

# Distribution Estimation under the Infinity Norm

**Amichai Painsky**

**Tel Aviv University, Israel**

**Joint work with Aryeh Kontorovich from Ben Gurion University**

# Motivation

- Consider a survey asking people for their favorite foods



# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs



# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs
- Recall Statistics 101:

Let  $X \sim \text{Bin}(n, \theta)$ . Wald's  $1 - \alpha$  level CI for  $\theta$  follows

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}$$



# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs
- Therefore, the common approach is to construct a binomial CI of level  $1 - \alpha$  for the most frequent food in the sample. For example:



$$\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{max}(1 - \hat{p}_{max})}{n}}$$

# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs
- Therefore, the common approach is to construct a binomial CI of level  $1 - \alpha$  for the most frequent food in the sample. For example:



$$\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{max}(1 - \hat{p}_{max})}{n}}$$

Pop-quiz: what does this guarantee?

# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs
- Therefore, the common approach is to construct a binomial CI of level  $1 - \alpha$  for the most frequent food in the sample. For example:



$$\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{max}(1 - \hat{p}_{max})}{n}}$$

Pop-quiz: what does this guarantee?

- If we repeat this process  $N$  times we get a coverage in  $(1 - \alpha)N$  of the experiments

# Motivation

- Consider a survey asking people for their favorite foods
- We would like to report the most popular food along with its CIs
- Therefore, the common approach is to construct a binomial CI of level  $1 - \alpha$  for the most frequent food in the sample. For example:



$$\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{max}(1 - \hat{p}_{max})}{n}}$$

Pop-quiz: what does this guarantee?

- If we repeat this process  $N$  times we get a coverage in  $(1 - \alpha)N$  of the experiments

**FALSE**

# Why is it Wrong?

A counter example:

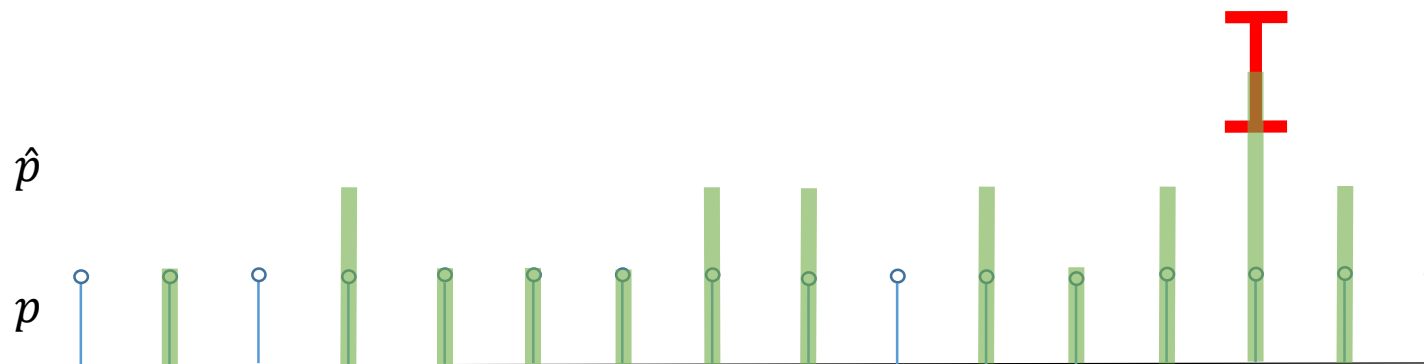
- Assume  $p$  is uniform over an alphabet size  $m$

# Why is it Wrong?

A counter example:

- Assume  $p$  is uniform over an alphabet size  $m$
- We construct a binomial CI for the most frequent event in the sample.

