Block Contextual MDPs for Continual Learning

Shagun Sodhani SODHANI@FB.COM
Franziska Meier FMEIER@FB.COM
Joelle Pineau JPINEAU@FB.COM
Facebook AI Research

Amy Zhang AMYZHANG@FB.COM
Facebook AI Research & UC Berkeley

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

In reinforcement learning (RL), when defining a Markov Decision Process (MDP), the environment dynamics are implicitly assumed to be stationary. This assumption of stationarity, while simplifying, can be unrealistic in many scenarios. In the continual reinforcement learning scenario, the sequence of tasks is another source of nonstationarity. In this work, we propose to examine this continual reinforcement learning setting through the *block contextual MDP* (BC-MDP) framework, which enables us to relax the assumption of stationarity. This framework challenges RL algorithms to handle both nonstationarity and rich observation settings and, by additionally leveraging smoothness properties, enables us to study generalization bounds for this setting. Finally, we take inspiration from adaptive control to propose a novel algorithm that addresses the challenges introduced by this more realistic BC-MDP setting, allows for zero-shot adaptation at evaluation time, and achieves strong performance on several nonstationary environments. [1].

Keywords: Reinforcement Learning, MDP, Block Contextual MDP, Continual Learning

1. Introduction

In the standard RL regime, many limiting assumptions are made to keep the problem setting tractable. A typical assumption is that the environment is stationary, i.e., the dynamics and reward do not change over time. However, from fluctuating traffic patterns to warehouse robots, most real-world settings do not conform to this assumption. Even the observation and action space can change over time in the more general case. These setups are commonly grouped under continual learning paradigm (Ring et al., 1994; Thrun, 1998; Hadsell et al., 2020) and non-stationarity is incorporated as a change in the task or environment distribution (that the agent operates in). The ability to handle non-stationarity is essential for developing continual learning agents (Khetarpal et al., 2020).

Real-life settings present an additional challenge: we can not rely on access to an interpretable and compact (if not minimal) state space. Often, we only have access to a rich and high-dimensional observation space. For example, when driving a car on a wet road, we only have access to the "view" around us and not the friction coefficient between the car and the road. Hence, we must account for irrelevant information in the observation when designing agents for nonstationary environments.

We propose to model this more realistic, rich observation, nonstationary setting as a Block Contextual MDP (BC-MDP) by combining two common assumptions: (i) the *block assumption* (Du et al.,

1. The appendix contains additional details and is available at https://shagunsodhani.com/docs/ZeusAppendix.pdf

2019) that addresses rich observations with irrelevant features and (ii) the *contextual MDP* (Hallak et al., 2015) assumption - MDPs with different dynamics and rewards share a common structure and a context that can describe the variation across tasks. We introduce the Lipschitz Block Contextual MDP framework that leverages results connecting Lipschitz functions to generalization (Xu and Mannor, 2010) and enables us to frame nonstationarity as a changing context over a family of stationary MDPs (thus modeling it as a contextual MDP). We propose a representation learning algorithm to enable the use of current RL algorithms (that rely on the prototypical MDP setting) in nonstationary environments. It works by constructing a context space that is Lipschitz with respect to the changes in dynamics and reward of the nonstationary environment. We show theoretically and empirically that the trained agent generalizes well to unseen contexts. We also provide value bounds based on this approximate abstraction which depend on some basic assumptions.

Our work is inspired from adaptive control (Slotine and Li, 1991), a control method that continuously performs parameter identification to adapt to nonstationary dynamics of a system. Adaptive control generally considers the "known unknowns," where the system properties are known, but their values are unknown. We focus on the "unknown unknowns" setting, where the agent neither knows the property nor its value in any task. While our setup is similar to meta-learning methods that "learn to learn,", meta-learning techniques generally require finetuning or updates on the novel tasks (Finn et al., 2017; Rakelly et al., 2019). In contrast *our method can adapt in a zero-shot manner without any parameter updates* and does not suffer from catastrophic forgetting (McCloskey and Cohen, 1989). This property is very critical when designing continual learning agents that operate in the real world. We refer to our proposed method as **Ze**ro-shot adaptation to **U**nknown **S**ystems (ZeUS).

Contributions. We 1) introduce the Lipschitz Block Contextual MDP framework for the continual RL setting, 2) provide theoretical bounds on adaptation and generalization ability to unseen tasks within this framework utilizing Lipschitz properties, 3) propose an algorithm (ZeUS) to perform online inference of "unknown unknowns" to solve a family of tasks (without performing learning updates at test time) and ensure the prior Lipschitz properties hold, and 4) empirically verify the effectiveness of ZeUS on environments with nonstationary dynamics or reward functions.

2. Related Work

In **System Identification and Adaptive Control** (Zadeh, 1956; Åström and Bohlin, 1965; Swevers et al., 1997; Ljung, 2010; Van Overschee and De Moor, 2012; Chiuso and Pillonetto, 2019; Ajay et al., 2019; Yu et al., 2017; Zhu et al., 2018) setup, the goal is to perform system identification of "known unknowns," where the environment properties are known, but their values are unknown. Applying this setup to the example of driving a car,, the agent knows that friction coefficient varies across tasks but does not know its value. The agent can infer the unknown value (from observed data) and condition its policy to solve a given task. We extend this setup to the "unknown unknowns" setting, where the agent neither knows the environment property nor its value in any task.

