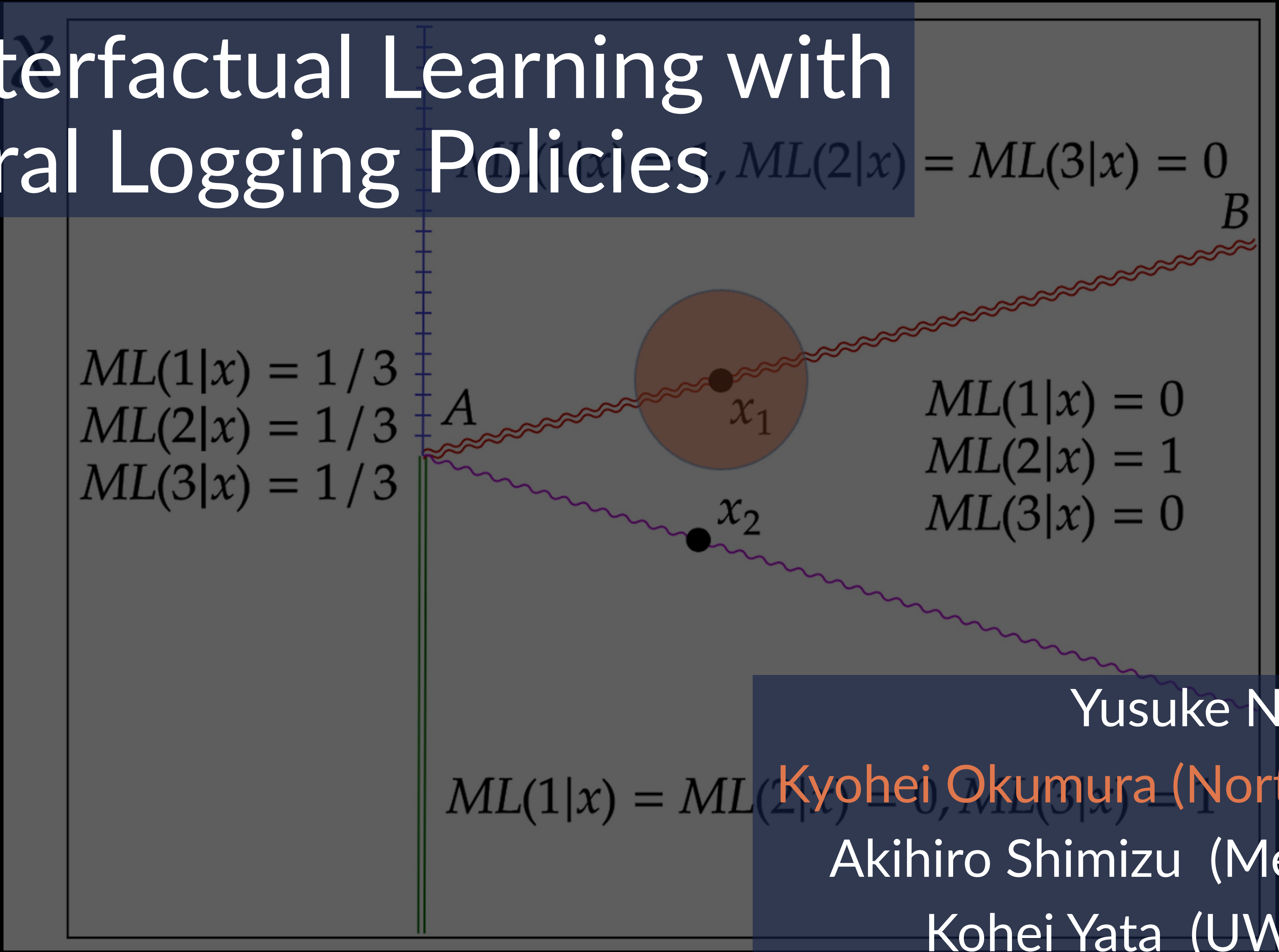
Counterfactual Learning with General Logging Policies



Yusuke Narita (Yale)
Kyohei Okumura (Northwestern)
Akihiro Shimizu (Mercari, Inc.)
Kohei Yata (UW Madison)

Algorithms are eating the world

Decision making by algorithms are everywhere in the world.

