データ駆動制度設計に向けて
—アルゴリズム生成データによる反実仮想予測—

奥村　恭平，清水　亮洋，成田　悠輔，矢田　紘平

公共政策からビジネスまで，アルゴリズムを利用した意思決定が広がっている．その際に重要なのが，過去に使用された方策（意思決定アルゴリズム）が蓄積したデータを用いて，過去に使われたことのない新しい方策の性能を予測することだ．「方策外評価」などと呼ばれるこの予測によって，データに基づいて意思決定・資源配分アルゴリズム・メカニズムを設計していくことが可能になる．本稿では，従来の手法では分析することの難しかった，より広いクラスの方策によって生成されたデータに適用可能な方策外評価手法を説明する．そして，提案手法をフリマアプリ・メルカリにおけるクーポン割当方策の評価に適用し，既存の方策を改善する方法を示す．

キーワード：因果効果，機械学習，反実仮想，方策外評価

1. はじめに

アルゴリズムによる意思決定・選択・推薦は今日世の至る所で行われている．動画やニュースなどのコンテンツ配信やソーシャルメディア，EC における広告配信はもちろんのこと，金融，裁判，監視に至るまで，アルゴリズムによる予測や分類を用いた意思決定が爆発的に広がっている．監視を例にとれば，監視カメラが捉えた画像データを画像認識し，そのうえで映っている人物が犯罪やテロに加担する可能性を機械学習アルゴリズムが予測する．そして，危険性が高いと予測された人物を追跡するという意思決定が行われている．

意思決定に用いられるアルゴリズムは，機械学習アルゴリズムに限らない．たとえば，世界各地の学校選択・入試制度や労働市場・臓器移植市場などではマッチングアルゴリズムが用いられている．また，国債市場・卸売市場やオンライン広告・中古品市場などでは，オークションアルゴリズムが活用されている．このようなマッチングやオークションなどの中央集権的な制度もまた，アルゴリズムを用いた意思決定である．アルゴリズムを用いた意思決定の例を表 1 にまとめた．

アルゴリズムによる意思決定を行ううえで重要なのが，まだ使われたことのない新しい意思決定アルゴリズム（**方策**とも呼ばれる）の性能を予測することだ．正確な性能予測があれば，着実にアルゴリズムを改善することができる．すぐに思い浮かぶ性能予測方法は，古いアルゴリズムと新しいアルゴリズムをランダムに人や地域に割り当てて比較するランダム化実験（RCT，A/B テスト）だろう．だが，RCT は工数も費用もかかるうえ，被験者に不公平感を与えて炎上しかねないという倫理的問題を抱えている [12]．

RCT に頼ることなく，過去のアルゴリズムが自然に生み出したデータだけで性能予測する方法はないだろうか？ データに基づくアルゴリズム的意思決定が当たり前になりつつある今，既存の制度が生み出したデータに基づいてより良い制度を逐次的に提案する**データ駆動型制度設計**の手法は今後ますます重要になっていくはずだ．

既存の方策が生成したデータを用いて新方策の性能を推定しようとする営みは，**方策外評価 (off-policy evaluation)** と呼ばれる．既存の方策が**確率的**であるとき，つまり，任意の入力に対し複数の選択肢を確率的に選ぶときについては，さまざまな方策外評価の手法が提案されてきた [13–21]．一方，既存の方策が**非確率的で確定的 (deterministic)** である場合，つまり，ある種の入力に対しては決まった一つの選択肢を確実に選ぶときについては，確立された手法が存在していなかった．

私たちは，非確率的な方策を含む幅広いクラスの既存方策に適用可能な方策外評価の手法を提案する．この方法は以下の観察に基づく．アルゴリズムが意思決定を行った場合，そこから生成されたデータには，意思決定がランダムに，あたかもサイコロを振ったかのように行われる自然実験がほぼ必ず含まれるという観察である．たとえば，多くの確率的な強化学習・バンディットアルゴリズムは選択（探索）をランダムに行うため，ほとんど RCT そのものである．また，教師

おくむら　きょうへい
ノースウェスタン大学
しみず　あきひろ
株式会社メルカリ
なりた　ゆうすけ
半熟仮想株式会社，イェール大学
yusuke.narita@yale.edu
やた　こうへい
イェール大学

