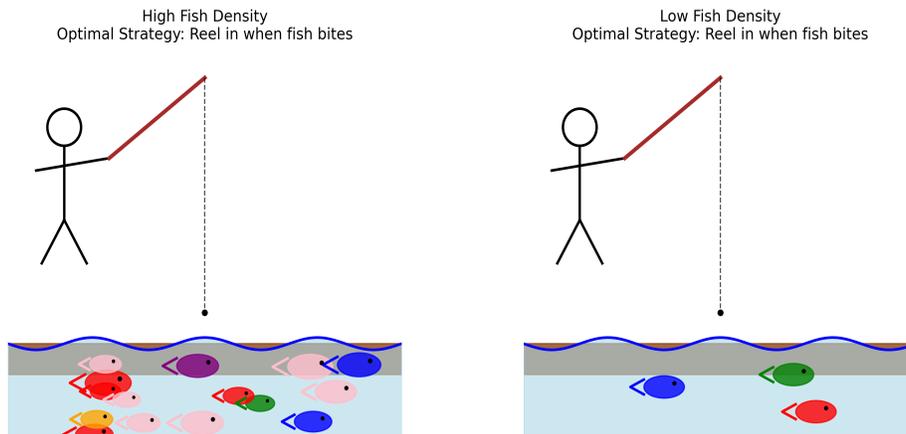

Figure 1: Fish abundance and type in the pond may affect the probability of reward. However, the optimal fishing strategy is the same regardless of fish abundance. Hence, fish abundance is relatively ignorable with respect to fishing strategy.

While Partially Observable Markov Decision Processes (POMDPs) address the challenge of partial observability [12], they require computationally expensive belief state maintenance, and convergence proofs typically require that the latency structure of the POMDP model is accurate. However, explicit modelling and estimation of all latent variables may not be necessary: Missing state components may not affect optimal decision-making, even when they violate the Markov property. Consider, for example, the fishing scenario described by Krause and Hübotter [8], where we wish to reel in the line only if there is a fish on the hook, and not otherwise. In this situation, we are rewarded for each successful catch of fish, with no reward acquired for reeling in an empty line. This problem does not explicitly model the composition of fish types in the pond, which may affect the overall rate of fish catching, since some fish may bite more frequently than others. However, the optimal strategy for line-reeling is to reel in

the line every time there is a fish on it, regardless of type; hence, the fish composition variable is not important for decision-making, and does not need to be included in the POMDP model. Figure 1 demonstrates this idea; regardless of how many fish are in the pond, one should still pull the reel in if a fish bites, and leave the hook in the water otherwise.

A similar phenomenon is present in the sepsis treatment model; here, an individual's insurance status does not affect the ranking of the potential treatments one could apply. Suppose, for example, that drug A is better than drug B for an insured individual. One might reasonably assume that if that same individual were uninsured, then drug A would still work better than drug B, even though the individual's probability of mortality would increase regardless of whether the individual received drug A or drug B; we say that insurance status is *relatively ignorable* with respect to the treatment action. With this in mind, it is unsurprising that the AI clinician which ignored patients' insurance status was highly effective despite the fact that standard Reinforcement Learning convergence proofs do not apply: Indeed, the AI clinician outperformed human clinicians by a substantial margin.

The current paper provides the mathematical scaffolding which justifies the use of Q-learning in [7]: Our main contribution is a novel concept of *relative ignorability*; this definition is foundational on established concepts developed in statistical literature on causal inference and missing data. Based on the novel definition, we provide a novel proof that Q-learning [18] converges under a marginal Bellman operator when missing components satisfy this condition. Finally, we discuss the potential utility of the relative ignorability concept beyond classic Q-learning, showing potential extensions to Deep Q-learning, as well as applications in clinical dosing and fault tolerance in distributed systems.

2   Background

Our novel concept of relative ignorability draws from missing data theory in causal inference [4, 11]. In causal inference, marginal structural models [14] handle time-varying confounding through reweighting schemas that share mathematical structure with POMDP estimation.

In classic statistical literature, there are three types of missing data: Missingness can be i) *missing completely at random*, where missingness is not related to outcome, ii) *missing at random*, where missingness and outcome are unrelated after controlling for the observed variables, or iii) *non-ignorable*, where missingness is associated with outcome even after controlling for observed covariates.

A seminal paper by Diggle and Kenward [4] identified non-ignorable missingness in three real-world datasets. One of these datasets was a trial which aimed to compare different diets for dairy cows: The outcome vector consisted of milk protein concentrations at a series of timepoints in the study. Some cows had to drop out of the study because they stopped producing milk, inducing a degree of missingness to the data.

This was shown to be *informative dropout*, which means that the resulting missingness in the data was non-ignorable. However, the informative dropout model yielded the same actionable insight as the model which ignored the informative dropout.

Actionable insights can be stable despite non-ignorable missingness if the missing variables affect treatment groups equally. Our novel concept of *relative ignorability* allows one to differentiate the stable action-insight case from other situations where inclusion of the missing data would lead us to a different actionable insight. We propose that it is only strictly necessary to model the missing information under the second case where missing variables are relatively ignorable, proving that sequential decision-making (Q-learning) based on incomplete information with relatively ignorable missingness can still yield an optimal policy. This theoretical result explains the efficacy of many practical applications such as the AI clinician discussed previously.

3   Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Over the probability space, we consider a stationary Markov Decision Process $(\mathcal{X}, \mathcal{A}, \Pi, \mu, \Gamma, \rho, \gamma)$ where:

- The countable set $\mathcal{X} \subset \mathbb{R}^d$ is the *state space*, and we denote the set of subsets of $\mathcal{X}$ by $\mathcal{B}(\mathcal{X})$.

- The *whole action space* is a countable set $\mathcal{A} \subset \mathbb{R}^{d'}$. We denote by $\mathcal{B}(\mathcal{A})$ the set of subsets of $\mathcal{A}$ and $\mathcal{P}(\mathcal{A})$ the set of probability measures on $\mathcal{A}$.

- We consider the collection of random variables $X_j : \Omega \to \mathcal{X}$, which represent the state, and $A_j : \Omega \to \mathcal{A}$, which represent the action.

