Analyzing the Variance of Policy Gradient Estimators for the Linear-Quadratic Regulator

James A. Preiss,[*] **Sébastien M. R. Arnold**,[*] **Chen-Yu Wei**[*]
Department of Computer Science
University of Southern California
Los Angeles, USA
{japreiss, seb.arnold, chenyu.wei}@usc.edu

Marius Kloft
Department of Computer Science
T.U. Kaiserslautern
Kaiserslautern, Germany
neu@cs.uni-kl.de

Abstract

We study the variance of the REINFORCE policy gradient estimator in environments with continuous state and action spaces, linear dynamics, quadratic cost, and Gaussian noise. These simple environments allow us to derive bounds on the estimator variance in terms of the environment and noise parameters. We compare the predictions of our bounds to the empirical variance in simulation experiments.

1 Introduction

Policy gradient (PG) algorithms are widely used for reinforcement learning (RL) in continuous spaces. PG methods construct an unbiased estimate of the gradient of the RL objective with respect to the policy parameters. They do so without the complication of intermediate steps of dynamics modeling or value function approximation (Sutton and Barto, 2018). However, the gradient estimate is known to suffer from high variance. This makes PG methods sample-inefficient with respect to environment interaction, creating an obstacle for applications to real physical systems.

In this paper, we seek a more detailed understanding of how the PG gradient estimate variance relates to properties of the continuous-space Markov decision process (MDP) that defines the RL problem instance. Such characterization is well-developed for discrete state spaces (e.g. Greensmith et al., 2004), but in continuous spaces, a detailed breakdown is not possible without further restrictions on the set of MDPs and the policy class. We choose to examine systems with linear dynamics, linear policy, quadratic cost, and Gaussian noise, known as LQR systems in control theory.

LQR systems are a popular case study for analyzing RL algorithms in continuous spaces. Fazel et al. (2018) show that the optimization landscape is neither convex nor smooth, but are still able to show global convergence and PAC-type complexity bounds for a zeroth-order optimization method that explores in parameter space. Malik et al. (2018) establish tighter error bounds with more detailed problem-dependence for the same class of algorithms, and extend the results to the noisy dynamics case. Yang et al. (2019) show convergence for an actor-critic method. Recht (2018) provides a broad overview including summaries of the author's prior work on model-based methods and value function approximation. Our variance bounds are similar to those of Malik et al. (2018), but we focus on characterizing the variance of policy gradient methods, which explore by taking random actions at each time step, rather than parameter-space exploration methods.

The earliest and simplest policy gradient algorithm is REINFORCE (Williams, 1992). More recent algorithms such as TRPO and PPO (Schulman et al., 2015, 2017) extend the basic idea of REINFORCE with techniques to inhibit the possibility of making very large changes in the policy action distribution in a single step. In a benchmark test (Duan et al., 2016), these algorithms generally learned better policies than REINFORCE, but their additional complexity makes them hard to analyze.

---

[*]Equal contribution.

Our primary contributions are derivations of bounds on the variance of the REINFORCE gradient estimate as an explicit function of the dynamics, reward, and noise parameters of the LQR problem instance. We validate our bounds with comparisons to the empirical gradient variance in random problems. We also explore the relationship between gradient variance and sample complexity, but find it to be less straightforward, as the problem parameters that affect variance also affect the optimization landscape. We emphasize that our goal is not to draw a conclusion about the utility of using REINFORCE to solve LQR problems, but rather to use LQR as an example system that is simple enough to allow us to "look inside" the REINFORCE policy gradient estimator.

2 Problem setting

In this section, we define the general finite-horizon reinforcement learning problem, the REINFORCE policy gradient estimator, and the LQR optimal control problem. We use the notation $\mathcal{P}(\mathcal{X})$ for the set of probability distributions over a measurable set $\mathcal{X}$. For arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, the set of all functions $\mathcal{X} \mapsto \mathcal{Y}$ is denoted as $\mathcal{Y}^{\mathcal{X}}$. Let $\| \cdot \|$ denote the $\ell_2 - \ell_2$ operator norm of a matrix or the ℓ_2 norm of a vector. The spectral radius (magnitude of largest eigenvalue) of a square matrix is denoted by $\rho(\cdot)$. The the set of positive semidefinite (resp. positive definite) $k \times k$ matrices is denoted by $\mathbb{P}_+^k$ (resp. $\mathbb{P}_{++}^k$). Finally, $\Sigma^{\frac{1}{2}}$ denotes the principal matrix square root of $\Sigma \in \mathbb{P}_+^k$.

RL problem statement. Reinforcement learning takes place in a Markov Decision Process (MDP) defined by state space $\mathcal{S}$, action space $\mathcal{A}$, initial state distribution $\varrho \in \mathcal{P}(\mathcal{S})$, state transition function $T : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathcal{S})$, and reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathbb{R})$. The agent's actions are drawn according to a stochastic policy $\pi : \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$. Let τ denote a state-action trajectory $s_1, a_1, s_2, a_2, \ldots, s_H, a_H$, of horizon $H \in \mathbb{N}$, and $\mathbb{T} = (\mathcal{S} \times \mathcal{A})^H$ the set of all such trajectories. The policy π induces a trajectory distribution $p_\pi \in \mathcal{P}(\mathbb{T})$, meaning $s_1 \sim \varrho$, $a_t \sim \pi(s_t)$, $s_{t+1} \sim T(s_t, a_t)$, $r_t \sim r(s_t, a_t)$. The RL optimization problem is defined over a tractable policy space $\Pi \subseteq \mathcal{P}(\mathcal{A})^{\mathcal{S}}$ as follows:

$$\underset{\pi \in \Pi}{\text{maximize}} \quad J(\pi) = \mathbb{E}_{\tau \sim p_\pi}[\phi(\tau)], \text{ where } \phi(\tau) = \sum_{t=0}^{H} r_t. \tag{1}$$

The MDP parameters (ϱ, T, r) are unknown to the RL algorithm. The algorithm must learn exclusively by sampling from p_π.

Policy gradient algorithms. Suppose π is parameterized by a real-valued vector θ. Analytical gradient descent of $\nabla_\theta J(\pi)$ is not possible because T and r are only accessible via sampling, with unknown gradients. There exist generic derivative-free optimization algorithms that solve such problems via perturbations in parameter space, but in RL problems, it is also possible to explore via stochastic actions instead. The simplest policy gradient algorithm is REINFORCE (Williams, 1992), which relies on the following identity (assuming sufficient regularity conditions):

$$\nabla_\theta J(\pi) = \nabla_\theta \int_{\mathbb{T}} p_\pi(\tau)\phi(\tau)d\tau = \mathbb{E}_{\tau \sim p_\pi} \left[\phi(\tau) \sum_{t=1}^{H} \nabla_\theta \log \pi(a_t|s_t) \right]. \tag{2}$$

An unbiased estimate of the latter expectation can be computed by executing π in the MDP for one full trajectory. Unfortunately, this estimate is known to have high variance. Variance reduction is possible by exploiting the Markov property (past rewards are independent of future actions) and/or using control variates (Greensmith et al., 2004), but we analyze plain REINFORCE here for simplicity.

