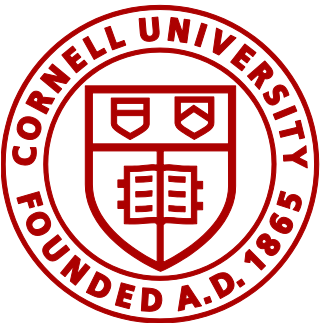


# Counterfactual Risk Minimization

## Learning from logged bandit feedback

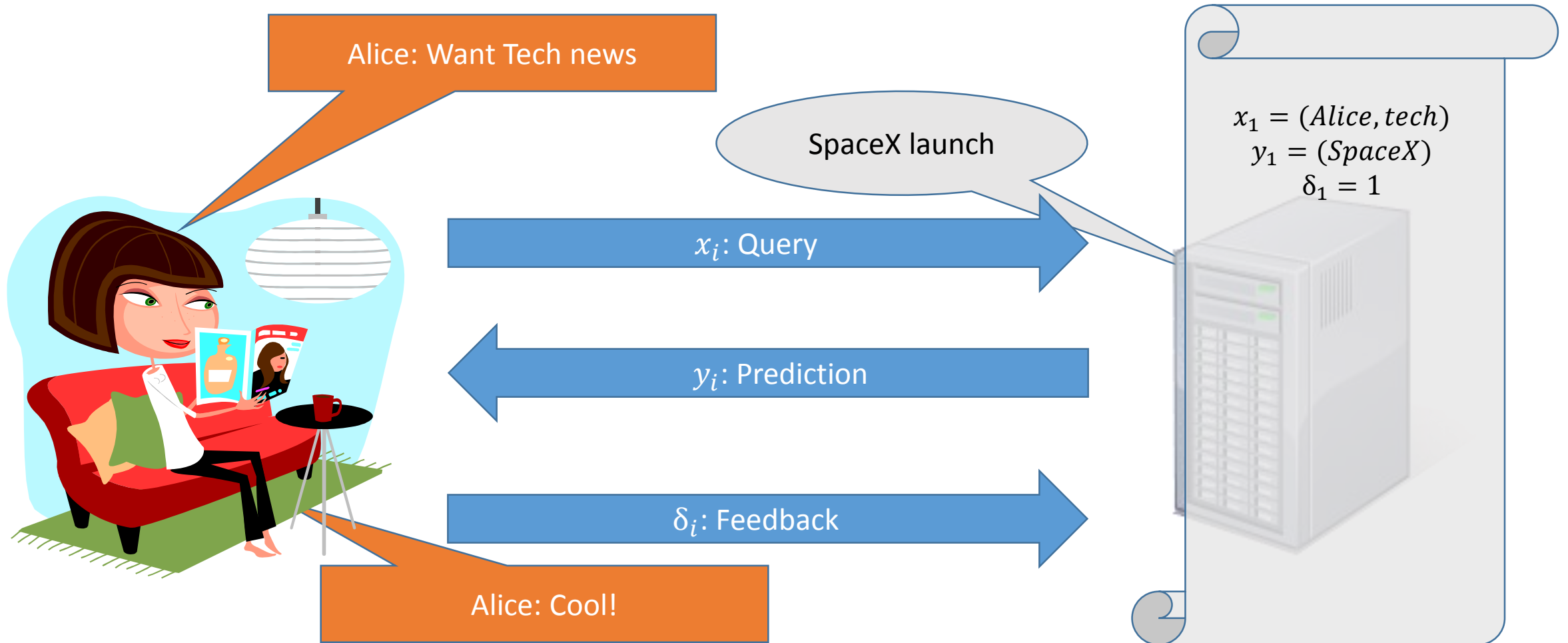
Adith Swaminathan, Thorsten Joachims



# Learning frameworks

$x$ $y$	Online	Batch
Full Information	Perceptron, ...	SVM, ...
Bandit Feedback	LinUCB, ...	?

