Reducing Variance in Gradient Bandit Algorithm using Antithetic Variates Method

Sihao Yu, Jun Xu*, Yanyan Lan, Jiafeng Guo, Xueqi Cheng
[1] University of Chinese Academy of Sciences, Beijing, China
[2] CAS Key Lab of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
yusihao@software.ict.ac.cn,{junxu,lanyanyan,guojiafeng,cxq}@ict.ac.cn

ABSTRACT

Policy gradient, which makes use of Monte Carlo method to get an unbiased estimation of the parameter gradients, has been widely used in reinforcement learning. One key issue in policy gradient is reducing the variance of the estimation. From the viewpoint of statistics, policy gradient with baseline, a successful variance reduction method for policy gradient, directly applies the control variates method, a traditional variance reduction technique used in Monte Carlo, to policy gradient. One problem with control variates method is that the quality of estimation heavily depends on the choice of the control variates. To address the issue and inspired by the antithetic variates method for variance reduction, we propose to combine the antithetic variates method with traditional policy gradient for the multi-armed bandit problem. Furthermore, we achieve a new policy gradient algorithm called Antithetic-Arm Bandit (AAB). In AAB, the gradient is estimated through coordinate ascent where at each iteration gradient of the target arm is estimated through: 1) constructing a sequence of arms which is approximately monotonic in terms of estimated gradients, 2) sampling a pair of antithetic arms over the sequence, and 3) re-estimating the target gradient based on the sampled pair. Theoretical analysis proved that AAB achieved an unbiased and variance reduced estimation. Experimental results based on a multi-armed bandit task showed that AAB can achieve state-of-the-art performances.

KEYWORDS

Policy gradient; Antithetic variates; Coordinate gradient

ACM Reference Format:
Sihao Yu, Jun Xu*, Yanyan Lan, Jiafeng Guo, Xueqi Cheng. 2018. Reducing Variance in Gradient Bandit Algorithm using Antithetic Variates Method. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210068

* Corresponding author: Jun Xu

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210068

1 INTRODUCTION

Reinforcement learning, including the Multi-Armed Bandit(MAB) [7] and Markov Decision Process(MDP) [5], have been successfully used in variant machine learning applications recently. Among the algorithms that solve the reinforcement learning problems, policy gradient [13] has shown its advantages in effectiveness in high-dimensional/continuous action spaces, fast convergence rate, and handling stochastic policies etc. Roughly speaking, policy gradient relies upon optimizing parametrized policies (a distribution over the agent actions) with respect to the expected return (long-term cumulative reward) by gradient ascent.

To calculate the parameter gradients at each optimization iteration, policy gradient algorithms such as REINFORCE [12, 13] usually adopt the Monte Carlo method [9] to estimate the expectation of the gradient. The gradient estimated by the Monte Carlo method is unbiased but usually has large variance, which hurts the efficiency and effectiveness of the traditional policy gradient algorithm. How to reduce the variance of the estimated gradient becomes a key issue in policy gradient algorithms.

A number of research has been conducted to reduce the gradient variance in policy gradient. For example, Policy gradient with baseline [3] is commonly used in real reinforcement learning tasks. In the method, a baseline variate, which is designed as the averaged rewards of the history steps, is first designed. Then, the real reward of the action minus the baseline is used as the reward for the gradient estimation and parameter updating. It also shows that variance of the newly estimated gradients is reduced while its expectation is identical to that of the traditional policy gradient. More methods on variance reduction for policy gradient please referred to [1, 8, 11]

From the viewpoint of statistics, the policy gradient with baseline is a direct application of control variates method [2], a variance reduction approach in Monte Carlo Method, to improve the traditional policy gradient. The baselines are implementations of the control variates in the reinforcement learning environment. In general, designing reliable control variates is critical for the success of control variates method. Inappropriate setting of the control variates may result in the raising of variance and hurt the estimation. When applied to reinforcement learning, though the policy gradient with baseline heuristically constructs the baselines, which is far away from the ideal control variates, it is difficult to achieve its optimal effect.

To get rid of this problem, antithetic variates method [4] is proposed. Every time antithetic variates method draws a pair of antithetic samples for the estimation. Since one antithetic sample in the pair is easily derived from another, the auxiliary functions (e.g., the control variates) is not a mandatory anymore.

