The Unintended Consequences of Discount Regularization: Improving Regularization in Certainty Equivalence Reinforcement Learning

Sarah Rathnam [1] **Sonali Parbhoo** [2] **Weiwei Pan** [1] **Susan A. Murphy** [1] **Finale Doshi-Velez** [1]

Abstract

Discount regularization, using a shorter planning horizon when calculating the optimal policy, is a popular choice to restrict planning to a less complex set of policies when estimating an MDP from sparse or noisy data (Jiang et al., 2015). It is commonly understood that discount regularization functions by de-emphasizing or ignoring delayed effects. In this paper, we reveal an alternate view of discount regularization that exposes unintended consequences. We demonstrate that planning under a lower discount factor produces an identical optimal policy to planning using any prior on the transition matrix that has the same distribution for all states and actions. In fact, it functions like a prior with stronger regularization on state-action pairs with more transition data. This leads to poor performance when the transition matrix is estimated from data sets with uneven amounts of data across state-action pairs. Our equivalence theorem leads to an explicit formula to set regularization parameters locally for individual state-action pairs rather than globally. We demonstrate the failures of discount regularization and how we remedy them using our state-action-specific method across simple empirical examples as well as a medical cancer simulator.

1. Introduction

In reinforcement learning (RL), planning under a shorter horizon is a common form of regularization. In the most extreme case, a discount factor of zero results in a contextual bandit setting. Using a reduced or zero discount factor for planning is common in real-world applications such as mobile health (Liao et al. (2020), Trella et al. (2022)), medicine (Oh et al. (2022), Awasthi et al. (2022), Durand et al. (2018)), and education (Cai et al. (2021), Qi et al. (2018)).

In this paper, we analyze discount regularization in the context of *certainty equivalence RL*. This means that the agent takes the estimated model as true when calculating the optimal policy (Goodwin & Sin, 1984). While planning using a reduced discount factor leads to better-performing policies in many cases (Jiang et al., 2015; Amit et al., 2020), our main contribution is to present a deeper conception of this method that reveals limitations. We do so by first proving that discount regularization produces the same optimal policy as averaging the transition matrix for each action with a transition matrix in which all rows are the same. This can also be viewed in terms of a prior on the transition matrix.

As further contributions, we utilize our reframing to expose unintended consequences. One such consequence is that the magnitude of the prior implied by discount regularization is higher for state-action pairs with more transition observations in the data and vice versa. This is generally not desirable as we want stronger regularization on states that we have observed less, and want to rely on the data in states where we have more. Another negative aspect we expose is the assumption of equal transition distributions for all state-action pairs, which is inappropriate in many contexts.

We also offer solutions to the problems exposed above in order to tailor regularization to the task at hand– both the data set and the environment. To mitigate the issue of inconsistent prior magnitudes in data sets with uneven exploration, we derive a state-action specific formula for the regularization parameter. Furthermore, the method by which we derive this parameter can be adapted to other priors to match the transition dynamics of the environment.

Finally, we demonstrate our results empirically on tabular examples and on a medical cancer dynamics simulator. First, we empirically confirm that discount regularization and a uniform prior on the transition matrix yield identical optimal policies. We then demonstrate that a uniform prior with fixed magnitude across state-action pairs outperforms discount regularization across environments. We also show that

[1]Harvard University, School of Engineering and Applied Sciences, Cambridge, MA USA [2]Imperial College London, London UK. Correspondence to: Sarah Rathnam <sarah_rathnam@g.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

our state-action-specific regularization parameter reduces loss without parameter tuning.

2. Related Works

Jiang et al. (2015) demonstrate that planning under a shorter horizon often yields policies that outperform ones learned using the true discount factor, even when both are evaluated in the true environment. They prove that using a lower planning discount factor restricts the planning to a less complex set of policies, thereby avoiding overfitting. They further demonstrate that the benefit of a lower discount factor is increasingly pronounced in cases where the model is estimated from a smaller data set. Amit et al. (2020) refer to this concept as "discount regularization," a term which we use here. Unlike these works, we provide means to connect discount regularization with placing a prior on the transition matrix.

While a Bayesian prior encodes expert knowledge, information from previous studies, or other outside information, we can also view a prior as a form of regularization since it forces the model not to overfit when data is limited (Poggio & Girosi, 1990; Ghavamzadeh et al., 2015). This is a flexible tool that allows us to regularize in a way that matches our prior knowledge and beliefs about the environment. In model-based Bayesian RL, the problem is often framed as a Bayes-Adaptive MDP (BAMDP), an MDP where the states are replaced by "hyperstates" that reflect the original state space combined with the posterior parameters of the transition function (Duff, 2002). In general, Bayesian RL algorithms do not explicitly address planner overfitting; rather they incorporate the probability distribution over models, causing the planner not to overfit to an uncertain model. For example, model-based Bayesian RL methods draw sample models from the posterior (Asmuth et al., 2012), sample hyperstates (Poupart et al., 2006), or apply an exploration bonus based on the amount of data (Kolter & Ng, 2009) or based on the variance of the parameters (Sorg et al., 2012). The BAMDP framework can also be extended to the case of partial observability (Ross et al., 2007; 2011). In this paper, we consider planning using the posterior mean of the transition matrix under a Dirichlet prior as a regularized form of the transition matrix, which is a common choice in model-based RL, e.g. Vlassis et al. (2012); O'Donoghue et al. (2020).

Previous works also discuss the limitations of a fixed discount factor and present approaches for more flexible discounting, for example state-dependent (Wei & Guo, 2011; Yoshida et al., 2013), state-action-dependent (Pitis, 2019), and transition-based discounting (White, 2017). We add to this work by demonstrating that discount regularization carries implicit assumptions of equal transition distributions for all state-action pairs and stronger regularization on those with more transition data.

Finally, in Arumugam et al. (2018), planning is conducted over the set of epsilon-greedy policies rather than deterministic policies. The additional stochasticity during planning prevents tailoring the policy too closely to the model. We show how the work by which we connect discount regularization to a Dirichlet prior by using a weighted average form of the transition matrix applies to epsilon-greedy regularization as well. This connection allows us to directly compare the methods in terms of transition matrix MSE to identify the right method for the environment. Like with Dirichlet prior, it also allows us to compute a state-action-specific parameter to control the amount of regularization.

3. Background and Notation

Markov Decision Process We consider a finite, discrete Markov Decision Process (MDP). An MDP M is characterized by $< S, A, R, T, \gamma >$, defined as follows. S: State space of size N. A: Action space. $R(s, a)$: Reward, as a function of state s and action a. $T(s, a)$: Transition function, mapping each state-action pair to a probability distribution over successor states. We assume T is unknown and estimated from the data. γ: Discount factor, $0 \leq \gamma < 1$.

Certainty Equivalence Certainty equivalence is a useful approach to offline model-based RL. The agent takes the estimated model as accurate when finding the optimal policy. It separates the estimation of the model from the policy optimization (Goodwin & Sin, 1984). The maximum likelihood estimate (MLE) is a natural choice for the model estimate, however maximum likelihood solutions can overfit, particularly in the case of small data sets (Murphy, 2012). Often, a better policy is obtained by regularizing the MDP before learning the certainty equivalence policy.

4. A Common Form: Regularization as a Weighted Average Transition Matrix

The analyses that follow stem from framing each method in a common form: a weighted average transition matrix. We demonstrate that discount regularization and the posterior mean of the transition matrix under a Dirichlet prior can both be expressed as a weighted average between the MLE transition matrix and a regularization matrix.

Dirichlet Prior on T As discussed in Sec. 2, a Dirichlet prior on the transition matrix T functions as a flexible form of regularization. Given a prior on T for state-action pair (s_n, a_k), $T_{\text{prior}}(s_n, a_k) \sim \text{Dirichlet}(\alpha_{n,k,1}, ..., \alpha_{n,k,N})$, the posterior mean functions as a regularized form. Though simple, this generates several important insights that deepen our understanding and facilitate better regularization.

Let $\langle c_{n,k,1}, ..., c_{n,k,N} \rangle$ be the transition count data observed from state s_n to states 1 through N under action a_k. It can be easily shown that posterior mean of the transition matrix $\hat{T}_{\substack{\text{post} \\ \text{mean}}}$ is equal to a weighted average of the MLE transition matrix and the mean of the prior:

$$\hat{T}_{\substack{\text{post} \\ \text{mean}}}(s_n, a_k) = (1-\epsilon)\hat{T}_{MLE}(s_n, a_k) + \epsilon T_{\substack{\text{prior} \\ \text{mean}}}(s_n, a_k) \tag{1}$$

where $\epsilon = \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}$.[1][2]

Discount Regularization Next we show that discount regularization is mathematically equivalent to replacing the transition matrix with the weighted average between that transition matrix and a matrix of zeros. Although this form is unusual as it is not a true transition matrix, we will show that it has utility in relating the amounts of regularization between methods.

To cast discount regularization in certainty-equivalence RL as a weighted average transition matrix, consider the Bellman equation for the value of each state under policy π, $V^\pi = R_\pi + \gamma T_\pi V^\pi$, where the vector V^π is the value of each state, R_π is the vector of rewards, and T_π is the transition matrix, all under policy π. Let $\gamma_p < \gamma$ be the planning discount factor, the lower discount factor used for regularization when calculating the certainty-equivalence policy. Then we have the Bellman equation $V^\pi = R_\pi + \gamma_p T_\pi V^\pi$. We rewrite the product $\gamma_p T_\pi$ from the Bellman equation as the product of true discount factor γ and a weighted average matrix: $\gamma_p T_\pi = \gamma[(1-\epsilon)T_\pi + \epsilon T_{\text{zeros}}]$, where T_{zeros} is an appropriately sized matrix of zeros and $\epsilon = \frac{\gamma - \gamma_p}{\gamma}$.

