Bayesian Counterfactual Risk Minimization

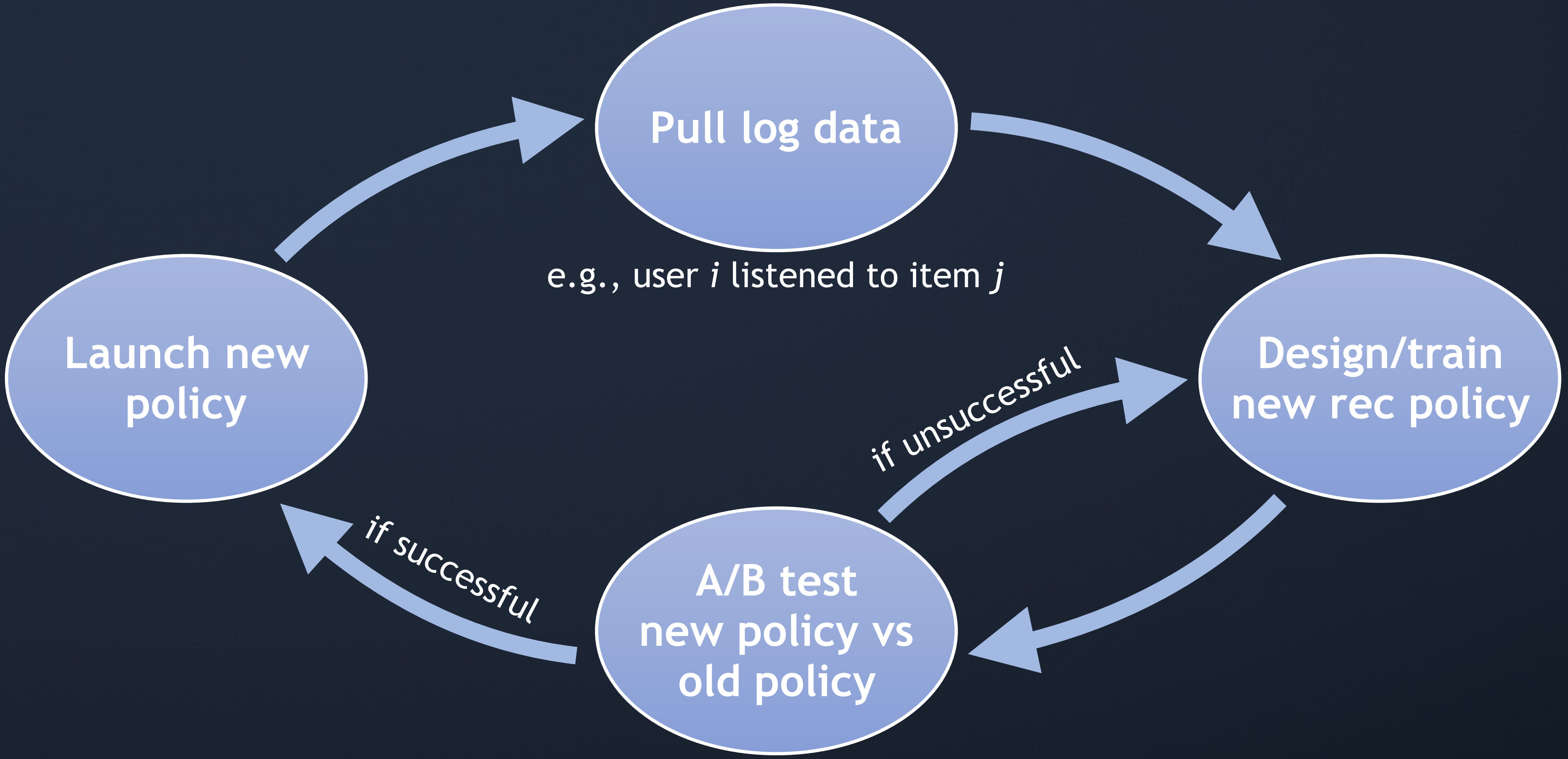
Ben London (blondon@) Amazon Music

Ted Sandler (sandler@) Amazon Music

International Conference on Machine Learning
Long Beach, CA, June 11, 2019

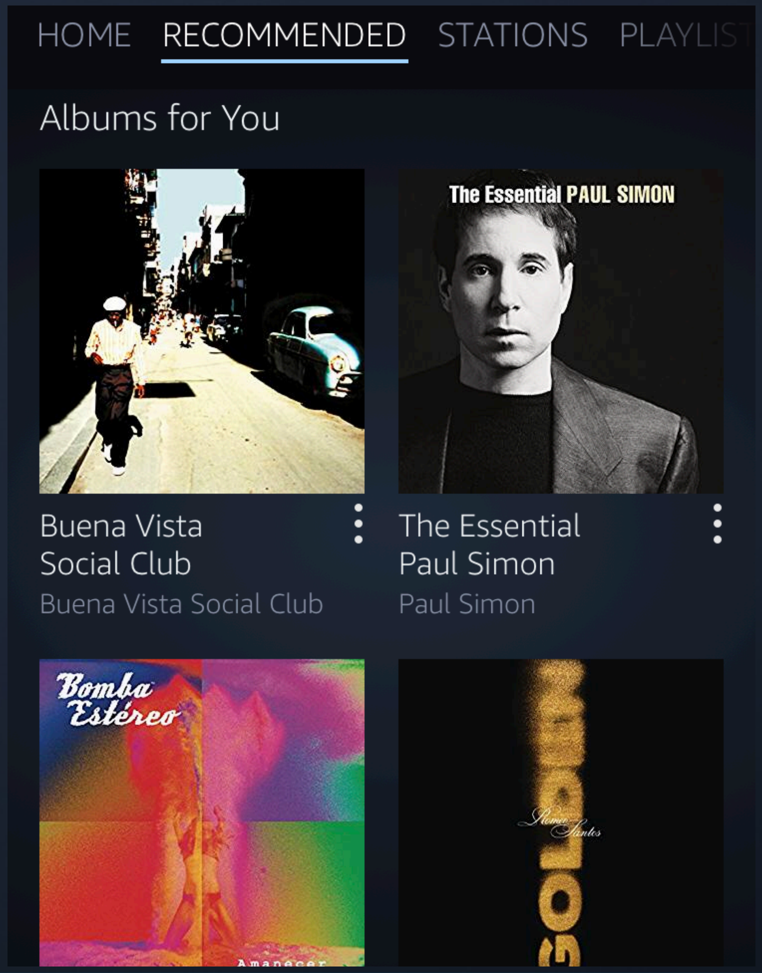
music

Learning from Logged Data



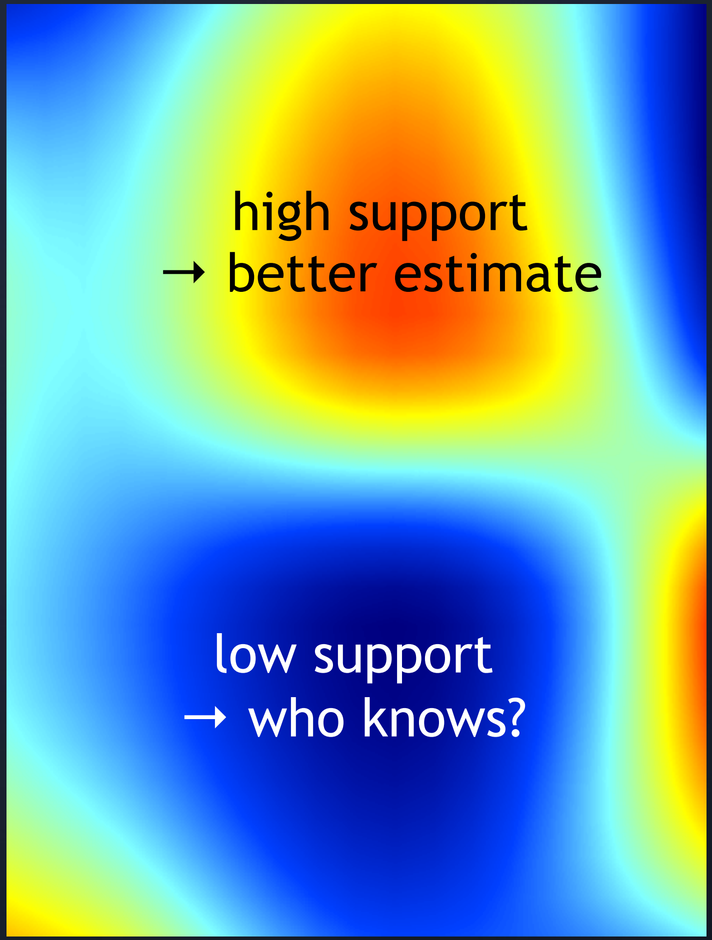
Problem 1: Bandit Feedback

- Only observe outcomes from actions taken
  - e.g., only get feedback on recommendations



Problem 2: Bias

- Logged data is *biased*
  - Policy typically not uniform distribution
  - User typically doesn't see everything
- Bias affects inferences
  - Self-fulfilling prophecies; "rich get richer"
  - Miss key insights due to insufficient support



music

IPS Policy Optimization

- Use ***inverse propensity score*** (IPS) estimator

$$\arg\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} -r_i \frac{\pi(a_i \mid x_i)}{p_i}$$

logged propensity $p_i = \pi_0(a_i \mid x_i)$

- IPS is an unbiased estimator of expected reward

$$\mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi(x)} [\rho(x,a)] \approx \frac{1}{n} \sum_{i=1}^{n} r_i \frac{\pi(a_i \mid x_i)}{p_i}$$