# Logged bandit feedback is everywhere!

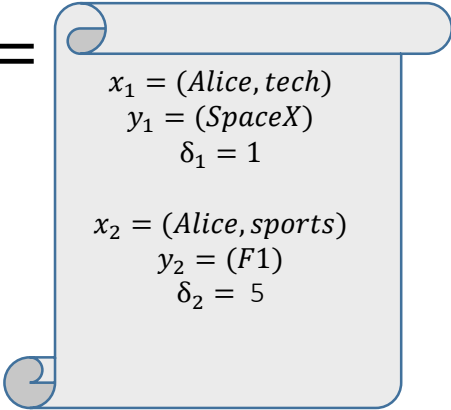


# Goal

- Risk of  $h: \mathbb{X} \mapsto \mathbb{Y}$

$$R(h) = \mathbb{E}_x[\delta(x, h(x))]$$

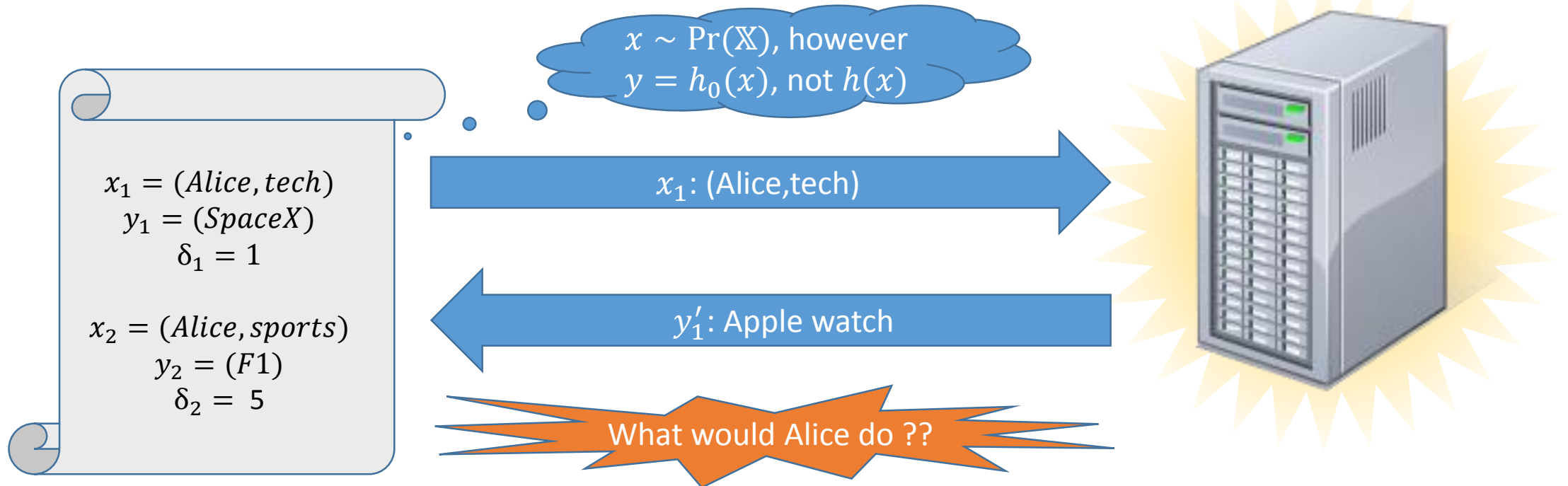
- Find  $h^* \in \mathcal{H}$  with minimum risk

- Can we find  $h^*$  using  $\mathcal{D} =$   collected from  $h_0$ ?

