CAB: Continuous Adaptive Blending for Policy Evaluation and Learning

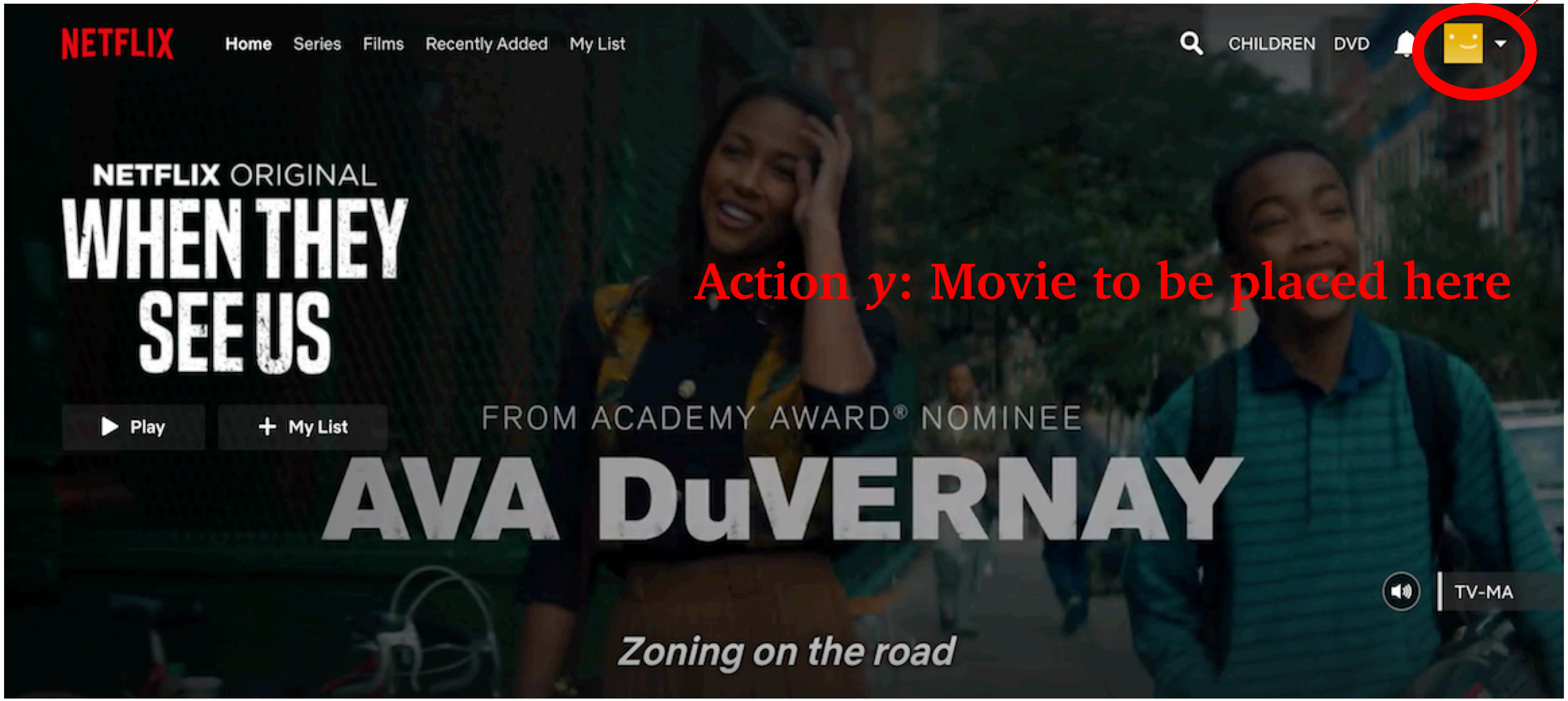
Yi Su*, Lequn Wang*, Michele Santacatterina and Thorsten Joachims