LQR systems. A discrete-time stochastic *linear-quadratic regulator* (LQR) system with state space $\mathcal{S} = \mathbb{R}^n$ and action space $\mathcal{A} = \mathbb{R}^m$ is defined by linear dynamics with additive Gaussian noise:

$$s_{t+1} = As_t + Ba_t + \epsilon_t^s, \quad \epsilon_t^s \sim \mathcal{N}(\mathbf{0}, \Sigma_s), \tag{3}$$

for dynamics matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, and noise covariance $\Sigma_s \in \mathbb{P}_+^n$. The reward function is given by $r_t = -(s_t^T Q s_t + a_t^T R a_t)$ for cost matrices $Q \in \mathbb{P}_+^n$, $R \in \mathbb{P}_{++}^m$. Intuitively, the goal is to drive the state towards zero without using too much control effort. The initial state s_1 follows an arbitrary zero-mean Gaussian distribution. A well-known result in control theory (Kwakernaak and Sivan, 1972) states that, if the system (A, B) is controllable, the infinite-horizon objective

$$\lim_{H \to \infty} \mathbb{E}_{s_1, \{\epsilon_t^s\}_{t=1}^H} \left[\sum_{t=1}^{H} r_t \right] \tag{4}$$

2

is maximized by a stationary linear policy $a_t = K^\star s_t$, $K^\star \in \mathbb{R}^{m \times n}$. The value of $K^\star$ depends on (A, B, Q, R), but not on the distributions of Σ_s or s_1. The same $K^\star$ is also the optimal controller for the deterministic version of the problem. $K^\star$ can be computed efficiently (Van Dooren, 1981).

To apply REINFORCE, the policy must be stochastic, so we consider linear *stochastic* policies

$$a_t = K s_t + \epsilon_t^a, \quad \epsilon_t^a \sim \mathcal{N}(\mathbf{0}, \Sigma_a), \tag{5}$$

for $K \in \mathbb{R}^{m \times n}, \Sigma_a \in \mathbb{P}_{++}^m$. The state noise Σ_s is an immutable property of the system, but the action noise Σ_a is not. Instead, it is usually chosen by the user of the RL algorithm, or learned as a parameter using (2). Genuine noise in the system actuators can be subsumed into $\Sigma_s' = \Sigma_s + B \Sigma_a$.

In the RL literature, action noise is usually seen as either 1) a tool for exploring of the state space, 2) a method of regularization to avoid converging on bad local optima, or 3) a consequence of a probabilistic interpretation of the RL problem (Levine, 2018). Its effect on the RL optimization algorithm is less frequently discussed, but in this work we find that it can be significant.

3 Main result: Variance bounds on the REINFORCE estimator

In this section, we present bounds on the variance of the REINFORCE estimator for LQR systems. The instantiation of REINFORCE (eq. (2)) for the system (eqs. (3) to (5)) using a single trajectory is:

$$\hat{g} = \left(\sum_{t=1}^{H} \Sigma_a^{-1} \epsilon_t^a s_t^\top \right) \left(\sum_{t=1}^{H} r_t \right) \in \mathbb{R}^{m \times n}. \tag{6}$$

The estimate $\hat{g}$ is a function of the independent random variables $\{\epsilon_t^a, \epsilon_t^s\}_{t=1}^H$. Although s_t is linear in $\{\epsilon_\tau^a, \epsilon_\tau^s\}_{\tau=1}^{t-1}$, r_t is quadratic in s_t, so the overall form of $\hat{g}$ is a product of a sum and a sum of products of sums. Therefore, while it is possible to apply matrix concentration inequalities (Tropp, 2015) to bound $\|s_t - \mathbb{E}[s_t]\|$ with high probability, it is more difficult to bound the dispersion of $\hat{g}$. Instead, we use a more specialized method to derive a bound on

$$\nu(\hat{g}) \triangleq \sum_{i=1}^m \sum_{j=1}^n \mathrm{Var}(\hat{g}_{i,j}) = \mathbb{E}\left[\mathbf{tr}(\hat{g}^\top \hat{g}) \right] - \mathbf{tr}(\mathbb{E}[\hat{g}]^\top \mathbb{E}[\hat{g}]),$$

which we simplify by bounding $\mathbb{E}\left[\mathbf{tr}(\hat{g}^\top \hat{g}) \right]$.

Theorem 1. *If $\rho(A + BK) < 1$, then $\mathbb{E}\left[\mathbf{tr}(\hat{g}^\top \hat{g}) \right] \leq O\left(\bar{n}^4 C_1^2 C_2^2 \right)$, where*

$$C_1 = \mu^2 \left\| \Sigma_a^{-\frac{1}{2}} \right\| \left(\|s_1\| + \sigma H \right) H',$$

$$C_2 = \|R\| \|\Sigma_a\| H + \mu^2 \left(\|Q\| + \|R\| \|K\|^2 \right) \left(\|s_1\|^2 + \sigma^2 H \right) H'^2,$$

$\bar{n} \triangleq \max\{n, m\}$, $\sigma \triangleq \left\| \Sigma_s^{\frac{1}{2}} \right\| + \left\| B \Sigma_a^{\frac{1}{2}} \right\|$, $H' \triangleq \min\left\{ H, \frac{1}{1 - \rho(A + BK)} \right\}$, *and μ is a constant bounding the transient behavior of $\|A + BK\|^t$, with more details provided in Appendix A.*

Proof Sketch.

1. Rewrite $\epsilon_t^a, \epsilon_t^s$ as $\Sigma_a^{\frac{1}{2}} \delta_t^a, \Sigma_s^{\frac{1}{2}} \delta_t^s$, where the δ_t^a, δ_t^s are unit-Gaussian random variables.

2. Bound $\mathbf{tr}(\hat{g}^\top \hat{g})$ by P, a polynomial function of the χ-distributed random variables $\{\|\delta_t^a\|, \|\delta_t^s\|\}_{t=1}^H$ with nonnegative coefficients.

3. Bound the sum of the coefficients of P by substituting 1 for all χ random variables.

4. Bound for the expectation of each monomial in P using the moments of the χ distribution.

A detailed proof of Theorem 1 is given in Appendix A.

For a special case of scalar states and actions, we also show a lower bound on $\mathbb{E}[\hat{g}^2]$. Since $\mathbb{E}[\hat{g}] = 0$ at a local optimum, this lower bound corresponds to the variance caused strictly by noise in the system when the policy is already optimal. Here, the matrices A, B, K, Q, R are denoted as a, b, k, q, r, and σ_s, σ_a denote the standard deviation (not variance) of state and action noise. This notation r is different from the notation r_t for reward.

