Learning Not to Learn:
Nature versus Nurture in Silico

Anonymous Author(s)
Affiliation
Address
email

Abstract

Animals are equipped with a rich innate repertoire of sensory, behavioral and motor skills, which allows them to interact with the world immediately after birth. At the same time, many behaviors are highly adaptive and can be tailored to specific environments by means of learning. In this work, we use mathematical analysis and the framework of meta-learning (or 'learning to learn') to answer when it is beneficial to learn such an adaptive strategy and when to hard-code a heuristic behavior. We find that the interplay of ecological uncertainty, task complexity and the agents' lifetime has crucial effects on the meta-learned amortized Bayesian inference performed by an agent. There exist two regimes: One in which meta-learning yields a learning algorithm that implements task-dependent information-integration and a second regime in which meta-learning imprints a heuristic or 'hard-coded' behavior. Further analysis reveals that non-adaptive behaviors are not only optimal for aspects of the environment that are stable across individuals, but also in situations where an adaptation to the environment would in fact be highly beneficial, but could not be done quickly enough to be exploited within the remaining lifetime. Hard-coded behaviors should hence not only be those that always work, but also those that are too complex to be learned within a reasonable time frame.

1 Introduction

The *'nature versus nurture'* debate (e.g., Mutti et al., 1996; Tabery, 2014) – the question which aspects of behavior are 'hard-coded' by evolution, and which are learned from experience – is one of the oldest and most controversial debates in biology. Evolutionary principles prescribe that hard-coded behavioral routines should be those for which there is no benefit in adaptation. This is believed to be the case for behaviors whose evolutionary advantage varies little among individuals of a species. Mating instincts or flight reflexes are general solutions that rarely present an evolutionary disadvantage. On the other hand, features of the environment that vary substantially for individuals of a species potentially ask for adaptive behavior (Buss, 2015). Naturally, the same principles should not only apply to biological but also to artificial agents. But how can a reinforcement learning agent differentiate between these two behavioral regimes? In this work, we use the meta-learning approach to acquire a qualitative understanding of which aspects of behavior should be hard-coded and which should be adaptive. Our hypothesis is that meta-learning can not only learn efficient learning algorithms, but can also decide not to be adaptive at all, and to instead apply a generic heuristic to the whole ensemble of tasks. Phrased in the language of biology, meta-learning can decide whether to hard-code a behavior or to render it adaptive, based on the range of environments the individuals of a species could encounter; i.e. it can provide a rigorous answer to the *'nature vs nurture'* debate. We show that the meta-learned algorithm depends on the interplay of three fundamental pillars of the biological reinforcement learning problem: *Ecological uncertainty*, *task* *complexity*, and *expected lifetime*. More specifically, our analytical and numerical analysis reveal that non-adaptive behaviors are optimal in two cases – when aspects of the environment are stable across the episodes of the agent and in situations where an investment in adaptation can not amortize itself fast enough given the limited lifetime of the agent. Our results are summarized as follows: The lifetime of an agent strongly affects the meta-learned adaptation strategy. The feasibility of meta-learning a learning algorithm depends not only on the task distribution but also the relative amount of entropy reduction that can occur within the agents' lifetime. The dynamics of RNN-based policies are capable of encoding heuristic and non-adaptive solutions deemed to be optimal during the meta-learning procedure. Hence, the meta-optimization may decide to imprint behaviors or to let the policy be shaped online. Meta-learned adaptation is not robust to a change in the adaptation timescale. Agents trained with a fixed lifetime may not generalize to settings with fewer or more time available. The results suggest that the design of the meta-task distribution has drastic effects on the meta-learned algorithm. This is essential for research questions that are interested in the conducted adaptation behavior. These include curriculum design, safe exploration as well as human-in-the-loop applications.

