Supplementary Material: Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems

Appendix

A Regret Analysis

We now prove the $\tilde{\mathcal{O}}\left(\sqrt{T\log\frac{1}{\delta}}\right)$ upper-bound on regret of Augmented RBMLE-UCB that was claimed in Section 4.

Lemma A.1. *The regret $R(T)$ of the Augmented RBMLE-UCB learning algorithm can be decomposed as $R(T) = R_1 + R_2 + R_3 + R_4$, where*

$$
\begin{aligned}
R_1 &:= \sum_{t=0}^{T}\left\{x_t^\intercal P(\theta_t)x_t - \mathbb{E}\left[x_{t+1}^\intercal P(\theta_{t+1})x_{t+1}|\mathcal{F}_t\right]\right\},\\
R_2 &:= \sum_{t=0}^{T}\mathbb{E}\big[x_t^\intercal\left(P(\theta_{t+1}) - P(\theta_t)\right)x_t|\mathcal{F}_t\big],\\
R_3 &:= \sum_{t=0}^{T}\left\{\left(A^\star x_t + B^\star u_t\right)^\intercal P(\theta_t)\left(A^\star x_t + B^\star u_t\right) - \left(A_t x_t + B_t u_t\right)^\intercal P(\theta_t)\left(A_t x_t + B_t u_t\right)\right\},\\
R_4 &:= \sum_{t=0}^{T}\left(J(\theta_t) - J(\theta^\star)\right).
\end{aligned}
\tag{19}
$$

Proof. Consider an algorithm that implements $u_t = K(\theta_t)x_t$ at time t. Note that $x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1}$. Define $\tilde{x}_{t+1}^u := A_t x_t + B_t u + w_{t+1}$. Then, the Bellman optimality equation for the Linear Quadratic control problem can be written as follows,

$$
\begin{aligned}
J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t &= \min_u\left(x_t^\intercal Q x_t + u^\intercal R u + \mathbb{E}\left[(\tilde{x}_{t+1}^u)^\intercal P(\theta_t)\tilde{x}_{t+1}^u|\mathcal{F}_t\right]\right)\\
&= \left(x_t^\intercal Q x_t + u_t^\intercal R u_t + \mathbb{E}\left[(\tilde{x}_{t+1}^{u_t})^\intercal P(\theta_t)\tilde{x}_{t+1}^{u_t}|\mathcal{F}_t\right]\right).
\end{aligned}
$$

Upon substituting the value of $\tilde{x}_{t+1}^{u_t}$ in the above, we get

$$
\begin{aligned}
&J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t\\
&= (x_t^\intercal Q x_t + u_t^\intercal R u_t) + \mathbb{E}\Big[(A_t x_t + B_t u_t + w_{t+1})^\intercal P(\theta_t)(A_t x_t + B_t u_t + w_{t+1})|\mathcal{F}_t\Big].
\end{aligned}
\tag{20}
$$

Note that $w_{t+1} = x_{t+1} - (A^\star x_t + B^\star u_t)$ and w_t is a martingale difference sequence (Assumption 2). Thus, the l.h.s. of (20) can be written as follows,

$$
\begin{aligned}
J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t - (x_t^\intercal Q x_t + u_t^\intercal R u_t) =& J^\star(\theta_t) - J^\star(\theta^\star) + x_t^\intercal P(\theta_t)x_t\\
&- (x_t^\intercal Q x_t + u_t^\intercal R u_t - J^\star(\theta^\star))\\
&= \mathbb{E}\left[x_{t+1}^\intercal P(\theta_{t+1})x_{t+1}|\mathcal{F}_t\right]\\
&- \mathbb{E}\big[x_t^\intercal\left(P(\theta_{t+1}) - P(\theta_t)\right)x_t|\mathcal{F}_t\big]\\
&- \left(A^\star x_t + B^\star u_t\right)^\intercal P(\theta_t)\left(A^\star x_t + B^\star u_t\right)\\
&+ \left(A_t x_t + B_t u_t\right)^\intercal P(\theta_t)\left(A_t x_t + B_t u_t\right).
\end{aligned}
$$

