An Information-Theoretic Analysis of OOD Generalization in Meta-Reinforcement Learning

Xingtu Liu
Simon Fraser University
rltheory@outlook.com

Abstract

In this work, we study out-of-distribution (OOD) generalization in meta-reinforcement learning from an information-theoretic perspective. We begin by establishing OOD generalization bounds for meta-supervised learning under two distinct distribution shift scenarios: standard distribution mismatch and a broad-to-narrow training setting. Building on this foundation, we formalize the generalization problem in meta-reinforcement learning and establish fine-grained generalization bounds that exploit the structure of Markov Decision Processes. Lastly, we analyze the generalization performance of a gradient-based meta-reinforcement learning algorithm.

1 Introduction

Meta-learning [Thrun and Pratt, 1998, Hospedales et al., 2021], or "learning to learn", is the study of algorithms that leverage prior experience across tasks to rapidly adapt to new tasks. By formalizing the process of learning from prior tasks, meta-learning provides a principled framework for developing models that are flexible, data-efficient, and robust to new environments. Since the goal of meta-learning is to generalize to new or even unseen tasks, it is crucial to develop a theoretical understanding of its generalization behavior. In reality, training tasks often suffer from selection bias and distribution shift over time. This challenge, known as out-of-distribution (OOD) generalization, raises fundamental questions about how prior experience should be leveraged when the testing environment deviates from the training environment. In this work, we address these questions by leveraging information-theoretic tools.

Building on our analysis of meta-supervised learning, we extend the study of OOD generalization to meta-reinforcement learning (meta-RL) [Gupta et al., 2018, Nagabandi et al., 2018, Yu et al., 2020, Beck et al., 2025]. Meta-RL adapts the meta-learning framework to sequential decision-making, where tasks are modeled as Markov decision processes (MDPs) and the task-specific objective is the cumulative reward. In real-world deployments (e.g., robotics, human interaction), agents often face shifts in rewards, dynamics, or state distributions beyond those seen in meta-training. Thus, understanding OOD generalization in meta-RL is crucial for ensuring reliable adaptation and decision-making [Beck et al., 2025]. Prior works such as Rimon et al. [2022] and Tamar et al. [2022] analyzed in-distribution generalization in meta-RL, while Simchowitz et al. [2021] provides a PAC-Bayesian analysis for Bayesian meta-RL. In contrast, our work develops a general information-theoretic analysis for OOD generalization in meta-RL and can be applied to subtask problems and gradient-based meta-RL algorithms. A key advantage of our framework lies in its ability to decouple the sources of OOD shift. Prior PAC-Bayesian analyses Simchowitz et al. [2021] bound generalization error using the total variation distance between global priors, which treats the whole environment as a black box. Our approach leverages the underlying structure of the MDPs, explicitly capturing the divergence between transition kernels, initial state distributions, and rewards.

Preprint.

Following the seminal work of Russo and Zou [2016] and Xu and Raginsky [2017], an information-theoretic framework has been developed to bound the generalization error of learning algorithms using the mutual information between the input dataset and the output hypothesis. This methodology formalizes the intuition that overfitted learning algorithms are less likely to generalize effectively. Unlike classical complexity-based approaches such as VC dimension or Rademacher complexity, the information-theoretic framework captures all aspects of the learning process, including the data distribution, hypothesis space, and learning algorithm. We therefore employ this framework to study generalization in meta-(reinforcement) learning under shifts between training and testing task distributions. In particular, we make the following contributions.

- In Section 3.1, we study the distribution mismatch setting, where the test environment differs from the training environment. In this case, we derive an upper bound $O\left(\mathrm{D} + \frac{\mathrm{MI}}{nm}\right)$ on the OOD generalization error, where D quantifies the environment mismatch, MI denotes the mutual information, n is the number of training tasks, and m is the number of samples per task.
- In Section 3.2, we consider the subtask setting, in which the meta-algorithm is trained on a broader environment. Here, the test environment is assumed to consist of tasks that form a strict subset of the training environment. The resulting generalization bound indicates that merely increasing the number of training tasks that are weakly related to the targeting tasks provides limited benefit.
- In Section 4, we extend our analysis to meta-RL, where both the hypothesis space and the objective differ from meta-supervised learning. We first extend the generalization bounds in Section 3.1 with a finer-grained analysis. The divergence term captures the distributional shifts across the initial state, transition dynamics, and reward function. To address the potential unboundedness issue of this divergence, we extend the analysis from Section 3.2 to the meta-RL setting. Finally, we derive a generalization bound for a two-level gradient-based meta-RL algorithm.

2 Problem Formulation

This section introduces preliminaries and problem formulation for meta-learning and meta-RL.

2.1 Preliminaries

Random variables are denoted by capital letters (e.g., X and Y), and their realizations are denoted by lowercase letters (e.g., x and y). For random variables X and Y, $P_{X,Y}$ denotes their joint distribution, P_X denotes the marginal distribution, and $P_{X|Y}$ denotes the conditional distribution. A random variable X is said to be σ-sub-Gaussian if $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$. For probability measures μ and ν, we use $D(\mu\|\nu)$ to denote the Kullback-Leibler (KL) divergence of μ with respect to ν. The mutual information between X and Y is defined as: $I(X;Y) := D(P_{X,Y}\|P_X \otimes P_Y)$. The conditional mutual information is defined as: $I(X;Y|Z) := \mathbb{E}_Z[D(P_{X,Y|Z}\|P_{X|Z} \otimes P_{Y|Z})]$. We denote by $\mathrm{cov}(X)$ and $H(X)$ the covariance matrix and entropy of a random variable X. For a matrix M, $\det(M)$ denotes its determinant. We use big-O notation to characterize the asymptotic upper bound of a function's growth rate.