Using this insight, when estimating the transition matrix from data, we can use the following weighted average transition matrix and the true discount factor γ for planning in place of the MLE transition matrix and lower discount factor γ_p.

$$\hat{T}_{\substack{\text{disc} \\ \text{reg}}}(s,a) = (1-\epsilon)\hat{T}_{\text{MLE}}(s,a) + \epsilon T_{\text{zeros}} \tag{2}$$

where $\epsilon = \frac{\gamma - \gamma_p}{\gamma}$.

Eq. 2 also provides another way to view discounting as "partial termination" (Sutton & Barto, 2018). According to this classic interpretation, the sum of discounted rewards can be viewed as the sum of undiscounted rewards partially terminating with degree 1 minus the discount factor at each step. Similarly, Eq. 2 with $\gamma = 1$ represents the agent terminating with probability $\epsilon = 1 - \gamma_p$ at each step.

In the following sections, we prove that discount regularization and averaging the transition matrix for each action with a transition matrix in which all rows are the same produce the same optimal policy for the same value of ϵ. Equating our two expressions for ϵ from Eqs. 1 and 2 generates a formula for the magnitude of an empirical Bayes prior on $T(s,a)$ implied by any reduced discount factor γ_p.

[1] Please see Appendix A.2 for derivation.

[2] State-action pair index on $\epsilon_{n,k}$ omitted for readability.

5. Equivalent Policy for Discount Regularization and Dirichlet Prior

5.1. Equivalence Theorem and Proof

The simplicity of Eqs. 1 and 2 allows for direct comparison between the two regularization methods. In fact, discount regularization produces the same optimal policy as averaging the transition matrix with a regularization matrix that is the same for all states and actions when both methods use the same value of ϵ. This result is stated more precisely in Thm. 1 and illustrated in Fig. 1.

Theorem 1. *Let M_1 and M_2 be finite-state, infinite horizon MDPs with identical state space, action space, and reward function and same discount rate $\gamma < 1$. Let $0 < \epsilon \leq 1$, and let $T_{reg}(s,a)$ be any matrix used for regularization that is the same for all (s,a) (i.e. identical rows).*

If M_1 has transition function T and uses discount rate $(1-\epsilon)\gamma$ in planning and M_2 has transition function $(1-\epsilon)T + \epsilon T_{reg}$, and uses discount rate γ in planning, then M_1 and M_2 have the same optimal policy.

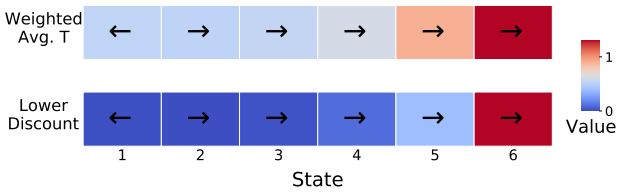


Figure 1: River Swim MDP described in Sec 7.1. Planning with lower discount rate or weighted average T yield different values (colors), but the same optimal policy (arrows).

Proof. The proof is structured as follows. (1) The optimal policy for all MDPs whose Bellman optimality equations differ only by added constant c are the same. (2) The Bellman optimality equation for an MDP in which the transition matrix is regularized by taking its weighted average with a matrix T_{reg} can be written in terms of a lower discount factor and an added constant. (3) Setting the constant from the previous step to 0, the optimal policy of the resulting MDP is the same as that of the original MDP. (4) The resulting Bellman equation is that of an MDP with the original unregularized transition matrix and reduced discount factor $(1-\epsilon)\gamma$.

(1) Consider Bellman's optimality equation for any arbitrary state s and action a for an MDP in which constant c is added

to every reward $r(s,a)$:

$$Q^*(s,a) = r(s,a)+c+\gamma \sum_{s'} T(s,a,s')\text{max}_{a'}(s',a')Q^*(s,a)$$

It is a known result that the optimal policy of an MDP is not affected by adding the same constant c to all rewards $r(s,a)$. (See, for example, Ng et al. (1999): "constant offets of the reward do not affect the optimal policy when $\gamma < 1$".)

It follows that the optimal policy $\pi_{\text{opt}}(s) = \text{argmax}_a Q^*(s,a)$ is the same for all values of c. So for all values of constant c, the MDP with the Bellman optimality equation above has the same optimal policy.

(2) Let $T_{reg}(s,a)$ be a transition matrix that is the same for all (s,a). We show that Bellman's optimality equation for a transition matrix regularized by taking its weighted average with the T_{reg} can be written in terms of a scaled discount factor and added constant.

$$\begin{aligned} Q^*(s,a) &= r(s,a) + \gamma \sum_{s'} [((1-\epsilon)T(s,a,s') \\ &\quad + \epsilon T_{reg}(s,a,s'))\text{max}_{a'} Q^*(s',a')] \end{aligned}$$

$$\begin{aligned} &= r(s,a) + \gamma(1-\epsilon) \sum_{s'} T(s,a,s')\text{max}_{a'} Q^*(s',a') \\ &\quad + \gamma\epsilon \sum_{s'} T_{reg}(s,a,s')\text{max}_{a'} Q^*(s',a') \end{aligned}$$

Letting $c(s,a) = \gamma\epsilon \sum_{s'} T_{reg}(s,a,s')\text{max}_{a'} Q^*(s',a')$, Bellman's optimality equation is:

$$\begin{aligned} Q^*(s,a) &= r(s,a) + c(s,a) \\ &+ \gamma(1-\epsilon) \sum_{s'} T(s,a,s')\text{max}_{a'} Q^*(s',a') \end{aligned}$$

By the assumptions of Thm. 1, $T_{reg}(s,a,s')$ is the same for all (s,a) and is therefore a function of s' only. $\text{max}_{a'} Q^*(s',a')$ is also a function of s' only. Therefore $c(s,a)$ is actually a constant number, which we can call c.

$$c = \gamma\epsilon \sum_{s'} \underbrace{T_{reg}(s,a,s')}_{\text{func. of s' only}} \underbrace{\text{max}_{a'} Q^*(s',a')}_{\text{func. of s' only}} = \text{constant}$$

(3) By (1), replacing c with 0, the resulting new MDP with Bellman optimality equation

$$Q^*(s,a) = r(s,a)+\gamma(1-\epsilon) \sum_{s'} T(s,a,s')\text{max}_{a'} Q^*(s',a')$$

has the same optimal policy.

(4) This resulting Bellman equation in (3) is that of the MDP with the original, unregularized transition matrix $T(s,a,s')$ and discount factor $\gamma(1-\epsilon)$. Therefore, the MDP with discount rate γ and transition matrix $(1-\epsilon)T(s,a,s') + \epsilon T_{reg}(s,a,s')$ and the MDP with discount rate $\gamma(1-\epsilon)$ and transition matrix $T(s,a,s')$ have identical optimal policies. □

Thm. 1 provides a deeper understanding of how discount regularization functions. At maximum regularization, $\gamma_p = 0$ or equivalently $\epsilon = 1$, it unites two views of the relationship between bandits and MDPs. One common view of a contextual bandit is an MDP with γ = 0 (Agarwal et al., 2019). Alternatively, a contextual bandit is an MDP in which "the transition probability is identical... for all states and actions" (Zanette & Brunskill, 2018). Our proof extends this equivalence beyond the bandit setting to all amounts of regularization.

Thm. 1 also reveals the limitations of discount regularization. First, the regularization matrix is the same regardless of the state and action, so it will be biased in environments where transition probabilities vary greatly based on the state and/or the action. Furthermore, as we demonstrate in the next section, this theorem leads to the result that discount regularization provides stronger regularization on state-action pairs with more data.

5.2. Dirichlet Prior Implied by Discount Regularization

We showed that discount regularization produces the same optimal policy as averaging the transition matrix with any matrix that is the same for all states and actions. Recall from Sec. 4 that a Dirichlet prior on $T(s,a)$ also results in a weighted average transition matrix form. In this section, we further expand on this relationship and will see that using state-action visitation rates from the data allows us to produce an empirical Bayes prior on $T(s,a)$ that results in the same optimal policy as discount regularization.

Using the equivalence in Thm. 1, we derive the prior magnitude that produces the same optimal policy for any planning discount rate. Since discount regularization employs the same planning discount rate and consequently the same value of ϵ for every state-action pair, the prior that produces an equivalent policy also has the same value of ϵ at every state-action pair. Using this equivalence and the two separate formulas for ϵ in Eqs. 1 and 2 yields a formula for the magnitude of prior implied by any value of planning discount factor γ_p. Setting the two formulas for ϵ equal and solving for $\sum_{i=1}^{N} \alpha_{n,k,i}$, we see that a lower planning discount factor implies a prior whose magnitude depends on the number of transitions from (s_n, a_k) in the data. [3]

$$\sum_{i=1}^{N} \alpha_{n,k,i} = \left(\frac{\gamma - \gamma_p}{\gamma_p}\right) \sum_{j=1}^{N} c_{n,k,j} \tag{3}$$

[3] Please see Appendix A.2.1 for details.

