EFFICIENT FULLY-OFFLINE META-REINFORCEMENT LEARNING VIA DISTANCE METRIC LEARNING AND BEHAVIOR REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the offline meta-reinforcement learning (OMRL) problem, a paradigm which enables reinforcement learning (RL) algorithms to quickly adapt to unseen tasks without any interactions with the environments, making RL truly practical in many real-world applications. This problem is still not fully understood, for which two major challenges need to be addressed. First, offline RL often suffers from bootstrapping errors of out-of-distribution state-actions which leads to divergence of value functions. Second, meta-RL requires efficient and robust task inference learned jointly with control policy. In this work, we enforce behavior regularization on learned policy as a general approach to offline RL, combined with a deterministic context encoder for efficient task inference. We propose a novel negative-power distance metric on bounded context embedding space, whose gradients propagation is detached from that of the Bellman backup. We provide analysis and insight showing that some simple design choices can yield substantial improvements over recent approaches involving meta-RL and distance metric learning. To the best of our knowledge, our method is the first model-free and end-to-end OMRL algorithm, which is computationally efficient and demonstrated to outperform prior algorithms on several meta-RL benchmarks.[1]

1 INTRODUCTION

Applications of reinforcement learning (RL) in real-world problems have been proven successful in many domains such as games (Silver et al., 2017; Vinyals et al., 2019; Ye et al., 2020) and robot control (Johannink et al., 2019). However, the implementations so far usually rely on interactions with either real or simulated environments. In other areas like healthcare (Gottesman et al., 2019), autonomous driving (Shalev-Shwartz et al., 2016) and controlled-environment agriculture (Binas et al., 2019) where RL shows promise conceptually or in theory, exploration in real environments is evidently risky, also building a high-fidelity simulator can be costly and challenging. Therefore a key step towards more practical RL algorithms is the ability to learn from fixed data (batch). Such paradigm, termed "offline RL" or "batch RL", would enable better generalization by incorporating diverse prior experience. Moreover, by leveraging and reusing previously collected data, off-policy algorithms such as SAC (Haarnoja et al., 2018) has been shown to achieve far better sample efficiency than on-policy methods. The same applies for offline RL algorithms since they are by nature off-policy.

The aforementioned design principles motivated a surge of recent works on offline/batch RL (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Siegel et al., 2020). These papers propose remedies by regularizing the trained policy to stay close to the training dataset of logged transitions, in order to mitigate the effect of bootstrapping error (Kumar et al., 2019) where errors in evaluation of out-of-distribution state-action pairs are never corrected and hence easily diverge due to inability to collect new data samples for feedback. There exist claims that offline RL can be implemented successfully without explicit correction for distribution mismatch given sufficiently large and diverse training data (Agarwal et al., 2020). However, we find such assumption unrealistic in many

[1]Source code and data are released at https://github.com/FOCAL-ICLR/FOCAL-ICLR/ and https://anonymous.4open.science/r/e657391f-0364-4670-a366-fb236d83f602/ respectively.

practices as well as in our experiments. In this paper, to tackle the out-of-distribution problem in offline RL in general, we follow the paradigm of behavior regularization in Wu et al. (2019).

For design of practical RL, besides the ability to learn without exploration, it's also ideal to have an algorithm that can generalize to a variety of scenarios. To solve real-world challenges in multi-task setting, such as treating different diseases, driving under various road conditions or growing diverse crops in autonomous greenhouses, a practical/data-efficient agent is expected to quickly transfer and adapt to unseen tasks, especially when the tasks share common structures. Meta-learning methods (Vilalta & Drissi, 2002; Thrun & Pratt, 2012; Finn et al., 2017) address this problem by learning an inductive bias from experience collected across a distribution of tasks, which can be naturally extended to reinforcement learning context. Under the umbrella of this so-called meta-RL, almost all current methods require on-policy data during either both meta-training and testing (Finn et al., 2017; Wang et al., 2016; Duan et al., 2016) or at least online exploration for task inference (Rakelly et al., 2019). An efficient method which incorporates both fully-offline learning and meta-learning in RL, despite few attempts (Li et al., 2019b; Dorfman & Tamar, 2020), to our knowledge, has not yet been fully developed and validated.

