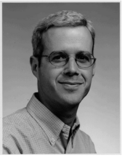
Learning Distributions from Samples

Sudeep Kamath

PRINCETON UNIVERSITY

Joint work with



Alon Orlitsky   Dheeraj Pichapati   Ananda Theertha Suresh

Simons Institute, 17 Mar 2015

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k})$$

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k})$$

Number of samples

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k}_{\boxed{k}})$$

Number of samples

Alphabet size

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k}_{\boxed{k}}) \in \boxed{\Delta_k}$$

Number of samples

Alphabet size

k-simplex

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k}_{\boxed{k}}) \in \boxed{\Delta_k}$$

Number of samples

Alphabet size

k-simplex

$L(p, q)$: loss for estimating p by q

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k}_{\boxed{k}}) \in \boxed{\Delta_k}$$

Number of samples

Alphabet size

k-simplex

$L(p, q)$: loss for estimating p by q

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

Learning probability distributions

$$\underbrace{X_1, X_2, \ldots, X_n}_{\boxed{n}} \sim \text{i.i.d. } p = (\underbrace{p_1, p_2, \ldots, p_k}_{\boxed{k}}) \in \boxed{\Delta_k}$$

Number of samples (n) — Alphabet size (k) — k-simplex (Δ_k)

$L(p, q)$: loss for estimating p by q

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

*Q: What is the **minimax risk** and the **optimal estimator** $\mathbf{q_{x^n}}$?*

Relevant measures of loss $L(p, q)$

Relevant measures of loss $L(p, q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction

Relevant measures of loss $L(p, q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction
- Chi squared distance: $\chi^2(p||q) = \left(\sum_i \frac{p_i^2}{q_i} \right) - 1$
  - multiple correspondence analysis

Relevant measures of loss $L(p, q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction
- Chi squared distance: $\chi^2(p||q) = \left(\sum_i \frac{p_i^2}{q_i} \right) - 1$
  - multiple correspondence analysis
- ℓ_1 distance: $||p - q||_1 = \sum_i |p_i - q_i|$
  - classification, machine learning

Relevant measures of loss $L(p,q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction
- Chi squared distance: $\chi^2(p||q) = \left(\sum_i \frac{p_i^2}{q_i}\right) - 1$
  - multiple correspondence analysis
- ℓ_1 distance: $||p - q||_1 = \sum_i |p_i - q_i|$
  - classification, machine learning
- Hellinger distance: $H(p||q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$
  - sequential and asymptotic statistics

Relevant measures of loss $L(p,q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction

f-divergence [Csiszár '63]:

$$D_f(p||q) = \sum_i q_i f\left(\frac{p_i}{q_i}\right)$$

where f is convex and $f(1) = 0$.

i

  - sequential and asymptotic statistics

Relevant measures of loss $L(p,q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$
  - compression, prediction
- Chi squared distance: $\chi^2(p||q) = \left(\sum_i \frac{p_i^2}{q_i}\right) - 1$
  - multiple correspondence analysis
- ℓ_1 distance: $||p-q||_1 = \sum_i |p_i - q_i|$
  - classification, machine learning
- Hellinger distance: $H(p||q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$
  - sequential and asymptotic statistics

Relevant measures of loss $L(p,q)$

- Relative entropy: $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$ (old story)
  - compression, prediction
- Chi squared distance: $\chi^2(p||q) = \left(\sum_i \frac{p_i^2}{q_i}\right) - 1$ (new story)
  - multiple correspondence analysis
- ℓ_1 distance: $||p-q||_1 = \sum_i |p_i - q_i|$
  - classification, machine learning
- Hellinger distance: $H(p||q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$
  - sequential and asymptotic statistics

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Empirical estimator is a bad idea
  - If $p_i > 0$ but $q_i = 0$, then $p_i \log \frac{p_i}{q_i} = \infty$

