Meta Learning MDPs with Linear Transition Models

Robert Müller
Technical University of Munich

Aldo Pacchiano
Microsoft Research, NYC

Abstract

We study meta-learning in Markov Decision Processes (MDP) with linear transition models in the undiscounted episodic setting. Under a task sharedness metric based on model proximity we study task families characterized by a distribution over models specified by a bias term and a variance component. We then propose BUC-MatrixRL, a version of the UC-Matrix RL algorithm (Yang and Wang, 2019) and show it can meaningfully leverage a set of sampled training tasks to quickly solve a test task sampled from the same task distribution by learning an estimator of the bias parameter of the task distribution. The analysis leverages and extends results in the learning to learn linear regression and linear bandit setting to the more general case of MDP's with linear transition models. We prove that compared to learning the tasks in isolation, BUC-Matrix RL provides significant improvements in the transfer regret for high bias low variance task distributions.

1 INTRODUCTION

Meta learning (Schmidhuber, 1987; Naik and Mammone, 1992) is a long-standing quest in machine learning. The goal is to use the experience gained in past tasks to solve new tasks, coming from the same task distribution, quickly. Since meta learning is a problem formulation, it can be combined with supervised learning, bandits or reinforcement learning (RL). While the move to bandits adds the challenge of exploration, the move to RL additionally introduces temporal dependence in the data.

To better understand which type of transfer is possible

in meta RL we ask the following questions: i) what is a family of MDP's? ii) how are the family members related? iii) how can we algorithmically leverage the shared structure? iv) given a characterisation of a family of MDP's, can we describe and quantify the gain of doing meta RL compared to standard RL on the individual tasks (**I**ndepedent **T**ask **RL**)? Recently, there has been a growing body of empirical work in meta RL (Zintgraf et al., 2020; Rakelly et al., 2019), addressing the question of how to algorithmically leverage shared structure. While these successes are impressive, it remains largely open to study conditions under which and to what extent transfer of knowledge is possible.

Depending on the setting, the definition of the task distribution is a delicate matter. For general MDP it remains an open question to quantify and formalise task similarity. Recently Müller et al. (2020) proposed different task embeddings, which are either coupled or independent of the existence of a differentiable policy, that give rise to a distance between tasks. Another distance metric between MDP is the maximum of state-action wise differences in transitions and reward. The well-known simulation lemma (Kearns and Singh, 2002) uses it to quantify the zero-shot performance of a policy in an unknown MDP which is a known distance apart from a known MDP.

An increasingly well-understood setting for sequential decision-making is to assume linear structure in the MDP, for example in the transition model (Yang and Wang, 2019; Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Cai et al., 2020; Zhou et al., 2020; He et al., 2020; Zhou et al., 2021; Jin et al., 2020). To come closer to study what and when meta RL is useful, we restrict ourselves in a first step to the subset of MDP with a linear transition model. This restriction allows us to represent each MDP uniquely via its transition core matrix; the space of matrix norms equips us with a natural notion of distance between tasks. In this restricted setting, we answer the question:

Can meta RL yield performance gains compared to ITRL in the space of MDP with linear transition model?

To quantify the gain of using meta RL we analyse the

transfer regret. Specifically, we propose and analyse in section 4 a biased version of UC-MatrixRL that follows the spirit of meta learning in ridge regression settings via learning a bias Denevi et al. (2018, 2019); Cella et al. (2020). While doing so, we tighten the analysis of UC-MatrixRL Yang and Wang (2019) by using a more careful log-determinant lemma. We quantify the transfer regret in an oracle setting and characterize properties of families of tasks for which meta RL improves compared to independent task learning. Since the oracle setting is not realistic, we conclude our paper by proposing two practical estimators of the mean task distribution transition core and analyse the resulting transfer regret.

Our paper leverages results from learning to learn in a linear regression setting. The basic concept of learning a bias vector or matrix on the meta train tasks to transfer knowledge via biased linear regression to the test tasks is applicable to all (provably efficient) algorithms based on linear regression in any MDP setting. For ease of presentation and because of its simplicity, we demonstrate the machinery on MDP with linear transition matrices.

2 RELATED WORK

While deep meta reinforcement learning has obtained large empirical successes, we wish to emphasize the lack of theoretical work in meta-learning for RL. A common approach in RL theory is to study tabular problems or MDP under linearity assumptions. This includes in particular linear mixture MDP, whose transition model is a linear function of an a priori given feature map of state-action-next state triples (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Cai et al., 2020; Zhou et al., 2020; He et al., 2020; Zhou et al., 2021), with the special case of MDP with linear transition model (Yang and Wang, 2019). This is the special case considered by our work.

Inspired from empirical successes, there is a growing body of literature on theoretical work on learning to learn in the supervised setting (Denevi et al., 2018, 2019; Khodak et al., 2019; Konobeev et al., 2021; Tripuraneni et al., 2021). Several works studied meta learning in a bandit setting (Azar et al., 2013; Deshmukh et al., 2017; Cella et al., 2020; Yang et al., 2021). Some works study multitask RL (Brunskill and Li, 2013; Calandriello et al., 2015; Lu et al., 2021). Conditions under which zero-shot RL is possible were recently investigated by Malik et al. (2021).

A particularly related line of work (Denevi et al., 2018, 2019) propose the idea of learning to learn via learning a bias in supervised learning. The knowledge acquired during the meta train phase is distilled into a bias

vector. At meta test time, the learner is regularised towards this learned bias vector. This inspiration is extended by Cella et al. (2020) to the stochastic bandit setting. Our paper is to Yang and Wang (2019) as Cella et al. (2020) is to Abbasi-Yadkori et al. (2011).

3 BACKGROUND

We use small bold letters to denote vectors (e.g. $\boldsymbol{x}$) and large bold letters to denote matrices (e.g. $\boldsymbol{X}$). Let $\boldsymbol{A} \in R^{d \times d}$ be a positive definite matrix. We denote the Mahalanobis norm of a vector as $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}$. Similarly, we denote the matrix version of the Malahonobis norm: $\|\boldsymbol{A}^{1/2} \boldsymbol{X}\|_F^2 = \text{tr}(\boldsymbol{X}^T (\boldsymbol{A}^{1/2})^T \boldsymbol{A}^{1/2} \boldsymbol{X}) = \text{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}) = \sum_j \|\boldsymbol{X}_j\|_{\boldsymbol{A}} = \|\boldsymbol{X}\|_{\boldsymbol{A}}^2$. Further, we denote the 2-1 matrix norm, which is the sum of the euclidean norms of the matrix columns: $\|\boldsymbol{X}\|_{2,1} = \sum_j \|\boldsymbol{X}_j\|_2 = \sum_j \langle \boldsymbol{X}_j, \boldsymbol{X}_j \rangle = \sum_j \sqrt{\sum_i \boldsymbol{X}_{i,j}^2}$. We abbreviate the set $\{1, \ldots, n\}$ as $[n]$. We denote by $\sum_{n',h}^{n,H}$ the double sum $\sum_{n'=1}^n \sum_{h=1}^H$.

In this section, we introduce the single task objective regret as well as the multi-task objective of meta transfer regret. We review the UC-MatrixRL algorithm Yang and Wang (2019) and derive a tighter regret bound using a more careful log determinant lemma.

3.1 RL in MDP's with linear transitions

We start by reviewing the undiscounted episodic setting. An MDP is given as sextuple: $\mathcal{M} = (S, A, r, P, H, \mu_0)$ with state-space S, action-space A, reward function $r : S \times A \to [0, 1]$, transition probabilities $p : S \times A \times S \to [0, 1]$, horizon H and starting distribution μ_0. To simplify our presentation we will assume a single starting state s_0. Extending our results beyond the setting of a Dirac initial distribution is straightforward. The learner is to play for N episodes, each consisting of H steps of interaction with the environment, for a total of $T = NH$ transitions. Throughout our discussion we will use both notations of time interchangeably and note the conversion: $t(n, h) = nH + h$. Thus $(n, h) + 1$ denotes either $(n, h + 1)$ if $h < H$ or otherwise the first step in the next episode $(n + 1, 1)$.

A policy is a mapping $\pi : S \times [H] \to A$ that maps each state, stage pair to an action. The value function of a policy π at stage h is given as: $V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^H r(s_t, \pi_t(s_t))\right]$ and the action value function as $Q_h^\pi(s, a) = r(s, a) + P(\cdot|s, a)V_{h+1}^\pi$. The optimal policy π^* is defined as the policy that maximizes the expected sum of future rewards: $\pi^* = \arg\max_{\pi \in \Pi} V^\pi(s) \ \forall s \in S$. The goal of the learner playing action a_t at time t

is to minimize the regret after $T = NH$ steps:

$$R_T(\mathcal{M}) = \sum_{n=1}^{N} \left[V^*(s_0) - \left(\sum_{h=1}^{H} r(s_{n,h}, a_{n,h}) \right) \right] . \quad (1)$$

A good learner achieves regret that is sublinear in T. Using regret as a performance metric allows us to compare different approaches.

In this work, we consider each MDP $\mathcal{M} = (S, A, r, P, H, \mu_0)$ as a task. A family of tasks is thus a finite or infinite set of MDPs. Within this paper we consider task families with shared S, A and r, so all the change happening between environments is in the transition probabilities P [1]. We restrict our analysis to the set of MDP with linear transition model.

Definition 3.1 (MDP with linear transition model). *A MDP $\mathcal{M}$ has a* linear transition model, *if for each $(s_t, a_t) \in S \times A, s_{t+1} \in S$ and a priori given feature maps $\phi(s_t, a_t) \in \mathbb{R}^d$ and $\psi(s_{t+1}) \in \mathbb{R}^{d'}$, there exists an unknown matrix $M^* \in \mathbb{R}^{d \times d'}$, which is referred to as* transition core, *such that:*

$$P(\tilde{s}|s, a) = \phi(s, a)^T M^* \psi(\tilde{s}). \quad (2)$$

Confining ourself to the space of MDP with linear transition model we see that each MDP $\mathcal{M}^*$ is uniquely characterised by its transition core M. Let $\mathcal{T}$ be a distribution of transition cores. A family of linear transition MDP's $\mathcal{M}$ with shared S, A and features ϕ and ψ can similarly be characterised as a set of transition cores $M \sim \mathcal{T}$ of its members. So, we write interchangeably $\mathbb{E}_{\mathcal{M} \sim \mathcal{M}}$ and $\mathbb{E}_{M \sim \mathcal{T}}$.

Our paper considers the meta learning setting. In a first phase the learner interacts with the training tasks from the training task distribution $\mathcal{M}_{train} \subseteq \mathcal{M}$. The goal of the meta learner is to reduce in the subsequent test phase the expected meta transfer regret after T steps on the test task distribution $\mathcal{M}_{test} \subseteq \mathcal{M}$:

$$\mathrm{Mtr}_T(\mathcal{M}_{test}) = \mathbb{E}_{\mathcal{M} \sim \mathcal{M}_{test}} Regret(T, \mathcal{M}) . \quad (3)$$

3.2 UC-MatrixRL algorithm

UC-MatrixRL is an algorithm for MDP with linear transitions proposed by Yang and Wang (2019). It switches between estimating the unknown transition core $\hat{M}_n$ on encountered transitions (s_t, a_t, s_{t+1}) and acting greedily with respect to an optimistic estimate of the Q-function build using $\hat{M}_n$. We describe both steps below.