- For a set $\mathcal{Z}$ ($\mathcal{X}$ or $\mathcal{A}$) and random variables $Z_k : \Omega \to \mathcal{Z}$ (X_j or A_j), we recall the notation of conditional probability

$$\mathbb{P}(Z_k \in \mathcal{Z}_k \mid Z_{k-1} = z_{k-1}, \ldots, Z_0 = z_0) = \frac{\mathbb{P}(Z_k^{-1}(\mathcal{Z}_k) \cap Z_{k-1}^{-1}(z_{k-1}) \cap \cdots \cap Z_0^{-1}(z_0))}{\mathbb{P}(Z_{k-1}^{-1}(z_{k-1}) \cap \cdots \cap Z_0^{-1}(z_0))},$$

  if

$$\mathbb{P}(Z_{k-1}^{-1}(z_{k-1}) \cap \cdots \cap Z_0^{-1}(z_0)) > 0.$$

- The class of *decision policies* Π is a set of functions $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$. The policy satisfies that the support of $\pi(\cdot \mid x)$ is a set $\mathcal{A}_x \subset \mathcal{A}$.

  The set $\mathcal{A}_x$ is the space of *allowable actions* for state x, meaning

$$\mathbb{P}(A_j \in \widetilde{\mathcal{A}} \mid X_j = x) > 0, \quad \forall j \geq 0, x \in \mathcal{X}, \widetilde{\mathcal{A}} \in \mathcal{B}(\mathcal{A}), \widetilde{\mathcal{A}} \cap \mathcal{A}_x \neq \emptyset.$$

  Since $\mathcal{X}, \mathcal{A}$ are countable, we have the following countable set

$$\mathcal{Y} = \bigcup_{x \in \mathcal{X}} \{x\} \times \mathcal{A}_x \subset \mathcal{X} \times \mathcal{A}.$$

The policy π represents the probability to take an action in $\mathcal{A}$ given state x, in the sense that

$$\mathbb{P}(A_j \in \widetilde{\mathcal{A}} \mid X_j = x_j) = \int_{\widetilde{\mathcal{A}}} \pi(da_j \mid x_j) = \pi(\widetilde{\mathcal{A}} \mid x_j)$$

for all $\widetilde{\mathcal{A}} \in \mathcal{B}(\mathcal{A}), x_j \in \mathcal{X}$.

- The distribution $\mu : \mathcal{B} \to [0, 1]$ is the *initial state distribution*, that is for $\widetilde{\mathcal{X}} \in \mathcal{B}(\mathcal{X})$

$$\mu(\widetilde{\mathcal{X}}) = \int_{\widetilde{\mathcal{X}}} \mu(dx_0) = \mathbb{P}(X_0 \in \widetilde{\mathcal{X}}).$$

We will consider $\mu(\{x\}) > 0$ for all $x \in \mathcal{X}$.

- The function $\Gamma : \mathcal{B}(\mathcal{X}) \times \mathcal{Y} \to [0, 1]$ is the *transition kernel*, that satisfies

$$\Gamma(\widetilde{\mathcal{X}} \mid x_j, a_j) = \int_{\widetilde{\mathcal{X}}} \Gamma(dx_{j+1} \mid x_j, a_j) = \mathbb{P}(X_{j+1} \in \widetilde{\mathcal{X}} \mid X_j = x_j, A_j = a_j)$$

for all $\widetilde{\mathcal{X}} \in \mathcal{B}(\mathcal{X}), (x_j, a_j) \in \mathcal{Y}$. Similar to the initial state distribution, we consider for all $j \geq 0, x_{j+1} \in \mathcal{X}$, there exists $(x_j, a_j) \in \mathcal{Y}$ such that $\Gamma(\{x_{j+1}\} \mid x_j, a_j) > 0$.

- The Markov property for the decision process includes:

$$\mathbb{P}(A_j \in \widetilde{\mathcal{A}} \mid X_j = x_j, A_{j-1} = a_{j-1}, \ldots, X_0 = x_0) = \mathbb{P}(A_j \in \widetilde{\mathcal{A}} \mid X_j = x_j),$$

and

$$\mathbb{P}(X_{j+1} \in \widetilde{\mathcal{X}} \mid X_j = x_j, A_j = a_j, \ldots, X_0 = x_0, A_0 = a_0) = \mathbb{P}(X_{j+1} \in \widetilde{\mathcal{X}} \mid X_j = x_j, A_j = a_j)$$

for all $j \geq 0$, where $\widetilde{\mathcal{A}} \in \mathcal{B}(\mathcal{A})$, $\widetilde{\mathcal{X}} \in \mathcal{B}(\mathcal{X})$, and $(x_k, a_k) \in \mathcal{Y}$ for $k = 0, \ldots, j$.

- The bounded function $\rho : \mathcal{Y} \to \mathbb{R}$ is the *reward function*.

- The constant $\gamma \in (0, 1)$ is the *discount factor*.

Example: We consider the simple system of 2 states $\mathcal{X} = \{0, 1\}$ and the action space $\mathcal{A} = \{0, 1, 2\}$.

We consider the allowable action spaces $\mathcal{A}_0 = \{0, 1\}$ and $\mathcal{A}_1 = \{2\}$. Thus, $\mathcal{Y} = \{(0, 0), (0, 1), (1, 2)\}$. The initial is given by $\mu(\{0\}) = 0.5$, and $\mu(\{1\}) = 0.5$. We want to transition from state 0 to state 1. Let the transition kernel be given by

$$\Gamma(\{0\} \mid 0, 0) = 0.6, \quad \Gamma(\{1\} \mid 0, 0) = 0.4, \quad \Gamma(\{0\} \mid 0, 1) = 0.1,$$

$$\Gamma(\{1\} \mid 0, 1) = 0.9, \quad \Gamma(\{1\} \mid (1, 2)) = 1.$$

We see that once we reach state 1, we stop. We consider a reward as follows:

$$\rho(0,0) = 1, \quad \rho(0,1) = 2, \quad \rho(1,2) = 0.$$

This means we give reward for any allowable action when in state 0 but give none when we already at state 1.

We can try to enforce the policy π, which is given by

$$\pi(\{0\} \mid 0) = 0.2, \quad \pi(\{1\} \mid 0) = 0.8, \quad \pi(\{2\} \mid 1) = 1.$$

Or, we consider π, which is given by

$$\pi(\{0\} \mid 0) = 0.8, \quad \pi(\{1\} \mid 0) = 0.2, \quad \pi(\{2\} \mid 1) = 1.$$

The class of policies Π consists of these two policies, and we want to know which is better given ρ and γ.