In the case of a uniform prior, which we take as the example prior in our simulations, the magnitude simplifies to

$$\alpha_{n,k,i} = \left(\frac{\gamma - \gamma_p}{\gamma_p}\right) \frac{\sum_{j=1}^{N} c_{n,k,j}}{N} \forall i$$

The relationship between uniform prior magnitude $\alpha_{n,k,i}$ and planning discount factor γ_p for an individual state-action pair is illustrated in Fig. 2. Furthermore, Eq. 3 shows us that, for any planning discount factor γ_p, the magnitude of the corresponding Dirichlet prior is higher for state-action pairs with more data. In other words, those (s, a) with more observations in the data are regularized more. Especially for data sets with uneven distribution of transition data, it may be better to use a more flexible regularization method. In Sec. 6, we use our framework to introduce state-action-specific regularization to mitigate this issue. Note also that the special case of $\gamma_p = 0$, the contextual bandit setting, presents an exception as the implied priors for all (s, a) are of infinite magnitude. This case is fundamentally different as the future is not just discounted but rather completely ignored.

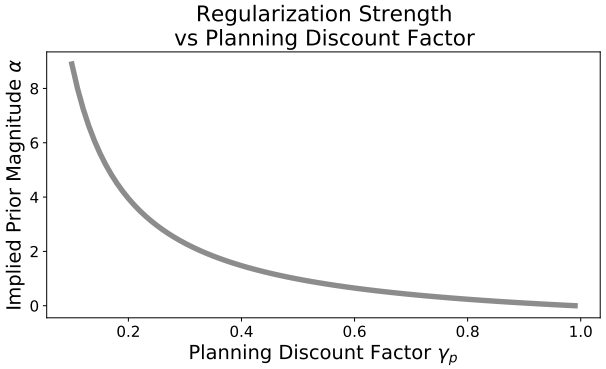


Figure 2: Magnitude of uniform Dirichlet prior implied by planning discount factor γ_p for MDP with 10 states, 20 transition observations per state, and γ = 0.99.

6. State-Action-Specific Regularization without Parameter Tuning

We exposed in Eq. 3 that discount regularization functions like a prior on the transition matrix with a potentially undesirable magnitude. To avoid this behavior, we return to the weighted average form introduced in Sec. 4 to derive a formula for state-action-specific regularization. Using this form, we calculate the MSE of the estimated transition matrix and identify the value of regularization paramter ϵ that minimizes the transition matrix MSE. While we recognize that a good transition matrix estimate does not guarantee a good policy, it is a reasonable step towards that goal.

We derive the closed-form expression of the MSE for the case of a uniform Dirichlet prior. We take $\text{MSE}(\hat{T}(s, a))$ to be the sum of the MSE of the individual elements. We provide the derivation using the bias-variance decomposition of MSE in Appendix B and the resulting form below. Let $\hat{T}_{\text{unif}}$ be the posterior mean of T under a uniform Dirichlet prior. Then,

$$\begin{aligned}&\text{MSE}[\hat{T}_{\text{unif}}(s_n, a_k)] = \\ &\underbrace{\sum_{i=1}^{N} (1-\epsilon)^2 \frac{1}{c_{n,k}} T(s_n, a_k, s_i)(1 - T(s_n, a_k, s_i))}_{variance} \\ &\underbrace{+ \epsilon \left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2}_{bias}\end{aligned} \tag{4}$$

where $c_{n,k}$ is the number of transition observations starting at state s_n under action a_k. Let ϵ^* to be the value of the regularization parameter ϵ calculated by minimizing the MSE equation. Then,

$$\epsilon^*(s_n, a_k) = \frac{K(s_n, a_k)}{K(s_n, a_k) + c_{n,k}} \tag{5}$$

where $K(s_n, a_k) = \frac{\sum_{i=1}^{N} T(s_n, a_k, s_i)(1 - T(s_n, a_k, s_i))}{\sum_{i=1}^{N} (\frac{1}{N} - T(s_n, a_k, s_i))^2}$.

The first term of Eq. 4 is the contribution of the MLE's variance to the error, in this case the only source of variance. The second term represents the bias introduced by regularization. The strength of regularization ϵ controls the trade-off between the bias and variance. The variance is driven by the amount of data $c_{n,k}$ both through its role in setting the amount of regularization ϵ^* and as a factor inversely impacting the variance term. Both bias and variance are impacted by the true transition distribution $T(s, a)$. A deterministic $T(s, a)$ maximizes bias for a given ϵ, but results in $\epsilon^* = 0$ (since $T(s, a, s')(1 - T(s, a, s')) = 0$ for all s'). At the other extreme, a $T(s, a)$ with uniform distribution maximizes the variance for a given ϵ but has no bias, so we default to $\epsilon^* = 1$. Intermediate values of ϵ^* trade off between bias and variance. Of course in practice, the true transition matrix T is generally not known.

A uniform prior on $T(s, a)$ with state-action-specific parameter ϵ^* improves upon discount regularization by setting the parameters locally for each state-action pair rather than forcing one global regularization parameter. Furthermore, there is no parameter tuning required, simply a plugin estimate for T (e.g. the MLE). In practice, we may worry that in the low data regimes in which regularization is required, the estimate of T will not be good enough to estimate ϵ^*. Nonetheless, our empirical examples in Sec. 7 demonstrate

that our formula for ϵ^* leads to a reduction in loss over a single global regularization parameter.

Note that the state-action-specific parameter ϵ^* combined with regularization matrix T_{reg} does not map directly to a state-action-specific discount factor. The expression for $c(s, a)$ in the proof of Thm 1 must be constant for the two methods to produce the same optimal policy and the state-action-specific discount factor breaks this equivalence.

7. Simulation Results

We have demonstrated that planning under a reduced discount factor functions as a prior on the transition matrix with higher magnitude for state-action pairs with more transition observations. We then proposed a better way to regularize by deriving an explicit formula for a uniform prior that minimizes that transition matrix MSE locally for each state-action pair. Next we confirm our results empirically.

First we demonstrate that the equality in Thm. 1 holds. We then compare the performance of (1) discount regularization, (2) a uniform prior on T with equal magnitude for all state-action pairs, and (3) our state-action-specific regularization on three simple tabular examples and a medical cancer simulator.

7.1. Tabular Environments

We demonstrate our results on three common environments from the RL literature. The first comes from the initial work proposing discount regularization. We choose this environment to demonstrate the limitations of discount regularization even in an environment where it is known to be beneficial. We choose the other two because of their differences in structure, connectivity, and rewards to ensure that our results hold in diverse environments.

10-State Random Chain The first environment is a distribution over MDPs and we sample one before generating each data set in the examples that follow. Jiang et al. (2015) empirically demonstrated the benefits of discount regularization on this randomly generated 10-state, 2-action MDP. For each state-action pair, 5 successor states are chosen at random to have nonzero transition probability. These probabilities are drawn independently from Uniform[0,1] and normalized to sum to one. The rewards are sampled independently from Uniform[0,1].

River Swim This common tabular environment described in Osband et al. (2013) consists of six states and two actions, as illustrated in Figure 3. The agent can attempt to swim right "against the current" towards the larger reward, or swim left with probability 1 towards the smaller reward.

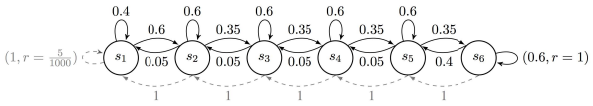


Figure 3: River Swim. Image from Osband et al. (2013).

Loop The "Loop" environment from Strens (2000) consists of nine states forming two loops, joined by a single state. Two actions "a" and "b" traverse the loops as indicated in Figure 4. A reward of 0, 1 or 2 is received at each time step, as indicated in Fig. 4. To add stochasticity to the transitions, we assume that at each time step the agent acts according to the desired action with probability .5 and chooses random between the actions with probability .5.

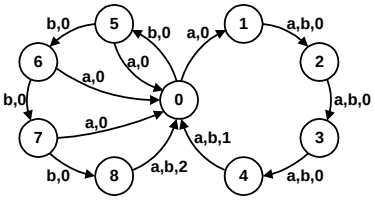


Figure 4: Loop Environment. Image from Strens (2000).

7.2. Procedure

To assess performance in each environment, we follow the procedure in Jiang et al. (2015). We repeatedly sample data sets from the true MDP. (A new MDP is sampled every time in the case of the 10-State Random Chain.) For each, we estimate the transition matrix from the data and assume the reward function is known. Then for a range of regularization strengths (ϵ or γ) we regularize the transition matrix separately using (1) discount regularization or (2) a uniform prior with constant magnitude across state-action pairs. We also regularize by (3) a uniform prior with state-action specific parameter. We then calculate the optimal policy. We compute the loss as the difference between the value of the true optimal policy in the true MDP and the value of the policy found in the estimated, regularized MDP, evaluated in the true MDP. The state-action-specific uniform prior is not dependent on a regularization parameter so we plot the single loss value horizontally.

7.3. Discount Regularization and Uniform Prior on Transition Matrix Yield Identical Optimal Policies

First, we empirically confirm our result from Thm. 1. When the implied value of ϵ is the same for all state-action pairs, a uniform prior on T will yield the same optimal policy as a planning discount factor of $\gamma(1 - \epsilon)$. As per Eq. 3, we

enforce equal ϵ across state-action pairs by sampling data sets with equal numbers of transition observations across state-action pairs. As demonstrated for the 10-State Random Chain environment in Fig. 5, loss is identical for both methods, as is expected for identical policies.

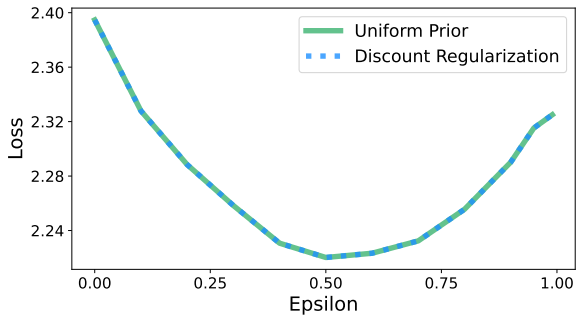


Figure 5: Discount regularization and a uniform prior on the transition matrix result in identical policies when transition count data are equal for all state-action pairs.

