Replacing Rewards with Examples: Example-Based Policy Search via Recursive Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement learning (RL) algorithms assume that users specify tasks by manually writing down a reward function. However, this process can be laborious and demands considerable technical expertise. Can we devise RL algorithms that instead enable users to specify tasks simply by providing examples of successful outcomes? In this paper, we derive a control algorithm that maximizes the future probability of these successful outcome examples. Prior work has approached similar problems with a two-stage process, first learning a reward function and then optimizing this reward function using another reinforcement learning algorithm. In contrast, our method directly learns a value function from transitions and successful outcomes, without learning this intermediate reward function. Our method therefore requires fewer hyperparameters to tune and lines of code to debug. We show that our method satisfies a new data-driven Bellman equation, where examples take the place of the typical reward function term. Experiments show that our approach outperforms prior methods that learn explicit reward functions.[1]

1 Introduction

In supervised learning settings, tasks are defined by data: what causes a car detector to detect cars is not the choice of loss function (which might be the same as for an airplane detector), but the choice of training data. Defining tasks in terms of data, rather than specialized loss functions, arguably makes it easier to apply machine learning algorithms to new domains. In contrast, reinforcement learning (RL) problems are typically posed in terms of a reward function, which are typically manually designed. Arguably, because designing reward functions is challenging, RL has been limited to applications with simple reward functions, and has been restricted to users who speak this language of mathematically-defined reward functions. Can we make task specification in RL similarly "data-driven"?

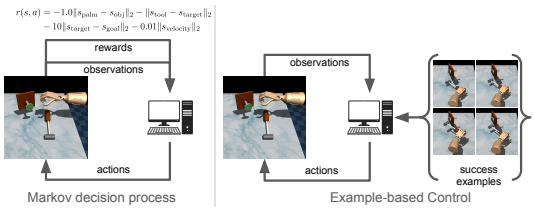

Figure 1: **Example-based control**: Whereas the standard MDP framework requires a user-defined reward function, example-based control specifies tasks via a handful of user-provided success examples.

Whereas the standard MDP formalism centers around predicting and maximizing the future reward, we will instead focus on the problem classifying whether a task will be solved in the future. The user will provide a collection of example success states, not a reward function. We call this problem setting *example-based control*. In effect, these examples tell the agent "What would the world look like if the task were solved?" For example, for the task of opening a door, success examples correspond to

[1] Project site with videos and code: `https://rce-anonymous.github.io/`

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

different observations of the world when the door is open. The user can find examples of success even for tasks that they themselves do not know how to solve. For example, the user could solve the task using actions unavailable to the agent (e.g., the user may have two arms, but a robotic agent may have only one) or the user could find success examples by searching the internet. As we will discuss in Sec. 3.1, this problem setting is different from imitation learning: we maximize a different objective function and only assume access to success examples, not entire expert trajectories.

Learning from examples is challenging because we must automatically identify when the agent has solved the task and reward it for doing so. Prior methods (either imitation learning from demonstrations or learning from success examples) take an indirect approach that resembles inverse RL: first learn a separate model to represent the reward function, and then optimize this reward function with standard RL algorithms. Our method is different from these prior methods because it learns to predict *future* success directly from transitions and success examples, without learning a separate reward function. This key difference has important algorithmic, theoretical, and empirical benefits. **Algorithmically**, our end-to-end approach removes potential biases in learning a separate reward function, reduces the number of hyperparameters, and simplifies the resulting implementation. **Theoretically**, we propose a method for classifying *future* events using a variant of temporal difference learning that we call recursive classification. This method satisfies a new Bellman equation, where success examples are used in place of the standard reward function term. We use this result to provide convergence guarantees. **Empirically**, we demonstrate that our method solves many complex manipulation tasks that prior methods fail to solve.

Our paper also addresses a subtle but important ambiguity in formulating example-based control. Some states might always solve the task while other states might rarely solve the task. But, without knowing how often the user visited each state, we cannot determine the likelihood that each state solves the task. Thus, an agent can only estimate the probability of success if they make an additional assumption about how the success examples were generated. We will discuss two choices of assumptions. The first choice of assumption is convenient from an algorithmic perspective, but sometimes violated in practice. A second choice is a worst-case approach, a problem setting that we call *robust example-based control*. Our analysis shows that the robust example-based control objective is equivalent to minimizing the squared Hellinger distance (an f-divergence).

In summary, this paper studies a data-driven framing of control, where reward functions are replaced by examples of successful outcomes. Our main contribution is an algorithm for off-policy example-based control. The key idea of the algorithm is to directly learn to predict whether the task will be solved in the future via *recursive classification*, without using separate reward learning and policy search procedures. Our analysis shows that our method satisfies a new Bellman equation where rewards are replaced by data (examples of success). Empirically, we demonstrate that our method excels at learning complex manipulation tasks solely from examples of success, solving tasks like bin picking that *none* of the baselines make any progress on. On all tasks, our method significantly outperforms state-of-the-art imitation learning methods (AIRL [9], DAC [18], and SQIL [28]) and recent methods that learn reward functions (VICE [10], PURL [38], and ORIL [41]). Our method learns many tasks that *none* of these baselines can solve. On tasks with image observations, we demonstrate that the notion of success learned by our method generalizes to new environments with varying shapes and goal locations.