$x_1 = (\text{Alice}, \text{tech})$   
 $y_1 = (\text{SpaceX})$   
 $\delta_1 = 1$

$x_2 = (\text{Alice}, \text{sports})$   
 $y_2 = (\text{F1})$   
 $\delta_2 = 5$

# Learning by replaying logs?



- Training/evaluation from logged data is counter-factual [Bottou et al]

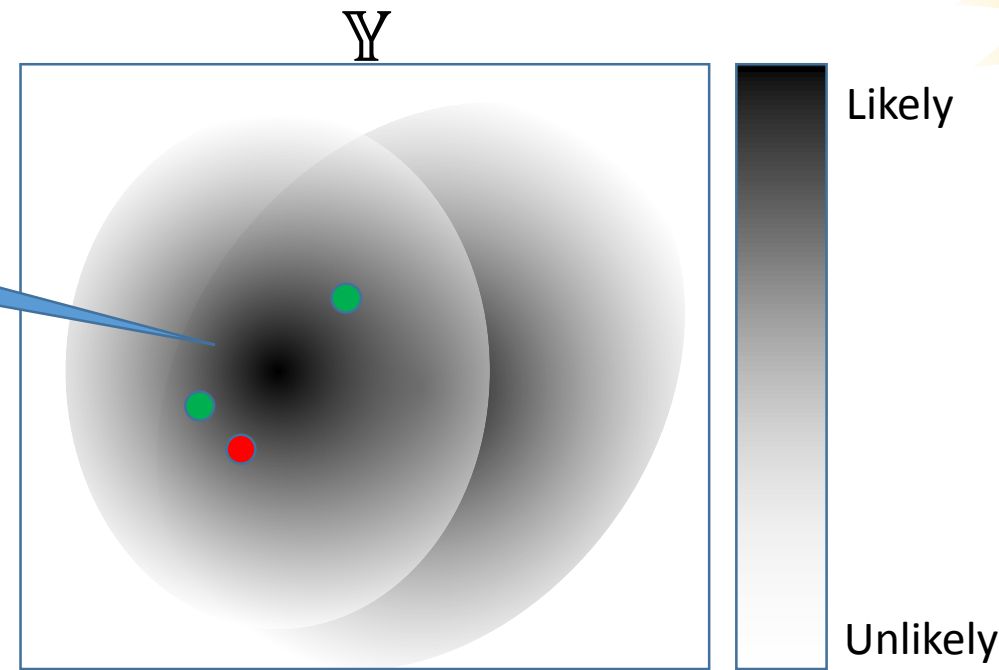
# Stochastic policies to the rescue!

- Stochastic policy:  $h: \mathbb{X} \mapsto \Delta(\mathbb{Y}), y \sim h(x)$   
 $R(h) = \mathbb{E}_x \mathbb{E}_{y \sim h(x)} [\delta(x, y)]$



$h_0$

SpaceX launch



$h$

# Counterfactual risk estimators

## Basic Importance Sampling [Owen]

$$\mathbb{E}_x \left[ \mathbb{E}_{y \sim h} [ \delta(x, y) ] \right] = \mathbb{E}_x \left[ \mathbb{E}_{y \sim h_0} \left[ \delta(x, y) \frac{h(y|x)}{h_0(y|x)} \right] \right]$$

Perf of new system

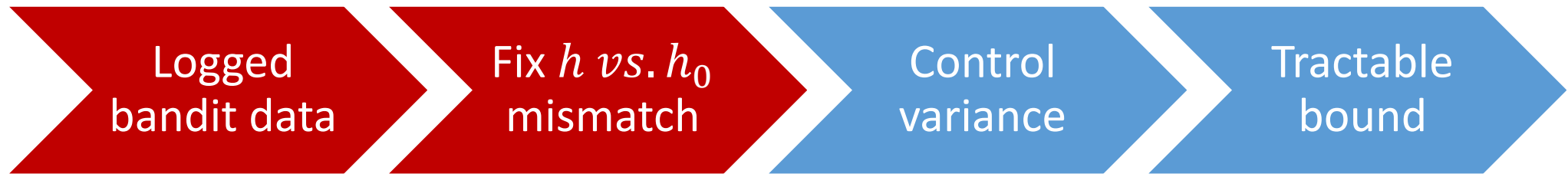
Samples from old system

Importance weight

- $\mathcal{D} = \{ (x_1, y_1, \delta_1, p_1), (x_2, y_2, \delta_2, p_2), \dots, (x_n, y_n, \delta_n, p_n) \}$
- $p_i = h_0(y_i|x_i)$  ... propensity [Rosenbaum et al]

$$\widehat{R}_{\mathcal{D}}(h) = \frac{1}{n} \sum_{i=1}^n \delta_i \frac{h(y_i|x_i)}{p_i}$$