In supervised learning, given a training dataset $Z = \{Z_i\}_{i=1}^n \sim \mu^{\otimes n}$, a learning algorithm $\mathcal{A}$ outputs a hypothesis $W = \mathcal{A}(Z)$ from the hypothesis space $\mathcal{W}$. Let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ denote the loss function. We define the empirical risk of W as $L(W) := \frac{1}{n}\sum_{i=1}^n \ell(W, Z_i)$, and define the in-distribution population risk as $L_\mu(W) := \mathbb{E}_{Z\sim\mu}[\ell(W, Z)]$. The expected in-distribution generalization error is defined as $\mathbb{E}_{Z\sim\mu}[L_\mu(W) - L(W)]$. Note that W is a function of Z. When the training distribution μ differs from the testing distribution ν, we have the OOD population risk $L_\nu(W) := \mathbb{E}_{Z'\sim\nu}[\ell(W, Z')]$. The expected OOD generalization is then defined as $\mathbb{E}_{Z\sim\mu}[L_\nu(W) - L(W)]$. In the following sections, we formally define OOD generalization for meta-learning and meta-RL, which is more intricate than in supervised learning.

2.2 Meta-Learning

In this work, we consider learning tasks sampled i.i.d. from the training environment $\mathcal{T}$. For a given task $\tau_i \sim \mathcal{T}$, let $Z_i = \{S_{i,j}\}_{j=1}^m \sim {\tau_i}^{\otimes m}$ denote its in-task training dataset. We assume each task is associated with a dataset of the same size m. Suppose there are n training tasks. We denote the meta-training dataset by $Z_{1:n} = \{Z_i\}_{i=1}^n$ and the task-specific parameters by $W = \{W_i\}_{i=1}^n$. The goal of meta-learning is to learn a meta-parameter θ that captures knowledge shared across tasks. Given θ and a new task, the meta-learner can adapt more efficiently and produce a task-specific hypothesis. During training, the meta-learner aims to minimize the empirical meta-risk

$$L_{Z_{1:n}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W_i \sim P_{W_i|Z_i,\theta}}[L(W_i, Z_i)] := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W_i \sim P_{W_i|Z_i,\theta}}\left[\frac{1}{m}\sum_{j=1}^{m} \ell(W_i, S_{i,j})\right]$$

where $L(W)$ denotes the task-specific empirical risk associated with parameter W, and $P_{W|Z,\theta}$ represents the distribution over task-specific parameters induced by the base-learner. The population meta-risk with respect to $\mathcal{T}$ is defined as

$$L_{\mathcal{T}}(\theta) := \mathbb{E}_{\tau\sim\mathcal{T}}\mathbb{E}_{Z\sim\tau^{\otimes m}}\mathbb{E}_{W\sim P_{W|Z,\theta}}[L_\tau(W)] := \mathbb{E}_{\tau\sim\mathcal{T}}\mathbb{E}_{Z\sim\tau^{\otimes m}}\mathbb{E}_{W\sim P_{W|Z,\theta}}\left[\mathbb{E}_{S\sim\tau}\ell(W,S)\right].$$

The quality of the learned meta-parameter θ can be evaluated via the (in-distribution) meta generalization error

$$\text{gen}_{\text{id}} := \mathbb{E}_{\theta, Z_{1:n}}[L_{Z_{1:n}}(\theta) - L_{\mathcal{T}}(\theta)].$$

However, in practice, the testing environment typically differs from the training environment, and the essence of meta-learning is to enable fast adaptation to new tasks. Thus, we focus on the OOD meta generalization error. Let $\mathcal{U}$ denote the targeting task environment and let $\mu \sim \mathcal{U}$ represent a test task,

$$\text{gen}_{\text{ood}} := \mathbb{E}_{\theta, Z_{1:n}}[L_{Z_{1:n}}(\theta) - L_{\mathcal{U}}(\theta)]$$

where

$$L_{\mathcal{U}}(\theta) = \mathbb{E}_{\mu\sim\mathcal{U}}\mathbb{E}_{Z\sim\mu^{\otimes m}}\mathbb{E}_{W\sim P_{W|Z,\theta}}\left[\mathbb{E}_{S\sim\mu}\ell(W,S)\right].$$

2.3 Meta-Reinforcement Learning

In supervised learning, the goal is to find a hypothesis that minimizes the empirical risk, whereas in standard RL, the goal is to find a policy that maximizes the cumulative reward. In meta-RL, the learner seeks to learn a meta-algorithm f_θ, parameterized by a meta-parameter θ, which serves as an adaptation rule. Given a new task, f_θ produces a task-specific policy π_ϕ parameterized by ϕ that aims to maximize the cumulative reward for that task.

The tasks in meta-RL are modeled as MDPs. An MDP is defined by $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, \rho, r, \gamma, H)$. We use $\Delta(\mathcal{X})$ to denote the set of the probability distribution over the set $\mathcal{X}$. $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, \rho, r, \gamma, H)$ is specified by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition function $\mathcal{P} : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, initial state distribution ρ, reward function $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, discount factor $\gamma \in [0, 1]$, and the horizon H. In this work, we consider a finite horizon for simplicity. We assume the reward is bounded, i.e. $r(s, a) \in [0, 1]$[1], $\forall(s, a)$.