Cornell University

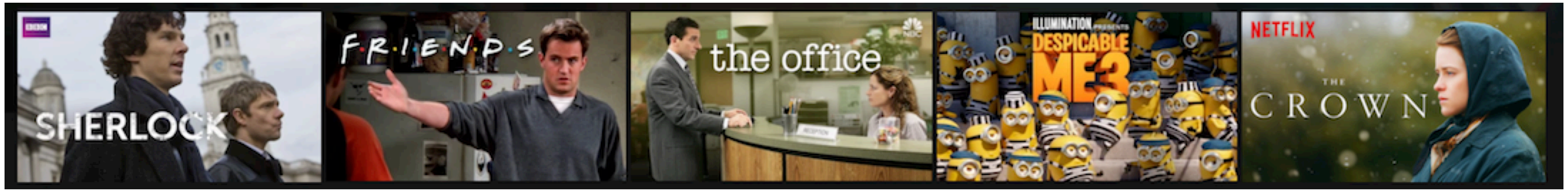
Cornell CIS
COMPUTING AND INFORMATION SCIENCE

Example: Netflix

Context x: User/History



Candidate:



Reward r: Whether user will click it

Goal: Off-Policy Evaluation and Learning

Evaluation: Expected performance for a new policy π

Online: A/B Testing

Draw S_1 from π_1 → $\hat{R}(\pi_1)$

Draw S_2 from π_2 → $\hat{R}(\pi_2)$

Draw S_3 from π_3 → $\hat{R}(\pi_3)$

Draw S_4 from π_4 → $\hat{R}(\pi_4)$

Draw S_5 from π_5 → $\hat{R}(\pi_5)$

Draw S_6 from π_6 → $\hat{R}(\pi_6)$

Offline: Off-policy evaluation

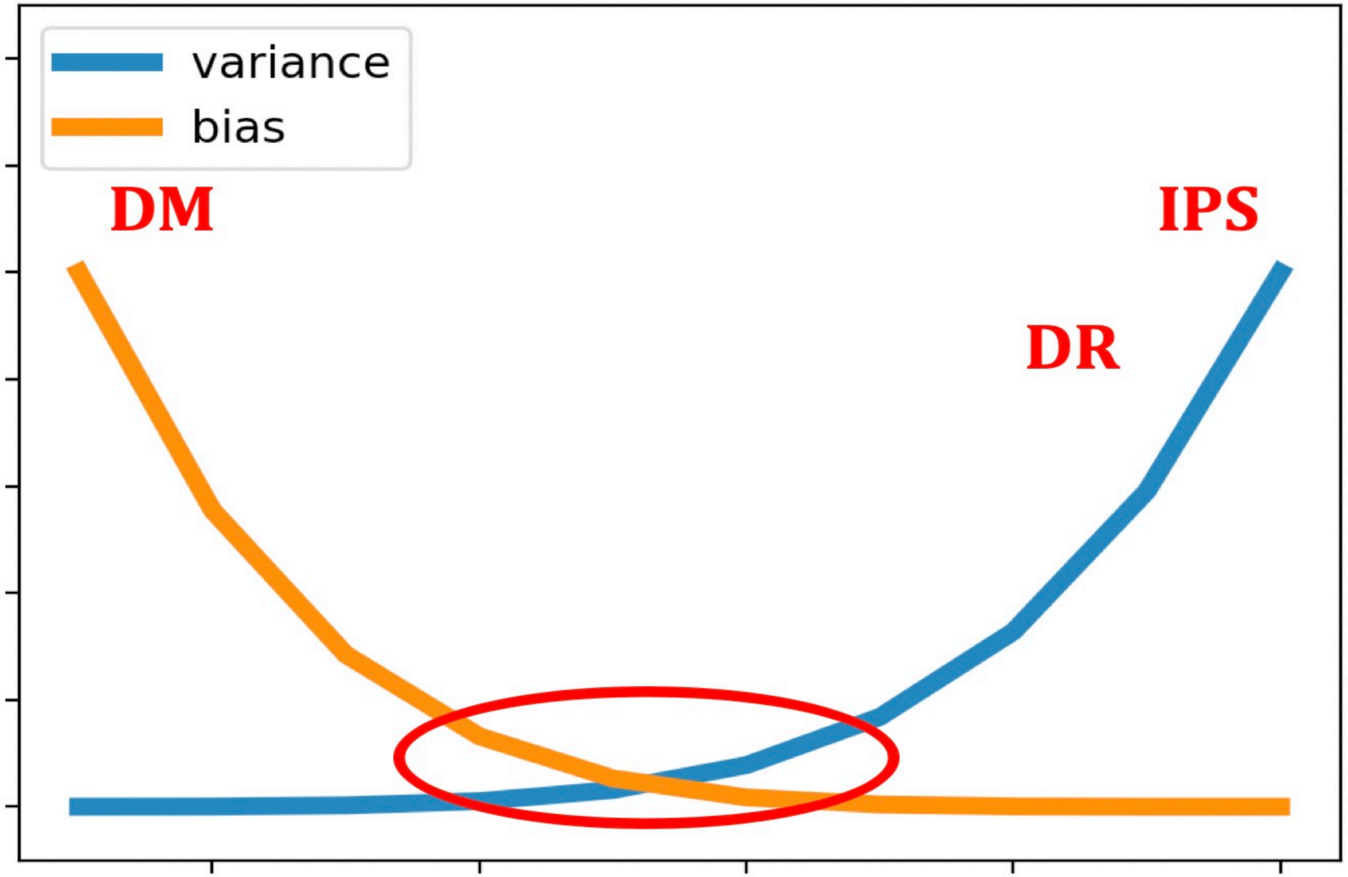
Draw S from π_0 → $\hat{R}(\pi_6)$ $\hat{R}(\pi_{12})$ $\hat{R}(\pi_{18})$