In this paper, under the first principle of maximizing practicality of RL algorithm, we propose an efficient method that integrates task inference with RL algorithms in a fully-offline fashion. Our fully-offline context-based actor-critic meta-RL algorithm, or FOCAL, achieves excellent sample efficiency and fast adaptation with only limited logged experience, on a range of deterministic continuous control meta-environments in MuJoCo (Todorov et al., 2012) as well as in a 2D navigation problem with sparse reward (Rakelly et al., 2019). The primary contribution of this work, to the best of our knowledge, is designing the first end-to-end and model-free offline meta-RL algorithm which is computationally efficient and effective without any prior knowledge of task identity or reward/transition functions. We show empirically that, for deterministic continuous control tasks, a deterministic context embedding is more effective than a probabilistic counterpart (Figure 4), contrary to online scenarios (Rakelly et al., 2019). To achieve efficient task inference, we propose an inverse-power loss for learning latent variables, in analogy with coulomb potential in electromagnetism, which is also unseen in previous work. Theoretical analyses and insights are provided to shed light on the specific design choices.

2 Related Work

Meta-RL Our work FOCAL builds upon the meta-learning (Vilalta & Drissi, 2002; Thrun & Pratt, 2012) framework in the context of reinforcement learning. Among all paradigms of meta-RL, this paper is most related to context-based and metric-based approaches. Context-based meta-RL employs models with memory such as recurrent (Duan et al., 2016; Wang et al., 2016; Fakoor et al., 2019), recursive (Mishra et al., 2017) or probabilistic (Rakelly et al., 2019) structures to achieve fast adaptation by aggregating experience into a latent representation on which the policy is conditioned. The design of recent methods usually leverages the temporal or Markovian properties of RL problems.

Metric-based meta-RL focuses on learning effective task representations to facilitate task inference and conditioned control policies, by employing techniques such as distance metric learning (Yang & Jin, 2006). Koch et al. (2015) proposed the first metric-based meta-algorithm for few-shot learning, in which a Siamese network (Chopra et al., 2005) is trained with triplet loss to compare the similarity between a query and supports in the embedding space. Matching networks (Vinyals et al., 2016) proposed the episodic training strategy and used the cross-entropy loss where the logits are the distances between a query and supports. Many metric-based meta-RL algorithms extend these works (Snell et al., 2017; Sung et al., 2018; Li et al., 2019a).

In contrast to context-based and metric-based methods, gradient-based meta-RL learns from aggregate experience using policy gradients (Finn et al., 2017; Nichol et al., 2018), meta-learned hyperparameters (Xu et al., 2018) or loss functions (Houthooft et al., 2018). These methods focus on on-policy meta-learning and are therefore much less sample efficient compared to off-policy context-based methods such as PEARL (Rakelly et al., 2019).

Among all aforementioned meta-learning approaches, this paper is most related to context-based PEARL algorithm (Rakelly et al., 2019) and metric-based prototypical networks (Snell et al.,

2017). PEARL achieves SOTA performance for off-policy meta-RL by introducing probabilistic permutation-invariant context encoder, along with a design which disentangles task inference and control by different sampling strategies. However, it requires exploration during meta-testing. The prototypical networks employ similar design of context encoder as well as an Euclidean distance metric on deterministic embedding space, but tackles meta-learning of classification tasks with squared distance loss as opposed to the inverse-power loss designed by FOCAL for the more complex offline meta-RL problem.

Offline/Batch RL To address the bootstrapping error (Kumar et al., 2019) problem of offline RL, this paper adopts behavior regularization directly from Wu et al. (2019), which provides a relatively unified framework of several recent offline or off-policy RL methods (Haarnoja et al., 2018; Fujimoto et al., 2019; Kumar et al., 2019). It incorporates a divergence function between distributions over actions in the actor-critic objectives. As with SAC (Haarnoja et al., 2018), one limitation of the algorithm is its sensitivity to reward scale and regularization strength. In our experiments, we indeed observed wide spread of optimal hyper-parameters across different meta-RL environments, shown in Appendix D.3.

Offline meta-RL To the best of our knowledge, despite attracting more and more attention, the offline meta-RL problem is still under-developed. We are aware of a few papers, mostly under review, that tackle the same problem as ours from different angles (Li et al., 2019b; Dorfman & Tamar, 2020). Li et al. (2019b) focuses on a specific scenario where biased datasets cause the task inference module to spuriously correlate only state-action pairs to the task identity, ignoring the reward information. This problem occurs when datasets of different tasks do not have significant overlap in their state-action visitation frequencies, and is exacerbated by sparse rewards. Even though we observed similar effects during our experiment, in this paper, we do not restrict our method design to such assumptions and instead try to robustify the algorithm in a more general setting, showing experimental results on scenarios for various training and testing datasets distributions (Appendix E.1). Another key difference is that our method is end-to-end and model free, whereas Li et al. (2019b) requires training of offline BCQ (Fujimoto et al., 2019) and reward function models for each task, which are computationally demanding. They also adopted the original PEARL strategy of probabilistic context encoder and couples learning of task inference and control by training the encoder to recover the state-action value function, for which we chose differently according to analytical and experimental evidence in Appendix E.

