Linear Stochastic Approximation: Constant Step-Size and Iterate Averaging

Abstract

We consider d-dimensional *linear inversion* problems (arising in machine learning) where the aim is to compute a $\theta_* \in \mathcal{R}^d$ such that $\theta_* = A^{-1}b$ using noisy samples of $A \in \mathcal{R}^{d\times d}$ and $b \in \mathcal{R}^d$. Linear stochastic approximation (LSA) is a widely used approach to solve such problems: the stochastic gradient descent and the temporal difference class of learning algorithms (such as TD(0) or GTD) for approximate value function estimation in reinforcement learning are LSA algorithms. An important parameter that affects the performance of LSA is the step-size or learning rate. In this paper, we look at a constant step-size averaged linear stochastic approximation approach (CALSA) to solve the linear inversion problem, ask whether a uniformly fast rate of $O(\frac{1}{t})$ is achievable for the mean square-error across all problem instances. We show that the answer to this question, in general, is *no*. However, we show that instance dependent finite-time rate of $O(\frac{1}{t})$ is achievable and for some interesting problem classes in reinforcement learning the constant step-size can be chosen in a problem independent manner.

1 Introduction

Many interesting machine learning problems such as linear least squares regression or approximate value function estimation (arising in RL) are *linear inversion* problems. Here, the aim is to compute $\theta_* = A^{-1}b$ using noisy samples of A and b. Typically, low iteration-cost incremental learning algorithms (e.g., variations of gradient descent) have been widely applied to such problems. A widely used incremental approach is linear stochastic approximation (LSAs in short) which can be expressed in the following general form:

$$\theta_t = \theta_{t-1} + \alpha_t(b_t - A_t\theta_{t-1}), \tag{1}$$

with $(\alpha_t)_t$ a positive step-size sequence chosen by the user and $(b_t, A_t) \in \mathcal{R}^d \times \mathcal{R}^{d\times d}$, $t \geq 0$, a sequence of identically distributed random variables such that $\mathbf{E}[b_t] = b$ and $\mathbf{E}[A_t] = A$. Some examples of LSA approaches include the stochastic gradient descent algorithm (SGD) for the problem of linear least-squares estimation (LSE) [**?** **?**], and the *temporal difference* (TD) class of learning algorithms in RL [**?** **?** **?** **?** **?** **?**]. A critical aspect of the design of these methods is the choice of learning rates i.e., the step-size sequence $(\alpha_t)_{t\geq 0}$: poor choices lead to slow convergence, or instability. In particular, learning rates can degrade, or the rate may depend on problem dependent constants, which can be very large. A useful choice has been the diminishing step-sizes [**?** **?** **?**], where $\alpha_t \to 0$ such that $\sum_{t\geq 0}\alpha_t = \infty$ and $\sum_{t\geq 0}\alpha_t^2 < \infty$. Here, $\alpha_t \to 0$ circumvents the need for guessing the magnitude of step-sizes that stabilize the updates, while the condition $\sum_{t\geq 0}\alpha_t = \infty$ ensures that initial conditions are forgotten and the condition $\sum_{t\geq 0}\alpha_t^2 < \infty$ ensures that the variance eventually goes to zero. An alternate idea, which we call constant-step size averaged LSA (CALSA) is to run (**??**) by choosing $\alpha_t = \alpha > 0$ $\forall t \geq 0$ with some $\alpha > 0$, and output the average $\hat{\theta}_t = \frac{1}{t+1}\sum_{i=0}^{t}\theta_i$. Thus, in CALSA, θ_t is an internal variable and $\hat{\theta}_t$ is the output of the algorithm (see **??** for a formal definition of CALSA). The idea is that the constant step-size leads to faster forgetting of initial conditions, while the averaging on the top reduces noise. This idea goes back to **?**] and **?**] who considered it in the context of stochastic approximation that LSA is a special case of.

Motivation: Recently, **?**] considered, what we call a constant step-size averaged stochastic gradient descent (CASGD)[1] for linear least squares regression problem (with *i.i.d.* sampling) and showed the following interesting results hold together: (*i*) there exists an *universal* constant step-size choice that can be calculated from only the knowledge of the bound on the magnitude of the noisy data; (*ii*) and the leading term

Preliminary work. Under review by AISTATS 2018. Do not distribute.

[1] SGD is an LSA of the form in (**??**).

as $t \to \infty$ in CASGD's (with the said constant step-size) mean-squared prediction error after t updates is at most $\frac{C}{t}$; where the constant $C > 0$ in the rate expression depends *only* on the bound on the data, the dimension d and is in particular independent of the eigen spectrum of $\mathbf{E}[A_t]$, a property which is not shared by other step-size tunings and variations of the basic SGD method. In simple terms, this means that there exists a constant universal step-size and uniform rates (that are independent of problem instance).

