Meta Learning MDPs with linear transition models

Robert Müller [1] **Aldo Pacchiano** [2] **Jack Parker-Holder** [3]

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 propose an algorithm that can meaningfully leverage learning in a set of sampled training tasks to quickly adapt to test tasks sampled from the same task distribution. We propose a biased version of the UC-MatrixRL algorithm (Yang and Wang, 2019). 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 study the effect of the bias on single task regret and expected regret over the task distribution. We prove that our algorithm provides significant improvements in the transfer regret for task distributions of low variance and high bias compared to learning the tasks in isolation. We outline and analyse two approaches to learn the bias.

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 previous tasks to solve future tasks quickly. More formally, the learner is given a training task distribution during training time and subsequently evaluated on a test task distribution. Since meta learning is a problem formulation it can be combined with supervised learning (Denevi et al., 2018; 2019; Khodak et al., 2019; Tripuraneni et al., 2021; Konobeev et al., 2021), bandits (Azar et al., 2013; Deshmukh et al., 2017; Cella et al., 2020) or reinforcement learning (RL). While the move to bandits adds the challenge of exploration, the move to RL introduces in addition temporal dependencies in the data. The meta learning protocol with a split in train and test tasks differs from the multitask setting with one distribution of tasks which was studied for RL (Brunskill and Li, 2013; Calandriello et al., 2015). Conditions under which zero-shot learning is possible were recently investigated by (Malik et al., 2021). Recently, there has been a growing body of empirical work in meta RL Zintgraf et al. (2020); Rakelly et al. (2019); Müller et al. (2020), addressing the question of how to algorithmically leverage shared structure. While this success is impressive, it remains largely open to study conditions under which and to what extent transfer of knowledge is possible.

An increasingly well understood setting for sequential decision-making is to assume linear structure in the MDP, for example, in the value function (Jin et al., 2020; Zhou et al., 2020; Cai et al., 2020) or in the transition matrix (Yang and Wang, 2019). To come closer to study what and when meta reinforcement learning is useful, we restrict ourselves in a first step to the subset of MDP with linear transition module. This restriction allows us to represent each MDP uniquely via its transition core matrix. The space of matrix norms equips us with a variety of natural task distances. We ask ourselves:

Under i) which characterisation of shared structure gives us meta RL ii) which improvement compared to independent single task learning and iii) how do we communicate the shared structure?

To quantify the gain of meta RL, we use the transfer regret as a performance metric for analysing a biased version of UC-MatrixRL that performs meta learning via learning a bias. 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. While the oracle setting is not realistic, we propose two practical estimators and prove the resulting transfer regret. Using a more careful log-determinant lemma allows to also tighten the analysis of UC-MatrixRL (Yang and Wang, 2019).

2. Biased UC-MatrixRL

2.1. Background

We consider the undiscounted episodic setting. A MDP is given as sixtupel: $\mathcal{M} = (S, A, r, P, H, \mu_0)$ with state-space S, action-space A, reward function $r : S \times A \to [0, 1]$,

*Equal contribution [1]Technical University of Munich [2]UC Berkeley [3]University of Oxford. Correspondence to: Robert Müller <2robert.mueller@gmail.com>.

Workshop on Reinforcement Learning Theory at ICML 2021,

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, an extension is straight forward. The learner has N episodes, each consisting of H steps to interact within the environment. Denote by $\mathcal{F}_{n,h}$ the filtration of fixed history up to stage (n, h), so $\mathcal{F}_{n,h} = \{s_{0,0}, a_{0,0}, \ldots, s_{n,h}, a_{n,h}\}$.

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 $V^{\pi}_{h+1}(s) = \mathbb{E}\left[\sum_{t=h}^{H} r(s_t, \pi_t(s_t))\right]$. The action value function is given as $Q^{\pi}_h(s, a) = r(s, a) + P(\cdot|s, a)V^{\pi}_{h+1}$. The goal of the learner is to find the optimal policy π^* which maximizes the expected sum of future rewards: $\pi^* = \arg\max_{\pi\in\Pi} V^{\pi}_h(s) \quad \forall s \in S, h \in [H]$. A learning algorithm K learns in an MDP $\mathcal{M}$ for N episodes, resulting in a total of $T = NH$ steps. 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 $(n + 1, 1)$. The regret of an algorithm after $T = NH$ steps in MDP $\mathcal{M}$ is defined as:

$$R(T, \mathcal{M}) = \mathop{\mathbb{E}}_{K} \sum_{n=1}^{N} \left[V^*(s_0) - \sum_{h=1}^{H} r(s_{n,h}, a_{n,h}) \right], \quad (1)$$

where the expectation is with respect to the path of state action pairs taken by the learner. Note that the algorithm determines in particular all actions a_t.