2 Related Work

Learning reward functions. A number of prior works have studied RL in settings where the task is specified either with examples of successful outcomes or complete demonstrations. These prior methods typically learn a reward function from data and then apply RL to this reward function (e.g., Fu et al. [10], Ziebart et al. [40]). Most inverse RL algorithms adopt this approach [1, 25, 27, 29, 36, 40], as do more recent methods that learn a *success classifier* to distinguishing successful outcomes from random states [10, 17, 31, 41]. Prior adversarial imitation learning methods [9, 14] can be viewed as *iteratively* learning a success classifier. Recent work in this area focuses on extending these methods to the offline setting [17, 41], incorporating additional sources of supervision [42], and learning the classifier via positive-unlabeled classification [15, 38, 41]. Many prior methods for robot learning have likewise used a classifier to distinguish success examples [3, 23, 34, 37]. Unlike these prior methods, our approach only requires examples of successful outcomes (not expert trajectories) and does not learn a separate reward function. Instead, our method learns a value function directly from examples, effectively "cutting out the middleman." Practically, our method removes hyperparameters

and potential bugs associated with learning a success classifier. Empirically, we demonstrate that our end-to-end approach outperforms these prior two-stage approaches. See Appendix D for more discussion of the relationship between our method and prior work.

Imitation learning without auxiliary classifiers. While example-based control is different from imitation learning, our method is similar to two prior imitation learning methods [19, 28]. ValueDICE [19], a method based on convex duality, uses full expert demonstrations for imitation learning. In contrast, our method learns from success examples, which are typically easier to provide than full expert demonstrations. SQIL [28] is a modification of SAC [13] that labels success examples with a reward of $+1$. The mechanics of our method are similar to SQIL [28], but key algorithmic differences (backed by stronger theoretical guarantees) result in better empirical performance. Our analysis in Sec. 4.2 highlights connections and differences between imitation learning and example-based control.

Goal-conditioned RL. Goal-conditioned RL is a special case of example-based control where the user provides a single success state (the goal). In discrete state spaces, the value functions learned by goal-conditioned Q-learning [30, 33] and successor features [2, 4, 20] corresponds to the probability of reaching this goal state. C-learning [6] shows how a similar bootstrapping procedure can be used to estimate the *density* of future states in continuous state spaces. The problem we focus on in this paper, example-based control, is a generalization of goal-conditioned RL. Indeed, the objective for C-learning [6] is a special case of example-based control applied to a single success example. However, example-based control allows users to indicate that tasks can be solved in many ways, allowing the agent to learn a more general notion of success. While our derivation builds off of C-learning, our algorithm is markedly different; for example, our method does not require hindsight relabeling, and learns a single policy rather than a goal-conditioned policy.

3 Example-Based Control via Recursive Classification

In contrast to standard RL, which uses reward functions, we aim to learn a policy that reaches states that are likely to solve the task (see Fig. 1). We start by formally describing the problem of learning from *examples* of success, which we will call *example-based control*. We then derive a method for solving this problem and provide convergence guarantees.

3.1 Problem Statement

Example-based control is defined by a controlled Markov process (i.e., an MDP without a reward function) with dynamics $p(\mathbf{s_{t+1}} \mid \mathbf{s_t}, \mathbf{a_t})$ and an initial state distribution $p_1(s_1)$, where $\mathbf{s_t} \in \mathcal{S}$ and $\mathbf{a_t}$ denote the time-indexed states and actions. The variable $s_{t+\Delta}$ denotes a state Δ steps in the future.

The agent is given a set of *success examples*, $\mathcal{S}^* = \{s^*\} \subseteq \mathcal{S}$. The binary random variable $\mathbf{e_t} \in \{0, 1\}$ indicates whether the task is solved at time t, and $p(\mathbf{e_t} \mid \mathbf{s_t})$ denotes the probability that the current state $\mathbf{s_t}$ solves the task. Given a policy $\pi_\phi(\mathbf{a_t} \mid \mathbf{s_t})$, we define the discounted future state distribution as

$$p^\pi(\mathbf{s_{t+}} \mid \mathbf{s_t}, \mathbf{a_t}) \triangleq (1-\gamma) \sum_{\Delta=0}^{\infty} p^\pi(s_{t+\Delta} = \mathbf{s_{t+}} \mid \mathbf{s_t}, \mathbf{a_t}). \tag{1}$$

Using this definition, we can write the probability of solving the task at a *future* step as

$$p^\pi(\mathbf{e_{t+}} \mid \mathbf{s_t}, \mathbf{a_t}) \triangleq \mathbb{E}_{p^\pi(\mathbf{s_{t+}} \mid \mathbf{s_t}, \mathbf{a_t})}[p(\mathbf{e_{t+}} \mid \mathbf{s_{t+}})]. \tag{2}$$

Example-based control maximizes the probability of solving the task in the (discounted) future:

Definition 1 (Example-based control)**.** *Given a controlled Markov process and distribution over success examples $p(\mathbf{s_t} \mid \mathbf{e_t} = 1)$, the example-based control problem is to find the policy that optimizes the likelihood of solving the task:*

$$\arg\max_\pi p^\pi(\mathbf{e_{t+}} = 1) = \mathbb{E}_{p_1(\mathbf{s_1}), \pi(\mathbf{a_1} \mid \mathbf{s_1})} \left[p^\pi(\mathbf{e_{t+}} = 1 \mid \mathbf{s_1}, \mathbf{a_1}) \right].$$

While objective is equivalent to the standard RL objective with the reward function $r(\mathbf{s_t}, \mathbf{a_t}) = p(\mathbf{e_t} = 1 \mid \mathbf{s_t})$, we assume the probabilities $p(\mathbf{e_t} \mid \mathbf{s_t})$ are unknown. Instead, we assume that we have samples of successful states, $\mathbf{s}^* \sim p_U(\mathbf{s_t} \mid \mathbf{e_t} = 1)$. Example-based control is different from

imitation learning because imitation learning requires full expert demonstrations. Goal-conditioned RL is a special case of example-based control where the user provides a single success state.

