# The Pitfalls of Regularization in Off-Policy TD Learning

**Gaurav Manek**
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
gmanek@cs.cmu.edu

**J. Zico Kolter**
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
zkolter@cs.cmu.edu

## Abstract

Temporal Difference (TD) learning is ubiquitous in reinforcement learning, where it is often combined with off-policy sampling and function approximation. Unfortunately learning with this combination (known as the *deadly triad*), exhibits instability and unbounded error. To account for this, modern RL methods often implicitly (or sometimes explicitly) assume that regularization is sufficient to mitigate the problem in practice; indeed, the standard deadly triad examples from the literature can be "fixed" via proper regularization. In this paper, we introduce a series of new counterexamples to show that the instability and unbounded error of TD methods is *not* solved by regularization. We demonstrate that, in the off-policy setting with linear function approximation, TD methods can fail to learn a non-trivial value function under *any* amount of regularization; we further show that regularization can induce divergence under common conditions; and we show that one of the most promising methods to mitigate this divergence (Emphatic TD algorithms) may also diverge under regularization. We further demonstrate such divergence when using neural networks as function approximators. Thus, we argue that the role of regularization in TD methods needs to be reconsidered, given that it is insufficient to prevent divergence and may itself introduce instability. There needs to be much more care in the practical and theoretical application of regularization to RL methods.

## 1 Introduction

Temporal Difference (TD) learning is a method for learning expected future-discounted quantities from Markov processes, using transition samples to iteratively improve estimates. This is most commonly used to estimate expected future-discounted rewards (the *value function*) in Reinforcement Learning (RL). Advances in RL allow us to use powerful function approximators, and also to use sampling strategies other than naively following the Markov process (MP). When TD, function approximation, and off-policy training are all combined, learned functions exhibit severe instability and divergence, as classically observed by Williams and Baird III [18], Tsitsiklis and Van Roy [15]. This combination is known in the literature as the *deadly triad* [11, pg. 264], and while many contemporary variants of TD are designed to converge despite the instability, the quality of the solution at convergence may be arbitrarily poor.

A common technique to avoid unbounded error is $\ell_2$ *regularization* [14], i.e. penalizing the squared norm of the weights in addition to the TD error. This is generally understood to bound the worst-case error in exchange for biasing the model and potentially increasing the error everywhere else. When used on three common examples of the deadly triad [6, 18, 11, pg.260], regularization appears to

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

mitigate the worst aspects of the divergence in practice. Consequently, it has become an essential assumption made by many RL algorithms [1, 10, 12, 19, 20, 21, 8] and is seen as routine and innocuous.

We argue that this perspective on regularization in off-policy TD is fundamentally mistaken. While regularization is indeed reasonably well-behaved and innocuous in classic fully-supervised contexts, the use of bootstrapping in TD means that even small amounts of model bias induced by regularization can cause divergence. This is an oft-ignored phenomenon in the literature, and so we introduce a series of new counterexamples (summarized in Table 1) to show how regularization can have counterintuitive and destructive effects in TD. We show that vacuous solutions and training instability are *not* solved by the use of regularization; that applying regularization can sometimes induce divergence and increase worst-case error; and that Emphatic TD algorithms–which are the most promising solution to this divergence–can themselves diverge when regularized. We finally also illustrate misbehaving regularization in the context of neural network value function approximation, demonstrating the general pitfalls of regularization possible in RL algorithms. Regularization needs to be treated cautiously in the context of RL, as it behaves differently than in supervised settings.

Our counterexamples demonstrate these core ideas:

**TD learning off-policy can be unstable and/or have unbounded error even when it converges.** Following well-established methods we show there is some off-policy distribution under which TD with linear value function approximation diverges *and* learns a model with unbounded error (even if it were able to converge to the TD fixed point). This concisely demonstrates key features of the training error: the error is small when the distribution is close to on-policy, but the error diverges around specific off-policy distributions. The intuition behind this, explained in Section 3, is that the off-policy[1] TD update involves a projection operation that depends on the sampling distribution and can be arbitrarily far away from the true value. This basic fact has already been established by past work [18, 6], but our example is based upon a particular simple three-state MP, drawn in Figure 1a.

**Regularization cannot always mitigate off-policy training error.** We next introduce regularization into our setting, and show how it changes the relationship between training error and off-policy training. As explained in Section 2, we penalize the $\ell_2$-norm of learned (linear) weights with some coefficient $\eta$; as $\eta$ increases, the learned weights approach zero. However, in **Example 1**, we show that there exists an off-policy distribution such that for any $\eta \geq 0 < \infty$, the regularized TD fixed point attains strictly higher approximation error than the zero solution (i.e., the infinitely regularized point). We call such examples *vacuous*. In other words, *vacuous value functions never do better than guessing zero for all states, for any amount of regularization*.

We further analyze this vacuous example in the context of the algorithm in [21]. In this work, the authors assume the use of regularization to derive bounds on the learned error under off-policy sampling. Although these bounds are technically correct in the case of our counterexample, they are very loose, at about 2000 times the threshold of vacuity. This highlights the challenge of formally relying on regularization to bound model error.

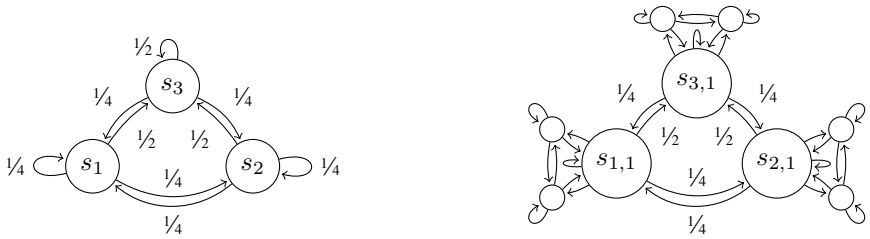
**Small amounts of regularization can cause model divergence or large errors.** There is a general implicit assumption in much ML literature that regularlization monotonically shrinks learned weights. This intuition comes from classic fully-supervised machine learning where it typically holds. But because TD bootstraps value estimates (i.e. learns values using its own output), it is possible for small amounts of bias to be arbitrarily magnified. We dub this phenomenon "small-eta error" and illustrate it in **Example 2**. We relate this to the presence of negative eigenvalues in an intermediate step of the solution and show that, in some settings, the error of the TD solution may be relatively small when applied with no regularization but adding regularization causes the model to have worse error than the zero solution.