Based on the observation, we propose a novel policy gradient method which uses antithetic variates to improve policy gradient for MAB. The proposed method, referred to as Antithetic-Arm Bandit (AAB), estimates the parameter gradients through sampling a pair of antithetic arms at each time. To achieve this, AAB adopts the coordinate ascent framework for the optimization where each coordinate corresponds an arm. At each iteration, the arms are sorted according to their estimated gradients. After that, a pair of antithetic arms are sampled on the basis of the sorted arms. The gradient of the target arm (determined with another sampling) is then re-estimated and updated.

Theoretical analysis showed that the gradients calculated with AAB was an unbiased estimation and the variance of the estimation was effectively reduced with high confidence.

Experiments were conducted to show the effectiveness of the proposed AAB. The experimental results based on an MAB task showed that AAB outperformed the baseline of traditional policy gradient and achieved comparable performances with the policy gradient with baseline.

2 BACKGROUND: VARIANCE REDUCTION IN POLICY GRADIENT

This section introduces the formulation of variance reduction methods in the policy gradient for the multi-armed bandit problem.

2.1 Gradient bandit algorithm

Suppose we are facing repeatedly with a choice among k different actions. After each choice, we receive a numerical reward chosen from a stationary probability distribution that depends on the selected action. Each action has an expected reward, called value. The objective is to maximize the expectation of total reward over some time periods. There are two targets in the game, that is, finding the best action which has the largest value and maximizing the accumulative reward in limited time periods.

Policy gradient aims to learn a numerical preference $H_t(a)$ for each action $a = \{1, 2, \cdots k\}$, to calculate the policy $\pi_t(a)$ on time step t. Denote the action selected on time t as A_t, the corresponding reward as R_t, and the value of selecting an action a as $q_*(a) = \mathbb{E}[R_t|A_t = a]$. The policy π_t (a distribution over the actions) is defined as the softmax over the preferences H_t:

$$\pi_t(a) = Pr(A_t = a) = \frac{\exp\{H_t(a)\}}{\sum_{a'=1}^{k} \exp\{H_t(a')\}}. \tag{1}$$

The policy in Equation (1) is used to play the bandit game, and the expected reward at time step t is $\mathbb{E}[R_t] = \sum_a \pi_t(a) q_*(a)$. In principle, the gradient of the expected reward $\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$ is used to update the preferences of the actions. However, it is difficult to calculate $\mathbb{E}[R_t]$ and its partial gradient to preference because $q_*(a)$ is unknown. Monte-Carlo method is used to estimate the gradient. The basic idea is the system gets a sample $q(a)$ as a reward when action a has been issued. Thus, the gradient is estimated as:

$$\begin{aligned} \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \sum_b \pi_t(b) q_*(b) (\mathbf{1}_{a=b} - \pi_t(a)) \\ &\overset{\text{sample}}{=} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} q(x_i)(\mathbf{1}_{a=x_i} - \pi_t(a))\right] = \mathbb{E}[\eta] \end{aligned} \tag{2}$$

where $n \geq 1$, the $\overset{\text{sample}}{=}$ samples $x_i \sim \pi_t$, and $\eta = \frac{1}{n} \sum_{i=1}^{n} q(x_i)(\mathbf{1}_{a=x_i} - \pi_t(a))$. The equation above shows that η is an unbiased estimation for the gradient of R_t (i.e., $\mathbb{E}[\eta] = \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$). Thus, one effective way to the accuracy of the estimation is to reduce the variance $\mathbb{V}[\eta]$.

2.2 Control variates method for policy gradient

Policy gradient with baseline is an effective method to reduce the variance of η. In the k-armed bandit problem, it calculates the gradient of $H_t(a)$ as $\frac{1}{n} \sum_{i=1}^{n} [q(x_i) \cdot (\mathbf{1}_{a=x_i} - \pi_t(a)) - R \cdot (\mathbf{1}_{a=x_i} - \pi_t(a))]$, where R is the averaged rewards of the history samplings.

From the viewpoint of statistics, policy gradient with baseline is an application of control variates in policy gradient. The original control variates method can be described as follows: considering the estimation θ of an integral of function $f(x)$, Monte Carlo method directly uses $\frac{1}{n} \sum_{i=1}^{n} (\xi_i)$, where $\xi_i \sim U(0, 1)$ and $n \geq 1$, as an estimation of θ, which is an obviously unbiased estimation.