In the examples that follow, we relax the requirement of equal data across state-action pairs to compare methods under a more realistic data distribution.

7.4. Exposing Problems with Discount Regularization

Discount Regularization performs poorly on data sets with uneven coverage across state-action pairs. In real-world conditions, it is unlikely that a data set will have equal numbers of transition observations across state-action pairs. In this case, recall that discount regularization functions as a prior with higher magnitude for state-action pairs with more data (Eq. 3). We compare this with a uniform prior on the transition matrix with equal magnitude for all state-action pairs. Fig. 6 shows the loss for each method across a range of values of ϵ (regularization strengths) for the three tabular environments. In these examples, the transition data is generated as tuples (s, a, r, s') with starting state and action chosen uniformly at random, but not enforced to be equal across state-action pairs. Even with transition data that is not heavily skewed away from uniform, the uniform prior with fixed magnitude generates policies that perform better (lower loss) in the true environment across a range of regularization strengths.

Discount regularization performs poorly when the transition distribution differs greatly across states and/or actions. In addition to poor performance in skewed data sets, discount regularization does not perform well in cases where the implied prior, equal for all (s, a), does not match the ground truth. For example, a domain expert may have knowledge that some state transitions are likely or others are impossible. Consider the case of River Swim. If a domain expert knows that Action 1 generally causes the agent to go left and Action 2 generally causes the agent to go right, we may choose a different prior on each action, where the prior on Action 1 deterministically moves the agent left and the prior on Action 2 deterministically moves the agent right. Fig. 7 compares the loss for this deterministic "left/right prior" with the other methods. Unsurprisingly, this hand-chosen prior results in lower loss than the methods which assume equal transition distributions for all states and actions.

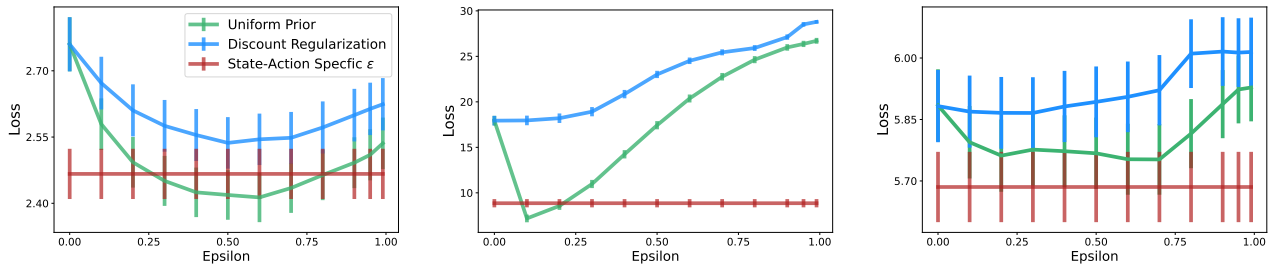
7.5. Simple and Flexible Parameter Tuning

Performance depends not only on choosing an appropriate regularizer for the data set and environment but also on setting the parameters correctly. We now show how the weighted average transition matrix view of regularization gives a straightforward way to set the regularization level, ϵ, in a state-action-specific way that is easily implemented without cross-validation.

Our method avoids parameter tuning. Minimizing the transition matrix MSE equation with respect to regularization parameter ϵ yields an explicit formula for the parameter ϵ^*, Eq. 5. This expression for ϵ^* depends inversely on the number of transition observations in the data, which allows for reduction in regularization with increased data. The only quantity we lack is an estimate for T, which can be approximated by the MLE. Alternatively, we can model T from the data then sample from the posterior, choosing ϵ to minimize the MSE (Eq. 4) across the sampled estimates of T. This is preferable to cross validation not only because it provides a simple, analytic form, but also because the situations in which regularization is beneficial generally involve few transition observations per state-action pair, resulting in insufficient amounts of data to divide into training and validation sets.

Our method remedies the issue of stronger regularization for state-action pairs with more data. Because the formula for ϵ^* is state-action-specific, it allows the flexibility to adjust the regularization amount separately across state-action pairs with different amounts of data and different transition distributions. This is particularly important as most real-world data sets have uneven distributions and requiring equal regularization across state-action pairs in that case impedes performance.

Returning to Fig. 6, we demonstrate that our state-action-specific regularization reduces loss without parameter tuning. The horizontal line for "State-action-specific ϵ" represents the loss when regularization parameter ϵ is set separately for each state-action pair. A state-action-specific regularization parameter yields loss that outperforms dis-



(a) 10-State Random Chain (b) River Swim (c) Loop

Figure 6: A uniform prior on the transition matrix outperforms discount regularization in all three environments. A state-action-specific uniform prior performs close to or better than a uniform prior with global regularization parameter ϵ.

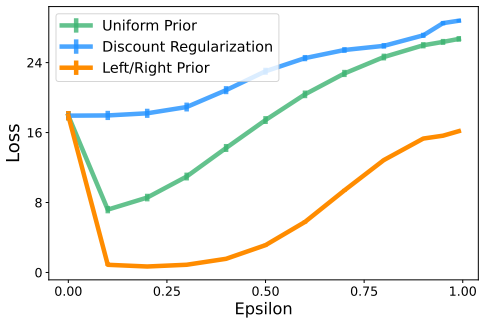


Figure 7: When a uniform prior is not appropriate, a prior chosen based on expert knowledge of the environment can perform better.

count regularization and is close to or outperforms a uniform prior of constant magnitude as well.

In conclusion, we replace cross-validation approaches to parameter tuning with a simple analytical formula that requires only a plug-in estimate of T. The regularization parameter is calculated for an individual state-action pair, which allows flexible regularization for uneven data distributions.

7.6. Cancer Simulation

We confirm our analysis in a larger, more realistic setting, using a cancer simulator developed by Ribba et al. (2012), as implemented by Gottesman et al. (2020). The simulator is based on data from patients with a type of tumor called low-grade gliomas (LGG). We use the version for chemotherapy drug TMZ. The structure of the model is based on 21 patients and parameters for the TMZ version are fit using data from 24 patients, with the remaining 96 held out for validation.

The state space consists of four dimensions: measurements for three different tumor tissue types and the drug concentration. We discretize the states by dividing each dimension into quartiles. The two actions represent whether or not to administer the chemotherapy drug TMZ at each time step, which represents one month. The reward at each time step is the reduction in total tumor size from the previous time step, minus a penalty for administering treatment at that time step. In the batch data, treatment at each time step is determined by a draw from the binomial distribution with treatment probability p. We compare regularization methods across a range of parameter choices: amount of stochasticity in the transition between states, magnitude of penalty to the reward for administering chemotherapy, noise in the starting state, and probability p of treatment in the batch data.

As in the previous examples, we compare the loss of the policies generated by discount regularization, a uniform prior on T, and state-specific uniform prior. Across variations in parameters, the two methods with global parameters performed similarly. A risk of both global methods in this case is that if ϵ is set incorrectly, the loss can be significantly higher. This makes state-action specific regularization particularly compelling, achieving loss near the minimum of all methods with the parameters set globally, but without tuning.

8. Discussion

Extension to Epsilon-Greedy Regularization One issue we address is that discount regularization's implicit prior does not match the ground truth of many environments. As discussed in the example of the "left/right prior" for River Swim, a uniform prior is not always appropriate. In cases like this, we can extend weighted average form to calculate the MSE and state-action-specific regularization parameter for other methods. One example is epsilon-greedy regularization, described in Sec. 2. For this method, the agent treats actions as epsilon-greedy during planning, transition-

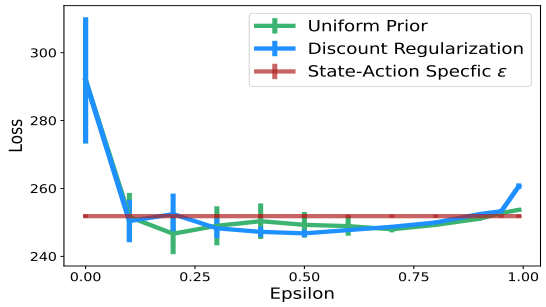


Figure 8: Cancer Simulator: State-action-specific regularization achieves near-minimum loss while avoiding the high loss resulting from incorrectly-set parameters.

ing according to the transition matrix for the greedy action with probability $(1-\epsilon)$ otherwise choosing uniformly at random between the transition matrices for all actions. Expressing epsilon-greedy regularization as a weighted average transition matrix allows us to calculate the MSE and state-action-specific parameter, which we do in Appendix D.2. The common form also facilitates comparison between regularization methods. We can easily compare methods by comparing the regularization matrix that is averaged with the MLE in each case, and choose the regularizer that best matches the environment.

Extension to Model-Free RL While we discussed Thm. 1 in the context of model-based algorithms, the proof applies to model-free methods as well. To extend our method, we can incorporate a weighted average transition matrix into a model-free method such as Q-learning by drawing random transitions from T_{reg} with probability $\epsilon^*(s,a)$ calculated from Eq. 5 and updating the Q function with the random transitions. A sample algorithm and results are presented in Appendix D.4. The algorithm achieves higher rewards compared to standard Q-learning in many environments.