One common solution to this problem is to lower-bound $\eta$ to guarantee that reguarlization behaves monotonically. However, we further show that such a lower bound may occur after the point of vacuity: a model that is not vacuous becomes vacuous for any regularization parameter above this lower bound. We also show that it is not always possible to select a single $\eta$ *a priori*, with examples of

[1] We consider a sampling distribution to be *on-policy* if it follows the stationary distribution of the MP; we do not explicitly consider a separate policy in this paper.

| | |
|---|---|
| **Example 1** | There exist off-policy distributions under which TD learns a *vacuous* model (one which–despite any amount of regularization–never does better than guessing zeros). |
| **Example 2** | Small values of the regularization parameter $\eta$ can make TD diverge in models that otherwise converge. This is an unavoidable effect of bootstrapping in TD, and setting a lower-bound to exclude this may render models vacuous. |
| **Example 3** | Emphatic-TD-inspired algorithms are a promising way to reweight samples and mitigate the effects of training off-policy. But if this reweighting is learned using TD, then using regularization can bias the emphasis model and cause the value model itself to diverge. |
| **Example 4** | Training instability and increased error due to the deadly triad also occur when neural networks are used. We construct an empirical example and draw qualitative comparisons. |

Table 1: Summary of theorems.




(a) Three-state MP.

(b) Nine-state MP.

Figure 1: Our three- and nine-state counterexample MPs. We use these to illustrate how TD models can fail despite common mitigating strategies in linear and neural network cases respectively.

mutually-incompatible off-policy distributions where there is no $\eta$ that achieves better than vacuous or nearly-vacuous results at different distributions.

**Emphatic-TD-based algorithms are vulnerable to instability from regularization.** Emphatic-TD [13] fundamentally solves the problem of training off-policy by resampling TD updates so they appear to be on-policy. This technique requires an emphasis model that decides how to scale each TD update, and learning this has been the key challenge preventing widespread adoption of Emphatic-TD. A recent paper [20] proposed learning this emphasis model using "reversed" TD while simultaneously learning the value model using regular TD. The resultant algorithm is called COF-PAC, and employs regularization to ensure that the two TD models eventually converge.

We show that regularization, while necessary, can be harmful for such models in **Example 3**. Specifically, we construct a model that converges to the correct solution without regularization but to an arbitrarily poor solution when regularized. The intuition behind this is that regularizing the emphasis model changes the effective distribution of the TD updates to the value model, which can cause the value model to have arbitrarily large error. We complete the example by showing that regularizing the value function separately does not restore performance.

**Regularization can cause model divergence in neural networks.** So far most analysis of the deadly triad in the literature focuses on the linear case. We extend our example to a nine-state Markov chain (shown in Figure 1b), and show how the previously identified problems persist into the neural network case in **Example 4**. We show two key similarities: first, models trained at certain off-policy distributions may be vacuous. Second, small amounts of regularization counterintuitively *increase* error. This illustrates Example 2 in the NN case.

## 2 Preliminaries and Notation

Consider the $n$-state Markov chain $(\mathcal{S}, P, R, \gamma)$, with state space $\mathcal{S}$, state-dependent reward $R : \mathcal{S} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1]$. $P \in \mathbb{R}^{n \times n}$ is the transition matrix, with $P_{ij}$ encoding the probability of moving from state $i$ to $j$. We wish to estimate the value function $V : \mathcal{S} \to R$, defined

as the expected discounted future reward of being in each state: $V(s) \doteq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| s_0 = s\right]$. A key property is that it follows the Bellman equation:

$$V = R + \gamma P V \tag{1}$$

Using linear function approximation to learn $V$, we assume a matrix of feature-vectors $\Phi \in \mathbb{R}^{n \times k}$ that is fixed, and a vector of parameters $w \in \mathbb{R}^k$ that is learned. The Bellman equation is then:

$$\Phi w = R + \gamma\, P\, \Phi w \tag{2}$$

When $w$ is learned with TD, this equation is only valid if the TD updates are *on-policy* (that is, they are distributed according to the steady-state probability of visiting each state, written as $\pi \in \mathbb{R}^n$). In the general case, where TD updates follow a (possibly) different distribution $\mu \in \mathbb{R}_0^n$, the TD solution is a fixed point of the Bellman operator followed by a projection [6]:

$$\Phi w = \Pi_\mu \left(R + \gamma P \Phi w\right) \tag{3}$$

where the matrix $\Pi_\mu = \Phi(\Phi^\top D \Phi)^{-1} \Phi^\top D$ projects the Bellman backup onto the columnspace of $\Phi$, reweighted by the diagonal matrix $D = \text{diag}(\mu)$. This yields the closed-form solution:

$$w = A^{-1} \vec{b} \tag{4}$$

Where $A = \Phi^\top D(I - \gamma P)\Phi$ and $\vec{b} = \Phi^\top D R$. When this solution is subject to $\ell_2$ regularization, some non-negative $\eta$ is added to ensure the matrix being inverted is positive definite:

$$w^*(\eta) = (A + \eta I)^{-1} \vec{b} \tag{5}$$

As will be important later, we note that as $\eta$ increases it drives $w^*(\eta)$ towards zero.

