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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of controlling a stochastic linear system with quadratic costs, when its system parameters are not known to the agent – called the adaptive LQ control problem. We re-examine an approach called "Reward-Biased Maximum Likelihood Estimate" (RBMLE) [1] that was proposed more than forty years ago, and which predates the "Upper Confidence Bound" (UCB) method as well as the definition of "regret" [2]. It simply added a term favoring parameters with larger rewards to the criterion for parameter estimation. We show how the RBMLE and UCB methods can be reconciled, and thereby propose an Augmented RBMLE-UCB algorithm that combines the penalty of the RBMLE method with the constraint of the UCB method [3], uniting the two approaches to optimism in the face of uncertainty. We establish that theoretically this method retains $\tilde{\mathcal{O}}(\sqrt{T})$ regret, the best known so far. We further compare the empirical performance of the proposed Augmented RBMLE-UCB and the standard RBMLE (without the augmentation) with UCB (OFU), Thompson Sampling (TS), Input Perturbation (IP), Randomized Certainty Equivalence (RCE) and StabL on many real-world examples that include flight control of Boeing 747 and Unmanned Aerial Vehicle. We perform extensive simulation studies showing that the Augmented RBMLE consistently outperforms OFU, TS and StabL by a huge margin, while it is marginally better than IP and moderately better than RCE.

1 Introduction

Consider a linear stochastic system $x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1}$, where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$ and w_t are the state, control applied, and state noise, respectively, at time t. We study the adaptive control/reinforcement learning problem [4, 5, 6, 7] where the controller's goal is to minimize the expected finite horizon quadratic cost $\mathbb{E}\sum_{t=1}^{T}\left(x_t^T Q x_t + {u_t}^T R u_t\right)$ by choosing the control input based on observing the past states and controls, without knowing the "true parameter" $\theta^\star := (A^\star, B^\star)^T$. The deviation from what would have been optimally possible had the true parameter been known is measured by the expected value of the "regret," [2], $R(T) := \sum_{t=1}^{T}\left(x_t^T Q x_t + u_t^T R u_t\right) - T\ J^\star(\theta^\star)$, where $J^\star(\theta^\star) := \limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\left(x_t^T Q x_t + u_t^T R u_t\right)$ is the optimal average cost achievable when $(A^\star, B^\star)$ is known.

For this adaptive linear-quadratic (LQ) control problem, there are broadly four classical approaches, the Reward-Biased Maximum Likelihood Estimate (RBMLE) approach [1], the Diminishing Excitation (DE) approach [8], the Upper Confidence Bound (UCB) approach [2], and the Thompson Sampling (TS) [9] approach based on sampling from the posterior distribution. The RBMLE and UCB approaches are "certainty equivalent" (CE) in the sense that they make an estimate $\hat{\theta}_t$ of the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

unknown true parameter, and then take an action $u_t = K(\hat{\theta}_t)x_t$ that would be optimal if the estimate $\hat{\theta}_t$ were indeed the true parameter. They only differ in what parameter estimate they choose. The DE approach applies $u_t = K(\hat{\theta}_t)x_t + v_t$, where v_t is an added "excitation," an independent noise, of diminishing variance and $\hat{\theta}_t$ is the maximum likelihood estimate. The recent scheme Randomized Certainty Equivalence (RCE) adds the excitation to the parameter estimate.

1.1 The Contributions

1. We unite the two approaches to "optimism under uncertainty," RBMLE [1] and UCB [2], by showing that the RBMLE method is a penalty version of the constrained optimization problem of UCB for the case of linear quadratic systems [10, 11, 12, 13, 3]. Based on this we propose an Augmented RBMLE (ARBMLE) method that combines the penalty and constrained versions so that on the one hand it retains the analytical tractability of UCB and on the other hand provides the performance of RBMLE.
2. We determine how to choose the biasing factor $\alpha(t)$ for ARBMLE, and establish a finite time regret bound $\tilde{\mathcal{O}}(\sqrt{T})$, the same as the OFU algorithm of [3], the best order available to date.
3. We perform extensive comparative simulation studies of the performance of:
   (a) ARBMLE and the standard RBMLE.
   (b) OFU [3], which is the UCB-approach adapted to the LQ problem.
   (c) TS [14] which is the Thompson sampling approach adapted to the LQ problem
   (d) Input Perturbation (IP) [15] which is a recent reincarnation of DE for the adaptive LQ problem that additionally assumes a piori knowlege of a stabilizing controller.
   (e) Stabl [16], a modified OFU that adds "excitation" to the input for initialization.
   (f) RCE [17], which adds excitation to the Least Squares Estimate (LSE).

   The examples used for our simulation study have been used in many recent papers [18, 19, 16], namely (a) the longitudinal flight control of Boeing 747 with linearized dynamics [16],(b) Unmanned Aerial Vehicle (UAV) [20, 16] (c) unstable Laplacian dynamics [18], and (d) large transient dynamics [18]. Our simulation results show that ARBMLE outperforms OFU, TS and StabL by a large margin, which is primarily due to lack of stabilization experienced by OFU and TS [16]. RCE also exhibits a higher regret than ARBMLE. While the empirical performance of IP is marginally worse than ARBMLE. The results show that the ARBMLE has the same performance as the original RBMLE [21], and they both outperform all the above other algorithms. Notably the choices made by ARBMLE and OFU within the confidence interval are very far apart, with ARBMLE outperforming OFU by a large margin.