Therefore by taking a sum from $t = 0$ to $t = T$ on both sides, we get $R(T) = R_1 + R_2 + R_3 + R_4$. □

Lemma A.2. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_1 \le 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B_\delta}$, where $B_\delta := b\log\left(\frac{4n\sqrt{b}}{v\delta}\right)$, $b := v + T(cDX_T)^2(1+c_0^2)$, $W = nL\sqrt{2n\log\left(\frac{8nT}{\delta}\right)}$ and D is as in* (10).

Proof. The proof is the same as that of Lemma 7 in [3], and hence omitted. □

Lemma A.3. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_2 \leq 2DX_T^2 \log_2 T$, where X_T is defined in* (17).

Proof. The term in the summation $\sum_{t=0}^{T} \mathbb{E}\left[x_t^\intercal \left(P(\theta_{t+1}) - P(\theta_t)\right) x_t | \mathcal{F}_t\right]$ (19) corresponding to time t is non-zero only when a change in the policy occurs at t. There are $(n+m)\log_2\left(1 + TX_T^2 \frac{1+c_0^2}{\lambda}\right)$ episodes till time T (Lemma 8, [3]). Therefore, there are $(n+m)\log_2\left(1 + TX_T^2 \frac{1+c_0^2}{\lambda}\right)$ non-zero terms and each of them is bounded by $2DX_T^2$. □

Lemma A.4. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_3 \leq \frac{8cX_T^2 D(1+c_0^2)}{\sqrt{\lambda}} \sqrt{T\beta_T\left(\frac{\delta}{8}\right) \log \frac{det(Z_T)}{det(\lambda I)}}$, where $Z_T := \max_{0\leq t\leq T} \|z_t\|$, and $\beta_T\left(\frac{\delta}{8}\right)$ is defined in* (15).

Proof. The proof is the same as that of Lemma 13 in [3], and hence omitted. □

Lemma A.5. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, $R_4 \leq \frac{1}{\alpha_0}\left(\beta_T\left(\frac{\delta}{4}\right) + \lambda c^2\right)\sqrt{T}$.*

Proof. As defined in Lemma A.1, we have,

$$R_4 = \sum_{t=0}^{T} \left(J^\star(\theta_t) - J^\star(\theta^\star)\right).$$

During the k-th episode, the algorithm chooses $u_t = K(\theta_{t_k})x_t,\ \forall\, t = t_k, t_k+1, \ldots, t_{k+1}$, where, θ_{t_k} is as in (18), and obtained by solving the corresponding optimization problem at the beginning of the episode at time t_k. Therefore R_4 can be written as :

$$R_4 = \sum_{k=0}^{K} \Delta_k, \text{ where } \Delta_k := \left(t_{k+1} - t_k\right)\left(J^\star(\theta_{t_k}) - J^\star(\theta^\star)\right).$$

Δ_k is bounded as follows:

$$\Delta_k = \left(t_{k+1} - t_k\right)\left(J(\theta_{t_k}) - J(\theta^\star)\right) \leq \frac{\left(t_{k+1} - t_k\right)}{\alpha(t_k)}\left(V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k})\right), \tag{21}$$

where the inequality holds since θ_{t_k} is a minimizer of $V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)$ (18). Moreover,

$$\begin{aligned} V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k}) &= V_{t_k}(\theta^\star) + V_{t_k}(\hat{\theta}_{t_k}) - V_{t_k}(\hat{\theta}_{t_k}) - V_{t_k}(\theta_{t_k}) \\ &\leq V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}), \end{aligned}$$

where the inequality follows since $\hat{\theta}_{t_k}$ is a minimizer of $V_{t_k}(\cdot)$.