Assume there are n MDPs $\{\mathcal{M}_i\}_{i=1}^n$ sampled i.i.d. from the training environment $\mathcal{T}$. Let $\phi = \{\phi_i\}_{i=1}^n$ denote the task-specific parameters. A key distinction between meta-RL and meta-supervised learning is that, in RL, we do not collect hypothesis-independent datasets. Instead, the data for each task consists of trajectories whose distribution depends on the task-specific parameters. Specifically, given ϕ and an MDP $\mathcal{M}$, a trajectory $\omega = \{s_h, a_h, r_h, s_{h+1}\}_{h=0}^H$ is distributed according to

$$P_{\omega|\mathcal{M},\phi} = \rho(s_0)\prod_{h=0}^{H} \pi_\phi(a_h|s_h)\mathcal{P}(s_{h+1}|s_h, a_h) \tag{1}$$

where π_ϕ is the task-specific policy, and ρ and $\mathcal{P}$ are the initial state distribution and transition kernel of $\mathcal{M}$. The meta-learner aims to maximize the empirical meta-RL objective

$$J_{\mathcal{M}_{1:n}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\phi_i\sim P_{\phi|\mathcal{M}_i,\theta}}[J(\pi_{\phi_i}, \mathcal{M}_i)] := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\phi_i\sim P_{\phi|\mathcal{M}_i,\theta}}\left[\mathbb{E}_{\omega\sim P_{\omega|\mathcal{M}_i,\phi_i}}\left[\sum_{j=0}^{H}\gamma^j r_j\right]\right] \tag{2}$$

[1] For rewards in $[R_{\min}, R_{\max}]$ simply rescale these bounds.

where $J(\pi_{\phi_i})$ denotes the expected discounted return of policy π_ϕ. Let $\mathcal{U}$ denote the targeting (testing) task environment. The population meta-RL objective is defined as

$$J_{\mathcal{U}}(\theta) := \mathbb{E}_{\mathcal{M}\sim\mathcal{U}}\mathbb{E}_{\phi\sim P_{\phi|\mathcal{M},\theta}}[J_{\mathcal{M}}(\pi_\phi)] := \mathbb{E}_{\mathcal{M}\sim\mathcal{U}}\mathbb{E}_{\phi\sim P_{\phi|\mathcal{M},\theta}}\left[\mathbb{E}_{\omega\sim P_{\omega|\mathcal{M},\phi}}\left[\sum_{j=0}^{H}\gamma^j r_j\right]\right].$$

Finally, the OOD meta-RL generalization error is given by

$$\text{gen}_{\text{ood}} := \mathbb{E}_{\theta,\mathcal{M}_{1:n}}[J_{\mathcal{M}_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)].$$

3 Out-of-Distribution Generalization in Meta-Learning

In this section, we investigate OOD generalization in meta-learning under two settings. The first is the mismatch setting, where the training environment $\mathcal{T}$ differs from the testing environment $\mathcal{U}$. The second is the subtask setting, where the tasks comprising $\mathcal{U}$ form a strict subset of those in $\mathcal{T}$. This latter case corresponds to a common meta-learning strategy which trains across a wide range of tasks to distill shared knowledge. For the first setting, we use mutual information to derive bounds on the OOD generalization error, while for the second setting, we apply conditional mutual information.

3.1 OOD Generalization in Meta-Learning via Mutual Information

We now present the main result. Let $P_{Z_{1:n}}$ denote the distribution of the meta-dataset under the training environment $\mathcal{T}$, and $Q_{Z_{1:n}}$ the corresponding distribution under the testing environment $\mathcal{U}$.

Theorem 1. *Suppose the loss function ℓ is σ-sub-Gaussian for any meta parameter θ, hypothesis W_i, and dataset Z_i. The OOD meta generalization error is upper-bounded by*

$$\mathbb{E}_{\theta,Z_{1:n}}[L_{Z_{1:n}}(\theta) - L_{\mathcal{U}}(\theta)] \leq \sqrt{\frac{2\sigma^2(I(\theta,W_{1:n};Z_{1:n}) + D(P_{Z_{1:n}}\|Q_{Z_{1:n}}))}{nm}}.$$

When the training and testing environments are identical, we have $D(P_{Z_{1:n}}\|Q_{Z_{1:n}}) = 0$, and the bound reduces to the standard in-distribution generalization bound. When the sampled tasks are i.i.d., we have $\frac{D(P_{Z_{1:n}}\|Q_{Z_{1:n}})}{nm} = \frac{D(P_Z\|Q_Z)}{m}$. Moreover, when the in-task samples are i.i.d., it follows that $\frac{D(P_Z\|Q_Z)}{m} = D(\tau\|\mu)$. Thus, strong generalization guarantees hold only if the distribution mismatch term is sufficiently small. By the chain rule of mutual-information, the bound can be further written as

$$\mathbb{E}_{\theta,Z_{1:n}}[L_{Z_{1:n}}(\theta) - L_{\mathcal{U}}(\theta)] \leq \sqrt{2\sigma^2 D(\tau\|\mu)} + \sqrt{\frac{2\sigma^2 I(\theta;Z_{1:n})}{nm}} + \sqrt{\frac{2\sigma^2\sum_{i=1}^n I(W_i;Z_i|\theta)}{nm}}.$$

The first term accounts for the distribution mismatch, the second term reflects the environmental uncertainty, and the third term reflects the task-level uncertainty.