The control variates method, on the other hand, estimates the integral with:

$$\begin{aligned} \theta &= \int_0^1 f(x)dx = \int_0^1 [f(x) - cg(x)]dx + c \int_0^1 g(x)dx \\ &= \int_0^1 [f(x) - cg(x)]dx + C, \end{aligned} \tag{3}$$

where c and $C = c \int_0^1 g(x)dx$ are constants. Define a new random variable $\zeta = f(\xi) - cg(\xi) + C$, where $\xi \sim U(0, 1)$ is a random number uniformly distributed over the interval $(0, 1)$. It is simple to calculate that $\mathbb{E}[\zeta] = \mathbb{E}[f(\xi)]$, and $\mathbb{V}[\zeta] \leq \mathbb{V}[f(\xi)]$ when $Corr[f(\xi), g(\xi)] \geq 0$, thus, ζ could be a better unbiased estimation with reduced variance for the integral.

In policy gradient with baseline, $R \cdot (\mathbf{1}_{a=x_i} - \pi_t(a))$ corresponds to $cg(x)$, $R \cdot \mathbb{E}[(\mathbf{1}_{a=x_i} - \pi_t(a))] = 0$ corresponds to C, and $Corr[R(\mathbf{1}_{a=x_i} - \pi_t(a)), q(x_i)(\mathbf{1}_{a=x_i} - \pi_t(a))] \geq 0$. Thus, policy gradient with baseline can be considered as an application of control variates method to policy gradient.

2.3 Antithetic Variates Method

In statistics, antithetic variates method is another approach to reducing the variance of Monte Carlo method. Still consider the problem of estimating $\theta = \int_0^1 f(x)dx$. Antithetic variates method adopts the stratified sampling strategy [10]. The sampler chooses a fixed set of numbers $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_n = 1$ and defines a new estimation ζ

$$\zeta = \sum_{j=1}^{n} (\alpha_j - \alpha_{j-1}) f[\alpha_{j-1} + (\alpha_j - \alpha_{j-1})\xi_j], \tag{4}$$

where the $\xi_j \in U(0, 1)(j = 1, \cdots, n)$. It can be shown that ζ is an unbiased estimation for θ and $\mathbb{V}(\zeta) \leq \mathbb{V}(\frac{1}{n} \sum_{i=1}^{n} f(\xi_i))$. Specifically, antithetic variates method constructs antithetic pair among the variables $f[\alpha_{j-1} + (\alpha_j - \alpha_{j-1})\xi_j]$ (i.e.,$Cov[\eta_1, \eta_2] < 0$ where $\eta_j = f[\alpha_{j-1} + (\alpha_j - \alpha_{j-1})\xi_j], j = 1, 2, \ldots, n$) to get the smaller $\mathbb{V}(\zeta)$.

3 OUR APPROACH: ANTITHETIC-ARM BANDIT

This section proposes Antithetic-Arm Bandit (AAB), a novel variance reduction policy gradient algorithm for multi-armed bandit, on the basis of antithetic variates method.

Algorithm 1 Antithetic-Armed Bandit (AAB)

Input: Action (arm) set $\mathbf{A} = \{1, \cdots, k\}$, number of iterations $\mathbf{T}$, stratified parameter m, learning rate λ

$H_t \leftarrow \mathbf{0}$
for $t = 1$ **to** $\mathbf{T}$ **do**
  $\pi_t \leftarrow \left\{ \frac{\exp H_t(a)}{\sum_{a'} \exp H_t(a')} \right\}_{a=1}^{k}$ {Equation (1)}
  Sample an arm $d \in \mathbf{A}$ according to π_t
  $\forall a \in \mathbf{A}, p(a) \leftarrow \begin{cases} 0 & a = d \\ \frac{\pi_t(a)}{(1-\pi_t(d))} & \text{otherwise} \end{cases}$
  $\forall a \in \mathbf{A}, \mathcal{I}(a) \leftarrow$Index of action a after descent sorting arms according to p
  $\alpha \leftarrow \sum_{i=1}^{m} \pi_t(\mathcal{I}(i)) + \frac{1}{2}\pi_t(\mathcal{I}(m+1))$ {Heuristics for calculating α}
  Sampling $\xi \sim U(0,1)$
  $\xi_1 \leftarrow \alpha\xi$, $\xi_2 \leftarrow 1 - (1-\alpha)\xi$
  $a_1 \leftarrow \sigma(\mathcal{I}, p, \xi_1)$, $a_2 \leftarrow \sigma(\mathcal{I}, p, \xi_2)$ {Algorithm 2}
  $g_d \leftarrow q(d)(1 - \pi_t(d))$
  $g_1 \leftarrow q(a_1)(-\pi_t(a_1)); g_2 \leftarrow q(a_2)(-\pi_t(a_2))$
  $\zeta_d \leftarrow \pi_t(d)g_d + (1 - \pi_t(d))(\alpha g_1 + (1-\alpha)g_2)$ {Equation (5)}
  $H_t(d) \leftarrow H_t(d) + \lambda\zeta_d$
end for

3.1 Antithetic-Arm Bandit

Suppose an k-armed bandit problem where the set of actions are $\mathbf{A} = \{1, 2, \cdots, k\}$ and each action $a \in \mathbf{A}$ represents the a-th arm of the bandit. AAB updates the gradient of different arms with coordinate ascent. Given an arm a, the estimation of gradient proposed by Monte Carlo method for a is

$$\eta = \frac{1}{n}\sum_{i=1}^{n} q(x_i)(\mathbf{1}_{a=x_i} - \pi_t(a))$$

where $x_i \sim \pi_t$, π_t is the current policy at time step t, and n is the number of samplings.

Denote the random sampled arm according to current policy π_t as $f_{\pi_t}(\xi)$ where $\xi \sim U(0,1)$. Also denote the function for calculating the gradient as G. Thus, after taking an action $x_i = f_{\pi_t}(\xi)$, the gradient of the d-th arm can be calculated as $G_d(x_i) = \mathbf{1}_{d=x_i} - \pi_t(d)$. At each iteration, AAB samples three arms: the first sampling chooses a target action $d \sim \pi_t$ and makes d to be the arm to update; the second and the third sample two antithetic arms for calculating the gradients for d. Specifically, given three sampled actions, the estimation of gradient for arm d can be calculated as:

$$\begin{aligned} \zeta_d = {} & \pi_t(d)G_d(d) + (1 - \pi_t(d))\alpha G_d f_{\pi_t}\sigma(\alpha\xi) \\ & + (1 - \pi_t(d))(1-\alpha)G_d f_{\pi_t}\sigma(1 - (1-\alpha)\xi) \end{aligned} \tag{5}$$

where the random variable $\xi \sim U(0,1)$, σ is the permutation function that sorts the arms so that the second and the third sampled arms are antithetic, and $\alpha \in (0,1)$ is the parameter which is heuristically set as $\sum_{i=1}^{m} \pi_t(\mathcal{I}(i)) + \frac{1}{2}\pi_t(\mathcal{I}(m+1))$ where m (named stratified parameter) represents the position of stratifying.

Algorithm 1 shows the AAB process and Algorithm 2 shows the function for sorting the arms.

At time step t, given the current policy π_t, an action d is sampled which corresponds the arm to update. Then the algorithm sort the actions (arms) with σ. After that, a pair of antithetic actions are sampled from $\mathbf{A} \setminus \{d\}$, and the two antithetic random variables g_1 and g_2 are constructed. It can be shown that g_1 and g_2 have high probability to be antithetic, because the arms were sorted. Finally the gradient for the chosen arm d is calculated and the policy is updated. The iteration is repeated until converge.

Algorithm 2 Sort function σ

Input: Sorted index $\mathcal{I}$, solved policy p, random variable ξ
Output: action a

for $a = 1$ **to** length of p **do**
  if $\xi - p(\mathcal{I}(a)) < 0$ **then**
    return $\mathcal{I}(a)$
  end if
end for

Intuitively, AAB constructs a monotonic compound function $G_d f_{\pi_t}\sigma$ which makes $Cov[g_1, g_2] < 0$ because $Cov[\xi_1, \xi_2] < 0$. In the next section, we show that AAB makes an unbiased estimation of the gradient and reduces the variance of the estimated gradients.

3.2 Theoretical analysis

AAB makes an unbiased estimation of the gradient, as shown in the following Theorem 3.1 because $G_d f_{\pi_t}(\xi)$ is an unbiased estimation obviously when $\xi \sim U(0,1)$.

Theorem 3.1. *$\forall t = 1, 2, \cdots$, and $\forall d \in \mathbf{A} = \{1, 2, \ldots, k\}$, the expectation of ζ_d in Equation (5) satisfies $\mathbb{E}[\zeta_d] = \mathbb{E}[G_d f_{\pi_t}(\xi)]$, where $\xi \sim U(0,1)$.*

Proof. Denote p_{ij} as the probability of choosing the actions i and j at the same time after action d being chosen. And we set $p_{ij} = 0$ if $i = d$, $j = d$ or $i > j$. We have

$$\begin{aligned} \mathbb{E}[\zeta_d] = {} & \pi_t(d)G_d(d) + (1 - \pi_t(d)) \sum_{i=1}^{k}\sum_{j=1}^{k}[\alpha p_{ij}G_d(i) + (1-\alpha)p_{ij}G_d(j)] \\ = {} & \pi_t(d)G_d(d) + (1 - \pi_t(d)) \sum_{i=1}^{k}(\sum_{j=1}^{k}\alpha p_{ij} + \sum_{j=1}^{k}(1-\alpha)p_{ji})G_d(i) \\ = {} & \pi_t(d)G_d(d) + \sum_{i \in A \setminus \{d\}} \pi_t(i)G_d(i) \\ = {} & \sum_{i \in \{A\}} \pi_t(i)G_d(i) = \mathbb{E}[G_d f_{\pi_t}(\xi)] \end{aligned}$$

□

AAB uses stratified sampling as the traditional antithetic variates method do. Given $\alpha \in (0,1)$, stratified sampling has lower variance than the primal Monte Carlo method. In coordinate ascent, the stratified sampling estimates the gradient of the arm d as:

$$\begin{aligned} \eta_d = {} & \pi_t(d)G_d(d) + (1 - \pi_t(d))\alpha G_d f_{\pi_t}(\pi_t(d) + \alpha\xi_1) \\ & + (1 - \pi_t(d))(1-\alpha)G_d f_{\pi_t}(\pi_t(d) + \alpha + (1-\alpha)\xi_2) \end{aligned} \tag{6}$$

where $\xi_1, \xi_2 \sim U(0,1)$ and $Cov[\xi_1, \xi_2] = 0$.

Moreover, it can be shown that the estimation of AAB has a smaller variance than stratified sampling, as shown in Theorem 3.2.

Theorem 3.2. *$\forall t = 1, 2, \cdots$, and $\forall d \in \mathbf{A} = \{1, 2, \ldots, k\}$, the variance of ζ_d in Equation (5) and η_d in Equation (6) satisfies $\mathbb{V}[\zeta_d] \leq \mathbb{V}[\eta_d]$*

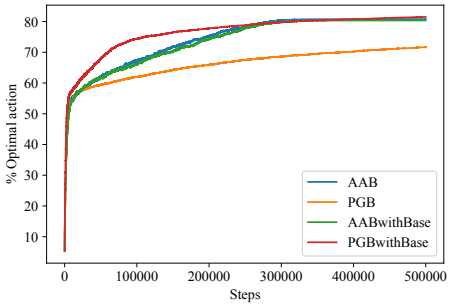

Figure 1: Performance curves of different methods.

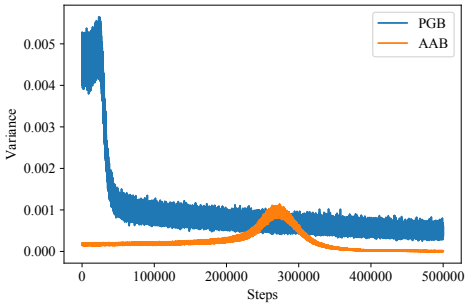

Figure 2: The variance of PGB and AAB.