Limitations As stated in Section 6, our method is based on learning a transition model with low MSE, which does not guarantee a good policy. In other words, it is possible to learn a good policy from a "bad" transition model and vice versa, in particular because certain errors in the model may affect the policy more than others. For example, the value equivalence literature demonstrates improved performance by learning a model that minimizes "value-equivalence loss"– a loss metric based on the Bellman operators induced by the model– rather than loss based on maximum likelihood estimation of the model (Grimm et al., 2020; 2021; 2022). We also note that our results are limited to a discrete state space as Thm. 1 applies only to the discrete case so we cannot make guarantees of similar results in a continuous state space. Naively discretizing by quantiles poses issues and can be improved upon by methods such as tile coding (Stone & Sutton, 2001). Extending our method to the continuous case is a topic of continuing work.

9. Conclusion

Discount regularization is a commonly used technique for dealing with noisy and sparse data. Although practitioners believe that they are simply ignoring delayed effects, we revealed through a simple reframing of discount regularization as a weighted average transition matrix that it implicitly assumes a prior on the transition matrix that has the same distribution for all states and actions. Problematically, the magnitude of the prior is higher for state-action pairs with more data. To remedy the issue, we used the weighted average form to derive an explicit formula for the regularization parameter that is calculated locally for each state-action pair rather than globally. Future work will explore the extension of our algorithm to model-free and continuous state space methodologies.

10. Acknowledgements

Research reported in this work was supported by NIH grants P50DA054039, P41EB028242, and 5R01MH123804-02.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2007076. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

Amit, R., Meir, R., and Ciosek, K. Discount factor as a regularizer in reinforcement learning. In *International conference on machine learning*, pp. 269–278. PMLR, 2020.

Arumugam, D., Abel, D., Asadi, K., Gopalan, N., Grimm, C., Lee, J. K., Lehnert, L., and Littman, M. L. Mitigating planner overfitting in model-based reinforcement learning. *arXiv preprint arXiv:1812.01129*, 2018.

Asmuth, J., Li, L., Littman, M. L., Nouri, A., and Wingate, D. A bayesian sampling approach to exploration in reinforcement learning. *arXiv preprint arXiv:1205.2664*, 2012.

Awasthi, R., Guliani, K. K., Khan, S. A., Vashishtha, A., Gill, M. S., Bhatt, A., Nagori, A., Gupta, A., Kumaraguru, P., and Sethi, T. Vacsim: Learning effective strategies for covid-19 vaccine distribution using reinforcement learning. *Intelligence-Based Medicine*, pp. 100060, 2022.

Cai, W., Grossman, J., Lin, Z. J., Sheng, H., Wei, J. T.-Z., Williams, J. J., and Goel, S. Bandit algorithms to personalize educational chatbots. *Machine Learning*, 110(9):2389–2418, 2021.

Duff, M. O. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pp. 67–82. PMLR, 2018.

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A., et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

Goodwin, G. and Sin, K. Adaptive filtering prediction and control,(book) prentice-hall. *Englewood Cliffs*, 6(7): 45, 1984.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pp. 3658–3667. PMLR, 2020.

Grimm, C., Barreto, A., Singh, S., and Silver, D. The value equivalence principle for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5541–5552, 2020.

Grimm, C., Barreto, A., Farquhar, G., Silver, D., and Singh, S. Proper value equivalence. *Advances in Neural Information Processing Systems*, 34:7773–7786, 2021.

Grimm, C., Barreto, A., and Singh, S. Approximate value equivalence. *Advances in Neural Information Processing Systems*, 35:33029–33040, 2022.

Jiang, N., Kulesza, A., Singh, S., and Lewis, R. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1181–1189. Citeseer, 2015.

Kolter, J. Z. and Ng, A. Y. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pp. 513–520, 2009.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

O'Donoghue, B., Osband, I., and Ionescu, C. Making sense of reinforcement learning and probabilistic inference. *arXiv preprint arXiv:2001.00805*, 2020.

Oh, S. H., Park, J., Lee, S. J., Kang, S., and Mo, J. Reinforcement learning-based expanded personalized diabetes treatment recommendation using south korean electronic health records. *Expert Systems with Applications*, 206:117932, 2022.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/6a5889bb0190d0211a991f47bb19a777-Paper.pdf.

Pitis, S. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7949–7956, 2019.

Poggio, T. and Girosi, F. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

Poupart, P., Vlassis, N., Hoey, J., and Regan, K. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 697–704, 2006.

Qi, Y., Wu, Q., Wang, H., Tang, J., and Sun, M. Bandit learning with implicit feedback. *Advances in Neural Information Processing Systems*, 31, 2018.

Ribba, B., Kaloshi, G., Peyre, M., Ricard, D., Calvez, V., Tod, M., Čajavec-Bernard, B., Idbaih, A., Psimaras, D., Dainese, L., et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapya tumor growth inhibition model for low-grade glioma. *Clinical Cancer Research*, 18(18):5071–5080, 2012.

Ross, S., Chaib-draa, B., and Pineau, J. Bayes-adaptive pomdps. *Advances in neural information processing systems*, 20, 2007.

Ross, S., Pineau, J., Chaib-draa, B., and Kreitmann, P. A bayesian approach for learning and planning in partially observable markov decision processes. *Journal of Machine Learning Research*, 12(5), 2011.

Sorg, J., Singh, S., and Lewis, R. L. Variance-based rewards for approximate bayesian reinforcement learning. *arXiv preprint arXiv:1203.3518*, 2012.

Stone, P. and Sutton, R. S. Scaling reinforcement learning toward robocup soccer. In *Icml*, volume 1, pp. 537–544, 2001.

Strens, M. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pp. 943–950, 2000.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*, pp. 113. MIT press, 2018.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255, 2022.

Vlassis, N., Ghavamzadeh, M., Mannor, S., and Poupart, P. Bayesian reinforcement learning. *Reinforcement learning*, pp. 359–386, 2012.

Wei, Q. and Guo, X. Markov decision processes with state-dependent discount factors and unbounded rewards/costs. *Operations Research Letters*, 39(5):369–374, 2011.

White, M. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, pp. 3742–3750. PMLR, 2017.

Yoshida, N., Uchibe, E., and Doya, K. Reinforcement learning with state-dependent discount factor. In *2013 IEEE third joint international conference on development and learning and epigenetic robotics (ICDL)*, pp. 1–6. IEEE, 2013.

Zanette, A. and Brunskill, E. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning*, pp. 5747–5755. PMLR, 2018.

A. Common Form Derivations

A.1. Discount Regularization

Consider the matrix form of the Bellman equation, using γ_p, the lower value of the discount factor used during planning for regularization: $V = R + \gamma_p TV$. By the steps below, we write the product $\gamma_p T$ from the Bellman equation as the product of true discount factor γ and a weighted average matrix.

$$\gamma_p T = [\gamma - (\gamma - \gamma_p)]T \qquad \text{(Add and subtract } \gamma\text{)}$$

$$\gamma_p T = \gamma(1 - \frac{(\gamma - \gamma_p)}{\gamma})T \qquad \text{(Pull out a factor of } \gamma\text{.)}$$

Let T_{zeros} be an appropriately sized matrix of zeros. Adding γT_{zeros} to the right hand side does not change the equality.

$$\gamma_p T = \gamma[(1 - \frac{\gamma - \gamma_p}{\gamma})T + T_{\text{zeros}}]$$

Multiply the T_{zeros} term inside the parentheses by $\frac{\gamma - \gamma_p}{\gamma}$. T_{zeros} is all zeros so a multiplier does not affect the equality.

$$\gamma_p T = \gamma[(1 - \frac{\gamma - \gamma_p}{\gamma})T + (\frac{\gamma - \gamma_p}{\gamma})T_{\text{zeros}}]$$

Let $\epsilon = \frac{\gamma - \gamma_p}{\gamma}$.

$$\gamma_p T = \gamma[(1 - \epsilon)T_{true} + \epsilon T_{\text{zeros}}]$$

We have replaced the product of the planning discount factor and the true transition matrix with the product of the true discount factor and a weighted average of the transition matrix and a matrix of zeros. To put this in our framework, consider regularizing the MLE transition matrix for state-action pair (s, a) via discount regularization, using planning discount factor γ_p. Using the proof in this section, our regularized estimated transition matrix $\hat{T}(s, a)$, is:

$$\hat{T}(s, a) = (1 - \epsilon)\hat{T}_{MLE}(s, a) + \epsilon T_{zeros}$$

where $\epsilon = \frac{\gamma - \gamma_p}{\gamma}$.

