
Behavior Alignment via Reward Function Optimization

Dhawal Gupta*
University of Massachusetts

Yash Chandak*†
Stanford University

Scott M. Jordan†
University of Alberta

Philip S. Thomas
University of Massachusetts

Bruno Castro da Silva
University of Massachusetts

Abstract

Designing reward functions for efficiently guiding reinforcement learning (RL) agents toward specific behaviors is a complex task. This is challenging since it requires the identification of reward structures that are not sparse and that avoid inadvertently inducing undesirable behaviors. Naively modifying the reward structure to offer denser and more frequent feedback can lead to unintended outcomes and promote behaviors that are not aligned with the designer’s intended goal. Although potential-based reward shaping is often suggested as a remedy, we systematically investigate settings where deploying it often significantly impairs performance. To address these issues, we introduce a new framework that uses a bi-level objective to learn *behavior alignment reward functions*. These functions integrate auxiliary rewards reflecting a designer’s heuristics and domain knowledge with the environment’s primary rewards. Our approach automatically determines the most effective way to blend these types of feedback, thereby enhancing robustness against heuristic reward misspecification. Remarkably, it can also adapt an agent’s policy optimization process to mitigate suboptimalities resulting from limitations and biases inherent in the underlying RL algorithms. We evaluate our method’s efficacy on a diverse set of tasks, from small-scale experiments to high-dimensional control challenges. We investigate heuristic auxiliary rewards of varying quality—some of which are beneficial and others detrimental to the learning process. Our results show that our framework offers a robust and principled way to integrate designer-specified heuristics. It not only addresses key shortcomings of existing approaches but also consistently leads to high-performing solutions, even when given misaligned or poorly-specified auxiliary reward functions.

1 Introduction

In this paper, we investigate the challenge of enabling reinforcement learning (RL) practitioners, who may not be experts in the field, to incorporate domain knowledge through heuristic auxiliary reward functions. Our goal is to ensure that such auxiliary rewards not only induce behaviors that align with the designer’s intentions but also allow for faster learning. RL practitioners typically model a given control problem by first designing simple reward functions that directly quantify whether (or how well) an agent completed a task. These could be, for instance, functions assigning a reward of +1 iff the agent reaches a specified goal state, and zero otherwise. However, optimizing a policy based on such a sparse reward function often proves challenging.

*Both authors contributed equally to this work. † Work done while at the University of Massachusetts. Corresponding author: Dhawal Gupta (dgupta@cs.umass.edu).

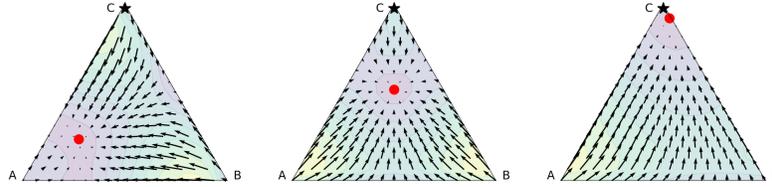


Figure 1: Auxiliary rewards can be used to convey to the agent how we (designers) *think* it should solve the problem. However, if not carefully designed, they can lead to policies that result in undesired behaviors. This figure provides a visual illustration of a toy example depicting how the proposed method works. The star represents the optimal policy, and the red dot represents the fixed point of a policy optimization process under a “sub-optimal” heuristic; i.e., one that, when naively combined with r_p , induces behaviors different from those under the optimal policy for r_p . **(Left)** Vector field of a policy optimization process converging to a sub-optimal policy. **(Middle and Right)** By changing the influence of auxiliary rewards, our method can dynamically *correct* the *entire policy optimization process* steering it towards a policy that results in the desired behavior.

To address this issue, designers often introduce auxiliary reward functions that supplement the original rewards. Auxiliary rewards are heuristic guidelines aimed at facilitating and speeding up the learning process. One could, e.g., augment the previously described reward function (which gives a reward of +1 upon reaching a goal state) with an auxiliary reward accounting for the agent’s distance to the goal. However, the effectiveness of using auxiliary reward functions largely depends on the problem’s complexity and the designer’s skill in crafting heuristics that, when combined with the original reward function, do not induce behaviors different than the ones originally intended [26, 27].

Existing methods like potential-based reward shaping [41] aim to incorporate domain knowledge without misaligning the behaviors induced by the resulting combined reward functions. However, as we discuss in Section 3, potential-based shaping has several limitations: (i) it is restricted to state-based functions; (ii) it amounts to a different initialization of the q -function; (iii) it does not alter policy gradients in expectation; and (iv) it can increase the variance in policy gradient methods.

To address these challenges, we introduce a scalable algorithm that empowers RL practitioners to specify potentially imperfect auxiliary reward functions. It ensures that the resulting optimization process will not inadvertently lead to unintended behaviors and that it will allow for faster learning. In particular, this paper addresses the following challenges:

- (1) How to incorporate auxiliary reward information:** We introduce a novel bi-level objective to analyze and automatically fine-tune designer-created auxiliary reward functions. It ensures they remain aligned with the original reward and do not induce behaviors different from those originally intended by the designer. Additionally, we formulate the problem to shape the optimization landscape, biasing our bi-level optimizer toward auxiliary reward functions that facilitate faster learning.
- (2) How to use auxiliary reward to mitigate algorithmic biases:** We show that our framework can automatically adjust how primary and auxiliary rewards are blended to mitigate limitations or biases inherent in the underlying RL algorithm (Section 4.1). For instance, many policy-gradient-based RL algorithms are subject to biases due to issues like discounting mismatch [59] or partial off-policy correction [51]. These biases can hinder the algorithm’s ability to identify near-optimal policies.
- (3) How to ensure scalability to high-dimensional problems:** We introduce an algorithm that employs *implicit gradients* to automatically adjust primary and auxiliary rewards, ensuring that the combined reward function aligns with the designer’s original expectations (see Figure 1). We evaluate our method’s efficacy across a range of tasks, from small-scale to high-dimensional control settings (see Section 6). In these tasks, we experiment with auxiliary rewards of varying quality; some accelerate learning, while others can be detrimental to finding an optimal policy.

2 Notation

In this paper, we investigate sequential decision-making problems modeled as Markov decision processes (MDPs). An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, p, r_p, r_{aux}, \gamma, d_0)$, where \mathcal{S} is the state set, \mathcal{A} is the action set, p is the transition function, $r_p: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the *primary* reward function, $r_{aux}: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the *auxiliary* reward function, γ is the discount factor, and d_0 is the initial state distribution.

$\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is an *optional* auxiliary reward function (possibly designed by a non-expert in machine learning, based on domain knowledge), and d_0 is the starting state distribution. Let $\pi_\theta : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ be any policy parameterized using $\theta \in \Theta$. For brevity, we will often use π_θ and θ interchangeably. Let S_t and A_t be the random variables for the state and action observed at the time t . As in the standard RL setting, the performance $J(\theta)$ of a policy π_θ is defined as the expected discounted return with respect to the (primary) reward function, r_p ; i.e., $J(\theta) := \mathbb{E}_\pi[\sum_{t=0}^T \gamma^t r_p(S_t, A_t)]$, where $T + 1$ is the episode length. An optimal policy parameter θ^* is defined as $\theta^* \in \arg \max_{\theta \in \Theta} J(\theta)$. A popular technique to search for θ^* is based on constructing sample estimates $\hat{\Delta}(\theta, r_p)$ of the (γ -dropped) policy gradient, $\Delta(\theta, r_p)$, given an agent's interactions with the environment for one episode [58, 59]. Then, using $\psi_\theta(s, a)$ as a shorthand for $d \ln \pi_\theta(s, a) / d\theta$, these quantities are defined as follows:

$$\Delta(\theta, r_p) = \mathbb{E}_{\pi_\theta} [\hat{\Delta}(\theta, r_p)] \quad \text{and} \quad \hat{\Delta}(\theta, r_p) := \sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j). \quad (1)$$

3 Limitations of Potential Based Reward Shaping

When the objective function $J(\theta)$ is defined with respect to a *sparse* reward function r_p (i.e., a reward function such that $r_p(s, a) := 0$ for most $s \in \mathcal{S}$ and $a \in \mathcal{A}$), searching for θ^* is challenging [24]. A natural way to provide more frequent (i.e., denser) feedback to the agent, in the hope of facilitating learning, is to consider an alternate reward function, $\tilde{r}_{\text{naive}} := r_p + r_{\text{aux}}$. However, as discussed earlier, r_{aux} may be a designer-specified auxiliary reward function not perfectly aligned with the objective encoded in r_p . In this case, using \tilde{r}_{naive} may encourage undesired behavior. An alternative way to incorporate domain knowledge to facilitate learning was introduced by Ng et al. [41]. They proposed using a *potential function*, $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ (analogous to r_{aux}), to define new reward functions of the form $\tilde{r}_\Phi(S_t, A_t, S_{t+1}) := r_p(S_t, A_t) + \gamma \Phi(S_{t+1}) - \Phi(S_t)$. Importantly, they showed that optimal policies with respect to the objective $\mathbb{E}[\sum_{t=0}^T \gamma^t \tilde{r}_\Phi(S_t, A_t, S_{t+1})]$ are also optimal with respect to $J(\theta)$.

While potential-based reward shaping can partially alleviate some of the difficulties arising from sparse rewards, Wiewiora [64] showed that q -learning using \tilde{r}_Φ produces the *exact same sequence of updates* as q -learning using r_p but with a different initialization of q -values. In what follows, we establish a similar result: we show that performing potential-based reward shaping has *no impact on expected policy gradient* updates; and that it can, in fact, even increase the variance of the updates.

Property 1. $\mathbb{E}[\hat{\Delta}(\theta, \tilde{r}_\Phi)] = \mathbb{E}[\hat{\Delta}(\theta, r_p)]$ and $\text{Var}(\hat{\Delta}(\theta, \tilde{r}_\Phi))$ can be higher than $\text{Var}(\hat{\Delta}(\theta, r_p))$.

All proofs are deferred to Appendix A. The above points highlight some of the limitations of potential-based shaping for policy gradients and q -learning—both of which form the backbone of the majority of model-free RL algorithms [58]. Furthermore, potential functions Φ *cannot* depend on actions [41], which restricts the class of eligible auxiliary rewards r_{aux} and heuristics functions that may be used. Finally, notice that Φ is designed independently of the agent's underlying learning algorithm. As we will show in the next sections, our method can autonomously discover auxiliary reward functions that not only facilitate learning but also help mitigate various types of algorithmic limitations and biases.

4 Behavior Alignment Reward Function

In this section, we introduce an objective function designed to tackle the primary challenge investigated in this paper: how to effectively leverage designer-specified auxiliary reward functions to rapidly induce behaviors envisioned by the designer. The key observation is that naively adding an auxiliary reward function r_{aux} to r_p may produce policies whose corresponding behaviors are misaligned with respect to the behaviors induced by r_p . In these cases, r_{aux} should be ignored during the search for an optimal policy. On the other hand, if r_p and r_{aux} may be combined in a way that results in the desired behaviors, then combinations that produce frequent and informative feedback to the agent should be favored, as they are likely to facilitate faster learning.

To tackle the challenges discussed above, we employ a bi-level optimization procedure. This approach aims to create a *behavior alignment reward* by combining r_{aux} and r_p using a parameterized function. Our method is inspired by the optimal rewards framework by Singh et al. [52, 53]. Let $\gamma_\varphi \in [0, 1)$ be

a discount rate value parameterized by $\varphi \in \Gamma$.² Let $r_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ be a *behavior alignment reward*: a function of both r_p and r_{aux} , parameterized by $\phi \in \Upsilon$, where Υ and Γ are function classes. One example of a behavior alignment reward function is $r_\phi(s, a) := f_{\phi_1}(s, a) + \phi_2 r_p(s, a) + \phi_3 r_{\text{aux}}(s, a)$, where $f_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\phi := (\phi_1, \phi_2, \phi_3)$. Let Alg be any gradient/semi-gradient/non-gradient-based algorithm that outputs policy parameters. To mitigate possible divergence issues arising from certain policy optimization algorithms like DQN [60, 1], we make the following simplifying assumption, which can generally be met with appropriate regularizers and step sizes:

Assumption 1. Given r_ϕ and γ_φ , the algorithm $\text{Alg}(r_\phi, \gamma_\varphi)$ converges to a fixed point θ — which we denote as $\theta(\phi, \varphi) \in \Theta$ to emphasize its indirect dependence on ϕ and φ through Alg , r_ϕ , and γ_φ .

Given this assumption, we now specify the following bi-level objective:

$$\phi^*, \varphi^* \in \arg \max_{\phi \in \Upsilon, \varphi \in \Gamma} J(\theta(\phi, \varphi)) - \lambda_\gamma \gamma_\varphi, \quad \text{where} \quad \theta(\phi, \varphi) := \text{Alg}(r_\phi, \gamma_\varphi). \quad (2)$$

Here, λ_γ serves as the regularization coefficient for the value of γ_φ , and Alg denotes a given policy optimization algorithm. Let, as an example, Alg be an on-policy gradient algorithm that uses samples to estimate the gradient $\Delta(\theta, r_p)$, as in (1). We can then define a corresponding variant of $\Delta(\theta, r_p)$ that is compatible with our formulation and objective, and which uses both r_ϕ and γ_φ , as follows:

$$\Delta_{\text{on}}(\theta, \phi, \varphi) := \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j, A_j) \right]. \quad (3)$$

Notice that the bi-level formulation in (2) is composed of three key components: **outer and inner objectives**, and an **outer regularization** term. In what follows, we discuss the need for these.