Our work is related to the **Continual (or Lifelong) RL** (Ring et al., 1994; Gama et al., 2014; Kaplanis et al., 2018; Aljundi et al., 2019; Javed and White, 2019; Hadsell et al., 2020). Specifically, we focus on the *passive nonstationarity* setup where the environment dynamics may change irrespective of the agent's behavior (Khetarpal et al., 2020). Unlike Lopez-Paz and Ranzato (2017); Chaudhry et al. (2019); Sodhani et al. (2020) which focus on challenges like catastrophic forgetting (McCloskey

and Cohen, 1989) [2], we focus on the ability to continually adapt (the policy) to unseen tasks (Hadsell et al., 2020).Unlike previous works like Xie et al. (2020) that uses a probabilistic hierarchical latent variable model to learn a representation of the environment and perform off-policy learning, we use task metrics to learn a context space and focus on generalization to unseen contexts.

Several works have focused on **modeling the environment context** from high-level pixel observations (Pathak et al., 2017; Chen et al., 2018; Xu et al., 2019). This context (along with the observation) is fed as input to the policy to enable it to adapt to unseen dynamics (by implicitly capturing the dynamics parameters). These approaches learn a single, global dynamics model conditioned on the output of a context encoder. Similar to these approaches, we also use a context encoder but introduce an additional loss to learn a context space with Lipschitz properties with respect to reward, and dynamics. Recently, Xian et al. (2021) proposed using HyperNetworks (Ha et al., 2017) that use the context to generate the weights of the *expert* dynamics model.

Other works on structured MDPs, that **leverage the Lipschitz properties**, include Modi et al. (2018) that assumes that the given contextual MDP is smooth and that the distance metric and Lipschitz constants are known. In contrast, we propose a method that constructs a new smooth contextual MDP, with bounds on downstream behavior based on the approximate-ness of the new contextual MDP. Modi and Tewari (2020) propose RL algorithms with lower bounds on regret but assume that the context is known and linear with respect to the MDP parameters. In contrast, we do not assume access to the context at train or test time or linearity with respect to MDP parameters.

Meta-reinforcement learning methods (Finn et al., 2017; Nagabandi et al., 2019a; Rakelly et al., 2019; Zhao et al., 2020) can be broadly classified as:i) Optimization-based methods (Finn et al., 2017; Zintgraf et al., 2019) that require updating model parameters for each task (and therefore suffer from catastrophic forgetting) and ii) Context-based methods (Nagabandi et al., 2019c) that perform online adaptation given a context representation. Follow-up work (Lee et al., 2020) introduced additional loss terms that encourage the context encoding to be useful for predicting both forward (next state) and backward (previous state) dynamics while being temporally consistent. In contrast, our objective is to learn a context space with Lipschitz properties with respect to reward and dynamics. Some works have proposed modeling meta-RL as task inference (Humplik et al., 2019; Kamienny et al., 2020) but assume access to some *privileged information* (like *task-id*) during training.

Our work is also related to the general problem of training a policy on Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998; Igl et al., 2018; Zhang et al., 2019; Hafner et al., 2019) that capture both nonstationarity and rich observation settings. Our experiments are performed in the POMDP setup where we train the agent using pixel observations. However, we focus on a specific class of POMDPs — the contextual MDP with hidden context, which enables us to obtain strong generalization performance to new environments. We discuss additional related works in multi-task RL, transfer learning, and MDP metrics in the Appendix [3].

3. Background & Notation

A **Markov Decision Process** (MDP) (Bellman, 1957; Puterman, 1995) is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, R, T, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $T : \mathcal{S} \times \mathcal{A} \to Dist(\mathcal{S})$ is the environment transition probability function, and $\gamma \in [0, 1)$ is the discount factor. At each time step, the learning agent perceives a state $s_t \in \mathcal{S}$, takes

2. Since our model does not perform parameter updates when transferring to unseen tasks, it does not suffer from catastrophic forgetting.
3. Available at: https://shagunsodhani.com/docs/ZeusAppendix.pdf

an action $a_t \in \mathcal{A}$ drawn from a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and with probability $T(s_{t+1}|s_t, a_t)$ enters next state s_{t+1}, receiving a numerical reward R_{t+1} from the environment. The value function of policy π is defined as: $V_\pi(s) = E_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s]$. The optimal value function V^* is the maximum value function over the class of stationary policies.

Contextual Markov Decision Process (Hallak et al., 2015) is an augmented Markov Decision Process that utilize *side information* as context, similar to contextual bandits. For example, the friction coefficient between car and road is a context variable that affects the environment dynamics.