A.2. Dirichlet Prior Derivation

Assume prior $P_{\text{prior}}(T(s_n, a_k)) = \text{Dirichlet}(\langle \alpha_{n,k,1}, ..., \alpha_{n,k,N} \rangle)$ on transition matrix $T(s_n, a_k)$ and let $\langle c_{n,k,1}, ..., c_{n,k,N} \rangle$ be the transition count data observed from state s_n to states 1 through N under action a_k. The posterior estimate of $T(s_n, a_k)$

follows a Dirichlet distribution with parameter $\langle c_{n,k,1} + \alpha_{n,k,1}, ..., c_{n,k,N} + \alpha_{n,k,N} \rangle$ and the posterior mean is:

$$
\begin{aligned}
T_{\substack{\text{post}\\\text{mean}}}(s_n, a_k) &= \langle \frac{c_{n,k,1} + \alpha_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{c_{n,k,N} + \alpha_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \rangle \\
&= \langle \frac{c_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{c_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \rangle + \langle \frac{\alpha_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{\alpha_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \rangle \\
&= \frac{\sum_{i=1}^{N} c_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \langle \frac{c_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{c_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \rangle \\
&+ \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} \alpha_{n,k,i}} \langle \frac{\alpha_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{\alpha_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \rangle \\
&= \frac{\sum_{i=1}^{N} c_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \langle \frac{c_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i}}, ..., \frac{c_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i}} \rangle + \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \langle \frac{\alpha_{n,k,1}}{\sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{\alpha_{n,k,N}}{\sum_{i=1}^{N} \alpha_{n,k,i}} \rangle
\end{aligned}
$$

Let $\hat{T}_{\text{MLE}}(s_n, a_k)$ be the MLE of $T(s_n, a_k)$. Then, $\hat{T}_{\text{MLE}}(s_n, a_k) = \langle \frac{c_{n,k,1}}{\sum_{i=1}^{N} c_{n,k,i}}, ..., \frac{c_{n,k,N}}{\sum_{i=1}^{N} c_{n,k,i}} \rangle$. Let $T_{\substack{\text{prior}\\\text{mean}}}(s_n, a_k)$ be the transition matrix implied by the prior for state s_n and action a_k, i.e. $T_{\substack{\text{prior}\\\text{mean}}}(s_n, a_k) = \langle \frac{\alpha_{n,k,1}}{\sum_{i=1}^{N} \alpha_{n,k,i}}, ..., \frac{\alpha_{n,k,N}}{\sum_{i=1}^{N} \alpha_{n,k,i}} \rangle$.

Using $\hat{T}_{\text{MLE}}(s_n, a_k)$ and $T_{\substack{\text{prior}\\\text{mean}}}(s_n, a_k)$, we can write $T_{\substack{\text{post}\\\text{mean}}}(s_n, a_k)$ as follows.

$$
T_{\substack{\text{post}\\\text{mean}}}(s_n, a_k) = \frac{\sum_{i=1}^{N} c_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}} \hat{T}_{\text{MLE}}(s_n, a_k) + \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{n=1}^{N} \alpha_{n,k,i}} T_{\substack{\text{prior}\\\text{mean}}}(s_n, a_k)
$$

Let $\epsilon = \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} c_{n,k,i} + \sum_{i=1}^{N} \alpha_{n,k,i}}$. Then we have:

$$
T_{\substack{\text{post}\\\text{mean}}}(s_n, a_k) = (1 - \epsilon) \hat{T}_{\text{MLE}}(s_n, a_k) + \epsilon T_{\substack{\text{prior}\\\text{mean}}}(s_n, a_k)
$$

A.2.1. Dirichlet Prior Implied by Discount Regularization

The equivalence proof above demonstrates that, for a given value of ϵ, averaging the transition matrix with either a matrix of the discrete uniform distribution or with the matrix of zeros yields the same policy. Since discount regularization applies the same ϵ to every state-action pair, the two methods are only exactly equivalent when the posterior transition matrix under a uniform prior has the same implied value of ϵ for all state-action pairs, i.e. $\epsilon = \frac{\sum_{i=1}^{N} \alpha_{n,k,i}}{\sum_{i=1}^{N} \alpha_{n,k,i} + \sum_{i=1}^{N} c_{n,k,i}}$ is the same for all state-action pairs, where $c_{n,k,i}$ and $\alpha_{n,k,i}$ are the transition count and prior from state n to state i under action k.

We can use the equivalence of the two methods for the same value of ϵ to solve for the magnitude of prior that is implied by a given discount factor in discount regularization. This is particularly interesting when $\sum_{i=1}^{N} c_{n,k,i}$ is unequal across state-action

pairs (s_n, a_k). In this case, we can use the equivalence of discount regularization and the uniform prior to compute the different priors implied by discount regularization across state-action pairs.

We will refer to $\sum_{i=1}^{N} c_{n,k,i}$ as $\sum c_i$ and $\sum_{i=1}^{N} \alpha_{n,k,i}$ as $\sum \alpha_i$ for brevity. Since we know both methods produce the same optimal policy for equal values of ϵ, we set the formulas for ϵ under each method equal to one another. We then solve for α_i to get the magnitude of prior that is implied by a given planning discount factor γ_p.

$$\begin{aligned}
\frac{\sum \alpha_i}{\sum \alpha_i + \sum c_i} &= \frac{\gamma - \gamma_p}{\gamma} \\
\gamma \sum \alpha_i &= (\sum \alpha_i + \sum c_i)(\gamma - \gamma_p) \\
\gamma \sum \alpha_i &= \gamma \sum \alpha_i - \gamma_p \sum \alpha_i + \gamma \sum c_i - \gamma_p \sum c_i \\
\gamma_p \sum \alpha_i &= \sum c_i(\gamma - \gamma_p) \\
\sum \alpha_i &= \sum c_i \frac{\gamma - \gamma_p}{\gamma_p}
\end{aligned}$$

For the uniform distribution, all α_i for a given state are the same, so substitute $\sum \alpha_i = N\alpha_i$.

$$\begin{aligned}
N\alpha_i &= \sum c_i \frac{\gamma - \gamma_p}{\gamma_p} \\
\alpha_i &= \frac{\sum c_i}{N} \frac{\gamma - \gamma_p}{\gamma_p}
\end{aligned}$$

So discount regularization functions like the Dirichlet prior:

$$T_{\text{prior}}(s_n, a_k) \sim \text{Dirichlet}(\frac{\sum c_i}{N} \frac{\gamma - \gamma_p}{\gamma_p}, ..., \frac{\sum c_i}{N} \frac{\gamma - \gamma_p}{\gamma_p}) \tag{6}$$

where again $\sum c_i$ is the total number of transitions observed in the data starting at state s_n under action a_k.

B. Uniform Prior MSE Calcuation

Let $\sum_{j=1}^{N} c_{n,k,j}$ be number of observations for state-action pair (s_n, a_k) in the data. We drop the index and write as $\sum c_j$ below for readability. Let $T(s_n, a_k)$ be the transition probability distribution under action a_k starting at state s_n. N is the number of states in the MDP.

$$\text{MSE}(\hat{T}(s_n, a_k)) = \sum_{i=1}^{N} \Big(\text{Variance}(\hat{T}(s_n, a_k, s_i)) + \text{Bias}^2(\hat{T}(s_n, a_k, s_i)) \Big)$$

$$
\begin{aligned}
\text{Variance}(\hat{T}_{\text{unif}}(s_n, a_k, s_i)) &= \text{Variance}\left((1-\epsilon)\hat{T}_{\text{MLE}}(s_n, a_k, s_i) + \epsilon\frac{1}{N}\right) \\
&= \text{Variance}\left((1-\epsilon)\hat{T}_{\text{MLE}}(s_n, a_k, s_i)\right) + \text{Variance}\left(\epsilon\frac{1}{N}\right) \\
&= (1-\epsilon)^2\text{Variance}(\hat{T}_{\text{MLE}}(s_n, a_k, s_i)) \\
&= (1-\epsilon)^2\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) \\
\text{Bias}(\hat{T}_{\text{unif}}(s_n, a_k, s_i)) &= \mathbb{E}\left[\hat{T}_{\text{unif}}(s_n, a_k, s_i)\right] - T(s_n, a_k, s_i) \\
&= \mathbb{E}\left[(1-\epsilon)\hat{T}_{\text{MLE}}(s_n, a_k, s_i) + \epsilon\frac{1}{N}\right] - T(s_n, a_k, s_i) \\
&= (1-\epsilon)T(s_n, a_k, s_i) + \epsilon\frac{1}{N} - T(s_n, a_k, s_i) \\
&= \epsilon\left(\frac{1}{N} - T(s_n, a_k, s_i)\right) \\
\text{MSE}(\hat{T}_{\text{unif}}(s_n, a_k)) &= \sum_{i=1}^{N}\left((1-\epsilon)^2\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + \epsilon^2\left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2\right)
\end{aligned}
$$

To solve for the optimal value of ϵ, set the first derivative equal to 0.

$$
\begin{aligned}
\frac{\partial\text{MSE}}{\epsilon} &= \sum_{i=1}^{N} -2(1-\epsilon)\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + 2\epsilon\left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2 \\
\sum_{i=1}^{N} -2\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) &+ 2\epsilon\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + 2\epsilon\left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2 = 0 \\
\sum_{i=1}^{N}\epsilon\left[\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + \left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2\right] &= \sum_{i=1}^{N}\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i))
\end{aligned}
$$

$$
\begin{aligned}
\epsilon &= \frac{\sum_{i=1}^{N}\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i))}{\sum_{i=1}^{N}\frac{1}{\sum c_j}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + \left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2} \\
&= \frac{\sum_{i=1}^{N}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i))}{\sum_{i=1}^{N}T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) + \sum c_j\left(\frac{1}{N} - T(s_n, a_k, s_i)\right)^2} \\
&= \frac{\sum_{i=1}^{N}[T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i))]/\sum_{i=1}^{N}[(\frac{1}{N} - T(s_n, a_k, s_i))^2]}{\sum_{i=1}^{N}[T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i))]/\sum_{i=1}^{N}[(\frac{1}{N} - T(s_n, a_k, s_i))^2] + \sum c_j}
\end{aligned}
$$

C. Empirical Example Details

C.1. Regularization Loss by ϵ

The following pseudocode summarizes the empirical example depicted in Figures 5 through 8 to compare the resulting loss for the the optimal policies found using each regularization method, for a range of values of ϵ.

