CAB: Continuous Adaptive Blending for Policy Evaluation and Learning

Yi Su*, Lequn Wang*, Michele Santacatterina and Thorsten Joachims
Example: Netflix

Action $y$: Movie to be placed here

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 $\pi$

Online: A/B Testing

- Draw $S_1$ from $\pi_1 \rightarrow \hat{R}(\pi_1)$
- Draw $S_2$ from $\pi_2 \rightarrow \hat{R}(\pi_2)$
- Draw $S_3$ from $\pi_3 \rightarrow \hat{R}(\pi_3)$
- Draw $S_4$ from $\pi_4 \rightarrow \hat{R}(\pi_4)$
- Draw $S_5$ from $\pi_5 \rightarrow \hat{R}(\pi_5)$
- Draw $S_6$ from $\pi_6 \rightarrow \hat{R}(\pi_6)$

Offline: Off-policy evaluation

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

Learning: ERM for batch learning from bandit feedback

$$\hat{\pi}^* = \arg\max_{\pi \in \Pi} \hat{R}(\pi)$$
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:

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

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

- **Model the bias**
  - $\beta_i = r(x_i, y_i)/\hat{\pi}_0(y_i|x_i)$
  - High variance, can be unbiased with known propensity

- **Control variate**
  - $\gamma_i = \hat{\delta}(x_i, y_i)/\hat{\pi}_0(y_i|x_i)$
  - Variance reduction, prohibited use in LTR
Contribution II: Continuous Adaptive Blending (CAB) Estimator

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

- 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