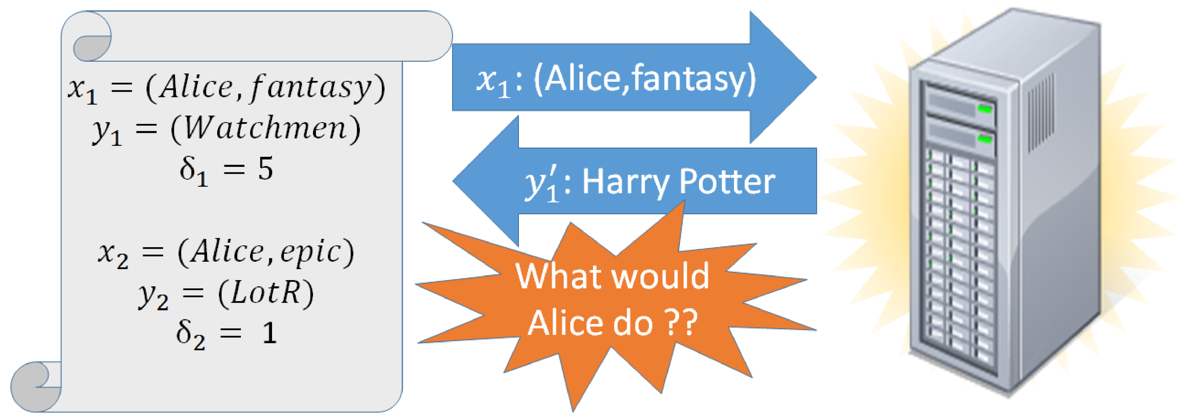
Counterfactual risk minimization: Learning from logged bandit feedback

Adith Swaminathan and Thorsten Joachims
Department of Computer Science, Cornell University

Aim: Offline learning for interactive systems

Can we re-use the interaction logs of deployed online systems (e.g. search engines, recommendation systems) to train better models offline?



Training using interaction logs is counter-factual [2].

- Logs are **biased** (actions favored by deployed system will be over-represented),
- and **incomplete** (no feedback for other plausible actions).

Our contribution

A learning principle — **C**ounterfactual **R**isk **M**inimization — and an efficient algorithm — **P**olicy **O**ptimizer for **E**xponential **M**odels — for this learning setting [1]. Our solution is to

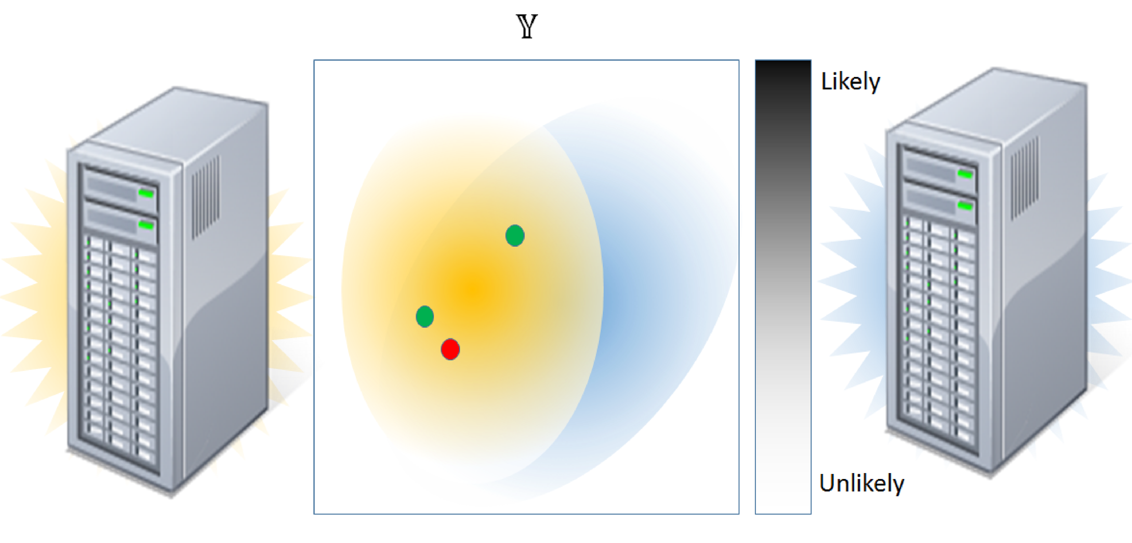
- predict by **sampling** and log **propensities**,
- use **counterfactual** risk estimators to fix bias,
- **regularize the variance**,
- and optimize a conservative bound using **majorization minimization**.