Algorithm 1 Regularization Loss Pseudocode

Input: MDP, list of ϵ values, regularization method
for $i = 1$ **to** [number of data sets] **do**
  Generate data set of n transition tuples
  Estimate $\hat{T}_{\text{MLE}}$ from data
  for ϵ in list **do**
    Regularize transition matrices by amount ϵ
    Calculate optimal policy π of regularized MDP
    Calculate loss comparing value of π vs. value of true optimal policy in true MDP
  end for
end for
Average loss by ϵ value across all data sets

Additional Details Tuples in batch data are collected with uniform probability across state-action pairs. Loss is calculated as the average difference in value across states for the policy found by value iteration in the estimated, regularized MDP as compared to the optimal policy, both evaluated in the true environment.

C.2. State-Specific Regularization Loss

The pseudocode below describes the procedure for the empirical demonstration of state-specific regularization depicted in Figure 6.

Algorithm 2 State-Specific Regularization Loss Pseudocode

Input: MDP, list of ϵ values, regularization method
for $i = 1$ **to** [number of data sets] **do**
  Generate data set of n transition tuples
  Estimate $\hat{T}_{\text{MLE}}$ from data
  for each state-action pair (s, a) **do**
    Calculate $\epsilon^*(s, a)$ that minimizes MSE using equations in Sec. 6
    Regularize transition matrix by taking the weighted average of $\hat{T}_{MLE}(s, a)$ and the appropriate regularization matrix, using weight $\epsilon^*(s, a)$
  end for
  Calculate optimal policy π of regularized MDP
  Calculate loss comparing value of π vs. value of true optimal policy in true MDP
end for
Average loss by ϵ value across all data sets

Additional Details In the step where we calculate ϵ^* using the Equation 4, two possibilities are to use $\hat{T}_{\text{MLE}}$ as a plug-in estimator for T or model the distribution of T from using the batch data and sample from the posterior. To generate the plot in Figure 6, we did the latter. We calculated the posterior $Dirichlet(\alpha_1, ...\alpha_N)$ from the batch data, then sample repeatedly from that distribution, calculating the MSE for each sample across values of ϵ. We then choose for ϵ^* the value of ϵ that had the lowest average loss across samples.

D. Extension to Epsilon-Greedy Regularization

D.1. Weighted-Average Transition Matrix Form

With the reward expressed as a function of state only (rather than state and action) is straightforward to write epsilon-greedy regularization in the weighted-average form.

Under epsilon-greedy action selection, the probability of next state s' given state s_n and greedy action a_k is:

$P_{\substack{\text{eps}\\\text{greedy}}}(s'|s_n, a_k) = P(a_k)P(s'|s_n, a_k) + P(\text{random action})P(s'|s_n, \text{random action})$

$P_{\substack{\text{eps}\\\text{greedy}}}(s'|s_n, a_k) = (1-\epsilon)P(s'|s_n, a_k) + \epsilon P(s'|s_n, \text{random action})$

The probability for each action in the last term is equal and their probability distributions are independent, so we can rewrite the term as a sum.

$P_{\substack{\text{eps}\\\text{greedy}}}(s'|s_n, a_k) = (1-\epsilon)P(s'|s_n, a_k) + \frac{\epsilon}{|A|}\sum_{a' \in A} P(s'|s_n, a')$

By definition, $P(s'|s_n, a_k)$ is defined by transition matrix T.

$T_{\substack{\text{eps}\\\text{greedy}}}(s_n, a_k, s') = (1-\epsilon)T(s_n, a_k, s') + \frac{\epsilon}{|A|}\sum_{a' \in A} T(s_n, a', s')$

In this setting, we are estimating $\hat{T}$ by the MLE, so we replace T with $\hat{T}_{\text{MLE}}$. This relationship holds for all next states, so we can extend the form above to the vector of next states, $\hat{T}_{\substack{\text{eps}\\\text{greedy}}}(s, a)$.

$\hat{T}_{\substack{\text{eps}\\\text{greedy}}}(s_n, a_k) = (1-\epsilon)\hat{T}_{MLE}(s_n, a_k) + \frac{\epsilon}{|A|}\sum_{a' \in A} \hat{T}_{MLE}(s_n, a')$

D.2. Epsilon-Greedy Transition Matrix MSE

Again, we decompose MSE into bias and variance and calculate each separately. As in the previous section, let $\sum_{j=1}^{N} c_{n,k,j}$ be number of observations for state-action pair (s_n, a_k) in the data. Let $T(s_n, a_k)$ be the transition probability distribution under action a_k starting at state s_n. N is the number of states in the MDP.

$$
\begin{aligned}
\text{Variance}(\hat{T}_{\substack{\text{eps}\\\text{greedy}}}(s_n, a_k, s_i)) &= \text{Variance}\left((1-\epsilon)\hat{T}_{\text{MLE}}(s_n, a_k, s_i) + \epsilon\frac{1}{|A|}\sum_{m=1}^{|A|}\hat{T}(s_n, a_m, s_i)\right) \\
&= \text{Variance}\left((1-\epsilon+\frac{\epsilon}{|A|})\hat{T}_{\text{MLE}}(s_n, a_k, s_i) + \epsilon\frac{1}{|A|}\sum_{m\neq k}\hat{T}(s_n, a_m, s_i)\right) \\
&= \left(1-\epsilon+\frac{\epsilon}{|A|}\right)^2 \text{Variance}(\hat{T}_{\text{MLE}}(s_n, a_k, s_i) + \left(\frac{\epsilon}{|A|}\right)^2 \sum_{m\neq k}\text{Variance}(\hat{T}(s_n, a_m, s_i)) \\
&= \left(1-\epsilon+\frac{\epsilon}{|A|}\right)^2 \frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) \\
&+ \left(\frac{\epsilon}{|A|}\right)^2 \sum_{m\neq k}\frac{1}{\sum_{j=1}^{N} c_{n,m,j}} T(s_n, a_m, s_i)(1-T(s_n, a_m, s_i)) \\
\text{Bias}(\hat{T}_{\substack{\text{eps}\\\text{greedy}}}(s_n, a_k, s_i)) &= \epsilon\frac{1}{|A|}\sum_{m\neq k}(T(s_n, a_m, s_i) - T(s_n, a_k, s_i)) \\
\text{MSE}(\hat{T}_{\substack{\text{eps}\\\text{greedy}}}(s_n, a_k)) &= \sum_{i=1}^{N}\left(1-\epsilon+\frac{\epsilon}{|A|}\right)^2 \frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n, a_k, s_i)(1-T(s_n, a_k, s_i)) \\
&+ \sum_{i=1}^{N}\left(\frac{\epsilon}{|A|}\right)^2 \sum_{m\neq k}\frac{1}{\sum_{j=1}^{N} c_{n,m,j}} T(s_n, a_m, s_i)(1-T(s_n, a_m, s_i)) \\
&+ \sum_{i=1}^{N}\left(\frac{\epsilon}{|A|}\right)^2 \left(\sum_{m\neq k} T(s_n, a_m, s_i) - T(s_n, a_k, s_i)\right)^2
\end{aligned}
$$

To solve for the value of ϵ that minimizes the MSE, set the first derivative equal to 0:

$$\begin{aligned}
\frac{\partial \text{MSE}}{\epsilon} = \sum_{i=1}^{N} 2\left(1-\epsilon+\frac{\epsilon}{|A|}\right)\left(\frac{1}{|A|}-1\right)\frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i)) \\
+ \frac{2\epsilon}{|A|^2}\sum_{m\neq k}\frac{1}{\sum_{j=1}^{N} c_{n,m,j}} T(s_n,a_m,s_i)(1-T(s_n,a_m,s_i)) \\
+ \frac{2\epsilon}{|A|^2}\sum_{m\neq k}\left(T(s_n,a_m,s_i)-T(s_n,a_k,s_i)\right)^2 \\
= \epsilon[\left(\frac{1}{|A|}-1\right)^2\frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i)) \\
+ \frac{1}{|A|^2}\sum_{m\neq k}\frac{1}{\sum_{j=1}^{N} c_{n,m,j}} T(s,a_m,s_i)(1-T(s_n,a_m,s_i)) \\
+ \frac{1}{|A|^2}\sum_{m\neq k}\left(T(s_n,a_m,s_i)-T(s_n,a_k,s_i)\right)^2] \\
+ \left(\frac{1}{|A|}-1\right)\frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i))
\end{aligned}$$

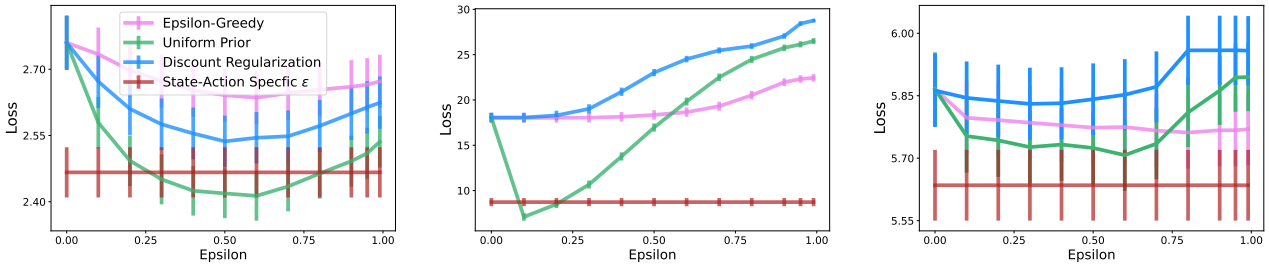
$$\epsilon = \frac{\sum_{i=1}^{N}(1-\frac{1}{|A|})\frac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i))}{\begin{aligned}\sum_{i=1}^{N}(\tfrac{1}{|A|}-1)^2\tfrac{1}{\sum_{j=1}^{N} c_{n,k,j}} T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i)) + \tfrac{1}{|A|^2}\sum_{m\neq k}\tfrac{1}{\sum_{j=1}^{N} c_{n,m,k}} T(s_n,a_m,s_i)(1-T(s_n,a_m,s_i)) \\ + \tfrac{1}{|A|^2}\sum_{m\neq k}\left(T(s_n,a_k,s_i)-T(s_n,a_m,s_i)\right)^2\end{aligned}}$$

