Online Policy Evaluation

nakajmiya

January 5, 2026

1 オフ方策評価

通常，機械学習はデータに基づいた正確な予測を可能にする技術であり，幅広く利用されている．実際に多くの研究論文では，解く意味があるとされているタスクにおいて，より高精度な予測をすることを目指している．例えば，天気予報の降水確率のように，予測値をそのまま活用するケースもある．しかし，Web 産業に目をむけると，機械学習の予測値をそのまま用いるでなく，予測値に基づいてなんらかの意思決定をしている場合が多い．一例を挙げるならば，ユーザーの商品のクリック確率を予測するだけでなく，その予測値に基づいて次にどの商品を推薦するかを決定する場合がある．このような問題は，**予測問題の最適化というよりもむしろ意思決定の最適化問題**といえる．

意思決定問題の性能評価の理想的な方法としては，方策を環境，あるいは，サービスを実際に実装することで結果を直接的に計測するオンライン実験，あるいは A/B テストを行うことである．しかし，オンライン実験には時間や大きなコストが伴う．また，悪い意思決定を実験しまったときに，実験期間においてユーザの気分を害したり収益を減らすことがある．したがって，オンライン実験をなるべく避け，意思決定方策の性能を安全に評価するために，既存の意思決定方策によりすでに収集されたログデータのみを利用して，新しい方策の性能を見積もりたいというモチベーションが生まれてくる．

このように，未だ実装したことない新たな意思決定方策の性能を，ログデータのみを用いて評価する問題を**オフ方策評価**という．正確なオフ方策評価により，新たな方策の性能を見積もることができれば，時間のかかるオンライン実験よりも，意思決定方策の改善サイクルを素早く回すことができる．

1.1 オフ方策評価の定式化

特徴量を $x \in \mathcal{X}$，行動を $a \in \mathcal{A}$，報酬を $r \in \mathbb{R}$ とする．Table 1 に特徴量，行動，　目的変数の例を示す．例えば，映画推薦を例に挙げると，ユーザーの年齢や性別などの属性情報やのこれまで視聴履歴が代表的な特徴量となり，行動 a は各映画を表し，報酬の例としては，視聴の有無や視聴時間の長さなど，最適化したい指標が状況に応じて設定される．

ここで，方策を π とし，方策 $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ を行動空間 $\mathcal{A}$ を上の条件付き確率分布として導入する（意思決定方策 $\pi(a|x)$ は我々自身が実装するため既知）．つまり，これはユーザ x に対して，各行動 a を選択する確率を表す．なお，特徴量分布 $p(x)$，報酬分布 $p(r|x,a)$ は未知で我々の制御の及ばない分布である．

方策 π による一連の意思決定プロセスをまとめる．まず，未知の確率分布 $p(x)$ に従う特徴量 x_i を観測する．次に，観測した x_i に基づき，方策 π が行動を選択する．これは特徴量 x_i で表されるユーザに対して，推薦すべき映画や配布すべきクーポン，投与すべき薬の種類を決めるといった意思決定を行う場面に対応する．最後に，特徴量 x_i と選択された行動 a_i の両方に依存して，報酬 r_i が未知の確率分布 $p(r|x_i, a_i)$ に従って観測される（例えばクリック率を r とすれば，r は誰に何を推薦したかに応じて変化するはず）．

ここで，意思決定方策 π の性能について定義を行う．

Definition 1. 意思決定方策 π の性能（*Policy Value*）は次のように定義される．

$$\begin{aligned} V(\pi) &\coloneqq \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}\left[r\right] \\ &= \mathbb{E}_{p(x)\pi(a|x)}\left[q(x,a)\right]. \end{aligned} \tag{1}$$

なお，$q(x,a) \coloneqq \mathbb{E}_{p(r|x,a)}\left[r\right]$ は特徴量 x と行動 a を条件づけた報酬の期待値であり，期待報酬関数と呼ばれる（例は Table 2）．

オフ方策評価の目的は，方策 π の性能 $V(\pi)$ をできるだけ正確に推定しようとすることである．方策 π を環境に一定期間実装するオンライン実験が可能ならば，その期間に観測される報酬の経験平均を計算すれば，$V(\pi)$ を正確に推定することが可能であるが，オンライン実験には困難が伴う場合が多い．よって，オンライン実験の良い代替，もしくはオンライン実験を行うべき少数の有望な方策を特定するためにの安全かつ効率的な方法が求められる．より具体的には，すでに環境に実装・運用されている意思決定方策 π_0（データ収集方策）により収集されたログデータのみを利用して，π_0 とは異なる新たな方策 π の性能 $V(\pi)$ を推定する統計的推定問題を考える．

ここで，オフ方策評価に用いることができるログデータ $\mathcal{D}$ は，次の独立同一分布からの抽出により与えられると想定する．

$$\mathcal{D} \coloneqq \{(x_i, a_i, r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x_i)\pi_0(a_i|x)_i p(r_i|x_i, a_i). \tag{2}$$

ログデータ $\mathcal{D}$ とは，データ収集方策 π_0 により収集された特徴量 x_i，行動 a_i，報酬 r_i からなるサイズ n の集合である．

オフ方策評価における主な研究目標は，データ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ のみを用いて，評価対象の方策 π（評価方策）の性能 $V(\pi)$ をより正確に推定できる推定量 $\hat{V}$ を構築することである．ここで，推定量 $\hat{V}$ は，ログデータ $\mathcal{D}$ を用いて計算される関数であり，真の性能 $V(\pi)$ にできるだけ近い値を取ることが望ましい．つまり，

$$V(\pi) \approx \hat{V}(\pi; \mathcal{D}).$$

Table 1: 特徴量，行動，目的変数の例.

応用例	特徴量 x	行動 a	報酬 r
映画推薦	年齢，性別，映画視聴履歴	映画の種類	クリック有無，視聴時間
投薬	年齢，性別，体重，検査結果	薬の種類	生存有無，血糖値

Table 2: 映画推薦の問題における期待報酬関数 $q(x,a)$ の例．それぞれの映画を推薦したときの視聴の有無を表す 2 値変数を報酬 $r \in \{0,1\}$ とすれば，$q(x,a)$ はユーザ x に映画 a を推薦した際に視聴される確率と等しい．

	タイタニック (a_1)	アバター (a_2)	スラムダンク (a_3)
ユーザ 1 (x_1)	$q(x_1,a_1)=0.2$	$q(x_1,a_2)=0.1$	$q(x_1,a_3)=0.5$
ユーザ 2 (x_2)	$q(x_2,a_1)=0.5$	$q(x_2,a_2)=0.7$	$q(x_2,a_3)=0.4$
ユーザ 3 (x_3)	$q(x_3,a_1)=0.3$	$q(x_3,a_2)=0.6$	$q(x_3,a_3)=0.9$

を達成したい．

推定量 $\hat{V}$ の性能は，一般的に平均二乗誤差（Mean Squared Error; MSE）により定量化される．

Definition 2. ある評価方策 π が与えられたとき，その方策の真の性能 $V(\pi)$ に対する推定量 $\hat{V}(\pi;\mathcal{D})$ の平均二乗誤差は次のように定義される．

$$\mathrm{MSE}\left[\hat{V}(\pi;\mathcal{D})\right] \coloneqq \mathbb{E}_{p(\mathcal{D})}\left[(V(\pi)-\hat{V}(\pi;\mathcal{D}))^2\right]. \quad (3)$$

平均二乗誤差とは，方策 π の真の性能 $V(\pi)$ と推定量 $\hat{V}(\pi;\mathcal{D})$ の二乗誤差の期待値である．平均二乗誤差が小さいほど，推定量 $\hat{V}$ は真の性能 $V(\pi)$ に対してより正確な推定を行っているといえる．なお，平均二乗誤差は，バイアス（squared bias）と分散（variance）の和に分解することができる．

$$\begin{aligned}
&\mathrm{MSE}\left[\hat{V}(\pi;\mathcal{D})\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[(V(\pi)-\hat{V}(\pi;\mathcal{D}))^2\right] \\
&= V(\pi)^2 - 2V(\pi)\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})\right] + \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})^2\right] \\
&= \left(\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})\right] - V(\pi)\right)^2 \\
&\quad + \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})^2 - \left(\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})\right]\right)^2\right] \\
&= \mathrm{Bias}\left[\hat{V}(\pi;\mathcal{D})\right]^2 + \mathrm{Var}\left[\hat{V}(\pi;\mathcal{D})\right]. \quad (4)
\end{aligned}$$