POEM

POEM is a simple, fast, stochastic optimizer for structured output prediction available at `http://www.cs.cornell.edu/~adith/poem`

It is as fast and expressive as Conditional Random Fields (CRFs), and trains using logged bandit feedback, without any supervised labels.

Counterfactual estimators



Learning from logged data without exploration is not possible.
Suppose the deployed system *sampled* $y \sim h_0(\mathcal{Y} \mid x)$.

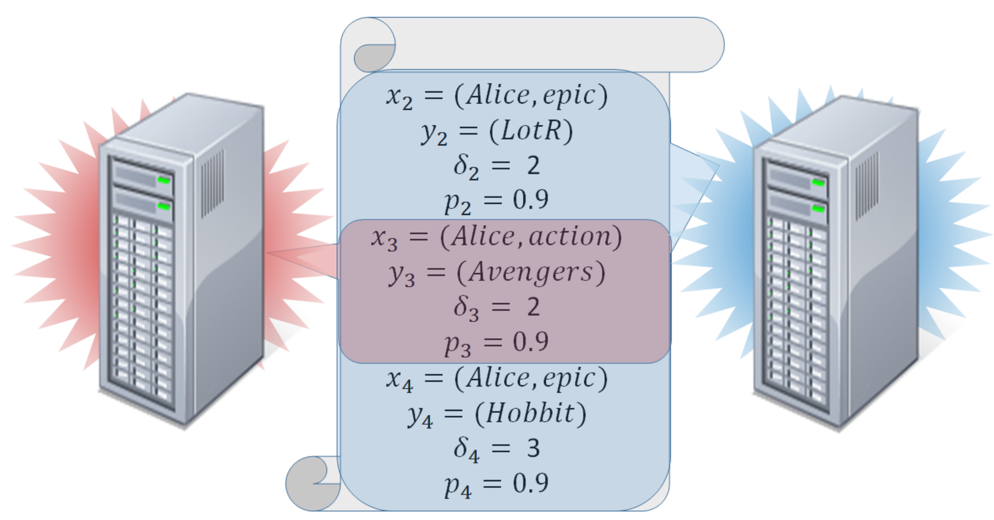
$$\underbrace{\mathbb{E}_x \mathbb{E}_{y\sim h(x)}\left[\delta(x,y)\right]}_{R(h),\text{ Risk of } h} = \underbrace{\mathbb{E}_x \mathbb{E}_{y\sim h_0(x)}}_{\text{Samples from deployed } h_0}\left[\delta(x,y)\underbrace{\frac{h(y\mid x)}{h_0(y\mid x)}}_{\text{Importance weight}}\right].$$

With $\mathcal{D} = \{(x_i, y_i, \delta_i, p_i)\}_{i=1}^n$, $p_i \equiv h_0(y_i \mid x_i)$,

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{h(y_i \mid x_i)}{p_i}.$$

This unbiased estimator has issues:

- Unbounded variance (think $p_i \simeq 0$).
- Degenerate minimizer (think $\delta_i \geq 0$).
- Importance sampling introduces variance.



Different effective sample sizes for different h!

Inverse propensity scoring,

$$h^{IPS} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{(\delta_i - \delta^{max}) \min\{M, \frac{h(y_i \mid x_i)}{p_i}\}}_{u_h^i}}_{\hat{R}^M(h)}.$$

fixes the first two issues. For the variance issue, we employ an empirical Bernstein argument [3].