To demonstrate dependence on the number of data observations, we put this in the form $\frac{K(s,a)}{K'(s,a)+\sum c_j}$, similar to the above. To do so, assuming equal observations C across all state-action pairs the above is equal to:

$$\begin{aligned}
\epsilon &= \frac{\sum_{i=1}^{N}(1-\frac{1}{|A|})T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i))}{\begin{aligned}\sum_{i=1}^{N}(\tfrac{1}{|A|}-1)^2 T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i)) + \tfrac{1}{|A|^2}\sum_{m\neq k} T(s_n,a_m,s_i)(1-T(s_n,a_m,s_i)) \\ + C\tfrac{1}{|A|^2}\sum_{m\neq k}\left(T(s_n,a_k,s_i)-T(s_n,a_m,s_i)\right)^2\end{aligned}} \\
&= \frac{\sum_{i=1}^{N}[(1-\frac{1}{|A|})T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i))]/[\frac{1}{|A|^2}\sum_{m\neq k}(T(s_n,a_k,s_i)-T(s_n,a_m,s_i))^2]}{\sum_{i=1}^{N}|A|^2\frac{(\frac{1}{|A|}-1)^2 T(s_n,a_k,s_i)(1-T(s_n,a_k,s_i))}{\sum_{m\neq k}(T(s_n,a_k,s_i)-T(s_n,a_m,s_i))^2} + \frac{\sum_{m\neq k} T(s_n,a_m,s_i)(1-T(s_n,a_m,s_i))}{\sum_{m\neq k}(T(s_n,a_k,s_i)-T(s_n,a_m,s_i))^2} + C}
\end{aligned}$$

D.3. Empirical Results for Epsilon-Greedy Regularization

D.3.1. Tabular Examples from Section 7



(a) 10-State Random Chain (b) River Swim (c) Loop

Figure 9: Comparison of epsilon-greedy regularization to uniform prior and discount regularization in our three tabular environments.

D.3.2. Controlled Environment

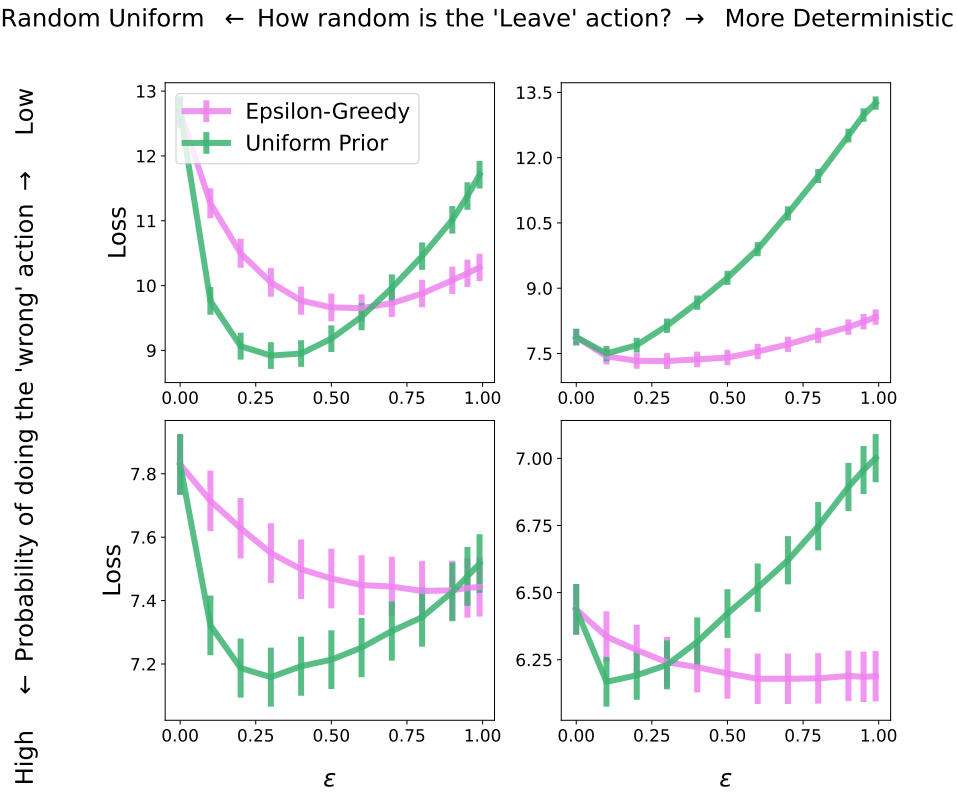


Figure 10: Additional empirical results comparing epsilon-greedy regularization to uniform prior on the environment described in this section. The method whose regularization matrix best aligns with the true environment generates lowest loss.

The second set of results come from a loop MDP with 10 states, $\gamma = .99$, and two actions which cause the agent to either stay in or leave the current state, described further below. We define the reward function to be a high reward region in the three adjacent states and zero elsewhere. We vary the transition dynamics along the following two axes to demonstrate the

relative performance of different regularizers across environments.

Transition Stochasticity, κ. In our experiments, we utilize a mixture of two transition matrices. The first is the "leave" transiton matrix. We vary the transition dynamics of "leaving" the current state, from deterministic (transition from state s to the adjacent state $s+1$ with probability 1) to uniform (transition with uniform probability across states). The second transition matrix is the "stay" transition matrix. The "stay" transition matrix causes the agent to remain in the current state with probability 0.75, and otherwise transition to a random state.

$$\text{leave} = \kappa * \text{deterministic } + (1-\kappa) * \text{uniform}$$
$$\text{stay} = 0.75 * \text{identity} + 0.25 * \text{uniform}$$

Action Similarity, λ. The agent's two actions are generated by mixing the "stay" and "leave" matrices together in the following way to control the action similarity.

$$\text{Action 1 ("probably stay")} = (1-\lambda) * \text{stay} + \lambda * \text{leave}$$
$$\text{Action 2 ("probably leave")} = (1-\lambda) * \text{leave} + \lambda * \text{stay}$$

λ varies from 0 (distinct actions) to 0.5 (identical actions).

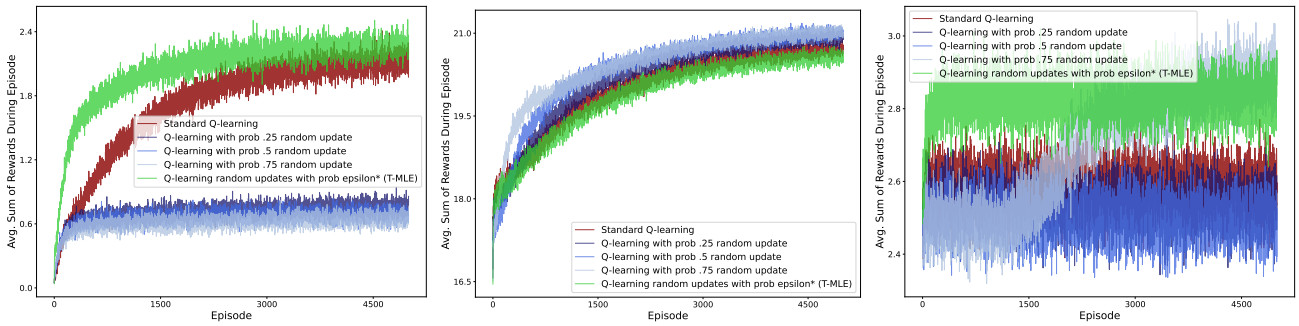
D.4. Extensions to Model-Free RL

D.4.1. Sample Algorithm

Algorithm 3 Q-learning with State-Action-Specific Regularization

Parameters: step size $\alpha \in (0, 1]$, regularization matrix $T_{reg}(s, a)$
Initialize $Q(s, a) = 0 \forall (s, a)$
for $e = 1$ **to** [number of episodes] **do**
  Choose initial state s randomly.
  while step_counter $<$ [steps per episode] **do**
    Choose action a from policy based on $Q(s, a)$ (e.g. epsilon-greedy based on current Q)
    Calculate ϵ^* from Eq. 5
    Draw $x \sim Bernoulli(\epsilon^*)$
    if $x = 1$ **then**
      Draw simulated next step s'_{sim} from $T_{reg}(s, a)$
      Update Q-function using s'_{sim}:
      $Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a) + \gamma max_a Q(s'_{sim}, A) - Q(s, a)]$
    else if $x = 0$ **then**
      Agent takes action a, observes next state s'
      $Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a) + \gamma max_a Q(s', A) - Q(s, a)]$
      step_counter += 1
      $s \leftarrow s'$
    end if
  end while
end for

D.4.2. RESULTS



(a) River Swim (b) 10-State Random Chain (c) Loop

Figure 11: Comparison of epsilon-greedy regularization to standard Q-learning in our three tabular environments. We also include as a baseline our algorithm with ϵ^* replaced by a constant probability.

Our simple modified Q-learning algorithm outperforms standard Q-learning on River Swim and Loop environments, but not on the random chain environment. Understanding where it performs best, why, and improving the extensions to model-free algorithms is a topic of ongoing research.