Instead of mutual information, one may also upper bound the generalization error using the Wasserstein or total variation distance when the loss function is bounded or Lipschitz continuous [Lopez and Jog, 2018, Wang and Mao, 2022, Liu et al., 2025]. We leave this extension for future work.

3.2 Subtask Generalization in Meta-Learning via Conditional Mutual Information

We now consider another setting in which the model is trained on a broad set of tasks but tested on a specific subset of them. This "broad-to-narrow" scenario results in a distribution shift arising from transitioning from a broad environment to a narrower one. We denote $\hat{\mathcal{T}}$ as the testing environment consisting of fewer tasks than $\mathcal{T}$. In this case, one could in principle apply the previous results by viewing $\mathcal{U}$ as $\hat{\mathcal{T}}$. However, in this case, the term $D(\tau|\mu)$ introduces a persistent bias, regardless of the sample size. To address this issue, we derive a new upper bound based on conditional mutual information [Steinke and Zakynthinou, 2020, Laakom et al., 2024].

The conditional mutual information approach was introduced by Steinke and Zakynthinou [2020], which normalizes the information content of each data point. Let $Z = \{Z_i^{\pm}\}_{i=1}^n$ consist of $2n$ samples drawn independently from P_Z. Let $U = \{U_i\}_{i=1}^n \in \{-1,1\}^n$ be uniformly random and independent

from Z. Denote $Z_i^{U_i}$ as the training sample selected from $Z_i^{\pm}$, and denote $Z_i^{-U_i}$ as the testing sample. Similarly, let $W = \{W_i^{\pm}\}_{i=1}^n$ denote the collection of task-specific hypotheses obtained by training on Z given the meta-parameter θ. Next, we define the subtask meta generalization error [Laakom et al., 2024] as

$$\text{gen}_{\text{sub}} := \mathbb{E}_{\hat{\tau}\sim\hat{\mathcal{T}}}\mathbb{E}_{\theta,Z_{1:n}}\left[\frac{1}{n_{\hat{\tau}}}\sum_{i=1}^{n}\mathbb{E}_{U_i}\left[\mathbb{1}\{\tau_i^{U_i}=\hat{\tau}\}\mathbb{E}_{W_i^{U_i}|\theta,Z_i^{U_i}}L(W_i^{U_i},Z_i^{U_i})\right.\right.$$
$$\left.\left.-\,\mathbb{1}\{\tau_i^{-U_i}=\hat{\tau}\}\mathbb{E}_{W_i^{-U_i}|\theta,Z_i^{-U_i}}L(W_i^{-U_i},Z_i^{-U_i})\right]\right] \tag{3}$$

where $L(W_i^{U_i}, Z_i^{U_i}) = \frac{1}{m}\sum_{j=1}^{m}\ell(W_i^{U_i}, S_{i,j}^{U_i})$.

The subtask generalization error gen_{sub} measures the expected difference in empirical training performance when adapting the meta-parameter to either of the paired datasets (Z_i^+ or Z_i^-), restricted to the target subtask environment. It quantifies how sensitive the empirical performance is to the specific data sample drawn for a task within the target subtask distribution. This definition is similar to those in the super-sample setting [Zhou et al., 2022], but is adapted for meta-learning.

Theorem 2. *Suppose the loss function ℓ is σ-sub-Gaussian. The subtask meta generalization error defined in eq. (3) is upper-bounded by*

$$\text{gen}_{\text{sub}} \le \mathbb{E}_{\hat{\tau}\sim\hat{\mathcal{T}}}\mathbb{E}_{Z_{1:n}}\left[\frac{1}{n_{\hat{\tau}}}\sum_{i=1}^{n}\sqrt{\frac{2\sigma^2 I(f_i;U_i|Z_{1:n})}{m}}\right]$$

where

$$f_i := \mathbb{1}\{\tau_i^- = \hat{\tau}\}\mathbb{E}_{W_i^-|\theta,Z_i^-}L(W_i^-,Z_i^-) - \mathbb{1}\{\tau_i^+ = \hat{\tau}\}\mathbb{E}_{W_i^+|\theta,Z_i^+}L(W_i^+,Z_i^+).$$

The conditional mutual information $I(f_i; U_i|Z_{1:n})$ quantifies how much information the choice of dataset (Z_i^+ vs. Z_i^-) provides about the resulting difference in expected empirical performance f_i on the target task. A lower conditional mutual information suggests a more stable adaptation process relevant to the target task.

This bound reflects the intuition behind meta-learning that more data per task helps. When m is sufficiently large, the contribution of meta-information becomes less significant. Besides, if the number of training tasks n increases without a corresponding rise in tasks that are closely related to the targeting distribution, the generalization bound may become even worse, as n grows while $n_{\hat{\tau}}$ remains fixed. In such cases, although the averaged mutual information $\frac{1}{n}\sum_{i=1}^{n}\sqrt{I(f_i;U_i|Z_{1:n})}$ may decrease, the term $\frac{1}{n_{\hat{\tau}}}\sum_{i=1}^{n}\sqrt{I(f_i;U_i|Z_{1:n})}$ can increase if the trade-off between enlarging the training set and improving the meta-parameter is not properly balanced. In contrast, when additional training tasks are collected and the number of samples per class of task increases proportionally, the model can better distill shared structure across tasks, allowing the meta-information to help generalization.