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Empirical estimator is a bad idea
  - If $p_i > 0$ but $q_i = 0$, then $p_i \log \frac{p_i}{q_i} = \infty$
- *Cromwell's rule:* Zero probability not assigned to any symbol

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Empirical estimator is a bad idea
  - If $p_i > 0$ but $q_i = 0$, then $p_i \log \frac{p_i}{q_i} = \infty$
- *Cromwell's rule:* Zero probability not assigned to any symbol

Oliver Cromwell
[1599-1658]

"I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Laplace estimator

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Laplace estimator

Probability $\propto$ no. of occurrences $+ 1$

$$\text{P(sunrise after } n \text{ days)} = \frac{n+1}{n+2}$$

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Laplace estimator

Probability $\propto$ no. of occurrences $+ 1$

P(sunrise after n days) $= \dfrac{n+1}{n+2}$

- Add-β estimator: *Probability* $\propto$ *no. of occurrences* $+ \beta$

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Laplace estimator

Probability $\propto$ no. of occurrences $+ 1$

$$\text{P(sunrise after } n \text{ days)} = \frac{n+1}{n+2}$$

- Add-β estimator: *Probability* $\propto$ *no. of occurrences* $+ \beta$
- *Cumulative risk or Minimax Redundancy*:

$$\min_{q_{x^i}} \max_{p \in \Delta_k} \sum_{i=0}^{n-1} \mathbb{E} D(p || q_{X^i})$$

Relative Entropy story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n})$$

- Laplace estimator

Probability $\propto$ no. of occurrences + 1

$$\text{P(sunrise after } n \text{ days)} = \frac{n+1}{n+2}$$

- Add-β estimator: *Probability* $\propto$ *no. of occurrences* $+ \beta$
- *Cumulative risk or Minimax Redundancy*:

$$\min_{q_{x^i}} \max_{p \in \Delta_k} \sum_{i=0}^{n-1} \mathbb{E} D(p || q_{X^i})$$

Add-$1/2$ is asymptotically optimal [Krichevsky-Trofimov '81]

Relative Entropy story

Relative Entropy story

$$\boxed{\frac{k-1}{2n} + o\left(\frac{1}{n}\right)} \leq r_{k,n} \leq \boxed{1.018 \times \frac{k-1}{2n} + o\left(\frac{1}{n}\right)}$$

Lower bound $\longleftarrow$ [Krichevsky '98] $\longrightarrow$ Best add-β rule

Relative Entropy story

$$\boxed{\frac{k-1}{2n} + o\left(\frac{1}{n}\right)} \leq r_{k,n} \leq \boxed{1.018 \times \frac{k-1}{2n} + o\left(\frac{1}{n}\right)}$$

Lower bound $\longleftarrow$ [Krichevsky '98] $\longrightarrow$ Best add-β rule

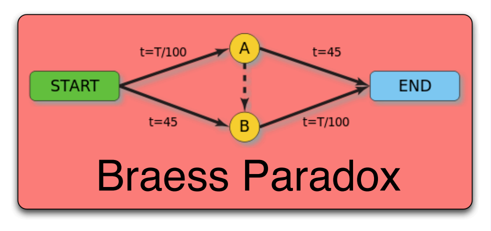
Theorem (Braess-Sauer '04)

Relative Entropy story

$$\boxed{\frac{k-1}{2n} + o\left(\frac{1}{n}\right)} \le r_{k,n} \le \boxed{1.018 \times \frac{k-1}{2n} + o\left(\frac{1}{n}\right)}$$

Lower bound ⟵ [Krichevsky '98] ⟶ Best add-β rule

Theorem (Braess-Sauer '04)



Relative Entropy story

$$\boxed{\frac{k-1}{2n} + o\left(\frac{1}{n}\right)} \leq r_{k,n} \leq \boxed{1.018 \times \frac{k-1}{2n} + o\left(\frac{1}{n}\right)}$$