1.2 Previous Works

Prior work on RBMLE has concentrated on establishing long-term average optimality [22, 23, 1, 10, 24, 25, 11, 12, 13, 26, 27, 28]. Recently, its regret performance has been addressed for Multi-Armed Bandits (MAB)s [29], linear contextual bandits [30], and MDPs [31]. Forced exploration techniques, somewhat similar in spirit to the ϵ_t-greedy learning algorithm of [7, 32], are studied in [33, 34, 35] vis-a-vis ensuring that they do not suffer from the insufficient exploration. An adaptive LQG control based on the UCB approach, called OFU, is proposed in [3] which establishes a regret of $\tilde{\mathcal{O}}(\sqrt{T})$. A similar algorithm to address the adaptive LQG control problem is also designed in [36]. A computationally efficient algorithm called ROBUST with a regret of $\tilde{\mathcal{O}}(T^{0.67})$ is proposed in [18]. An alternative approach for designing learning algorithms is Thompson sampling [9], and [37, 38] have established an expected regret of $\tilde{\mathcal{O}}(\sqrt{T})$ in a Bayesian context. More recently, new DE-based algorithms, IP [15] and RCE [17], have been proposed that make an additional assumption, not made in ARBMLE, OFULQ or TS, that one has access to a stabilizing controller for the unknown system, and establish $\tilde{O}(\sqrt{T})$ regret.

1.3 The RBMLE Approach

The RBMLE approach, proposed four decades ago [1], was the first approach not resorting to forced choices as in [39]. We begin by giving an informal description of it in the context of the adaptive

LQ problem. Since the true parameter $\theta^\star := (A^\star, B^\star)^T$ is not known one can make a Least-Squares Estimate (LSE) $\hat{\theta}_t = (\hat{A}_t, \hat{B}_t)^T$ of them:

$$\hat{\theta}_t \in \text{ArgMin}_{\theta=(A,B)^T} \sum_{s=0}^{t-1} ||x_{s+1} - Ax_s - Bu_s||^2. \tag{1}$$

Under the certainty-equivalent approach, the control input applied is $u_t = K(\hat{\theta}_t)x_t$ where $K(\theta)$ is the optimal linear feedback gain for the LQ problem when the system is described by $\theta = (A, B)^T$. Suppose now that these estimates were to converge as $t \to \infty$ to $\hat{\theta}_\infty = (\hat{A}_\infty, \hat{B}_\infty)^T$. Then, asymptotically, the input applied is $u_t \approx K(\hat{\theta}_\infty)x_t$. The closed-loop system therefore settles down to behaving according to

$$x_{t+1} \approx (A^\star + B^\star K(\hat{\theta}_\infty))x_t + w_{t+1}. \tag{2}$$

As it does so, one loses the ability to identify the matrices $(A^\star, B^\star)$, and asymptotically one can only identify the closed-loop gain $A^\star + B^\star K(\hat{\theta}_\infty)$. The problem is that as the parameter estimates begin to converge to $\hat{\theta}_\infty$, the control gain converges to $K(\hat{\theta}_\infty)$ and further exploration ceases, and one ends up only identifying the behavior of the system under the limiting gain $K(\hat{\theta}_\infty)$ being applied to the system. Since the limiting policy $u = K(\hat{\theta}_\infty)x$ need not be optimal for the long-term average cost for the true system $(A^\star, B^\star)$, the CE rule leads to a sub-optimal performance. Indeed this problem goes by various names in different fields – the dual control problem [40, 41], the closed-identifiability problem [42, 43, 44], or the problem of exploration vs. exploitation [45].

This problem was resolved in [1] without resorting to forced exploration as in [39]. The key observation made there was that under CE the limiting parameter estimate $\hat{\theta}(\infty)$ has a one-sided bias. Specifically, $J^\star(\hat{\theta}_\infty) \geq J^\star(\theta^\star)$, i.e., the limiting parameter estimate has an optimal cost that is larger than the optimal cost of $\theta^\star$. To see this, denote by $J(K, \theta)$ the long-term cost of using the control gain K when the parameter is θ. Then the fact that the models $\hat{\theta}_\infty$ and $\theta^\star$ have identical behavior under the gain $K(\hat{\theta}_\infty)$ implies that $J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta^\star})$. Now note that $J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J^\star(\hat{\theta}_\infty)$ since the gain $K(\hat{\theta}(\infty))$ is optimal for $\hat{\theta}_\infty$. However the gain $K(\hat{\theta}_\infty)$ is not necessarily optimal for $\theta^\star$, and so $J(K(\hat{\theta}_\infty), \hat{\theta^\star}) \geq J^\star(\theta^\star)$. Therefore

$$J^\star(\hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta^\star}) \geq J^\star(\theta^\star). \tag{3}$$

Following from this observation, it was reasoned in [1] that if one could slightly bias the MLE to favor models with lower optimal costs $J^\star(\theta)$ so as to obtain $J^\star(\hat{\theta}_\infty) \leq J^\star(\theta^\star)$, then one would have equality throughout (3), to obtain $J(K(\hat{\theta}_\infty), \hat{\theta^\star}) = J^\star(\theta^\star)$, yielding the desired result that the gain $K(\hat{\theta}_\infty)$ is optimal for $\theta^\star$. So motivated, [1] proposed RBMLE[1] which in the context of the LSE suggests choosing the parameter estimate $\hat{\theta}_{RBMLE}(t) = (\hat{A}_{RBMLE}, \hat{B}_{RBMLE})^T$ as:

$$\hat{\theta}_{RBMLE}(t) \in \text{ArgMin}_{\theta=(A,B)^T} \Big[\alpha(t)J^\star(\theta) + \sum_{t=0}^{T-1} ||x_{t+1} - Ax_t - Bu_t||^2\Big]. \tag{4}$$

Generally $\alpha(t) = o(t)$ with the precise growth rate of $\alpha(t) \nearrow +\infty$ dependent on the context.