Advertisement	Law	Security
		

Decision making by algorithms

Input	Action (Decision)	Outcome
X	A	Y

$$X \xrightarrow{\text{policy } \pi} A$$

e.g.	Individual Characteristics (Age, Gender, etc.)	Give a coupon or not	Purchase Value

How can we choose policy π that achieves better outcomes？

Off-Policy Evaluation

- **Goal:** Estimate the performance of a new counterfactual policy
- **Solution 1:** A/B Testing
  - Problem: Costly and/or risky
- **Solution 2:** Off-Policy Evaluation
  - Use the log data generated by the existing policy
  - Improve the system without conducting A/B tests

Full Support vs. Deficient Support Policies

- **Full support policy**: for all covariates, all actions are chosen w.p. >0

$$\forall x \forall a, \pi(a \mid x) > 0$$

- Real-world decision-making often uses deficient support policies
  - e.g. Give a coupon iff one's covariate is in some region

- **Problem:** Hard to conduct estimation using the log data generated by deficient support policies

Our Contribution

- Propose an OPE estimator applicable to data generated by a broad class of policies including deficient support policies.
  - **Theory**: the estimator has consistency:
    the prediction converges in probability to the true performance of a counterfactual policy as the sample size increases.
  - **Real-world Application:** evaluate coupon targeting policies by a major online platform, Mercari.

    - How much more do people spend when they get a coupon?
    - Should the company allocate more coupons or not?

Overview

Framework

Key Concept: APS

Results

Conclusion

Framework

- Action $a \in \mathcal{A} := \{1, \ldots, m\}$.
- Potential rewards $\left(Y(a)\right)_{a \in \mathcal{A}}$
  - Action a is chosen → Reward $Y(a)$ is observed
- Context X. $\mathcal{X} := \mathrm{supp}(X) \subseteq \mathbb{R}^p$
- Logging policy $ML: \mathcal{X} \to \Delta(\mathcal{A})$
  - $ML(a \mid x)$ = proba of taking action a for indiv with context x.
  - The researcher can simulate ML (i.e., knows $ML(a \mid x)$).

DGP of log data $(Y_i, X_i, A_i)_{i=1}^n$

- Logging policy ML generates log data $(Y_i, X_i, A_i)_{i=1}^n$
- For each i,
  1. $\left(\left(Y_i(a)\right)_a, X_i\right)$ is i.i.d.-drawn from an unknown distribution
  2. Action A_i is chosen w.p. $ML(A_i \mid X_i)$
  3. Reward $Y_i := Y_i(A_i)$ is recorded.

Input | Action (Decision) | Outcome

$$\boxed{X_i \xrightarrow{\;ML\;} A_i} \longrightarrow Y_i = Y_i(A_i)$$

Goal

Using log data $(Y_i, X_i, A_i)_{i=1}^{n}$,
estimate the performance of a counterfactual policy π

