ODM with SGD

Haoyu Chen

January 2019

1 Set-up

Posit a parametric model for the Q-function

$$\mathbb{E}(Y|A,X) = \mu(A,X;\beta), \tag{1.1}$$

where $\beta \in \mathcal{B} \subseteq \mathbb{R}^d$. Assume $\mathbb{E}(Y|A,X) = \mu(A,X;\beta_0)$ for some $\beta_0 \in \mathcal{B}$. The optimal oracle decision is to choose action

$$A = I\{\mu(1,X,\beta_0) > \mu(0,X,\beta_0)\}. \tag{1.2}$$

If we can estimate β_0 using some process to get $\hat{\beta}$. The estimated optimal decision is

$$A = I\{\mu(1,X,\hat{\beta}) > \mu(0,X,\hat{\beta})\}. \tag{1.3}$$

In online decision-making, we can update the parameter estimator at each decision step and obtain a sequence of estimators $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_T$. It is suggested by Polyak and Juditsky to use the average $\bar{\beta}_t = t^{-1}\sum_{i=1}^{t}\hat{\beta}_i$ as the final estimator to accelerate the approximation. To address the exploration-and-exploitation dilemma, we adopt ε-greedy policy to make online decisions. At each decision step t, the propensity score $\pi(X) = P(A=1|X)$ is calculated using

$$\hat{\pi}_t(X) = (1-\varepsilon_t)I\{\mu(1,X,\bar{\beta}_t) > \mu(0,X,\bar{\beta}_t)\} + \frac{\varepsilon_t}{2}. \tag{1.4}$$

Algorithm 1: Online Decision-Making with SGD

Input: $\hat{\beta}_0 = \bar{\beta}_0 = 0$, $\hat{\pi}_0 = 1/2$, α_t, ε_t

for $t = 1$ **to** T **do**

Observe X_t

Sample A_t from Bernoulli$(\hat{\pi}_{t-1}(X_t))$

Observe Y_t, $O_t = (X_t, A_t, Y_t)$

Calculate the IPW gradient

$$g(\hat{\beta}_{t-1}; O_t) = \frac{\nabla \ell(\hat{\beta}_{t-1}; O_t) I\{A = 1\}}{2\hat{\pi}_{t-1}(X_t)} + \frac{\nabla \ell(\hat{\beta}_{t-1}; O_t) I\{A = 0\}}{2(1 - \hat{\pi}_{t-1}(X_t))} \tag{1.5}$$

Update $\hat{\beta}_t = \hat{\beta}_{t-1} - \alpha_t g(\hat{\beta}_{t-1}; O_t)$

Update $\bar{\beta}_t = (\hat{\beta}_t + (t-1)\bar{\beta}_{t-1})/t$

Update

$$\hat{\pi}_t(X) = (1 - \varepsilon_t) I\{\mu(1, X, \bar{\beta}_t) > \mu(0, X, \bar{\beta}_t)\} + \frac{\varepsilon_t}{2}$$

end

Define

$$\beta^* = \arg\min_{\beta} L(\beta) \equiv \mathbb{E}_{O \sim \mathcal{P}_O^r} \ell(\beta; O), \tag{1.6}$$

where $\mathcal{P}_O^r$ is the joint distribution of $O = (X, A, Y)$ under the random policy which selects $A \sim$ Bernoulli$(1/2)$ independently of $X \sim \mathcal{P}_X$, and $Y|X, A \sim \mathcal{P}_{Y|X,A}$. In comparison, denote $\mathcal{P}_O^\pi$ the joint distribution of O_t under the proposed policy in Algorithm 1. Check that

$$\begin{aligned}
&\mathbb{E}_{O_t \sim \mathcal{P}_O^\pi}(g(\hat{\beta}_{t-1}; O_t) | \mathcal{F}_{t-1}) \\
=&\mathbb{E}[\mathbb{E}(g(\hat{\beta}_{t-1}; O_t) | \mathcal{F}_{t-1}, X_t) | \mathcal{F}_{t-1}] \\
=&\mathbb{E}\left[\frac{\nabla \ell(\hat{\beta}_{t-1}; 1, X_t, \mu(1, X_t; \beta_0) \mathbb{E}(I\{A = 1\} | \mathcal{F}_{t-1}, X_t)}{2\hat{\pi}_{t-1}(X_t)} \middle| \mathcal{F}_{t-1} \right] \\
&+ \mathbb{E}\left[\frac{\nabla \ell(\hat{\beta}_{t-1}; 0, X_t, \mu(0, X_t; \beta_0)(1 - \hat{\pi}_{t-1}(X_t))}{2(1 - \hat{\pi}_{t-1}(X_t))} \middle| \mathcal{F}_{t-1} \right] \\
=&\mathbb{E}_{O_t \sim \mathcal{P}_O^r}[\ell(\beta; X_t, A_t, Y_t) | \mathcal{F}_{t-1}] \\
=&\nabla L(\hat{\beta}_{t-1}).
\end{aligned}$$