## 3 Our Counterexamples

When deadly triad conditions are present, TD may learn a value function with arbitrarily large error even if the true value function can be represented with low error. Consider the three-state MP in Figure 1a, which we instantiate with the value function $V = [1,\ 2.2,\ 1.05]^\top$ and discount factor $\gamma = 0.99$. The reward function is computed as $R \leftarrow (I - \gamma P)V$. We choose a basis $\Phi$ with small representation error $\|\Pi_\mu V - V\| \leq \epsilon$:

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & -2.2 \\ 1/2(1.05+\epsilon) & -1/2(1.05+\epsilon) \end{bmatrix} \qquad \text{where } \epsilon > 0 \tag{6}$$

We first consider the unregularized ($\eta = 0$) case, closely following the derivation in [6]. We wish to show there is some sampling distribution $\mu$ such that error in the learned value function is unbounded. To do this, we set $\mu = [0.56(1-p), 0.56p, 0.44]$, where $p \in (0, 1)$. We set $\epsilon = 10^{-4}$ and find $p$ around which $A$ is ill-conditioned by solving $\det(A) = 0$:

$$p = 0.102631 \qquad \vee \qquad p = 0.807255 \tag{7}$$

$A^{-1}$ (and consequently the error) can be made arbitrarily large by selecting $p$ close to these values, which completes the introductory example. Now we look at the behavior of TD under regularization, which is the main contribution of our paper.

### 3.1 Regularization cannot always mitigate off-policy training error.

There is a belief in the literature that regularization is a trade-off between reducing the blow-up of asymptotic errors and accurately learning the value function everywhere else [1, 21]. However, this belief does not accurately capture the nature of regularization: we show that it is possible to learn models that never perform better than always guessing zero despite any amount of regularization. That is, the TD error at all $\eta$ is at least as much as the error as $\eta \to \infty$. We call such models *vacuous*.

**Example 1.** When TD is regularized, there may exist some off-policy distribution at which TD learns a vacuous model. In notation:

$$\|\Phi w^*(\eta) - V\| \geq \lim_{\eta\to\infty} \|\Phi w^*(\eta) - V\| = \|\Phi\vec{0} - V\| = \|V\| \qquad \forall \eta \in \mathbb{R}_0^+ \tag{8}$$

*Details.* We use the same setting as in Section 3. A detailed derivation is provided in Appendix B.2.

We begin by noting that we can easily find the solution $\hat{w}$ that minimizes the least-squares error $\|\Phi\hat{w} - V\|$. If we consider this solution as a vector (as drawn in Figure 2a), we can immediately see that there is an $\ell_2$-ball around $\hat{w}$ corresponding to the set of $w^*(\eta)$ with no more than $\|V\|$ error.

Similarly, we can trace the trajectory that the TD solution $w^*(\eta)$ takes as $\eta$ is increased from 0 to $\infty$. We know that, as $\eta \to \infty$, $w^*(\eta)$ is crushed to zero and so all trajectories must eventually terminate at the origin. When regularized models are not vacuous, the trajectory intersects the non-vacuous-error ball. We see this in trajectory 2, where the error briefly dips below $\|V\|$ in Figure 2b.

Intuitively, a sufficient condition for a solution to be vacuous is that it remains in the half-space that is tangent to and excludes the non-vacuous parameter ball. This is equivalent to finding some distribution $\mu$ such that $\hat{w}^\top w^*(\eta) \leq 0$ for all $\eta$, which we numerically solve to obtain the model in trajectory 1. From Figure 2a we can see the trajectory remains in the half-space, and from Figure 2b we can see that the error is never less than $\|V\|$. Trajectory 1 is a vacuous example. □

We observe that Example 1, because it remains entirely in the halfspace $\hat{w}^\top w^*(\eta) \leq 0$, could easily be generalized to other forms of regularization. We leave this for future work.

***Breaking the Deadly Triad*** **and our counterexample.** In light of our example 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 and show that their method may produce loose bounds on TD solutions, and so doesn't quite break the deadly triad:

$$\|\Phi w^*(\eta) - V\| \leq \frac{1}{\xi}\left(\frac{\sigma_{\max}(\Phi)^2}{\sigma_{\min}(\Phi)^4\sigma_{\min}(D)^{2.5}} \cdot \|V\|\eta + \|\Pi_D V - V\|\right) \tag{9}$$

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

$$\eta > \arg\inf_{\eta} \|\Phi - C_0\| = {}^{0.177}/_{(1-\xi)^2} \tag{10}$$

$$\inf_{\xi} b(\xi, \eta) = 5.20 \times 10^4 \approx 2000 * \|V\| \tag{11}$$

Their method bounds the error in our example by $2000 * \|V\|$, which is tremendously loose. (We analyze a different example in Appendix B.3, showing a still-loose but improved bound of $8 * \|V\|$.) This illustrates the risk of relying on regularization to formally bound model error.

### 3.2 Small amounts of regularization can cause large increases in training error.

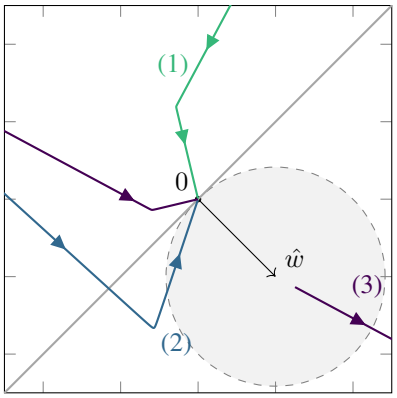
There is a general assumption in the literature that $\ell_2$ regularization monotonically shrinks the learned weights. While this is true in classification, regression, and other non-bootstrapping contexts, this is not true in TD. Because TD bootstraps values, it is possible for model bias to be arbitrarily magnified.

This can be understood in terms of the eigenvalues of the matrix $A$ in Equation 5. By increasing values along the diagonal, $\ell_2$ regularization increases eigenvalues of the matrix $(A + \eta I)$ to ensure it is positive definite. Under off-policy distributions, it is possible for $A$ to have eigenvalues that are negative or zero. This implies that there are $\eta$ for which $\det(A + \eta I) = 0$, and selecting $\eta$ close to these values allows us to achieve arbitrarily high error. We show one such case in Example 2. This is not merely theoretical–we demonstrate this in the neural network case in Section 3.4.

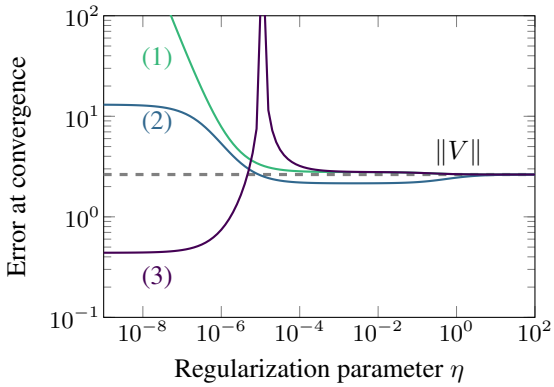
**Example 2.** When TD is regularized, the model may diverge around (typically small) values of $\eta$. Lower-bounding $\eta$, a common mitigation, can make well-behaved models vacuous. It is not always possible to select a single value of $\eta$ that makes models vacuous at different sampling distributions.

*Details.* Using our three-state example, we set $\mu = [0.05, 0.05, 0.9]$ and solve for $\det(A + \eta I) = 0$:

$$0 = \det(A + \eta I) = \eta^2 + 5.45 \times 10^{-2}\eta - 7.47 \times 10^{-3} \quad \Longrightarrow \quad \eta = 0.0634 \tag{12}$$




(a) As $\eta$ increases, $w^*(\eta)$ traces different trajectories at different $\mu$. $\hat{w}$ minimizes the error, and we shade the area with TD error less than $\|V\|$.




(b) We plot the error curves corresponding to the three $w^*(\eta)$ trajectories, along with $\|V\|$. Trajectory 1 is vacuous because the error is at least $\|V\|$ for all $\eta$.

Figure 2: Plotting the trajectory of the parameters on the left and the errors on the right, we show how our counterexample 1 is never better than $\|V\|$ because it remains in halfspace where $\hat{w}^\top w^*(\eta) \leq 0$. For comparison, we show trajectory 2 that is improved by regularization, and 3, which exhibits small-$\eta$ errors. (The trajectories are stretched, so the errors in the two plots are not directly comparable.)

As in the introductory example, the error can be made arbitrarily large by setting $\eta \approx 0.0634$.

This small-$\eta$ divergence effect can appear in several ways, illustrated in Figure 3a. Typically, this appears as one or more points at which TD error diverges before the region at which regularization reduces the model error below $\|V\|$. The first and second plot in Figure 3a show two such cases, where the error increases sharply at two and one points respectively.

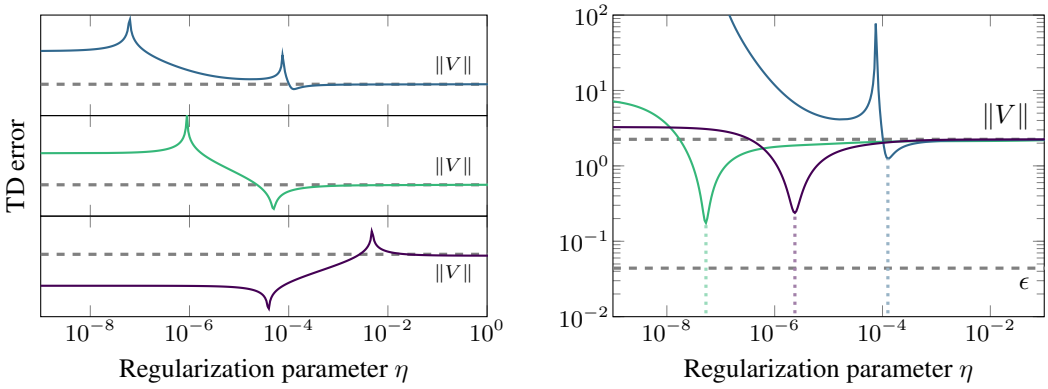
In the literature, it is commonly assumed that $A$ is "nearly" positive definite, where only a few eigenvalues are non-positive, and those are close to zero. This gives rise to the common mitigation of setting a lower-bound $\eta_0$ such that $(A + \eta I)$ is positive definite for $\eta > \eta_0$. This may render an otherwise well-behaved model vacuous. The third plot in Figure 3a illustrates this: the model is not vacuous when unregularized, but is vacuous in the domain $\eta > 10^{-2}$ where divergence is prohibited.

A common practice in the literature is to set $\eta$ before training, without regard for the sampling distribution. This is ill advised, as the value may be under- or over-regularizing depending on the sampling distribution. One such example is illustrated in Figure 3b, where selecting an $\eta$ that minimizes the error for one distribution will lead to vacuous or nearly-vacuous results in the other two. A second example in Figure 2b has no single $\eta$ for which trajectories 2 and 3 are both non-vacuous. This is especially relevant as regularization is commonly used to permit distribution drift during training, as discussed in Section 4. If the training distribution changes while $\eta$ is fixed, then algorithms that can be proven to converge to good solutions under some original distribution may converge to poor solutions as the distribution drifts. □

### 3.3 Emphatic approaches and our counterexample

Emphatic-TD eliminates instability from off-policy sampling by reweighting incoming data (via an importance function) so it appears to be on-policy. There is considerable interest in making this more practical, especially by learning the importance and value models simultaneously. A leading example of this work is COF-PAC [20], which uses $\ell_2$-regularized versions of GTD2 [12] to learn both the value and emphasis models. The authors rely on regularization, particularly because the target policy changes during learning. This makes COF-PAC vulnerable to regularization-caused error. We illustrate this with Example 3 in which COF-PAC learns correctly when unregularized, but has large error when regularized. (Mathematical details are deferred to Appendix B.5.)

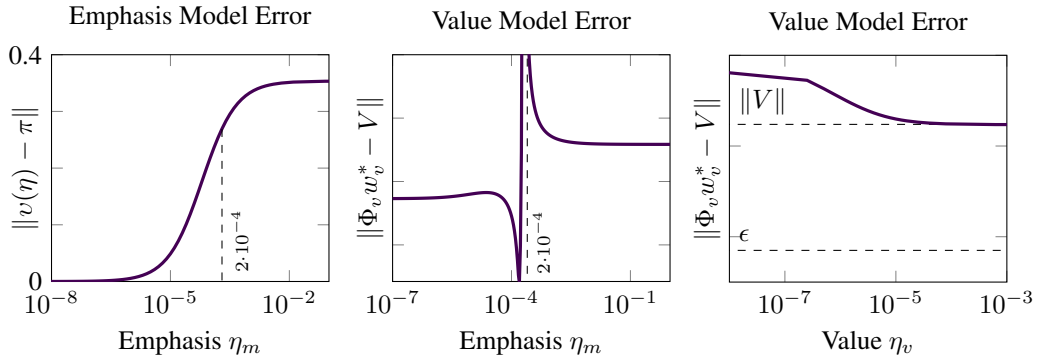
**Example 3.** COF-PAC may learn the value function with low error when unregularized, but with arbitrarily high error when regularized.
*Details.* Conceptually, COF-PAC maintains two separate models that are each updated by TD: the emphasis and the value models. This emphasis model is used to reweight TD updates to the value




(a) Different MPs at off-policy distributions selected to show small-$\eta$ error. The error may increase at multiple $\eta$, and may even occur *after* the optimal $\eta$.

(b) Three off-policy distributions with mutually incompatible $\eta$. There is no $\eta$ at which all models are not vacuous or nearly vacuous.

Figure 3: We plot TD error against $\eta$ to show small-$\eta$ errors (left) and mututally-incompatible $\eta$ (right). We also plot the error at the limit of vacuity $\|V\|$ and the representation error $\epsilon$.




(a) $\eta_m$ distorts the emphasis model.

(b) $\eta_m$ distorts value.

(c) $\eta_v$ can't fix this.

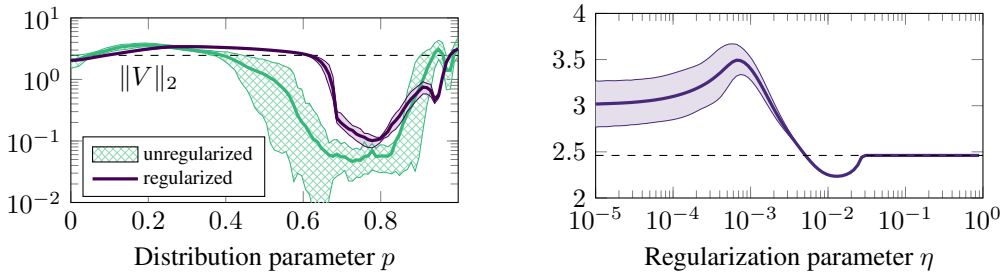
Figure 4: Regularization on the emphasis model ($\eta_m$) distorts the effective distribution (Figure 4a). Specific values of $\eta_m$ induce the value function to diverge (Figure 4b). The resultant value function is vacuous (Figure 4c). Under COF-PAC, regularization can greatly increase model error.

function so they appear to come from the on-policy distribution. Our strategy is to first show how regularization biases the emphasis model, and then how this bias causes the value model to diverge. We begin with our three-state MP, noting its on-policy distribution is $\pi = [.25\ .25,\ .5]$. We wish to learn the values using COF-PAC while sampling off-policy at $\mu = [.2\ .2\ .6]$.

Now we introduce a key conceptual tool: $\upsilon(\eta_m)$, which is the effective distribution seen by the TD-updates, influenced by the emphasis regularization parameter $\eta_m$. Unregularized, COF-PAC is able to resample off-policy updates to the on-policy distribution: $\upsilon(0) \equiv \pi$. If the model is regularized, then the effective distribution moves away from $\pi$. Figure 4a illustrates the distance between $\upsilon(\eta_m)$ and $\pi$ as the regularlization parameter increases.

We can use the effective distribution to compute the error in the value model. Plotting the relationship between the value function error and $\eta_m$ in Figure 4b, we can see the value function has asymptotic error around $\eta_m = 2 \times 10^{-4}$. This shows how COF-PAC may diverge with specific regularization.

COF-PAC also allows for the value function to be separately regularized with parameter $\eta_v$. We show the effect of this in Figure 4c, where the value function never does much better than $\|V\|$ making it (nearly) vacuous. We can conclude that regularizing the emphasis model may cause the value model to diverge, and this cannot be fixed by regularizing the value function separately. □




(a) Mean and $10^{\text{th}}$–$90^{\text{th}}$ percentile errors of 100 NN value models trained to convergence.

(b) The relationship between error and $\eta$ at the off-policy distribution $p = 0.31$.

Figure 5: We illustrate how regularization interacts with NN value functions, showing that the problems identified in this paper persist in the NN case.

COF-PAC makes the strong assumption that Kolter's relaxed-contraction condition [6, eqn. 10] holds in both the emphasis and value models [20, asm. 4]. We discuss this in Appendix B.5.1.

### 3.4 Applied to multi-layer networks

We use a 9-state variant of our example to study the deadly triad in multi-layer neural networks (NNs). A deterministic observation function is chosen so we can control the degree of function approximation. We train a simple two-layer neural network with 3 neurons in the hidden layer. The value function is assigned pseudo-randomly in range $[-1, 1]$. (See Appendix C for details.)

**Example 4.** Vacuous models and small-$\eta$ error also occur in neural network conditions.

*Details.* We train 100 models using simple semi-gradient TD updates under a fixed learning rate. We plot the mean and the $10^{\text{th}}$–$90^{\text{th}}$ percentile range in Figure 5a, with and without regularization. TD is known to exhibit high variance, and regularization is the traditional remedy for that. We corroborate this by noting that the performance of the unregularized model varies widely, but regularization leads to similar performance across initializations at the cost of increased error.

First, we show that vacuous models may exist in the neural network case. In Figure 5a, note how there are some off-policy distributions under which both the regularized and unregularized models perform worse than the threshold of vacuity. We can numerically verify that vacuous models exist. Second, we show the small-$\eta$ error problem in the neural network case in Figure 5b, where we plot the TD error against $\eta$ at a fixed off-policy distribution. We observe that around $\eta \approx 10^{-3}$ the TD Error unexpectedly *increases* before decreasing, which clearly illustrates this phenomenon. □

These qualitative links show a clear connection between the neural network case and the linear case, and highlights the importance of correctly handling off-policy sampling.

## 4 Related Work

Three examples of the deadly triad are common in the literature: the classic Tsitsiklis and Van Roy $(w, 2w)$ example [11, p. 260], Kolter's example [6], and Baird's counterexample which shows how training instability can exist despite overparameterization [18].

$\ell_2$ regularization is common when proving that an algorithm converges under a changing sampling policy. This is seen in GTD (analyzed in [19]), GTD2 [12], RO-TD [10], and COF-PAC [20]. This assumption may also be used to ensure convergence when training with a target network [21]. Despite the prevalence of regularization, the induced bias from using it is not well studied. It is often dismissed as a mere technical assumption, as in [1]. We contradict that: using regularization for convergence proofs may induce catastrophic bias. By showing concrete examples, this work hopes to inspire further investigation into regularization-induced bias in the same vein as [19].

**Alternatives to regularization and TD** We focus on $\ell_2$ regularization in this paper, which penalizes the $\ell_2$-norm of the learned weights; it is also possible to use $\ell_1$ regularization with a proximal

operator/saddle point formulation as in [10], or any convex regularization term under a fixed target policy [19]. Instead of directly regularizing the weights, COP-TD uses a discounted update [4]. DisCor [7] propagates bounds on Q-value estimates to quickly converge TD learning in the face of large bootstrapping error; it is not clear if DisCor can overcome off-policy sampling. A separate primal-dual saddle point method has also been adapted to $\ell_2$ regularization [2] and is known to converge under deadly triad conditions, and recent work [17] has derived error bounds with improved scaling properties in the linear setting, offering a promising line of research.

Emphatic-TD [13] fixes the fundamental problem in off-policy TD by reweighting updates so they appear on-policy. The core idea underlying these techniques is to estimate the "followon trace" for each state, the (weighted, $\lambda$- and $\gamma-$discounted) probability mass of all states whose value estimates it influences. This trace is then used to estimate the emphasis, which is the reweighting factor for each update. While this family of methods is provably optimal in expectation, it is subject to tremendous variance in theory and practice, especially when the importance is estimated using Monte-Carlo sampling.[2] In practice, these methods learn the follow-on trace using TD [5, 20] or similar [16], which makes them vulnerable to bias induced by the use of regularization.

## 5 Conclusion

There is a tremendous focus in the RL literature on proving convergence of novel algorithms, but not on the error at convergence. Papers like [21] are laudable because they provide error bounds; even if the current bounds are loose, future work will no doubt tighten them. In this work, we show that the popular technique of $\ell_2$ regularization does not always prevent singularities and could even introduce catastrophic divergence. We show this with a new counterexample that elegantly illustrates the problems with learning off-policy and how it persists into the NN case.

Even though regularization can catastrophically fail in the ways we illustrate, it remains a reasonable method that may offer a fair tradeoff – as long as we are careful to check that we are not running afoul of the failure modes we explain in the paper. It may be possible to design an adaptive regularization scheme that can avoid these pathologies. For now, testing the model performance over a range of regularization parameters (spanning several orders of magnitude) is the best option we have to detect such pathological behavior.

Emphatic-TD is perhaps the most promising area of research for mitigating off-policy TD-learning. The key problem preventing its widespread adoption is the difficulty in estimating the emphasis function, but future work in this area may be able to overcome this. Our example shows the risk of relying on regularization in practical implementations of such methods. It is absolutely critical that Emphatic algorithms correctly manage regularization to avoid the risks that we highlight in this paper.

[2] Sutton and Barto's textbook [11] says about Emphatic-TD that "it is nigh impossible to get consistent results in computational experiments." (when applied to Baird's example).

## References


[1] R. B. Diddigi, C. Kamanchi, and S. Bhatnagar. A convergent off-policy temporal difference algorithm. *arXiv preprint arXiv:1911.05697*, 2019.

[2] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[3] W. Fedus, P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, and W. Dabney. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*, pages 3061–3071. PMLR, 2020.

[4] C. Gelada and M. G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

[5] R. Jiang, S. Zhang, V. Chelu, A. White, and H. van Hasselt. Learning expected emphatic traces for deep rl. *arXiv preprint arXiv:2107.05405*, 2021.

[6] J. Kolter. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24:2169–2177, 2011.

[7] A. Kumar, A. Gupta, and S. Levine. Discor: Corrective feedback in reinforcement learning via distribution correction. *arXiv preprint arXiv:2003.07305*, 2020.

[8] A. Kumar, R. Agarwal, T. Ma, A. Courville, G. Tucker, and S. Levine. DR3: Value-based deep reinforcement learning requires explicit regularization. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=POvMvLi91f`.

[9] S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020. URL `https://arxiv.org/abs/2005.01643`.

[10] S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, and J. Liu. Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *arXiv preprint arXiv:1405.6757*, 2014.

[11] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, second edition edition, 2020.

[12] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000, 2009.

[13] R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17:2603–2631, 2016.

[14] A. N. Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

[15] J. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *Rep. LIDS-P-2322). Lab. Inf. Decis. Syst. Massachusetts Inst. Technol. Tech. Rep*, 1996.

[16] H. van Hasselt, S. Madjiheurem, M. Hessel, D. Silver, A. Barreto, and D. Borsa. Expected eligibility traces. *arXiv preprint arXiv:2007.01839*, 2021.

[17] A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22384–22429. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/wagenmaker22a.html`.

[18] R. J. Williams and L. C. Baird III. Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems. Technical report, Citeseer, 1993.

[19] H. Yu. On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*, 2017.

[20] S. Zhang, B. Liu, H. Yao, and S. Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pages 11204–11213. PMLR, 2020.

[21] S. Zhang, H. Yao, and S. Whiteson. Breaking the deadly triad with a target network. *CoRR*, abs/2101.08862, 2021. URL `https://arxiv.org/abs/2101.08862`.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We clearly specify that this paper contains counterexamples illustrating a possible outcome.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The total compute time to replicate all results in this paper is less than one CPU-hour.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] We only use academic writing, not code with a separate license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# A Notation

| Symbol | | Description |
|---|---|---|
| $n$ | $\in \mathbb{Z}^+$ | Number of states. |
| $m$ | $\in \mathbb{Z}^+$ | Number of features. |
| $\pi$ | $\in \mathbb{R}^n$ | on-policy distribution. |
| $\mu$ | $\in \mathbb{R}^n$ | sampling distribution, may be on- or off-policy. |
| $\upsilon$ | $: \mathbb{R}^+ \to \mathbb{R}^n$ | apparent distribution induced by $\eta$-regularizing the emphatic correction of off-policy $\mu$ to on-policy $\pi$ |
| $\eta$ | $\in \mathbb{R}_0^+$ | $\ell_2$ regularization parameter |
| $\eta_m$ | $\in \mathbb{R}_0^+$ | $\ell_2$ regularization parameter for emphasis model in COF-PAC (the Emphatic algorithm we analyze) |
| $\eta_v$ | $\in \mathbb{R}_0^+$ | $\ell_2$ 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. |
| $\Phi$ | $\in \mathbb{R}^{[n \times m]}$ | 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 $\Phi$. |
| $w^*(\eta)$ | $\in \mathbb{R}^{[m \times 1]}$ | 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 |
| $\|x\|$ | $\in \mathbb{R}_0^+$ | $\ell_2$-norm of vector or matrix $x$, equal to $\sqrt{x^\top x}$ |
| $\|x\|_D$ | $\in \mathbb{R}_0^+$ | $\ell_2$-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 $\ell_2$ regularization, our linear model fails with asymptotic error. As this penalizes the $\ell_2$-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 $\eta$, 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 $\Phi$ 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 $\eta$ 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 $\eta$ 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 $\eta$. □

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.

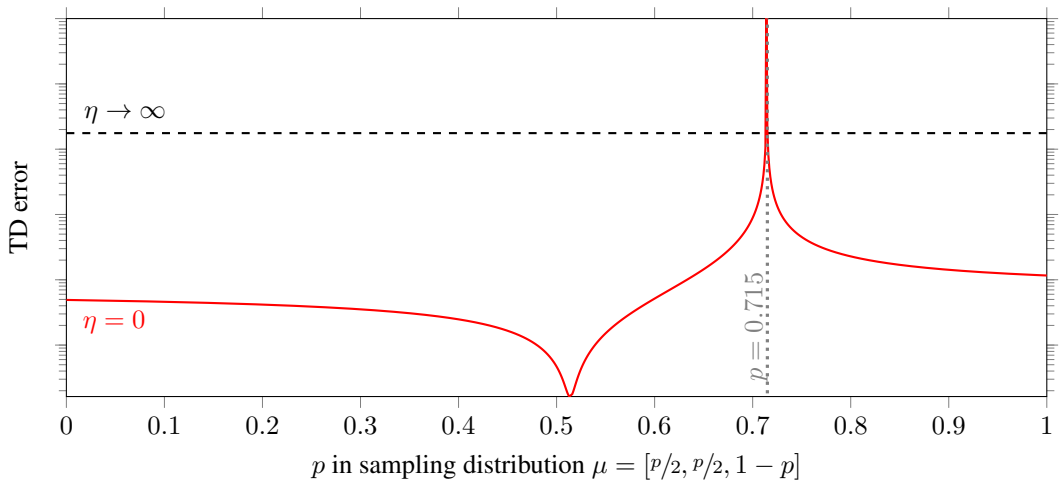



Figure 6: We plot TD error against $p$ for our three-state MP with $\epsilon = 10^{-4}$. This shape is similar to that in [6]. There is a minima close to $\pi$ ($p \approx 0.5$), and an asymptote at the singularity ($p \approx 0.715$). At different levels of regularization the error function moves between 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 $\eta$, and therefore also $b$:

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

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

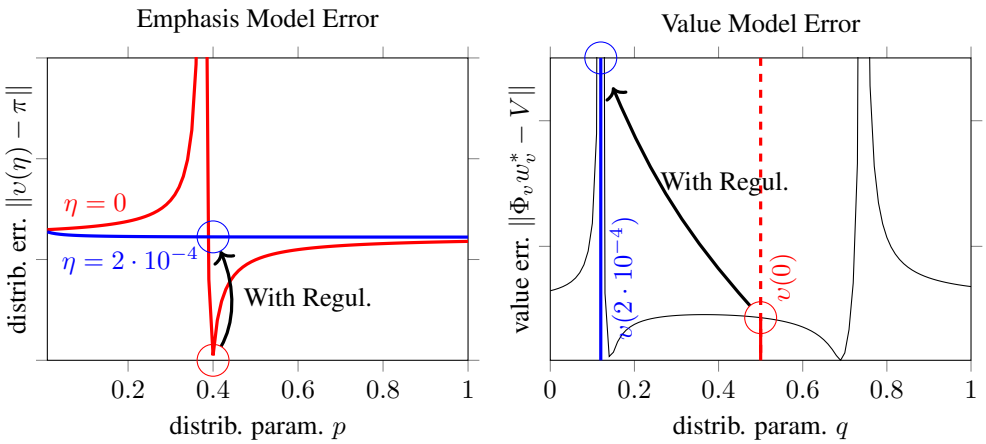
Under our example, their method bounds the error at no more than $7.86 * \|V\|$, which is a very loose bound that permits vacuous solutions. This illustrates the risk of trying to regularize away singularities, particularly in theoretical work.

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 $\eta$.




(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 $\eta$ close to $4.83 \times 10^{-3}$. 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 $\Phi_m$ and $\Phi_v$ 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 $\Phi_m$ 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 $\upsilon$, 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 $\Phi_m$. This allows $\eta$ 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 $\upsilon$ 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 $\mu$ and value model at $\upsilon$. Kolter's condition selects a convex subset of distributions under which one-step transition followed by projection onto $\Phi$ 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.

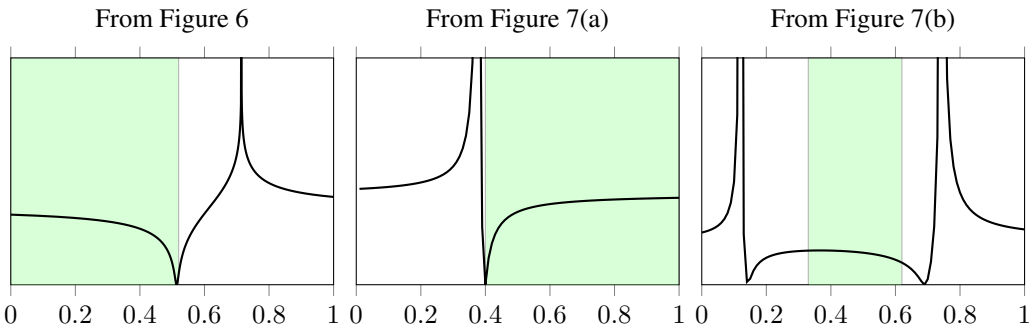



Figure 8: Kolter's non-expansion condition holds in the shaded region of each graph.

# C Applied to multi-layer networks

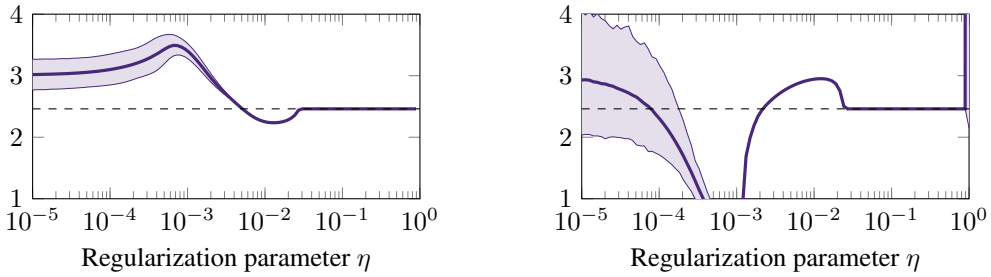
We also use a variant of our example to study how the deadly triad appears in multi-layer networks. As illustrated in Figure 11, we replace each self-loop with two additional states, forming a clique with the original state. The resultant MP has $n = 9$ states; we define a deterministic obervation function $o : \mathcal{S} \to \mathbb{B}^6$. 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 $\eta$ 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-$\eta$ divergence.




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

(b) $p = 0.95$

Figure 9: The relationship between error and $\eta$ 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-$\eta$ 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 $\ell_2$ 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.

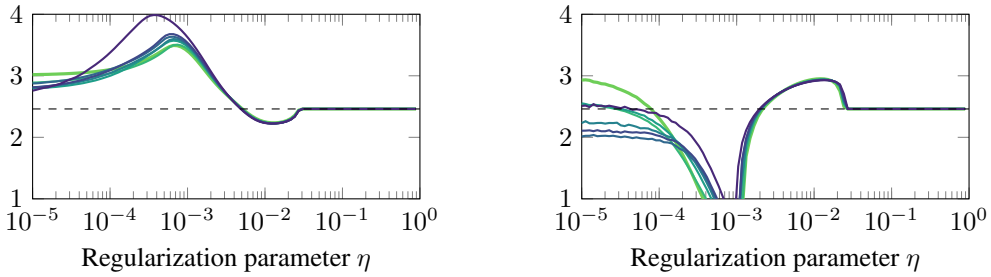
## D Markov Processes

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 $\eta$ and error with different amount of model parameterization (with 3, 5, 7, 9, 11, and 13 nodes in the hidden layer, corresponding to darkening colors.)

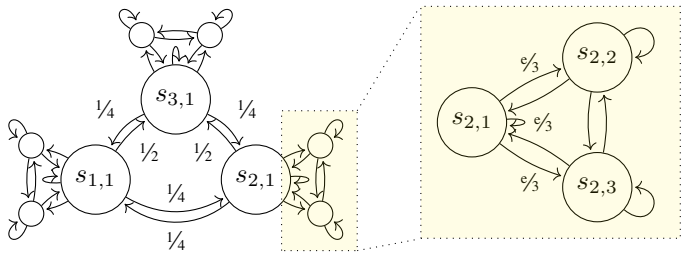



Figure 11: Our three-state counter-example MP is extended to nine states to illustrate how the deadly triad problem could manifest in multi-layer neural networks. The self-loop in the original example is replaced with a clique with uniform transitions except as labelled with the original edge weight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

### 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. $p \to 0.02, q \to 0.2$