A Reinforcement Learning Approach for Health-Behavioural Recommendations to Reduce Cancer Risk

Gloria Desideri
Politecnico di Milano
gloria.desideri@mail.polimi.it

Andrés Pasinetti
Politecnico di Milano
andres.pasinetti@polimi.it

Abstract

In the past years, the adoption of specific lifestyle choices, e.g., healthy food, reduced smoke and alcohol consumption, has been revealed to be a crucial tool to reduce cancer mortality in several studies. However, digital health interventions to make people adopt such behaviours require personalization to ensure long-term engagement and effectiveness for specific users. Indeed, their design is challenged by the variability in users' capabilities, learning patterns, and fatigue dynamics. In the literature, the development of such systems has been held back by the scarcity of available longitudinal, individual-level health behaviour datasets, which do not allow the use of classical reinforcement learning (RL) techniques for learning an effective personalized intervention strategy. In this work, we tackle the intervention recommendation problem using a meta-RL approach to provide personalized intervention suggestions to users and Model-Agnostic Meta-Learning (MAML) with Actor Critic (AC) policies to enable rapid policy adaptation to new users from minimal interaction data. We conduct empirical studies on cross-task adaptation showing that our approach adapts with limited data per user and outperforms both the chosen baselines.

1 Introduction

In the past decade, there has been an effort at a national level over the EU [19] to improve cancer prevention and treatment due to the large incidence of such a disease over the entire population. Modifiable behavioural factors, such as tobacco use, alcohol consumption, and physical inactivity, are estimated to account for 42% of all cancers diagnosed and 45% of all cancer deaths in recent years (Islami et al., 2018) [10]. Promoting sustainable changes in these behaviours is therefore a public health priority.

One of the most promising approaches to deliver behavioural suggestions is the use of smartphone apps specifically designed to give behavioural recommendations [1]. These apps would provide tailored guidance and nudges toward healthier lifestyles. Indeed, smartphone notifications constitute a frequent delivery method for these interventions, prompting users to take small yet impactful actions in their daily lives. More specifically, behavioural recommendations interventions can encourage users to engage in regular exercise, maintain balanced diets, or practice mindfulness techniques, all of which contribute to overall well-being and disease prevention.

Traditional behavioural interventions often rely on generic recommendations, predetermined by clinical teams based on general guidelines rather than individual user needs [12]. While such standardized approaches can provide a foundation for behaviour change, they may fail to account for individual differences in motivation, lifestyle constraints, or progress over time. Users may receive recommendations that are either too frequent, making them feel pressured or overwhelmed,

Submitted to the Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025). Do not distribute.

or too challenging, leading to frustration and disengagement. Without personalization, even well-intentioned health interventions risk becoming counterproductive, as users may quickly lose interest in applications that do not align with their capabilities or preferences.

Advances in reinforcement learning (RL) offer promising solutions to this challenge by enabling dynamic, personalized behavioural recommendations. Rather than relying on static rules, RL-driven interventions adapt online to user feedback, learning from responses and adjusting recommendations accordingly. This personalization helps to keep interventions engaging and achievable, improving adherence and long-term impact. A recent review by Weimann and Gißke (2024) [33] highlights growing interest in RL for behavioural recommendations. However, few studies explicitly target achieving a goal by selecting activities of appropriate difficulty. As argued by Guadagnoli and Lee (2004) [9], task difficulty critically shapes practice effectiveness. They distinguish *nominal difficulty*, an inherent property of the task, from *functional difficulty*, which reflects how demanding the task is for a given individual in a particular context. Learning is most effective near the *optimal challenge point*: tasks that are too easy yield minimal progress, whereas overly difficult tasks erode engagement. In digital coaching, this makes task difficulty a natural control variable: by adjusting difficulty in response to the user's state, the system can maintain functional difficulty near the optimal challenge point, supporting sustained learning and adherence.

The *cold-start problem* [13] is another challenge for RL in this context. Many RL techniques require large amounts of expensive interaction data to develop a different strategy for every user. Reliable policies cannot be trained from beginning since new users usually only give a small number of interactions. This motivates the application of *meta-learning*, which aims to train models that can quickly adjust to new users using less data.

This study focuses on developing a Reinforcement Learning (RL) system for personalized behavioural recommendations, aiming to enhance long-term user engagement and adherence. The first objective is to create a system that dynamically adjusts the difficulty of the suggested tasks based on the user's evolving state, ensuring that interventions remain relevant and effective while minimizing the risk of disengagement or early dropout. A second objective is to enable generalization across users with diverse and unobservable characteristics without requiring extensive individual experience.

To summarize, the main contributions of this work are:

- **MDP user model.** We model behavioral change as a task-specific Markov Decision Process (MDP) capturing how fatigue and motivation shape engagement and learning.
- **Meta-RL with MAML.** We cast the problem of providing personalized recommendations as a meta-RL over a distribution of user MDPs and use a Model Agnostic Meta Learning (MAML) based actor–critic to adapt the policy to a target user group with a few gradient steps.
- **Empirical study.** We demonstrate generalization to unseen users and effective adaptation when inner-loop data come from similar (non-identical) users (cross-task adaptation).

2 Related Works

We start by reviewing the existing solution for behavioural intervention personalization. Since most of them would require the availability of specific datasets for their application, we also provide a snapshot of the available data for our specific field of interest, i.e., relationships between behaviours and cancer risks and effects of interventions for cancer risk reduction.

Among the most related works, Wang et al. [32] optimize context-aware notification timing to increase weekly running while capping message volume. A dynamic Bayesian network (DBN) simulates user cognition (e.g., memory accessibility, urge to run) and context. A REINFORCE policy decides whether or not to send a notification to the user. Khanshan et al. [11] address dropout and burden in Experience Sampling Method (ESM) studies by simulating participant responses with a DBN. The simulator evaluates adaptive prompting policies before deployment, mitigating RL's cold-start burden. Just-in-Time Adaptive Interventions (JITAIs) tailor support based on momentary vulnerability or opportunity, aiming to avoid user fatigue through data-driven timing and content selection [18]. While effective for timing, the methods proposed above neither adapt notification content nor generalize easily to domains where contextual signals are sparse (e.g., nutrition, mental health).