Need for Outer- and Inner-Level Objectives: The **outer-level objective** in Equation (2) serves a critical role: it evaluates different parameterizations, denoted by ϕ , for the behavior alignment reward function. These parameterizations influence the induced policy $\theta(\phi, \varphi)$, which is evaluated using the performance metric J . Recall that this metric quantifies the alignment of a policy with the designer’s primary reward function, r_p . In essence, the outer-level objective seeks to optimize the behavior alignment reward function to produce policies that are effective according to r_p . This design adds robustness against any misspecification of the auxiliary rewards.³ In the inner-level optimization, by contrast, Alg identifies a policy $\theta(\phi, \varphi)$ that is optimal or near-optimal with respect to r_ϕ (which combines r_{aux} through the behavior alignment reward). In the **inner-level optimization**, the algorithm Alg works to identify a policy $\theta(\phi, \varphi)$ that is optimal or near-optimal in terms of r_ϕ (which incorporates r_{aux} via the behavior alignment reward). By employing a bi-level optimization structure, several benefits emerge. When r_{aux} is well-crafted, r_ϕ can exploit its detailed information to give Alg frequent/dense reward feedback, thus aiding the search for an optimal θ^* . Conversely, if r_{aux} leads to sub-optimal policies, then the influence of auxiliary rewards can be modulated or decreased accordingly by the optimization process by adjusting r_ϕ . Consider, for example, a case where the behavior alignment reward function is defined as $r_\phi(s, a) := f_{\phi_1}(s, a) + \phi_2 r_p(s, a) + \phi_3 r_{\text{aux}}(s, a)$. In an adversarial setting—where the designer-specified auxiliary reward r_{aux} may lead to undesired behavior—the bi-level optimization process has the ability to set ϕ_3 to 0. This effectively allows the behavior alignment reward function r_ϕ to exclude r_{aux} from consideration. Such a bi-level approach to optimizing the parameters of behavior alignment reward functions can act as a safeguard against the emergence of sub-optimal behaviors due to a misaligned auxiliary reward, r_{aux} . This design is particularly valuable because it allows the objective in (2) to leverage the potentially dense reward structure of r_{aux} to provide frequent action evaluations when the auxiliary reward function is well-specified. At the same time, the approach maintains robustness against possible misalignments.

Need for Outer Regularization: The bi-level optimization problem (2) may have multiple optimal solutions for ϕ —including the trivial solution where r_{aux} is always ignored. The goal of regularizing the outer-level objective (in the form of the term $\lambda_\gamma \gamma_\varphi$) is to incorporate a prior that adds a preference for solutions, ϕ^* , that provide useful and frequent evaluative feedback to the underlying RL algorithm. In the next paragraphs, we discuss the need for such a regularizer and motivate its mathematical form. First, recall that *sparse* rewards can pose challenges for policy optimization. An intuitive solution

²Our framework can be generalized to support state-action dependent discount rates, γ .

³“Misspecification” indicates that an optimal policy for $r_p + r_{\text{aux}}$ may not be optimal for r_p alone.

to this problem could involve biasing the optimization process towards *denser* behavior alignment reward functions, e.g., by penalizing for sparsity of r_ϕ . Unfortunately, the distinction between sparse and dense rewards alone may not fully capture the nuances of what designers typically consider to be a “good” reward function. This is the case because *a reward function can be dense and still may not be informative*; e.g., a reward function that provides -1 to the agent in every non-goal state is dense but fails to provide useful feedback regarding how to reach a goal state. A better characterization of how useful (or informative) a reward function is may be constructed in terms of how *instructive* and *instantaneous* the evaluation or feedback it generates is. We consider a reward function to be *instructive* if it produces rewards that are well-aligned with the designer’s goals. A reward function is *instantaneous* if its corresponding rewards are dense, rather than sparse, and are more readily indicative of the optimal action at any given state.⁴ Reward functions that are both instructive and instantaneous can alleviate issues associated with settings with sparse rewards and long horizons. To bias our bi-level optimization objective towards this type of reward function, we introduce a regularizer, γ_φ . This regularizer favors solutions that can generate policies with high performance (i.e., high expected return J with respect to r_p) *even when the discount factor γ_φ is small*. To see why, first notice that this regularizer encourages behavior alignment reward functions that provide more instantaneous feedback to the agent. This has to be the case; otherwise, it would be challenging to maximize long-term reward should the optimized alignment reward function be sparse. Second, the regularizer promotes instructive alignment reward functions—i.e., functions that facilitate learning policies that maximize J . This is equally crucial: effective policies under the metric J are the ones that align well with the designer’s objectives as outlined in the original reward function, r_p .

4.1 Overcoming Imperfections of Policy Optimization Algorithms

The advantages of the bi-level formulation in (2) extend beyond robustness to sub-optimality from misspecified r_{aux} . Even with a well-specified r_{aux} , RL algorithms often face design choices, such as the bias-variance trade-off, that can induce sub-optimal solutions. Below we present examples to show how *bias* in the underlying RL algorithm may be mitigated by carefully optimizing r_ϕ and γ_φ .

4.1.1 Bias in policy gradients: Recall that the popular “policy gradient” $\Delta(\theta, r_p)$ is not, in fact, the gradient of any function, and using it in gradient methods may result in biased and sub-optimal policies [44]. However, policy gradient methods based on $\Delta(\theta, r_p)$ remain vastly popular in the RL literature since they tend to be sample efficient [59]. Let $\Delta_\gamma(\theta, r_p)$ denote the *unbiased* policy gradient, where $\Delta_\gamma(\theta, r_p) := \mathbb{E}[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j)]$. We can show that with a sufficiently expressive parameterization, optimized r_ϕ and γ_φ can effectively mimic the updates that would have resulted from using the *unbiased* gradient $\Delta_\gamma(\theta, r_p)$, even if the underlying RL algorithm uses the *biased* “gradient”, $\Delta_{\text{on}}(\theta, \phi, \varphi)$, as defined in (3). Detailed proofs are in Appendix A.

Property 2. *There exists $r_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\gamma_\varphi \in [0, 1)$ such that $\Delta_{\text{on}}(\theta, \phi, \varphi) = \Delta_\gamma(\theta, r_p)$.*

4.1.2 Off-policy learning without importance sampling: To increase sample efficiency when evaluating a given policy π_θ , it is often useful to use off-policy data collected by a different policy, β . Under the assumption that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \frac{\pi_\theta(s,a)}{\beta(s,a)} < \infty$, importance ratios $\rho_j := \prod_{k=0}^j \frac{\pi_\theta(s_k, a_k)}{\beta(s_k, a_k)}$ can be used to adjust the updates and account for the distribution shift between trajectories generated by β and π_θ . However, to avoid the high variance stemming from ρ_j , many methods tend to drop most of the importance ratios and thus only partially correct for the distribution shift—which can lead to bias [51]. We can show (given a sufficiently expressive parameterization for the behavior alignment reward function) that this type of bias can also be mitigated by carefully optimizing r_ϕ and γ_φ .

Let us denote the unbiased off-policy update with full-distribution correction as $\Delta_{\text{off}}(\theta, r_p) := \mathbb{E}_\beta[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \sum_{j=t}^T \rho_j \gamma^j r_p(S_j, A_j)]$. Now consider an extreme scenario where off-policy evaluation is attempted *without any correction for distribution shift*. In this situation, and with a slight abuse of notation, we define $\Delta_{\text{off}}(\theta, \phi, \varphi) := \mathbb{E}_\beta[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j, A_j)]$.

Property 3. *There exists $r_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\gamma_\varphi \in [0, 1)$ such that $\Delta_{\text{off}}(\theta, \phi, \varphi) = \Delta_{\text{off}}(\theta, r_p)$.*

Remark 1. *Our method is capable of mitigating various types of algorithmic biases and imperfections in underlying RL algorithms, without requiring any specialized learning rules. Additionally, thanks*

⁴E.g., if $r_\phi \approx q^*$, then its corresponding rewards are instantly indicative of the optimal action at any state.

to the γ_φ^* regularization, it favors reward functions that lead to faster learning of high-performing policies aligned with the designer’s objectives, as outlined in the original reward function r_p .

5 BARFI: Implicitly Learning Behavior Alignment Rewards

Having introduced our bi-level objective and discussed the benefits of optimizing r_ϕ and γ_φ , an important question arises: Although $\theta(\phi, \varphi)$ can be optimized using any policy learning algorithm, how can we efficiently identify the optimal ϕ^* and φ^* in equation (2)? Given the practical advantages of gradient-based methods, one would naturally consider using them for optimizing ϕ^* and φ^* as well. However, a key challenge in our setting lies in computing $dJ(\theta(\phi, \varphi))/d\phi$ and $dJ(\theta(\phi, \varphi))/d\varphi$. These computations require an analytical characterization of the impact that r_ϕ and γ_φ have on the *entire optimization process* of the inner-level algorithm, Alg.

In addressing this challenge, we initially focus on an Alg that employs policy gradients for updating π_θ . Similar extensions for other update rules can be derived similarly. We start by re-writing the expression for $dJ(\theta(\phi, \varphi))/d\phi$ using the chain rule:

$$\frac{dJ(\theta(\phi, \varphi))}{d\phi} = \underbrace{\frac{dJ(\theta(\phi, \varphi))}{d\theta(\phi, \varphi)}}_{(a)} \underbrace{\frac{d\theta(\phi, \varphi)}{d\phi}}_{(b)}, \quad (4)$$

where (a) is the policy gradient at $\theta(\phi, \varphi)$, and (b) can be computed via implicit bi-level optimization, as discussed below.

Implicit Bi-Level Optimization: We compute (4) by leveraging implicit gradients [14, 34, 19], an approach previously employed, e.g., in few-shot learning [38, 49] and model-based RL algorithms [50]. First, observe that when Alg converges to $\theta(\phi, \varphi)$, then it follows that

$$\Delta(\theta(\phi, \varphi), \phi, \varphi) = 0. \quad (5)$$

Let ∂f denote the partial derivative with respect to the immediate arguments of f , and df be the total derivative as before. That is, if $f(x, g(x)) := xg(x)$, then $\frac{\partial f}{\partial x}(x, g(x)) = g(x)$ and $\frac{df}{dx}(x, g(x)) = g(x) + x\frac{\partial g}{\partial x}(x)$. Therefore, taking the total derivative of (5) with respect to ϕ yields

$$\frac{d\Delta(\theta(\phi, \varphi), \phi, \varphi)}{d\phi} = \frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\phi} + \frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\theta(\phi, \varphi)} \frac{\partial\theta(\phi, \varphi)}{\partial\phi} = 0. \quad (6)$$

By re-arranging terms in (6), we obtain the term (b) in (4). In particular,

$$\frac{\partial\theta(\phi, \varphi)}{\partial\phi} = - \left(\frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\theta(\phi, \varphi)} \right)^{-1} \frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\phi}. \quad (7)$$

Furthermore, by combining (7) and (4) we obtain the desired gradient expression for ϕ :

$$\frac{dJ(\theta(\phi, \varphi))}{d\phi} = - \frac{\partial J(\theta(\phi, \varphi))}{\partial\theta(\phi, \varphi)} \underbrace{\left(\frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\theta(\phi, \varphi)} \right)^{-1}}_{\mathbf{H}} \underbrace{\frac{\partial\Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial\phi}}_{\mathbf{A}}. \quad (8)$$

Similarly, a gradient expression for φ can be derived; the full derivation is detailed in Appendix E. Using θ^* as shorthand for $\theta(\phi, \varphi)$, we find that the terms **A** and **H** can be expressed as

$$\mathbf{A} = \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \gamma_\varphi^{j-t} \frac{\partial r_\phi(S_j, A_j)}{\partial\phi} \right)^\top \right], \quad \mathbf{H} = \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \frac{\partial\psi_{\theta^*}(S_t, A_t)}{\partial\theta^*} \left(\sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j, A_j) \right) \right].$$

When working with the equations above, we assume the inverse of \mathbf{H}^{-1} exists. To mitigate the risk of ill-conditioning, we discuss regularization strategies for Alg in Appendix D. Notice that equations (8) and (15) are the key elements needed to calculate the updates to ϕ and φ in our bi-level optimization’s outer loop. However, computing **A** and **H** directly can be impractical for high-dimensional problems due to the need for outer products and second derivatives. To address this, we employ two strategies: (1) We approximate \mathbf{H}^{-1} using the Neumann series [40], and (2) we calculate (8) and (15) via Hessian-vector products [47], which are readily available in modern auto-diff libraries [46]. These methods eliminate the need for explicit storage or computation of **H** or **A**.

The mathematical approach outlined above results in an algorithm with linear compute and memory footprint, having $O(d)$ complexity, where d is the number of parameters for both the policy and the reward function. Details can be found in Appendix C. We refer to our method as BARFI, an acronym for *behavior alignment reward function's implicit* optimization.⁵ BARFI is designed to iteratively solve the bi-level optimization problem defined in (2). With policy regularization, the updates to r_ϕ and γ_φ incrementally modify $\theta(\phi, \varphi)$. This enables us to initialize Alg using the fixed point achieved in the previous inner optimization step, further reducing the time for subsequent inner optimizations.

6 Empirical Analyses

Our experiments serve multiple purposes and include detailed ablation studies. First, we demonstrate our bi-level objective's efficacy in discovering behavior alignment reward functions that facilitate learning high-performing policies. We focus especially on its robustness in situations where designers provide poorly specified or misaligned auxiliary rewards that could disrupt the learning process (Section 6.1). Second, we present a detailed analysis of the limitations of potential-based reward shaping, showing how it can lead to suboptimal policies (Section 6.2). We then provide a qualitative illustration of the behavior alignment reward function learned by BARFI (Section 6.3). Finally, we evaluate how well BARFI scales to problems with high-dimensional, continuous action and state spaces (Section 6.4).

In the sections that follow, we examine a range of methods and reward combinations for comparison:

- *Baseline RL methods:* We consider baseline RL methods that employ a naive reward combination strategy: $\tilde{r}_{\text{naive}}(s, a) := r_p(s, a) + r_{\text{aux}}(s, a)$. In this case, the auxiliary reward from the designer is simply added to the original reward without checks for alignment. Both the REINFORCE and Actor-Critic algorithms are used for optimization.
- *Potential-based shaping:* To assess how well potential-based reward shaping performs, we introduce variants of the baseline methods. Specifically, we investigate the effectiveness of the reward function $\tilde{r}_\Phi(s, a, s') := r_p(s, a) + \gamma r_{\text{aux}}(s') - r_{\text{aux}}(s)$.
- **BARFI:** We use REINFORCE as the underlying RL algorithm when implementing BARFI and define $r_\phi(s, a) := \phi_1(s, a) + \phi_2(s)r_p(s, a) + \phi_3(s)r_{\text{aux}}(s, a)$. Our implementation includes a warm-up period wherein the agent collects data for a fixed number of episodes, using \tilde{r}_{naive} , prior to performing the first updates to ϕ and φ (See Appendix 5 for the complete algorithm).

We evaluate each algorithm across four distinct environments: GridWorld, MountainCar [58], Cart-Pole [16], and HalfCheetah-v4 [9]. These domains offer increasing levels of complexity and are intended to assess the algorithms' adaptability. Furthermore, we examine their performance under a variety of auxiliary reward functions, ranging from well-aligned to misaligned with respect to the designer's intended objectives.