Focus: We are interested in the setting of *policy evaluation* [**?**] using linear value function approximation from experience replay [**?**] in a batch setting [**?**] in RL using the TD class of algorithms [**? ? ? ? ?**]. Here, data is presented in the form of a sequence $D_t = (\phi_t, \phi'_t, r_t)$ (A_t and b_t are appropriate functions of D_t), where $\phi_t, \phi'_t \in \mathcal{R}^d$ are the *linear features* of successive states visited. Our aim here is to repeat the feat of [**?**], in that we want to design TD algorithms in a problem independent fashion. We want to explore whether it is possible to have universal step-size and uniform rates for TD algorithms applied to RL problems.

Setup: In this paper, we study general CALSA (thereby extending the scope of prior work by **?**] from CASGD to general CALSAs thereby covering TD algorithms as well). Our restrictions on the common distribution is that the "noise variance" should be bounded (as we consider squared errors), and that the matrix $\mathbf{E}[A_t]$ must be Hurwitz, i.e., all its eigenvalues have positive real parts.

Contributions: The important results are as under:

1. **Finite-time Instance Dependent Bounds** (**??**): For a given linear inversion problem P, we measure the performance of CALSA (that solves P) in terms of the mean square error (MSE) given by $\mathbf{E}_P\left[\|\hat{\theta}_t - \theta_*\|^2\right]$. For the first time in the literature, we show that (under our stated assumptions) there exists an $\alpha_P > 0$ such that for any $\alpha \in (0, \alpha_P)$, the MSE is at most $\frac{C_{P,\alpha}}{t} + \frac{C_{P',\alpha}}{t^2}$ with some positive constants $C_{P,\alpha}, C_{P',\alpha}$ that we explicitly compute from P.

2. **Negative Results:** We show that *i*) in general in it not possible to achieve a uniformly fast finite-time rate of $O(\frac{1}{t})$ that holds for all problem P chosen form a given problem class $\mathcal{P}$; *ii*) in general we cannot choose a constant step-size that is universal across a given class of problem $\mathcal{P}$.

3. **Reinforcement Learning:** We show universality of step-size in the following interesting scenarios: *i*) we consider the class of problem with *second order feature stationarity* where $\mathbf{E}\phi_t\phi_t^\top = \mathbf{E}\phi'_t\phi'_t\top$. CA-TD(0) for class of RL problems where there is ; *ii*) CA-GTD for any class of RL problems.

2 Notations and Definitions

We denote the sets of real and complex numbers by $\mathcal{R}$ and $\mathbb{C}$, respectively. For $x \in \mathbb{C}$ we denote its modulus and complex conjugate by $|x|$ and $\bar{x}$, respectively. We denote d-dimensional vector spaces over $\mathcal{R}$ and $\mathbb{C}$ by $\mathcal{R}^d$ and $\mathbb{C}^d$, respectively, and use $\mathcal{R}^{d\times d}$ and $\mathbb{C}^{d\times d}$ to denote $d \times d$ matrices with real and complex entries, respectively. We denote the transpose of C by $C^\top$ and the conjugate transpose by $C^* = \bar{C}^\top$ (and of course the same notation applies to vectors, as well). We will use $\langle \cdot, \cdot \rangle$ to denote the inner products: $\langle x, y \rangle = x^* y$. We use $\|x\| = \langle x, x \rangle^{1/2}$ to denote the 2-norm. For $x \in \mathbb{C}^d$, we denote the general quadratic norm with respect to a positive definite (see below) Hermitian matrix C (i.e., $C = C^*$) by $\|x\|_C^2 \doteq x^* C x$. The norm of the matrix A is given by $\|A\| \doteq \sup_{x\in\mathbb{C}^d:\|x\|=1} \|Ax\|$. We use $\kappa(A) = \|A\|\|A^{-1}\|$ to denote the condition number of matrix A. We denote the spectral radius of a matrix A by $\Lambda(A) = \{\max_i |\lambda_i| : \lambda_i \text{ is an eigenvalue of } A\}$. We denote the identity matrix in $\mathbb{C}^{d\times d}$ by $\mathcal{I}$ and the set of invertible $d\times d$ complex matrices by $\mathrm{GL}(d)$. For a positive real number $B > 0$, we define $\mathbb{C}_B^d = \{b \in \mathbb{C}^d \mid \|b\| \leq B\}$ and $\mathbb{C}_B^{d\times d} = \{A \in \mathbb{C}^{d\times d} \mid \|A\| \leq B\}$ to be the balls in $\mathbb{C}^d$ and $\mathbb{C}^{d\times d}$, respectively, of radius B. We use $Z \sim P$ to denote the fact that Z (which can be a number, or vector, or matrix) is distributed according to probability distribution P; $\mathbf{E}$ denotes mathematical expectation.