Early work on sedentary type-2 diabetes patients showed RL-personalized motivational texts improved walking and glycaemic control over static reminders [37]. The DIAMANTE trial [2] confirmed RL-tailored daily messages boosted step counts versus random or non-personalized texts. Tragos et al. [31] moved beyond messaging, using RL to recommend exercise sequences, improving retention and enjoyment. However, these studies either optimize message framing or focus narrowly on workout planning with fatigue models specific to physical exertion, i.e., schemes that are less transferable to broader habit formation. CarpeDiem [20] uses gamification for nutrition through weekly *missions*, e.g., vegetable servings, grounded in behaviour-change theory.

Another direction explored in behavioural change is to personalize how an action is communicated to the user. Computer-tailored persuasive messaging can improve attitudes but does not guarantee behaviour change. A study by d'Hondt, Nuijten, and Van Gorp (2019) [6] used personality-based profiling guides message framing via probabilistic user models. It showed that adaptive persuasive systems might induce elevated attitudes toward persuasive approaches, but these systems do not necessarily cause a change in health behaviour.

Matthews et al. [16] model fatigue as recoverable and unrecoverable components affecting effort-based choices, embedded in an MDP. Fatigue raises effort discounting, narrowing the value gap between acting and resting. While insightful, this binary work/rest framing and irreversible accumulation do not fit our setting, where no explicit rest action exists, and missions need graded difficulty with acceptance probabilities rather than binary choices. Our model directly estimates mission acceptance likelihood and adapts task difficulty to sustain engagement over time.

2.1 Data availability

At the outset, we considered whether existing longitudinal datasets could provide a basis for our study. For our purposes, such data would need to capture both behavioural outcomes and how these evolve in response to recommendations delivered over time. In particular, we surveyed European and national repositories that monitor cancer incidence alongside lifestyle factors, and datasets related to behavioural interventions. Large cohort studies such as EPIC [24], and population surveys like EHIS [1] and EU-SILC [7], contain rich information on diet, physical activity, alcohol, smoking, and socio-economic status. Several national registries (e.g., Sweden's COSM and SMC, Spain's *Sistema Nacional de Salud*, Slovenia's Cancer Registry) add long-term incidence and mortality records. Yet access to these sources is typically restricted, and most release only aggregated statistics, providing neither patient-level nor continuously updated data that link habits to cancer outcomes. Conversely, intervention-oriented datasets: CAPTURE-24 [5] (wearable activity tracking), LifeSnaps [36] (multimodal smartwatch study), and GLOBEM [35] (four-year well-being panel) offer fine-grained behavioural signals. However, they focus on short-term engagement or sensor outputs rather than verified behavioural change and cancer risk. None simultaneously supplies longitudinal, individual-level trajectories of both health behaviour and disease incidence that our user-model calibration demands.

The main datasets available outside Europe are the NIH–AARP Diet & Health Study [25], a prospective cohort of over 560,000 AARP members aged 50–71 recruited in 1995–1996 with diet and lifestyle questionnaires and cancer outcomes tracked via state registries and the National Death Index, the Nurses' Health Studies I, II, and 3 [3], long-running cohorts beginning in 1976 and 1989 with repeated questionnaires on diet, lifestyle, and medications and outcomes confirmed through medical records and registries, and the American Cancer Society's Cancer Prevention Studies CPS-II [4] Nutrition and CPS-3 [21], large cohorts initiated in 1992 and 2006–2013 that collect lifestyle data and biospecimens and follow participants for incident cancers through registry linkages. As for European datasets, the main U.S. cohorts face similar constraints: exposures are updated intermittently rather than continuously, for example biennial questionnaires in the Nurses' Health Studies and triennial surveys in CPS-3, and cancer outcomes are ascertained by state registry linkages that carry roughly a two-year reporting delay. In addition, sampling frames limit generalization, such as older AARP members in NIH–AARP and predominantly female, mostly white nurses in earlier NHS waves, and the availability of individual-level data is limited since access typically requires approved collaboration or data-use agreements rather than open microdata.

[1] Resource available at: https://ec.europa.eu/eurostat/web/microdata/european-health-interview-survey

Owing to these gaps, we elected to generate synthetic, longitudinal data with a virtual-user simulator, which lets us control the distribution of habits, fatigue dynamics, and intervention responses while maintaining full transparency and reproducibility for meta-learning experiments.

3 Problem Formulation

We model the problem of behavioural change as a Markov Decision Process (MDP) [22] where the user receives suggestions of tasks of varying difficulty to improve in one or more pillars. To represent observable and latent user factors, we build on the model of Lu et al. (2025) [15], which mathematically describes the acquisition of a new skill through actions of varying difficulty. In the original setting, all user characteristics and the policy that prescribes task difficulty are fixed. We opt to map their notion of *skill* to a behavioral *pillar to improve* (e.g., activity, sleep, diet) and adopt the same transition dynamics to model user change. For the sake of simplicity, in the following, we consider a single pillar, but an extension to multiple pillars is straightforward.

Formally, we consider an MDP $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma, \mu_0, T)$, composed by state space $\mathcal{S}$, action space $\mathcal{A}$, transition matrix P, reward function R, discount coefficient γ, and initial state distribution μ_0. More specifically, the elements of the state space $s_t \in \mathcal{S}$, visited by the agent at step t, is represented by a vector $s_t := (M_t, F_t, L_t)$, where M_t represents the motivation of the user, F_t is their fatigue, and L_t is the level of the pillar. The action that an agent performs, at round t, $a_t \in \mathcal{A} = \mathbb{R}^+$ represents the mission difficulty and the instantaneous reward r_t is represented by the pillar increment, i.e., $r_{t+1} = L_{t+1} - L_t$ and episodes are finite with stochastic transitions. We assume that the initial state of each user $\mu_0 = (M_0, F_0, L_0)$ is fixed and that the MDP has a known finite time horizon T. As common in the RL field, the objective is to learn a policy $\pi(s_t)$ that selects action a_t to perform at round t, i.e., selects the mission at the current round, to maximize the cumulative reward that in our case (using $\gamma = 1$) is the increment of the pillar level at the end of the time horizon w.r.t. the initial one, formally:

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T}(L_{t+1} - L_t)\right] = \mathbb{E}\left[L_T - L_0\right],$$

where the expectation is w.r.t. the stochasticity of the MDP, induced by the unknown transition model P which depends on the user model, and the one present in the policy π.