2 Related Work & Background

Previous work has shown that LSTM-based meta-learning is capable of distilling a sequential integration algorithm akin to amortized Bayesian inference (Ortega et al., 2019; Rabinowitz, 2019). Here we investigate when the integration of information might not be the optimal strategy to meta-learn. We analytically characterize a task regime in which not adapting to sensory information is optimal. Rabinowitz (2019) previously studied the outer loop learning dynamics and found differences across several tasks, the origin of which is however not fully understood. This work may provide an explanation for these different meta-learning dynamics and the dependence on the task distribution as well as the time horizon of adaptation. Our work is most closely related to Pardo et al. (2017) and Zintgraf et al. (2019). Pardo et al. (2017) study the impact of fixed time limits and time-awareness on deep reinforcement learning agents. They propose using a timestamp as part of the state representation in order to avoid state-aliasing and the non-Markovianity resulting from a finite horizon treatment of an infinite horizon problem. Our setting differs in several aspects. First, we study the case of meta-reinforcement learning where the agent has to learn within a single lifetime. Second, we focus on a finite horizon perspective with limited adaptation. Zintgraf et al. (2019), on the other hand, do investigate meta reinforcement-learning for Bayes-adaptive Markov Decision Processes and introduce a novel architecture that disentangles task-specific belief representations from policy representations. Similarly to our work, Zintgraf et al. (2019) are interested in using the meta-learning framework to distill Bayes optimal exploration behavior. While their adaptation setup extends over multiple episodes, we focus on single lifetime adaption and analytically analyze when it is beneficial to learn in the first place.

3 Learning not to Learn

To disentangle the influence of ecological uncertainty, task complexity, and lifetime on the nature of the meta-learned strategy, we first focus on a simple two-arm Gaussian bandit task, which allows for an analytical solution. At the beginning of an episode the mean reward of the first bandit arm is sampled from a Gaussian distribution with mean -1 and standard deviation σ_p, i.e. $\mu \sim \mathcal{N}(-1, \sigma_p^2)$. It remains constant for the lifetime T of the agent. The second arm is deterministic and a pull yields a reward of 0. Within an episode and if the agent chooses to pull from the first arm, its resulting pull reward is sampled from a second Gaussian, $r \sim \mathcal{N}(\mu, \sigma_l)$. This task formulation then allows us to disentangle two sources of uncertainty:

- **Epistemic uncertainty** σ_p: The amount of ecological uncertainty that can be reduced over the lifetime of the agent by a meta-learned algorithm that integrates information. This maps onto the entropy of the respective ecological niche.
- **Aleatoric uncertainty** σ_l: The irreducible sensory uncertainty that sets the time scale on which information needs to be integrated in order to learn an optimal strategy. We use this as a proxy for task complexity.

In this simple setting, the optimal meta-learned strategy can be calculated analytically. The optimal exploration strategy is to initially explore the stochastic arm for a given trial number n. Afterwards, it chooses the best arm based on its maximum a posteriori-estimate of the remaining episode return. The optimal amount of exploration trials $n^\star$ can then be derived analytically (see appendix):

$$n^\star = \arg\max_n \mathbb{E}[\sum_{t=1}^{T} r_t | n, T, \sigma_l, \sigma_p] = \arg\max_n \left[-n + \mathbb{E}_{\mu,r}\left[(T-n) \times \mu \times p(\hat{\mu} > 0)\right]\right] ,$$

where $\hat{\mu}$ is the estimate of the mean reward of the first arm after the n exploration trials. We find two distinct types of behavior (left-hand side of figure 1): A regime in which learning via exploration is effective and a second regime in which not learning is the optimal behavior. It may be optimal not to learn for two reasons: First, the ecological uncertainty may be so small that it is very unlikely that the stochastic first arm is better. Second, if the sensory uncertainty is too large relative to the range of potential ecological niches it may simply not be possible to integrate sufficient information given a limited lifespan. We make two observations:

1. There exists a hard threshold between learning and not learning behaviors described by the ratio of σ_l and σ_p. If σ_l is too large, the value of exploration (or the reduction in uncertainty) is too small to be profitable within the remaining lifetime of the agent. Instead, it is advantageous to hard-code a heuristic choice.
2. The two regimes consistently exist across different lifetimes. As the lifetime grows, the learning regime becomes more and more prevalent. Given a sufficient amount of time, learning by exploring the uncertain arm is the best strategy.