Definition 1 (Contextual Markov Decision Process) *A contextual Markov decision process (CMDP) is defined by tuple $\langle \mathcal{C}, \mathcal{S}, \mathcal{A}, \mathcal{M} \rangle$ where $\mathcal{C}$ is the context space, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space. $\mathcal{M}$ is a function which maps a context $c \in \mathcal{C}$ to MDP parameters $\mathcal{M}(c) = \{R^c, T^c\}$.*

In the real world, we typically operate in a "rich observation" setting without access to a compressed state representation and the learning agent has to learn a mapping from the observation to the state. This additional relaxation of the original CMDP definition as a form of Block MDP (Du et al., 2019) was previously introduced in Sodhani et al. (2021) for the multi-task setting where the agent focuses on a subset of the space for a specific task, which we present here again for clarity:

Definition 2 (Block Contextual Markov Decision Process (Sodhani et al., 2021)) *A block contextual Markov decision process (BC-MDP) is defined by tuple $\langle \mathcal{C}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{M} \rangle$ where $\mathcal{C}$ is the context space, $\mathcal{S}$ is the state space, $\mathcal{O}$ is the observation space, $\mathcal{A}$ is the action space. $\mathcal{M}$ is a function which maps a context $c \in \mathcal{C}$ to MDP parameters and observation space $\mathcal{M}(c) = \{R^c, T^c, \mathcal{O}^c\}$.*

The continual learning setting differs from sequential multi-task learning as there is no delineation of tasks when c changes, causing nonstationarity in the environment. We make an additional assumption that the change in c is smooth over time and the BC-MDP itself is smooth, as shown in Definition 3. We now define a Lipschitz MDP for the MDP family we are concerned with.

Definition 3 (Lipschitz Block Contextual MDP) *Given a BC-MDP $\langle \mathcal{C}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{M} \rangle$ and distance metric $d(\cdot, \cdot)$ over context space, if for any two contexts $c_1, c_2 \in \mathcal{C}$, we have the following constraints,*

$$\forall (s, a), W(T^{c_1}(s, a), T^{c_2}(s, a)) \leq L_p d(c_1, c_2),$$
$$\forall (s, a), \|R^{c_1}(s, a) - R^{c_2}(s, a)\| \leq L_r d(c_1, c_2),$$

then the BC-MDP is referred to as a Lipschitz BC-MDP with smoothness parameters L_p and L_r.

Here W denotes the Wasserstein distance. Note that Definition 3 is not a limiting assumption because we do not assume access to the context variables c_1 and c_2, and they can therefore be chosen so that the Lipschitz condition is always satisfied. In this work, we focus on a method for learning a context space that satisfies the above property.

4. Generalization Properties of Lipschitz BC-MDPs

The key idea behind the proposed method (presented in full in Section 5) is to construct a context space $\mathcal{C}$ with Lipschitz properties with respect to dynamics and reward, and therefore, optimal value functions across tasks. In this section, we show how this Lipschitz property aids generalization. The following results hold for any given observation (or state) space and are not unique to Block MDPs, so we use notation with respect to states $s \in \mathcal{S}$ without loss of generality. Since we do not have access to the true context space, in Section 5, we describe how to learn a context space with the

desired characteristics. To construct a context space that is Lipschitz with respect to tasks, notably the optimal value functions across tasks, we turn to metrics based on state abstractions, and define a task distance metric for the continual RL setting.

Definition 4 (Task Metric) *Given two tasks sampled from a BC-MDP, identified by context c_i & c_j. Let $\Updownarrow\rceil\sqcup$ be the space of bounded pseudometrics on context space $\mathcal{S}$. We define $F : \Updownarrow\rceil\sqcup \mapsto \Updownarrow\rceil\sqcup$ by:*

$$F(h)(c_i, c_j) := \max_{s,a \in \{S,A\}} \Big[\big| R^{c_i}(s,a) - R^{c_j}(s,a) \big| + W(h)\big(T^{c_i}(s,a), T^{c_j}(s,a)\big) \Big], \tag{1}$$

where $W(h)$ is the Wasserstein distance between transition probability distributions. This iterative update has a unique fixed point which is our metric d_{task}.

We can now show that the dynamics, reward, and optimal value function are all also Lipschitz with respect to d_{task}. The first two are clear results from Definition 4.

Corollary 1 (V_c^* is Lipschitz with respect to d_{task}) *Let V^* be the optimal, universal value function for a given discount factor γ and context space $\mathcal{C}$. Then V^* is Lipschitz continuous with respect to d_{task} with Lipschitz constant $\frac{1}{1-\gamma}$ for any $s \in \mathcal{S}$,*

$$|V^*(s,c) - V^*(s,c')| \leq \frac{1}{1-\gamma} d_{task}(c,c').$$

The proof can be found in the Appendix. Applying Theorem 1 to a continual RL setting assumes that the context be identifiable from a limited number of environment interactions.

Assumption 1 (Identifiability) *Let k be some constant number of steps the agent takes in a new environment with context c. There exists an $\epsilon_c > 0$ such that a context encoder ψ can take those transition tuples $(s_i, a_i, s_i', r_i), i \in \{1, ..., k\}$ and output a predicted context $\hat{c}$ that is ϵ_c-close to c.*

There are two key assumptions wrapped up in Assumption 1. The first is that the new environment is uniquely identifiable from k transitions, and the second is that we have a context encoder that can approximately infer that context. While this assumption can be strong for many high-dimensional environments where it may be difficult for a random policy to identify the environment within k steps, we are not trying to identify some ground truth context, but merely some notion of context as it affects the optimal policy in a new environment. Thus, given that we are deploying learned agents in these environments, by construction we only care about environment changes that affect that policy and are noticeable within k steps. In practice, we use neural networks for modeling ψ and verify that neural networks can indeed learn to infer the context, as shown in the Appendix.

Why do we care about the Lipschitz property? Xu and Mannor (2010) established that Lipschitz continuous functions are robust, i.e. the gap between test and training error is bounded. This result is only useful when the problem space is Lipschitz, which is often not the case in RL. However, we have shown that any BC-MDP is Lipschitz continuous with respect to metric d_{task}. We now define a general supervised learning setup to bound the error of learning dynamics and reward models. The following result requires that the data-collecting policy is ergodic, i.e. a Doeblin Markovian chain (Doob, 1953; Meyn and Tweedie, 1993), defined as follows.

