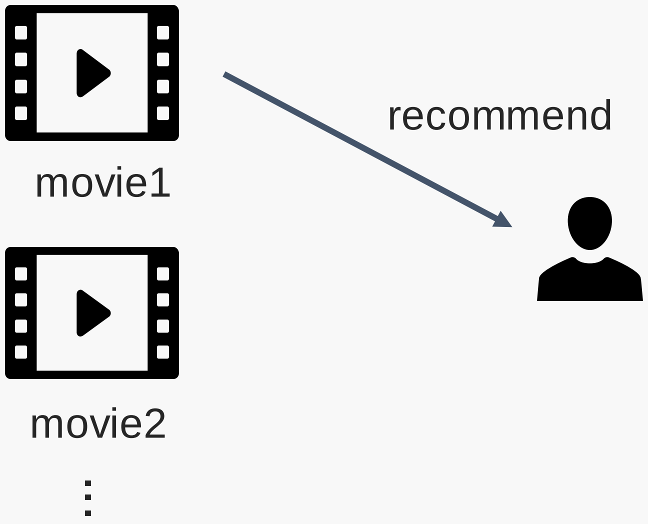
Cross-Domain Off-Policy Evaluation and Learning for Contextual Bandits

Yuta Natsubori[1], Masataka Ushiku[1], Yuta Saito[2]

[1]Hakuhodo DY Holdings, Inc., [2]Cornell University

Off-Policy Evaluation (OPE)

Movie recommendation



Logging policy π_0 makes decisions to recommend movie

Logged dataset collected by logging policy

$$\mathcal{D} \sim \pi_0$$

OPE

We aim to evaluate the performance of a target policy (new system) π using only the logged dataset

Issues of an existing estimator

Many recent estimators are based on Inverse propensity scoring (IPS)

$$\hat{V}_{\mathrm{IPS}}(\pi; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}}_{w(x_i, a_i)} r_i$$

Importance weight

The use of importance weighting often causes

severe variance issues, particularly when the sample size is small.

Issues of an existing estimator

IPS relies on the *common support* assumption to provide a low-bias estimate.

$$\pi(a|x) > 0 \implies \pi_0(a|x) > 0, \ \forall a \in \mathcal{A}, \forall x \in \mathcal{X}$$

We can only evaluate a target policy regarding actions that have already been sufficiently explored by the logging policy

When logging policy is deterministic or there are new actions, existing estimators cannot evaluate under-explored or new actions at all.

A New problem setup:
Cross-domain OPE

Cross-domain OPE problem setup

We can have access not only to the logged data collected from ***target* domain** but also the data from ***source* domains**.

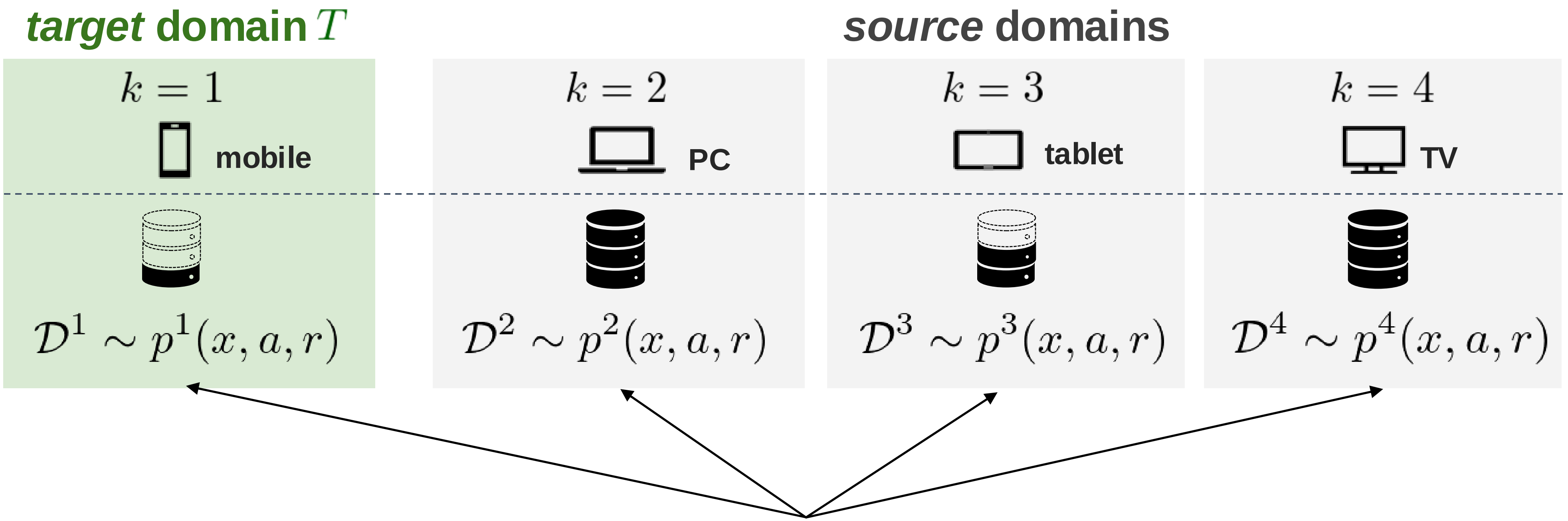
	***target* domain T**	***source* domains**		
domain k (e.g., device)	$k=1$ **mobile**	$k=2$ **PC**	$k=3$ **tablet**	$k=4$ **TV**
logged data $\mathcal{D}$	**sample size : small** $\mathcal{D}^1 \sim p^1(x,a,r)$	**sample size : large** $\mathcal{D}^2 \sim p^2(x,a,r)$	**sample size : medium** $\mathcal{D}^3 \sim p^3(x,a,r)$	**sample size : large** $\mathcal{D}^4 \sim p^4(x,a,r)$
actions in logged data (e.g., movie)	movie1 movie2 movie3	movie1 movie2 movie6	movie2 movie3 movie5	movie3 movie4 movie6

target policy → π

actions that can be taken under a target policy: movie1, movie3, movie2, **movie4** (**new action**)

We can leverage these data even when the target domain has <u>less logged data</u>, <u>a deterministic logging policy</u>, and <u>new actions</u>.

Important remark on the use of data from different domains



Different data generation process

Naively integrating the source domains data for estimation of the target policy introduces substantial bias.

Key idea: Reward function decomposition

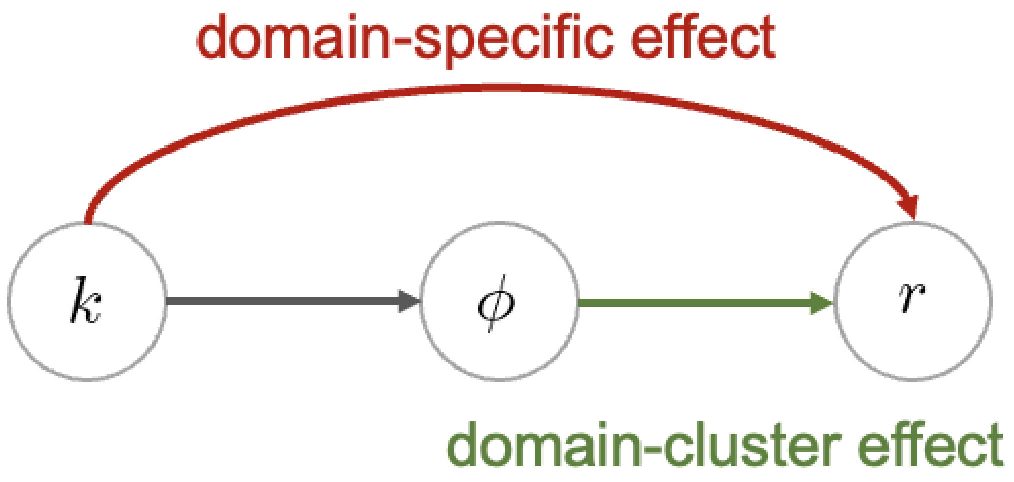
We consider the following decomposition of the reward function

$$\underbrace{q^k(x,a)}_{\text{expected reward in domain } k} = \underbrace{g(x,a,\phi(k))}_{\text{domain-cluster effect}} + \underbrace{h(x,a,k)}_{\text{domain-specific effect}}$$

Where ϕ is a function to cluster similar domains (e.g., similar device)

The domain-cluster effect :
Effect that domains in the same cluster have in common.

The domain-specific effect :
Effect that depends on each domain k .



The COPE estimator

The COPE estimator leverages data explored in source domains and is defined as

$$\hat{V}_{\mathrm{COPE}}(\pi; \mathcal{D}^{\phi(T)}) := \underbrace{\frac{1}{n^{\phi(T)}} \sum_{k \in \phi(T)} \sum_{i=1}^{n^k} \frac{\pi(a_i^k | x_i^k)}{p^{\phi(T)}(a_i^k | x_i^k)} (r_i^k - \hat{q}^T(x_i^k, a_i^k))}_{\text{estimate the domain-cluster effect}} + \underbrace{\frac{1}{n^T} \sum_{i=1}^{n^T} \sum_{a^T \in \mathcal{A}} \pi(a^T | x_i^T) \hat{q}^T(x_i^T, a^T)}_{\text{estimate the domain-specific effect}}$$

The denominator of the importance weight of COPE is defined by

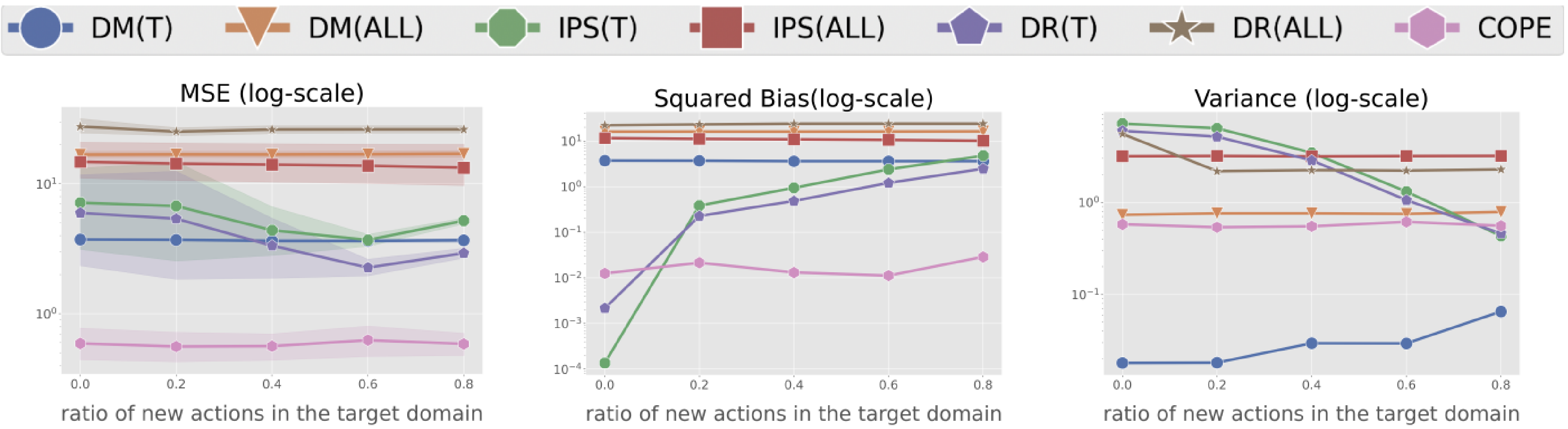
$$p^{\phi(T)}(a|x) := \frac{1}{n^{\phi(T)}} \sum_{k \in \phi(T)} n^k \frac{p^k(x)}{p^T(x)} \pi_0^k(a|x)$$

COPE has a low bias by estimating the value of deficient and new actions by using data from source domains that are in the same cluster as the target domain, i.e., $\phi(T)$.

COPE also provides substantial variance reduction by transferring information from source domains.

Experiment Results

OPE experiments under varying ratios of new actions



COPE consistently performs the best without being affected by the presence of new actions.

Please check out the paper for the details.