Let us now state some definitions that will be useful for presenting our main results.

Definition 1. For a probability distribution P over $\mathbb{C}^d \times \mathbb{C}^{d\times d}$, we let P^V and P^M denote the respective marginals of P over $\mathbb{C}^d$ and $\mathbb{C}^{d\times d}$. By *abusing notation* we will often write $P = (P^V, P^M)$ to mean that P is a distribution with the given marginals. Define $A_P = \int M\, dP^M(M), C_P = \int M^* M\, dP^M(M), b_P = \int v\, dP^V(v)\,, \rho_d(\alpha, P) \doteq \inf_{x\in\mathbb{C}^d:\ \|x\|=1} \langle x, ((A_P + A_P^*) - \alpha A_P^* A_P)\, x\rangle, \rho_s(\alpha, P) \doteq \inf_{x\in\mathbb{C}^d:\ \|x\|=1} \langle x, ((A_P + A_P^*) - \alpha C_P)\, x\rangle\,.$

Note that $\rho_d(\alpha, P) \geq \rho_s(\alpha, P)$. Here, subscripts s and d stand for *stochastic* and *deterministic* respectively.

Definition 2. Let $P = (P^V, P^M)$ as in **??**; $b \sim P^V$ and $A \sim \mathcal{P}^M$ be random variables distributed according to P^V and P^M. For $U \in \mathrm{GL}(d)$ define P_U to be the distribution of $(U^{-1}b, U^{-1}AU)$. We also let (P_U^V, P_U^M) denote the corresponding marginals.

Definition 3. We call a matrix $A \in \mathbb{C}^{d\times d}$ *Hurwitz*

(H) if all eigenvalues of A have positive real parts. We call a matrix $A \in \mathbb{C}^{d\times d}$ *positive definite* (PD) if $\langle x, Ax\rangle > 0, \forall x \neq 0 \in \mathbb{C}^d$. If $\inf_x \langle x, Ax\rangle \geq 0$ then A is *positive semi-definite* (PSD). We call a matrix $A \in \mathcal{R}^{d\times d}$ to be *symmetric positive definite* (SPD) is it is symmetric i.e., $A^\top = A$ and PD.

Note that SPD implies that the underlying matrix is real.

Definition 4. We call the distribution P in **??** to be H/PD/SPD if A_P is H/PD/SPD.

Though $\rho_s(\alpha, P)$ and $\rho_d(\alpha, P)$ depend only on P^M, we use P instead of P^M to avoid notational clutter.

Example 1. The matrices $\begin{bmatrix}0.1 & -1\\ 1 & 0.1\end{bmatrix}$, $\begin{bmatrix}0.1 & 0.1\\ 0 & 0.1\end{bmatrix}$ and $\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}$ are examples of H, PD and SPD matrices, respectively, and they show that while SPD implies PD, which implies H, the reverse implications do not hold.

3 Problem Setup

We are interested in what we call as *linear inversion* problems where the aim is to compute a $\theta_* \in \mathcal{R}^d$ such that

$$\theta_* = A^{-1}b, \tag{2}$$

where $A \in \mathcal{R}^{d\times d}$ and $b \in \mathcal{R}^d$. The interesting case, is when A and b in (**??**) are available only as noisy samples. In this paper, we study what we call the *constant step-size averaged linear stochastic approximation* algorithm (CALSA) to solve (**??**), and is given by

$$\text{LSA:} \qquad \theta_t = \theta_{t-1} + \alpha(b_t - A_t\theta_{t-1})\,, \tag{3a}$$

$$\text{PR-Average:} \qquad \hat{\theta}_t = \frac{1}{t+1}\sum\nolimits_{s=0}^{t} \theta_i\,. \tag{3b}$$

The algorithm updates a pair of parameters $\theta_t, \bar{\theta}_t \in \mathcal{R}^d$ incrementally, in discrete time steps $t = 1, 2, \ldots$ based on data $b_t \in \mathcal{R}^d$, $A_t \in \mathcal{R}^{d\times d}$. Here $\alpha > 0$ is a positive step-size parameter; an input to the algorithm besides the initial value θ_0. The iterate θ_t is treated as an internal state of the algorithm, while $\hat{\theta}_t$ is the output at time step t. The update of θ_t alone is considered a form of constant step-size LSA. Sometimes A_t will have a special form and then the matrix-vector product $A_t\theta_{t-1}$ can also be computed in $O(d)$ time, a scenario common in reinforcement learning[**?** **?** **?** **?** **?**]. This makes the algorithm particularly attractive in large-scale computations when d is in the range of thousands, or millions, or more, as may be required by modern applications (e.g., [**?**])

In what follows, for $t \geq 1$ we make use of the σ-fields $\mathcal{F}_{t-1} \doteq \sigma\{\theta_0, A_1, \ldots, A_{t-1}, b_1, \ldots, b_{t-1}\}$; $\mathcal{F}_{-1}$ is the trivial σ algebra, and P denotes a single problem instance (which the CALSA solves) and $\mathcal{P}$ is a class of problems. Note that a given problem instance is characterized by the distribution $P = (P^V, P^M)$ (see **??**).We are interested in the behaviour of (**??**) under the following assumption:

Assumption 1.