Definition 5 (Doeblin chain) *A Markov chain $\{s_i\}_{i=1}^{\infty}$ on a state space $\mathcal{S}$ is a Doeblin chain (with α and t) if there exists a probability measure ρ on $\mathcal{S}$, $\alpha > 0$, an integer $t \geq 1$ such that*

$$P(s_t \in H | s_0 = s) \geq \alpha \rho(H); \quad \forall \text{ measurable } H \subseteq \mathcal{S}; \forall s \in \mathcal{S}.$$

Let $\hat{\mathcal{L}}(\cdot)$ denote expected error and $\mathcal{L}_{\text{emp}}(\cdot)$ denote training error of an algorithm $\mathcal{A}$ on training data $\mathbf{s} = \{s_1, ..., s_n\}$ and evaluated on points $z \in \mathcal{Z}$ sampled from distribution μ:

$$\hat{\mathcal{L}}(\cdot) := \mathbb{E}_{z\sim\mu}\mathcal{L}(\mathcal{A}_{\mathbf{s}}, z); \quad \mathcal{L}_{\text{emp}}(\cdot) := \frac{1}{n}\sum_{s_i \in \mathbf{s}} \mathcal{L}(\mathcal{A}_{\mathbf{s}}, s_i).$$

Here, $\mathcal{A}_{\mathbf{s}}$ denotes the instantiation of the learned algorithm trained on data $\mathbf{s}$ whereas $\mathcal{A}$ refers to the general learning algorithm. We can now bound the generalization gap using a result from Xu and Mannor (2010) using an additional assumption about the algorithm.

Theorem 1 (Generalization via Lipschitz Continuity (Xu and Mannor, 2010)) *If the test error, given a learning algorithm $\mathcal{A}$, is $\frac{1}{1-\gamma}$-Lipschitz and the training data $\mathbf{s} = \{s_1, ..., s_n\}$ are the first n outputs of a Doeblin chain with constants α, t, then for any $\delta > 0$ with probability at least $1 - \delta$,*

$$\left|\hat{\mathcal{L}}(\mathcal{A}_{\mathbf{s}}) - \mathcal{L}_{emp}(\mathcal{A}_{\mathbf{s}})\right| \leq \frac{\epsilon}{1-\gamma} + M\left(\frac{8t^2(K\ln 2 + \ln(1/\delta))}{\alpha^2 n}\right)^{1/4}.$$

K denotes the ϵ-covering number of the state space. ϵ controls the granularity at which we discretize, or partition, that space. If ϵ is larger, K is smaller. M is a scalar that uniformly upper-bounds the loss $\mathcal{L}$. Once we learn a smooth context space, this result bounds the generalization error of supervised learning problems like learned dynamics and reward models. These learned models allow us to construct a new MDP that is $\epsilon_R, \epsilon_T, \epsilon_c$-close to the original. We can now show how this error propagates when learning a policy.

Theorem 2 (Generalization Bound) *Without loss of generality we assume all tasks in a given BC-MDP family have reward bounded in $[0, 1]$. Given two tasks $\mathcal{M}_{c_i}$ and $\mathcal{M}_{c_j}$, we can bound the difference in Q^π between the two MDPs for a given policy π learned under an $\epsilon_R, \epsilon_T, \epsilon_{c_i}$-approximate abstraction of $\mathcal{M}_{c_i}$ and applied to $\mathcal{M}_{c_j}$,*

$$\left\|Q^\pi_{\mathcal{M}_{c_j}} - [Q^\pi_{\bar{\mathcal{M}}_{\hat{c}_i}}]_{\mathcal{M}_{c_j}}\right\|_\infty \leq \epsilon_R + \gamma\left(\epsilon_T + \epsilon_{c_i} + \|c_i - c_j\|_1\right)\frac{1}{2(1-\gamma)}.$$

Proof in the Appendix Theorem 2 shows that if we learn an ϵ-optimal context-conditioned policy for context c_i and encounter a new context c_j at evaluation time where c_j is close to c_i, then the context-conditioned policy will be ϵ-optimal for the new task by leveraging the Lipschitz property. While these results do not scale well with the dimensionality of the state space and discount factor γ, they show that representation learning is a viable approach to developing robust world models (Theorem 1), which translates to tighter bounds on the suboptimality of learned Q functions (Theorem 2).

5. Zero-shot Adaptation to Unknown Systems

Based on the findings in Section 4, we can improve generalization by constructing a context space that is Lipschitz with respect to the changes in dynamics and reward of the nonstationary environment. In practice, computing the maximum Wasserstein distance over the entire state-action space is computationally infeasible. We relax this requirement by taking the expectation over Wasserstein distance with respect to the marginal state distribution of the behavior policy. This leads us to a representation learning objective that leverages this relaxed version of the task metric in Definition 4:

$$\mathcal{L}(\phi, \psi, T, R) = MSE\underbrace{\left(\left|\left|\psi(H_1) - \psi(H_2)\right|\right|_2, d(c_1, c_2)\right)}_{\text{context loss}} + \underbrace{\mathcal{L}_{\mathcal{D}}(\phi, \psi, T, R)}_{\text{dynamics loss}} + \underbrace{\mathcal{L}_{\mathcal{R}}(\phi, \psi, T, R)}_{\text{reward loss}}, \tag{2}$$

