Transition Noise Facilitates Interpretability

Zachery Boner
zachery.boner@duke.edu
Department of Computer Science
Duke University

Harry Chen
harry.chen084@duke.edu
Department of Computer Science
Duke University

Michal Moshkovitz
michal.moshkovitz@mail.huji.ac.il
Bosch center for AI

Ronald Parr
parr@cs.duke.edu
Department of Computer Science
Duke University

Cynthia Rudin
cynthia@cs.duke.edu
Department of Computer Science and
Department of Electrical and Computer Engineering
Duke University

Lesia Semenova
lesia.semenova@duke.edu
Department of Computer Science
Duke University

Abstract

Recent research in supervised learning has demonstrated that noise in data generation processes leads to the existence of accurate and simpler/interpretable machine learning models. However, the implications of this effect in the context of reinforcement learning, specifically in Markov Decision Processes (MDPs), have not been thoroughly explored. This paper investigates how noise influences the interpretability of MDPs. For two different types of transition noise, adding noise is provably equivalent to solving a noiseless MDP with a smaller discount factor. Regardless of the value function or policy function representation, problems with shorter planning horizons may be more conducive to interpretable solutions, simply because short-term effects and consequences tend to have more concise representations.

1 Introduction

Reinforcement learning (RL) is inherently complex. The algorithms rely on learning from vast amounts of data generated through interactions with the environment, which often results in highly complex models. This complexity can make it difficult to interpret or understand the reasoning behind specific decisions made by the RL agent, making the system very challenging to debug, adjust to domain knowledge, or trust – all crucial aspects for applications in high-stakes decision domains such as medical treatment or autonomous vehicle control.

Designing interpretable models is not always straightforward and might require complicated optimization techniques or extensive domain knowledge. Therefore, it is important to assess when interpretability is possible in RL systems. For example, recent results in supervised learning (Semenova et al., 2023; 2022), have demonstrated that noise in data generation processes is the practical and theoretical motivator for the increased interpretability of optimal models.

This paper takes an agnostic approach to policy or value function representation, as well as an agnostic approach to solution algorithms.

For a discounted, infinite horizon Markov Decision Processes (MDP), the planning horizon is nominally infinite, but discounting attenuates the impact of future rewards. For any discount factor and

any ϵ, one can compute a horizon length τ beyond which future choices can be ignored with penalty of no more then ϵ on the quality of decisions made. This so-called ϵ-horizon Kearns et al. (1999), τ, can be used to determine a maximum number of steps of value iteration needed, or the depth of a search needed to determine (or possibly explain) the decision taken from a single state.

The discount factor, γ, in MDPs is often viewed as a measure of future uncertainty. One interpretation of the discounting factor is that it represents a $1-\gamma$ probability of death (transition to a state with value 0) at every time step. Anecdotally, researchers have likened other types of uncertainty to discounting but, to our knowledge, this connection has not been formalized to the extent done in this paper.

The main technical contributions of this paper are to describe two specific noise models where adding noise is exactly equivalent to solving the noiseless version of the problem with a larger discount factor. To some extent, these results align with longstanding intuitions among MDP and reinforcement learning (RL) practitioners. The somewhat surprising results are the precise nature of the equivalence to a noiseless case.

2 Preliminaries

2.1 MDP notation

An MDP is a tuple, $M = (S, A, T, R, \gamma)$, where:

- S is the state space. We will assume a finite state space of size n, but our results should generalize to continuous state spaces.
- A is the action space. We assume a discrete set of actions, $a \in A$ are possible in each state.
- T is a transition model, specifying the probability of next states, $P(s'|s, a)$ for combinations of states, next states, and actions.
- R is a reward function. For simplicity, we will assume R is defined over states. We will express R as a column vector, with R_i as entry i in the vector.
- γ is a discount factor $0 \leq \gamma < 1$.
- $\mathbf{1}_{m\times n}$ is an $m \times n$ matrix of 1s. If only one dimension is specified, then this is a vector of the (context dependent) conformable orientation.