1.4 The UCB Approach

The 'Upper Confidence Bound" (UCB) approach was first proposed in the context of Multi-Armed Bandit (MAB) Problems in [2]. In the context of Bernoulli bandits, it essentially consists of constructing, for each arm i at each time t, a confidence interval $(\theta(i, 1; t), \theta(i, 2; t))$ of its payoff probability with confidence $(1 - \delta(t))$, for $\delta(t) = o(\frac{1}{t})$, and playing the arm i with the highest value of $\theta(i, 1; t)$. This approach has been generalized to a variety of contexts including linear contextual bandits [46], Gaussian Processes [47], MDPs [48, 49], and LQ systems [3].

[1] It has been called the Cost-Biased MLE, as in [11], or Reward-Biased MLE (RBMLE) as in [29], depending on whether one minimizes a cost or maximizes a reward.

Particularized to LQ systems, the OFU algorithm of [3] that is based on the UCB approach suggests choosing $\hat{\theta}_{UCB}(t) = (\hat{A}_{UCB}, \hat{B}_{UCB})^T$ as the minimizer of:

$$\min_{\theta \in \mathcal{C}_t(\delta)} J^\star(\theta), \text{ where} \tag{5}$$

$$\mathcal{C}_t(\delta) := \left\{ \theta = (A, B)^T : \sum_{t=0}^{T-1} ||x_{t+1} - Ax_t - Bu_t||^2 \leq \gamma_T(\delta) \right\} \tag{6}$$

is a $(1-\delta)$-high confidence set of parameters for an appropriate choice of $\gamma_T(\delta)$ [3].

2 The Augmented RBMLE-UCB Method

The UCB approach has been called "Optimism in the Face of Uncertainty" (OFU), since it chooses an optimistic arm after calculating the confidence intervals. The RBMLE version is also the same, though it does it in a different way by directly giving preference to parameters that can yield better rewards. It arrives at optimism in a very systematic way by noticing that closed-loop identification leads to the chain of inequalities (3), which could then be made into equalities by ensuring that $J^\star(\hat{\theta}_\infty) \leq J^\star(\theta^\star)$.

We now show how one may reconcile the two approaches to optimism in the face of uncertainty in the context of the LQ problem, and then combine them to obtain a method that has experimentally superior empirical performance while also allowing a proof that it achieves the currently best known order of regret.

It can be seen that the RBMLE approach (4) can be considered an unconstrained penalty version of the constrained optimization problem (5)-(6), with a penalty factor $\frac{1}{\alpha(t)}$ for constraint violation. This provides a justification for employing "optimism" in the UCB approach.

This raises the question of whether we can take advantage of these synergies to fashion a superior algorithm, that is superior from the viewpoint of being able to provide a theoretical guarantee of the best known order of regret to date, as well as superior from the point of view of providing the best experimental performance of the algorithms to date. One can draw inspiration here from the Augmented Lagrangian Method of [50, 51]. Given a constrained problem as in (5)-(6), it adds a penalty for constraint violation, and also adjoins the constraint through a Lagrange multiplier, thus involving the constraint twice. In our case, we add the penalty and also retain the constraint. This leads to the Augmented RBMLE-UCB method:

$$\hat{\theta}_{\text{Aug}}(t) \in \text{ArgMin}_{\theta \in \mathcal{C}_t(\delta)} \Big[\alpha(t) J^\star(\theta) + \sum_{t=0}^{T-1} ||x_{t+1} - Ax_t - Bu_t||^2 \Big]. \tag{7}$$

The advantage of retaining the constrained optimization is theoretical, in that we can make use of the bounds on parameter estimates from concentration inequalities, as we shall show in the sequel. We thereby prove in Section 5 that it also has $\tilde{\mathcal{O}}(\sqrt{T})$ order of regret, which is the best known so far [3]. Moreover, simulations reported in Section 6 show that its performance is also much improved. In all cases, the simulation performance of ARBMLE is the same as RBMLE, raising the open problem of proving that the original RBMLE for the LQ problem also has the best known order of regret.

3 Problem Formulation

We make the following assumptions on $\theta^\star$.

Assumption 1. *There exists a known positive constant c such that $\theta^\star \in \mathcal{S} = \mathcal{S}_0 \cap \mathcal{S}_1$ where, $\mathcal{S}_0 = \{\theta = (A, B)^T \mid (A, B)$ is stabilizable and $\left(A, Q^{1/2}\right)$ is detectable$\}$[2] [54], and $\mathcal{S}_1 = \{\theta = (A, B)^T \mid trace(\theta^T \theta) \leq c\}$.*

[2] We note that even though [3] requires each θ from $\mathcal{S}_0$ to satisfy the stronger reachability-observability assumption, the results therein hold true if this is replaced by the weaker stabilizability-detectability assumption [52, p.61]. In fact all that is required for the proofs to go through, is that for each $\theta \in \mathcal{S}_0$, the corresponding Riccati equation to have a unique solution. It is well-known [53] that stabilizability and detectability, as above, is sufficient for this.