Since interacting with the environment to collect experience is expensive in many settings, we define *off-policy example-based control* as the version of this problem where the agent learns from environment interactions collected from other policies. In this setting, the agent learns from two distinct datasets: **(1)** transitions, $\{(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}}) \sim p(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}})\}$, which contain information about the environment dynamics; and **(2)** success examples, $\mathcal{S}^* = \{\mathbf{s}^* \sim p(\mathbf{s_t} \mid \mathbf{e_t} = 1)\}$, which specify the task that the agent should attempt to solve. Our analysis will assume that these two datasets are fixed. The main contribution of this paper is an algorithm for off-policy example-based control.

An assumption on success examples. The probability of solving the task at state $\mathbf{s_t}$, $p(\mathbf{e_t} = 1 \mid \mathbf{s_t})$, cannot be uniquely determined from success examples and transitions alone. For example, a state which rarely solves the task might be included as a success example. To elucidate this ambiguity, we define $p_U(\mathbf{s_t})$ as the state distribution visited by the user; note that the user may be quite bad at solving the task themselves. Then, the probability of solving the task at state $\mathbf{s_t}$ depends on how often a user visits state $\mathbf{s_t}$ versus how often the task is solved when visiting state $\mathbf{s_t}$:

$$p(\mathbf{e_t} = 1 \mid \mathbf{s_t}) = \frac{p_U(\mathbf{s_t} \mid \mathbf{e_t} = 1)}{p_U(\mathbf{s_t})} p_U(\mathbf{e_t} = 1). \tag{3}$$

For example, the user may complete a task using two strategies, but we cannot determine which of these strategies is more likely to succeed unless we know how often the user attempted each strategy. Thus, any method that learns from success examples must make an additional assumption on $p_U(\mathbf{s_t})$. We will discuss two choices of assumptions. The **first choice** is to assume that the user visited states with the same frequency that they occur in the dataset of transitions. That is,

$$p_U(\mathbf{s_t}) = \iint p(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}}) d\mathbf{a_t} d\mathbf{s_{t+1}}. \tag{4}$$

Intuitively, this assumption implies that the user has the same capabilities as the agent. Prior work makes this same assumption [10, 24, 31] without stating it explicitly. Experimentally, we find that our method succeeds even in cases where this assumption is violated.

However, the user may not visit the same distribution of states, especially if the dataset of transitions $\{(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}})\}$ was collected using a robot with different capabilities from the human. For example, the human user may prefer ordering takeout for dinner over cooking for themselves, but the dataset of transitions might consist entirely of a robot cooking dinner (we assume the robot cannot order takeout). The handle these difference in capabilities, the **second choice** is to use a *worst-case* formulation, which optimizes the policy to be robust to *any* choice of $p_U(\mathbf{s_t})$. Surprisingly, this setting admits a tractable solution, as we discuss in Sec. 4.2.

3.2 Predicting Future Success by Recursive Classification

We now describe our method for example-based control. We start with the more standard **first choice** for the assumption on $p_U(\mathbf{s_t})$ (Eq. 4); we discuss the second choice in Sec. 4.2. Our approach estimates the probability in Eq. 2 indirectly via a future success classifier. This classifier, $C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t})$, discriminates between "positive" state-action pairs which lead to successful outcomes (i.e., sampled from $p(\mathbf{s_t}, \mathbf{a_t} \mid \mathbf{e_{t+}} = 1)$) and random "negatives" (i.e., sampled from the marginal distribution $p(\mathbf{s_t}, \mathbf{a_t})$). We will use different class-specific weights, using a weight of $p(\mathbf{e_{t+}} = 1)$ for the "positives" and a weight of 1 for the "negatives" Bayes-optimal classifier is

$$C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t}) = \frac{p^\pi(\mathbf{s_t}, \mathbf{a_t} \mid \mathbf{e_{t+}} = 1) p(\mathbf{e_{t+}} = 1)}{p^\pi(\mathbf{s_t}, \mathbf{a_t} \mid \mathbf{e_{t+}} = 1) p(\mathbf{e_{t+}} = 1) + p(\mathbf{s_t}, \mathbf{a_t})}. \tag{5}$$

These class specific weights let us predict the probability of future success using the optimal classifier:

$$\frac{C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t})}{1 - C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t})} = p^\pi(\mathbf{e_{t+}} = 1 \mid \mathbf{s_t}, \mathbf{a_t}). \tag{6}$$

Importantly, the resulting method will not actually require estimating the weight $p(\mathbf{e_{t+}} = 1)$. We would like to optimize these parameters using maximum likelihood estimation:

$$\mathcal{L}^\pi(\theta) \triangleq p(\mathbf{e_{t+}} = 1)\, \mathbb{E}_{p(\mathbf{s_t}, \mathbf{a_t} \mid \mathbf{e_{t+}} = 1)}[\log C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t})] + \mathbb{E}_{p(\mathbf{s_t}, \mathbf{a_t})}[\log(1 - C^\pi_\theta(\mathbf{s_t}, \mathbf{a_t}))]. \tag{7}$$