$$V(\pi) := \mathbb{E}\left[\sum_{a \in \mathcal{A}} Y(a)\pi(a \mid X)\right]$$

1. Is it possible to estimate $V(\pi)$? → **Identification**

In the ideal world where we could have an infinite amount of data

2. How can we estimate $V(\pi)$ given finite data? → **Estimation**

- Overview
- Framework
- **Key Concept: APS**
- Results
- Conclusion

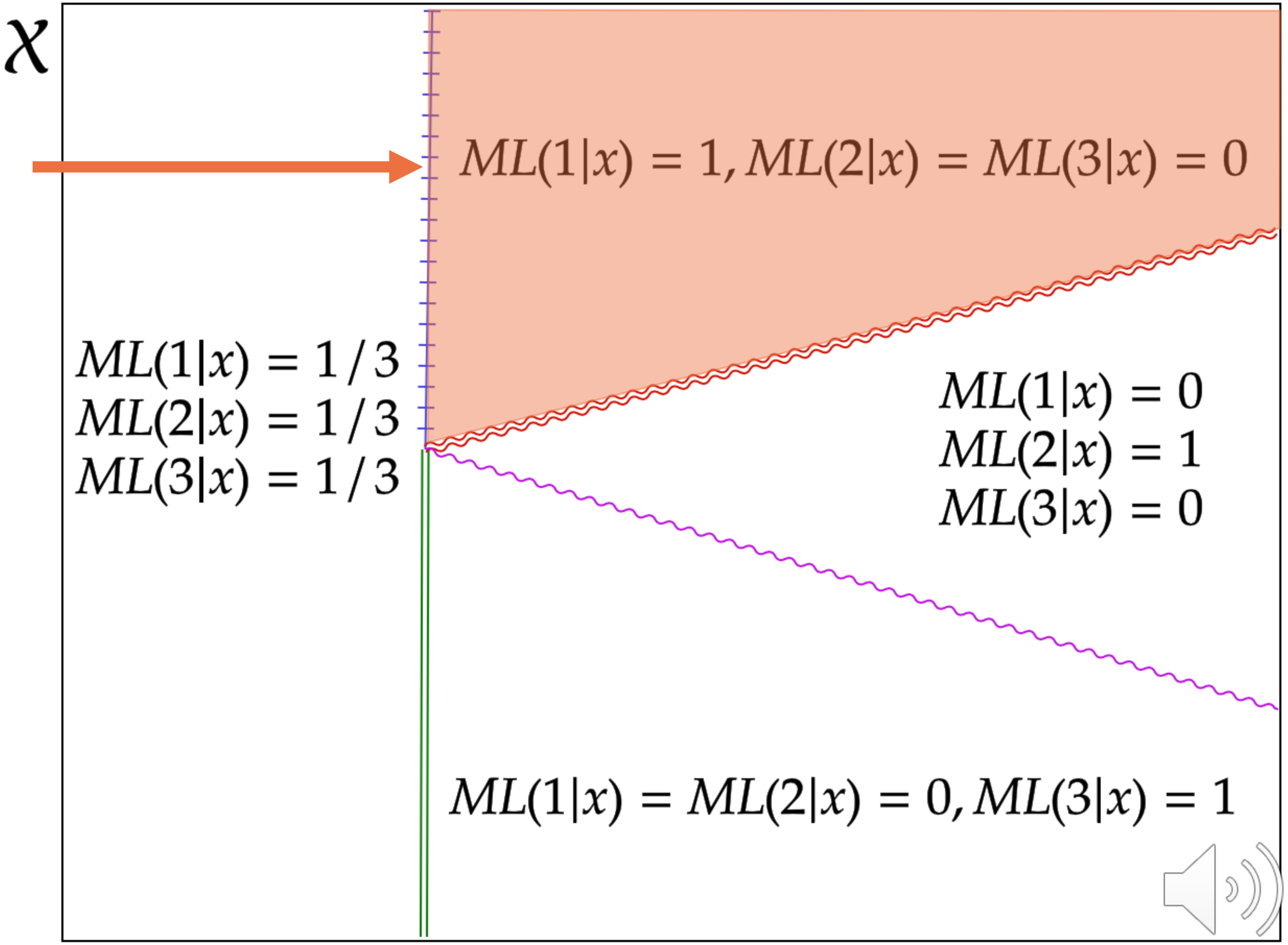
Estimation with deficient support policy is hard

- **Full-support logging policy**: For any x and a, $ML(a \mid x) \in (0,1)$
  - In this case, $V(\pi)$ is identified using propensity score.
- **Deficient support logging policy**: For some $\overline{a}, \overline{x}$, the logging policy produce no data on $Y(\overline{a}) \mid X = \overline{x}$
  - $\mathbb{E}[Y(\overline{a}) \mid X = \overline{x}]$ is not directly identified.

Ex. Deficient support $\mathcal{X} \subseteq \mathbb{R}^2, \mathcal{A} = \{1,2,3\}$

- Logging policy ML only chooses $a = 1$ in this region.

- Since there is no data, $\mathbb{E}[Y(2) \mid x], \mathbb{E}[Y(3) \mid x]$ are not easily identified.



Q: Can we estimate some causal effects using log data generated by a deficient support policy?

A: Yes, we can.