Assumption 2. *Let $\mathcal{F}_t := \sigma((x_0, u_0), (x_1, u_1), \cdots, (x_t, u_t))$ denote the history of states and inputs until time t. The state noise $\{w_t\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, with $\mathbb{E}[w_{t+1} w_{t+1}^T | \mathcal{F}_t] = I_n$, and is element-wise sub-Gaussian, i.e., for any $\gamma \in \mathbb{R}$, there exists $L > 0$ with $\mathbb{E}\left[exp(\gamma w_{t,j}) | \mathcal{F}_t\right] \leq exp\left(\frac{\gamma^2 L^2}{2}\right)$ $\forall\, j$ and $\forall\, t$.*

We assume that $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times m}$ are known positive semi-definite and positive definite matrices. Define $z_t^T := (x_t^T, u_t^T)$. Then, the linear stochastic system can equivalently be written as

$$x_{t+1} = \theta^{\star T} z_t + w_{t+1}. \tag{8}$$

For a matrix M, we use $\|M\|$ to denote its operator norm induced from the ℓ_2-norm.

3.1 Controller Design for a Known LQG System

For a discrete time linear system, $x_{t+1} = \theta^T z_t + w_{t+1}$, where $\theta = (A, B)^T \in \mathcal{S}$, there is an unique positive semidefinite matrix $P(\theta)$ that satisfies the Riccati equation (see [6]) $P(\theta) = Q + A^T P(\theta) A - A^T P(\theta) B (B^T P(\theta) B + R)^{-1} B^T P(\theta) A$. The optimal control law which minimizes the long-term average quadratic cost is $u_t = K(\theta) x_t$, where the "gain matrix" is $K(\theta) := -(B^T P(\theta) B + R)^{-1} B^T P(\theta) A$. The optimal average cost, $J^\star(\theta)$ is equal to $trace(P(\theta))$. As a consequence of Assumption 1, the parameter set $\mathcal{S}$ is bounded, hence one can show that $P(\theta)$ is bounded as well:

$$D := \sup_{\theta \in \mathcal{S}} ||P(\theta)|| < \infty. \tag{9}$$

Assumption 3.

$$\rho := \sup_{\theta \in \mathcal{S}} ||A + BK(\theta)|| < 1, \textit{ and} \tag{10}$$

$$c_0 := \sup_{\theta \in \mathcal{S}} ||K(\theta)|| < +\infty. \tag{11}$$

3.2 Construction of Confidence Interval

The squared fitting error of θ is defined as $V_t(\theta) := \sum_{s=0}^{t-1} ||(x_{s+1} - \theta^T z_s||^2$. The ℓ_2-regularized squared fitting error with parameter $\lambda > 0$ is $E_t(\theta) = \lambda ||\theta||_2^2 + V_t(\theta)$. Let $\hat{\theta}_t$ be the ℓ_2 regularized least-squares estimate of $\theta^\star$, i.e., $\hat{\theta}_t \in \arg\min_{\theta \in \mathcal{S}} E_t(\theta)$. Next, given the history $\mathcal{F}_t$, we construct a "high-probability confidence ball," $\mathcal{C}_t(\delta)$ around $\hat{\theta}_t$, i.e., a set of plausible system parameters that contains the true parameter $\theta^\star$ with a high probability. Let $\mathcal{C}_t(\delta) := \left\{\theta : trace\left((\theta - \hat{\theta}_t)^T Z_t (\theta - \hat{\theta}_t)\right) \leq \beta_t(\delta)\right\}$, where $Z_t := \lambda I_{n+m} + \sum_{s=0}^{t} z_s z_s^T$, and

$$\beta_t(\delta) := \left(nL \sqrt{2 \log\left(\frac{\sqrt{det(Z_t) det(\lambda I)}}{\delta}\right)} + \sqrt{\lambda} c \right)^2. \tag{12}$$

Define

$$E_1(t) = \{\theta^\star \in \mathcal{C}_s(\delta/4),\ \forall\, s = 1, 2, \cdots, t\} \text{ and } E_2(t) = \{||x_s|| \leq d_t,\ \forall s \leq t\}, \tag{13}$$

where d_t is defined in the Appendix. Let $E_1 := E_1(T)$ and $E_2 := E_2(T)$. Then, as in [3], $\mathbb{P}(E_1 \cap E_2) \geq 1 - \frac{\delta}{2}$. Moreover, on the event $E_2(t)$ defined in (13), the following holds,

$$\max_{1 \leq s \leq t} \|x_s\| \leq X_t = Y_t^{n+m+1}\ \forall t = 1, 2, \ldots, \text{where,} \tag{14}$$

$$Y_t := \max\left(e, \lambda(n+d)(e-1) 4 \left(c_1 \log\left(\frac{1}{\delta}\right) + c_2 \log\left(\frac{t}{\delta}\right)\right)\right) \times \left(\log^2\left(4c_1 \log(\frac{1}{\delta}) + 4c_2 \log(\frac{t}{\delta})\right)\right),$$

with c_1 and c_2 being problem dependant constants.

