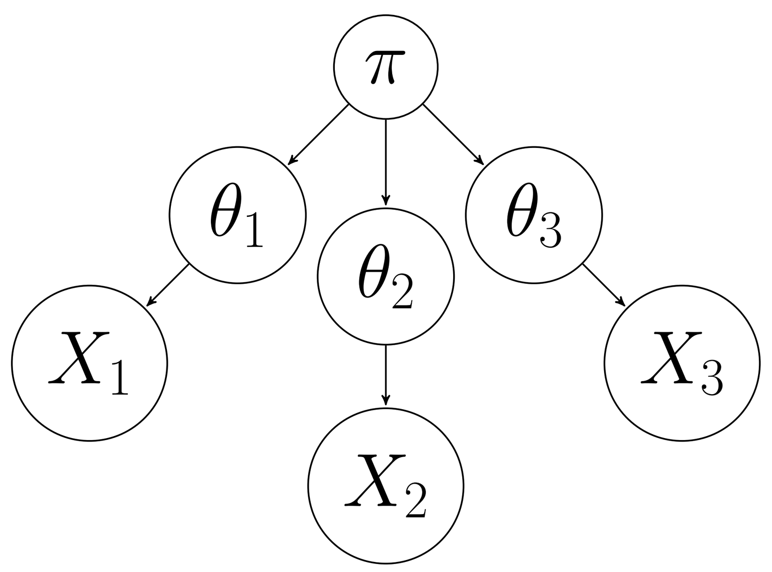
Empirical Bayes via ERM and Rademacher complexities: the Poisson model

Soham Jana [1], Yury Polyanksiy [2], Anzo Teh [2], Yihong Wu [3]

[1]Princeton University [2]MIT [3]Yale University

Problem Formulation

$$\theta_i \overset{\text{iid}}{\sim} \pi \qquad X_i \sim \text{Poi}(\theta_i) \qquad p_\pi(x) = \int \frac{e^{-\theta}\theta^x}{x!} d\pi(\theta)$$



Goal: estimate $\hat{f}$ that minimizes $\mathbb{E}[(\hat{f}(X) - f_\pi(X))^2]$.

Bayes estimator: $f_\pi(x) = \mathbb{E}[\theta | X = x] = (x+1)\frac{p_\pi(x+1)}{p_\pi(x)}$.

Regularity constraint: $f_\pi(x) \le f_\pi(x+1), \forall x \ge 0$.

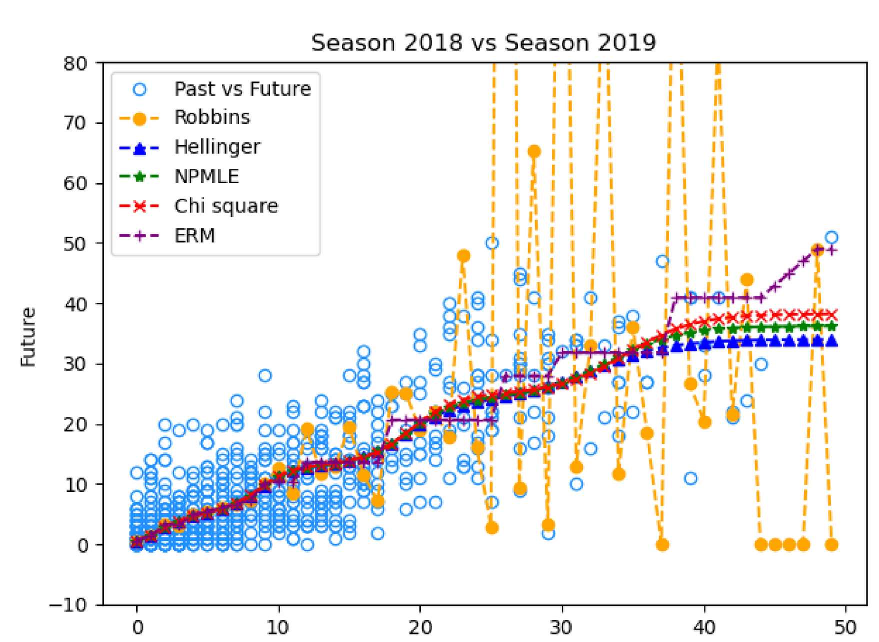
Related Work

f-modelling: Robbins estimator $f_{\text{Rob}}(x) \triangleq (x+1)\frac{N(x+1)}{N(x)+1}$