Key: Approximate Propensity Score (APS)

Approximate Propensity Score

Ball w/ center $x \in \mathcal{X}$, radius $\delta > 0$

$$p_\delta^{ML}(a \mid x) := \frac{\int_{\boxed{B(x,\delta)}} ML(a \mid x^*) dx^*}{\int_{B(x,\delta)} dx^*}$$

Average probability that action $a \in \mathcal{A}$ is chosen by ML in the neighborhood of point $x \in \mathcal{X}$

Approximate Propensity Score

Approximate Propensity Score (APS)

$$p^{ML}(a \mid x) := \lim_{\delta \downarrow 0} p_{\delta}^{ML}(a \mid x)$$

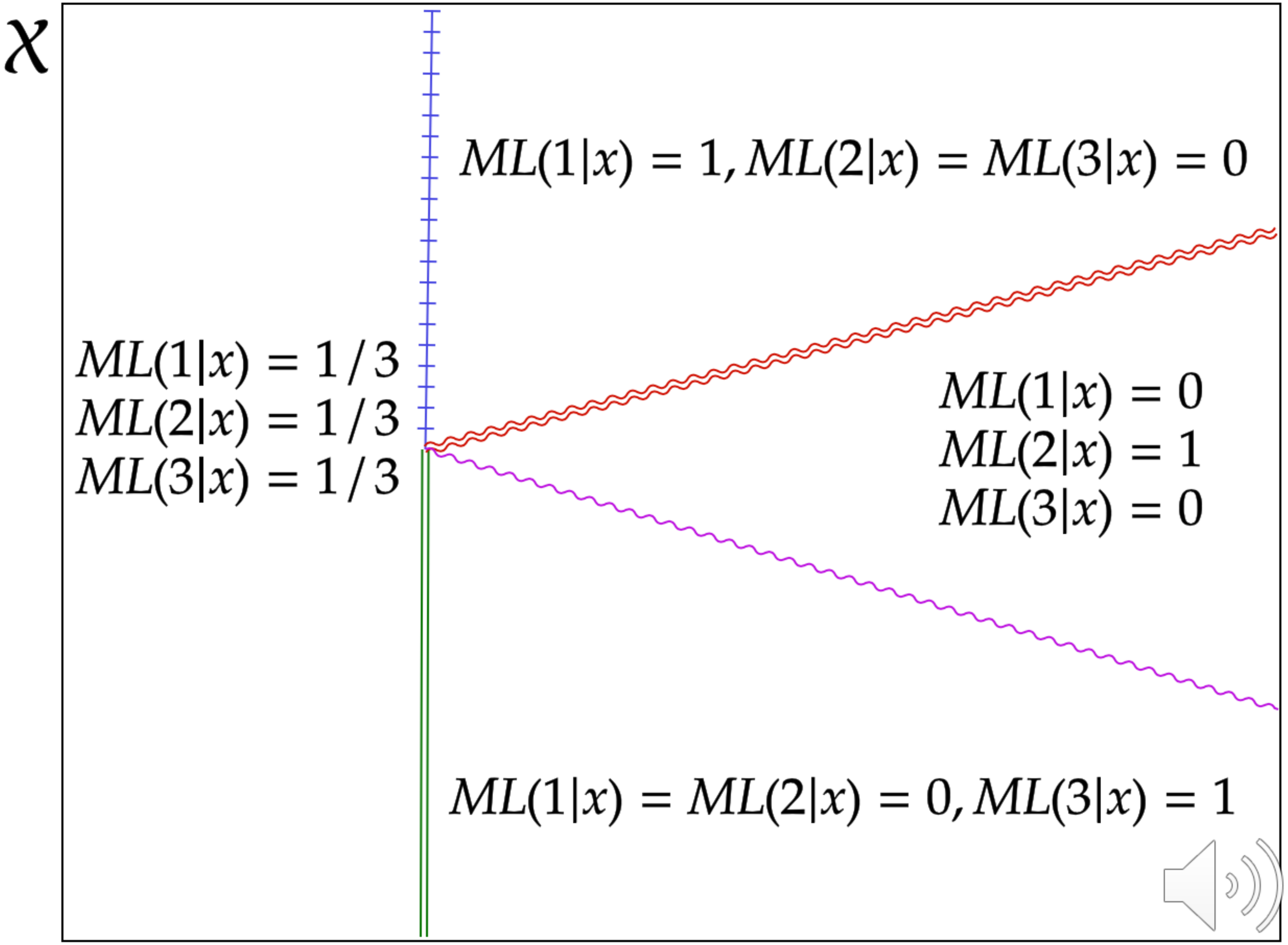
Average probability that action $a \in \mathcal{A}$ is chosen by ML
in the neighborhood of point $x \in \mathcal{X}$

