Variance Reduction in Actor Critic Methods (ACM)

Eric Benhamou

Variance Reduction in Actor Critic Methods (ACM)

Eric Benhamou*[1,2]

[1]A.I Square Connect, Lamsade PSL

July 23, 2019

Abstract

After presenting Actor Critic Methods (ACM), we show ACM are control variate estimators. Using the projection theorem, we prove that the Q and Advantage Actor Critic (A2C) methods are optimal in the sense of the L^2 norm for the control variate estimators spanned by functions conditioned by the current state and action. This straightforward application of Pythagoras theorem provides a theoretical justification of the strong performance of QAC and AAC most often referred to as A2C methods in deep policy gradient methods. This enables us to derive a new formulation for Advantage Actor Critic methods that has lower variance and improves the traditional A2C method.

keywords: Actor critic method, Variance reduction, Projection, Deep RL.

1 Introduction

Recently ACM have emerged as the state of the art methods in Deep Reinforcement Learning (DRL) problems, Jaderberg et al. (2016) or Espeholt et al. (2018). The introduction of Deep Reinforcement Learning methods have enabled to enlarge the scope of RL to a wide variety of domains through trial and error learning: atari games Mnih et al. (2016) Go : Silver et al. (2016), image recognition Zoph et al. (2017), physics tasks, including classic problems such as cartpole swing-up, dexterous manipulation, legged locomotion and car driving Lillicrap et al. (2015), traffic signal control Mannion et al. (2016a), electricity generator scheduling Mannion et al. (2016b), water resource management Mason et al. (2016), controlling robots in real environments Levine et al. (2016) and other transport problem Talpaert et al. (2019) and control and manipulation tasks for robotics Barth-Maron et al. (2018); Horgan et al. (2018).

Advantage Actor Critic (A2C) and Asynchronous Advantage Actor Critic (A3C) methods Mnih et al. (2016) have been originally derived from the REINFORCE algorithm Williams (1992). Sutton et al. (1999) has derived in full generality the policy gradient descent method while Mnih et al. (2016) (respectively Babaeizadeh et al. (2017)) have introduced the deep learning approach as well as capacity to distribute computations across CPUs (respectively GPUs). Currently, Deep Advantage Actor Critic methods achieve state of the art performance in Deep RL. The efficiency of these techniques has been mostly verified from an experimental point of view with multiple experiments on test-beds environment such Open AI gym and Atari games. However, theoretical considerations have been lacking sofar to express why these methods outperform other DRL methods. In this paper, we try precisely to answer this challenging question and to provide a theoretical justification for the efficiency of Advantage Actor Critic methods. After presenting the key motivation for A2C methods and in particular the concept of baseline, we prove that there exist optimal baselines according to the L^2 norm. We explain the concept of variance reduction and relate Q and Advantage Actor Critic methods to conditional expectations that can be interpreted as projection in the L^2 norm of the initial logathmic gradient in REINFORCE.

2 Related Work

Presenting and comparing ACM has been the subject of multiple papers such as Grondman et al. (2012), but also recent papers like Jaderberg et al. (2016) Lillicrap et al. (2015); Zoph et al. (2017); Barth-Maron et al. (2018); Horgan et al. (2018) or Espeholt et al. (2018). Actor critic are examined from a baseline point of view and the stream of quoted papers above aims at prov-

*{eric.benhamou@}{dauphine.fr}/{aisquareconnect.com}

ing mostly the convergence of these methods seen as a enhancement of REINFORCE.

Another variety of research has been to find efficient optimization from a GPU and CPU perspective as in Mnih et al. (2016), Babaeizadeh et al. (2017) and lately Espeholt et al. (2018). Mnih et al. (2016) have implied an asynchronous version of the A2C algorithm, having a set of actors that learns from a set of critics. Computation workload is spread against multiple CPUs allowing efficient computation parallelism. Babaeizadeh et al. (2017) have made further progress by spreading computation across multiple GPUs leveraging the fact that the learning process in actor critic method can be split in a few highly parallel tasks that can be efficiently done with GPUs. Espeholt et al. (2018) have designed a new algorithm entitled V trace for learners to be able to learn asynchronously from different workers.

However, the core of the idea of Actor Critic method, namely a variance reduction technique has not been very well examined and quite overlooked. This paper aims at precisely showing the real importance of variance reduction in these methods to emphasize the logic behind these methods and provides new techniques. This extends the work of **?** that only consider one dimensional control variates. We show that we can do multi dimensional control variates. Our work also extends the work of Schulman et al. (2016) that provides a generalized advantage estimation using trust region optimization procedure for both the policy and the value function, which are represented by neural networks.