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: $\hat{\boldsymbol{M}}_n = \hat{\boldsymbol{W}}_0$, $(\boldsymbol{V}_0^{\lambda})^{-1} = \frac{1}{\lambda}\boldsymbol{I}$
for *episode* $n = 1, \ldots, N$*:* **do**
  Build optimistic Q-function using M_n, β_n **for** *step* $h = 1, \ldots, H$: **do**
    select greedy action $a_{n,h} = \arg\max_{a\in A} Q_{n,h}(s_{n,h}, a)$
    record next state $s_{n,h+1}$
  end
  update feature matrix: $\boldsymbol{V}_{n+1} \leftarrow \boldsymbol{V}_n + \sum_{h\le H} \phi_{n,h}\phi_{n,h}^T$
  recompute ellipsoid radius β_{n+1} like in equation 8
  possibly update bias estimate $\hat{\boldsymbol{W}}_{n+1}$
  recompute transition core estimate $\boldsymbol{M}_{n+1}$ using equation 6
end

For a family of MDP $\mathscr{M}$ the expected transfer regret is:

$$R(T, \mathscr{M}) = \mathop{\mathbb{E}}_{\mathcal{M}\sim\mathscr{M}} R(T, \mathcal{M}) \quad (2)$$

While we focus on the expected transfer regret an alternative approach could be to examine worst case transfer regret. We restrict our analysis to the set of MDP's that have a linear transition core.

Definition 2.1 (Linear transition core MDP)**.** *An MDP $\mathcal{M}$ has a linear transition core, if for each $(s_t, a_t) \in S \times A, s_{t+1} \in S$, 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 $\boldsymbol{M}^* \in \mathbb{R}^{d\times d'}$ s.t.:*

$$P(\tilde{s}|s, a) = \phi(s, a)^T \boldsymbol{M}^* \psi(\tilde{s}). \quad (3)$$

The matrix $\boldsymbol{M}^*$ is referred to as transition core.

In the space of MDP with linear transition model, we see that each MDP $\mathcal{M}$ is uniquely characterised by its transition core $\boldsymbol{M}$. Thus, we can characterise a family of linear transition MDP's $\mathscr{M}$ over the same state and action space and with similar features ϕ and ψ as the set of transition cores of its members. Let $\mathcal{T}$ be a distribution of transition cores $\boldsymbol{M}$ we have $\mathscr{M} = \{\mathcal{M}|\boldsymbol{M}(\mathcal{M}) \in \mathcal{T}\} = \{\boldsymbol{M}|\mathcal{M}(\boldsymbol{M}) \in \mathscr{M}\}$. Thus, we can write interchangeably $\mathbb{E}_{\mathcal{M}\sim\mathscr{M}}$ and $\mathbb{E}_{\boldsymbol{M}\sim\mathscr{T}}$.

2.2. Learning a model via regression

Upper Confidence-MatrixRL (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 $\hat{M}_n$. In this subsection, we describe a biased ridge regression setup that is a generalisation of the setup behind the original UC-MatrixRL.

The feature maps $\phi_{n,h} = \phi(s_{n,h}, a_{n,h}) \in \mathbb{R}^d$ and $\psi_{n,h} = \psi(s_{(n,h)+1}) \in R^{d'}$ are fixed and a priory given. Using the identity matrix $\boldsymbol{I} \in R^{d\times d}$, we define $\boldsymbol{K}_{\psi} = \sum_{\tilde{s}\in S} \psi(\tilde{s})\psi(\tilde{s})^T$ and $\boldsymbol{V}_n = \sum_{n'<n,h\le H} \phi_{n',h}\phi_{n',h}^T$. Letting $\boldsymbol{V}_n^{\lambda} = \lambda\boldsymbol{I} + \boldsymbol{V}_n$, in particular $\boldsymbol{V}_0^{\lambda} = \lambda\boldsymbol{I}$, we have.