With a single user, or equivalently assuming that the population of the users is responding in the same way to missions with the same difficulty level a_t, this problem presents characteristics suitable for the use of classical RL techniques. However, since we cannot assume that all users respond the same way to our recommendations, the policy should adapt to users with different characteristics. Moreover, even learning a policy for a single type of user would require a significant number of samples. These issues lead us to exploit a meta-RL [30] approach, to have the ability to quickly adjust to new user characteristics and require only a limited amount of interactions. In our approach, we build on gradient-based meta-RL, seeking an initialization that can be fine-tuned to novel tasks in just a few gradient descent steps. The initialization policies are obtained by a literature-supported behavioural model that will be described in the following section.

4 Simulation Model

Due to the lack of real-world behavioral recommendation data, we relied on a synthetically defined user model to generate samples. In the following, we report a brief description of the model proposed by Lu et al. (2025). Formally, skill evolution is modeled as follows:

$$L_{t+1} = L_t + p_t f_S(a_t), \tag{1}$$

where $p_t \in \{0, 1\}$ is a boolean variable denoting whether the individual is working on that skill at time t ($p_t = 1$) or not ($p_t = 0$), and $f_S(a_t)$ is the function regulating the rate of skill acquisition depending on the mission difficulty a_t. More specifically, let us define the net drive to be $D_t := M_t - F_t$; then p_{t+1} is computed as follows:

$$p_{t+1} := \mathbb{1}\left\{(p_t = 0 \wedge D_t \geq \theta_t) \vee (p_t = 1 \wedge D_t > 0)\right\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, $\theta_t = \max\{a_t, \theta_{\min}\}$ is a start/keep threshold, and $\theta_{\min}$ is a given minimum difficulty level. In practice, if the user did not follow the suggested mission ($p_t = 0$),

i.e., was resting in the previous round, they start their mission if the drive is larger than θ_t. Conversely, if the user was already following the suggested mission in the previous round ($p_t = 1$), the user continues the mission only if they have positive drive $D_t > 0$. Thus, a harder task (i.e., large actions a_t) raises the barrier to start following the mission, while $\theta_{\min}$ enforces a baseline activation cost even for easy tasks. Instead, continuing to follow the same mission only requires positive drive D_t, creating a hysteresis between starting and continuing.

Regarding the skill acquisition function $f_S(\cdot)$, the selected model assumes a triangular shape as follows:

$$f_S(a) = \begin{cases} 0 & \text{if } a < 0 \\ \frac{a\overline{y}}{\overline{a}} & \text{if } 0 \leq a \leq \overline{a} \\ \frac{(-a+2\overline{a})\overline{y}}{\overline{a}} & \text{if } \overline{a} < a \leq 2\overline{a} \\ 0 & \text{if } a > 2\overline{a} \end{cases}, \tag{2}$$

where $\overline{a}$ is the user's optimal learning point, $\overline{y}$ caps the maximum per-step skill increase, and thus the overall learning speed.

Let us define the execution rate as follows:

$$E_t \;=\; p_t f_E(a_t), \tag{3}$$

where

$$f_E(a) = \begin{cases} \overline{y} & \text{if } x < 0 \\ \frac{(\overline{a}-a)\overline{y}}{\overline{a}} & \text{if } 0 \leq a \leq \overline{a} \\ 0 & \text{if } a > \overline{a} \end{cases}. \tag{4}$$

In practice, the function $f_E(\cdot)$ says that the execution rate is at a maximum when the demands of the task are fully met by one's skill, and it decreases linearly as relative task difficulty increases. Based on the execution rate, we can model the motivation M_t evolution of a user as follows:

$$\begin{aligned} M_{t+1} &= M_t + c_1(1 - M_t) + c_2\left[w\,\Delta L_t + (1-w)\,L_t\,E_t\right] \\ &= M_t + c_1(1 - M_t) + c_2\,p_t\left[w\,f_S(a_t) + (1-w)\,L_t\,f_E(a_t)\right], \end{aligned} \tag{5}$$

where $c_1 \in \mathbb{R}^+$ is a parameter that sets how fast motivation reverts toward its baseline, $c_2 \in \mathbb{R}^+$ scales the extra motivation change driven by learning/performance when working, $w \in [0, 1]$ controls the mix between progress-driven and performance-driven motivation gains, and E_t is the rate of execution, i.e., proportion of skill level required to conduct the task. Notice that the second part of the equation allows us to differentiate between learning-driven and performance-driven individuals. Indeed, motivation increases more whenever the individual is learning faster or when their performance is high, that is, if the user is highly skilled and executing a task with maximum execution.

Finally, the fatigue F_t is modeled assuming that it increases linearly whenever the individual is working, and it decreases to zero when the user is resting. Formally:

$$F_{t+1} \;=\; \begin{cases} F_t + c_3 & \text{if } p_t = 1 \\ F_t - c_4\,F_t & \text{if } p_t = 0 \end{cases}, \tag{6}$$

where $c_3 \in \mathbb{R}^+$ is the per-step fatigue accumulation rate and $c_4 \in \mathbb{R}^+$ is the recovery rate.

5 Meta learning approach

Based on the above model, we have that every choice of the parameter vector $v = (\overline{a}, \overline{y}, w)$ induces a distinct MDP characterizing a user that would require a specific optimal policy $\pi^*(v)$ to be learned.[2] Therefore, each parameter vector $v = (\overline{a}, \overline{y}, w)$ constitutes a separate task.