For example  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$

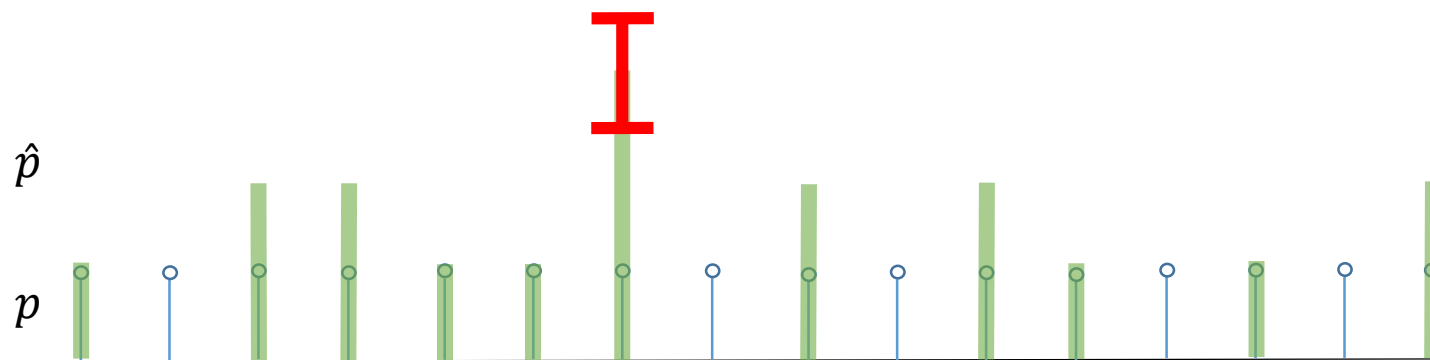


# Why is it Wrong?

A counter example:

- Assume  $p$  is uniform over an alphabet size  $m$
- We construct a binomial CI for the most frequent event in the sample.

For example  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$



# Why is it Wrong?

A counter example:

- Assume  $p$  is uniform over an alphabet size  $m$
- We construct a binomial CI for the most frequent event in the sample.

For example  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$



# Selective Inference

- The classical inference regime:

$$P(\cap_u \theta_u \in T_u(X^n)) \geq 1 - \alpha$$

# Selective Inference

- The classical inference regime:

$$P(\cap_u \theta_u \in T_u(X^n)) \geq 1 - \alpha$$

- Selective inference generalizes this framework and considers a subset of parameters of interest, selected during the experiment

# Selective Inference

- The classical inference regime:

$$P(\cap_u \theta_u \in T_u(X^n)) \geq 1 - \alpha$$

- Selective inference generalizes this framework and considers a subset of parameters of interest, selected during the experiment
- For example,

$$P(\cap_u \{\theta_u \in T_u(X^n), \theta_u \text{ is selected}\}) \geq 1 - \alpha$$

# Selective Inference

- The classical inference regime:

$$P(\cap_u \theta_u \in T_u(X^n)) \geq 1 - \alpha$$

- Selective inference generalizes this framework and considers a subset of parameters of interest, selected during the experiment
- For example,

$$P(\cap_u \{\theta_u \in T_u(X^n), \theta_u \text{ is selected}\}) \geq 1 - \alpha$$

- Denoted as *Simultaneous over Selected* (Benjamini et al., 2019)

# Notation and Problem Statement

- Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$

# Notation and Problem Statement

- Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$
- Let  $X^n$  be a sample of  $n$  i.i.d. observations from  $p$

# Notation and Problem Statement

- Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$
- Let  $X^n$  be a sample of  $n$  i.i.d. observations from  $p$
- Let  $u^* = \operatorname{argmax}_{u \in \mathcal{X}} N_u(X^n)/n$  be the most frequent symbol in the sample

# Notation and Problem Statement

- Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$
- Let  $X^n$  be a sample of  $n$  i.i.d. observations from  $p$
- Let  $u^* = \operatorname{argmax}_{u \in \mathcal{X}} N_u(X^n)/n$  be the most frequent symbol in the sample
- Let  $\hat{p}_{u^*} = \max_{u \in \mathcal{X}} N_u(X^n)/n = \hat{p}_{max}$