Variance regularization

With high probability in $\mathcal{D} \sim h_0$, $\forall h \in \mathcal{H}$,

$$\underbrace{R(h)}_{\text{True risk}} \leq \underbrace{\hat{R}^M(h)}_{\text{Empirical risk}} + \mathcal{O}\Big(\underbrace{\sqrt{\frac{\hat{Var}(\boldsymbol{u}_h)}{n}}}_{\text{Variance control}}\Big) + M\cdot\mathcal{O}\Big(\underbrace{\frac{\mathcal{N}_\infty(\mathcal{H})}{n}}_{\text{Capacity control}}\Big).$$

Counterfactual risk minimization

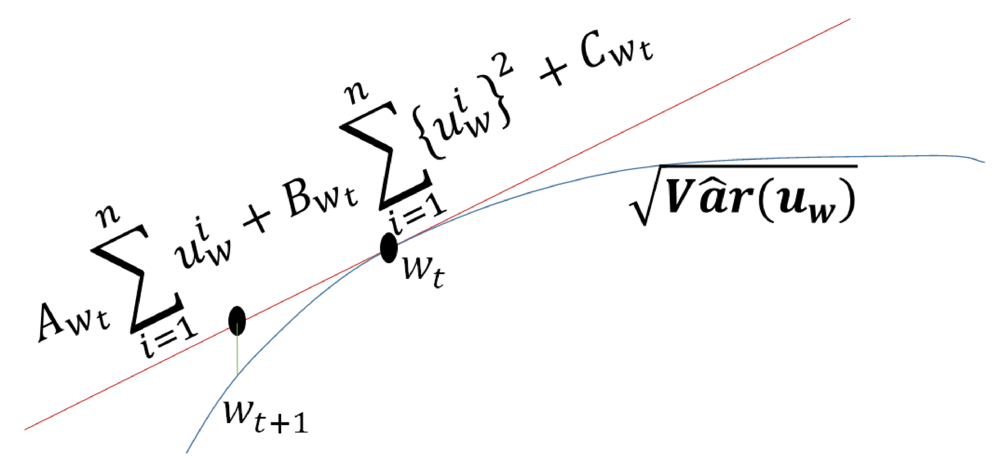
$$h^{CRM} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ \hat{R}^M(h) + \lambda\sqrt{\frac{\hat{Var}(h)}{n}}.$$

Deriving POEM from CRM

For CRFs, $h_w(y \mid x) \propto e^{w\cdot\phi(x,y)}$.

$$w^{CRM} = \underset{w}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} u_w^i + \lambda\sqrt{\frac{\hat{Var}(\boldsymbol{u}_w)}{n}} + \mu\|w\|^2.$$

To optimize at scale, we Taylor-approximate.



- Adagrad with $\nabla u_w^i\{1+\lambda\sqrt{n}(A_{w_t} + 2B_{w_t}u_w^i)\}$.
- After epoch, $w_{t+1} \leftarrow w$, compute $A_{w_{t+1}}$, $B_{w_{t+1}}$.

