The limit of model-free off-policy evaluation and adaptive oracle-assisted estimators

Yu-Xiang Wang

Carnegie Mellon University

Mentors: Alekh Agarwal, Miro Dudik and the rest of ML team

Off-Policy Evaluation: Answering the "what-if" question

- Targeted advertisement
  - A "policy" decides which ad to show based on "context"
  - Then the user may click or not click
  - The click-through rate measures how good the policy is

- **What if** I ran a different policy instead?
  - a.k.a., Counterfactual reasoning

Many applications

- For safe policy deployment
- For policy optimization

Contextual bandits

- Contexts:
  - drawn iid, possibly infinite domain
- Actions:
  - Taken by a randomized policy μ
- Reward:
  - revealed only for the action taken
- Value:
  - Expected reward of a policy

- We collect data $\{x_i, a_i, r_i\}_{i=1}^n$, we also know μ completely.

- **What if** we use a different policy π?
  - How to we estimate the value of π, using $\mu, \pi, \{x_i, a_i, r_i\}_{i=1}^n$,

Importance sampling/Inverse propensity scoring

Importance weight: ρ_i

$$\hat{v}^{\pi}_{\mathrm{IPS}} = \sum_{i=1}^{n} \frac{\pi(a_i, x_i)}{\mu(a_i, x_i)} r_i.$$

Pros:

- Model-free
- Unbiased
- Computationally efficient

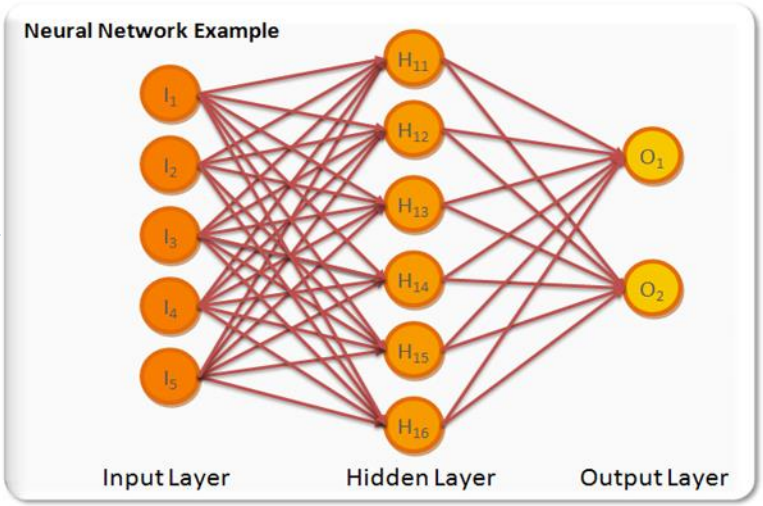
Cons:

- High variance because the weight is large

Model-based approach

Feature

Context × Action



Label

Reward

Pros:

- Low-variance. Predict E(r|x,a) directly.
- Can evaluate on unseen contexts

Cons:

- Often high bias
- The model can be wrong/hard to learn

Many estimators are proposed. Are they optimal? How good is good enough?

- The need of a minimax theory

$$\inf_{\hat{v} \in \text{all estimators}} \sup_{\text{A class of contextual bandits problems}} \mathbb{E}_{\mu}(\hat{v} - v^{\pi})^2$$

- Minimax risk/rate
  - The best “worst-case” risk / rate
  - That’s information-theoretically possible