Dorfman & Tamar (2020) on the other hand, completely circumvents the out-of-distribution problem of offline RL by assuming sufficient data samples. They also formulate the OMRL as a Bayesian RL (Ghavamzadeh et al., 2016) problem and employs a probabilistic approach. Therefore we consider their methodology tangential to ours.

3 PRELIMINARIES

3.1 NOTATIONS AND PROBLEM STATEMENT

We consider fully-observed Markov Decision Process (MDP) (Puterman, 2014) in deterministic environments such as MuJoCo (Todorov et al., 2012). An MDP can be modeled as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \rho_0, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $P(s'|s,a)$, bounded reward function $R(s,a)$, initial state distribution $\rho_0(s)$ and discount factor $\gamma \in (0,1)$. The goal is to find a policy $\pi(a|s)$ to maximize the cumulative discounted reward starting from any state. We introduce the notion of multi-step state marginal of policy π as $\mu_\pi^t(s)$, which denotes the distribution over state space after rolling out π for t steps starting from state s. The notation $R_\pi(s)$ denotes the expected reward at state s when following policy π: $R_\pi(s) = \mathbb{E}_{a\sim\pi}[R(s,a)]$. The state-value function (a.k.a. value function) and action-value function (a.k.a Q-function) are therefore

$$V_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim \mu_\pi^t(s)}[R(s_t)] \tag{1}$$

$$Q_\pi(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)}[V_\pi(s')] \tag{2}$$

Q-learning algorithms are implemented by iterating the Bellman optimality operator $\mathcal{B}$, defined as:

$$(\mathcal{B}\hat{Q})(s,a) := R(s,a) + \gamma \mathbb{E}_{P(s'|s,a)}[\max_{a'} \hat{Q}(s',a')] \tag{3}$$

when the state space is large/continuous, $\hat{Q}$ is used as a hypothesis from the set of function approximators (e.g. neural networks).

In the offline context of this work, given a distribution of tasks $p(\mathcal{T})$ where every task is an MDP, we study off-policy meta-learning from collections of static datasets of transitions $\mathcal{D}_i = \{(s_{i,t}, a_{i,t}, s'_{i,t}, r_{i,t}) | t = 1, ..., N\}$ generated by a set of behavior policies $\{\beta_i(a|s)\}$ associated with each task index i. A key underlying assumption of meta-learning is that the tasks share some common structures. Accordingly, by definition of MDP, in this paper we restrict our attention on tasks with shared state and action space, but differ in transition and reward functions.

We define the meta-optimization objective as

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \left[\mathcal{L}_{\mathcal{T}_i}(\theta) \right] \tag{4}$$

where $\mathcal{L}_{\mathcal{T}_i}(\theta)$ is the objective evaluated on transition samples drawn from task $\mathcal{T}_i$. A common choice of $p(\mathcal{T})$ is the uniform distribution on the set of given tasks $\{\mathcal{T}_i | i = 1, ..., n\}$. In this case, the meta-training procedure turns into minimizing the average losses across all training tasks

$$\hat{\theta}_{\text{meta}} = \arg\min_{\theta} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[\mathcal{L}_k(\theta) \right] \tag{5}$$

3.2 Behavior Regularized Actor Critic

Similar to SAC, to constrain the out-of-distribution error in offline RL, for each individual task $\mathcal{T}_i$, behavior regularization (Wu et al., 2019) introduces a divergence measure between the learner π_θ and the behavior policy π_b and in both value and target Q-functions. For simplicity, we ignore task index in this section:

$$V_\pi^D(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim \mu_\pi^t(s)} \left[R_\pi(s) - \alpha D(\pi_\theta(\cdot|s_t), \pi_b(\cdot|s_t)) \right] \tag{6}$$

$$\bar{Q}_\psi^D(s, a) = \bar{Q}_\psi(s, a) - \alpha \hat{D}(\pi_\theta(\cdot|s), \pi_b(\cdot|s)) \tag{7}$$

where $\bar{Q}$ denotes a target Q function without gradients and $\hat{D}$ denotes a sample-based estimate of the divergence function D. In actor-critic framework, the loss functions of Q-value and policy learning are given by, respectively,

