Canonical Estimation in a Rare Events Regime

Mesrob Ohannessian*, **Vincent Y. F. Tan**†, Munther Dahleh*

† Department of ECE, University of Wisconsin-Madison

*LIDS, MIT

SILO (Oct 2011)

Challenge of Data Sparsity

- We have more and more data

Challenge of Data Sparsity

- We have more and more data
- Today's Google Corpus (1 billion words)

Challenge of Data Sparsity

- We have more and more data
- Today's Google Corpus (1 billion words)
- Yet we do not have enough data

Challenge of Data Sparsity

- We have more and more data
- Today's Google Corpus (1 billion words)
- Yet we do not have enough data
- Classical statistics involve an alphabet $\mathcal{A}$ and a pmf $p \in \mathcal{P}(\mathcal{A})$

$$\mathcal{P}(\mathcal{A}) := \left\{ \mathbf{p} \in \mathbb{R}^{|\mathcal{A}|} : p(a) \geq 0, \sum_{a \in \mathcal{A}} p(a) = 1 \right\}$$

Challenge of Data Sparsity

- We have more and more data
- Today's Google Corpus (1 billion words)
- Yet we do not have enough data
- Classical statistics involve an alphabet $\mathcal{A}$ and a pmf $p \in \mathcal{P}(\mathcal{A})$

$$\mathcal{P}(\mathcal{A}) := \left\{ \mathbf{p} \in \mathbb{R}^{|\mathcal{A}|} : p(a) \geq 0, \sum_{a \in \mathcal{A}} p(a) = 1 \right\}$$

- $X_1, \ldots, X_n$ are independent samples from p
- Law of large numbers:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}X$$

The WVK model for large alphabets

- Sequence of alphabets $\mathcal{A}_n$ and a sequence of pmfs $p_n \in \mathcal{P}(\mathcal{A}_n)$

The WVK model for large alphabets

- Sequence of alphabets $\mathcal{A}_n$ and a sequence of pmfs $p_n \in \mathcal{P}(\mathcal{A}_n)$
- To each n, we have an i.i.d. sequence of samples

$$X_{1,n}, X_{2,n}, \ldots, X_{n,n} \sim \prod_{i=1}^{n} p_n(x_{n,i})$$

The WVK model for large alphabets

- Sequence of alphabets $\mathcal{A}_n$ and a sequence of pmfs $p_n \in \mathcal{P}(\mathcal{A}_n)$
- To each n, we have an i.i.d. sequence of samples

$$X_{1,n}, X_{2,n}, \ldots, X_{n,n} \sim \prod_{i=1}^{n} p_n(x_{n,i})$$

- Let P_n be the law of $np_n(X_n)$ where $X_n \sim p_n$

The WVK model for large alphabets

- Sequence of alphabets $\mathcal{A}_n$ and a sequence of pmfs $p_n \in \mathcal{P}(\mathcal{A}_n)$
- To each n, we have an i.i.d. sequence of samples

$$X_{1,n}, X_{2,n}, \ldots, X_{n,n} \sim \prod_{i=1}^{n} p_n(x_{n,i})$$

- Let P_n be the law of $np_n(X_n)$ where $X_n \sim p_n$

Definition (Wagner, Viswanath and Kulkarni, IT-Trans 2011)

We say that $\{(\mathcal{A}_n, p_n)\}_{n\in\mathbb{N}}$ is a rare-events source if

$$\frac{\check{c}}{n} \le p_n(a) \le \frac{\hat{c}}{n}, \qquad \forall\, a \in \mathcal{A}_n$$

and $\exists\, P$ such that $P_n \Rightarrow P$.

The WVK model for large alphabets

- Sequence of alphabets $\mathcal{A}_n$ and a sequence of pmfs $p_n \in \mathcal{P}(\mathcal{A}_n)$
- To each n, we have an i.i.d. sequence of samples

$$X_{1,n}, X_{2,n}, \ldots, X_{n,n} \sim \prod_{i=1}^{n} p_n(x_{n,i})$$

- Let P_n be the law of $np_n(X_n)$ where $X_n \sim p_n$

Definition (Wagner, Viswanath and Kulkarni, IT-Trans 2011)

We say that $\{(\mathcal{A}_n, p_n)\}_{n\in\mathbb{N}}$ is a rare-events source if

$$\frac{\check{c}}{n} \le p_n(a) \le \frac{\hat{c}}{n}, \qquad \forall\, a \in \mathcal{A}_n$$

and $\exists\, P$ such that $P_n \Rightarrow P$. Note $\mathrm{supp}(P) \subseteq \mathcal{C} := [\check{c}, \hat{c}]$.