Since $\theta^\star \in C_{t_k}(\delta)$, it follows from the definition of the confidence ball that $V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}) \leq \beta_{t_k}\left(\frac{\delta}{4}\right)$. Since $\beta_t(\delta/4) \leq \beta_T(\delta/4)$, we have $V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}) \leq \beta_T\left(\frac{\delta}{4}\right)$. Therefore,

$$V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k}) \leq \beta_T\left(\frac{\delta}{4}\right). \tag{22}$$

Setting $\alpha(t) = a_0\sqrt{T}$, we get

$$\sum_{k=1}^{K} \Delta_k \leq \beta_T\left(\frac{\delta}{4}\right) \sum_{k=1}^{K} \frac{t_{k+1} - t_k}{\alpha_0\sqrt{T}} = \frac{1}{\alpha_0}\beta_T\left(\frac{\delta}{4}\right)\sqrt{T}.$$

□

A.1 Proof of Theorem 4.1

Proof. To analyze regret on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we substitute individual bounds on $R_1,\ R_2,\ R_3$ and R_4 in order to obtain

$$R(T) \leq 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B_\delta} + 2DX_T^2\log_2 T + \frac{8X_T^2SD(1+C^2)}{\sqrt{\lambda}}\sqrt{T\beta_T\left(\frac{\delta}{8}\right)\log\frac{det(Z_T)}{det(\lambda I)}} + \frac{1}{\alpha_0}\beta_T\left(\frac{\delta}{4}\right)\sqrt{T}.$$

□

B Definition of d_t

The quantity d_t in the definition of $\mathcal{E}_2(t)$ in (16) is defined as follows,

$$\begin{aligned}
d_t &:= \frac{1}{1-\rho}\left(\frac{\eta}{\rho}\right)^{n+m}\left[2L\sqrt{n\log\frac{4nt(t+1)}{\delta}} + G\mathcal{Z}_T^{\frac{n+m}{n+m+1}}\beta_t\left(\frac{\delta}{4}\right)^{(2(n+d+1))^{-1}}\right],\\
\eta &:= \max\left\{1, \sup_{\theta\in\mathcal{S}}||A^\star + B^\star K(\theta)||\right\},\\
\mathcal{Z}_T &:= \max_{0\leq t\leq T}\|z_t\|,\\
G &:= 2\left(\frac{2c(n+m)^{n+m+0.5}}{\sqrt{U}}\right)^{(n+m+1)^{-1}},\\
U &:= \frac{U_0}{H},\\
U_0 &:= \frac{1}{16^{n+m-2}\max\{1, c^{2(n+m-2)}\}},\\
H &\text{ is constant such that } H > \max\left\{16, \frac{4c^2M^2}{(n+m)U_0}\right\},\\
M &:= \sup_{Y\geq 0}\frac{nL\sqrt{(n+m)\log\left(\frac{1+\frac{TY}{\lambda}}{\delta}\right)} + \lambda^{1/2}c}{Y}.
\end{aligned}$$

C Simulation Experiments

In this section, we provide the details on the simulation experiments, along with some additional results. The code and instructions for replicating the presented results are provided in the supplementary material.

1. We begin by describing the linear systems used for our experiments in Section 5.
   (a) **Unmanned Aerial Vehicle (UAV):** This system represents a linearized dynamics of an unmanned aerial vehicle (UAV) in a two-dimensional plane, which has been recently studied in the context of reinforcement learning in [17, 20]. The first and third states represent the positions, while the second and fourth states represent the velocities in each dimension. The inputs are accelerations in each dimension.

$$A^\star = \begin{bmatrix}1 & 0.5 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.5\\ 0 & 0 & 0 & 1\end{bmatrix}, B^\star = \begin{bmatrix}0.125 & 0\\ 0.5 & 0\\ 0 & 0.125\\ 0 & 0.5\end{bmatrix}, Q = diag(1, 0.1, 2, 0.2), R = I_2.$$