4 Out-of-Distribution Generalization in Meta-Reinforcement Learning

Meta-RL provides a framework for learning transferable priors across RL tasks. Building on our OOD generalization analysis in meta-supervised learning, we now extend these results to meta-RL. The meta-RL setting introduces two key distinctions: the objective shifts from empirical loss minimization to maximizing expected cumulative reward, and the data is no longer a fixed, hypothesis-independent dataset. Instead, it consists of trajectories generated through the agent's own interactions with the MDPs, making the data distribution inherently dependent on the learning parameters. We begin in Section 4.1 by deriving general OOD bounds for meta-RL under both the distribution mismatch and subtask settings. We then apply this framework in Section 4.2 to analyze the generalization performance of a specific gradient-based meta-RL algorithm. Finally, in Section 4.3, we demonstrate how these generalization error bounds can be directly related to the suboptimality gap in the target task environment, considering both standard and offline meta-RL settings.

4.1 Generalization Bounds for Meta-Reinforcement Learning

We first present the following result, which can be viewed as an application of Theorem 1 to the meta-RL setting.

Theorem 3. *The OOD meta-RL generalization error is upper-bounded by*

$$\mathbb{E}_{\theta,\mathcal{M}_{1:n}}[J_{\mathcal{M}_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)] \leq \sqrt{\frac{2I(\theta,\phi_{1:n};\mathcal{M}_{1:n}) + 2D(P_{\mathcal{M}_{1:n}}\|Q_{\mathcal{M}_{1:n}})}{n(1-\gamma)^2}}.$$

As before, this bound can be further decomposed as follows

$$\mathbb{E}_{\theta,\mathcal{M}_{1:n}}[J_{\mathcal{M}_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)] \leq \sqrt{\frac{2D(P_{\mathcal{M}}\|Q_{\mathcal{M}})}{(1-\gamma)^2}} + \sqrt{\frac{2I(\theta;\mathcal{M}_{1:n})}{n(1-\gamma)^2}} + \sqrt{\frac{2\sum_{i=1}^{n} I(\phi_i;\mathcal{M}_i|\theta)}{n(1-\gamma)^2}}. \tag{4}$$

The first term measures the distribution mismatch between the training and target environments, which vanishes when there is no distribution shift. This term can be further decomposed into three parts, which measure the divergence between the initial state distribution, transition kernel, and reward.

Lemma 4. *The KL divergence between two environments decomposes into the respective KL divergences of their initial state distributions, transition kernels, and reward functions:*

$$D(P_{\mathcal{M}}\|Q_{\mathcal{M}}) \leq D(\rho\|\rho') + (H+1)\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}(D(\mathcal{P}(\cdot|s,a)\|\mathcal{P}'(\cdot|s,a)) + D(r(s,a)\|r'(s,a))).$$

The decomposition established in this lemma highlights a fundamental advantage of our approach over existing PAC-Bayesian bounds Simchowitz et al. [2021], which relies on the total variation distance to measure the discrepancy between latent priors. Because total variation lacks a natural chain rule, those derived bounds inherently couple all sources of environmental variation together. In contrast, our KL-divergence based formulation exploits the underlying structure of the MDPs. This structural awareness allows our bounds to pinpoint exactly which component of the distribution shift is driving the generalization error. Thus, it avoids unnecessary pessimism when only a subset of the MDP components shifts.

The second term in eq. (4) captures the environmental uncertainty, reflecting how strongly the learned meta-parameters depend on the particular set of training environments. This term decreases as the number of training tasks increases, indicating improved robustness with more training tasks.

The third term in eq. (4) quantifies the information each adapted task-level policy ϕ_i retains about its corresponding MDP $\mathcal{M}_i$ given the shared meta-parameter. A small mutual information implies that the policy primarily focuses on maximizing rewards rather than memorizing task-specific details. Unfortunately, as in meta-learning, this term remains nonzero even with infinitely many training tasks, since each task inherently involves adaptation uncertainty conditioned on θ. This residual bias explains why meta-RL cannot be perfectly unbiased, even when there is no distribution shifts. Finally, the discount factor γ amplifies all these effects, as tasks with longer effective horizons are more sensitive to distributional and uncertainty-related deviations.

Note that the divergence terms in Lemma 4 can be potentially unbounded. To avoid such issue, we now consider the subtask problem in meta-RL. We define the subtask meta-RL generalization error as

$$\begin{aligned}\text{gen}_{\text{sub}} := \mathbb{E}_{\hat{\mathcal{M}}\sim\hat{\mathcal{T}}}\mathbb{E}_{\theta,\mathcal{M}_{1:n}}\Bigg[\frac{1}{n_{\hat{\mathcal{M}}}}\sum_{i=1}^{n}\mathbb{E}_{U_i}\Big[\mathbb{1}\{\mathcal{M}_i^{U_i}=\hat{\mathcal{M}}\}\mathbb{E}_{\phi_i^{U_i}|\theta,\mathcal{M}_i^{U_i}}J(\pi_{\phi_i^{U_i}},\mathcal{M}_i^{U_i}) \\ - \mathbb{1}\{\mathcal{M}_i^{-U_i}=\hat{\mathcal{M}}\}\mathbb{E}_{\phi_i^{-U_i}|\theta,\mathcal{M}_i^{-U_i}}J(\pi_{\phi_i^{-U_i}},\mathcal{M}_i^{-U_i})\Big]\Bigg]\end{aligned} \tag{5}$$

where $J(\pi_{\phi_i^{U_i}},\mathcal{M}_i^{U_i}) = \mathbb{E}_{\omega\sim P_{\omega|\mathcal{M}_i^{U_i},\phi_i^{U_i}}}\left[\sum_{j=0}^{H}\gamma^j r_j\right]$.

