Rectifying Regression in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Keywords: Value-based methods, Regression, Loss functions.

Summary

This paper investigates the impact of the loss function in value-based methods for reinforcement learning through an analysis of underlying prediction objectives. We theoretically show that mean absolute error is a better prediction objective than the traditional mean squared error for controlling the learned policy's suboptimality gap. Furthermore, we present results that different loss functions are better aligned with these different regression objectives: binary and categorical cross-entropy losses with the mean absolute error and squared loss with the mean squared error. We then provide empirical evidence that algorithms minimizing these cross-entropy losses can outperform those based on the squared loss in linear reinforcement learning.

Contribution(s)

1. We demonstrate certain cross entropy losses can accelerate convergence under certain structural assumptions, supported by negative results for the purely mean-focused squared loss.
   Context: We build upon a recent line of theoretical (Foster & Krishnamurthy, 2021; Ayoub et al., 2024; Wang et al., 2024) and empirical (Bellemare et al., 2017; Dabney et al., 2018; Farebrother et al., 2024) research showing that value learning with certain loss functions can yield faster convergence rates under specific structural assumptions, such as the optimal policy achieving the maximum possible value or having low variance returns. We complement these findings by providing lower bounds that link these convergence rates to the chosen regression objective—in this case mean absolute error and mean squared error.
2. We provide empirical results showing that value-based methods using log-loss (and its reparameterized multi-class variant) can outperform squared-loss methods in a linear batch reinforcement learning setting (inverted pendulum with Fourier features).
   Context: The work of Lyle et al. (2019) suggest that, in linear reinforcement learning, cross-entropy losses (e.g., binary or categorical) perform on par with squared loss and that their advantages appear primarily in deep reinforcement learning settings. However, our theoretical and empirical findings suggest a more subtle situation: in linear reinforcement learning, cross-entropy losses can outperform the canonical squared loss. Our experiments are limited to a single environment (inverted pendulum) with Fourier features.

Rectifying Regression in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

This paper investigates the impact of the loss function in value-based methods for reinforcement learning through an analysis of underlying prediction objectives. We theoretically show that mean absolute error is a better prediction objective than the traditional mean squared error for controlling the learned policy's suboptimality gap. Furthermore, we present results that different loss functions are better aligned with these different regression objectives: binary and categorical cross-entropy losses with the mean absolute error and squared loss with the mean squared error. We then provide empirical evidence that algorithms minimizing these cross-entropy losses can outperform those based on the squared loss in linear reinforcement learning.

1 Introduction

Value-based methods are ubiquitous in reinforcement learning (RL) (Sutton & Barto, 2018) and contextual bandits (Lattimore & Szepesvári, 2020), where the goal is to predict rewards and then choose actions that maximize expected returns. The "natural objective function" (Chapter 9.2 of (Sutton & Barto, 2018)) in these settings is the *mean squared error* (MSE). Traditionally in RL, value-based algorithms that leverage function approximation aim to minimize the MSE between a learned value function and observed returns, as popularized by algorithms such as Q-learning and its extension to function approximation (Ernst et al., 2005; Riedmiller, 2005; Mnih et al., 2015). Despite its success in many practical scenarios, recent theoretical (Foster & Krishnamurthy, 2021; Ayoub et al., 2024) and empirical (Farebrother et al., 2024) findings highlight that minimizing *MSE* can yield worse decision-making performance (in terms of expected returns) than alternative empirical losses. In particular, minimizing alternative losses can achieve faster learning when the optimal value $v^\star$ is close to the maximum possible value.

A promising alternative is the *log-loss* (cross-entropy loss), which we argue aligns more closely with the mean absolute error (MAE)—a tighter surrogate for decision quality in many applications. Indeed, it has been shown that, under certain problem structures, controlling the MAE can lead to faster convergence (Foster & Krishnamurthy, 2021) and a smaller suboptimality gap (Ayoub et al., 2024) than controlling MSE. Building on these insights, we analyze *value-based methods* that employ log-loss and demonstrate their advantages over those using squared loss. Our exposition initially restricts attention to the offline contextual bandit (or reward-sensitive classification (Elkan, 2001)) setting, where one aims to learn a policy from a fixed dataset of context–reward pairs. However, the core ideas and proofs naturally extend to more general RL problems with function approximation (Antos et al., 2007; Chen & Jiang, 2019; Ayoub et al., 2024).

We propose a *reparameterized* version of the categorical cross-entropy loss (ℓ_{cat}), which learns a multi-category distribution over possible outcomes while still accurately recovering the mean of a bounded random variable as was similarly done by Lyle et al. (2019). This addresses the irreducible bias found in purely bin-based approaches. Our analysis highlights that modeling distributional properties—not just the mean—can accelerate learning when the distribution has low variance or other favorable characteristics.

In summary, our work contributes new insights into why certain losses such as log-loss are better suited than squared loss in certain decision-making settings, highlights methods to reparameterize cross entropy losses that can be utilized simultaneously for classification and regression, and highlights broader implications for distributional RL. Although we focus on contextual bandits for clarity, the same techniques extend naturally to RL, offering an explanation as to why certain loss function can improve value-based algorithms.

2 Definitions

In this section, we formally define regression and classification problems. We then introduce a solution to the classification problem that leverages a regression oracle, highlighting the connection between these two paradigms. By distinguishing problems from their solutions, we establish a formal framework for understanding how an appropriately chosen regression objective can guide the development of more effective decision-making algorithms.