---

[1]Note that, Yang and Wang (2019) point out, that the assumption of a shared reward function could be lifted by adding an optimistic reward estimation step like in LINUCB Dani et al. (2008)

The feature maps $\phi_{n,h} = \phi(s_{n,h}, a_{n,h}) \in \mathbb{R}^d$ and $\psi_{n,h} = \psi(s_{(n,h)+1}) \in \mathbb{R}^{d'}$ are fixed and a priori given. Using the identity matrix $I \in \mathbb{R}^{d \times d}$, we define the following matrices: $K_\psi = \sum_{\tilde{s} \in S} \psi(\tilde{s}) \psi(\tilde{s})^T$, $V_n = \sum_{n' \le n, h \le H} \phi_{n',h} \phi_{n',h}^T$ and $V_n^\lambda = \lambda I + V_n$. Note that V_n is a symmetric matrix, $V_0^\lambda = \lambda I$ and we abbreviate $M_{n,H} = M_n$. We have furthermore:

$$\mathbb{E}\left[\phi_{n,h}\psi_{n,h}^T K_\psi^{-1}|s_{n,h}, a_{n,h}\right] = \phi_{n,h}\phi_{n,h}^T M^* . \quad (4)$$

M_n can be estimated via the ridge regression problem:

$$\arg\min_{M} \sum_{n',h}^{n,H} \|\psi_{n',h}^T K_\psi^{-1} - \phi_{n',h}^T M\|_2^2 + \lambda\|M\|_F^2 , \quad (5)$$

which has the solution:

$$M_n = (V_n^\lambda)^{-1} \sum_{n',h}^{n,H} \phi_{n',h}\psi_{n',h}^T K_\psi^{-1} . \quad (6)$$

UC-MatrixRL constructs optimistic value function estimates based on a matrix ball, $B_n = B_{\beta_n}(M_n)$ of radius, β_n, centred at the current estimate of the transition core $\hat{M}_n$. In episode n, the optimistic Q-function is recursively constructed as:

$$Q_{n,H+1}(s, a) = 0 \; \forall (s, a) \in S \times A$$
$$Q_{n,h}(s, a) = r(s, a) + \max_{M \in B_n} \phi(s, a)^T M \psi^T V_{n,h+1} \; h \in [H] ,$$

where the value function satisfies $V_{n,h}(s) = \max_a Q_{n,h}(s, a)$. Let $\Psi = \left[\psi(s_1), \ldots, \psi(s_{|S|})\right]^T \in R^{S \times d'}$ be the matrix of concatenated ψ features of all states. In order to show regret bounds for the algorithm, we require regularity conditions in feature space.

Assumption 3.1. *(Feature regularity) For positive constants $C_M, C_\phi, C_\psi, C'_\psi$ we have:*

1. *$\|\phi(s, a)\|_2^2 \le C_\phi \; \forall (s, a) \in S \times A$,*

2. *$\|\Psi K_\psi^{-1}\|_{2,\infty} \le C'_\psi$*

3. *$\|\Psi^T v\|_2 \le C_\psi \|v\|_\infty \; \forall v \in \mathbb{R}^S$*

4. *$\|M^*\|_F^2 \le C_M d$*

These assumptions are almost the same as assumptions $2'$ in Yang and Wang (2019), only 1. differs in that they assume $\|\phi(s, a)\|_2^2 \le C_\phi d$. The set of assumptions leads to a construction of a matrix ball of Frobenius norm.

3.3 Regret bound of UC-MatrixRL

Yang and Wang (2019) show in theorem 2 that under assumption 3.1, with first item

modified as described in the previous section, UC-MatrixRL has regret upper bounded in $O\left[C_\psi\sqrt{\|\boldsymbol{M}^*\|_F^2 + C_\psi'^2 \ln(1 + \frac{NHC_\phi}{\lambda})}\right]dH^2\sqrt{T}$. The proof relies on a version of the log-determinant lemma (lemma 8) on $|\phi_{n,h}\|_{V_n^{-1}}$. While the remainder of the paper could be done based on this analysis, we reanalyse this self normalised norm of the summed features, which improves the horizon dependence by a factor of $\sqrt{H}$ and allows emphasising the impact of the bias.

For notational convenience, we will define: $D := 1 + \frac{nHC_\phi}{\lambda d}$. In appendix B we prove the following version of the log determinant lemma.

Lemma 3.1 (Alternative Log Determinant Lemma). *The following inequality holds,*

$$\sum_{n=1}^N \sum_{h=1}^H \|\phi_{n,h}\|_{V_n^{-1}} \leq \sum_{n=1}^N \sum_{h=1}^H 2\|\phi_{n,h}\|_{V_{n,h}^{-1}} + \frac{C_\phi}{\lambda} d\log(D).$$

Intuitively, this means that the self normalised norm of the sum of feature matrices grows significantly slower than H. Despite using an inverse norm on the left hand side defined via the V_n matrices and lacking all the inter-episode outer products, we can show the sum of these inverse norms can be upper bounded by the a constant multiple of the sum of the inverse norms that defined by the inter-episode $V_{n,h}$ matrices plus a term of the form $O(d\log(n))$. This lemma allows us to apply the determinant lemma and save a factor of $\sqrt{H}$ in the regret.

Let us define $C_{\phi,\lambda} := (4 + C_\phi/\lambda)$. Using our version of the log determinant lemma results in the following regret bound.

Theorem 3.2 (Regret UC-Matrix RL). *The regret of the UC-MatrixRL algorithm, under assumptions 3.1, after $T = NH$ steps in MDP $\mathcal{M}$ is upper bounded with probability at least $1 - 1/(NH)$ as:*

$$R_T(\boldsymbol{M}^*) \leq \left(C_\psi'\sqrt{d'd\log(NHD)} + \sqrt{\lambda}\|\boldsymbol{M}^*\|_F\right)$$
$$2C_\psi H\sqrt{C_{\phi,\lambda}Td\ln(D)} \quad (7)$$

The theorem is a special case of theorem 4.2, the regret of biased UC-MatrixRL, with bias matrix $\boldsymbol{W} = 0$. Our bound allows replacing a $\sqrt{2H}$ factor by $\sqrt{C_{\phi,\lambda}} = \sqrt{(4 + C_\phi/\lambda)}$ which is independent of the horizon H.

4 BIASED UC-MATRIX RL WITH ORACLE ACCESS

In this section we describe and analyse a biased version of the UC-MatrixRL algorithm from Yang and Wang

(2019). We will highlight how the optimal bias improves the regret bound.

Given a fixed bias matrix $\boldsymbol{W} \in \mathbb{R}^{d \times d'}$, the learner faces at the end of the n-th episode the ridge regression problem:

$$\boldsymbol{M}_n = \arg\min_{\boldsymbol{M}} \sum_{n',h}^{n,H} \|\psi_{n',h}^T \boldsymbol{K}_\psi^{-1} - \phi_{n',h}^T \boldsymbol{M}\|_2^2$$
$$+ \lambda\|\boldsymbol{M} - \boldsymbol{W}\|_F^2 . \quad (8)$$

We refer to $\boldsymbol{W}$ as the bias matrix or bias transition core. Using ridge regression with the change of variables $\boldsymbol{M} = \boldsymbol{B} + \boldsymbol{W}$ and solving for $\boldsymbol{B}$ yields for $\boldsymbol{M}_n$ the solution:

$$\boldsymbol{W} + \left(\boldsymbol{V}_n^\lambda\right)^{-1} \sum_{n',h}^{n,H} \phi_{n,h}\left(\psi_{n',h}^T\boldsymbol{K}_\psi^{-1} - \phi_{n',h}^T\boldsymbol{W}\right). \quad (9)$$

The resulting algorithm is given in algorithm 1.

Algorithm 1 Within Task Biased Upper-Confidence Matrix RL (BUC-MatrixRL)

Input: MDP(S, A, P, r, s_0, H), features $\phi : S \times A \to \mathbb{R}^d$, $\psi : S \to \mathbb{R}^{d'}$, $\lambda > 0$, $\hat{W}_0$, N
Initialize: $\hat{\boldsymbol{M}}_0 = \hat{\boldsymbol{W}}_0$, $(\boldsymbol{V}_0^\lambda)^{-1} = \frac{1}{\lambda}\boldsymbol{I}$
for episode $n = 1, \ldots, N$: **do**
    Build optimistic Q-function using $\boldsymbol{M}_n, \beta_n$
    for step $h = 1, \ldots, H$: **do**
        select greedy action $a_{n,h}$ using $Q_{n,h}$
        record next state $s_{n,h+1}$
    end for
    update feature matrix $\boldsymbol{V}_{n+1} = \boldsymbol{V}_n + \sum_h^H \phi_{n,h}\phi_{n,h}^T$
    recompute ellipsoid radius β_{n+1} like in eq. 11
    possibly update bias estimate $\hat{\boldsymbol{W}}_{n+1}$
    recompute core estimate $\hat{\boldsymbol{M}}_{n+1}$ using eq. 9
end for

4.1 Analysis of biased UC-MatrixRL

We begin by analysing the biased UC-MatrixRL algorithm and analyse afterwards ways to obtain an estimate of the bias. The proof works in two steps, first we show that the optimal transition core $\boldsymbol{M}^*$ is with high probability in an over time shrinking ball around M_n and secondly we bound the resulting regret. The employed proof techniques are primarily based on Yang and Wang (2019), Abbasi-Yadkori et al. (2011) and Cella et al. (2020).

4.1.1 Naive construction of confidence balls

In order to construct confidence balls we will use self normalizing inequalities. We use them to bound the

deviation of the optimal transition core from the algorithm's estimate $\hat{\boldsymbol{M}}_n$.

Theorem 4.1 (Confidence Ellipsoid BUC-MatrixRL)**.** *Let $\delta > 0$. BUC-MatrixRL with bias matrix $\boldsymbol{W}$, produces under the assumptions 3.1 with probability at least $1 - \delta$ for all $t > 0$ estimates $\hat{\boldsymbol{M}}_n$, of the core matrix, that satisfy:*

$$\|\hat{\boldsymbol{M}}_n - \boldsymbol{M}^*\|_F \leq C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log(D)} \tag{10}$$
$$+ \sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}^*\|_F =: \beta_n^{\boldsymbol{W}}(\delta) \ .$$

Thus for all times $t > 1$ and recalling the regularity of the feature map ϕ we have with probability at least $1 - \delta$, that the true transition core $\boldsymbol{M}^*$ is contained in the ellipsoid $C_n(\delta) = C_{\beta_n^{\boldsymbol{W}}(\delta)}^{\boldsymbol{W}}(\hat{\boldsymbol{M}}_n) = \{\boldsymbol{M} \in \mathbb{R}^{d \times d'} | \|\hat{\boldsymbol{M}}_n - \boldsymbol{M}\|_F \leq \beta_n^{\boldsymbol{W}}(\delta)\}$ of radius $\beta_n^{\boldsymbol{W}}(\delta)$ and centroid $\hat{\boldsymbol{M}}_n$. Clearly it holds $\beta_n^{\boldsymbol{W}}(\delta) \leq \beta_n^{\boldsymbol{0}}(\delta)$, which combined with item 4 of assumption 3.1 yields the radius $\beta_n^{\boldsymbol{0}}(\delta)$:

$$C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log(D)} + \sqrt{\lambda}dC_m \tag{11}$$

It is a common choice to select $\delta = 1/(NH)$. We defer the proof to the appendix C.1. Having obtained the confidence ball, we are now ready to obtain a regret bound for BUC-MatrixRL.

4.1.2 Regret bound

Using the fact that an ellipsoid of shrinking volume around the current core matrix estimate contains the true core matrix with high probability, results in the following regret:

Theorem 4.2 (Regret BUC-Matrix RL)**.** *Under the assumptions of theorem 4.1 after $T = NH$ steps, choosing the ellipsoid radius $\beta_n^{\boldsymbol{W}}(\delta)$ as in 4.1 BUC-MatrixRL abides with probability at least $1 - 1/(NH)$ the following bound on the regret:*

$$R_T(\boldsymbol{M}^*) \leq \left(C'_\psi \sqrt{d'd \log(TD)} + \sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}^*\|_F\right)$$
$$2C_\psi H \sqrt{C_{\phi,\lambda} T d \ln(D)}$$

We defer the proof to appendix C.2. We wish to emphasise the behaviour of biased UC-MatrixRL for two special choices of bias transition cores, the origin $\boldsymbol{0} \in \mathbb{R}^{d \times d'}$ and the true transition core $\boldsymbol{M}^*$:

1. Choosing an **uninformed transition core** $W = \boldsymbol{0} \in \mathbb{R}^{d \times d'}$ recovers unbiased UC-MatrixRL, thus the regret bound from theorem 3.2 applies.