Is the common meta-learning framework capable of reproducing these different qualitative behaviors and performing Bayes optimal amortized inference across the entire spectrum of meta-task distributions? Or differently put: Can memory-based meta-learning yield agents that do not only learn to learn but that also learn *not* to learn? To answer this question, we train LSTM-based RL^2 (Wang et al., 2016) agents with the standard synchronous actor-critic (Mnih et al., 2016) setup on the same grid of aleatoric and epistemic uncertainties. Importantly, we optimize the initial condition of the hidden state, which is reset after each episode. We obtain the amount of meta-learned exploration by testing the RL^2 agents on hold-out bandits for which we set $\sigma_p = 0$ and only vary σ_l. Thereby, it is ensured that the deterministic arm is the better arm. We can then define the number of exploration trials as the pulls from the suboptimal stochastic arm. We observe that meta-learning is capable of yielding agents that behave according to our derived theory of a Bayes optimal agent, which explicitly knows the given lifetime as well as uncertainties σ_l, σ_p (figure 1). Importantly, the meta-learned behavior also falls into two regimes: A regime in which the meta-learned strategy resembles a learning algorithm and a regime in which the recurrent dynamics encode a hard-coded choice of the deterministic arm. Furthermore, the trailing edge between the two meta-learned regimes shifts with the agent's lifetime as predicted by the Bayesian theory. As the lifetime increases, wider ecological niches at higher levels of task complexity become solvable and the strategy of learning profitable.

The two behavioral regimes are characterized by distinct recurrent dynamics of the trained LSTM agents. The two left-most columns of figure 2 display the policy entropy and hidden state statistics for a network trained on a σ_l, σ_p-combination associated with the regime in which learning is the optimal behavior. We differentiate between the case in which the deterministic arm is the better one ($\mu < 0$) and the case in which the second arm should be preferred ($\mu > 0$). In both cases the agent first explores in order to identify the better arm. We examined how these strategies evolve over the course of meta-training and find that there are two phases: After an initial period of universal random behavior across all conditions, the distinct behavioral regimes emerge (figure 1). We note that this observation may be partially caused by the linear annealing of the entropy regularization coefficient in the actor-critic objective which we found to be crucial in training the networks. Moreover, the hidden dynamics appear to display two different attractors, which correspond to either of the arms being the better choice. The better arm can clearly be identified from the PCA-dimensionality reduced hidden state dynamics (bottom row of figure 2). The two right-most columns of figure 2, on the other hand, depict the same statistics for a network that was meta-trained on the regime in which the optimal strategy is not to learn. Indeed, the agent always chooses the deterministic arm, regardless of whether it is the better choice. Accordingly, the network dynamics seem to fall into a single attractor. In summary, we observe that the meta-learning process is capable of providing a quantitative model for the *'nature versus nurture'* trade-off depending on the ecological uncertainty of the environment, the task complexity defined by the sensory precision and the lifetime of the agent. Next, we investigate whether this insight generalizes to more complex domains by studying spatial reasoning.