2.1 Problem: Regression

Consider a supervised learning problem where the set of contexts is denoted by $\mathcal{X}$. The learner is provided with a dataset $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, consisting of n independently and identically distributed (i.i.d.) context-label pairs, sampled from a distribution $\mathcal{P}^{\otimes n}$. Each sample (X_i, Y_i) satisfies $X_i \in \mathcal{X}$ and $Y_i \in [0,1]$, where $\mathcal{P}$ is an element of $\mathcal{M}_1(\mathcal{X} \times [0,1])$, the space of probability measures over $\mathcal{X} \times [0,1]$. We assume $\mathcal{M}_1(\cdot)$ is defined over an appropriately equipped σ-algebra.

Define the conditional expectation of the label given the context as

$$f^\star(x) = \mathbb{E}[Y_1 | X_1 = x], \quad \forall x \in \mathcal{X}.$$

A score function f maps contexts to predicted labels in $[0,1]$, formally defined as $f : \mathcal{X} \to [0,1]$.

Throughout this paper, we assume that we are given a *realizable class* (Assumption 2.1) of score functions $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$. Let $\mathcal{P}_X \in \mathcal{M}_1(\mathcal{X})$ denote the marginal distribution over contexts.

Assumption 2.1 (Realizability). $f^\star \in \mathcal{F}$.

For $p \geq 1$, the prediction error of a function $f \in \mathcal{F}$ is defined as

$$\mathsf{VE}_p(f) = \mathbb{E}\big|f(X) - f^\star(X)\big|^p = \int |f(x) - f^\star(x)|^p \, P_X(dx).$$

Thus, $\mathsf{VE}_p(f)$ quantifies the deviation of f from the optimal predictor $f^\star$. The objective in regression is to learn a function $\hat{f}$ that minimizes the prediction error $\mathsf{VE}_p(\hat{f})$. Special cases of this error metric include the mean squared error when $p = 2$

$$\mathsf{VE}_2(f) = \mathbb{E}[(f(X) - f^\star(X))^2],$$

and the mean absolute error when $p = 1$

$$\mathsf{VE}_1(f) = \mathbb{E}\big|f(X) - f^\star(X)\big|.$$

2.2 Problem: Classification

Consider a (reward-sensitive) classification problem, where the set of contexts is denoted by $\mathcal{S}$, and the set of actions is a finite set $\mathcal{A}$ with cardinality $A = |\mathcal{A}| < \infty$. The learner is provided with a dataset $D_n = ((S_1, R_1), \ldots, (S_n, R_n))$, consisting of n i.i.d. context-reward vector pairs, sampled from $\mathcal{P}^{\otimes n}$. Each sample satisfies $S_i \sim \mathcal{S}$ and $R_i \in [0,1]^{\mathcal{A}}$, where $\mathcal{P} \in \mathcal{M}_1(\mathcal{S} \times [0,1]^{\mathcal{A}})$ represents the joint distribution over contexts and reward vectors.

Define the expected reward function as

$$r(s,a) = \mathbb{E}[R_1(a) \mid S_1 = s], \quad \forall s \in \mathcal{S},\, a \in \mathcal{A}.$$

A policy (classifier) π maps contexts to probability distributions over actions, formally defined as $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$. We use $\pi(a \mid s)$ to denote the probability assigned by π to action $a \in \mathcal{A}$ when the context is $s \in \mathcal{S}$. Let $\mathcal{P}_S \in \mathcal{M}_1(\mathcal{S})$ denote the marginal distribution over contexts. The expected return of following policy π is

$$v^\pi = \int \sum_{a \in \mathcal{A}} \pi(a \mid s)\, r(s,a)\, P_S(ds).$$

An optimal policy $\pi^\star$ is one that maximizes the expected return across all policies:

$$v^{\pi^\star} = v^\star = \max_\pi v^\pi.$$

Define the *suboptimality gap* (regret) of using policy π instead of the optimal policy as

$$\text{Sub-opt}(\pi) = v^\star - v^\pi.$$

The objective in classification is to learn a policy $\hat{\pi}$ that maximizes the expected return $v^{\hat{\pi}} = \hat{v}$, or equivalently, minimizes the suboptimality gap $\text{Sub-opt}(\hat{\pi})$.

2.3 Solution: Value-Based Methods for Classification

In statistical learning theory (Vapnik, 2013), the objective of classification is often to learn a policy $\hat{\pi}$ that minimizes the suboptimality gap $\text{Sub-opt}(\hat{\pi})$. Given a class of policies $\Pi \subset \mathcal{M}_1(\mathcal{A})^{\mathcal{S}}$, a natural approach is the principle of empirical risk minimization on the dataset D_n (Vapnik, 1991):

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \sum_{i=1}^n \pi(\cdot \mid S_i)^\top R_i\,.$$

However, directly optimizing this empirical risk is generally computationally intractable (NP hard), even for relatively simple policy classes (Ben-David et al., 2003; Feldman et al., 2012). This motivates the use of *value-based methods* for classification, which leverage a class of candidate value functions to reformulate the problem as a regression task—often solvable via gradient descent.

Formally, given a dataset $D_n = ((S_1, R_1), \ldots, (S_n, R_n))$, a realizable class (Assumption 2.1) of candidate value functions $\mathcal{F} \subseteq [0,1]^{\mathcal{S}\times\mathcal{A}}$, and a loss function ℓ, we solve the following regression problem:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{a \in \mathcal{A}} \ell\big(f(S_i, a), R_i(a)\big).$$

The learned function $\hat{f}$ is then used to define the greedy policy:

$$\hat{\pi}(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \hat{f}(s,a).$$

We refer to any method that minimizes a regression loss and then selects actions greedily with respect to the minimizer as a value-based method.