ここで，

$$\mathrm{Bias}\left[\hat{V}(\pi;\mathcal{D})\right] \coloneqq \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})\right] - V(\pi) \quad (5)$$

$$\mathrm{Var}\left[\hat{V}(\pi;\mathcal{D})\right] \coloneqq \mathbb{E}_{p(\mathcal{D})}\left[\left(\hat{V}(\pi;\mathcal{D}) - \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}(\pi;\mathcal{D})\right]\right)^2\right] \quad (6)$$

はそれぞれ推定量 $\hat{V}$ のバイアスとバリアンスを指す．一般にバイアスとバリアンスはトレードオフの関係にあるため，バイアスを下げるとバリアンスが上がり，逆にバリアンスを下げるとバイアスが上がることが多い．オフ方策評価において，より良い平均二乗誤差を達成するためには，バイアスとバリアンスの双方を低く抑えることが重要なテーマとなる．

2 標準的な推定量とその性質

2.1 オンライン実験による方策性能推定

初めに，オンライン実験を通じた方策性能推定を行う．オンライン実験とは，評価方策 π そのものを環境に実装することで得たログデータを用いて，$V(\pi)$ を推定することを目指す．すなわちオンライン実験では，次のログデータを用いて推定を行う．

$$\mathcal{D}_{\mathrm{online}} \coloneqq \{(x_i,a_i,r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x_i)\underbrace{\pi(a_i|x_i)}_{\text{評価方策}}p(r_i|x_i,a_i). \quad (7)$$

ここで，$\mathcal{D}_{\mathrm{online}}$ は，評価方策 π 自身により収集されたログデータである．評価方策 π そのものが形成する同時分布 $p(x)\pi(a|x)p(r|x,a)$ からデータが生成されている点に注意されたい．オンライン実験を通じた方策の性能推定では，次に AVG 推定量がよく用いられる．

Definition 3. 評価方策 π のオンライン実験により収集したログデータ $\mathcal{D}_{\mathrm{online}}$ が与えられたとき，評価方策 π の性能 $V(\pi)$ に対する AVG 推定量は次のように定義される．

$$\hat{V}_{\mathrm{AVG}}(\pi;\mathcal{D}_{\mathrm{online}}) \coloneqq \frac{1}{n}\sum_{i=1}^n r_i. \quad (8)$$

AVG 推定量は，ログデータとして観測された報酬 $\{r_i\}_{i=1}^n$ の単純な平均値で定義される．

次に，AVG 推定量の性質について考察する．

Theorem 1. 評価方策 π のオンライン実験により収集したデータ $\mathcal{D}_{\mathrm{online}}$ を用いたとき，AVG 推定量は，真の性能 $V(\pi)$ に対する不偏推定量になる．すなわち，

$$\mathbb{E}_{p(\mathcal{D}_{\mathrm{online}})}\left[\hat{V}(\pi;\mathcal{D}_{\mathrm{online}})\right] = V(\pi). \quad (9)$$
$$\left(\Longrightarrow \mathrm{Bias}\left[\hat{V}(\pi;\mathcal{D}_{\mathrm{online}})\right] = 0\right)$$

Proof.

$$
\begin{aligned}
&\text{LHS of (9)} \\
&= \mathbb{E}_{p(\mathcal{D}_{\text{online}})}\left[\hat{V}_{\text{AVG}}(\pi;\mathcal{D}_{\text{online}})\right] \\
&= \mathbb{E}_{p(x_i)\pi(a_i|x_i)p(r_i|x_i,a_i)}\left[\frac{1}{n}\sum_{i=1}^{n} r_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x_i)\pi(a_i|x_i)p(r_i|x_i,a_i)}\left[r_i\right] \\
&= \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}\left[r\right] \because (x,a,r) \overset{\text{i.i.d.}}{\sim} p(x)\pi(a|x)p(r|x,a) \\
&= V(\pi) \\
&= \text{RHS of (9)} \qquad (10)
\end{aligned}
$$

□

つまり，仮にオンライン実験を行うことができるのであれば，単に AVG 推定量を考えるだけで，バイアスを生じない不偏推定が可能になる．オンライン実験の際に不偏推定量が多くの場合に用いられるのはこれに起因する．次に，バリアンスについて考察する．バリアンスは，全分散の公式

$$
\begin{aligned}
&\mathbb{V}_{p(x,y)}\left[f(x,y)\right] \\
&= \mathbb{E}_{p(x,y)}\left[(f(x,y))^2\right] - \left(\mathbb{E}_{p(x,y)}\left[(f(x,y))\right]\right)^2 \\
&= \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[(f(x,y))^2\right]\right] - \left(\mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[(f(x,y))\right]\right]\right)^2 \\
&= \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[(f(x,y))^2\right] - \left(\mathbb{E}_{p(y|x)}\left[(f(x,y))\right]\right)^2\right] \\
&+ \mathbb{E}_{p(x)}\left[\left(\mathbb{E}_{p(y|x)}\left[(f(x,y))\right]\right)^2\right] - \left(\mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[(f(x,y))\right]\right]\right)^2 \\
&= \mathbb{E}_{p(x)}\left[\mathbb{V}_{p(y|x)}\left[f(x,y)\right]\right] + \mathbb{V}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[f(x,y)\right]\right] \qquad (11)
\end{aligned}
$$

を複数回適用することで次のように求まる．

$$
\begin{aligned}
&\text{Var}\left[\hat{V}_{\text{AVG}}(\pi;\mathcal{D}_{\text{online}})\right] \\
&= \mathbb{V}_{p(\mathcal{D}_{\text{online}})}\left[\frac{1}{n}\sum_{i=1}^{n} r_i\right] \\
&= \frac{1}{n}\mathbb{V}_{p(x)\pi(a|x)p(r|x,a)}\left[r\right] \because (x,a,r) \overset{\text{i.i.d.}}{\sim} p(x)\pi(a|x)p(r|x,a) \\
&= \frac{1}{n}\left(\mathbb{E}_{p(x)\pi(a|x)}\left[\mathbb{V}_{p(r|x,a)}[r]\right] + \mathbb{V}_{p(x)\pi(a|x)}\left[\mathbb{E}_{p(r|x,a)}[r]\right]\right) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi(a|x)}\left[\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi(a|x)}\left[q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi(a|x)p(r|x,a)}\left[r\right]\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi(a|x)}\left[\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi(a|x)}\left[q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi(a|x)}\left[q(x,a)\right]\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi(a|x)}\left[\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi(a|x)}\left[q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]). \qquad (12)
\end{aligned}
$$

ここで，報酬の条件付き分散を $\sigma^2(x,a) \coloneqq \mathbb{V}_{p(r|x,a)[r]}$，簡略化のために期待報酬関数 $q(x,a)$ の方策 π に関する期待値を，$q(x,\pi) \coloneqq \mathbb{E}_{\pi(a|x)}[q(x,a)] = \mathbb{E}_{\pi(a|x)p(r|x,a)}[r]$ として表す．

以上までの導出をもとに AVG 推定量のバリアンスを整理すると，次のようになる．

Theorem 2. ある方策 π のオンライン実験により収集したデータ $\mathcal{D}_{\text{online}}$ を用いるとき，AVG 推定量は次の平均二乗誤差をもつ．

$$
\begin{aligned}
&\text{MSE}\left[\hat{V}(\pi;\mathcal{D}_{\text{online}})\right] \\
&= \text{Var}\left[\hat{V}(\pi;\mathcal{D}_{\text{online}})\right] \because \text{Bias}\left[\hat{V}(\pi;\mathcal{D}_{\text{online}})\right] = 0 \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi(a|x)}\left[\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi(a|x)}\left[q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]). \qquad (13)
\end{aligned}
$$

定理 2 は，オンライン実験を通じた方策の平均二乗誤差が，

1. 実験により収集したデータ数 n
2. 報酬ノイズ $\sigma^2(x,a)$
3. 各ユーザ内の期待報酬関数のばらつき度合い（の期待値）$\mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi(a|x)}\left[q(x,a)\right]\right]$
4. 異なるユーザ間の期待報酬関数のばらつき度合い $\mathbb{V}_{p(x)}\left[q(x,\pi)\right]$

によって決まることを示している．データ数 n が大きければ，平均二乗誤差は小さくなる一方で，報酬ノイズやユーザ内・ユーザ間の期待報酬関数のばらつき度合いといった我々に制御できない環境依存の要素が大きいと，オンライン実験を行なったとしても平均二乗誤差は大きくなる．