# Notation and Problem Statement

- Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$
- Let  $X^n$  be a sample of  $n$  i.i.d. observations from  $p$
- Let  $u^* = \operatorname{argmax}_{u \in \mathcal{X}} N_u(X^n)/n$  be the most frequent symbol in the sample
- Let  $\hat{p}_{u^*} = \max_{u \in \mathcal{X}} N_u(X^n)/n = \hat{p}_{max}$
- We seek  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$

# Related Result

**Asymptotic Result** (Shifeng and Guoying, 2004)

$$\sqrt{n}(\hat{p}_{max} - p_{max}) \xrightarrow{D} \xi_{[1]}$$

where  $(\xi_1, \dots, \xi_k)^T \sim \mathcal{N}\left(0, \left(p_{max}(\delta_{ij} - p_{max})\right)_{k \times k}\right)$  and  $k = \#(p_i = p_{max})$

# Related Result

## Asymptotic Result (Shifeng and Guoying, 2004)

$$\sqrt{n}(\hat{p}_{max} - p_{max}) \xrightarrow{D} \xi_{[1]}$$

where  $(\xi_1, \dots, \xi_k)^T \sim \mathcal{N}\left(0, \left(p_{max}(\delta_{ij} - p_{max})\right)_{k \times k}\right)$  and  $k = \#(p_i = p_{max})$

- For  $k = 1$  the  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$  should **asymptotically** work

# Related Result

## Asymptotic Result (Shifeng and Guoying, 2004)

$$\sqrt{n}(\hat{p}_{max} - p_{max}) \xrightarrow{D} \xi_{[1]}$$

where  $(\xi_1, \dots, \xi_k)^T \sim \mathcal{N}\left(0, \left(p_{max}(\delta_{ij} - p_{max})\right)_{k \times k}\right)$  and  $k = \#(p_i = p_{max})$

- For  $k = 1$  the  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$  should **asymptotically** work
- Very slow convergence rate for uniform or nearly uniform distributions

# Related Result

## Asymptotic Result (Shifeng and Guoying, 2004)

$$\sqrt{n}(\hat{p}_{max} - p_{max}) \xrightarrow{D} \xi_{[1]}$$

where  $(\xi_1, \dots, \xi_k)^T \sim \mathcal{N}\left(0, \left(p_{max}(\delta_{ij} - p_{max})\right)_{k \times k}\right)$  and  $k = \#(p_i = p_{max})$

- For  $k = 1$  the  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$  should **asymptotically** work
- Very slow convergence rate for uniform or nearly uniform distributions
- $k$  is unknown

# Related Result

## Asymptotic Result (Shifeng and Guoying, 2004)

$$\sqrt{n}(\hat{p}_{max} - p_{max}) \xrightarrow{D} \xi_{[1]}$$

where  $(\xi_1, \dots, \xi_k)^T \sim \mathcal{N}\left(0, \left(p_{max}(\delta_{ij} - p_{max})\right)_{k \times k}\right)$  and  $k = \#(p_i = p_{max})$

- For  $k = 1$  the  $\hat{p}_{max} \pm z_{\alpha/2} \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})/n}$  should **asymptotically** work
- Very slow convergence rate for uniform or nearly uniform distributions
- $k$  is unknown

**Additional related results** by Gupta and Nagel (1967), Gelfand et al. (1992), Glaz and Sison (1999), Fithian (2015), Dyer and Owen (2012), Zrnic and Fithian (2022)

# So What Do We Suggest?

- We seek  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$

# So What Do We Suggest?

- We seek  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$
- Possible relaxation  $P\left(\max_{u \in \mathcal{X}} |\hat{p}_u - p_u| \leq U(X^n)\right) \geq 1 - \alpha$

