CAB: Continuous Adaptive Blending for Policy Evaluation and Learning

Yi Su*, Lequn Wang*, Michele Santacatterina, Thorsten Joachims *equal contribution

Cornell University

ICML | 2019
Thirty-sixth International Conference on Machine Learning

Motivation

Contextual Bandit Protocol

- Context x drawn *i.i.d* from $P(\mathcal{X})$
- Given logging policy π_0, for a particular context x, action y is drawn from $\pi_0(y|x)$.
- System receives feedback $r(x,y)$ with mean $\delta(x,y) = \mathbb{E}_r[r(x,y)]$.
- Logged data is in the format

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

Challenges:

- partial information (only see the feedback for the chosen action)
- biased feedback (by the choices of the policy that logged the data)

Applications: search engines, recommender systems, persenalized medicine, ad-placement system

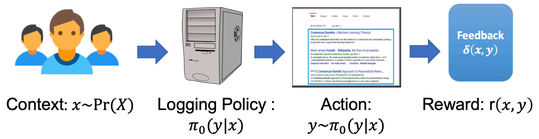

Figure 1:Contextual Bandit Protocol

Goal

- **Off policy evaluation**: using $\mathcal{S}$ to evaluate the performance of a new policy π:

$$R(\pi) = \mathbb{E}_{x\sim P(\mathcal{X})}\mathbb{E}_{y\sim\pi(\mathcal{Y}|x)}\mathbb{E}_r[r(x,y)] \tag{2}$$

- **Off policy learning**: finding an optimal policy π^* that maximizes the expected reward:

$$\pi^* = \underset{\pi\in\Pi}{\operatorname{argmax}}[R(\pi)] \tag{3}$$

In practice, $R(\pi)$ is unknown, ERM for Batch Learning from Bandit Feedback (BLBF) gives us:

$$\hat{\pi}^* = \underset{\pi\in\Pi}{\operatorname{argmax}}[\hat{R}(\pi)] \tag{4}$$

Interpolated Counterfactual Estimator (ICE) Family

Notation. Let $\hat{\delta}(x,y)$ be the estimated reward and let $\hat{\pi}_0$ be the estimated logging policy. Let $\zeta(x,y) := 1 - \frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}$ be the multiplicative deviation of the propensity estimates, and $\Delta(x,y) = \hat{\delta}(x,y) - \delta(x,y)$ be the additive deviation of reward model.

ICE Family
$\hat{R}^{\mathbf{w}}(\pi) = \mathbf{DM} + \mathbf{IPS} + \mathbf{CV}$

Given a triplet $\mathbf{w} = (w^\alpha, w^\beta, w^\gamma)$ of weighting function:

$$\hat{R}^{\mathbf{w}}(\pi) = \frac{1}{n}\sum_{i=1}^n \Big\{ \sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x_i) w^\alpha_{i\bar{y}} \alpha_{i\bar{y}} \\ \pi(y_i|x_i) w_i^\beta \beta_i + \pi(y_i|x_i) w_i^\gamma \gamma_i \Big\} \tag{5}$$

- **DM** (Direct Model): $\alpha_{i\bar{y}} := \alpha(x_i, \bar{y}) = \hat{\delta}(x_i, \bar{y})$. Typically enjoys low variance but can have high bias.
- **IPS** (Inverse Propensity Scoring): $\beta_i := \beta(x_i, y_i) = \frac{r(x_i,y_i)}{\hat{\pi}_0(y_i|x_i)}$. Unbiased if propensity is known, but have large variance due to policy mismatch.
- **CV** (Control Variate): $\gamma_i := \gamma(x_i, y_i) = \frac{\hat{\delta}(x_i,y_i)}{\hat{\pi}_0(y_i|x_i)}$. Used as control variate, like in Doubly Robust (DR) Estimator.

Bias of ICE. Under common support condition, the bias of ICE estimator with weighting function $\mathbf{w} = (w^\alpha, w^\beta, w^\gamma)$ is

$$\mathbb{E}_x\mathbb{E}_{y\sim\pi}[w^\alpha\Delta - w^\beta\zeta\delta + w^\gamma(\Delta - \zeta(\delta+\Delta)) \\ + (w^\alpha + w^\beta + w^\gamma)\delta - \delta] \tag{6}$$

Variance of ICE.

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[w^\alpha\Delta - w^\beta\zeta\delta + w^\gamma(\Delta - \zeta(\delta+\Delta)) \\ + (w^\alpha + w^\beta + w^\gamma)\delta]\Big) + \mathbb{E}_x\mathbb{E}_\pi\Big[(w^\beta)^2 c(1-\zeta)^2\sigma_r^2\Big] \\ + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(w^\beta c(1-\zeta)\delta + w^\gamma c(1-\zeta)(\delta+\Delta))\Big]\Big\} \tag{7}$$

Continuous Adaptive Blending

Desirable Properties.