# Story so far





# Importance sampling causes non-uniform variance!

$x_1 = (Alice, sports)$   
 $y_1 = (F1)$   
 $\delta_1 = 5$   
 $p_1 = 0.9$

Want: Error bound that captures  
variance of importance sampling

$x_3 = (Star, ...)$   
 $\delta_3 = 2$   
 $p_3 = 0.9$

$x_4 = (Alice, tech)$   
 $y_4 = (Tesla)$   
 $\delta_4 = 1$   
 $p_4 = 0.9$

$h_1$   
 $\widehat{R}_D(h_1) = 1$

$h_2$   
 $\widehat{R}_D(h_2) = 1.33$

# Counterfactual Risk Minimization

- W.h.p. in  $\mathcal{D} \sim h_0$

$$\forall h \in \mathcal{H}, \quad R(h) \leq \widehat{R}_{\mathcal{D}}(h) + O\left(\sqrt{\widehat{\text{Var}}_{\mathcal{D}}(h)/n}\right) + O(\mathcal{N}_{\infty}(\mathcal{H})/n)$$

\*conditions apply. Refer [Maurer et al]

Empirical risk

Variance regularization

Capacity control

Learning objective

$$h^{CRM} = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}_{\mathcal{D}}(h) + \lambda \sqrt{\frac{\widehat{\text{Var}}_{\mathcal{D}}(h)}{n}}$$

# POEM: CRM algorithm for structured prediction

- CRFs:  $h_w \in \mathcal{H}_{lin}$ ;  $h_w(y|x) = \frac{\exp(w\phi(x,y))}{Z(x;w)}$
- **P**olicy **O**ptimizer for **E**xponential **M**odels :

$$w^* = \operatorname{argmin}_w \left[ \frac{1}{n} \sum_{i=1}^n \delta_i \frac{h_w(y_i|x_i)}{p_i} + \lambda \sqrt{\frac{\widehat{Var}(h_w)}{n}} + \mu \|w\| \right]$$

Good: Gradient descent, search over infinitely many  $w$

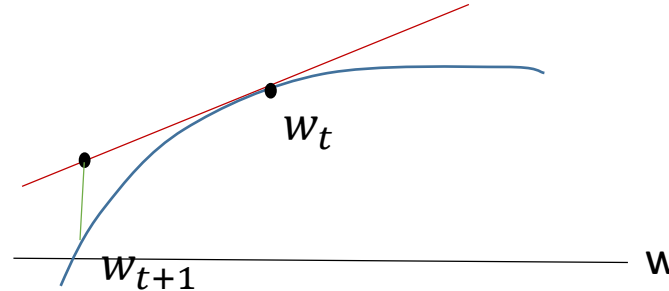
Bad: Not convex in  $w$

Ugly: Resists stochastic optimization

# Stochastically optimize $\sqrt{\widehat{Var}(\mathbf{h}_w)}$ ?

- Taylor-approximate!

$$\sqrt{\widehat{Var}(\mathbf{h}_w)} \leq A_{w_t} \sum_{i=1}^n h_w^i + B_{w_t} \sum_{i=1}^n \{h_w^i\}^2 + C_{w_t}$$



