Polynomial Estimators for Data Stream Computations

Sumit Ganguly

Indian Institute of Technology, Kanpur

PROBLEM 1: pTH POWER SPARSE RECOVERY

- Assume integral $p \geq 2$. Let $x^p = (x_1^p, \ldots, x_n^p)$.
- Recover z from Ax such that

$$\|z - x^p\|_2^2 \leq D \min_{z' \text{ k } -sparse} \|z' - x^p\|_2^2$$

  where, D is constant or $1 + \epsilon$.
- For $p = 1$, same as the ℓ_2/ℓ_2 problem.

Problem 2: High Frequency Moment Estimation

- $$F_p = \|x\|_p^p = \sum_{i \in [n]} |x_i|^p, p > 2$$

  Given $\epsilon \in (0,1)$, return $\hat{F}_p$ satisfying

  $$\left|\hat{F}_p - F_p\right| \leq \epsilon F_p$$

  with probability at least $3/4$.

Polynomial Estimator: Motivation

- Applications require to estimate $\psi(\mathbb{E}[X])$, X is a random variable. $\mu = \mathbb{E}[X]$.
- E.g., for pth moments problem and pth power sparse recovery, need to estimate $|x_i|^p$.
- Usually given an estimate λ of μ

$$|\lambda - \mu| \leq \sigma \quad .$$

Motivation: Polynomial Estimators

- Simplest estimate for $\psi(\mathbb{E}[X])$ is $\psi(\lambda)$.
- Problem: May not be accurate enough.
- E.g., $|\mu^p - \lambda^p| = O(p\mu^{p-1}\sigma)$.
- There exist estimators ϑ with $|\mu^p - \vartheta| = O(p^2\mu^{p-2}\sigma^2)$. More accurate for $\mu \gg p\sigma$.

Taylor Polynomial Estimator

- Let $\{X_j\}_{j=1}^k$ be k independent copies of X with expectation μ and variance σ^2.
- Consider k-term Taylor's series expansion for $\psi(\mu)$ around λ.

$$\psi(\mu) = \psi(\lambda) + \gamma_1(\lambda)(\mu - \lambda) + \ldots + \gamma_k(\lambda)(\mu - \lambda)^k$$

- Replace $(\mu - \lambda)^v$ by $(X_1 - \lambda)(X_2 - \lambda) \ldots (X_v - \lambda)$.
- This gives the k-term Taylor polynomial estimator

$$\vartheta(\psi, \lambda, k) = \sum_{v=0}^{k} \gamma_v(\lambda)(X_1 - \lambda) \cdots (X_v - \lambda)$$

EXPECTATION

By independence of the X_j's,

$$\begin{aligned}\mathbb{E}[\vartheta] &= \sum_{v=0}^{k} \gamma_v(\lambda)\mathbb{E}[(X_1-\lambda)\cdots(X_v-\lambda)] \\ &= \sum_{v=0}^{k} \gamma_v(\lambda)(\mu-\lambda)^j \\ &= \psi(\mu) - \gamma_{k+1}(\zeta)(\mu-\lambda)^{k+1},\end{aligned}$$

for some $\zeta \in (\mu, \lambda)$.

Expectation: Example

For the function $\psi(\mu) = \mu^p$,

- For p integral and $k \geq p$, $\mathbb{E}[\vartheta] = \mu^p$.
- For $k \geq p$,

$$\left|\mathbb{E}[\vartheta] - \mu^p\right| < (9p)^{-(k+1)}\mu^p$$

VARIANCE

- Let $\eta^2 = 2\sigma^2$, η positive.

$$\text{Var}\left[\vartheta\right] = \left(\sum_{v=1}^{k} |\gamma_v(\lambda)|\eta^v\right)^2$$

- For $\psi(x) = x^p$

$$\text{Var}\left[\vartheta\right] \leq p^2\eta^2(|\lambda| + \eta)^{2p-2}$$

For $\mu > 2p\sigma$,

$$\text{Var}\left[\vartheta\right] \leq 2p^2\sigma^2\mu^{2p-2}$$

Motivation

- TPEST keeps k independent copies $X_1, \ldots, X_k$. Instead suppose we keep $16k$ copies.
- Choose "fair" number of k-subsets from these. Create TPEST from these subsets and take their average.
- Expectation is correct.
- Could variance reduce? By how much?
- By factor of $O(k)$ if k-subsets are "sufficiently distant".