2. **Oracle Prior** When the algorithm has access to an oracle, it is possible to set the bias transition core of BUC-MatrixRL to the true transition core, $\boldsymbol{W} = \boldsymbol{M}^*$. As a result the term $\sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}^*\|_F$ vanishes. Recalling the definition of D as $1 + \frac{nHC_\phi}{\lambda d}$, it is clear that the regret goes to 0 as $\lambda \to \infty$. Thus, the bigger the regularisation strength, λ, the smaller the regret we suffer. This is since for large λ, $\bar{\boldsymbol{M}}_n = \boldsymbol{M}^*$, thus the learner selects action greedily with respect to the true Q-function obtained by using the true dynamics.

4.2 Meta Transfer Regret Bound

The goal of meta learning to generalise from the training tasks to all tasks of the task distribution at hand. While we bounded the single task regret in the previous section, what we are really interested in is the transfer regret, thus $E_{\mathcal{M} \sim \mathscr{M}} R_T(\mathcal{M})$, where $\mathscr{M}$ is the distribution of all MDP. To do so, we define the mean and variance of the task distribution. As each sampled transition core M gives rise to an MDP $\mathcal{M} \in \mathscr{M}$, we have an equivalence of sampling transition cores M from the distribution of transition cores $\mathscr{T}$ or sampling MDP $\mathcal{M}$ from the distribution of MDPs $\mathscr{M}$. Thus, we can write interchangeably $\mathbb{E}_{\mathcal{M} \sim \mathscr{M}}$ and $\mathbb{E}_{\boldsymbol{M} \sim \mathscr{T}}$. Recall the transfer regret from equation 3 as being the expectation of the individual task regret across the distribution of considered MDPs $\mathscr{M}$.

We are interested to study the impact of the bias matrix $\boldsymbol{W}$ on the incurred transfer regret. Let us define the variance and the mean absolute distance of the bias matrix relative to the distribution of environments $\mathcal{T}$:

$$\text{Var}_W(\mathcal{T}) = \mathbb{E}_{\boldsymbol{M} \sim \mathcal{T}}\left[\|\boldsymbol{M} - \boldsymbol{W}\|_F^2\right]$$
$$\text{Mad}_W(\mathcal{T}) = \mathbb{E}_{\boldsymbol{M} \sim \mathcal{T}}\left[\|\boldsymbol{M} - \boldsymbol{W}\|_F\right] \ .$$

We skip the argument, the task distribution $\mathcal{T}$, whenever it is clear form the context that it is either the training task or test task distribution. We are now ready to state a first bound on the Mtr:

Theorem 4.3 (Meta Transfer Regret BUC-MatrixRL)**.** *Under the assumptions of theorem 4.2 we have with probability at least $1 - 1/(NH)$ for a task distribution $\mathcal{T}$ the following Mtr after T steps per task (where we absorb constant factors into C):*

$$\text{Mtr}_T(\mathcal{T}) \leq CC_\psi H C'_\psi d \sqrt{d'TC_{\phi,\lambda} \log(TD) \ln(D)}$$
$$+ CC_\psi H \text{Mad}_{\boldsymbol{W}} \sqrt{\lambda TC_{\phi,\lambda} d \ln(D)}$$
$$\leq CC_\psi H C'_\psi d \sqrt{d'TC_{\phi,\lambda} \log(TD) \ln(D)}$$
$$+ CC_\psi H \sqrt{\text{Var}_{\boldsymbol{W}} \lambda TC_{\phi,\lambda} d \ln(D)}$$

We will now interpret the obtained upper bound on the transfer regret for different choices of regularisation strength λ and bias matrix $\boldsymbol{W}$:

1. We obtain in the limit of an infinite regularisation strength (omitting a constant 4 from $C_{\phi,\lambda}$):

$$\lim_{\lambda \to \infty} \mathrm{Mtr}_T(\mathcal{T})$$
$$\leq \lim_{\lambda \to \infty} CC_\psi H \sqrt{\mathrm{Var}_{\boldsymbol{W}} \lambda T d \ln \left(1 + \frac{nHC_\phi}{\lambda d} \right)}$$
$$= CC_\psi H \sqrt{\mathrm{Var}_{\boldsymbol{W}} T^2 C_\phi}$$

a $\sqrt{\mathrm{Var}_{\boldsymbol{W}}}$ dependence of the transfer regret on the variance of the task distribution with respect to the bias matrix $\boldsymbol{W}$.

2. It turns out that for a particular choice of λ we can actually improve the $\sqrt{\mathrm{Var}_{\boldsymbol{W}}}$ dependence of the transfer regret. For a matrix $\boldsymbol{Q} \in \mathbb{R}^{d \times d'}$, we define the function $\mathrm{DVar} : R^{d \times d'} \to \mathbb{R}$; $\mathrm{DVar}(Q) := \left(1 + \frac{T^2 \mathrm{Var}_{\boldsymbol{Q}} C_\phi}{d} \right)$. Let $\lambda = \frac{1}{T \mathrm{Var}_W}$, which is motivated by the following two aspects. Firstly, the reliance on the prior should shrink the more samples are seen in a particular task, thus the $1/T$ part. Moreover, we want the learner to rely less on the prior $\boldsymbol{W}$ if the task distribution is very broad as indicated by a high task variance $\mathrm{Var}_{\boldsymbol{W}}$, thus the $1/\mathrm{Var}_{\boldsymbol{W}}$. The resulting transfer regret is:

$$\mathrm{Mtr}_T(\mathcal{T}) = \left(1 + C_\psi' \sqrt{d' dT \log\left(T\mathrm{DVar}(\boldsymbol{W}) \right)} \right)$$
$$CC_\psi H \sqrt{C_{\phi,\lambda} d \ln\left(\mathrm{DVar}(\boldsymbol{W}) \right)} .$$

This means that the choice of regularisation strength $\frac{1}{T\mathrm{Var}_{\boldsymbol{W}}}$ yields a $\sqrt{\log(1 + \mathrm{Var}_{\boldsymbol{W}})}$ dependence, whereas the limit of $\lambda \to \infty$ yields a worse dependence of $\sqrt{\mathrm{Var}_{\boldsymbol{W}}}$.

3. Keeping $\lambda = \frac{1}{T\mathrm{Var}_{\boldsymbol{W}}}$ the transfer regret BUC-MatrixRL with the bias matrix chosen as mean transition core $\bar{\boldsymbol{M}}$ goes to zero as the variance of the task distribution goes to 0.

4. **Oracle BUC-MatrixRL** improves against individual task learning, whenever the variance of the task distribution is much lower than its offset from the origin:

$$\mathrm{Var}_{\bar{\boldsymbol{M}}} = \mathbb{E}_{\boldsymbol{W} \sim \mathcal{T}} \| \boldsymbol{M} - \bar{\boldsymbol{M}} \|_F^2$$
$$\ll \mathbb{E}_{\boldsymbol{M} \sim \mathcal{T}} \| \boldsymbol{M} \|_F^2 = \mathrm{Var}_0 .$$

Choosing an appropriate step size of $\lambda = \frac{1}{T\mathrm{Var}_{\boldsymbol{W}}}$ allows for a dependence of the upper bound of the transfer regret of $\sqrt{\ln(1 + \mathrm{Var}_{\boldsymbol{W}})}$. In the limit of no variance of the task distribution we recover the single task regret bound from theorem 4.2.

5 PRACTICAL META LEARNING WITH BIASED UC-MATRIX RL

So far we assumed access to an oracle of the optimal transition core $\boldsymbol{M}^*$ and showed its usefulness in terms of incurred transfer regret. In any practical setting we do however have no access to an oracle, thus we want to transfer knowledge from previous tasks. Following the meta learning protocol, we have access to G many training tasks. Each task g was created by sampling $M_g \sim \mathcal{T}$. We run biased UC-MatrixRL with a bias transition core based on the transition core estimates of previous tasks for N episodes a H steps in MDP $\mathcal{M}_g$. To enable generalisation across the full distribution of tasks, we use at meta test time BUC-Matrix RL with a bias distilling the transferable knowledge from meta training. The full training protocol is given in algorithm 2. In the case of sequentially arriving tasks $1, \ldots, g$ we can always consider all previous tasks $1, \ldots, g-1$ as training tasks, and the test set as the one element set containing only the g-th task. Thus, our approach caters also to the continual learning setting.

Algorithm 2 Meta Train
___
    Input: set of training tasks $\mathcal{M}_1, \ldots, \mathcal{M}_G$, features $\phi : S \times A \to \mathbb{R}^d$, $\psi : S \to \mathbb{R}^{d'}$, $\lambda > 0$, bias for the first task $\hat{\boldsymbol{W}}_0$, episode number N, horizon H
    Initialize: $\hat{M}_{0,0} = \hat{\boldsymbol{W}}_0$, $\boldsymbol{V}_0^{-1} = \frac{1}{\lambda} \boldsymbol{I}$
    for train task $\mathcal{M}_g \in \{\mathcal{M}_1, \ldots, \mathcal{M}_G\}$: **do**
        run alg. 1 on MDP $\mathcal{M}_g$ with bias $\boldsymbol{W} = \hat{\boldsymbol{W}}_{g-1}$
    end for

Inspired from the bias estimators in Cella et al. (2020) we describe now two methods to estimate the bias transition core matrix in episode n of task g, $\boldsymbol{W}_{g,n}$, based on previous experience.

To account for the multitask setting we add a subscript indicating the task index $g \in [G]$. Recall our assumption that all tasks share the same feature maps ϕ and ψ and the interchangeability of the T and nh time within a task. For the g-th MDP, we have the transition core M_g, feature matrix $\boldsymbol{V}_{g,t} = \boldsymbol{V}_{g,t//H,t\%H} = \sum_{p \leq t} \phi(s_{g,p}) \phi(s_{g,p})^T = \sum_{p \leq t} \phi_{g,p} \phi_{g,p}^T$, the concatenation of all features in episode n: $\Phi_{g,n} = [\phi_{g,n,1}, \ldots, \phi_{g,n,H}]^T$ and $\Psi_{g,n} = [\psi_{g,n,1}, \ldots, \psi_{g,n,H}]^T$ and the radius of the confidence ellipsoid at stage t as $\beta_{g,t}^{\mathbf{0}}$. A quantity central to our analysis is the mean estimation error of the bias $\hat{\boldsymbol{W}}_{g,n,h}$ at stage (n, h) of the g-th task with respect to the true mean transition core $\bar{\boldsymbol{M}}$:

$$\epsilon_{g,n,h}(\mathcal{M}) = \| \bar{\boldsymbol{M}} - \hat{\boldsymbol{W}}_{g,n,h} \|_F^2 . \quad (12)$$

As the learner faces the tasks in a sequential manner there is an inherent estimation error, due to the fact

that it has only samples from the task distribution. We define $\bar{W}_{G,t} = \frac{1}{GT+t}\left(\sum_{g=1}^{G} T M_g + t M_{G+1}\right)$ the mean transition core of the observed MDP and denote the estimation error relative to the true mean transition core $\bar{M}$ as $H_{\mathscr{M}}(G+1, \bar{M}) = \|\bar{M} - \bar{W}_{G,t}\|_F$.