What are we interested in?

- $\mathcal{A}_n, p_n, P_n, P$ are all unknown

What are we interested in?

- $\mathcal{A}_n, p_n, P_n, P$ are all unknown
- Using the samples $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$, WVK estimated

Probabilities	$p_n^n(X_{n,1}, \ldots, X_{n,n})$
Entropy	$H(p_n)$
Relative Entropy	$D(p_n \| q_n)$

What are we interested in?

- $\mathcal{A}_n, p_n, P_n, P$ are all unknown
- Using the samples $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$, WVK estimated

Probabilities	$p_n^n(X_{n,1}, \ldots, X_{n,n})$
Entropy	$H(p_n)$
Relative Entropy	$D(p_n \vert\vert q_n)$

- Note the independence from reordering of the symbols of $\mathcal{A}_n$

What are we interested in?

- $\mathcal{A}_n, p_n, P_n, P$ are all unknown
- Using the samples $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$, WVK estimated

Probabilities	$p_n^n(X_{n,1}, \ldots, X_{n,n})$
Entropy	$H(p_n)$
Relative Entropy	$D(p_n \| q_n)$

- Note the independence from reordering of the symbols of $\mathcal{A}_n$
- Other quantities?

Alphabet size	$\lvert\mathcal{A}_n\rvert$
Range of probabilities	$\mathcal{C} := [\check{c}, \hat{c}]$

What are we interested in?

- $\mathcal{A}_n, p_n, P_n, P$ are all unknown
- Using the samples $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$, WVK estimated

$$\begin{array}{rl} \text{Probabilities} & p_n^n(X_{n,1}, \ldots, X_{n,n}) \\ \text{Entropy} & H(p_n) \\ \text{Relative Entropy} & D(p_n || q_n) \end{array}$$

- Note the independence from reordering of the symbols of $\mathcal{A}_n$
- Other quantities?

$$\begin{array}{rl} \text{Alphabet size} & |\mathcal{A}_n| \\ \text{Range of probabilities} & \mathcal{C} := [\check{c}, \hat{c}] \end{array}$$

- Can we estimate all reasonable quantities in a universal manner?

Canonical Estimation Problems

Let $\{Y_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued random variables such that

- There exists continuous $f_n(x)$ that converge to $f(x)$ pointwise on $\mathcal{C}$

$$\mathbb{E}[Y_n] = \int_{\mathcal{C}} f_n(x)\, dP_n(x)$$

- $|Y_n - \mathbb{E}[Y_n]| \to 0$ almost surely

Canonical Estimation Problems

Let $\{Y_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued random variables such that

- There exists continuous $f_n(x)$ that converge to $f(x)$ pointwise on $\mathcal{C}$

$$\mathbb{E}[Y_n] = \int_{\mathcal{C}} f_n(x)\, dP_n(x)$$

- $|Y_n - \mathbb{E}[Y_n]| \to 0$ almost surely

By Skorohod's representation theorem, $Y_n \to \int_{\mathcal{C}} f(x)\, dP(x)$ a.s.

Canonical Estimation Problems

Let $\{Y_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued random variables such that

- There exists continuous $f_n(x)$ that converge to $f(x)$ pointwise on $\mathcal{C}$

$$\mathbb{E}[Y_n] = \int_{\mathcal{C}} f_n(x)\, dP_n(x)$$

- $|Y_n - \mathbb{E}[Y_n]| \to 0$ almost surely

By Skorohod's representation theorem, $Y_n \to \int_{\mathcal{C}} f(x)\, dP(x)$ a.s.

Definition

An estimator $\{\hat{Y}_n : \mathcal{A}_n^n \to \mathbb{R}\}_{n\in\mathbb{N}}$ is consistent if

$$\hat{Y}_n(X_{n,1}, \ldots, X_{n,n}) \to \int_{\mathcal{C}} f(x)\, dP(x)$$

almost surely.