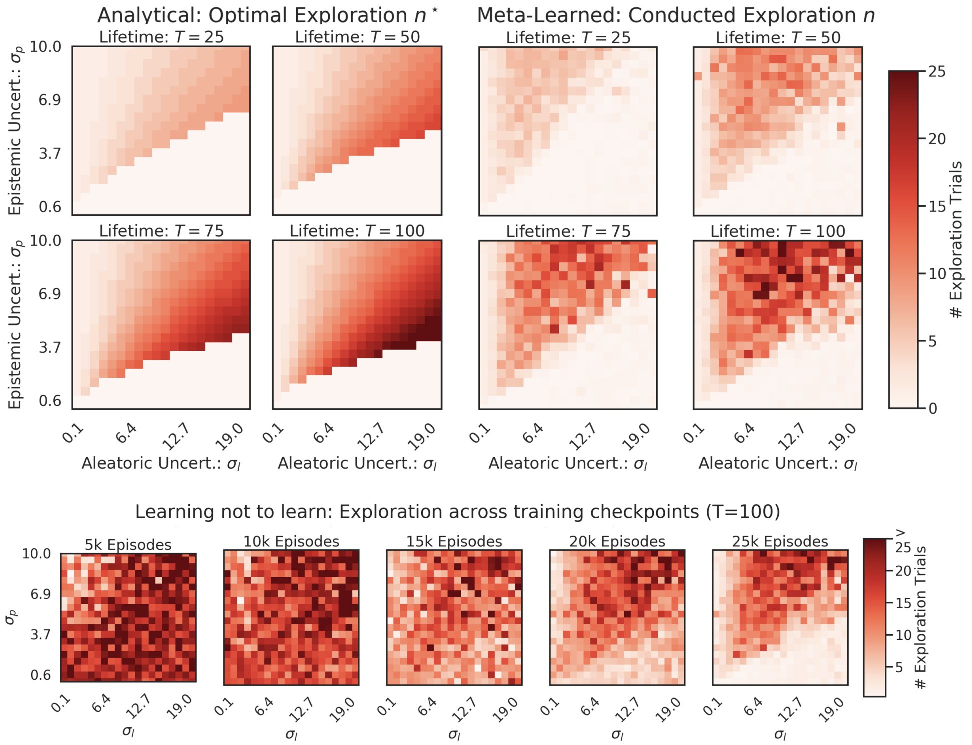

Figure 1: Theory and meta-learned exploration in a two-arm Gaussian bandit. **Top - Left**: Bayes optimal exploration behavior for different lifetimes and across uncertainty conditions σ_l, σ_p. **Top - Right**: Meta-learned exploration behavior using the RL^2 (Wang et al., 2016) framework. **Bottom row**. The meta-learned exploration strategy is visualized for different increasing checkpoints throughout meta-learning ($T = 100$). The amount of meta-learned exploration is averaged both over 5 training runs and 100 episodes.

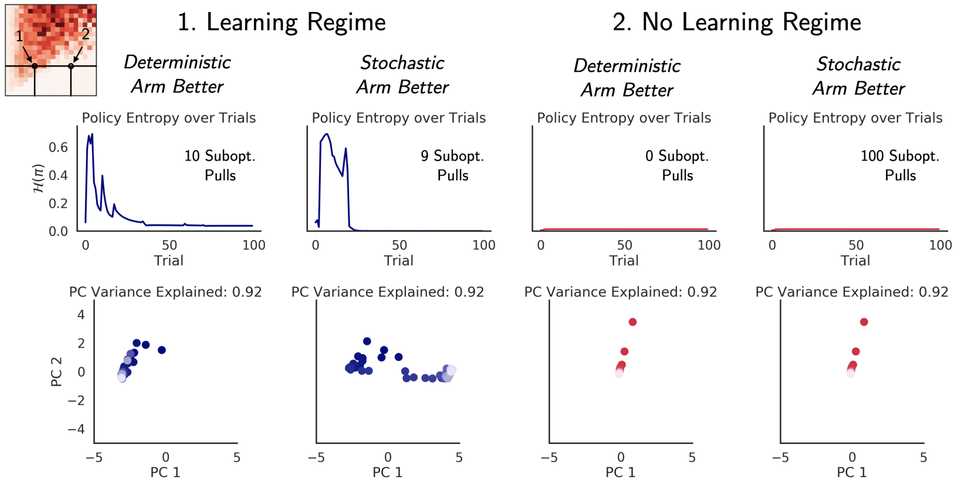

Figure 2: Meta-learned recurrent dynamics of learning (blue) and not learning (red) for a lifetime $T = 100$. **First two columns**. Bandit for which an adaptive strategy is predicted by the theory. **Last two columns**. Bandit for which the heuristic choice of the deterministic arm is the Bayes optimal behavior. A lighter color indicates later episode trials.

4 Time Horizons, Meta-Learned Strategies & Entropy Reduction

While the simple bandit task provides an analytical perspective on the trade-off of learning versus hard-coded behavior, it is not obvious that the obtained insights generalize to more complex situations, i.e., to distributions of finite-horizon MDPs. To investigate this, we studied exploration behavior in an ensemble of grid worlds task. We hypothesize that meta-learning yields qualitatively different spatial exploration strategies depending on the lifetime of the agent. For short a lifetime, the agent should opt for small rewards that are easy to find. For longer lifetimes, the agent can spend time to explore the environment and identify higher rewards that are harder to find.