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

Given a bias matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d'}$, we define $\boldsymbol{B} := \sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}\|_F$ and $\boldsymbol{B}^* := \sqrt{\lambda}\|\boldsymbol{W} - \boldsymbol{M}^*\|_F$. $\boldsymbol{M}$ can be estimated via the following ridge regression problem:

$$\arg\min_{\boldsymbol{M}} \sum_{n',h}^{n,H} \|\psi_{n',h}^T\boldsymbol{K}_{\psi}^{-1} - \phi_{n',h}^T\boldsymbol{M}\|_2^2 + \boldsymbol{B}^2 . \quad (5)$$

We call $\boldsymbol{W}$ the bias matrix or bias transition core. Using ridge regression and solving for $\boldsymbol{B}$ gives the solution:

$$\left(\boldsymbol{V}_n^{\lambda}\right)^{-1} \sum_{n',h}^{n,H} \phi_{n,h}\left(\psi_{n',h}^T\boldsymbol{K}_{\psi}^{-1} - \phi_{n',h}^T\boldsymbol{W}\right) + \boldsymbol{W} . \quad (6)$$

Here we define $\boldsymbol{M}_0 = \boldsymbol{W}$. When writing $\boldsymbol{M}_{n,h}$ we talk about the estimate after h steps in the n episode and abbreviate $\boldsymbol{M}_{n,H} = \boldsymbol{M}_n$. 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 $\hat{M}_n$. For episode n the optimistic Q-function is recursively constructed as:

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

where the value function is defined as $V_{n,h}(s) = \Pi_{h' \in [0,H]} [\max_a Q_{n,h'}(s,a)]$. 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. We require the following regularity conditions:

Assumption 2.1. *(Feature regularity) For positive constants $C_M, C_\phi, C_\psi, C'_\psi$ we have:*

1. $\|\phi(s,a)\|_2^2 \leq C_\phi \; \forall (s,a) \in S \times A$,
2. $\|\Psi K_\psi^{-1}\|_{2,\infty} \leq C'_\psi$
3. $\|\Psi^T v\|_2 \leq C_\psi \|v\|_\infty \; \forall v \in R^S$
4. $\|M^*\|_F^2 \leq C_M d$

These assumptions are almost the same as assumptions $2'$ in (Yang and Wang, 2019), only 1. differs in that they assume $\|\phi(s,a)\|_2^2 \leq C_\phi d$.

2.3. Regret bound of MatrixRL

Under assumption 2.1, with first item modified as described in the previous section, (Yang and Wang, 2019) proof for UC-MatrixRL a regret bound of: $O\left[C_\psi \sqrt{\|\boldsymbol{M}^*\|_F^2 + C'^2_\psi \ln(1 + \frac{NHC_\phi}{\lambda})}\right] dH^2\sqrt{T}$.

We reanalyse the MatrixRL algorithm using self normalising techniques from (Abbasi-Yadkori et al., 2011) to directly bound the noise term when constructing a matrix ball. 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 2.1. *The following inequality holds,*

$$\sum_{n,h}^{N,H} \|\phi_{n,h}\|_{(\boldsymbol{V}_n^\lambda)^{-1}} \leq \sum_{n,h}^{N,H} 2\|\phi_{n,h}\|_{(\boldsymbol{V}_{n,h}^\lambda)^{-1}} + \frac{L_\phi}{\lambda} d \log(D) \, .$$

This allows us to tighten the analysis by a factor of $\sqrt{H}$ compared to using lemma 8 in (Yang and Wang, 2019) at the expense of an increase from 2 to $C_{\phi,\lambda} := (4 + C_\phi/\lambda)$ in front of the log determinant term.

2.4. Construction of confidence balls

We will now study how different bias matrices influence the regret of transfer regret of biased UC-MatrixRL. Firstly, we construct modified confidence sets.

Theorem 2.2. *For any $\delta > 0$ we have under the assumption 2.1 with probability at least $1 - \delta$ for all $t > 0$ and bias matrix $\boldsymbol{W}$ that $\|\hat{\boldsymbol{M}}_n - \boldsymbol{M}^*\|_F$ is less or equal to:*

$$C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log(D)} + \boldsymbol{B}^* \, . \tag{7}$$

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 $C_n(\delta) = C_{\beta_n(\delta)}(\hat{\boldsymbol{M}}_n) = \{\boldsymbol{M} \in \mathbb{R}^{d \times d'} | \|\hat{\boldsymbol{M}}_n - \boldsymbol{M}\|_F \leq \beta_n(\delta)\}$ of radius $\beta_n(\delta)$ and centroid $\hat{\boldsymbol{M}}_n$, with:

$$\beta_n(\delta) = C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log(D)} + \boldsymbol{B}^* \, . \tag{8}$$

It is a common choice to select $\delta \leq 1/(NH)$.

We defer the proof to the appendix C.1.