(a) Relationship between Σ_s and Σ_a (b) Control authority $|B|$ (c) Stability $\rho(A + BK)$
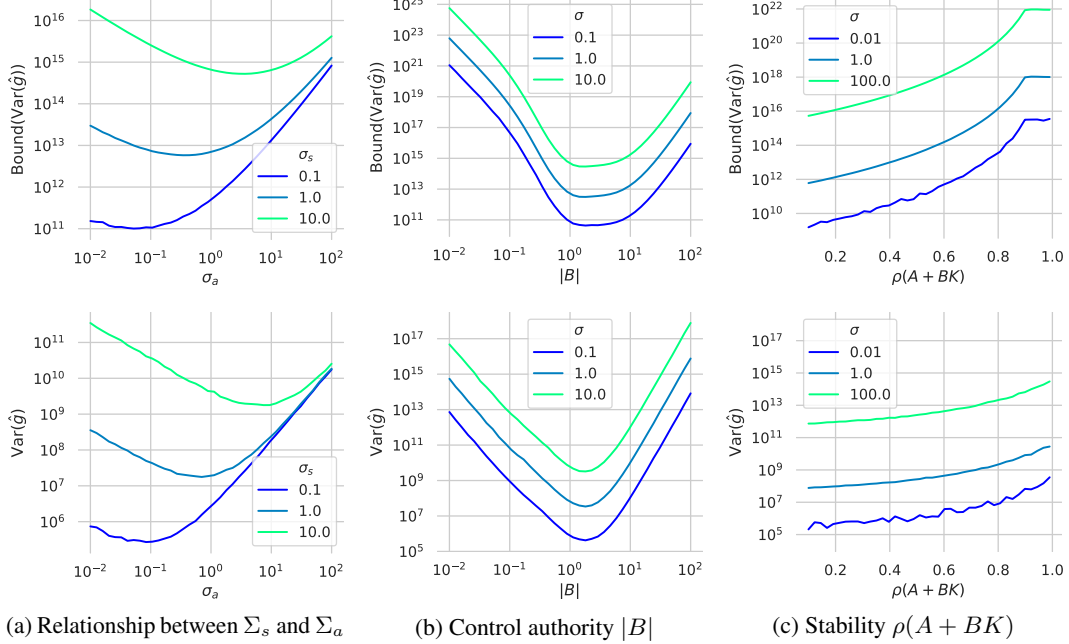
Figure 1: Comparison between our upper bound from Theorem 1 (top) and the empirically measured variance (bottom) as they relate to various parameters of the LQR problem. Behavior is qualitatively similar for action noise covariance (a) and control authority (b), but less similar for the stability (c) where our bounds are loose. Further discussion is in Sections 4 to 4.

Theorem 2. *If $m = n = 1$ and $0 \leq a + bk < 1$, then $\mathbb{E}[\hat{g}^2] \geq \Omega(c_1^2 c_2^2)$, with*

$$c_1 = \frac{1}{\sigma_a} \left(|s_1| + \sigma\sqrt{H} \right) \sqrt{h'},$$
$$c_2 = r\sigma_a^2 H + (q + rk^2)(s_1^2 + \sigma^2 H)h',$$

where $\sigma \triangleq \sigma_s + b\sigma_a$ and $h' \triangleq \min\left\{ H, \frac{1}{1-(a+bk)^2} \right\}$.

A detailed proof of Theorem 2 is given in Appendix B. If we reduce the upper bound of Theorem 1 to its scalar case, all terms match with the notable exception of the horizon-related terms H and H' (compare to h'), which appear squared in several places in Theorem 1 compared to the equivalent term in Theorem 2. There is another gap in the denominators of H' and h' since $\frac{1}{1-x^2} < \frac{1}{1-x}$ on the domain $x \in (0,1)$. We conjecture our upper bound can be tightened by fully exploiting the independence of the noise variables $\epsilon_s, \ \epsilon_a$.

4 Experiments

In the first set of experiments, we compare our upper bound of $\nu(\hat{g})$ to its empirical value when executing REINFORCE in randomly generated LQR problems. Our results show qualitative similarity in the parameters for which our upper and lower bounds match. On the other hand, the gap with respect to stability-related parameters is also visible. For each experiment shown here, we repeated the experiment with different random seeds and observed qualitatively identical results.

We generate random LQR problems with the following procedure. We sample each entry in A and B i.i.d. from $\mathcal{N}(0, \sigma=n^{-1/2})$ and $\mathcal{N}(0, \sigma=m^{-1/2})$ respectively. To construct a random $k \times k$ positive definite matrix, we sample from the $\mathrm{Wishart}(k^{-1}I, k)$ distribution by computing $Q = X^T X$ for X i.i.d. analogous to A. The scale factor k^{-1} ensures that if the vector x is distributed by $\mathcal{N}(0, k^{-1}I)$, such that $\mathbb{E}[\|x\|^2] = 1$, then $\mathbb{E}[x^T Q x] = 1$. We sample Q, R, Σ_s, and Σ_a this way.

In each experiment, we vary some of these parameters systematically while holding the others constant, allowing us to visualize the impact of each parameter on $\nu(\hat{g})$. We plot the upper bound

4

of Theorem 1 on the top row of Figure 1, and the empirical estimate of $\nu(\hat{g})$ on the bottom row. In both cases, since the variance depends on the initial state s_1, we sample $N = 100$ initial states s_1 from $\mathcal{N}(0, n^{-1}I)$ and plot $\mathbb{E}_{s_1 \sim \mathcal{N}(0, n^{-1}I)} \nu(\hat{g})$. We estimate $\nu(\hat{g})|_{s_1=s}$ for a particular initial state s by sampling 30 trajectories with random ϵ^a, ϵ^s.

Effect of Σ_a. In this experiment, we generate a random LQR problem and replace Σ_a with $\sigma_a I$ for σ_a geometrically spaced in the range $[10^{-2}, 10^2]$. Using a scaled identity matrix is common practice when applying RL to a problem where there is no *a priori* reason to correlate the noise between different action dimensions. We evaluate the variance at the value $K = K^\star$, where $K^\star$ is the infinite-horizon optimal controller computed using traditional LQR synthesis, as described in Section 2. Using $K^\star$ ensures that $\rho(A + BK) < 1$, a required condition to apply Theorem 1.

Results are shown in Figure 1a. The separate line plots correspond to scaling the random Σ_s by the values $\{0.1, 1, 10\}$, while the x−axis corresponds to the value of σ_a. For each value of σ_s, there appears to be a unique σ_a that minimizes $\nu(\hat{g})$, and this value of σ_a increases with σ_s. This phenomenon appears in both the bound and empirical variance.

Effect of $\|B\|$. In this experiment, we generate a random problem where $m = n$ and replace B with bI for b geometrically spaced in the range $[10^{-2}, 10^2]$. The resulting system essentially gives the policy direct control over each state. For each B, we compute a separate infinite-horizon optimal K and sample the variance for different s_1. Results are shown in Figure 1b. The separate line plots correspond to scaling both the random Σ_s and the random Σ_a by the values $\{0.1, 1, 10\}$. The x−axis corresponds to the value of b. Again, there appears to be a unique $\|B\|$ that minimizes $\nu(\hat{g})$, but its value changes minimally for different magnitudes of Σ_s, Σ_a.