Lower bound ⟵ [Krichevsky '98] ⟶ Best add-β rule

Theorem (Braess-Sauer '04)

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p || q_{X^n}) = \frac{k-1}{2n} + o\left(\frac{1}{n}\right)$$

Relative Entropy story

$$\boxed{\frac{k-1}{2n} + o\left(\frac{1}{n}\right)} \leq r_{k,n} \leq \boxed{1.018 \times \frac{k-1}{2n} + o\left(\frac{1}{n}\right)}$$

Lower bound ⟵ [Krichevsky '98] ⟶ Best add-β rule

Theorem (Braess-Sauer '04)

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} D(p||q_{X^n}) = \frac{k-1}{2n} + o\left(\frac{1}{n}\right)$$

Asymptotically optimum: varying-add-β rule,
β varying with no. of occurrences

$$\beta_0 = 1/2, \qquad \beta_1 = 1, \qquad \beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p || q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p||q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Cromwell's rule applies: no estimate q_i may be zero

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p||q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Cromwell's rule applies: no estimate q_i may be zero

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

For any $k \geq 2, n \geq 1$,

$$\frac{k-1}{n+k+1} - \frac{k(k-1)\left[\log(n+1)+1\right]}{4(n+k)(n+k+1)} \leq r_{k,n} \leq \frac{k-1}{n+1}$$

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p||q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Cromwell's rule applies: no estimate q_i may be zero

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

For any $k \geq 2, n \geq 1$,

$$\frac{k-1}{n+k+1} - \frac{k(k-1)\left[\log(n+1)+1\right]}{4(n+k)(n+k+1)} \leq r_{k,n} \leq \frac{k-1}{n+1}$$

For fixed k,

$$r_{k,n} = \frac{k-1}{n} + O\left(\frac{\log n}{n^2}\right)$$

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p||q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Cromwell's rule applies: no estimate q_i may be zero

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

For any $k \geq 2, n \geq 1$,

$$\frac{k-1}{n+k+1} - \frac{k(k-1)\left[\log(n+1)+1\right]}{4(n+k)(n+k+1)} \leq r_{k,n} \leq \frac{k-1}{n+1}$$

For fixed k,

$$r_{k,n} = \frac{k-1}{n} + O\left(\frac{\log n}{n^2}\right)$$

Upper bound:
Laplace estimator


BLACK
MAGIC

Chi squared distance story

$$r_{k,n} = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E}\, \chi^2(p||q_{X^n}) = \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} \sum_i \frac{p_i^2}{q_i} - 1$$

Cromwell's rule applies: no estimate q_i may be zero

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

For any $k \geq 2, n \geq 1$,

$$\frac{k-1}{n+k+1} - \frac{k(k-1)\left[\log(n+1)+1\right]}{4(n+k)(n+k+1)} \leq r_{k,n} \leq \frac{k-1}{n+1}$$

For fixed k,

$$r_{k,n} = \frac{k-1}{n} + O\left(\frac{\log n}{n^2}\right)$$

Upper bound:
Laplace estimator

Chi squared distance story: $k = 2$

$\text{Alphabet} = \{1, 2\}$

Chi squared distance story: $k = 2$

$$\text{Alphabet} = \{1, 2\}$$
$$p = P(X = 1),$$
$$1 - p = P(X = 2)$$

Chi squared distance story: $k = 2$

$$\text{Alphabet} = \{1, 2\}$$

$$p = P(X = 1),$$

$$1 - p = P(X = 2)$$

Expected loss under Laplace estimator =

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+1}{n+2}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+1}{n+2}\right)} - 1 \right)$$

Chi squared distance story: $k = 2$

$$\text{Alphabet} = \{1, 2\}$$

$$p = P(X = 1),$$

$$1 - p = P(X = 2)$$

Expected loss under Laplace estimator =

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+1}{n+2}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+1}{n+2}\right)} - 1 \right) \leq \frac{1}{n}$$

what if?