Our meta-learning framework builds on *Model-Agnostic Meta-Learning* (MAML) [8] by specializing it to our setting. In the following, we will detail only the main difference points from the original

[2]Notice that we opt for characterizing a user only with these three variables since they are the ones that mostly impact the simulation model above. A more general approach would also take into account the $c_1, \ldots, c_4$ constants. In the following, we will use the values of these coefficients as prescribed by [15], whose values are reported in the supplementary material.

version, deferring a detailed description of the framework to the paper [8]. In the original setup, a linear feature baseline estimated returns and advantages in the inner loop. However, from preliminary empirical analysis, we notice that this baseline underfits and produces noisy advantages. Therefore, we opt to use a Multi Layer Perceptron value baseline and train it together with the policy, which adds a value fitting term to the inner loop objective. To preserve exploration, we also included an entropy bonus. Finally, to improve the quality of the advantage estimate, we use generalized advantage estimation [28]. Together, these choices define an actor-critic loss for the inner loop and yield more stable and more sample-efficient updates across tasks.

The meta-learning approach is divided in two phases. In the *meta-training phase* we first leverage the simulator described in Section 4 to generate a diverse set of tasks $v_i = (\overline{a}_i, \overline{y}_i, w_i)$: we sample $\overline{a}_i$, $\overline{y}_i$, and w_i independently from bounded uniform distributions reported in Table 2. During meta-policy training, at every outer-loop iteration we roll out $N \in \mathbb{N}$ trajectories in the simulator under a sampled batch of tasks. These trajectories are then used to update the meta-policy following [8]. After this *meta-training* phase the *adaptation phase* is performed. We adapt the resulting meta-policy to new simulated users $v'_1, \dots v'_K$ (tasks), $K \in \mathbb{N}$ whose characteristics are unobserved: we collect on-policy trajectories for each new user with the unadapted (i.e., post-training phase) policy and use them exclusively to perform task-specific *inner-loop* updates.

To improve sample efficiency, an online-learning approach, specifically contextual bandits [14] was also considered at the beginning of our analysis. Although contextual bandits promise sample efficiency, our setting is an MDP: actions change future states, and motivation and fatigue accumulate, requiring long-horizon planning. With context restricted to observable features and key transition parameters $(\overline{a}, \overline{y}, w)$ unobserved, a bandit tends to converge to a single difficulty per user, learns myopically, and transfers poorly across users. Even though instantaneous learning peaks at $x = \overline{a}$, optimal actions may differ because high x raises the engagement threshold and can drive p to zero when $M - F$ is low, while smaller x sustains work and progress. Motivation depends on both learning progress and performance, so performance-oriented users (small w) benefit from easier tasks with higher execution, and fatigue dynamics can justify modulating x to avoid premature disengagement or to schedule recovery. Parameter uncertainty also requires exploratory variation in x to identify user characteristics, and near the horizon conservative choices can preserve remaining steps. These considerations motivate an RL formulation within users and a meta-RL approach across users, rather than a bandit model.

6 Experimental results

Our experimental setting is designed to evaluate (i) the overall effectiveness of MAML-based adaptation corresponding to the baselines introduced later in this section and (ii) a more realistic scenario in which the trajectories used during the *adaptation phase* are collected from users with similar characteristics. We trained our meta-learning algorithm, from now on addressed as `MAML`, over an environment simulating users.[3] In particular, similarly to what has been proposed in [8], to avoid overly aggressive updates during learning, the inner-loop learning rate is scheduled to decay linearly with the number of inner steps. Moreover, we use two separate entropy coefficients: a smaller one during the inner-loop adaptation and a larger one during the outer meta-update. The higher entropy weight at the meta step promotes broader exploration across tasks, while the lower weight in the inner loop keeps task-specific learning stable. For both the *meta-training* and the *adaptation* phases we sample $N = 30$ trajectories at each inner-loop step. We also assume our episode has a time horizon of $T = 20$.[4][5]

6.1 Baseline comparison

In the first setting, we compared our approach against three baselines: a randomly initialized policy `Random`; an Advantage Actor–Critic (A2C) policy, called `A2C pretrained`, in which the parameter

[3] The parameters used for the algorithm are reported in Table 1 in the supplementary material, and the ones for setting the environment in Table 2.

[4] In a real setting for behavioural mission recommendation, rounds t usually correspond to weeks.

[5] The code to fully reproduce the experiments will be released with the final paper version.

vector $(\overline{a}, \overline{y}, w)$ is re-sampled uniformly at each environment reset. We also add to the optimal policy, called `Oracle`, trained with A2C on a single task with full knowledge of its parameters.[6]

We aim to assess **meta-learning effectiveness**, i.e., whether a MAML-trained initialization improves performance on a broad set of held-out tasks after only a few adaptation steps. Second, we aim to evaluate whether the meta-training procedure yields a better starting point than the random policy. Third, we examine whether MAML outperforms the *mean policy* learned by `A2C pretrained` that for every environment reset, samples a task at random and performs task-specific updates without the meta-objective. We sample a total of 100 unseen tasks that we use for the comparison of the MAML algorithm and the baselines. For each task we perform the adaptation steps, sampling the same number of trajectories per step of the *meta-training phase*.

The main metrics we analyze are the average cumulative reward of the sampled trajectories over all tasks in the *adaptation phase* $\overline{R}^{\text{post}}$ and the improvement with respect to the adaptation step $\Delta\overline{R}$. Let us define for each new user $k \in \{1, \dots, K\}$ the trajectories generated before and after the adaptation provided by an algorithm as follows

$$\mathcal{H}_{\text{pre}}^{(k)} = \{\tau_{k,i}^{\text{pre}}\}_{i=1}^{N}, \qquad \mathcal{H}_{\text{post}}^{(k)} = \{\tau_{k,i}^{\text{post}}\}_{i=1}^{N},$$

respectively, where each trajectory is

$$\tau_{k,i}^{\bullet} = \big((s_{k,i,1}^{\bullet}, a_{k,i,1}^{\bullet}, r_{k,i,1}^{\bullet}), \dots, (s_{k,i,T}^{\bullet}, a_{k,i,T}^{\bullet}, r_{k,i,T}^{\bullet})\big),$$

where $\bullet \in \{\text{pre}, \text{post}\}$, and T is the horizon of the i^{th} trajectory for task k. For a specific task k, we define its average cumulative reward is defined as:

$$\overline{R}_k^{\bullet} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} r_{k,i,t}^{\bullet}.$$

