Comparison and Unification of Three Regularization Methods in Batch Reinforcement Learning

Sarah Rathnam [1] **Susan A. Murphy** [2 3] **Finale Doshi-Velez** [2]

Abstract

In batch reinforcement learning, there can be poorly explored state-action pairs resulting in poorly learned, inaccurate models and poorly performing associated policies. Various regularization methods can mitigate the problem of learning overly-complex models in Markov decision processes (MDPs), however they operate in technically and intuitively distinct ways and lack a common form in which to compare them. This paper unifies three regularization methods in a common framework– a weighted average transition matrix. Considering regularization methods in this common form illuminates how the MDP structure and the state-action pair distribution of the batch data set influence the relative performance of regularization methods. We confirm intuitions generated from the common framework by empirical evaluation across a range of MDPs and data collection policies.

1. Introduction

In certainty-equivalence reinforcement learning, the estimated model is treated as accurate when finding the optimal policy, without taking into account model uncertainty (Goodwin & Sin, 2014). Consequently, when acting according to certainty-equivalence control, we risk finding a policy tailored to a model that is overly-expressive for the amount of data. This is especially problematic in a batch setting, as further exploration is not possible to improve the model.

Many regularization methods address the problem of overfitting, for example reducing the planning horizon, using the posterior mean transition matrix under a Bayesian prior, or adding stochasticity to policies during planning. However, a challenge arises in understanding how they relate and choosing between them because regularization methods act on different elements of the MDP. In the methods listed above, a reduced planning horizon modifies the discount factor, planning using the posterior mean transition matrix modifies the transition matrix, and planning over the set of stochastic policies modifies the set of policies over which we optimize. Furthermore, their interpretations differ, for instance in the previously mentioned cases: decreasing the planning horizon, infusing outside information into the model, and planning over a stochastic set of policies.

Given certain constraints, the posterior mean transition matrix under a Bayesian prior is equivalent to a weighted average of the maximum likelihood estimator (MLE) transition matrix and the transition matrix implied by the prior. Similarly, we express the other two regularization methods above as a weighted average between the MLE transition matrix and a regularization matrix of another form. In this common Bayesian-like form, instead of comparing across disparate elements of the MDP, we can simply compare the form of the regularization matrix in each case and select the one that is most appropriate for the situation. This framing suggests that a uniform Bayesian prior performs better in an MDP with densely-interconnected states, a lower discount factor performs better when balancing goals of different timescales, and planning over stochastic policies is preferable to avoid a catastrophic outcome. Simulations confirm that these hypotheses hold in many cases, but also underscore the need to take the data collection policy as well as the MDP into account when selecting a regularization method.

2. Regularization in Certainty-Equivalence RL: Background and Related Work

Bayesian Prior as Regularization A prior encodes expert knowledge, information from previous studies, or other outside information. We can also view a prior on the transition function as a form of regularization since it forces the model not to overfit when data is limited (Poggio & Girosi, 1990). 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.

[1]Department of Applied Mathematics, Harvard University [2]Department of Computer Science, Harvard University [3]Department of Statistics, Harvard University. Correspondence to: Sarah Rathnam <sarah_rathnam@g.harvard.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Discount Regularization Jiang et al. (2015) demonstrate that using a lower discount factor often leads to learning a policy that performs better than the one learned using the true discount factor. They prove that a lower discount factor restricts 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 will use here.

Planning over ϵ-Greedy Policies Arumugam et al. (2018) propose a regularization method where planning is conducted over the set of ϵ-greedy policies rather than deterministic policies. The added stochasticity prevents tailoring the policy too closely to the model. Like discount regularization, planning over ϵ-greedy policies restricts the class of policies that can be optimal (Arumugam et al., 2018).

Related Work: Other Regularization Methods Beyond the methods included in our unified framework, state aggregation maps the true MDP to a simpler, abstract representation. States are grouped by characteristics such as action-value function or optimal action (Li et al., 2006). Another method, L_2 regularization, introduces a complexity penalty, balancing a simpler model against one that fits the data more closely. For example, Amit et al. (2020) provide a framework to unify discount regularization with L_2 regularization in TD learning. Their use of L_2 regularization penalizes large value estimates, encouraging consistent value estimates across state-action pairs. In contrast, we frame methods as regularizing the transition matrix, thereby restricting model complexity.

3. Notation and Definitions

Methods in this paper are applied in a finite MDP setting. An MDP M is characterized by $< S, A, R, T, \gamma >$, defined as follows. S: State space of size N. A: Action space. $R(s)$: Reward function. R generally maps each state-action pair to a real-valued reward. In this paper, we consider rewards as a function of states only. $T(s'|s, a)$: Transition function, mapping each state-action pair to a probability distribution over successor states. γ: Discount factor, $0 \leq \gamma < 1$. We assume T and R are unknown and estimated from the data.

4. Unification: Regularization as a Weighted Average Transition Matrix

Each method above modifies a different element of the MDP. To compare, we frame each as a weighted average of the MLE transition matrix and a matrix of another form. In this framework, we can compare by analyzing the matrix that is averaged with the MLE in each case.

Dirichlet Prior We consider a Dirichlet distribution over the vector of successor state probabilities for a state-action pair, $T(s, a) = \langle p_1, ..., p_N \rangle$. We assume prior $P_{prior}(T(s, a)) = \text{Dirichlet}(\langle \alpha_1, ..., \alpha_N \rangle)$. The posterior mean can be expressed as a weighted average of $\hat{T}_{MLE}(s, a)$, the MLE of $T(s, a)$, and $T_{\text{prior mean}}(s, a)$, the transition matrix implied by the prior:

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

where $\epsilon = \frac{\sum \alpha_i}{\sum c_i + \sum \alpha_i}$ and c_i is the transition count from state s to state i in the data set.

The expression above is written for a single state s. To express the matrix $T_{\text{post mean}}(a)$ as a weighted average of the MLE and the prior transition matrix, ϵ must be equal for all states. If we assume (1) $\sum c_i$ equal across all states for given action a (uniform visits), and (2) $\sum \alpha_i$ equal across all states for given action a (identical priors), then we can write the matrix of posterior means as

$$\hat{T}(a) = (1 - \epsilon)\hat{T}_{MLE}(a) + \epsilon T_{\text{prior mean}}(a) \qquad (1)$$

(Full derivation in Appendix A.1.) Condition (2) holds for the choice of a uniform prior in empirical examples, however condition (1), uniform visits, is restrictive and unrealistic. We consequently do not enforce uniform visits in examples, however the weighted average form still provides insight in comparing this regularization form to others.

Discount Regularization To express discount regularization in the form of Equation 1, consider the matrix form of the Bellman equation $V = R + \gamma TV$. Let $\gamma_l < \gamma$ be the lower value of the discount factor used for regularization. We write $\gamma_l T$ from the Bellman equation under discount regularization as the product of γ, the true discount factor for the MDP, and a weighted average matrix:

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

where T_{zeros} is a matrix of zeros and $\epsilon = \frac{\gamma - \gamma_l}{\gamma}$.

Hence using a lower discount factor is equivalent to using γ, the true value of the discount factor for the MDP, and replacing the transition matrix with its weighted average with a matrix of zeros. Applying this to our unified framework, we replace the MLE transition matrix for action a with the regularized form:

$$\hat{T}(a) = (1 - \epsilon)\hat{T}_{MLE}(a) + \epsilon T_{zeros} \qquad (2)$$

(Full derivation in Appendix A.2.)

Planning over ϵ-Greedy Policies Finally, we frame planning over the set of ϵ-greedy policies as a weighted average transition matrix. When finding the optimal policy from the

estimated MDP by policy iteration, all policies are treated as ϵ-greedy. Then we perform the greedy, deterministic policy that had the best ϵ-greedy performance.

When following an ϵ-greedy policy, for greedy action a, the agent transitions according to transition matrix $T(a)$ with probability $(1-\epsilon)$ and chooses uniformly at random between the transition matrices for all actions with probability ϵ. Estimating each transition matrix by its MLE, the transitions under an ϵ-greedy policy corresponds to:

$$\hat{T}(a) = (1-\epsilon)\hat{T}_{MLE}(a) + \epsilon\frac{1}{|\mathcal{A}|}\sum_{a'}\hat{T}_{MLE}(a') \quad (3)$$

Recall that we restrict our consideration to the case of state-dependent rewards $R(s)$. Under this assumption, planning over the set of ϵ-greedy policies is equivalent to replacing the MLE transition matrix for each action with Equation 3 before computing the optimal greedy policy.

5. Discussion of Unified Framework

With the methods expressed in a common form, we can now make predictions about their relative performance.

Uniform Prior and Discount Regularization Connection A surprising result revealed by the unified form is that, when constrained to the weighted average form (uniform exploration and equal priors), a uniform prior produces the same optimal policy as discount regularization for the same value of ϵ.

Theorem 1. *Let M_1 and M_2 be finite-state MDPs with identical state space, action space, and reward function. Let M_1 have transition function T and discount factor $(1-\epsilon)\gamma$. Let M_2 have discount factor γ and transition function $(1-\epsilon)T + \epsilon T_{unif}$, where T_{unif} is the uniform transition matrix. Then M_1 and M_2 have the same optimal policy.*

Proof. See Appendix A.3.1 for proof. □

Impact of MDP Structure *A uniform Dirichlet prior is a good regularizer in a dense world.* With a uniform prior, the posterior transition matrix is not constrained by the connections between states in the true MDP. If the MDP has a high level of connectivity between states, the connectivity of a uniform prior is appropriate, however in the case of a sparsely connected MDP, assuming all states are linked is unlikely to be optimal.

Discount regularization balances between planning lengths. The discount factor determines planning horizon, prioritizing shorter- versus longer-term rewards. The weighted average view of discount regularization is consistent with the view of discounting as causing the agent to act as if it transitions according to the true transition matrix with probability $1-\epsilon$ and exit the MDP (represented by the matrix of zeros) with probability ϵ (Sutton & Barto, 2018, p. 113). Faced with the prospect of exit, the agent prioritizes closer rewards. We predict that this is beneficial when balancing the trade-offs of differently sized rewards at different distances.

ϵ-greedy planning avoids catastrophic outcomes. In the ϵ-greedy case, averaging the transition matrices of all actions causes the agent to act as if there is more stochasticity in the transitions. We hypothesize that the added randomness during planning will cause the agent to find a more conservative policy and perform better in MDPs with catastrophic outcomes.

Impact of Data Collection Policy In the unified form for discount regularization, the regularization matrix is the same for all state-action pairs. In contrast, the regularization matrix for planning over ϵ-greedy policies is the same for all actions, but differ by state. Finally, a Dirichlet prior, when not constrained to uniform visits, regularizes each state-action pair separately. Therefore, for data sets with uneven counts across states and/or actions, we expect a Dirichlet prior to perform best, followed by ϵ-greedy planning then discount regularization because of the ability to separately tailor the regularization to the state-action pair.

Furthermore, examining the equivalence between discount regularization and the weighted average form of the uniform Dirichlet prior reveals that discount regularization functions like a Dirichlet prior with all parameters of magnitude $\frac{\gamma-\gamma_l}{\gamma_l}\frac{\sum c_i}{N}$ (see Appendix A.3.2 for derivation). This underscores the limitations of discount regularization under uneven data collection. The magnitude of the prior is higher for state-action pairs with more data, which is not desirable.

6. Empirical Examples

Equipped with a common framework, we implement the three regularization methods across simple tabular examples. We explore the impact of following characteristics on the loss of the resulting policy: MDP structure, probability that an action in the data set is generated by the optimal policy, starting state of trajectories in the data set, and data set size.

6.1. MDP Types

Dense world: Interconnected Grid This example MDP illustrates dense and complex connections between states. For each action, the agent transitions according to the arrows in Figure 1(a) with equal probability. Rewards are normally distributed with means as indicated.

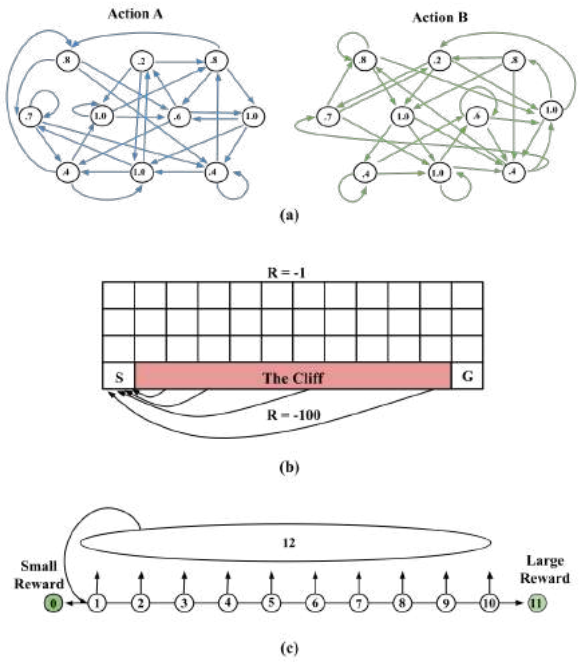

Figure 1. (a) Interconnected Grid (b) Cliff Walk (c) Two Goals

Catastrophic Outcome: Cliff Walk The Cliff Walk example from Sutton & Barto (2018) represents an agent that moves left, right, up, and down, with added noise. Mean rewards are -100 in the cliff and -1 for all other transitions. After reaching state **G**, the agent receives no further rewards.

Different Planning Lengths: Two Goals The final example MDP depicts differently sized rewards on opposite ends of a linear grid. The agent moves left, right, or up, each with noise. Transitioning to state 0 results in a reward of mean 0.10 and transitioning to state 11 results in a reward of mean 1. Rewards for all other transitions have mean 0 and after reaching either the small or large reward, the agent receives no further rewards.

6.2. Implementation

For each MDP, for a range of data collection policies and sizes, we generate trajectories from the true MDP. We estimate the transition and reward matrices as the MLE. For a range of ϵ between 0 and 1 (or prior magnitude from 0 to 1000), we regularize the transition matrix for each action. We find the optimal policy via policy iteration. The policy that is optimal in the estimated, regularized MDP and the optimal policy for the true MDP are compared in terms of performance on the true MDP. Details are in Appendix B.

6.3. Results

To compare regularization methods across MDPs and data sets, we plot the loss of the resulting policy. We also plot mean squared error (MSE) of the estimated, regularized transition matrix compared to the true transition matrix to investigate the extent to which better approximating the true transition matrix drives lower loss.

Hypothesized interactions between MDP structure and regularization are confirmed, although mediated by data collection policy. In Figure 2, comparing results by MDP confirms our predictions under the condition of uniformly random data collection, although less pronounced in the case of Interconnected Grid. Deviating from a uniformly random data collection policy by generating an increasing percentage of the actions from the true optimal policy impacts which method minimizes loss as well as whether regularization is beneficial at all. We note a reversal in which regularization method minimizes loss for the Cliff Walk when data is not fully random, and with Two Goals, no regularization method is beneficial as actions in the data set are generated increasingly from the optimal policy.

Data set size and starting state demonstrate less impact on relative loss. Number and length of trajectories do not considerably impact relative performance. The impact of trajectory starting state varies by MDP. While the ordering of methods by loss is not dramatically shifted across starting states, the shape of the loss curves indicate a differing impact of regularization. Results are in Appendices C.2 and C.3.

Impact of uneven data collection is inconclusive. We hypothesized that regularizers that are more flexible in allowing different amounts of regularization across state-action pairs outperform under uneven exploration. Both generating the data set from the optimal policy and restricting starting state cause uneven exploration, yet we do not clearly observe the hypothesized relationship. Further investigation is needed to isolate the impact of uneven exploration.

Lower loss does not consistently correspond to lower transition matrix MSE. Although loss is partially driven by the ability to accurately replicate the true transition matrix, there are other factors impacting loss to be identified.

To summarize, choosing between regularization methods can be viewed in terms of choosing the regularization matrix from the weighted average form that best aligns with the context at hand in terms of both the data collection policy and MDP structure. The empirical examples in this section provide evidence that both of these factors impact relative loss. Further work remains to formalize the conditions in which each regularizer is preferred.

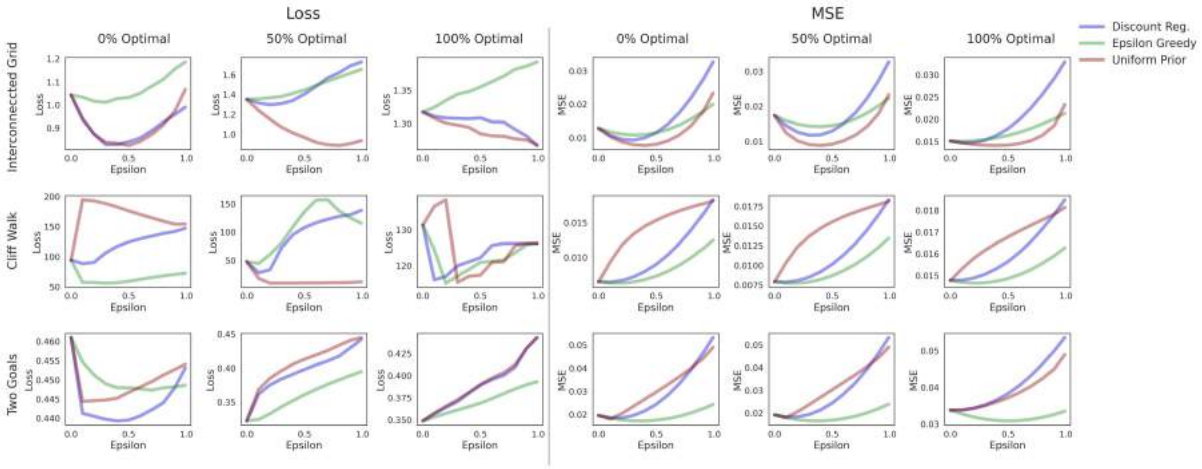

Figure 2. Loss and MSE by Optimality of Data Collection Policy. Predictions by MDP hold for random data collection, but results vary when data partially- or fully-generated from optimal policy. Random start states; 15 trajectories of length 10 each for Interconnected Grid and Two Goals; 25 trajectories of length 20 for Cliff Walk.

7. Conclusion

We have unified three MDP regularization methods into a common framework that helps us to predict and understand their performance in different settings. The common form and empirical examples demonstrate that it is vital to consider both the the MDP structure and the data collection policy when deciding between regularization methods. The unified form also motivates viewing discount regularization as replacing the maximum likelihood estimate transition matrix with its posterior mean under a uniform prior, when the data set is constrained to uniform state visitation. In future work, we will leverage the unified framework to systematically characterize MDPs and data sets to select a regularization method.

Acknowledgements

Research reported in this work was supported by the National Institute Of Biomedical Imaging And Bioengineering and the Office of the Director of the National Institutes of Health under award number P41EB028242, the National Institute on Alcohol Abuse and Alcoholism under award number R01AA023187.

SR and FDV acknowledge support from NSF project 2007076. Research reported in this work was also supported by the National Institute Of Biomedical Imaging And Bioengineering and the Office of the Director of the National Institutes of Health under award number P41EB028242, the National Institute on Alcohol Abuse and Alcoholism under award number R01AA023187.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

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. Mitigating planner overfitting in model-based reinforcement learning. *ArXiv*, abs/1812.01129, 2018.

Goodwin, G. C. and Sin, K. S. *Adaptive filtering prediction and control*. Courier Corporation, 2014.

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

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for mdps. *ISAIM*, 4:5, 2006.

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

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

APPENDIX

A. Full Derivation of Unified Form

A.1. Dirichlet Prior

Assume prior $P_{prior}(T(s,a)) = \text{Dirichlet}(\langle \alpha_1, ..., \alpha_N \rangle)$ on transition matrix $T(s,a)$ and let $\langle c_1, ..., c_N \rangle$ be the transition count data observed from state s to states 1 to N under action a. The posterior of $T(s,a)$ follows a Dirichlet distribution with parameter $\langle c_1 + \alpha_1, ..., c_N + \alpha_N \rangle$ and the posterior mean is:

$$T_{\text{post mean}}(s,a) = \langle \frac{c_1 + \alpha_1}{\sum c_i + \sum \alpha_i}, ..., \frac{c_N + \alpha_N}{\sum c_i + \sum \alpha_i} \rangle$$

$$T_{\text{post mean}}(s,a) = \langle \frac{c_1}{\sum c_i + \sum \alpha_i}, ..., \frac{c_N}{\sum c_i + \sum \alpha_i} \rangle + \langle \frac{\alpha_1}{\sum c_i + \sum \alpha_i}, ..., \frac{\alpha_N}{\sum c_i + \sum \alpha_i} \rangle$$

Multiply each term by 1.

$$\begin{aligned} T_{\text{post mean}}(s,a) &= \frac{\sum c_i}{\sum c_i} \langle \frac{c_1}{\sum c_i + \sum \alpha_i}, ..., \frac{c_N}{\sum c_i + \sum \alpha_i} \rangle + \frac{\sum \alpha_i}{\sum \alpha_i} \langle \frac{\alpha_1}{\sum c_i + \sum \alpha_i}, ..., \frac{\alpha_N}{\sum c_i + \sum \alpha_i} \rangle \\ &= \frac{\sum c_i}{\sum c_i + \sum \alpha_i} \langle \frac{c_1}{\sum c_i}, ..., \frac{c_N}{\sum c_i} \rangle + \frac{\sum \alpha_i}{\sum c_i + \sum \alpha_i} \langle \frac{\alpha_1}{\sum \alpha_i}, ..., \frac{\alpha_N}{\sum \alpha_i} \rangle \end{aligned}$$

Let $\hat{T}_{MLE}(s,a)$ be the MLE of $T(s,a)$: $\hat{T}_{MLE}(s,a) = \langle \frac{c_1}{\sum c_i}, ..., \frac{c_N}{\sum c_i} \rangle$.

Let $T_{\text{prior mean}}(s,a)$ be the transition matrix implied by the prior for state s and action a. $T_{\text{prior mean}}(s,a) = \langle \frac{\alpha_1}{\sum \alpha_i}, ..., \frac{\alpha_N}{\sum \alpha_i} \rangle$.