All previous examples were canonical

- Probabilities

$$Y_n = \frac{1}{n} \log p_n^n(X_{n,1}, \ldots, X_{n,n}) + \log n, \qquad f(x) = \log x$$

All previous examples were canonical

- Probabilities

$$Y_n = \frac{1}{n} \log p_n^n(X_{n,1}, \ldots, X_{n,n}) + \log n, \qquad f(x) = \log x$$

- Entropies

$$Y_n = H(p_n) - \log n, \qquad f(x) = -\log x$$

All previous examples were canonical

- Probabilities

$$Y_n = \frac{1}{n} \log p_n^n(X_{n,1}, \ldots, X_{n,n}) + \log n, \qquad f(x) = \log x$$

- Entropies

$$Y_n = H(p_n) - \log n, \qquad f(x) = -\log x$$

- Alphabet size

$$Y_n = \frac{|\mathcal{A}_n|}{n}, \qquad f(x) = \frac{1}{x}$$

All previous examples were canonical

- Probabilities

$$Y_n = \frac{1}{n} \log p_n^n(X_{n,1}, \ldots, X_{n,n}) + \log n, \qquad f(x) = \log x$$

- Entropies

$$Y_n = H(p_n) - \log n, \qquad f(x) = -\log x$$

- Alphabet size

$$Y_n = \frac{|\mathcal{A}_n|}{n}, \qquad f(x) = \frac{1}{x}$$

- Range of probabilities

$$C = [\check{c}, \hat{c}], \qquad f(x) = x^q,$$

All previous examples were canonical

- Probabilities

$$Y_n = \frac{1}{n} \log p_n^n(X_{n,1}, \ldots, X_{n,n}) + \log n, \qquad f(x) = \log x$$

- Entropies

$$Y_n = H(p_n) - \log n, \qquad f(x) = -\log x$$

- Alphabet size

$$Y_n = \frac{|\mathcal{A}_n|}{n}, \qquad f(x) = \frac{1}{x}$$

- Range of probabilities

$$C = [\check{c}, \hat{c}], \qquad f(x) = x^q, \qquad \hat{c} = \lim_{q\to\infty} \left[\int_{\mathcal{C}} x^q \, dP(x) \right]^{1/q}$$

Estimate by imitating the source

- Our strategy is to estimate the shadow P_n, the law of $np_n(X_n)$

Estimate by imitating the source

- Our strategy is to estimate the shadow P_n, the law of $np_n(X_n)$
- If we can construct a sequence of measures $\hat{P}_n$ such that

$$\hat{P}_n(X_{n,1}, \ldots, X_{n,n}) \Rightarrow P$$

almost surely...

Estimate by imitating the source

- Our strategy is to estimate the shadow P_n, the law of $np_n(X_n)$
- If we can construct a sequence of measures $\hat{P}_n$ such that

$$\hat{P}_n(X_{n,1}, \ldots, X_{n,n}) \Rightarrow P$$

almost surely...

- Then by integrating against the correct limiting function f,

$$\hat{Y}_n(X_{n,1}, \ldots, X_{n,n}) = \int_{\mathcal{C}} f(x)\, d\hat{P}_n(x)$$

we have a consistent estimator.

Estimate by imitating the source

- Our strategy is to estimate the shadow P_n, the law of $np_n(X_n)$
- If we can construct a sequence of measures $\hat{P}_n$ such that

$$\hat{P}_n(X_{n,1}, \ldots, X_{n,n}) \Rightarrow P$$

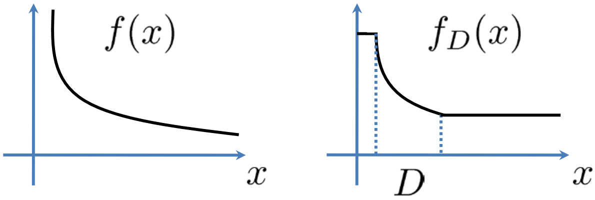
  almost surely...
- Then by integrating against the correct limiting function f,

$$\hat{Y}_n(X_{n,1}, \ldots, X_{n,n}) = \int_{\mathcal{C}} f(x)\, d\hat{P}_n(x)$$

  we have a consistent estimator.
- Problems:
  - Support $\mathcal{C}$ isn't known
  - $f(x)$ doesn't have to be bounded everywhere
  - How to get the estimate $\hat{P}_n(x)$?