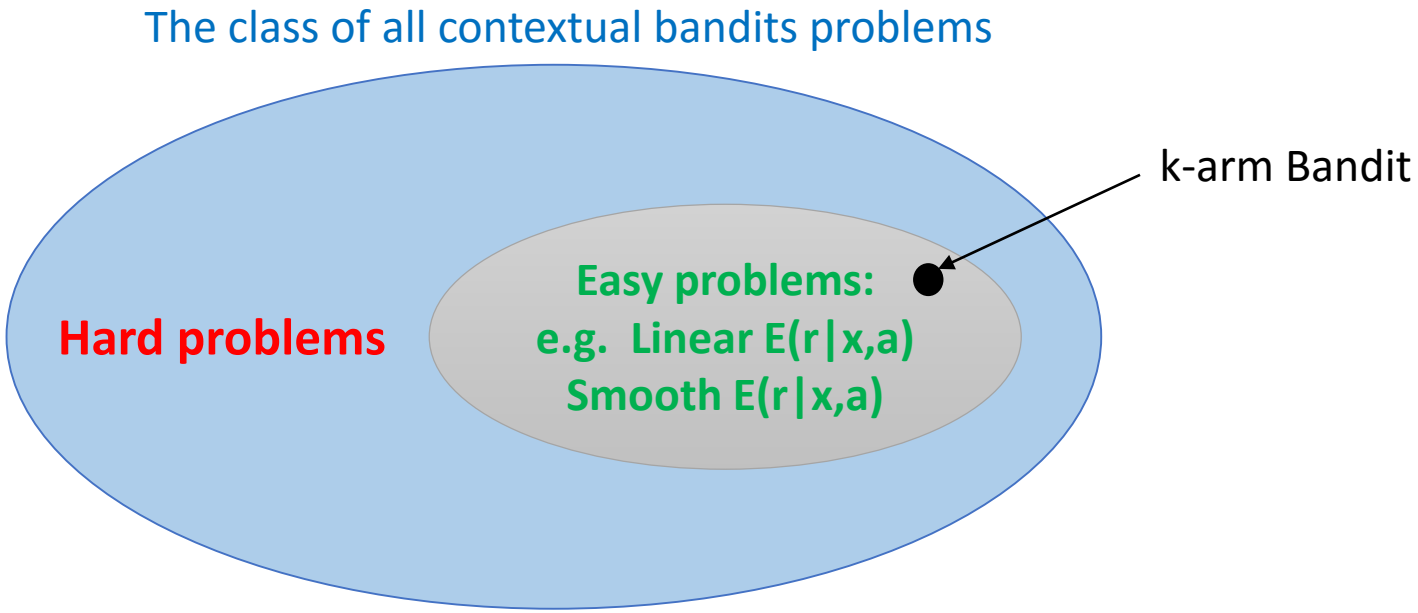
We show that: IPS is optimal!

- The high variance is required.
  - In contextual bandits with large context spaces.
  - Model-free approach is fundamentally limited.

- Somewhat surprising because:
  - Lihong showed that in k-arm bandit (when the number of contexts is small), IPS is strictly suboptimal.

Li, Lihong, Rémi Munos, and Csaba Szepesvári. "Toward Minimax Off-policy Value Estimation." *AISTATS*. 2015.

The pursuit of adaptive estimators



- Perform minimax optimal on hard problems.
- Perform better on easier problems.

Suppose we are given an oracle

- Could be very good, or completely off.
- How to **make the best use** of the predictions?

Why not just use doubly robust?

- An elegant way of combining IPS/model-based.

  Dudık, Langford and Li. "Doubly Robust Policy Evaluation and Learning." *ICML*-11.

  Jiang and Li. "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning." *ICML*-2016.

- We show that: DR is Strictly suboptimal even with perfect oracle:

$$\hat{r}|x,a = \mathbb{E}(r|x,a).$$

Oracle-Assisted estimator

- Recall that IPS is bad because: $\hat{v}_{\mathrm{IPS}}^{\pi} = \sum_{i=1}^{n} \frac{\pi(a_i, x_i)}{\mu(a_i, x_i)} r_i.$

- Oracle-Assisted estimator:

For each $i = 1, \ldots, n$, for each action $a \in A$:

if $\frac{\pi(a,x_i)}{\mu(a,x_i)} \leq \tau$:

use IPS (or DR)