- Unbiased under correct model for reward and propensity.
- Applicable to a broad range of partial information setting, LTR requires the γ_i term to be 0.
- Achieve low MSE, with bias-variance trade-off.
- Could be used in gradient-based learning, the weighting function needs to be sub-differentiable.

CAB Estimator

$\hat{R}_{CAB}(\pi) = \hat{R}^{\mathbf{w}}(\pi)$ with weighting function:

$$\begin{cases} w^\alpha_{i\bar{y}} = 1 - \min\Big\{M\frac{\pi_0(|x_i)}{\pi(|x_i)}, 1\Big\} \\ w_i^\beta = \min\Big\{M\frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\Big\} \\ w_i^\gamma = 0 \end{cases}$$

Bias improvements over cIPS and DM. under logged propensities:

- CAB: $\mathbf{B}(\hat{R}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[\Delta(1 - \frac{M}{c})\mathbb{1}\{c > M\}]$
- cIPS: $\mathbf{B}(\hat{R}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[-\delta(1 - \frac{M}{c})\mathbb{1}\{c > M\}]$
- DM: $\mathbf{B}(\hat{R}(\pi)) = \mathbb{E}_x\mathbb{E}_\pi[\Delta]$

Variance improvements over IPS and DR. under logged propensities:

- CAB:

$$\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta + \Delta(1 - \frac{M}{c})\mathbb{1}\{c > M\}]\Big) \\ + \mathbb{E}_x\mathbb{E}_\pi\Big[c\sigma_r^2\mathbb{1}\{c \le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{c > M\}\Big] \\ + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}[(c\delta)\mathbb{1}\{c \le M\} + M\delta\mathbb{1}\{c > M\}]\Big]\Big\}. \tag{8}$$

- DR: $\frac{1}{n}\{\mathbb{V}_x(\mathbb{E}_\pi[\delta]) + \mathbb{E}_x\mathbb{E}_\pi[c\sigma_r^2] + \mathbb{E}_x[\mathbb{V}_{\pi_0}(c\Delta)]\}$
- IPS: $\frac{1}{n}\{\mathbb{V}_x(\mathbb{E}_\pi[\delta]) + \mathbb{E}_x\mathbb{E}_\pi[c\sigma_r^2] + \mathbb{E}_x[\mathbb{V}_{\pi_0}(c\delta)]\}$

Extension to CAB-DR. CAB-DR inherits the weight design from CAB, with the new weighting function defined by: $w^\alpha_{i\bar{y}} = 1, w_i^\beta = \min\{M\frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\}$ and $w_i^\gamma = -\min\{M\frac{\pi_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\}$.

Experiments

BLBF: UCI multi-class datasets. Logging Policy: Logistic regression. Reward Model: Logistic Regression.

Ranking with biased user feedback: Yahoo! LTR. Logging Policy: SVM-rank. Reward Model: Gradient Boosted Tree.

Real-world dataset: Amazon Music contextual bandit problem. Logging Policy: Thompson Sampling Contextual Bandit Algorithm. Reward model: Model learned by Thompson Sampler.

CAB MSE curve with hyper-parameter M

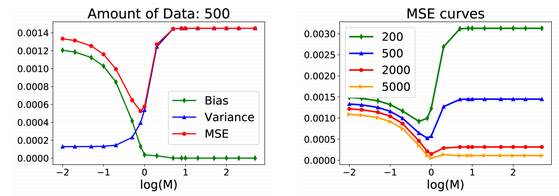

Figure 2:MSE curve on Satimage

CAB Evaluation Performance

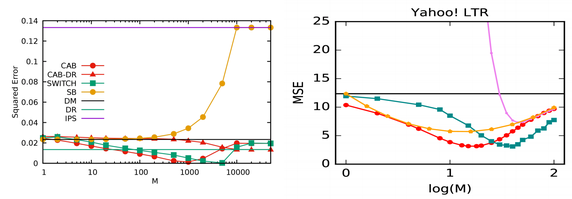

Figure 3:Error of estimates on the Amazon Contextual Bandit Problem (left) and MSE on Yahoo!LTR (right).

CAB Learning Performance

DATA	LETTER	OPTDIGITS	SATIMAGE	PENDIGITS	LTR
DM	0.6372	0.0649	0.3083	0.1133	11.25
DR	0.6852	0.0471	0.2762	0.1191	-
IPS	0.8969	0.0695	0.3266	0.2748	11.17
cIPS	0.8504	0.0447	**0.2415**	0.1228	11.39
SB	0.6091	0.0460	0.2481	0.0949	10.98
SWITCH	-	-	-	-	10.95
CAB	**0.5740**	**0.0445**	0.2442	**0.0917**	**10.83**
CAB-DR	0.5877	0.0461	0.2762	0.0946	-

Acknowledgements

This research was supported in part by NSF Awards IIS-1615706 and IIS-1513692, as well as a gift from Amazon.