   (b) **Unstable Laplacian Dynamics** This represents a Laplacian system where the adjacent nodes are weakly connected. The lack of stablility (i.e., $\lambda_{max}(A^\star) \geq 1$) makes it a

challenging example for system identification and hence it has been studied recently in [18, 19, 55, 58, 17]. The system matrices are as follows:

$$A^\star = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, B^\star = I_3, Q = I_3, R = I_3.$$

(c) **Large transient dynamics:** We also consider the following unstable system which additionally exhibits large transients.

$$A^\star = \begin{bmatrix} 1 & 0 & 0 \\ 1.1 & 1 & 0 \\ 0 & 1.1 & 1 \end{bmatrix}, B^\star = I_3, Q = I_3, R = I_3.$$

(d) **Longitudinal Flight Control of Boeing 747:** This represents the linearized dynamics of Boeing 747 at 40,000 ft altitute and speed of 774 ft/sec, which was first introduced in [59]. The empirical performance of OFULQ, TS and StabL for this system was recently studied in [17]. The four states represent velocity of aircraft along the body axis, velocity perpendicular to the body axis, angle of the body axis with horizontal and the angular velocity. The inputs are elevator angle and thrust of the aircraft. The system matrices are as follows:

$$A^\star = \begin{bmatrix} 0.99 & 0.03 & -0.02 & -0.32 \\ 0.01 & 0.47 & 4.7 & 0 \\ 0.02 & -0.06 & 0.4 & 0 \\ 0.01 & -0.04 & 0.72 & 0.99 \end{bmatrix}, B^\star = \begin{bmatrix} 0.01 & 0.99 \\ -3.44 & 1.66 \\ -0.83 & 0.44 \\ -0.47 & 0.25 \end{bmatrix}, Q = I_4, R = I_4.$$

2. In our experiments, we compared the empirical performance of Augmented RBMLE-UCB and RBMLE with following algorithms: (1) OFULQ [3], (2) Thompson Sampling [15], (3) StabL [17], (4) Randomized Certainty Equivalence (RCE) [10], and (5) Input Perturbations [16]. The pseudo-code for all of the implemented algorithms is given in Algorithm 2, where the choice of θ_{t_k} and u_t made by each algorithm are described in Table 2. The optimization problems for ARBMLE, RBMLE, OFULQ and StabL described in Table 2 are non-convex problems. We used projected gradient descent to solve the optimization problems. Expression for gradient of the RBMLE objective with respect to θ can be obtained explicitly as in [52].

Algorithm 2 Reinforcement Learning for LQ systems.

Initialize: $t = 0$, $Z_0 = \lambda I_{n+m}$
for $k = 0, 1, \cdots$ **do**
  if $det(Z_t) > 2det(Z_{t_{k-1}})$ **then**
    Calculate θ_t as defined by the RL algorithm (See Table 2).
  else
    $\theta_t = \theta_{t-1}$
  end if
  $u_t = f(K(\theta_t), x_t)$ (See Table 2).
  $Z_{t+1} = Z_t + z_s z_s^T$
end for

3. Initially, the controls are chosen as follows in order to obtain a initial estimates of the system:

$$u_t = K_{\text{init}} x_t + \eta_t \text{ for } 0 \leq t \leq T_{\text{init}} \text{ and } \eta_t \text{ is } \mathcal{N}(0, 1),$$

i.i.d., K_{init} is a stabilizing controller and $x_0 = 0$. The noise is pre-generated, ensuring that initialization is uniform across algorithms. The definition of confidence interval in ARBMLE, TS and StabL depends on the choice of confidence parameter δ and a constant c such that $trace(\theta^{\star T}\theta^\star) < c^2$. StabL algorithm uses an excitation $\mathcal{N}(0, \sigma_w^2)$ for $T < T_w$. The values of various hyper-parameters used in our experiments are described in Table 3.