# So What Do We Suggest?

- We seek  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$
- Possible relaxation  $P\left(\max_{u \in \mathcal{X}} |\hat{p}_u - p_u| \leq U(X^n)\right) \geq 1 - \alpha$
- In other words, find the minimal  $U(X^n)$  such that  $\|\hat{p} - p\|_\infty \leq U(X^n)$   
w.p.  $1 - \alpha$

# Classical Results

- A simple application of McDiarmid's inequality suggests (Boucheron, Lugosi and Bousquet, 2003)

$$\|\hat{p} - p\|_\infty \leq \frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{1}{\alpha}\right)}{2n}}$$

# Classical Results

- A simple application of McDiarmid's inequality suggests (Boucheron, Lugosi and Bousquet, 2003)

$$\|\hat{p} - p\|_\infty \leq \frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{1}{\alpha}\right)}{2n}}$$

- On the other hand, for the binomial case  $Y \sim \text{Bin}(n, \theta)$  we know that (Bousquet et al., 2003, Dasgupta and Hsu, 2008b)

$$|\theta - \hat{\theta}| \leq \sqrt{\frac{5\hat{\theta}(1 - \hat{\theta})}{n} \log \frac{2}{\alpha}} + \frac{5}{n} \log \frac{2}{\alpha}$$

# Classical Results

- A simple application of McDiarmind's inequality suggests (Boucheron, Lugosi and Bousquet, 2003)

$$\|\hat{p} - p\|_\infty \leq \frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{1}{\alpha}\right)}{2n}}$$

- On the other hand, for the binomial case  $Y \sim \text{Bin}(n, \theta)$  we know that (Bousquet et al., 2003, Dasgupta and Hsu, 2008b)

$$|\theta - \hat{\theta}| \leq \sqrt{\frac{5\hat{\theta}(1 - \hat{\theta})}{n} \log \frac{2}{\alpha}} + \frac{5}{n} \log \frac{2}{\alpha}$$

- Can we expect something like

$$\|\hat{p} - p\|_\infty \stackrel{?}{\lesssim} \sqrt{\frac{\max_u \hat{p}_u(1 - \hat{p}_u)}{n} \log \frac{1}{\alpha}} + \frac{1}{n} \log \frac{1}{\alpha}$$

# Main Result

**Theorem 3 (P. and Kontorovich 2024, Theorem 3)**

$$\|\hat{p} - p\|_\infty \leq 2 \sqrt{\frac{\max_u p_u(1 - p_u)}{n} \log \frac{2}{\alpha} + \frac{V^*}{n} + \frac{4}{3n} \log \frac{2(n+1)}{\alpha} + \frac{\log(n)}{n}}$$

*w.p.  $1 - \alpha$ , where  $V^* = \max_{u \in \mathbb{N}} p_{[u]} (1 - p_{[u]}) \log(1 + u)$*

# Main Result

## Theorem 3 (P. and Kontorovich 2024, Theorem 3)

$$\|\hat{p} - p\|_\infty \leq 2 \sqrt{\frac{\max_u p_u(1 - p_u)}{n} \log \frac{2}{\alpha} + \frac{V^*}{n} + \frac{4}{3n} \log \frac{2(n+1)}{\alpha} + \frac{\log(n)}{n}}$$

w.p.  $1 - \alpha$ , where  $V^* = \max_{u \in \mathbb{N}} p_{[u]} (1 - p_{[u]}) \log(1 + u)$

- Comparing to our dream result  $\|\hat{p} - p\|_\infty \stackrel{?}{\lesssim} \sqrt{\frac{\max_u \hat{p}_u(1 - \hat{p}_u)}{n} \log \frac{1}{\alpha} + \frac{1}{n} \log \frac{1}{\alpha}}$

