Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

POLICY EVALUATION WITH STOCHASTIC GRADIENT ESTIMATION TECHNIQUES

Yi Zhou

Department of Mathematics
Institute for Systems Research
University of Maryland
8223 Paint Branch Dr
College Park, MD 20742, USA

Michael C. Fu

Robert H. Smith School of Business
Institute for Systems Research
University of Maryland
7699 Mowatt Ln
College Park, MD 20742, USA

Ilya O. Ryzhov

Robert H. Smith School of Business
University of Maryland
7699 Mowatt Ln
College Park, MD 20742, USA

ABSTRACT

In this paper, we consider policy evaluation in a finite-horizon setting with continuous state variables. The Bellman equation represents the value function as a conditional expectation, which can be further transformed into a ratio of two stochastic gradients. By using the finite difference method and the generalized likelihood ratio method, we propose new estimators for policy evaluation and show how the value of any given state can be estimated using sample paths starting from various other states.

1 INTRODUCTION

Many real-world business problems can be modeled using finite-horizon stochastic process. Reinforcement learning (RL) techniques have been successful in solving these problems, such as optimal asset allocation (Neuneier 1995), inventory management in supply chains (Oroojlooyjadid et al. 2022) and option pricing and hedging (Longstaff and Schwartz 2001). There are many well-known reinforcement learning techniques, including Q-learning (Tsitsiklis 1994), policy gradient method (Sutton et al. 1999), etc. For a general introduction to RL techniques, one can refer to Sutton and Barto (2018). In this paper, we focus on the problem of policy evaluation, where the task is to estimate a value function for a given action policy. The action policy can then be iteratively improved by finding the optimal action based on the estimated value function. This procedure is known as the policy improvement step.

Two widely-used approaches for policy evaluation are the Monte Carlo (MC) method and the temporal difference (TD) method (Sutton 1988). The TD label encompasses a broader class of methods, such as TD(λ), introduced by Tsitsiklis and Van Roy (1997). The value of a given state can be represented by a conditional expectation, and both MC and TD estimate it by simulating the trajectory of the process starting from that same state. The main difference between them is that MC requires the simulation to continue until termination, while TD simulates only a single transition and calculates an estimate based on the Bellman equation. In this paper, we investigate ways to estimate the value of a state using sample paths that start in other states. There has been some recent work along these lines, such as Daveloose et al. (2019) and Zhou et al. (2021), and the latter one is a generalization of the former.

Zhou et al. (2021) have shown that under appropriate assumptions, a conditional expectation can be represented as a ratio of the gradients of two expectations. The expressions inside these expectations involve nondifferentiable indicator functions. We can use this ratio representation as long as stochastic estimators for these gradients are available. The area of stochastic gradient estimation (SGE) have been studied extensively and many techniques have been developed: finite difference (FD) methods, infinitesimal perturbation analysis (IPA) (Glasserman 1991; Ho and Cao 1983), the likelihood ratio (LR) method (Glynn 1987), and the generalized likelihood ratio (GLR) method (Peng et al. 2018). FD is a simple but effective method, but a major drawback is that it generally produces a biased estimator, which could impair the convergence rate of policy evaluation. Another important disadvantage of FD is that when we use a smaller perturbation parameter to decrease the bias of the estimator, a larger number of sample paths are needed to reduce the variance. IPA and LR yield unbiased estimators when applicable, but for our problem, both fail due to the existence of the indicator functions in the ratio representation (Peng et al. 2018). It is possible to obtain an unbiased estimator using GLR, but this requires an explicit form for the distribution of the initial state. Though we have the flexibility to specify this density function, it is not clear how to do this "well". Moreover, the application of GLR generally requires differentiability of the value function, thus has many limitations.

With a FD-based or GLR-based estimator of the value function, we can use stochastic gradient descent to minimize the Bellman error, analogously to TD. The asymptotic convergence of such a procedure can be obtained by extending the available theory for TD methods, which is mainly based on the ODE method for stochastic approximation (Kushner and Yin 2003). This approach, however, does not provide finite-time performance guarantees. Bhandari et al. (2018) investigated the finite-time performance of TD in infinite-horizon, discrete-state problems with linear function approximation. However, for the algorithms associated with our proposed new FD and GLR estimators, the finite-time analysis could become more complicated, since there are many parameters that should be determined by the user. For example, the number of sample paths used to obtain the FD estimator could have a great impact on the performance of the algorithm, and a large number of sample paths would lead to high simulation costs, which is important when we consider budget-dependent convergence rate results (L'Ecuyer and Yin 1998) for these algorithms.

The paper is organized as follows. In Section 2, we review the problem of policy evaluation in a finite horizon setting along with two classical methods, MC and TD. In Section 3, we propose new estimators for the value function using stochastic gradient estimation techniques. In Section 4, we describe the detailed algorithms and give preliminary insights into the convergence analysis. Numerical experiments are conducted in Section 5.

2 PROBLEM FORMULATION

We consider the policy evaluation problem for a Markov decision process (MDP) in a finite-horizon setting where the state variables take continuous values. Let T be the termination time and $\mathscr{X}$ be the state space. Suppose at time step $t = 0,\ldots,T$, the state variable is $x_t \in \mathscr{X}$, and the next state x_{t+1} is generated by the transition

$$x_{t+1} = h_t(x_t, \pi(x_t,t), Z_{t+1}), \tag{1}$$

where π is the given action policy that we aim to evaluate, i.e., $\pi(x_t,t)$ gives the action taken at t, and the quantity Z_{t+1} represents the randomness of the transition, assumed to be independent of the state and the action. Let $R_{t+1}(x_t,x_{t+1})$ be a deterministic function representing the one-step reward obtained between steps t and $t+1$. The value function associated with the MDP, denoted by V^π is then given by

$$V^\pi(x,t) = \mathbb{E}\left[\sum_{i=t}^{T-1} \gamma^{i-t} R_{i+1}(x_i,x_{i+1}) | x_t = x, t\right],$$

where $\gamma \in (0,1)$ is a discount factor. It is well known that V^π satisfies the Bellman equations $V^\pi(x,T) = 0$, and $V^\pi(x,t) = \mathscr{L}V^\pi(x,t+1)$ for $t = 0,\ldots,T-1$, where the operator $\mathscr{L}$ is defined by

$$\mathscr{L}V(x,t+1) = \mathbb{E}\left[R_{t+1}(X_t,X_{t+1}) + \gamma V(X_{t+1},t+1)|X_t = x,t\right], \quad t = 0,\ldots,T-1.$$