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{\substack{(s,a,r,s') \sim \mathcal{D} \\ a' \sim \pi_\theta(\cdot|s')}} \left[\left(r + \gamma \bar{Q}_\psi^D(s', a') - Q_\psi(s, a) \right)^2 \right] \tag{8}$$

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\mathbb{E}_{a'' \sim \pi_\theta(\cdot|s)} [Q_\psi(s, a'')] - \alpha \hat{D} \right] \tag{9}$$

3.3 Context-Based meta-RL

Context-based meta-RL algorithms aggregate context information, typically in form of task-specific transitions, into a latent space $\mathcal{Z}$. Adaptation at test time in meta-RL can be viewed as a special case of RL in a partially-observed MDP (POMDP, (Kaelbling et al., 1998)) by including the latent representation z as the unobserved part of the state. Once given complete information of z and s combined as the full state, the reinforcement learning of policy $\pi_\theta(s, z)$ becomes RL in regular MDP, and properties of regular RL such as the existence of optimal policy and value functions hold naturally. We therefore formulate the context-based meta-RL problem as solving a *task-augmented MDP* (TA-MDP). The formal definitions are provided in Appendix B.

4 Method

Based on our formulation of context-based meta-RL problem, FOCAL first learns an effective representation of meta-training tasks on latent space $\mathcal{Z}$, then solves the offline RL problem on TA-MDP with behavior regularized actor critic method. We summarize our training procedure in Figure 1 and describe the detailed algorithm in Appendix A. Our method consists of two key design choices: deterministic context encoder and distance metric learning for task inference.

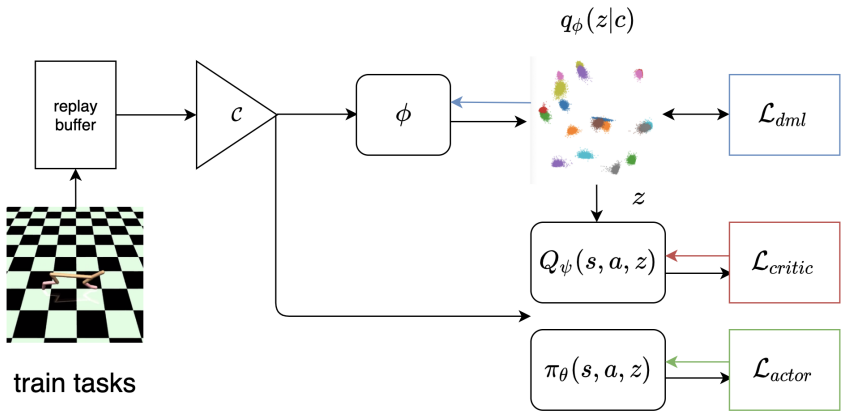


Figure 1: **Meta-training procedure.** The inference network q_ϕ uses context data to determine the latent context variable z, which conditions the actor and critic, and is optimized by the DML objective. The actor and critic objectives are optimized in parallel without any mixing of gradients.

4.1 Deterministic Context Encoder

Similar to Rakelly et al. (2019), we introduce an *inference network* $q_\phi(z|c)$, parameterized by ϕ, to encode task identity from context c. In terms of the context encoder design, recent meta-RL methods either employ recurrent neural networks (Duan et al., 2016; Wang et al., 2016) to capture the temporal property of MDP, or use probabilistic models (Rakelly et al., 2019) for reasoning of uncertainties. These design choices are proven effective in on-policy or partially-offline off-policy algorithms. However, since our approach aims to address the fully-offline meta-RL problem, we argue that a deterministic context encoder works better in this scenario, given a few assumptions. First, we consider only *deterministic MDP* in this paper, where the transition function is a Dirac delta distribution. We assume that all meta-learning tasks in this paper are deterministic MDPs, which is satisfied by common RL benchmarks such as MuJoCo. The formal definitions are detailed in Appendix B. Additionally, all tasks share the same state and action space, while each is characterized by a unique set of transition and reward functions. Mathematically, this means there exists an injective function $f : \mathcal{T} \to \mathcal{P} \times \mathcal{R}$, where $\mathcal{P}$ and $\mathcal{R}$ are functional spaces of transition probability $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \{0, 1\}$ and bounded reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ respectively. A stronger condition of this injective property is that for every task $\mathcal{T} \sim p(\mathcal{T})$ and any state-action pair (s, a), the corresponding transition and reward are unique across all tasks, which brings the *Task-Transition Correspondence* assumption:

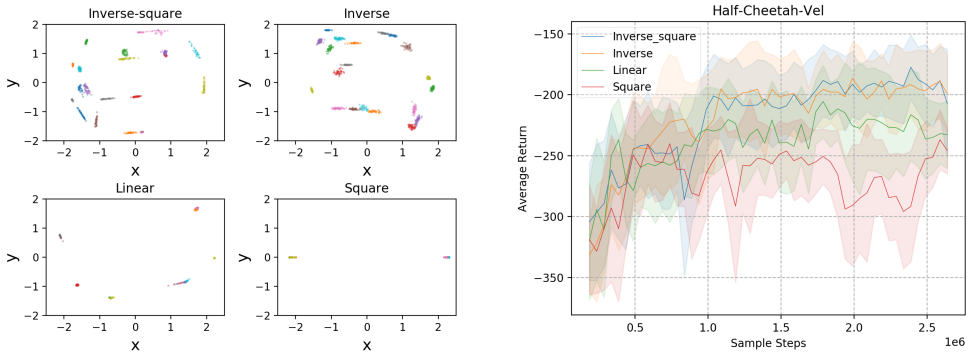
Assumption 1 (Task-Transition Correspondence) *We consider meta-RL with a task distribution $p(\mathcal{T})$ to satisfy task-transition correspondence if and only if $\forall \mathcal{T}_1, \mathcal{T}_2 \sim p(\mathcal{T}), (s, a) \in \mathcal{S} \times \mathcal{A}$:*

$$P_1(\cdot|s, a) = P_2(\cdot|s, a), R_1(s, a) = R_2(s, a) \iff \mathcal{T}_1 = \mathcal{T}_2 \tag{10}$$

Under the deterministic MDP assumption, the transition probability function $P(\cdot|s, a)$ is associated with the transition map $t : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. The task-transition correspondence suggests that, given the action-state pair (s, a) and task $\mathcal{T}$, there exists a unique transition-reward pair (s', r). Based on these assumptions, one can define a task-specific map $f_\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times \mathbb{R}$ on the set of transitions $\mathcal{D}_\mathcal{T}$:

$$f_\mathcal{T}(s_t, a_t) = (s'_t, r_t), \quad \forall \mathcal{T} \sim p(\mathcal{T}), (s_t, a_t, s'_t, r_t) \in \mathcal{D}_\mathcal{T} \tag{11}$$

Recall that all tasks defined in this paper share the same state-action space, hence $\{f_\mathcal{T} | \mathcal{T} \sim p(\mathcal{T})\}$ forms a function family defined on the transition space $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R}$, which is also by definition the context space $\mathcal{C}$. This lends a new interpretation that as a task inference module, the context encoder $q_\phi(z|c)$ enforces an embedding of the task-specific map $f_\mathcal{T}$ on the latent space $\mathcal{Z}$, i.e. $q_\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R} \to \mathcal{Z}$. Following Assumption 1, every transition $\{s_i, a_i, s'_i, r_i\}$ corresponds to a unique task $\mathcal{T}_i$, which means in principle, task identity can be inferred from any single transition datum. This implies the context encoder should be permutation-invariant and deterministic, since



(a) context embedding

(b) test-task performance

Figure 2: (a) A 2D projection of the 5D embedding vectors of samples drawn from 20 randomized tasks on Half-Cheetah-Vel by principal component analysis (PCA). Trials cover four DML losses of different powers. The inverse-power losses separate all embeddings effectively whereas linear and square losses exhibit degenerate behaviors. Data points are color-coded according to task identity. (b) FOCAL trained with inverse-power losses also outperform the linear and square distance losses.

the embedding of context does not depend on the order of the transitions nor involve any uncertainty. **This observation is crucial since it provides theoretical basis for few-shot learning (Snell et al., 2017; Sung et al., 2018) in our settings**. In particular, when learning in a fully-offline fashion, any meta-RL algorithm at test-time cannot perform adaptation by exploration. The theoretical guarantee that a few randomly-chosen transitions can enable effective task inference ensures that FOCAL is feasible and efficient.