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

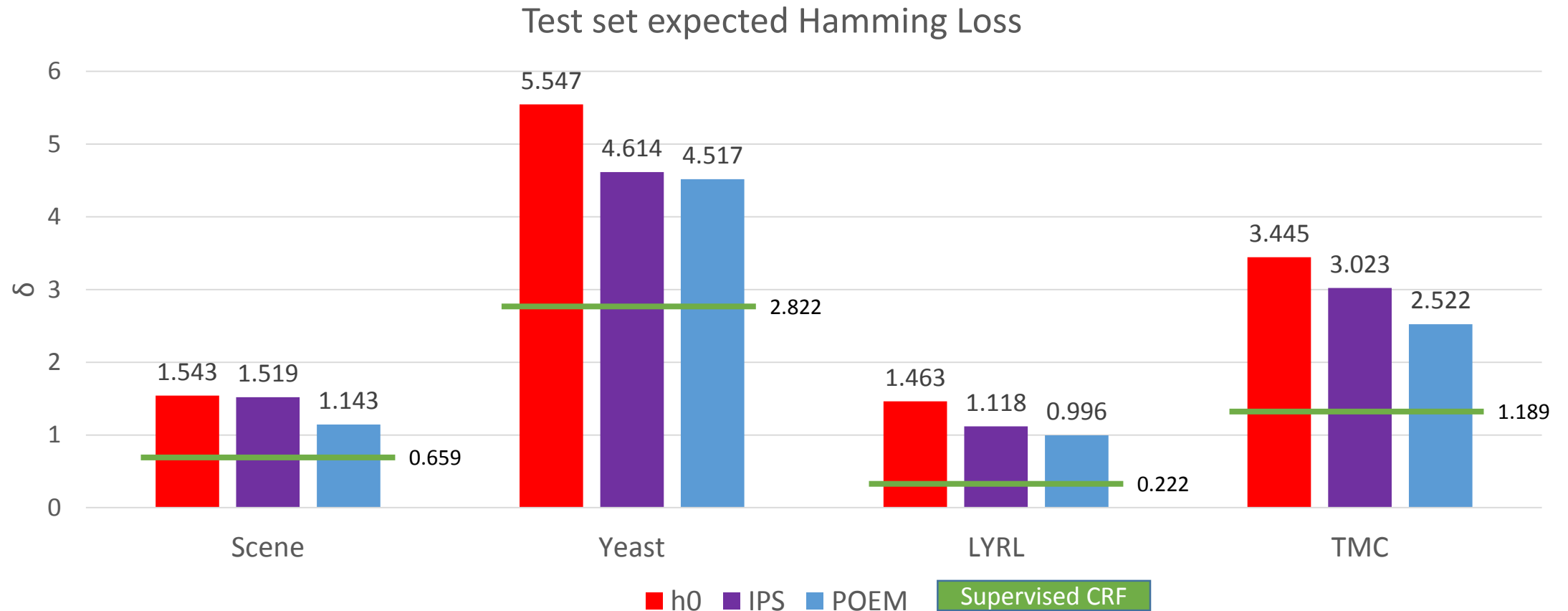
# Experiment

- Supervised  $\rightarrow$  Bandit **MultiLabel** [Agarwal et al]
- $\delta(x, y) = \text{Hamming}(y^*(x), y)$  (smaller is better)
- LibSVM Datasets
  - Scene (few features, labels and data)
  - Yeast (many labels)
  - LYRL (many features and data)
  - TMC (many features, labels and data)
- Validate hyper-params  $(\lambda, \mu)$  using  $\hat{R}_{\mathcal{D}_{val}}(h)$
- Supervised test set expected Hamming loss

# Approaches

- Baselines
  - $h_0$ : Supervised CRF trained on 5% of training data
- Proposed
  - IPS (No variance penalty) (extends [Bottou et al])
  - POEM
- Skylines
  - Supervised CRF (independent logit regression)