By [3, Proposition 7.28], there exists unique probability measure μ_j^π on $(\mathcal{X} \times \mathcal{A})^j \times \mathcal{X}$, which is supported in $\mathcal{Y}^j \times \mathcal{X}$, such that for all $\widetilde{\mathcal{X}}_j \in \mathcal{B}(\mathcal{X})$ and $\widetilde{\mathcal{A}}_j \in \mathcal{B}(\mathcal{A})$, we have

$$\mu_j^\pi(\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{A}}_{j-1} \times \widetilde{\mathcal{X}}_j) = \int_{\widetilde{\mathcal{X}}_0} \int_{\widetilde{\mathcal{A}}_0} \cdots \int_{\widetilde{\mathcal{X}}_j} \Gamma(dx_j \mid x_{j-1}, a_{j-1}) \pi(da_{j-1} \mid x_{j-1}) \ldots \pi(da_0 \mid x_0) \mu(dx_0).$$

There also exists unique probability measure $\bar{\mu}_j^\pi$ on $(\mathcal{X} \times \mathcal{A})^{j+1}$, which is supported in $\mathcal{Y}^{j+1}$, such that

$$\bar{\mu}_j^\pi(\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j) = \int_{\widetilde{\mathcal{X}}_0} \int_{\widetilde{\mathcal{A}}_0} \cdots \int_{\widetilde{\mathcal{A}}_j} \pi(da_j \mid x_j) \Gamma(dx_j \mid x_{j-1}, a_{j-1}) \ldots \pi(da_0 \mid x_0) \mu(dx_0).$$

According to [3, Proposition 7.28], by Kolmogorov extension theorem, there exists a unique extension $\bar{\mu}_\infty^\pi$ of μ_j^π and $\bar{\mu}_j^\pi$ on $(\mathcal{X} \times \mathcal{A})^{\mathbb{N}}$, supported in $\mathcal{Y}^{\mathbb{N}}$. It is an extension in the sense that

$$\mu_j^\pi(\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{A}}_{j-1} \times \widetilde{\mathcal{X}}_j) = \int_{\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{X}}_j \times \mathcal{A} \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N} \setminus \{0,1,\ldots,j\}}} \bar{\mu}_\infty^\pi, \quad \forall j \geq 0;$$

and

$$\bar{\mu}_{j+1}^\pi(\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j) = \int_{\widetilde{\mathcal{X}}_0 \times \widetilde{\mathcal{A}}_0 \times \cdots \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N} \setminus \{0,1,\ldots,j\}}} \bar{\mu}_\infty^\pi, \quad \forall j \geq 0.$$

For $\bar{\mu}_\infty^\pi$, we have the following lemma.

Lemma 1. *For $\widetilde{\mathcal{X}}_j \subset \mathcal{X}, \widetilde{\mathcal{A}}_j \subset \mathcal{A}$ and $\widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \cap \mathcal{Y} \neq \emptyset$, we have*

$$\bar{\mu}_\infty^\pi((\mathcal{X} \times \mathcal{A})^j \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N} \setminus \{0,1,\ldots,j\}}) > 0.$$

Proof. It suffices to show the proof for the case $\widetilde{\mathcal{X}}_j = \{x\}, \widetilde{\mathcal{A}}_j = \{a\}$ for $(x, a) \in \mathcal{Y}$. For the mentioned case, we have

$$\bar{\mu}_\infty^\pi((\mathcal{X} \times \mathcal{A})^j \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,j\}})$$
$$= \bar{\mu}_{j+1}^\pi((\mathcal{X} \times \mathcal{A})^j \times \{x\} \times \{a\})$$
$$= \mathbb{P}(X_j = x, A_j = a)$$
$$= \mathbb{P}(A_j = a \mid X_j = x)\mathbb{P}(X_j = x)$$
$$= \mathbb{P}(A_j = a \mid X_j = x) \sum_{(x',a')\in\mathcal{Y}} \mathbb{P}(X_j = x \mid X_{j-1} = x', A_{j-1} = a')\mathbb{P}(X_{j-1} = x', A_{j-1} = a').$$

From the equality, we can obtain the positivity by induction. $\qquad\square$

For $\widetilde{\mathcal{X}}_j \subset \mathcal{X}, \widetilde{\mathcal{A}}_j \subset \mathcal{A}$ and $\widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \cap \mathcal{Y} \neq \emptyset$, we can define a probability measure $\bar{\mu}_{j+1,\infty}^\pi(\cdot \mid \widetilde{\mathcal{X}}_j, \widetilde{\mathcal{A}}_j)$ over $(\mathcal{X} \times \mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,j\}}$ as follows

$$\bar{\mu}_{j+1,\infty}^\pi(\widetilde{\mathcal{Y}}_{j+1,\infty} \mid \widetilde{\mathcal{X}}_j, \widetilde{\mathcal{A}}_j) = \frac{\bar{\mu}_\infty^\pi((\mathcal{X} \times \mathcal{A})^j \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \times \widetilde{\mathcal{Y}}_{j+1,\infty})}{\bar{\mu}_\infty^\pi((\mathcal{X} \times \mathcal{A})^j \times \widetilde{\mathcal{X}}_j \times \widetilde{\mathcal{A}}_j \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,j\}})},$$

for all Borel measurable $\widetilde{\mathcal{Y}}_{j+1,\infty}$ of $(\mathcal{X} \times \mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,j\}}$.

For bounded measurable function $g : \mathcal{Y}^{\mathbb{N}\setminus\{0,\dots,j\}} \to \mathbb{R}$, we define

$$\mathbb{E}_{j+1,\infty}^\pi[g \mid \widetilde{\mathcal{X}}_j, \widetilde{\mathcal{A}}_j] = \int_{(\mathcal{X}\times\mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,j\}}} g\bar{\mu}_{j+1,\infty}^\pi(\cdot \mid \widetilde{\mathcal{X}}_j, \widetilde{\mathcal{A}}_j).$$

We now discuss the missing data model. In the Markov Decision Process, for each $j \geq 0$, we get the state $x_j \in \mathcal{X}$, the action $a_j \in \mathcal{A}$ is then chosen to determine the reward ρ and transition Γ. In the missing data model, we only have partial information of the state. At time j, we consider a countable set $\mathcal{X}_j^o \subset \mathbb{R}^{d_j}$, where $d_j \leq d$ as the observed state space. The observation is characterized by the index set $I_j^o \subset \{1, \dots, d\}$ and a projection map $\mathcal{O}_j : \mathcal{X} \to \mathcal{X}_j^o, (x^i)_{i=1}^d \mapsto (x^i)_{i\in I_j^o}$.

Now, we introduce the new concepts of *partially ignorability* and *relative ignorability*.