2.5. Regret bound of BUC-MatrixRL

Theorem 2.3. *BUC-MatrixRL suffers under the assumptions of theorem 2.2 after $T = NH$ steps, choosing the ellipsoid radius $\beta_n(\delta)$ as in 2.2 and $\delta < 1/(NH)$, the following regret:*

$$R(T, \boldsymbol{M}^*) \leq 2C_\psi H \Big(C'_\psi \sqrt{d'd \log(NHD)} + \boldsymbol{B}^* \Big) \sqrt{C_{\phi,\lambda} T d \ln(D)} \, . \tag{9}$$

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 $0 \in \mathbb{R}^{d \times d'}$ recovers the standard UC-Matrix RL algorithm and whereas the regret for the true transition core $\boldsymbol{M}^*$, recalling the definition of D as $1 + \frac{nHC_\phi}{\lambda d}$, is 0 as $\lambda \to \infty$. Thus, the larger the regularisation strength λ, the smaller the suffered regret.

2.6. Transfer Regret bound of BUC-MatrixRL

The goal of meta learning is to generalise from the training tasks to all tasks of the task distribution at hand. We proceed to bound the transfer regret $R(R, \mathscr{M})$. We define the mean absolute distance $Mad_W = \mathbb{E}_{\boldsymbol{M} \sim \mathcal{T}} [\|\boldsymbol{M} - \boldsymbol{W}\|_F]$ and variance $Var_W = \mathbb{E}_{\boldsymbol{M} \sim \mathcal{T}} \left[\|\boldsymbol{M} - \boldsymbol{W}\|_F^2\right]$ of the task distribution relative to a bias matrix $\boldsymbol{W}$.

Theorem 2.4. *Under the assumptions of theorem 2.3 and $\lambda = 1/(NHVar_{\boldsymbol{W}})$ we have for a task distribution $\mathcal{T}$ the following transfer regret after T steps per task:*

$$\begin{aligned} R(T, \mathcal{T}) &\leq CC_\psi H C'_\psi d \sqrt{d'TC_{\phi,\lambda} \log(NHD) \ln(D)} \\ &\quad + CC_\psi H Mad_{\boldsymbol{W}} \sqrt{\lambda T H d \ln(D)} \\ &\leq \left(1 + C'_\psi \sqrt{d'dT \log(TDVar_{\boldsymbol{W}})}\right) CC_\psi H \\ &\quad \sqrt{C_{\phi,\lambda} d \ln(DVar_{\boldsymbol{W}})} \end{aligned}$$

The choice of mean transition core $\bar{\boldsymbol{M}}$ as bias matrix reduces the transfer regret to the independent learning setting, whenever $Var_0 < Var_{\bar{\boldsymbol{M}}}$. This is the case whenever the task distribution is not centred in the origin. If the task distribution is biased and has in addition low variance $Var_{\bar{M}} \approx 0$, thus we have $\log(1+0)$ under the second root, which sends the regret to zero. Generally, the oracle biased UC-MatrixRL improves against individual task learning whenever the variance of the task distribution is much lower than the second moment: $\mathrm{Var}_{\bar{\boldsymbol{M}}} = \mathbb{E}_{\boldsymbol{W}\sim\mathcal{T}} \|\boldsymbol{M} - \bar{\boldsymbol{M}}\|_F^2 \ll \mathbb{E}_{\boldsymbol{M}\sim\mathcal{T}} \|\boldsymbol{M} - \bar{\boldsymbol{M}}\|_F^2 = \mathrm{Var}_0$. So, the oracle transfer regret improves against the single task regret whenever the spread of the transition cores is small compared to the distance of the mean transition core to the origin.

3. Meta Learning with BUC-Matrix RL

So far, we assumed access to an oracle of the optimal transition core M^* and showed its usefulness in terms of incurred transfer regret. In any practical setting, we 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, a T steps per task. Each task g was created by sampling a transition core $\boldsymbol{M}_g \sim \mathcal{T}$. 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 biased UC-Matrix RL with a bias matrix containing the transferable knowledge from meta training. The meta training protocol is given in algorithm 2.

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: $\hat{M}_{0,0} = \hat{W}_0$, $V_0^{-1} = \frac{1}{\lambda} I$
for *train task* $\mathcal{M}_g \in \{\mathcal{M}_1, \ldots, \mathcal{M}_G\}$*:* **do**
  run algorithm 1 on MDP $\mathcal{M}_g$ with bias $\boldsymbol{W} = \boldsymbol{W}_{g-1}$
  possibly update bias estimate $\hat{\boldsymbol{W}}_g$
end
return estimated mean transition model $\boldsymbol{W}_G$

Subsequently, the learning algorithm, with the bias extracted from the training phase, is evaluated on its transfer regret on the test task distribution. In general, we are interested in the regime of only a few interactions per test task but many training tasks, so small T and large G.