# Main Result

## Theorem 3 (P. and Kontorovich 2024, Theorem 3)

$$\|\hat{p} - p\|_\infty \leq 2 \sqrt{\frac{\max_u p_u(1 - p_u)}{n} \log \frac{2}{\alpha} + \frac{V^*}{n} + \frac{4}{3n} \log \frac{2(n+1)}{\alpha} + \frac{\log(n)}{n}}$$

w.p.  $1 - \alpha$ , where  $V^* = \max_{u \in \mathbb{N}} p_{[u]} (1 - p_{[u]}) \log(1 + u)$

- Comparing to our dream result  $\|\hat{p} - p\|_\infty \stackrel{?}{\lesssim} \sqrt{\frac{\max_u \hat{p}_u(1 - \hat{p}_u)}{n} \log \frac{1}{\alpha} + \frac{1}{n} \log \frac{1}{\alpha}}$
- We get extra  $\log(n)/n$  and  $V^*/n$  terms. We show that these are inevitable by proving matching lower bounds

# Main Result

## Theorem 3 (P. and Kontorovich 2024, Theorem 3)

$$\|\hat{p} - p\|_\infty \leq 2 \sqrt{\frac{\max_u p_u(1 - p_u)}{n} \log \frac{2}{\alpha} + \frac{V^*}{n} + \frac{4}{3n} \log \frac{2(n+1)}{\alpha} + \frac{\log(n)}{n}}$$

w.p.  $1 - \alpha$ , where  $V^* = \max_{u \in \mathbb{N}} p_{[u]}(1 - p_{[u]}) \log(1 + u)$

- Comparing to our dream result  $\|\hat{p} - p\|_\infty \stackrel{?}{\lesssim} \sqrt{\frac{\max_u \hat{p}_u(1 - \hat{p}_u)}{n} \log \frac{1}{\alpha} + \frac{1}{n} \log \frac{1}{\alpha}}$ 
  - We get extra  $\log(n)/n$  and  $V^*/n$  terms. We show that these are inevitable by proving matching lower bounds
  - We can replace  $p_u$  with  $\hat{p}_u$  at some relatively small cost

# How Good is this Proxy?

**Theorem 4 (P. and Kontorovich 2024, Theorem 8)** *assume there exists  $U(X^n)$  such that  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$ . Then,*

$$\mathbb{E}(U(X^n)) \geq z_{\alpha/2} \sqrt{\frac{\max_u p_u(1 - p_u)}{n}} + o\left(\frac{1}{n}\right)$$