Ex. APS

$$\mathcal{X} \subseteq \mathbb{R}^2, \mathcal{A} = \{1,2,3\}$$

$$p^{ML}(\, a \mid x \,) := \lim_{\delta \downarrow 0} p_{\delta}^{ML}(a \mid x)$$

$$p_{\delta}^{ML}(a \mid x) := \frac{\int_{B(x,\delta)} ML(a \mid x^*) dx^*}{\int_{B(x,\delta)} dx^*}$$

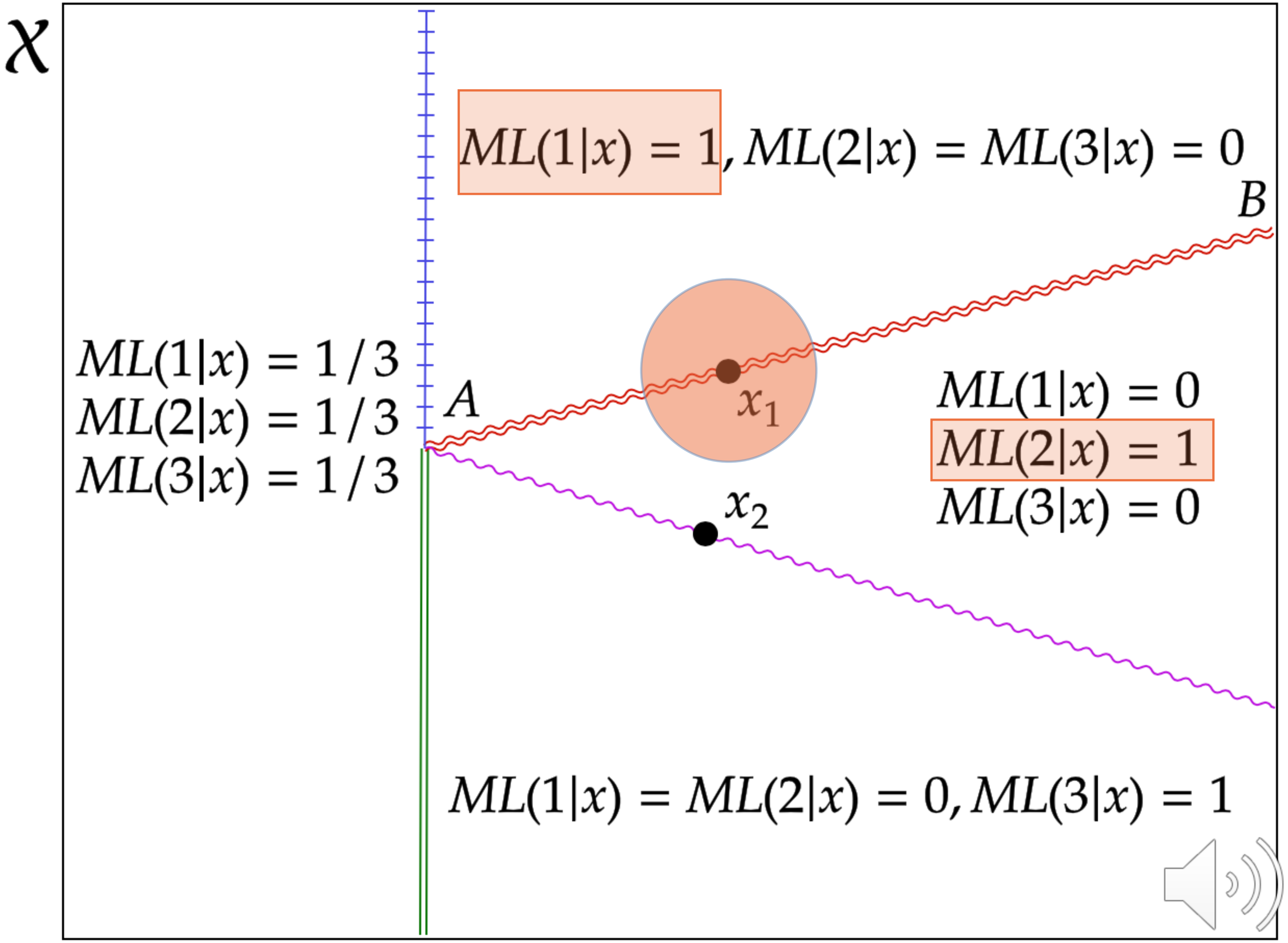


Ex. APS

$$\mathcal{X} \subseteq \mathbb{R}^2, \mathcal{A} = \{1,2,3\}$$

$$p^{ML}(\, a \mid x \,) := \lim_{\delta \downarrow 0} p_\delta^{ML}(a \mid x)$$

$$p_\delta^{ML}(a \mid x) := \frac{\int_{B(x,\delta)} ML(a \mid x^*) dx^*}{\int_{B(x,\delta)} dx^*}$$