Definition 1 (Partial Ignorability). *The Markov Decision Process is called partially ignorable at time $j + 1$ if the following conditions are met:*

- *There exists a partition $\{1, \dots, d\}$ into I_U, I_W so that the random variable $X_{j+1} = (X_{j+1}^i)_{i=1}^d$, where $U = (X_{j+1}^i)_{i\in I_U}, W = (X_{j+1}^i)_{i\in I_W}$ are two independent random variables. The independence in here means that there exist distributions Γ_U, Γ_W such that*

  $$\Gamma_U(\mathcal{U} \mid x, a)\Gamma_W(\mathcal{W} \mid x, a) = \Gamma((\mathcal{U}, \mathcal{W}) \mid x, a)$$

  for $\mathcal{U}, \mathcal{W}$ such that $(\mathcal{U}, \mathcal{W}) = \{x \in \mathbb{R}^d, (x^i)_{i\in I_U} \in \mathcal{U}, (x^i)_{i\in I_W} \in \mathcal{W}\} \subset \mathcal{X}$ and $(x, a) \in \mathcal{Y}$.
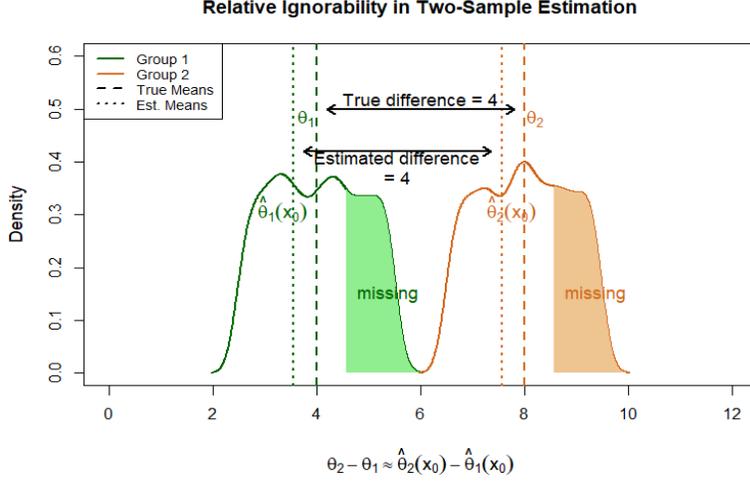
7

Figure 2: A heuristic example of relative ignorability in a group mean difference estimation problem. Note that while each estimated group mean is biased, the estimated difference in means is accurate, because missingness affects both groups equivalently.

- *The action can be decided by U, that is for all $\pi \in \Pi$, there exists π_U such that*

$$\pi(\{a\} \mid (x^i)_{i=1}^d) = \pi_U(\{a\} \mid (x^i)_{i \in I_U})$$

  for all $(x^i)_{i=1}^d \in \mathcal{X}$.

- *Finally, we have*

$$\Gamma_U(\mathcal{U} \mid x, a) = \Gamma_U(\mathcal{U} \mid x', a), \quad \forall (x, a), (x', a) \in \mathcal{Y}, \mathcal{O}_j(x) = \mathcal{O}_j(x'),$$

  and for all $(\mathcal{U}, \mathcal{W}) \subset \mathcal{X}$.

For a bounded $g : \mathcal{Y} \to \mathbb{R}$, we can embed g into a function $(g)_k$, which maps $\mathcal{Y}^{\mathbb{N} \setminus \{0,1,\dots,j\}}$, defined by

$$(g)_k((x_l, a_l)_{l \geq j+1}) = g(x_k, a_k).$$

Definition 2 (Relative Ignorability). *We consider the observed operators $(\mathcal{O}_j)_{j \geq 0}$ with the observed state space $\mathcal{X}_j^o$. Let $g : \mathcal{Y} \to \mathbb{R}$ be a bounded function. We say that the missing model is relatively ignorable with respect to g at time j for policy π if*

$$\mathbb{E}_{j+1,\infty}^\pi \Big[(g)_{j+1} \mid \{x\}, \{a\}\Big] = \mathbb{E}_{j+1,\infty}^\pi \Big[(g)_{j+1} \mid \{x'\}, \{a\}\Big],$$

for all $(x, a), (x', a) \in \mathcal{Y}$ and $\mathcal{O}_j(x) = \mathcal{O}_j(x')$.

4 Main Results

4.1 Bellman Operator

We assume the following:

(A1) The Markov Decision Process is partially ingorable for some fixed I_U at any time $j + 1$, where $j \geq 0$.

(A2) The missing model is relatively ignorable to ρ at any time $j \geq 0$ and for all the policy π in Π.

Example: Suppose we have $\mathcal{X} \subset \mathbb{R}^4$, $\mathcal{A} \subset \mathbb{R}$. For a $j \geq 0$, we have $X_{j+1} = ((X_{j+1}^i)_{i \in \{1,2\}}, (X_{j+1}^i)_{i \in \{3,4\}})$, and the Markov Decision Process is $(X_{j+1}^i)_{i \in \{1,2\}}$–partially ignorable. We consider the observed operator at any time j is either

$$(x^1, x^2, x^3, x^4) \mapsto (x^1, x^2) \text{ or } (x^1, x^2, x^3, x^4) \mapsto (x^1, x^2, x^3).$$

In other words, in this process, the first two components of the next states is decided by the first two components of the previous states, and the action is also decided based on the first two component, missing x^3 or x^4 is ignorable.

If the reward is

$$\rho(x^1, x^2, x^3, x^4, a) = \frac{|ax_1|}{|ax_1| + 1},$$

then the missing model is relatively ignorable for ρ at any time $j \geq 0$ since ρ depends only on the first component of the state.

We back to the general settings, the value function $V_j : \mathcal{Y}^{\mathbb{N} \setminus \{0,1,\ldots,j-1\}} \to \mathbb{R}$ at time j is defined by

$$V_j((x_k, v_k)_{k \geq j}) = \sum_{k=j}^{\infty} \gamma^{k-j} \rho(x_k, a_k).$$

Since ρ is bounded and $\gamma \in (0, 1)$, the series converges absolutely for given $(x_k, v_k) \in \mathcal{Y}$.

For $j \geq 0$, the *action-value* function $q_j^\pi : \mathcal{Y} \to \mathbb{R}$ is defined by

$$q_j^\pi(x, a) = \rho(x, a) + \gamma \mathbb{E}_{j+1,\infty}^\pi [V_{j+1} \mid \{x\}, \{a\}].$$

We will have

$$q_j^\pi(x, a) = \rho(x, a) + \gamma \mathbb{E}_{j+1,\infty}^\pi [(q_{j+1}^\pi)_{j+1} \mid \{x\}, \{a\}].$$

Definition 3 (Bellman Operator). *For $j \geq 0$ and $Q : \mathcal{Y} \to \mathbb{R}$ is a bounded function , the Bellman operator of policy π at time j is defined by*

$$T_j^\pi Q(x, a) := \rho(x, a) + \gamma \mathbb{E}_{j+1,\infty}^\pi [(Q)_{j+1} \mid \{x\}, \{a\}],$$

for $(x, a) \in \mathcal{Y}$. We also define the Bellman optimality operator at time j by

$$T_j^* Q(x, a) := \rho(x, a) + \gamma \sup_{\pi \in \Pi} \mathbb{E}_{j+1, \infty}^\pi [(Q)_{j+1} \mid \{x\}, \{a\}],$$

for $(x, a) \in \mathcal{Y}$.