For K validation tasks the average cumulative reward is:

$$\overline{R}^{\bullet} = \frac{1}{K} \sum_{k=1}^{K} \overline{R}_k^{\bullet}.$$

Finally, the improvement of the reward is formally defined as follows:

$$\Delta\overline{R} = \overline{R}^{\text{post}} - \overline{R}^{\text{pre}}.$$

Results In Figure 1 we report the evolution, as a function of the number of adaptation steps, of $\Delta\overline{R}$ and $\overline{R}^{\text{post}}$. The gap in terms of returns between adapted and unadapted trajectories is substantially larger than that achieved by `Random` policy. Moreover, the cumulative rewards obtained by `MAML` exceed those of the `A2C-Pretrained` policy after only one step, while the pretrained policy does not show a significant improvement post-adaptation. Figure 2 provides an example of the evolution of the adaptation steps of the post-adaptation results for different values of $\overline{a} \in \{1, 2, 3\}$ (keeping $\overline{y}$ and w fixed). The gap between `MAML` and `A2C-Pretrained` is larger for small values of $\overline{a}$, indicating that the average policy favors higher-value actions whose effects materialize only when $\overline{a}$ is also large, thereby penalizing tasks with low $\overline{a}$. However, we notice that it gets sufficiently close to the `Oracle` policy only for large values of $\overline{a}$, and the adaptation steps after the first do not significantly improve the performance.

6.2 Cross task adaptation

We investigate the **cross-task adaptation** of our methods, checking whether a meta-trained policy adapts to a target task using trajectories from slightly different tasks, rather than relying solely on on-task data. In this second experiment, we simulate an adaptation scenario in which data from similar participants entered the study simultaneously. We started from baseline values of the three parameters, which were randomized: $\overline{a} \in \{1, 2, 3\}$, $\overline{y} \in \{0.1, 0.5, 0.9\}$ and $w \in \{0, 0.5, 1\}$. For each

[6] Initially we also considered a contextual bandit as a baseline due to the fast capabilities it generally has for learning with a few samples. However, the learned policy was not able to generalize well across users, most likely due to the MDP nature of the problem itself.

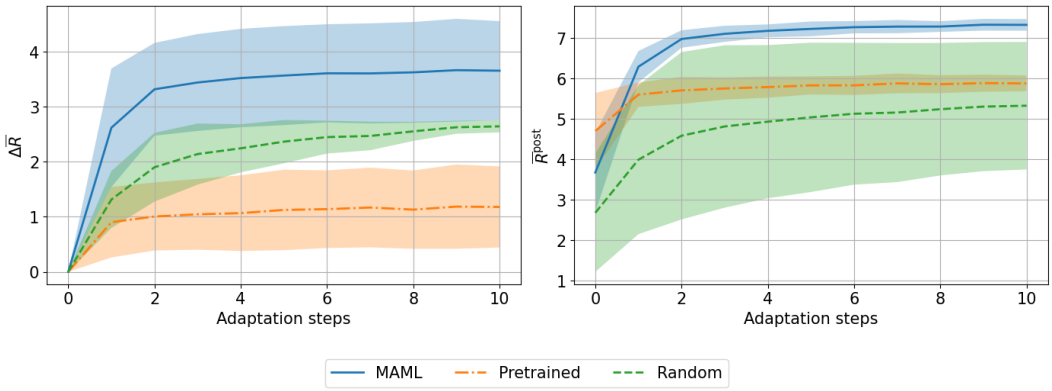

Figure 1: Evolution of mean $\Delta\overline{R}$ per adaptation step (left) and evolution of $\overline{R}^{\text{post}}$ per adaptation step across adapted tasks (right). Shaded areas represent the standard deviation computed on 5 different seeds, for both the meta-training and testing phase.

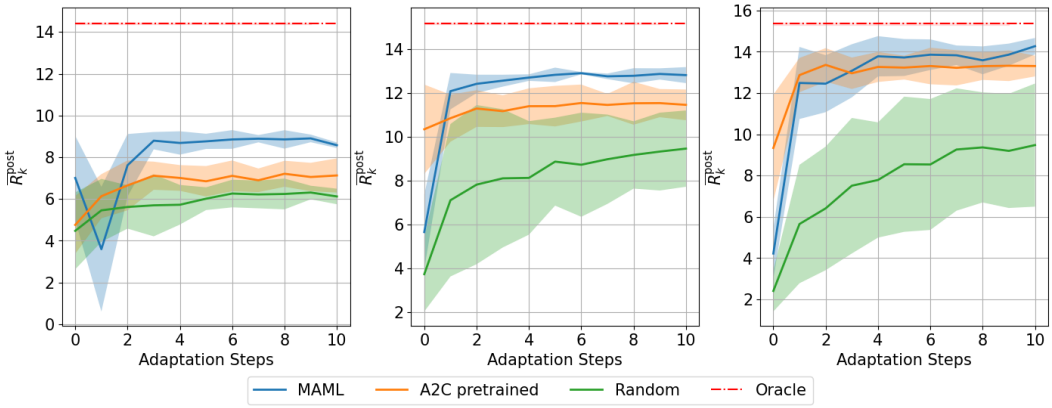

Figure 2: Evolution of $\overline{R}_k^{\text{post}}$ per adaptation step for different values of $\overline{a} = 1$ (left), $\overline{a} = 2$ (center), $\overline{a} = 3$ (right), keeping $\overline{y} = 0.9$ and $w = 0.5$ fixed. Shaded areas represent the standard deviation computed on 5 different seeds, for both the meta-training and testing phase.

adaptation step from 1 to 3, we sampled values of $\overline{a}$, $\overline{y}$ and w uniformly from the interval $i \pm \epsilon_i$, where $i = \{\overline{a}, \overline{y}, w\}$ and $\epsilon = (\epsilon_{\overline{a}}, \epsilon_{\overline{y}}, \epsilon_w)$. In total, for each adaptation step of each base task, we generate 30 similar tasks, the same number of trajectories used during the *meta-training phase* for each adaptation step of the inner loop. From each of these 30 tasks, we sample one trajectory, constructing a batch of N trajectories from similar tasks. We use this batch of trajectories from similar tasks to perform the inner loop steps. Furthermore, we compared performances for different values of $\epsilon_{\overline{a}}$, keeping the other two $\epsilon_{\overline{y}}$ and ϵ_w fixed, selecting this parameter for our analysis as it has the greatest impact on the obtained results.