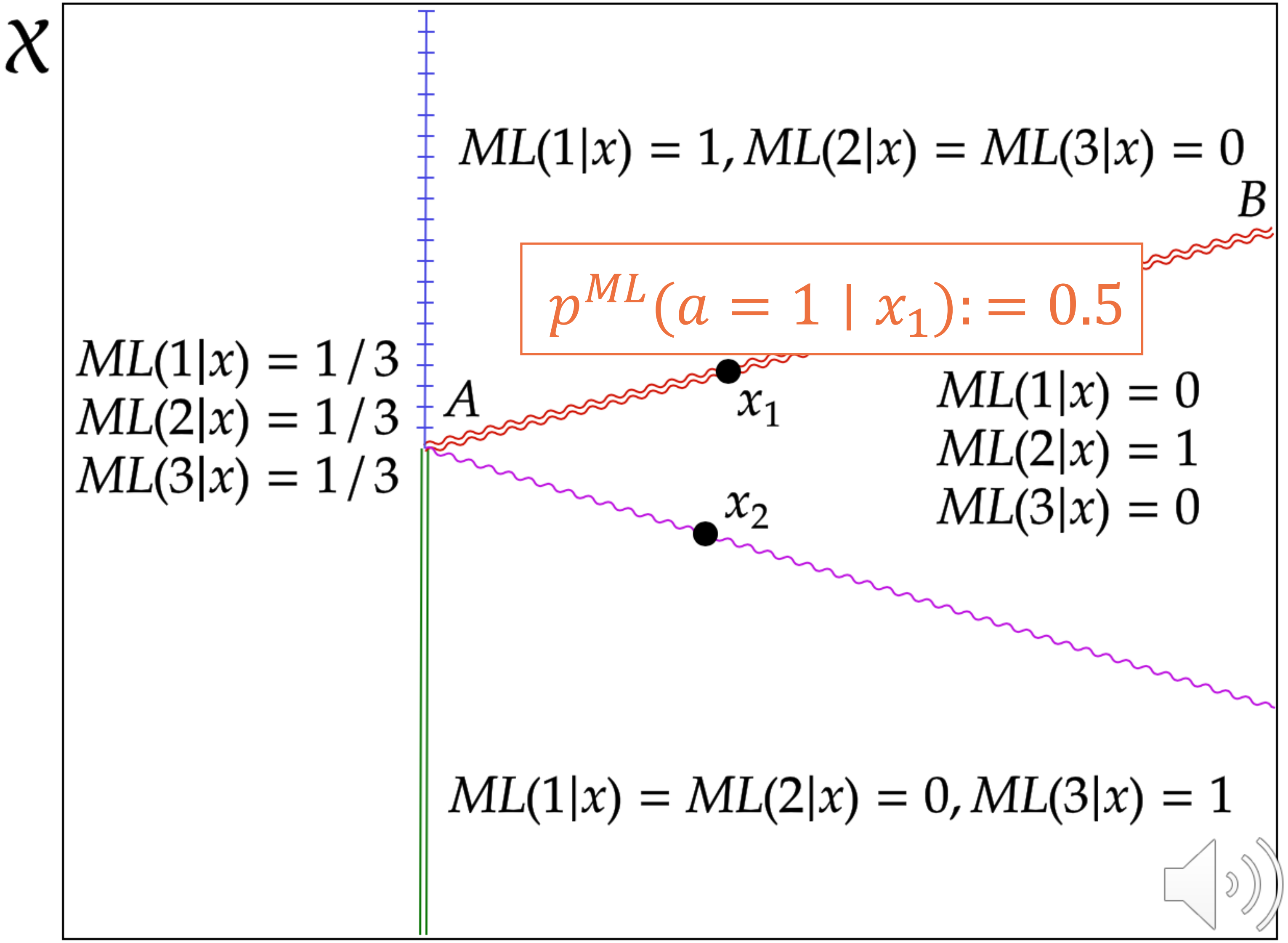


Ex. APS

$$\mathcal{X} \subseteq \mathbb{R}^2, \mathcal{A} = \{1,2,3\}$$

$$p^{ML}(a \mid x) := \lim_{\delta \downarrow 0} p_\delta^{ML}(a \mid x)$$

$$p_\delta^{ML}(a \mid x) := \frac{\int_{B(x,\delta)} ML(a \mid x^*) dx^*}{\int_{B(x,\delta)} dx^*}$$