We have the following lemma:

Lemma 2. *T_j^π and T_j^* are contractive mappings with factor γ in the sup-norm.*

Proof. For any two bounded function $Q_1, Q_2 : \mathcal{Y} \to \mathbb{R}$, we estimate

$$\|T_j^\pi Q_1(x, a) - T_j^\pi Q_2(x, a)\|_\infty = \gamma \sup_{(x,a) \in \mathcal{Y}} \mathbb{E}_{j+1, \infty}^\pi [(Q_1 - Q_2)_{j+1} \mid \{x\}, \{a\}]$$

$$\leq \gamma \sup_{(x,a) \in \mathcal{Y}} \mathbb{E}_{j+1, \infty}^\pi [\|Q_1 - Q_2\|_\infty \mid \{x\}, \{a\}] = \gamma \|Q_1 - Q_2\|_\infty.$$

Then, for Bellman optimality operator, we get that

$$\|T_j^* Q_1(x, a) - T_j^* Q_2(x, a)\|_\infty = \|\sup_{\pi \in \Pi} T_j^\pi Q_1(x, a) - \sup_{\pi \in \Pi} T_j^\pi Q_2(x, a)\|_\infty$$

$$\leq \sup_{\pi \in \Pi} \|T_j^\pi Q_1(x, a) - T_j^\pi Q_2(x, a)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty.$$

$\square$

Lemma 3. *Given Assumptions (A1) and (A2), if the missing model is relatively ignorable with respect to Q at some time $j \geq 1$ and for some $\widetilde{\pi} \in \Pi$, and if $I_U \supset I_j^o$, then the missing model is relatively ignorable with respect to $T_j^{\widetilde{\pi}} Q$ at any time $k \geq 0$ for all policy $\pi \in \Pi$.*

As a consequence, if the missing model is relatively ignorable with respect to Q at some time $j \geq 1$ for all $\pi \in \Pi$, and if $I_U \supset I_j^o$, then the missing model is relatively ignorable with respect to $T_j^* Q$ at any time $k \geq 0$ for all policy $\pi \in \Pi$.

Proof. Since the missing model is relatively ignorable with respect to ρ at any time for all policy, we need to show that the missing model is also relative ignorable with respect to

$$E_j^{\widetilde{\pi}} Q(x, a) = \mathbb{E}_{j+1, \infty}^{\widetilde{\pi}} [(Q)_{j+1} \mid \{x\}, \{a\}]$$

at any time and for all policy.

Since the missing model is relatively ignorable with respect to Q at time j for policy $\widetilde{\pi}$, we have

$$\mathbb{E}_{j+1, \infty}^{\widetilde{\pi}} [(Q)_{j+1} \mid \{x\}, \{a\}] = \mathbb{E}_{j+1, \infty}^{\widetilde{\pi}} [(Q)_{j+1} \mid \{x'\}, \{a\}],$$

for all $(x, a), (x', a) \in \mathcal{Y}$ and $\mathcal{O}_j(x) = \mathcal{O}_j(x')$.

Thus, we get

$$E_j^{\widetilde{\pi}} Q(x, a) = E_j^{\widetilde{\pi}} Q(x', a),$$

10

for all $(x, a), (x', a) \in \mathcal{Y}$ and $\mathcal{O}_j(x) = \mathcal{O}_j(x')$. Because $I_U \supset I_j^o$, when $(x^i)_{i \in I_W}$ varies, $T_j^{\widetilde{\pi}}(x, a)$ stays fixed.

For $(x, a), (x', a) \in \mathcal{Y}$ and $\mathcal{O}_k(x) = \mathcal{O}_k(x')$, we compute

$$\mathbb{E}_{k+1,\infty}^{\pi}[(E_j^{\widetilde{\pi}}Q)_{k+1} \mid \{x\}, \{a\}]$$

$$= \frac{\bar{\mu}_j^{\pi}((\mathcal{X} \times \mathcal{A})^k \times \{x\} \times \{a\})}{\bar{\mu}_\infty^{\pi}((\mathcal{X} \times \mathcal{A})^k \times \{x\} \times \{a\} \times (\mathcal{X} \times \mathcal{A})^{\mathbb{N}\setminus\{0,1,\dots,k\}})} \int_{\mathcal{X} \times \mathcal{A}} E_j^{\widetilde{\pi}}(\hat{x}, \hat{a}) \pi(d\hat{a} \mid \hat{x}) \Gamma(d\hat{x} \mid x, a)$$

$$= \int_{\mathcal{X} \times \mathcal{A}} E_j^{\widetilde{\pi}}(\hat{x}, \hat{a}) \pi_U(d\hat{a} \mid (\hat{x}^i)_{i \in I_U}) \Gamma_U(d(\hat{x}^i)_{i \in I_U} \mid x, a) \Gamma_W(d(\hat{x}^i)_{i \in I_W} \mid x, a)$$

$$= \int_{\mathcal{X} \times \mathcal{A}} E_j^{\widetilde{\pi}}(\hat{x}, \hat{a}) \pi_U(d\hat{a} \mid (\hat{x}^i)_{i \in I_U}) \Gamma_U(d(\hat{x}^i)_{i \in I_U} \mid x', a) \Gamma_W(d(\hat{x}^i)_{i \in I_W} \mid x', a)$$

$$= \mathbb{E}_{k+1,\infty}^{\pi}[(E_j^{\widetilde{\pi}}Q)_{k+1} \mid \{x'\}, \{a\}].$$

If the missing model is relatively ignorable with respect to Q for all $\pi \in \Pi$ then

$$\sup_{\pi \in \Pi} \mathbb{E}_{j+1,\infty}^{\pi}[(Q)_{j+1} \mid \{x\}, \{a\}] = \sup_{\pi \in \Pi} \mathbb{E}_{j+1,\infty}^{\pi}[(Q)_{j+1} \mid \{x'\}, \{a\}],$$

for all $(x, a), (x', a) \in \mathcal{Y}$ and $\mathcal{O}_j(x) = \mathcal{O}_j(x')$. We perform the previous computation again to obtain that the missing model is relatively ignorable with respect to T_j^*Q at any time $k \geq 0$ for all policy $\pi \in \Pi$. $\square$

The lemma means that the Bellman operator is relatively ignorable if we have the additional assumption $I_U \supset I_j^o$. Formally, we assume

(A3) For U is the component in Assumption (A1), there exists $j \geq 1$ such that $I_U \supset I_j^o$.