In our experiments, we investigate different types of auxiliary reward functions for each environment: some are action-dependent, while others are designed to reward actions aligned with either effective or ineffective known policies. These functions, therefore, vary in their potential to either foster rapid learning or inadvertently mislead the agent away from the designer's primary objectives, hindering the efficiency of the learning process. Comprehensive details of each environment and their corresponding auxiliary reward functions can be found in Appendix F.

6.1 BARFI's Robustness to Misaligned Auxiliary Reward Functions

In this section, we evaluate the performance of various methods for reward combination, particularly in scenarios where auxiliary reward functions can either be well-aligned with a designer's intended goals or be misaligned or poorly specified, thus inadvertently hindering efficient learning. We introduce two types of auxiliary reward functions for CartPole. First, we used domain knowledge to design an r_{aux} that provides bonuses when the agent's actions align with a known effective policy in this domain. Second, we designed an adversarial example where the auxiliary reward function rewards actions that are consistent with a particularly poorly performing policy. For MountainCar, we first leveraged knowledge about an *energy pumping policy* (i.e., a well-known effective policy [18]) to

⁵“BARFI” commonly refers to a type of south-Asian sweet confectionery, typically pronounced as ‘bur-fee’.

Table 1: Summary of the performance of various reward combination methods and types of r_{aux}

Method for Reward Combination	CartPole		MountainCar	
	Well-aligned r_{aux}	Misaligned r_{aux}	Well-aligned r_{aux} (w.r.t. <i>energy policy</i>)	Partially-aligned r_{aux} (w.r.t. <i>high velocity policy</i>)
BARFI (<i>our method</i>)	487.2 \pm 9.4	475.5 \pm 15.5	0.99 \pm 0.0	0.90 \pm 0.1
\tilde{r}_{naive} (<i>naive reward combination</i>)	498.9 \pm 1.0	9.04 \pm 0.2	0.99 \pm 0.0	0.63 \pm 0.1
\tilde{r}_{Φ} (<i>potential-based shaping</i>)	8.98 \pm 0.2	500 \pm 0.0	0.00 \pm 0.0	0.00 \pm 0.0

BARFI’s performance compared to two baselines that use: \tilde{r}_{naive} and \tilde{r}_{Φ} , respectively. CartPole uses an action-dependent r_{aux} function that either rewards agents when actions align with a known effective policy (*well-aligned* r_{aux}) or with a poorly-performing policy (*misaligned* r_{aux}). MountainCar uses either an action-dependent function aligned with an energy-pumping policy [18] or a partially-aligned function incentivizing higher velocities. BARFI consistently achieves near-optimal performance across scenarios, even if given poorly specified/misaligned auxiliary rewards. Competitors, by contrast, often induce suboptimal policies. Performances significantly below the optimal are shown in red, above.

craft an auxiliary reward function that provides bonuses for actions in line with such a control strategy. We also experimented with a partially-aligned auxiliary function that rewards high velocity—a factor not particularly indicative of high performance.

Table 1 summarizes the results across different reward functions and combination methods. The results suggest that if auxiliary rewards provide positive feedback when agents’ actions align with effective policies, naive combination methods perform well. In such cases, auxiliary rewards effectively “nudge” agents towards emulating expert actions. However, our experimental results also indicate that *all* baseline methods are susceptible to poor performance when auxiliary rewards are not well aligned with the designer’s goals. We provide more discussion for potential-based shaping in Section 6.2.

The key takeaway from the experimental results in Table 1 is that BARFI consistently performs well across various domains and under different types of auxiliary rewards. Specifically, when designer-specified feedback is appropriate and can assist in accelerating learning, BARFI efficiently exploits it to produce high-performing policies. Conversely, if auxiliary rewards are misaligned with the designer’s intended goals, BARFI is capable of adapting and effectively dismissing “misleading rewards”. This adaptability ensures that high-performing policies can be reliably identified. Other methods, by contrast, succeed only in some of these scenarios. Importantly, the unpredictability of whether a given auxiliary reward function will aid or hinder learning makes such alternative methods less reliable, as they may fail to learn effective policies.

6.2 Pitfalls of potential-based reward shaping

We now turn our attention to the (possibly negative) influence of action-dependent auxiliary rewards, particularly when used in combination with potential-based reward shaping. Our results in Table 1 reveal a key limitation: potential-based shaping struggles to learn efficient policies even when auxiliary rewards are well-aligned with effective strategies. This shortcoming is attributable to the action-dependent nature of the auxiliary rewards, which compromises the potential shaping technique’s guarantee of policy optimality.

As there is no prescribed way for designing potential shaping when r_{aux} is action-dependent, we use a direct extension of the original formulation [41] by considering $\tilde{r}_{\Phi}(s, a, s', a') := r_p(s, a) + \gamma r_{\text{aux}}(s', a') - r_{\text{aux}}(s, a)$. Furthermore, we designed an auxiliary reward function, r_{aux} , that is well aligned: it provides positive reward signals of fixed magnitude both for (s, a) and (s', a') whenever the agent’s actions coincide with the optimal policy. Notice, however, that if $\gamma < 1$, the resultant value from $\gamma r_{\text{aux}}(s', a') - r_{\text{aux}}(s, a)$ is *negative*. Such a negative component may deter the agent from selecting actions that are otherwise optimal, depending on how r_{aux} and r_p differ in magnitude. Conversely, potential-based shaping can also occasionally perform well under misaligned rewards. In these cases, the shaping function may yield a *positive* value whenever the agent selects an optimal action, which could induce well-performing behaviors.

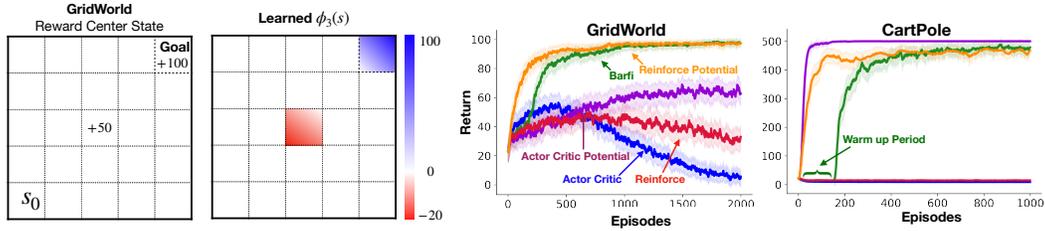


Figure 2: (Left) Misaligned r_{aux} (+50) for entering the center state and r_p (+100) at the goal state. (Center Left) Optimized weight $\phi_3(s)$ depicting how the learned reward function penalizes the agent for entering the center state. (Center Right) Performances in the GridWorld domain under a misspecified r_{aux} . (Right) Performances in the CartPole domain under a r_{aux} providing positive feedback for mimicking a known ineffective policy.

6.3 What does r_ϕ learn?

We now investigate BARFI’s performance and robustness in the GridWorld when operating under misspecified auxiliary reward functions. Consider the reward function depicted in Figure 2 [left]. This reward function provides the agent with a bonus for visiting the state at the center, akin to providing intermediate feedback to the agent when it makes progress towards the goal. However, such intermediate positive feedback can lead to behaviors where the agent repeatedly cycles around the middle state (i.e., behaviors that are misaligned with the original objective of reaching the goal state at the top right corner of the grid). Importantly, BARFI is capable of autonomously realizing that it should disregard such misleading incentives (Figure 2 [center left]), thereby avoiding poorly-performing behaviors that focus on revisiting irrelevant central states. Similarly, when Cartpole operates under a misspecified r_{aux} (Figure 2 [right]), BARFI is capable of rapidly adapting (after a warm-up period) and effectively disregarding misleading auxiliary reward signals. These results highlight once again BARFI’s robustness when faced with reward misspecification.

6.4 Scalability to High-Dimensional Continuous Control

One might wonder whether computing implicit gradients for ϕ and φ would be feasible in high-dimensional problems, due to the computational cost of inverting Hessians. To address this concern, we leverage Neumann series approximation with Hessian-vector products (See Appendix C) and conduct further experiments, as shown in Figure 3. These experiments focus on evaluating the scalability of BARFI in control problems with high-dimensional state spaces and continuous actions—scenarios that often rely on neural networks for both the policy and critic function approximators. For a more comprehensive evaluation, we also introduced an alternative method named BARFI unrolled. Unlike BARFI, which uses implicit bi-level optimization, BARFI unrolled employs path-wise bi-level optimization. It maintains a complete record of the optimization path to determine updates for ϕ and φ . Further details regarding this alternative method can be found in Appendix C.6.

We conducted experiments on the HalfCheetah-v4 domain and investigated, in particular, a reward function comprising two components with varying weights. This empirical analysis was designed to help us understand how different weight assignments to each reward component could influence the learning process. Specifically, in HalfCheetah-v4, the agent receives a positive reward r_p proportional to how much it moved forward. It also incurs a small negative reward (concretely, an auxiliary reward, $r_{\text{aux}}(s, a) := c\|a\|_2^2$, known as a *control cost*) for the torque applied to its joints. A hyperparameter c determines the balance between these rewards. The naive combination of such primary and auxiliary rewards is defined as $\tilde{r}_{\text{naive}}(s, a) = r_p(s, a) + r_{\text{aux}}(s, a)$. Figure 3 [left] shows that the baselines and both variants of BARFI appear to learn effectively. With alternative reward weighting schemes, however, only BARFI and BARFI unrolled show learning progress, as seen in Figure 3 [middle]. It is worth noting that path-wise bi-level optimization can become impractical as the number of update steps in (4) increases, due to growing computational and memory requirements (Figure 3 [right]). Although we do not recommend BARFI unrolled, we include its results for completeness. Additional ablation studies on (a) the effect of the inner optimization step; (b) Neumann approximations; (c) decay of γ ; and (d) returns based on r_ϕ , are provided in Appendix H.

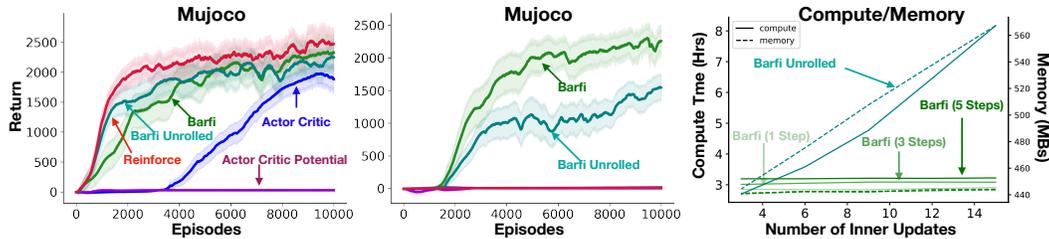


Figure 3: Results for MuJoCo environment. **(Left)** Auxiliary reward is defined to be $-c\|a\|_2^2$, where c is a positive hyperparameter and a is the continuous high-dimensional action vector. **(Middle)** Similar setting as before, but uses an amplified variant of the auxiliary reward: $-4c\|a\|_2^2$. It is worth highlighting that even under alternative reward weighting schemes, both variants of our (behavior-aligned) bi-level optimization methods demonstrate successful learning. Learning curves correspond to mean return over 15 trials, and the shaded regions correspond to one standard error. **(Right)** Required compute and memory for `BARFI_unrolled`, compared to `BARFI`, as a function of the number of inner-optimization updates. This figure also showcases `BARFI`'s characteristics under various orders of Neumann approximation.

7 Related work

This paper focuses primarily on how to efficiently leverage auxiliary rewards r_{aux} . Notice, however, that in the absence of r_{aux} , the resulting learned behavior alignment rewards r_ϕ may be interpreted as *intrinsic rewards* [70, 71]. Furthermore, several prior works have investigated meta-learning techniques, which are methods akin to the bi-level optimization procedures used in our work. Such prior works have employed meta-learning in various settings, including automatically inferring the effective return of trajectories [68, 62, 7, 71], parameters of potential functions [72, 28, 17], targets for TD learning [69], rewards for planning [54, 23], and even fully specified reinforcement learning update rules [33, 45]. Additionally, various other relevant considerations to effectively learning rewards online have been discussed by Armstrong et al. [5]. Our work complements these efforts by focusing on the reward alignment problem, specifically in settings where auxiliary information is available. An extended discussion on related works can be found in Appendix B. It is worth mentioning that among the above-mentioned techniques, most rely on path-wise meta-gradients. As discussed in Section 6.4, this approach can be disadvantageous as it often performs only one or a few inner-optimization steps, which limits its ability to fully characterize the result of the inner optimization [67]. Further, it requires caching intermediate steps, which increases computational and memory costs. `BARFI`, by contrast, exploits implicit gradients to alleviate these issues by directly characterizing the fixed point of `Alg` induced by learned behavior alignment rewards.

Finally, it is also important to highlight that a concurrent work on reward alignment using bi-level optimization was made publicly available after our manuscript was submitted for peer-reviewing at NeurIPS [10]. While our work analyses drawbacks of potential-based shaping and establishes different forms of correction that can be performed via bi-level optimization, this concurrent work provides complementary analyses on the convergence rates of bi-level optimization, as well as a discussion on its potential applications to Reinforcement Learning from Human Feedback (RLHF).

8 Conclusion and Future Work

In this paper, we introduced `BARFI`, a novel framework that empowers RL practitioners—who may not be experts in the field—to incorporate domain knowledge through heuristic auxiliary reward functions. Our framework allows for more expressive reward functions to be learned while ensuring they remain aligned with a designer's original intentions. `BARFI` can also identify reward functions that foster faster learning while mitigating various limitations and biases in underlying RL algorithms. We empirically show that `BARFI` is effective in training agents in sparse-reward scenarios where (possibly poorly-specified) auxiliary reward information is available. If the provided auxiliary rewards are determined to be misaligned with the designer's intended goals, `BARFI` autonomously adapts and effectively disincentivizes their use as needed. This adaptability results in a reliable pathway to identifying high-performing policies. The conceptual insights offered by this work provide RL practitioners with a structured way to design more robust and easy-to-optimize reward functions. We believe this will contribute to making RL more accessible to a broader audience.

Acknowledgement and Funding Disclosures

We thank Andy Barto for invaluable discussions and insightful feedback on an earlier version of this manuscript, which significantly improved the quality of our work.

This work is partially supported by the National Science Foundation under grant no. CCF-2018372 and by a gift from the Berkeley Existential Risk Initiative.