1. $(b_t, A_t) \sim P$, $t \geq 0$ is an *i.i.d.* sequence. We let A_P be the expectation of A_t, b_P be the expectation of b_t, as in **??**. We assume that P is Hurwitz.
2. The martingale difference sequences[2] $M_t \doteq A_t - A_P$ and $N_t \doteq b_t - b_P$ associated with A_t and b_t satisfy the following

   $$\mathbf{E}_P\left[\|M_t\|^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma^2_{A_P}\,, \mathbf{E}_P\left[\|N_t\|^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma^2_{b_P}\,.$$

   with some $\sigma^2_{A_P}$ and $\sigma^2_{b_P}$. Further, we assume $\mathbf{E}_P[M_t N_t] = 0$
3. A_P is invertible and thus the vector $\theta_* = A_P^{-1}b_P$ is well-defined.

Performance Metric: We are interested in the behavior of the mean squared error (MSE) at time t given by $\mathbf{E}\left[\|\hat{\theta}_t - \theta_*\|^2\right]$. More generally, one can be interested in $\mathbf{E}_P\left[\|\hat{\theta}_t - \theta_*\|^2_C\right]$, the MSE with respect to a PD Hermitian matrix C. Since in general it is not possible to exploit the presence of C unless it is connected to P in a special way (see **??**), here we restrict ourselves to $C = \mathcal{I}$. In what follows, for the sake of brevity, we drop the subscript P in the quantities $\mathbf{E}_P[\cdot]$, $\sigma^2_{A_P}$ and $\sigma^2_{b_P}$.

We now present two important *linear inversion* problems in machine learning.

Example 2 (Linear Least Squares Regression)**.** Let $(x_t, y_t) \in \mathcal{R}^d \times \mathcal{R}, t \geq 0$ be *i.i.d.* such that $\mathbf{E}\|x_t\|^2$ and $\mathbf{E}[y_t^2]$ are finite. The linear least-square minimization problem is the minimization of

$$f(\theta) = \frac{1}{2}\mathbf{E}(\langle x_t, \theta\rangle - y_t)^2 \tag{4}$$

The solution $\theta_* \in \mathcal{R}^d$ to (**??**) is also the solution to the *linear inversion* problem of the form in (**??**) with $A = \mathbf{E}x_t x_t^\top$ and $b = \mathbf{E}y_t x_t$. The stochastic gradient descent (SGD) with a constant step-size $\alpha > 0$ and iterate averaging (see **?**]) for solving (**??**) is an instance of CALSA in
eqrefeq:lsa with $A_t = x_t x_t^\top$, $b_t = y_t x_t$. Here, we are

[2]That is, $\mathbf{E}[M_t|\mathcal{F}_{t-1}] = 0$ and $\mathbf{E}[N_t|\mathcal{F}_{t-1}] = 0$ and M_t, N_t are $\mathcal{F}_t$ measurable, $t \geq 0$.

interested in the MSE for either *estimation* or *prediction* problems which are given by $\mathbf{E}\left\|\hat{\theta}_t - \theta_*\right\|^2$ and $\mathbf{E}\left\|\hat{\theta}_t - \theta_*\right\|_A^2$ respectively.

Example 3 (Approximate Policy Evaluation in Reinforcement Learning)**.** Consider a *Markov Decision Process* given by the tuple $\langle S, A, P, R, \gamma\rangle$, where S is the state space, A is the action space, P is the probability transition kernel which specifies the probability $p_a(s, s')$ of transitioning from state $s \in S$ to state $s' \in S$ under an action $a \in A$, and $R = (r_a(s), a \in a, s \in S)$ specifies the reward $r_a(s)$ of performing action a in state s. Formally, a policy is a map π that specifies a probability distribution $\pi(\cdot|s)$ over A for any given S. The policy evaluation problem deals with computing $V_\pi(s) = \mathbf{E}[\sum_{t=0}^{\infty}\gamma^t r_{a_t}(s_t)|s_o = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim p_{a_t}(s_t, \cdot)]$. It is known that $V_\pi = R_\pi + \gamma P_\pi V_\pi$, where $R_\pi = (\mathbf{E}_\pi[r_a(s)], s \in \mathcal{S}) \in \mathcal{R}^{|\mathcal{S}|}$ is the reward vector and $P_\pi(s, s') = \mathbf{E}_\pi p_a(s, s')$ is the probability transition matrix of the Markov chain under policy π. Thus computing the value function is a *linear inversion* problem i.e., $V_\pi = (I - \gamma P_\pi)^{-1} R_\pi$.