Algorithm 1 Recursive Classification of Examples

Input: success examples $\mathcal{S}^*$
Initialize policy $\pi_\phi(\mathbf{a_t} \mid \mathbf{s_t})$, classifier $C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t})$, replay buffer $\mathcal{D}$
while not converged **do**
  Collect a new trajectory: $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tau \sim \pi_\phi\}$
  Sample success examples: $\{\mathbf{s_t}^{(1)} \sim \mathcal{S}^*, \mathbf{a_t}^{(1)} \sim \pi_\phi(\mathbf{a_t}|\mathbf{s_t}^{(1)})\}$
  Sample transitions: $\{(\mathbf{s_t}^{(2)}, \mathbf{a_t}^{(2)}, \mathbf{s_{t+1}}) \sim \mathcal{D}, \mathbf{a_{t+1}} \sim \pi_\phi(\mathbf{a_{t+1}} \mid \mathbf{s_{t+1}})\}$
  $w \leftarrow \frac{C_\theta^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})}{1 - C_\theta^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})}$ ▷ Eq. 9
  $\mathcal{L}(\theta) \leftarrow (1-\gamma)\mathcal{CE}(C_\theta(\mathbf{s_{t+1}}^{(1)}, \mathbf{a_t}^{(1)}); y = 1) + (1+\gamma w)\mathcal{CE}(C_\theta(\mathbf{s_t}^{(2)}, \mathbf{a_t}^{(2)}); y = \frac{\gamma w}{1+\gamma w})$
  Update classifier: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{L}(\theta)$ ▷ Eq. 8
  Update policy: $\phi \leftarrow \phi + \eta \nabla_\phi \mathbb{E}_{\pi_\phi(\mathbf{a_t}|\mathbf{s_t})}[C_\theta(\mathbf{s_t}, \mathbf{a_t})]$
return π_ϕ

However, we cannot directly optimize this objective because we cannot sample from $p(\mathbf{s_t}, \mathbf{a_t} \mid \mathbf{e_{t+}} = 1)$. We convert Eq. 7 into an equivalent loss function that we can optimize using three steps; see Appendix A for a detailed derivation. The **first step** is to factor the distribution $p(\mathbf{s_t}, \mathbf{a_t}, \mathbf{e_{t+}} = 1)$. The **second step** is to decompose $p^\pi(\mathbf{e_{t+}} = 1 \mid \mathbf{s_t}, \mathbf{a_t})$ into two terms, corresponding to the probabilities of solving the task at time $t' = t+1$ and time $t' > t+1$. We can estimate the probability of solving the task at the next time step using the set of success examples. The **third step** is to estimate the probability of solving the task at time $t' > t+1$ by evaluating the classifier at the next time step. Combining these three steps, we can *equivalently* express the objective function in Eq. 7 using off-policy data:

$$\mathcal{L}^\pi(\theta) = (1-\gamma)\mathbb{E}_{\substack{p_U(\mathbf{s_t}|\mathbf{e_t}=1) \\ p(\mathbf{a_t}|\mathbf{s_t})}}[\underbrace{\log C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t})}_{(a)}] + \mathbb{E}_{p(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}})}[\underbrace{\gamma w \log C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t})}_{(b)} + \underbrace{\log(1 - C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t}))}_{(c)}], \tag{8}$$

where

$$w = \mathbb{E}_{p(\mathbf{a_{t+1}}|\mathbf{s_{t+1}})}\left[\frac{C_\theta^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})}{1 - C_\theta^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})}\right] \tag{9}$$

is the classifier's prediction (ratio) at the next time step. Our resulting method can be viewed as a temporal difference [33] approach to classifying future events. We will refer to our method as **recursive classification of examples (RCE)**. This equation has an intuitive interpretation. The first term **(a)** trains the classifier to predict 1 for the success examples themselves, and the third term **(c)** trains the classifier to predict 0 for random transitions. The important term is the second term **(b)**, which is analogous to the "bootstrapping" term in temporal difference learning [32]. For term **(b)**, the classifier is trained to predict that the probability of future success succeed depends on the probability of success at the *next* time step, as inferred using the classifier's own predictions.

Our resulting method is similar to existing actor-critic RL algorithms. To highlight the similarity to existing actor-critic methods, we can combine the (b) and (c) terms in the classifier objective function (Eq. 8) to express the loss function in terms of two cross entropy losses:

$$\min_\theta (1-\gamma)\mathbb{E}_{\substack{p(\mathbf{s_t}|\mathbf{e_t}=1), \\ \mathbf{a_t} \sim \pi(\mathbf{a_t}|\mathbf{s_t})}}[\mathcal{CE}(C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t}); y = 1)] + (1+\gamma w)\mathbb{E}_{p(\mathbf{s_t}, \mathbf{a_t}, \mathbf{s_{t+1}})}\Big[\mathcal{CE}\big(C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t}); y = \frac{\gamma w}{\gamma w + 1}\big)\Big]. \tag{10}$$

These cross entropy losses update the classifier to predict $y = 1$ for the success examples and to predict $y = \frac{\gamma w}{1+\gamma w}$ for other states.

Algorithm summary. Alg. 1 summarizes our method, which alternates between updating the classifier, updating the policy, and (optionally) collecting new experience. We update the policy to choose actions that maximize the classifier's confidence that the task will be solved in the future: $\max_\phi \mathbb{E}_{\pi_\phi(\mathbf{a_t}|\mathbf{s_t})}[C_\theta^\pi(\mathbf{s_t}, \mathbf{a_t})]$. Following prior work [7, 35]), we regularized the policy updates by adding an entropy term with coefficient $\alpha = 1e-4$. Implementing our method on top of existing methods such as SAC [13] or TD3 [11] requires only changing the standard Bellman loss with the loss in Eq. 10. See Appendix E for implementation details; code is available on the project website.

4 Analysis

Similarly to how RL algorithms satisfy certain convergence and optimality guarantees for reward-based MDPs, RCE satisfies many of the same convergence and optimality guarantees for example-based control. We include a further discussion of how RCE relates to prior work in Appendix D.

4.1 Bellman Equations and Convergence Guarantees

RCE satisfies a new Bellman equation, a result we use to prove that RCE converges to the Bayes-optimal classifier and the optimal policy (Def. 1). These results are important for showing that example-based control retains the theoretical guarantees of reward-based control. Proofs of all results are given in Appendix B.