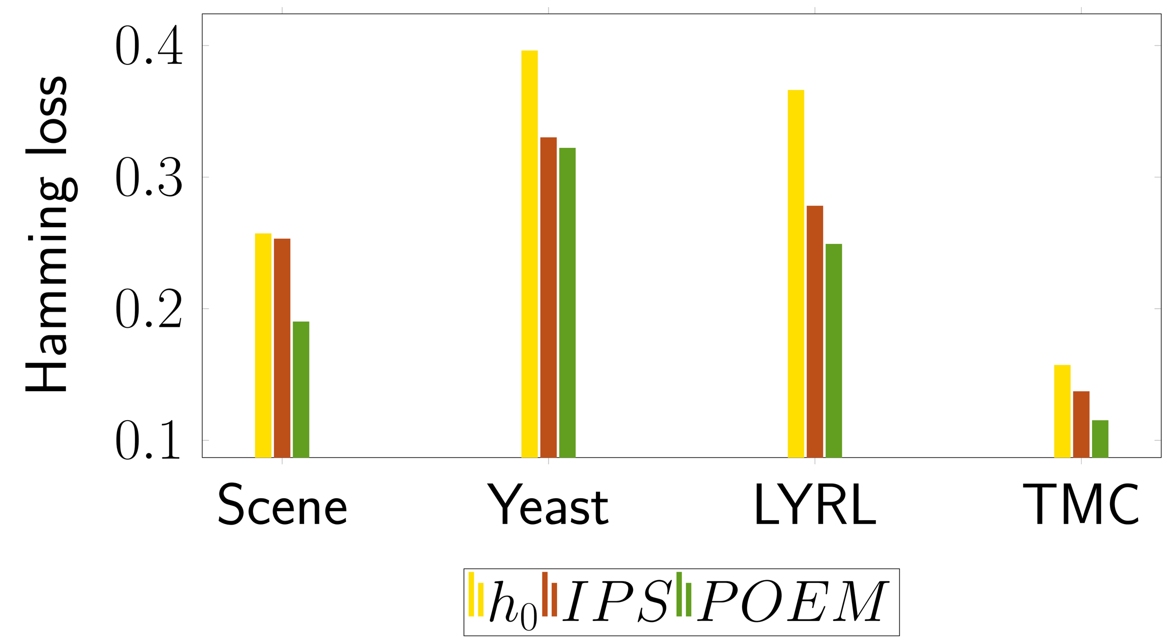
Experiments

Supervised $\mapsto$ Bandit Multi-Label classification with $\delta \equiv$ Hamming loss on four datasets.

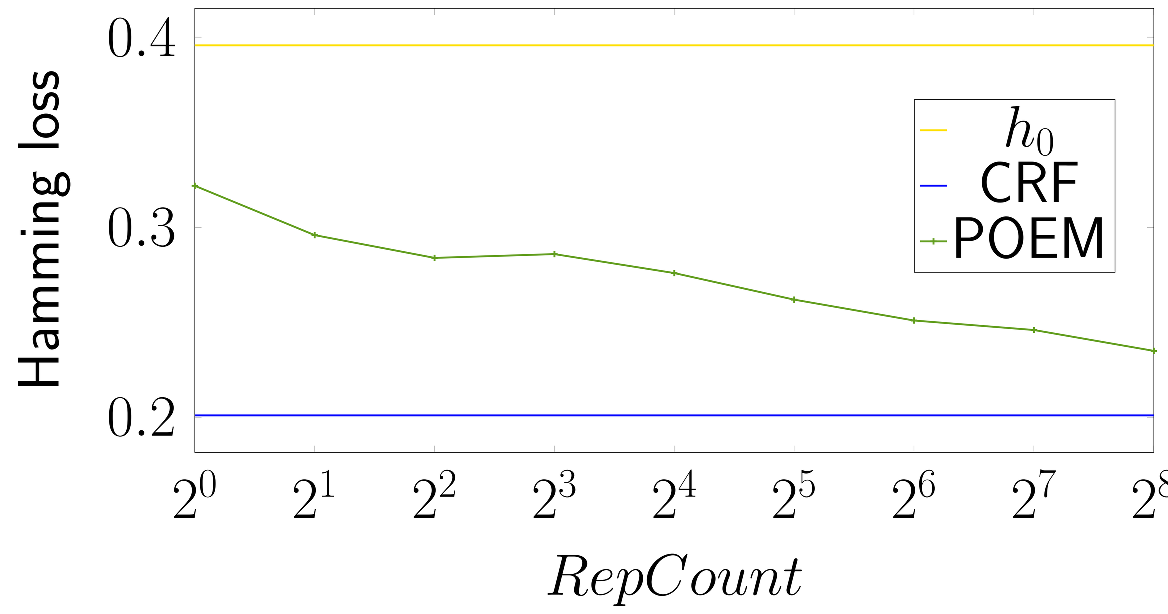
POEM is computationally efficient versus batch L-BFGS and compares favorably with CRF of *scikit-learn*.

Avg. Time (s)	Scene	Yeast	TMC	LYRL
POEM(L-BFGS)	75.20	94.16	949.95	561.12
POEM	4.71	5.02	276.13	120.09
CRF	4.86	3.28	99.18	62.93

POEM is statistically significantly better ($p = 0.05$) than IPS and h_0 (CRF trained on 5% of train set) on all datasets.



POEM recovers supervised performance as $n \to \infty$ (simulated by $\mathcal{D} \sim h_0$ *RepCount* many times on the *Yeast* dataset).



- Non-convex objective, but good local optima.
- Even with poor h_0, POEM achieves good loss.
- Sweet spot for stochasticity of h_0.
- MAP predictions from POEM work well.

References

[1] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. *ICML*, 2015.

[2] Léon Bottou, Jonas Peters, Joaquin Q. Candela, Denis X. Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[3] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. *COLT*, 2009.

Acknowledgment

This research was funded through NSF Award IIS- 1247637, IIS-1217686, JTCII Cornell-Technion Research Fund, and a gift from Bloomberg.

Cornell University