Following Theorem 2 of Polyak and Juditsky, write

$$\hat{\beta}_t = \hat{\beta}_{t-1} - \alpha_t g(\hat{\beta}_{t-1}; O_t) = \hat{\beta}_{t-1} - \alpha_t (R(\hat{\beta}_{t-1}) - \xi_t),$$

where $R(\beta) = \nabla L(\beta)$ and $\xi_t = R(\hat{\beta}_{t-1}) - g(\hat{\beta}_{t-1}; O_t)$. Then ξ_t is a martingale difference process since

$$\mathbb{E}_{O_t \sim \mathcal{P}_O^\pi}(\xi_t | \mathcal{F}_{t-1}) = \nabla L(\hat{\beta}_{t-1}) - \mathbb{E}_{O_t \sim \mathcal{P}_O^\pi}(g(\hat{\beta}_{t-1}; O_t) | \bar{O}_{t-1}) = 0.$$

A1. *$L(\beta)$ is continuously differentiable and strongly convex with constant $\lambda > 0$.*

A2. *$\nabla L(\beta)$ is L_0-Lipchitz continuous.*

A3. *The Hessian matrix $H(\beta) = \nabla^2 L(\beta)$ exists and is continuous in $\{\beta : \|\beta - \beta^*\|_2 < \delta\}$, and $H = H(\beta^*) \succ 0$.*

A4. *Let $v(\beta) = \mathbb{E}_{O\sim\mathcal{P}_O^r}[\|\nabla\ell(\beta; O)\|_2^2]$. There exists $C_1 > 0$ such that*

$$v(\beta) \le C_1(1 + \|\beta\|_2^2).$$

And there exists $C_2 > 0$ such that for all $\beta \in \mathcal{B}$,

$$\mathbb{E}_{O\sim\mathcal{P}_O^r}[\|\nabla\ell(\beta; O) - \nabla\ell(\beta^*; O)\|_2^2] \le C_2\|\beta - \beta^*\|_2^2.$$

Now, check the assumptions of Theorem 2 of Polyak and Juditsky. Let $V(\Delta) = L(\beta^* + \Delta) - L(\beta^*)$, then

- $V(0) = 0$ and $\nabla V(0) = \nabla L(\beta^*) = 0$.
- By A1, $L(\beta^* + \Delta) \ge L(\beta^*) + \nabla L(\beta^*)^\top\Delta + \lambda\|\Delta\|_2^2$, thus $V(\Delta) \ge \lambda\|\Delta\|_2^2$.
- $\|\nabla V(\beta^* + \Delta_1) - \nabla V(\beta^* + \Delta_2)\|_2^2 \le L_0\|\Delta_1 - \Delta_2\|_2^2$ by A2.
- $\nabla V(\beta - \beta^*)^\top R(\beta) = R(\beta)^\top R(\beta) > 0$ for all $\beta \neq \beta^*$.
- There exists l_0 such that $R(\beta)^\top R(\beta) = \nabla L(\beta)^\top \nabla L(\beta) > l_0 V(\beta - \beta^*) = l_0[L(\beta) - L(\beta^*)]$ for $\|\beta - \beta^*\|_2 < \delta$ by A3.

Therefore Assumption 1 of Theorem 2 of Polyak and Juditsky is verified. By A3, there exists $K_1 > 0$ such that

$$\|\nabla L(\beta) - H^*(\beta - \beta_0)\|_2^2 \le K_1\|\beta - \beta_0\|_2^2.$$

So Assumption 2 is satisfied. Decompose the noise vector as $\xi_t = \xi_t^* + \zeta_t(\hat\beta_{t-1})$ where $\xi_t^* = -g(\beta^*; O_t)$ and $\zeta_t(\hat\beta_{t-1}) = R(\hat\beta_{t-1}) - [g(\hat\beta_{t-1}; O_t) - g(\beta^*; O_t)]$. Then

$$\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}(\xi_t^*|\mathcal{F}_{t-1}) = -\nabla L(\beta^*) = 0.$$