MOTIVATION

- TPEST keeps k independent copies $X_1, \ldots, X_k$. Instead suppose we keep $16k$ copies.
- Choose “fair” number of k-subsets from these. Create TPEST from these subsets and take their average.
- Expectation is correct.
- Could variance reduce? By how much?
- By factor of $O(k)$ if k-subsets are “sufficiently distant”.

Construction

- Keep $s = O(\log n)$ independent copies X. Let $s = 8k$
- Choose a family Y of s-dimensional binary vectors, each consisting of exactly k ones, such that $|Y| \geq 16k$ and Hamming distance between two vectors in Y is at least $3k/2$.
- By Gilbert-Varshamov theorem, such a Y exists.
- $Y = \{y_1, \ldots, y_r\}$, $r = 16k$. Each $y_i \subset [s]$ of size k.

Averaged Polynomial Estimator

- For $y \in Y$, choose a random $\pi : [k] \to [k]$.
- Let $y = (y(1), \ldots, y(k))$ in ascending order of indices
- Reorder elements of y by π, that is $\pi(y) \equiv (y(\pi(1)), y(\pi(2)), \ldots, y(\pi(k)))$.
- Use this ordering for TPEST using y.

$$\vartheta(y) = \sum_{v=0}^{k} \gamma_v(\lambda) \prod_{l=1}^{v} (X_{y(\pi(l))} - \lambda)$$

Averaged Polynomial Estimator

- $\bar{\vartheta}$ = average of the $\vartheta(y)$'s for $y \in Y$.

$$\mathrm{Var}\left[\bar{\vartheta}\right] = \frac{1}{k}\left(\sum_{v=1}^{k}|\gamma_v(\lambda)|(2\sigma)^v\right)^2$$

- For $\psi(\mu) = \mu^p$,

$$\mathrm{Var}\left[\bar{\vartheta}\right] \leq (1/k)\mathrm{Var}\left[\vartheta\right]$$

pTH POWER SPARSE RECOVERY: OUR RESULTS

Recover z from Ax such that

$$\|z - x^p\|_2^2 \leq (1+\epsilon) \min_{x' \text{ k } -sparse} \left\|x'^p - x^p\right\|_2^2$$

- Requires $\Omega(\epsilon^{-1} n^{1-1/p} k^{1/p} \log(n/k))$ rows in A
- We cannot match this bound.
- Notation: $\left\|x^{\text{res}(k)}\right\|_2^2$ —second moment of all but top-k items.
  $\left\|x^{\text{res}(H)}\right\|_2^2$ is the second moment of x except for the contribution of items in H.

Upper Bound

- For $C = O(p^2 2^{2p} \epsilon^{-1} n^{1-1/p} k^{1/p})$, return a $2k$-sparse vector z with support H satisfying

$$\begin{aligned} \|z_H - x^p\|_2^2 \le & \left\|x^{\mathrm{res}(k)}\right\|_{2p}^{2p} + \epsilon^p A_1^{p-1} \left\|x^{\mathrm{res}(C)}\right\|_{2p}^{2p} \\ & + \epsilon A_1 k^{-1/p} \left\|x^{\mathrm{res}(C)}\right\|_{2p}^{2} \|x_H\|_{2p-2}^{2p-2} \\ & + \epsilon A_2 k^{-1/p} \left\|x^{\mathrm{res}(C)}\right\|_{2p}^{2} \cdot \left\|x_{H \setminus \mathrm{TOP}(k)}\right\|_{2p-2}^{2p-2} \end{aligned}$$

  where, $A_1 = O((p^2 2^{2p})^{-1})$ and $A_2 = O(1)$.
- Lower Bound:

$$\Omega\left((\epsilon(A_1 + A_2)^{-1} n^{1-1/p} k^{1/p}\right)$$

Sparse Recovery: Special Case

- Let $C = O(p^2 2^{2p+2} \epsilon^{-1} n^{1-2/p} k^{1/p})$. Then, algorithm returns $2k$-sparse z with support H s.t.