Since T_j^* is a contractive mapping for all $j \geq 0$. We start with $Q = 0$, which is relatively ignorable. By the Banach Fixed-Point Theorem, the sequence

$$Q, T_j^*Q, (T_j^*)^2Q, (T_j^*)^3Q, \dots$$

converges in L^∞-norm to the fixed-point Q_j^* of T_j^* and this sequence is a sequence of relatively ignorable functions.

4.2 Convergence Under Relative Ignorability

In Q-learning, we consider the learing factor $\alpha_n \in (0, 1)$, which satisfies the following assumption:

(A4)
$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

For a given $j \geq 0$, the algorithm for Q-learning rule is given by

$$Q^n(x,a) = (1 - \alpha_n)Q^{n-1}(x,a) + \alpha_n T_j^* Q^{n-1}(x,a),$$

for the initial function $Q^0 = 0$.

Theorem 1 (Q-Learning Convergence). *For the Markov Decision Process with Assumption (A2) and the Q-learning with Assumption (A4), the sequence $Q^n(X_j, A_j)$ converges to $Q_j^*(X_j, A_j)$ with probability 1 for any policy $\pi \in \Pi$.*

Proof. Fix a policy $\pi \in \Pi$, we consider

$$\delta^n(x,a) = T_j^* Q^{n-1}(x,a) - Q_j^*(x,a),$$

and

$$\Delta^n(x,a) = Q^{n-1}(x,a) - Q_j^*(x,a).$$

From the Q-learning rule, we get

$$\Delta^{n+1}(x,a) = (1 - \alpha_n)\Delta^n(x,a) + \alpha_n \delta^n(x,a).$$

We need to show that $\Delta^n(X_j, A_j)$ converges to 0 with probability 1.

Since Q_j^* is a fixed point of T_j^*, and T_j^* is a contractive map with constant γ, we estimate that

$$\begin{aligned}
\|\delta^n\|_\infty &= \|T_j^* Q^{n-1} - T_j^* Q_j^*\|_\infty \\
&\leq \gamma \|Q^{n-1} - Q_j^*\|_\infty = \gamma \|\Delta^n\|_\infty.
\end{aligned}$$

As $\mathbb{P}(X_j = x, A_j = a) > 0$ for all $(x,a) \in \mathcal{Y}$, we obtain that

$$\|\delta^n(X_j, A_j)\|_{L^\infty(\Omega)} \leq \gamma \|\Delta^n(X_j, A_j)\|_{L^\infty(\Omega)}.$$

As $\alpha_n, \gamma \in (0,1)$, we observe that

$$\|\Delta^{n+1}(X_j, A_j)\|_{L^\infty(\Omega)} \leq (1 - (1-\gamma)\alpha_n)\|\Delta^n(X_j, A_j)\|_{L^\infty(\Omega)} \leq e^{-(1-\gamma)\alpha_n}\|\Delta^n(X_j, A_j)\|_{L^\infty(\Omega)}.$$

Hence, inductively, we obtain

$$\|\Delta^{m+1}(X_j, A_j)\|_{L^\infty(\Omega)} \leq e^{-(1-\gamma)\sum_{n=1}^m \alpha_n}\|\Delta^1(X_j, A_j)\|_{L^\infty(\Omega)}.$$

As $\sum_{n=1}^\infty \alpha_n = \infty$, we get

$$\|\Delta^{m+1}(X_j, A_j)\|_{L^\infty(\Omega)} \to 0$$

as $m \to \infty$. This leads to $Q^n(X_j, A_j)$ converges with probability 1 to $Q_j^*(X_j, A_j)$. $\qquad \square$

Figure 3 visualizes a POMDP. States x transition according to dynamics Γ, but the agent only observes partial information x_o (with the remaining components x_m missing). The agent maintains belief states $b(x)$ representing probability distributions over possible underlying states. Under relative ignorability conditions, the belief-space policy π can still converge to the optimal policy π^* through successive applications of the Q-learning update equation[18] which we denote B_0.
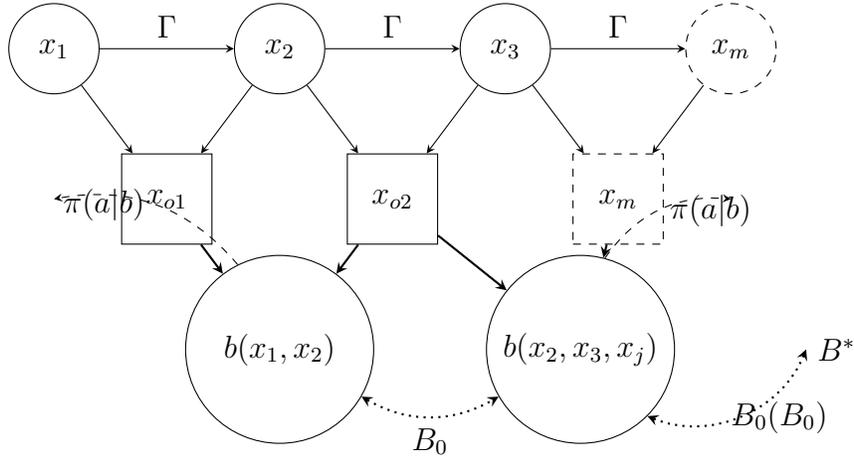
Figure 3: Visualization of a POMDP.

4.3 Examples of Relative Ignorability

Example 1: Clinical Dosing Strategy. Consider a clinical decision problem where patient state $X_j = (X_{j,o}, X_{j,m})$ includes observed symptoms and unobserved genetic markers, with action space $\mathcal{A} = \{\text{chemotherapy, immunotherapy}\}$. Let $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be the progression indicating function.

If genetic markers affect disease progression but not treatment response, then for all possible genetic marker values $x_m^{(1)}, x_m^{(2)} \in \Omega_m$:

$$\arg\max_{a \in A} Q((x_{j,o}, x_m^{(1)}), a) = \arg\max_{a \in A} Q((x_{j,o}, x_m^{(2)}), a) = a^*(x_{j,o}).$$

In this case, the optimal treatment depends only on the observed symptoms. Thus, one example of a policy π is

$$\pi^o(a^*(x_{j,o}) \mid x_{j,o}) = \pi(a^*(x_{j,o}) \mid (x_{j,o}, x_m^{(1)})) = 1$$

for any $x_m^{(1)} \in \Omega_m$. If the next observed symptoms depend only on the current symptoms and the treatment, then the model is partially ignorable.