References

- [1] Joshua Achiam, Ethan Knight, and Pieter Abbeel. Towards characterizing divergence in deep q-learning. *arXiv preprint arXiv:1903.08894*, 2019.
- [2] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and Zico Kolter. Differentiable convex optimization layers. *arXiv preprint arXiv:1910.12430*, 2019.
- [3] Ron Amit, Ron Meir, and Kamil Ciosek. Discount factor as a regularizer in reinforcement learning. In *International Conference on Machine Learning*, 2020.
- [4] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, 2017.
- [5] Stuart Armstrong, Jan Leike, Laurent Orseau, and Shane Legg. Pitfalls of learning a reward function online. *arXiv preprint arXiv:2004.13654*, 2020.
- [6] Karl J Åström and Tore Hägglund. Pid control. *IEEE Control Systems Magazine*, 2006.
- [7] Sarah Bechtle, Artem Molchanov, Yevgen Chebotar, Edward Grefenstette, Ludovic Righetti, Gaurav Sukhatme, and Franziska Meier. Meta-learning via learned loss. *arXiv preprint arXiv:1906.05374*, 2019.
- [8] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 2000.
- [9] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- [10] Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Dinesh Manocha, Huazheng Wang, Furong Huang, and Mengdi Wang. Aligning agent policy with externalities: Reward design via bilevel rl. *arXiv preprint arXiv:2308.02585*, 2023.
- [11] Xu Chu Dennis Ding, Stephen L Smith, Calin Belta, and Daniela Rus. LTL control in uncertain environments with probabilistic satisfaction guarantees. *IFAC Proceedings Volumes*, 2011.
- [12] Chuong B Do, Chuan-Sheng Foo, and Andrew Y Ng. Efficient multiple hyperparameter learning for log-linear models. In *Advances in Neural Information Processing Systems*, 2007.
- [13] Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, 2012.
- [14] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.
- [15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In , *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017.
- [16] Razvan V Florian. Correct equations for the dynamics of the cart-pole system. *Center for Cognitive and Neural Studies (Coneural), Romania*, 2007.
- [17] Zhao-Yang Fu, De-Chuan Zhan, Xin-Chun Li, and Yi-Xing Lu. Automatic successive reinforcement learning with multiple auxiliary rewards. In *IJCAI*, 2019.

- [18] Sina Ghiassian, Banafsheh Rafiee, Yat Long Lo, and Adam White. Improving performance in reinforcement learning by breaking generalization in neural networks. In , *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, 2020.
- [19] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.
- [20] Edward Grefenstette, Brandon Amos, Denis Yarats, Artem Molchanov, Franziska Meier, and Kyunghyun Cho. higher: A pytorch meta-learning library. 2020.
- [21] Ricardo Grunitzki, Bruno C da Silva, and Ana LC Bazzan. A flexible approach for designing optimal reward functions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.
- [22] Ricardo Grunitzki, Bruno C da Silva, and LC Ana Bazzan. Towards designing optimal reward functions in multi-agent reinforcement learning problems. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.
- [23] Xiaoxiao Guo, Satinder Singh, Richard Lewis, and Honglak Lee. Deep learning for reward design to improve monte carlo tree search in atari games. *arXiv preprint arXiv:1604.07095*, 2016.
- [24] Joshua Hare. Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*, 2019.
- [25] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 2012.
- [26] Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. Teaching with rewards and punishments: Reinforcement or communication? In *CogSci*, 2015.
- [27] Mark K Ho, Fiery Cushman, Michael L Littman, and Joseph L Austerweil. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, 2019.
- [28] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 2020.
- [29] Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, 2018.
- [30] Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. Learning reward machines for partially observable reinforcement learning. *Advances in Neural Information Processing Systems*, 2019.
- [31] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *arXiv preprint arXiv:2010.03950*, 2020.
- [32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In , *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [33] Louis Kirsch, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Improving generalization in meta reinforcement learning using learned objectives. *arXiv preprint arXiv:1910.04098*, 2019.
- [34] Steven G Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012.
- [35] Hadas Kress-Gazit, Georgios E Fainekos, and George J Pappas. Temporal-logic-based reactive mission and motion planning. *IEEE Transactions on Robotics*, 2009.

- [36] Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 2013.
- [37] Jan Larsen, Lars Kai Hansen, Claus Svarer, and M Ohlsson. Design and regularization of neural networks: the optimal use of a validation set. In *Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, 1996.
- [38] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [39] Michael L Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. Environment-independent task specifications via GLTL. *arXiv preprint arXiv:1704.04341*, 2017.
- [40] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- [41] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, 1999.
- [42] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, 2018.
- [43] Scott Niekum, Andrew G Barto, and Lee Spector. Genetic programming for reward function search. *IEEE Transactions on Autonomous Mental Development*, 2010.
- [44] Chris Nota and Philip S. Thomas. Is the policy gradient a gradient? *arXiv preprint arXiv:1906.07073*, 2020.
- [45] Junhyuk Oh, Matteo Hessel, Wojciech M Czarnecki, Zhongwen Xu, Hado van Hasselt, Satinder Singh, and David Silver. Discovering reinforcement learning algorithms. *arXiv preprint arXiv:2007.08794*, 2020.
- [46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., 2019.
- [47] Barak A Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 1994.
- [48] Silviu Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.
- [49] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, 2019.
- [50] Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. *arXiv preprint arXiv:2004.07804*, 2020.
- [51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [52] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2009.
- [53] Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2010.
- [54] Jonathan Sorg, Richard L Lewis, and Satinder Singh. Reward design via online gradient ascent. *Advances in Neural Information Processing Systems*, 2010.

- [55] Jonathan Sorg, Satinder P Singh, and Richard L Lewis. Internal rewards mitigate agent boundedness. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010.
- [56] Jonathan Sorg, Satinder P Singh, and Richard L Lewis. Optimal rewards versus leaf-evaluation heuristics in planning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.
- [57] Jonathan Daniel Sorg. *The Optimal Reward Problem: Designing Effective Reward for Bounded Agents*. PhD thesis, University of Michigan, 2011.
- [58] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [59] Philip Thomas. Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*, 2014.
- [60] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997.
- [61] Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. *arXiv preprint arXiv:1912.02175*, 2019.
- [62] Yufei Wang, Qiwei Ye, and Tie-Yan Liu. Beyond exponentially discounted sum: Automatic learning of return function. *arXiv preprint arXiv:1905.11591*, 2019.
- [63] Martha White. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, 2017.
- [64] Eric Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 2003.
- [65] Eric M Wolff, Ufuk Topcu, and Richard M Murray. Robust control of uncertain markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012.
- [66] David H Wolpert and Kagan Tumer. Optimal reward functions in distributed reinforcement learning. In *Intelligent agent technology: Research and development*. World Scientific, 2001.
- [67] Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger Grosse. Understanding short-horizon bias in stochastic meta-optimization. *arXiv preprint arXiv:1803.02021*, 2018.
- [68] Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.
- [69] Zhongwen Xu, Hado P van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, and David Silver. Meta-gradient reinforcement learning with an objective discovered online. *Advances in Neural Information Processing Systems*, 2020.
- [70] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems*, 2018.
- [71] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado Van Hasselt, David Silver, and Satinder Singh. What can learned intrinsic rewards capture? In *International Conference on Machine Learning*, 2020.
- [72] Haosheng Zou, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*, 2019.

Behavior Alignment via Reward Function Optimization (Supplemental Material)

Table 2: Notations

Symbol	Description
θ	Parameters for policy π
ϕ	Parameters for reward function
φ	Parameters for learned γ
$\pi_\theta, r_\phi, \gamma_\varphi$	Functional form of policy, reward and γ with their respective parameters
$\alpha_\theta, \alpha_\phi, \alpha_\varphi$	Step sizes for the respective parameters
$\lambda_\theta, \lambda_\phi, \lambda_\varphi$	Regularization for policy, reward and γ function
δ	Number of on-policy samples collected between subsequent updates to ϕ, φ
η	Neumann Approximator Eigen value scaling factor
n	Number of loops used in Neumann Approximation
optim	Any standard optimizer like Adam, RMSprop, SGD, which takes input as gradients and outputs the appropriate update
E	Total Number of episodes to sample from the environment
N_i	Number of updates to be performed for updating the π by Alg
N_0	Number of initial updates to be performed
τ	Sample of a trajectory from a full episode

A Proofs for Theoretical Results

In this section, we provide proofs for Property 1, Property 2, and Property 3. For the purpose of these proofs, we introduce some additional notation. To have a unified MDP notation for goal-based and time-based tasks, we first consider that in the time-based task, time is a part of the state such that Markovian dynamics is ensured.

The (un-normalized) discounted and (un-normalized) undiscounted visitation probability is denoted as

$$d_\gamma^\pi(s, a) := \sum_{t=0}^T \gamma^t \Pr(S_t = s, A_t = a; \pi), \quad (9)$$

$$\bar{d}^\pi(s, a) := \sum_{t=0}^T \Pr(S_t = s, A_t = a; \pi). \quad (10)$$

We can normalize it so that it is a distribution as follows :

$$d^\pi(s, a) := \frac{\bar{d}^\pi(s, a)}{\sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \bar{d}^\pi(s', a')}.$$

Property 1. *The expected update performed by the biased policy gradient update is same when using the primary reward and reward modified with potential-based shaping, i.e., $\Delta(\theta, \tilde{r}) = \Delta(\theta, r_p)$. Further, the variance of the update when using potential-based reward shaping can be higher than the variance of the update performed using the primary reward, i.e., $\text{Var}(\hat{\Delta}(\theta, \tilde{r})) \geq \text{Var}(\hat{\Delta}(\theta, r_p))$.*

Proof. Part 1: Equality of the expected update

$$\begin{aligned}
 \Delta(\theta, \tilde{r}) &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} \tilde{r}(S_j, A_j) \right] \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \left(\sum_{j=t}^T \gamma^{j-t} (r_p(S_j, A_j) + \gamma \Phi(S_{j+1}) - \Phi(S_j)) \right) \right] \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j) \right] + \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} (\gamma \Phi(S_{j+1}) - \Phi(S_j)) \right] \\
 &= \Delta(\theta, r_p) + \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} (\gamma \Phi(S_{j+1}) - \Phi(S_j)) \right] \\
 &\stackrel{(a)}{=} \Delta(\theta, r_p) + \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) (\gamma^{T-t+1} \Phi(S_{T+1}) - \Phi(S_t)) \right] \\
 &\stackrel{(b)}{=} \Delta(\theta, r_p) + \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) (\gamma^{T-t+1} c - \Phi(S_t)) \right] \\
 &\stackrel{(c)}{=} \Delta(\theta, r_p) + \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T (\gamma^{T-t+1} c - \Phi(S_t)) \mathbb{E}_{\pi_\theta} [\psi_\theta(S_t, A_t) | S_t] \right] \\
 &\stackrel{(d)}{=} \Delta(\theta, r_p),
 \end{aligned}$$

where (a) holds because on the expansion of future return, intermediate potential values cancel out, (b) holds because S_{T+1} is the terminal state and potential function is defined to be a fixed constant c for any terminal state [41], (c) holds from the law of total expectation, and (d) holds because,

$$\mathbb{E}_{\pi_\theta} [\psi_\theta(S_t, A_t) | S_t] = \sum_{a \in \mathcal{A}} \pi_\theta(S_t, a) \frac{\partial \ln \pi_\theta(S_t, a)}{\partial \theta} = \sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta(S_t, a)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{a \in \mathcal{A}} \pi_\theta(S_t, a) = 0.$$

In the stochastic setting, i.e., when using sample average estimates instead of the true expectation, $\gamma^{T-t+1} c - \Phi(S_t)$ is analogous to a state-dependent baseline for the sum of discounted future primary rewards. It may reduce or increase the variance of $\Delta(\theta, r_p)$, depending on this baseline's co-variance with $\sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j)$.

Note: As we encountered the potential at the terminal state to be c , as it is a constant, we will use the value of $c = 0$ in accordance with [41].

Part 2: Variance characterization

For this result that discusses the possibility of the variance being higher when using potential-based reward shaping, we demonstrate the result using a simple example. We will consider the single-step case wherein an episode lasts for one time step. That is, the agent takes an action A_0 at the starting state S_0 and then transitions to the terminal state. Hence, the stochastic update $\hat{\Delta}(\theta, \tilde{r})$ can be written as:

$$\begin{aligned}
 \hat{\Delta}(\theta, \tilde{r}) &= \psi_\theta(S_0, A_0) \tilde{r}(S_0, A_0) \\
 &= \psi_\theta(S_0, A_0) (r_p(S_0, A_0) - \Phi(S_0)),
 \end{aligned}$$

wherein, we assume that Φ for terminal states is 0 and similarly, $\hat{\Delta}(\theta, r_p) = \psi_\theta(S_0, A_0) r_p(S_0, A_0)$.

For the purpose of this proof we will consider the case wherein we have a scalar θ , i.e., $\theta \in \mathbb{R}$, such that, $\psi_\theta(\cdot, \cdot) \in \mathbb{R}$.

Hence, $\text{Var}(\hat{\Delta}(\theta, \tilde{r}))$ can be written as:

$$\begin{aligned} \text{Var}(\hat{\Delta}(\theta, \tilde{r})) &= \mathbb{E}[\hat{\Delta}(\theta, \tilde{r})^2] - \mathbb{E}[\hat{\Delta}(\theta, \tilde{r})]^2 \\ &= \mathbb{E}[(\psi_\theta(S_0, A_0)(r_p(S_0, A_0) - \Phi(S_0)))^2] - \mathbb{E}[\psi_\theta(S_0, A_0)(r_p(S_0, A_0) - \Phi(S_0))]^2 \\ &\stackrel{(a)}{=} \mathbb{E}[\psi_\theta(S_0, A_0)^2(r_p(S_0, A_0)^2 + \Phi(S_0)^2 - 2\Phi(S_0)r_p(S_0, A_0))] - \mathbb{E}[\psi_\theta(S_0, A_0)(r_p(S_0, A_0))]^2 \\ &= \mathbb{E}[\psi_\theta(S_0, A_0)^2(\Phi(S_0)^2 - 2\Phi(S_0)r_p(S_0, A_0))] + \\ &\quad \underbrace{\mathbb{E}[\psi_\theta(S_0, A_0)^2 r_p(S_0, A_0)^2] - \mathbb{E}[\psi_\theta(S_0, A_0)(r_p(S_0, A_0))]^2}_{\text{Var}(\hat{\Delta}(\theta, r_p))}. \end{aligned}$$

Therefore,

$$\text{Var}(\hat{\Delta}(\theta, \tilde{r})) - \text{Var}(\hat{\Delta}(\theta, r_p)) = \mathbb{E}[\psi_\theta(S_0, A_0)^2(\Phi(S_0)^2 - 2\Phi(S_0)r_p(S_0, A_0))].$$

Subsequently, variance of $\hat{\Delta}(\theta, \tilde{r})$ will be higher than that of $\hat{\Delta}(\theta, r_p)$ if $\mathbb{E}[\psi_\theta(S_0, A_0)^2\Phi(S_0)^2] - 2\mathbb{E}[\psi_\theta(S_0, A_0)^2\Phi(S_0)r_p(S_0, A_0)] > 0$.