We now briefly describe the problem of approximate policy evaluation in a batch setting with *i.i.d.* resampling. The data is sampled from the stationary distribution d_{π_b} of a behaviour policy π_b and is presented as the sequence $(s_t, s'_t, r_t, a_t), t \geq 0$, where $s'_t \sim p_{\pi_b}(s, \cdot)$, $r_t = R(s_t)$, $a_t \sim \pi_b(\cdot|s_t)$and $s_t \sim d_{\pi_b}$. The aim is to approximate $V_\pi \approx \Phi\theta_*$, where $\Phi \in \mathcal{R}^{|S|\times d}$ is a feature matrix and θ_* is a weight vector to be learnt. Typically, θ_* is obtained as a solution a linear inversion problem and algorithms to compute θ_* are known as *temporal difference* learning algorithms. Two important TD algorithms that we are interested in this paper namely TD(0) and gradient temporal difference (GTD) are given as follows:

TD(0)	GTD
$\delta_t = r_t + (\gamma\phi'_t - \phi_t)^\top\theta_t,$ $\theta_{t+1} = \theta_t + \alpha\rho_t\phi_t(\delta_t),$ $\hat{\theta}_t = \frac{1}{t+1}\sum_{s=0}^{t}\theta_s$	$\Delta_t = \phi_t(\gamma\phi'_t - \phi_t)^\top,$ $y_{t+1} = y_t + \alpha\rho_t(\phi_t r_t + \Delta_t\theta_t - y_t),$ $\theta_{t+1} = \theta_t + \alpha\Delta_t^\top y_{t+1}$ $\hat{\theta}_t = \frac{1}{t+1}\sum_{s=0}^{t}\theta_s, \hat{y}_t = \frac{1}{t+1}\sum_{s=0}^{t} y_s,$

Table 1: Shows the TD(0) and GTD algorithms. Here $\rho_t \doteq \frac{\pi(a_t|s_t)}{\pi_b(a_t|s)}$ is the *importance sampling* ratio that corrects for the differences in the *behaviour* policy π_b and the *target* policy π whose values function V_π is of interest.

4 Problem Landscape

In this section we try to understand following are possible questions:

1. (Q1) Is is possible to obtain uniformly fast asymptotic rates, i.e., given a problem class $\mathcal{P}$ does the MSE converge to zero at a rate $O(\frac{1}{t})$[3] as $t \to \infty$?
2. (Q2) Is it possible to obtain uniform fast rates in finite-time, i.e., given a problem class $\mathcal{P}$ we want the MSE converges to zero at a rate $\frac{C}{t}$, where the C is independent of $\mathcal{P}$?

It turns out that the answers to both Q1 and Q2 is *negative*, and we present (informal) arguments for the same.

Q2- Uniformly fast finite-time rate: Consider the following special case of CALSA with additive noise:

$$\text{LSA:} \qquad \theta_t = \theta_{t-1} + \alpha(b_t - A_P\theta_{t-1}), \tag{5a}$$

$$\text{PR-Average:} \qquad \hat{\theta}_t = \frac{1}{t+1}\sum_{s=0}^{t}\theta_i. \tag{5b}$$

Notice that (**??**) is a special case of (**??**) with $A_t = A_P, \forall t \geq 0$. For the *error* variable defined as $\hat{e}_t \doteq \hat{\theta}_t - \theta_*$, the following proposition holds.

Proposition 1. For all $\alpha_P > 0$ such that $\Lambda(I - \alpha_P A) < 1$, we have

$$\hat{e}_t = \frac{1}{t+1}(\alpha_P A)^{-1}\big([I - (I - \alpha_P A)^{t+1}e_0 + \alpha_P\sum_{s=1}^{t}[I - (I - \alpha_P A)^{t+1-s}]N_s]\big)$$

In what follows, we would like to understand what CALSA holds for the problem class $\mathcal{P}_{SPD}$ where A_P is real symmetric and positive definite with $\Lambda(A_P) < 1$ for all $P \in P_{SPD}$. In particular, we want to know for a fixed t and adversarial choice of problem instance from $\mathcal{P}_{SPD}$ whether it is possible to choose a constant step-size $\alpha > 0$ so that we get desirable rates of convergence for the MSE. We state our observation in the following remarks:

- **Step-Size**: We can choose any $\alpha \in (0, 1)$ and it follows that $\Lambda(I - \alpha A) < 1, \forall P \in \mathcal{P}_{SPD}$. Thus, there exists a universal step-size choice for $\mathcal{P}_{SPD}$.
- **Estimation Case:** For small t, i.e., t such that $\Lambda(\alpha t A_P) < 1$ we have $\mathbf{E}\|\hat{e}_t\|^2 \approx \frac{1}{(t+1)^2}(\mathcal{B}(t+1)^2 + \alpha^2\sigma_b^2 O(t^3)) = \mathcal{B} + \alpha^2\sigma_b^2 O(t)$. Thus for a fixed $t > 0$ and bias $\mathcal{B}$, it follows that no matter how we choose α, we would suffer a $O(\mathcal{B})$ error.
- **Prediction Case:** For small t, i.e., t such that $\Lambda(\alpha t A_P) < 1$ we have $\mathbf{E}\|\hat{e}_t\|_A^2 \approx O(\frac{\mathcal{B}}{\alpha t}) + O(\sigma_b^2\alpha)$, where we can balance the bias and the variance terms by choosing $\alpha = \frac{1}{\sqrt{t}}$.

[3] $\frac{1}{t}$ is the statistical rate.

- **Scaling Noise Case:** For small t, i.e., t such that $\Lambda(\alpha t A_P) < 1$, and $\sigma_b^2 \leq \Lambda(A_P)$ we have $\mathbf{E}\|\hat{e}_t\|_A^2 \approx O(\frac{\mathcal{B}}{\alpha t}) + O(\frac{1}{t})$, and by choosing $\alpha = 1$ we achieve a rate of $O(\frac{1}{t})$.

Thus, it turns out that a uniformly fast finite-time rate of $O(\frac{1}{t})$ can be achieved only under special structure on the noise and we way in which the error is measured. For instance, we have pressing evidence to suspect that in the additive noise case, the linear least squares problem achieves uniform rate of $O(\frac{1}{t})$. We believe the question in the multiplicative noise is still open (for more see [**?**]).

Q1 Uniformly fast asymptotic rate: A necessary condition to achieve an uniform asymptotic rate is to be choose a step-size $\alpha > 0$ that is problem independent. We show a problem class $\mathcal{P}$ where such an instance independent choice is not possible: consider the deterministic class $\mathcal{P}_{det}$ where $b_t = \mathbf{0}$, $A_t = A_P, \forall t \geq 0$ and $A_P = \left\{ \begin{bmatrix} u & v \\ -v & u \end{bmatrix} : \right\}$ and the class $\mathcal{P}_{det}$ is generated by the set $\{(u, v) : u^2 + v^2 \leq B\}$. It is clear that the data from $\mathcal{P}_{det}$ is bounded. Is it also true that there exists a universal step-size for $\mathcal{P}_{det}$? The answer is no:

Proposition 2. There does not exist α such that $\Lambda(I - \alpha A_P) < 1, \forall P \in \mathcal{P}$.

A sufficient Condition: Going by the negative answers for Q1 and Q2, we shift our focus to obtaining uniform asymptotic rates. However, in order to do so we need a verifiable condition that implies existence of a constant step-size across a given problem class $\mathcal{P}$. To this end, we introduce the notion of what we call *admissibility*.

Definition 5. Call a set of distributions $\mathcal{P}$ over $\mathbb{C}^d \times \mathbb{C}^{d\times d}$ *admissible* if there exists $\alpha_{\mathcal{P}} > 0$ such that $\rho_s(\alpha, P) > 0$ holds for all $P \in \mathcal{P}$ and $\alpha \in (0, \alpha_{\mathcal{P}})$.

It is easy to see that $\alpha \mapsto \rho_s(\alpha, P)$ is decreasing, hence if $\alpha_{\mathcal{P}} > 0$ witnesses that $\mathcal{P}$ is admissible then any $0 < \alpha' \leq \alpha_{\mathcal{P}}$ is also witnessing this. In simple terms, the class $\mathcal{P}$ admits a universal step-size choice. We believe that the usefulness of **??** will be more clear in the discussion that follows the instance dependent bound in **??**.

5 Instance Dependent Bounds

We break the question of achieving a uniformly fast asymptotic rate of $O(\frac{1}{t})$ into two parts *i*) showing that an instance dependent rate of $O(\frac{1}{t})$ exists for any problem P, and *ii*) showing that for specific problem classes a universal step-size choice can be made.

In this section, we derive instance dependent bounds that are valid for a given problem P (satisfying **??**). Here, we only present the main results followed by a discussion. The detailed proofs can be found in **??**. We start with a lemma, which is needed to meaningfully state our main result:

Lemma 1. Let P be a distribution over $\mathcal{R}^d \times \mathcal{R}^{d\times d}$ satisfying **??**. Then there exists an $\alpha_{P_U} > 0$ and $U \in \mathrm{GL}(d)$ such that $\rho_d(\alpha, P_U) > 0$ and $\rho_s(\alpha, P_U) > 0$ holds for all $\alpha \in (0, \alpha_{P_U})$.