5.1 Averaging previous transition core estimates - a low bias estimator

The first approach is to use an weighted average of previous transition core estimates as bias. The motivation is that the knowledge acquired in previous MDP $\mathcal{M}_1$, ..., $\mathcal{M}_G$ is distilled in the respective final estimate of the transition core $\hat{W}_{G,T}$. Knowledge transfer between the tasks is achieved by aggregation of the individual estimated transition cores:

$$\hat{W}_{G,n,h} = \sum_{g=1}^{G-1} \frac{T}{Z} \hat{M}_{g,T} + \frac{nH+h}{Z} \hat{M}_{G,n,h}, \quad (13)$$

with normalisation factor $Z = T(G-1) + nH + h$. This choice of bias estimator results in the following bound on the transfer regret.

Theorem 5.1. *BUC-MatrixRL incurs after T interactions in G previous tasks, using the bias estimator $\hat{W}_{G,n,h}$ as in equation 13, $\lambda = \frac{1}{T Var_{\hat{W}_{G,n,h}}}$ and under assumptions 3.1, with probability at least $1 - 1/(NH)$ at most the following meta transfer regret:*

$$\mathrm{Mtr}_T(\mathcal{M}_{G+1}) \leq CC_\psi H d C'_\psi$$

$$\sqrt{C_{\phi,\lambda} d' T \log\left(T + \frac{T^3 C_\phi \left(Var_{\bar{M}} + \epsilon_{G,T}(\mathscr{M})\right)}{d}\right)}$$

The mean estimation error can be upper bounded as:

$$\sqrt{\epsilon_{G,T}(\mathscr{M})} \leq H_{\mathscr{M}}(G+1, \bar{M}) + \max_{g \in [G]} \frac{\beta^{\mathbf{0}}_{g,T}(1/NH)}{\lambda_{min}^{1/2}(V_{g,T}^\lambda)}$$

The proof is based on theorem 4.3 and is deferred to appendix D. Analysing the bound on the noise we show that, by a Frobenius matrix norm version of Bennetts inequality, the first term goes for an increasing number of tasks to zero. This means the estimation error is dominated by the second term. Recall our choice of $\lambda = 1/(TVar_{\hat{W}})$, we see that the estimation error increases with the variance of our bias matrix estimator. The meta learning procedure comes with an additional storage requirement of the size of the transition core in which the estimates of the training tasks are averaged.

5.2 Global ridge regression - a high bias estimator

The previous estimator shared knowledge between MDP's $\mathcal{M}_g$ via the final estimated transition cores.

Here we present an estimator that instead builds features on all transitions seen in all previous MDP and performs one global ridge regression to estimate the transition core matrix which we use as bias in biased UC-MatrixRL. This approach is inspired by the high bias estimator in (Cella et al., 2020) and is in line with previous estimators in the multitask bandit literature. The knowledge transfer between tasks works thus in form of feature embeddings of the observed (s, a, s') transitions instead of aggregated objects. Let $\tilde{V}_{G,n,h} := \sum_{g=1}^{G} V_{G,T} + V_{G+1,n,h}$. Global ridge regression uses the estimator:

$$\hat{W}_{G,n,h} = (\tilde{V}_{G,n,h}^\lambda)^{-1} \left[\sum_{g=1}^{G-1} \sum_{n,h}^{N,H} \phi_{g,n,h} \psi_{g,n,h} K_\psi^{-1} + \sum_{n',h'}^{n,h} \phi_{G,n',h'} \psi_{G,n',h'} K_\psi^{-1} \right]. \quad (14)$$

This yields the following meta transfer regret:

Theorem 5.2. *BUC-MatrixRL incurs after T interactions in G previous tasks, using the bias estimator $\hat{W}_{G,n,h}$ as in equation 14, $\lambda = \frac{1}{T Var_{\hat{W}_{G,n,h}}}$ and assumptions 3.1, with probability at least $1 - 1/(NH)$ at most the following meta transfer regret:*

$$\mathrm{Mtr}_T(\mathcal{M}_{G+1}) \leq CC_\psi H d C'_\psi$$

$$\sqrt{C_{\phi,\lambda} d' T \log\left(T + \frac{T^3 C_\phi \left(Var_{\bar{M}} + \epsilon_{G,T}(\mathcal{T})\right)}{d}\right)}$$

Let $\nu_{\min} = \lambda_{\min}\left(\tilde{V}_{G,n,h}\right)$ be the minimal singular value of the global feature matrix. Then the mean estimation error can be upper bounded as:

$$\sqrt{\epsilon_{G,T}(\mathscr{M})} \leq H_{\mathscr{M}}(G+1, \bar{M})$$

$$+ \underbrace{\frac{dC_M}{\lambda + \nu_{\min}} + C'_\psi \sqrt{\frac{2}{\lambda + \nu_{\min}} \log\left(NH + \frac{GN^2 H^2 C_\phi}{\lambda d}\right)}}_{\frac{\beta^{\mathbf{0}}(1/(GNH))}{\lambda + \nu_{\min}}}$$

$$+ 2(G+1) \max_{g \in [G+1]} \tilde{H}(G+1, M_g).$$

$\tilde{H}(G+1, M_g)$ is a weighted version of the estimation error $H_{\mathscr{M}}(G+1, \bar{M})$ towards the current transition core M_g:

$$\tilde{H}(G, M_g) = H_{\mathscr{M}}(g, M_g) \sigma_{\max}\left(V_{g,T} \tilde{V}_{G,N,H}^{-1}\right), \quad (15)$$

where $\sigma_{\max}\left(V_{g,T} \tilde{V}_{G,N,H}^{-1}\right)$ quantifies the misalignment of task g to the other tasks observed so far.

The regret bound follows from theorem 4.3 and is derived in appendix E. Comparing this to the low

bias solution in theorem 5.1 we see that the variance is now $\frac{\beta^{\mathbf{0}}(1/(GNH))}{\lambda+\nu_{\min}}$ instead of $\frac{\beta^{\mathbf{0}}_{g,T}(1/NH)}{\lambda^{1/2}_{min}(V^{\lambda}_{g,T})}$. From observing, $\nu_{min} \geq \frac{G}{dd'}\lambda_{min}(\boldsymbol{V}_g) \ \forall g \in [G]$ it follows that we shrink the variance by a factor $(dd')/G$. As the number of training tasks G goes to infinity, the variance goes to zero. This comes however at the price of increased bias $2(G+1)\max_{g\in[G+1]}\tilde{H}(G+1, \boldsymbol{M}_g)$ which increases proportional to the task misalignment, $\sigma_{\max}\left(V_{g,T}\tilde{V}^{-1}_{G,N,H}\right)$. We illustrate its behaviour on two corner cases.

First, assume that the task distribution $\mathcal{T} = \{\boldsymbol{M}\}$, so our task distribution, consists of a transition single core matrix. As we face always the same task, we expect extremely favourable transfer. In particular, we suffer in this case no bias as $\tilde{H}(G+1, M_g) = 0$.

The other extreme are completely unrelated tasks. The minimal relatedness we can generate is if corresponding vectors of the transition cores are orthogonal to each other. Consider the task distribution, where each column of the transition cores is a d-dimensional basis vector. In this case one can do meta learning on $d-1$ mutually different training tasks without seeing an improved meta transfer regret on the d-th task.

Note that the linear MDP models requires feature maps ϕ and ψ as input. The transition cores live in a feature map dependent space, making it hard to connect properties of the transition cores of a task family to interpretable properties of the environments.

5.3 ITRL vs MTRL

We can now compare the gain of doing meta learning compared to independent task learning. In the independent learner scenario, the transfer regret depends on the variance of the tasks relative to the origin, $\mathrm{Var}_{\mathbf{0}}$. Doing meta learning allows us to obtain estimators that replace the variance with respect to the origin by the variance to the mean task $\mathrm{Var}_{\bar{\boldsymbol{M}}}$ plus an additional error term. In the oracle case, the error term is zero. Since in practice there is no oracle, we bound the additional error term for two different bias estimation methods. For any task distribution of small variance but large offset, the obtained transfer regret incurred by meta learning is lower than from ITRL.

6 DISCUSSION

Our paper gives an affirmative answer to the initial question of the usefulness of meta RL. Using the one-to-one correspondence of MDP with linear transition core and the core matrix, we have a notion of distance between tasks. This allows the characterisation of any distribution of linear transition core MDPs via its offset

and variance. We prove a decrease in transfer regret of meta RL compared to independent task learning whenever the variance of the task distribution is small compared to the offset from the origin. While we show this improvement first in a setting with access to an oracle that reveals the offset of the transition core distribution, we extend the result to two practical estimators.

One estimator performs knowledge transfer between the tasks using an aggregation of distilled knowledge obtained on previous tasks in the form of estimated transition cores. This method suffers, however, a possibly large error due to the direct proportionality of the estimation error, thus transfer regret with the variance of the transition core estimator. The second proposed estimator suffers bias proportional to the task misalignment in a trade-off for a variance that goes to zero as the number of tasks goes to infinity. Here, the transfer of knowledge happens via the sharing of embeddings of the observed transitions. Note that similar meta learning via learning a bias estimators can be combined with any (provably efficient) RL algorithm based on linear regression for similar gains compared to ITRL.

Our work chooses meta transfer regret at performance measure, which weights all test tasks equally. An interesting avenue of future work is to analyse alternative objectives, for example the worst case transfer regret within the test distribution.

A major limitation of the framework of linear transition core MDP is the assumption of known feature maps ϕ and ψ. This allows studying the usefulness of meta learning for rapid learning in a newly encountered task. In empirically successful meta deep RL works, the feature embeddings are, however, not given but instead learned. Thus, the effectiveness of meta learning could lie within learning a good initialisation/bias or in learning a set of reusable features. For the case of model agnostic meta learning (Finn et al., 2017) the empirical study (Raghu et al., 2019) finds that feature reuse is the dominant factor in the examined few shot classification and RL tasks. To fully understand the usefulness of meta learning in general MDP's, it is thus necessary to also take feature learning into account.

We lastly wish to emphasize that each linear MDP can be trivially embedded as its generating vector/matrix. The space of vector-/ matrix norms gives immediately rise to distances between tasks. It remains however an open question how to embed general MDP and to develop distance metrics between the task embeddings that are meaningful, for example in the sense that learning in nearby tasks aids learning in target tasks.

Acknowledgements

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvari, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. F. (2020). Model-based reinforcement learning with value-targeted regression.

Azar, M. G., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models.

Brunskill, E. and Li, L. (2013). Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization.

Calandriello, D., Lazaric, A., and Restelli, M. (2015). Sparse multi-task reinforcement learning. *Intelligenza Artificiale*, 9(1):5–20.

Cella, L., Lazaric, A., and Pontil, M. (2020). Meta-learning with stochastic linear bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1360–1370. PMLR.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. (2019). Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575. PMLR.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Learning to learn around a common mean. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018)*, volume 31. NIPS Proceedings.

Deshmukh, A. A., Dogan, U., and Scott, C. (2017). Multi-task learning for contextual bandits.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

He, J., Zhou, D., and Gu, Q. (2020). Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.11566*.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In *L4DC*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. *ArXiv*, abs/2101.02195.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232.

Khodak, M., Balcan, M.-F., and Talwalkar, A. (2019). Provable guarantees for gradient-based meta-learning.

Konobeev, M., Kuzborskij, I., and Szepesvári, C. (2021). A distribution-dependent analysis of meta-learning.

Lai, T. L., Wei, C. Z., et al. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lu, R., Huang, G., and Du, S. S. (2021). On the power of multitask representation learning in linear mdp.

Malik, D., Li, Y., and Ravikumar, P. (2021). When is generalizable reinforcement learning tractable?

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR.

Müller, R., Parker-Holder, J., and Pacchiano, A. (2020). Taming the herd: Multi-modal meta-learning with a population of agents.

Naik, D. K. and Mammone, R. J. (1992). Meta-neural networks that learn by learning. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pages 437–442 vol.1.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. (2019). Rapid learning or feature reuse? towards understanding the effectiveness of maml. *ArXiv*, abs/1909.09157.

Rakelly, K., Zhou, A., Finn, C., Levine, S., and Quillen, D. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5331–5340, Long Beach, California, USA. PMLR.