Let $\Sigma = \mathbb{E}_{O\sim\mathcal{P}_O^r}\{\nabla\ell(\beta^*; O)[\nabla\ell(\beta^*; O)]^\top\}$ be the covariance matrix. Then

$$\begin{aligned}
&\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}(\xi_t^*\xi_T^{*\top}|\mathcal{F}_{t-1})\\
=&\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}(g(\beta^*; O_t)[g(\beta^*; O_t)]^\top|\mathcal{F}_{t-1})\\
=&\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}\left(\frac{\nabla\ell(\beta^*; O_t)[\nabla\ell(\beta^*; O_t)]^\top I\{A_t = 1\}}{2\hat\pi_{t-1}(X_t)} + \frac{\nabla\ell(\beta^*; O_t)[\nabla\ell(\beta^*; O_t)]^\top I\{A_t = 0\}}{2[1 - \hat\pi_{t-1}(X_t)]}\middle|\mathcal{F}_{t-1}\right)\\
=&\mathbb{E}_{O_t\sim\mathcal{P}_O^r}\{\nabla\ell(\beta^*; O_t)[\nabla\ell(\beta^*; O_t)]^\top|\mathcal{F}_{t-1}\}\\
=&\Sigma.
\end{aligned}$$

Similarly, $\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}(\|\xi_t^*\|_2^2|\mathcal{F}_{t-1}) = \mathbb{E}_{O_t\sim\mathcal{P}_O^r}[\|\nabla\ell(\beta^*;O_t)\|_2^2] = \mathrm{tr}(\Sigma)$. Thus

$$\sup_t \mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}(\|\xi_t^*\|_2^2 I\{\|\xi_t^*\|_2 > C\}|\mathcal{F}_{t-1}) \xrightarrow{p} 0 \text{ as } C\to\infty.$$

Note that $\|R(\hat\beta_{t-1})\|_2^2 = \|R(\hat\beta_{t-1}) - R(\beta^*)\|_2^2 \le L_0\|\hat\beta_{t-1}-\beta^*\|_2^2$ by A2, we have

$$\begin{aligned}
\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}[\|\zeta_t(\hat\beta_{t-1})\|_2^2|\mathcal{F}_{t-1}] &\le 2\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}[\|g(\hat\beta_{t-1};O_t) - g(\beta^*;O_t)\|_2^2|\mathcal{F}_{t-1}] + 2\|R(\hat\beta_{t-1})\|_2^2 \\
&= 2\mathbb{E}_{O_t\sim\mathcal{P}_O^r}[\|\nabla\ell(\hat\beta_{t-1};O_t) - \nabla\ell(\beta^*;O_t)\|_2^2] + 2\|R(\hat\beta_{t-1})\|_2^2 \\
&\le C\|\hat\beta_{t-1}-\beta^*\|_2^2
\end{aligned}$$

by A4. Finally,

$$\begin{aligned}
&\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}[\|\xi_t(\hat\beta_{t-1})\|_2^2|\mathcal{F}_{t-1}] + \|R(\hat\beta_{t-1})\|_2^2 \\
\le& 2\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}[\|\xi_t^*\|_2^2|\mathcal{F}_{t-1}] + 2\mathbb{E}_{O_t\sim\mathcal{P}_O^\pi}[\|\zeta_t(\hat\beta_{t-1})\|_2^2|\mathcal{F}_{t-1}] + \|R(\hat\beta_{t-1})\|_2^2 \\
\le& 2\mathrm{tr}(\Sigma) + 2C\|\hat\beta_{t-1}-\beta^*\|_2^2 + L_0\|\hat\beta_{t-1}-\beta^*\|_2^2 \\
\le& K_2(1+\|\hat\beta_{t-1}-\beta^*\|_2^2)
\end{aligned}$$

for some $K_2 > 0$. Therefore Assumption 3 is satisfied. Applying Theorem 2 of Polyak and Juditsky, we have

$$\sqrt{t}(\bar\beta_t - \beta^*) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = H^{-1}\Sigma(H^{-1})^\top$. If the loss function ℓ is well chosen, $\beta_0 = \beta^*$. The plugin estimators for Σ and H are

$$\hat\Sigma_t = \frac{1}{t}\sum_{s=1}^t g(\hat\beta_s;O_s)[g(\hat\beta_s;O_s)]^\top$$

and

$$\hat H_t = \frac{1}{t}\sum_{s=1}^t \nabla^2\ell(\hat\beta;O_s)\left[\frac{I\{A_s=1\}}{2\hat\pi_{s-1}(X_s)} + \frac{I\{A_s=0\}}{2(1-\hat\pi_{s-1}(X_s))}\right]$$