Proof. Let $\theta_1 = \alpha G_d f_{\pi_t} \sigma(\alpha\xi)$, $\theta_2 = (1-\alpha)G_d f_{\pi_t}\sigma(1-(1-\alpha)\xi)$, $\theta_3 = \alpha G_d f_{\pi_t}(\pi_t(d) + \alpha\xi_1)$, and $\theta_4 = (1-\alpha)G_d f_{\pi_t}(\pi_t(d) + \alpha + (1-\alpha)\xi_2)$ where $\xi, \xi_1, \xi_2 \sim U(0,1)$.
$\mathbb{V}[\zeta_d] = (1-\pi_t(d))^2(\mathbb{V}[\theta_1] + \mathbb{V}[\theta_2] + Cov[\theta_1, \theta_2])$
$\mathbb{V}[\eta_d] = (1-\pi_t(d))^2(\mathbb{V}[\theta_3] + \mathbb{V}[\theta_4] + Cov[\theta_3, \theta_4])$
$\because$ θ_1 and θ_3 are I.I.D., and θ_2 and θ_4 are I.I.D.
$\therefore$ $\mathbb{V}[\theta_1] = \mathbb{V}[\theta_3], \mathbb{V}[\theta_2] = \mathbb{V}[\theta_4]$
$\because$ $Cov[\theta_1, \theta_2] \leq 0, Cov[\theta_3, \theta_4] = 0$
$\therefore$ $\mathbb{V}[\zeta_d] \leq \mathbb{V}[\eta_d]$ □

4 EXPERIMENTS

We conducted experiments to test the proposed AAB algorithm. Following the practices in [6], "one real competitor" in [6] was used as our experiments. As for the reward in the k-armed bandit, the reward distributions were set to Bernoulli and the expected rewards of actions were set as $p_1 = 0.5$, $p_2 = 0.5 - \frac{1}{10k}$ and $p_i = 0.4, i = 3, \ldots, k$. The number of arms k was set to $k = 20$. The parameter m for calculating α in AAB was set as $m = 4$.

Figure (1) shows the performance curves of different methods in terms of the ratio of choosing the optimal action. From the results, we can see that our approaches (AAB and AABwithBase[1]) performed better than the baseline method of policy gradient (PGB), and have similar performance to the method of policy gradient with baseline (PGBwithBase). Note that PGB and PGBwithBase only update the first of 3 sampled arms at each iteration. From the results we can see that 1) AAB can effectively reduce the variance, making it outperform PGB; 2) both AAB and policy gradient with baseline can effectively reduce the gradient variance, leading to similar performances; and 3) combining AAB with policy gradient with baseline (AABwithBase) can marginally improve the performances.

We also tested the variance curves of AAB and PGB, as shown in Figure (2). The variances of the estimated gradient by AAB are in general smaller than that of by PGB in all of the iterations, showing the effectiveness of AAB in reducing the gradient variance.

5 CONCLUSION

In this paper, we propose a novel variance reduction method for policy gradient in multi-armed bandit problem, called Antithetic-Arm Bandit (AAB). Compared with existing method of policy gradient with baseline which can be viewed as a control variates method in statistics, AAB resorts to the antithetic variants method for the task. Algorithms were proposed to conduct the estimation and the theoretical analysis showed that the gradients estimated by AAB are unbiased and the variance is smaller than that of by the conventional Monte Carlo methods. Experimental results also showed that AAB can achieve the state-of-the-art performances.

[1]Note that AAB can be combined with the policy gradient with baseline for further reducing the variance.

6 ACKNOWLEDGMENTS

This work was funded by the National Key R&D Program of China under Grants No. 2016QY02D0405, the 973 Program of China under Grant No. 2014CB340401, the National Natural Science Foundation of China (NSFC) under Grants No. 61773362, 61425016, 61472401, 61722211, and 20180290, and the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102.

REFERENCES

[1] Jonathan Baxter and Peter L Bartlett. 2001. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research* 15, 1 (2001), 319–350.
[2] E.C.FIELLER and H.O.HARTLEY. 1954. SAMPLING WITH CONTROL VARIABLES. *Biometrika* 41, 3/4 (1954), 494–501.
[3] Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research* 5, Nov (2004), 1471–1530.
[4] J. M. Hammersley and K. W. Morton. 1956. A new monte carlo technique: Antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society* 52, 3 (1956), 449–475.
[5] Ronald A Howard. 1960. Dynamic programming and markov processes. (1960).
[6] Zohar Karnin, Tomer Koren, and Oren Somekh. 2013. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 1238–1246.
[7] Michael N Katehakis and Arthur F Veinott Jr. 1987. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research* 12, 2 (1987), 262–268.
[8] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*. 1008–1014.
[9] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.
[10] Jerzy Neyman. 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* 97, 4 (1934), 558–625.
[11] R. S Sutton. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Submitted to Advances in Neural Information Processing Systems* 12 (1999), 1057–1063.
[12] R. J. Williams. 1988. Towards a theory of reinforcement-learning connectionist systems. *Issues in Education* 4, 1 (1988), 1–94.
[13] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.