Schmidhuber, J. (1987). Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172.

Tripuraneni, N., Jin, C., and Jordan, M. I. (2021). Provable meta-learning of linear representations.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. (2021). Impact of representation learning in linear bandits.

Yang, L. F. and Wang, M. (2019). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound.

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes.

Zhou, D., He, J., and Gu, Q. (2020). Provably efficient reinforcement learning for discounted mdps with feature mapping.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. (2020). Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*.

SUPPLEMENTARY MATERIAL

Organization of Supplementary Material

A NOTATION

Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{A} \in R^{d \times d}$ a positive definite matrix. We define the Mahalanobis norm $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d'}$ we have the Frobenius norm and the inducing matrix inner product norm $\|\boldsymbol{X}\|_F = \sqrt{tr(\boldsymbol{X}^T \boldsymbol{X})} = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X} \rangle_F} = \sqrt{\sum_{i,j} |\boldsymbol{X}_{i,j}|^2}$. We have further the Mahalanobis version of the Frobenius norm for a symmetric positive definite $\boldsymbol{A} \in R^{d \times d}$:

$$\|\boldsymbol{A}^{1/2} \boldsymbol{X}\|_F^2 = \mathrm{tr}(\boldsymbol{X}^T (\boldsymbol{A}^{1/2})^T \boldsymbol{A}^{1/2} \boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}) = \sum_j \|\boldsymbol{X}_j\|_{\boldsymbol{A}} = \|\boldsymbol{X}\|_{\boldsymbol{A}}^2 \tag{16}$$

We denote the column indices of $\boldsymbol{X}$ as $j \in \{1, \ldots, d'\}$ and row indices $i \in \{1, \ldots, d\}$. We denote the 2-1 matrix norm, which is the sum of the euclidean norms of the matrix columns: $\|\boldsymbol{X}\|_{2,1} = \sum_j \|\boldsymbol{X}_j\|_2 = \sum_j \langle \boldsymbol{X}_j, \boldsymbol{X}_j \rangle = \sum_j \sqrt{\sum_i \boldsymbol{X}_{i,j}^2}$. Similar we can define the $\boldsymbol{A} - 1$ norm, which is the sum of the Mahalanobis norms of the individual columns: $\|\boldsymbol{X}\|_{\boldsymbol{A},1} = \sum_j \|\boldsymbol{X}_j\|_{\boldsymbol{A}} = \sum_j \sqrt{\boldsymbol{X}_j^T \boldsymbol{A} \boldsymbol{X}_j}$. We have further:

$$\|\boldsymbol{A}^{1/2} \boldsymbol{X}\|_{2,1} = \sum_j \|\boldsymbol{A}^{1/2} \boldsymbol{X}_j\|_2 = \sum_j \sqrt{\boldsymbol{X}_j^T \boldsymbol{A} \boldsymbol{X}_j} = \sum_j \|\boldsymbol{X}_j\|_{\boldsymbol{A}} = \|\boldsymbol{X}\|_{\boldsymbol{A},1} \ .$$

We will also use the $2 - \infty$ norm of a matrix $\|\boldsymbol{X}\|_{2,\infty} = \max_j \|\boldsymbol{X}_j\|_2$

B USEFUL LEMMAS

Throughout the proof we will need at different locations the elliptical potential lemma:

Lemma B.1 (Lemma 19.4 in (Lattimore and Szepesvári, 2020)). *Let $\boldsymbol{V}_0 \in \mathbb{R}^{d,d}$ positive definitive and $\phi_1, \ldots, \phi_h \in \mathbb{R}^d$ a sequence of vectors with $\|\phi_t\|_2^2 \leq C_\phi^2 \ \forall t \in [h]$, $\boldsymbol{V}_h = V + \sum_{t \leq h} \phi_t^T \phi_t$. Then:*

$$\sum_{t=1}^h \min(1, \|\phi_t\|_{V_t^{-1}}^2) \leq 2 \log \left(\frac{\det \boldsymbol{V}_h}{\det \boldsymbol{V}_0} \right) \leq 2d \log \left(\frac{tr \boldsymbol{V}_0 + hL^2}{d \det \boldsymbol{V}_0^{1/d}} \right) \tag{17}$$

Recalling our feature regularity assumption $\|\phi(s,a)\|_2^2 \leq C_\phi$ we get:

$$2\log\left(\frac{\det \boldsymbol{V}_h}{\det \boldsymbol{V}_0}\right) \leq 2d\log\left(\frac{\lambda d + hC_\phi}{d\lambda}\right) = 2d\log\left(1 + \frac{HC_\phi}{\lambda d}\right) . \tag{18}$$

Note that (Yang and Wang, 2019)(lemma 10) would get here by virtue of choosing $\|\phi(s,a)\|_2^2 \leq dC_\phi$ and $\lambda = 1$:

$$2\log\left(\frac{\det \boldsymbol{V}_h}{\det \boldsymbol{V}_0}\right) \leq 2d\log\left(\frac{\lambda d + hdC_\phi}{d\lambda}\right) = 2d\log\left(1 + hC_\phi\right) . \tag{19}$$

Lemma B.2. *If* $\mathbf{B} \succeq \mathbf{C} \succ \mathbf{0}$ *be* $d \times d$ *dimensional matrices then,*

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} \leq \frac{\det(\mathbf{B})}{\det(\mathbf{C})}.$$

Proof. Given any $\mathbf{y} \in \mathbb{R}^d$ let $\mathbf{x} = \mathbf{C}^{-1/2}\mathbf{y}$. Then

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} = \sup_{\mathbf{y}\neq 0} \frac{\mathbf{y}^\top \mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\mathbf{y}}{\|\mathbf{y}\|_2^2} = \left\|\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\right\|_{op}$$

by the definition of the operator norm. Recall that by assumption $\mathbf{B} - \mathbf{C} \succeq 0$ therefore $\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2} - \mathbf{I} \succeq 0$, and hence all the eigenvalues of $\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}$ are at least 1. Thus

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} \leq \left\|\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\right\|_{op} \leq \det(\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}) = \frac{\det(\mathbf{B})}{\det(\mathbf{C})},$$

where the last equality follows since $\frac{\det(\mathbf{B})}{\det(\mathbf{C})} = \det(\mathbf{C}^{-1/2})\det(\mathbf{B})\det(\mathbf{C}^{-1/2}) = \det(\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2})$. This completes the proof. $\square$

Recall that $V_{n,h} = V_n + \sum_{h'<h} \phi_{n,h'}\phi_{n,h'}^\top$. We are now ready to proof the doubling log-determinant lemma.

Lemma 3.1 (Alternative Log Determinant Lemma)**.** *The following inequality holds,*

$$\sum_{n=1}^{N}\sum_{h=1}^{H} \|\phi_{n,h}\|_{V_n^{-1}} \leq \sum_{n=1}^{N}\sum_{h=1}^{H} 2\|\phi_{n,h}\|_{V_{n,h}^{-1}} + \frac{C_\phi}{\lambda}d\log\left(D\right).$$

Proof of Lemma 3.1. Define $e_{n,h} = \mathbf{1}\left(\|\phi_{n,h}\|_{(\boldsymbol{V}_{n,h}^\lambda)^{-1}} \leq 2\|\phi_{n,h}\|_{V^\lambda)_n^{-1}}\right)$. We define $e_{n,h}^c = 1 - e_{n,h}$.

$$\sum_{n=1}^{N}\sum_{h=1}^{H} \|\phi_{n,h}\|_{(\boldsymbol{V}_n^\lambda)^{-1}} \leq \sum_{n=1}^{N}\sum_{h=1}^{H} 2\|\phi_{n,h}\|_{(\boldsymbol{V}_{n,h}^\lambda)^{-1}} + \sum_{n=1}^{N}\sum_{h=1}^{H} e_{n,h}^c \frac{C_\phi}{\lambda}.$$

If $e_{n,h}^c = 1$ and as a consequence of Lemma B.2,

$$2 \leq \frac{\|\phi_{n,h}\|_{(\boldsymbol{V}^\lambda)_{n,h}^{-1}}}{\|\phi_{n,h}\|_{(\boldsymbol{V}^\lambda)_n^{-1}}} \leq \frac{\det(\boldsymbol{V}_{n,h})}{\det(\boldsymbol{V}_n)}.$$

And therefore it must be that,

$$2^{\sum_{n=1}^{N}\sum_{h=1}^{H} e_{n,h}^c} \leq \frac{\det(\boldsymbol{V}_N)}{\det(\lambda\mathbf{I})}$$

Furthermore, as a consequence of Lemma B.1,

$$\log\left(\frac{\det(\boldsymbol{V}_N)}{\det(\lambda\mathbf{I})}\right) \leq d\log\left(1 + \frac{nHC_\phi^2}{\lambda d}\right).$$

We therefore conclude that,

$$\sum_{n=1}^{N}\sum_{h=1}^{H} e_{n,h}^c \leq d\log\left(1 + \frac{HnC_\phi^2}{\lambda d}\right).$$

The result follows. □

Intuitively this means that the self normalised norm of the sum of feature matrices grows significantly slower than H. Despite using an inverse norm on the left hand side defined via the V_n matrices and lacking all the inter-episode outer products, we can show the sum of these inverse norms can be upper bounded by the a constant multiple of the sum of the inverse norms that defined by the inter-episode $V_{n,h}$ matrices plus a term of the form $O(d\log(n))$.

C   ANALYSING THE REGRET OF (BIASED) UC-MATRIX RL

The analysis of UC-MatrixRL works in two steps. First we show that the optimal transition core $\boldsymbol{M}^*$ is with high probablility contained in an ellipsoid around the current estimate $\boldsymbol{M}_n$. Subsequently we proof the regret for the case that $\boldsymbol{M}^*$ is with high probability in the constructed confidence ellipsoid.

We give the proof for the general biased case. The unbiased case follows immediately by choosing the zero matrix as bias, $W = 0$.

C.1   Constructing Confidence Sets

Theorem 4.1 (Confidence Ellipsoid BUC-MatrixRL). *Let $\delta > 0$. BUC-MatrixRL with bias matrix $\boldsymbol{W}$, produces under the assumptions 3.1 with probability at least $1 - \delta$ for all $t > 0$ estimates $\hat{\boldsymbol{M}}_n$, of the core matrix, that satisfy:*

$$\begin{aligned}
\|\hat{\boldsymbol{M}}_n - \boldsymbol{M}^*\|_F \leq &C'_\psi\sqrt{2d'\log\left(\frac{1}{\delta}\right) + d'd\log(D)} \\
&+ \sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}^*\|_F =: \beta_n^{\boldsymbol{W}}(\delta).
\end{aligned} \tag{10}$$

Thus for all times $t > 1$ and recalling the regularity of the feature map ϕ we have with probability at least $1 - \delta$, that the true transition core $\boldsymbol{M}^*$ is contained in the ellipsoid $C_n(\delta) = C_{\beta_n^{\boldsymbol{W}}(\delta)}^{\boldsymbol{W}}(\hat{\boldsymbol{M}}_n) = \{\boldsymbol{M} \in \mathbb{R}^{d\times d'} | \|\hat{\boldsymbol{M}}_n - \boldsymbol{M}\|_F \leq \beta_n^{\boldsymbol{W}}(\delta)\}$ of radius $\beta_n^{\boldsymbol{W}}(\delta)$ and centroid $\hat{\boldsymbol{M}}_n$. Clearly it holds $\beta_n^{\boldsymbol{W}}(\delta) \leq \beta_n^{\boldsymbol{0}}(\delta)$, which combined with item 4 of assumption 3.1 yields the radius $\beta_n^{\boldsymbol{0}}(\delta)$:

$$C'_\psi\sqrt{2d'\log\left(\frac{1}{\delta}\right) + d'd\log(D)} + \sqrt{\lambda}dC_m \tag{11}$$

Proof. The proof leverages the confidence ellipsoid from theorem 2 in (Abbasi-Yadkori et al., 2011).