Effect of $\rho(A + BK)$. Here we measure the change in variance with respect to the closed-loop spectral radius $\rho(A+BK)$. To synthesize controllers K such that $\rho(A + BK)$ obtains a specified value, we use the pole placement algorithm of Tits and Yang (1996). A pole placement algorithm $\mathscr{P}$ is a function

$$K = \mathscr{P}(A, B, \lambda_1, \ldots, \lambda_n), \quad \lambda_i \in \mathbb{C},$$

such that the eigenvalues of $A + BK$ are $\lambda_1, \ldots, \lambda_n$. We sample a "prototype" set of n eigenvalues with $\lambda_1, \ldots, \lambda_{\lceil n/2 \rceil}$ as complex conjugate pairs $\lambda_i, \lambda_{i+1} = re^{\pm i\varphi}$, where $r \sim \text{Uniform}([0,1])$ and $\varphi \sim \text{Uniform}([0,\pi))$, and sample the remaining real λ_i from $\text{Uniform}([-1,1])$. Then, for each desired ρ, we compute $K_\rho = \mathscr{P}(A, B, \rho\lambda_1, \ldots, \rho\lambda_n)$. By rescaling the same set of λ_i instead of sampling a new set for each ρ, we avoid confounding effects from changing other properties of K.

Figure 2: Scatter plot of empirical $\nu(\hat{g})$ (x−axis) and upper bound from Theorem 1 (y−axis) with varying state dimensionality n and time horizon H. Each point represents one random LQR problem.

Results are shown in Figure 1c. Again, we repeat the experiment for different magnitudes of Σ_s and Σ_a. Unlike the previous two experiments, here we see qualitatively different behavior between our upper bound and the empirical variance. The bound begins to increase rapidly near $\rho = 1$, corresponding to the growth of $1/(1 - \rho)$ in the term H', but at $\rho = 0.9$ the H term becomes active in H', and the bound suddenly flattens. In contrast, the empirical variance grows more moderately and does not explode near the threshold of system instability. This provides further evidence that the upper bound of Theorem 1 can be tightened to match the $\sqrt{H'}$ and $\sqrt{H}$ terms in the special-case lower bound of Theorem 2.

Dimensionality parameters. In all of the preceding experiments, we arbitrarily chose the state and action dimensions $n = 5$, $m = 3$ and time horizon $H = 10$. To visualize the variance for other values of these parameters, we generate 1000 random LQR problems with n and H each varying over the set $\{3, 10, 30\}$. We fix $m = \lceil n/2 \rceil$. Results are shown in Figure 2. The overall positive trend with a slope greater than 1 shows that the bound grows superlinearly with respect to the empirical, as
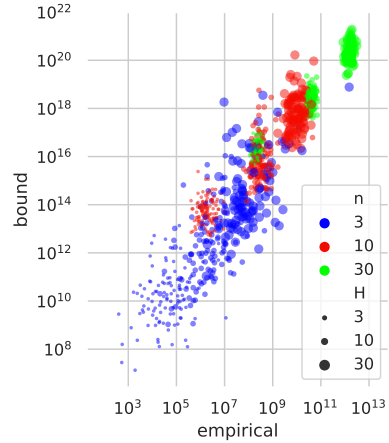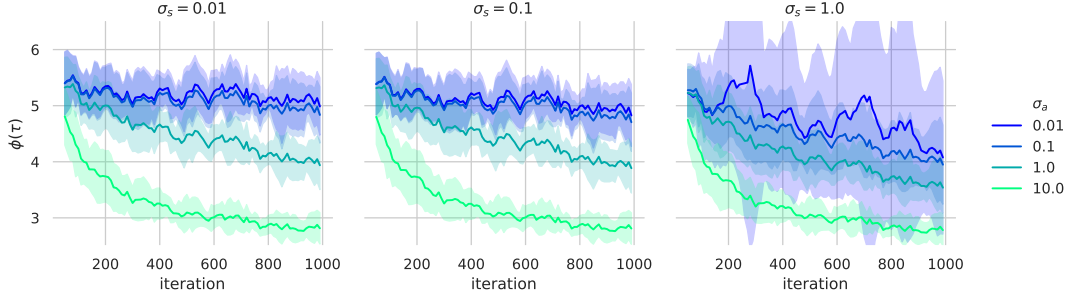
5

Figure 3: Learning curves of REINFORCE for a random LQR problem with varying scales of action noise σ_a and environment noise σ_s. Larger σ_a strictly improves learning, despite larger variance of $\hat{g}$.

expected. One interesting property is the tighter clustering for large values of n. This may be due to several eigenvalue distribution results in random matrix theory which state that, as $n \to \infty$, our random LQR problems tend to become similar up to a basis change (Tao, 2012).

4.1 RL learning curves for varying Σ_a

The results in Section 4 suggest that, for a fixed Σ_s, the magnitude of Σ_a has a significant effect on $\nu(g)$. This is of practical interest because Σ_a is usually under control of the RL practitioner. It is therefore natural to ask if the change in variance corresponds to a change in the rate of convergence of REINFORCE. We test this empirically by executing REINFORCE in variants of one random LQR problem with different values of Σ_a and Σ_s. To avoid a confounding effect from larger Σ_a incurring greater penalty from the $-a_t^T R a_t$ term in r_t, we evaluate the trained policies in a modified version of the problem where $\Sigma_a = \Sigma_s = \mathbf{0}$. As discussed in Section 2, the optimal $K^\star$ for the stochastic problem is also optimal for the deterministic problem, so each problem variant should converge to the same evaluation returns in the limit.

We initialize K by perturbing the elements of the LQR-optimal controller with i.i.d. Gaussian noise and scaling the perturbation until $\rho(A + BK) \approx 0.98$. After every 10 iterations of REINFORCE, we evaluate the current policy in the noise-free environment. For each (Σ_a, Σ_s) pair, we repeat this experiment 10 times with different random seeds. The random seed only affects the $\epsilon_t^a, \epsilon_t^s$ and s_1 samples. The aggregate data are shown in Figure 3. Shaded bands correspond to one standard deviation across the separate runs of REINFORCE. The lowercase σ_a, σ_s refer to scaling factors applied to the initial samples of Σ_a, Σ_s in the random LQR problem.

The effect is quite different than one would predict from variance alone. For all values of Σ_s in the experiment, problems with larger Σ_a converge faster—whereas Figure 1a would suggest that the "optimal" value of Σ_a changes with respect to Σ_s. The fact that larger Σ_a tends to make REINFORCE converge faster is not obvious, given the Σ_a^{-1} term in $\hat{g}$ (6). Also, when Σ_a is very small and Σ_s is very large, the algorithm becomes unstable and sees large variations across different random seeds. For the middle values $\Sigma_a \in \{0.1, 1.0\}$, we observe that larger Σ_s causes faster convergence.