Theorem 1. Let P be a distribution over $\mathcal{R}^d \times \mathcal{R}^{d\times d}$ satisfying **??**. Then, for $U \in \mathrm{GL}(d)$ and $\alpha_{P_U} > 0$ as in **??**, for all $\alpha \in (0, \alpha_{P_U})$ and for all $t \geq 0$,

$$\mathbf{E}\left[\|\hat{\theta}_t - \theta_*\|^2\right] \leq \nu \left\{ \frac{\|\theta_0 - \theta_*\|^2}{(t+1)^2} + \frac{v^2}{t+1} \right\},$$

where $\nu = \left(1 + \frac{2}{\alpha\rho_d(\alpha, P_U)}\right) \frac{\kappa(U)^2}{\alpha\rho_s(\alpha, P_U)}$ and $v^2 = \alpha^2(\sigma_A^2 \|\theta_*\|^2 + \sigma_b^2) + \alpha(\sigma_A^2 \|\theta_*\|) \|\theta_0 - \theta_*\|$.

Note that ν depends on P_U and α, while v^2 in addition also depends on θ_0. The dependence, when it is essential, will be shown as a subscript.

Theorem 2 (Lower Bound)**.** There exists a distribution P over $\mathcal{R}^d \times \mathcal{R}^{d\times d}$ satisfying **??**, such that there exists $\alpha_P > 0$ so that $\rho_s(\alpha, P) > 0$ and $\rho_d(\alpha, P) > 0$ hold for all $\alpha \in (0, \alpha_P)$ and for any $t \geq 1$, $\mathbf{E}\left[\|\hat{\theta}_t - \theta_*\|^2\right] \geq \frac{1}{\alpha^2 \rho_d(\alpha,P)\rho_s(\alpha,P)} \left\{ \frac{\beta_t \|\theta_0 - \theta_*\|^2}{(t+1)^2} + \frac{v^2 \sum_{s=1}^t \beta_{t-s}}{(t+1)^2} \right\}$, where $\beta_t = \left(1 - (1 - \alpha\rho_s(\alpha, P))^t\right)$ and v^2 is as in **??**.

Note that $\beta_t \to 1$ as $t \to \infty$. Hence, the lower bound essentially matches the upper bound. In what follows, we discuss the specific details of these results.

Role of U: U is helpful in transforming the recursion in θ_t to $\gamma_t = U^{-1}\theta_t$, which helps in ensuring $\rho_s(\alpha, P_U) > 0$. Such similarity transformation have also been considered in analysis of RL algorithms [**?**]. More generally, one can always take U in the result that leads to the smallest bound.

Role of $\rho_s(\alpha, P)$ **and** $\rho_d(\alpha, P)$: When P is positive definite, we can expand the MSE as We can expand the MSE as

$$\mathbf{E}\left[\|\hat{e}_t\|^2\right] = \frac{1}{(t+1)^2} \left\langle \sum_{s=0}^t e_s, \sum_{s=0}^t e_s \right\rangle, \qquad (6)$$

where $\hat{e}_t = \hat{\theta}_t - \theta_*$ and $e_t = \theta_t - \theta_*$. The inner product in (**??**) is a summation of *diagonal* terms $\mathbf{E}[\langle e_s, e_s\rangle]$ and *cross* terms of $\mathbf{E}[\langle e_s, e_q\rangle]$, $s \neq q$. The growth of the diagonal terms and the cross terms depends on the spectral norm of the random matrices $H_t = I - \alpha A_t$ and that of the deterministic matrix $H_P = I - \alpha A_P$,

respectively. jFor the MSE to be bounded, we need the spectral norms to be less than unity, implying the conditions $\rho_s(\alpha, P) > 0$ and $\rho_d(\alpha, P) > 0$. If P is Hurwitz, we can argue on similar lines by first transforming P into a positive definite problem P_U and replacing $\rho_s(\alpha, P)$ and $\rho_d(\alpha, P)$ by $\rho_s(\alpha, P_U)$ and $\rho_d(\alpha, P_U)$, and introducing $\kappa(U)$ to account for the forward ($\gamma = U^{-1}\theta$) and reverse ($\theta = U\gamma$) transformations using U^{-1} and U respectively.

Constants α, $\rho_s(\alpha, P)$ and $\rho_d(\alpha, P)$ do not affect the exponents $\frac{1}{t}$ for variance and $\frac{1}{t^2}$ for bias terms. This property is not enjoyed by all step-size schemes, for instance, step-sizes that diminish at $O(\frac{c}{t})$ are known to exhibit $O(\frac{1}{t^{\mu c/2}})$ (μ is the smallest real part of eigenvalue of A_P), and hence the exponent of the rates are not robust to the choice of $c > 0$ [**?** **?**].