We start with the solution of biased UC-Matrix RL and write out the regression target to illustrate the impact of the noise:

$$\boldsymbol{M}_n = \left(\boldsymbol{V}_n^\lambda\right)^{-1}\left[\sum_{n' < n, h \leq H} \phi_{n,h}\left(\psi_{n,h}K_\psi^{-1} - \phi_{n,h}^T W\right)\right] + \boldsymbol{W} \tag{20}$$

$$= \left(\boldsymbol{V}_n^\lambda\right)^{-1}\left[\sum_{n' < n, h \leq H} \phi_{n,h}\left(\phi_{n,h}\boldsymbol{M}^* + \eta_{n,h} - \phi_{n,h}^T W\right)\right] + \boldsymbol{W}. \tag{21}$$

We denote by $\eta_{n,h} = K_\psi^{-1}\psi_{n,h} - (\boldsymbol{M}^*)^T\phi_{n,h} \in R^{d'}$ the noise vector. Using assumption 3.1, 2., we have:

$$\left\|(\boldsymbol{M}^*)^T\phi_{n,h}\right\|_2 = \left\|\mathbb{E}\left[K_\psi^{-1}\psi_{n,h}|\mathcal{F}_{n,h}\right]\right\|_2 \leq \mathbb{E}\left[\|K_\psi^{-1}\psi_{n,h}\|_2|\mathcal{F}_{n,h}\right] \leq C'_\psi. \tag{22}$$

As a result we have $\mathbb{E}\left[\eta_{n,h} = 0\right]$ and $\|\eta_{n,h}\|_2^2 \leq 2C'_\psi$, thus our noise is $2C'_\psi$ subgaussian. Using the identity $\sum_{n' < n, h \leq H} \phi_{n,h}\phi_{n,h}^T = (V_n^\lambda) - \lambda I$ on the terms involving M^* and W we can write:

$$M_n - M^* = (V_n^\lambda)^{-1} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} + \lambda (V_n^\lambda)^{-1}(W - M^*) \tag{23}$$

Using Cauchy-Schwarz we have for any $X \in R^{d,d'}$:

$$\langle X, M_n - M^* \rangle_F = \left\langle X, (V_n^\lambda)^{-1} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} \right\rangle_F + \lambda \left\langle X, (V_n^\lambda)^{-1}(W - M^*) \right\rangle_F \tag{24}$$

$$= \left\langle (V_n^\lambda)^{-1/2}X, (V_n^\lambda)^{-1/2} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} \right\rangle_F + \lambda \left\langle (V_n^\lambda)^{-1/2}X, (V_n^\lambda)^{-1/2}(W - M^*) \right\rangle_F \tag{25}$$

$$\leq \left\| (V_n^\lambda)^{-1/2}X \right\|_F \left(\left\| (V_n^\lambda)^{-1/2} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} \right\|_F + \lambda \|(V_n^\lambda)^{-1/2}(W - M^*)\|_F \right). \tag{26}$$

Inserting the choice $X = (V_n^\lambda)(M_n - M^*)$, using the symmetry of (V_n^λ) to split it and observing $\|(V_n^\lambda)^{-1/2}A\|_F^2 \leq 1/\lambda_{min}(V_n^\lambda)\|A\|_F^2 \leq 1/\lambda\|A\|_F^2$ yields:

$$\|(V_n^\lambda)^{1/2}(M_n - M^*)\|_F^2 \leq \|(V_n^\lambda)^{1/2}(M_n - M^*)\|_F \left(\left\| (V_n^\lambda)^{-1/2} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} \right\|_F + \sqrt{\lambda}\|(W - M^*)\|_F \right). \tag{27}$$

Division by $\|(V_n^\lambda)^{1/2}(M_n - M^*)\|_F$ yields:

$$\|M_n - M^*\|_F \leq \left(\left\| (V_n^\lambda)^{-1/2} \sum_{n' < n, h \leq H} \phi_{n,h}\eta_{n,h} \right\|_F + \sqrt{\lambda}\|(W - M^*)\|_F \right). \tag{28}$$

We can write the Frobenius norm of a matrix as the root of the squared two norms of the columns of the matrix, thus $\|A\|_F = \sqrt{\sum_j \|(A)_j\|_2^2}$. Note that each $\phi_{n,h}(\eta_{n,h})_j$ term is a vector.

$$\|M_n - M^*\|_F \leq \left(\sqrt{\sum_{j=1}^{d'} \left\| \sum_{n' < n, h \leq H} \phi_{n,h}(\eta_{n,h})_j \right\|_{(V_n^\lambda)^{-1}}^2} + \sqrt{\lambda}\|(W - M^*)\|_F \right). \tag{29}$$

We apply the self-normalising bound for vector-values martingales from theorem 1 in (Abbasi-Yadkori et al., 2011)/ theorem 20.4 (Lattimore and Szepesvári, 2020) to the noise term and obtain:

$$\|M_n - M^*\|_F \leq \left(\sqrt{d'2(C'_\psi)^2 \log \frac{\det((V_n^\lambda))^{1/2}}{\delta \det(V_0)^{1/2}}} + \sqrt{\lambda}\|(W - M^*)\|_F \right). \tag{30}$$

Invoking the elliptical potential (lemma B.1, equation 18) on this yields:

$$\|M_n - M^*\|_F \leq C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log\left(1 + \frac{nHC_\phi}{d\lambda}\right)} + \sqrt{\lambda}\|W - M^*)\|_F. \tag{31}$$

Here we used the assumption $\|\Phi(s,a)\|_2^2 \leq C_\phi$. Using the dC_ϕ constraint from Yang and Wang (2019) would have given an extra in the nominator which would have cancelled with the denominator.

$\square$

C.2   Doing RL using the estimated matrix ball

We start by showing that the estimation error is along the directions of the exploration.

Lemma C.1 (similar to lemma 5 in Yang and Wang (2019)). *For any $M \in B_n$ we have*

$$\|\phi(s,a)^T(M - \boldsymbol{M}_n)\|_1 \le \beta_n \sqrt{\phi(s,a)^T((\boldsymbol{V}_n^\lambda))^{-1}\phi(s,a)} \tag{32}$$

Proof.

$$\|\phi(s,a)^T(M - \boldsymbol{M}_n)\|_1 = \|\phi(s,a)^T(\boldsymbol{V}_n^\lambda)^{-1/2}(\boldsymbol{V}_n^\lambda)^{1/2}(M - \boldsymbol{M}_n)\|_1 \tag{33}$$

$$\le \|\phi(s,a)^T(\boldsymbol{V}_n^\lambda)^{-1/2}\|_2\|(\boldsymbol{V}_n^\lambda)^{1/2}(M - \boldsymbol{M}_n)\|_{2,1} \tag{34}$$

$$= \|\phi(s,a)^T(\boldsymbol{V}_n^\lambda)^{-1/2}\|_2\|(M - \boldsymbol{M}_n)\|_{(\boldsymbol{V}_n^\lambda),1} \tag{35}$$

$$\le \beta_n\|\phi(s,a)^T(\boldsymbol{V}_n^\lambda)^{-1/2}\|_2 \tag{36}$$

□

While Abbasi-Yadkori et al. (2011) directly decompose the regret, in case of MDPs with linear transition modules we show first that the value iteration error per step is not too large which allows us to subsequently bound the regret. For notational purposes it wil be convinient to define:

$$w_{n,h} = \sqrt{\phi(s_{n,h}, a_{n,h})^T((\boldsymbol{V}_n^\lambda))^{-1}\phi(s_{n,h}, a_{n,h})} = \sqrt{\phi_{n,h}^T((\boldsymbol{V}_n^\lambda))^{-1}\phi_{n,h}}\ . \tag{37}$$

Lemma C.2 (Like Lemma 6 in Yang and Wang (2019)).

$$Q_{n,h}(s_{n,h}, a_{n,h}) - [r(s_{n,h}, a_{n,h}) + P(\cdot|s_{n,h}, a_{n,h})V_{n,h+1}] \le 2C_\psi H \beta_n w_{n,h} \tag{38}$$

Proof. Similar to Yang and Wang (2019) and using the updated bound in lemma C.1 □

Using this we refine lemma 7 of Yang and Wang (2019) as follows:

Lemma C.3 (lemma 7 in Yang and Wang (2019)). *Assumptions hold, $1 \le \beta_1 \le \cdots \le \beta_N$, then we can bound the regret at terminal time $T = NH$:*

$$Regret(T) \le 2C_\psi H \beta_n \mathbb{E}\left[\sum_{n=1}^N \sum_{h=1}^H \sqrt{\min(1, w_{n,h}^2)}\right] + \sum_{n=1}^N H\mathbb{P}[E_n = 0] \tag{39}$$

The only term that remains to be bounded is:

$$\sum_{n=1}^N \sum_{h=1}^H \sqrt{\min(1, w_{n,h}^2)} \le \sqrt{HN \sum_{n=1}^N \sum_{h=1}^H \min(1, w_{n,h}^2)} \tag{40}$$

Looking at the structure of $\sum_{n=1}^N \sum_{h=1}^H \min(1, w_{n,h}^2)$ it is tempting to reapply the classic elliptical potential lemma 41. However, this uses at step n, h the $V_{n,h}$ induced Malahonobis norm. We shall however have a fixed Malahonobis norm $\|.\|_{V_n}$ throughout episode n. We derive a related log determinant lemma, which is quite similar to lemma B.1, however we need to pay a factor H.

Lemma C.4 (Analogous to lemma 8 in (Yang and Wang, 2019)).

$$\sum_{n=1}^N \sum_{h=1}^H \min(1, w_{n,h}^2) \le 2H \ln \frac{\det(V_{N+1})}{\det(V_0)} \le 2Hd \ln \frac{NHC_\phi + \lambda}{\lambda} \tag{41}$$

Proof. Note that for any $u \geq 0$ it holds $\min(1, u) \leq 2 \ln(1 + u)$. Thus we get for the l.h.s. of 41:

$$\sum_{n=1}^{N} \sum_{h=1}^{H} \min(1, w_{n,h}^2) \leq 2 \sum_{n=1}^{N} \sum_{h=1}^{H} \ln(1 + w_{n,h}^2) . \tag{42}$$

We are now going to bound the r.h.s. by a log determinant ratio. Recalling the structure of $(\boldsymbol{V}_n^\lambda)$ we have:

$$V_{n+1} = (\boldsymbol{V}_n^\lambda) + \sum_{h=1}^{H} \phi_{n,h} \phi_{n,h}^T . \tag{43}$$

The determinant is a multiplicative map, thus we can decompose $\det(V_{n+1})$ as:

$$\det(V_{n+1}) = \det(V_n) \times \det(I + (\boldsymbol{V}_n^\lambda)^{-1/2} \sum_{h=1}^{H} \phi_{n,h} \phi_{n,h}^T (\boldsymbol{V}_n^\lambda)^{-1/2}) . \tag{44}$$

Thus we get:

$$\det(V_{n+1}) \geq \det((\boldsymbol{V}_n^\lambda)) \prod_{h=1}^{H} (1 + w_{n,h}^2)^{1/H} \geq \det(V_{n-1}) \dots \tag{45}$$

$$\geq \det(V_0) \prod_{n=1}^{N} \prod_{h=1}^{H} (1 + w_{n,h}^2)^{1/H} . \tag{46}$$

Putting things together and recalling eq. 18 we get:

$$\sum_{n=1}^{N} \sum_{h=1}^{H} \ln(1 + w_{n,h}^2) \leq 2H \ln \frac{\det(V_{N+1})}{\det(V_0)} \leq 2Hd \ln \left(1 + \frac{NHC_\phi}{\lambda d}\right) . \tag{47}$$

$\square$

It remains to combine this with the confidence ellipsoid we constructed in theorem 4.1. We begin by restating the regret bound.