4.2 Distance Metric Learning (DML) of Latent Variables

In light of our analysis on the context encoder design, the goal of task inference is to learn a robust and effective representation of context for better discrimination of task identities. Unlike PEARL, which requires gradients from the Bellman update to train the inference network, our main insight is to disentangle the learning of context encoder from the learning of control policy. As explained in previous reasoning about the deterministic encoder, the latent variable is a representation of the task properties involving only transition probability and reward, which in principle should be completely captured by the datasets. **Given continuous neural networks as function approximators, the learned value functions conditioned on latent variable z cannot account for the difference between tasks if the corresponding embedding vectors are too close (Appendix C)**. Therefore for implementation, we formulate the latent variable learning problem as obtaining the map $q_\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R} \rightarrow \mathcal{Z}$ of transition data $\mathcal{D}_i = \{(s_{i,t}, a_{i,t}, s'_{i,t}, r_{i,t}) | t = 1, ..., N\}$ that clusters similar data points (same task) while pushes away dissimilar data points (different tasks) on the embedding space $\mathcal{Z}$, which is essentially distance metric learning (DML) (Sohn, 2016). A common loss function in DML is contrasitive loss (Chopra et al., 2005; Hadsell et al., 2006). Given input data $x_i, x_j \in \mathcal{X}$ and output label $y \in \{1, ..., L\}$, it is written as

$$\mathcal{L}^m_{cont}(x_i, x_j; q) = \mathbf{1}\{y_i = y_j\}||q_i - q_j||_2^2 + \mathbf{1}\{y_i \neq y_j\}\max(0, m - ||q_i - q_j||_2)^2 \tag{12}$$

where m is a margin parameter imposing the distance between examples from different classes to be larger than m, q_i is the embedding vector of x_i. For data point of different tasks, contrastive loss rewards the distance between their embedding vectors by L^2 norm, which is weak when the distance is small, as in the case when z is normalized and q_ϕ is randomly initialized. Empirically, we observe that objectives with positive powers of distance lead to degenerate representation of tasks, forming clusters of embedding vectors of different tasks (Figure 2(a)). Theoretically, this is due to the fact that an accumulative L^2 loss of distance between data points is proportional to the dataset variance, which may lead to degenerate distribution such as Bernoulli distribution. This is proven in Appendix B. To build robust and efficient task inference module, it's crucial to ensure every embedding vector to be separated from each other (Appendix C). We therefore introduce a negative-power variant of

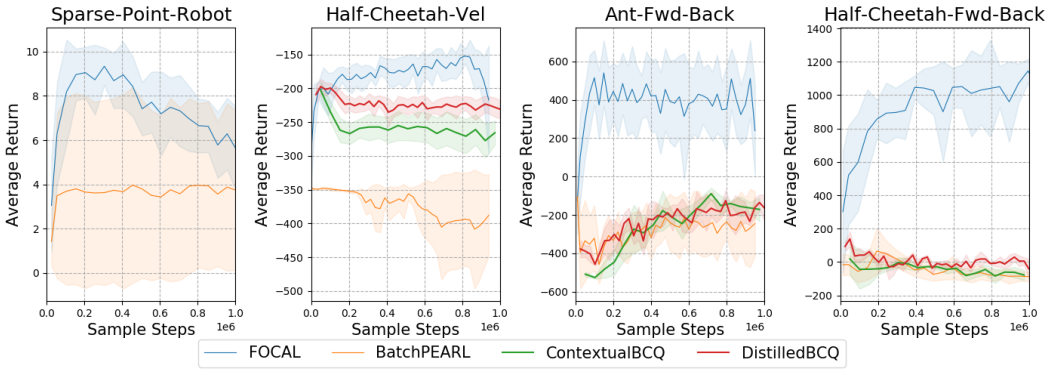


Figure 3: Average episodic testing return of FOCAL vs. other baselines on four meta-environments. FOCAL significantly outperforms other baselines in terms of asymptotic performance and sample efficiency. The only drawback is the over-fitting behavior on Sparse-Point-Robot, which is due to the sparse reward and offline RL with limited data.

contrastive loss as follows:

$$\mathcal{L}^m_{dml}(x_i, x_j; q) = \mathbf{1}\{y_i = y_j\}||q_i - q_j||_2^2 + \mathbf{1}\{y_i \neq y_j\}\beta \cdot \frac{1}{||q_i - q_j||_2^n + \epsilon} \tag{13}$$

where $\epsilon > 0$ is a small coefficient to avoid division by zero, the power n can be any non-negative number. We experimented with 1 and 2 in this paper.