5 Discussion

In this work, we derived bounds on the variance of the REINFORCE policy gradient estimator in the stochastic linear-quadratic control setting. Our upper bound is fully general, while our lower bound applies to the scalar case at a stationary point. The bounds match with respect to all system parameters except the time horizon H and stability $\rho(A + BK)$. We compared our bound prediction to the empirical variance in a variety of experimental settings, finding a close qualitative match in the parameters for which the bounds are tight.

Our experiments in Section 4.1 plotting the empirical convergence rate of REINFORCE suggest that the effect of action noise Σ_a on the overall RL performance is not fully captured by its effect on the variance. An interesting direction for future work would be to investigate the role of Σ_a more closely and attempt to disentangle its effect on gradient magnitude, variance, exploration, and regularization. Such an analysis could lead to improved variance reduction methods or algorithms that manipulate Σ_a to speed up the RL optimization.

References

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. *CoRR*, abs/1604.06778, 2016.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for linearized control problems. *CoRR*, abs/1801.05039, 2018.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471–1530, 2004.

Huibert Kwakernaak and Raphael Sivan. *Linear optimal control systems*, volume 1. Wiley-Interscience New York, 1972.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *CoRR*, abs/1805.00909, 2018.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, and Martin J. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *CoRR*, abs/1812.08305, 2018.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *CoRR*, abs/1806.09460, 2018.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 2 edition, 2018.

Terence Tao. *Topics in random matrix theory*. Graduate studies in mathematics. American Mathematical Society, Providence, RI, 2012. URL `http://cds.cern.ch/record/2122643`.

André L Tits and Yaguang Yang. Globally convergent algorithms for robust pole assignment by state feedback. *IEEE Transactions on Automatic Control*, 41(10):1432–1452, 1996.

Lloyd N Trefethen and Mark Embree. *Spectra and pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Paul Van Dooren. A generalized eigenvalue approach for solving riccati equations. *SIAM Journal on Scientific and Statistical Computing*, 2(2):121–135, 1981.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *CoRR*, abs/1907.06246, 2019.

A Proof of Theorem 1

In this appendix, we provide the detailed derivation of the upper bound stated in Theorem 1. We define the following notations: let δ_t^s and δ_t^a be independent random vectors that follow $\mathcal{N}(\mathbf{0}, I_n)$ and $\mathcal{N}(\mathbf{0}, I_m)$ respectively for all t. The ϵ_t^s and ϵ_t^a defined in Equations (3) and (5) are thus written as $\epsilon_t^s = \Sigma_s^{\frac{1}{2}} \delta_t^s$ and $\epsilon_t^a = \Sigma_a^{\frac{1}{2}} \delta_t^a$. We will use the following steps to upper-bound $\mathbb{E}\left[\mathbf{tr}(\hat{g}^\top \hat{g})\right]$:

1. Bound $\mathbf{tr}(\hat{g}^\top \hat{g})$ by P, a polynomial of $d_t^s \triangleq \|\delta_t^s\|$ and $d_t^a \triangleq \|\delta_t^a\|$ for $t = 1, 2, \ldots, H$. We restrict that P should have only nonnegative coefficients. Since we assume δ_t^s and δ_t^a are independent random vectors with standard normal distribution, d_t^s and d_t^a will be independent random variables following the $\chi(n)$ and $\chi(m)$ distributions respectively.

2. Bound the sum of the (already nonnegative) coefficients of P by substituting all of the d_t^s, d_t^a with one. More formally, let

$$P(\{d_t^s, d_t^a\}_{t=1}^H) = \sum_i C_i \prod_{t=1}^H (d_t^s)^{\alpha_{t,i}} \prod_{t=1}^H (d_t^a)^{\beta_{t,i}},$$

where $\prod_{t=1}^H (d_t^a)^{\alpha_{t,i}} \prod_{t=1}^H (d_t^s)^{\beta_{t,i}}$ is the i-th monomial in P, and C_i is its nonegative coefficient. Then we calculate $\sum_i C_i$ by substituting all d_t^s, d_t^a with 1. We use the notation $\mathbb{1}(P)$ to denote this operation (we define the $\mathbb{1}(\cdot)$ operator analogously for expressions other than P itself):

$$\mathbb{1}(P) = P|_{d_t^s=1, d_t^a=1, \forall t} = \sum_i C_i.$$

3. Bound for the expectation of all monomials in P, i.e., find M such that

$$\mathbb{E}\left[\prod_{t=1}^H (d_t^s)^{\alpha_{t,i}} \prod_{t=1}^H (d_t^a)^{\beta_{t,i}}\right] \le M, \quad \forall i.$$

With the three steps above, we can then bound $\mathbb{E}\left[\mathbf{tr}(\hat{g}^\top \hat{g})\right] \le \mathbb{E}[P] \le \sum_i C_i M = M \mathbb{1}(P)$. To calculate M, we can use the known formula for the k-th moment of a χ random variable.

Recall that

$$\hat{g} = \left(\sum_{t=1}^H \Sigma_a^{-1} \epsilon_t^a s_t^\top\right) \left(\sum_{t=1}^H r_t\right),$$

and thus

$$\mathbf{tr}(\hat{g}^\top \hat{g}) = \mathbf{tr} \underbrace{\left[\left(\sum_{t=1}^H \Sigma_a^{-1} \epsilon_t^a s_t^\top\right)^\top \left(\sum_{t=1}^H \Sigma_a^{-1} \epsilon_t^a s_t^\top\right)\right]}_{\textbf{Term}_1} \underbrace{\left(\sum_{t=1}^H r_t\right)^2}_{\textbf{Term}_2}.$$

Thanks to the property $\mathbb{1}(P) = \mathbb{1}(P_1)\mathbb{1}(P)$ for polynomial $P \equiv P_1 P$, and $\mathbb{1}(P) = \mathbb{1}(P_1) + \mathbb{1}(P)$ for $P \equiv P_1 + P$, we do not need to directly find P and bound $\mathbb{1}(P)$. Instead, we can bound $\mathbb{1}(P')$ for some smaller-order component P' of P, and then using the above addition/multiplication operations to obtain $\mathbb{1}(P)$. Below in Section A.1, we first bound $\|s_t\|$ by a polynomial $\lceil s_t \rceil$ of $\{d_\tau^s, d_\tau^a\}$, and then find $\mathbb{1}(\lceil s_t \rceil)$. In Section A.2 and A.3, we further upper bound $\mathbb{1}(\textbf{Term}_1)$ and $\mathbb{1}(\textbf{Term}_2)$ with the help of $\mathbb{1}(\lceil s_t \rceil)$. Then finally we obtain an upper bound for $\mathbb{E}[\mathbf{tr}(\hat{g}^\top \hat{g})]$ as $M \mathbb{1}(\textbf{Term}_1)\mathbb{1}(\textbf{Term}_2)$.