$S = \{x_i, y_i, r_i, \pi_0(y_i|x_i)\}_{i=1}^{n}$

Learning: ERM for batch learning from bandit feedback

$$\widehat{\pi^*} = argmax_{\pi \in \Pi} \hat{R}(\pi)$$

Main Approaches



Contribution I: Present a family of counterfactual estimators.

Contribution II: Design a new estimator that inherits desirable properties.

Contribution I: Interpolated Counterfactual Estimator Family

Notation: $\hat{\delta}(x, y)$ be the estimated reward for action y given context x. Let $\hat{\pi}_0$ be the estimated (known) logging policy.

Interpolated Counterfactual Estimator (ICE) Family

Given a triplet $\mathcal{W} = (w^\alpha, w^\beta, w^\gamma)$ of weighting functions:

$$\hat{R}^w(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} \pi(y|x_i)\, w_{iy}^{\alpha}\alpha_{iy} + \frac{1}{n}\sum_{i=1}^{n} \pi(y_i|x_i) w_i^{\beta}\beta_i + \frac{1}{n}\sum_{i=1}^{n} \pi(y_i|x_i) w_i^{\gamma}\gamma_i$$

Model the world
$\alpha_{iy} = \hat{\delta}(x_i, y)$
High bias, small variance

Model the bias
$\beta_i = r(x_i, y_i)/\widehat{\pi_0}(y_i|x_i)$
High variance, can be unbiased with known propensity

Control variate
$\gamma_i = \hat{\delta}(x_i, y_i)/\widehat{\pi_0}(y_i|x_i)$
Variance reduction, prohibited use in LTR

Contribution II: Continuous Adaptive Blending (CAB) Estimator

$$\hat{R}_{CAB}(\pi) = \hat{\mathrm{R}}^{\mathbf{w}}(\pi) \text{ with } \begin{cases} \mathrm{w}_{\mathrm{i}\bar{\mathrm{y}}}^{\alpha} = 1 - \min\left\{M \frac{\pi_0(\bar{y}|x_i)}{\pi(\bar{y}|x_i)}, 1\right\} \\ \mathrm{w}_{\mathrm{i}}^{\beta} = \min\left\{M \frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\right\} \\ \mathrm{w}_{\mathrm{i}}^{\gamma} = 0 \end{cases}$$

- ✓ Can be sustainably less biased than clipped IPS and DM.
- ✓ While having low variance compared to IPS and DR.
- ✓ Subdifferentiable and capable of gradient based learning: POEM (Swaminathan & Joachims, 2015a), BanditNet (Joachims et.al., 2018)
- ✓ Unlike DR, can be used in off-policy Learning to Rank (LTR) algorithms. (Joachims et.al., 2017)

See our poster at **Pacific Ballroom #221**
Thursday (Today) 6:30-9:00pm