... we had used an add-β estimator ... ?

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples
- p

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples
- $p = P(X = 1)$

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples
- $p = P(X = 1)$
- β

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples
- $p = P(X = 1)$
- β: add-β estimator

What if we had used add-β?

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

- n: number of samples
- $p = P(X=1)$
- β: add-β estimator

If i 1's and $(n-i)$ 2's, probability estimate for $X=1$ is

$$\frac{i+\beta}{n+2\beta}$$

What if we had used add-β?

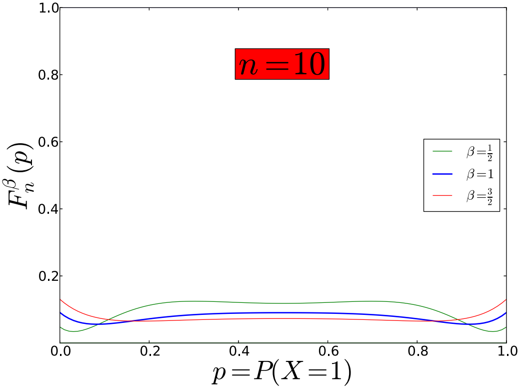
$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

What if we had used add-β?

$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$

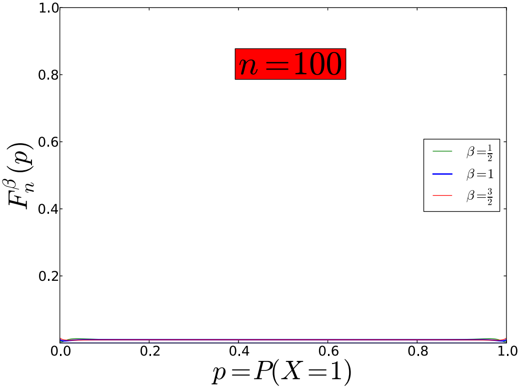
What if we had used add-β?

$$F_n^{\beta}(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



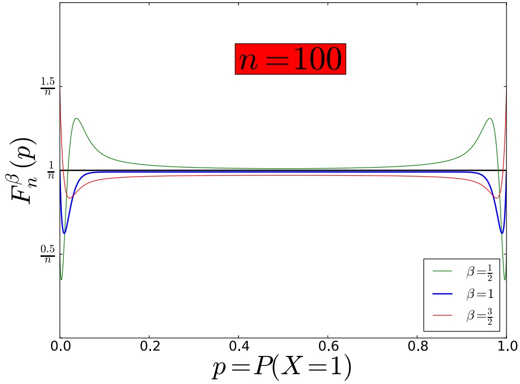
What if we had used add-β?

$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



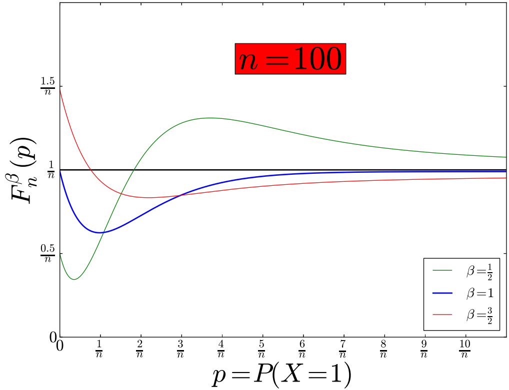
What if we had used add-β?

$$F_n^{\beta}(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



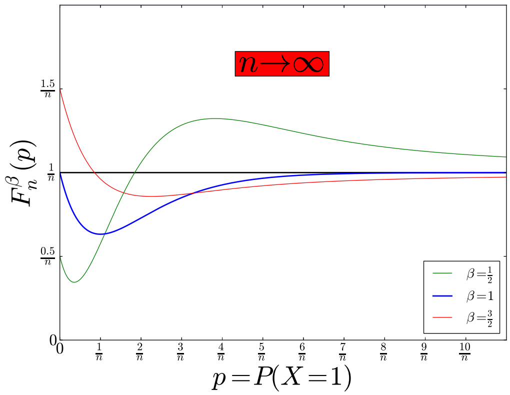
What if we had used add-β?