A.1 Bounding $\|s_t\|$

In this section, we bound $\|s_t\|$ from above. Although the spectral radius $\rho(A + BK)$ determines the asymptotic stability of the closed-loop system, it guarantees little about the transient behavior–for example, for any $x > 1$ and $0 < \epsilon < 1$, the matrix

$$A = \begin{bmatrix} \epsilon & x \\ 0 & \epsilon \end{bmatrix}$$

has the properties $\rho(A) < 1$, $\|A\| > x$. Therefore, while the state magnitude $\|s_t\|$ is bounded by the operator norm $\|A + BK\|^{t-1}$, it is too restrictive to require $\|A + BK\| < 1$. Instead, we will use the following result from the literature:

Lemma 3 (Trefethen and Embree (2005)). *Let $A \in \mathbb{R}^{n \times n}$ be a matrix with $\rho(A) < 1$. Then there exists $\mu > 0$ such that, for all $k \in \mathbb{N}$,*

$$\|A^k\| \leq \mu \rho(A)^k. \tag{7}$$

μ is bounded by the "resolvent condition" $\mu \leq 2enr(A)$, where e is the exponential constant and

$$r(A) = \sup_{z \in \mathbb{C},\, |z| > 1} (|z| - 1)\|(zI - A)^{-1}\|. \tag{8}$$

The derivation and interpretation of (8) is a deep subject related to the matrix *pseudospectrum*, covered extensively by Trefethen and Embree (2005). Intuitively, $r(A)$ is large if a small perturbation $\epsilon \in \mathbb{R}^{n \times n}$ would cause $\rho(\epsilon + A) > 1$.

Expanding the state transition function in Equation (3) with the linear stochastic policy in Equation (5), we get

$$s_t = (A + BK)^{t-1} s_1 + \sum_{\tau=1}^{t-1} (A + BK)^{t-\tau-1}(\Sigma_s^{\frac{1}{2}} \delta_\tau^s + B\Sigma_a^{\frac{1}{2}} \delta_\tau^a). \tag{9}$$

Recall that $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the $\ell_2 - \ell_2$ operator norm for matrices. By Lemma 3, there exists μ such that $\|(A + BK)^k\| \leq \mu\rho(A)^k$. Let $f = \rho(A + BK)$, $\sigma_s^2 = \|\Sigma_s\|$, $\sigma_a^2 = \|\Sigma_a\|$, and $b = \|B\|$. By repeatedly applying the triangle inequality,

$$\begin{aligned}
\|s_t\| &= \left\| (A + BK)^{t-1} s_1 + \sum_{\tau=1}^{t-1} (A + BK)^{t-\tau-1}(\Sigma_s^{\frac{1}{2}} \delta_\tau^s + B\Sigma_a^{\frac{1}{2}} \delta_\tau^a) \right\| \\
&\leq \left\| (A + BK)^{t-1} s_1 \right\| + \sum_{\tau=1}^{t-1} \left\| (A + BK)^{t-\tau-1}(\Sigma_s^{\frac{1}{2}} \delta_\tau^s + B\Sigma_a^{\frac{1}{2}} \delta_\tau^a) \right\| \\
&\leq \mu f^{t-1} \|s_1\| + \mu \sum_{\tau=1}^{t-1} f^{t-\tau-1}(\sigma_s d_\tau^s + b\sigma_a d_\tau^a).
\end{aligned} \tag{10}$$

We denote the final bound in (10) as $\lceil s_t \rceil$. The bound $\lceil s_t \rceil$ is linear in the random variables $\{d_t^s, d_t^a\}_{t=1}^H$ with only positive coefficients. Furthermore,

$$\begin{aligned}
\mathbb{1}(\lceil s_t \rceil) &= \mu f^{t-1} \|s_1\| + \mu \sum_{\tau=1}^{t-1} f^{t-\tau-1}(\sigma_s + b\sigma_a) \\
&= \mu f^{t-1} \|s_1\| + \mu \sum_{\tau=0}^{t-2} f^\tau (\sigma_s + b\sigma_a).
\end{aligned} \tag{11}$$

A.2 Bounding Term$_1$

Lemma 4.

$$\mathbf{tr}\left[\left(\sum_{t=1}^H \Sigma_a^{-1} \epsilon_t^a s_t^\top \right)^\top \left(\sum_{t=1}^H \Sigma_a^{-1} \epsilon_t^a s_t^\top \right) \right] \leq \|\Sigma_a^{-1}\| \left(\sum_{t=1}^H d_t^a \|s_t\| \right)^2.$$

Proof. Let $\xi_t = \Sigma_a^{-1} \epsilon_t^a$. Then

$$
\begin{aligned}
\mathbf{tr}\left[\left(\sum_{t=1}^H \xi_t s_t^\top \right)^\top \left(\sum_{t=1}^H \xi_t s_t^\top \right) \right] &= \mathbf{tr}\left[\sum_{1 \le i,j \le H} s_i \xi_i^\top \xi_j s_j^\top \right] \\
&= \sum_{1 \le i,j \le H} \xi_i^\top \xi_j s_i^\top s_j \\
&\le \sum_{1 \le i,j \le H} |\xi_i^\top \xi_j| |s_i^\top s_j| \\
&= \sum_{1 \le i,j \le H} |\delta_i^{a\top} \Sigma_a^{-1} \delta_j^a| |s_i^\top s_j| \\
&\le \sum_{1 \le i,j \le H} \|\Sigma_a^{-1}\| \|\delta_i^a\| \|\delta_j^a\| \|s_i\| \|s_j\| \\
&= \|\Sigma_a^{-1}\| \left(\sum_{t=1}^H \|\delta_t^a\| \|s_t\| \right)^2 \\
&\le \|\Sigma_a^{-1}\| \left(\sum_{t=1}^H d_t^a \lceil s_t \rceil \right)^2,
\end{aligned}
\tag{12}
$$

in which we make use of the Cauchy-Schwarz inequality and the fact that $\Sigma^{\frac{1}{2}} \Sigma^{-2} \Sigma^{\frac{1}{2}} = \Sigma^{-1}$ for positive semidefinite Σ. $\qquad\square$

A.3   Bounding Term$_2$

We now bound $C \triangleq \sum_{t=1}^H -r_t$ from above. Note that $-r_t \ge 0$, since $Q \succeq 0$ and $R \succ 0$. Let $q = \|Q\|$, $r = \|R\|$, and $k = \|K\|$. Then

$$
\begin{aligned}
C = \sum_{t=1}^H -r_t &= \sum_{t=1}^H s_t^\top Q s_t + a_t^\top R a_t \\
&\le \sum_{t=1}^H q\|s_t\|^2 + r\|a_t\|^2 = \sum_{t=1}^H q\|s_t\|^2 + r\|K s_t + \Sigma_a^{\frac{1}{2}} \delta_t^a\|^2 \\
&\le \sum_{t=1}^H q\|s_t\|^2 + r(k\|s_t\| + \sigma_a\|\delta_t^a\|)^2 \\
&\le \sum_{t=1}^H q\|s_t\|^2 + 2rk^2\|s_t\|^2 + 2r\sigma_a^2\|\delta_t^a\|^2 \\
&\le (q + 2rk^2) \sum_{t=1}^H \lceil s_t \rceil^2 + 2r\sigma_a^2 \sum_{t=1}^H (d_t^a)^2,
\end{aligned}
\tag{13}
$$

where the triangle inequality and the fact $(a+b)^2 \le 2(a^2 + b^2)$ are used above. This bound on C is a quadratic polynomial in the d^s, d^a.