2.2 Direct Method（DM）推定量

本セクションから，オフ方策評価の問題を取り扱う．具体的には，評価方策 π とは異なるデータ収集方策 π_0 が収集したログデータ $\mathcal{D}$ を用いて，評価方策 π の性能 $V(\pi)$ を推定する問題を考える．まずはじめに，オフ方策評価における基本推定量の 1 つである Direct Method（DM）推定量を導入する．

Definition 4. データ収集方策 π が収集したログデータ $\mathcal{D}$ が与えられたとき，評価方策 π の性能 $V(\pi)$ に対する DM 推定量は次のように定義される．

$$
\begin{aligned}
\hat{V}_{\text{DM}}(\pi;\mathcal{D},\hat{q}) &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\hat{q}(x_i,\pi) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi(a|x_i)\hat{q}(x_i,a). \qquad (14)
\end{aligned}
$$

なお，$\hat{q}(x,a)$ は，期待報酬関数 $q(x,a)$ の推定モデルである．これは，報酬関数 r を目的変数とした次の教師あり学習問題を解くことで得られる．

$$
\hat{q}(x,a) = \arg\min_{q'\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}\ell_r\left(r_i, q'(x_i,a_i)\right). \qquad (15)
$$

ここで，ℓ_r は報酬 r に対する損失関数（交差エントロピー，二乗誤差等）であり，$\mathcal{Q}$ は報酬関数の推定モデル

Table 3: 映画推薦の問題における期待報酬関数の推定モデル $\hat{q}(x,a)$ の例.

	タイタニック (a_1)	アバター (a_2)	スラムダンク (a_3)
ユーザ 1 (x_1)	$\hat{q}(x_1,a_1)=0.1$	$\hat{q}(x_1,a_2)=0.2$	$\hat{q}(x_1,a_3)=0.2$
ユーザ 2 (x_2)	$\hat{q}(x_2,a_1)=0.4$	$\hat{q}(x_2,a_2)=0.3$	$\hat{q}(x_2,a_3)=0.3$
ユーザ 3 (x_3)	$\hat{q}(x_3,a_1)=0.5$	$\hat{q}(x_3,a_2)=0.8$	$\hat{q}(x_3,a_3)=0.9$

の仮説空間（リッジ回帰，ニューラルネットワーク，ランダムフォレスト等）である．

DM 推定量は，ログデータ $\mathcal{D}$ を用いて，報酬 r に対する予測誤差を最小化する基準で，評価方策 π の期待報酬関数 $q(x,a)$ を推定する（Table 3）．仮に報酬を精度良く予測できる推定モデル $\hat{q}(x,a)$ を得ることができれば，それを方策の性能の定義に代入することで，正確な方策評価を行うことができるはずである．この考えに基づいて，真の期待報酬関数 $q(x,a)$ をその推定モデル $\hat{q}(x,a)$ で代替すると，

$$\begin{aligned} V(\pi) &\approx \mathbb{E}_{p(x)\pi(a|x)}\left[\hat{q}(x,a)\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi(a|x_i)}\left[\hat{q}(x_i,a)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{q}(x_i,\pi) \quad \because q(x,\pi) \coloneqq \mathbb{E}_{\pi(a|x)}\left[q(x,a)\right] \\ &=: \hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q}) \end{aligned} \tag{16}$$

となり，DM 推定量の定義を得る．

ここから，DM 推定量の性質について考察する．まずバイアスについてだが，以下の定理が成立する．

Theorem 3. あるデータ収集方策 π_0 が収集したログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，DM 推定量は，真の性能 $V(\pi)$ に対する次のバイアスを持つ．

$$\mathrm{Bias}\left[\hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q})\right] = \mathbb{E}_{p(x)\pi(a|x)}\left[\Delta_{q,\hat{q}}(x,a)\right]. \tag{17}$$

なお，$\Delta_{q,\hat{q}}(x,a) \coloneqq \hat{q}(x,a) - q(x,a)$ は，期待報酬関数の推定モデル $\hat{q}(x,a)$ の予測誤差を表す．

Proof.

$$\begin{aligned} &\text{LHS of (17)} \\ &= \mathrm{Bias}\left[\hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q})\right] \\ &= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q})\right] - V(\pi) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}\left[\hat{q}(x,\pi)\right] - V(\pi) \\ &\quad \because (x,a,r) \overset{\text{i.i.d.}}{\sim} p(x)\pi(a|x)p(r|x,a) \\ &= \mathbb{E}_{p(x)\pi(a|x)}\left[\hat{q}(x,a) - q(x,a)\right] \\ &= \text{RHS of (17)}. \end{aligned} \tag{18}$$

□

この定理から，DM 推定量のバイアスが，期待報酬関数の推定モデルの予測性能 $\Delta_{q,\hat{q}}(x,a)$ で決定することがわかる（正確には評価方策 π に対する期待値となっており，評価方策 π がより高い確率で選択する行動 $a \in \mathcal{A}$ に関する予測誤差が，DM 推定量におけるバイアスへの寄与度が大きい）．すなわち，期待報酬関数の推定モデルが真の期待報酬関数に対してどれだけ予測誤差を持つかが，DM 推定量のバイアスに影響を与える．続いて DM 推定量のバリアンスについて計算する．

$$\begin{aligned} \mathrm{Var}\left[\hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q})\right] &= \mathbb{V}_{p(\mathcal{D})}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{q}(x_i,\pi)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\hat{q}(x,\pi)\right] \\ &= \frac{1}{n}\mathbb{V}_{p(x)}\left[\hat{q}(x,\pi)\right]. \end{aligned} \tag{19}$$

バイアスとバリアンスの計算をまとめると，次の定理が成立する．

Theorem 4. あるデータ収集方策 π_0 が収集したログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，DM 推定量は，真の性能 $V(\pi)$ に対する次の平均二乗誤差を持つ．

$$\begin{aligned} &\mathrm{MSE}\left[\hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q})\right] \\ &= \mathbb{E}_{p(x)\pi(a|x)}\left[\Delta_{q,\hat{q}}(x,a)\right]^2 + \frac{1}{n}\mathbb{V}_{p(x)}\left[\hat{q}(x,\pi)\right]. \end{aligned} \tag{20}$$

定理 4 から，DM 推定量の平均二乗誤差が，

1. ログデータの大きさ n
2. 期待報酬関数の推定モデルの予測誤差 $\Delta_{q,\hat{q}}(x,a)$
3. 異なるユーザ間の期待報酬関数の推定モデルのばらつき度合い $\mathbb{V}_{p(x)}\left[\hat{q}(x,\pi)\right]$

によって決定されることがわかる．ログデータのサイズ n が大きかったり期待報酬関数の予測値 $\hat{q}(x,a)$ のばらつきが小さければ，バリアンス項は小さくなる．しかし，これらに関係なく推定モデルの予測誤差が大きい場合，バイアス項が大きいままであるため，DM 推定量の平均二乗誤差は大きくなる．一般にすべての行動 $a \in \mathcal{A}$ について精度良く近似することは難しいため，DM 推定量はバイアスが生じやすい欠点を備えた推定量であるとされている．

2.3 Inverse Propensity Score（IPS）推定量

Inverse Propensity Score（IPS）推定量は，DM 推定量とは異なる考えに基づいて設計されたオフ方策評価の手法である．具体的には IPS 推定量は次のように定義される．

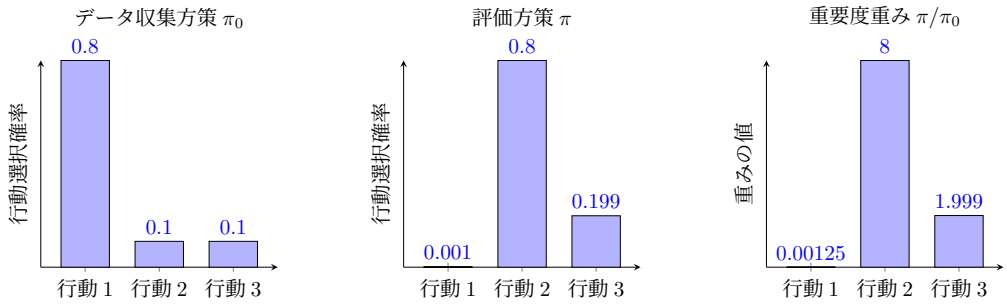

Figure 1: 重要度重み $w(x,a) = \pi(a|x)/\pi_0(a|x)$ が計算される様子.