Lemma 4.1. *The Bayes-optimal classifier C^π for policy π satisfies the following identity:*

$$\frac{C^\pi(\mathbf{s_t}, \mathbf{a_t})}{1 - C^\pi(\mathbf{s_t}, \mathbf{a_t})} = (1-\gamma)p(\mathbf{e_t} = 1 \mid \mathbf{s_t}) + \gamma \mathbb{E}_{\substack{p(\mathbf{s_{t+1}}|\mathbf{s_t},\mathbf{a_t}) \\ \pi(\mathbf{a_{t+1}}|\mathbf{s_{t+1}})}} \left[\frac{C^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})}{1 - C^\pi(\mathbf{s_{t+1}}, \mathbf{a_{t+1}})} \right]. \tag{11}$$

The proof combines the definition of the Bayes-optimal classifier with the assumption from Eq. 4. This Bellman equation is analogous to the standard Bellman equation for Q-learning, where the reward function is replaced by $(1-\gamma)p(\mathbf{e_t} = 1 \mid \mathbf{s_t})$ and the Q function is parametrized as $Q^\pi_\theta(\mathbf{s_t}, \mathbf{a_t}) = \frac{C^\pi_\theta(\mathbf{s_t},\mathbf{a_t})}{1-C^\pi_\theta(\mathbf{s_t},\mathbf{a_t})}$. While we do not know how to compute this reward function, the update rule for RCE is equivalent to doing value iteration using that reward function and that parametrization of the Q-function:

Lemma 4.2. *In the tabular setting, the **expected** updates for RCE are equivalent to doing value iteration with the reward function $r(\mathbf{s_t}) = (1-\gamma)p(\mathbf{e_t} = 1 \mid \mathbf{s_t})$ and a Q-function parametrized as $Q^\pi_\theta(\mathbf{s_t}, \mathbf{a_t}) = \frac{C^\pi_\theta(\mathbf{s_t},\mathbf{a_t})}{1-C^\pi_\theta(\mathbf{s_t},\mathbf{a_t})}$.*

This result tells us that RCE is equivalent to doing standard RL with a reward function $(1-\gamma)p(\mathbf{e_t} = 1 \mid \mathbf{s_t})$, but does not require knowing the probability that each state solves the task. Since value iteration converges in the tabular setting, an immediate consequence of Lemma 4.2 is that tabular RCE also converges:

Corollary 4.2.1. *RCE converges in the tabular setting.*

So far we have analyzed the training process for the classifier for a fixed policy. We conclude this section by showing that optimizing the policy w.r.t. the classifier improves the policy's performance.

Lemma 4.3. *Let policy $\pi(a \mid s)$ and success examples $\mathcal{S}^*$ be given, and let $C^\pi(\mathbf{s_t}, \mathbf{a_t})$ denote the corresponding Bayes-optimal classifier. Define the improved policy as acting greedily w.r.t. C^π: $\pi'(\mathbf{a_t} \mid \mathbf{s_t}) = \mathbb{1}(\mathbf{a} = \arg\max_a C^\pi(\mathbf{s_t}, \mathbf{a}))$. Then the improved policy is at least as good as the old policy at solving the task: $p^{\pi'}(\mathbf{e_{t+}} = 1) \geq p^\pi(\mathbf{e_{t+}} = 1)$.*

4.2 Robust Example-based Control

In this section, we derive a principled solution for the case where $p_U(\mathbf{s_t})$ is not known, which will correspond to modifying the objective function for example-based control. However, we will argue that, in some conditions, the method proposed in Sec. 3.2 is *already* robust to unknown $p_U(\mathbf{s_t})$, if that method is used with online data collection. The goal of this discussion is to provide a theoretical relationship between our method and a robust version of example-based control that makes fewer assumptions about $p_U(\mathbf{s_t})$. This discussion will also clarify how changing assumptions on the user's capabilities can change the optimal policy.

When introducing example-based control in Sec. 3.1, we emphasized that we *must* make an assumption to make the example-based control problem well defined. The exact probability that a success example solves the task depends on how often the user visited that state, which the agent does not know. Therefore, there are many valid hypotheses for how likely each state is to solve the task. We can express the set of valid hypotheses using Bayes' Rule:

$$\mathcal{P}_{\mathbf{e_t}|\mathbf{s_t}} \triangleq \left\{ \hat{p}(\mathbf{e_t} = 1 \mid \mathbf{s_t}) = \frac{p_U(\mathbf{s_t} \mid \mathbf{e_t} = 1)p(\mathbf{e_t} = 1)}{p_U(\mathbf{s_t})} \right\}.$$

Previously (Sec. 3.2), we resolved this ambiguity by assuming that $p_U(\mathbf{s_t})$ was equal to the distribution over states in our dataset of transitions. However, many common settings violate this assumption, especially when the user has different dynamics constraints than the agent. For example, a human user collecting success examples for a cleaning task might usually put away objects on a shelf at eye-level, whereas transitions collected by a robot interact with the ground-level shelves more frequently. Under our previous assumption, the robot would assume that putting objects away on higher shelves is more satisfactory than putting them away on lower shelves, even though doing so might be much more challenging for the robot.