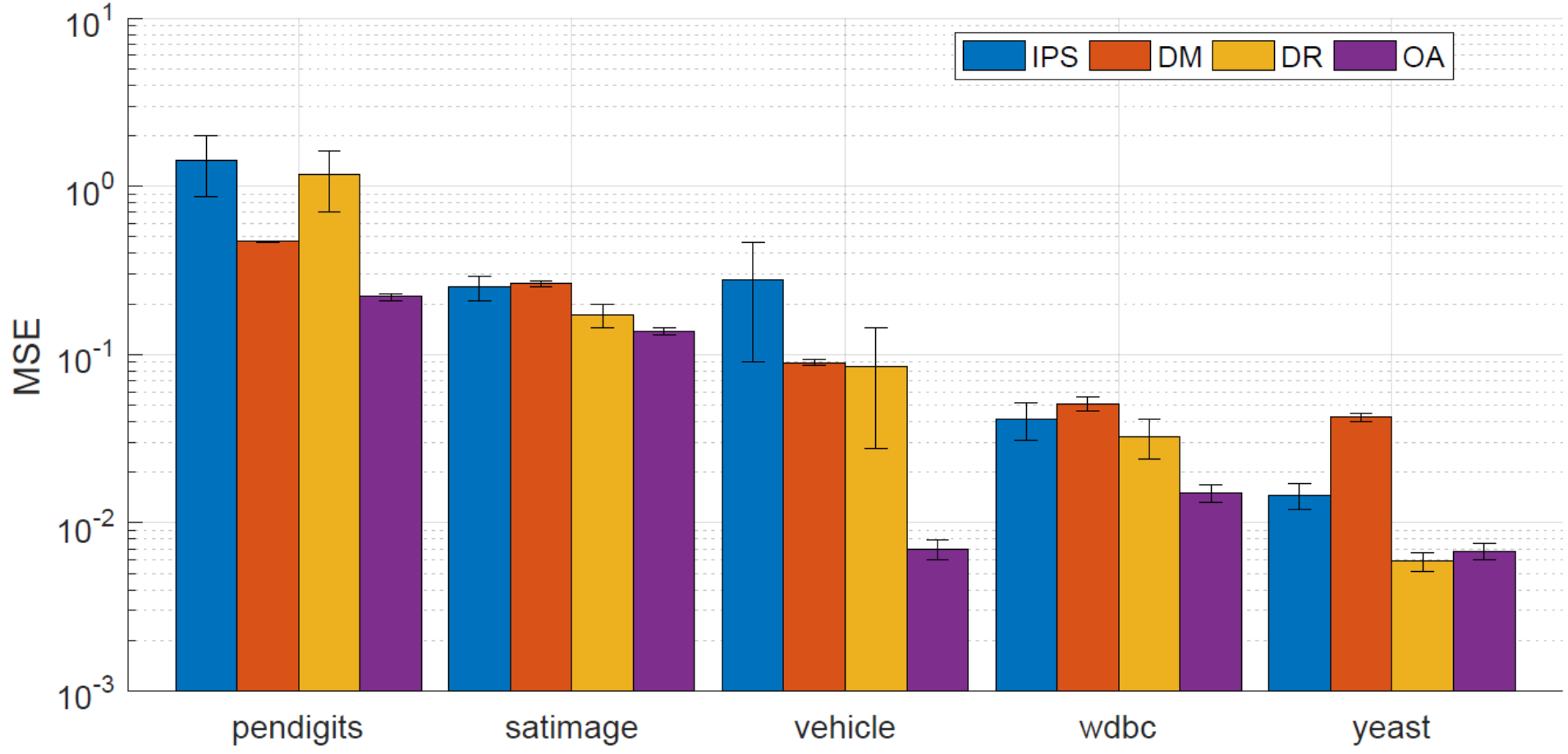
else:

use the ORACLE estimator

Oracle-Assisted estimator

- For appropriately chosen τ
  - **Matches the lower bound when oracle is perfect.**
  - **Minimax when oracle is horrible.**
  - **Robust to large importance weight.**
- Auto-tune parameter with bias/variance upper bounds.

Experiments (UCI multiclass classification => Contextual Bandit)



Conclusion

- IPS is optimal.
  - Need to go beyond the model-free approach.
- DR is unsatisfactory.
- We propose an estimator
  - That is great in theory
  - Perform quite well in practice.

Thank you! Any questions?

The limit of the model-free off-policy evaluation

- The class of problems:

$$\left\{ r|x,a \;\middle|\; \begin{array}{l} 0 \le \mathbb{E}(r|x,a) \le R_{\max} \\ 0 \le \mathrm{Var}(r|x,a) \le \sigma^2 \end{array} \right\}$$

- Our main theorem: the minimax risk

$$\Theta\left[\frac{1}{n}\left(\mathbb{E}_\mu \rho^2 \sigma^2 + \mathbb{E}_\mu \rho^2 R_{\max}^2\right)\right]$$

  - IPS is optimal! High variance is required…
  - Very different from multi-arm bandit! (Li et. al., 2015)

Error bounds for OA

$$\mathrm{MSE}(\hat{v}_{\mathrm{OA}}) \leq \underbrace{\frac{\mathbb{E}R_{\max}^2}{2n}}_{(1)} + \underbrace{\frac{2}{n}\mathbb{E}_\mu \sigma^2 \rho^2 1_{\{a \in A_x\}} + \frac{1}{2n}\mathbb{E}_\mu R_{\max}^2 \rho^2 1_{\{a \in A_x\}}}_{(2)} + \underbrace{|\mathbb{E}_{\pi'} \epsilon(a,x)|^2 \, \mathbb{P}_\pi(a \in A_x^c)^2}_{(3)}.$$

1) Required even with optimal oracle
2) Reduced IS variance. (from IPS or DR)
3) Bias from ORACLE.

Asymptotically, we should be able to improve this by a constant factor of 2.

Conservative autotuning

- Minimize estimated theoretical upper bounds

$$\begin{aligned}\mathrm{Var}(\hat{v}_{\mathrm{OA}}) &\leq \frac{R_{\max}^2}{2n} + \frac{R_{\max}^2/2 + 2\sigma^2}{n}\mathbb{E}_\mu \rho^2 1_{\{a \in A_x\}} \\ &\approx \frac{R_{\max}^2}{2n} + \frac{R_{\max}^2/2 + 2\sigma^2}{n^2}\sum_{i=1}^{n}\sum_{a \in A_x}\frac{\pi^2(x_i,a)}{\mu(x_i,a)}1_{\{a \in A_{x_i}\}}.\end{aligned}$$

$$\mathrm{Bias}^2(\hat{v}_{\mathrm{OA}}) \leq R_{\max}^2 \mathbb{P}_\pi(a \in A_x^c)^2 \approx R_{\max}^2 \left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_\pi(a \in A_{x_i}^c)\right]^2.$$