# (1) Does variance regularization help?



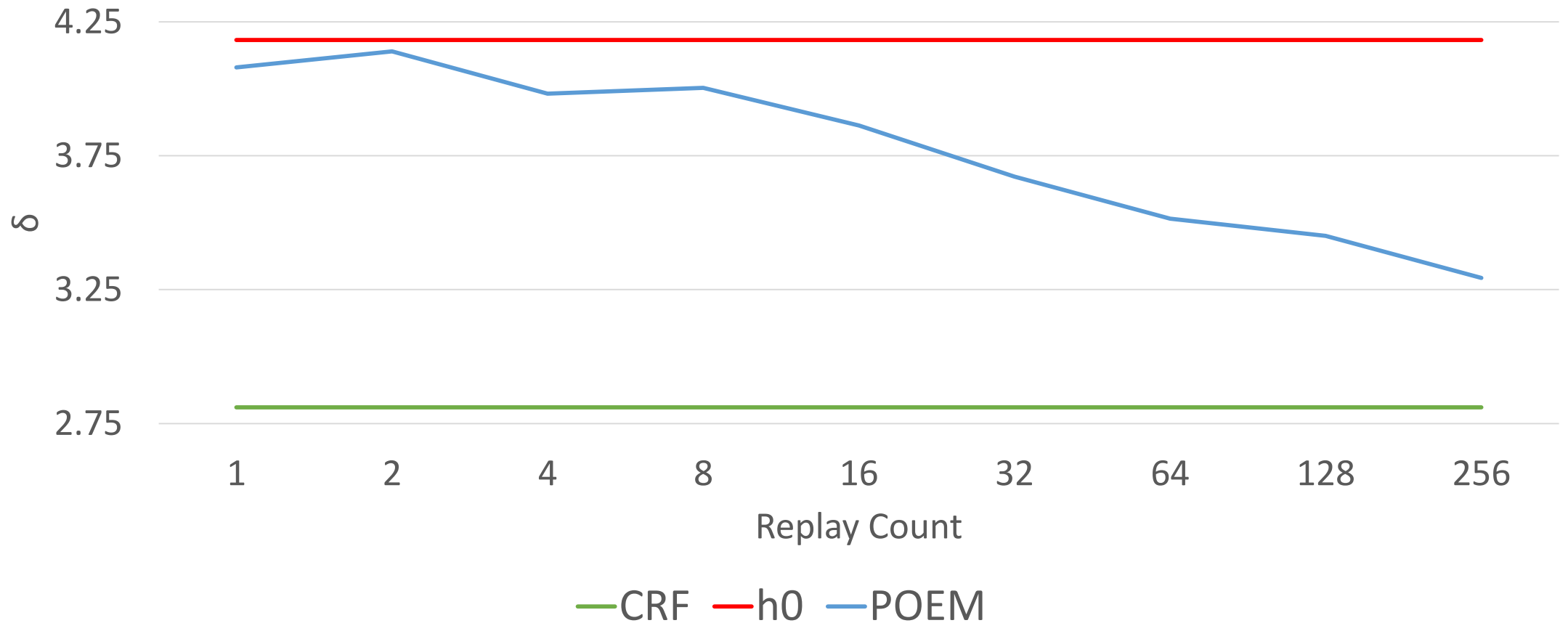
## (2) Is it efficient?

Avg Time (s)	Scene	Yeast	LYRL	TMC
POEM(B)	75.20	94.16	561.12	949.95
POEM(S)	4.71	5.02	120.09	276.13
CRF	4.86	3.28	62.93	99.18