To test this hypothesis, we train a RL^2-based meta learner to explore a maze with three different types of goal locations (top row of figure 3): g_h (green object), g_m (yellow object) and g_s (pink object) with transition rewards $R(g_h) > R(g_m) > R(g_s)$. During an episode/lifetime the goal locations are fixed. At the beginning of the episode g_m and g_h are randomly sampled. The location of g_s, on the other hand, remains fixed across all training episodes. We sample the possible locations for g_h from the outermost column and row (11 locations) while g_m varies along the third row and column (excluding the borders, 5 locations). Thereby, the three goals encode destinations with varying degrees of spatial uncertainty and payoff. The agent can move up, down, left and right. After it (red circle) transitions into a goal location, it receives the associated reward and is teleported back to the initial location in the bottom left corner. Importantly, the agent does not observe the goal locations but instead has to infer the locations based on the observed transition rewards. Depending on the lifetime of the agent during the inner loop adaptation, we find that meta-learning can imprint 3 qualitatively different strategies (figure 3 middle and right column): For small lifetimes the agent executes a hard-coded policy that repeatedly walks to the safe, low-reward pink object. As the lifetime and consequently inner loop adaptation is increased, we find that the agent's meta-learned policy starts to explore a broader range of locations int the maze, first exploring possible locations of the medium-reward object and – for long lifetimes – the distant and uncertain high-reward object (figure 3 middle column). Consistently, the agent exploits increasingly uncertain rewards with increasing lifetime (figure 3 right column).

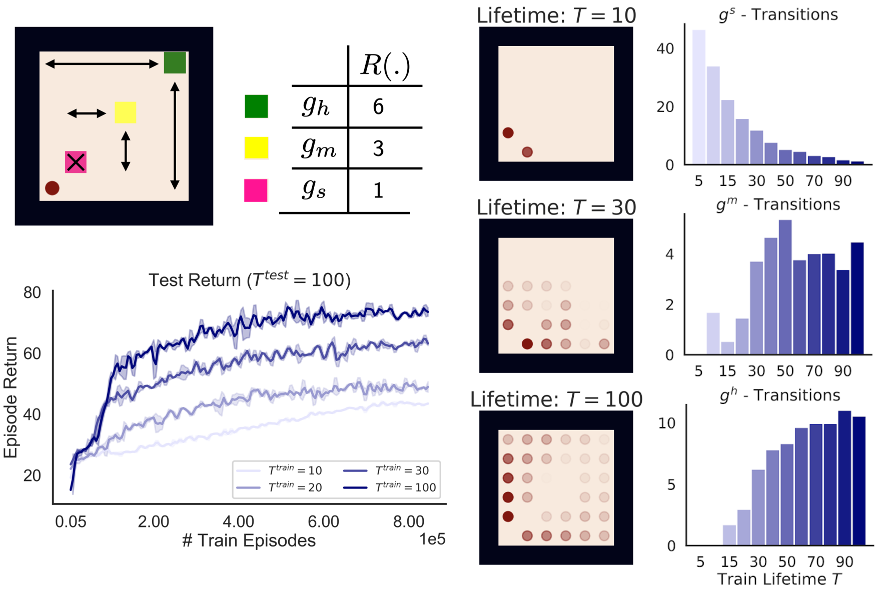

Figure 3: Grid navigation task with 3 different rewards, which differ in the amount of reward and the uncertainty in location. **Top-Left**: Task formulation. **Bottom-Left**: Learning curves ($T^{test} = 100$) for different training lifetime, averaged over 10 independent training runs and 10 evaluation episodes. **Middle**: Relative state occupancy of meta-learned exploration strategies for different training lifetimes during meta-learning (averaged over 100 episodes of length 100). **Right**: Lifetime dependence of the performance and the visitation counts of the goal locations for an RL^2 agent trained on random 6×6 grid worlds and evaluated on $T^{test} = 100$.

Furthermore, we investigated how meta-learned strategies generalize across different timescales of adaptation. More specifically, we trained an agent to learn (or not) with a given lifetime and tested how the learned behavior performed in a setting where there is more or less time available. As predicted, we find that the test time normalized return of the agents decreased with the discrepancy between training and test lifetime (figure 4). This can be problematic in settings where the agents does not have access to its exact lifetime and highlights the lack of time-robustness of meta-adaptation. The agents displayed clear hallmarks of model-based behavior and behavioral changes over their lifetime (figure 5). When the agent has encountered the high reward once, it resorts to a deterministic exploitation strategy that follows a shorter trajectory through the environment than the one initially used during exploration. Furthermore, the adaptive policies identify when there is not enough time left in the episode to reach the previously exploited goal location. In that case the policies switch towards the easier to reach small goal location. The oscillating policy entropy (columns three and four of figure 5) is state-specific and indicates that the meta-learned strategies have correctly learned a transition model of the relevant parts of the environment. If an action does not affect the overall length of the trajectory to a goal, this is reflected in the entropy of the policy. Finally, we analyzed the distinct recurrent dynamics for the three different strategies (final column of figure 5). We find that the dimensionality of the dynamics increases with the adaptivity of the behavior. As the training lifetime increases, the participation ratio (Gao et al., 2017) of the hidden state dynamics increases and the explained variance of the first three principal components drops.