Algorithm 1 Augmented RBMLE-UCB (ARBMLE)

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**
    solve the following optimization to obtain θ_{t_k},

$$\theta_t \in \arg\min_{\theta \in \mathcal{S} \cap \mathcal{C}_{t_k}(\delta)} \{V_t(\theta) + \alpha(t) J^\star(\theta)\},$$

  else
    $\theta_t = \theta_{t-1}$
  end if
  $u_t = K(\theta_t) x_t$
  $Z_{t+1} = Z_t + z_s z_s^T$
end for

4 The Augmented RBMLE-UCB Algorithm

We employ a version of the Augmented RBMLE-UCB (ARBMLE) algorithm that proceeds in an episodic manner. Let t_k denote the starting time of the k-th episode. Then during episode k, it implements the control policy, $u_t = K(\theta_{t_k}) x_t \; \forall \, t \in [t_k, t_{k+1})$. where, θ_{t_k} is given by solving the following optimization problem.

$$\theta_{t_k} \in \arg\min_{\theta \in \mathcal{S} \cap \mathcal{C}_{t_k}(\delta)} \{V_t(\theta) + \alpha(t_k) J^\star(\theta)\}, \tag{15}$$

where the bias-term, $\alpha(t) = \alpha_0 \sqrt{T}, \forall t$, for $\alpha_0 > 0$.

5 Regret Analysis

In this section, we provide an upper bound on the regret $R(T)$. In Lemma 5.1, we decompose the regret $R(T)$ into easy-to-analyze terms. Then, we provide upper bounds on each of these four terms, which allows us to get a upper bound on the regret.

Lemma 5.1. [3] *The regret of the Augmented RBMLE-UCB learning algorithm can be decomposed as $R(T) = R_1 + R_2 + R_3 + R_4$, where*

$$\begin{aligned}
R_1 &:= \sum_{t=0}^{T} \left\{ x_t^T P(\theta_t) x_t - \mathbb{E}\left[x_{t+1}^T P(\theta_{t+1}) x_{t+1} | \mathcal{F}_t \right] \right\}, \\
R_2 &:= \sum_{t=0}^{T} \mathbb{E}\big[x_t^T \left(P(\theta_{t+1}) - P(\theta_t) \right) x_t | \mathcal{F}_t \big], \\
R_3 &:= \sum_{t=0}^{T} \left\{ \left(A^\star x_t + B^\star u_t \right)^T P(\theta_t) \left(A^\star x_t + B^\star u_t \right) - \left(A_t x_t + B_t u_t \right)^T P(\theta_t) \left(A_t x_t + B_t u_t \right) \right\}, \\
R_4 &:= \sum_{t=0}^{T} \left(J(\theta_t) - J(\theta^\star) \right).
\end{aligned} \tag{16}$$

We bound R_1, R_2, R_3 and R_4 on the event $E_1 \cap E_2$.

Lemma 5.2. *On the event $E_1 \cap E_2$, $R_1 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n\sqrt{B_\delta}$, 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* (9).

Lemma 5.3. *On the event $E_1 \cap E_2$, $R_2 \leq 2DX_T^2 \log_2 T$, where X_T is defined in* (14).

[3] All proofs are presented in the Appendix.

Lemma 5.4. *On the event* $E_1 \cap E_2$, $R_3 \leq \frac{8cX_T^2 D(1+c_0^2)}{\sqrt{\lambda}} \sqrt{T\beta_T\left(\frac{\delta}{8}\right) \log \frac{det(Z_T)}{det(\lambda I)}}$, *where* $Z_T := \max_{0\leq t\leq T} \|z_t\|$, *and* $\beta_T\left(\frac{\delta}{8}\right)$ *is defined in* (12).

Lemma 5.5. *On the event* $E_1 \cap E_2$, $R_4 \leq \frac{1}{\alpha_0}\left(\beta_T\left(\frac{\delta}{4}\right) + \lambda c^2\right)\sqrt{T}$.

We now use these bounds in conjunction with Lemma 5.1 to bound the regret of Augmented RBMLE-UCB.

Theorem 5.6. *For any* $\delta \in (0,1)$ *and* $T > 0$, *the regret of the RBMLE Algorithm is upper-bounded by* $R(T) \leq \tilde{\mathcal{O}}\left(\sqrt{T \log \frac{1}{\delta}}\right)$ *with a probability at least* $(1-\delta)$.

6 Empirical Performance

We evaluate the empirical performance of ARBMLE as well as standard (unaugmented) RBMLE. We compare these algorithms with OFU [3], Thompson Sampling (TS) [14], Input Perturbations (IE) [15], Randomized Certainty Equivalence (RCE) [17], and Stabl [16]. The results shown here are for the following examples of linear system that have appeared in the recent literature on adaptive control of linear systems:

1. Longitudinal Flight Control of Boeing 747 [16].
2. Unmanned Aerial Vehicle (UAV) [20, 16]
3. Unstable Laplacian dynamics [18, 16, 19].
4. Large transient dynamics [18].

The details of these examples are provided in the Appendix.

Each simulation experiment is performed for a time horizon of 500, and repeated 100 times. The reported results are the averaged values over the 100 runs. In Figure 1, we compare the $\log_{10}$ of the averaged regret of ARBMLE, OFU, TS and Stabl. In Figure 2, we plot the averaged regret of RBMLE, ARBMLE, IP and RCE. We summarize the results using the regret values at $T = 500$, averaged over the 100 runs,in Table 1. More results for more examples are provided in the Appendix. The details of the implementations can be found in the Appendix.