Furthermore, if we define $g(x_{j,o}, x_m^{(1)}) = \max_{a \in A} Q\left((x_{j,o}, x_m^{(1)}), a\right)$, then g is relatively ignorable.

If genetic markers strongly influence treatment effectiveness, then there exist genetic marker values $x_m^{(1)}, x_m^{(2)}$ such that:

$$\arg\max_{a \in A} Q((x_{j,o}, x_m^{(1)}), a) \neq \arg\max_{a \in A} Q((x_{j,o}, x_m^{(2)}), a)$$

For example, if $x_m^{(1)}$ indicates a mutation making immunotherapy optimal while $x_m^{(2)}$ indicates wild-type genes making chemotherapy optimal, then the same observed symptoms require different treatments based on genetics. In this counterexample, missingness is relatively non-ignorable, and standard Q-learning may converge to a suboptimal

13

point. This scenario might occur if the presence of a wild-type gene is also associated with a phenotype that influences the prescriber's decision.

Example 2: Split-brain Tolerance in Distributed Systems Consider a distributed system with nodes containing customer data, where a network partition separates nodes into two groups that cannot communicate. Each partition may continue processing transactions, potentially creating conflicting states. Let $X_j = (X_{j,1}, X_{j,2})$ represent the system state, where $X_{j,1}$ contains data from partition 1 and $X_{j,2}$ contains data from partition 2. During a split brain event, each partition observes only its local state: partition 1 observes $X_{j,o} = X_{j,1}$ while $X_{j,m} = X_{j,2}$ is missing, and vice versa.

Not all inconsistencies matter equally for different application functions: Only data that is both conflicting and used by the specific application function being evaluated is relatively non-ignorable with respect to that function's decision-making.

Systems architects can use relative ignorability to implement *selective degradation*: during split brain scenarios, functions that depend on relatively non-ignorable conflicting data can be temporarily disabled or routed to require manual approval, while functions that operate on relatively ignorable conflicts can continue operating normally.

Let $\hat{g}(X_j, A_j)$ represent an application function (such as "approve customer transaction" or "calculate account balance") that takes system state and proposed action as inputs. The missing components $X_{j,m}$ are relatively ignorable with respect to $\hat{g}$ if Equation 1 holds (where $X_{j,o}$ is the observed information and Ω_m is the set of possible values $X_{j,m}$ can take), even when $X_{j,m}$ contains conflicting information.

$$\hat{g}((X_{j,o}, X_{j,m}), A_j) = \hat{g}((X_{j,o}, X'_{j,m}), A_j) \forall X_{j,m}, X'_{j,m} \in \Omega_m \tag{1}$$

Assuming one has knowledge of the data dependencies for each application function, one could track which data elements are conflicting across partitions, and disable only the functions which use the conflicting information. This selective termination process would allow unaffected functions to continue to run, thereby minimizing the impact to the application while also mitigating the impact of information errors.

Consider, for example, a payment processing system which includes the following functions: i) Fraud detection, which uses the customer's transation history, and ii) Marketing recommendations, which is based on transaction history as well as the user profile (age, gender, address, etc.) iii) other capabilities such as purchasing, which uses neither transaction history or demographic information. Let us use x'_j to denote the transaction history at time j and x''_j to denote the user's demographic data.

Suppose that x''_j is stored on both partitions and differs due to the split-brain problem. Traditional methods for handling this situation might include temporarily disabling the user's account to avoid compounding errors, since x''_j is classically non-ignorable with respect to the application function for that user. However, this might not be necessary: Since the fraud detection capability and other functionality does not depend on x''_j, we might simply disable the marketing recommendations and allow

the rest of the application functions to run normally. We can do this because x_j'' is *relatively ignorable* with respect to fraud detection (and other functions).

5   Discussion

Classical Q-learning convergence theory [18, 6, 17], and its extensions to function approximation [10] assume complete state observability. We have developed, here, a framework of relative ignorability that relaxes this fundamental requirement while preserving convergence guarantees. Our result eliminates the need for explicit POMDP modelling and estimation in certain cases, by specifying when such complexity is unnecessary.

Recent work on causal reinforcement learning [19] shares similar motivations but focuses on confounding variables rather than general missing data patterns; our relative ignorability framework provides a more general unifying perspective on when partial observability can be safely ignored. Advantage learning [5] also naturally connects to relative ignorability. Advantage learning focuses on learning the advantage function $A(x, a) = Q(x, a) - V(x)$, which represents relative action values rather than absolute Q-values. From a relative ignorability perspective, if missing components affect all actions equally, they may bias individual Q-values while preserving advantage rankings. The results described here could also be extended to show that advantage learning is more robust to certain types of missing data than standard Q-learning. Our results show that the Markov assumption can be relaxed when violations do not affect decision-making. This suggests that the observability requirements arise dynamically relative to task demands.

There are a variety of future directions for our work. First, our theoretical result might be extended to Deep Q-learning, a popular modification of classical Q-learning which leverages Deep Neural Networks in order to estimate the Q-function. To probe the feasibility of this direction, we conducted a simulation which compared the performance of Deep Q-Learning under various relative ignorability conditions. The simulated environment consisted of a classic 2x2 gridworld with one goal square as well as a trap square, and potential actions consisted of moving left, right, up or down. In addition, we added a latent "mode" variable which affected the layout: The complete state is $X_j = (position, mode)$ where $position \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ and $mode \in \{0, 1\}$. The agent only observes position; mode is missing. We consider two forms of the latent mode, one which is relatively ignorable, and one which is relatively non-ignorable. The relatively ignorable mode affects reward values only: if mode=1, the goal square yields 10 reward (8 if mode=0), and trap yields -10 penalty (-8 if mode =0). Note that since mode affects the reward even after conditioning on the observed state, the unobserved mode is classically non-ignorable, though relatively ignorable. Conversely, the relatively non-ignorable mode swaps the location of the goal with the location of the trap, consistently returning 10 reward at the goal and -10 at the trap. Additionally, the agent was penalized with -0.1 reward for each step, to incite efficient

progress toward the goal.