  - Caveat: logging policy must have full support

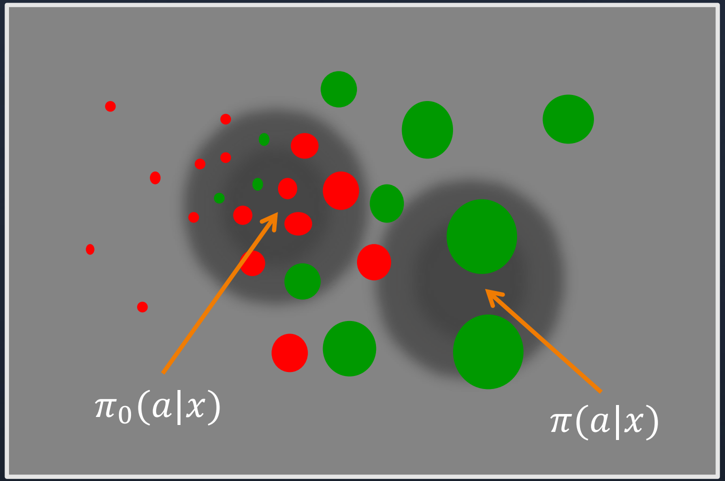
IPS Policy Optimization

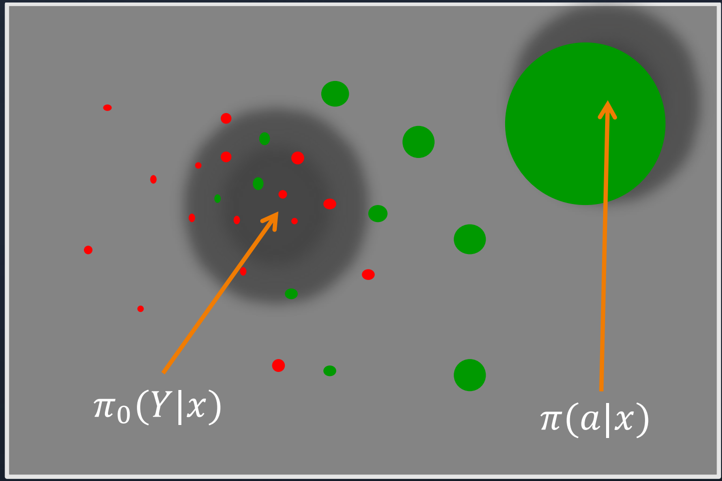
- Use ***inverse propensity score*** (IPS) estimator

$$\arg\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} -r_i \frac{\pi(a_i \,|\, x_i)}{p_i}$$

logged propensity $p_i = \pi_0(a_i \,|\, x_i)$

- Problem: IPS has ***high variance***





CRM Principle

- *Counterfactual Risk Minimization* (CRM) principle

$$\arg\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} -r_i \frac{\pi(a_i \mid x_i)}{p_i} + \lambda \sqrt{\hat{\mathrm{Var}}(\pi, S)}$$

variance regularization

- Motivated by PAC risk analysis
- Stochastic optimization of variance regularizer is tricky
  - *Policy optimization for exponential models* (POEM) algorithm

[Swaminathan & Joachims, ICML 2015]

Bayesian CRM Principle

- ***Bayesian Counterfactual Risk Minimization*** (CRM) principle

$$\arg\min_{\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} -r_i \frac{\pi_{\mathbb{Q}}(a_i \,|\, x_i)}{p_i} \quad + \quad \lambda\, D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P})$$

KL div. from prior to posterior

- Bayesian policy: $\pi_{\mathbb{Q}}(a \,|\, x) = \Pr_{h \sim \mathbb{Q}}\{h(x) = a\}, \quad h : \mathcal{X} \to \mathcal{A}$
- Motivated by PAC-Bayes risk analysis
- Takeaway: posterior should stay close to the prior
  - What should the prior be? ***How about the logging policy!***

[London & Sandler, ICML 2019]

Application to Mixed Logits

- ***Mixed logit model***

$$h_{w,\gamma}(x) = \arg\max_a \; w \cdot \phi(x,a) + \gamma_a$$

$$w \sim \mathcal{N}(\mu, \Sigma) \qquad \gamma \sim \text{Gumbel}(0,1)^k$$

  - Like a softmax with normally distributed parameters

$$\pi_{\mathbb{Q}}(a \,|\, x) = \mathbb{E}_{(w,\gamma)\sim\mathbb{Q}}\left[\mathbb{1}\{h_{w,\gamma}(x) = a\}\right] = \mathbb{E}_w\left[\frac{\exp(w \cdot f(x,a))}{\sum_{a'} \exp(w \cdot f(x,a'))}\right]$$

- Risk bound motivates logging policy regularization

$$D_{\text{KL}}(\mathbb{Q} \| \mathbb{P}) = \text{O}(\|\mu - \mu_0\|^2)$$

L2 distance to logging policy parameters

[London & Sandler, ICML 2019]

music

Contributions

- PAC Risk bounds for Bayesian policies
- Application to mixed logits
- Logging policy prior motivates new regularizer
- Two new learning objectives (one convex) that minimize bounds
- Experiments show proposed methods gain up to 76% more reward than POEM, while also simpler/more efficient

Visit poster #113 in Pacific Ballroom