$$\begin{aligned} \|z_H - x^p\|_2^2 \leq & \|x^{\text{res}(k)}\|_{2p}^{2p} + \\ & \epsilon B_1 k^{-1/p} \|x^{\text{res}(C)}\|_p^2 \, \|x_H\|_{2p-2}^{2p-2} + \\ & \epsilon B_2 k^{-1/p} \|x^{\text{res}(C)}\|_p^2 \cdot \|x_{H \setminus T}\|_{2p-2}^{2p-2} + \\ & \epsilon^p B_1^{p-1} \|x^{\text{res}(C)}\|_p^{2p} \end{aligned}$$

  where, $B_1 = O((p^2 2^{2p+2})^{-1})$, $B_2 = O(1)$.
- Lower Bound of $\Omega((\epsilon B_1 (p \log p))^{-1} n^{1-2/p} k^{1/p})$.

pTH POWER SPARSE RECOVERY: STRUCTURES

1. HH: COUNTSKETCH structure for heavy-hitters. Table heights $O(p^2C)$, number of tables $O(p + \log n)$.
2. TPEST: COUNTSKETCH-like structure for averaged TP estimator (same height, width). Only hash functions are $O(1)$-wise independent (instead of pair-wise independent).
3. Will keep $O(\log(1/\delta))$ copies of TPEST for taking median to boost confidence of the $\bar{\vartheta}$. (Probably excessive!)

Recovery Algorithm: basics

1. $H =$ set of top-$2k$ items by $|\hat{x}_i|$ from HH.
2. For $i \in H$, choose estimates x'_{li} for x_i from at least $s/2$ tables in TPEST where NOCOLL(i) holds, that is, i does not collide with any other item of H.
3. If less than $s/2$ tables satisfy NOCOLL(i) then FAIL.
4. NOCOLL(i) fails with probability $e^{-\Omega(s)}$.

WHY USE NOCOLL?

Reason 1: Reduces Variance.

1. Fact: $\left\|x^{\text{res}(H)}\right\|_2^2 \le 9\left\|x^{\text{res}(2k)}\right\|_2^2$ [CorMuthu06, others].
2. Gives $\sigma^2 = \text{Var}\left[x'_{il} \mid \text{NOCOLL}\right] \le O(\left\|x^{\text{res}(2k)}\right\|_2^2/C)$.

Reason 2: Conditional on $\forall i \in H, \text{NOCOLL}(i)$,

1. For $i, j \in H$, $\text{Var}\left[\bar{\vartheta}_i + \bar{\vartheta}_j\right] = \text{Var}\left[\bar{\vartheta}_i\right] + \text{Var}\left[\bar{\vartheta}_j\right]$

Recovery Algorithm: First Attempt

1. H be the set of top-$2k$ items by $|\hat{x}_i|$.
2. For $i \in H$, choose estimates x'_{li} for x_i from at least $s/2$ tables in TPEST where NOCOLL(i) holds, that is, i does not collide with any other item of H.
3. For $i \in H$, set $z_i = \bar{\vartheta}_i$ corresponding to the function $\psi(t) = t^p$ and using $\{x'_{li} : l \in \text{NOCOLL}(i)\}$). Other coordinates are set to 0. Return z_H.