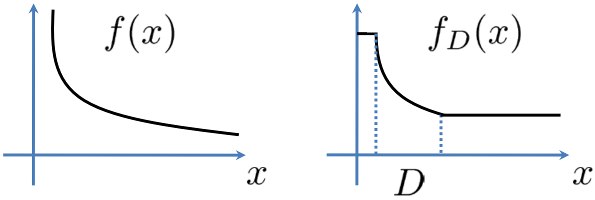
Estimate the function

Consider a tapered version of f



Estimate the function

Consider a tapered version of f



Lemma (Ohannessian-Tan-Dahleh)

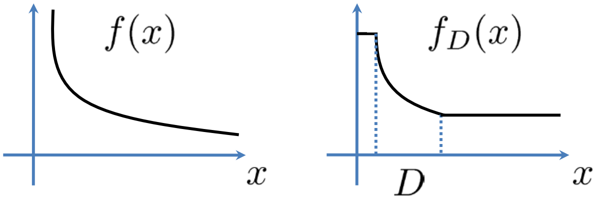
If $\mathcal{C} \subset \mathcal{D}$, then

$$\hat{Y}_n := \int_{\mathcal{C}} f_D(x)\, d\hat{P}_n(x)$$

is a consistent estimator

Estimate the function

Consider a tapered version of f



Lemma (Ohannessian-Tan-Dahleh)

If $\mathcal{C} \subset \mathcal{D}$, then

$$\hat{Y}_n := \int_{\mathcal{C}} f_D(x)\, d\hat{P}_n(x)$$

is a consistent estimator

If $\mathcal{C}$ is unknown, just let $\mathcal{D}$ grow gradually with n

Estimate with rates

Recall the Wasserstein distance

$$d_W(P, Q) = \sup_{h \in Lip(1)} \left| \int_{\mathbb{R}^+} h \, dP - \int_{\mathbb{R}^+} h \, dQ \right|$$

Estimate with rates

Recall the Wasserstein distance

$$d_W(P,Q) = \sup_{h \in Lip(1)} \left| \int_{\mathbb{R}^+} h\, dP - \int_{\mathbb{R}^+} h\, dQ \right|$$

Lemma (Ohannessian-Tan-Dahleh)

If

$$Lip(f_{D_n})\, d_W(\hat{P}_n, P) \to 0,$$

then,

$$\hat{Y}_n := \int_{\mathbf{R}^+} f_{D_n}(x)\, d\hat{P}_n(x)$$

is consistent.

Estimate with rates

Recall the Wasserstein distance

$$d_W(P,Q) = \sup_{h \in Lip(1)} \left| \int_{\mathbb{R}^+} h\, dP - \int_{\mathbb{R}^+} h\, dQ \right|$$

Lemma (Ohannessian-Tan-Dahleh)

If

$$Lip(f_{D_n})\, d_W(\hat{P}_n, P) \to 0,$$

then,

$$\hat{Y}_n := \int_{\mathbf{R}^+} f_{D_n}(x)\, d\hat{P}_n(x)$$

is consistent.

How to estimate the shadow P_n?

Pseudo-Empirical Measure

Good-Turing estimator:

- Denote the set of symbols that appear k times as $\mathcal{B}_{n,k} \subset \mathcal{A}_n$
- Denote their probabilities as

$$\gamma_{n,k} = p_n(\mathcal{B}_{n,k}) = \sum_{a \in \mathcal{B}_{n,k}} p_n(a)$$

Pseudo-Empirical Measure

Good-Turing estimator:

- Denote the set of symbols that appear k times as $\mathcal{B}_{n,k} \subset \mathcal{A}_n$
- Denote their probabilities as

$$\gamma_{n,k} = p_n(\mathcal{B}_{n,k}) = \sum_{a \in \mathcal{B}_{n,k}} p_n(a)$$