A.4   Combining bounds

Combining Section A.2 and Section A.3, we have

$$
\mathbf{tr}(\hat{g}^\top \hat{g}) \le P = C^2 \|\Sigma_a^{-1}\| \left(\sum_{t=1}^H d_t^a \lceil s_t \rceil \right)^2.
\tag{14}
$$

For brevity, let $\alpha = \|\Sigma_a^{-1}\|$, $\beta = q + 2rk^2$, $\gamma = 2r\sigma_a^2$, $\sigma = \sigma_s + b\sigma_a$, and $H' \triangleq \min\left\{ H, \frac{1}{1-f} \right\}$. H' reflects the stability of the closed-loop system: if highly stable ($f \ll 1$), we have $H' \ll H$, but when approaching instability ($f \to 1$), H' approaches H.

We expand P and substitute $d_t^s = 1$, $d_t^a = 1$ for all t to compute the sum of P's coefficients, using the notation $\mathbb{1}(\cdot)$ for the transformation of replacing all d with 1. We first bound $\mathbb{1}(C^2)$:

$$\mathbb{1}(C^2) \leq \left(\gamma H + \beta \sum_{t=1}^{H} \mathbb{1}(\lceil s_t \rceil)^2\right)^2 \leq \left(\gamma H + \beta \mu^2 \sum_{t=1}^{H} \left[f^{t-1}\|s_1\| + \sigma H'\right]^2\right)^2$$

$$\leq \left(\gamma H + 2\beta \mu^2 \sum_{t=1}^{H} \left[f^{2t-2}\|s_1\|^2 + \sigma^2 H'^2\right]\right)^2 \quad (15)$$

$$\leq \left(\gamma H + 2\beta \mu^2 (H'\|s_1\|^2 + \sigma^2 H H'^2)\right)^2$$

where the result is obtained by repeatedly applying the fact $(a+b)^2 \leq 2(a^2 + b^2)$. Next,

$$\mathbb{1}\left(\left(\sum_{t=1}^{H} d_t^a \lceil s_t \rceil\right)^2\right) \leq \left(\mu \sum_{t=1}^{H} f^{t-1}\|s_1\| + \sigma H'\right)^2 \quad (16)$$

$$\leq \mu^2 H'^2 (\|s_1\| + \sigma H)^2.$$

Finally,

$$\mathbb{1}(P) \leq \alpha \mu^2 H'^2 \left(\gamma H + 2\beta \mu^2 (H'\|s_1\|^2 + \sigma^2 H H'^2)\right)^2 (\|s_1\| + \sigma H)^2$$

$$= 4\|\Sigma_a^{-1}\|\mu^2 H'^2 \left(r\sigma_a^2 H + \mu^2(q + 2rk^2)(H'\|s_1\|^2 + \sigma^2 H H'^2)\right)^2 (\|s_1\| + \sigma H)^2 \quad (17)$$

$$\triangleq \overline{\mathbb{1}(P)}.$$

$\overline{\mathbb{1}(P)}$ is an order-8 polynomial in the d^s, d^a. The formula for the 8th moment of a $\chi(n)$ random variable is

$$\mathbb{E}[X^8] = n(n+2)(n+4)(n+6),$$

so $\mathbb{E}[\mathbf{tr}(\hat{g}^\top \hat{g})] \leq \overline{\mathbb{1}(P)} \cdot O(\bar{n}^4)$, where $\bar{n} = \max(n, m)$.

(Since we are summing the variances of $O(\bar{n}^2)$ random variables in $\hat{g}$, we would expect scaling of no less than $\bar{n}^2$ compared to the scalar case.)

B $\quad \mathbb{E}[\hat{g}^2]$'s lower bound in the scalar case

In this section, we make the assumption that $m = n = 1$, i.e., the states and actions are both scalars. The matrices A, B, K, Q, R are thus denoted as a, b, k, q, r here (notice that this notation r is different from the notation r_t for reward). Other notations follow the definitions in Appendix A. The aim in this appendix is to derive a lower bound for $\mathbb{E}[\hat{g}^2]$ in the special case when $0 \leq a + bk < 1$.

Lemma 5.

$$\mathbb{E}\left[\left(\sum_{t=1}^{H} \frac{\epsilon_t^a s_t}{\sigma_a^2}\right)^2 \left(\sum_{t=1}^{H} r_t\right)^2\right] \geq \mathbb{E}\left[\left(\sum_{t=1}^{H} \frac{\epsilon_t^a s_t}{\sigma_a^2}\right)^2\right] \mathbb{E}\left[\left(\sum_{t=1}^{H} r_t\right)^2\right]. \quad (18)$$

Proof. When $a + bk \geq 0$, all terms have positive coefficients. Rename the $2H$ random variables $\{\delta_t^s, \delta_t^a\}_{t=1}^{H}$ as $x_1, \ldots, x_{2H}$. We can see that a monomial on the right-hand side:

$$\mathbb{E}\left[x_1^{\alpha_1} \cdots x_{2H}^{\alpha_{2H}}\right] \mathbb{E}\left[x_1^{\beta_1} \cdots x_{2H}^{\beta_{2H}}\right]$$

corresponds to the monomial on the left-hand side:

$$\mathbb{E}\left[x_1^{\alpha_1 + \beta_1} \cdots x_{2H}^{\alpha_{2H} + \beta_{2H}}\right].$$

The x_i are independent zero-mean normal random variables, so the property $\mathbb{E}[x_i^\alpha]\mathbb{E}[x_i^\beta] \leq \mathbb{E}[x_i^{\alpha+\beta}]$ holds for any non-negative integers α, β. Combining with the fact that all coefficients are non-negative shows the lemma. $\qquad \square$

In the following two subsections, we lower bound the two terms on the right-hand side of Eq. (18) separately. Note that the first term can be simplified as

$$\left(\sum_{t=1}^{H} \frac{\epsilon_t^a s_t}{\sigma_a^2}\right)^2 = \left(\sum_{t=1}^{H} \frac{\delta_t^a s_t}{\sigma_a}\right)^2 = \frac{1}{\sigma_a^2}\left(\sum_{t=1}^{H} \delta_t^a s_t\right)^2.$$