3 Background

We consider the standard reinforcement learning framework. A learning agent interacts with an envi-ronment $\mathcal{E}$ to decide rational or optimal actions and receives in returns rewards. These rewards are not necessarily only positive. These rewards act as a feedback for finding the best action. Using the established formalism of Markov decision process, we assume that there exists a discrete time stochastic control process represented by a 4-tuple defined by $(\mathcal{S}, \mathcal{A}, P_a, R_a)$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of actions, $P_a(s, s') = \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$ the transition probability that action a in state s at time t will lead to state s' and finally, $R_a(s, a)$ the immediate reward received after state s and action a.

The requirement of a Markovian state is not a strong constraint as we can stack observations to enforce that the Markov property is satisfyed.

Following Mnih et al. (2016) or Jaderberg et al. (2016), we introduce the concept of observations and pile them to coin states. In this setting, the agent perceives at time t an observation o_t along with a reward r_t. The agent decides an action a_t. The agent's state s_t is a function of its experience until time t, $s_t = f(o_1, r_1, a_1, ..., o_t, r_t)$. This setup guarantees that states are Markovian. In practice, we do not keep track of the full history but only a limited set of historical experiences. The n-step return $R_{t:t+n}$ at time t is defined as the discounted sum of rewards, $R_{t:t+n} = \sum_{i=1}^{n} \gamma^i r_{t+i}$, where $\gamma \in [0, 1)$ is the discounted factor.

The value function is the expected return from state s, $V^\pi(s) = \mathbb{E}\left[R_{t:\infty} | s_t = s, \pi\right]$, when actions are selected accorded to a policy $\pi(a|s)$. The action-value function $Q^\pi(s, a) = \mathbb{E}\left[R_{t:\infty} | s_t = s, a_t = a, \pi\right]$ is the expected return following action a from state s. The goal of the agent is to find a policy π that maximizes the value function.

4 Variance Reduction

When states and or actions are in high dimensions, we parametrize our policy by parameters denoted θ. Typically these parameters are the ones of the parameters of a deep network (the weight of all layers of our deep network). The intuition of ACM is to leverage a policy gradient descent with a reduced variance. Let us show this precisely. Recall that the policy gradient is given by the policy logarithmic gradient weighted by the sum of future discounted rewards (see Williams (1992) and Sutton et al. (1999))

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_\tau \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | \, s_t) \underbrace{\left(\sum_{t'=t+1}^{T} \gamma^{t'-t} r_{t'} \right)}_{R_t} \right] \\ &= \mathbb{E}_\tau \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R_t \right] \end{aligned} \tag{1}$$

where τ represents a trajectory, R_t the sum of future discounted rewards, and $\gamma \in [0, 1)$ the discount factor. Hence, equation (1) shows that we update the policy deep network parameters through Monte Carlo updates computed as an expectation. We should stop for a while on this expectation as this is a critical part in the update. If the estimation of this expectation is not very accurate because of a large variance of our

estimator provided by the standard empirical mean, we would incur high variability in our gradient update, hence a slow converging gradient policy method. It therefore makes a lot of sense to see if we could find another expression of our policy gradient with lower variance. This approach of finding a modified expression inside the expectation that has the same expected value but a lower variance is referred to as variance reduction Hammersley (1964). A typical method variance reduction method is to use control variate(s) (see for instance Ross (2002)). A lower variance in the gradient will produce less noisy gradient and cause less unstable learning leading to a policy distribution skewing to the optimal direction more rapidly.

4.1 Control Variate

To present control variate, let us make thing quite general. Suppose we try to estimate a quantity μ defined as an expectation $\mu = \mathbb{E}[\hat{m}]$ of an estimator $\hat{m}$ and suppose we know another statistic (or estimator) $\hat{t}$ such that we not only know its expectation $\tau = \mathbb{E}[\hat{t}]$, but also its correlation with our initial estimator denoted by $\rho_{\hat{m},\hat{t}}$. We can build an unbiased and better estimator of μ as follows. We compute the 'control variate' estimator of $\hat{m}$ by subtracting a zero expectation term: $\hat{m'} = \hat{m} - \alpha(\hat{t} - \tau)$ for $\alpha \in \mathbb{R}$. We have the following control variate proposition

Proposition 4.1. *Optimal Control Variate - The control variate estimator $\hat{m'}$ is unbiased for any value of α. Among all possible values of α, the optimal one (in the sense that it produces the estimator with minimum variance) is given by*

$$\alpha^* = \frac{\mathbb{C}\text{ov}(\hat{m}, \hat{t}\,)}{\mathbb{V}\text{ar}(\hat{t})} = \frac{\sigma_{\hat{m}}}{\sigma_{\hat{t}}}\rho_{\hat{m},\hat{t}}$$