We highlight the key observations from above experiments:

- ARBMLE and RBMLE were always found to have the same empirical performance. More specifically, both these algorithms choose the same estimate θ_t, and this suggests that $\frac{1}{\alpha(t)}$ needs to be greater than the Lagrange multiplier for the ball constraint $\theta \in \mathcal{C}_{t_k}(\delta)$. This also motivates future study of the regret of the (unaugmented) RBMLE.
- As can be seen in Figure 1, ARBMLE/RBMLE outperform OFU and TS by a huge margin. We conjecture that this is due to temporary instability when OFU, TS are employed [16]. ARBMLE/ RBMLE also has signficantly lower regret compared to StabL.
- Figure 2 shows that ARBMLE/RBMLE also outperforms RCE moderately and IP marginally.

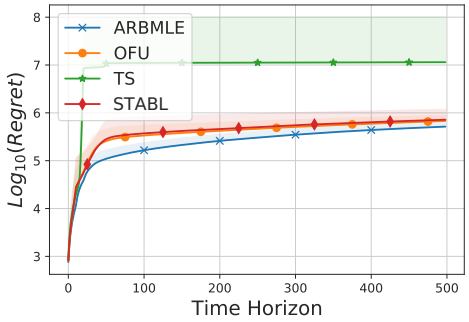
Remark 1. *Our simulations for ARBMLE, OFU and TS are based on the confidence interval* $\beta_t(\delta)$ *as defined in* (12). *Instead, recent works [16, 18] use* $\beta_t(\delta) := trace\left((\theta^\star - \hat{\theta}_t)^T Z_t (\theta^\star - \hat{\theta}_t)\right)$. *However, one may note that* $\theta^\star$ *in* $\beta_t(\delta)$ *is not known to the learning agent, and so such a definition of* $\beta_t(\delta)$ *is not a viable implementation. The effect of the choice of* $\beta_t(\delta)$ *on the regret performance is shown in the appendix.*

7 Concluding Remarks

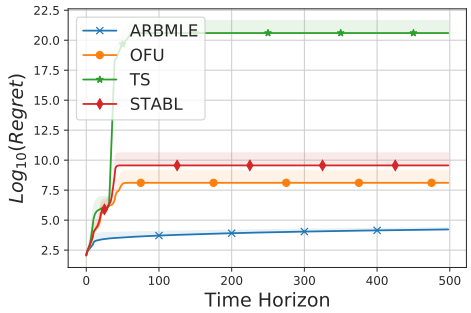
We reconcile the RBMLE and UCB approaches by showing that RBMLE is an unconstrained penalty version of the constrained optimization problem of UCB. Showing that OFU is an constrained version of RBMLE also explains why the optimism embodied in UCB-based schemes is

Ex.	RBMLE	ARBMLE	OFU	TS	IP	RCE	STABL
(a)	515668	515668	678521	1.1×10^7	525384	1.1×10^{14}	718608
(b)	16755	16755	1.2×10^8	3.9×10^{20}	16793	131055	3.6×10^9
(c)	3237	3237	8574	4077	3253	3420	5945
(d)	5951	5951	1.98×10^8	25211	5971	6468	390022

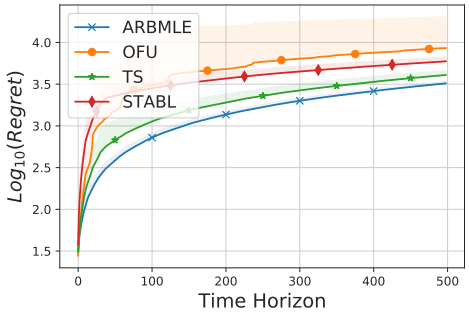
Table 1: Average Regret Performance at $T = 500$.



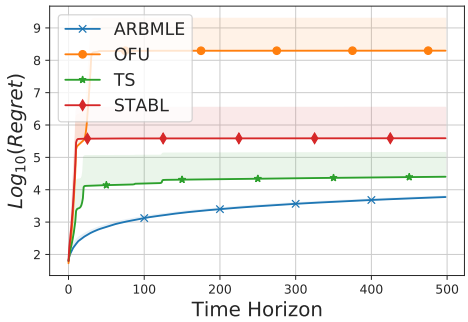
(a) Boeing 747



(b) UAV



(c) Unstable Laplacian



(d) Large Transient Dynamics

Figure 1: Logarithm of the Averaged Regret over 100 runs of RBMLE, ARBMLE, OFU and TS for various example systems.

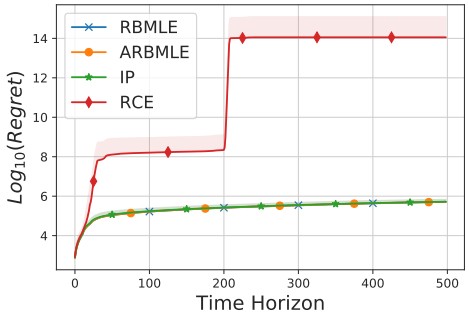
justified. In particular, it is justified by the goal of nullifying the one-sided bias that results from the closed-loop identification of dynamics.

Building on this, we propose an Augmented RBMLE-UCB method that not only matches the best known order of regret, $\tilde{\mathcal{O}}(\sqrt{T})$, to date – that of the OFU algorithm – but also outperforms OFU, TS and Stabl by a significant margin in simulation experiments. In fact, for the UAV experiment, the regret of OFU is $\approx 10^4$, TS $\approx 10^{16}$ and StabL $\approx 10^5$ times that of RBMLE. It outperforms RCE also, but the gains are moderate, while with respect to IP the gains are smaller. The simulations were carried out on real-world examples such as Boeing 747, UAV, Unstable Laplacian and Large Transient Dynamics which were taken from recent works [18, 55, 19, 16, 20] on adaptive LQG control.