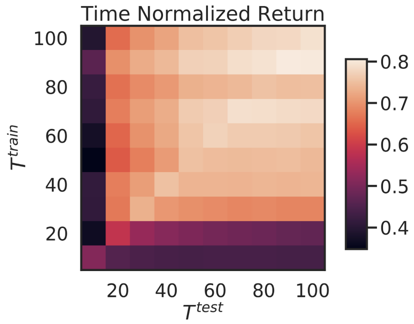

Figure 4: Episode return for agents trained on T^{train} and tested on T^{test}. The returns are normalized by the test lifetime. The statistics are averaged over 5 training runs and 500 test episodes.

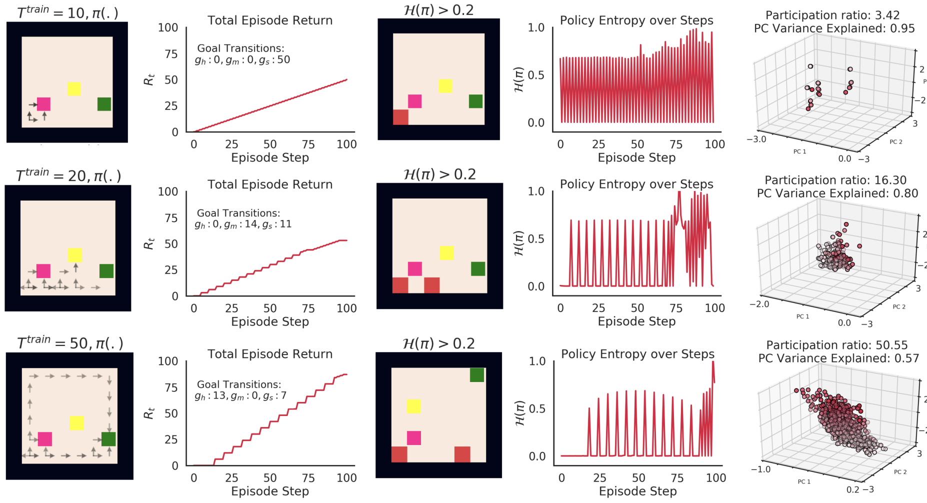

Figure 5: Characteristic trajectories for three different types of meta-learned strategies. **Top to bottom**: Episode rollouts ($T^{test} = 100$) for inner loop training lifetimes $T^{train} = \{10, 20, 50\}$. **First four columns**: The agents' trajectories, episode return, states with high average policy entropy (red squares) and the policy entropy for the same sampled environment. **Final column**. PCA-dimensionality reduced hidden state dynamics for different 100 rollout episodes and the agents' respective training lifetimes. A lighter color indicates later episode trials.

Conclusion. This work has investigated the interplay of three considerations when designing meta-task distributions: The diversity of the task distribution, task complexity and training lifetime. Depending on these, traditional meta-learning algorithms are capable of flexibly interpolating between distilling a learning algorithm and hard-coding a heuristic behavior. The different regimes emerge in the outer loop of meta-learning and are characterized by distinct recurrent dynamics shaping the hidden activity. Meta-learned strategies were not capable of generalizing to timescales they were not trained on, emphasizing the importance of the training lifetime in meta-learning.

References

BUSS, D. (2015): *Evolutionary psychology: The new science of the mind*, Psychology Press.

GAO, P., E. TRAUTMANN, B. YU, G. SANTHANAM, S. RYU, K. SHENOY, AND S. GANGULI (2017): "A theory of multineuronal dimensionality, dynamics and measurement," *BioRxiv*, 214262.

MNIH, V., A. P. BADIA, M. MIRZA, A. GRAVES, T. LILLICRAP, T. HARLEY, D. SILVER, AND K. KAVUKCUOGLU (2016): "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 1928–1937.

MUTTI, D. O., K. ZADNIK, AND A. J. ADAMS (1996): "Myopia. The nature versus nurture debate goes on." *Investigative ophthalmology & visual science*, 37, 952–957.