Example 1: Linear model

$$\mu(A, X; \beta) = u(A, X, \beta),$$

where $u(\cdot)$ is a linear function of β. Consider the mean square loss

$$\ell(\beta; O) = \frac{1}{2}(Y - \mu(A, X; \beta))^2.$$

The gradient is

$$\nabla \ell(\beta; O) = (u(A, X; \beta) - Y)\nabla u(A, X, \beta)$$

and

$$\begin{aligned}
\nabla L(\beta) &= \mathbb{E}[(u(A, X; \beta) - u(A, X; \beta_0))\nabla u(A, X, \beta)] \\
&= \mathbb{E}[P(A = 1|X)(u(1, X; \beta) - u(1, X; \beta_0))\nabla u(1, X, \beta) \\
&\quad + P(A = 0|X)(u(0, X; \beta) - u(0, X; \beta_0))\nabla u(0, X, \beta)].
\end{aligned}$$

While

$$\begin{aligned}
\mathbb{E}(\nabla \ell(\beta; O_t)|\bar{O}_{t-1}) = \mathbb{E}[&P(A_t = 1|\bar{O}_{t-1}, X_t)(u(1, X_t; \beta) - u(1, X_t; \beta_0))\nabla u(1, X_t, \beta) \\
&+ P(A_t = 0|\bar{O}_{t-1}, X_t)(u(0, X_t; \beta) - u(0, X_t; \beta_0))\nabla u(0, X_t, \beta)].
\end{aligned}$$

In order for $\mathbb{E}(\xi_t|\mathcal{F}_{t-1}) = 0$, either

$$P(A = 1|X) = P(A_t = 1|\bar{O}_{t-1}, X_t),$$

or

$$(u(1, X; \beta) - u(1, X; \beta_0))\nabla u(1, X, \beta) = (u(0, X; \beta) - u(0, X; \beta_0))\nabla u(0, X, \beta).$$

Example 2: Logistic model

$$\mu(A, X; \beta) = \frac{1}{1 + e^{-u(A, X, \beta)}},$$

where $u(\cdot)$ is a linear function of β. Consider the cross entropy loss

$$\ell(\beta; O) = -Y \log \mu(A, X; \beta) - (1 - Y)\log(1 - \mu(A, X; \beta)).$$

The gradient is

$$\nabla \ell(\beta; O) = (\mu(A, X; \beta) - Y)\nabla u(A, X, \beta)$$

and

$$\nabla L(\beta) = \mathbb{E}[(\mu(A, X; \beta) - \mu(A, X; \beta_0))\nabla u(A, X, \beta)]$$

1.1 Definition of $L(\beta)$

Definition 1:

$$L(\beta) = \mathbb{E}\ell(\beta; O) = \int \ell(\beta; x, a, y)p(y|x, a)p(a|x)p(x)dydadx,$$

where $p(a|x)$ is determined by a fixed policy $\pi(x) = P(A = 1|X = x)$. It could be $\pi(x) = 1/2$ or $\pi(x) = P(\mu(1, x; \beta_0) > \mu(0, x; \beta_0))$. The point is the policy does not change with previous collected data.

Definition 2:

$$L(\beta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(\ell(\beta; O_t)|\bar{O}_{t-1}),$$

Following definition 2,

$$\begin{aligned}
\mathbb{E}(\xi_t|\mathcal{F}_{t-1}) &= \nabla L(\hat{\beta}_{t-1}) - \mathbb{E}(\nabla\ell(\hat{\beta}_{t-1}; O_t)|\bar{O}_{t-1}) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{s=1}^{T} \nabla\mathbb{E}(\ell(\hat{\beta}_{t-1}; O_s)|\bar{O}_{s-1}) - \nabla\mathbb{E}(\ell(\hat{\beta}_{t-1}; O_t)|\bar{O}_{t-1}) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{s=1, s\neq t}^{T} \Big[\nabla\mathbb{E}(\ell(\hat{\beta}_{t-1}; O_s)|\bar{O}_{s-1}) - \nabla\mathbb{E}(\ell(\hat{\beta}_{t-1}; O_t)|\bar{O}_{t-1})\Big]
\end{aligned}$$