g-modelling: minimum distance estimator

$$\hat{\pi} = \arg\min_{\pi} d(p^{\text{emp}} || p_\pi) \qquad \hat{f}(x) \triangleq (x+1)\frac{p_{\hat{\pi}}(x+1)}{p_{\hat{\pi}}(x)}$$

Modelling	Empirical performance	Regularity	Speed	Optimality
f-modelling	Bad	No	Fast	Yes
g-modelling	Good	Yes	Slow	Yes
Our method	Good	Yes	Fast	Yes



ERM Estimator Derivation

Summation by parts: $\mathbb{E}[\theta f(X)] = \mathbb{E}[(X+1)\frac{p_\pi(X+1)}{p_\pi(X)} f(X)] = \mathbb{E}[X f(X-1)]$;

$$f_\pi = \arg\min_f \mathbb{E}[(f(X) - \theta)^2] = \arg\min_f \mathbb{E}[f(X)^2 - 2Xf(X-1)].$$

$$f_{\text{erm}} \in \arg\min_{f \in \mathcal{F}} \hat{\mathbb{E}}[f(X)^2 - 2Xf(X-1)] \qquad \mathcal{F} = \{f : f(x) \le f(x+1), \forall x \ge 0\}.$$

Mixture	$p(X\|\theta)$	ERM Objective
Geo(θ)	$\theta^X(1-\theta)$	$\hat{\mathbb{E}}[f(X)^2 - 2f(X) + 2f(X-1)\mathbf{1}_{\{X>0\}}]$
NB(r,θ)	$\binom{k+r-1}{k}(1-\theta)^r\theta^k$	$\hat{\mathbb{E}}[f(X)^2 - 2\frac{X+1}{X+r}f(X-1)\mathbf{1}_{\{X>0\}}]$
$\mathcal{N}(\theta,1)$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(X-\theta)^2}{2}\right)$	$\hat{\mathbb{E}}[f(X)^2 - 2Xf(X) + 2f'(X)]$
Exp(θ)	$\theta\exp(-\theta X)$	$\hat{\mathbb{E}}[f(X)^2 - 2f'(X)]$

Table 1. ERM objectives for other mixture models.

ERM Algorithm (Generalized Isotonic Regression)

Iterative interval partitioning (stop at $b_m = X_{\max} + 1$).

$$b_i = \begin{cases} 1 & i = 0 \\ 1 + \arg\min_{b_{i-1} \le i^* \le X_{\max}} \frac{\sum_{i=b_{i-1}}^{i^*}(a_i+1)N(a_i+1)}{\sum_{i=b_{i-1}}^{i^*} N(a_i)} & i \ge 1 \end{cases}$$

$$x \in [b_m, b_{m+1} - 1] : \hat{f}_{\text{erm}}(x) = \frac{\sum_{i=b_m}^{b_{m+1}-1}(a_i+1)N(a_i+1)}{\sum_{i=b_m}^{b_{m+1}-1} N(a_i)}.$$

ERM lemma: $\hat{f}_{\text{erm}} \le X_{\max} \in O(\log n)$ w.h.p.

One-Dimensional Optimality

Bounded prior: $\pi \in \mathcal{P}([0,h])$: $\text{Regret}_\pi(\hat{f}_{\text{erm}}) \le O\left(\frac{\max\{1,h\}^3}{n}\left(\frac{\log n}{\log\log n}\right)^2\right)$.

