Stochastic Variance Reduction Methods for Policy Evaluation

Simon S. Du*, Jianshu Chen†, Lihong Li†, Lin Xiao†, Dengyong Zhou†
*Carnegie Mellon University †Microsoft Research

Policy Evaluation Background

Markov Reward Process

- S: State Space.
- A: Action Space.
- π: Policy $S \to A$.
- $P(s'|s,a)$: Transition Probability.
- $R(s)$: Reward function.

Value Function

- Long term payoff of a given policy.
- Def: $V^\pi(s) = E\left[r_1 + \gamma r_2 + \ldots | s_1 = s\right]$.
- Linear Policy Evaluation:

$$\widehat{V}(s) = \phi(s)^\top \theta$$

  - $\phi(s)$: d-dimensional feature vector.
  - $\theta \in \mathbb{R}^d$: coefficients to be learned.

Objective Function and Reformulation

Data: trajectory of n steps: $(s_1, r_1, s_2, r_2, \ldots, s_n, r_n)$ with $\phi_t = \phi(s)$.
Goal: Minimize Empirical MSPBE (Mean Square Projected Bellman Error):

$$\min_\theta \frac{1}{2}\left|\left|A\theta - b\right|\right|^2_{C^{-1}} + \frac{\rho}{2}\left|\left|\theta\right|\right|^2_2$$

$$A = \frac{1}{n}\sum_{t=1}^n A_t = \phi_t\left(\phi_t - \phi_{t+1}\right)^\top, b = \frac{1}{n} b_t = r_t\phi_t, C = \frac{1}{n}\phi_t\phi_t^\top$$

Challenges:

- The objective does NOT have the finite sum structure (preclude modern methods).
- Even computing the gradient requires $O(nd^2)$ complexity.

Key Observation: minimizing Regularized Empirical MSPBE is equivalent to

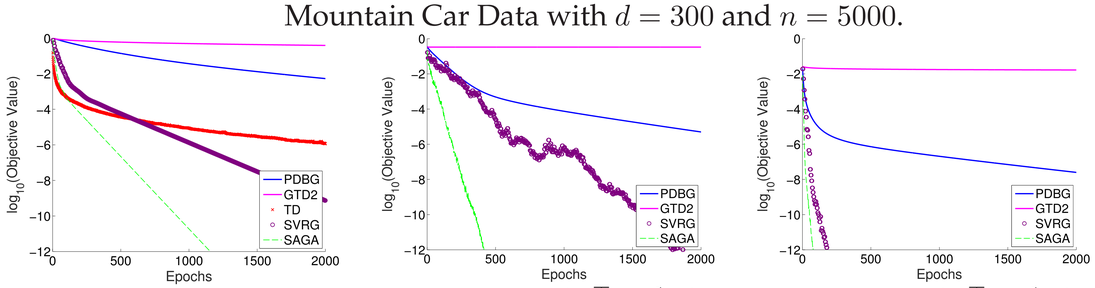
$$\min_\theta \max_w \left\{ w^\top(b - A\theta) - \frac{1}{2}\left|\left|w\right|\right|^2_C + \frac{\rho}{2}\left|\left|\theta\right|\right|^2_2 \right\}$$

$$= \min_\theta \max_w \frac{1}{n}\sum_{t=1}^n \left\{ w^\top(b_t - A_t\theta) - \frac{1}{2}\left(w^\top C_t\right)^2 + \frac{\rho}{2}\left|\left|\theta\right|\right|^2_2 \right\}$$

- Has the FINITE SUM structure.
- Stochastic gradient $B_t(\theta, w) = \begin{bmatrix} \rho\theta - A_t w \\ A_t\theta - b_t + C_t w \end{bmatrix}$ can be computed in $O(d)$ time and $B(\theta, w) = \sum_{t=1}^n B_t(\theta, w)$ can be computed in $O(nd)$ time.

Experiment on Mountain Car Data

Mountain Car Data with $d = 300$ and $n = 5000$.



$\rho = 0$. $\rho = 0.01\lambda_{\max}(A^\top C^{-1} A)$. $\rho = 0.01\lambda_{\max}(A^\top C^{-1} A)$.

Algorithms, Complexities, Extensions

Algorithm: Primal-Dual-Batch-Gradient (PDBG)

Input: initial point (θ, w), step sizes σ_θ and σ_w, and number of epochs M.

For $i = 1$ **to** M: $\begin{bmatrix}\theta \\ w\end{bmatrix} \leftarrow \begin{bmatrix}\theta \\ w\end{bmatrix} - \begin{bmatrix}\sigma_\theta & 0 \\ 0 & \sigma_w\end{bmatrix} B(\theta, w)$

Algorithm: SVRG for Policy Evaluation

Input: initial point (θ, w), step sizes $\{\sigma_\theta, \sigma_w\}$, outer iterations M, and inner iterations N.
For $m = 1$ **to** M:
  Initialize $(\tilde{\theta}, \tilde{w}) = (\theta, w)$ and compute $B(\tilde{\theta}, \tilde{w})$.
  For $j = 1$ **to** N:
    Sample an index t_j from $\{1, \cdots, n\}$ and update using variance-reduced gradient:

$$\begin{bmatrix}\theta \\ w\end{bmatrix} \leftarrow \begin{bmatrix}\theta \\ w\end{bmatrix} - \begin{bmatrix}\sigma_\theta & 0 \\ 0 & \sigma_w\end{bmatrix}\left(B_{t_j}(\theta, w) + B(\tilde{\theta}, \tilde{w}) - B_{t_j}(\tilde{\theta}, \tilde{w})\right).$$

Algorithm: SAGA for Policy Evaluation

Input: initial point (θ, w), step sizes σ_θ and σ_w, and number of iterations M.
Compute each $g_t = B_t(\theta, w)$ for $t = 1, \ldots, n$ and $B = B(\theta, w) = \frac{1}{n}\sum_{t=1}^n g_t$.
For $m = 1$ **to** M:
  Sample an index t_m from $\{1, \cdots, n\}$, compute $h_{t_m} = B_{t_m}(\theta, w)$ and update:

$$\begin{bmatrix}\theta \\ w\end{bmatrix} \leftarrow \begin{bmatrix}\theta \\ w\end{bmatrix} - \begin{bmatrix}\sigma_\theta & 0 \\ 0 & \sigma_w\end{bmatrix}(B + h_{t_m} - g_{t_m}), B \leftarrow B + \frac{1}{n}(h_{t_m} - g_{t_m}), g_{t_m} \leftarrow h_{t_m}.$$

Algorithm	Total Complexity $O(\cdot)$
SVRG / SAGA (this paper)	$nd(1 + \frac{\kappa_1}{n})\log\left(1/\epsilon\right)$
GTD2 (Sutton et al, 2009)	$d\kappa_2/\epsilon$
PDBG (this paper)	$nd\kappa_3\log(1/\epsilon)$
LSTD (Boyan, 2002)	nd^2

$\kappa_1, \kappa_2, \kappa_3$ are condition numbers.

Our algorithms extend directly to:

- off-policy PE: data sampled from a different distribution, i.e.,"covariate shift".
- PE with eligibility traces, which smooth the gap between MSPBE and MSE.

The only difference is in how A, B,C are formed.

Conclusion and Future Directions

Conclusions:

1. A saddle-point formulation of (batch) policy evaluation.
2. First-order algorithms with linear convergence rate

Future Directions:

1. Extension to nonlinear value-function approximation.
2. Extension to the controlled case (i.e., policy optimization).