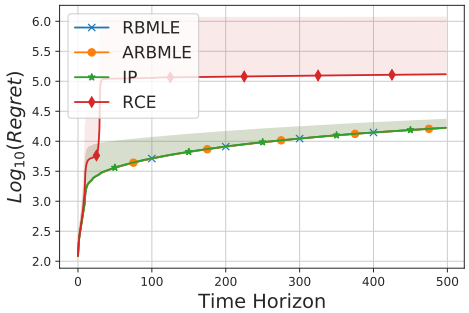
This work further extends recent studies of RBMLE for MDPs [31], stochastic Multi-Armed Bandits [29] and Contextual bandits [30] establishing state-of-art regret results as well as empirically good performance.

There remain several open questions.

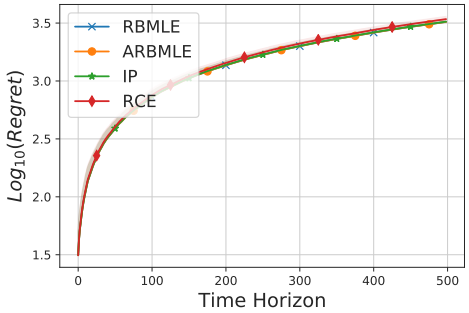
Currently we study only an augmented form of RBMLE, i.e. ARBMLE, in which the agent searches for a parameter value that optimizes a certain reward-biased maximum likelihood objective function within a "high-confidence ball." However, (unaugmented) RBMLE optimizes this over the larger set $\mathcal{S}$ (Assumption 1) that is known to contain the true parameter with high probability. The reason



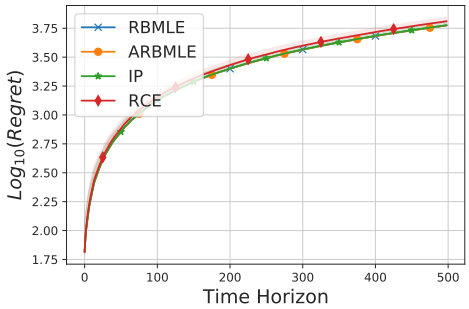
(a) Boeing 747



(b) UAV



(c) Unstable Laplacian



(d) Large Transient Dynamics

Figure 2: Logarithm of the Averaged Regret over 100 runs of RBMLE, ARBMLE, IP and RCE for various example systems. Table 1 provides the quantative results for a finer comparative evaluation of the regrets.

for this augmentation is simply that currently we are unable to prove regret bounds for RBMLE without it. By including it, we can capitalize on the nice technical results in [3] in the analysis of OFU. Simulations show that the constraint is loose, and performance is the same with/without the constraint. In fact, as shown by the simulations, the choices of θ_{t_k} made by the standard RBMLE, and the Augmented RBMLE are the same. It remains to be seen whether similar regret guarantees can be obtained by the (unaugmented) RBMLE, which remains an open problem.

One may note that ARBMLE and OFU differ in their choices of decisions. Denoting their estimates of the unknown parameter at a given time by θ_{ARBMLE} and θ_{OFU} respectively, θ_{OFU} always lies at the boundary of the UCB-ball constraint. This is easily seen to be true for bandits and MDPs, and empirically observed to be so for LQ systems. In contrast, θ_{ARBMLE} is most often in the interior of the UCB-ball. Moreover, while OFU treats all models within the UCB-ball equally and only assesses them by their cost, ARBMLE prefers models that are closer to the Least Squares Estimate (θ_{LSE}). For example, if $J^\star(\theta_1) = J^\star(\theta_2)$, then ARBMLE prefers the θ_i that is closer to θ_{LSE}. We conjecture that this is the reason ARBMLE has significantly better performance than OFU – but we are unable to prove it.

It has been shown in previous works on adaptive LQG control [10, 11] that under appropriate formulations RBMLE is guaranteed to stabilize an unknown LQG system over an infinite horizon, i.e., $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\|x(t)\|^2 + \|u(t)\|^2) < \infty$ a.s.. Morover, the sample path performance cost $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(x^T(t)Qx(t) + u^T(t)Ru(t))$ is (a.s.) equal to the optimal performance that could be attained if the system parameters $(A^\star, B^\star)$ were known. In fact, stability is a prerequisite before one can establish the latter type of result, as the decades of work on stochastic adaptive control in the nineteen seventies to the nineties has shown. While recent work has tended to study the finer performance measure of regret, it is usually over a finite horizon, and more attention to the stability of the learning process over an infinite horizon appears well deserved. Similarly, robustness which also was rigorously formulated and addressed in earlier decades earlier needs to re-examined.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] P. R. Kumar and A Becker. A new family of optimal adaptive controllers for markov chains. IEEE Transactions on Automatic Control, 27(1):137–146, 1982.

[2] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1):4–22, 1985.

[3] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In Proceedings of the 24th Annual Conference on Learning Theory, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[4] Karl J Åström and Björn Wittenmark. Adaptive control. Courier Corporation, 2013.