$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



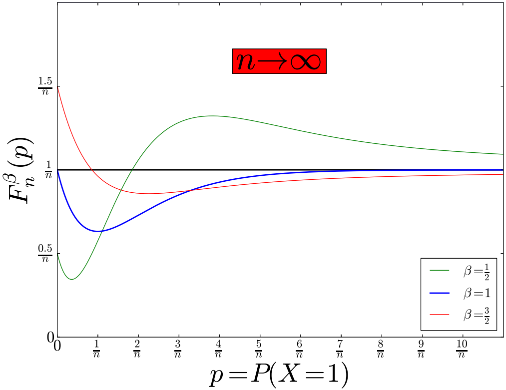
What if we had used add-β?

$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



What if we had used add-β?

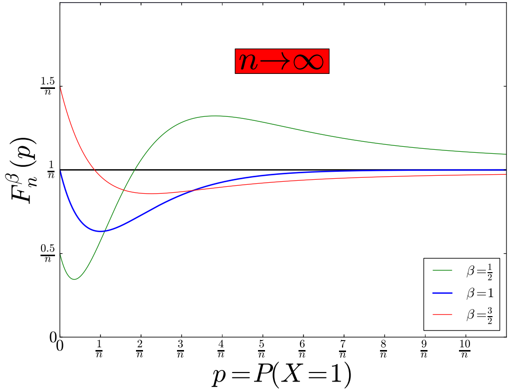
$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



$$\max_{p \in [0,1]} F_n^\beta(p) \sim \frac{c(\beta)}{n}$$

What if we had used add-β?

$$F_n^\beta(p) := \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left(\frac{p^2}{\left(\frac{i+\beta}{n+2\beta}\right)} + \frac{(1-p)^2}{\left(\frac{n-i+\beta}{n+2\beta}\right)} - 1 \right)$$



$$\max_{p \in [0,1]} F_n^\beta(p) \sim \frac{c(\beta)}{n}$$

$$c(\beta) \begin{cases} = 1 & \beta = 1 \\ > 1 & \beta \neq 1 \end{cases}$$

How many more stories to tell?

How many more stories to tell?

Relative entropy

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

Why simpler?

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

Why simpler? Luck!

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

Why simpler? Luck!

- Expected loss near boundaries depends erratically on loss function and estimator

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

Why simpler? Luck!

- Expected loss near boundaries depends erratically on loss function and estimator
- A coherent understanding of optimal estimator for all f-divergence loss?

How many more stories to tell?

Relative entropy

Asymptotically optimum:
varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum:
Laplace estimator
(add-1 rule)

Why simpler? Luck!

- Expected loss near boundaries depends erratically on loss function and estimator
- A coherent understanding of optimal estimator for all f-divergence loss? Highly challenging!

How many more stories to tell?

Relative entropy

Asymptotically optimum: varying-add-β rule

$$\beta_0 = 1/2, \qquad \beta_1 = 1,$$
$$\beta_2 = \beta_3 = \ldots = 3/4.$$

Chi squared distance

Asymptotically optimum: Laplace estimator (add-1 rule)

Why simpler? Luck!

- Expected loss near boundaries depends erratically on loss function and estimator
- A coherent understanding of optimal estimator for all f-divergence loss? Highly challenging!
- Eg. for Hellinger loss, best add-β estimator can't meet natural lower bound

How many more stories to tell?

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i.

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i. Let f be convex, thrice-differentiable, sub-exponential with $f(1) = 0$.