Algorithm	θ_{t_k}	$u_t,\ \forall\, t\ \in \{t_k, t_k+1, \ldots, t_{k+1}-1\}$
ARBMLE	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} \{V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)\}$,	$K(\theta_{t_k})x_t$
RBMLE	$\arg\min_{\theta\in\mathcal{S}} \{V_{t_k}(\theta) + \alpha(t)J^\star(\theta)\}$,	$K(\theta_{t_k})x_t$
OFULQ	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} J^\star(\theta)$	$K(\theta_{t_k})x_t$
TS	$\hat{\theta}_{t_k} + \beta_{t_k}(\delta)Z_{t_k}^{-1/2}\mathcal{N}(0,1)$	$K(\theta_{t_k})x_t$
IP	$\hat{\theta}_{t_k}$	$K(\theta_{t_k})x_t + \eta_t^{IP}$
RCE	$\hat{\theta}_{t_k} + \eta_t^{RCE}$	$K(\theta_{t_k})x_t$
StabL	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} J^\star(\theta)$	$\begin{cases} K(\theta_{t_k})x_t + \mathcal{N}(0,\sigma_w^2) \text{ if } T < T_w \\ K(\theta_{t_k})x_t \text{ otherwise} \end{cases}$

Table 2: Choices of θ_{t_k} and u_t for various algorithms.

Parameter	T_{init}	T	δ	λ	α_0	σ_w	T_w
Value	50	500	10^{-4}	10^{-4}	10^{-2}	2	35

Table 3: Values of various parameters

4. We provide simulation results for following additional examples. Figure 3 includes the comparison between ARBMLE, OFULQ, TS and StabL. Figure 4 provides a comparison between ARBMLE, RBMLE, RCE and IP. The average regret values for these systems at $T = 500$ are shown in Table 4.
5. (a) Stabilizable but Not Controllable System: We consider a system studied in [17] which is stabilizable but not controllable. Lack of controllability is challenging for system identification. ARBMLE/RBMLE outperforms OFULQ, TS, StabL and RCE by a significant margin.

$$A^\star = \begin{bmatrix} -2 & 0 & 1.1 \\ 1.5 & 0.9 & 1.3 \\ 0 & 0 & 0.5 \end{bmatrix}, B^\star = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, Q = R = I_3.$$

(b) Chained Integrator Dynamics: We consider a simple chained integrator system with 2-dimensional states and 2-dimensional input.

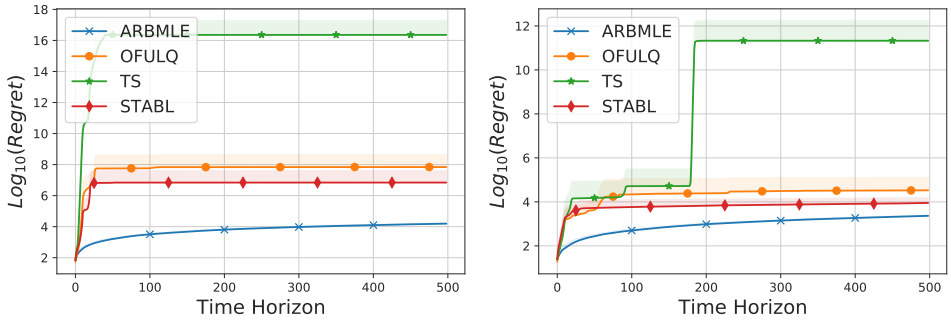
$$A^\star = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, B^\star = I_2, Q = I_2, R = I_2.$$

Ex.	RBMLE	ARBMLE	OFULQ	TS	IP	RCE	STABL
(a)	15665	15663	6.9×10^7	2.2×10^{16}	15628	39593	6.9×10^6
(b)	2322	2322	33449	2.1×10^{11}	2337	2402	8927

Table 4: Average Regret Performance at $T = 500$.