5 Experiments

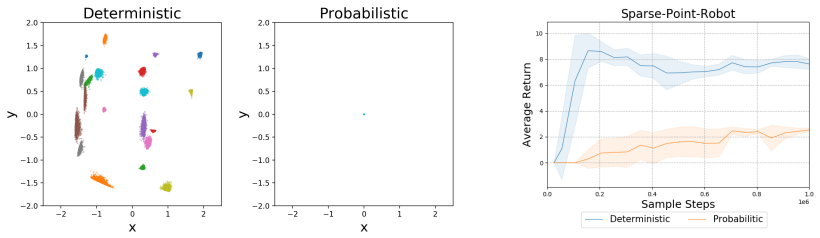
In our experiments, we assess the performance of FOCAL by comparing it with several baseline algorithms on meta-RL benchmarks. Specific design choices are examined through ablations. All trials are averaged over 3 random seeds. Supplementary experiments are provided in Appendix E.

5.1 Sample Efficiency and Performance

We evaluate FOCAL on four continuous control meta-environments of robotic locomotion, three of which are simulated via the MuJoCo simulator (Todorov et al., 2012), with the exception of a 2D navigation problem with sparse reward called Sparse-Point-Robot. All environments require adaptation across reward functions (walking direction for Half-Cheetah-Fwd-Back and Ant-Fwd-Back, target velocity for Half-Cheetah-Vel and target location for Sparse-Point-Robot). These meta-Rl benchmarks were previously introduced by Finn et al. (2017); Rakelly et al. (2019) and are described in Appendix D. We train stochastic SAC models for every single task and roll out policies saved at every checkpoint to collect trajectories. The offline training datasets can be collected as a selection of the saved trajectories. This setup facilitates tuning of the performance level and state-action distributions of the datasets for each task.

For OMRL, there are two natural baselines. The first is by modifying PEARL to train and test from solely logged data without exploration. The second one is Contextual BCQ introduced by Li et al. (2019b). By incorporating a latent space z in the state space for task inference, it can be interpreted as the task-augmented variant of offline BCQ algorithm Fujimoto et al. (2019). Lastly, we include comparison with the proposed method in Li et al. (2019a) which we term "Distilled BCQ". Although as discussed earlier, Distilled BCQ is a model-based, two-stage method as opposed to our model-free and end-to-end approach, we consider it by far the most competitive and related OMRL algorithm to FOCAL, due to the lack of other OMRL methods.

As shown in Figure 3, we observe that FOCAL significantly outperforms prior offline meta-RL methods across all domains in terms of both sample efficiency and performance. Most interestingly, when compared to Batch PEARL, a vanilla version of FOCAL without behavior regularization or



(a) context embedding (b) test-task performance

Figure 4: (a) 2D projection of the 5D embedding vectors of samples drawn from 20 randomized tasks on Walker-2D-Params by PCA. The deterministic inverse-square encoder separates all embeddings effectively whereas the probabilistic variant exhibits no separation. Data points are color-coded according to task identity. (b) FOCAL trained with deterministic encoder also outperforms the probabilistic variant on Sparse-Point-Robot.

distance metric learning, the gap in performance is evident due to divergence of value functions as well as suboptimal representation of the context encoder (Appendix E). For each meta environment, we maintain the same datasets for training FOCAL and the other baselines, most of which were generated by expert-level policies. However, for some tasks such as Ant-Fwd-Back, we observed that a diverse training sets (generated by expert, medium, random policies) result in a better meta policy. We conjecture that due to higher difficulty of tasks, the out-of-distribution problem exacerbates as the diversity of data required to train a successful policy increases. Indeed, among all our environments, the regularization strength required to train a successful OMRL policy on Ant-Fwd-Back is much greater than other tasks. The experimental details are included in Appendix D.

5.2 Ablations

Based on our previous analysis, we examine and validate two key design choices of FOCAL by the following ablations.

5.2.1 Deterministic vs. Probabilistic Context Encoder

Despite abundance successes of probabilistic/variational inference models in previous work (Kingma & Welling, 2013; Alemi et al., 2016), which is also adopted in the SOTA off-policy meta-RL algorithm PEARL, we observe experimentally that a deterministic context encoder performs significantly better on both simple (Sparse-Point-Robot) and complex (Walker-2D-Params) tasks in fully offline settings, shown in Figure 4. As discussed in our methodology, this is due to

1. Offline meta-RL does not require exploration. Also when the *Task-Transition Correspondence* is satisfied, there is not need to reason about uncertainty for adaption.
2. The deterministic context encoder in FOCAL is trained with carefully designed distance metric learning objective, detached from the Bellman update, which provides better efficiency and stability for meta-learning.