$$\mathcal{L}_{\mathcal{D}}(\phi, \psi, T, R) = MSE\Big(T(\phi(o_t^{c_1}), a_t^{c_1}, \psi(H_1)), \phi(o_{t+1}^{c_1})\Big) + MSE\Big(T(\phi(o_t^{c_2}), a_t^{c_2}, \psi(H_2)), \phi(o_{t+1}^{c_2})\Big),$$

$$\mathcal{L}_{\mathcal{R}}(\phi, \psi, T, R) = MSE\Big(R(\phi(o_t^{c_1}), a_t^{c_1}, \psi(H_1)), r_{t+1}^{c_1}\Big) + MSE\Big(R(\phi(o_t^{c_2}), a_t^{c_2}, \psi(H_2)), r_{t+1}^{c_2}\Big).$$

where red indicates stopped gradients. $H_1 := \{o_t^{c_1}, a_t, r_t, o_{t+1}^{c_1}, ...\}$ and $H_2 := \{o_t^{c_2}, a_t, r_t, o_{t+1}^{c_2}, ...\}$ are transition sequences from two environments with contexts c_1 and c_2 respectively. During training, the transitions are uniformly sampled from a replay buffer. We do not require access to the true context for computing $d(c_1, c_2)$ (in Equation (2)) as we can approximate $d(c_1, c_2)$ using Definition 4. Specifically, we train a transition dynamics model and a reward model (via supervised learning) and use their output to approximate $d(c_1, c_2)$. In practice, we scale the context learning error, our task metric loss, using a scalar value denoted as α_ψ.

We describe the architecture of ZeUS in Figure 1. We have an observation encoder ϕ that encodes the pixel-observations into real-valued vectors. A buffer of interaction-history is maintained for computing the context. The context encoder first encodes the individual state-action transition pairs and then aggregates the representations using standard operations: *sum*, *mean*, *concate*, *product*, *min* and *max* [4]. All the components are instantiated using feedforward networks and trained end-to-end. During inference, assume that the agent is operating in some environment denoted by (latent) context c_1. At time t, the agent gets an observation $o_t^{c_1}$ which is encoded into $s_t^{c_1} := \phi(o_t^{c_1})$[5]. The context encoder ψ encodes the last k interactions (denoted as H_1) into a context encoding $c_1 := \psi(H_1)$[6]. The observation and context encodings are concatenated and fed to the policy to get the action.

During training, we sample a batch of interaction sequences from the buffer. For sake of exposition, we assume that we sample only 2 sequences H_1 and H_2. Similar to the inference pipeline, we compute $\phi(o_t^{c_1}), \psi(H_1), \phi(o_t^{c_2})$ and $\psi(H_2)$ and the loss (Equation (2)). We highlight that the algorithm does not know if the two (sampled) interactions correspond to the same context or not. Hence, in a small percentage of cases, H_1 and H_2 could correspond to the same context and the context loss will be equal to 0. For implementing the loss in equation Equation (2), we do not need access to the true context as the distance between the contexts can be approximated using the learned transition and reward models using Definition 4. The pseudo-code is provided in the Appendix. Since ZeUS is a representation learning algorithm, it must be paired with a policy optimization algorithm for end-to-end training. In the scope of this work, we use SAC-AE Yarats et al. (2021)), though ZeUS can be used with any policy optimization algorithm.

6. Experiments

We design our experiments to answer the following questions: **i)** How well does ZeUS perform when training over a family of tasks with varying dynamics? **ii)** Can ZeUS adapt and generalize to unseen environments (with novel dynamics or reward) without performing any gradient updates? (see Figure 2 and Figure 3), **iii)** Can ZeUS learning meaningful context representations when training over a family of tasks with varying dynamics? (see Figure 4)

6.1. Setup

Similar to the setups from Zhou et al. (2019); Lee et al. (2020); Zhang et al. (2021), we start with standard RL environments and extend them by modifying parameters that affect the dynamics (e.g.

4. We experiment with these aggregation operators for all the baselines and not just ZeUS.
5. We overload notation here since the true state space is latent.
6. We again overload notation here since the true context space is also latent.

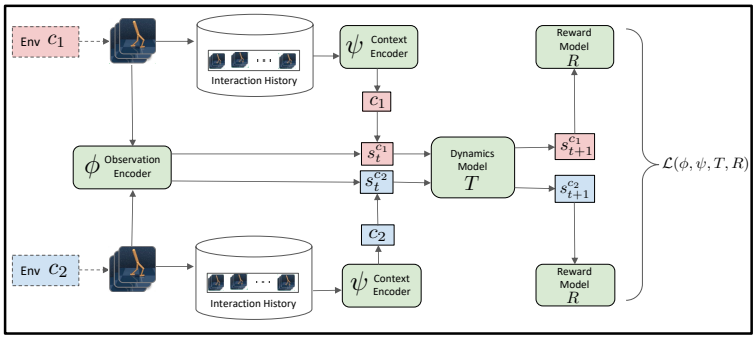


Figure 1: Proposed ZeUS algorithm. The components shown in green (i.e. observation encoder, context encoder, dynamics model and reward model) are shared across tasks. Components/representations in red or blue belong to separate tasks.