Given a fixed initial state x_0 and the fixed policy π, the distribution $F_t(x)$ of the state variable X_t is fixed. Since the state variables are assumed to take continuous values, we suppose that the value function V^π is approximated by a parametrized model $V_\theta(x,t)$. In general, the parametrized model could take different forms for different t. For each $t = 0,\ldots,T$, it is desirable to find θ_t^* such that

$$\theta_t^* = \arg\min_{\theta_t \in \Theta} \frac{1}{2} \|V_{\theta_t}(x,t) - V^\pi(x,t)\|_{F_t}^2 \tag{2}$$

where Θ is a closed and convex parameter space, and $\|\cdot\|_{F_t}^2$ is defined by $\|V(x,t)\|_{F_t}^2 := \mathbb{E}_{X\sim F_t}\left[|V(X,t)|^2\right]$. Let Π_t represent the projection operator that gives the best approximation of a given function with respect to F_t. Then the solution to (2) can be written as $\Pi_t V^\pi(x,t)$. For simplicity, we assume that θ_t^* exists and is unique for each t.

2.1 Monte Carlo (MC) and Temporal Difference (TD) Methods

In the following, we give a brief review of the Monte Carlo (MC) and temporal difference (TD) methods for fitting the value of a fixed policy π.

Suppose that at iteration $n+1$, we simulate a sample path $\{x_0,x_1,R_1,x_2,R_2,\ldots,x_T,R_T\}$. Then, the MC and the TD estimators for $V^\pi(x,t)$ are given by $\hat{V}^{MC}(x,t) = \sum_{i=t}^{T-1} \gamma^{i-t} R_{i+1}$ and $\hat{V}^{TD}(x,t) = R_{t+1} + \gamma V_{\theta_{t+1}^n}(x_{t+1},t+1)$, respectively, where θ^n is the parameter obtained in iteration n. In practice, for both methods, estimators are calculated in a backward direction. The MC estimator at iteration $n+1$ does not depend on θ^n, while the TD estimator is affected by θ^n. Therefore, in TD, instead of solving (2) as in the MC method, we are essentially solving a different problem for θ_t,

$$\min_{\theta_t \in \Theta} \frac{1}{2} \left\|V_{\theta_t}(x,t) - \mathscr{L}V_{\theta_{t+1}}(x,t+1)\right\|_{F_t}^2. \tag{3}$$

Consequently, the error in the approximation of $V^\pi(x,t)$ by $V_{\theta_t^*}(x,t)$ depends on the approximation error induced by θ_{t+1}^*. If $V_{\theta_{t+1}^*} = \Pi_{t+1}V^\pi$, then the approximation error of $V^\pi(x,t)$ by $V_{\theta_t^*}(x,t)$ is simply a projection error, which can be seen from

$$\begin{aligned}
\left\|V_{\theta_t^*}(x,t) - V^\pi(x,t)\right\|_{F_t} &= \left\|\Pi_t \mathscr{L}V_{\theta_{t+1}^*}(x,t+1) - V^\pi(x,t+1)\right\|_{F_t} \\
&\le \left\|\Pi_t \mathscr{L}V_{\theta_{t+1}^*}(x,t+1) - \Pi_t \mathscr{L}V^\pi(x,t+1)\right\|_{F_t} + \|\Pi_t V^\pi(x,t) - V^\pi(x,t)\|_{F_t} \\
&\le \gamma \left\|V_{\theta_{t+1}^*}(x,t+1) - V^\pi(x,t+1)\right\|_{F_{t+1}} + \|\Pi_t V^\pi(x,t) - V^\pi(x,t)\|_{F_t},
\end{aligned}$$

where the first inequality is obtained from the Bellman equation, and the second equality follows because the projection operator Π_t is non-expansive and the Bellman operator $\mathscr{L}$ is a contraction.

The optimal parameters θ are searched using a stochastic gradient descent algorithm for the objectives (2) and (3), where the gradient estimator can be written in a general form:

$$g_t(\theta_t,\theta_{t+1}) = (\hat{V}^\pi(x,t) - V_{\theta_t}(x,t))\nabla_{\theta_t} V_{\theta_t}(x,t), \tag{4}$$

Therefore, the key difference between MC and TD lies in the form of $\hat{V}^\pi(x,t)$. In the next section, we propose new formulations of $\hat{V}^\pi(x,t)$ derived by stochastic gradient estimation (SGE) techniques.

3 NEW ESTIMATORS OF THE VALUE FUNCTION

In the following, we present a ratio representation of a conditional expectation, and then propose two new estimators of the value function.

3.1 Ratio Representation of a Conditional Expectation

The Bellman equation shows that the problem of estimating $V^\pi(x,t)$ can be reduced to the problem of estimating a particular conditional expectation. Furthermore, Zhou et al. (2021) have shown that a conditional expectation can be represented by a ratio of the gradients of two expectations. For our purpose, this technique can be applied in policy evaluation. To simplify the presentation, we consider the case where the state variable x is a one-dimensional variable.