Theorem 4.2 (Regret BUC-Matrix RL). *Under the assumptions of theorem 4.1 after $T = NH$ steps, choosing the ellipsoid radius $\beta_n^{\boldsymbol{W}}(\delta)$ as in 4.1 BUC-MatrixRL abides with probability at least $1 - 1/(NH)$ the following bound on the regret:*

$$R_T(\boldsymbol{M}^*) \leq \left(C_\psi' \sqrt{d'd \log(TD)} + \sqrt{\lambda} \|\boldsymbol{W} - \boldsymbol{M}^*\|_F\right)$$

$$2C_\psi H \sqrt{C_{\phi,\lambda} Td \ln(D)}$$

Proof. By theorem 4.1 we choose:

$$\beta_n = C_\psi' \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log(D)} + \sqrt{\lambda} \|W - \boldsymbol{M}^*\|_F . \tag{48}$$

Then the bad event, thus the optimal laying being outside the confidence ellipsoid, occurs only with small probability:

$$\mathbb{P}\left[\forall n \leq N | E_n = 1\right] \geq 1 - \delta \tag{49}$$

Thus we get by lemma C.3 when using lemma C.4 :

$$Regret(T, \mathcal{M}) \leq 2C_\psi H \beta_N \sqrt{2THd \ln\left(1 + \frac{NHC_\phi}{\lambda d}\right)} \tag{50}$$

Using lemma 3.1 instead of lemma C.4 yields:

$$Regret(T, \mathcal{M}) \leq 2C_\psi H \beta_N \sqrt{C_{\phi,\lambda} T d \ln\left(1 + \frac{NHC_\phi}{\lambda d}\right)} \tag{51}$$

In the remainder of the paper we use the regret bound obtained by using lemma 3.1. $\qquad\square$

C.3 Meta Transfer Regret

Theorem 4.3 (Meta Transfer Regret BUC-MatrixRL)**.** *Under the assumptions of theorem 4.2 we have with probability at least $1 - 1/(NH)$ for a task distribution $\mathcal{T}$ the following* Mtr *after T steps per task (where we absorb constant factors into C):*

$$\begin{aligned} \mathrm{Mtr}_T(\mathcal{T}) \leq &CC_\psi HC'_\psi d\sqrt{d'TC_{\phi,\lambda}\log(TD)\ln(D)} \\ &+ CC_\psi H\mathrm{Mad}_{\boldsymbol{W}}\sqrt{\lambda TC_{\phi,\lambda}d\ln(D)} \\ \leq &CC_\psi HC'_\psi d\sqrt{d'TC_{\phi,\lambda}\log(TD)\ln(D)} \\ &+ CC_\psi H\sqrt{\mathrm{Var}_{\boldsymbol{W}}\lambda TC_{\phi,\lambda}d\ln(D)} \end{aligned}$$

Proof. The first inequality is a result of applying theorem 4.2 with task distribution $\mathcal{T}$ and recalling the definition of transfer regret in equation 3. Jensen's inequality yields the second inequality. $\qquad\square$

D PROOF OF THE TRANSFER REGRET FOR THE LOW BIAS ESTIMATOR

This analysis is inspired form the analysis of the low bias case in (Cella et al., 2020). We start by stating the following vectorial version of Bennett's inequality which we shall need to bound the estimation error.

Lemma D.1 (Vectorial Version of Bennett's inequality; lemma 2 (Smale and Zhou, 2007)/lemma 3 (Cella et al., 2020))**.** *Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_G$ be G independent random vectors in $\mathbb{R}^d$ drawn from a joint distribution $\mathcal{T}$ with mean $\bar{n}$ and variance σ_m. Assume a bounded 2-norm at all stages: $\|\boldsymbol{m}_g\|_2 \leq C_M \; \forall g \in [G]$. For any $\delta \in (0,1)$ it holds with confidence $1 - \delta$:*

$$\frac{1}{G}\sum_{g=1}^G \|\boldsymbol{m}_g - \bar{\boldsymbol{m}}\|_2 \leq \frac{2\log(2/\delta)C_M}{G} + \sqrt{\frac{2\log(2/\delta)\sigma_M}{G}} \tag{52}$$

Using the interpretation of the Frobenius norm of a matrix as the 2-norm of a flattened version of the matrix we obtain the following corollary:

Corollary D.1.1 (Frobenius Version of Bennett's inequality)**.** *Let $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_G$ be G independent random matrices in $\mathbb{R}^{d\times d'}$ sampling from a joint distribution $\mathcal{T}$ with mean $\bar{M}$ and variance σ_M. Assume a bounded Frobenius norm at all stages: $\|\boldsymbol{M}_g\|_F \leq C_M \; \forall g \in [G]$. For any $\delta \in (0,1)$ it holds with confidence $1 - \delta$:*

$$\frac{1}{G}\sum_{g=1}^G |\boldsymbol{M}_g - \bar{\boldsymbol{M}}|_F \leq \frac{2\log(2/\delta)C_M}{G} + \sqrt{\frac{2\log(2/\delta)\sigma_M}{G}} \tag{53}$$

Recalling the definition of the empirical mean task $\bar{\boldsymbol{W}}_{G,t} = \frac{1}{NT+t}\left(\sum_{g=1}^G TM_g + tM_{G+1}\right)$ and the estimation error relative to the true mean transition core $\bar{\boldsymbol{M}}$ as $H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}}) = \|\bar{\boldsymbol{M}} - \bar{\boldsymbol{W}}_{G,t}\|_F$ we get from corollary D.1.1

$$\lim_{G\to\infty} H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}}) = 0 \,. \tag{54}$$

We deduce that the observation error is dominated by the variance term $\max_{g\in[G]} \frac{\beta_{g,T}(1/NH)}{\lambda_{min}^{1/2}(V_{g,T}^\lambda)}$. From standard linear regression results Lai et al. (1982) we know that $\lambda_{min}(V_{g,T}) \geq \log T$. By construction of the V_j^λ we have as

such $\lambda_{min}(V_{g,T}^\lambda) \geq \lambda + \lambda_{min}(V_{g,T}) \geq \lambda + \log T$. We see that the smaller T the larger the impact of the size of λ. Recall further our regularisation strength schedule: $\lambda = \frac{1}{TVar_{\hat{W}}}$. We see that if the variance of our estimator is large, the resulting λ is small. This leads to smaller $\lambda_{min}(V_{g,T})$, which yields to a higher variance, thus we are left with a circle of potentially self reinforcing variance. This is particularly problematic for task distributions $\mathscr{M}$ of high variance.

Theorem 5.1. *BUC-MatrixRL incurs after T interactions in G previous tasks, using the bias estimator $\hat{W}_{G,n,h}$ as in equation 13, $\lambda = \frac{1}{TVar_{\hat{W}_{G,n,h}}}$ and under assumptions 3.1, with probability at least $1 - 1/(NH)$ at most the following meta transfer regret:*

$$\mathrm{Mtr}_T(\mathcal{M}_{G+1}) \leq CC_\psi H d C'_\psi$$
$$\sqrt{C_{\phi,\lambda} d'T \log\left(T + \frac{T^3 C_\phi \left(Var_{\bar{M}} + \epsilon_{G,T}(\mathscr{M})\right)}{d}\right)}$$

The mean estimation error can be upper bounded as:

$$\sqrt{\epsilon_{G,T}(\mathscr{M})} \leq H_{\mathscr{M}}(G+1, \bar{M}) + \max_{g \in [G]} \frac{\beta_{g,T}^{\mathbf{0}}(1/NH)}{\lambda_{min}^{1/2}(V_{g,T}^\lambda)}$$

Proof. Recall the upper bound on the meta transfer regret from theorem: 4.3:

$$MTR_T(\mathcal{T}) \leq CC_\psi H C'_\psi d \sqrt{d'TC_{\phi,\lambda} \log(TD) \ln(D)} + CC_\psi H \sqrt{\mathrm{Var}_W \lambda TC_{\phi,\lambda} d \ln(D)} . \tag{55}$$

Inserting $\lambda = \frac{1}{TVar_W}$ and focussing for the ease of presentation on the dominant term we get:

$$MTR_T(\mathcal{T}) \leq CC_\psi H d C'_\psi \sqrt{C_{\phi,\lambda} d'T \log\left(T + \frac{T^3 C_\phi Var_W}{d}\right)} \tag{56}$$

Recall the choice of bias matrix estimator from equation 13

$$\hat{W}_{G,n,h} = \sum_{g=1}^{G-1} \frac{T}{Z} \hat{M}_{g,T} + \frac{nH+h}{Z} \hat{M}_{G,n,h}.$$

The regret upper bound for this choice of bias matrix amounts to:

$$MTR_T(\mathcal{T}) \leq CC_\psi H d C'_\psi \sqrt{C_{\phi,\lambda} d'T \log\left(T + \frac{T^3 C_\phi Var_{\hat{W}_{G,n,h}}}{d}\right)} \tag{57}$$

By the triangle inequality we have:

$$\sqrt{Var_{\hat{W}_{G,n,h}}} = \sqrt{\mathbb{E}_{M \sim \mathcal{T}}\left[\|M - \hat{W}_{G,n,h}\|_F\right]} \leq \sqrt{Var_{\bar{M}}} + \sqrt{\epsilon_{G,n,h}(\mathcal{T})} .$$

Plugging this into equation 57 yields the desired meta transfer regret bound. It remains to analyse the estimation

error $\epsilon_{G,T}(\mathcal{M})$:

$$
\begin{aligned}
\sqrt{\epsilon_{G,T}(\mathcal{M})} &= \|\bar{\boldsymbol{M}} - \hat{\boldsymbol{W}}^\lambda_{G,T}{}_F\| \\
&\leq \|\bar{\boldsymbol{M}} - \bar{\boldsymbol{W}}_{G,T}\|_F + \|\bar{\boldsymbol{W}}_{N,T} - \hat{\boldsymbol{W}}^\lambda_{G,T}\|_F \\
&= H_{\mathcal{M}}(G+1, \bar{\boldsymbol{M}}) + \|\bar{\boldsymbol{W}}_{N,T} - \hat{\boldsymbol{W}}^\lambda_{G,T}\|_F \\
&\leq H_{\mathcal{M}}(G+1, \bar{\boldsymbol{M}}) + \max_{g \in [G]} \|\boldsymbol{M}_g - \hat{\boldsymbol{M}}_{g,T}\|_F \\
&\leq H_{\mathcal{M}}(G+1, \bar{\boldsymbol{M}}) + \max_{g \in [G]} \frac{\|\left(\boldsymbol{M}_g - \hat{\boldsymbol{M}}_{g,T}\right)(\boldsymbol{V}^\lambda_{g,T})^{1/2}\|_F}{\lambda^{1/2}_{\min}(\boldsymbol{V}^\lambda_{g,T})} \\
&\leq H_{\mathcal{M}}(G+1, \bar{\boldsymbol{M}}) + \max_{g \in [G]} \frac{\beta^\lambda_j(1/T)}{\lambda^{1/2}_{\min}(\boldsymbol{V}^\lambda_{g,T})}.
\end{aligned}
$$

$\square$

E   PROOF OF THE TRANSFER REGRET FOR THE HIGH BIAS ESTIMATOR

The estimator as well as the way to prove its properties is adapted from section 5 in (Cella et al., 2020) to the matrix case. To this end we introduce firstly an additional variable:

$$
\bar{\boldsymbol{W}}'_{G,t+1} = \left(\tilde{\boldsymbol{V}}_{G,t}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{V}_{g,T}\boldsymbol{M}_g + \boldsymbol{V}_{N+1,t}\boldsymbol{M}_{G+1}\right).
$$

Using the triangle inequality we can now separate error causes. We get the *estimation error* $\hat{\boldsymbol{W}}^\lambda_{G,t+1} - \bar{\boldsymbol{W}}'_{G,t+1}$ which we handle in E.1 and the *estimation bias* $\bar{\boldsymbol{W}}'_{G,t+1} - \bar{\boldsymbol{W}}_{G,t+1}$ which we cover in lemma E.2.