the friction between agent and the ground) or the reward (e.g. target velocity) such that they exhibit the challenging nonstationarity and rich-observation conditions of our BC-MDP setting. We create environments with varying transition dynamics by varying some physical properties in the following Mujoco (Todorov et al., 2012) based environments from the DM Control Suite (Tassa et al., 2018): Cheetah-Run-v0 (length of cheetah's torso), Walker-Walk-v0 (friction coefficient between walker and the ground), Walker-Walk-v1 (length of walker's foot) and Finger-Spin-v0 task (size of the finger). For environments with varying reward function, we use the Cheetah-Run-v1 environment (vary agent's target velocity) and Sawyer-Peg-v0 environment (vary the goal position) from Zhao et al. (2020) and assume access to the reward function, as done in Zhao et al. (2020).

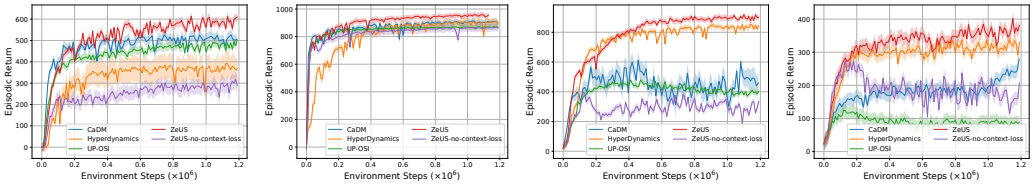
For all environments, we pre-define a range of parameters to train and evaluate on. For the case of nonstationary dynamics, we create two set of parameters for evaluation: *interpolation* (and *extrapolation*) where the parameters are sampled from a range that lies within (and outside) the range of parameters during training. For the case of nonstationary reward, we sample the parameters for the test environments from the same range as the training environments. We report the evaluation performance of the best performing hyper-parameters for all algorithms (measured using the training performance). We run all experiments with 10 seeds and report both mean and standard error (denoted by the shaded area on the plots).

6.2. Baselines

We select representative baselines from different areas of related work (Section 2): *UP-OSI* (Yu et al., 2017) is a system identification approach that infers the true parameters and conditioning the policy on the inferred parameters. *Context-aware Dynamics Model , CaDM* (Lee et al., 2020) is a context modelling based approach that outperforms Gradient and Recurrence-based meta learning approaches (Nagabandi et al., 2019b). *HyperDynamics*(Xian et al., 2021) generates the weights of the dynamics model (for each environment) by conditioning on a context vector and is shown to outperform both ensemble of experts and meta-learning based approaches (Nagabandi et al., 2019b). We also consider a *Context-conditioned Policy* where the context encoder is trained using the one-step forward dynamics loss. This approach can be seen as an ablation of the ZeUS algorithm without the context learning error (from Equation (2)). We refer to it as *Zeus-no-context-loss*.

6.3. Adapting and generalizing to unseen environments

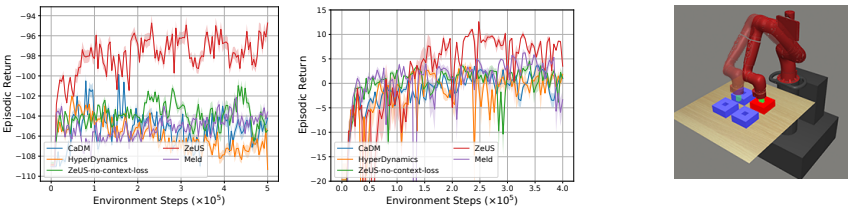
In Figure 2, we compare ZeUS's performance on the heldout *extrapolation* evaluation environments which the agent has not seen during training. The transition dynamics varies across these tasks. *Hy-*



(*a*) Cheetah-Run-v0 (*b*) Finger-Spin-v0 (*c*) Walker-Walk-v0 (*d*) Walker-Walk-v1

Figure 2: We compare the performance of the proposed ZeUS algorithm with *CaDM*, *UP-OSI*, *HyperDynamics* and *ZeUS-no-context-loss* algorithms on the heldout evaluation environments (extrapolation) for four families of tasks with different dynamics parameters.

perDynamics performs well on some environments but requires more resources to train (given that it generates the weights of dynamics models for each transition in the training batch). *UP-OSI* uses privileged information (in terms of the extra supervision). Both *CaDM* and ZeUS are reasonably straightforward to implement (and train) though ZeUS outperforms the other baselines. The context loss (Equation (2)) is an important ingredient for the generalization performance as observed by the performance of *Zeus-no-context-loss*. The corresponding plots for performance on the training environments and heldout *interpolation* evaluation environments are given in the Appendix. For additional ablation results for these environments, refer to the Appendix.



(*a*) Cheetah-Run-v1 (*b*) Sawyer-Peg-v0 (*c*) Sawyer-Peg-v0

Figure 3: (a), (b): We compare the performance of the proposed ZeUS algorithm with *CaDM*, *HyperDynamics*, *ZeUS-no-bisim* and *Meld* algorithms on environments with different reward functions. (c): Illustration of the Sawyer-Peg-V0 task.

In Figure 3, we compare ZeUS's performance with the baselines when the reward function varies across tasks. Since all the models have access to the reward, we do not compare with UP-OSI which is trained to infer the reward. Instead we include an additional baseline, Meld (Zhu et al., 2020), a meta-RL approach that performs inference in a latent state model to adapt to a new task. Like before, ZeUS outperforms the other baselines.