Bias and Variance: The MSE at time t is bounded by a sum of two terms. The first *bias* term is given by $\mathcal{B} = \nu \frac{\|\theta_0 - \theta_*\|^2}{(t+1)^2}$, bounding how fast the initial error $\|\theta_0 - \theta_*\|^2$ is forgotten. The second *variance* term is given by $\mathcal{V} = \nu \frac{v^2}{t+1}$ and captures the rate at which noise is rejected.

Behaviour for extreme values of α: As $\alpha \to 0$, the bias term blows up, due to the presence of α^{-1} there. This is unavoidable (see also **??**) and is due to the slow forgetting of initial conditions for small α. Small step-sizes are however useful to suppress noise, as seen from that in our bound α is seen to multiply the variances σ_A^2 and σ_b^2. In quantitative terms, we can see that the α^{-2} and α^2 terms are trading off the two types of errors. For larger values of α with α_P chosen so that $\rho_s(\alpha, P) \to 0$ as $\alpha \to \alpha_P$ (or α_{P_U} as the case may be), the bounds blow up again.

The lower bound of **??** shows that the upper bound of **??** is tight in a number of ways. In particular, the coefficients of both the $1/t$ and $1/t^2$ terms inside $\{\cdot\}$ are essentially matched. Further, we also see that the $(\rho_s(\alpha, P)\rho_d(\alpha, P))^{-1}$ appearing in $\nu = \nu_{P_u,\alpha}$ cannot be removed from the upper bound. Note however that there are specific examples, such as SGD for linear least-squares, where this latter factor can in fact be avoided (for further remarks see **??**).

We now state results which show that certain important class of RL problems are *admissible*.

Theorem 3. Define constants $B \doteq \max_{ij} |A_{ij}|$ and $\rho_{\max} \doteq \max_{t \geq 0} \rho_t$. We have

- For TD(0) in **??**, the class of RL problem with $\mathbf{E}\phi_t\phi_t^\top = \mathbf{E}\phi'_t\phi'^\top_t$ is admissible with witness $\alpha = \frac{1}{B^2 d \rho_{\max}}$.
- For GTD in **??**, RL problems are admissible with witness $\alpha = \frac{1}{2B^4 d^2 \rho_{\max}}$.

Remarks:

- The condition for the constant step-size is related to the maximum possible operator norm of the random matrices A_ts involved. However, since A_ts themselves are rank-1 in the features, it turns out that the result in **??** hold for normalized features with step-size of $\alpha = 1$ in the case ofTD(0) and $\alpha = \frac{1}{2}$ in the case of GTD.
- From **??** we observe that the constant step-size for GTD is *squared* that of the TD(0) (sans the factor of 2) and hence can be very conservative. This is because the GTD can be thought of the CALSA to solve the linear inverse problem $A^\top A\theta_* = A^\top b$, using two variables (the primal and the dual). Thus the step-size also gets *squared* relative to TD(0), and is the indirect price paid for having a stable scheme irrespective of the distribution mismatch (*on* versus *off* policy).

6 Related Work

Other Step-Size Methods: It is clear that for the LSA in (**??**) to be stable α_t should be non-increasing. In this paper, we showed that results for LSA with a constant step-size and averaging of the iterates. This brings us to a brief discussion on the other two non-increasing choices for step-size strategies namely the diminishing and adaptive strategies. An immediate choice could be $\alpha_t = \frac{1}{t}$, however it fails very badly even in the following simple example: Consider the LSA in 1-dimensional given by $\theta_{t+1} = \theta_t + \alpha_t(b - a\theta_t)$, with $0 < b = a << 1$. The solution to the linear inversion problem is $\theta_* = 1$, however, one can show via elementary arguments that the MSE converges only at a rate of $O(\frac{1}{t^{2a}})$, which is very bad for small values of $a > 0$. The more practical approach has been to use $\alpha_t = \frac{cc_0}{c+t}$, for some $c >> 0$, so that initially for $t << c$ there is a constant step-size of c_0 and later on the step-sizes diminish at a rate of $\frac{1}{t}$. Typically, both c and c_0 are tuned and only the best values are reported (without accounting cost of searching for the useful c and c_0). **?**] introduce a new adaptive step-size tuning method for TD algorithm and show that it performs better in comparison to various other adaptive approaches. The adaptive step-size rule suggested[4] is $\alpha_t = \min(\alpha_{t-1}, \| \phi_t^\top(\gamma\phi'_t - \phi_t) \|^{-1})$. It turns out that once the *minimum* is attained after all the states transition pairs (s, s') are visited at least once, the rule stops to adapt and stays constant. However, it is evident that a constant step-size alone is not enough

[4]The case when eligibility traces are not used.