As a metric, we report $\overline{R}^{\text{post}}$ computed by sampling post-adaptation trajectories $\mathcal{H}_{\text{post}}^{(k)}$ on the base tasks only.

Results Figure 3 shows the results for the different combinations of baseline tasks with $\epsilon_{\overline{a}} = 0.2$, $\epsilon_{\overline{y}} = 0.2$, $\epsilon_w = 0.2$. We observe that the main differences across the three algorithms occur when $\overline{a} = 1$. In the other two cases, we can observe that `MAML` is significantly better than the random policy but has no significant advantage over the `A2C pretrained` policy. This also comes from the fact that the pretrained policy is already effective with large values of $\overline{a}$, so a small variation in the parameters does not result in significant underperformance. Furthermore, no performance degradation is observed when using adaptation with similar users. Figure 4 shows the results related to the second case in which we keep $\epsilon_{\overline{y}}$ and ϵ_w fixed at 0.2 and vary $\epsilon_{\overline{a}} \in \{0.2, 1, 1.5\}$. The most significant performance differences are mostly present when $\overline{a} = 1$, where greater user diversity

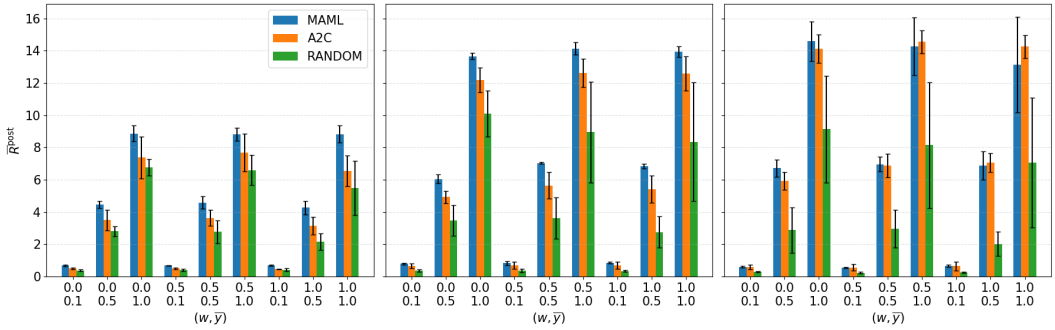

Figure 3: $\overline{R}_k^{\text{post}}$ after 3 adaptation steps with randomly generated similar tasks used in the adaptation process with $\overline{a} = 1$ (left), $\overline{a} = 2$ (center), $\overline{a} = 3$ (right).

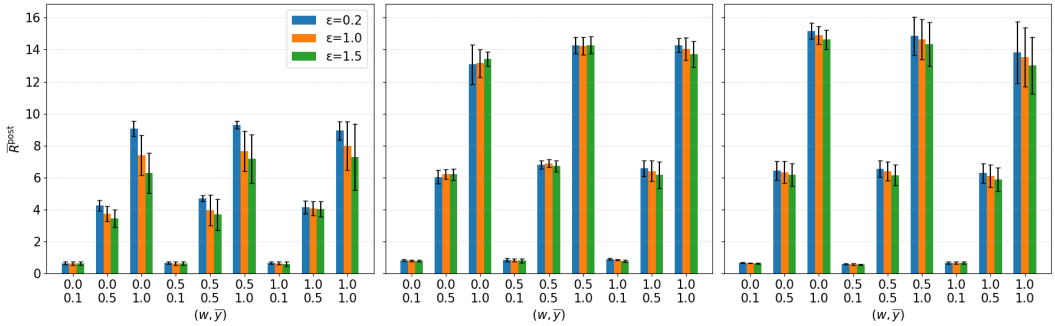

Figure 4: $\overline{R}_k^{\text{post}}$ after 3 adaptation steps for $\epsilon_{\overline{a}} \in \{0.2, 1, 1.5\}$ with $\overline{a} = 1$ (left), $\overline{a} = 2$ (center), $\overline{a} = 3$ (right).

corresponds to lower performance. This effect might be due to the fact that as $\epsilon_{\overline{a}}$ increases, the sampling neighborhood $[\overline{a} - \epsilon_{\overline{a}}, \overline{a} + \epsilon_{\overline{a}}]$ widens, placing more mass on values with $\overline{a} < 1$. For smaller $\overline{a}$, the set of actions yielding positive reward, $\{a : \mathbb{E}[R \mid a, \overline{a}] > 0\}$, shrinks. Consequently, the reward signal weakens and the policy's adaptation becomes substantially more difficult.

7 Conclusion and future works

In this study, we introduce a meta-reinforcement learning framework for tailored behavioral suggestions that mimics a variety of participant profiles using a synthetic user model. Two basic variables, fatigue and motivation, were included in the formulation of the behavioral suggestions problem. The task difficulty adaptation for users with diverse features was subsequently presented as a meta-RL problem in which we want to minimize the amount of data and, consequently, the number of user interactions required for personalization. Finally, our experiments demonstrated that gradient-based meta-learning can adapt effectively to new user configurations within just a few adaptation steps, and that leveraging trajectories from similar users can maintain performance without requiring on-task data. We note that the benefits of the meta-learning approach are contingent on the user model producing substantial variation in optimal policy structure.