Theorem 5. *The subtask meta-RL generalization error defined in eq. (5) is upper-bounded by*

$$\text{gen}_{\text{sub}} \leq \mathbb{E}_{\hat{\mathcal{M}}\sim\hat{\mathcal{T}}}\mathbb{E}_{\mathcal{M}_{1:n}}\left[\frac{1}{n_{\hat{\mathcal{M}}}}\sum_{i=1}^{n}\sqrt{\frac{2I(f_i;U_i|\mathcal{M}_{1:n})}{(1-\gamma)^2}}\right]$$

where

$$f_i := \mathbb{1}\{\mathcal{M}_i^- = \hat{\mathcal{M}}\}\mathbb{E}_{\phi_i^-|\theta,\mathcal{M}_i^-} J(\pi_{\phi_i^-}, \mathcal{M}_i^-) - \mathbb{1}\{\mathcal{M}_i^+ = \hat{\mathcal{M}}\}\mathbb{E}_{\phi_i^+|\theta,\mathcal{M}_i^+} J(\pi_{\phi_i^+}, \mathcal{M}_i^+).$$

This bound for the subtask meta-RL setting provides an interpretation analogous to that of Theorem 2. A low conditional mutual information implies that the learned meta-parameter provides a strong prior, making the adapted policy's performance less sensitive to the specific trajectories experienced during adaptation. Besides, simply increasing the total number of training tasks without a proportional increase in tasks relevant to the target environment may fail to improve generalization. Lastly, unlike Theorem 2, this result focuses on task-level samples and does not involve the within-task sample size m.

4.2 Generalization Bounds for Gradient-Based Meta-RL Algorithm

Many existing meta-RL methods, particularly black-box and task-inference approaches, struggle to generalize effectively outside their training distribution. While parameterized policy gradient methods like MAML have theoretical potential to adapt eventually as they retain the structure of a standard RL algorithm, a theoretical understanding of their performance is limited.

In this subsection, we analyze a variant of the meta-gradient RL algorithm, where both the policy gradient and meta-gradient updates are perturbed with isotropic Gaussian noise. Meta-gradient RL [Xu et al., 2018, 2020] optimizes task-specific and meta parameters through a two-level optimization process: an inner loop that updates task-specific parameters using the current meta parameter and collected trajectories, and an outer loop that updates the meta parameter after inner-loop updates. In this variant, both updates are performed using noisy gradients, where isotropic Gaussian perturbations are added to each step. This approach can be viewed as an application of stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011] to the meta-RL setting. The injected noise helps the algorithm escape local minima and converge to global optima under sufficiently regular non-convex objectives. Additionally, the stochastic perturbations naturally encourage exploration in reinforcement learning [Ishfaq et al., 2025].

Suppose $\theta \in \mathbb{R}^d$, and the inner loop performs T updates. Given a specific task $\mathcal{M}_i$, under MAML [Finn et al., 2017], the inner-loop is a policy gradient algorithm whose initial parameter ϕ_i^0 is the meta-parameters θ. At iteration t, the algorithm collects data

$$\omega_{i,t} = \{s_h, a_h, r_h, s_{h+1}\}_{h=0}^{H} \sim \rho_i(s_0) \prod_{h=0}^{H} \pi_{\phi_i^t}(a_h|s_h)\mathcal{P}_i(s_{h+1}|s_h, a_h)$$

where ρ_i and $\mathcal{P}_i$ are associated with $\mathcal{M}_i$. We assume each step uses one trajectory of horizon H. For a trajectory $\omega_{i,t}$, the inner gradient estimator can be computed by

$$\nabla J(\pi_{\phi_i^t}, \mathcal{M}_i) = \sum_{h=0}^{H-1} \gamma^h r_h \sum_{k=0}^{H-1} \nabla \log \pi_{\phi_i^t}(a_k|s_k).$$

We then update the task parameter by

$$\phi_i^{t+1} = \phi_i^t + \beta_t \nabla J(\pi_{\phi_i^t}, \mathcal{M}_i) + \zeta_t$$

where β_t is the learning rate and $\zeta_t \sim \mathcal{N}(\mathbf{0}, \kappa_t^2 \mathbf{1}_d)$ is an isotropic Gaussian noise. Suppose the outer loop performs M updates. At iteration m, we sample a batch $\mathcal{B}_m \subseteq \mathcal{M}_{1:n}$ of size b. The meta parameter is then updated as

$$\theta^{m+1} = \theta^m + \alpha_m \frac{1}{b} \sum_{\mathcal{M}_i \in \mathcal{B}_m} \nabla J_{\mathcal{M}_i}(\theta^m) + \xi_m$$

where α_m is the learning rate and $\xi_m \sim \mathcal{N}(\mathbf{0}, \tilde{\kappa}_m^2 \mathbf{1}_d)$. Explicit formulations of the meta-gradient estimators can be found in Al-Shedivat et al. [2017], Stadie et al. [2018]. In the previous sections, we identified task-level and environmental uncertainties in the upper bound. Beyond improving exploration and optimization, the injected noise also help mitigating these two sources of uncertainty.