Define $W(x,x',t;V) = R_t(x,x') + \gamma V(x',t)$. Then,

$$\begin{aligned}
\mathscr{L}V(x,t+1) =& \mathbb{E}\left[W(X_t,X_{t+1},t+1;V)|X_t = x,t\right] \\
=& \lim_{\varepsilon\to 0}\mathbb{E}\left[W(X_t,X_{t+1},t+1;V)|X_t \in (x-\varepsilon,x+\varepsilon)\right] \\
=& \frac{\lim_{\varepsilon\to 0}(1/2\varepsilon)\mathbb{E}\left[W(X_t,X_{t+1},t+1;V)\mathbf{1}\left\{X_t \in (x-\varepsilon,x+\varepsilon)\right\}\right]}{\lim_{\varepsilon\to 0}(1/2\varepsilon)\mathbb{E}\left[\mathbf{1}\left\{X_t \in (x-\varepsilon,x+\varepsilon)\right\}\right]} \qquad (5) \\
=& \left.\frac{\frac{d}{d\xi}\mathbb{E}\left[W(X_t,X_{t+1},t+1;V)\mathbf{1}\left\{X_t \le \xi\right\}\right]}{\frac{d}{d\xi}\mathbb{E}\left[\mathbf{1}\left\{X_t \le \xi\right\}\right]}\right|_{\xi=x}, \qquad (6)
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. If we can obtain estimators of the two stochastic gradients in (6), then $\mathscr{L}V(x,t+1)$ can be estimated by taking the ratio. For higher-dimension problems, the representation would involve higher-order partial derivatives. By the derived ratio representation, when estimating the value function at a specific point (x,t), we do not have to simulate the next states starting at exactly this fixed point as in the MC or the TD method. Instead, we could use sample paths from $(\tilde{x},t)$, where $\tilde{x}$ can be different from x. In other words, given t, any sample paths starting from any state can be used to estimate the value function for any x. Therefore, the ratio representation allows much more flexibility in simulating sample paths and offers new ways of constructing different estimators of V.

3.2 Stochastic Gradient Estimation (SGE)

By the ratio representation (6), stochastic gradient estimation (SGE) techniques are needed to estimate the conditional expectation. There are many widely used SGE techniques, which include finite difference (FD) methods, infinitesimal perturbation analysis (IPA) and the likelihood ratio (LR) method. Among those, FD is the one of the most straightforward methods. Suppose $\mathscr{D} = \{(X_t^j, X_{t+1}^j, R_{t+1}^j), j = 1,\ldots,M\}$ are i.i.d random pairs, where $X_t^j \sim F_t$, the corresponding X_{t+1}^j is obtained by the transition function (1), and $R_{t+1}^j = R_{t+1}(X_t^j, X_{t+1}^j)$. Then, from (5), a symmetric FD estimator is given by

$$\hat{V}^{FD}(x,t;V,\mathscr{D},\varepsilon) = \frac{\sum_{j=1}^M W(X_t^j,X_{t+1}^j,t+1;V)\mathbf{1}\left\{X_t^j \in (x-\varepsilon,x+\varepsilon)\right\}}{\sum_{j=1}^M \mathbf{1}\left\{X_t^j \in (x-\varepsilon,x+\varepsilon)\right\}}, \qquad (7)$$

where $\varepsilon > 0$ is a small fixed perturbation parameter. Although FD is easy to implement, it has several drawbacks. First, the FD estimator is biased, which usually leads to a slower convergence rate compared to unbiased estimators. Moreover, for a small ε, the event $\{X_t \in (x-\varepsilon,x+\varepsilon)\}$ could be a rare event, thus a large number of sample paths may be required to make the denominator in (7) nonzero. To overcome these drawbacks, alternative SGE techniques that could yield unbiased estimators, such as IPA and LR might be preferred, but since both the numerator and denominator in (6) involve an indicator function

which introduces discontinuities, both IPA and LR fail. Peng et al. (2018) have proposed a generalized likelihood ratio (GLR) method, which can overcome this difficulty and obtain unbiased estimators for (6). The application of GLR can be justified under Assumption 1.

Assumption 1 The following statements hold for all t:

(i) Z_{t+1} is independent of X_t.
(ii) $f_t(x)$, the density of X_t, has unbounded support.
(iii) $\ln f_t(x)$ is differentiable for all $x \in \mathbb{R}$.
(iv) $\mathbb{E}\left[\left(\partial_x \ln f_t(X_t)\right)^2\right] < \infty$.
(v) $\mathbb{E}\left[\left(W(X_t, X_{t+1}, t+1; V)\right)^2\right] < \infty$.

Under Assumption 1, the numerator in (6) can be represented as an ordinary expectation according to

$$\begin{aligned}
&\frac{d}{d\xi}\mathbb{E}\left[W(X_t, X_{t+1}, t+1; V)\mathbf{1}\left\{X_t \le \xi\right\}\right] \\
=&\mathbb{E}\left[\mathbf{1}\left\{X_t \le \xi\right\}\left(\frac{d}{dX_t}W(X_t, X_{t+1}, t+1; V) + W(X_t, X_{t+1}, t+1; V)\frac{d}{dX_t}\ln f_t(X_t)\right)\right]. \qquad (8)
\end{aligned}$$

Here, we briefly discuss the derivation of (8); please see Peng et al. (2018) for complete technical details. Since the indicator function $\mathbf{1}\{z \le 0\}$ involves a discontinuity, which prevents us from interchanging the expectation and the differentiation operator, we first replace it by a sequence of approximated continuous functions $\chi_\varepsilon(\cdot)$, i.e., $\lim_{\varepsilon\to 0}\chi_\varepsilon(z) = \mathbf{1}\{z \le 0\}$. Then we have

$$\begin{aligned}
&\frac{d}{d\xi}\mathbb{E}\left[W(X_t, X_{t+1}, t+1; V)\chi_\varepsilon(X_t - \xi)\right] \\
=&\mathbb{E}\left[W(X_t, X_{t+1}, t+1; V)\frac{d}{d\xi}\chi_\varepsilon(X_t - \xi)\right] \\
=&-\int_{\mathbb{R}} W \cdot \frac{d}{dx}\chi_\varepsilon(x-\xi)f(x)dx \\
=&-W \cdot \chi_\varepsilon(x-\xi)f_t(x)|_{-\infty}^{\infty} + \int_{\mathbb{R}}\chi_\varepsilon(x-\xi)\frac{d}{dx}\left(W \cdot f_t(x)\right)dx \\
=&\int_{\mathbb{R}}\chi_\varepsilon(x-\xi)\left(\frac{d}{dx}W + W \cdot \frac{d}{dx}\ln f_t(x)\right)f_t(x)dx, \qquad (9)
\end{aligned}$$

where the first equality holds under integrability assumptions in Assumption 1, the second equality is obtained by a change of variable, and the third equality is obtained by the integration by parts formula. Letting $\varepsilon \to 0$, (9) converges to (8).