A policy, π, for an MDP is a mapping from states to actions. Every policy corresponds to a transition matrix, P^π. The value function for a policy π satisfies:

$$V^\pi = R + \gamma P^\pi V^\pi = (I - \gamma P^\pi)^{-1} R = \sum_{i=0}^{\infty} \gamma^i P^i R.$$

A policy π^* is optimal if:

$$\forall \pi : V^{\pi *} \geq V^\pi.$$

2.2 MDP Properties

Another convenient property of MDPs is that shifting the reward function by a uniform constant shifts the value function by a scaled version of this constant, i.e.,

$$(I - \gamma P^\pi)^{-1}(R + c) = \frac{c}{1-\gamma} + (I - \gamma P^\pi)^{-1} R = V^\pi + \frac{c}{1-\gamma}.$$

This allows us to assume that all rewards are non-negative without loss of generality. When all rewards are non-negative, multiplying the reward by a positive constant shifts the entire value function by the same constant:

$$(I - \gamma P^\pi)^{-1} \delta R = \delta (I - \gamma P^\pi)^{-1} R = \delta V^\pi.$$

Since shifting or scaling by a positive constant does not change an $argmax$, the optimal policy is invariant to these changes as well.

One way to think about discounting is that the agent makes a transition to a state with value fixed at 0 at each time step with probability $1 - \gamma$. This effectively reduces the probability mass the agent can control by a factor of γ at each time step. We can derive the ϵ-horizon in terms of the probability mass the agent gets to control after time τ. If we assume that after time τ all of the agent's controllable mass is allocated towards the worst possible choices (value 0) instead of achieving the highest possible value ($R = 1$), then the suboptimality of failing to plan beyond horizon depth τ is:

$$\sum_{i=\tau}^{\infty} \gamma^i (1 - 0) = \gamma^\tau \sum_{i=0}^{\infty} \gamma^i = \frac{\gamma^\tau}{1 - \gamma}.$$

The ϵ-horizon is obtained by solving for τ:

$$\begin{aligned} \frac{\gamma^\tau}{1-\gamma} &\leq \epsilon \\ \tau &\geq \log_\gamma \epsilon + \log_\gamma (1 - \gamma). \end{aligned}$$

3 Sticky Noise

We augment the definition of MDPs to M_β, where β determines a tendency to stay in the same state that is added to every policy. We assume this is true for all policies, π:

$$P^\pi_\beta = \beta P^\pi + (1 - \beta) I.$$

Thus, when $\beta = 1$ we have our original M, and when $\beta = 0$, it is impossible to leave whatever state the agent starts in. One realization of sticky noise can be actions that have a small probability of failing at any time. Another realization could be an artifact of discretization: A discretized environment always violates the Markov property a little in that it treats all parts of the state space that lie within a cell as equivalent. In practice, if an agent starts in a corner of a cell and executes an action designed to move to an adjacent cell, that action may not always succeed; sometimes it could move the agent to an opposing edge or vertex of the cell it is currently in rather than the next cell.

3.1 Finite Horizon Case

We build our intuitions for the effects of this type of noise by starting with the finite horizon case for a fixed policy, π. Denote by i_j the time the agent arrived at the j-th state and by Δ_j the number of time steps the agent was at state j. Thus, $i_{j+1} = i_j + \Delta_j$. Note that these are random variables that are independent of each other and independent of the state; they solely depend on the noise.

The value function of the noisy MDP is the expectation of the following terms:

$$R_0 + \gamma R_0 + \gamma^2 R_0 + \ldots + \gamma^{\Delta_0 - 1} R_0 + \gamma^{i_1} R_1 + \gamma^{i_1 + 1} R_1 + \ldots \tag{1}$$

If we group together all terms that are associated with the same state, and use the formula for the sum of a geometrical series we get that Equation 1 is equal to:

$$\begin{aligned} &R_0 (1 + \gamma + \gamma^2 + \ldots + \gamma^{\Delta_0 - 1}) + R_1 \gamma^{i_1} (1 + \gamma + \gamma^2 + \ldots) \\ &= R_0 \frac{1 - \gamma^{\Delta_0}}{1 - \gamma} + R_1 \gamma^{i_1} \frac{1 - \gamma^{\Delta_1}}{1 - \gamma} + \ldots \end{aligned}$$

Taking the expectation, and exploiting linearity of expectation we can focus on each term in the addition separately:

$$\mathbb{E}\left[R_j \gamma^{i_j} \frac{1-\gamma^{\Delta_j}}{1-\gamma}\right].$$

The three terms R_j, γ_{i_j}, and $\frac{1-\gamma^{\Delta_j}}{1-\gamma}$ are independent, so we can analyze the expectation of each of the three terms independently. Let us start with $\mathbb{E}\left[\frac{1-\gamma^{\Delta_j}}{1-\gamma}\right]$. Importantly, the expectation of γ^{Δ_j} is equal for all states, and we will denote it by $\mathbb{E}[\gamma^{\Delta}]$.