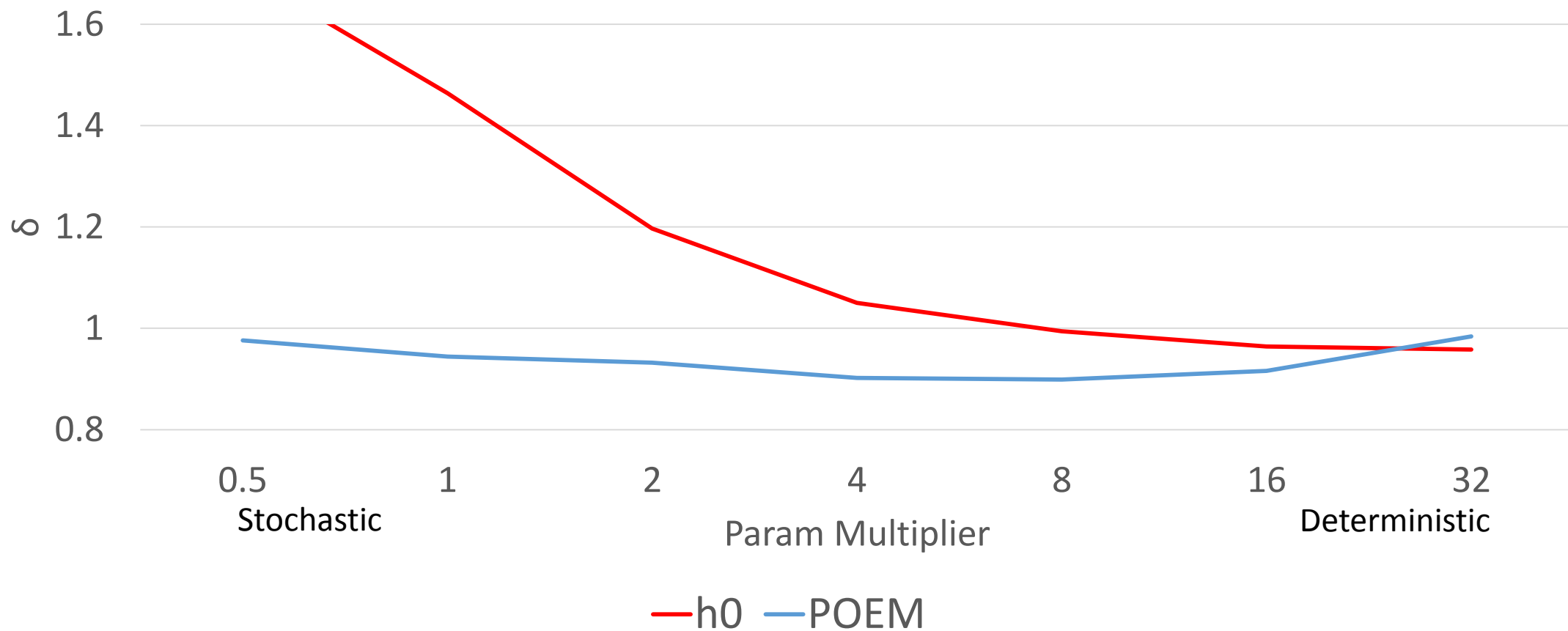
- POEM recovers same performance at fraction of L-BFGS cost
- Scales as supervised CRF, learns from bandit feedback



(3) Does generalization improve as  $n \rightarrow \infty$ ?



# (4) Does stochasticity of $h_0$ affect learning?



# Conclusion

- CRM principle to learn from logged bandit feedback
  - Variance regularization
- POEM for structured output prediction
  - Scales as supervised CRF, learns from bandit feedback
- Contact: [adith@cs.cornell.edu](mailto:adith@cs.cornell.edu)
- POEM available at <http://www.cs.cornell.edu/~adith/poem/>
- Long paper: Counterfactual risk minimization – Learning from logged bandit feedback, <http://jmlr.org/proceedings/papers/v37/swaminathan15.html>
- Thanks!

# References

1. Art B. Owen. 2013. *Monte Carlo theory, methods and examples*.
2. Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70. 41-55.
3. Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.* 14, 1, 3207-3260.
4. Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li and Robert Schapire. 2014. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. Proceedings of the 31<sup>st</sup> International Conference on Machine Learning. 1638-1646.
5. Andreas Maurer and Massimiliano Pontil. 2009. Empirical Bernstein bounds and sample-variance penalization. Proceedings of the 22nd Conference on Learning Theory.
6. Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning.