A Notation

Symbol		Description
n	$\in \mathbb{Z}^+$	Number of states.
m	$\in \mathbb{Z}^+$	Number of features.
π	$\in \mathbb{R}^n$	on-policy distribution.
μ	$\in \mathbb{R}^n$	sampling distribution, may be on- or off-policy.
υ	$: \mathbb{R}^+ \to \mathbb{R}^n$	apparent distribution induced by η-regularizing the emphatic correction of off-policy μ to on-policy π
η	$\in \mathbb{R}_0^+$	ℓ_2 regularization parameter
η_m	$\in \mathbb{R}_0^+$	ℓ_2 regularization parameter for emphasis model in COF-PAC (the Emphatic algorithm we analyze)
η_v	$\in \mathbb{R}_0^+$	ℓ_2 regularization parameter for value model in COF-PAC (the Emphatic algorithm we analyze)
p	$\in [0, 1]$	distribution parameter used to express a family of possible sampling distributions.
Φ	$\in \mathbb{R}^{[n \times m]}$	Feature basis for the value function
$\hat{w}$	$\in \mathbb{R}^{[m \times 1]}$	Linear weights for value function, fit using least-squares regression of V on Φ.
$w^*(\eta)$	$\in \mathbb{R}^{[m \times 1]}$	Linear weights for value function, learned using TD.
$\Phi w^*(\eta)$	$\in \mathbb{R}^{[n \times 1]}$	Learned value function
V	$\in \mathbb{R}^{[n \times 1]}$	True value function
$\|V\|$	$\in \mathbb{R}$	Error from guessing zeros, equivalent to the threshold for a vacuous example
$\|x\|$	$\in \mathbb{R}_0^+$	ℓ_2-norm of vector or matrix x, equal to $\sqrt{x^\top x}$
$\|x\|_D$	$\in \mathbb{R}_0^+$	ℓ_2-norm of vector or matrix x under D, equal to $\sqrt{x^\top D x}$

B Example Details

We provide a more detailed explanation of our examples.

B.1 "Vacuous" models

Without ℓ_2 regularization, our linear model fails with asymptotic error. As this penalizes the ℓ_2-norm of the learned weights, this removes the asymptote and so we can no longer use the existence of an asymptote as evidence of failure. Instead, we propose a different definition of failure by noting that, in the limiting case, regularization drives the learned weights to zero ($\lim_{\eta \to \infty} w^*(\eta) = \vec{0}$). The learned value function $\Phi \cdot \vec{0} = \vec{0}$ has no information about the true value function. We argue that if the error with any $\eta \in \mathbb{R}^+$ is never better than this case then the model is vacuous and hence adopt the threshold error of $\|\Phi \cdot \vec{0} - V\| = \|V\|$ to call a model vacuous. This explains the failure condition in Equation 8.

B.2 Details of regularization example

This provides numeric details for Example 1.

Example 5. When TD is regularized, there may exist some off-policy distribution at which TD learns a vacuous model. In notation:

$$\|\Phi w^*(\eta) - V\| \geq \lim_{\eta \to \infty} \|\Phi w^*(\eta) - V\| = \|\Phi \vec{0} - V\| = \|V\| \qquad \forall \eta \in \mathbb{R}_0^+ \tag{13}$$

Details. We use the same setting as in Section 3. We observe that $\hat{w} = [1, -1]^\top$ minimizes the least-squares error $\|\Phi \hat{w} - V\|$, and further observe that a sufficient (but not necessary) condition for

a solution to be vacuous is that $\hat{w}^\top w^*(\eta) \leq 0$. Solving:

$$0 = \hat{w}^\top w^*(\eta) = \frac{\eta p - 0.233\eta - 0.304p^2 + 0.276p - 0.025}{\eta^2 + 1.44\eta p + 0.215\eta - 0.193p^2 + 0.175p - 0.016} \implies p \in \{0.102636, \ldots\} \tag{14}$$

We verify that TD is vacuous at $p = 0.102636$ by computing the TD error at convergence:

$$\|\Phi w^*(\eta) - V\|^2\big|_{p=\tilde{p}} = \frac{\eta^2(0.148 + 0.744\eta + \eta^2)}{\eta^2(0.132 + 0.727\eta + \eta^2)}\|V\|^2 \geq \|V\|^2 \quad (\forall \eta \in \mathbb{R}^+) \tag{15}$$

Since the fraction term in Equation 15 is obviously improper, we can conclude that our example will always have at least $\|V\|$ error over all η, and is therefore vacuous. □

We note that the error is not defined at $\eta = 0$ because this corresponds to a model divergence similar to our introductory example. In practice, the TD fixed point will still converge to a vacuous solution:

$$\lim_{\eta->0} \|\Phi w^*(\eta) - V\|^2 = {}^{0.148}\!/_{0.132}\|V\|^2 > \|V\|^2 \tag{16}$$

B.3 Additional simple regularization example

We present a second example where the error is stationary with respect to the regularization parameter. This is worse than Example 5 because we are able to show that the point the model converges to is *independent* of regularization.

Example 6. When TD is regularized, there may exist some off-policy distribution at which the TD fixed point is independent of the regularization parameter.

Details. We use the same setting as in Section 3, except the value function is $V = [1,\ 1,\ 1.05]^\top$ and basis Φ selected to have small representation error $\|\Pi_D V - V\| \leq \epsilon$:

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ {}^1\!/_2(1.05 + \epsilon) & -{}^1\!/_2(1.05 + \epsilon) \end{bmatrix} \qquad \text{where } \epsilon > 0 \tag{17}$$

. We set $\epsilon = 10^{-4}$ and write down $w^*(\eta)$ in terms of g, a scalar function of η and p:

$$w^*(\eta) = (A + \eta I)^{-1}\vec{b} = \frac{(2\eta + p)(0.925 - 1.29p)}{100\eta^2 + 47.4p\eta + 1.85\eta - 1.30p^2 + 0.927p} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} \equiv g(p, \eta) \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{18}$$

When $g(p, \eta) \leq 0$, the TD solution is vacuous. We show that directly:

$$\|\Phi w^*(\eta) - V\| = \|g(p, \eta)\Phi * [1, -1]^\top - \Phi * [1, -1]^\top\| = \|g(\eta) - 1\| \cdot \|V\| \tag{19}$$

When $g(p, \eta) \leq 0$, then $\|g(p, \eta) - 1\| \geq 1$ for all η and the TD solution is vacuous. We find such a solution by noting the numerator has two roots in p, one of which corresponds to a vacuous solution: $g(0.715083, \eta) = 0\ (\forall \eta)$, and this completes the example! In this setting, when TD updates follow the sampling distribution $p \approx 0.715083$, the error of the model at convergence is always $\|V\|$ regardless of regularization. Our example converges to the same vacuous value regardless η. □

In Figure 6, we can see that the TD error intersects the $\eta \to \infty$ line immediately before and after the singularity. Our counterexample corresponds to the second root (that is, the intersection point at higher p.) This is because that corresponds to the stationary point between the asymptote that is crushed and the error on the right that increases. If our simpler derivation proved unsatisfying, we can also derive this counterexample using this fact:

$$0 = \frac{d}{d\eta}\hat{w}^\top w^*(\eta) = \frac{p(p - 0.715083)}{p(p - 0.714303)^2} \tag{20}$$

From this, we can easily see that the counterexample is at $p = 0.715083$. And this completes the example! We have discovered some p at which the TD error is always at least $\|V\|$, regardless of regularization, and so our example learns a vacuous value function.

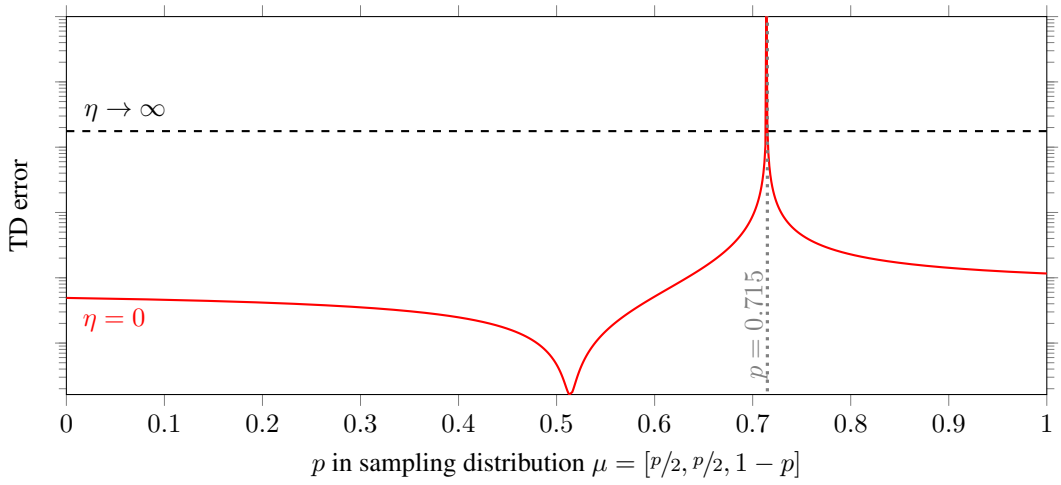


Figure 6: We plot TD error against p for our three-state MP with $\epsilon = 10^{-4}$. This shape is similar to that in [6]. There is a minima close to π ($p \approx 0.5$), and an asymptote at the singularity ($p \approx 0.715$). At different levels of regularization the error function moves between the unregularized case ($\eta = 0$) and the limiting case ($\eta \to \infty$), as analyzed in Section 3.1. We show that there is some p at which the error is never below the $\eta \to \infty$ line.

B.3.1 *Breaking the Deadly Triad* and our counterexample.

In light of our counterexample we examine the work of [21] in which the authors derive a bound for the regularized TD error under a novel double-projection update rule. We apply our example to their bound b to show that their method permits vacuous TD solutions and doesn't quite break the deadly triad. Starting from Equation 9:

$$\|\Phi w^*(\eta) - V\| \leq b(\eta, \xi) = \frac{1}{\xi}\left(\frac{\sigma_{\max}(\Phi)^2}{\sigma_{\min}(\Phi)^4 \sigma_{\min}(D)^{2.5}} \cdot \|V\|\eta + \|\Pi_D V - V\|\right) \tag{21}$$

$$= {}^1\!/\!_\xi \cdot (38.0\eta + 8.07 \times 10^{-5}) \tag{22}$$

for $\xi \in [0, 1]$, where $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular value respectively. Theorem 2 from [21] bounds η, and therefore also b:

$$\eta > \arg\inf_\eta \|\Phi - C_0\| = 0.367(6.86 - 13.7\xi + 6.86\xi^2)^{-1} \tag{23}$$

$$\inf_\xi b(\xi, \eta) = 13.8 = 7.86 * \|V\| \tag{24}$$

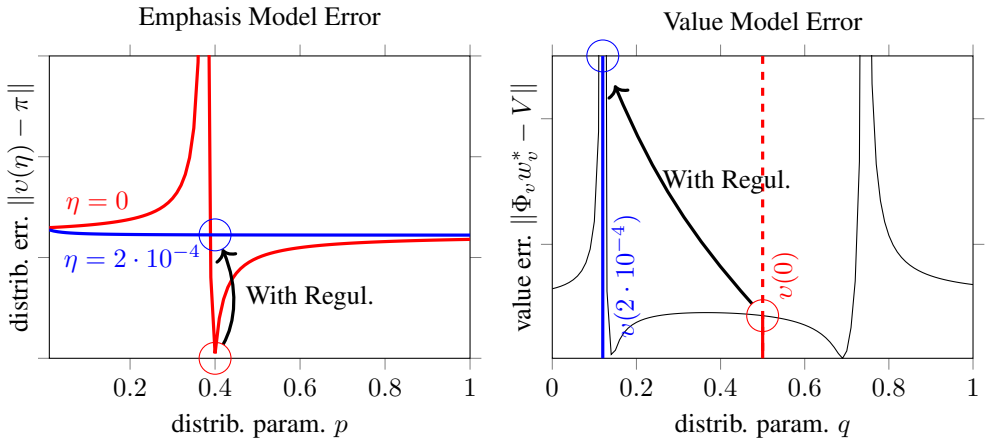
Under our example, their method bounds the error at no more than $7.86 * \|V\|$, which is a very loose bound that permits vacuous solutions. This illustrates the risk of trying to regularize away singularities, particularly in theoretical work.

Investigating the cause of the loose bounds reveals that the presence of $\sigma_{\min}(D)^{2.5}$ in 9 is largely responsible. As D is a diagonal matrix encoding the sampling distribution, $\sigma_{\min}(D)$ is the smallest sampling rate of any state, and so the bound must be at least $\frac{\eta}{\xi n^{2.5}}$ for any perfectly representable n-state MP. Unfortunately, this appears to be fundamental limit caused by finding a linear bound to an error that scales non-linearly, and following their derivation in the appendix does not readily admit a way to improve this.

B.4 Small-Eta Error

Our simplified example allows us to show this easily.

Example 7. When TD is regularized, the model may diverge around (typically small) values of η.



(a) distribution is $[p/2,\ p/2,\ (1-p)]$

(b) distribution is $[(1-q)/2,\ q/2,\ 0.5]$

Figure 7: Regularization distorts the emphasis model (left), which induces the value function (right) to move to a singularity. Unregularized models are shown in red, regularized models in blue. Regularization can interact with emphasis models to significantly worsen learned value functions.

Details. We set $p = 0.9$ and solve for $\det(A + \eta I) = 0$:

$$0 = 100\eta^2 + 47.4p\eta + 1.85\eta - 1.30p^2 + 0.927p \tag{25}$$

$$\eta = 0.00482577 \quad \vee \quad \eta = -0.45 \tag{26}$$

Note that the denominator of $g(p, \eta)$ is proportional to $\det(A + \eta I)$, and so $g(0.9, \eta)$–and the error at the TD fixed point–can be made arbitrarily large by selecting η close to 4.83×10^{-3}. As this is the only positive root, the model does not diverge at other values.

B.5 Emphatic approaches and our counterexample

We use an MP with the same transition function as in Figure 1a, with separate bases Φ_m and Φ_v for the emphasis and value stages respectively. We assume that our interest in all states is uniformly $i = 1$.

We begin by setting the off-policy sampling distribution of $\mu = [.2\ .2\ .6]$, used as the diagonal matrix $D_\mu = \text{diag}(\mu)$. Thanks to the simple structure of our example, we know the emphasis is $m = \frac{i}{1-\gamma} \cdot \pi D_\mu^{-1} \propto (5/4, 5/4, 5/6)$. We select a basis that allows us to represent this emphasis:

$$\Phi_m = \begin{bmatrix} 5/4 & 0 \\ 0 & -1/100 \cdot 5/4 \\ 5/12 & -1/100 \cdot 5/12 \end{bmatrix} \tag{27}$$

We deliberately choose Φ_m to have a poor condition number for reasons that will become apparent later. We can represent $c \cdot (5/4, 5/4, 5/6)$ exactly for any constant c:

$$\Phi_m \cdot (1, -100) \cdot c = c \cdot (5/4, 5/4, 5/6) \tag{28}$$

Using Equation 5 from [20], we define the matrices:

$$C_m = \Phi_m^\top D_\mu \Phi_m = \begin{bmatrix} 0.417 & -1.04 \times 10^{-3} \\ -1.04 \times 10^{-3} & 4.17 \times 10^{-5} \end{bmatrix} \tag{29}$$

$$A_m = \Phi_m^\top (I - \gamma P^\top) D_\mu \Phi_m = \begin{bmatrix} 0.159 & 1.536 \times 10^{-3} \\ 1.536 \times 10^{-3} & 1.59 \times 10^{-5} \end{bmatrix} \tag{30}$$

And we apply these to the formulation in Lemma 3 and compute the emphasis weights as a function of the regularization $w_m : \mathbb{R}_0^+ \to \mathbb{R}^+$:

$$w_m^*(\eta) = (A_m^\top C_m^{-1} A_m + \eta I)^{-1} A_m^\top C_m^{-1} \Phi_m^\top D i \tag{31}$$

We can then use this to compute the new apparent distribution υ, which is the effective distribution that the updates to the value model see, and it is equal to the emphasis multiplied by the off-policy distribution.

$$\upsilon(\eta) = \Phi_m \cdot w_m^*(\eta) \cdot D \tag{32}$$

Without any regularization, this should be exactly equal to the on-policy distribution.

$$\upsilon(0) = [0.25\ 0.25\ 0.5] \equiv \pi \tag{33}$$

When we compute this value with a small amount of regularization $\eta = 2 \times 10^{-4}$, we observe that the apparent distribution drifts far away from the on-policy distribution.

$$\upsilon(2 \times 10^{-4}) = [0.44\ 0.06\ 0.5] \tag{34}$$

The proximate cause of this is the poor condition number of C, caused by the $\frac{1}{100}$ scale factor applied to the second column of Φ_m. This allows η to affect different columns by different (relative) amounts in the definition of $w^*(\eta)$, which pushes it away from the symmetric solution. See this error shift in Figure 7a.

So far, we have shown how regularization causes a shift in the apparent distribution that the TD updates see. To complete the example we show how this moves the fixed point of the value function away from a stable point into an asymptote where it may grow without bounds. This second phase follows in the same pattern as the first phase, starting with the desired value function: $V = [1\ 2.69\ 1.05]$ and a basis that can almost exactly represent the value function:

$$\Phi_v = \begin{bmatrix} 1 & 0 \\ 0 & -2.69 \\ 1/2(\epsilon + 1.05) & -1/2(\epsilon + 1.05) \end{bmatrix}$$
$$\epsilon = 2 \times 10^{-4}$$

We use this basis to compute the state-rewards $R = (I - \gamma P)V = [-0.43\ 1.26\ -0.38]$ and define the matrices A_v and C_v and the solution $w_v^*(\eta)$:

$$A_v = \Phi_v^\top (I - \gamma P^\top) D \Phi_v$$
$$C_v = \Phi_v^\top D \Phi_v$$
$$w_v^*(\eta) = (A_v^\top C_v^{-1} A_v + \eta I)^{-1} A_v^\top C_v^{-1} \Phi_v^\top D R$$

We can use this solution to compute the error between the value function and the true values, $\|\Phi_v w_v(\eta) - V\|$. First, under the corrected distribution without regularization $\upsilon(0) \equiv \pi$:

$$\Phi_v w_v^*(0)|_{D=\text{diag}(\upsilon(0))} = 0.000865$$

Then, with regularization in the value function (but not in the emphasis function):

$$\Phi_v w_v^*(2 \times 10^{-4})|_{D=\text{diag}(\upsilon(0))} = 0.0162$$

Then, under the apparent distribution υ induced by use of regularization in the emphasis function, without and with regularization:

$$\Phi_v w_v^*(0)|_{D=\text{diag}(\upsilon(2\times10^{-4}))} = 418.601$$
$$\Phi_v w_v^*(2 \times 10^{-4})|_{D=\text{diag}(\upsilon(2\times10^{-4}))} = 3.00$$

It is immediately obvious that the use of regularization in the emphasis function causes the learned value function to be incorrect. Including a regularizing term in the value estimate is not sufficient to fix the value function. This completes the example. □

B.5.1 Kolter's non-expansion condition and our counterexample.

In the construction of COF-PAC, a key assumption made is that both the emphasis and value models are not subject to runaway TD [20, asm. 4]. Specifically, they make the strong assumption that Kolter's relaxed-contraction condition [6, eqn. 10] holds in the emphasis model at μ and value model at υ. Kolter's condition selects a convex subset of distributions under which one-step transition followed by projection onto Φ is non-expansive. We illustrate these regions in Figure 8. Even in the one-dimensional parameterization shown, this condition only holds in a small sub-region of the space and therefore appears to be a very strong condition. Empirically determining if such a condition holds (or training models to enforce it) may be possible with TD-DO [6, sec 4.1], but it is not clear how that method interacts with regularization.

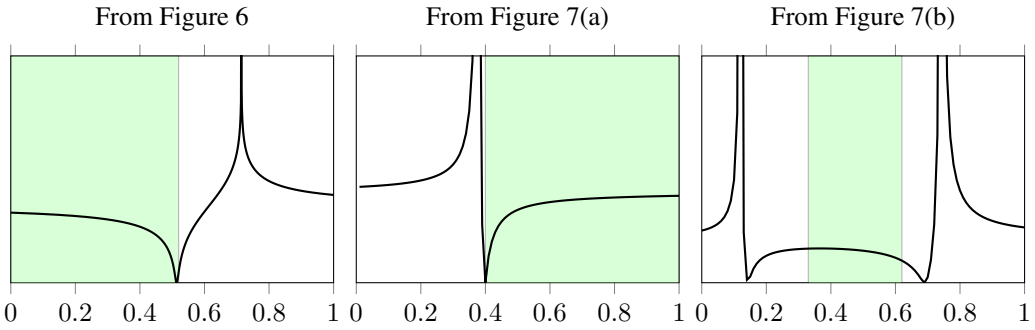


Figure 8: Kolter's non-expansion condition holds in the shaded region of each graph.

C Applied to multi-layer networks

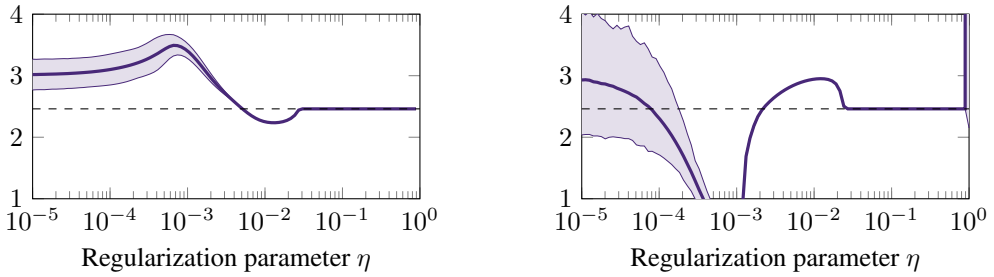
We also use a variant of our example to study how the deadly triad appears in multi-layer networks. As illustrated in Figure 11, we replace each self-loop with two additional states, forming a clique with the original state. The resultant MP has $n = 9$ states; we define a deterministic obervation function $o : \mathcal{S} \to \mathbb{B}^6$. where each state is encoded as the concatenation of the one-hot vector of its subscripts. The value function is assigned pseudo-randomly in range $[-1, 1]$, and a consistent reward function is assigned. We select the family of sampling distributions $\mu \propto [4p,\ 1p,\ 1p, 4p,\ 1p,\ 1p, 8(1-p),\ 4(1-p),\ 4(1-p)]$, where the on-policy distribution is at $p = 0.5$.

We wish to learn the model with a two-layer network with $k < n$ nodes in the inner layer. We define the network as $f(o(s_{i,j})) = \tan^{-1}(o(s_{i,j}) * \omega_1) * \omega_2$. The parameters $\omega_1 \in \mathbb{R}^{6\times k}, \omega_2 \in \mathbb{R}^{k\times 1}$ are trained to convergence using simple TD updates with semi-gradient updates, a fixed learning rate, and without a target network.

In addition to the example in Figure 5b, we present an additional example in Figure 9. The same Markov process, at a different off-policy distribution, attains a curve where the non-vacuous region lies before the divergent region, similar to the second row in Figure 3a. An added observation is that these two graphs are mutually incompatible – there is no fixed η that can simultaneously do better than vacuity in both, which promotes the idea of testing multiple regularization parameters or using an adaptive regularization scheme.

C.1 Overparameterization does not solve this problem

Baird's counterexample [18] shows how, in the linear case, that off-policy divergence can also happen under overparameterization, as long as some amount of function approximation occurs. It is not obvious that this conclusion persists in the neural network case, so we include an additional example showing that the parameterization doesn't resolve small-η divergence.



(a) $p = 0.31$ (Same as Figure 5b)

(b) $p = 0.95$

Figure 9: The relationship between error and η at different off-policy distributions, showing mutually incompatible regularization behavior. The shaded range indicates the region between the 5th and 95th percentile of 100 differently-initialized models.

In Figure 10 we plot models with 3 to 13 nodes in the hidden layer. For reference, the MDP has 9 states, so some models under-parameterize and some models over-parameterize. We observe that, in the low-regularization regime, increasing the number of parameters improves the error slightly. However, increasing the number of parameters in the hidden layer does not change the behavior in the the small-η divergence region.

C.2 Relationship to modern RL algorithms

It is still not obvious how strongly this instability affects modern RL algorithms, which are also sensitive to a variety of other failure modes. Unlike our analysis, the sampling distribution changes during training, and regularization mechanisms are more complex than simple ℓ_2 penalities. The exact relationship between the instabilities we study and RL algorithms is an open problem, but we offer two pieces of indirect evidence suggesting there is a link.

First, in the offline/batch RL literature, it is well-known that online RL algorithms naively applied can catastrophically fail if the learned policy is not consistent with the data distribution. This is known as the distribution shift problem, [9, p. 26] and offline RL algorithms are generally constructed to explicitly address this. Second, when using experience replay buffers in online RL algorithms, policy quality generally improves when older transitions are more quickly evicted [3]. However, there are multiple factors at work here, and it is not possible to separate out the instability from off-policy sampling from the remaining factors.

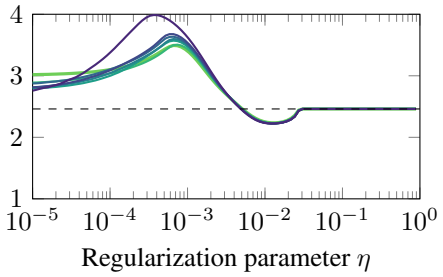
D Markov Processes

We use a three-, five- and nine-state Markov Processes to generate examples for this paper. Here we give details of the construction of each example. Mathematica code for all examples is included in the supplementary material.

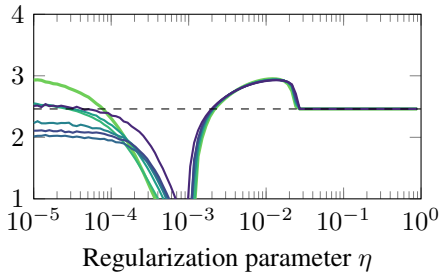
D.1 Three-state

The construction of the three-state MDP is described in Section 3 and illustrated in Figure 1a. This example is used in Examples 1 and 3 For completeness, the transition matrix is:

$$\frac{1}{4}\begin{bmatrix}1 & 1 & 2\\ 1 & 1 & 2\\ 1 & 1 & 2\end{bmatrix} \tag{35}$$



(a) $p = 0.31$



(b) $p = 0.95$

Figure 10: The relationship between η and error with different amount of model parameterization (with 3, 5, 7, 9, 11, 13, and 64 nodes in the hidden layer, corresponding to darkening colors.)

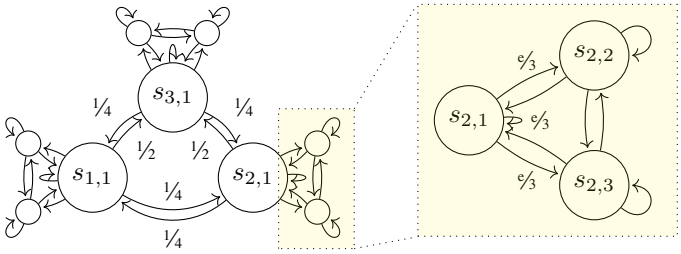


Figure 11: Our three-state counter-example MP is extended to nine states to illustrate how the deadly triad problem could manifest in multi-layer neural networks. The self-loop in the original example is replaced with a clique with uniform transitions except as labelled with the original edge weight e.

D.2 Nine-state

This example is used to train neural networks. The construction is based on the three-state example and the construction is illustrated in Figures 1b and 11. The transition matrix (with omitted zeros) is:

$$\begin{bmatrix} 1 & 1 & 1 & 3 & & & 6 & & \\ 4 & 4 & 4 & & & & & & \\ 4 & 4 & 4 & & & & & & \\ 3 & & & 1 & 1 & 1 & 6 & & \\ & & & 4 & 4 & 4 & & & \\ & & & 4 & 4 & 4 & & & \\ 3 & & & 3 & & & 2 & 2 & 2 \\ & & & & & & 4 & 4 & 4 \\ & & & & & & 4 & 4 & 4 \end{bmatrix} \tag{36}$$

and the observation function that forces the neural network to approximate is:

$$o : \mathcal{S} \to \mathbb{R}^6 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{37}$$

D.3 Five-state

We use this to generate Figure 3a. The transition matrix is:

$$\begin{bmatrix} .4 & .4 & .2 & 0 & 0 \\ .4 & .4 & .2 & 0 & 0 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & 0 & .2 & .4 & .4 \\ 0 & 0 & .2 & .4 & .4 \end{bmatrix} \tag{38}$$

We set value function $V = [1, 1, 1.05, 1, 1]^\top$, $\gamma = 0.99$, $\epsilon = 0.05$ and set basis:

$$\Phi = \frac{1}{3} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 3 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \tag{39}$$

We also parameterize the off-policy distribution as:

$$D = \frac{1}{2}\text{diag}([pq, pq, 2(1-p), p(1-q), p(1-q)]) \tag{40}$$

where $p, q \in (0, 1)$. We verify that $\sum D = 1$ over this domain. The on-policy distribution is $\pi = {}^{1}/_{12}[2, 3, 2, 3, 2]$. The plots correspond to the off-policy distributions:

1. $p \to 0.77, q \to 0.85$
2. $p \to 0.4, q \to 0.9$
3. $p \to 0.02, q \to 0.2$