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

$$T_{\text{post mean}}(s,a) = \frac{\sum c_i}{\sum c_i + \sum \alpha_i} \hat{T}_{MLE}(s,a) + \frac{\sum \alpha_i}{\sum c_i + \sum \alpha_i} T_{\text{prior mean}}(s,a)$$

Let $\epsilon = \frac{\sum \alpha_i}{\sum c_i + \sum \alpha_i}$. Consequently, we have:

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

The expression above is for a single state s. To write $T_{\text{post mean}}(a)$ as a matrix for all states for a given action, we must pull out the same factor of ϵ and $(1-\epsilon)$ for all states. Hence we assume:

1. $\sum c_i$ equal across all states for a given action a, and
2. $\sum \alpha_i$ equal across all states for a given action a,

Assuming the conditions above and taking the matrix of posterior means as our estimate of $T(a)$:

$$\hat{T}(a) = (1-\epsilon)\hat{T}_{MLE}(a) + \epsilon T_{\text{prior mean}}(a)$$

A.2. Discount Regularization

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

First add and subtract γ.

$$\gamma_l T = [\gamma - (\gamma - \gamma_l)]T$$

Pull out a factor of γ.

$$\gamma_l T = \gamma(1 - \frac{(\gamma - \gamma_l)}{\gamma})T$$

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

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

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

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

Let $\epsilon = \frac{\gamma-\gamma_l}{\gamma}$,

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

We have replaced the product of the regularization 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 the unified framework, consider regularizing the MLE transition matrix for action a via discount regularization. Using the proof in this section, our regularized estimated transition matrix for action a, $\hat{T}(a)$, is:

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

A.3. Discount Regularization - Uniform Prior Connection

A.3.1. Proof of Theorem 1: Equivalence of Weighted Average Form

First we show that the optimal policy is not affected by adding the same constant x to all rewards $r(s, a)$. Let $Q^\pi_x(s, a)$ be the action-value function for policy π when adding constant x to all rewards. Then,

$Q^\pi_x(s, a) = \mathbb{E}_\pi[\sum_{k\geq 0}^{\infty} \gamma^k (r(s_k, a_k) + x)|s_0 = s, a_o = a] = \mathbb{E}_\pi[\sum_{k\geq 0}^{\infty} \gamma^k r(s_k, a_k)|s_0 = s, a_o = a] + \frac{x}{1-\gamma}$

and the action-value function of the optimal policy is, $Q^*_x(s, a) = \max_\pi[\mathbb{E}_\pi[\sum_{k\geq 0}^{\infty} \gamma^k r(s_k, a_k)|s_0 = s, a_o = a] + \frac{x}{1-\gamma}]$

The optimal action at state s is $\pi_{opt}(s) = \text{argmax}_a Q^*_x(s, a)$. The first term of the expression for $Q^*_x(s, a)$ does not contain x and the second does not depend on a, therefore π_{opt} is not affected by the choice of any constant added to $r(s, a)$.

Next, observe that $Q^*_x(s, a)$ is the solution to Bellman's optimality equation, $Q_x(s, a) = r(s, a) + x + \gamma \sum_{s'} T(s, a, s')\text{max}_{a'} Q_x(s', a')$. From above, we established that $\pi_{opt} = \text{argmax}_a Q^*_x(s, a)$ does not depend on x. Therefore the solution to Bellman's optimality equation also does not depend on x.

Bellman's optimality equation for a transition matrix regularized by averaging with the uniform transition matrix can be written in terms of a scaled discount factor and added constant. In this case, Bellman's optimality equation is
$Q^*(s, a) = r(s, a) + \gamma \sum_{s'}[((1 - \epsilon)T(s, a, s') + \epsilon\frac{1}{n})\text{max}_{a'} Q^*(s', a')]$
$Q^*(s, a) = r(s, a) + \gamma(1 - \epsilon) \sum_{s'} T(s, a, s')\text{max}_{a'} Q^*(s', a') + \gamma\frac{\epsilon}{n} \sum_{s'} \text{max}_{a'} Q^*(s', a')$
Letting $x = \gamma\frac{\epsilon}{n} \sum_{s'} \text{max}_{a'} Q^*(s', a')$, Bellman's optimality equation is:
$Q^*(s, a) = r(s, a) + x + \gamma(1 - \epsilon) \sum_{s'} T(s, a, s')\text{max}_{a'} Q^*(s', a')$

x is constant with respect to a, so by this first section of the proof, it does not affect the optimal policy. Therefore we can write the expression for the optimal policy at state s as:
$\pi_{opt}(s) = \text{argmax}_a Q^*(s, a)$
$\pi_{opt}(s) = \text{argmax}_a (r(s, a) + x + \gamma(1 - \epsilon) \sum_{s'} T(s, a, s')\text{max}_{a'} Q^*(s', a'))$
$\pi_{opt}(s) = \text{argmax}_a (r(s, a) + \gamma(1 - \epsilon) \sum_{s'} T(s, a, s')\text{max}_{a'} Q^*(s', a'))$

This is the optimal policy for the MDP with the original transition matrix and discount factor $(1 - \epsilon)\gamma$. This is equivalent to discount regularization, and matches the value of epsilon $\epsilon = \frac{\gamma-\gamma_l}{\gamma}$ that we derived in the previous section.

A.3.2. Dirichlet Prior Implied by Discount Regularization

The equivalence proof above demonstrates that, for a given value of ϵ, averaging the transition matrix with the uniform matrix or with the matrix of zeros yields the same policy. Averaging with the uniform matrix is only exactly equivalent to a

uniform prior if the sum of the transition counts is equal for all starting states, for a given action. Recall that $\epsilon = \frac{\sum \alpha_i}{\sum \alpha_i + \sum c_i}$, where c_i are the transition counts from the data and α_i are the parameters of the Dirichlet prior. We can solve to find the prior magnitude α_i implied by the choice of ϵ and observed transition counts $\sum c_i$ in the weighted average form. This reveals what Dirichlet prior we are implicitly using when we regularize by the weighted average uniform form, and consequently the Dirichlet prior implied by discount regularization.

From $\epsilon = \frac{\sum \alpha_i}{\sum \alpha_i + \sum c_i}$, observe $\sum \alpha_i = \frac{\epsilon}{1-\epsilon} \sum c_i$. We assume N states. For the uniform distribution, all α_i for a given state are the same, so substitute $\sum \alpha_i = N\alpha_i$ to get $\alpha_i = \frac{\epsilon}{1-\epsilon} \frac{\sum c_i}{N}$. Therefore, using a lower discount rate yields the same optimal policy as setting a uniform Dirichlet prior over each row of the transition matrix with magnitude $\frac{\epsilon}{1-\epsilon} \frac{\sum c_i}{N}$.

We can relate this back to the value of γ. Recall $\epsilon = \frac{\gamma - \gamma_l}{\gamma}$, where γ is the true value of the discount factor and γ_l is the lower value used for regularization. Plugging this into the expression above yields $\alpha_i = \frac{\gamma - \gamma_l}{\gamma_l} \frac{\sum c_i}{N}$. So discount regularization functions like a Dirichlet prior

$$T_{prior}(s,a) \sim \text{Dirichlet}(\frac{\gamma - \gamma_l}{\gamma_l} \frac{\sum c_i}{N}, ..., \frac{\gamma - \gamma_l}{\gamma_l} \frac{\sum c_i}{N}) \tag{4}$$

where again $\sum c_i$ is the total number of transitions in the data starting at state s.

B. Implementation Details

Algorithm 1 Regularization Loss Pseudocode

```
Input: MDP, epsilon list, regularization method
for i = 1 to 5000 do
    Generate data set: n trajectories of length l
    Estimate MDP from data
    for ε in epsilon list do
        Regularize transition matrices by amount ε
        Calculate optimal policy π of regularized MDP
        Calculate loss comparing π vs. true optimal policy in true MDP
    end for
end for
Average loss by ε value across all data sets
```

To compare policies resulting from different regularization methods, we implement the following procedure, summarized in Algorithm 1. Separately for each of the three example MDPs, we repeatedly generate data sets of trajectories from the true MDP. We estimate the transition and reward matrices as the MLE of the data. The estimate of the reward function at state-action pair (s, a) is then the mean of the observed rewards at (s, a). The estimated probability of transition from state s to state s' given action a is the number of times the transition from s to s' is observed given action a divided by the number of times state-action pair (s, a) is observed in the data set. For state-action pairs that are not observed in the data, we assume equal transition probabilities to all states and reward of 0.50.