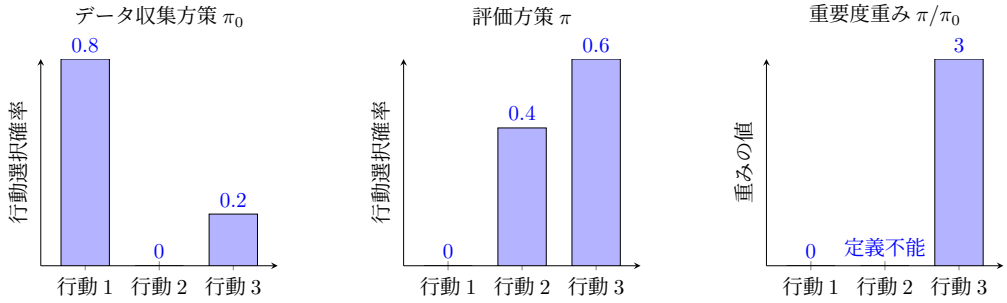

Figure 2: Assumption 1（共通サポート）が満たされない例.

Definition 5. データ収集方策 π_0 が収集したログデータ $\mathcal{D}$ が与えられたとき，評価方策 π の性能 $V(\pi)$ に対する IPS 推定量は次のように定義される.

$$\begin{aligned}\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D}) &:= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i \\ &= \frac{1}{n}\sum_{i=1}^{n} w(x_i,a_i)r_i. \end{aligned} \tag{21}$$

ただし，評価方策 π とデータ収集方策 π_0 による行動選択確率の比であり，重み $w(x,a)$ は次のように定義される.

$$w(x,a) := \frac{\pi(a|x)}{\pi_0(a|x)}. \tag{22}$$

この重み w は，重要度重み（importance weight）と呼ばれる.

推定モデル $\hat{q}$ を構築するような DM 推定量と比べると，単なる重み付け平均で定義されていることから，容易に実装可能な推定量である．重み付け平均として用いられるのは，データ収集方策 π_0 と行動選択確率 $\pi_0(a_i|x_i)$ の比である.

ここで，Figure 1 に重要度重みが計算される様子を示す．これをみると，データ収集方策 π_0 と評価方策 π が似ている場合には，重要度重みが 1 に近い値を取る．一方で，データ収集方策 π_0 と評価方策 π が大きく異なる場合には，重要度重みのばらつきが大きくなることが読み取れる．直感的には，この重要度重みを活用し，データ収集方策 π_0 によるログデータ $\mathcal{D}$ から，評価方策 π に関する情報を得ようとするものだと考えればよい.

では，IPS 推定量もバイアスやバリアンスを調べて，良い推定量か確かめていくが，事前準備として共通サポートと呼ばれる仮定を導入する.

Assumption 1. 任意の $x \in \mathcal{X}, a \in \mathcal{A}$ に対して，

$$\pi(a|x) > 0 \Longrightarrow \pi_0(a|x) > 0 \tag{23}$$

を満たすとき，データ収集方策 π_0 は，評価方策 π に対して共通サポートを持つという.

共通サポートは，評価方策 π が正の確率で選択する行動 a に対して，データ収集方策 π_0 も同様の行動 a を選択する確率が 0 でないことを仮定している．共通サポートが成り立つ例は，先に示した Figure 1 となっており，共通サポートの破れが発生する例は Figure 2 に示した.

この仮定のもとで，IPS 推定量のバイアスを計算すると次のようになる.

Theorem 5. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いるとき，共通サポートの仮定のもとで，IPS 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する不偏推定量である．すなわち，

$$\begin{aligned}\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] = V(\pi). \\ \left(\Longrightarrow \mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] = 0\right)\end{aligned} \tag{24}$$

Proof.

$$
\begin{aligned}
&\text{LHS of (24)}\\
&= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]\\
&= \mathbb{E}_{p(x_i)\pi_0(a_i|x_i)p(r_i|x_i,a_i)}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i\right]\\
&= \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{\pi(a|x)}{\pi_0(a|x)}r\right]\\
&= \mathbb{E}_{p(x)p(r|x,a)}\left[\sum_{a\in\mathcal{A}}\cancel{\pi_0(a|x)}\frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}r\right]\\
&= \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}\left[r\right]\\
&= V(\pi)\\
&= \text{RHS of (24)} \qquad (25)
\end{aligned}
$$

□

よって，オンライン実験ができない場合に，バイアスの意味においてはIPS推定量がより優れた推定量であることが言える．一方で，IPS推定量のバリアンスの計算をすると，

$$
\begin{aligned}
&\mathbb{V}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]\\
&= \mathbb{V}_{p(x_i)\pi_0(a_i|x_i)p(r_i|x_i,a_i)}\left[w(x_i,a_i)r_i\right]\\
&= \frac{1}{n}\mathbb{V}_{p(x)\pi_0(a|x)p(r|x,a)}\left[w(x,a)r\right]\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[\mathbb{V}_{p(r|x,a)}\left[w(x,a)r\right]\right]\\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\mathbb{E}_{p(r|x,a)}\left[w(x,a)r\right]\right]) \quad \because (11)\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]\\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\mathbb{E}_{p(r|x,a)}\left[w(x,a)r\right]\right])\\
&\qquad \because \sigma^2(x,a) \coloneqq \mathbb{V}_{p(r|x,a)}\left[r\right]\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]\\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[w(x,a)q(x,a)\right])\\
&\qquad \because q(x,a) \coloneqq \mathbb{E}_{p(r|x,a)}\left[r\right]\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]\\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right]\\
&\qquad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right]) \quad \because (11)\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]\\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]) \qquad (26)
\end{aligned}
$$

$$
\because q(x,\pi) \coloneqq \mathbb{E}_{\pi(a|x)}\left[q(x,a)\right] = \mathbb{E}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right].
$$

となる．このことから，IPS推定量の平均二乗誤差に関する次の定理を導ける．

Theorem 6. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いるとき，共通サポートの仮定のもとで，IPS推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次の平均二乗誤差を持つ．

$$
\begin{aligned}
&\mathrm{MSE}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]\\
&= \mathrm{Var}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] \quad \because \mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] = 0\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]\\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]). \qquad (27)
\end{aligned}
$$

Equation (27) から，IPS推定量の平均二乗誤差が，オンライン実験と同様に，n が大きくなるにつれて減少していく一方で，報酬ノイズ $\sigma^2(x,a)$ や期待報酬関数 $q(x,\pi)$ のばらつきが大きいとき，推定精度が悪化することがわかる．また，AVG推定量やDM推定量と比べると，Equation (27) には重要度重み $w(x,a)$ が関わる分散の項 $\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]$ が出現しているという相違点がある．これは，IPS推定量はデータ収集方策と大きく異なった挙動を持つ方策を評価する際に，バイアスが生じやすいという問題が起こる原因となる．まとめると，DM推定量がバイアスに関する欠点があったのに対し，IPS推定量はバリアンスに関する欠点がある．

さらに，意思決定の性能とその推定量に関してヘフディングの不等式から次の定理を導ける．

Theorem 7. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$（報酬は，$r\in[0,1]$）を用いるとき，共通サポートの仮定のもとで，IPS推定量に関する次の不等式が，$1-\delta$ 以上の確率で成り立つ．

$$
|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D}) - V(\pi)| \leq w_{\max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}. \qquad (28)
$$

ただし，$w_{\max}$ は，重要度重み $w(x,a) \coloneqq \max_{x,a}(w(x,a))$ である．

Proof. i.i.d. 確率変数 $X_1, X_2, \cdots, X_n$ に対する標本平均を $\bar{X}_n$ とする．このとき，X_i が区間 $[a_i,b_i]$ に制限された場合で成立する以下の不等式を，ヘフディングの不等式と呼ぶ．

$$
\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}[\bar{X}_n]\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right). \qquad (29)
$$

ここで，$^\forall i\in[n], r_i\in[0,1]$ とし，ヘフディングの不等式を使えば，

$$
\begin{aligned}
&\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D}) - \mathbb{E}_{p(\mathcal{D})}[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})]\right| \geq \varepsilon\right]\\
&= \mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D}) - V(\pi)\right| \geq \varepsilon\right] \quad \because (24)\\
&\leq 2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(w_{\max}\cdot 1 - 0)^2}\right)\\
&\leq 2\exp\left(-\frac{2n\varepsilon^2}{w_{\max}^2}\right) =: \delta.
\end{aligned}
$$

と評価できる．このとき，ε について

$$
\varepsilon = w_{\max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \qquad (30)
$$

のように整理できるので，

$$\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)\right|\geq w_{\max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\right]\leq\delta$$
$$\Longrightarrow\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)\right|\leq w_{\max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\right]$$
$$\geq 1-\delta. \tag{31}$$