Using the GLR method, we can obtain an unbiased estimator for the numerator of (6). However, the representation (8) requires explicit knowledge of the density function $f_t(\cdot)$, which is not available in practice. At the same time, the derivation of the ratio representation (6) does not require that X_t is generated from F_t, i.e., the expectations in (6) can be taken with respect to a user-specified distribution F_t^s. Following this, to apply GLR, we need Assumption 1 to hold where F_t and f_t are replaced by F_t^s and f_t^s, respectively. For the denominator of (6), we can apply the same GLR technique to obtain a corresponding ordinary expectation, but the ratio of two sample means would still be a biased estimator for $\mathscr{L}V(x,t+1)$. Therefore, another benefit of using a user-specified distribution F_t^s is that the denominator of (8) simply becomes $f_t^s(x)$, which leads to an unbiased estimator for $\mathscr{L}V(x,t+1)$ when we use sample means of i.i.d.

observations to estimate the numerator. Thus, the final form of the GLR estimator is given by

$$\hat{V}^{GLR}(x,t;V,\tilde{\mathscr{D}}) = \frac{1}{M \cdot f_t^s(x)} \sum_{j=1}^{M} \left[\mathbf{1}\left\{ \tilde{X}_t^j \le x \right\} \left(\frac{d}{d\tilde{X}_t^j} W(\tilde{X}_t^j, \tilde{X}_{t+1}^j, t+1; V) + W \cdot \frac{d}{d\tilde{X}_t^j} \ln f_t^s(\tilde{X}_t^j) \right) \right], \quad (10)$$

where $\tilde{\mathscr{D}} = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1, \ldots, M\}$ are i.i.d random pairs, where $\tilde{X}_t^j \sim F_t^s$, the corresponding $\tilde{X}_{t+1}^j$ is obtained by (1), and $\tilde{R}_{t+1}^j = R_{t+1}(\tilde{X}_t^j, \tilde{X}_{t+1}^j)$. Although the GLR estimator is unbiased, it is more complicated to implement and has some limitations compared to other estimators. First, it is not clear which choices of $F_t(s)$ are "good". Moreover, since we are aiming to estimate the value function at $X_t \sim F_t$, extra simulation is needed to simulate these evaluation points. Moreover, if W is not differentiable w.r.t the state variable, (10) would introduce additional conditional expectation terms conditioning on points at which W is not differentiable. In such cases, the GLR method could be ineffective.

4 ALGORITHMS AND THEORETICAL ANALYSIS

In general, we update the parameters θ using

$$\theta_t^{n+1} = \theta_t^n + \alpha^n g_t^n \quad t = T, \ldots, 0, \quad (11)$$

where α_n is the step-size and $g_t^n = g_t(\theta_t^n, \theta_{t+1}^n)$ is the gradient estimator at iteration n. For different methods, we only need to replace $\hat{V}^\pi(x,t)$ in (4) by the corresponding estimators. For example, for FD, the gradient estimator is

$$g_t^{FD}(\theta_t^n, \theta_{t+1}^n; \varepsilon) = (\hat{V}^{FD}(x,t; V_{\theta_{t+1}^n}, \varepsilon) - V_{\theta_t^n}(x,t)) \nabla_{\theta_t^n} V_{\theta_t^n}(x,t). \quad (12)$$

The resulting policy evaluation algorithms with the FD and GLR estimators are given in Algorithms 1 and 2.

The convergence analysis of algorithms using FD and GLR estimators would be similar to that of TD and MC, which has been studied extensively in the literature. One common approach is to use ODE techniques to establish asymptotic convergence by showing that the ODE has a global asymptotically stable equilibrium (Tsitsiklis and Van Roy 1997). Moreover, under appropriate assumptions, convergence rate results can also be established (Bhandari et al. 2021). Here, we provide some insights into the proofs of the proposed algorithms by using previous results shown for stochastic approximation (SA) problems (L'Ecuyer and Yin 1998).

First, we define

$$\begin{aligned} \bar{\psi}_t(\theta_t) &= \mathbb{E}_{X_t \sim F_t} \left[\left(\mathscr{L} V_{\theta_{t+1}^*}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t) \right) \nabla V_{\theta_t}(X_t, t) \right], \\ \bar{g}_t(\theta_t, \theta_{t+1}) &= \mathbb{E}_{X_t \sim F_t} \left[\left(\mathscr{L} V_{\theta_{t+1}}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t) \right) \nabla V_{\theta_t}(X_t, t) \right]. \end{aligned}$$

Let (X_t, X_{t+1}) be a random pair, where $X_t \sim F_t$ and X_{t+1} is its corresponding next state, then

$$\begin{aligned} \psi_t(\theta_t) &= (\hat{V}(X_t, t; V_{\theta_{t+1}^*}) - V_{\theta_t}(X_t, t)) \nabla_{\theta_t} V_{\theta_t}(X_t, t), \\ g_t(\theta_t, \theta_{t+1}) &= (\hat{V}(X_t, t; V_{\theta_{t+1}}) - V_{\theta_t}(X_t, t)) \nabla_{\theta_t} V_{\theta_t}(X_t, t) \end{aligned}$$

are stochastic estimators of $\bar{\psi}_t(\theta_t)$ and $\bar{g}_t(\theta_t, \theta_{t+1})$, respectively. It is easily seen that $\bar{\psi}_t(\theta_t)$ is the negative of the gradient of the objective in,

$$\min_{\theta_t} \frac{1}{2} \left\| \left(\mathscr{L} V_{\theta_{t+1}^*}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t) \right)^2 \right\|_{F_t}, \quad (13)$$

Algorithm 1: Policy Evaluation with FD.

Input: fixed policy π, value function approximation model V_θ, initial number of sample paths M_0, simulation budget M in each iteration, number of evaluation points for each t in each iteration, N_0, step sizes sequences $\{\alpha_t^n\}$, initial parameters θ^0, initial permutation parameter $\varepsilon^0 > 0$.

Simulate M_0 sample paths $\mathscr{D}^0 \leftarrow \{X_t^j, j = 1,\ldots,M_0, t = 0,...,T\}$ under policy π.

for $n = 0, 1, \ldots$ **do**

  Simulate M sample paths $\mathscr{S}^n = \{X_t^j, j = 1,\ldots,M, t = 0,...,T\}$ under policy π, and $\mathscr{D}^{n+1} = \mathscr{D}^n \bigcup \mathscr{S}^n$.

  $\varepsilon^n \leftarrow \varepsilon^0 n^{-1/6}$.

  for $t = T, T-1, \ldots, 0$ **do**

    $\mathscr{C}^{n+1} \leftarrow N_0$ numbers randomly selected from $\{1,\ldots,|\mathscr{D}^{n+1}|\}$ with equal probability.

    Evaluation: For each $X_t^i, i \in \mathscr{C}^{n+1}$, calculate the FD estimator $\hat{V}^{FD}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \mathscr{D}^{n+1}, \varepsilon^n)$.

    Batch Gradient Estimator: For each $X_t^i, i \in \mathscr{C}^{n+1}$, calculate the gradient estimator by (12), denoted by $g_t^{n,i}$, the batch estimator $g_t^n \leftarrow (1/|\mathscr{C}^{n+1}|)\sum_{i\in\mathscr{C}^{n+1}} g_t^{n,i}$.

    Update: $\theta_t^{n+1} \leftarrow \theta_t^n + \alpha_t^n g_t^n$.

  end

end

which is very similar to (3) except that in (13) the optimal θ_{t+1}^* is given. Using the update $\theta_t^{n+1} = \theta_t^{n+1} + \alpha_t^n \psi_t(\theta_t^n)$ with $\alpha_t^n = O(n^{-1})$, an $O\left(n^{-\min\{2\beta,1+\nu\}}\right)$ convergence rate of $\mathbb{E}\left[\|\theta_t^n - \theta_t^*\|^2\right]$ can be established under assumptions on the decay rates of (14) and (15), along with other assumptions on $\bar{\psi}$ (see Theorem 3.1 in L'Ecuyer and Yin (1998)).

$$\|\bar{\psi}(\theta_t^n) - \mathbb{E}[\psi_t^n|\theta_t^n]\| \le K_\beta n^{-\beta} \ w.p.1 \tag{14}$$

$$\mathbb{E}\left[\|\bar{\psi}(\theta_t^n) - \mathbb{E}[\psi_t^n|\theta_t^n]\|^2\right] \le K_\nu n^{-\nu}. \tag{15}$$

Similar results could be expected to hold for our update (11) under similar assumptions, i.e., we would require the decaying behavior of the conditional bias and variance of the estimator g_t^n to follow (14) and (15). In practice, for FD, the conditional bias can be reduced by using a decreasing perturbation parameter along with an increasing number of sample paths, while for GLR, the bias is 0, since the estimator is unbiased.

5 NUMERICAL EXPERIMENTS

In this section, we compare four algorithms with different gradient estimators: the MC estimator, the temporal difference estimator, the FD estimator and the generalized likelihood ratio estimator. Algorithms are tested on two benchmark problems. In both experiments, we use a 2nd-degree polynomial model to approximate the value function.

5.1 Optimal Asset Allocation and Consumption

We consider policy evaluation in the problem of finding optimal asset allocation and consumption to maximize aggregated utility of consumption (Rao and Jelvis 2022). Suppose we have one risky asset and one riskless asset. The value of the risky asset S_t follows a geometric Brownian process $dS_t = \mu S_t dt + \sigma S_t dZ_t$, where μ is a drift constant and σ is a volatility constant, Z_t is a standard Brownian motion. In addition, the riskless asset $\tilde{S}_t$ follows $d\tilde{S}_t = r\tilde{S}_t dt$, where r is the riskfree rate. Let X_t be the wealth at time t, and denote

Algorithm 2: Policy Evaluation with GLR.

Input: fixed policy π, value function approximation model V_θ, initial number of sample paths M_0 and $\tilde{M}_0$, simulation budget M and $\tilde{M}$ in each iteration, step sizes sequences $\{\alpha_t^n\}$, initial parameters θ^0, user-specified simulation distributions $F_t^s(\cdot)$, $t = 0,\ldots,T$.

construct the initial set of evaluation points: simulate M_0 sample paths $\mathscr{D}^0 \leftarrow \{X_t^j, j = 1,\ldots,M_0, t = 0,...,T\}$ under policy π.

For each $t = 0,\ldots,T$, simulate $\tilde{M}_0$ sample paths $\tilde{\mathscr{D}}_t^0 = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1,\ldots,\tilde{M}_0\}$ under policy π, where $\tilde{X}_t^j \sim F_t^s$.

for $n = 0,1,\ldots$ **do**

Simulate M_0 sample paths $\mathscr{S}^n \leftarrow \{X_t^j, j = 1,\ldots,M_0, t = 0,...,T\}$ under policy π following the dynamics. $\mathscr{D}^{n+1} \leftarrow \mathscr{D}^n \bigcup \mathscr{S}^n$.

For each $t = 0,\ldots,T$, simulate $\tilde{M}$ sample paths $\tilde{\mathscr{S}}_t^n = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1,\ldots,\tilde{M}\}$ under policy π, where $\tilde{X}_t^j \sim F_t^s$, and $\tilde{\mathscr{D}}_t^{n+1} \leftarrow \tilde{\mathscr{D}}_t^n \bigcup \tilde{\mathscr{S}}_t^n$.

for $t = T, T-1,\ldots,0$ **do**

$\mathscr{C}^{n+1} \leftarrow N_0$ numbers randomly selected from $\{1,\ldots,|\mathscr{D}^{n+1}|\}$ with equal probability.

Evaluation: For each $X_t^i, i \in \mathscr{C}^{n+1}$, calculate the GLR estimator $\hat{V}^{GLR}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \tilde{\mathscr{D}}_t^{n+1})$.

Batch Gradient Estimator: For each $X_t^i, i \in \mathscr{C}^{n+1}$, calculate the gradient estimator, denoted by $g_t^{n,i}$, the batch estimator $g_t^n \leftarrow (1/|\mathscr{C}^{n+1}|)\sum_{i\in\mathscr{C}^{n+1}} g_t^{n,i}$.

Update: $\theta_t^{n+1} \leftarrow \theta_t^n + \alpha_t^n g_t^n$.

end

end

by $p(X_t,t)$ the percentage of wealth allocated to the risky asset and $1-p(X_t,t)$ the percentage of wealth allocated to the riskless asset. Let $c(X_t,t)$ be the wealth consumption per unit time. Then the wealth X_t evolves according to

$$dX_t = ((p(X_t,t)(\mu - r) + r)X_t - c(X_t,t))dt + p(X_t,t)\sigma X_t dZ_t.$$

The optimal value function is given by

$$V^*(x,t) = \max_{p,c} \mathbb{E}\left[\int_t^T e^{-\rho(s-t)}U(c(X_s,s))ds + e^{-\rho(s-t)}B(T)U(X_T)|X_t = x,t\right]$$

where $B(T) = \varepsilon^\eta$ is a fixed function, $\varepsilon, \eta \in (0,1)$, $U(x)$ is the utility of consumption function, $\rho \geq 0$ is the utility discount rate. Suppose $U(x) = \frac{x^{1-\eta}}{1-\eta}$, and the optimal action policy and the associated optimal value function have closed forms. Thus, we can apply policy evaluation algorithms for the optimal action policy (p^*, c^*) and compare the estimated value function with the optimal value function. First, we approximate the original continuous-time problem by discretization. The one-step transition thus can be written as

$$X_{t+1} = X_t + ((p_t^*(X_t,t)(\mu - r) + r)X_t - c^*(X_t,t))\Delta t + p_t^*(X_t,t)\sigma X_t \cdot \tilde{Z}_{t+1}, \tag{16}$$

where Δt is a small time interval, $\tilde{Z}_{t+1} \sim \mathscr{N}(0, \sqrt{\Delta t})$. The Bellman operator can be approximated by

$$\mathscr{L}V(x,t+1) = \mathbb{E}\left[e^{-\rho\cdot\Delta t}U(c^*(X_t,t))\Delta t + e^{-\rho\cdot\Delta t}V(X_{t+1},t+1)|X_t = x,t\right]. \tag{17}$$

The MC, TD and FD estimators can be easily derived from previous discussions. For GLR, we have to specify a distribution F_t^s. Intuitively, we would hope F_t^s be close to F_t (see Figure 1), but this usually

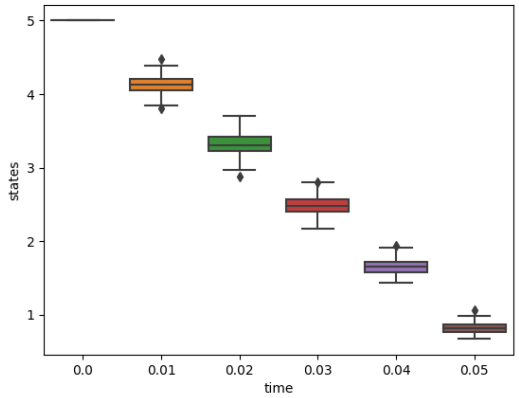

Figure 1: Boxplot of $F_t(\cdot)$. Parameters used: $T = 0.05$, $\Delta t = 0.01$, $x_0 = 10$, $\mu = 1$, $\sigma = 2$, $r = 0.8$, $\rho = 1.2$, $\eta = 0.4$, $\varepsilon = 0.01$.

cannot be achieved in practice. Instead, we set F_t^s to be $exp(\lambda_t)$, where $\lambda_t = 1/(at+b)$, and a and b are parameters fitted by a linear regression of the sample means $\{\bar{X}_t, t = 0,\ldots,T\}$ that is calculated from some pre-simulated sample paths starting from an initial wealth x_0. We use an exponential distribution for F_t^s, since the support of the X_t is non-negative under the optimal action policy. However, from Assumption 1, f_t^s should have unbounded support for GLR to be valid. Therefore, we need to perform a change of variable, e.g., $Y_t = \ln(X_t)$ in (17). Above all, in nth iteration, the GLR estimator (see Algorithm 2) is:

$$\begin{aligned}
&\hat{V}^{GLR}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \tilde{\mathscr{D}}_t^{n+1}) \\
=&\frac{e^{-\rho\cdot\Delta t}}{|\tilde{\mathscr{D}}_t^{n+1}|\cdot f_t^s(X_t^i)\cdot X_t^i}\sum_{j=1}^{|\tilde{\mathscr{D}}_t^{n+1}|}\left[\mathbf{1}\left\{\tilde{X}_t^j \le X_t^i\right\}\left(\frac{dV_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1)}{d\tilde{Y}_t^j} + V_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1)\cdot q(\tilde{Y}_t^j)\right)\right] \\
&+ e^{-\rho\cdot\Delta t}U(c^*(X_t^i, t))\Delta t,
\end{aligned}$$

where $f_t^s(x) = \lambda e^{-\lambda x}$, $\tilde{X}_t^j \sim exp(\lambda_t)$, $\tilde{Y}_t^j = \ln(\tilde{X}_t^j)$ with density $f_Y(y) = \lambda e^{y-\lambda e^y}$, and $q(y) := \frac{d\ln(f_y(y))}{dy} = 1 - \lambda e^y$. The gradient $\frac{dV_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1)}{d\tilde{Y}_t^j}$ can be calculated by taking derivatives of the parameterized approximation model and the transition function (16) and applying the chain rule.

Numerical results are shown in Figure 2, from which we can see the algorithm with FD has the best performance as it obtains a closest approximation of the true value function, and is the most stable. The GLR estimator has the worst performance, but becomes comparable to other methods after many epochs. It is worth mentioning that for the GLR method, there are many hyperparameters that can be tuned, such as F_t^s. The choice of F_t^s could have a great impact on the performances and how to design a "good" one still remains a challenging problem.

5.2 (s,S) Inventory Policy Evaluation

We consider a classic periodic review inventory control problem with zero lead time, i.e., an order arrives as soon as it is placed. Let X_t be the inventory level at the beginning of day t, $t = 0,\ldots,T$, a_t be the order amount between t and $t+1$, c_h be the unit holding cost, c_s be the unit shortage cost, $c_s > c_h$, c be the unit order cost and $K > 0$ be the fixed reorder cost. The cost R_{t+1} incurred between t and $t+1$ is given by

$$c_h \max\{X_t, 0\} + c_s \max\{-X_t, 0\} + ca_t + K\mathbf{1}\{a_t > 0\}.$$

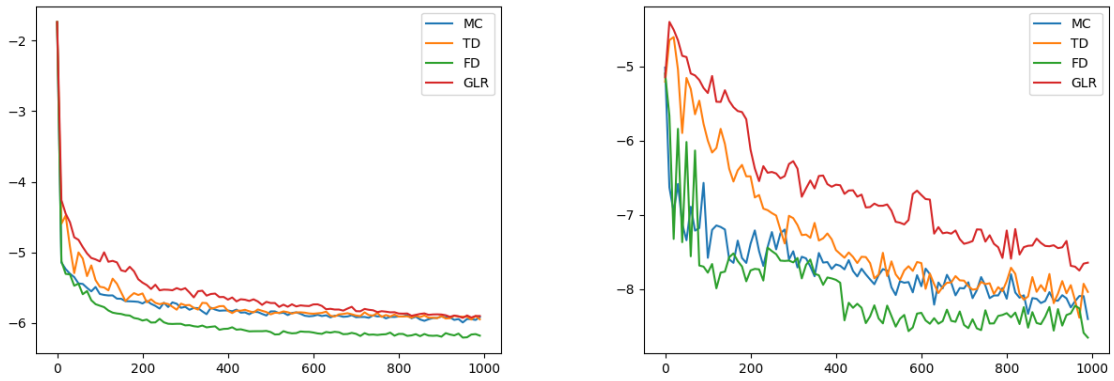

Figure 2: The left and right plots represent the logarithmic of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Results are reported every 10 epochs. For FD, $M_0 = 200$, $M = 1$, $N_0 = 200$. For GLR, $M_0 = \tilde{M}_0 = 200$, $M = \tilde{M} = 1$, $N_0 = 200$.

Assume that the demand Z_{t+1} between t and $t+1$, $t = 0, \ldots, T-1$ are i.i.d. $\exp(\lambda)$ random variables. Then the inventory X_{t+1} is obtained by the following transition

$$X_{t+1} = X_t + a_t - Z_{t+1}.$$

The policy we would like to evaluate is the well-known (s,S) policy, $S > s$, that is

$$a_t(X_t) = \begin{cases} S - X_t & X_t \leq s \\ 0 & X_t > s \end{cases}.$$

For this problem, the action policy $a_t(x)$ is not differentiable at $x = s$, therefore, the GLR method would yield an additional conditional expectation conditioning on $X_t = s$. Considering this issue, we only compare MC, TD and FD for this experiment. Numerical results are shown in Figure 3, from which we can see that FD achieves a higher accuracy compared to MC and TD after around 350 epochs. Besides, the stability of the three methods are comparable. In practice, in one iteration, FD requires a higher computational cost than MC and TD, since FD uses a set of sample paths instead of a single one. However, the benefit is that this set of sample paths can be used to construct estimators of the value function for many points in one iteration. This feature makes the policy evaluation algorithm with FD suitable for parallel computing, which could greatly reduce the computational time.

6 CONCLUSION

We have proposed two new estimators for policy evaluation, the FD estimator and the GLR estimator, based on the key observation that the conditional expectation in the Bellman error can be represented as a ratio of two ordinary expectations. These two estimators allow us the flexibility to estimate the value of any given state using sample paths starting from other states. However, there are still many practical issues that need further investigation. For example, the number of sample paths used in each iteration could greatly affect the performance of the algorithms. Moreover, for GLR, although theoretically it can yield an unbiased estimator, how to choose an appropriate F_t^s remains another practical challenge. One possible way is to generate sample paths following F_t as the iteration proceeds and dynamically adjusting F_t^s based on collected replications. GLR is theoretically appealing, but has many challenges for practical implementation and is more suitable when the action policy and the value function are differentiable.