B.1 Lower bounding $\mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a s_t\right)^2\right]$

By the expansion of s_t in Eq. (9), we have

$$\mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a s_t\right)^2\right] = \mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a \left((a+bk)^{t-1} s_1 + L_t\right)\right)^2\right]$$

$$\geq \mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a (a+bk)^{t-1} s_1\right)^2\right] + \mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a L_t\right)^2\right],$$

where $L_t \triangleq \sum_{\tau=1}^{t-1} (a+bk)^{t-1-\tau}(\sigma_s \delta_\tau^s + b\sigma_a \delta_\tau^a)$. The first term is equal to $\sum_{t=1}^{H} (a+bk)^{2t-2} s_1^2 = \frac{1-(a+bk)^{2H}}{1-(a+bk)^2} s_1^2$. The second term can be further written as

$$\mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a \sum_{\tau=1}^{t-1} (a+bk)^{t-1-\tau}(\sigma_s \delta_\tau^s + b\sigma_a \delta_\tau^a)\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{H} (\delta_t^a)^2 \left(\sum_{\tau=1}^{t-1} (a+bk)^{t-1-\tau}(\sigma_s \delta_\tau^s + b\sigma_a \delta_\tau^a)\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{H} \left(\sum_{\tau=1}^{t-1} (a+bk)^{t-1-\tau}(\sigma_s \delta_\tau^s + b\sigma_a \delta_\tau^a)\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{H} \sum_{\tau=1}^{t-1} (a+bk)^{2t-2-2\tau}(\sigma_s^2 + b^2\sigma_a^2)\right]$$

$$= (\sigma_s^2 + b^2\sigma_a^2)\mathbb{E}\left[\sum_{t=1}^{H} \frac{1-(a+bk)^{2t-2}}{1-(a+bk)^2}\right]$$

$$= (\sigma_s^2 + b^2\sigma_a^2)\left(\frac{H}{1-(a+bk)^2} - \frac{1-(a+bk)^{2H}}{(1-(a+bk)^2)^2}\right).$$

In the above several equalities, we use the independence among δ_t^a, δ_t^s. Combining two terms, we get

$$\mathbb{E}\left[\left(\sum_{t=1}^{H} \delta_t^a s_t\right)^2\right] \geq \frac{s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)}{1-(a+bk)^2} - \frac{(a+bk)^{2H}}{1-(a+bk)^2} s_1^2 - (\sigma_s^2 + b^2\sigma_a^2)\frac{1-(a+bk)^{2H}}{(1-(a+bk)^2)^2}$$

$$= \left(\frac{1-(a+bk)^{2H}}{1-(a+bk)^2} s_1^2 + \left(H - \frac{1-(a+bk)^{2H}}{1-(a+bk)^2}\right)\frac{\sigma_s^2 + b^2\sigma_a^2}{1-(a+bk)^2}\right)$$

$$\approx \begin{cases} \frac{s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)}{1-(a+bk)^2} & \text{when } H \gg \frac{1}{1-(a+bk)^2} \\ H\left(s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)\right) & \text{when } H \ll \frac{1}{1-(a+bk)^2} \end{cases}$$

$$\approx \min\left\{H, \frac{1}{1-(a+bk)^2}\right\}\left(s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)\right)$$

$$\approx H'\left(s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)\right).$$

B.2   Lower bounding $\mathbb{E}\left[\left(\sum_{t=1}^{H} r_t\right)^2\right]$

We first lower bound this term by

$$\mathbb{E}\left[\left(\sum_{t=1}^{H} r_t\right)^2\right] \geq \mathbb{E}\left[\sum_{t=1}^{H} r_t\right]^2 = \mathbb{E}\left[\sum_{t=1}^{H} qs_t^2 + r(ks_t + \sigma_a\delta_t^a)^2\right]^2$$

$$= \mathbb{E}\left[\sum_{t=1}^{H}(q + rk^2)s_t^2 + 2rks_t\sigma_a\delta_t^a + r\sigma_a^2\delta_t^{a2}\right]^2$$

$$= \left((q + rk^2)\mathbb{E}\left[\sum_{t=1}^{H} s_t^2\right] + Hr\sigma_a^2\right)^2$$

$$\mathbb{E}\left[\sum_{t=1}^{H} s_t^2\right] = \mathbb{E}\left[\sum_{t=1}^{H}\left((a + bk)^{t-1}s_1 + L_t\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{H}\left((a + bk)^{2t-2}s_1^2 + 2(a + bk)^{t-1}s_1 L_t + L_t^2\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{H}(a + bk)^{2t-2}s_1^2 + \sum_{t=1}^{H}\left(\sum_{\tau=1}^{t-1}(a + bk)^{t-1-\tau}(\sigma_s\delta_t^s + b\sigma_a\delta_t^a)\right)^2\right]$$

(the middle term $\mathbb{E}[(a + bk)^{t-1}s_1 L_t]$ is zero because L_t is a sum of zero-mean RVs)

$$= \frac{1 - (a + bk)^{2H}}{1 - (a + bk)^2}s_1^2 + \mathbb{E}\left[\sum_{t=1}^{H}\sum_{\tau=1}^{t-1}(a + bk)^{2t-2-2\tau}(\sigma_s^2 + b^2\sigma_a^2)\right]$$

$$= \frac{1 - (a + bk)^{2H}}{1 - (a + bk)^2}s_1^2 + (\sigma_s^2 + b^2\sigma_a^2)\left(\frac{H}{1 - (a + bk)^2} - \frac{1 - (a + bk)^{2H}}{(1 - (a + bk)^2)^2}\right)$$

$$\approx \min\left\{H, \frac{1}{1 - (a + bk)^2}\right\}\left(s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)\right)$$

$$\approx H'\left(s_1^2 + H(\sigma_s^2 + b^2\sigma_a^2)\right).$$

The second-to-last approximation is obtained similarly as in the previous subsection.

B.3   Combining

Combining the results in previous two subsections, we get the final lower bound on $\mathbb{E}[\hat{g}^2]$:

$$\mathbb{E}[\hat{g}^2] = \mathbb{E}\left[\left(\sum_{t=1}^{H}\frac{\delta_t^a s_t}{\sigma_a}\right)^2\left(\sum_{t=1}^{H} r_t\right)^2\right] \geq \mathbb{E}\left[\left(\sum_{t=1}^{H}\frac{\delta_t^a s_t}{\sigma_a}\right)^2\right]\mathbb{E}\left[\left(\sum_{t=1}^{H} r_t\right)^2\right]$$

$$\geq \Omega(c_1^2 c_2^2),$$

where (recall $\sigma \triangleq \sigma_s + b\sigma_a$)

$$c_1 = \frac{1}{\sigma_a}\left(|s_1| + \sigma\sqrt{H}\right)\sqrt{H'},$$

$$c_2 = r\sigma_a^2 H + (q + rk^2)(s_1^2 + \sigma^2 H)H'.$$