Example: Let us look at an example where the above condition can be true. Let us consider an MDP with a single state and a single-step horizon. In that case, we can consider the variance of the update to the policy at the said state, i.e.,

$$\text{Var}_\pi(\hat{\Delta}(\theta, \tilde{r})) - \text{Var}_\pi(\hat{\Delta}(\theta, r_p)) = \Phi(s)^2\mathbb{E}_\pi[\psi_\theta(s, A)^2] - 2\Phi(s)\mathbb{E}_\pi[\psi_\theta(s, A)^2(r_p(s, A))],$$

where s is the fixed state. Hence, the variance of the potential-based method might be more than the variance from using only the primary reward when

$$\begin{aligned} \Phi(s)^2\mathbb{E}_\pi[\psi_\theta(s, A)^2] - 2\Phi(s)\mathbb{E}_\pi[\psi_\theta(s, A)^2(r_p(s, A))] &> 0 \\ \Phi(s)^2\mathbb{E}_\pi[\psi_\theta(s, A)^2] &> 2\Phi(s)\mathbb{E}_\pi[\psi_\theta(s, A)^2(r_p(s, A))]. \end{aligned}$$

Further, let us consider the case where $\Phi(s) \neq 0$, because otherwise the variance of the update for those states would be same, and $\Phi(s) > 0$.

$$\begin{aligned} \Phi(s)^2\mathbb{E}_\pi[\psi_\theta(s, A)^2] &> 2\Phi(s)\mathbb{E}_\pi[\psi_\theta(s, A)^2(r_p(s, A))] \\ \Phi(s)\mathbb{E}_\pi[\psi_\theta(s, A)^2] &> 2\mathbb{E}_\pi[\psi_\theta(s, A)^2(r_p(s, A))]. \end{aligned}$$

We can see that the above condition can be satisfied by choosing a potential function that might be overly optimistic about the average reward of the state s , i.e. any $\Phi(s)$, s.t. $\Phi(s) > 2r_p(s, a) \forall a$ would lead to an increase in variance. A common place where this might be true is the use of an optimal value function (as hinted by [41]) as a baseline for a bad/mediocre policy initially. \square

Property 2. There exists $r_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\gamma_\phi \in [0, 1)$ such that $\Delta_{on}(\theta, \phi, \varphi) = \Delta_\gamma(\theta, r_p)$.

Proof. Recall the definition of $\Delta_\gamma(\theta, r_p)$ from Section 4.1:

$$\Delta_\gamma(\theta, r_p) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j) \right].$$

Using the law of total expectation,

$$\begin{aligned}
 \Delta_\gamma(\theta, r_p) &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \mathbb{E}_{\pi_\theta} \left[\sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j) \middle| S_t, A_t \right] \right] \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) q^{\pi_\theta}(S_t, A_t) \right] \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \sum_{t=0}^T \gamma^t \Pr(S_t = s, A_t = a; \pi_\theta) \psi_\theta(s, a) q^{\pi_\theta}(s, a) \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \psi_\theta(s, a) q^{\pi_\theta}(s, a) \sum_{t=0}^T \gamma^t \Pr(S_t = s, A_t = a; \pi_\theta) \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \psi_\theta(s, a) q^{\pi_\theta}(s, a) d_\gamma^{\pi_\theta}(s, a). \tag{11}
 \end{aligned}$$

Notice from (9) and (10) that for any (s, a) pair, if $d_\gamma^{\pi_\theta}(s, a) > 0$, then $\bar{d}^{\pi_\theta}(s, a) > 0$ since $\gamma \geq 0$. Therefore, dividing and multiplying by $\bar{d}^{\pi_\theta}(s, a)$ leads to:

$$\begin{aligned}
 \Delta_\gamma(\theta, r_p) &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \bar{d}^{\pi_\theta}(s, a) \psi_\theta(s, a) q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^{\pi_\theta}(s, a)} \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \sum_{t=0}^T \Pr(S_t = s, A_t = a; \pi_\theta) \psi_\theta(s, a) q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^{\pi_\theta}(s, a)} \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) q^{\pi_\theta}(S_t, A_t) \frac{d_\gamma^{\pi_\theta}(S_t, A_t)}{\bar{d}^{\pi_\theta}(S_t, A_t)} \right].
 \end{aligned}$$

Now, notice that if $\gamma_\phi = 0$ and $r_\phi(s, a) = q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^{\pi_\theta}(s, a)}$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$\Delta_{\text{on}}(\theta, \phi, \varphi) = \Delta_\gamma(\theta, r_p).$$

□

Property 3. There exists $r_\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\gamma_\phi \in [0, 1)$ such that $\Delta_{\text{off}}(\theta, \phi, \varphi) = \Delta_{\text{off}}(\theta, r_p)$.

Proof. This proof follows a similar technique as the proof for Property 2. Recall the definition of $\Delta_{\text{off}}(\theta, r_p)$:

$$\begin{aligned}
 \Delta_{\text{off}}(\theta, r_p) &:= \mathbb{E}_\beta \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} \rho_j r_p(S_j, A_j) \right] \\
 &:= \mathbb{E}_\beta \left[\sum_{t=0}^T \gamma^t \rho_t \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma^{j-t} \rho_{j-t} r_p(S_j, A_j) \right].
 \end{aligned}$$

Now using the law of total expectations,

$$\begin{aligned}
 \Delta_{\text{off}}(\theta, r_p) &= \mathbb{E}_\beta \left[\sum_{t=0}^T \gamma^t \rho_t \psi_\theta(S_t, A_t) \mathbb{E}_\beta \left[\sum_{j=t}^T \gamma^{j-t} \rho_{j-t} r_p(S_j, A_j) \middle| S_t, A_t \right] \right] \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) \mathbb{E}_{\pi_\theta} \left[\sum_{j=t}^T \gamma^{j-t} r_p(S_j, A_j) \middle| S_t, A_t \right] \right] \\
 &= \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t \psi_\theta(S_t, A_t) q^{\pi_\theta}(S_t, A_t) \right] \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \psi_\theta(s, a) q^{\pi_\theta}(s, a) d_\gamma^{\pi_\theta}(s, a),
 \end{aligned}$$

where the last line follows similar to (11). Now, notice that for any (s, a) pair, the assumption that $\pi_\theta(s, a)/\beta(s, a) < \infty$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, implies $d_\gamma^{\pi_\theta}(s, a)/d_\gamma^\beta(s, a) < \infty$. Further, if $d_\gamma^\beta(s, a) > 0$ it has to be that $d^\beta(s, a) > 0$ as well. Therefore, $d_\gamma^{\pi_\theta}(s, a)/d^\beta(s, a) < \infty$ as well. Multiplying and dividing by $d^\beta(s, a)$ results in:

$$\begin{aligned}
 \Delta_{\text{off}}(\theta, r_p) &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \bar{d}^\beta(s, a) \psi_\theta(s, a) q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^\beta(s, a)} \\
 &= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \sum_{t=0}^T \Pr(S_t = s, A_t = a; \beta) \psi_\theta(s, a) q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^\beta(s, a)} \\
 &= \mathbb{E}_\beta \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) q^{\pi_\theta}(S_t, A_t) \frac{d_\gamma^{\pi_\theta}(S_t, A_t)}{\bar{d}^\beta(S_t, A_t)} \right].
 \end{aligned}$$

Finally, notice that if $\gamma_\phi = 0$ and $r_\phi(s, a) = q^{\pi_\theta}(s, a) \frac{d_\gamma^{\pi_\theta}(s, a)}{\bar{d}^\beta(s, a)}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\Delta_{\text{off}}(\theta, \phi, \varphi) = \Delta_{\text{off}}(\theta, r_p).$$

□

Remark 2. Notice that as with any optimization problem, issues of realizability and identifiability of the desired r_ϕ must be taken into account. The examples provided in this section aim to highlight the capability of optimized behavior alignment reward functions. In particular, they not only improve and accelerate the learning process but are also capable of inducing updates capable of ‘fixing’ imperfections in the underlying RL algorithm.

B Extended Related Works

The bi-level objective draws inspiration from the seminal work of Singh et al. [52, 53] that provides an optimal-rewards framework for an agent. Prior works have built upon it to explore search techniques using evolutionary algorithms [43, 21], develop extensions for multi-agent setting [66, 22], and mitigate sub-optimality due to use of inaccurate models [55–57]. Our work also builds upon this direction and focuses on various aspects of leveraging auxiliary rewards r_{aux} , while staying robust against its misspecification.

Apart from specifying auxiliary rewards r_{aux} directly, other techniques for reward specification include linear temporal logic [35, 11, 65, 39] or reward machines [29–31] that allow exposing the reward functions as a white-box to the agent.

Recent works also explore γ that is state-action dependent [63, 48], or establishes connection between γ and value function regularization in TD learning Amit et al. [3]. These ideas are complementary to our proposed work and combining these with BARFI remains interesting directions for the future.

The concept of path-based meta-learning was initially popularized for few-shot task learning in supervised learning [15, 42]. Similar path-based approaches have been adopted in reinforcement

learning (RL) in various forms [28, 62, 69, 71]. Initially designed for stochastic gradient descent, these methods have been extended to other optimizers such as Adam [32] and RMSprop [25], treating them as differentiable counterparts [20].

C Algorithm

In this section we discuss the algorithm for the proposed method. As the proposed method does behavior alignment reward function's implicit optimization, we name it BARFI. Pseudo-code for BARFI is presented in Algorithm 5. We will first build on some preliminaries to understand the concepts.

C.1 Vector Jacobian Product

Let us assume that, $x \in \mathbb{R}^d, y \in \mathbb{R}^m, f(x, y) \in \mathbb{R}$. Then, we know that $\partial f(x, y)/\partial x \in \mathbb{R}^d, \partial f(x, y)/\partial y \in \mathbb{R}^m, \partial^2 f(x, y)/\partial y \partial x \in \mathbb{R}^{d \times m}$. Let us also assume that we have a vector $v \in \mathbb{R}^d$, and if we need to calculate the following, we can pull the derivative outside as shown:

$$\underbrace{\underbrace{v}_{\mathbb{R}^d} \underbrace{\frac{\partial^2 f(x, y)}{\partial y \partial x}}_{\mathbb{R}^{d \times m}}}_{\mathbb{R}^m} = \frac{\partial}{\partial y} \underbrace{\left\langle \underbrace{v}_{\mathbb{R}^d}, \underbrace{\frac{\partial f(x, y)}{\partial x}}_{\mathbb{R}^d} \right\rangle}_{\mathbb{R}^1}.$$

As we can see, the vector Jacobian product can be broken down into differentiating a vector product but shifting the place of multiplication, in which case we assume that the gradient passes through v w.r.t. y and hence we don't ever have to deal with large multiplications. Also note that the outer partial w.r.t. can easily be handled by `autodiff` packages. A pseudo-code is show in Algorithm 1.

Algorithm 1: Jacobian Vector Product

- 1 **Input:** $f(x, y) \in \mathbb{R}^1, x \in \mathbb{R}^d, y \in \mathbb{R}^m, v \in \mathbb{R}^d$
 - 2 $f' \leftarrow \text{grad}(f(x, y), x)$
 - 3 $\text{jvp} \leftarrow \text{grad}(f', y, \text{grad_outputs} = v)$
 - 4 **Return:** `jvp`
-

C.2 Neumann Series Approximation for Hessian Inverse

Recall, that for a given real number $\beta \in \mathbb{R}$, such that $0 \leq \beta < 1$, we know that the geometric series of this has a closed form solution, i.e.,

$$\begin{aligned} s &= 1 + \beta^1 + \beta^2 + \beta^3 + \dots + \\ &= \frac{1}{1 - \beta}. \end{aligned}$$

Similarly, given we have a value α such that $\beta = 1 - \alpha$, we can write α^{-1} as follows:

$$\begin{aligned} \frac{1}{1 - \beta} &= 1 + \beta + \beta^2 + \beta^3 + \dots + \\ \frac{1}{1 - (1 - \alpha)} &= 1 + (1 - \alpha) + (1 - \alpha)^2 + (1 - \alpha)^3 + \dots + \\ \alpha^{-1} &= 1 + (1 - \alpha) + (1 - \alpha)^2 + (1 - \alpha)^3 + \dots + \\ \alpha^{-1} &= \sum_{i=0}^{\infty} (1 - \alpha)^i. \end{aligned}$$

The same can be generalized for a matrix, i.e., given a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we can write \mathbf{A}^{-1} as follows:

$$\mathbf{A}^{-1} = \sum_{i=0}^{\infty} (\mathbf{I} - \mathbf{A})^i.$$

Note for the above to hold, matrix \mathbf{A} , where we represent $\text{eig}(\mathbf{A})$ as the eigenvalues of matrix \mathbf{A} , we should have the following condition to hold, $0 < \text{eig}(\mathbf{A}) < 1$. Note here we would regularize \mathbf{A} to ensure that all eigenvalues are positive, and then we can always scale the matrix \mathbf{A} , by its biggest eigenvalue to ensure that the above condition holds. Let say $\eta = 1/\max \text{eig}(\mathbf{A})$. Then we can write the following:

$$\begin{aligned}\mathbf{A}^{-1} &= \frac{\eta}{\eta} \mathbf{A}^{-1} \\ &= \eta(\eta\mathbf{A})^{-1} \\ &= \eta \sum_{i=0}^{\infty} (\mathbf{I} - \eta\mathbf{A})^i.\end{aligned}$$

As $\eta\mathbf{A}$ would always satisfy the above condition.