ORTEGA, P. A., J. X. WANG, M. ROWLAND, T. GENEWEIN, Z. KURTH-NELSON, R. PASCANU, N. HEESS, J. VENESS, A. PRITZEL, P. SPRECHMANN, ET AL. (2019): "Meta-learning of sequential strategies," *arXiv preprint arXiv:1905.03030*.

PARDO, F., A. TAVAKOLI, V. LEVDIK, AND P. KORMUSHEV (2017): "Time limits in reinforcement learning," *arXiv preprint arXiv:1712.00378*.

RABINOWITZ, N. C. (2019): "Meta-learners' learning dynamics are unlike learners'," *arXiv preprint arXiv:1905.01320*.

TABERY, J. (2014): *Beyond versus: The struggle to understand the interaction of nature and nurture*, MIT Press.

WANG, J. X., Z. KURTH-NELSON, D. TIRUMALA, H. SOYER, J. Z. LEIBO, R. MUNOS, C. BLUNDELL, D. KUMARAN, AND M. BOTVINICK (2016): "Learning to reinforcement learn," *arXiv preprint arXiv:1611.05763*.

ZINTGRAF, L., K. SHIARLIS, M. IGL, S. SCHULZE, Y. GAL, K. HOFMANN, AND S. WHITESON (2019): "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning," *arXiv preprint arXiv:1910.08348*.

A Supplementary Materials

A.1 Mathematical Derivation of Optimal Exploration

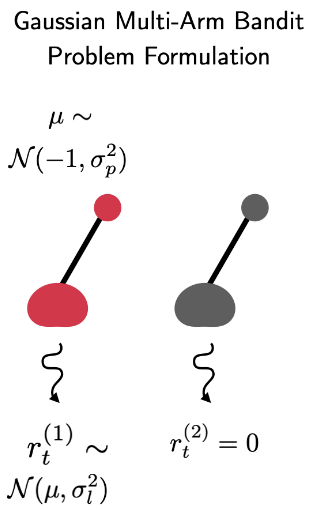
Gaussian Multi-Arm Bandit
Problem Formulation
$\mu \sim \mathcal{N}(-1, \sigma_p^2)$
$r_t^{(1)} \sim \mathcal{N}(\mu, \sigma_l^2)$
$r_t^{(2)} = 0$

Figure 6: Two-arm Gaussian bandit.

In the following section we describe the Gaussian two-arm bandit setting analyzed in section 3. The first arm generates stochastic rewards r according to hierarchical Gaussian emissions. Between individual episodes the mean reward μ is sampled from a Gaussian distribution with standard deviation σ_p. The second arm, on the other hand, arm generates a deterministic reward of 0. More specifically, the generative process for rewards resulting from a pull of the first arm is described as follows:

$$\mu \sim \mathcal{N}(-1, \sigma_p^2);\ r \sim \mathcal{N}(\mu, \sigma_l^2)$$
$$p(\mu|\sigma_p^2) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left\{-\frac{(\mu+1)^2}{2\sigma_p^2}\right\}$$
$$p(r|\mu, \sigma_l^2) = \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left\{-\frac{(r-\mu)^2}{2\sigma_l^2}\right\} .$$

Our Bayesian agent is assumed to spend a fixed amount of trials n of the overall lifetime T exploring the second stochastic arm. This assumption is justified since our corresponding meta-learning agent may easily encode the deterministic nature of the fixed arm 0 and therefore only explore the second non-deterministic arm.[1] The expected cumulative reward of such a two-phase exploration-exploitation policy can then be factorized as follows:

$$\begin{aligned}\mathbb{E}_{\mu,r}\left[\sum_{t=1}^{T} r_t \middle| n\right] &= \mathbb{E}_{\mu,r}\left[\sum_{t=1}^{n} r_t \middle| n\right] + \mathbb{E}_{\mu,r}\left[\sum_{t=n+1}^{T} r_t \middle| n, \hat{\mu}\right] \\ &= (-1) \times n + \mathbb{E}_{\mu}\left[(T-n)p(\hat{\mu} > 0)\mu\right]\end{aligned}$$

where $\hat{\mu}$ denotes the maximum a posteriori (MAP) estimate of μ after n trials:

$$\hat{\mu} = \frac{1}{P_{tot}}\left[-1 \times P_p + n \times P_l \times \bar{r}\right] .$$

with $P_p = \frac{1}{\sigma_p^2}$, $P_l = \frac{1}{\sigma_l^2}$, $P_{tot} = P_p + nP_l$ and $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$. For a fixed μ (within a lifetime) this random variable follows a univariate Gaussian distribution with sufficient statistics given by:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{P_{tot}} = (P_p + nP_l\bar{r})\,; Var[\hat{\mu}] = \frac{1}{nP_l} .$$