Inspired by the bias estimators in (Cella et al., 2020) we outline two approaches to practically instantiate the bias estimation procedure. One works by aggregating final transition cores, the other by aggregating over all observed transitions on all tasks. 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 $\boldsymbol{V}_{g,t} = \boldsymbol{V}_{g,t//H,t\%H} = \sum_{p\leq t} \phi(s_{g,p})\phi(s_{g,p})^T = \sum_{p\leq t} \phi_{g,p}\phi_{g,p}^T$ and the radius of the confidence ellipsoid at state t as $\beta_{g,t}$. A crucial quantity is the mean estimation error of the bias $\hat{\boldsymbol{W}}_{g,n,h}$ at stage (n,h) of the $g+1$-st task with respect to the true mean transition core $\bar{\boldsymbol{M}}$:

$$\epsilon_{g,n,h}(\mathscr{M}) = \|\bar{\boldsymbol{M}} - \hat{\boldsymbol{W}}_{g,n,h}\|_F^2 \,. \tag{10}$$

As we face the tasks in a sequential manner there is an inherent estimation error, due to the fact that the learner has only samples from the task distribution. We define $\bar{W}_{G,t} = \frac{1}{NT+t}\left(\sum_{g=1}^{G} T\boldsymbol{M}_g + t\boldsymbol{M}_{G+1}\right)$ the mean transition core of the observed MDP and denote the estimation error relative to the true mean transition core $\bar{\boldsymbol{M}}$ as $H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}}) = \|\bar{\boldsymbol{M}} - \bar{\boldsymbol{W}}_{G,t}\|_F$.

3.1. 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 $\mathcal{M}_g$ is distilled in the respective final estimate of the transition core $\boldsymbol{M}_{g,T}$. Knowledge transfer between the tasks is achieved by aggregation of the individual transition cores:

$$\hat{\boldsymbol{W}}_{G,t} = \frac{1}{TG+t}\left(\sum_{g=1}^{G} T\hat{\boldsymbol{W}}_{g,T} + t\hat{\boldsymbol{W}}_{G+1,t}\right) \,. \tag{11}$$

Note that this approach is not updating the estimated bias matrix $\hat{\boldsymbol{W}}_g$ in algorithm 2 based on previous tasks. This choice of bias estimator results in the following bound on the transfer regret.

Theorem 3.1. *Biased UC-MatrixRL incurs after T interactions in G previous tasks, using the bias $\hat{\boldsymbol{W}}_{G,n,h}$ as in equation 11 and assumptions as in theorem 2.4, the transfer regret $R(T, \mathcal{M}_{G+1})$ that is upper bounded by:*

$$CC_\psi HdC'_\psi \sqrt{C_{\phi,\lambda} d' T} \sqrt{\log\left(T + \frac{T^3 C_\phi \left(Var_{\bar{M}} + \epsilon_{G,T}(\mathscr{M})\right)}{d}\right)} \,. \tag{12}$$

The root of the mean estimation error $\sqrt{\epsilon_{G,T}(\mathscr{M})}$ can be upper bounded by:

$$H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}}) + \max_{g\in[G]} \frac{\beta_{g,T}(1/NH)}{\lambda_{min}^{1/2} V_{g,T}^{\lambda}} \,. \tag{13}$$

The proof is in appendix D. We show with a Frobenius matrix norm version of Bennetts inequality that $H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}})$ goes for an increasing number of tasks to zero. This means the estimation error is dominated by the second term, the variance of the worst task so far. Recall our choice

of $\lambda = 1/(NHVar_{\boldsymbol{W}})$, 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.

3.2. 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 bias estimator in (Cella et al., 2020) and is in line with previous estimators in the multi-task 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. We define $\tilde{\boldsymbol{V}}_{G,n,h} \coloneqq \sum_{g=1}^{G} \boldsymbol{V}_{G,T} + \boldsymbol{V}_{G+1,n,h}$ and propose to use the following estimator:

$$\hat{\boldsymbol{W}}_{G,n,h} = (\tilde{\boldsymbol{V}}^{\lambda}_{G,n,h})^{-1}\Big(\sum_{g=1}^{G}\sum_{n,h}^{N,H} \phi_{g,n,h}\psi_{g,n,h}\boldsymbol{K}_{\psi}^{-1} + \sum_{n',h}^{n,H} \phi_{g,n',h}\psi_{g,n',h}\boldsymbol{K}_{\psi}^{-1}\Big) \tag{14}$$

This estimator is the transition core returned by meta training in algorithm 2. The transfer regret of this procedure can be upper bounded as:

Theorem 3.2. *BUC-MatrixRL incurs after T interactions in G previous tasks, using the bias $\hat{\boldsymbol{W}}_{G,n,h}$ as in equation 14 and assumptions as in theorem 2.4, the transfer regret as in equation 12. 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 root of the mean estimation error can be upper bounded by:*

$$H_{\mathscr{M}}(G+1, \bar{\boldsymbol{M}}) + 2(G+1)\max_{g\in[G+1]} \tilde{H}(G+1, \boldsymbol{M}_g) + \underbrace{\frac{dC_M}{\lambda+\nu_{\min}} + C'_{\psi}\sqrt{\frac{2}{\lambda+\nu_{\min}}\log\left(NH + \frac{GN^2H^2C_{\phi}}{\lambda d}\right)}}_{\frac{\beta^{\lambda}(1/(GNH))}{\lambda+\nu_{\min}}}.$$

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

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

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

The regret bound follows from theorem 2.4. and the proof is defered to appendix E.

Comparing this to the low bias solution in theorem 3.1 we see that the variance is now $\frac{\beta^{\lambda}(1/(GNH))}{\lambda+\nu_{\min}}$ instead of $\frac{\beta_{g,T}(1/NH)}{\lambda_{min}^{1/2}V^{\lambda}_{g,T}}$. 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 this 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. For a detailed analysis, we refer to appendix E.1. This approach requires to store GNH transitions to compute the aggregated bias matrix of meta training.

4. 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 to characterise 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 settings with 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. Creating and analysing communication protocols for knowledge transfer in the multitask setting is an interesting avenue for future research.

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.

References

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

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.

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.

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

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.

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

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

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*.

Appendix

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 $\boldsymbol{A} \in R^{d\times d}$:

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

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 $V_0 \in \mathbb{R}^{d,d}$ positive definitive and $\phi_1, \ldots, \phi_h \in \mathbb{R}^d$ a sequence of vectors with $\|\phi_t\|_2^2 \le L^2 \ \ \forall t \in [h]$, $V_h = V + \sum_{h\le t} \phi_h^T\phi_h$. Then:*

$$\sum_{t=1}^{h} \min(1, \|\phi_t\|_{V_t^{-1}}^2) \le 2\log\left(\frac{\det V_h}{\det V_0}\right) \le 2d\log\left(\frac{trV_0 + hL^2}{d\det V_0^{1/d}}\right) \tag{17}$$

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

$$2\log\left(\frac{\det V_h}{\det V_0}\right) \le 2d\log\left(\frac{\lambda d + hC_\phi}{d\lambda}\right) = 2d\log\left(1 + \frac{HC_\phi}{\lambda d}\right) \ . \tag{18}$$

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

$$2\log\left(\frac{\det V_h}{\det V_0}\right) \le 2d\log\left(\frac{\lambda d + hdC_\phi}{d\lambda}\right) = 2d\log\left(1 + HC_\phi\right) \ . \tag{19}$$

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

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

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. □

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

Proof of Lemma 2.1. 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 $e_{n,h}^c = 1 - e_{n,h}$.

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

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

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

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,

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

We therefore conclude that,

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

The result follows. □

C. Analysing the regret of (biased) UC-MatrixRL

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

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

C.1. Constructing Confidence Sets

Proof. of theorem 2.2

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:

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

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

We denote by $\eta_{n,h} = K_{\psi}^{-1}\psi_{n,h} - (M^*)^T\phi_{n,h} \in R^{d'}$ the noise vector. Using assumption 2.1, 2., we have:

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

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 $\sum_{n'<n,h\leq H} \phi_{n,h}\phi_{n,h}^T = V_n - \lambda I$ on the terms involving M^* and W we can write:

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

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

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

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

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

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

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

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

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

Recalling equation 16 we rewrite:

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

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:

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

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.

Finally, we highlight the impact of a particularly important choice of $\delta < 1/(NH)$:

$$\|M_n - M^*\|_F \leq C'_\psi \sqrt{d'd \log\left(NH + \frac{N^2 H^2 C_\phi}{d\lambda}\right)} + \sqrt{\lambda} \|W - M^*\|_F . \tag{32}$$

□

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*

$$\|\phi(s,a)^T (M - M_n)\|_1 \leq \beta_n \sqrt{\phi(s,a)^T (V_n)^{-1} \phi(s,a)} \tag{33}$$

Proof.