C.3 Neumann Approximation for Hessian Vector Product

Given we have seen how we can approximate the Inverse of a matrix without relying $O(d^3)$ operations, through Neumann approximation, lets look what needs to be done for our updates. Recall that the update ϕ, φ (8) and (15) were,

$$\frac{\partial J(\theta(\phi, \varphi))}{\partial \phi} = - \underbrace{\frac{\partial J(\theta(\phi, \varphi))}{\partial \theta(\phi, \varphi)}}_v \left(\underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)}}_{\mathbf{H}} \right)^{-1} \underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \phi}}_{\mathbf{A}}$$

and

$$\frac{\partial (J(\theta(\phi, \varphi)) - \frac{1}{2} \|\gamma_\varphi\|^2)}{\partial \varphi} = - \underbrace{\frac{\partial J(\theta(\phi, \varphi))}{\partial \theta(\phi, \varphi)}}_v \left(\underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)}}_{\mathbf{H}} \right)^{-1} \underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \varphi}}_{\mathbf{B}} - \frac{\partial \gamma_\varphi}{\partial \varphi}.$$

Let us look closely at the update for ϕ and we can generalize the updates easily for the case of φ .

$$\frac{\partial J(\theta(\phi, \varphi))}{\partial \phi} = -v\mathbf{H}^{-1}\mathbf{A}$$

We first look at how we can approximate the value of $v\mathbf{H}^{-1}$ efficiently, as we can always make use of the Jacobian Vector product later to get $(v\mathbf{H}^{-1})\mathbf{A}$, as $v\mathbf{H}^{-1}$ becomes a vector. Let us assume we wish to run the Neumann approximation up to n steps, i.e., we want to approximate \mathbf{H}^{-1} up to n order Neumann expansion,

$$\eta(\eta\mathbf{H}^{-1}) \approx \eta \sum_{i=0}^n (\mathbf{I} - \eta\mathbf{H})^i \tag{12}$$

Here we are assuming that the outer optimization for update (1) is for the function $J(\theta(\phi, \varphi))$ and the inner optimization which is represented by the update (3) is $f(\theta(\phi, \varphi), \phi, \varphi)$, i.e.,

$$\Delta(\theta, r_p) = \frac{\partial J(\theta(\phi, \varphi))}{\partial \theta}, \quad \Delta(\theta, \phi, \varphi) = \frac{\partial f(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta}.$$

The most common form in which $f(; B)$ is usually defined is the following:

$$f(\theta, \phi, \varphi; B) := \frac{1}{|B|} \sum_{\tau \in B} \left[\sum_{t=0}^T \log(\pi_\theta(S_t^\tau, A_t^\tau)) \sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j^\tau, A_j^\tau) \right].$$

Algorithm 5: BARFI: Behavior Alignment Reward Function’s Implicit optimization

```

1 Input:  $J, f, \alpha_\theta, \alpha_\phi, \alpha_\varphi, \eta, n, \delta, \text{optim}, E, N_i, N_0,$ 
2 Initialize:  $\pi_\theta, r_\phi, \gamma_\varphi$ 
3 Initialize:  $\text{optim}_\theta \leftarrow \text{optim}(\alpha_\theta), \text{optim}_\phi \leftarrow \text{optim}(\alpha_\phi), \text{optim}_\varphi \leftarrow \text{optim}(\alpha_\varphi)$ 
4  $\mathcal{D}_{\text{off}} \leftarrow []$ 
   # Collect a batch of data for warmup period
5 for  $e \in [1, N_0)$  do
6   | Generate  $\tau_e$  using  $\pi_\theta$ 
7   | Append  $\tau_e$  to  $\mathcal{D}_{\text{off}}$ 
   # Initial training steps using warmup data
8 for  $i \in [0, N_i + N_0)$  do
9   | Sample a batch of trajectories  $B$  from  $\mathcal{D}_{\text{off}}$ 
   | # Update policy
10  |  $\theta \leftarrow \theta + \text{optim}_\theta(\text{grad}(f(\theta, \phi, \varphi; B), \theta))$ 
   # Start reward alignment
11 for  $e \in [N_0, E)$  do
   # Collect a batch of on-policy data
12   $\mathcal{D}_{\text{on}} \leftarrow []$ 
13  for  $j \in [0, \delta)$  do
14    | Generate trajectory  $\tau_{e+j}$  using  $\pi_\theta$  and append in  $\mathcal{D}_{\text{on}}$ 
15   $e \leftarrow e + \delta$ 
   # Update  $r_\phi$  and  $\gamma_\varphi$ 
16   $\Delta_\phi \leftarrow \text{Algorithm 3}(\theta, \phi, \varphi, J, f, n, \eta, \mathcal{D}_{\text{off}}, \mathcal{D}_{\text{on}})$ 
17   $\Delta_\varphi \leftarrow \text{Algorithm 4}(\theta, \phi, \varphi, J, f, n, \eta, \mathcal{D}_{\text{off}}, \mathcal{D}_{\text{on}})$ 
18   $\phi \leftarrow \phi + \text{optim}_\phi(\Delta_\phi)$ 
19   $\varphi \leftarrow \varphi + \text{optim}_\varphi(\Delta_\varphi)$ 
20   $\mathcal{D}_{\text{off}} \leftarrow \mathcal{D}_{\text{off}} + \mathcal{D}_{\text{on}}$ 
   # Learn policy for new reward function, initializing from the last
21  for  $i \in [0, N_i)$  do
22    | Sample a batch of trajectories  $B$  from  $\mathcal{D}_{\text{off}}$ 
   | # Update policy
23    |  $\theta \leftarrow \theta + \text{optim}_\theta(\text{grad}(f(\theta, \phi, \varphi; B), \theta))$ 

```

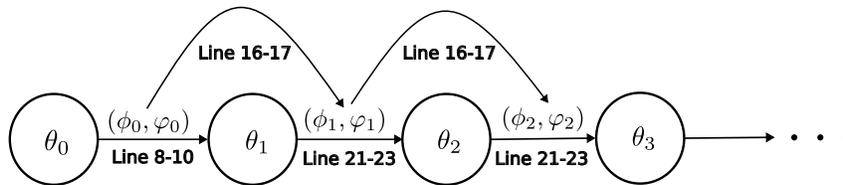


Figure 4: **Algorithm Flow:** The change in different parameters

In Lines 13–14, a new batch \mathcal{D}_{on} of data is sampled using the policy returned by the inner-optimization process. This data is used to compute $\partial J(\theta(\phi, \varphi)) / \partial \theta(\phi, \varphi)$. Existing data \mathcal{D}_{off} that was used in the inner-optimization process is then used to compute $\partial \theta(\phi, \varphi) / \partial \phi$ and $\partial \theta(\phi, \varphi) / \partial \varphi$. Using these in (8) and (15), the parameters for r_ϕ and γ_φ are updated in Lines 16 and 17, respectively.

Finally, the new data \mathcal{D}_{on} is merged into the existing data \mathcal{D}_{off} and the entire process continues.

C.5 Note on Approximation

An important limitation of the methods discussed above is that $\theta(\phi, \varphi)$ is considered such that $\Delta(\theta(\phi, \varphi), \phi, \varphi) = 0$, i.e., the Alg is run to convergence. In practice, we only execute Alg for a predetermined number of update steps that need not result in convergence to an optimum *exactly*. However, the impact of this approximation can be bounded by assuming convergence to an ϵ -neighborhood of the optima [49]. Furthermore, due to smoothness in the functional space, slight

changes to ϕ and φ should result in slight shifts in the optimum $\theta(\phi, \varphi)$. The continuity property allows for improvements in the optimization process: it suffices to initialize the parameters of each inner-loop optimization problem with the final parameters of the approximate fixed point solution, (ϕ, φ) , identified in the previous iteration of the inner loop. The complete resulting algorithm is presented in the appendix as Algorithm 5.

C.6 Path-wise Bi-level Optimization

An alternative approach for computing the term **(b)** in (4) is possible. The formulation of BARFI described above, based on implicit bi-level optimization, is agnostic to the optimization path taken by Alg. For the sake of completeness, let us also consider a version of BARFI that does take into account the path followed by the inner optimization loop. This is advantageous because it allows us to eliminate the need for the convergence criteria (5). We call this variant BARFI *unrolled*. The main difference, in this case, is that when computing the term **(b)** in (4), we now consider each inner update step until the point $\theta(\phi, \varphi)$ is reached—where the sequence of steps depends on the specific Alg used for the inner updates. Details are deferred to Appendix C. Notice that this approach results in a path-wise optimization process that can be more demanding in terms of computation and memory. We further discuss this issue, and demonstrate the efficacy of this alternative approach, in the empirical analyses section (Section ??).

D Smoothing the objective

To understand why $J(\theta(\phi, \varphi))$ might be ill-conditioned is to note that, often a small perturbation in the reward function doesn't necessarily lead to a change in the corresponding optimal policy. This can lead lack of gradient directions in the neighborhood of ϕ, φ for gradient methods to be effective. This issue can be addressed by employing common regularization techniques like L2 regularization of the policy parameters or entropy regularization for the policy⁶. We discuss two ways to regularize the objective in the upcoming sections.

D.1 L2 Regularization

To understand how severely ill-conditioned $J(\theta(\phi, \varphi))$ can be, notice that a small perturbation in the reward function often does not change the corresponding optimal policies or the outcome of a policy optimization algorithm Alg. Therefore, if the parameters of the behavior alignment reward are perturbed from ϕ to ϕ' , it may often be that $J(\theta(\phi, \varphi)) = J(\theta(\phi', \varphi))$ and this limits any gradient based optimization for ϕ as $\partial J(\theta(\phi, \varphi)) / \partial \phi$ is 0. Similarly, minor perturbations in φ may result in no change in $J(\theta(\phi, \varphi))$ either.

Fortunately, there exists a remarkably simple solution: incorporate regularization for the *policy parameters* θ in objective for Alg in the inner-level optimization. For example, the optimal policy for the following regularized objective $\mathbb{E}_{\pi_\theta} [\sum_{t=0}^T \gamma_\varphi^t r_\phi(S_t, A_t)] - \frac{\lambda}{2} \|\theta\|^2$ varies smoothly to trade-off between the regularization value of θ and the magnitude of the performance characterized by (r_ϕ, γ_φ) , which changes with the values of r_ϕ and γ_φ . See Figure 5 for an example with L2 regularization.

D.2 Entropy Regularized

In Section D.1, smoothing of $J(\theta(\phi, \varphi))$ was done by using L2 regularization on the policy parameters θ in the inner-optimization process. However, alternate regularization methods can also be used. For example, in the following we present an alternate update rule for θ based on entropy regularization,

$$\Delta(\theta, \phi, \varphi) := \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_\theta(S_t, A_t) \sum_{j=t}^T \gamma_\varphi^{j-t} (r_\phi(S_j, A_j) - \lambda \ln \pi_\theta(S_j, A_j)) \right].$$

⁶This regularization is performed so as to avoid a noninvertible Hessian as we had discussed in Section 5

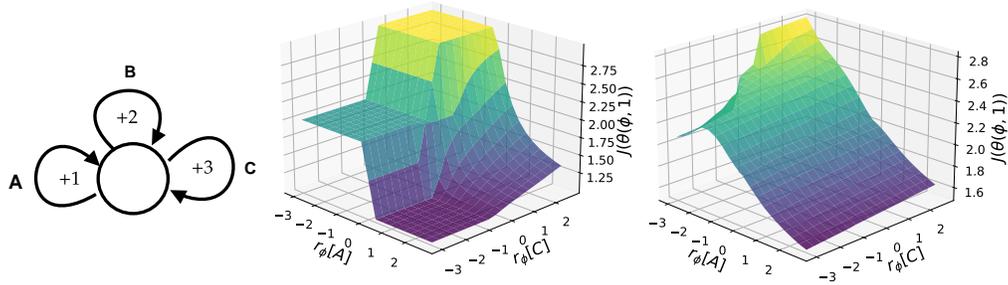


Figure 5: **(Left)** A bandit problem, where the data is collected from a policy β that samples action A mostly. **(Middle)** Each point on the 3D surface corresponds to the performance of $\theta(\phi, 1)$ returned by an Alg that uses the update rule $\Delta_{\text{off}}(\theta, \phi, 1)$ corresponding to the value of r_ϕ for actions A and C in the bottom axes; r_θ for action B is set to 0 to avoid another variable in a 3D plot. Notice that small perturbation in r_ϕ may lead to no or sudden changes in $J(\theta(\phi, 1))$. **(Right)** Performance of $\theta(\phi, 1)$ returned by an Alg that uses the update rule $\Delta_{\text{off}}(\theta, \phi, 1) - \theta$ that incorporates gradient of the L2 regularizer. Vector fields in Figure 1 were also obtained from this setup.

Notice that new update rule for ϕ and φ can be obtained from steps (4) to (15) with the following \mathbf{A} , \mathbf{B} , and \mathbf{H} instead, where for shorthand $\theta^* = \theta(\phi, \varphi)$,

$$\begin{aligned} \mathbf{A} &= \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j, A_j) \right)^\top \right], \\ \mathbf{B} &= \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \frac{\partial \gamma_\varphi^{j-t}}{\partial \varphi} (r_\phi(S_j, A_j) - \lambda \ln \pi_{\theta^*}(S_j, A_j)) \right) \right], \\ \mathbf{H} &= \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \frac{\partial \psi_{\theta^*}(S_t, A_t)}{\partial \theta^*} \left(\sum_{j=t}^T \gamma_\varphi^{j-t} (r_\phi(S_j, A_j) - \lambda \ln \pi_{\theta^*}(S_j, A_j)) \right) - \lambda \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \gamma_\varphi^{j-t} \psi_{\theta^*}(S_j, A_j)^\top \right) \right]. \end{aligned}$$

E Meta Learning via Implicit Gradient: Derivation

The general technique of implicit gradients [14, 34, 19] has been used in a vast range of applications, ranging from energy models [13, 36], differentiating through black-box solvers [61], few-shot learning [38, 49], model-based RL [50], differentiable convex optimization neural-networks layers [4, 2], to hyper-parameter optimization [37, 8, 12, 40]. In this work, we show how implicit gradients can also be useful to efficiently leverage auxiliary rewards r_a and overcome various sub-optimality.

Taking total derivative in (5) with respect to ϕ ,

$$\frac{d\Delta(\theta(\phi, \varphi), \phi, \varphi)}{d\phi} = \frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \phi} + \frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)} \frac{\partial \theta(\phi, \varphi)}{\partial \phi} = 0. \quad (13)$$

Let us try to understand why the above is true, considering the finite difference approach for this derivative,

$$\begin{aligned} \frac{d\Delta(\theta(\phi, \varphi), \phi, \varphi)}{d\phi} &= \lim_{\|d\phi\| \rightarrow 0} \frac{\Delta(\theta(\phi + d\phi, \varphi), \phi + d\phi, \varphi) - \Delta(\theta(\phi, \varphi), \phi, \varphi)}{d\phi} \\ &= \frac{0 - 0}{d\phi} = 0, \end{aligned}$$

$\Delta(\theta(\phi + d\phi, \varphi), \phi + d\phi, \varphi) = \Delta(\theta(\phi, \varphi), \phi, \varphi) = 0$, as $\theta(\cdot, \cdot)$ defines convergence to fixed point.