For each of a list of values of ϵ between 0 and 1 (or for uniform prior, a list of multipliers between 0 and 1000), we regularize the estimated transition matrix for each action. We then find the optimal policy via policy iteration. Separately, we calculate the true optimal policy using the known, true MDP. We then evaluate the policy found from the estimated, regularized MDP and the policy from the true MDP, both in the true MDP. We compute loss as the weighted average difference in values of the two policies across all states, weighted by the starting state distribution.

To explore the impact on different aspects of the data set, we vary the trajectory starting state, the length and number of trajectories, and the probability of an action being generated from the optimal policy versus a random policy. For the Cliff Walk, trajectory starting states considered are uniformly random, start at S, or start within 2 states of G. For the Interconnected Grid, starting states are either uniformly random, limited to 5 of the 10 states, or limited to 1 state. In the case of Two Goals, starting states are uniformly random, starting in state 1 (next to the small reward) or starting in state 10 (next to the large reward).

The MSE is the squared difference in transition probabilities between the true and estimated transition matrices, averaged across all state-action pairs. For discount regularization, the weighted average form is not a true transition matrix. There is an implicit absorbing state that the agent enters with probability ϵ at each step. For discount regularization, we calculate the MSE in relation to the augmented transition matrix with the absorbing state and also without it.

C. Additional Results

In the case of discount regularization, the regularized form is not a true transition matrix, so we also plot its MSE taking into account the implicit absorbing state. We display plots of the MSE without the absorbing state as well because the scale allows for viewing the differences in detail.

C.1. Results by Distance from Optimal

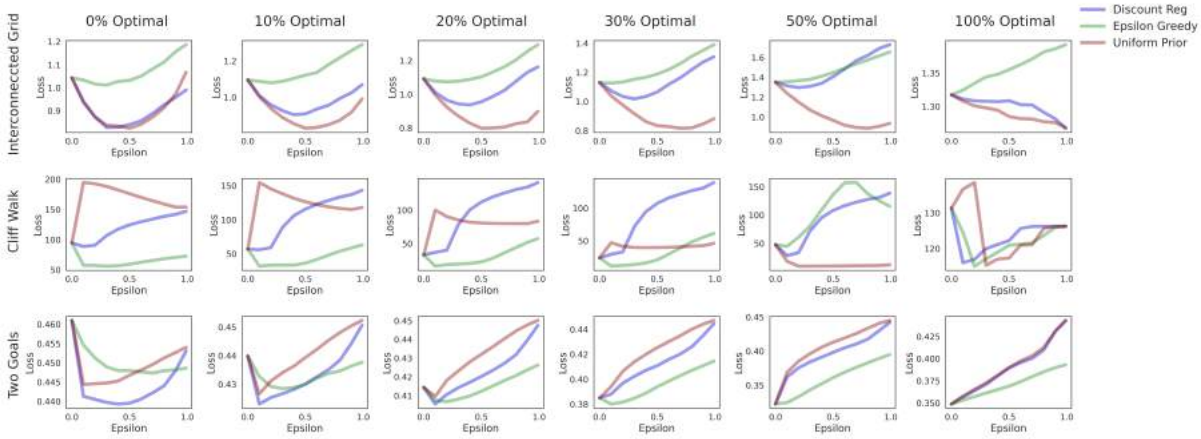

Figure 3. **Loss, varying probability that actions in the data set are drawn from optimal policy.** Random start states; 15 trajectories of length 10 each for Interconnected Grid and Two Goals; 25 trajectories of length 20 for Cliff Walk. Percentages chosen to show change in shape of curve.

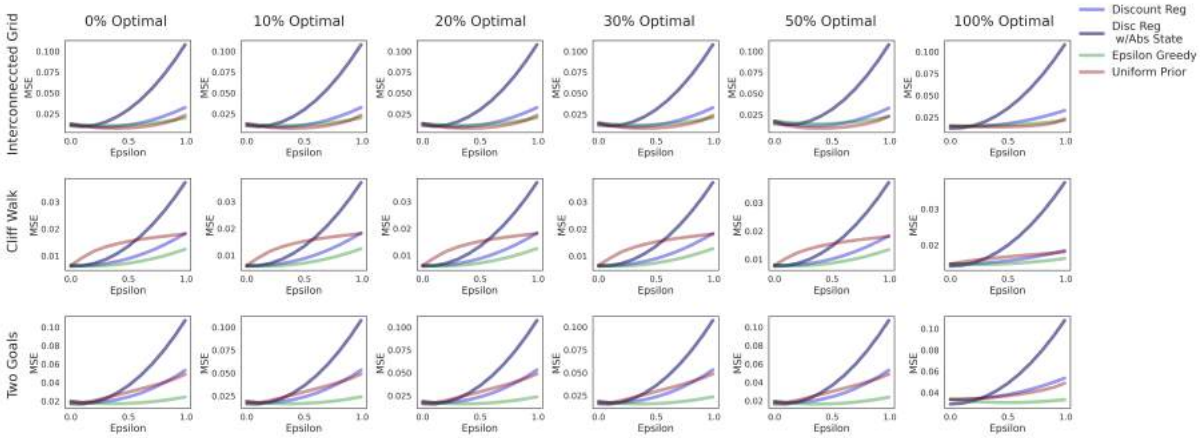

Figure 4. **MSE, varying probability that actions in the data set are drawn from optimal policy.** Random start states; 15 trajectories of length 10 each for Interconnected Grid and Two Goals; 25 trajectories of length 20 for Cliff Walk. Percentages chosen to show change in shape of curve.

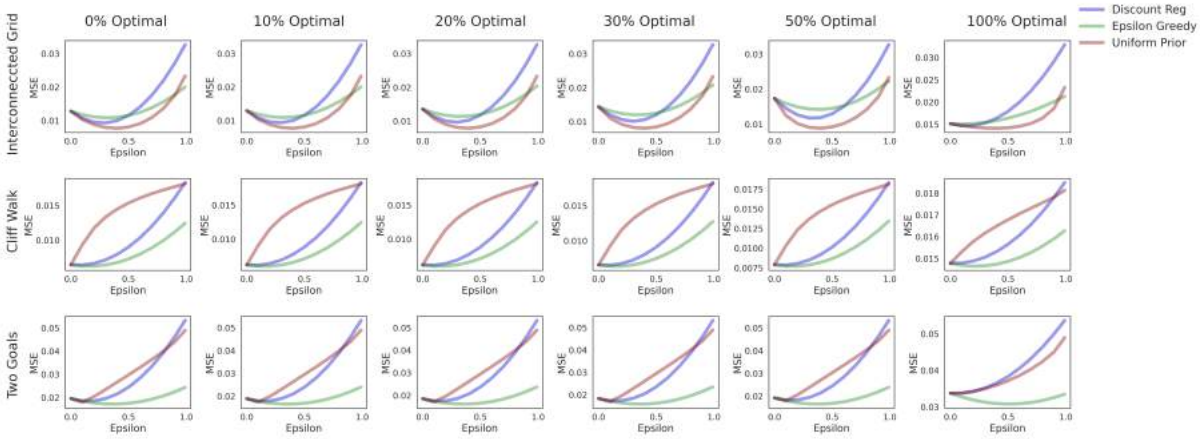

Figure 5. **MSE, varying probability that actions in the data set are drawn from optimal policy.** Random start states; 15 trajectories of length 10 each for Interconnected Grid and Two Goals; 25 trajectories of length 20 for Cliff Walk. Percentages chosen to show change in shape of curve. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.2. Results by Data Set Size

C.2.1. INTERCONNECTED GRID

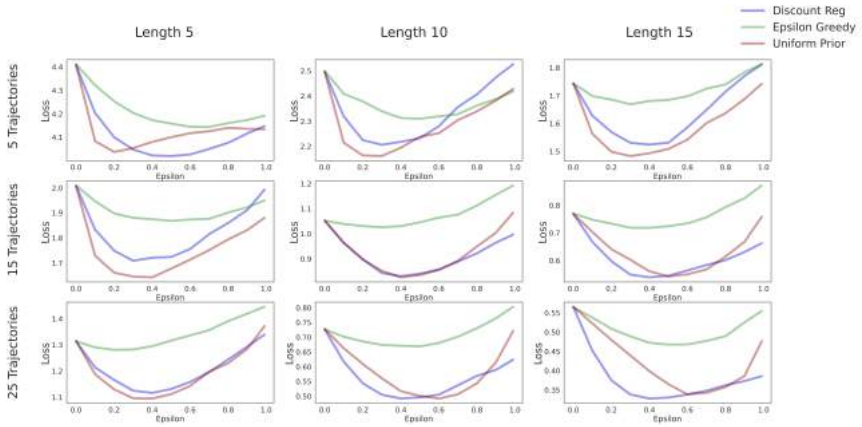

Figure 6. **Interconnected Grid Loss** varying number and length of trajectories in data set. Random start states, random policy.