The implication is that the value function in the noisy MDP is scaled by $\mathbb{E}\left[\frac{1-\gamma^{\Delta}}{1-\gamma}\right]$.

Now let us analyze the term $\mathbb{E}[\gamma^{i_j}]$.

Claim 3.1.

$$\mathbb{E}[\gamma^{i_j}] = \mathbb{E}[\gamma^{\Delta}]^j.$$

Proof. We will prove the claim by induction. For the basis we have $\mathbb{E}[\gamma^{i_0}] = \mathbb{E}[\gamma^0] = 1$, where the first equality follows from the fact that $i_0 = 0$. For the induction step we have,

$$\mathbb{E}[\gamma^{i_{j+1}}] = \mathbb{E}[\gamma^{i_j + \Delta_j}] = \mathbb{E}[\gamma^{i_j}]\mathbb{E}[\gamma^{\Delta}] = \mathbb{E}[\gamma^{\Delta}]^{j+1}.$$

□

The implication is that the value function in the noisy MDP has an effective discount factor of $\mathbb{E}[\gamma^{\Delta}]$.

3.2 Infinite Horizon Case

We now state and prove the effect of sticky noise on general, infinite horizon MDPs.

Theorem 3.2. *For MDP* $M = (S, A, T, R, \gamma)$*, and sticky version of* M*,* M_β*, with sticky noise parameter* β*, there exists an MDP,* $M' = (S, A, T, R, \frac{\beta\gamma}{(1-\gamma(1-\beta))})$*, with the same optimal policy as* M_β*.*

Proof. The fixed point for the value function for policy π in the sticky MDP must obey:

$$\begin{aligned}
V_\beta^\pi &= R + \gamma P_\beta^\pi V_\beta^\pi \\
V_\beta^\pi &= R + \gamma(\beta P^\pi + (1-\beta)I)V_\beta^\pi \\
V_\beta^\pi - \gamma(1-\beta)V_\beta^\pi - \beta\gamma P^\pi V_\beta^\pi &= R \\
(1-\gamma(1-\beta))V_\beta^\pi - \beta\gamma P^\pi V_\beta^\pi &= R \\
V_\beta^\pi - \frac{\beta\gamma}{(1-\gamma(1-\beta))}P^\pi V_\beta^\pi &= \frac{R}{(1-\gamma(1-\beta))} \\
(I - \frac{\beta\gamma}{(1-\gamma(1-\beta))}P^\pi)V_\beta^\pi &= \frac{R}{(1-\gamma(1-\beta))} \\
V_\beta^\pi &= (I - \frac{\beta\gamma}{(1-\gamma(1-\beta))}P^\pi)^{-1}\frac{R}{(1-\gamma(1-\beta))}.
\end{aligned}$$

Observe that this is equivalent to solving for the value function of M *with the original transition matrix* but replacing γ with with a scaled discount factor, $\frac{\beta\gamma}{(1-\gamma(1-\beta))}$ and reward scaled by $\frac{1}{(1-\gamma(1-\beta))}$. This can be replaced with the unscaled discount factor without changing the optimal policy, completing the proof. □

3.3 Reconciling the infinite and finite horizon cases

If we extend the effective discount factor for the finite horizon case to the infinite horizon, we get:

$$\mathbb{E}[\gamma^{\Delta}] = \sum_{i=1}^{\infty} \beta(1-\beta)^{i-1}\gamma^i = \beta\gamma \sum_{i=0}^{\infty} [\gamma(1-\beta)]^i = \frac{\beta\gamma}{1-\gamma(1-\beta)}.$$

This shows that, in the limit of an infinite number of time steps, the finite horizon analysis is consistent with the infinite horizon case – as it should be. (Note that the first summation starts from 1 because γ^{Δ} includes the waiting time in the first state.)