In the absence of any prior knowledge about $p_U(\mathbf{s_t})$ (e.g., knowledge about the user's capabilities), we can instead optimize for solving the task assuming the *worst* possible choice of $p_U(\mathbf{s_t})$. This approach will make the agent robust to imperfect knowledge of the user's abilities and to mislabeled success examples. Formally, we define the *robust example-based control* problem as

$$\max_{\pi} \min_{\hat{p}(\mathbf{e_t}|\mathbf{s_t}) \in \mathcal{P}_{\mathbf{e_t}|\mathbf{s_t}}} \mathbb{E}_{p^\pi(\mathbf{s_{t+}})}[\hat{p}(\mathbf{e_{t+}} = 1 \mid \mathbf{s_{t+}})] = \max_{\pi} \min_{p_U(\mathbf{s_t})} \mathbb{E}_{p^\pi(\mathbf{s_{t+}})} \left[\frac{p_U(\mathbf{s_t} \mid \mathbf{e_t} = 1)}{p_U(\mathbf{s_t})} p(\mathbf{e_t} = 1) \right]. \tag{12}$$

Rewriting the objective on the right-hand side, we observe that this objective is equivalent to having the adversary assign a weight of $1/p_U(\mathbf{s_t})$ to each success example. The optimal adversary will assign lower weights to success examples that the policy frequently visits and higher weights to less-visited success examples. Intuitively, the optimal policy should try to reach many of the success examples, not just the ones that are easiest to reach. Thus, such a policy will continue to succeed even if certain success examples are removed, or are later discovered to have been mislabeled. Surprisingly, solving this two-player game corresponds to minimizing an f-divergence:

Lemma 4.4. *Define $H^2[p(x), q(x)] = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$ as the squared Hellinger distance, an f-divergence. Robust example-based control (Eq. 12) is equivalent to minimizing the squared Hellinger distance between policy's discounted state occupancy measure and the* ***conditional*** *distribution $p(\mathbf{s_t} \mid \mathbf{e_t} = 1)$:*

$$\min_{\hat{p}(\mathbf{e_t}|\mathbf{s_t}) \in \mathcal{P}_{\mathbf{e_t}|\mathbf{s_t}}} p^{\pi,\hat{p}}(\mathbf{e_{t+}}) = 1 - \frac{1}{2} H^2[p(\mathbf{s_t}|\mathbf{e_t} = 1), p^\pi(\mathbf{s_{t+}} = \mathbf{s_t})].$$

The main idea of the proof (found in Appendix C) is to compute the worst-case distribution $p_U(\mathbf{s_t})$ using the calculus of variations. Preliminary experiments (Fig. 5 in Appendix C) show that a version of RCE with online data collection finds policies that perform well on the robust example-based control objective (Eq. 12). In fact, under somewhat stronger assumptions, we can show that the solution of robust example-based control is a fixed point of *iterated* RCE (see Appendix C.2). Therefore, in our experiments, we use RCE with online data collection.

5 Experiments

Our experiments study how effectively RCE solves example-based control tasks, especially in comparison to prior methods that learn an explicit reward function. Both RCE and the prior methods receive only the success examples as supervision; no method has access to expert trajectories of reward functions. Additional experiments in Sec. 5.2 study whether RCE can solve tasks using image observations. These experiments test whether RCE can solve tasks in new environments that are different from those where the success examples were collected, and test whether RCE learns policies that learn a general notion of success rather than just memorizing the success examples. We include videos of learned policies online[2] and include implementation details, hyperparameters, ablation experiments, and a list of failed experiments in the Appendix.

We compare RCE against prior methods that infer a reward function from the success examples and then apply an off-the-shelf RL algorithm; some baselines iterate between these two steps. AIRL [9] is a popular adversarial imitation learning method. VICE [10] is the same algorithm as AIRL, but intended to be applied to success examples rather than full demonstrations. We will label this method as "VICE" in figures, noting that it is the same algorithm as AIRL. DAC [18] is a more recent, off-policy variant of AIRL. We also compared against two recent methods that learn rewards from *demonstrations*: ORIL [41] and PURL [38]. Following prior work [17], we also compare against

[2] `https://rce-anonymous.github.io/`

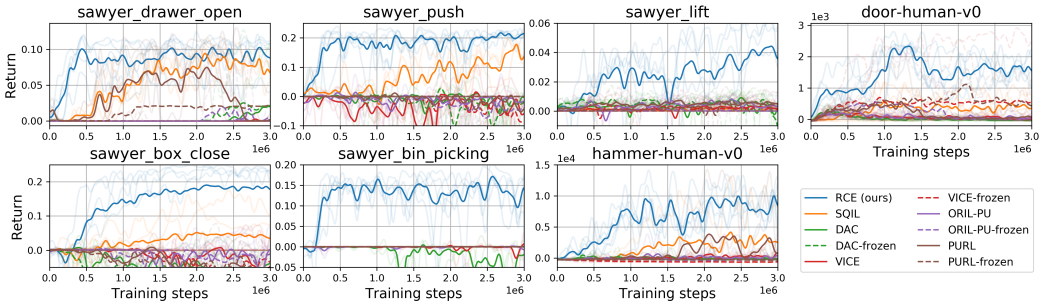

Figure 3: **Recursive Classification of Examples for learning manipulation tasks**: We apply RCE to a range of manipulation tasks, each accompanied with a dataset of success examples. For example, on the `sawyer_lift` task, we provide success examples where the object has been lifted above the table. We use the cumulative task return (↑ is better) solely for evaluation. Our method (blue line) outperforms prior methods across all tasks.

"frozen" variants of some baselines that first train the parameters of the reward function and then apply RL to that reward function without updating the parameters of the reward function again. Our method differs from these baselines in that we do not learn a reward function from the success examples and then apply RL, but rather learn a policy directly from the success examples. Lastly, we compare against SQIL [28], an imitation learning method that assigns a reward of $+1$ to states from demonstrations and 0 to all other states. SQIL does not learn a separate reward function and structurally resembles our method, but is derived from different principles (see Sec. 2.).

