ECEN 689
Special Topics in Data Science for Communications Networks

Nick Duffield
Department of Electrical & Computer Engineering
Texas A&M University

Lecture 3
Estimation and Bounds
Estimation from Packet Sampled Flow Records

- Original flows: (key, #packets) = (k, p)
- Generic subset sum query:
  - True value $X(K) = \#\text{packets in flows with key } k \in K$
- How to estimate from packet sampled flow records:
  - Intuition: sampling yields about $1/N^{th}$ of the original packets
  - Multiply number of sampled packets by $N$ to estimate #original packets
- Key idea: can apply the idea to any subset of packets:
  - Estimate: $X(K)$ by $X'(K) = N \times \#\text{packets in sampled flows with key } k \in K$
Unbiased Estimation

- Consider a single item (e.g. packet)
- Equipped with a size $x$ (e.g. #bytes)
- Sampled with probability $q$
- The **Horvitz-Thompson Estimator** of $x$ is

\[
    x' = \begin{cases} 
    x/q & \text{if the item is selected} \\
    0 & \text{otherwise} 
    \end{cases}
\]

- **Theorem**: $x'$ is **unbiased** estimator of $x$: Expectation $E[x'] = x$
  
  $E[x] = \text{sum of estimate values weighted by their probabilities}$
  
  $= \Pr[\text{item sampled}] \times x/q + \Pr[\text{item not sampled}] \times 0$
  
  $= q \times x/q + (1-q) \times 0$
  
  $= x$
Subset Sum Estimation with Horvitz-Thompson

- Set of items with weights \( \{x_1, x_2, \ldots, x_n\} \)
- Individual sampling probabilities \( \{q_1, q_2, \ldots, q_n\} \)
- Subset sum \( X(S) = \sum_{i \in S} x_i \) for any subset \( S \subset \{1,2,\ldots,n\} \)
- Estimate subset sum using sum of constituent estimates
  \[
  X'(S) = \sum_{i \in S} x'_i \text{ where } x'_i = \frac{x_i}{q_i} \text{ if } i \text{ selected, } 0 \text{ otherwise}
  \]
- Theorem: \( X'(S) \) is an unbiased estimator of \( X(S) \)
  - Proof uses linearity of expectation
    \[
    E[X'(S)] = E[\sum_{i \in S} x'_i] = \sum_{i \in S} E[x'_i] = \sum_{i \in S} x_i = X(S)
    \]
    - Doesn’t require independent sampling

- More on Horvitz-Thompson: Kolaczyk. Ch. 5.2
- See also Duffield review article
Subset Sum Estimation for Packets (and Bytes)

- $X(K) = \#\{ \text{packets in samples flows with key } k \in K \}$
  
  
  $= \sum_{i \in S(K)} x_i \ , \ x_i = 1$; $S(K) = \text{packets with key } k \in K$

- For brevity will mostly write $X$ for $X(K)$ and $S$ for $S(K)$

- Packets sampled with probability $1/N$

- $X' = N^* \#\{ \text{packets in sampled flows with key } k \in K \}$

  
  $= \sum_{i \in S(K)} x'_i \ \text{with } x'_i = N \text{ if packet is sampled, } 0 \text{ otherwise}$

- Recognize $X'$ as the HT estimator of $X$: $E[X'] = X$

- Can make similar construction for byte estimation
Estimation Accuracy

• Think of independent $1$ in $N$ sampling from $X$ packets
• Number of samples has binomial $B(X, 1/N)$ distribution
• Estimate $X'$ is distributed as $N \times B(X, 1/N)$ i.e.

$$
\Pr[X'=xN] = \binom{X}{x} N^{-x} (1 - 1/N)^{x-x}
$$

• Example:
  - $X = 5,000$,
  - $N = 100$,
  - 1,000 trials of $X'$
Accuracy and Bounds

- HT estimation: weights \( \{x_1, \ldots, x_n\} \) & probabilities \( \{q_1, q_2, \ldots, q_n\} \)
- Initially examine accuracy of subsets sums \( X = X(K) \) via variance:

\[
\text{Var}(X') = E[ (X' - X)^2 ] = E[ (X')^2 ] - X^2
\]

- Assume independent sampling for the moment: \( x'_i \) are independent

\[
\text{Var}(X') = \text{Var}(\sum_{i \in S} x'_i) = \sum_{i \in S} \text{Var}(x'_i)
\]

- Compute variance for HT estimate of individuals \( x'_i \):

- Exercise:

\[
\text{Var}(x'_i) = x_i^2 (1/p_i - 1) \text{ and so } \text{Var}(X') = \sum_{i \in S} x_i^2 (1/p_i - 1)
\]
Confidence Intervals for HT estimates?