We performed Deep Q-learning using a two-layer neural network, with ReLU activation between the two layers. Each layer consisted of 64 nodes. For comparison with POMDP learning, we also consider a situation where the agent has access to a noisy signal of the mode. We generated this noisy signal as a random normal variable, updated at each timestep, with the mean equal to the value of the latent mode (0 or 1), and variance $\sigma^2 = 0.15$. The code to implement the environment step update follows:

```
def step(self, action_idx):
      action = ACTIONS[action_idx]
      dx, dy = ACTION_TO_DELTA[action]
      new_x = np.clip(self.pos[0] + dx, 0, GRID_SIZE - 1)
      new_y = np.clip(self.pos[1] + dy, 0, GRID_SIZE - 1)
      self.pos = (new_x, new_y)

      reward = -0.1
      done = False
      if self.pos == (0, 1):  # Goal
          reward = (10 if self.mode == 0 else 9) if self.relative_ignorability
              else (-10 if self.mode == 0 else 10)
          done = True
      elif self.pos == (1, 0):  # Trap
          reward = -10 if self.relative_ignorability
              else (10 if self.mode == 0 else -10)
          done = True

      # observe new evidence
      self.observed_variable = np.random.normal(loc=self.mode, scale=0.15)

      # update belief based on new observation
      self.update_belief(obs=self.observed_variable)

      return np.array([*self.pos, self.belief_mode],

      dtype=np.float32), reward, done
```

For POMDP estimation, we performed a classic Bayesian belief update at each step as follows. Here, the belief mode is the agent's prior belief of what the current mode is, where prior belief is intialized agnostically to 0.5.

```
def update_belief(self, obs):
    p1 = np.exp(-(obs - 1)**2 / 2)
    p0 = np.exp(-(obs - 0)**2 / 2)
```

```
prior = self.belief_mode
self.belief_mode = (p1 * prior) / (p1 * prior + p0 * (1 - prior) + 1e-8)
```

We trained all models for 1000 episodes, using $\gamma = 0.9$, $\alpha = .001$, and epsilon decay rate 0.995. To ensure stability of the results, we repeated the simulation with 5 different random seeds, and averaged the resulting reward curve for each mode (0 or 1, with 1 corresponding to the relatively non-ignorable case), model (POMDP or Vanilla), and training episode (1 to 1000) across the seeds.

Figure 5 shows the 50-timepoint rolling mean of the 5-seed average reward curve from each algorithm and mode across training epochs. Despite the nonignorably missing "mode" information, rewards obtained by vanilla Q-learning converge to 9: Note that 9 is the maximum possible reward, since it is the average of 8 and 10 (the goal rewards under each mode in the relatively ignorable setting). Moreover, vanilla Q-learning appears to be more efficient than POMDP in the relatively ignorable case, converging in fewer epochs. The slower convergence of POMDP Q-learning is understandable from an information theoretic perspective: In POMDP learning, the information in each observation must be shared across two estimation procedures - belief update and Q update - while in Vanilla Q-learning there is only one function to estimate. However, if the mode is non-ignorable, then Vanilla Q-learning does not converge, while POMDP estimation can still find the optimal policy.
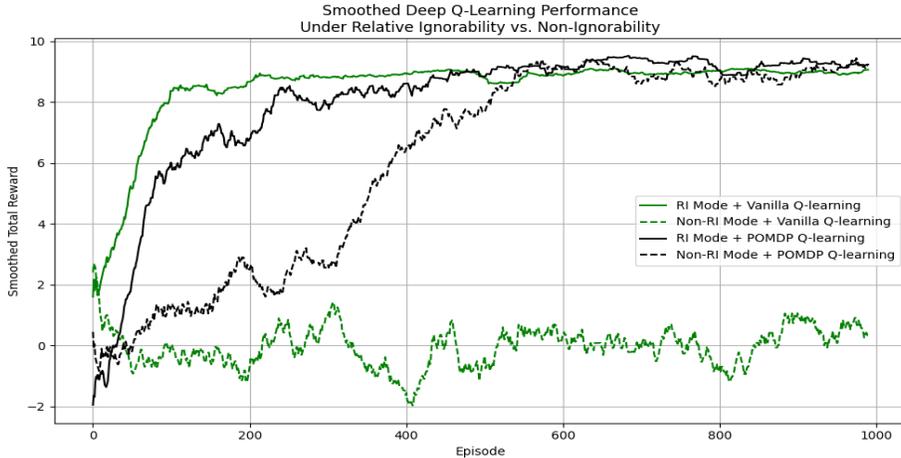

Figure 4: Reward curves from Vanilla and POMDP Q-learning under relatively ignorable and relatively non-ignorable latent mode. Vanilla Q-learning converges to the maximum obtainable reward faster than POMDP Q-learning under relative ignorability.

In addition to Deep Q-learning, our framework could also extend to continuous state spaces, function approximation, and policy gradient methods. Connections to optimal control theory [2] suggest further theoretical developments. In practice, determining relative ignorability requires domain knowledge or empirical validation. Future

17

work should develop algorithms for automatically detecting when this condition holds. Methods based on the index of sensitivity to nonignorability developed by Troxel, Ma, and Heitjan [16] seem compelling.

6   Conclusion

We have introduced relative ignorability as a condition under which Q-learning converges despite missing state components. This framework relaxes classical assumptions while maintaining theoretical guarantees, offering a middle ground between full observability and complex POMDP solutions. Our results suggest that the curse of dimensionality in agentic AI may be mitigated by focusing on decision-relevant information rather than complete state reconstruction.

References

[1] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[2] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[3] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978.

[4] Peter Diggle and Michael G. Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 43(1):49–73, 1994.

[5] Mance E. Harmon and Leemon C. Baird. Multi-player residual advantage learning with general function approximation. *Machine Learning*, 23:165–187, 1996.

[6] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

[7] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[8] Andreas Krause and Jonas Hübotter. Probabilistic artificial intelligence. *arXiv preprint arXiv:2502.05244*, 2025.

[9] Gagan Kumar, Amit Taneja, Tilottama Majumdar, Elizabeth R Jacobs, Jeff Whittle, Rahul Nanchal, et al. The association of lacking insurance with outcomes of severe sepsis: retrospective analysis of an administrative database. *Critical care medicine*, 42(3):583–591, 2014.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[11] Karthika Mohan, Judea Pearl, and Jin Tian. Missing data as a causal inference problem. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2013.

[12] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[13] Chanu Rhee, Travis M Jones, Yasir Hamad, Anupam Pande, Jack Varon, Cara O'Brien, Deverick J Anderson, David K Warren, Raymund B Dantes, Lauren Epstein, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in us acute care hospitals. *JAMA network open*, 2(2):e187571–e187571, 2019.

[14] James M. Robins, Miguel A. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[15] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[16] Andrea B Troxel, Guoguang Ma, and Daniel F Heitjan. An index of local sensitivity to nonignorability. *Statistica Sinica*, pages 1221–1237, 2004.

[17] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[18] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

[19] Junzhe Zhang and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 33, 2020.