5.1 Evaluating RCE for Example-Based Control.

We evaluate each method on five Sawyer manipulation tasks from Meta-World Yu et al. [39] and two manipulation tasks from Rajeswaran et al. [26]. Fig. 2 illustrates these tasks. On each task, we provide the agent with 200 successful outcomes to define the task. For example, on the `open_drawer` task, these success examples show an opened drawer. We emphasize that these success examples only reflect the final state where the task is solved and are not full expert trajectories. This setting is important in practical use-cases: it is often easier for humans to arrange the workspace into a successful configuration than it is to collect an entire demonstration. See Appendix E.3 for details on how success examples were generated for each task. While these tasks come with existing user-defined reward functions, these rewards are not provided to any of the methods in our experiments and are used solely for evaluation (↑ is better). We emphasize that this problem setting is exceedingly challenging: the agent provided only with examples of success states (e.g., an observation where an object has been placed in the correct location). Most prior methods that tackle similar tasks employ hand-designed reward functions or distance functions, full demonstrations, or carefully-constructed initial state distributions.

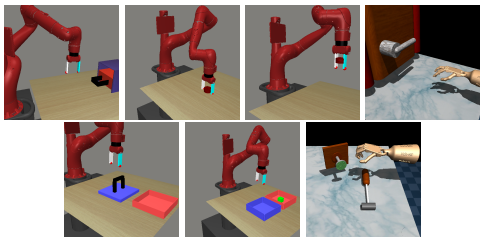

Figure 2: **Manipulation Environments**

The results in Fig. 3 show that RCE significantly outperforms prior methods across all tasks. RCE solves many tasks, such as bin picking and hammering, that *none* of the baselines make *any* progress towards solving. The most competitive baseline, SQIL, only makes progress on the easiest two tasks; even on those tasks, SQIL learns more slowly than RCE and achieves lower asymptotic return. To check that all baselines are implemented correctly, we confirm that all can solve a very simple reaching task described in the next section.

5.2 Example-Based Control from Images

Our second set of experiments studies whether RCE can learn image-based tasks and assesses the generalization capabilities of our method. We designed three *image-based* manipulation tasks. The `reach_random_position` task entails reaching a red puck, whose position is randomized in each episode. The `reach_random_size` task entails reaching a red object, but the actual shape of that object varies from one episode to the next. Since the agent cannot change the size of the object and

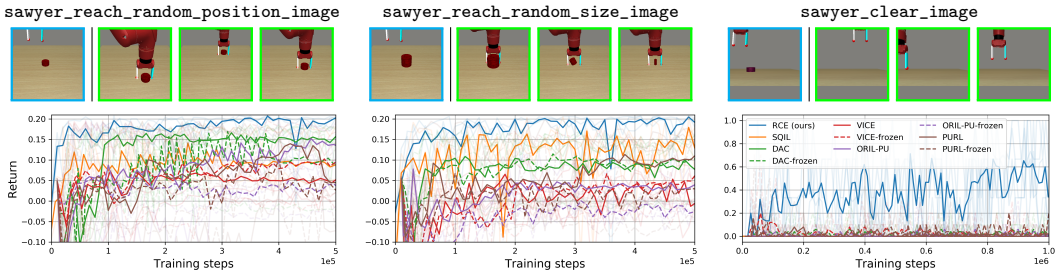

Figure 4: **Example-based control from images**: We evaluate RCE on three manipulation tasks using image-observations. *(Top)* We show examples of the initial state and success examples for each task. *(Bottom)* RCE (blue line) outperforms prior methods, especially on the more challenging clearing task. For the `random_size` task (*center*), this entails reaching for new objects that have different sizes from any seen in the success examples.

the size is randomized from one episode to the next, it is impossible to reach any of the previously-observed success examples. To solve this task, the agent must learn a notion of "success" that is more general than reaching a fixed goal state. The third task, `sawyer_clear_image`, entails clearing an object off the table, and is mechanically more challenging than the reaching tasks.

Fig. 4 shows results from these image-based experiments, comparing RCE to the same baselines. We observe that RCE has learned to solve both reaching tasks, reaching for the object regardless of the location and size of the object. This task is mechanically easier than the state-based tasks in Fig. 3, and all the baselines make some progress on this task, but learn more slowly than our method. The good performance of RCE on the `reach_random_size` task illustrates that RCE can solve tasks in a new environment, where the object size is different from that seen in the success examples. We hypothesize that RCE learns faster than these baselines because it "cuts out the middleman," learning a value function directly from examples rather than indirectly via a separate success classifier. To support this hypothesis, we note SQIL, which also avoids learning an intermediary classifier, learns faster than other baselines on the second variation. On the more challenging clearing task, only our method makes progress, suggesting that RCE is a more effective algorithm for learning these image-based control tasks. In summary, these results show that RCE outperforms prior methods at solving example-based control tasks from image observations, and highlights that RCE learns a policy that solves tasks in new environments that look different from any of the success examples.

5.3 Ablation Experiments

We ran seven additional experiments to study the importance of hyperparameters and design decisions. Appendix F provides full details and figures. These experiments highlight that RCE is not an imitation learning method: RCE fails when applied to full expert trajectories, which are typically harder to provide than success examples. Other ablation experiments underscore the importance of using n-step returns and validate the approximation made in Sec. 3.2.

6 Conclusion

In this paper, we proposed a data-driven approach to control, where examples of success states are used in place of a reward function. Our method estimates the probability of reaching a success example in the future and optimizes a policy to maximize this probability of success. Unlike prior imitation learning methods, our approach is "end-to-end" and does not require learning an auxiliary classifier or reward function. Our method is therefore simpler, with fewer hyperparameters and fewer lines of code to debug. Our analysis rests on a new data-driven Bellman equation, where example success states replace the typical reward function term. We use this Bellman equation to prove convergence of our classifier and policy.