By re-arranging terms in (13) we obtain the term (b) in (4),

$$\frac{\partial \theta(\phi, \varphi)}{\partial \phi} = - \left(\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)} \right)^{-1} \frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \phi}. \quad (14)$$

By combining (14) with (4) we obtain the desired gradient expression for ϕ ,

$$\frac{\partial J(\theta(\phi, \varphi))}{\partial \phi} = -\frac{\partial J(\theta(\phi, \varphi))}{\partial \theta(\phi, \varphi)} \underbrace{\left(\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)} \right)^{-1}}_{\mathbf{H}} \underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \phi}}_{\mathbf{A}},$$

and following similar steps, it can be observed that the gradient expression for φ ,

$$\frac{\partial (J(\theta(\phi, \varphi)) - \frac{1}{2} \|\gamma_\varphi\|^2)}{\partial \varphi} = -\frac{\partial J(\theta(\phi, \varphi))}{\partial \theta(\phi, \varphi)} \underbrace{\left(\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \theta(\phi, \varphi)} \right)^{-1}}_{\mathbf{H}} \underbrace{\frac{\partial \Delta(\theta(\phi, \varphi), \phi, \varphi)}{\partial \varphi}}_{\mathbf{B}} - \frac{\partial \gamma_\varphi}{\partial \varphi},$$

where using θ^* as a shorthand for $\theta(\phi, \varphi)$ the terms \mathbf{A} , \mathbf{B} and \mathbf{H} can be expressed as

$$\mathbf{A} = \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \gamma_\varphi^{j-t} \frac{\partial r_\phi(S_j, A_j)}{\partial \phi} \right)^\top \right], \quad \mathbf{B} = \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \psi_{\theta^*}(S_t, A_t) \left(\sum_{j=t}^T \frac{\partial \gamma_\varphi^{j-t}}{\partial \varphi} r_\phi(S_j, A_j) \right) \right], \quad (15)$$

$$\mathbf{H} = \mathbb{E}_{\mathcal{D}} \left[\sum_{t=0}^T \frac{\partial \psi_{\theta^*}(S_t, A_t)}{\partial \theta^*} \left(\sum_{j=t}^T \gamma_\varphi^{j-t} r_\phi(S_j, A_j) \right) \right] - \lambda.$$

These provide the necessary expressions for updating ϕ and φ in the outer loop. As \mathbf{A} involves an outer product and \mathbf{H} involves second derivatives, computing them *exactly* might not be practical when dealing with high-dimensions. Standard approximation techniques like conjugate-gradients or Neumann series can thus be used to make it more tractable [40]. In our experiments, we made use of the Neumann approximation to the Hessian Inverse vector product ($\mathbf{A}\mathbf{H}^{-1}$), which requires the same magnitude of resources as the baseline policy gradient methods that we build on top off.

Algorithm: Being based on implicit gradients, we call our method BARFI, shorthand for *behavior alignment reward function's implicit optimization*. Overall, BARFI iteratively solves the bi-level optimization specified in (2) by alternating between using (3) till approximate converge of `Alg` to $\theta(\phi, \varphi)$ and then updating r_ϕ and γ_φ . Importantly, being based on (3) for sample efficiency, `Alg` leverages only the past samples and does *not* sample any new trajectories for the inner level optimization. Further, due to policy regularization which smoothens the objective as discussed in D, updates in r_ϕ and γ_φ changes the policy resulting from `Alg` gradually. Therefore, for compute efficiency, we start `Alg` from the policy obtained from the previous inner optimization, such that it is in proximity of the new fixed point. This allows BARFI to be both sample and compute efficient while solving the bi-level optimization iteratively online. Pseudo-code for BARFI and more details on the approximation techniques can be found in Appendix C.

F Environment & Reward Details

The first environment is a **GridWorld** (GW), where the start state is in the bottom left corner and a goal state is in the top right corner. The agent receives an r_p of +100 on reaching the goal followed by termination of the episode. The second environment is **MountainCar** (MC) [58], wherein we make use of the sparse reward variant, wherein the agent receives a +1 reward on reaching on top of the hill and 0 otherwise. The third environment is **CartPole** (CP) [16]. Finally, to assess the scalability we pick HalfCheetah-v4 from **Mujoco** (MJ) suite of OpenAI Gym [9].

For each environment, we define two auxiliary reward functions. For GridWorld, we define the functions: $r_{\text{aux, GW}}^1 := -(s - s_{\text{goal}})^2$, which provides the negative L2 squared distance from the goal position, and $r_{\text{aux, GW}}^2 := 50 \times \mathbf{1}_{s \in S_{\text{center}}}$, which provides an additional bonus of +50 to the agent along the desired path to the goal state (i.e. the center states). In MountainCar the state is composed of two components: the position x , and velocity v . The first auxiliary reward function, $r_{\text{aux, MC}}^1(s, a) := |v|$, encourages a higher absolute velocity of the car, and the second, $r_{\text{aux, MC}}^2(s, a) := \mathbf{1}_{\text{sign}(v) = a}$, encourages the direction of motion to increase the magnitude of the velocity (also known as the *energy pumping policy* [18]). For CartPole, we consider a way to reuse knowledge from a hand crafted policy. CartPole can be solved using a Proportional Derivate (PD) controller [6], hence we tune a PD controller, $\text{PD}^* : \mathcal{S} \rightarrow \mathcal{A}$, to solve CartPole for the max possible return. We design two auxiliary

reward functions which make use of this PD controller. The first, $r_{\text{aux,CP}}^1(s, a) := 5 \times \mathbf{1}_{\text{PD}^*(s)=a} - (1 - \mathbf{1}_{\text{PD}^*(s)=a})$, encourages the agent to match the action of the optimal PD controller, and penalizes it for not matching. The second auxiliary reward function, $r_{\text{aux,GW}}^1(s, a) := -r_{\text{aux,CP}}^1(s, a)$, encourages the agent to do the opposite. In the case of Mujoco, the reward function provided by the environment is itself composed of multiple different functions. We explain the same and the respective auxiliary functions for this case later.

We have considered several forms of information encoded as auxiliary rewards for these experiments. We have heuristic-based reward functions (i.e., $r_{\text{aux,GW}}^1, r_{\text{aux,GW}}^1, r_{\text{aux,MC}}^1$). Reward functions that encode a guess of an optimal policy (i.e., $r_{\text{aux,MC}}^1, r_{\text{aux,CP}}^1$) and reward functions that change the optimal policy (i.e., $r_{\text{aux,GW}}^1, r_{\text{aux,CP}}^1$). We also have rewards that only depend on states (i.e., $r_{\text{aux,GW}}^1, r_{\text{aux,GW}}^1, r_{\text{aux,MC}}^1$) as well as ones that depend on both state and actions (i.e., $r_{\text{aux,MC}}^1, r_{\text{aux,CP}}^1, r_{\text{aux,CP}}^1$). Therefore, we can test if BARFI can overcome misspecified auxiliary reward functions and does not hurt performance when well-specified.

Mujoco Environment In this experiment, we investigate the scalability of BARFI in learning control policies for high-dimensional state spaces with continuous action spaces. In HalfCheetah-v4 the agent’s task is to move forward, and it receives a reward based on its forward movement (denoted as r_p). Additionally, there is a small cost associated with the magnitude of torque required for action execution (denoted as $r_{\text{aux}}(s, a) := c|a|_2^2$). The weighting between the main reward and the control cost is pre-defined as c for this environment, and we form the reward as $\tilde{r}(s, a) = r_p(s, a) + r_{\text{aux}}(s, a)$. We explore how an arbitrary weighting choice can cause the agent to fail in learning, while BARFI is still able to adapt and learn the appropriate weighting, remaining robust to possible misspecification. We consider two different weightings for the control cost: the first weighting, denoted as $r_{\text{aux,MJ}}^1(s, a) := -c|a|_2^2$, uses the default setting, while the second weighting, denoted as $r_{\text{aux,MJ}}^1(s, a) := -4c|a|_2^2$, employs a scaled variant of the first weighting. Additionally, we implement the path-wise bi-level optimization variant i.e., BARFI Unrolled. In these experiments, we keep the value of γ fixed to isolate the agent’s capability to adapt and recover from an arbitrary reward weighting. We will also measure what computational and performance tradeoffs we might have to make between using the implicit version i.e., BARFI against, the path-wise version i.e., BARFI Unrolled.

G Details for the Empirical Results

G.1 Implementation Details

In this section we will briefly describe the implementation details around the different environments that were used.

GridWorld (GW): In the case of GridWorld we made use of the Fourier basis (of Order = 3) over the raw coordinates of agent position in the GridWorld. Details about this could be found in the `src/Utils/Basis.py` file.

MountainCar (MC): For this environment, to reduce the limitation because of the function approximator we used TileCoding [58], which offers a suitable representation for the MountainCar problem. We used 4 Tiles and Tilings of 5.

CartPole (CP): For CartPole also make use of Fourier Basis of (Order = 3), with linear function approximator on top of that.

MuJoco (MJ): For this we made use of a neural network with 1 hidden layer of 32 nodes and ReLU activation as the function approximator over the raw observations. The output of the policy is continuous actions, hence we used a Gaussian representation, where the policy outputs the mean of the multivariate Gaussian and we used a fixed diagonal standard deviation, fixed to $\sigma = 0.1$.

General Details: All the outer returns are evaluated without any discounting, whereas all the inner optimizations were initialized with $\gamma_\varphi = 0.99$. Hence to do this we made φ a single bias unit, initialized to 4.6, and passed through a sigmoid (i.e., $\sigma(4.6) = 0.99$).

For GW, CP and MC r_ϕ is defined as below

$$r_\phi(s, a) = \phi_1(s) + \phi_2(s)r_p + \phi_3(s)r_a$$

Wherein ϕ_1, ϕ_2, ϕ_3 are scalar outputs of a 3-headed function, in this case simply a linear layer over the states inputs.

Whereas in the case of MJ, we have

$$r_\phi(s, a) = \phi_1 + r_p + \phi_3 r_a$$

Wherein ϕ_1 is initialized to zero and ϕ_3 is 1.0 act like bias units.

Gradient normalization was used for all the cases where neural nets were involved (i.e., MJ), and also for MJ we modified the Baseline (REINFORCE) update to subtract the running average of the performance as a baseline to get acceptable performance for the baseline method.

G.2 Hyper-parameter Selection

As different make use of different function approximators the range of hyper-params can vary we talk about all the above over here.

Best-performing Parameters for different methods and environments are listed where

Table 3: Hyper-parameters for GridWorld

Hyper Parameter	BARFI Value	REINFORCE Value	Actor-Critic Value
α_θ	1×10^{-3}	1×10^{-3}	1×10^{-3}
α_ϕ	5×10^{-3}	—	—
α_φ	5×10^{-3}	—	—
optim	RMSprop	RMSprop	RMSprop
λ_θ	0.25	0.25	0.25
λ_ϕ	0.0625	—	—
λ_φ	4.0	—	—
Buffer	1000	—	—
Batch Size	1	1	1
η	0.0005	—	—
δ	3	—	—
n	5	—	—
N_0	150	—	—
N_i	15	—	—

Table 4: Hyper-parameters for MountainCar

Hyper Parameter	BARFI Value	REINFORCE Value	Actor-Critic Value
α_θ	0.015625	0.125	0.03125
α_ϕ	0.0625	—	—
α_φ	0.0625	—	—
optim	RMSprop	RMSprop	RMSprop
λ_θ	0.0	0.0	0.25
λ_ϕ	0.0	—	—
λ_φ	0.25	—	—
Buffer	50	—	—
Batch Size	1	1	1
η	0.001	—	—
δ	3	—	—
n	5	—	—
N_0	50	—	—
N_i	15	—	—

Hyperparameter Sweep : Here we list the details about how we swept the values for different hyper-params. We used PyTorch [46] for all our implementations. We usually used an optimizer between RMSProp or Adam with default parameters as provided in Pytorch. For $\alpha_\theta \in \{5 \times$

Table 5: Hyper-parameters for CartPole

Hyper Parameter	BARFI Value	REINFORCE Value	Actor-Critic Value
α_θ	1×10^{-3}	1×10^{-3}	5×10^{-4}
α_ϕ	1×10^{-3}	—	—
α_φ	5×10^{-3}	—	—
optim	RMSprop	RMSprop	RMSprop
λ_θ	1.0	1.0	0.0
λ_ϕ	0.0	—	—
λ_φ	4.0	—	—
Buffer	10000	—	—
Batch Size	1	1	1
η	0.0005	—	—
δ	3	—	—
n	5	—	—
N_0	150	—	—
N_i	15	—	—

Table 6: Hyper-parameters for MuJoCo

Hyper Parameter	BARFI Value	REINFORCE Value	Actor-Critic Value
α_θ	7.5×10^{-5}	5×10^{-4}	2.5×10^{-4}
α_ϕ	2.5×10^{-3}	—	—
α_φ	0.0	—	—
optim	Adam	Adam	Adam
λ_ϕ	0.0625	—	—
λ_φ	0.0	—	—
Buffer	50	—	—
Batch Size	1	1	1
η	0.0005	—	—
δ	3	—	—
n	5	—	—
N_0	30	—	—
N_i	15	—	—

$10^{-3}, 2.5 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}, 1 \times 10^{-4}, 7.5 \times 10^{-5}$ }, we use similar ranges for $\alpha_\phi, \alpha_\varphi$ (which tend to be larger). For λ_θ and λ_ϕ , we swept from $[0, 0.25, 0.5, 1.0]$ and for λ_γ we swept from $[0, 0.25, 1.0, 4.0, 16.0]$. We simply list ranges for different values and later we present sensitivity curves showing that these values are usually robust for BARFI across different methods as we can see from the tables above. $\delta \in [1, 3, 5]$, $n \in [1, 3, 5]$, $N_i \in [1, 3, 6, 9, 12, 15]$, $\eta \in [1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}]$, $N_0 \in [30, 50, 100, 150]$, $\text{buffer} \in [25, 50, 100, 1000]$. α for Tilecoding was adopted from [58] and hence similar ranges were swept in that case. Most sweeps were done with around 10 seeds, and later the parameter ranges were reduced and performed with more seeds.