$$\begin{aligned}
\|\phi(s,a)^T (M - M_n)\|_1 &= \|\phi(s,a)^T V_n^{-1/2} V_n^{1/2} (M - M_n)\|_1 && (34)\\
&\leq \|\phi(s,a)^T V_n^{-1/2}\|_2 \|V_n^{1/2}(M - M_n)\|_{2,1} && (35)\\
&= \|\phi(s,a)^T V_n^{-1/2}\|_2 \|(M - M_n)\|_{V_n,1} && (36)\\
&\leq \beta_n \|\phi(s,a)^T V_n^{-1/2}\|_2 && (37)
\end{aligned}$$

□

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:

$$w_{n,h} = \sqrt{\phi(s_{n,h}, a_{n,h})^T (V_n)^{-1} \phi(s_{n,h}, a_{n,h})} = \sqrt{\phi_{n,h}^T (V_n)^{-1} \phi_{n,h}} . \tag{38}$$

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

$$Q_{n,h}(s_{n,h}, a_{n,h}) - [r(s_{n,h}, a_{n,h}) + P(\cdot|s_{n,h}, a_{n,h}) V_{n,h+1}] \leq 2C_\psi H \beta_n w_{n,h} \tag{39}$$

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,* $1 \leq \beta_1 \leq \cdots \leq \beta_N$*, then we can bound the regret at terminal time* $T = NH$*:*

$$Regret(T) \leq 2C_\psi H \beta_n \, \mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=1}^{H}\sqrt{\min(1, w_{n,h}^2)}\right] + \sum_{n=1}^{N} H\mathbb{P}\left[E_n = 0\right] \tag{40}$$

The only term that remains to be bounded is:

$$\sum_{n=1}^{N}\sum_{h=1}^{H}\sqrt{\min(1, w_{n,h}^2)} \leq \sqrt{HN\sum_{n=1}^{N}\sum_{h=1}^{H}\min(1, w_{n,h}^2)} \tag{41}$$

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 42. 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))**.**

$$\sum_{n=1}^{N}\sum_{h=1}^{H}\min(1, w_{n,h}^2) \leq 2H \ln\frac{\det(V_{N+1})}{\det(V_0)} \leq 2Hd\ln\frac{NHC_\phi + \lambda}{\lambda} \tag{42}$$

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 42:

$$\sum_{n=1}^{N}\sum_{h=1}^{H}\min(1, w_{n,h}^2) \leq 2\sum_{n=1}^{N}\sum_{h=1}^{H}\ln(1 + w_{n,h}^2)\,. \tag{43}$$

We are now going to bound the r.h.s. by a log determinant ratio. Recalling the structure of V_n we have:

$$V_{n+1} = V_n + \sum_{h=1}^{H}\phi_{n,h}\phi_{n,h}^T\,. \tag{44}$$

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

$$\det(V_{n+1}) = \det(V_n) \times \det(I + V_n^{-1/2}\sum_{h=1}^{H}\phi_{n,h}\phi_{n,h}^T V_n^{-1/2})\,. \tag{45}$$

Thus we get:

$$\det(V_{n+1}) \geq \det(V_n)\prod_{h=1}^{H}(1 + w_{n,h}^2)^{1/H} \geq \det(V_{n-1})\ldots \tag{46}$$

$$\geq \det(V_0)\prod_{n=1}^{N}\prod_{h=1}^{H}(1 + w_{n,h}^2)^{1/H}\,. \tag{47}$$

Putting things together and recalling eq. 18 we get:

$$\sum_{n=1}^{N}\sum_{h=1}^{H}\ln(1 + w_{n,h}^2) \leq 2H\ln\frac{\det(V_{N+1})}{\det(V_0)} \leq 2Hd\ln\left(1 + \frac{NHC_\phi}{\lambda d}\right)\,. \tag{48}$$

□

It remains to combine this with the confidence ellipsoid we constructed in theorem 2.2 which yields the regret bound:

Proof of theorem 2.3. By theorem 2.2 we choose:

$$\beta_n = C'_\psi \sqrt{2d' \log\left(\frac{1}{\delta}\right) + d'd \log\left(D\right)} + \sqrt{\lambda}\|W - M^*\|_F \ . \tag{49}$$

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

$$\mathbb{P}\left[\forall n \leq N | E_n = 1\right] \geq 1 - \delta \tag{50}$$

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

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

Using lemma 2.1 instead of lemma C.4 yields:

$$Regret(T, \mathcal{M}) \leq 2C_\psi H \beta_N \sqrt{C_{\phi,\lambda} Td \ln\left(1 + \frac{NHC_\phi}{\lambda d}\right)} \tag{52}$$

In the remainder of the paper we use the regret bound obtained by using lemma 2.1. □