• Common to think of confidence intervals in terms of the normal distribution
  – Because the width of the distribution is characterized by the variance

• But distribution of HT estimator not normal; so either
  1. Find confidence intervals using variance without normal assumption
  2. Exploit simple form of the distribution of the estimator
Towards Variance Based Bounds

• Theorem (Markov inequality)
  – Let $Y$ be non-negative random variable with $E[Y] < \infty$
  – Then $\Pr[ Y \geq a ] \leq E[Y] / a$

• Proof:
  
  $$E[Y] = E[Y | Y < a] \cdot \Pr[ Y < a] + E[Y | Y \geq a] \cdot \Pr[ Y \geq a]$$
  
  $$\geq 0 + a \cdot \Pr[ Y \geq a]$$

• Use: take $a = n \cdot E[Y]$. Then $\Pr[ Y \geq n \cdot E[Y] ] \leq 1/n$
  
  – Probability $Y$ exceeds its expectation by factor $n$ is less than $1/n$

• Very general, but does not depend on (or exploit) variance

• More on bounds: Mitzenmacher & Upfal, Ch. 3
Variance Based Bounds

• Markov Inequality: \( \Pr[ Y \geq a ] \leq E[Y] / a \), for \( Y > 0, E[Y] < \infty \)
• Given r.v. \( Z \), set \( Y = (Z - E[Z])^2 \) and \( a = b^2 \)
• Apply Markov inequality (assume \( E[(Z - E[Z])^2 = \text{Var}(Z) < \infty] \))

\[
\Pr[ (Z - E[Z])^2 \geq b^2 ] \leq E[ (Z - E[Z])^2 ] / b^2
\]

• In other words (this is called Chebychev’s Inequality)

\[
\Pr[ |Z - E[Z]| \geq b ] \leq \text{Var}[Z] / b^2
\]

• Interpretation: set \( b = n \ast \text{Sqrt}(\text{Var}[Z]) = n \ast \text{StdDev}(Z) \)
  
  – Probability that \( Z \) further than \( n \ast \text{StdDev} \) away from mean is \( \leq 1/n^2 \)
Application to Packet Sampling

- Estimate packet counts \( (x_i = 1) \); uniform sampling prob \( 1/N \)

- Chebychev Inequality for \( X' = X'(K) \)
  \[
  \text{Var}(X') = \sum_{i \in S} \text{Var}(x'_i) = X^*(N-1)
  \]
  \[
  \text{Prob}( | X' - X | \geq b ) \leq X^*(N-1) / b^2
  \]

- Rewriting with \( b = n \times X \)
  - a large deviation (of \( X'(K) \) proportional to its expectation \( X \))
    
    \[
    \text{Prob}( | X' - X | \geq n \times X ) \leq (N-1) / (n^2 \times X)
    \]

- Behavior of bound
  - Increasing in \( N \): less certainty if fewer sampled packets
  - Decreasing in \( X \): more certainty if estimating a larger subset
Exponential Bounds

• Recall Markov Inequality:
  \[ \Pr[ Y \geq a ] \leq \frac{E[Y]}{a} \text{ for } Y \geq 0 \text{ with } E[Y] < \infty \]

• Obtained Chebychev using \( Y = (Z-E(Z))^2 \). Try similar.

• For a r.v. \( Z \), let \( Y = e^{tZ} \) for \( t \geq 0 \), and set \( a = \exp(tz_0) \).
  \{Z \geq z_0\} = \{Y \geq \exp(tz_0)\} \text{ since } \exp(tz_0) \text{ is increasing in } z_0 \text{ when } t \geq 0

• Apply Markov Inequality: \( \Pr[Z \geq z_0] \leq \exp(-tz_0)E[e^{tZ}] \)

• Which value of \( t \) to take?
  – The value of \( t \) that minimizes the upper bound!
  \[ \Pr[Z \geq z_0] \leq \min_{t \geq 0} \exp(-tz_0)E[e^{tZ}] \]
Exponential Bounds and Packet Sampling

• Independent packet sampling with probability $q = 1/N$
• Exponential bound $\Pr[Z \geq z_0] \leq \min_{t \geq 0} \exp(tz_0) \ E[\exp(tZ)]$
  
  Take $Z = X'=X'(K)$ and $z = (1+\delta)X$ (large deviation)