Theorem 6. *The OOD meta-RL generalization error for the noisy iterative meta-gradient RL algorithm is upper-bounded by*

$$\mathbb{E}_{\theta,\mathcal{M}_{1:n}}[J_{\mathcal{M}_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)] \leq \sqrt{\frac{2D(P_{\mathcal{M}_{1:n}}\|Q_{\mathcal{M}_{1:n}}) + \mathcal{E}_1 + \mathcal{E}_2}{n(1-\gamma)^2}}$$

where

$$\mathcal{E}_1 = \sum_{m=0}^{M-1} \mathbb{E}_{\theta^m}\left[\log\left(\det\left(\frac{\alpha_m^2}{\tilde{\kappa}_m^2}\tilde{\Sigma}_m + \mathbf{1}_d\right)\right)\right],$$

$$\mathcal{E}_2 = \sum_{m=0}^{M-1}\sum_{t=1}^{T} \mathbb{E}_{\theta^m,\phi_{1:n}^{T(m+1)-t}}\left[\log\left(\det\left(\frac{\beta_{T(m+1)-t}^2}{\kappa_{T(m+1)-t}^2}\Sigma_{T(m+1)-t} + \mathbf{1}_{nd}\right)\right)\right],$$

$$\tilde{\Sigma}_m = \mathrm{cov}\left(\frac{1}{b}\sum_{\mathcal{M}_i\in\mathcal{B}_m}\nabla J_{\mathcal{M}_i}(\theta^m)\right), \quad \textit{and} \quad \Sigma_k = \mathrm{cov}\left(\begin{bmatrix}\nabla J(\pi_{\phi_1^k},\mathcal{M}_1)\\ \vdots \\ \nabla J(\pi_{\phi_n^k},\mathcal{M}_n)\end{bmatrix}\right).$$

We have established an OOD generalization bound for a noisy iterative meta-gradient RL algorithm. The term $\mathcal{E}_1$ reflects environmental uncertainty and depends on the meta-learning rate, the injected noise variance, and the covariance of the meta-gradients. The term $\mathcal{E}_2$ reflects task-level uncertainty and depends on the inner-loop learning rate, the noise variance, and the covariance of the inner-loop gradients. The SGLD algorithm analyzed in Chen et al. [2021], Liu et al. [2025] assumes bounded gradients. However, policy gradients are often unbounded in practice. Thus, we incorporate the gradient variances $\tilde{\Sigma}_m$ and Σ_k into the upper bound [Wen et al., 2025], which captures the sharpness of the optimization landscape [Jiang et al., 2019].

4.3 From Generalization to Suboptimality

We now demonstrate how the meta-RL generalization error can be related to the suboptimality gap in the target task environment. Let $\mathrm{gen}_{\mathrm{ood}}(\theta)$ denote the OOD meta-RL generalization error associated with the meta-parameter θ. Let $\hat{\theta} = \mathcal{A}(\mathcal{M}_{1:n})$ be the meta-parameter output by a meta-learner that maximizes the empirical meta-RL objective. Define $\theta^* \in \arg\max_\theta J_{\mathcal{U}}(\theta)$ as the optimal meta-parameter with respect to the population objective. Then, by Theorem 3, we have

$$\begin{aligned}\mathbb{E}[J_{\mathcal{U}}(\theta^*) - J_{\mathcal{U}}(\hat{\theta})] &\leq \mathbb{E}[J_{\mathcal{U}}(\theta^*) - J_{\mathcal{M}_{1:n}}(\theta^*) + J_{\mathcal{M}_{1:n}}(\theta^*) - J_{\mathcal{M}_{1:n}}(\hat{\theta}) + J_{\mathcal{M}_{1:n}}(\hat{\theta}) - J_{\mathcal{U}}(\hat{\theta})] \\ &\leq \mathrm{gen}_{\mathrm{ood}}(\hat{\theta}) - \mathrm{gen}_{\mathrm{ood}}(\theta^*). \end{aligned} \tag{6}$$

Next, we turn to the discussion of the offline RL setting. Note that the analyzed generalization error thus far is task-level. The same framework can be extended to the offline RL setting, where the generalization error additionally accounts for within-task samples. In particular, for the offline meta-RL setting under episodic MDPs with horizon H, the expected discounted return in eq. (2) can be replaced by an unbiased estimator of the Bellman error. For a given task $\mathcal{M}_i$, let $\phi_i = (\phi_{i,j})_{j=1}^H$ denote the estimated optimal Q-value functions, and let π_{ϕ_i} denote the policy induced by ϕ_i. The Bellman error is then defined as

$$\mathcal{E}(\phi_i) = \frac{1}{H}\sum_{h=1}^{H}\|\phi_{i,h} - \mathcal{T}_h^*\phi_{i,h+1}\|^2$$

where $\mathcal{T}_h^*$ is the Bellman operator that propagates value functions forward by one step under the greedy policy. Let $Z_i = \{S_{i,j}\}_{j=1}^m$ denote the in-task training dataset for task $\mathcal{M}_i$. An unbiased estimator of the mean squared empirical Bellman error can be constructed using the double-sampling trick (i.e., sampling two next states s' and s'' for each transition) [Duan et al., 2021]

$$L(\phi_i, Z_i) := \frac{1}{mH}\sum_{(s,a,r,s',s'',h)\in Z_i}\left[(\phi_{i,h}(s,a) - r - V_{\phi_{i,h+1}}(s'))^2 - \frac{1}{2}(V_{\phi_{i,h+1}}(s') - V_{\phi_{i,h+1}}(s''))^2\right].$$