[5] P. R. Kumar. A survey of some results in stochastic adaptive control. SIAM Journal on Control and Optimization, 23(3):329–380, 1985.

[6] P. R. Kumar and Pravin Varaiya. Stochastic systems: Estimation, identification, and adaptive control. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[7] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[8] Han-Fu Chen and Lei Guo. Identification and stochastic adaptive control, volume 5. Springer Science & Business Media, 1991.

[9] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.

[10] P. R. Kumar. Optimal adaptive control of linear-quadratic-gaussian systems. SIAM Journal on Control and Optimization, 21(2):163–178, 1983.

[11] Marco C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. SIAM Journal on Control and Optimization, 36(6):1890–1907, 1998.

[12] Maria Prandini and Marco C Campi. Adaptive lqg control of input-output systems—a cost-biased approach. SIAM Journal on Control and Optimization, 39(5):1499–1519, 2000.

[13] Sergio Bittanti, Marco C Campi, et al. Adaptive control of linear time invariant systems: the "bet on the best" principle. Communications in Information & Systems, 6(4):299–320, 2006.

[14] Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In Artificial Intelligence and Statistics, pages 1246–1254. PMLR, 2017.

[15] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. arXiv preprint arXiv:1811.04258, 2018.

[16] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 5354–5390. PMLR, 28–30 Mar 2022.

[17] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear-quadratic regulators, 2018.

[18] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. arXiv preprint arXiv:1805.09388, 2018.

[19] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 3108–3117. PMLR, 2019.

[20] Feiran Zhao, Keyou You, and Tamer Basar. Infinite-horizon risk-constrained linear quadratic regulator with average cost, 2021.

[21] Marco C Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. SIAM Journal on Control and Optimization, 36(6):1890–1907, 1998.

[22] P. R. Kumar. Adaptive control with a compact parameter set. SIAM Journal on Control and Optimization, 20(1):9–13, 1982.

[23] P. R. Kumar and Woei Lin. Optimal adaptive controllers for unknown markov chains. IEEE Transactions on Automatic Control, 27(4):765–774, 1982.

[24] P. R. Kumar. Simultaneous identification and adaptive control of unknown systems over finite parameter sets. IEEE Transactions on Automatic Control, 28(1):68–76, 1983.

[25] VS Borkar. The Kumar-Becker-Lin scheme revisited. Journal of optimization theory and applications, 66(2):289–309, 1990.

[26] Łukasz Stettner. On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model. Applied Mathematics and Optimization, 27(2):161–177, 1993.

[27] VS Borkar. Self-tuning control of diffusions without the identifiability condition. Journal of optimization theory and applications, 68(1):117–138, 1991.

[28] TE Duncan, B Pasik-Duncan, and L Stettner. Almost self-optimizing strategies for the adaptive control of diffusion processes. Journal of optimization theory and applications, 81(3):479–507, 1994.

[29] Xi Liu, Ping-Chun Hsieh, Yu Heng Hung, Anirban Bhattacharya, and P. R. Kumar. Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits. In International Conference on Machine Learning, pages 6248–6258. PMLR, 2020.

[30] Yu-Heng Hung, Ping-Chun Hsieh, Xi Liu, and P. R. Kumar. Reward-biased maximum likelihood estimation for linear stochastic bandits. arXiv preprint arXiv:2010.04091, 2020.

[31] Akshay Mete, Rahul Singh, Xi Liu, and P. R. Kumar. Reward biased maximum likelihood estimation for reinforcement learning. In Learning for Dynamics and Control, pages 815–827. PMLR, 2021.

[32] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2):235–256, 2002.

[33] Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. The Annals of Statistics, 10(1):154–166, 1982.

[34] Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. SIAM Journal on Control and Optimization, 25(2):466–481, 1987.

[35] Han-Fu Chen and Lei Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. SIAM Journal on Control and Optimization, 25(4):845–867, 1987.

[36] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In NIPS, pages 2645–2653, 2012.

[37] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In UAI, pages 1–11. Citeseer, 2015.

[38] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with thompson sampling. In 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 1198–1205. IEEE, 2017.

[39] Herbert Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

[40] Aleksandr Aronovich Feldbaum. Dual control theory. i. Avtomatika i Telemekhanika, 21(9):1240–1249, 1960.

[41] Björn Wittenmark. Adaptive dual control methods: An overview. Adaptive Systems in Control and Signal Processing 1995, pages 67–72, 1995.

[42] Vivek Borkar and P. Varaiya. Adaptive control of markov chains, i: Finite parameter set. IEEE Transactions on Automatic Control, 24(6):953–957, 1979.

[43] Donald A Berry and Bert Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). London: Chapman and Hall, 5(71-87):7–7, 1985.

[44] John Gittins, Kevin Glazebrook, and Richard Weber. Multi-armed bandit allocation indices. John Wiley & Sons, 2011.

[45] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

[46] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[47] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995, 2009.

[48] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In Advances in neural information processing systems, pages 49–56, 2007.

[49] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(4), 2010.

[50] Magnus R Hestenes. Multiplier and gradient methods. Journal of optimization theory and applications, 4(5):303–320, 1969.

[51] Michael JD Powell. A method for nonlinear constraints in minimization problems. Optimization, pages 283–298, 1969.

[52] Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. PhD thesis, 2013.

[53] Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. Optimal control. John Wiley & Sons, 2012.

[54] Thomas Kailath. Linear systems, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

[55] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. Foundations of Computational Mathematics, 20(4):633–679, 2020.