2.4 The Choice of the Loss Function

In the context of reinforcement learning, value-based methods commonly minimize the squared loss (Sutton & Barto, 2018; Szepesvári, 2022):

$$\ell_{\text{sq}}(x,y) = (x-y)^2, \quad \text{where } x, y \in [0,1].$$

This contrasts with traditional classification tasks, where the cross-entropy loss is mostly used. We remark that while “cross-entropy loss” is typically associated with the negative log-likelihood under Bernoulli or categorical models, the squared loss is the cross entropy loss that arises under a univariate Gaussian model. We introduce the terminology *log-loss* and *cat-loss* to denote the cross-entropy losses for Bernoulli and categorical distributions, respectively.

To develop intuition, we first focus on the log-loss, as it represents the simplest special case of the more general cat-loss.

We define the log-loss as

$$\ell_{\log}(x, y) = y \log \frac{1}{x} + (1 - y) \log \frac{1}{1 - x}, \quad \text{where } x, y \in [0, 1] ,$$

with the convention $0 \log \infty = \lim_{u \to 0} u \log \frac{1}{u} = 0$. The following proposition establishes that the minimizer of the log-loss is the population mean.

Proposition 2.2. *Let Y be a bounded random variable with $Y \in [0, 1]$ and mean $\mathbb{E}[Y] = \mu$. Then, for any $x \in [0, 1]$, the expected log-loss satisfies:*

$$\mathbb{E}[\ell_{\log}(x, Y)] \geq \mathbb{E}[\ell_{\log}(\mu, Y)].$$

Proof. Recall the binary Kullback–Leibler (KL) divergence

$$\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

Then, observe that

$$\begin{aligned} \mathbb{E}[\ell_{\log}(x, Y)] - \mathbb{E}[\ell_{\log}(\mu, Y)] &= \mu \log \frac{1}{x} + (1 - \mu) \log \frac{1}{1 - x} - \mu \log \frac{1}{\mu} - (1 - \mu) \log \frac{1}{1 - \mu} \\ &= \mathrm{kl}(\mu, x). \end{aligned}$$

Since the KL divergence is always nonnegative, with equality if and only if $x = \mu$, the result follows. □

This proposition implies that, given an infinite number of observations of a bounded random variable $Y \in [0, 1]$, minimizing the log-loss *recovers the mean* of Y. In later sections, we will extend this insight to the cat-loss—the canonical loss for multi-class classification.

3 Mean Absolute Error as a More Natural Objective for Decision Making

We now present the central insight of this paper: *the mean absolute error (MAE), rather than the mean squared error (MSE), is a more suitable regression objective for decision-making problems* such as reinforcement learning and classification. Formally, let $\hat{\pi}$ denote the greedy policy with respect to some learned function $\hat{f} \in [0, 1]^{\mathcal{S} \times \mathcal{A}}$. For a score function $f \in [0, 1]^{\mathcal{S} \times \mathcal{A}}$ and a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$, define

$$f(s, \pi) = f\big(s, \pi(s)\big).$$

We now bound the suboptimality gap $\text{Sub-opt}(\hat{\pi})$. Analyses of value-based methods for classification typically reduce a classification objective (i.e., $\text{Sub-opt}(\hat{\pi})$) to a regression objective (i.e., $\overline{\mathrm{VE}}_p(\hat{f})$); see, for instance, Antos et al. (2007); Chen & Jiang (2019); Ayoub et al. (2024) for batch RL and Ayoub et al. (2020); Jin et al. (2021; 2023) for online RL. In doing so, one obtains both a mean absolute error (MAE) term and a root mean squared error (rMSE) term. Indeed, observe that

$$
\begin{aligned}
\text{Sub-opt}(\hat{\pi}) &= \int r(s,\pi^\star) - r(s,\hat{\pi})\, P_S(ds) \\
&= \int r(s,\pi^\star) - \hat{f}(s,\hat{\pi}) + \hat{f}(s,\hat{\pi}) - r(s,\hat{\pi})\, P_S(ds) \\
&\le \int r(s,\pi^\star) - \hat{f}(s,\pi^\star) + \hat{f}(s,\hat{\pi}) - r(s,\hat{\pi})\, P_S(ds) \qquad (1) \\
&\le \int \big| r(s,\pi^\star) - \hat{f}(s,\pi^\star) \big|\, P_S(ds) + \int \big| \hat{f}(s,\hat{\pi}) - r(s,\hat{\pi}) \big|\, P_S(ds) \qquad \text{(MAE)} \\
&\le \sqrt{\int \big(r(s,\pi^\star) - \hat{f}(s,\pi^\star) \big)^2 P_S(ds)} + \sqrt{\int \big(r(s,\hat{\pi}) - \hat{f}(s,\hat{\pi}) \big)^2 P_S(ds)}\,. \qquad \text{(rMSE)}
\end{aligned}
$$

The first inequality in (1) uses the fact that $\hat{\pi}$ is greedy with respect to $\hat{f}$. The last inequality applies Jensen's inequality. Notice that the key reduction from a policy-space comparison ($\hat{\pi}$ vs. $\pi^\star$) to a function-space comparison ($\hat{f}$ vs. r) occurs in (1).