4 Reset Noise

In this section, we consider a noise model that is identical for each state. For all states and actions, suppose there is a noise distribution μ that is mixed with each state's next state distribution. One way to think of this as some sort of reset noise, by which with some probability $1-\alpha$, the agent is reset to some distribution over states. This probability and the reset state distribution are independent of the agent's current state and action. This can be thought of as a probability of the robot getting "kidnapped" by a human and moved to another location, or a game being reset to a starting configuration. Other realizations of this could include a random outcome in a game (such as going to jail in Monopoly$^{\text{TM}}$) that can happen from any state, the bat in Hunt the Wumpus.

Following the convention of the previous section, we augment the definition of MDPs to M_α, where α determines an amount of reset noise that is added to the transition matrix for every policy. Define P^μ to be an $n \times n$ matrix with μ in each row. For any π and any α, we have:

$$V_\alpha^\pi = \alpha P^\pi + (1-\alpha)P^\mu.$$

Thus, when $\alpha = 1$ we have our original M, and when $\alpha = 0$, all states transition uniformly to all other states under any policy.

Theorem 4.1. *For MDP $M = (S, A, T, R, \gamma)$, and sticky version of M, M_α, with reset noise parameter α, there exists an MDP, $M' = (S, A, T, R, \alpha\gamma)$, with the same optimal policy as M_α.*

Proof. The value function for a policy in this M_α satisfies:

$$\begin{aligned}
V_\alpha^\pi &= R + \gamma P_\alpha^\pi V_\alpha^\pi \\
V_\alpha^\pi &= R + \gamma(\alpha P^\pi + (1-\alpha)P^U)V_\alpha^\pi \\
V_\alpha^\pi &= R + \gamma\alpha P^\pi V_\alpha^\pi + \gamma(1-\alpha)P^\mu V_\alpha^\pi \\
V_\alpha^\pi - \gamma\alpha P^\pi V_\alpha^\pi - \gamma(1-\alpha)P^\mu V_\alpha^\pi &= R \\
(I - \gamma\alpha P^\pi - \gamma(1-\alpha)P^\mu)V_\alpha^\pi &= R \\
V_\alpha^\pi &= (I - \gamma\alpha P^\pi - \gamma(1-\alpha)P^\mu)^{-1}R
\end{aligned}$$

Before simplifying this further, we review the Sherman-Morrison formula:

$$(D + uv^T)^{-1} = D^{-1} - \frac{D^{-1}uv^TD^{-1}}{1 + v^TD^{-1}u}$$

To align with our matrix inversion problem, we consider:

- $D = (I - \gamma\alpha P^\pi)$
- $u = -\gamma(1-\alpha)\mathbf{1}_n$
- $v = \mu$

Given this, we observe that:

- $D\mathbf{1}_n = (1-\gamma\alpha)\mathbf{1}_n \implies D^{-1}\mathbf{1}_n = \frac{1}{1-\gamma\alpha}\mathbf{1}_n$
- $uv^T = -\gamma(1-\alpha)P^\mu$
- $D^{-1}u = -\gamma(1-\alpha)D^{-1}\mathbf{1}_n = -\gamma\frac{1-\alpha}{1-\gamma\alpha}\mathbf{1}_n$
- $v^T D^{-1}u = -\gamma\frac{1-\alpha}{1-\gamma\alpha}\mu^T\mathbf{1}_n = -\gamma\frac{1-\alpha}{1-\gamma\alpha}$
- $1 + v^T D^{-1}u = 1 - \gamma\frac{1-\alpha}{1-\gamma\alpha} = \frac{1-\gamma\alpha-\gamma+\gamma\alpha}{1-\gamma\alpha} = \frac{1-\gamma}{1-\gamma\alpha}$
- $D^{-1}uv^T = -\gamma\frac{1-\alpha}{1-\gamma\alpha}\mathbf{1}_n\mu^T = -\gamma\frac{1-\alpha}{1-\gamma\alpha}P^\mu$
- $\frac{D^{-1}uv^T}{1+v^T D^{-1}u} = \frac{-\gamma\frac{1-\alpha}{1-\gamma\alpha}P^\mu}{\frac{1-\gamma}{1-\gamma\alpha}} = -\gamma\frac{1-\alpha}{(1-\gamma)}P^\mu$.