Lemma E.1 (Estimation error of global ridge regression)**.** *The following rewriting holds:*

$$
\hat{\boldsymbol{W}}^\lambda_{G,t+1} - \bar{\boldsymbol{W}}'_{G,t+1} = \left(\tilde{\boldsymbol{V}}^\lambda_{G,t}\right)^{-1} \left(\sum_{g=1}^{G}\sum_{s=1}^{T} \phi_{g,s}\eta_{g,s} + \sum_{s=1}^{t} \phi_{G+1,s}\eta_{G+1,s}\right) - \lambda \left(\tilde{\boldsymbol{V}}^\lambda_{G,t}\right)^{-1} \bar{\boldsymbol{W}}'_{G,t+1}
$$

Proof. Follows immediately from (Cella et al., 2020). $\square$

Lemma E.2 (Estimation bias of global ridge regression)**.** *From Section D, we use:*

$$
\bar{\boldsymbol{W}}_{G,t+1} = \frac{1}{GT+t} \left(\sum_{g=1}^{G} T\boldsymbol{M}_g + t\boldsymbol{M}_{G+1}\right).
$$

Differently from $\bar{\boldsymbol{W}}'_{G,t}$ this definition is a weighted average of the transition cores of the G previously encountered tasks. Thus, we have:

$$
\begin{aligned}
\|\bar{\boldsymbol{M}} - \bar{\boldsymbol{W}}'_{G,t}\|_F &\leq \frac{1}{NT+t} \sum_{g=1}^{G} \left[\|\bar{\boldsymbol{M}} - \bar{\boldsymbol{W}}_{G,t}\|_F + (GT+t)\|\bar{\boldsymbol{W}}_{G,t} - \bar{\boldsymbol{W}}'_{G,t}\|_F\right] \\
&= H_{\mathcal{T}}(G+1, \bar{\boldsymbol{M}}) + \|\bar{\boldsymbol{W}}_{G,t} - \bar{\boldsymbol{W}}'_{G,t}\|_F
\end{aligned}
$$

where we have denoted with $H_{\mathcal{T}}(G+1, \bar{\boldsymbol{M}})$ according to what we have done in subsection 5.1. We can now focus

on the term $\|\bar{\boldsymbol{W}}'_{G,t} - \bar{\boldsymbol{W}}_{G,t}\|_F$ which can be equivalently rewritten as:

$$
\begin{aligned}
\|\bar{\boldsymbol{W}}'_{G,t+1} - \bar{\boldsymbol{W}}_{G,t+1}\|_F &= \left\| \left(\tilde{\boldsymbol{V}}_{G,t}\right)^{-1} \sum_{g=1}^{G} \left(\boldsymbol{V}_{g,T}\boldsymbol{M}_g + \boldsymbol{V}_{G+1,t}\boldsymbol{M}_{G+1}\right) - \bar{\boldsymbol{W}}_{G,t} \right\|_F \\
&\leq \sum_{g=1}^{G} \left|\tilde{\boldsymbol{V}}_{G,t}^{-1}\boldsymbol{V}_{g,T}\right| \|\boldsymbol{M}_g - \bar{\boldsymbol{W}}_{G,t}\|_F + \left|\tilde{\boldsymbol{V}}_{G,t}^{-1}\boldsymbol{V}_{G+1,t}\right| \|\boldsymbol{M}_{G+1} - \bar{\boldsymbol{W}}_{G,t}\|_F \\
&\leq \sum_{g=1}^{G} H_{\mathcal{T}}(G+1,\boldsymbol{M}_g)\left|\tilde{\boldsymbol{V}}_{G,t}^{-1}\boldsymbol{V}_{g,T}\right| + H_{\mathcal{T}}(G+1,\boldsymbol{M})\left|\tilde{\boldsymbol{V}}_{G,t}^{-1}\boldsymbol{V}_t\right| \\
&= \sum_{g=1}^{G} H_{\mathcal{T}}(G+1,\boldsymbol{M}_g)\sigma_{\max}\left(\boldsymbol{V}_{g,t}\tilde{\boldsymbol{V}}_{G,t}^{-1}\right) + H_{\mathcal{T}}(G+1,\boldsymbol{M}_{G+1})\sigma_{\max}\left(\boldsymbol{V}_{g,t}\tilde{\boldsymbol{V}}_{G,t}^{-1}\right) \\
&\leq (G+1)\max_{g=1,\dots,G+1}\left(H_{\mathcal{T}}(G+1,\boldsymbol{M}_j)\sigma_{\max}\left(\boldsymbol{V}_{g,t}\tilde{\boldsymbol{V}}_{G,t}^{-1}\right)\right) \\
&= (G+1)\max_{g=1,\dots,G+1}\tilde{H}(G+1,\boldsymbol{M}_g)
\end{aligned}
$$

We have used the fact that the matrix norm of a given matrix A induced by the Euclidean norm corresponds to the spectral norm, which is the largest singular value of the matrix $\sigma_{\max}(A)$.

Theorem 5.2. *BUC-MatrixRL incurs after T interactions in G previous tasks, using the bias estimator $\hat{\boldsymbol{W}}_{G,n,h}$ as in equation 14, $\lambda = \frac{1}{T \, Var_{\hat{\boldsymbol{W}}_{G,n,h}}}$ and assumptions 3.1, with probability at least $1 - 1/(NH)$ at most the following meta transfer regret:*

$$
\mathrm{Mtr}_T(\mathcal{M}_{G+1}) \leq CC_{\psi}HdC'_{\psi}
$$
$$
\sqrt{C_{\phi,\lambda}d'T\log\left(T + \frac{T^3 C_{\phi}\left(Var_{\bar{M}} + \epsilon_{G,T}(\mathcal{T})\right)}{d}\right)}
$$

Let $\nu_{\min} = \lambda_{\min}\left(\tilde{\boldsymbol{V}}_{G,n,h}\right)$ be the minimal singular value of the global feature matrix. Then the mean estimation error can be upper bounded as:

$$
\sqrt{\epsilon_{G,T}(\mathcal{M})} \leq H_{\mathcal{M}}(G+1,\bar{\boldsymbol{M}})
$$
$$
+ \underbrace{\frac{dC_M}{\lambda + \nu_{\min}} + C'_{\psi}\sqrt{\frac{2}{\lambda + \nu_{\min}}\log\left(NH + \frac{GN^2H^2C_{\phi}}{\lambda d}\right)}}_{\frac{\beta^{\mathbf{0}}(1/(GNH))}{\lambda + \nu_{\min}}}
$$
$$
+ 2(G+1)\max_{g\in[G+1]}\tilde{H}(G+1,\boldsymbol{M}_g) .
$$

$\tilde{H}(G+1,\boldsymbol{M}_g)$ *is a weighted version of the estimation error $H_{\mathcal{M}}(G+1,\bar{\boldsymbol{M}})$ towards the current transition core $\boldsymbol{M}_g$:*

$$
\tilde{H}(G,\boldsymbol{M}_g) = H_{\mathcal{M}}(g,\boldsymbol{M}_g)\sigma_{\max}\left(\boldsymbol{V}_{g,T}\tilde{\boldsymbol{V}}_{G,N,H}^{-1}\right) , \tag{15}
$$

where $\sigma_{\max}\left(\boldsymbol{V}_{g,T}\tilde{\boldsymbol{V}}_{G,N,H}^{-1}\right)$ quantifies the misalignment of task g to the other tasks observed so far.

Proof. The bound on the meta transfer regret is obtained similar to the first part of the proof of theorem 5.1 by using the triangle inequality:

$$
\sqrt{Var_{\hat{\boldsymbol{W}}_{G,n,h}}} = \sqrt{\mathbb{E}_{\boldsymbol{M}\sim\mathcal{T}}\left[\|\boldsymbol{M} - \hat{\boldsymbol{W}}_{G,T}^{\lambda}\|_F^2\right]} \leq \sqrt{Var_{\bar{M}}} + \sqrt{\epsilon_{G,t}(\mathcal{T})} .
$$

It thus remains to bound the estimation error. According to lemma E.2 we rewrite:

$$\sqrt{\epsilon_{G,t}(\mathcal{T})} \leq H_{\mathcal{T}}(G+1, \bar{M}) + (G+1) \max_{g=1,\ldots,G+1} \tilde{H}(G+1,j) + \|\bar{W}'_{G,T} - \hat{W}^{\lambda}_{G,T}\|_F$$

Recalling the definition of the Frobenius Malahonobis norm for matrices (equation 16) it remains only to apply lemma E.1 which gives:

$$
\begin{aligned}
\|\bar{W}'_{G,T} - \hat{W}^{\lambda}_{G,T}\|_F &= \left\| \left(\tilde{V}^{\lambda}_{G,T}\right)^{-1} \left(\sum_{g=1}^{G}\sum_{s=1}^{T} \phi_{g,s}\eta_{g,s} + \sum_{s=1}^{T}\phi_s\eta_s \right) \right\|_F + \left\| \lambda \left(\tilde{V}^{\lambda}_{G,T}\right)^{-1} \bar{W}'_{G,T} \right\|_F \\
&\leq \left\| \sum_{g=1}^{G}\sum_{s=1}^{T}\phi_{g,s}\eta_{g,s} + \sum_{s=1}^{T}\phi_s\eta_s \right\|_{\left(\tilde{V}^{\lambda}_{G,T}\right)^{-2}} + \lambda\|\bar{W}'_{G,T}\|_{\left(\tilde{V}^{\lambda}_{G,T}\right)^{-2}} \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\tilde{V}^{\lambda}_{G,T})} \left\| \sum_{g=1}^{G}\sum_{s=1}^{T}\phi_{g,s}\eta_{g,s} + \sum_{s=1}^{T}\phi_s\eta_s \right\|_{\left(\tilde{V}^{\lambda}_{G,T}\right)^{-1}} + \frac{1}{\lambda_{\min}(\tilde{V}^{\lambda}_{G,T})}\|\bar{W}'_{G,T}\|_F \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\tilde{V}^{\lambda}_{G,T})} C'_{\psi} \sqrt{2\log\left(T + \frac{(GT+T)TL^2}{\lambda d}\right)} + \|\bar{W}'_{G,T} - \bar{W}_{G,T}\|_F + \frac{1}{\lambda_{\min}(\tilde{V}^{\lambda}_{G,T})}\|\bar{W}_{G,T}\|_F \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\tilde{V}^{\lambda}_{G,T})} C'_{\psi} \sqrt{2\log\left(T + \frac{(GT+T)TL^2}{\lambda d}\right)} + \|\bar{W}'_{G,T} - \bar{W}_{G,T}\|_F + \frac{S}{\lambda_{\min}(\tilde{V}^{\lambda}_{G,T})} \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\tilde{V}^{\lambda}_{G,T})} C'_{\psi} \sqrt{2\log\left(T + \frac{(GT+T)TL^2}{\lambda d}\right)} + (G+1)\max_{g=1,\ldots,G+1}\tilde{H}(G+1,j) \\
&\quad + \frac{S}{\lambda_{\min}(\tilde{V}^{\lambda}_{G,T})}
\end{aligned}
$$

In the last inequality we used again lemma E.2. We can now introduce $\nu_{\min} = \lambda_{\min}(\tilde{V}_{G,T})$ as eigenvalue of the global feature matrix. It follows as desired:

$$\sqrt{\epsilon_{G,T}(\mathcal{M})} \leq H_{\mathcal{M}}(G+1, \bar{M}) + 2(G+1)\max_{g\in[G+1]}\tilde{H}(G+1, M_g) \tag{58}$$

$$+ \frac{dC_M}{\lambda + \nu_{\min}} + C'_{\psi}\sqrt{\frac{2}{\lambda + \nu_{\min}}\log\left(GH + \frac{GN^2H^2C_{\phi}}{\lambda d}\right)}. \tag{59}$$

$\square$