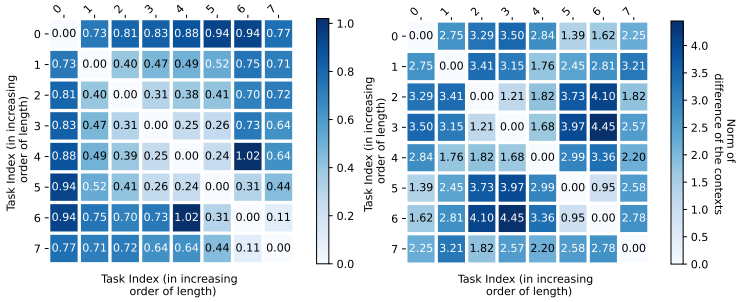


Figure 4: Norm of pairwise differences of contexts for different tasks for Cheetah-Run-v0 setup when trained with context loss (left) and without context loss (right).

6.4. Learning a Meaningful Context Representation

We want to evaluate if the context representation constructed by ZeUS contains meaningful information about the true context of the BC-MDP. We compute the norm of pairwise difference of the learned contexts (corresponding to different tasks) for the Cheetah-Run-v0 setup (length of cheetah's torso varies across the tasks) when it is trained with and without the context loss (Equation (2)). As shown in Figure 4 (left), when training with the context loss, tasks that are closer in terms of torso length are generally closer in the context space. We also report the Spearman's rank correlation coefficient between the ranking (of distance) between the learned contexts and the ground truth context. Training with context loss results in a much higher correlation (0.60) than training without(0.23), showing that the context loss is useful for capturing relationship across tasks.

7. Limitations

A theoretical limitation of this work is the inability to provide guarantees based on the likelihood of the model learning the correct causal dynamics. By structuring the context space to be Lipschitz, we can give guarantees only for those dynamics and reward where the context is close to the contexts seen at training time. While this result flows directly from Theorem 2, it is important to be aware of this limitation, namely that ZeUs may have poor performance when the distance between the training and evaluation contexts is high. We demonstrate an example in Figure 5, where we plot the performance of the agent for different values of target velocities (for Cheetah-Run-v1). While ZeUS outperforms the other methods, its performance also degrades as we move away from the training distribution.

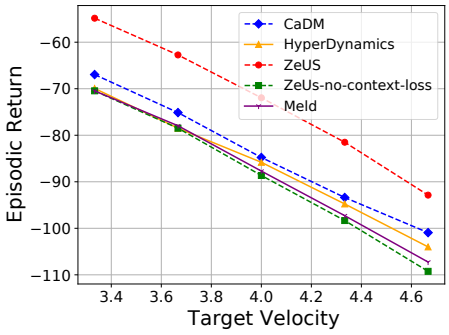


Figure 5: The performance of all the algorithms (on Cheetah-Run-v1) degrades as we move away from the training distribution.

Empirically, the performance of our algorithm also relies on dense reward signal to distinguish across tasks. However, many real world environments do not naturally have dense reward. One simple extension of our method to mitigate this issue in sparse reward environments is to use a learned value function as a dense reward substitute.

8. Discussion

In this work, we propose to use the Block Contextual MDP framework to model the nonstationary, rich observation, RL setting. We provide theoretical bounds on adaptation and generalization ability to unseen tasks within this framework and propose a representation learning algorithm (ZeUS) for performing online inference of "unknown unknowns". We empirically verify the effectiveness of ZeUS on environments with nonstationary dynamics and reward functions.

There are several interesting directions for further research. One way to tighten the generalization bounds is by constraining the neural networks used in ZeUS to have smaller Lipschitz constants. This is known to be able to improve generalization bounds (Neyshabur et al., 2015). We can also consider improving the algorithm to infer the underlying causal structure of the dynamics, as discussed in Section 7. This is a much harder problem than constructing a context space and inferring context in new environments. Another direction to extend ZeUS is to account for *active nonstationarity*, where the agent's actions can affect the environment. ZeUS would work for this setting, but there is clearly an additional structure that can be leveraged for improved performance.

References

Anurag Ajay, Maria Bauza, Jiajun Wu, Nima Fazeli, Joshua B Tenenbaum, Alberto Rodriguez, and Leslie P Kaelbling. Combining physical simulators and object-based networks for control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3217–3223. IEEE, 2019.

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *ArXiv preprint*, abs/1908.04742, 2019. URL https://arxiv.org/abs/1908.04742.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *ArXiv preprint*, abs/1902.10486, 2019. URL https://arxiv.org/abs/1902.10486.

Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9355–9366, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/b8cfbf77a3d250a4523ba67a65a7d031-Abstract.html.

Alessandro Chiuso and Gianluigi Pillonetto. System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:281–304, 2019.

J. L. Doob. *Stochastic processes*. John Wiley & Sons, New York, 1953. MR 15,445b. Zbl 0053.26802.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 2019. URL http://proceedings.mlr.press/v97/du19b.html.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. URL http://proceedings.mlr.press/v70/finn17a.html.

João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), 2014. ISSN 0360-0300. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 2020.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR, 2019. URL http://proceedings.mlr.press/v97/hafner19a.html.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes, 2015.

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *ArXiv preprint*, abs/1905.06424, 2019. URL https://arxiv.org/abs/1905.06424.