Future work will focus on three parallel directions. First, we aim to validate our approach with empirical datasets, linking real-world behavioral trajectories to health outcomes, to assess the realism and transferability of our simulator-based findings. Second, we plan to extend the environment so that fatigue dynamics also depend on the magnitude of the chosen action, allowing for more nuanced modeling of user burden. Third, we will investigate offline meta-RL methods such as PEARL [23] to fully exploit user similarity and historical data without extensive online interaction.

References

[1] Sasan Adibi, editor. *Mobile health; A technology road map*. Springer Series in Bio-/Neuroinformatics. Springer International Publishing, January 2015.

[2] Adrian Aguilera, Marvyn Arévalo Avalos, Jing Xu, Bibhas Chakraborty, Caroline Figueroa, Faviola Garcia, Karina Rosales, Rosa Hernandez-Ramos, Chris Karr, Joseph Williams, Lisa Ochoa-Frongia, Urmimala Sarkar, Elad Yom-Tov, and Courtney Lyles. Effectiveness of a digital health intervention leveraging reinforcement learning: Results from the diabetes and mental health adaptive notification tracking and evaluation (DIAMANTE) randomized clinical trial. *J. Med. Internet Res.*, 26:e60834, October 2024.

[3] Ying Bao, Monica L Bertoia, Elizabeth B Lenart, Meir J Stampfer, Walter C Willett, Frank E Speizer, and Jorge E Chavarro. Origin, methods, and evolution of the three nurses' health studies. *Am. J. Public Health*, 106(9):1573–1581, September 2016.

[4] Eugenia E Calle, Carmen Rodriguez, Eric J Jacobs, M Lyn Almon, Ann Chao, Marjorie L McCullough, Heather S Feigelson, and Michael J Thun. The american cancer society cancer prevention study II nutrition cohort: rationale, study design, and baseline characteristics. *Cancer*, 94(9):2490–2501, May 2002.

[5] Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data*, 11(1):1135, 2024.

[6] Jens E d'Hondt, Raoul C Y Nuijten, and Pieter M E Van Gorp. Evaluation of computer-tailored motivational messaging in a health promotion context. In *Modeling and Using Context*, Lecture notes in computer science, pages 120–133. Springer International Publishing, Cham, 2019.

[7] Eurostat. Eu statistics on income and living conditions (eu-silc), scientific use files. Microdata via Eurostat, 2020-2024.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017.

[9] Mark A Guadagnoli and Timothy D Lee. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *J. Mot. Behav.*, 36(2):212–224, June 2004.

[10] Farhad Islami, Ann Goding Sauer, Kimberly D Miller, Rebecca L Siegel, Stacey A Fedewa, Eric J Jacobs, Marjorie L McCullough, Alpa V Patel, Jiemin Ma, Isabelle Soerjomataram, W Dana Flanders, Otis W Brawley, Susan M Gapstur, and Ahmedin Jemal. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the united states. *CA Cancer J. Clin.*, 68(1):31–54, January 2018.

[11] Alireza Khanshan, Pieter Van Gorp, and Panos Markopoulos. Simulating participant behavior in experience sampling method research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, April 2023. ACM.

[12] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychol.*, 34S(Suppl):1220–1228, December 2015.

[13] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, ICUIMC '08, page 208–211, New York, NY, USA, 2008. Association for Computing Machinery.

[14] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *CoRR*, abs/1003.0146, 2010.

[15] Mingzhen Lu, Tyler Marghetis, and Vicky Chuqiao Yang. A first-principles mathematical model integrates the disparate timescales of human learning. *Npj Complex.*, 2(1), May 2025.

[16] Julian Matthews, M Andrea Pisauro, Mindaugas Jurgelis, Tanja Müller, Eliana Vassena, Trevor T-J Chong, and Matthew A J Apps. Computational mechanisms underlying the dynamics of physical and cognitive fatigue. *Cognition*, 240(105603):105603, November 2023.

[17] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.

[18] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann. Behav. Med.*, 52(6):446–462, May 2018.

[19] OECD. *Beating cancer inequalities in the EU*. OECD Health Policy Studies. OECD, January 2024.

[20] Silvia Orte, Carolina Migliorelli, Laura Sistach-Bosch, Meritxell Gómez-Martínez, and Noemi Boqué. A tailored and engaging mhealth gamified framework for nutritional behaviour change. *Nutrients*, 15(8):1950, April 2023.

[21] Alpa V Patel, Eric J Jacobs, Daniela M Dudas, Peter J Briggs, Cari J Lichtman, Elizabeth B Bain, Victoria L Stevens, Marjorie L McCullough, Lauren R Teras, Peter T Campbell, Mia M Gaudet, Elizabeth G Kirkland, Melissa H Rittase, Nance Joiner, W Ryan Diver, Janet S Hildebrand, Nancy C Yaw, and Susan M Gapstur. The american cancer society's cancer prevention study 3 (CPS-3): Recruitment, study design, and baseline characteristics. *Cancer*, 123(11):2014–2024, June 2017.

[22] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[23] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *CoRR*, abs/1903.08254, 2019.

[24] Elio Riboli, KJ Hunt, Nadia Slimani, Pietro Ferrari, Teresa Norat, M Fahey, UR Charrondiere, B Hemon, C Casagrande, J Vignat, et al. European prospective investigation into cancer and nutrition (epic): study populations and data collection. *Public health nutrition*, 5(6b):1113–1124, 2002.

[25] Arthur Schatzkin, Amy F Subar, Frances E Thompson, Linda C Harlan, Joseph Tangrea, Albert R Hollenbeck, Paul E Hurwitz, Linda Coyle, Nicki Schussler, Dominique S Michaud, et al. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health–american association of retired persons diet and health study. *American journal of epidemiology*, 154(12):1119–1125, 2001.

[26] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May 1987.

[27] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.

[28] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

[29] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[30] Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009.

[31] Elias Tragos, Diarmuid O'Reilly-Morgan, James Geraci, Bichen Shi, Barry Smyth, Cailbhe Doherty, Aonghus Lawlor, and Neil Hurley. Keeping people active and healthy at home using a reinforcement learning-based fitness recommendation framework. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6237–6245, California, August 2023. International Joint Conferences on Artificial Intelligence Organization.