Proof of theorem 2.4. We start by recalling the definition of transfer regret in equation 2. The first inequality is a result of applying theorem 2.3 with task distribution $\mathcal{T}$. Jensen's inequality and reordering yields the second inequality. □

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 choose the estimator of the transition core as defined in equation 11. By the triangle inequality we have:

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

Using the equivalent task descriptions $\mathcal{T}$ and $\mathscr{M}$ and recalling the estimation error from equation 13:

$$\sqrt{\epsilon_{G,T}(\mathscr{M})} \leq H_{\mathscr{M}}(G+1, \bar{M}) + \max_{g\in[G]} \frac{\beta_{g,T}(1/NH)}{\lambda^{1/2}_{min} V^{\lambda}_{g,T}} \ . \tag{53}$$

To analyse the estimation error term we start by restating the following vectorial version of Bennett's inequality:

Lemma D.1 (Vectorial Version of Bennett's inequality; lemma 2 (Smale and Zhou, 2007)/lemma 3 (Cella et al., 2020))**.** *Let $m_1, \ldots, m_G$ 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: $\|M_g\|_2 \leq C_M \ \forall g \in [G]$. For any $\delta \in (0,1)$ it holds with confidence $1-\delta$:*

$$\|\frac{1}{G}\sum_{g=1}^{G}\left[m_g - \bar{m}\right]\|_2 \leq \frac{2\log(2/\delta)C_M}{G} + \sqrt{\frac{2\log(2/\delta)\sigma_M}{G}} \tag{54}$$

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 corollay:

Corollary D.1.1 (Frobenius Version of Bennett's inequality)**.** *Let $M_1, \dots, M_G$ 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: $\|M_g\|_F \leq C_M \ \forall g \in [G]$. For any $\delta \in (0,1)$ it holds with confidence $1-\delta$:*

$$\|\frac{1}{G}\sum_{g=1}^{G}\left[M_g - \bar{M}\right]\|_F \leq \frac{2\log(2/\delta)C_M}{G} + \sqrt{\frac{2\log(2/\delta)\sigma_M}{G}} \tag{55}$$

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

$$\lim_{G\to\infty} H_{\mathscr{M}}(G+1,\bar{M}) = 0\,. \tag{56}$$

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 $\lambda_{min}(V_{j,T}) \geq \log T$. By construction of the V_j^{λ} we have as such $\lambda_{min}(V_{j,T}^{\lambda}) \geq \lambda + \lambda_{min}(V_{j,T}) \geq \lambda + \log T$. We see that the smaller T the larger the impact of the size of λ. Recall further our regularisation strength schedule: $\lambda = \frac{1}{TVar_{\bar{W}}}$. We see that if the variance of our estimator is large, the resulting λ is small. This leads to smaller $\lambda_{min}(V_{j,T})$, 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.

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). □

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:

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

where we have denoted with $H_{\mathcal{T}}(G+1, \bar{\boldsymbol{M}})$ according to what we have done in subsection 3.1. We can now focus on the term $\|\bar{\boldsymbol{W}}'_{G,t} - \bar{\boldsymbol{W}}_{G,t}\|_F$ which can be equivalently rewritten as:

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

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 $\sigma_{\max}(A)$.

proof of theorem 3.2. By the triangle inequality we have:

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

According to lemma E.2 we rewrite:

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

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

which gives:

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

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

$$
\sqrt{\epsilon_{G,T}(\mathscr{M})} \leq H_{\mathscr{M}}(G+1,\bar{M}) + 2(G+1)\max_{g\in[G+1]}\tilde{H}(G+1,M_g) \tag{57}
$$

$$
+ \frac{dC_M}{\lambda+\nu_{\min}} + C'_{\psi}\sqrt{\frac{2}{\lambda+\nu_{\min}}\log\left(GH + \frac{GN^2H^2C_{\phi}}{\lambda d}\right)}\,. \tag{58}
$$

□

E.1. Impact of Task misalignment

The reduction in variance we obtain with the global ridge regression comes at the price of an increased bias. The bias term increases with the misalignment of the tasks, $\sigma_{\max}\left(V_{g,T}\tilde{V}^{-1}_{G,N,H}\right)$. We illustrate its behavior via two corner cases.

First, assume that the task distribution $\mathcal{T} = \{M\}$, so our task distribution consists of a transition core matrix. As we face always the same task we expect extremely favorable 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. So we can consider the task distribution, where each column of the transition cores is a basis vector. The larger the variance and dissimilarity in the set of tasks we have, the larger is the number of tasks G we require for an improvement in transfer regret.