Motivation for MAE vs. rMSE. As shown, both MAE and rMSE naturally arise in bounding $\text{Sub-opt}(\hat{\pi})$. However that MAE is a *tighter* approximation to $\text{Sub-opt}(\hat{\pi})$ than rMSE, since

$$
\int |f - g| \le \sqrt{\int (f-g)^2}\,.
$$

In practical settings, there are cases where $\int |r - \hat{f}|$ is significantly smaller than $\sqrt{\int (r - \hat{f})^2}$, implying that algorithms designed to control rMSE (such as squared-loss minimization) can incur larger suboptimality than algorithms targeted toward controlling MAE (such as log-loss minimization). Since the ultimate goal in decision making is to select good actions, it is natural to adopt a regression metric (and loss) that is more closely aligned with suboptimality. We now present a set of results that confirm our intuition.

3.1 Positive Results

We highlight that minimizing the log-loss (i.e., $\ell_{\log}$) yields bounds that scale with $(1 - v^\star)$, which can be small in problems where the optimal policy achieves a reliable goal or accumulates near-maximal return. The following lemma adapts the result of Foster & Krishnamurthy (2021) to the rewards-based setting.

Lemma 3.1. *Assume $r \in \mathcal{F}$ and define*

$$
\hat{f}_{\log} \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \ell_{\log}\big(f(S_i,a),\, R_i(a)\big)\,.
$$

Let $\hat{\pi}_{\log}$ be the greedy policy w.r.t. $\hat{f}_{\log}$. Then with probability $1-\delta$,

$$
\text{Sub-opt}(\hat{\pi}_{\log}) \le 16\sqrt{\frac{2\,(1-v^\star)\,A\,\log\big(|\mathcal{F}|/\delta\big)}{n}} + \frac{136\,A\,\log\big(|\mathcal{F}|/\delta\big)}{n}\,,
$$

and

$$
\begin{aligned}
&\int \big| r\big(s,\pi^\star\big) - \hat{f}\big(s,\pi^\star\big) \big| P_S(ds) + \int \big| \hat{f}\big(s,\hat{\pi}\big) - r\big(s,\hat{\pi}\big) \big| P_S(ds) \\
&\qquad \le 16\sqrt{\frac{2\,(1-v^\star)\,A\,\log\big(|\mathcal{F}|/\delta\big)}{n}} + \frac{136\,A\,\log\big(|\mathcal{F}|/\delta\big)}{n}\,.
\end{aligned}
$$

The proof of Lemma 3.1 can be found in Appendix B. This result first appeared for cost-sensitive classification (Foster & Krishnamurthy, 2021) and was extended to batch RL (with costs) in Ayoub et al. (2024). By adapting their proofs, one obtains the same bound for batch RL with rewards. Observe that their analysis also proceeds by reducing the classification (or RL) objective to bounding the MAE between $\hat{f}_{\log}$ and r. We now show that, under certain conditions ($v^\star \approx 1$), one can achieve small MAE while the rMSE remains large.

3.2 Negative Results

In this section, we argue that rMSE is not the most "natural objective function" (Chapter 9.2 of (Sutton & Barto, 2018)) for RL. We make two claims. The first claim (Proposition 3.2) is that there are problems where the rMSE of $\hat{f}_{\log}$ is at least $\frac{1}{\sqrt{n}}$, while its MAE is as small as $\frac{1}{n}$. The second claim (Lemma 3.3) is that if one directly minimizes the squared loss ℓ_{sq}, then there exist problems for which the resulting estimator suffers a large ($\gtrsim 1/\sqrt{n}$) MAE (and hence rMSE), even when $v^\star \approx 1$.

Proposition 3.2. *Let $\mathcal{X} = \{x, x'\}$. For every $n \geq 1$, there exists a realizable (Assumption 2.1) function class $\mathcal{F} : \mathcal{X} \to [0,1]$ with $|\mathcal{F}| = 2$ and a data distribution $\mathcal{P}$ such that $1 - v^\star = 1 - \int f^\star \, P_X(dx) < \frac{1}{n}$. However, with probability at least $1/(2e)$*

$$\sqrt{\int \left(f^\star(x) - \hat{f}_{\log}(x)\right)^2 P_X(dx)} = \frac{1}{2\sqrt{n}} \, .$$

In the construction of Proposition 3.2, Lemma 3.1 implies that

$$\int \left|f^\star(x) - \hat{f}_{\log}(x)\right| P_X(dx) \lesssim \frac{1}{n},$$

yet the rMSE remains at order $\frac{1}{\sqrt{n}}$. Intuitively, context x' appears with low probability ($1/n$), so the log-loss estimator $\hat{f}_{\log}$ might be quite noisy on x' but accurate on the high-probability context x. Since rMSE weights the rare event x' more heavily than MAE does, it overestimates the error relevant for *decision making* (which focuses on mainline states). The formal details can be found in Appendix C.

Finally, we show that this "over-weighting of rare events" carries over to *minimizing the squared loss* directly, as commonly done by value-based RL algorithms with function approximation (Sutton & Barto, 2018).