The probability of exploiting arm 2 after n exploration trials is then given by the following integral:

$$p(\hat{\mu} > 0) = \int_0^{\infty} p(\hat{\mu})d\bar{\mu} .$$

$\mathbb{E}_{\mu,r}[\sum_{t=1}^{T} r_t|n]$ may then be evaluated by numerical integration and the resulting optimal $n^\star$ is obtained by searching over a range of $n = 1, \dots, T$.

[1] The second column of figure 2 validates this assumption since the policy entropy quickly vanishes after an initial exploration phase.

A.2 Experimental Details

A.2.1 Memory-Based Meta-Reinforcement Learning

We follow the standard RL^2 paradigm (Wang et al., 2016) and train an LSTM-based actor-critic architecture using the A2C objective (Mnih et al., 2016):

$$\begin{aligned}
\mathcal{L}^{AC} &= \mathcal{L}^{\pi} + \beta_v \mathcal{L}^{v} - \beta_e \mathcal{L}^{e} \\
\mathcal{L}^{\pi} &= \mathbb{E}_{\pi}\left[\log \pi(a_t|x_t)[R_t - V(x_t)]\right] \\
\mathcal{L}^{v} &= \mathbb{E}_{\pi}\left[(R_t - V(x_t))^2\right] \\
\mathcal{L}^{e} &= \mathbb{E}_{\pi}\left[\mathcal{H}(\pi(a_t|x_t))\right] \\
R_t &= \sum_{i=0}^{T-t-1} \gamma^i r_{t+i},
\end{aligned}$$

where R_t denotes the cumulative discounted reward resulting from the rollout of the episode. The agent interacts with a sampled environment for a single episode. Between episodes a new environment is sampled. Unless otherwise stated we ensure a proper scaling of the timestamp input and follow Pardo et al. (2017) by normalizing the time input to lie between -1 and 1.

A.2.2 Gaussian Multi-Arm Bandits: Hyperparameters

All results of section 3 for the two-arm Gaussian bandit setting (and all σ_l, σ_p-combinations) may be reproduced using the following set of hyperparameters:

Parameter	Value	Parameter	Value	Parameter	Value
Training episodes	30k	Learning rate	0.001	L_2 Weight decay λ	$3e-06$
Clipped gradient norm	10	Optimizer	Adam	Workers	2
γ_T	0.999	$\beta_{e,T}$	0.005	β_v	0.05
γ_0	0.4	$\beta_{e,0}$	1	LSTM hidden units	48
γ Anneal time	27k Ep.	β_e Anneal time	30k Ep.	Learned hidden init.	✓
γ Schedule	Exponential	β_e Schedule	Linear	Forget gate bias init.	1
-	-	-	-	Orthogonal weight init.	✓

Table 1: Hyperparameters (architecture & training procedure) of the bandit A2C agent.

We use the same set of hyperparameters for all σ_l, σ_p and T combinations. Our theoretical results are derived for the case of $\gamma = 1$. This poses a challenge when training recurrent policies. The difficulty of the temporal credit assignment problem is implicitly defined by the effective length of the time window. We observed that starting with a large γ often times hindered the network in learning. Hence, we discount factor as a form of implicit curriculum and annealed it accordingly.

A.2.3 Gridworld Navigation Task: Hyperparameters

Parameter	Value	Parameter	Value	Parameter	Value
Training episodes	1M	Learning rate	0.001	L_2 Weight decay λ	0
Clipped gradient norm	10	Optimizer	Adam	Workers	7
γ	0.99	$\beta_{e,T}$	0.5	β_v	0.1
β_e Schedule	Linear	$\beta_{e,0}$	0.01	LSTM hidden units	256
-	-	β_e Anneal time	700k	Learned hidden init.	✓

Table 2: Hyperparameters (architecture & training procedure) of the gridworld A2C agent.

In order to train the agents that generated the state occupancies in figure 3 and the rollout trajectories in figure 5 we additionally annealed the discount factor starting at 0.8 to 1 within the first 800k episodes.