VARIATION

- Estimator gives good accuracy when $|x_i| > 2p\sigma$.

$$\mathrm{Var}\left[\vartheta\right] = p^2\sigma^2(|x_i| + 2\sigma)^{2p-2}, \ \sigma = \left\|x^{\mathrm{res}(C)}\right\|_2/\sqrt{C}$$

- For items in H with $|x_i| < p\sigma$, errors increase.
- Poly. estimator θ_{p-2j} for $\psi(\mu) = \mu^{p-2j}$ gives good accuracy for $|x_i| > 2(p-2j)\sigma$, $j = 1, 2, \ldots, p/2 - 1$. Take median of $O(\log n)$ such estimators and evaluate $z_i = \theta_{p-2j}^{p/(p-2j)}$. Then,

$$|z_i - \mu^p|^2 \leq O(\mu^{2p-2}\sigma^2) \ .$$

Recovery Algorithm

1. $H =$ set of top-$2k$ items by $|\hat{x}_i|$ from HH.
2. For $i \in H$, choose estimates x'_{li} for x_i from at least $s/2$ tables in TPEST where NOCOLL(i) holds, that is, i does not collide with any other item of H.
3. If less than $s/2$ tables satisfy NOCOLL(i) then FAIL.
4. Estimate $\left\|x^{\text{res}(2C)}\right\|_2$ using standard algorithms to within $1 + 0.1$ factors. Estimate σ.
5. Order items in H by absolute values of estimates $|\hat{x}_i|$.
6. Use polynomial estimator for μ^{p-2j} if $|\hat{x}_i| \in \hat{\sigma}(p-2j, p-2j+2)$. Take medians to boost confidence. Set $z_i = \theta^{p/(p-2j)}$.
7. Recovered vector is z with support H.

Main Property

- Returns $2k$-sparse vector z with support H, $C > 4k$, satisfying

$$\begin{aligned}\|z_H - x^p\|_2^2 \le & \|x^{\mathrm{res}(k)}\|_{2p}^{2p} + \\ & A_1\left(\frac{\|x^{\mathrm{res}(C)}\|_2^2}{C}\right)\|x_H\|_{2p-2}^{2p-2} + \\ & A_2\left(\frac{\|x^{\mathrm{res}(C)}\|_2^2}{C}\right)\|x_{H\setminus T}\|_{2p-2}^{2p-2} + \\ & A_2 k\left(\frac{\|x^{\mathrm{res}(C)}\|_2^2}{C}\right)^p\end{aligned}$$

where, $A_1 = O(p^2)$, $A_2 = O(p^2 2^{2p+2})$.

pth frequency moments: $F_p = \|x\|_p^p$, $p > 2$.
Estimate F_p to within factors of $1 \pm \epsilon$.

Summary: pth Frequency Moments

Upper Bounds, space in words:

- AMS 96. $O(n^{1-1/p}\epsilon^{-2})$.
- Indyk-Woodruff 05. $O(p^2 n^{1-2/p} poly(\epsilon^{-1}, \log n))$.
- BGKS06. $O(p^2 n^{1-2/p}\epsilon^{-2-4/p} \log^2(n) \log(m))$.
- Andoni-Krauthgamer-Onak 10.

$$O(p^2 n^{1-2/p}\epsilon^{-2-4/p}(\log n) E_{p,n})$$

where, $E_{p,n} = (1-2/p)^{-1}(1-n^{-4(1-2/p)})$. This is $O(1)$ for $p = 2+\Omega(1)$ and $O(\log n)$ for $p = 2 + o(1)$.

Upper Bounds contd.

- Braverman-Ostrovsky 2010.
  $O(p^2 n^{1-2/p} \epsilon^{-2-4/p} (\log n) G_{p,n}), \quad G(p,n) \geq E(p,n)$.
- Current work. $O(p^2 n^{1-2/p} \epsilon^{-2} \log(n))$.

Lower Bounds

Space in bits.

- AMS 96: $\Omega(n^{1-5/p})$.
- Bar-Yossef Jayram Kumar Sivakumar 02: $\Omega(n^{1-2/p-\delta}\epsilon^{-2/p})$
- Chakrabarti-Khot-Sun 03: $\Omega(n^{1-2/p}\epsilon^{-2/p})$.
- Woodruff-Zhang 11. $\Omega(n^{1-2/p}\epsilon^{-4/p}/\log^{O(1)}(n))$.
- G 11. $\Omega(n^{1-2/p}\epsilon^{-2}/\log(n))$.