Lemma 3.3. *Let $\mathcal{X} = \{x, x'\}$. For any $n \geq 1$, there exists a realizable (Assumption 2.1) function class $\mathcal{F} : \mathcal{X} \to [0,1]$ with $|\mathcal{F}| = 2$ and a data distribution $\mathcal{P}$ such that $1 - v^\star = 1 - \int f^\star \, P_X(dx) < \frac{1}{n}$. However, with probability at least $1/(2e)$,*

$$\int \left|\hat{f}_{\text{sq}}(x) - f^\star(x)\right| P_X(dx) \geq \frac{1}{3\sqrt{n}} \, ,$$

where

$$\hat{f}_{\text{sq}} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell_{\text{sq}}\big(f(X_i), Y_i\big).$$

The proof of Lemma 3.3 can be found in Appendix C. In cost-sensitive classification, Foster & Krishnamurthy (2021) showed an analogous phenomenon: the greedy policy with respect to the squared-loss minimizer can incur $\frac{1}{\sqrt{n}}$-sized suboptimality even though $v^\star \geq 1 - \frac{1}{n}$. Our lemma complements their result by demonstrating that the squared-loss minimizer itself fails to achieve a $O(1/n)$ decay in its mean absolute error, whereas the log-loss minimizer can achieve $O(1/n)$ decay in similar settings. Collectively, Lemma 3.1, Proposition 3.2, and Lemma 3.3 highlight that the MAE (and hence the log-loss objective) is more tightly coupled to $\text{Sub-opt}(\hat{\pi})$ than rMSE is. Moreover, algorithms specifically designed to control the MAE—such as minimizing $\ell_{\log}$ when $v^\star \approx 1$—can adapt more effectively to problem structure than those designed around controlling rMSE (via ℓ_{sq}).

4 Reparameterizing the Categorical Cross Entropy Loss

We have seen that the log-loss can outperform the squared loss in decision-making tasks, particularly when $v^\star$ is close to 1. A natural next step is to seek a multi-category version of log-loss that retains its ability to learn the mean with sufficient data. This can be accomplished by *Reparameterizing* the categorical cross-entropy loss so that it serves as both a "classification" and a "regression" loss.

Recall that the canonical categorical cross-entropy (cat-loss) can be written as the negative log-likelihood of an exponential family (Brown, 1987). Let $y \in [0,1]$ be a *scalar* and $\theta \in \mathbb{R}^K$. In its canonical form, the cat-loss can be (naively) used for value learning in reinforcement learning,

$$\ell(\theta, y) = \log\Big(\sum_{i=1}^{K} \exp(\theta_i)\Big) - T(y)^\top \theta,$$

where $T(y)$ "bins" y into one of K discrete categories. Concretely,

$$T(y) = \left[\mathbb{I}\{\, 0 \leq y \leq \nu_1\}, \mathbb{I}\{\nu_1 < y \leq \nu_2\}, \ldots, \mathbb{I}\{\nu_{K-1} < y \leq 1\}\right],$$

where $0 < \nu_1 < \nu_2 < \ldots < \nu_{K-1} < 1$. Unfortunately, this form of the cat-loss introduces an irreducible *projection bias* for regression tasks since the exact location of y within the bin is lost. This bias prevents accurate value function estimation.

Reparameterized Cat-Loss. To remove this bias while retaining the multi-category structure, we reparameterize the loss to incorporate y directly into the sufficient statistic. Define

$$\ell_{\text{cat}}(\theta, y) = \underbrace{\log\Big(1 + \sum_{i=1}^{K-1} \exp\big(\nu_i\, \theta_i\big) + \exp(\theta_K)\Big)}_{= A(\theta)} - y\, T(y)^\top \theta.$$

Here, $A(\theta)$ is the log-partition function, and the sufficient statistic is $y\,T(y)$. Thus, ℓ_{cat} remains the negative log-likelihood of an exponential family, but one that *does not* lose all fine-grained information about y through binning. The following proposition shows that ℓ_{cat} preserves the mean of any bounded scalar random variable.

Proposition 4.1. *Let Y be a bounded random variable taking values in $[0,1]$ with mean $\mu = \mathbb{E}[Y]$, and let P be its distribution. Then*

$$\theta_\star \in \operatorname*{argmin}_{\theta \in \mathbb{R}^K} \int \ell_{\text{cat}}(\theta, y)\, P(dy) \iff \big(\nabla A(\theta_\star)\big)^\top \mathbf{1} = \mu.$$

Proof. Since ℓ_{cat} is convex in θ, first-order optimality conditions imply

$$\theta_\star \in \operatorname*{argmin}_{\theta \in \mathbb{R}^K} \int \ell_{\text{cat}}(\theta, y)\, P(dy) \iff \nabla\Big(\int \ell_{\text{cat}}(\theta_\star, y)\, P(dy)\Big) = 0\,.$$

Differentiating under the integral, we obtain

$$\nabla A(\theta_\star) = \int y\, T(y)\, P(dy).$$

Next, observe that $T(y)$ is a one-hot bin indicator, so

$$\Big(\int y\, T(y)\, P(dy)\Big)^\top \mathbf{1} = \int \big(y\, T(y)^\top \mathbf{1}\big)\, P(dy) = \int y\, P(dy) = \mu\,.$$

It follows that $\big(\nabla A(\theta_\star)\big)^\top \mathbf{1} = \mu$, establishing the claim. □

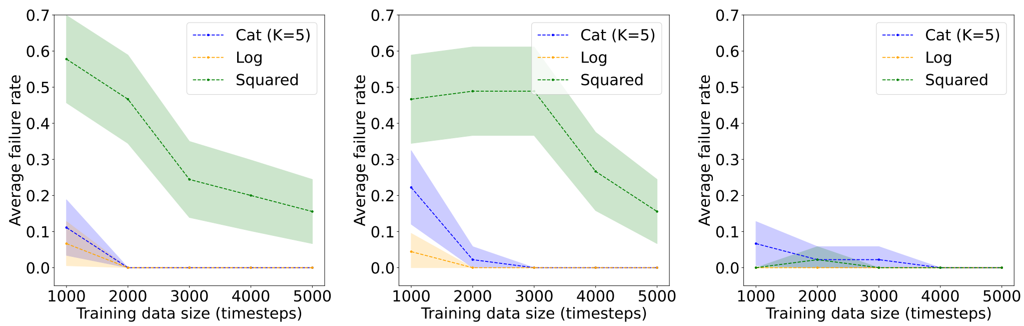

Figure 1: Failure rates for inverted pendulum as a function of the size of the batch dataset. Results are averaged over 45 independently collected datasets, and fitted Q-iteration was run for 50 iterations. We report 90% confidence intervals via the shaded regions. The LEFT and MIDDLE figures use Fourier features of order 2, the RIGHT figure uses Fourier features of order 3. The LEFT figure uses 5 uniformly spaced points as the support for the CAT, while the MIDDLE and RIGHT figures use 5 non-uniformly spaced points as the support.