Using the Sherman-Morrison formula, we get:

$$\begin{aligned}
V_\alpha^\pi &= (I - \gamma\alpha P^\pi - \gamma(1-\alpha)P^\mu)^{-1}R \\
&= (D + uv^T)^{-1}R \\
&= \left(D^{-1} - \frac{D^{-1}uv^T D^{-1}}{1 + v^T D^{-1}u}\right)R \\
&= D^{-1}R - \frac{D^{-1}uv^T}{1 + v^T D^{-1}u}D^{-1}R \\
&= (I - \gamma\alpha P^\pi)^{-1}R + \gamma\frac{1-\alpha}{1-\gamma}P^\mu(I - \gamma\alpha P^\pi)^{-1}R.
\end{aligned}$$

For $M' = (S, A, T, R, \gamma\alpha)$, and V'^π as the value function for policy π in M', we have:

$$\begin{aligned}
V_\alpha^\pi &= V'^\pi + \gamma\frac{1-\alpha}{1-\gamma}P^\mu V'^\pi \\
V_\alpha^\pi &= V'^\pi + \gamma\frac{1-\alpha}{1-\gamma}\overline{v_\mu'^\pi}
\end{aligned}$$

where $\overline{v_\mu'^\pi}$ is the μ-weighted mean state value in V'^π.

So far, we have made a connection between M_α and an MDP with a reduced discount M', but these are not identical MDPs. Since $\overline{v_\mu'^\pi}$ is a policy-dependent offset, we still need to show that a policy that is optimal for M' must also be optimal for M_α. To see this, we first make the following observation about about any average of state values for two policies. Let ρ be column vector corresponding to a distribution over states, then for any π_1 and π_2, where π_1 weakly dominates π_2, and positive constant c, we have:

$$V^{\pi_1} \geq V^{\pi_2} \rightarrow c\rho^T V^{\pi_1} \geq c\rho^T V^{\pi_2}.$$

Now, consider $\rho = \mu$, and $c = \gamma\frac{1-\alpha}{1-\gamma}$, π'^* optimal for M', and any other π:

$$\begin{aligned}
V'^{\pi'^*} &\geq V'^\pi \\
V'^{\pi'^*} + c\mu^T V'^{\pi'^*} &\geq V'^\pi + c\mu^T V'^{\pi'^*} \\
V'^{\pi'^*} + \overline{v_\mu'^{\pi'^*}} &\geq V'^\pi + c\mu^T V'^\pi \\
V'^{\pi'^*} + \overline{v_\mu'^{\pi'^*}} &\geq V'^\pi + \overline{v_\mu'^\pi} \\
V_\alpha^{\pi'^*} &\geq V_\alpha^\pi.
\end{aligned}$$

Therefore π'^* must also be optimal for M_α. □

In summary, we have shown that solving a MDP M_α, which adds reset noise parameterized by α to MDP M is equivalent to solving another MDP M' which has the same transition model T as M, but has the discount factor reduced from γ to $\gamma\alpha$.

Since sticky noise and reset noise each result in an MDP with the same optimal policy but a smaller discount factor, the analysis can be combined sequentially to model MDPs that contain both types of noise.

5 Discussion

We introduced two models for MDP transition noise, and demonstrated that adding these types of noise is *exactly equivalent* to solving a noiseless version of the MDP with a smaller discount factor. Smaller discount factors imply shorter ϵ-horizons, which we believe facilitates the discovery of interpretable policies or value functions. These results parallel recent findings in supervised learning, where noise increases the Rashomon set, thereby making it easier to find interpretable hypotheses.

One might wonder if strong statements like these can be made about more general noise models. The simulation lemma for MDPs (Kearns & Singh, 2002) suggests that it may be difficult to make such claims about arbitrary noise models. The simulation lemma bounds the effects of small changes in transition probabilities on the infinite horizon value function. The bounds are quite large, indicating a maximum change in value that scales with the 1-norm of the transition probability difference, and $\frac{1}{(1-\gamma)^2}$. Getting the tighter and more informative results in this paper required making very specific assumptions about the form of the noise. An interesting question for future work is whether the two noise models considered here are the only two non-trivial models for which this tight coupling between noise and discounting exists.

References

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2023.