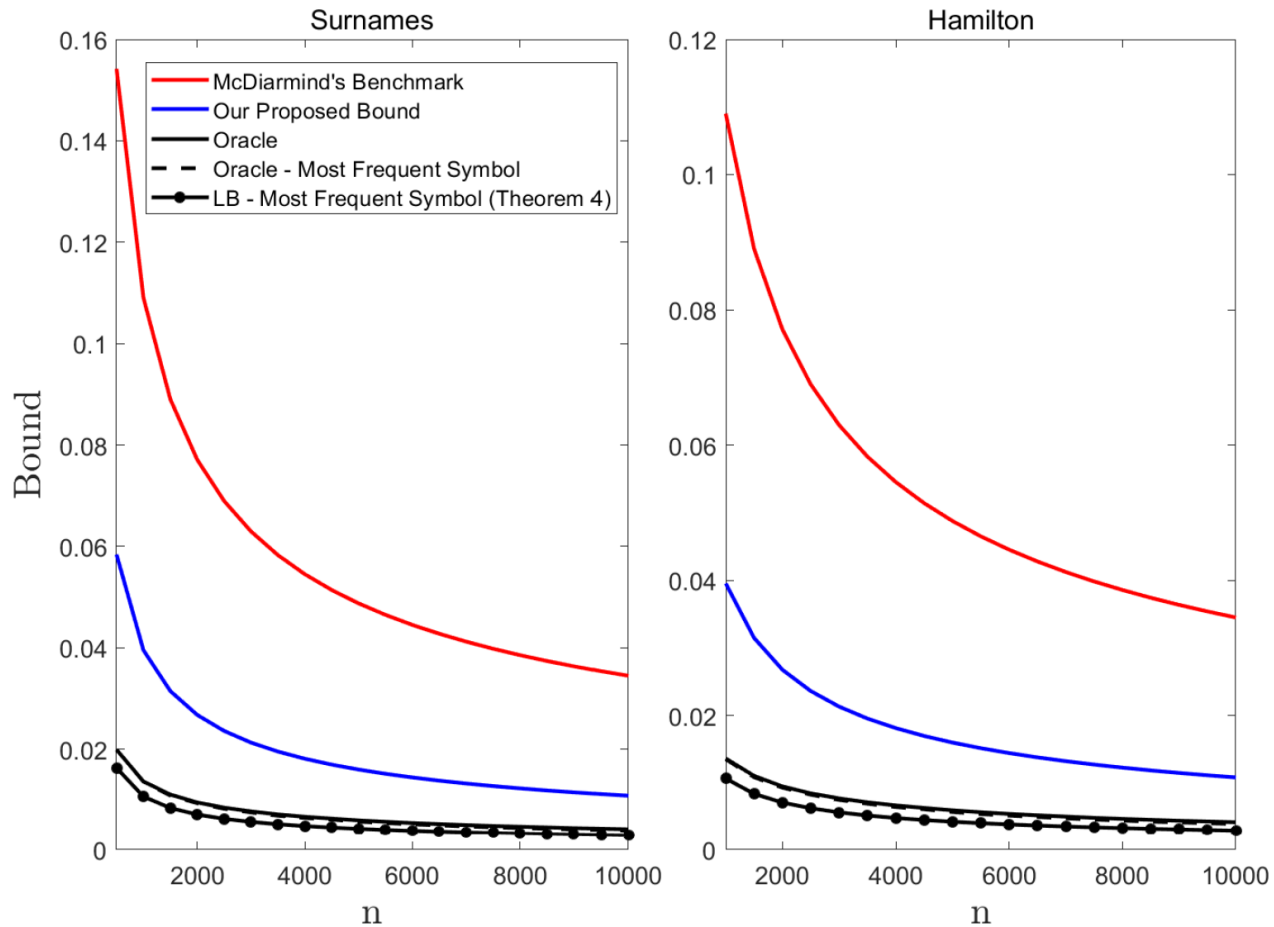
# How Good is this Proxy?

**Theorem 4 (P. and Kontorovich 2024, Theorem 8)** *assume there exists  $U(X^n)$  such that  $P(|\hat{p}_{u^*} - p_{u^*}| \leq U(X^n)) \geq 1 - \alpha$ . Then,*

$$\mathbb{E}(U(X^n)) \geq z_{\alpha/2} \sqrt{\frac{\max_u p_u(1 - p_u)}{n}} + o\left(\frac{1}{n}\right)$$

- The minimal expected length over our desired CI is almost the same as our “dream” infinity norm bound

# How Good is this Proxy?



# Are We Done?

- Distribution estimation under infinity norm for selective inference

# Are We Done?

- Distribution estimation under infinity norm for selective inference
- Improve our understanding of the most frequent events in the sample

# Are We Done?

- Distribution estimation under infinity norm for selective inference
- Improve our understanding of the most frequent events in the sample
  - Is the infinity norm the best proxy we can find?

# Are We Done?

- Distribution estimation under infinity norm for selective inference
- Improve our understanding of the most frequent events in the sample
  - Is the infinity norm the best proxy we can find?
  - Can we obtain tighter bounds? Improved (and simpler) schemes?

# Are We Done?

- Distribution estimation under infinity norm for selective inference
- Improve our understanding of the most frequent events in the sample
  - Is the infinity norm the best proxy we can find?
  - Can we obtain tighter bounds? Improved (and simpler) schemes?

# Thank you!