Hence, minimizing ℓ_{cat} recovers the mean of any bounded scalar random variable Y while the structure of $T(y)$ still allows multi-category classification. In essence, by choosing the sufficient statistic to be $y\,T(y)$, we unify classification and regression without the loss of granularity inherent in a purely bin-based approach.

Since the cat-loss is a generalization of the log-loss, a first-order bound, similar to that of Lemma 3.1, can be shown for the cat-loss. Thus the cat-loss also does well when the optimal return $v^\star \approx 1$, albeit with an additional factor of K scaling the bound due to concentration arguments (Zhang, 2006; Grünwald & Mehta, 2020).

Extensions. This mean-preserving strategy can be generalized to other distribution-based losses derived from exponential families. For instance, *HL Gauss* (Imani & White, 2018), which has been shown to have very low test MAE (Imani et al., 2024), can be reparameterized similarly to retain its distributional modeling benefits for classification, while accurately recovering the mean of continuous targets. In a broader sense, *any* exponential-family log-likelihood can be adapted for dual "classification"-regression usage by carefully selecting sufficient statistics that embed the raw target y, thus allowing us to harness the benefits of classification[1] losses (i.e., categorical cross entropy) for reinforcement learning and continue regressing (Farebrother et al., 2024).

5 Numerical Experiments

We evaluate fitted Q-iteration (Ernst et al., 2005; Szepesvári, 2022) trained with squared loss (SQUARE), log-loss (LOG), and cat-loss (CAT) on the inverted pendulum environment (Lagoudakis & Parr, 2003; Riedmiller, 2005), where the goal is to keep an inverted pendulum balanced by applying the correct forces. The state space is two-dimensional (angle and angular velocity), and there are three actions (left, right, do nothing). The environment dynamics follow Lagoudakis & Parr (2003), with two modifications: (i) when the pendulum falls below horizontal, the state terminates, and (ii) the angular momentum is clipped to [-5,5] to facilitate the use of Fourier features (Konidaris et al., 2011) of orders 2 and 3. The agent receives a reward of 0 for staying upright and -1 for falling, with a discount factor of $\gamma = 0.99$. All datasets are collected by a policy that selects actions uniformly at random until failure (which typically occurs after 6 steps). We then evaluate whether the learned policies keep the inverted pendulum above horizontal after 3000 steps and report the policy's *failure rate*.

[1]Cross-entropy losses

For minimizing SQUARE, we use its closed-form solution; for minimizing LOG, we apply Newton's method (Sun & Tran-Dinh, 2019); and for minimizing CAT, we use a limited-memory BFGS method (Schmidt et al., 2009). All three methods are guaranteed to converge to their respective optima superlinearly.

In this environment, there is a policy that achieves the maximum possible return. As shown in Section 5, both LOG and CAT outperform SQUARE with order-2 Fourier features, while all three perform similarly with order-3 features, though LOG learns fastest. While LOG and SQUARE were relatively straightforward to implement, CAT proved more sensitive to both the choice of ℓ_2 regularization and the spacing and number of bins.

6 Conclusion

In this paper, we examined how the choice of regression objective can influence the design of value-based methods in reinforcement learning. We showed that losses aligning with mean absolute error, such as log-loss and a reparameterized categorical loss, can yield stronger theoretical guarantees and better empirical outcomes than squared loss, especially when the optimal policy is near maximum return. We also presented negative examples illustrating that purely MSE-based approaches can learn slowly in such scenarios. Finally, our experiments in a linear batch reinforcement learning setting reinforce these conclusions.

References

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in Neural Information Processing Systems*, 2007.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020.

Alex Ayoub, Kaiwen Wang, Vincent Liu, Samuel Robertson, James McInerney, Dawen Liang, Nathan Kallus, and Csaba Szepesvári. Switching the loss reduces the cost in batch reinforcement learning. In *International Conference on Machine Learning*, 2024.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017.

Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 2003.

Lawrence D. Brown. *Fundamentals of statistical exponential families : with applications in statistical decision theory.* Institute of Mathematical Statistics, 1987.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2018.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, 2001.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taiga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, et al. Stop regressing: training value functions via classification for scalable deep rl. In *International Conference on Machine Learning*, 2024.

Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 2012.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 2021.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *Journal of Machine Learning Research*, 2020.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *International Conference on Machine Learning (ICML)*, 2018.

Ehsan Imani, Kai Luedemann, Sam Scholnick-Hughes, Esraa Elelimy, and Martha White. Investigating the histogram loss in regression. *arXiv preprint arXiv:2402.13425*, 2024.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: new rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 2023.

George Konidaris, Sarah Osentoski, and Philip Thomas. Value function approximation in reinforcement learning using the fourier basis. In *AAAI Conference on Artificial Intelligence*, 2011.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 2003.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, 2005.