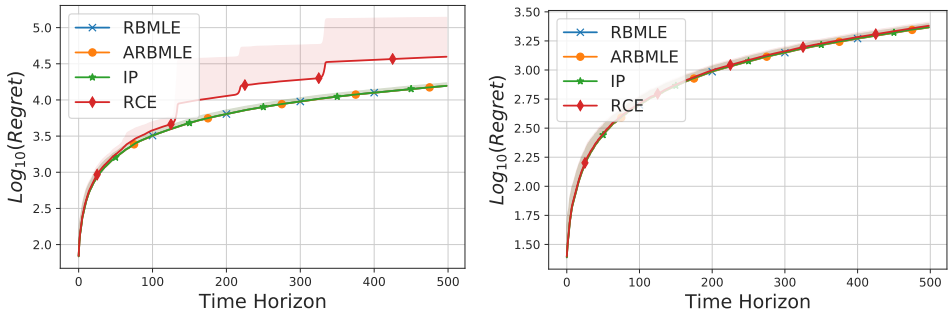
6. Additonal Remarks:

- We use a stabilizing controller K_{init} for initialization of our simulation experiments similar to [18]. Note that our theoretical regret analysis does not assume knowledge of a stabilizing controller, unlike some recent works on adaptive control of LQG systems including [18, 10, 56].
- As demonstrated in the simulation results, OFULQ and Thompson Sampling have a very large initial regret indicating poor initial estimates of system parameters (also highlighted in [17]). ARBMLE/RBMLE, IP and RCE show much better initial regret performance compared to OFULQ and TS.
- Implementation of ARBMLE, OFULQ, StabL and TS involve definition $\beta_t(\delta)$ which denotes boundary of confidence interval. Our simulations for ARBMLE, OFULQ, StabL and TS are based on $\beta_t(\delta)$ as defined in (15). Instead, recent works [18, 17] use $\beta_t(\delta) := trace\left((\theta^\star - \hat{\theta}_t)^\intercal Z_t(\theta^\star - \hat{\theta}_t)\right)$. However, one may note that $\theta^\star$ in $\beta_t(\delta)$ is not known to the learning agent, and so such a definition of $\beta_t(\delta)$ is not a viable for implementation. The effect of the choice of $\beta_t(\delta)$ on the regret performance is shown in the Figure 5.



(a) Stabilizable, not controllable (b) Chained Integrator Dynamics

Figure 3: Logarithm of the Averaged Regret over 50 runs of ARBMLE, OFULQ, TS. and StabL.

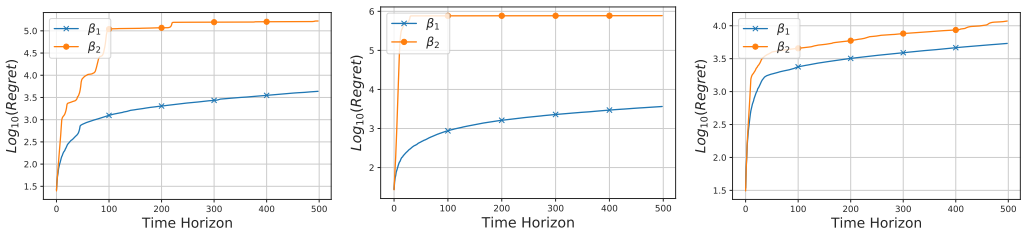


(a) Stabelizable, not controllable (b) Chained Integrator Dynamics

Figure 4: Logarithm of the Averaged Regret over 50 runs of ARBMLE, RBMLE, RCE, TS

- The estimates of OFULQ lies on boundary, while the estimates of ARBMLE/RBMLE, IP and RCE are closer to the least squared estimate. Note that RBMLE can be seen as Lagrangian version of OFULQ, indicating that $\alpha(t)$ may be much smaller than the implicit Lagrange multiplier for OFULQ.

7. The code for our simulation experiments is provided in the supplementary material. The seed values for random number generation are set appropriately for replication of the results. The instruction for the code are provided in supplementary material.



(a) OFULQ (b) TS (c) Stabl

Figure 5: Effect of choice of confidence interval definition on performance. β_1: confidence interval as defined in 15. β_2: confidence interval as defined in [18]