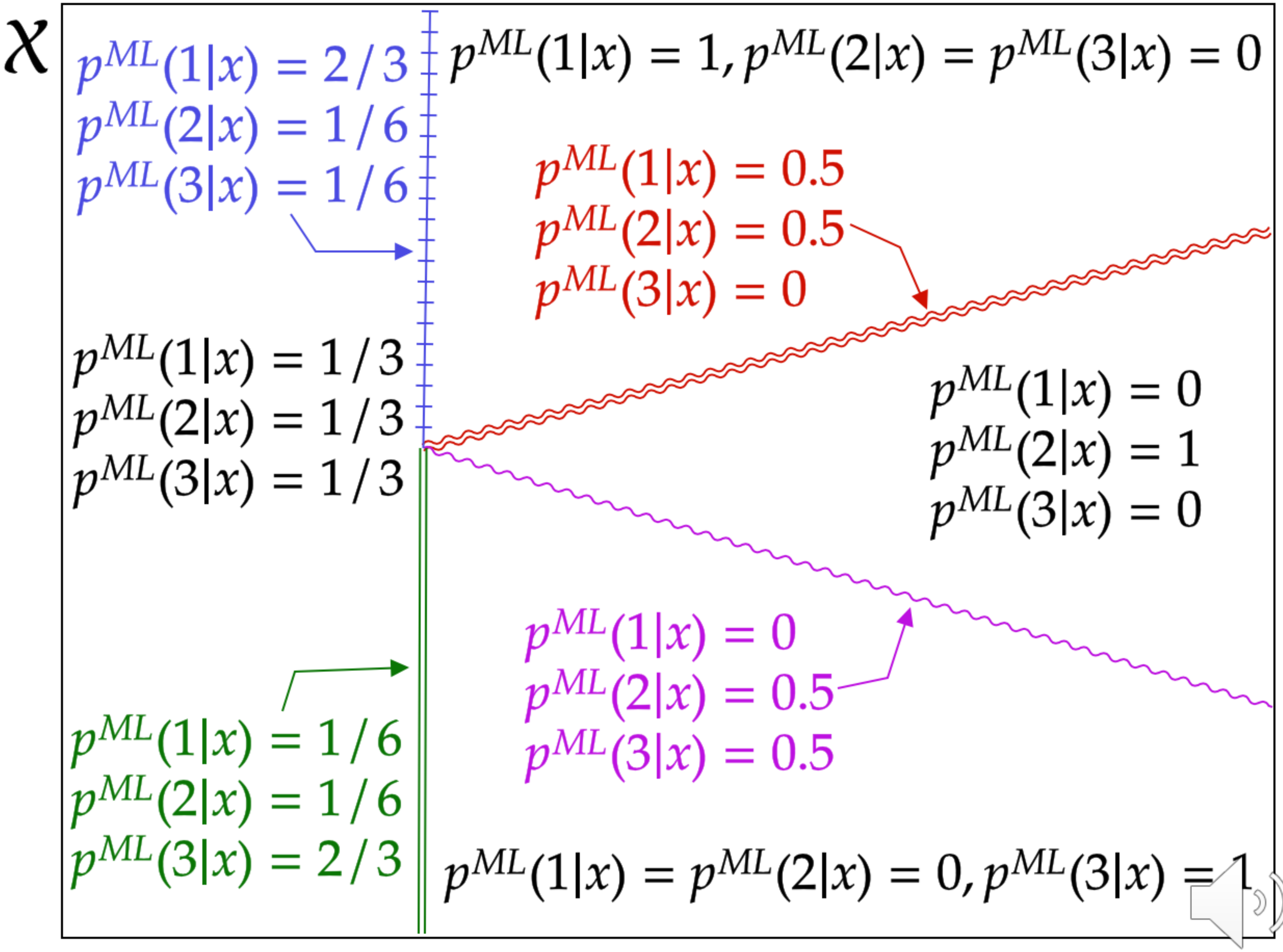
Ex. APS

$\mathcal{X} \subseteq \mathbb{R}^2, \mathcal{A} = \{1,2,3\}$

Under certain cond.'s,
For any a, x s.t.

$$p^{ML}(a \mid x) > 0,$$

$\mathbb{E}[\,Y(a) \mid x\,]$ is identified.