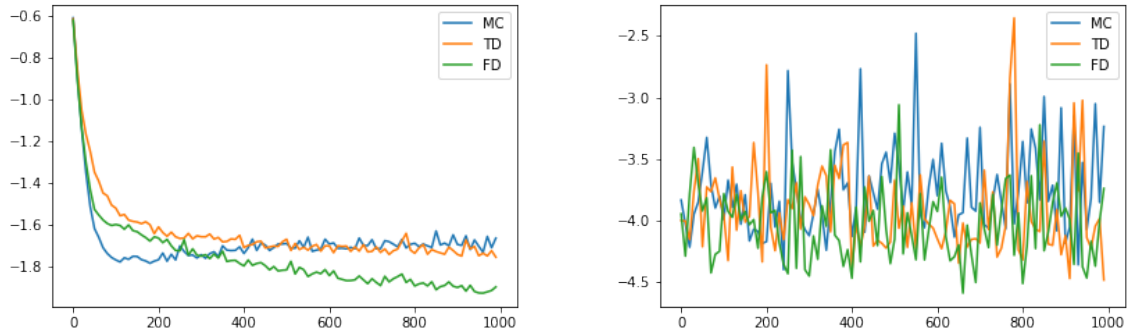

Figure 3: (s,S) policy evaluation. The left and right plots represent the logarithmic of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Parameters used: $T=5$, $\gamma=0.9$, $\lambda=0.2$, $c_h=3$, $c_s=5$, $c=3$, $K=5$. Results are reported every 10 epochs. For FD, $M_0=200$, $M=1$, $N_0=200$.

ACKNOWLEDGMENTS

This work was supported in part by the Air Force Office of Scientific Research under Grant FA95502010211.

REFERENCES

Bhandari, J., D. Russo, and R. Singal. 2021. "A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation". *Operations Research* 69(3):950–973.

Daveloose, C., A. Khedher, and M. Vanmaele. 2019. "Representations for Conditional Expectations and Applications to Pricing and Hedging of Financial Products in Lévy and Jump-diffusion Setting". *Stochastic Analysis and Applications* 37(2):281–319.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. New York: Springer.

Glynn, P. W. 1987. "Likelilood Ratio Gradient Estimation: An Overview". In *Proceedings of the 1987 Winter Simulation Conference*, edited by A. Thesen, H. Grant, and W. D. Kelton, 366–375. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ho, Y. C., and X. Cao. 1983. "Perturbation Analysis and Optimization of Queueing Networks". *Journal of Optimization Theory and Applications* 40(4):559–582.

Kushner, H., and G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. New York: Springer.

L'Ecuyer, P., and G. Yin. 1998. "Budget-dependent Convergence Rate of Stochastic Approximation". *SIAM Journal on Optimization* 8(1):217–247.

Longstaff, F. A., and E. S. Schwartz. 2001. "Valuing American Options by Simulation: A Simple Least-squares Approach". *The Review of Financial Studies* 14(1):113–147.

Neuneier, R. 1995. "Optimal Asset Allocation Using Adaptive Dynamic Programming". In *Advances in Neural Information Processing Systems, Vol 8*, edited by D. Touretzky, M. Mozer, and M. Hasselmo, 952–958. Cambridge, Massachusetts: MIT Press.

Oroojlooyjadid, A., M. Nazari, L. V. Snyder, and M. Takáč. 2022. "A Deep Q-Network for the Beer Game: Deep Reinforcement Learning for Inventory Optimization". *Manufacturing & Service Operations Management* 24(1):285–304.

Peng, Y., M. C. Fu, J.-Q. Hu, and B. Heidergott. 2018. "A New Unbiased Stochastic Derivative Estimator for Discontinuous Sample Performances with Structural Parameters". *Operations Research* 66(2):487–499.

Rao, Ashwin and Jelvis, Tikhon 2022. "Foundations of Reinforcement Learning with Applications in Finance". https://stanford.edu/~ashlearn/RLForFinanceBook/book.pdf, accessed 12nd March 2022.

Sutton, R. S. 1988. "Learning to Predict by the Methods of Temporal Differences". *Machine Learning* 3(1):9–44.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. 2 ed. Cambridge, Massachusetts: A Bradford Book.

Sutton, R. S., D. McAllester, S. Singh, and Y. Mansour. 1999. "Policy Gradient methods for Reinforcement Learning with Function Approximation". In *Advances in Neural Information Processing Systems, Vol 12*, edited by S. Solla, T. Leen, and K. Müller, 1057–1063. Cambridge, Massachusetts: MIT Press.

Tsitsiklis, J., and B. Van Roy. 1997. "An Analysis of Temporal-difference Learning with Function Approximation". *IEEE Transactions on Automatic Control* 42(5):674–690.

Tsitsiklis, J. N. 1994. "Asynchronous Stochastic Approximation and Q-learning". *Machine Learning* 16(3):185–202.

Zhou, Y., M. C. Fu, and I. O. Ryzhov. 2021. "Estimating a Conditional Expectation with the Generalized Likelihood Ratio Method". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

YI ZHOU is a Ph.D. candidate in Applied Mathematics, Statistics and Scientific Computation at the Department of Mathematics at the University of Maryland, College Park. Her research focuses on stochastic optimization for operations research problems. Her email is yzh@umd.edu.

MICHAEL C. FU holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland, College Park. His research interests include stochastic gradient estimation, simulation optimization, and applied probability. He served as WSC2011 Program Chair, NSF Operations Research Program Director, *Management Science* Stochastic Models and Simulation Department Editor, and *Operations Research* Simulation Area Editor. He received the INFORMS Simulation Society's Distinguished Service Award in 2018. He is a Fellow of INFORMS and IEEE. His e-mail addresses is mfu@umd.edu.

ILYA O. RYZHOV is an Associate Professor of Operations Management in the Decision, Operations and Information Technologies department of the Robert H. Smith School of Business at the University of Maryland. His research primarily focuses on simulation optimization and statistical learning, with applications in business analytics, revenue management, and nonprofit/humanitarian operations. He is a coauthor of the book *Optimal Learning* (Wiley, 2012). His work was recognized in WSC's Best Theoretical Paper Award competition on three separate occasions (winner in 2012, finalist in 2009 and 2016), and he received I-SIM's Outstanding Publication Award in 2017. His email address is iryzhov@umd.edu.