Oracle-Assisted estimator

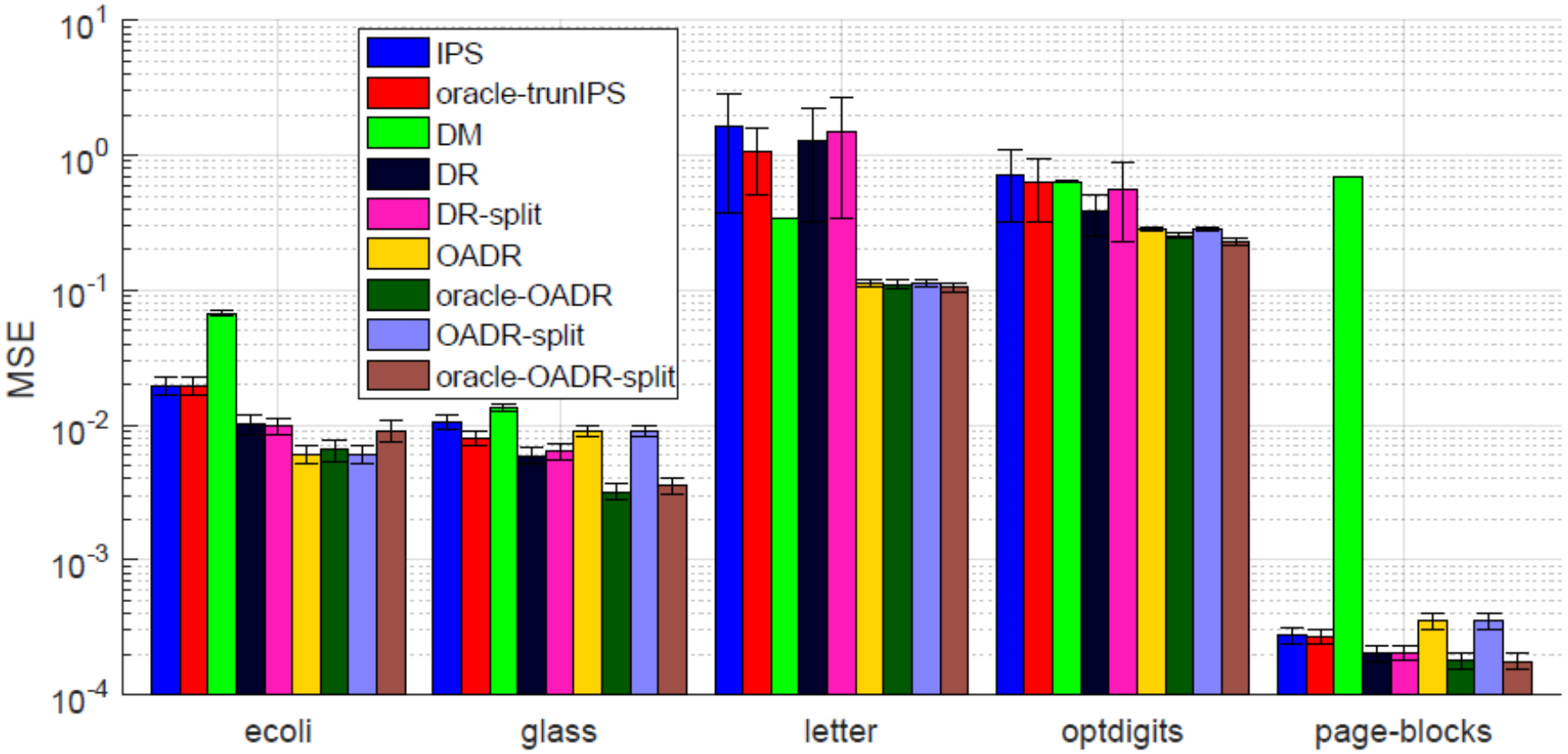
- Divide and conquer

$$v^\pi = \underbrace{\mathbb{E}_\pi r 1_{\{\rho \le \tau\}}}_{\text{Easy part}} + \underbrace{\mathbb{E}_\pi r 1_{\{\rho > \tau\}}}_{\text{Hard part of the problem}}$$

$$\hat{v}_{\mathrm{OA}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[r_i \rho_i 1_{\{a_i \in A_i\}}\right]}_{\text{Use IPS or DR}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{a \in A_i^c} \hat{r}(x_i, a)\pi(a|x_i)}_{\text{Use the oracle}}.$$

- **Matches the lower bound when oracle is perfect.**
- **Minimax when oracle is horrible.**
- **Robust to large weight.**

More baselines



More baselines