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i. Let f be convex, thrice-differentiable, sub-exponential with $f(1) = 0$.

$$r_{k,n}(\alpha) = \min_{q_{x^n}} \max_{p \in \Delta_k, p_i \geq \alpha} D_f(p||q_{X^n})$$

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i. Let f be convex, thrice-differentiable, sub-exponential with $f(1) = 0$.

$$r_{k,n}(\alpha) = \min_{q_{x^n}} \max_{p \in \Delta_k, p_i \geq \alpha} D_f(p || q_{X^n}) = \frac{(k-1)f''(1)}{2n} + o\left(\frac{1}{n}\right)$$

How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

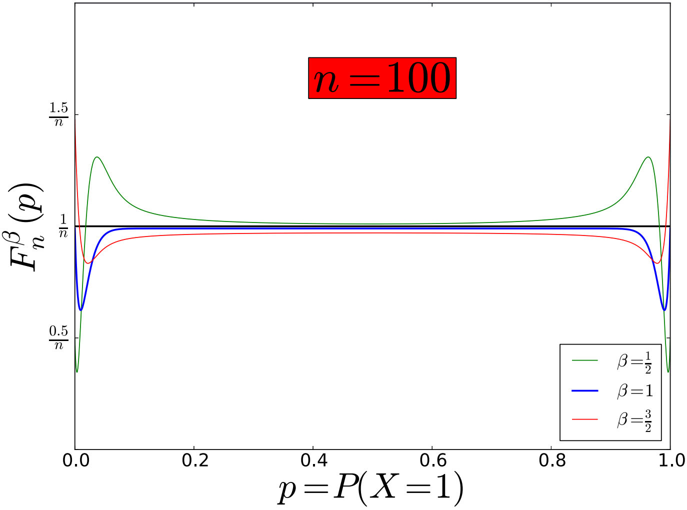
Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i. Let f be convex, thrice-differentiable, sub-exponential with $f(1) = 0$.

$$r_{k,n}(\alpha) = \min_{q_{x^n}} \max_{p \in \Delta_k, p_i \geq \alpha} D_f(p || q_{X^n}) = \frac{(k-1)f''(1)}{2n} + o\left(\frac{1}{n}\right)$$

Any add-β rule for $\beta > 0$ is asymptotically optimal.

How many more stories to tell?



How many more stories to tell?

Avoid simplex boundaries and look at f-divergence loss

Theorem (K.-Orlitsky-Pichapati-Suresh '15)

Suppose $p = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq \alpha > 0$ for all i. Let f be convex, thrice-differentiable, sub-exponential with $f(1) = 0$.

$$r_{k,n}(\alpha) = \min_{q_{x^n}} \max_{p \in \Delta_k, p_i \geq \alpha} D_f(p || q_{X^n}) = \frac{(k-1)f''(1)}{2n} + o\left(\frac{1}{n}\right)$$

Any add-β rule for $\beta > 0$ is asymptotically optimal.

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

- Non-asymptotic bounds for chi squared loss

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

- Non-asymptotic bounds for chi squared loss
- General formula for arbitrary f-divergence

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

- Non-asymptotic bounds for chi squared loss
- General formula for arbitrary f-divergence: difficult!

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

- Non-asymptotic bounds for chi squared loss
- General formula for arbitrary f-divergence: difficult!
- $p_i \geq \alpha > 0$: any add-β estimator optimal and simple formula for minimax risk

Summary

Minimax Risk $r_{k,n}$

$$r_{k,n} := \min_{q_{x^n}} \max_{p \in \Delta_k} \mathbb{E} L(p, q_{X^n})$$

q_{x^n} is an estimator for p based on $x^n = (x_1, x_2, \ldots, x_n)$

- Non-asymptotic bounds for chi squared loss
- General formula for arbitrary f-divergence: difficult!
- $p_i \geq \alpha > 0$: any add-β estimator optimal and simple formula for minimax risk

"On learning distributions from their samples"
- coming soon on arxiv