が成立する．□

また，ヘフディングの不等式とは別に，ベルシュタインの不等式と呼ばれる集中不等式を使うことで次の定理も導ける．

Theorem 8. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$（報酬は $r\in[0,1]$）を用いるとき，共通サポートの仮定のもとで，IPS 推定量に関する次の不等式が，$1-\delta$ 以上の確率で成り立つ．

$$|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)|\leq\frac{2w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}. \tag{32}$$

ただし，$w(x,a):=\max_{x,a}(w(x,a))$，$\sigma^2:=\mathrm{Var}[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})]$ と定義した．

Proof. i.i.d. 確率変数 $X_1,X_2,\cdots,X_n$ に対する標本平均を $\bar{X}_n$ とする．X_i を区間 $[a_i,b_i]$ に制限された場合で成立する以下の不等式を，ベルシュタインの不等式と呼ぶ．

$$\mathbb{P}\left[\left|\bar{X}_n-\mathbb{E}[\bar{X}_n]\right|\geq\varepsilon\right]$$
$$\leq 2\exp\left(-\frac{n\varepsilon^2}{2\mathrm{Var}[\bar{X}_n]+\frac{2}{3}\varepsilon\max_{i\in[n]}(b_i-a_i)}\right). \tag{33}$$

ここで，先に導入したベルシュタインの不等式を使うことにより，以下の不等式評価ができる．$\forall i\in[n],r_i\in[0,1]$ とすることで以下の不等式が成立する．

$$\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-\mathbb{E}_{p(\mathcal{D})}[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})]\right|\geq\varepsilon\right]$$
$$=\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)\right|\geq\varepsilon\right]$$
$$\leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2+\frac{2}{3}\varepsilon w_{\max}}\right)=:\delta.$$

このとき，ε について整理すると，ε に関する二次方程式を構成できる．

$$n\varepsilon^2-\frac{2}{3}w\log\frac{2}{\delta}\varepsilon-2\sigma^2\log\frac{2}{\delta}=0. \tag{34}$$

解の公式を使ってこれを解くと，

$$\varepsilon=\frac{w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\left(-\frac{w_{\max}}{3n}\log\frac{2}{\delta}\right)^2+\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\quad\because\varepsilon>0$$
$$<\frac{w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\left(-\frac{w_{\max}}{3n}\log\frac{2}{\delta}\right)^2}$$
$$+\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\quad\because\sqrt{a+b}<\sqrt{a}+\sqrt{b}$$
$$=\frac{2w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}$$

とできる．よって，

$$\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)\right|\geq\frac{2w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\right]\leq\delta$$
$$\Longrightarrow\mathbb{P}\left[\left|\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})-V(\pi)\right|\leq\frac{2w_{\max}}{3n}\log\frac{2}{\delta}+\sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\right]$$
$$\geq 1-\delta.$$

が成立する．□

上の定理から，IPS 推定量のバイアスがないことが示されたが，バリアンスが大きいとき，推定量の精度が悪化することがわかる．

■ 共通サポートの仮定が成り立たない場合

これまでの分析は共通サポート仮定が成り立っている上での議論であったが，共通サポート仮定が成り立たない場合には，IPS 推定量のバイアスが生じることが知られている．共通サポートが成り立たないケースにおける分析を行うため，事前準備として次のような集合を導入する．

$$\mathcal{U}(x,\pi,\pi_0):=\{a\in\mathcal{A}\mid\pi_0(a|x)=0,\pi(a|x)>0\}. \tag{35}$$

$\mathcal{U}(x,\pi,\pi_0)$ は，評価方策 π のもとでは選択される可能性がある（$\pi(a|x)>0$）行動がある一方で，データ収集方策 π_0 のもとでは選択されない（$\pi_0(a|x)=0$）行動の集合である．仮に共通サポートの仮定が成り立っている場合には，$\mathcal{U}(x,\pi,\pi_0)=\emptyset$ となる．逆に，共通サポートの仮定が成り立っていない場合には，$^{\exists}x\in\mathcal{X},\mathcal{U}(x,\pi,\pi_0)\neq\emptyset$ となる．

以上より，共通サポートの仮定が成り立たない場合における IPS 推定量のバイアスに関する次の定理が成立する．

Theorem 9. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いると，IPS 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次のバイアスを持つ．

$$\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]=-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)}\pi(a|x)q(x,a)\right]. \tag{36}$$

なお，共通サポートが成り立つ場合には，任意の $x\in\mathcal{X}$

に対して $\mathcal{U}(x,\pi,\pi_0)=\emptyset$ となり，$\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]=0$ を導くため，Theorem 5 に整合する．

Proof.

$$
\begin{aligned}
&\text{LHS of (36)}\\
&=\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]\\
&=\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right]-V(\pi)\\
&=\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)^c}\cancel{\pi_0(a|x)}\frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}q(x,a)\right]\\
&\quad-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi(a|x)q(x,a)\right]\\
&=-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi(a|x)q(x,a)-\sum_{a\in\mathcal{U}(x,\pi,\pi_0)^c}\pi(a|x)q(x,a)\right]\\
&=-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)}\pi(a|x)q(x,a)\right]\\
&\quad\because\mathcal{A}\backslash\mathcal{U}(x,\pi,\pi_0)^c=\mathcal{U}(x,\pi,\pi_0)\\
&=\text{RHS of (36)}.
\end{aligned}
\tag{37}
$$

□

これより，共通サポートが成り立たない場合に IPS 推定量を用いてしまうと，データ収集方策 π_0 のもとで選択されない行動に関して情報が得られず評価が出来ないため，それらの行動の期待報酬の分だけ過小評価されてしまう．

■ データ収集方策 π_0 が未知の場合

これまでの IPS 推定量はデータ収集方策 π_0 が既知であることを案に仮定してきたが，実際の問題ではデータ収集方策 π_0 に関して完全な知識をもっていない場合がある（意思決定者とデータ分析者は異なることも多い）．このような状況において，データ収集方策 π_0 が未知の場合に IPS 推定量を適用することができるかを考える．よくある解決策として，IPS 推定量を用いてデータ収集方策 π_0 を推定し，その推定されたデータ収集方策 $\hat{\pi}_0$ を用いて IPS 推定量を計算する方法がある．このとき，IPS 推定量の定義は次のようになる．

Definition 6. データ収集方策 π_0 が未知である場合，評価方策 π の性能 $V(\pi)$ に対する IPS 推定量は次のように定義される．

$$
\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{\pi}_0)\coloneqq\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)}r_i. \tag{38}
$$

なおデータ収集方策の推定モデル $\hat{\pi}_0$ は，次のような教師あり分類問題などを解くことで求まる．

$$
\hat{\pi}_0(a|x)=\underset{\pi'(a|x)\in\Pi}{\arg\max}\frac{1}{n}\sum_{i=1}^{n}\ell_a\left(a_i,\pi'(a_i|x_i)\right). \tag{39}
$$

なお，ℓ_a は行動 a に対する損失関数であり，Π はデータ収集方策の仮説集合である．

ここで，データ収集方策 π_0 が未知の場合に IPS 推定量のバイアスについて以下の定理が成立する．

Theorem 10. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いると，データ収集方策 π_0 を推定モデル $\hat{\pi}_0$ で代替した場合の IPS 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次のバイアスを持つ．

$$
\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{\pi}_0)\right]=\mathbb{E}_{p(x)\pi(a|x)}\left[\delta(x,a)q(x,a)\right]. \tag{40}
$$

ただし，$\delta(x,a)$ はデータ収集方策の推定誤差であり，次のように定義する．

$$
\delta(x,a)\coloneqq\frac{\pi(a|x)}{\hat{\pi}_0(a|x)}-1. \tag{41}
$$

もし，$\hat{\pi}(a|x)=\pi(a|x)$ ならば，すべての $x\in\mathcal{X},a\in\mathcal{A}$ に対して $\delta(x,a)=0$ となるため，$\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{\pi}_0)\right]=0$ となり，これは Theorem 5 に整合する．

Proof.

$$
\begin{aligned}
&\text{LHS of (40)}\\
&=\mathrm{Bias}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{\pi}_0)\right]\\
&=\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{\pi}_0)\right]-V(\pi)\\
&=\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\frac{\pi_0(a|x)}{\hat{\pi}_0(a|x)}\pi(a|x)q(x,a)\right]\\
&\quad-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi(a|x)q(x,a)\right]\\
&=\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\left(\frac{\pi(a|x)}{\hat{\pi}_0(a|x)}-1\right)\pi(a|x)q(x,a)\right]\\
&=\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\delta(x,a)\pi(a|x)q(x,a)\right]\\
&=\mathbb{E}_{p(x)\pi(a|x)}\left[\delta(x,a)q(x,a)\right]\\
&=\text{RHS of (40)}.
\end{aligned}
\tag{42}
$$

□

この事実から，データ収集方策 π_0 の代替である推定モデル $\hat{\pi}_0$ による IPS 推定量を用いると，データ収集方策の乖離度 $\delta(x,a)$ に依存したバイアスが生じることがわかる．真のデータ収集方策 π_0 が既知であれば，IPS 推定量はバイアスがないことを示したが，データ収集方策 π_0 を推定した結果，その推定誤差 $\delta(x,a)$ によるバイアスが生じることは直感的である．

2.4 Clipped Inverse Propensity Score（CIPS）推定量

Clipped Inverse Propensity Score（CIPS）推定量は，IPS 推定量のバリアンスの問題を解決するために提案された手法であり，次のように定義される．

Definition 7. データ収集方策 π_0 が収集したログデータ $\mathcal{D}$ と評価方策 π に対して，Clipped Inverse Propensity Score（CIPS）推定量は次のように定義される．

$$\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \min\left\{w(x_i,a_i),\lambda\right\} r_i. \tag{43}$$

なお，$\lambda \geq 0$ は，バイアスとバリアンストレードオフを調整するハイパーパラメータ，また，$w(x_i,a_i) \coloneqq \pi(a_i|x_i)/\pi_0(a_i|x_i)$ は，重要度重みである．ちなみに，$\lambda = \infty$ としたとき，IPS 推定量に一致するため，CIPS 推定量は IPS 推定量の一般化になっている．

CIPS 推定量は，IPS 推定量の重要度重み $w(x,a)$ をクリッピングすることで，重要度重みの大きな値を抑制することでバリアンスの問題を解決しようとする手法である．λ を十分大きな値に設定したときは，IPS 推定量と大きな違いは生まれないため，バイアスの大きさは抑えられるものの，同時にバリアンスは大きくなってしまう．逆に，λ を小さく設定したときは，バリアンスの減少効果は見込めるものの，バイアスが発生しかねない．

では，この事実を確かめるために，まず CIPS 推定量のバイアスについて分析する．

Theorem 11. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いると，CIPS 推定量は，共通サポートのもとで，評価方策 π の真の性能 $V(\pi)$ に対する次のバイアスを持つ．

$$\begin{aligned}&\mathrm{Bias}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right]\\ &= \mathbb{E}_{p(x)\pi(a|x)}\left[\left(\frac{\lambda}{w(x,a)}-1\right)\mathbb{1}\left\{w(x,a)>\lambda\right\} q(x,a)\right].\end{aligned} \tag{44}$$

ただし，$\mathbb{1}\{\cdot\}$ は指示関数である．

Proof.

$$\begin{aligned}
&\text{LHS of (44)}\\
&= \mathrm{Bias}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right]\\
&= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right] - V(\pi)\\
&= \mathbb{E}_{p(x)\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right] - \mathbb{E}_{p(x)\pi(a|x)}\left[q(x,a)\right]\\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi_0(a|x)\min\left\{w(x,a),\lambda\right\} q(x,a) - \pi(a|x)q(x,a)\right]\\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi(a|x)\left(\frac{1}{w(x,a)}\min\left\{w(x,a),\lambda\right\}-1\right) q(x,a)\right]\\
&= \mathbb{E}_{p(x)\pi(a|x)}\left[\left(\min\left\{1,\frac{\lambda}{w(x,a)}\right\}-1\right) q(x,a)\right]\\
&= \mathbb{E}_{p(x)\pi(a|x)}\left[\left(\frac{\lambda}{w(x,a)}-1\right)\mathbb{1}\left\{w(x,a)>\lambda\right\} q(x,a)\right].\\
&= \text{RHS of (44)}
\end{aligned} \tag{45}$$

□

次に，バリアンスについても分析を行う．

Theorem 12. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いると，CIPS 推定量は，共通サポートのもとで，評価方策 π の真の性能 $V(\pi)$ に対する次のバリアンスを持つ．

$$\begin{aligned}
&\mathrm{Var}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right]\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\}^2\sigma^2(x,a)\right]\\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right]\right]\\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right]\right]).
\end{aligned} \tag{46}$$

Proof.

$$\begin{aligned}
&\text{LHS of (46)}\\
&= \mathrm{Var}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right]\\
&= \mathbb{V}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{CIPS}}(\pi;\mathcal{D},\lambda)\right]\\
&= \frac{1}{n}\mathbb{V}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\min\left\{w(x,a),\lambda\right\} r\right]\\
&= \frac{1}{n}\mathbb{E}_{p(x)\pi_0(a|x)}\left[\mathbb{V}_{p(r|x,a)}\left[\min\left\{w(x,a),\lambda\right\} r\right]\right.\\
&\quad \left. + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\mathbb{E}_{p(r|x,a)}\left[\min\left\{w(x,a),\lambda\right\} r\right]\right]\right) \quad \because (11)\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\}^2\sigma^2(x,a)\right]\\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right])\\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\}^2\sigma^2(x,a)\right]\\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right]\right]\\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_0(a|x)}\left[\min\left\{w(x,a),\lambda\right\} q(x,a)\right]\right]) \quad \because (11)\\
&= \text{RHS of (46)}.
\end{aligned} \tag{47}$$

□

以上の議論を踏まえると，

1. λ を大きな値に設定しておけば，$\mathbb{1}\{w(x,a) > \lambda\} = 0$ が成り立ちやすくなり，バイアスがを小さくできるが，バリアンスが大きくなる．
2. λ を小さな値に設定すると，$\min\{w(x,a), \lambda\} = \lambda \ll w(x,a)$ が成り立ちやすくなり，バイアスが大きくなるが，バリアンスを小さくできる．

ということが言える．また，ログデータのサイズ n に応じて適切な λ の値は変化するので，問題設定ごとに適切な λ を選択することが理想的である．

2.5 Doubly Robust（DR）推定量

ここまでを振り返ると，DM 推定量はバイアスが大きいがバリアンスが小さく，IPS 推定量はバイアスがないがバリアンスが大きいという特徴を持っていた．そこで考案されたのが，これらの推定量を利点をうまく組み合わせた Doubly Robust（DR）推定量である．DR 推定量は次のように定義される．

Definition 8. データ収集方策 π_0 が収集したログデータ $\mathcal{D}$ が与えられたとき，評価方策 π の性能 $V(\pi)$ に対する DR 推定量は次のように定義される．

$$\begin{aligned}
&\hat{V}_{\mathrm{DR}}(\pi; \mathcal{D}, \hat{q}) \\
&:= \frac{1}{n}\sum_{i=1}^{n}\{\hat{q}(x_i, \pi) + w(x_i, a_i)(r_i - \hat{q}(x_i, a_i))\} \\
&= \hat{V}_{\mathrm{DM}}(\pi; \mathcal{D}, \hat{q}) + \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)(r_i - \hat{q}(x_i, a_i)). \quad (48)
\end{aligned}$$

ただし，$\hat{q}(x,a)$ は期待報酬関数の推定モデルである．また，重要度重み

$$w(x,a) := \pi(a|x)/\pi_0(a|x) \quad (49)$$

は，DM 推定量のときに用いたものと同様である．

DR 推定量では，第一項において，DM 推定量の期待報酬関数の推定値 $\hat{q}(x_i, \pi)$ を用いつつも，第二項において，IPS 推定量の重要度重み $w(x_i, a_i)$ を用いて，報酬 r_i と期待報酬関数の推定値 $\hat{q}(x_i, a_i)$ の差を補正している．仮に推定モデル $\hat{q}(x,a)$ がある程度の推定精度をもっていれば，重要度重み $w(x,a)$ による補正項がゼロに近い値を持つ．よって，

$$\hat{V}_{\mathrm{DR}}(\pi; \mathcal{D}, \hat{q}) \approx \hat{V}_{\mathrm{DM}}(\pi; \mathcal{D}, \hat{q})$$

となるため，$\hat{q}(x_i, \pi)$ の性能が高い場合における DM 推定量のバイアスは小さくなる．その上，重要度重みの大きさが不安定になることに起因した IPS 推定量のバリアンスの問題を軽減することができる．

ではまず，DR 推定量のバイアスについて分析する．

Theorem 13. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，DR 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する不偏推定量である．すなわち，

$$\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DR}}(\pi; \mathcal{D}, \hat{q})\right] = V(\pi). \quad (50)$$
$$\left(\Longrightarrow \mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi; \mathcal{D}, \hat{q})\right] = 0\right)$$

Proof.

$$\begin{aligned}
&\text{LHS of (50)} \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DR}}(\pi; \mathcal{D}, \hat{q})\right] \\
&= \mathbb{E}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\hat{q}(x,\pi) + w(x,a)(r - \hat{q}(x,a))\right] \\
&= \mathbb{E}_{p(x)p(r|x,a)}\left[\hat{q}(x,\pi) + \sum_{a\in\mathcal{A}} \cancel{\pi_0(a|x)} \frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}(r - \hat{q}(x,a))\right] \\
&= \mathbb{E}_{p(x)p(r|x,a)}\left[\cancel{\hat{q}(x,\pi)} - \cancel{\hat{q}(x,\pi)} + \sum_{a\in\mathcal{A}} \pi(a|x) r\right] \\
&= \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}\left[r\right] \\
&= V(\pi) \\
&= \text{RHS of (50)}. \quad (51)
\end{aligned}$$

□

Theorem 13 から，DR 推定量は，DM 推定量の原因となっていた期待報酬関数の推定モデル $\hat{q}(x,a)$ を用いているのにも，バイアスが生じないことがわかる．

また，いままでの推定量と同様にして全分散の公式 (11) を

使い，DR 推定量のバリアンスについても計算すると，

$$
\begin{aligned}
&\mathbb{V}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] \\
&= \frac{1}{n}\mathbb{V}_{p(x)\pi_0(a|x)p(r|x,a)}\left[\hat{q}(x,\pi)+w(x,a)(r-\hat{q}(x,a))\right] \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[\mathbb{V}_{p(r|x,a)}\left[\hat{q}(x,\pi)+w(x,a)(r-\hat{q}(x,a))\right]\right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\mathbb{E}_{p(r|x,a)}\left[\hat{q}(x,\pi)+w(x,a)(r-\hat{q}(x,a))\right]\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\hat{q}(x,\pi)+w(x,a)\left\{q(x,a)-\hat{q}(x,a)\right\}\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\hat{q}(x,\pi)-w(x,a)\Delta_{q,\hat{q}}(x,a)\right] \\
&\qquad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\hat{q}(x,\pi)+w(x,a)\left\{q(x,a)-\hat{q}(x,a)\right\}\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{V}_{p(x)\pi_0(a|x)}\left[\hat{q}-w(x,a)\Delta_{q,\hat{q}}(x,a)\right]) \\
&\qquad \because \Delta_{q,\hat{q}}(x,a) \coloneqq \hat{q}(x,a)-q(x,a) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[\hat{q}(x,\pi)-w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] \\
&\qquad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_0(a|x)}\left[\hat{q}(x,\pi)+w(x,a)\left\{q(x,a)-\hat{q}(x,a)\right\}\right]\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] \\
&\qquad + \mathbb{V}_{p(x)}\left[\hat{q}(x,\pi)+\mathbb{E}_{\pi(a|x)}\left[q(x,a)-\hat{q}(x,a)\right]\right]) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]) \qquad (52)
\end{aligned}
$$

となる．これまでの分析結果に基づくと，DR 推定量の平均二乗誤差に関する次の定理が成立する．

Theorem 14. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いるとき，共通サポートの仮定のもとで，DR 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次の平均二乗誤差を持つ．

$$
\begin{aligned}
&\mathrm{MSE}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] \\
&= \mathrm{Var}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] \quad \because \mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right]=0 \\
&= \frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right] \\
&+ \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] + \mathbb{V}_{p(x)}\left[q(x,\pi)\right]). \qquad (53)
\end{aligned}
$$

DR 推定量の平均二乗誤差を見ると，ログデータのサイズ n が大きなるほど推定精度が向上する一方，報酬ノイズ $\sigma^2(x,a)$ や期待報酬関数の推定モデル $\hat{q}(x,a)$ の予測誤差 $\Delta_{q,\hat{q}}(x,a)$ が大きいとき，推定精度が悪化することがわかる．DR 推定量と IPS 推定量の差を知るために，平均二乗誤差の差分を計算すると，

$$
\begin{aligned}
&\mathrm{MSE}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{q})\right] - \mathrm{MSE}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D})\right] \\
&= \mathrm{Var}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D},\hat{q})\right] - \mathrm{Var}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D})\right] \\
&= \frac{1}{n}(\cancel{\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]} \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] + \cancel{\mathbb{V}_{p(x)}\left[q(x,\pi)\right]}) \\
&- \frac{1}{n}(\cancel{\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\sigma^2(x,a)\right]} \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] + \cancel{\mathbb{V}_{p(x)}\left[q(x,\pi)\right]}) \\
&= \frac{1}{n}(\mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\Delta_{q,\hat{q}}(x,a)\right]\right] \qquad (54)
\end{aligned}
$$

となる．この値が大きいほど，DR 推定量の平均二乗誤差が IPS 推定量のそれよりも正確であると言える．式を観察すると，IPS 推定量の平均二乗誤差には，報酬期待関数 $q(x,a)$ 自体のバリアンスが出現している一方で，DR 推定量の平均二乗誤差には期待報酬関数に対する推定モデルの誤差 $\Delta_{q,\hat{q}}(x,a)$ が現れている．すなわち，期待報酬関数に対する推定モデルの誤差が期待報酬関数それ自体よりも小さくなる程度の精度 ($|\Delta_{q,\hat{q}}(x,a)| \leq q(x,a), {}^{\forall}(x,a)$) を有する推定モデル $\hat{q}(x,a)$ さえ得ることができれば，DR 推定量は IPS 推定量よりも推定精度が高くなる．このことから，DR 推定量は DM 推定量と IPS 推定量の長所をうまく組み合わせた，良い平均二乗誤差を達成できる良い推定量となっている．

■ 共通サポートの仮定が成り立たない場合

IPS 推定量と同様に，共通サポートの仮定が成り立たない場合における DR 推定量のバイアスについて分析を行う．

Theorem 15. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，DR 推定量は，評価方策 π の真の性能 $V(\pi)$ に対して次のバイアスを持つ．

$$
\begin{aligned}
&\mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)}\pi(a|x)\Delta_{q,\hat{q}}(x,a)\right]. \qquad (55)
\end{aligned}
$$

なお，仮に共通サポートの仮定が満たされている場合は，すべての $x\in\mathcal{X}$ に対して $\mathcal{U}(x,\pi,\pi_0)=\emptyset$ になることから，バイアスがゼロとなるため，結果が整合する．

Proof.

$$
\begin{aligned}
&\text{LHS of (55)} \\
&= \mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q})\right] - V(\pi) \\
&= \mathbb{E}_{p(x)\pi_0(a|x)}\left[\hat{q}(x,\pi) + w(x,a)(q(x,a)-\hat{q}(x,a))\right] \\
&\quad - \mathbb{E}_{p(x)}\left[q(x,\pi)\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \Delta_{q,\hat{q}}(x,a)\right] \\
&\quad - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)^c} \cancel{\pi_0(a|x)}\,\frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}\Delta_{q,\hat{q}}(x,a)\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{U}(x,\pi,\pi_0)} \pi(a|x)\Delta_{q,\hat{q}}(x,a)\right] \\
&\quad \because \mathcal{A}\backslash\mathcal{U}(x,\pi,\pi_0)^c = \mathcal{U}(x,\pi,\pi_0) \\
&= \text{RHS of (55)}. \qquad (56)
\end{aligned}
$$

□

■ データ収集方策 π_0 が未知の場合

真のデータ分布 π_0 に対しての事前知識がなく，推定モデル $\hat{\pi}_0$ で代替した場合の DR 推定量は，

$$
\begin{aligned}
&\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q},\hat{\pi}_0) \\
&\coloneqq \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{q}(x_i,\pi) + \frac{\pi(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)}(r_i - \hat{q}(x_i,a_i))\right\}. \qquad (57)
\end{aligned}
$$

この場合，DR 推定量で発生するバイアスは次のようになる．

Theorem 16. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，データ収集方策 π_0 が未知である場合において，推定モデル $\hat{\pi}_0$ で代替した場合の *DR* 推定量は，評価方策 π の真の性能 $V(\pi)$ に対して次のバイアスを持つ．

$$
\begin{aligned}
&\mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q},\hat{\pi}_0)\right] \\
&= -\mathbb{E}_{p(x)\pi(a|x)}\left[\delta(x,a)\Delta_{q,\hat{q}}(x,a)\right]. \qquad (58)
\end{aligned}
$$

なお，$\delta(x,a) \coloneqq \pi(a|x)/\hat{\pi}_0(a|x) - 1$ は，データ収集方策 π_0 の推定誤差である．

Proof.

$$
\begin{aligned}
&\text{LHS of (58)} \\
&= \mathrm{Bias}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q},\hat{\pi}_0)\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\mathrm{DR}}(\pi;\mathcal{D},\hat{q},\hat{\pi}_0)\right] - V(\pi) \\
&= \mathbb{E}_{p(x)\pi_0(a|x)}\left[\hat{q}(x,\pi) + \frac{\pi(a|x)}{\hat{\pi}_0(a|x)}(q(x,a)-\hat{q}(x,a))\right] \\
&\quad - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi(a|x)q(x,a)\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi(a|x)\left\{\Delta_{q,\hat{q}}(x,a) - \frac{\pi_0(a|x)}{\hat{\pi}_0(a|x)}\Delta_{q,\hat{q}}(x,a)\right\}\right] \\
&= -\mathbb{E}_{p(x)\pi(a|x)}\left[\left\{\frac{\pi_0(a|x)}{\hat{\pi}_0(a|x)} - 1\right\}\Delta_{q,\hat{q}}(x,a)\right] \\
&= -\mathbb{E}_{p(x)\pi(a|x)}\left[\delta(x,a)\Delta_{q,\hat{q}}(x,a)\right] \\
&= \text{RHS of (58)}. \qquad (59)
\end{aligned}
$$

□

期待報酬関数かデータ収集方策のいずれかが，正しく推定できていれば，bias がゼロになるという興味深い性質がある．

2.6 Switch Doubly Robust（Switch-DR）推定量

Switch Doubly Robust（Switch-DR）推定量は，DR 推定量と DM 推定量を重要度重みの大きさに応じて使い分けることで．DR 推定量バイアスの問題を解決するために提案された手法であり，次のように定義される．

Definition 9. データ収集方策 π_0 が収集したログデータ $\mathcal{D}$ が与えられたとき，評価方策 π の性能 $V(\pi)$ に対する Switch-DR 推定量は次のように定義される．

$$
\begin{aligned}
&\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q}) \\
&\coloneqq \frac{1}{n}\sum_{i=1}^{n}\hat{q}(x_i,\pi) + w(x_i,a_i)\mathbb{1}\left\{w(x_i,a_i)\le\lambda\right\}(r_i - \hat{q}(x_i,a_i)) \\
&= \hat{V}_{\mathrm{DM}}(\pi;\mathcal{D},\hat{q}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} w(x_i,a_i)\mathbb{1}\left\{w(x_i,a_i)\le\lambda\right\}(r_i - \hat{q}(x_i,a_i)).
\end{aligned} \qquad (60)
$$

なお，$\lambda \ge 0$ はバイアス・バリアンストレードオフを調整するハイパーパラメータである．

Switch-DR 推定量は，重要度重みが小さくなる（$w(x,a) \le \lambda$）ようなデータ i については，バリアンスの懸念が小さいため DR 推定量を採用する（$\lambda = \infty$ のとき，DR 推定量に一致）一方で，重要度重みが大きくなる（$w(x,a) > \lambda$）ようなデータ i については，バリアンスの問題を避けるため DM 推定量を採用する（$\lambda = 0$ のとき，DM 推定量に一致）というアイデアに基づいて設計されている．では，具体的に Switch-DR 推定量のバイアスとバリアンスについて分析する．

Theorem 17. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ と真の期待報酬関数 $q(x,a)$ の推定モデル $\hat{q}(x,a)$ を用いたとき，共通サポートのもとで，Switch-DR 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次のバイアスを持つ．

$$\begin{aligned}&\mathrm{Bias}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]\\&=\mathbb{E}_{p(x)\pi_0(a|x)}\left[\mathbb{1}\left\{w(x,a)>\lambda\right\}\Delta_{q,\hat{q}}(x,a)\right].\end{aligned}\tag{61}$$

Proof.

$$\begin{aligned}&\text{LHS of (61)}\\&=\mathrm{Bias}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]\\&=\mathbb{E}_{p(\mathcal{D})}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]-V(\pi)\\&=\mathbb{E}_{p(x)\pi(a|x)}\left[\hat{q}(x,a)\right]\\&\quad-\mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\cancel{\pi_0(a|x)}\frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}\mathbb{1}\left\{w(x,a)\le\lambda\right\}\Delta_{q,\hat{q}}(x,a)\right]\\&\quad-\mathbb{E}_{p(x)\pi(a|x)}\left[q(x,a)\right]\\&=\mathbb{E}_{p(x)\pi(a|x)}\left[\Delta_{q,\hat{q}}(x,a)\right]\\&\quad-\mathbb{E}_{p(x)\pi(a|x)}\left[\mathbb{1}\left\{w(x,a)\le\lambda\right\}\Delta_{q,\hat{q}}(x,a)\right]\\&=\mathbb{E}_{p(x)\pi(a|x)}\left[\left(1-\mathbb{1}\left\{w(x,a)\le\lambda\right\}\right)\Delta_{q,\hat{q}}(x,a)\right]\\&=\mathbb{E}_{p(x)\pi(a|x)}\left[\mathbb{1}\left\{w(x,a)>\lambda\right\}\Delta_{q,\hat{q}}(x,a)\right]\\&=\text{RHS of (61)}.\end{aligned}\tag{62}$$

□

次に，バリアンスについて分析する．

Theorem 18. あるデータ収集方策 π_0 により収集されたログデータ $\mathcal{D}$ を用いるとき，共通サポートのもとで，Switch-DR 推定量は，評価方策 π の真の性能 $V(\pi)$ に対する次のバリアンスを持つ．

$$\begin{aligned}&\mathrm{Var}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]\\&=\frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\mathbb{1}\left\{w(x,a)\le\lambda\right\}\sigma^2(x,a)\right]\\&\quad+\mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[\mathbb{1}\left\{w(x,a)\le\lambda\right\}q(x,a)\right]\right]\\&\quad+\mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi(a|x)}\left[\mathbb{1}\left\{w(x,a)\le\lambda\right\}q(x,a)\right]\right]).\end{aligned}\tag{63}$$

Proof.

$$\begin{aligned}&\text{LHS of (63)}\\&=\mathrm{Var}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]\\&=\mathbb{V}_{p(\mathcal{D})}\left[\hat{V}_{\text{Switch-DR}}(\pi;\mathcal{D},\hat{q})\right]\\&=\frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\mathbb{1}\left\{w(x,a)\le\lambda\right\}\sigma^2(x,a)\right]\\&+\mathbb{V}_{p(x)\pi(a|x)}\left[\hat{q}(x,\pi)-w(x,a)\mathbb{1}\left\{w(x,a)\le\lambda\right\}\Delta_{q,\hat{q}}(x,a)\right])\\&=\frac{1}{n}(\mathbb{E}_{p(x)\pi_0(a|x)}\left[w^2(x,a)\mathbb{1}\left\{w(x,a)\le\lambda\right\}\sigma^2(x,a)\right]\\&\quad+\mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}\left[w(x,a)\mathbb{1}\left\{w(x_i,a_i)\le\lambda\right\}\Delta_{q,\hat{q}(x,a)}\right]\right])\\&\quad+\mathbb{V}_{p(x)}\left[\hat{q}(x,\pi)-\mathbb{E}_{\pi(a|x)}\left[\mathbb{1}\left\{w(x_i,a_i)\le\lambda\right\}\Delta_{q,\hat{q}(x,a)}\right]\right])\\&=\text{RHS of (63)}.\end{aligned}\tag{64}$$

□

以上の議論を踏まえると，

1. λ を大きな値に設定しておけば，$\mathbb{1}\left\{w(x,a)>\lambda\right\}=0$ が成り立ちやすくなり，バイアスがを小さくできるが，バリアンスが大きくなる．
2. λ を小さな値に設定すると，$\min\left\{w(x,a),\lambda\right\}=\lambda\ll w(x,a)$ が成り立ちやすくなり，バイアスが大きくなるが，バリアンスを小さくできる．

ということが言える．また，ログデータのサイズ n に応じて適切な λ の値は変化するので，問題設定ごとに適切な λ を選択することが理想的である．

3 実験

...

References

[1] 齋藤優太. 反実仮想機械学習. 技術評論社, 2024.