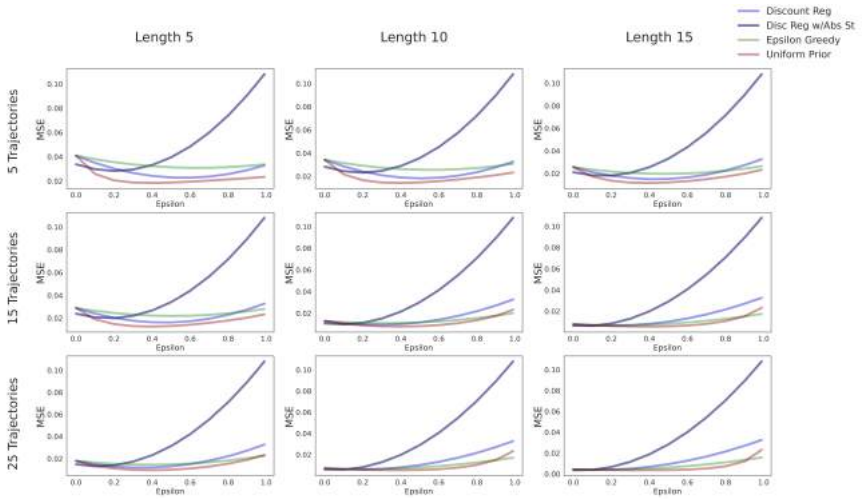

Figure 7. **Interconnected Grid MSE** varying number and length of trajectories in data set. Random start states, random policy.

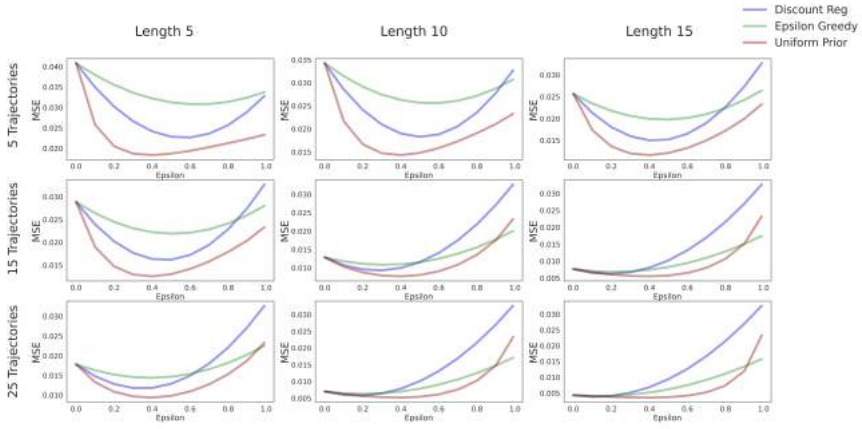

Figure 8. **Interconnected Grid MSE** varying number and length of trajectories in data set. Random start states, random policy. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.2.2. CLIFF WALK

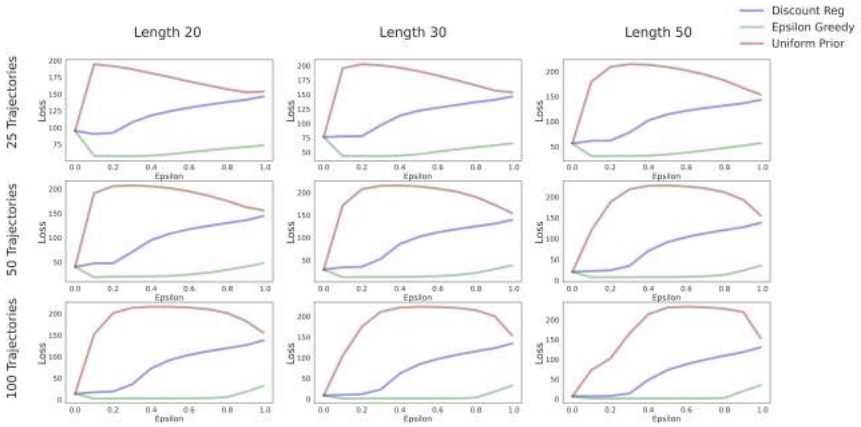

Figure 9. **Cliff Walk Loss** varying number and length of trajectories in data set. Random start states, random policy.

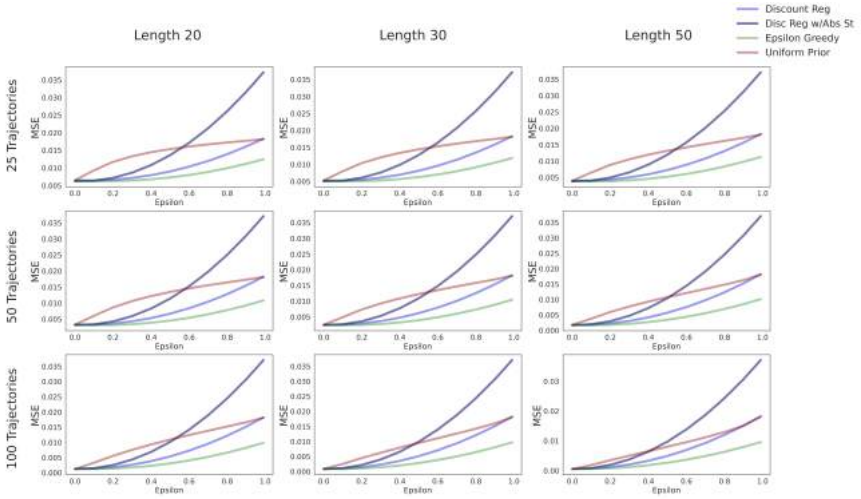

Figure 10. **Cliff Walk MSE** varying number and length of trajectories in data set. Random start states, random policy.