Maximilian Igl, Luisa M. Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2122–2131. PMLR, 2018. URL http://proceedings.mlr.press/v80/igl18a.html.

Khurram Javed and Martha White. Meta-learning representations for continual learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1818–1828, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f4dd765c12f2ef67f98f3558c282a9cd-Abstract.html.

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.

Pierre-Alexandre Kamienny, Matteo Pirotta, Alessandro Lazaric, Thibault Lavril, Nicolas Usunier, and Ludovic Denoyer. Learning adaptive exploration strategies in dynamic environments through informed policy regularization. *ArXiv preprint*, abs/2005.02934, 2020. URL https://arxiv.org/abs/2005.02934.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2502–2511. PMLR, 2018. URL http://proceedings.mlr.press/v80/kaplanis18a.html.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *ArXiv preprint*, abs/2012.13490, 2020. URL https://arxiv.org/abs/2012.13490.

Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 5757–5766. PMLR, 2020.

Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2009.12.001. URL https://www.sciencedirect.com/science/article/pii/S1367578810000027.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6467–6476, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. URL /brokenurl#probability.ca/MT.

Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 829–838. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/modi20a.html.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019a. URL https://openreview.net/forum?id=HyztsoC5Y7.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=HyztsoC5Y7.

Anusha Nagabandi, Chelsea Finn, and Sergey Levine. Deep online learning via meta-learning: Continual adaptation for model-based RL. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019c. URL https://openreview.net/forum?id=HyxAfnA5tm.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1376–1401. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Neyshabur15.html.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2778–2787. PMLR, 2017. URL http://proceedings.mlr.press/v70/pathak17a.html.

Martin L Puterman. Markov decision processes: Discrete stochastic dynamic programming. *Journal of the Operational Research Society*, 1995.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5331–5340. PMLR, 2019. URL http://proceedings.mlr.press/v97/rakelly19a.html.

Mark Bishop Ring et al. *Continual learning in reinforcement environments*. PhD thesis, University of Texas at Austin Austin, Texas 78712, 1994.

J.J.E. Slotine and W. Li. *Applied Nonlinear Control*. Prentice Hall, 1991. ISBN 978-0-13-040890-7. URL https://books.google.com/books?id=cwpRAAAAMAAJ.

Shagun Sodhani, Sarath Chandar, and Yoshua Bengio. Toward Training Recurrent Neural Networks for Lifelong Learning. *Neural Computation*, 32(1):1–35, 2020. ISSN 0899-7667. doi: 10.1162/neco_a_01246. URL https://doi.org/10.1162/neco_a_01246.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9767–9779. PMLR, 2021. URL http://proceedings.mlr.press/v139/sodhani21a.html.

Jan Swevers, Chris Ganseman, D Bilgin Tukel, Joris De Schutter, and Hendrik Van Brussel. Optimal robot excitation and identification. *IEEE transactions on robotics and automation*, 13(5):730–740, 1997.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. Technical report, DeepMind, 2018.

Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.

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*, pages 5026–5033. IEEE, 2012.

Peter Van Overschee and BL De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

Zhou Xian, Shamit Lal, Hsiao-Yu Tung, Emmanouil Antonios Platanios, and Katerina Fragkiadaki. Hyperdynamics: Generating expert dynamics models by observation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pHXfe1cOmA.

Annie Xie, James Harrison, and Chelsea Finn. Deep reinforcement learning amidst lifelong non-stationarity, 2020.

Huan Xu and Shie Mannor. Robustness and generalization. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 503–515. Omnipress, 2010. URL http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=511.

Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. *ArXiv preprint*, abs/1906.03853, 2019. URL https://arxiv.org/abs/1906.03853.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *AAAI*, 2021.

Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. In *Robotics: Science and Systems*, 2017.

L Zadeh. On the identification problem. *IRE Transactions on Circuit Theory*, 3(4):277–281, 1956.

Amy Zhang, Zachary C. Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments. *The Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2019.

Amy Zhang, Shagun Sodhani, Khimya Khetarpal, and Joelle Pineau. Learning robust state abstractions for hidden-parameter block mdps. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=fmOOI2a3tQP.

Tony Z. Zhao, Anusha Nagabandi, Kate Rakelly, Chelsea Finn, and Sergey Levine. Latent state models for meta-reinforcement learning from images. In *4th Annual Conference on Robot Learning, CoRL 2020, Proceedings*, Proceedings of Machine Learning Research. PMLR, 2020.

Wenxuan Zhou, Lerrel Pinto, and Abhinav Gupta. Environment probing interaction policies. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryl8-3AcFX.

Shaojun Zhu, Andrew Kimmel, Kostas E. Bekris, and Abdeslam Boularias. Fast model identification via physics engines for data-efficient policy search. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3249–3256. ijcai.org, 2018. doi: 10.24963/ijcai.2018/451. URL https://doi.org/10.24963/ijcai.2018/451.

Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *ArXiv preprint*, abs/2009.07888, 2020. URL https://arxiv.org/abs/2009.07888.

Luisa M. Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7693–7702. PMLR, 2019. URL http://proceedings.mlr.press/v97/zintgraf19a.html.

Karl Johan Åström and Torsten Bohlin. Numerical identification of linear dynamic systems from normal operating records. In *Proc. IFAC Conference on Self-Adaptive Control Systems*, 1965.