• Product form: $\exp(tX') = \exp(t\sum_{i \in S} x'_i) = \prod_{i \in S} \exp(t x'_i)$
• $x'_i$ are independent, so expectation preserves product
  
  $E[\exp(t X')] = \prod_{i \in S} E[\exp(t x'_i)] = E[\exp(t x'_1)]^X$

• Exponential moment for single $x'_i$
  
  $E[\exp(t x'_1)] = 1^* \Pr[x'_1=0] + e^{t/q} * \Pr[x'_1=1/q] = (1-q) + q*e^{t/q}$

• Combining:
  
  $\Pr[ X' \geq (1+\delta)X ] \leq f(q, \delta)^X$ with $f(q, \delta) = \min_{t \geq 0} e^{-t(1+\delta)} (1+q (e^{t/q}-1))$
Minimization and Chernoff Bounds

• Common to use further bound \(1+q \left( e^{t/q} - 1 \right) \leq \text{Exp} \ q(e^{t/q} - 1)\)
  – More convenient final result, especially if non-equal probabilities \(q_i\)

• \(f(q, \delta) \leq g(q, \delta) = \min_{t \geq 0} e^{-t(1+\delta)} \text{Exp} \ q(e^{t/q} - 1)\)

• Minimized when \(t = q \log(1+\delta)\), yielding
  \(g(q, \delta) = \left( \frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^q\)

• Combining: \(\Pr[X' \geq (1+\delta)X] \leq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^q X\)

• Bound depends only on
  – \(\delta\): the size of the large deviation
  – \(qX = X/N\): exponentially small in the average number of samples

• Simpler bound using \(\frac{e^\delta}{(1+\delta)^{1+\delta}} < \text{Exp}[{-\delta^2/3}]\) for \(0<\delta<1\)
Illustration: 2 Bounds

\[ \text{Exp}[-\delta^2 / 3] \]

\[ e^\delta / (1+\delta)^{1+\delta} \]
Illustration: Exponentially small in $E[\#\text{samples}]=n$

$$(e^{\delta}/(1+\delta)^{1+\delta})^n$$
Corresponding Lower Bounds

- \( \Pr[X' \leq (1-\delta) X] \leq (e^{-\delta} / (1-\delta)^{1-\delta})^{qX} \)

- Simplified bound using \( e^{-\delta} / (1-\delta)^{1-\delta} < \text{Exp}[-\delta^2 / 2] \) for \( 0 < \delta < 1 \)

- Exercise: verify
Confidence Bounds for $X$

- So far: given $X$, we can compute bounds on distribution of $X'$
  - e.g. $\Pr[X' \geq (1+\delta)X] \leq \epsilon$ for some small $\epsilon = \exp(-\delta^2 X/(3N))$
- In practice, we are given estimate $X'$, want to bound true $X$
- Idea for lower bound:
  - Given some small target $\epsilon$, and our estimate $X'$
  - Find largest possible value $X_{\min}$ for $X < X'$ such that:

$$\Pr[\text{estimate is } X' \text{ or larger}] \leq \epsilon$$

$$\exp(-\delta^2 X/(3N))$$
Confidence Bounds for $X$

• Find largest possible value $X_{\min}$ for $X < X'$ such that

$$Pr[\text{estimate is } X' \text{ or larger}] \leq \varepsilon \quad \text{i.e.} \quad Pr[X'' \geq X' = (1 + \delta) X] \leq \varepsilon \quad (*)$$

where $X''$ has same distribution as $X'$ but we now consider $X'$ fixed

• Use (any) upper bound e.g. $Pr[X'' \geq (1 + \delta)X] \leq \exp(-\delta^2 X/(3N))$

• For (*) to hold it suffices that:

$$\exp(-\delta^2 X/(3N)) \leq \varepsilon \text{ with } \delta = X'X - 1 \quad \text{i.e.} \quad - (X'/X - 1)^2 X/(3N) \leq \log(\varepsilon) \quad (**)$$

• $X_{\min}$ is largest such $X < X'$ for which (**) is true

• Exercise: $X_{\min} = X' + (\beta - (\beta(\beta + 4X'))^{1/2})/2$ where $\beta = -3N \log(\varepsilon)$

• Similar upper bound $X_{\max}$ for $X$, using Chernoff Lower bound
Summary of HT and Bounds

• HT tells us how to make unbiased estimates from samples
  – Size of bytes in flows from packet sampled netflow
• How accurate is estimate
  – Chebychev Inequality and Chernoff Bounds
  – General results, but can be applied to HT estimates
• Confidence intervals for true quantities from estimates