- Overview
- Framework
- Key Concept: APS
- **Results**
- Conclusion

Identification(Learning w/ infinite data)

Prop. 1 (Identification of the performance)

Under A1-3, the performance $V(\pi)$ of counterfactual policy π is identified.

A2 (Constant Conditional Mean Differences)

There exists $\beta: \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ s.t. $\mathbb{E}[Y(a) \mid X] - \mathbb{E}[Y(a') \mid X] = \beta(a, a')$

See the paper for discussion on A2
(and def. of A1&A3)

From Identification to Estimation

$$V(\pi) := \mathbb{E}\left[\sum_{a\in\mathcal{A}} Y(a)\pi(a \mid X)\right]$$

$$= V(ML) + \mathbb{E}\left[\sum_{a=2}^{m} \beta(a,1)(\pi(a \mid X) - ML(a \mid X))\right]$$

From Identification to Estimation

$$V(\pi) := \mathbb{E}\left[\sum_{a \in \mathcal{A}} Y(a)\pi(a \mid X)\right]$$

$$= V(ML) + \mathbb{E}\left[\sum_{a=2}^{m} \beta(a,1)\boxed{(\pi(a \mid X) - ML(a \mid X))}\right]$$

Known to the researcher

From Identification to Estimation

$$V(\pi) := \mathbb{E}\left[\sum_{a \in \mathcal{A}} Y(a)\pi(a \mid X)\right]$$

$$= \boxed{V(ML)} + \mathbb{E}\left[\sum_{a=2}^{m} \beta(a,1)(\pi(a \mid X) - ML(a \mid X))\right]$$

Can be estimated by $\frac{1}{n}\sum_{i=1}^{n} Y_i$

From Identification to Estimation

$$V(\pi) := \mathbb{E}\left[\sum_{a \in \mathcal{A}} Y(a)\pi(a \mid X)\right]$$

$$= V(ML) + \mathbb{E}\left[\sum_{a=2}^{m} \boxed{\beta(a,1)}(\pi(a \mid X) - ML(a \mid X))\right]$$

Can be estimated via OLS with APS control

(See the paper for more details)

Estimation (Learning w/ finite data)

$$V(\pi) = \boxed{V(ML)} + \mathbb{E}\left[\sum_{a=2}^{m} \boxed{\beta(a,1)}(\pi(a \mid X) - ML(a \mid X))\right]$$

$$\hat{V}(\pi) = \boxed{\frac{1}{n}\sum_{i=1}^{n} Y_i} + \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a=2}^{m} \boxed{\hat{\beta}_a}(\pi(a \mid X) - ML(a \mid X))\right]$$

Estimation (Learning w/ finite data)

Thm. 1 (Consistency)

Under certain cond.'s, $\hat{V}(\pi)$ converges in proba. to $V(\pi)$ as $n \to \infty$

If the sample size is sufficiently large,

the proposed method correctly estimates the performance of the counterfactual policy π

Application: Coupon Targeting Policy at Mercari

- Logging policy (current policy used by the company)
  - Use data from a past A/B test, train a prediction model τ
  - $\dim(\mathcal{X}) > 200$
  - Offer a coupon to those with a high predicted effect (top 80% of the distribution)

$$ML(a \mid x) = \boxed{1\{\tau(x) \geq q_{0.2}\}}$$

Deficient support

- **Result:** it would be profitable to expand the campaign.

- Overview
- Framework
- Key Concept: APS
- Results
- **Conclusion**

Summary

- This paper proposes an OPE method applicable to log data generated by a broad class of logging policies, including deficient support ones.
  - Based on approximated propensity score(APS)
  - The estimator has consistency.

- Apply the proposed method to the coupon targeting policy at Mercari.
  - An example of natural logging policies with deficient support
  - **Result**: it would be profitable to expand the campaign.