Mark Schmidt, Ewout Berg, Michael Friedlander, and Kevin Murphy. Optimizing costly functions with simple constraints: a limited-memory projected quasi-newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, 2009.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: an introduction*. The MIT Press, 2018.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer Nature, 2022.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1991.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: second-order bounds for reinforcement learning. In *International Conference on Machine Learning*, 2024.

Tong Zhang. From ε-entropy to kl-entropy: analysis of minimum information complexity density estimation. *The Annals of Statistics*, 2006.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Additional Notation

In this section, we introduce additional notation we will find useful in stating our theoretical results. For a distribution P over the reals, let $\mathbb{E}(P)$ denote the mean of P and $\mathrm{Var}(P)$ denote the variance of P (assuming they exist). Furthermore for $p, q \in [0, 1]$ define the pointwise triangular deviation between p, q as

$$\Delta(p, q) = \frac{(p-q)^2}{p+q}$$

and the binary Hellinger distance of p and q as

$$\mathrm{h}^2(p, q) = \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + \frac{1}{2}(\sqrt{1-p} - \sqrt{1-q})^2\,.$$

Furthermore for any $x \in [0, 1]$, define

$$L(x) = 1 - x\,.$$

B Proof of Lemma 3.1

We begin by bounding the suboptimality gap of the value-based method that minimizes the log-loss $\ell_{\log}$.

Proof of Lemma 3.1. By Lemma D.1, we have

$$\begin{aligned}\text{Sub-opt}(\hat{\pi}) &= v^\star - \hat{v} = L(\hat{v}) - L(v^\star)\\ &\le 8\sqrt{L(v^\star)\,\mathbb{E}\Big[\sum_{a\in\mathcal{A}} \Delta\big(L(f^\star(S,a)),\, L(\hat{f}(S,a))\big)\Big]} + 17\,\mathbb{E}\Big[\sum_{a\in\mathcal{A}} \Delta\big(L(f^\star(S,a)),\, L(\hat{f}(S,a))\big)\Big].\end{aligned}$$

Next, by Lemma D.2,

$$\Delta\big(L(f^\star(S,a)),\, L(\hat{f}(S,a))\big) \le 4\ \mathrm{h}^2\big(f^\star(S,a),\, \hat{f}(S,a)\big).$$

Applying Theorem D.3, we get

$$\mathbb{E}\Big[\sum_{a\in\mathcal{A}} \Delta\big(L(f^\star(S,a)),\, L(\hat{f}(S,a))\big)\Big] \le \frac{2A\,\log\big(|\mathcal{F}|/\delta\big)}{n}.$$

The second part of the lemma follows from the argument in the proof of Lemma 1 in Foster & Krishnamurthy (2021), where bounding the mean absolute error appears as an intermediate step. □

C Proof of the Negative Results

In this section, we show that minimizing the empirical squared loss does not always achieve a $1/n$-rate for the mean absolute error (MAE) when $\int f^\star(x)\,P_X(dx) \approx 1$. By contrast, Lemma 3.1 implies that minimizing the empirical log-loss *does* achieve such a rate under the same conditions.

Setup. Given a dataset $\{(X_i, Y_i)\}_{i=1}^n \sim \mathcal{P}^{\otimes n}$ with $Y_i \in [0, 1]$ and $X_i \in \mathcal{X}$, define the empirical squared loss of a function $f : \mathcal{X} \to [0, 1]$ by

$$\hat{L}(f) = \sum_{i=1}^n \big(f(X_i) - Y_i\big)^2.$$

Let P_X denote the distribution of the contexts. Given a fixed function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$, define the empirical risk minimizer (ERM) for squared loss as

$$\hat{f}_{\text{sq}} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}(f).$$

We now construct a problem instance where $\hat{f}_{\text{sq}}$ does not achieve $1/n$-rate convergence for MAE, even though $f^\star \approx 1$. Recall that by Lemma 3.1, the empirical log-loss minimizer *can* achieve a $1/n$-rate under similar conditions.

Construction. Let $\mathcal{X} = \{x, x'\}$ and set $P_X(x) = 1 - 1/n$ and $P_X(x') = 1/n$. The labels Y have the following conditional distributions:

1. $Y \mid x = 1 - \frac{1}{2n}$ almost surely, so $\mathbb{E}[Y \mid X = x] = f^\star(x) = 1 - \frac{1}{2n}$.
2. $Y \mid x' \sim \text{Bernoulli}(1/2)$, so $\mathbb{E}[Y \mid X = x'] = f^\star(x') = 1/2$.

We take the function class $\mathcal{F} = \{f^\star, \psi\}$, where

$$\psi(x) = 1 - \frac{1}{2n} - \frac{1}{3\sqrt{n}}, \quad \text{and} \quad \psi(x') = 0.$$

This class satisfies the realizability assumption (i.e., $f^\star \in \mathcal{F}$).