[32] Shihan Wang, Chao Zhang, Ben Kröse, and Herke van Hoof. Optimizing adaptive notifications in mobile health interventions systems: Reinforcement learning from a data-driven behavioral simulator. *J. Med. Syst.*, 45(12):102, October 2021.

[33] Thure Georg Weimann and Carola Gißke. Unleashing the potential of reinforcement learning for personalizing behavioral transformations with digital therapeutics: A systematic literature review. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, 2024.

[34] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3-4):229–256, May 1992.

[35] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in neural information processing systems*, 35:24655–24692, 2022.

[36] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Sci. Data*, 9(1):663, October 2022.

[37] Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *J. Med. Internet Res.*, 19(10):e338, October 2017.

A Background

In this section, we introduce the necessary background on reinforcement learning and meta learning.

A.1 Reinforcement Learning

Reinforcement Learning (RL) [29] models decision making as an agent interacting with an environment to maximize cumulative reward. At each step t, the agent observes state s_t, samples action $a_t \sim \pi_\theta(\cdot \mid s_t)$, then receives reward r_{t+1} and transitions to s_{t+1}. Under the Markov assumption, future outcomes depend only on the current state–action pair.

Asynchronous Advantage Actor-Critic (A2C) [17] is an on-policy actor–critic method that maintains:

- **Actor** $\pi_\theta(a \mid s)$: a parameterized stochastic policy.
- **Critic** $V_\phi(s)$: a value network estimating the expected discounted return.

During training, the agent runs its policy for n steps (or until termination) and computes the ***n-step return***

$$R_t \;=\; \sum_{i=0}^{n-1} \gamma^i \, r_{t+i+1} \;+\; \gamma^n \, V_\phi\big(s_{t+n}\big)\,,$$

from which the *advantage* is obtained:

$$\hat{A}_t = R_t - V_\phi(s_t)\,.$$

The networks are updated jointly by minimizing the combined loss

$$\mathcal{L}(\theta,\phi) = -\,\mathbb{E}_t\big[\log \pi_\theta(a_t \mid s_t)\,\hat{A}_t\big] \;+\; \alpha\,\mathbb{E}_t\big[\big(V_\phi(s_t) - R_t\big)^2\big] \;-\; \beta\,\mathbb{E}_t\big[\mathcal{H}(\pi_\theta(\cdot \mid s_t))\big]\,,$$

where the first term improves the policy toward positive advantages, the second fits the value estimate, and the third (weighted by β) encourages exploration via entropy regularization.

A.2 Meta Learning

Meta-learning [26], often referred to as "learning to learn", seeks to endow models with the ability to rapidly adapt to new tasks using only a small amount of data. In the classical supervised setting, a model is trained on a single task by minimizing a loss function over many examples. In contrast, a meta-learner is trained over a *distribution of tasks* $\mathcal{T} \sim p(\mathcal{T})$, so that after meta-training it can quickly fine-tune to an unseen task $\mathcal{T}_i$ using only a handful of samples.

Formally, each task $\mathcal{T}_i$ is defined by

$$\mathcal{T}_i = \big\{\mathcal{L}_i,\; q_i(\mathbf{x}),\; q_i(\mathbf{x}' \mid \mathbf{x}, a),\; H_i\big\},$$

where $\mathcal{L}_i$ is the task-specific loss, $q_i(\mathbf{x})$ the initial data distribution, $q_i(\mathbf{x}' \,|\, \mathbf{x}, a)$ the transition (if any), and H_i the horizon or episode length. During meta-training, one alternates between:

1. **Inner update (task adaptation):** For each sampled task $\mathcal{T}_i$, update the model parameters θ to θ_i' by taking one or more gradient steps on $\mathcal{L}_i$ using K examples (the "K-shot" setting).
2. **Meta-update:** Adjust the original parameters θ so as to minimize the loss of the adapted models θ_i' across the task batch.

A powerful yet simple meta learning algorithm that we will use in this work is Model-Agnostic Meta-Learning (MAML) [8] instantiates this idea in a task and model agnostic way by directly optimizing θ for *fast adaptivity*. In one step of gradient descent on a task $\mathcal{T}_i$, MAML computes

$$\theta_i' = \theta - \alpha\,\nabla_\theta \mathcal{L}_i\big(f_\theta\big)\,,$$

and then performs the meta-update

$$\theta \leftarrow \theta - \beta \sum_i \nabla_\theta \mathcal{L}_i\big(f_{\theta_i'}\big)\,.$$

By differentiating through the inner update, MAML learns an initialization θ from which only a few gradient steps are required to achieve strong performance on a new task. An adaptation of this algorithm for reinforcement learning is proposed in the original paper and employs REINFORCE [34] for the inner loop and TRPO [27] for the outer loop.

B Training hyperparameters

Tables 1 and 2 contain the parameters used in the experimental study.

Parameter	Value
Discount Factor (γ)	0.99
GAE Lambda	0.95
First-Order	False
Hidden Sizes	[64, 64]
Nonlinearity	tanh
Inner Loop Batch Size (N)	30
Number of Inner Steps	3
Inner Loop Learning Rate	0.1
Number of Batches	200
Outer Loop Batch Size (K)	100
Maximum KL Divergence	0.1
CG Iterations	10
CG Damping	1
Line Search Max Steps	15
Line Search Backtrack Ratio	0.5
Critic coefficient	0.25
Inner Entropy Coefficient	0.01
Outer Entropy Coefficient	0.1
Initial Step Size of Line Search	0.1

Table 1: Parameters for the meta learning algorithm training

Parameter	Value
S_0	0
M_0	0.5
p_0	1
F_0	0
E_0	0
c_1	0.5
c_2	1.2
c_3	0.1
c_4	0.1
T	20
$\theta_{\min}$	0.005
w	$\mathcal{U}([0, 1])$
$\overline{a}$	$\mathcal{U}([1, 3])$
$\overline{y}$	$\mathcal{U}([0.1, 1])$

Table 2: Environment parameters. w, $\overline{a}$ e $\overline{y}$ sono campionati da distribuzioni uniformi sui rispettivi intervalli.