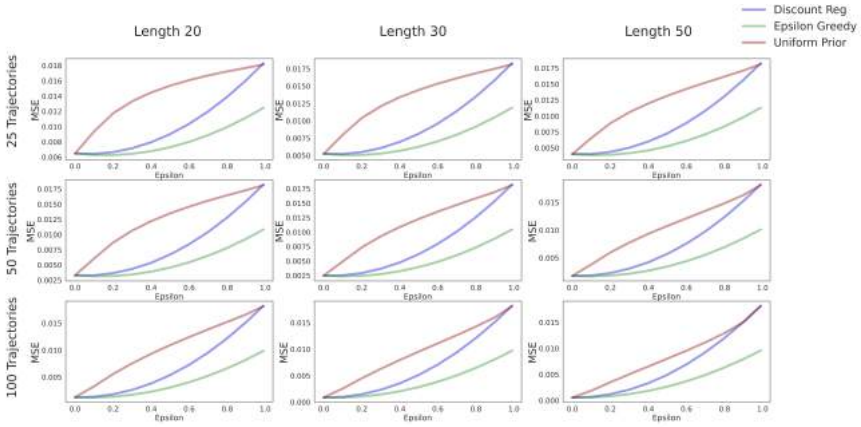

Figure 11. **Cliff Walk MSE** varying number and length of trajectories in data set. Random start states, random policy. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.2.3. TWO GOALS

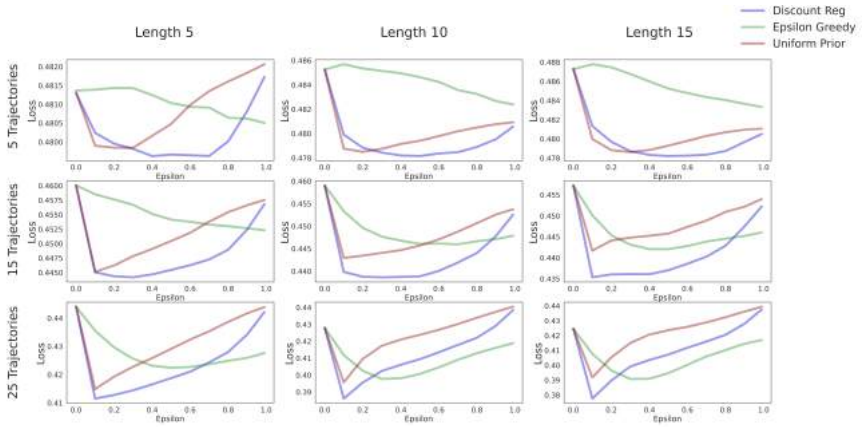

Figure 12. **Two Goals Loss** varying number and length of trajectories in data set. Random start states, random policy.

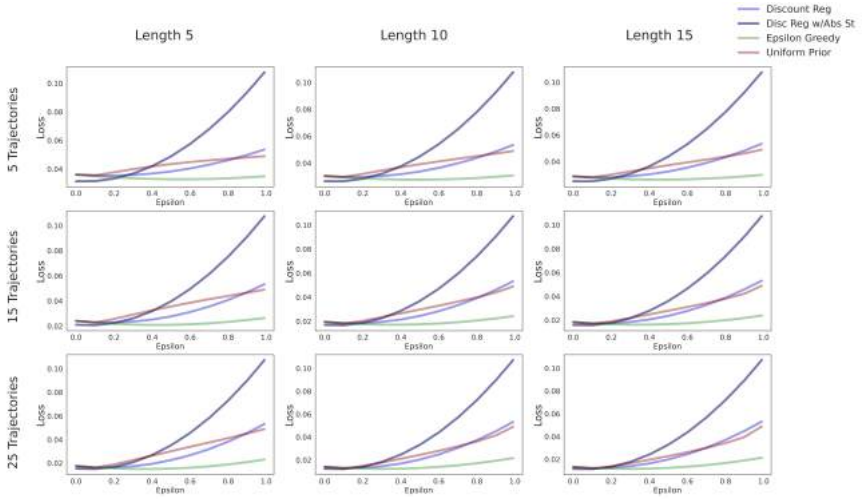

Figure 13. **Two Goals MSE** varying number and length of trajectories in data set. Random start states, random policy.

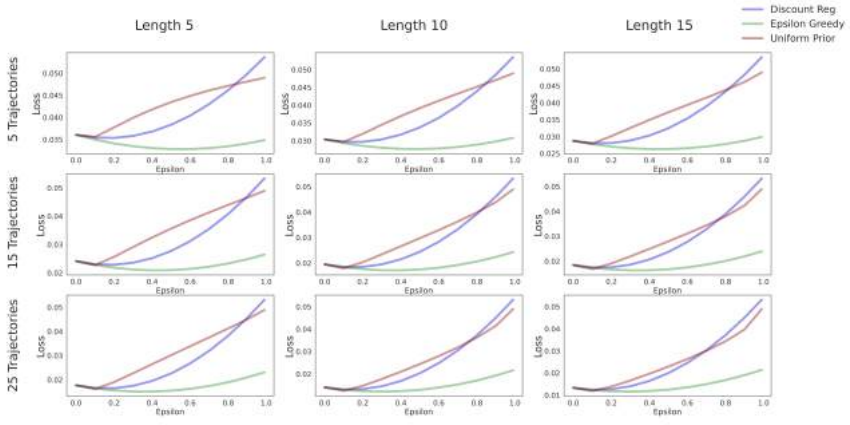

Figure 14. **Two Goals and MSE** varying number and length of trajectories in data set. Random start states, random policy. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.3. Results by Trajectory Start State

C.3.1. Interconnected Grid

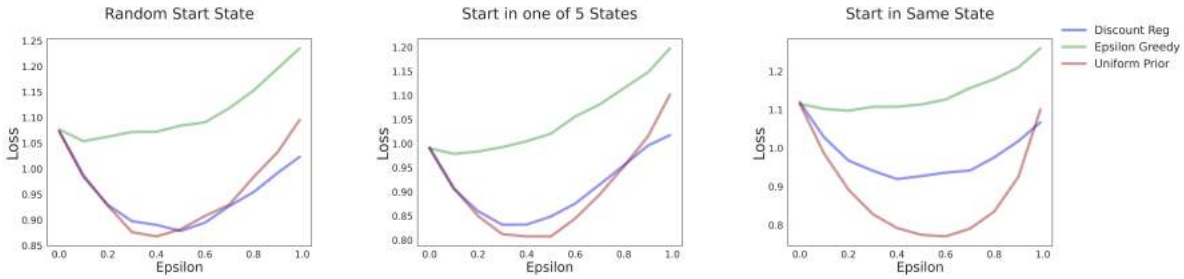

Figure 15. **Interconnected Grid Loss** varying start state. Random policy, 15 trajectories of length 10.

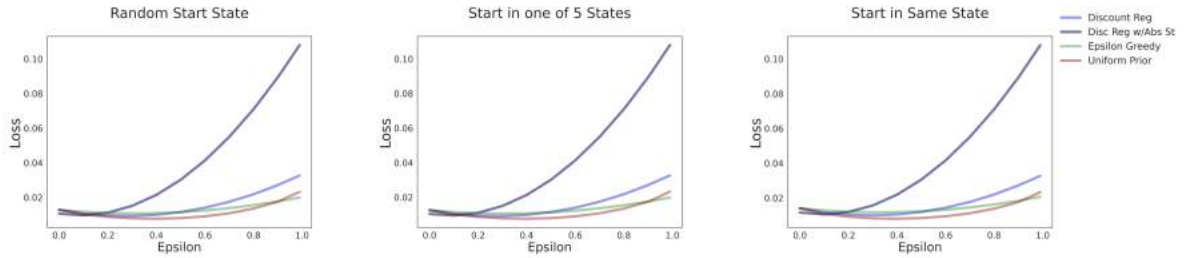

Figure 16. **Interconnected Grid MSE** varying start state. Random policy, 15 trajectories of length 10.