However, the deterministic context encoder which is built upon the *Task-Transition Correspondence* assumption, is less effective in presence of sparse reward. In this case, a non-negligible portion of samples violate Assumption 1 and provide confounding information for task inference. Consequently, the algorithm may learn false correlation between tasks and transitions and prone to overfitting. Indeed, among the four experiments in Figure 3, we observe much significant over-fitting on Sparse-Point-Robot compared to others.

5.2.2 Power Law of Distance Metric Loss

As examined in Figure 2, we tested context embedding loss with different powers of distance, from L^{-2} to L^2. A 2D PCA projection view of the high-dimensional embedding space in Figure 2 demonstrates that, distance metric loss with negative powers are more effective in separating embedding

vectors of different tasks, where positive powers exhibit degenerate behaviors, leading to less robust and effective conditioned policies. By a physical analogy, the inverse-power losses provide "repulsive forces" that drive apart all data points, regardless of the initial distribution. In electromagnetism, consider the latent space as a 3D metal cube and the embedding vectors as positions of positive charges. By Gauss's law, at equilibrium state, all "charges" are distributed on the surface of the cube with densities positively related to the local curvature of the surface. Indeed, despite the latent space in Fig. 2 is of dimension 5, we observe from the "Inverse-square" and "Inverse" trials that almost all vectors are located near the edge of the latent space, with higher concentration around the vertices, which have the highest local curvatures (Appendix E.2).

Loss	RMS	ESR
Inverse-square	1.282	0.430
Inverse	1.217	0.420
Linear	1.385	0.409
Square	1.415	0.253

Table 1: Embedding Statistics on Half-Cheetah-Vel

To evaluate the effectiveness of different form of DML loss, we define a metric called *effective separation rate* (ESR) which computes the percentage of embedding vector pairs of different tasks whose distance on $\mathcal{Z}$ is larger than the expectation of randomly distributed vector pairs, i.e., $\sqrt{\frac{2}{3}l}$ on $(-1, 1)^l$. Table 1 demonstrates that DML losses of negative power are more effective in maintaining distance between embeddings of different tasks, while no significant distinction is shown in terms of RMS distance, which is aligned with our insight that RMS or effectively classical L^2 objective, can be optimized by degenerate distributions (Appendix B.1). **This is the core challenge addressed by our proposed inverse-power loss**.

6 Conclusion

In this paper, we propose a novel fully-offline meta-Rl algorithm, FOCAL, in pursuit of more practical RL. Our method involves independent distance metric learning of a deterministic context encoder for efficient task inference, combined with an actor-critic apparatus with regularization to learn effective policy from static data. By re-formulating the meta-RL tasks as *task-augmented MDPs* with *Task-Transition Correspondence* assumption, we shed light on the effectiveness of our design choices in both theory and experiments. Our approach achieves superior results compared to existing OMRL algorithms on a diverse set of continous control meta-RL domains. Despite the success, we also discuss the breaking down of our fundamental assumptions in extreme scenarios such as sparse reward or stochastic environments, which opens up avenues for future work of more general and sophisticated OMRL algorithms.

References

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Jonathan Binas, Leonie Luginbuehl, and Yoshua Bengio. Reinforcement learning for sustainable agriculture. In *ICML 2019 Workshop Climate Change: How Can AI Help*, 2019.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

Ron Dorfman and Aviv Tamar. Offline meta reinforcement learning. *arXiv preprint arXiv:2008.02598*, 2020.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl˜2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, and Alexander J Smola. Meta-q-learning. *arXiv preprint arXiv:1910.00125*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*, 2016.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25(1):16–18, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Rein Houthooft, Yuhua Chen, Phillip Isola, Bradly Stadie, Filip Wolski, OpenAI Jonathan Ho, and Pieter Abbeel. Evolved policy gradients. In *Advances in Neural Information Processing Systems*, pp. 5400–5409, 2018.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6023–6029. IEEE, 2019.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2019a.

Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Iskov Christensen, and Hao Su. Multi-task batch reinforcement learning with metric learning. *arXiv*, pp. arXiv–1909, 2019b.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340, 2019.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pp. 1857–1865, 2016.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. In *Advances in neural information processing systems*, pp. 2396–2407, 2018.

Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4, 2006.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *AAAI*, pp. 6672–6679, 2020.