The corresponding control variate estimator is given by $\hat{m^*} = \hat{m} - \alpha^*(\hat{t} - \tau)$ and has a variance given by

$$\mathbb{V}\text{ar}(\hat{m^*}) = (1 - \rho^2_{\hat{m},\hat{t}})\sigma^2_{\hat{m}} \leq \sigma^2_{\hat{m}}$$

Hence the best control variate estimators are obtained for highly positively correlated ($\rho_{\hat{m},\hat{t}} \approx 1$) or negatively correlated ($\rho_{\hat{m},\hat{t}} \approx -1$) control variates

Proof. Refer to Appendix section 7.1 □

Intuitively, the more correlated (positively or negatively) the estimators $\hat{m}$ and $\hat{t}$, the better we can exploit the knowledge of our control variate zero expectation estimator $\hat{t} - \tau$ to reduce the variance of our initial estimator $\hat{m}$. If we stop for a minute, this is quite trivial. For our control variates, we know the true expectation. If by any chance our estimator is very correlated (positively or negatively) to our control variates, we can exploit this knowledge to correct our estimator.

As a matter of fact, through control variate, we exploit information about the errors in estimates of known quantities ($\hat{t} - \tau$) to reduce the error of an estimate of an unknown quantity ($\hat{m}$). We can also analyze the optimal control variate weight $\alpha^* = \frac{\mathbb{C}\text{ov}(\hat{m},\hat{t}\,)}{\mathbb{V}\text{ar}(\hat{t})}$ as a regression coefficient of our unknown estimator $\hat{m}$ over the space of linear combination of the zero expectation estimator $\hat{t} - \tau$. Hence control variate consists in just focusing on the orthogonal part of our unknown estimator $\hat{m}$ against the space of linear combination of the zero expectation estimator $\hat{t} - \tau$. If by any chance our unknown estimator is highly correlated to the control variate, this orthogonal part is close to zero and we achieve a much lower variance estimator.

In practice, in deep RL, we know neither the correlation between the two estimators $\rho_{\hat{m},\hat{t}}$ nor the variances of our two estimators: $\sigma_{\hat{m}}$ or $\sigma_{\hat{t}}$. Hence, instead of being able to define strictly a control variate estimator, we create a pseudo control variate estimator given by

$$\hat{m} - \hat{t} \tag{2}$$

provided $\mathbb{E}[\hat{t}] = 0$. The latter formulation takes a control variate coefficient of $\alpha = 1$, which implicitly assumes that $\mathbb{C}\text{ov}(\hat{m}, \hat{t}\,) \approx \mathbb{V}\text{ar}(\hat{t})$. We will use this setting to interpret various Actor Critic (AD) methods as control variate estimators for standard AC method for policy gradients. This is summarized by the proposition below:

Proposition 4.2. *Actor Critic Methods - The following estimators of the policy gradient are unbiased and can be analyzed as control variate estimators of REINFORCE policy gradient estimator given by $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) R_t\right]$:*

- $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) Q(s,a)\right]$ *(Q-AC)*
- $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) A(s,a)\right]$ *(A-AC)*
- $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) TD(s)\right]$ *(TD-AC)*

where the Advantage function (A) is defined as

$$A(s,a) = \mathbb{E}_{\pi_\theta}[r_{t+1} + \gamma V(s_{t+1}) \mid s_t = s, a_t = a] - V(s_t)$$

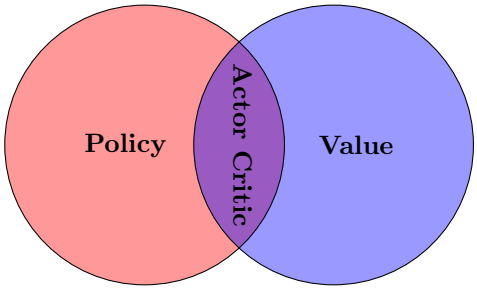

Figure 1: The different RL methods. ACM combine policy and value computations, making them mixed methods

and the Temporal Difference function (TD) as

$$TD(s_t) = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

Proof. Refer to Appendix section 7.2 □

Traditionally, methods for solving Reinforcement Learning (RL) are either categorized as policy method if they aim to find the optimal policy $\pi^\star$ or as value methods if they aim to find the optimal 'Q' function. Somehow, Q-AC and AAC methods aim to find the optimal policy thanks to gradient ascent computation but use in their gradient ascent term a 'Q' function, making these methods a mix between policy and value methods. This is summarized by figure 1.

4.2 Conditional expectation, projection and optimality

Sofar, we have revisited AC methods as control variates. There is however a strong connection with conditional expectation and projection. Let us make the link. But let us first recall a basic property of conditional expectation that states that the conditional expectation with respect to a sub σ-algebra $\mathcal{G} includedin \mathcal{F}$ of a stochastic variable X, $\mathcal{L}^2$ measurable on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, is the best prediction of the sub-space spanned by this sub σ-algebra:

Proposition 4.3. *Conditional expectation and Pythagoras - Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, $X : \Omega \to \mathbb{R}^n$ a random variable on that probability space square integrable and $\mathcal{G} \subseteq \mathcal{F}$ is a sub σ-algebra of $\mathcal{F}$, then we have that*

- *$X - \mathbb{E}[X \mid \mathcal{G}]$ is orthogonal to any element Y of $\mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ where $\mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ is the space of random variable on the sub σ-algebra $(\Omega, \mathcal{G}, \mathcal{P})$ that are square integrable.*
- *$\mathbb{E}[X \mid \mathcal{G}]$ is the best prediction in the sense that $\mathbb{E}[X \mid \mathcal{G}]$ minimizes its variance with X: $\mathbb{E}[(X - Y)^2]$ among any element $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$.*

Proof. Refer to Appendix section 7.3 □

Combining proposition 4.1, 4.2 and 4.3 enables us comparing the various ACM from a variance point of view. It is worth looking at the intuition of the function involved in the different ACM. TD-AC relies on the Temporal Difference which is the projection of the cumulated discounted rewards on the sub σ-algebra generated by the knowledge of the state s_{t+1} while Q-AC relies on the action value 'Q' function, which is its projection on the sub σ-algebra generated by the knowledge of the state s_t, at and the A-AC on the advantage function which is the difference between the action value 'Q' function and the value function which is of the cumulated discounted rewards on the sub σ-algebra generated by the knowledge of the state s_t. Intuitively, as the various sub σ algebras are bigger (in the sense that each one is contained by the next one), the corresponding AC methods should become more effective, meaning TD-AC should be improved by Q-AC that should in turn be improved by the A-AC method. This is the subject of proposition 4.4

Proposition 4.4. *AC methods Comparison - From a variance point of view, TD-AC is less efficient than A-AC, and REINFORCE is less efficient than Q-AC.*

Proof. Refer to Appendix section 7.4 □

5 Towards new AC methods

In order to create new AC methods, it is useful to notice a property of the policy gradient computation that provides additional control variates.

Proposition 5.1. *If a function $\Phi(s, a)$ is such that when integrating with respect to the policy $\pi_\theta(s, a)$, it does not depend on the parameter θ, which means that $\nabla_\theta \int \pi_\theta(s, a)\Phi(s, a) = 0$, then its gradient policy term is null:*

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s, a)\Phi(s, a)\right] = 0 \tag{3}$$

Proof. Refer to Appendix section 7.5 □

Remark 5.1. *In particular if the function $\Phi(s, a)$ writes as a function of s only: $\Phi(s, a) = \Psi(s)$ and if*

$\Psi(s)$ is stationary in the sense that $\int \pi_\theta(s,a)\Psi(s)d\theta = \Psi(s)$, than its policy gradient term is equal to zero:

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a)\Psi(s)\right] = 0 \tag{4}$$

Typical stationary functions are the state value function $V(s)$ and the constant function: 1

Equipped with these additional control variates, it is useful to examine multi dimensional control variates estimators that is the subject of the following proposition

Proposition 5.2. *Multi Dimensional Control Variates Estimators - Let us have a collection of random variables $\hat{t}_1, \ldots, \hat{t}_d$ for which we know the expectation $\tau_i = \mathbb{E}[\hat{t}_i]$ (for any $i = 1, \ldots, d$). Let us denote by*

$$T = (\hat{t}_1 - \mathbb{E}[\hat{t}_1], \ldots, \hat{t}_d - \mathbb{E}[\hat{t}_d])^T$$

the vector of control variates, λ a d-dimensional real vector and build the multi dimensional control variates estimator as follows:

$$\hat{m'} = \hat{m} - \lambda^T T \tag{5}$$

As in the one dimensional case, the control variate estimator $\hat{m'}$ is unbiased for any value of $\lambda \in \mathbf{R}^d$. Assuming that $\mathbb{E}[T\,T^T]$ is non singular, among all the possible values of λ, the optimal one (in the sense that it produces the estimator with minimum variance) is given by

$$\lambda^* = \mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T] \tag{6}$$

The corresponding control variate estimator is given by $\hat{m^*} = \hat{m} - (\lambda^\star)^T T$ and has a variance given by

$$\mathbb{V}\mathrm{ar}(\hat{m^*}) = \mathbb{V}\mathrm{ar}(\hat{m}) - \mathbb{E}[\hat{m}T]^T\mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T] \tag{7}$$

Proof. Refer to Appendix section 7.6 □

Equations (6) and (7) are just generalization of the one dimensional case. The condition that $\mathbb{E}[T\,T^T]$ is non singular means that we have control variates that are independent in the sense that none of them is a linear combination of some of the other control variates. This condition is important as to use multi dimensional control variates we really need to find control variates that operate in another dimension space.

6 Conclusion

In this paper, we have revisited AC methods. We have shown that these methods can be interpreted as control variate estimators of REINFORCE. We have also proved using the property of conditional expectation, that the Q and Advantage Actor Critic are optimal control variate estimators. We have invented a new method that combines optimally the general advantage function to establish a new Actor Critic method that is optimal from a control variate point of view.

References

Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., and Kautz, J. (2017). Reinforcement learning thorugh asynchronous advantage actor-critic on a gpu. In *ICLR*.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., TB, D., Muldal, A., Heess, N., and Lillicrap, T. (2018). Distributional policy gradients. In *International Conference on Learning Representations*.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1407–1416, Stockholmsmässan, Stockholm Sweden. PMLR.

Grondman, I., Busoniu, L., Lopes, G. A. D., and Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *Trans. Sys. Man Cyber Part C*, 42(6):1291–1307.

Hammersley, J. M., D. C. H. (1964). *Monte Carlo Methods*. John Wiley & Sons, New York.

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., van Hasselt, H., and Silver, D. (2018). Distributed prioritized experience replay. In *International Conference on Learning Representations*.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *CoRR*, abs/1611.05397.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(1):1334–1373.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *ArXiv e-prints*.

Mannion, P., Duggan, J., and Howley, E. (2016a). *An Experimental Review of Reinforcement Learning Algorithms for Adaptive Traffic Signal Control*, pages 47–66.

Mannion, P., Mason, K., Devlin, S., Duggan, J., and Howley, E. (2016b). Multi-objective dynamic dispatch optimisation using multi-agent reinforcement learning.

Mason, K., Mannion, P., Duggan, J., and Howley, E. (2016). Applying multi-agent reinforcement learning to watershed management.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *ICML*, volume 48, pages 1928–1937, New York, New York, USA. PMLR.

Ross, S. (2002). *Simulation*. Academic Press.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pages 1057–1063, Cambridge, MA, USA. MIT Press.

Talpaert, V., Sobh, I., Kiran, B. R., Mannion, P., Yogamani, S., El-Sallab, A., and Perez, P. (2019). Exploring applications of deep reinforcement learning for real-world autonomous driving systems. *arXiv e-prints*.

Williams, R. J. (1992). *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, volume 8. Springer.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2017). Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012.

7 Appendix

7.1 Proof of proposition 4.1

We have

$$\mathbb{E}[\hat{m'}] = \mathbb{E}[\hat{m}] - \alpha\mathbb{E}[\hat{t} - \tau] = \mathbb{E}[\hat{m}] = \mu,$$

since $\mathbb{E}[\hat{t}] = \tau$. This proves that the control variate estimator is unbiased for any value of α. The variance of this estimator is easy to compute and is given by:

$$\mathbb{V}\text{ar}(\hat{m'}) = \mathbb{V}\text{ar}(\hat{m}) - 2\alpha\mathbb{C}\text{ov}(\hat{m}, \hat{t}) + \alpha^2\mathbb{V}\text{ar}(\hat{t})$$

that is a second order parabola function of α that is minimum for

$$\alpha^* = \frac{\mathbb{C}\text{ov}(\hat{m}, \hat{t}\,)}{\mathbb{V}\text{ar}(\hat{t})} = \frac{\sigma_{\hat{m}}}{\sigma_{\hat{t}}}\rho_{\hat{m},\hat{t}}$$

Its minimum value is given by

$$\mathbb{V}\text{ar}(\hat{m^*}) = (1 - \rho^2_{\hat{m},\hat{t}})\sigma^2_{\hat{m}} \leq \sigma^2_{\hat{m}}$$

that is closed to zero for highly positively correlated ($\rho_{\hat{m},\hat{t}} \approx 1$) or negatively correlated ($\rho_{\hat{m},\hat{t}} \approx -1$) control variates. □

7.2 Proof of proposition 4.2

Let us tackle one by one the various estimators given in proposition 4.2.

The **Q Actor Critic (Q-AC)** estimator is given by $\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) Q(s,a)\right]$. It writes also as

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) (\underbrace{R(s)}_{\hat{m}} - \underbrace{(R(s) - Q(s,a))}_{\hat{t}})\right]$$

which shows that it is a control variate estimator (as explained in equation (2)) provided we prove that

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a)(Q(s,a) - R(s))\right] = 0 \quad (8)$$

The equation (8) is trivially verified as one of the definitions of the state action value function $Q(s,a)$ (often referred to as the 'Q' function) is the following:

$$Q(s,a) = \mathbb{E}_{\pi_\theta}[R(s) \mid s_t = s, a_t = a]$$

The law of total expectation states that if X is a random variable and Y any random variable on the same probability space, then $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ or in other words $\mathbb{E}[\mathbb{E}[X \mid Y] - X] = 0$, which concludes the proof for the Q Actor Critic (Q-AC) method with $X = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) R(s)$ and $Y = (s_t = s, a_t = a)$.

As for the **Advantage Actor Critic (A-AC)** method, the same reasoning shows that it suffices to prove that

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a)(A(s,a) - R(s))\right] = 0$$

to show that this is also a control variate method. Recall that the advantage function is given by $A(s,a) = Q(s,a) - V(s)$. Using the result previously proved for the Q AC method (equation (8)), it suffices to prove that

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) V(s)\right] = 0 \quad (9)$$

But this is a straight consequence of the more general proposition 5.1 and its remark 5.1.

Finally, let us prove that the **TD Actor Critic (TD-AC)** method is also a control variate. Recall that the Temporal Difference term is given by

$$TD(s_t) = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

Using equation (9), it suffices to prove that

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a)(r_{t+1} + \gamma V(s_{t+1}) - R(s_t))\right] = 0 \quad (10)$$

This is again a straightforward application of the law of total expectation as

$$V(s_{t+1}) = \mathbb{E}_{\pi_\theta}[\sum_{s=t+2}^{T} \gamma^{s-(t+2)} r_s \mid s_{t+1}]$$

while

$$R(s_t) = \sum_{s=t+1}^{T} \gamma^{s-(t+1)} r_s$$

Hence we can apply the law of total expectation with $X = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s,a) \sum_{s=t+2}^{T} \gamma^{s-(t+1)} r_s$ and $Y = s_{t+1}$. This concludes the proof. □

7.3 Proof of proposition 4.3

Let us first prove that for any $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$, we have

$$\mathbb{E}[\dot{Y}(X - \mathbb{E}[X \mid \mathcal{G}]] = 0$$

This is straight application of the law of total expectation (also referred to as the law of iterated expectation or also the tower property) as follows

$$\begin{aligned} & \mathbb{E}[Y \cdot (X - \mathbb{E}[X \mid \mathcal{G}])] \\ = \; & \mathbb{E}[\mathbb{E}[Y \cdot (X - \mathbb{E}[X \mid \mathcal{G}]) \mid \mathcal{G}]] \\ = \; & \mathbb{E}\left[Y \cdot \Big(\underbrace{\mathbb{E}[X \mid \mathcal{G}] - \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{G}]}_{=0} \Big) \right] \\ = \; & 0 \end{aligned}$$

which proves the orthogonality.

For $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$, we have

$$\begin{aligned} & \mathbb{E}[(X - Y)^2] \\ = \; & \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{G}] + \mathbb{E}[X \mid \mathcal{G}] - Y)^2] \\ = \; & \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{G}])^2 + (\mathbb{E}[X \mid \mathcal{G}] - Y)^2] \\ & +2 \underbrace{\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{G}])(\mathbb{E}[X \mid \mathcal{G}] - Y)]}_{=0} \\ \geq \; & \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{G}])^2] \end{aligned}$$

which concludes the proof that the squared L^2 norm of $X - Y$ for any element Y of $\mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ is lower bounded by the squared L^2 norm of $X - \mathbb{E}[X \mid \mathcal{G}]$ □

7.4 Proof of proposition 4.4

As explained in the proposition 4.1, the variance of the control variate is due to the residuals computed as the initial estimator minus the control variate. The control variate is the orthogonal projection of the initial estimator on the linear subspace spanned by the control variate. The proposition 4.3 shows that within the possible sub σ the conditional expectation is the orthogonal projection and is the best estimator. Recall that the Temporal difference writes as

$$r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{11}$$

Recall also that the value function is a conditional expectation

$$V(s_{t+1}) = \mathbb{E}[R_{t+1} \mid s_{t+1}] \tag{12}$$

while the state action value 'Q' function is also a conditional expectation but with respect to s_t, a_t:

$$Q(a_t, s_t) = \mathbb{E}[R_t \mid s_t, a_t] \tag{13}$$

Last but not least, Advantage function writes as

$$A(a_t, s_t) = Q(a_t, s_t) - V(s_t) \tag{14}$$

From equations (11), (13) and (14), we can deduce that A AC has lower variance than TD AC as the corresponding sub σ-algebra obtained by conditioning with respect to s_t, a_t, used for A-AC, contains the one obtained by conditioning with respect to s_{t+1} and used for TD-AC.

As 'Q' function is a conditional expectation of the future discounted rewards, Q-AC should perform better (have a lower variance) than just REINFORCE.

7.5 Proof of proposition 5.1

Proof. : Recall that the logarithmic gradient writes as

$$\nabla_\theta \log \pi_\theta(s, a) = \frac{\nabla_\theta \pi_\theta(s, a)}{\pi_\theta(s, a)}$$

Assuming smooth functions such that we can interchange expectation (integration) and gradient (derivation), we have the following:

$$\begin{aligned} & \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s, a) \Phi(s, a) \right] \\ = \; & \sum_{t=0}^{T-1} \mathbb{E}_{\pi_\theta}\left[\frac{\nabla_\theta \pi_\theta(s, a)}{\pi_\theta(s, a)} \Phi(s, a) \right] \\ = \; & \sum_{t=0}^{T-1} \int \frac{\nabla_\theta \pi_\theta(s, a)}{\pi_\theta(s, a)} \Phi(s, a) \pi_\theta(s, a) d\theta \\ = \; & \sum_{t=0}^{T-1} \int \nabla_\theta \pi_\theta(s, a) \Phi(s, a) d\theta \\ = \; & \sum_{t=0}^{T-1} \nabla_\theta \int \pi_\theta(s, a) \Phi(s, a) d\theta \\ = \; & 0 \end{aligned}$$

which concludes the proof □

7.6 Proof of proposition 5.2

We can provide multiple proofs and interpretation of this result.

proof 1 Let us use traditional variation calculus. It is immediate that for any value $\lambda \in \mathbf{R}^d$ it is unbiased as the additional term has a null expectation: $\mathbb{E}[\lambda^T T] = 0$. We can compute the variance of the control variate estimator $\widetilde{m}(\lambda) = \hat{m} - \lambda^T T$, given by

$$\mathbb{V}\text{ar}(\widetilde{m}(\lambda)) = \mathbb{V}\text{ar}(\hat{m}) - 2\lambda^T \mathbb{E}[\hat{m} T] + \lambda^T \mathbb{E}[T\, T^T] \lambda \tag{15}$$

Assuming the covariance matrix $\mathbb{E}[T\, T^T]$ is non singular, our minimum variance problem lies in finding the

minimum of a defined parabolla, hence its minimum is given by first order optimality (see Boyd and Vandenberghe (2004) A.13)

$$\lambda^* = \mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T] \tag{16}$$

with the minimum given by:

$$\mathbb{V}\mathrm{ar}(\widetilde{m}(\lambda^*)) = \mathbb{V}\mathrm{ar}(\hat{m}) - \mathbb{E}[\hat{m}T]^T\mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T]$$

proof 2 Another way to demonstrate this result is to look at the L^2 space of all square integrable random variables defined on the same probability space as $\hat{m}$ equiped with the canonical inner product $< X_1, X_2 > = \mathbb{E}[X_1X_2]$ and the implied Hilbertian norm $\|X\| = \mathbb{E}[X^2]^{1/2}$ for any $X, X_1, X_2 \in L^2$. Let $\mathcal{G}$ be the linear compact and closed space subspace spanned by any linear combination of C: $W \in L^2$ such that $W = \mu^T T$ for some $\mu \in \mathbb{R}^d$. Using the fact that the expectation of $\widetilde{m}(\lambda)$ does not depend on λ, the variance minimum in $\mathbb{R}^d$ can be cast as a minimum distance problem of $\hat{m}$ with $\mathcal{G}$. This is because

$$\begin{aligned}
&\arg\min_{\lambda\in\mathbb{R}^d} \mathbb{V}\mathrm{ar}(\widetilde{m}(\lambda)) \\
&= \arg\min_{\lambda\in\mathbb{R}^d} \mathbb{E}[\widetilde{m}(\lambda)^2] - (\mathbb{E}[\widetilde{m}(\lambda)])^2 \\
&= \arg\min_{\lambda\in\mathbb{R}^d} \mathbb{E}[\widetilde{m}(\lambda)^2] \\
&= \arg\min_{\lambda\in\mathbb{R}^d} \|\hat{m} - \lambda^T T\| \\
&= \arg\min_{g\in\mathcal{G}} \|\hat{m} - g\|
\end{aligned}$$

This is nice property as the minimum distance problem can be solved with the Hilbert space projection theorem that states that the closest point of $\mathcal{G}$ to $\hat{m}$ is given by its orthogonal projection $g^* \in \mathcal{G}$. This orthogonal projection $g^* = (\lambda^*)^T T$ is characterized uniquely by

$$< \hat{m} - g^*, g >= 0 \quad \text{for any} \quad g \in \mathcal{G}$$

which writes as $\mathbb{E}[\hat{m}g] - \mathbb{E}[g^*g] = 0$. Hence for any $\lambda \in \mathbb{R}^d$, we have $\mathbb{E}[\hat{m}\,T^T]\lambda = (\lambda^*)^T\mathbb{E}[T\,T^T]\lambda$, which implies $\mathbb{E}[\hat{m}\,T^T] = (\lambda^*)^T\mathbb{E}[T\,T^T]$ or equivalently

$$(\lambda^*)^T = \mathbb{E}[\hat{m}\,T^T]\mathbb{E}[T\,T^T]^{-1}$$

leading to the following solution:

$$\lambda^* = \mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T]$$

proof 3 Let us show how we can decouple the problem into d one dimensional control variate problems using diagonalization. We are looking for the minimum variance control variate spanned by our initial control variates basis $(\hat{t}_1 - \mathbb{E}[\hat{t}_1], \ldots, \hat{t}_d - \mathbb{E}[\hat{t}_d])^T$. Obviously, we can use any equivalent basis. In particular, if we use Gram Schmidt orthogonalization, we can ensure that the d components of our control variate basis are orthogonal for the implied L^2 inner product. If some of the component are non independent we can retrieve the corresponding control variate vectors to ensure they are all independent. The orthogonality of the d components of our control variate basis ensures that the covariance matrix $\mathbb{E}[T\,T^T]$ is symmetric and non negative definitive. Hence, because of the Takagi's factorization, there exists a diagonal matrix D with non negative diagonal terms and U an unitary matrix $UU^T = I$ such that $\mathbb{E}[T\,T^T] = U^TDU$. Let define $W = U, T$. We have $\mathbb{E}[W\,W^T] = D$, so that the components of W are orthogonal with variance given by diagonal terms $W_{ii} = D_{ii} > 0$ for $i = 1, \ldots, d$. Because of the equivalence of basis, our optimization problem decouples as it can be reformulated as follows:

$$\mathrm{minimize}_{\gamma\in\mathbb{R}^d}\mathbb{V}\mathrm{ar}(\hat{m} - \gamma^T W)$$

which is equivalent to $\mathrm{minimize}_{\gamma_i\in\mathbb{R}}\mathbb{V}\mathrm{ar}(\hat{m}_{ii} - \gamma W_{ii})$ for any $i = 1, \ldots, d$, where $\hat{m}_{ii}$ is the i coordinate of the random variable $\hat{m}$ in the basis implied by W that is spanned by the orthorgonal vectors W_i that is fill with zero except for coordinate i equal to W_{ii}. These decoupled and independent optimization problems are now just one dimensional control variate optimization problem whose solution are

$$\gamma_i^* = \frac{\mathbb{E}(\hat{m}_{ii}, W_{ii})}{\mathbb{E}(W_{ii}^2)} = \frac{< \hat{m}, W_i >}{< W_i, W_i >}$$

so that the solution is given by

$$(\gamma^*)^T W = \sum_{i=1,\ldots,d} \gamma_i^* W_i = \sum_{i=1,\ldots,d} \frac{< \hat{m}, V_i >}{< W_i, W_i >} V_i$$

Noticing that the term $\frac{<\hat{m},V_i>}{<V_i,V_i>}$ is the i^{th} term of the $\mathbb{E}[\hat{m}\,T^T]\mathbb{E}[T\,T^T]^{-1}$, this can be rewritten as

$$(\gamma^*)^T = \mathbb{E}[\hat{m}\,T^T]\mathbb{E}[T\,T^T]^{-1}$$

leading to the following solution:

$$\lambda^* = \mathbb{E}[T\,T^T]^{-1}\mathbb{E}[\hat{m}T]$$

which concludes the third proof. □