G.3 Compute

The computer is used for a cluster where the CPU class is Intel Xeon Gold 6240 CPU @2.60GHz. The total compute required for GW was around 3 CPU years⁷, CP also required around 3 CPU years, and MC required around 4 CPU years. For MJ we needed around 5-6 CPU years. In total we utilized around 15-16 CPU years, where we needed around 1 GB of memory per thread.

⁷1 CPU year := Compute equal to running a CPU thread for a year.

H Extra Results & Ablations

H.1 Experiment on partially misspecified r_{aux}

In these set of experiments, we consider the case where auxiliary reward signals could be helpful only in a few—possibly arbitrary—state-action pairs. In general, we anticipate that solutions in this scenario would be such that assigns weightings, allowing the agent to avoid regions where might be misspecified. Meanwhile, the agent would still make use of the places where is well specified and useful.

We consider another r_{aux} in the GridWorld domain in which the auxiliary reward is misspecified for a subset of states near the starting position. Meanwhile, it is still well-specified for states near the goal (Figure 6 (a)). Figures 6 (b) and (c) illustrate the learned and the weighting on, showcasing the agent’s ability to depict the expected behavior described above.

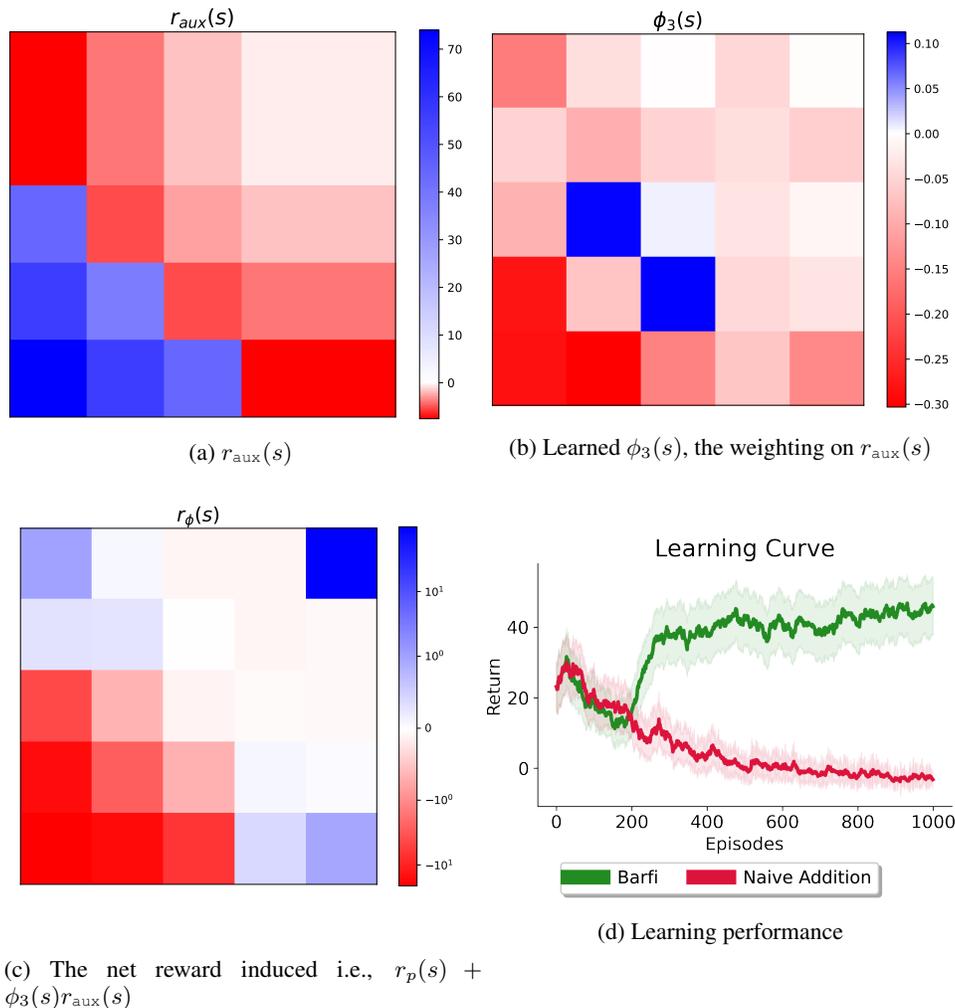


Figure 6: 40 random seeds were used to generate the plots. The starting state is at the bottom left and the goal state is at the top right corner. The primary reward r_p is +100 when the agent reaches the goal and 0 otherwise. (a) A state-dependent r_{aux} function that is partially misspecified (in the blue region r_{aux} provides a value equal to the **Manhattan distance**, thereby incentivizing the agent to stay near the start), and partially well specified (in the red region, it is the **negative Manhattan distance** and encourages movement towards the goal). (b) The state-dependent weighting $\phi_3(s)$ learned by BARFI negates the positive value from r_{aux} near the start state. (c) The effective reward function $r_p(s) + \phi_3(s)r_{\text{aux}}(s)$ learned by BARFI. (d) Learning curves for BARFI, and the baseline that uses a naive addition ($r_p(s) + r_{\text{aux}}(s)$) of the above auxiliary reward.

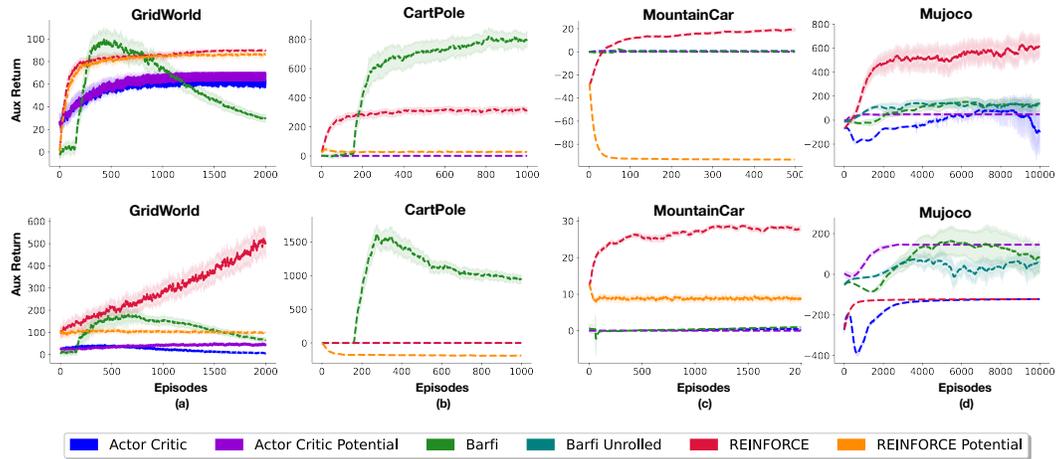


Figure 7: **Return induced by learned reward functions:** This figure illustrates the aux return collected by agent based on the learned r_ϕ , the curves are chosen based on best-performing curves on r_p , and averaged over 20 runs (except 40 for GW). **(a) Top** $- r_{\text{aux,GW}}^1$, **Bottom** $- r_{\text{aux,GW}}^2$, **(b) Top** $- r_{\text{aux,CP}}^1$, **Bottom** $- r_{\text{aux,CP}}^2$, **(c) Top** $- r_{\text{aux,MC}}^2$, **Bottom** $- r_{\text{aux,MC}}^1$, **(d) Top** $- r_{\text{aux,MJ}}^1$, **Bottom** $- r_{\text{aux,MJ}}^2$.

H.2 Return based on learned r_ϕ and γ_ϕ

Figure 7 and Figure 8 summarize the achievable return based on r_ϕ and the γ learned by the agent across different domains and reward specification. We observe that REINFORCE often optimizes the naive combination of reward for sure, but that doesn't really lead to a good performance on r_p , whereas BARFI does achieve an appropriate return on r_ϕ , but is also able to successively decay γ as the learning progress across different domains. Particularly notice Figure 7 (a) Bottom, where REINFORCE does optimize aux return a lot, but actually fails to solve the problem, as it simply learns to loop around the center state.

H.3 Ablations

Figure 9 represents the ablation of BARFI on GridWorld with the misspecified reward for its different params. We can see that usually having $\eta = 0.001, 0.0005, n = 5$ works for the approximation.

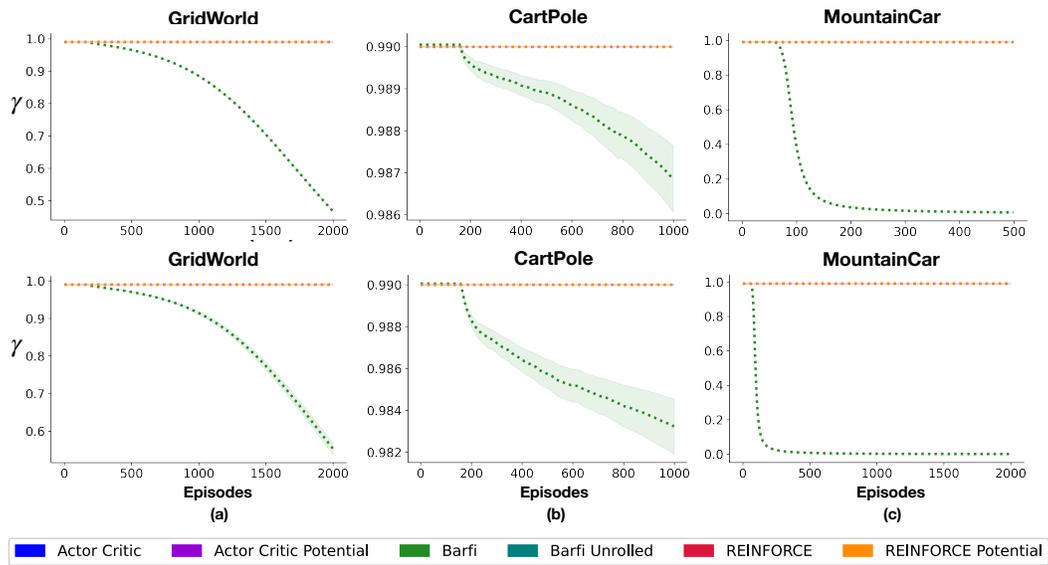


Figure 8: **Learned discounting** γ_φ : This figure illustrates the learned γ_φ for BARFI and normal γ for other methods, the curves are chosen based on best-performing curves on r_p , and averaged over 20 runs (except 40 for GW). **(a) Top** – $r_{aux,GW}^1$, **Bottom** – $r_{aux,GW}^2$, **(b) Top** – $r_{aux,CP}^1$, **Bottom** – $r_{aux,CP}^2$, **(c) Top** – $r_{aux,MC}^2$, **Bottom** – $r_{aux,MC}^1$. Mujoco is not included as the γ was not learned in that case. We can observe that the agents start to learn to decay γ at the appropriate pace. Note that the curves for methods other than BARFI and BARFI Unrolled are overlapping as the baselines don't change the value of γ .

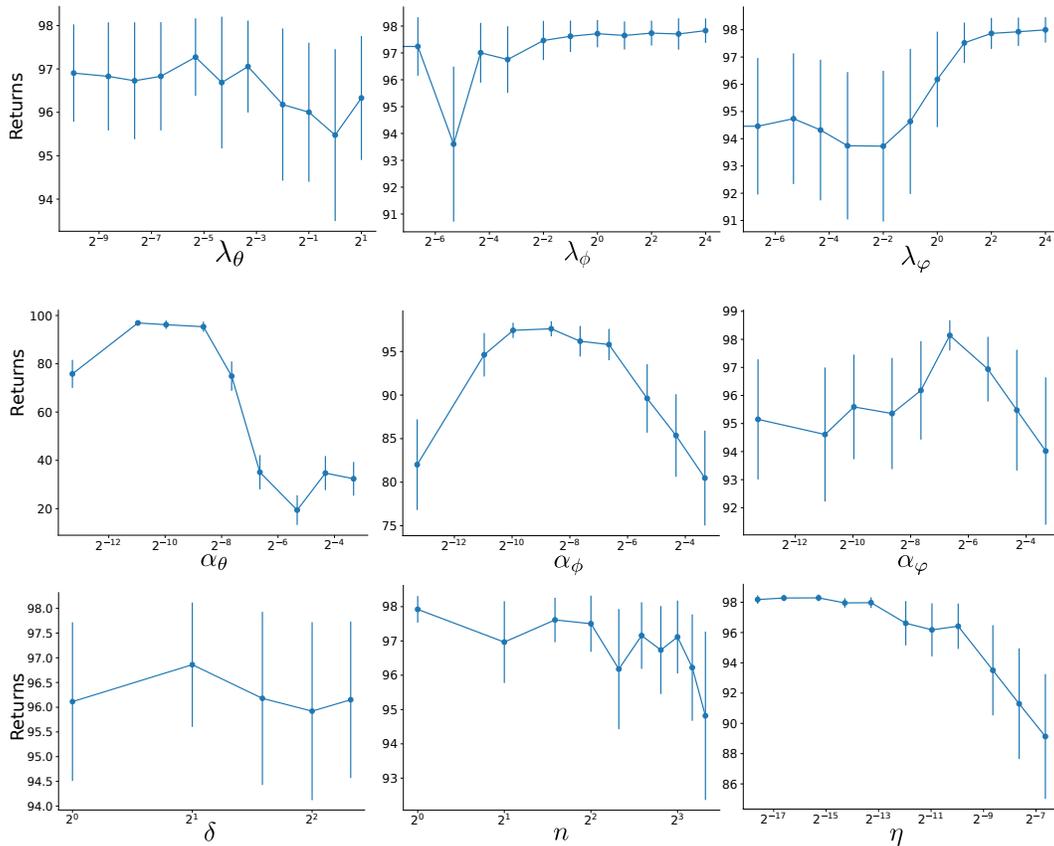


Figure 9: **Sensitivity Curves:** The set of graphs representing the sensitivity of different hyper-params keeping all the other params fixed. The sensitivity is for BARFI in GW with $r_{\text{aux}, \text{GW}}^2$, i.e., the misspecified reward. We choose the best-performing parameters and vary each parameter to see its influence. The curves are obtained for 50 runs (seeds) in each case, and error bars are standard errors. We can notice that α_θ and α_ϕ can have a large influence, and tend to stay around similar values. $\lambda_{\theta, \phi, \varphi}$ tends to help but doesn't really influence a lot in terms of its magnitude, except larger values of λ_φ seem to do better. Smaller values of η seems to work fine, hence something around 5×10^{-4} , 1×10^{-3} usually should suffice. n, δ can be chosen to around 5 and 3, and usually workout fine. We also defined $N_i = 5 \times \delta$ in this case.