Algorithm: pth Frequency Moments

- Hss structure (Variation of Indyk-Woodruff technique). Keep HH and TPEST structures for levels $l = 0, 1, \ldots, L$. $L = \log(n/C)$, where C is space parameter.
- $O(1)$-wise independent hash functions $g_1, g_2, \ldots, g_L : [n] \rightarrow [0, 1]$ and

$$\Pr\left[g_j(i) = 1\right] = q$$

- All items map to level 0, i maps to level l if

$$g_1(i) = g_2(i) = \ldots = g_l(i) = 1$$

with probability q^l. (q will be a constant).

Departure from HSS

- Sampling is done per level with probability q.
- Use TPEST structure for μ^p at each level along with COUNTSKETCH.
- Level l structure has tables with height $C_l = \alpha^l C$, sizes are geometrically decreasing. [GLPS10] idea.
- Rest (grouping, thresholds, etc..) are analogous.

PROPERTIES

Variations over Hss properties and functions

- $x_{(l)}$: frequency vector of sub-stream at level l.
- $\left\| x_{(l)}^{\text{res}(C_l)} \right\|_2^2 \leq q^{-l} \left\| x^{\text{res}(q^{-l}C_l/2)} \right\|_2^2$, with probability $1 - e^{-C_l/6}$.
- Threshold of group G_l is

$$T_l^2 = \frac{1}{C_l} \max(q^{-l} \left\| x^{\text{res}(C)} \right\|_2^2, \log(n))$$

- $G_l = \{i : T_l \leq |x_i| < T_{l-1}\}, l = 0, 1, \ldots, L$.
- Estimate for $\hat{x}_i$ is the estimate from the lowest level where it maps and also crosses the threshold.
- $\bar{G}_l : \{i \text{ maps to level } l \text{ and } T_l \leq \hat{x}_i < T_{l-1}\}$.

Properties contd.

- Estimate:

$$\hat{F}_p = \sum_{l=0}^{L} \sum_{i \in \bar{G}_l} q^{-l} \bar{\vartheta}_i \ .$$

-

$$\left| \mathbb{E}\left[\hat{F}_p\right] - F_p \right| \leq n^{-\Omega(1)} F_p \ .$$

$$\mathrm{Var}\left[\hat{F}_p\right] = \sum_{l=0}^{L} q^{-l} \cdot \frac{p^2 \left\| x^{\mathrm{res}(C)} \right\|_2^2}{C_l} \cdot \left\| x_{G_l} \right\|_{2p-2}^{2p-2}$$

- Separate sum for $l = 0$ and $l \geq 1$.

VARIANCE CALCULATION

- Let $C = O(p^2\epsilon^{-2}n^{1-2/p}/q)$. So

$$\left\|x^{\text{res}(C)}\right\|_2^2/C \le \epsilon^2 q \left\|x^{\text{res}(C)}\right\|_p^2/(4p^2) \ .$$

Sum of terms from $l = 1, 2, \ldots, L$

$$\le \frac{\epsilon^2 \left\|x^{\text{res}(C)}\right\|_p^2}{4q} \times \sum_{l=1}^{L} q^{-l} \left\|x_{G_l}\right\|_p^p \left(\frac{q^{l-1}\left\|x^{\text{res}(C)}\right\|_p^2}{\alpha^{l-1}p^2}\right)^{(p-2)/2}$$

$$\le \frac{\epsilon^2 \left\|x^{\text{res}(C)}\right\|_p^p}{4} \sum_{l\ge 1} \left\|x_{G_l}\right\|_p^p \left(\frac{\epsilon^2 q}{\alpha p^2}\right)^{(p-2)(l-1)/2}$$

$$\le \frac{\epsilon^2 \left\|x\right\|_p^{2p}}{4}$$

CONCLUSIONS

- An overview of polynomial estimators and its applications in sparse recovery and estimating F_p for $p > 2$.
- Future work: Estimators based on Fourier or other series.

THANK YOU!

CONCLUSIONS

- An overview of polynomial estimators and its applications in sparse recovery and estimating F_p for $p > 2$.
- Future work: Estimators based on Fourier or other series.

THANK YOU!