- Estimate these using the Good-Turing estimator:

$$\phi_{n,k} := (k+1)\frac{|\mathcal{B}_{n,k+1}|}{n}$$

Pseudo-Empirical Measure

Good-Turing estimator:

- Denote the set of symbols that appear k times as $\mathcal{B}_{n,k} \subset \mathcal{A}_n$
- Denote their probabilities as

$$\gamma_{n,k} = p_n(\mathcal{B}_{n,k}) = \sum_{a \in \mathcal{B}_{n,k}} p_n(a)$$

- Estimate these using the Good-Turing estimator:

$$\phi_{n,k} := (k+1) \frac{|\mathcal{B}_{n,k+1}|}{n}$$

- E.g.: Probability of missing mass $\approx \phi_{n,0}$ [Budianu and Tong 2004]

Pseudo-Empirical Measure

Strong law of large numbers gives:

Lemma (WVK)

Let the P-Poisson mixture be

$$\lambda_k^P = \int_{\mathbb{R}^+} \frac{e^{-x} x^k}{k!} \, dP(x), \qquad k = 0, 1, \ldots$$

Pseudo-Empirical Measure

Strong law of large numbers gives:

Lemma (WVK)

Let the P-Poisson mixture be

$$\lambda_k^P = \int_{\mathbb{R}^+} \frac{e^{-x} x^k}{k!} \, dP(x), \qquad k = 0, 1, \ldots$$

Then, $\|\gamma_n - \lambda^P\|_1 \to 0$ *and* $\|\phi_n - \lambda^P\|_1 \to 0$ *almost surely.*

Pseudo-Empirical Measure

Strong law of large numbers gives:

Lemma (WVK)

Let the P-Poisson mixture be

$$\lambda_k^P = \int_{\mathbb{R}^+} \frac{e^{-x} x^k}{k!} \, dP(x), \qquad k = 0, 1, \ldots$$

Then, $\|\gamma_n - \lambda^P\|_1 \to 0$ *and* $\|\phi_n - \lambda^P\|_1 \to 0$ *almost surely.*

Theorem (Ohannessian-Tan-Dahleh)

For "most natural" rare event sources, there exist an $s > 0$ *such that*

$$n^s \sup_{k \in \mathbb{N}} |F_{\phi_n}(k) - F_{\lambda^P}(k)| \to 0, \qquad a.s.$$

(Kolmogorov-Smirnov convergence)

Estimation of $\hat{P}_n(x)$ via mixture distribution learning

Theorem

The (pseudo) maximum-likelihood estimator

$$\hat{P}_n^{ML} = \arg\min_Q \; D(\phi_n \,||\, Q)$$

is a valid construction, i.e., $\hat{P}_n^{ML} \Rightarrow P$ almost surely.

Estimation of $\hat{P}_n(x)$ via mixture distribution learning

Theorem

The (pseudo) maximum-likelihood estimator

$$\hat{P}_n^{ML} = \arg\min_Q D(\phi_n \,||\, Q)$$

is a valid construction, i.e., $\hat{P}_n^{ML} \Rightarrow P$ *almost surely.*

Theorem

The minimum distance estimator

$$\hat{P}_n^{MD} = \arg\min_Q \sup_{k \in \mathbb{N}} \left| F_{\phi_n}(k) - F_{Poi(Q)}(k) \right|$$

is also valid. Furthermore, there exists $s > 0$ *such that* $n^s d_W(\hat{P}_n, P) \to 0$ *almost surely (with some technical conditions).*

Estimating Entropies

- Normalized entropy: $Y_n = H(p_n) - \log n$

Estimating Entropies

- Normalized entropy: $Y_n = H(p_n) - \log n$
- Canonical with $f(x) = -\log x$

Estimating Entropies

- Normalized entropy: $Y_n = H(p_n) - \log n$
- Canonical with $f(x) = -\log x$
- f is D_n-Lipschitz on $\mathcal{D} := [D_n^{-1}, D_n]$

Estimating Entropies

- Normalized entropy: $Y_n = H(p_n) - \log n$
- Canonical with $f(x) = -\log x$
- f is D_n-Lipschitz on $\mathcal{D} := [D_n^{-1}, D_n]$