Subexponential prior: $\pi \in \text{SubE}(s) = \left\{G : G([t,\infty])) \le 2e^{-t/s}, \forall t > 0\right\}$.

$$\text{Regret}_\pi(\hat{f}_{\text{erm}}) \le O\left(\frac{\max\{1,s\}^3}{n}(\log n)^3\right).$$

Multi-dimensional ERM

$$\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} \boldsymbol{\pi} \qquad X_{ij} \overset{\text{ind}}{\sim} \text{Poi}(\theta_{ij}), j = 1, \cdots, d.$$

$$\hat{\boldsymbol{f}}_{\text{erm}} = \arg\min_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}} \quad \hat{\mathbb{E}}\left[||\boldsymbol{f}(\boldsymbol{X})||^2 - 2\sum_{j=1}^d X_j f_j(\boldsymbol{X} - \boldsymbol{e}_j)\right],$$

$$\boldsymbol{\mathcal{F}} = \{\boldsymbol{f} : \mathbb{Z}_+^d \to \mathbb{R}_+^d : f_i(\boldsymbol{x}) \le f_i(\boldsymbol{x} + \boldsymbol{e}_i), \forall i = 1, \cdots, d, \forall \boldsymbol{x} \in \mathbb{Z}_+^d\}.$$

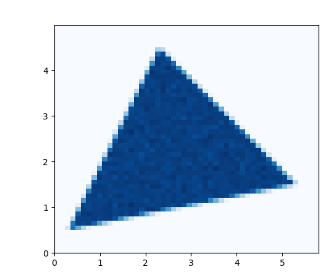


Fig. 1. $\boldsymbol{\theta}_i \overset{\text{Unif}}{\sim}$ triangle.

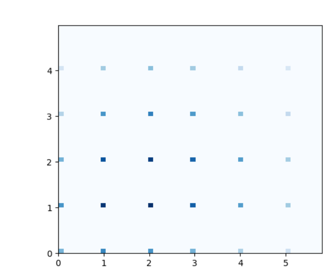


Fig. 2. $\boldsymbol{X}_i \overset{\text{ind}}{\sim} \text{Poi}(\boldsymbol{\theta}_i)$

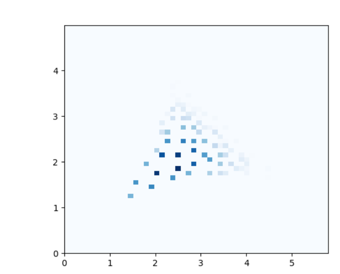


Fig. 3. Denoised via $\hat{\boldsymbol{f}}_{\text{erm}}$.

$$\boldsymbol{\pi} \in \mathcal{P}([0,h]^d) : \text{Regret}_\pi(\hat{\boldsymbol{f}}_{\text{erm}}) \le O\left(\frac{d}{n}\max\{c_1, c_2 h\}^{d+2}\left(\frac{\log(n)}{\log\log(n)}\right)^{d+1}\right)$$

$$\pi_1, \cdots, \pi_d \in \text{SubE}(s) : \text{Regret}_\pi(\hat{\boldsymbol{f}}_{\text{erm}}) \le O\left(\frac{d}{n}(\max\{c_3, c_4 s\}\log(n))^{d+2}\right)$$

Proof Techniques: Localization, Offset Rademacher

Localized function class: $\mathcal{F}_* \triangleq \{f \in \mathcal{F} : f \le X_{\max} \vee X'_{\max}\}$.

$$\text{Regret}_\pi(\hat{f}) \le \frac{3}{n}T_1(n) + \frac{2}{n}T_2(n)$$

$$T_1(n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}_*} \sum_{i=1}^n (\epsilon_i - \frac{1}{6})(f(X_i) - f^*(X_i))^2\right],$$

$$T_2(n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}_*} \sum_{i=1}^n \left\{2\epsilon_i(f^*(X_i)(f^*(X_i) - f(X_i)) - X_i(f^*(X_i - 1) - f(X_i - 1))) - \frac{1}{4}(f^*(X_i) - f(X_i))^2\right\}\right].$$

- $\pi \in \mathcal{P}([0,h])$: $T_1(n), T_2(n) \lesssim \max\{1,h^2\}M + \max\{1,h\}M^2$; $M \in O(\mathbb{E}[X_{\max}])$;
- $\pi \in \text{SubE}(s)$: prior truncation $\to \mathcal{P}([0, c_s \log n])$.

Conference on Learning Theory, 2023

anzoteh@mit.edu