表 1　アルゴリズムに基づく意思決定の例

	アルゴリズムが用いる変数 (X)	アルゴリズムの意思決定 (Z)	結果変数 (Y)	アルゴリズム例
ウェブ企業	利用者の閲覧履歴，アクセスの時間・場所	表示コンテンツ	利用者が表示コンテンツにアクセスしたかどうか	バンディットなどの強化学習 [1]
自動車共有サービス	利用者がアプリを開いた時点における周辺地域の需要と供給	サービスの価格	利用者がサービスを利用したかどうか	価格上昇・動的価格決定 [2]
裁判官	被告人の犯罪歴，年齢などの属性	釈放すべきか否か	被告人が再犯したかどうか	教師あり学習 [3]
学校選択制・中央集権入試	家庭の学校への選好，学校での優先権	学校への割当・入学権	将来の成績や収入など	受入保留アルゴリズムなどの割当アルゴリズム [4–9]
オークション	入札者の入札額	入札者が落札したか	入札者の将来の経済パフォーマンス	オークション・アルゴリズム [10, 11]

付き学習で予測された何らかの変数がある基準値を上回るかどうかで選択を決めるアルゴリズムを考えた場合，基準値の近くでは，ほぼ同じ状況であるにもかかわらず，基準値をたまたま上回ったかどうかというほとんど偶然の要因で異なった意思決定が行われる．これも局所的な自然実験とみなせる．

こういった自然実験はさまざまな目的のために使える．意思決定のうちどれが効果的かを測るために使えるし，新たな意思決定アルゴリズムを導入するとどのような性能を発揮しそうかを予測するためにも使える．私たちは，この観察を一般の機械学習アルゴリズムについて定式化し，**アルゴリズムが自然に生成したデータを用いてアルゴリズムを改善する手法を開発する．**

この手法が使える場面は，ビジネスから政策まで幅広い．具体的な応用として，フリマアプリ・メルカリにおけるクーポン配信方策の評価を行い，改善案を提示する．

2.　因果効果の学習

以下では，まず方策外評価の問題を定式化したうえで，アルゴリズムがどのような条件を満たせば自然実験が存在し，まだ見ぬ方策の性能の反実仮想予測が可能になるかを分析していく．なお，本稿では骨組みを素描するに留める．技術的詳細の説明は文献 [17, 22] を参照されたい．

2.1　アルゴリズムによる意思決定の定式化

$\mathcal{A} \coloneqq \{1, 2, \ldots, m\}$ を意思決定者が選べる**行動**の集合とする．$Y(a)$ で行動 a が選択された場合に観測される**潜在結果**を表す．潜在結果 $Y_i(a)$ は個人 i について行動 a が選択された場合に観測される結果で，現実には $Y_i(1), \ldots, Y_i(m)$ のどれか一つだけ観察される．たとえば，i にクーポンを配布した場合，i にクーポンを配布しなかった場合の i の購入額については観測することができない．意思決定者は，行動を選択する前に，**文脈** $X \in \mathcal{X} \subseteq \mathbb{R}^p$ を観測する．文脈は，たとえば個々人の年収・職業・過去の行動履歴などの属性に対応する．本稿では，簡単のため，文脈の各要素は連続値をとると仮定する．

データを生成する既存の方策を**記録方策 (logging policy)** と呼び，$ML\colon \mathbb{R}^p \to \Delta(\mathcal{A})$ で表す．$ML(a \mid x)$ は，文脈 x を観測したとき記録方策が行動 a を選ぶ確率を表す．意思決定者は記録方策を知っており，それを完全にシミュレートできるとする．つまり，各 a, x について $ML(a \mid x)$ の値が既知で計算可能だと仮定する．任意の a, x について $ML\,(a|x) \in (0, 1)$ である場合，記録方策は**確率的**だという．そうでない場合，記録方策は**非確率的**であるという．

記録方策により，ログデータ $(Y_i, X_i, A_i)_{i=1}^n$ が次のように生成される：(1) まず $(Y_i(\cdot), X_i)$ が未知の分布から i.i.d. に引かれる，(2) 次に，行動 A_i が確率分布 $ML(\cdot \mid X_i)$ に従って選ばれる，(3) 最後に，報酬 $Y_i = Y_i(A_i)$ が観測される．私たちの目的は，ログデータを元に，与えられた反実仮想方策 $\pi\colon \mathbb{R}^p \to \Delta(\mathcal{A})$ の性能

$$V(\pi) \coloneqq E\left[\sum_{a \in \mathcal{A}} Y(a)\pi(a|X)\right]$$

を予測・学習することである．

2.2　近似傾向スコア

まず，データが無限に入手可能だという理想的な状況下で，反実仮想方策 π の性能 $V(\pi)$ の値が求められる条件を見つけよう．このような無限データを使った学習をよく識別 (identification) と呼ぶ．図 1 で表される記録方策を例に考えてみよう．

まず初めに，方策が確率的である場合は反実仮想方策

の性能を予測できることを見よう．図 1 の左部分の領域に着目してほしい．この領域内の任意の点 x を取る．この x に対し，既存方策 ML は行動 1, 2, 3 をすべて正の確率で選ぶため，どの行動 a についても $(A_i, X_i) = (a, x)$ となるログデータが存在し，$E[Y(a)|X = x]$ を識別することができる．すると，この点 x における反実仮想方策 $\pi(a \mid x)$ の性能 $E[\sum_{a\in\mathcal{A}} Y(a)\pi(a|X) \mid X = x]$ が識別可能になる．

一方で，記録方策が非確率的である場合は反実仮想方策の性能を予測することが難しい．仮に，新しい反実仮想方策は，図 1 の色付きの領域内の点 x において，行動 2 を一定の確率で選択するとする．このとき，x において行動 2 を選んだ場合の報酬の期待値 $E[Y(2) \mid X = x]$ については，$(A_i, X_i) = (2, x)$ となるようなデータがログデータ内に存在しないため，この値を直接識別することができない．そのため，非確率的な記録方策のデータを元に新しい方策の性能を予測するには，何らかの工夫が必要になる．

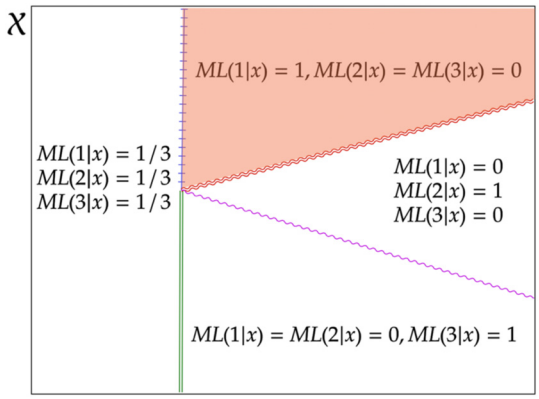

図 1　記録方策の例
文脈の空間は二次元であり，4 分割されている．たとえば色のついている右上の領域を見てみよう．図は，この領域内の文脈 x が観測されたとき，記録方策は確率 1 で行動 1 を選択することを表している．

この壁を乗り越えるために鍵となるのが，**近似傾向スコア（Approximate Propensity Score，以下 APS と略す）**という新たに提案する概念である．まず，数学的な定義を与え，次にその直観的な意味を説明する．まず，$x \in \mathcal{X}$ を中心とした半径 $\delta > 0$ の球を $B(x, \delta)$ で表す．そして，

$$p_\delta^{ML}(a|x) \coloneqq \frac{\int_{B(x,\delta)} ML(a|x^*)dx^*}{\int_{B(x,\delta)} dx^*}$$

とし，文脈 x における行動 a の近似傾向スコアを次のように定義する：

$$p^{ML}(a|x) \coloneqq \lim_{\delta\downarrow 0} p_\delta^{ML}(a|x).$$

近似傾向スコアは，文脈 x の限りなく小さな近傍で，行動 a が選ばれる平均確率と解釈できる．先ほどの例を用いて，近似傾向スコアを理解してみよう．まず，図 2(1) における点 x_1 に着目してみる．この点の十分小さな近傍を考えると，行動 1 は円の上半分においては確率 1 で選ばれ，下半分においては確率 0 で選ばれる．よって，この近傍において行動 1 は確率的に選ばれており，文脈 x_1 における行動 1 の近似傾向スコアは 0.5 となる．同様の手順を繰り返すことで，空間内の各点における近似傾向スコアが計算でき，その結果は図 2(2) のようになる．

2.3　識別（無限データによる学習）

近似傾向スコアは，反実仮想方策の性能を学習できるのはいつかを教えてくれるリトマス試験紙になる．x における近似傾向スコアが 0 と 1 以外の値であれば，x

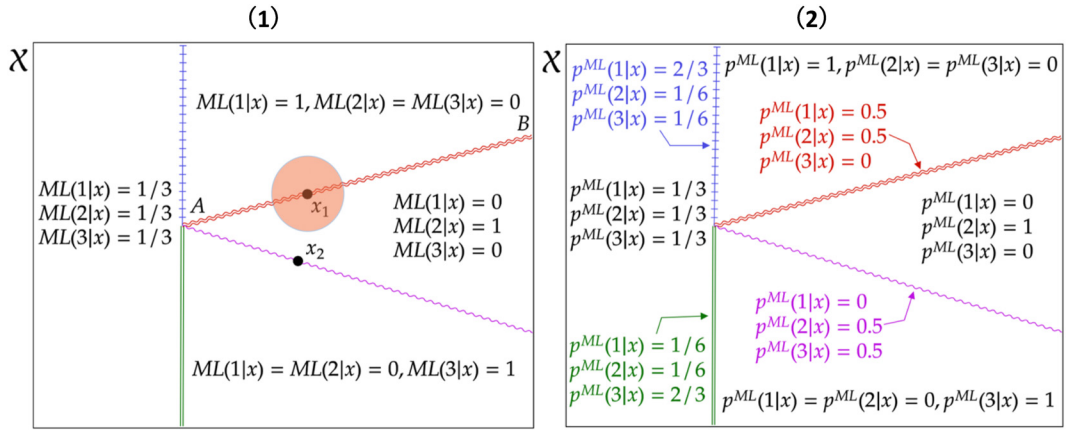

図 2　近似傾向スコアの例

の付近で，処置が行われる場合と行われない場合のどちらもが正の確率で存在する．彼らはほとんど同じ属性をもつほとんど同じ人たちなので，そのような x の付近では自然実験が発生していると考えられる．よって，x の付近で処置を受けた人と受けていない人を比べれば，因果効果を学習できそうだ．実際，以下の補題が成り立つ：

仮定 1（潜在結果の条件付き期待値の局所連続性）． $p^{ML}(a|x) > 0, ML(a|x) = 0$ ならば，$E[Y(a)|X = x]$ は x において連続である．

補題 1. 仮定 1 の下で，$x \in int(\mathcal{X}), p^{ML}(a|x) > 0$ であるならば，$E[Y(a)|X = x]$ が識別可能である．ただし，$int(S)$ は集合 S の内部を表す．

ここで，ある因果効果のパラメーターが識別 (identify) できるとは，その因果効果のパラメーターが (Y_i, X_i, A_i) の同時分布から一意に定まることをさす．つまり，仮に無限大のデータがあり (Y_i, X_i, A_i) の同時分布がわかれば，その因果効果のパラメーターが学習できることを意味する．上の補題は，たとえある行動が記録方策に選ばれなくても，近似傾向スコアが正でありさえすればその因果効果を識別できることを示している．

ただし，補題 1 だけでは，点 x において $p^{ML}(a|x) = 0$ であるような行動 a については $E[Y(a)|X = x]$ が識別できず，結果 π の性能を識別することができない．π の性能の識別については，追加で仮定を課す必要がある．実際，以下の補題が成立することが示せる：

仮定 2（因果的効果は一定）． ある関数 $\beta\colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ が存在し，$E[Y(a)|X] - E[Y(a')|X] = \beta(a, a')$.

仮定 3（非ゼロな APS 列の存在）． 任意の $a \geq 2$ に対し，以下を満たすある選択の列 $1 = a_1, a_2, \ldots, a_L = a$ が存在する：任意の $l \in \{1, \ldots, L-1\}$ に対し，ある $x_l \in int(\mathcal{X})$ が存在し，$p^{ML}(a_{l+1}|x_l) > 0, p^{ML}(a_l|x_l) > 0$.

補題 2. 仮定 1，2，3 の下で，任意の a, x について，$E[Y(a)|X = x]$ が識別される．

仮定 2，3 がどう補題 2 の成立に寄与するのか，再び例を用いて見てみよう．

まず，図 3(2) の点 $\overline{x}$ を考える．この点において行動 1，2 の近似傾向スコアは 0 であるため，$E[Y(1)|\overline{x}]$, $E[Y(2)|\overline{x}]$ を直接は識別できない．どうしたらこれらの因果効果を識別できるだろうか．

次に図 3(1) を見てみよう．色のついた領域内の点については，$p^{ML}(1|x) > 0$，$p^{ML}(2|x) > 0$ が成立している．よって，補題 1 よりこの領域内のサンプル（たとえば図 3(2) の点 x_1）を用いて，$E[Y(1) - Y(2)|x] = \beta(1, 2)$ が識別される．同様にして，$p^{ML}(2|x) > 0$, $p^{ML}(3|x) > 0$ なるサンプル（たとえば図 3(2) の点 x_2）を用いて，$E[Y(2) - Y(3)|x] = \beta(2, 3)$ が識別される．また，点 $\overline{x}$ において行動 3 は選ばれるため，$E[Y(3) \mid \overline{x}]$ もログデータから識別可能である．

さて，これらの結果と，仮定 2，3 から $E[Y(1) \mid \overline{x}]$ が識別されることをみよう．（$E[Y(2)|\overline{x}]$ についても同様に示せる．）まず，簡単な変形により，以下が成立する．

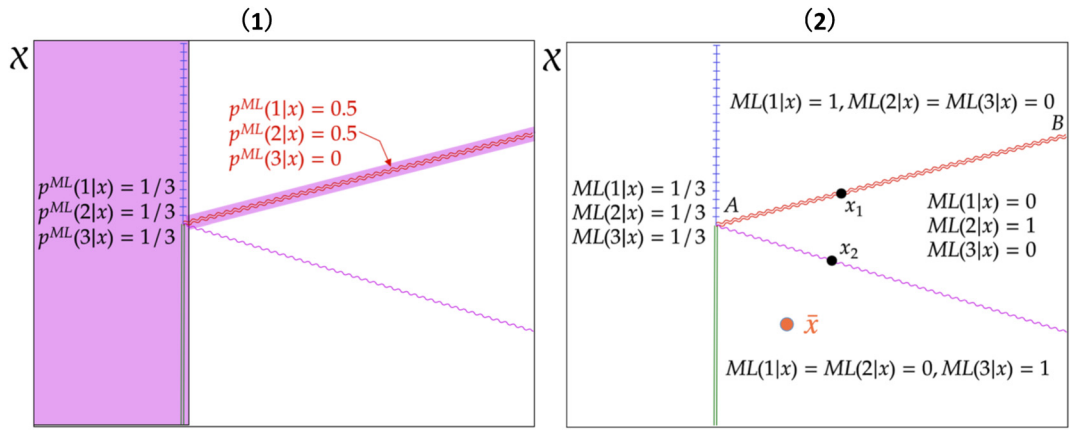

図 3 任意の点における因果効果の識別

$$
\begin{aligned}
E[Y(1)|\overline{x}] &= E[(Y(1)-Y(2))+(Y(2)-Y(3)) \\
&\quad + Y(3)|\overline{x}] \\
&= E[Y(1)-Y(2)|\overline{x}] \\
&\quad + E[Y(2)-Y(3)|\overline{x}] + E[Y(3)|\overline{x}]
\end{aligned}
$$

仮定 2 より，最後の式は以下と等価である．

$$\beta(1,2)+\beta(2,3)+E[Y(3)|\overline{x}]$$

仮定の下，これらの値はいずれも識別可能であることは既に見た．よって，$E[Y(1)|\overline{x}]$ が識別されることがわかった．

以上の結果を組み合わせて，次の命題を示すことができる：

命題 1（反実仮想方策の性能の識別）**.** 仮定 1，2，3 の下，反実仮想方策 π の性能 $V(\pi)$ が識別できる．

2.4 推定（有限データを用いた学習）

命題 1 では，仮に記録方策が生み出したデータが無限に存在した場合の学習（識別）を扱った．では，実世界の有限なデータをどのように分析すれば因果効果を推定できるだろうか？

まず，計算により，以下が成立することを示せる：

$$
\begin{aligned}
V(\pi) &= E\left[\sum_a Y(a)\pi(a|X)\right] \\
&= V(ML) \\
&\quad + E\left[\sum_{a\geq 2}\beta(a,1)(\pi(a|X)-ML(a|X))\right]
\end{aligned}
$$

最右辺の式中の各値は，実は $\beta(a,1)$ 以外のものについては，推定量を自然に構成することができる．実際，$\pi(a \mid x), ML(a \mid x)$ については意思決定者はシミュレートすることが可能だし，$V(ML)$ は $\sum_i Y_i/n$ で推定できることが示せる．よって，$\beta(a,1)$ の推定量をどう構成するかが焦点となる．以降，簡単のため以下を仮定する：

仮定 4. $\forall a,\ \exists x, p^{ML}(a|x)>0,\ p^{ML}(1|x)>0.$

n 人の個人を含むデータ $(Y_i, X_i, A_i)_{i=1}^n$ が与えられたとする．まず，δ を小さい値に設定し，それぞれの個人について $p_\delta^{ML}(a|x)$ を計算する．$p_\delta^{ML}(a|x)$ は，人間の手で解析的に求めるか，$p_\delta^{ML}(a|x)$ の定義式の右辺の積分をシミュレーションで近似すれば計算できる．次に，$A_i \in \{1,a\}, p_\delta^{ML}(a|X_i) \in (0,1)$ を満たす一部の標本 I_a のみを抽出したうえで，以下の値を計算する．

$$q_\delta^{ML}(a|x) \coloneqq \frac{p_\delta^{ML}(a|x)}{p_\delta^{ML}(a|x)+p_\delta^{ML}(1|x)}$$

$q_\delta^{ML}(a|x)$ は，一部の標本 I_a における a の APS と解釈できる．そのうえで，以下の回帰式を最小二乗法 (Ordinary Least Square) で推定する．

$$Y_i = \alpha_a + \beta_a 1\{A_i=a\} + \gamma_a q_\delta^{ML}(a|X_i)$$

この最小二乗法では，近似傾向スコアを制御したうえで，結果変数を処置変数に回帰している．前節での議論が示唆するように，近似傾向スコアを共有する個人の間では処置が行われるかどうかがほとんどランダムに決まると考えられる．そのため，上の最小二乗法を用いて同じ近似傾向スコアを共有した人の中で処置を受けた人と受けなかった人を比べれば，処置の因果効果を測れると期待できる．最小二乗推定量 $\widehat{\beta_a}$ が $\beta(a,1)$ の推定量であり，$V(\pi)$ の推定量は，

$$\hat{V}(\pi) \coloneqq \frac{1}{n}\sum_i Y_i + \frac{1}{n}\sum_{a\geq 2}\widehat{\beta_a}(\pi(a|X_i)-ML(a|X_i))$$

となる．すると，次の事実が成り立つ：

定理 1（反実仮想方策の性能の一致推定）**.** 適当な仮定の下，$\hat{V}(\pi)$ は $V(\pi)$ に $n\to\infty$ で確率収束する．

つまり，標本数が十分大きいとき，提案手法は記録方策 ML によって生成されたデータを用いて，反実仮想手法 π の性能を正しく推定できる．どんなに X_i が高次元で記録方策 ML が複雑であっても，上の単純な最小二乗法さえ回せば因果効果が学習できるのは嬉しい．

3. フリマアプリ・メルカリにおけるクーポン配信方策のデザイン

上述した提案手法を使ってフリマアプリ・メルカリにおけるクーポン配信方策を分析してみよう．ここでは分析の概要を述べるに留め，分析の詳細は Narita et al. [20, 23] に譲る．

まず，メルカリにおいて用いられていた既存のクーポン配信方策（上述の枠組みにおける記録方策に相当）を説明する．クーポン配信の対象となるのは，4 日前にメルカリに登録したものの，まだ何も購入していないユーザーである．まず，過去の A/B テストのデータを用いて，ユーザーの特徴を入力としてクーポン効果の予測値を返すクーポン配布効果の予測モデル $\tau\colon \mathbb{R}^p \to \mathbb{R}$

表 2　クーポン配布がユーザー行動に与える効果の推定値

	Our Proposed Method with APS Controls					Mean
	$\delta = 0.4$	$\delta = 0.8$	$\delta = 1.2$	$\delta = 2.0$	$\delta = 3.0$	Differences
	(1)	(2)	(3)	(4)	(5)	(6)
Effect on Purchase Value	0.35	0.82	0.92	0.54	0.72	−0.17
	(0.59)	(0.39)	(0.30)	(0.28)	(0.21)	(0.11)
Effect on # of Transactions	0.43	0.47	0.66	0.49	0.74	−0.07
	(0.50)	(0.34)	(0.28)	(0.25)	(0.19)	(0.10)
Effect on Point Usage	0.37	0.71	0.57	0.47	0.64	0.68
	(0.42)	(0.29)	(0.26)	(0.22)	(0.17)	(0.04)
Coupon Cost Effectiveness Measure	79.57	96.35	134	93.51	92.07	—
	(130)	(48.97)	(61.97)	(49.33)	(28.45)	
N	2758	4688	6016	8085	9602	89486

各 $\delta \in \{0.4, 0.8, 1.2, 2.0, 3.0\}$ について APS を計算し，$\beta(1,0) = E[Y(1) - Y(0)|x]$ を推定した（なお，ここでは行動の集合を $\mathcal{A} := \{0,1\}$ としている）．第 1 行は購入額を，第 2 行は購入回数を，第 3 行はクーポン使用料を結果の変数とした分析結果である．いずれの値も，標本平均で除し標準化してある．第 5 行の値は，各 δ について APS が非ゼロであった標本の数を表している．第 6 列の値は，$A_i = 1$ である標本と $A_i = 0$ である標本の平均の差を計算した．（　）内の数字は標準偏差または標準誤差を表す．

を構成する．そのうえで，クーポン効果の予測値 $\tau(X_i)$ が上位 8 割以内であるようなユーザーに対し，900 円分のクーポンを配布した．このクーポン配布方策は，

$$ML(x) := 1\{\tau(x) \geq q_{0.2}\}$$

（ただし，$q_{0.2}$ は 20%分位点）と表すことができ，非確率的な方策であることに注意されたい．この記録方策の生成したデータに提案手法を応用し，反実仮想方策の性能を推定する．結果の指標としては，クーポン配布後 18 日以内における (1) 購入額，(2) 購入回数，(3) クーポン使用量を使った．

結果は表 2 のようにまとめられる．

表 2 の 1 列目から 5 列目は，異なる δ について計算した APS を用いて推定した因果効果の値を表している．機密上の理由から，各値は標本平均で除して標準化してある．分析結果は，クーポンを配布した場合に，クーポンを配布しなかった場合と比較して，購入額・購入回数・クーポン使用料がそれぞれ平均の 35–92%，43–74%，37–71%増加することを示している．これはクーポンを配布された人と配布されなかった人の平均購入額の単純な差である第 6 列の値が負であることと対照的である．第 6 列の値が負であることは，記録方策が，購入しにくい人にクーポンを配布していたことを示唆している．

さて，クーポンを配布されると人々はより商品を購入するようになり，クーポン使用額も増えることがわかった．では，メルカリとしてはクーポンを配布することは得なのだろうか？ メルカリは，購入額の 10%を収入として得ている．一方で，使用されたクーポン分はメルカリの支出となる．よって，次の条件が満たされるとき，クーポンをより多く配布することが利益増につながることになる．

$$\text{クーポン配布による購入額の増加} \times 0.1 \geq \text{クーポン配布によるクーポン使用料の増加}$$

これは次の式と等価である．

$$\frac{\text{購入額の増加分}}{\text{クーポン使用料の増加分}} \geq 10$$

表 2 の第 4 行は，上式の左辺の推定値になっている．いずれの値も 10 を大きく上回っている．これは，今より多くの人にクーポンを配布することで，メルカリの利益が上がることを示唆している．

4. 結論と展望

アルゴリズムに基づく意思決定が当たり前となりつつある今日，既存のアルゴリズムが生成したデータを基にアルゴリズムを改善していくための手法・枠組みが重要になっている．アルゴリズムは単に意思決定を行うだけでなく，運用時に新たなデータを生成していく．「アルゴリズムの運用により生成されたデータを用いて，アルゴリズムの性能評価・改善を行い，更新されたアルゴリズムを運用することでよりよい意思決定をしつつ新たなデータの生成も行う．そして新たなデータを用いてアルゴリズムを再び評価・改善する…」といった「運用 → データ生成 → 評価・改善」の流れが今後の社会では当たり前になっていくと予想される．方策外評価はそのような問題の分析・解決に資することを目的とした営みである．本稿では，方策外評価の

枠組みを説明したうえで，その適用範囲を拡張，フリマアプリ・メルカリの業務データに実用した．

まだまだ研究すべき課題は残っている．本稿の分析に直接関わるものに限っても，次のような疑問がすぐに考えられる：

- 「条件付き期待値の差が一定」という仮定（仮定 2）は成立しない場合が多いだろう．そのような場合に本稿の手法を拡張し，なんらかの因果効果を推定することは可能だろうか？
- APS の定義において，近傍の半径 δ はどう定めるべきか？

より大きな問いとして，複数回の方策の更新が可能であるときに，どのように方策を更新していくべきかという問いが考えられる．今回紹介した枠組みでは，あくまで更新の機会は一回であったが，データ駆動型制度設計が真に社会に定着した折には，方策を一定間隔で逐次更新していくことになる．その際にどう方策を更新するべきかについての指針を与えるための枠組み・手法の開発も，今後の重要なテーマになると考えられる．

謝辞　本稿を完成させるにあたり，杉山侑吏さんからご意見やご助言をいただきました．ありがとうございました．

参考文献

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Bradford Book, 2018.
[2] P. Z. Cohen, R. W. Hahn, J. Hall, S. D. Levitt and R. Metcalfe, "Using big data to estimate consumer surplus: The case of Uber," *NBER Working Paper*, 22627, 2016.
[3] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig and S. Mullainathan, "Human decisions and machine predictions," *Quarterly Journal of Economics*, **133**, pp. 237–293, 2017.
[4] 安田洋祐（編），『学校選択制のデザイン―ゲーム理論アプローチ―』，NTT 出版，2010.
[5] A. Abdulkadiroglu, J. D. Angrist, Y. Narita and P. A. Pathak, "Research design meets market design: Using centralized assignment for impact evaluation," *Econometrica*, **85**, pp. 1373–1432, 2017.
[6] A. Abdulkadiroglu, J. D. Angrist, Y. Narita, P. A. Pathak and R. Zarate, "Regression discontinuity in serial dictatorship: Achievement effects at Chicago's exam schools," *American Economic Review*, **107**, pp. 240–245, 2017.
[7] A. Abdulkadiroglu, J. D. Angrist, Y. Narita and P. A. Pathak, "Breaking ties: Regression discontinuity design meets market design," *Econometrica*, forthcoming.
[8] Y. Narita, "A theory of quasi-experimental evaluation of school quality," *Management Science*, **67**, pp. 4643–5300, 2020.
[9] M. Tanaka, Y. Narita, C. Moriguchi, "Meritocracy and its discontent: Long-run effects of repeated school admission reforms," *RIETI Discussion Paper Series*, 20-E-002, 2020.
[10] K. Kawai and J. Nakabayashi, "Detecting large-scale collusion in procurement auctions," *Journal of Political Economy*, forthcoming.
[11] S. Chawla, J. D. Hartline and D. Nekipelov, "Mechanism redesign," *arXiv preprint*, arXiv:1708.04699, 2017.
[12] Y. Narita, "Incorporating ethics and welfare in randomized experiments," In *Proceedings of the National Academy of Sciences*, 2020.
[13] D. Precup, "Eligibility traces for off-policy policy evaluation," In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766, 2000.
[14] L. Li, W. Chu, J. Langford and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," In *Proceedings of the 19th international conference on World wide web (WWW)*, pp. 661–670, 2010.
[15] F. Amat, A. Chandrashekar, T. Jebara and J. Basilico, "Artwork personalization at netflix," In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 487–488, 2018.
[16] Y. Narita, S. Yasui and K. Yata, "Efficient counterfactual learning from bandit feedback," In *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, pp. 4634–4641, 2019.
[17] Y. Narita and K. Yata, "Algorithm is experiment: machine learning, market design, and policy eligibility rules," *RIETI Discussion Paper Series*, 21-E-057, 2020.
[18] Y. Saito, S. Aihara, M. Matsutani and Y. Narita, "Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation," In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*, 2021.
[19] Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita and K. Tateno, "Evaluating the robustness of off-policy evaluation," In *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, pp. 114–123, 2021.
[20] Y. Narita, K. Okumura, A. Shimizu and K. Yata, "Counterfactual learning with general data-generating policies," *Working Paper*, 2021.
[21] H. Kiyohara, Y. Saito, T. Matsuhiro, Y. Narita, N. Shimizu and Y. Yamamoto, "Doubly robust off-policy evaluation for ranking policies under the cascade behavior model," *International Conference on Web Search and Data Mining*, 2022.
[22] 成田悠輔，栗飯原俊介，齋藤優太，松谷恵，矢田紘平，"すべての機械学習は A/B テストである，" 人工知能，**35**，pp. 517–525, 2020.
[23] Y. Narita, S. Yasui and K. Yata, "Debiased off-policy evaluation for recommendation systems," In *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 372–379, 2021.