Limitations and future work. Empirically, we observe that the classifier's predictions were not well calibrated but nonetheless produced effective policies. This issue resembles the miscalibration in Q-functions observed in prior work [11, 22]. In future work we aim to develop better off-policy evaluation techniques for example-based control to lift this limitation. We believe that formulating control problems in terms of data, rather than the reward-centric MDP, better captures the essence of many real-world control problems and suggests a new set of attractive learning algorithms.

References

[1] Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

[2] Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. (2017). Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pages 4055–4065.

[3] Calandra, R., Owens, A., Upadhyaya, M., Yuan, W., Lin, J., Adelson, E. H., and Levine, S. (2017). The feeling of success: Does touch sensing help predict grasp outcomes? In *Conference on Robot Learning*, pages 314–323. PMLR.

[4] Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624.

[5] Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220.

[6] Eysenbach, B., Salakhutdinov, R., and Levine, S. (2021). C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*.

[7] Fox, R., Pakman, A., and Tishby, N. (2016). Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 202–211.

[8] Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

[9] Fu, J., Luo, K., and Levine, S. (2018a). Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*.

[10] Fu, J., Singh, A., Ghosh, D., Yang, L., and Levine, S. (2018b). Variational inverse control with events: A general framework for data-driven reward definition. *Advances in neural information processing systems*, 31:8538–8547.

[11] Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR.

[12] Guadarrama, S., Korattikara, A., Ramirez, O., Castro, P., Holly, E., Fishman, S., Wang, K., Gonina, E., Wu, N., Kokiopoulou, E., Sbaiz, L., Smith, J., Bartók, G., Berent, J., Harris, C., Vanhoucke, V., and Brevdo, E. (2018). TF-Agents: A library for reinforcement learning in tensorflow. [Online; accessed 25-June-2019].

[13] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

[14] Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *NIPS*, pages 4565–4573.

[15] Irpan, A., Rao, K., Bousmalis, K., Harris, C., Ibarz, J., and Levine, S. (2019). Off-policy evaluation via off-policy classification. In *Advances in Neural Information Processing Systems*, pages 5437–5448.

[16] Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201.

[17] Konyushkova, K., Zolna, K., Aytar, Y., Novikov, A., Reed, S., Cabi, S., and de Freitas, N. (2020). Semi-supervised reward learning for offline reinforcement learning. *arXiv preprint arXiv:2012.06899*.

[18] Kostrikov, I., Agrawal, K. K., Dwibedi, D., Levine, S., and Tompson, J. (2018). Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations*.

[19] Kostrikov, I., Nachum, O., and Tompson, J. (2019). Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*.

[20] Kulkarni, T. D., Saeedi, A., Gautam, S., and Gershman, S. J. (2016). Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*.

[21] Laskin, M., Srinivas, A., and Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR.

[22] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *ICLR (Poster)*.

[23] Lu, Q., Van der Merwe, M., Sundaralingam, B., and Hermans, T. (2020). Multifingered grasp planning via inference in deep neural networks: Outperforming sampling by learning differentiable models. *IEEE Robotics Automation Magazine*, 27(2):55–65.

[24] Nasiriany, S. (2020). *DisCo RL: Distribution-Conditioned Reinforcement Learning for General-Purpose Policies*. PhD thesis.

[25] Pomerleau, D. A. (1989). Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313.

[26] Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. (2018). Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*.

[27] Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. (2006). Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736.

[28] Reddy, S., Dragan, A. D., and Levine, S. (2020). {SQIL}: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*.

[29] Ross, S., Gordon, G., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635.

[30] Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In *International conference on machine learning*, pages 1312–1320.

[31] Singh, A., Yang, L., Finn, C., and Levine, S. (2019). End-to-end robotic reinforcement learning without reward engineering. In *Robotics: Science and Systems*.

[32] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

[33] Sutton, R. S. (1995). Td models: Modeling the world at a mixture of time scales. In *Machine Learning Proceedings 1995*, pages 531–539. Elsevier.

[34] Vecerik, M., Sushkov, O., Barker, D., Rothörl, T., Hester, T., and Scholz, J. (2019). A practical approach to insertion with variable socket position using deep reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 754–760. IEEE.

[35] Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

[36] Wulfmeier, M., Ondruska, P., and Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

[37] Xie, A., Singh, A., Levine, S., and Finn, C. (2018). Few-shot goal inference for visuomotor learning and planning. In *Conference on Robot Learning*, pages 40–52. PMLR.

[38] Xu, D. and Denil, M. (2019). Positive-unlabeled reward learning. *arXiv preprint arXiv:1911.00459*.

[39] Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. (2020). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100. PMLR.

[40] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA.

[41] Zolna, K., Novikov, A., Konyushkova, K., Gulcehre, C., Wang, Z., Aytar, Y., Denil, M., de Freitas, N., and Reed, S. (2020). Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*.

[42] Zolna, K., Reed, S., Novikov, A., Colmenarej, S. G., Budden, D., Cabi, S., Denil, M., de Freitas, N., and Wang, Z. (2019). Task-relevant adversarial imitation learning. *arXiv preprint arXiv:1910.01077*.

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] See Sec. 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not foresee plausible negative impacts of this work.
   (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] See Appendix B.
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] Code is available at:
   `https://github.com/rce-anonymous/rce-anonymous.github.io/tree/main/code`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A and provided code.
   (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)? [No] Each experiment took approximately one day on a standard CPU server. The exact compute resources are proprietary.
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? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (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]