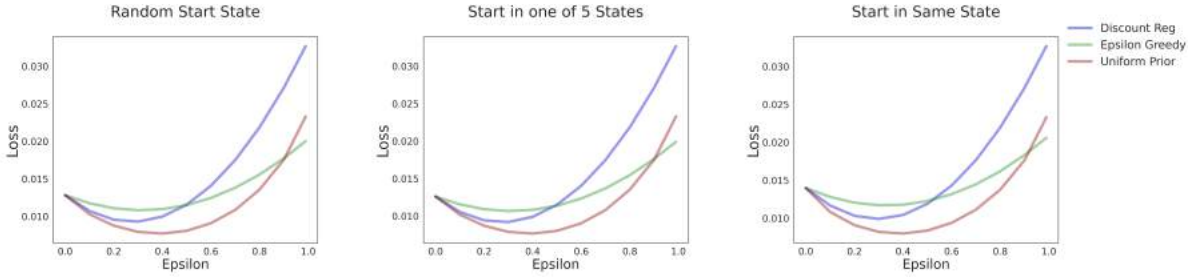

Figure 17. **Interconnected Grid MSE** varying start state. Random policy, 15 trajectories of length 10. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.3.2. CLIFF WALK

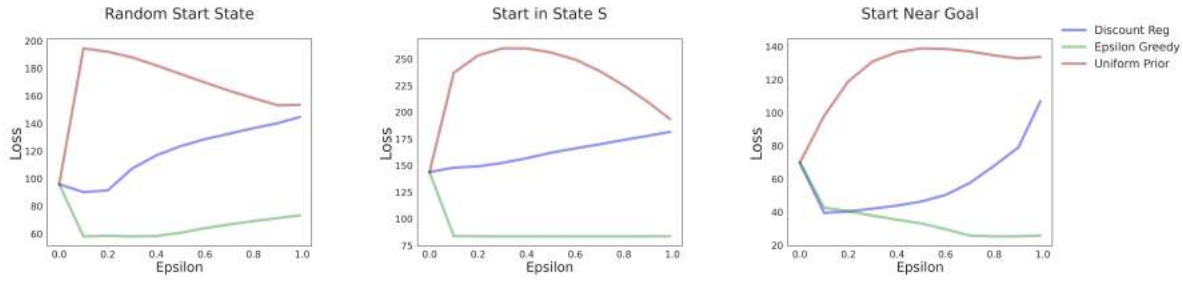

Figure 18. **Cliff Walk Loss** varying start state. Random policy, 25 trajectories of length 20.

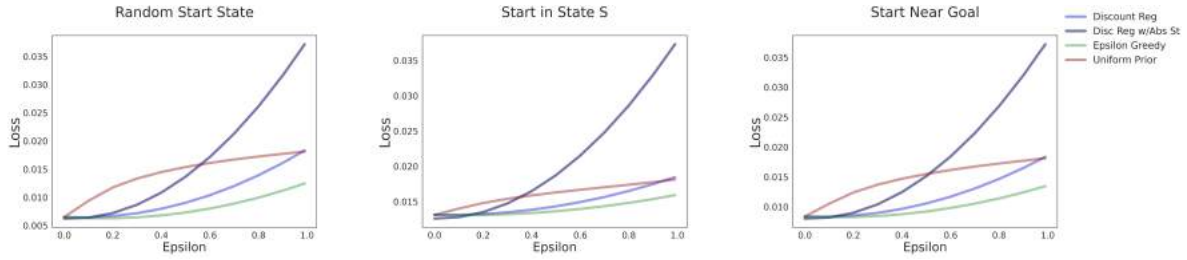

Figure 19. **Cliff Walk MSE** varying start state. Random policy, 25 trajectories of length 20.

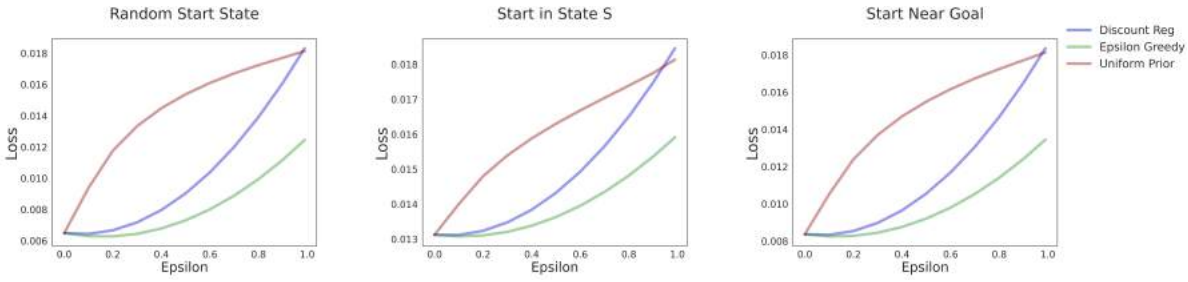

Figure 20. **Cliff Walk MSE** varying start state. Random policy, 25 trajectories of length 20. Discount regularization with absorbing state removed for scale, to show detail on other curves.

C.3.3. TWO GOALS

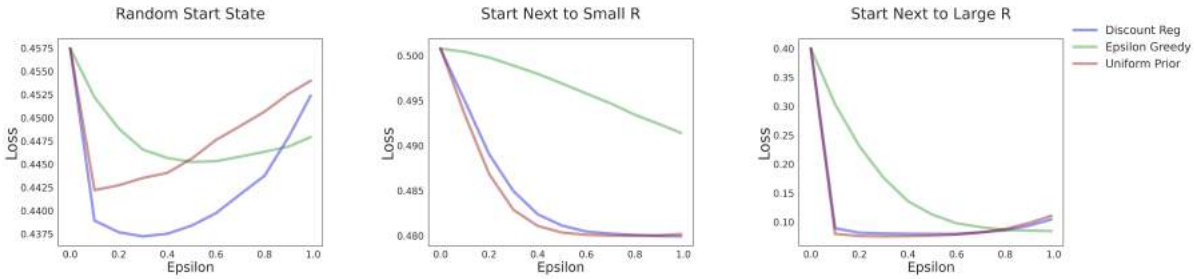

Figure 21. **Two Goals Loss** varying start state. Random policy, 15 trajectories of length 10.

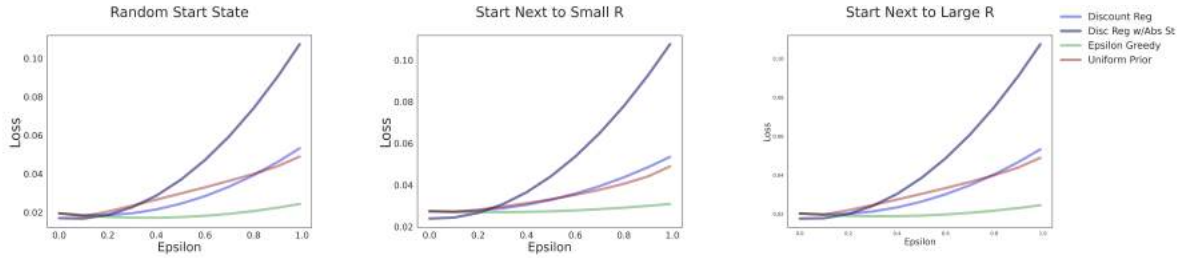

Figure 22. **Two Goals MSE** varying start state. Random policy, 15 trajectories of length 10.

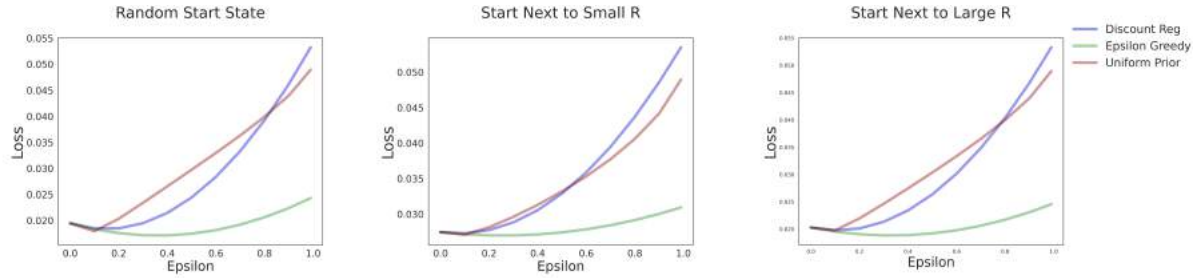

Figure 23. **Two Goals MSE** varying start state. Random policy, 15 trajectories of length 10. Discount regularization with absorbing state removed for scale, to show detail on other curves.