Lemma

With $D_n = o(n^s)$,

$$\hat{Y}_n := \int_{\mathbb{R}^+} (-\log x)_{D_n} \, d\hat{P}_n(x)$$

is consistent

Estimating Alphabet Size

- Normalized alphabet size: $Y_n = |\mathcal{A}_n|/n$

Estimating Alphabet Size

- Normalized alphabet size: $Y_n = |\mathcal{A}_n|/n$
- Canonical with $f(x) = \frac{1}{x}$ [Bhat and Sporat 2004]

Estimating Alphabet Size

- Normalized alphabet size: $Y_n = |\mathcal{A}_n|/n$
- Canonical with $f(x) = \frac{1}{x}$ [Bhat and Sporat 2004]
- f is D_n^2-Lipschitz on $\mathcal{D} := [D_n^{-1}, D_n]$

Estimating Alphabet Size

- Normalized alphabet size: $Y_n = |\mathcal{A}_n|/n$
- Canonical with $f(x) = \frac{1}{x}$ [Bhat and Sporat 2004]
- f is D_n^2-Lipschitz on $\mathcal{D} := [D_n^{-1}, D_n]$

Lemma

With $D_n = o(n^{s/2})$,

$$\hat{Y}_n := \int_{\mathbb{R}^+} \left(\frac{1}{x}\right)_{D_n} d\hat{P}_n(x)$$

is consistent

Estimating support $\mathcal{C} = [\check{c}, \hat{c}]$

- Not quite canonical but close.
- Let Z be the weak limit of $Z_n := np_n(X_n)$ and let $P := Z_*(\mathbb{P})$. Then

$$\hat{c} = \operatorname*{esssup}_{\omega} Z(\omega) = \lim_{q \to \infty} \left[\int_{\mathcal{C}} x^q \, dP(x) \right]^{1/q}$$

Estimating support $\mathcal{C} = [\check{c}, \hat{c}]$

- Not quite canonical but close.
- Let Z be the weak limit of $Z_n := np_n(X_n)$ and let $P := Z_*(\mathbb{P})$. Then

$$\hat{c} = \operatorname*{esssup}_{\omega} Z(\omega) = \lim_{q \to \infty} \left[\int_{\mathcal{C}} x^q \, dP(x) \right]^{1/q}$$

- Thus, "$f(x) = x^q$". Let q grow with n too!

Estimating support $\mathcal{C} = [\check{c}, \hat{c}]$

- Not quite canonical but close.
- Let Z be the weak limit of $Z_n := np_n(X_n)$ and let $P := Z_*(\mathbb{P})$. Then

$$\hat{c} = \operatorname*{esssup}_{\omega} Z(\omega) = \lim_{q \to \infty} \left[\int_{\mathcal{C}} x^q \, dP(x) \right]^{1/q}$$

- Thus, "$f(x) = x^q$". Let q grow with n too!

Lemma

With $q_n = \frac{\log n}{\log \log n}$ and $D_n = o(n^{\frac{s}{2q_n}})$,

$$\hat{Y}_n := \left[\int_{\mathbb{R}^+} (x^{q_n})_{D_n} \, d\hat{P}_n(x) \right]^{1/q_n}$$

is consistent for estimating $\hat{c}$

Conclusions

- Challenge: Large alphabets and data scarcity

Conclusions

- Challenge: Large alphabets and data scarcity
- Model: WVK's rare-events regime model

Conclusions

- Challenge: Large alphabets and data scarcity
- Model: WVK's rare-events regime model
- Problems: Canonical estimation

Conclusions

- Challenge: Large alphabets and data scarcity
- Model: WVK's rare-events regime model
- Problems: Canonical estimation
- Solution: Imitating the source

Conclusions

- Challenge: Large alphabets and data scarcity
- Model: WVK's rare-events regime model
- Problems: Canonical estimation
- Solution: Imitating the source
  - Abstract methods
  - Concrete constructions

Conclusions

- Challenge: Large alphabets and data scarcity
- Model: WVK's rare-events regime model
- Problems: Canonical estimation
- Solution: Imitating the source
  - Abstract methods
  - Concrete constructions
- Future work: Further analysis of convergence rates