Proof of Lemma 3.3. Suppose $\hat{L}(\psi) \le \hat{L}(f^\star)$. Then $\hat{f}_{\text{sq}} = \psi$ and thus

$$\int \left|\hat{f}_{\text{sq}}(x) - f^\star(x)\right| P_X(dx) \ge \left| \left(1 - \frac{1}{2n} - \frac{1}{3\sqrt{n}}\right) - \left(1 - \frac{1}{2n}\right) \right| = \frac{1}{3\sqrt{n}}.$$

It remains to show that $\hat{L}(\psi) \le \hat{L}(f^\star)$ holds with constant probability. Let N_2 be the number of times $X_i = x'$ in the dataset. Then

$$\mathbb{P}(N_2 = 1) = \sum_{i=1}^{n} \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} = \frac{1}{1 - \frac{1}{n}}\left(1 - \frac{1}{n}\right)^{n-1} \ge \frac{1}{e},$$

for all $n \ge 1$. Conditioning on the event $N_2 = 1$, we have exactly one observation of x'. Since $Y = 0$ in that observation with probability $1/2$, it follows that with probability at least $1/(2e)$ we observe a single $(x', 0)$ point in the dataset. On this event,

$$\hat{L}(f^\star) - \hat{L}(\psi) = \frac{1}{4} - (n-1)\left(\frac{1}{3\sqrt{n}}\right)^2 \ge \frac{1}{4} - \frac{1}{9} > 0.$$

Hence $\hat{f}_{\text{sq}} = \psi$ with probability at least $1/(2e)$ for all $n \ge 1$. □

Proof of Proposition 3.2. We use a similar construction to prove that the root mean squared error (rMSE) of $\hat{f}_{\log}$ can remain at $\Omega(1/\sqrt{n})$ even when $\int f^\star(x)\, P_X(dx) \approx 1$. We slightly modify the function class to

$$\mathcal{F}_{\log} = \{\, f^\star,\, \phi\}, \quad \text{where } \phi(x) = f^\star(x), \phi(x') = 0.$$

Define

$$\hat{L}_{\log}(f) = \sum_{i=1}^{n} \ell_{\log}\big(f(X_i), Y_i\big), \quad \hat{f}_{\log} = \operatorname*{argmin}_{f \in \mathcal{F}_{\log}} \hat{L}_{\log}(f).$$

If $\hat{L}_{\log}(\phi) \le \hat{L}_{\log}(f^\star)$, then $\hat{f}_{\log} = \phi$, and

$$\sqrt{\int \big(\hat{f}_{\log}(x) - f^\star(x)\big)^2 P_X(dx)} = \frac{1}{2\sqrt{n}}.$$

Let N_2 be the number of times $X_i = x'$. As in the proof of Lemma 3.3,

$$\mathbb{P}(N_2 = 1) \geq \frac{1}{e}.$$

Conditioning on $N_2 = 1$, we have $Y = 0$ at x' with probability $1/2$, which occurs with probability $1/(2e)$. On this event,

$$\hat{L}_{\log}(f^\star) - \hat{L}_{\log}(\phi) = \log \frac{1}{1 - 0.5} - \log \frac{1}{1} = \log(2) > 0,$$

so $\hat{f}_{\log} = \phi$. Hence with probability at least $1/(2e)$, the rMSE between $\hat{f}_{\log}$ and $f^\star$ remains $\frac{1}{2\sqrt{n}}$.

Thus, we have shown that there are problems where $v^\star \approx 1$ but the rMSE of $\hat{f}_{\log}$ decays no faster than $1/\sqrt{n}$.

D Technical Results

Lemma D.1 (Lemma 1 of Foster & Krishnamurthy (2021))**.** *For any function $f : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and policy π that is greedy with respect to f,*

$$\begin{aligned} &L(v^\pi) - L(v^\star) \\ &\quad \leq 8 \sqrt{L(v^\star) \int \sum_{a \in \mathcal{A}} \Delta\big(L^\star(s, a), \hat{L}(s, a)\big)\, P_S(ds)} + 17 \int \sum_{a \in \mathcal{A}} \Delta\big(L^\star(s, a), \hat{L}(s, a)\big)\, P_S(ds) \end{aligned}$$

where $L^\star(s, a) = L(f^\star(s, a))$ and $\hat{L}(s, a) = L(\hat{f}(s, a))$.

Lemma D.2 (Lemma A.1 of Ayoub et al. (2024))**.** *For all $p, q \in [0, 1]$, we have*

$$\frac{1}{4}\Delta(p, q) \leq \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 \leq \mathrm{h}^2(p, q)\,.$$

D.1 Concentration for the Log-Loss Estimator

Fix a context set $\mathcal{X}$. Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. samples from a distribution $\nu \in \mathcal{M}_1(\mathcal{X} \times [0, 1])$. Define the regression function

$$f^\star(x) = \mathbb{E}\big[Y_1 \,\big|\, X_1 = x\big].$$

Suppose we have a finite class of candidate functions $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$. Recall that the log-loss estimator is given by

$$\hat{f}_{\log} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \ell_{\log}\big(f(X_i),\, Y_i\big),$$

where for $x, y \in [0, 1]$,

$$\ell_{\log}(x, y) = y\, \log \frac{1}{x} + (1 - y)\, \log \frac{1}{1 - x},$$

with the convention $0 \log \infty = \lim_{u \to 0} u \log \frac{1}{u} = 0$.

Foster & Krishnamurthy (2021) establish the following concentration result for $\hat{f}_{\log}$. We restate it here for completeness.

Theorem D.3. *Suppose $f^\star \in \mathcal{F}$. Let $D_n = \{(X_i, Y_i)\}_{i=1}^n \sim \nu^{\otimes n}$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\int \mathrm{h}^2\big(\hat{f}_{\log}(x),\, f^\star(x)\big)\, \nu_X(dx) \leq \frac{2\, \log\big(|\mathcal{F}|/\delta\big)}{n},$$

where ν_X is the marginal distribution of $X_1, \ldots, X_n$.

Proof. The result follows directly from the last equation on page 24 of the arXiv version of Foster & Krishnamurthy (2021), taking $A = 1$. ∎