Under this setting, the meta-learner aims to maximize the empirical meta-RL objective

$$J_{Z_{1:n}}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\phi_i\sim P_{\phi|Z_i,\theta}}[L(\phi_i, Z_i)],$$

while the corresponding population meta-risk is defined as

$$J_{\mathcal{U}}(\theta) := \mathbb{E}_{\mathcal{M}\sim\mathcal{U}}\mathbb{E}_{Z\sim\mathcal{M}}\mathbb{E}_{\phi\sim P_{\phi|Z,\theta}}[\mathcal{E}(\phi)].$$

Consequently, the OOD generalization error is given by $\text{gen}_{\text{ood}} = \mathbb{E}_{\theta,Z_{1:n}}[J_{Z_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)]$. Since $L(\phi_i, Z_i) \in [-2H^2, 4H^2]$, it follows under our proof framework that

$$\mathbb{E}_{\theta,Z_{1:n}}[J_{Z_{1:n}}(\theta) - J_{\mathcal{U}}(\theta)] \leq \sqrt{\frac{64H^2(I(\theta,\phi_{1:n};Z_{1:n}) + D(P_{Z_{1:n}}\|Q_{Z_{1:n}}))}{nm}}.$$

Let $V^*_{\mathcal{M}}(s_1)$ and $V^{\pi}_{\mathcal{M}}(s_1)$ denote, respectively, the optimal value function and the value function induced by policy π at the initial state s_1 and step 1 in MDP $\mathcal{M}$. By relating the Bellman error to value suboptimality [Duan et al., 2021], we obtain

$$\mathbb{E}_{\mathcal{M}\sim\mathcal{U}}\mathbb{E}_{Z\sim\mathcal{M}}\mathbb{E}_{\phi\sim P_{\phi|Z,\theta}}[V^*_{\mathcal{M}}(s_1) - V^{\pi_\phi}_{\mathcal{M}}(s_1)] \leq 2H\sqrt{C\cdot\text{gen}_{\text{ood}}} + 2H\sqrt{C\cdot\mathbb{E}_{\theta,Z_{1:n}}[J_{Z_{1:n}}(\theta)]} \tag{7}$$

where C is the concentrability coefficient, which quantifies the adequacy of dataset coverage for off-policy evaluation. Thus, from eq. (6) and eq. (7), we establish the connection between the meta-RL generalization error and the suboptimality gap in both standard and offline settings.

5 Conclusion

In this paper, we provided an information-theoretic analysis for OOD generalization in both meta-supervised learning and meta-RL. We investigated two distinct scenarios: the classical distribution mismatch setting and a "broad-to-narrow" subtask setting. Our generalization bound for meta-RL explicitly decouples the distribution shift, measuring the divergence across the initial state distribution, transition kernel, and reward function. Besides, we derived a generalization bound for a gradient-based meta-RL algorithm and discussed the connection between the generalization error and the suboptimality gap in both standard and offline meta-RL.

References

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. *arXiv preprint arXiv:1710.03641*, 2017.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, Shimon Whiteson, et al. A tutorial on meta-reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(2-3):224–384, 2025.

Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34:25878–25890, 2021.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta-reinforcement learning of structured exploration strategies. *Advances in neural information processing systems*, 31, 2018.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. *arXiv preprint arXiv:2501.17827*, 2025.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Firas Laakom, Yuheng Bu, and Moncef Gabbouj. Class-wise generalization error: an information-theoretic analysis. *arXiv preprint arXiv:2401.02904*, 2024.

Wenliang Liu, Guanding Yu, Lele Wang, and Renjie Liao. An information-theoretic framework for out-of-distribution generalization with applications to stochastic gradient langevin dynamics. *IEEE Transactions on Information Theory*, 2025.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 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. *arXiv preprint arXiv:1803.11347*, 2018.

Zohar Rimon, Aviv Tamar, and Gilad Adler. Meta reinforcement learning with finite training tasks-a density estimation approach. *Advances in Neural Information Processing Systems*, 35: 13640–13653, 2022.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel J Hsu, Thodoris Lykouris, Miro Dudik, and Robert E Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in neural information processing systems*, 34:26382–26394, 2021.

Bradly C Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Aviv Tamar, Daniel Soudry, and Ev Zisselman. Regularization guarantees generalization in bayesian reinforcement learning through algorithmic stability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8423–8431, 2022.

Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998.

Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. *arXiv preprint arXiv:2210.00706*, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Wen Wen, Tieliang Gong, Yuxin Dong, Yong-Jin Liu, and Weizhan Zhang. Towards sharper information-theoretic generalization bounds for meta-learning. *arXiv preprint arXiv:2501.15559*, 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Zhongwen Xu, Hado P van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, and David Silver. Meta-gradient reinforcement learning with an objective discovered online. *Advances in Neural Information Processing Systems*, 33:15254–15264, 2020.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. *IEEE Transactions on Information Theory*, 68(5):3304–3316, 2022.