Median Calculation

Yuval Shavitt
School of Electrical Engineering
TEL AVIU UNIVERSITY אוניברסיטת תל-אביב

Median Estimation

Algorithm	Calculation Complexity	Storage Complexity
Classical Blum, Floyd, Pratt, Rivest, and Tarjan [1974]	O(N)	O(N)
Improvements [1975-2001]	Improves by a factor	O(N)
Munro and Patterson [1980]	p passes	$O(N^{-1/p})$
Battiato et al. [2000]	O(N)	in-place O(N)
Rousseeuw et al. [1990] *Remedian*[1]	O(N *log*(N))	O(*log*(N))
Additional works in article [1997-2002][1]	O(N)	$O(N^{1/2})$;$O(log^2(N))$
Greenwald & Khanna [2001]	O(N)	O(*log*(εN)/ε)
FAME	O(N)	2

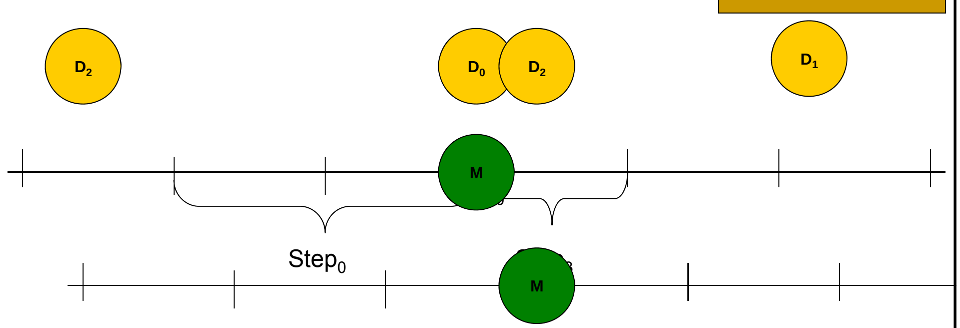
[1] Requires a-priori knowledge of number of samples.

FAME:
Fast Algorithm for Median Estimation

- Linear execution time
- Constant Memory (2 doubles)
- Works on dataset of any size
- Convergence rate adopts to data variance
- Can be easily integrated in hardware, SQL, Java
- Variations
  - Windowed
  - No overshoots

FAME Flow

M_3=M_2-$Step_2$
$Step_3$=$Step_2$/2



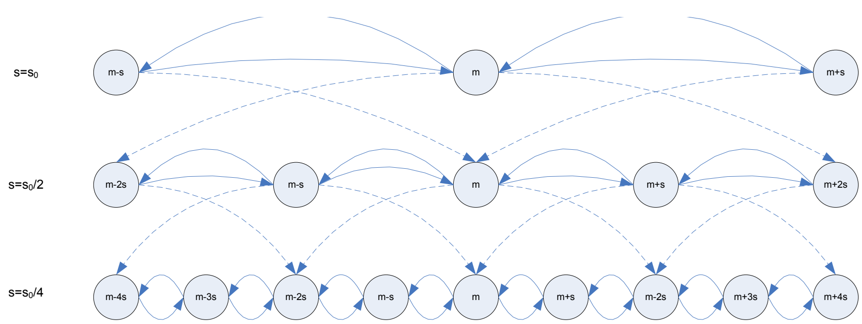
Fame Formal Description

Algorithm 1 : Fast Algorithm for Median Estimation

Initialization:
$M = data(1)$
$Step = max(|data(1)/2|, b)$ // b is a minimal initial step

For each new item i:
if $M > data(i)$ **then**
    $M = M - step$
else if $M < data(i)$ **then**
    $M = M + step$
end if
if $|data(i) - M| < step$ **then**
    $step = step/2$
end if

Proof: 1-D Markov to FAME

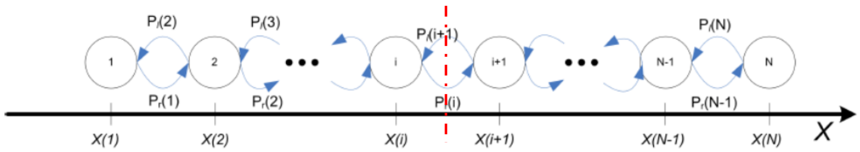


Proof of correctness

- We define a random process

$$X_{n+1} = X_n + \text{step} \cdot \text{sign}(x - X_n)$$

- x~P(x<X)



$$P_l(i) = P(x < X(i))$$

$$P_r(i) = 1 - P(x < X(i))$$

$$\pi(i) \cdot P_r(i) = \pi(i+1) \cdot P_l(i+1)$$

Note: We assume i.i.d. input and differentiable P(x<X)

Proof of correctness

- Lemma I
  Let C be a Markov chain as defined above, then for $\Delta \rightarrow 0$, the steady state probability distribution in the median surrounding behaves as:

$$\pi_x(x) \sim \exp(-2P'(x_m)/\Delta \cdot (x - x_m)^2)$$

Assumption: i.i.d. input, P(x<X) is differentiable
Other cases can be treated as well

Proof: Lemma I

We recall that $\pi_x(x)$ is defined as

$$\pi_x(x)(1 - P(x)) = \pi_x(x + \Delta)P(x + \Delta)$$

For $\Delta \to 0$ it can be approximated as:

$$\pi_x(x)(1 - P(x)) = [\pi_x(x) + \Delta\pi'(x)][P(x) + \Delta P'(x)]$$

And then rearranged:

$$\pi_x'(x) \underbrace{- \left[\frac{1 - 2P(x) - \Delta P'(x)}{\Delta P(x) + \Delta^2 P'(x)}\right]}_{g(x)} \pi_x(x) = 0$$

Proof: Lemma I

We define:

$$g(x) \equiv - \left[\frac{1 - 2P(x) - \Delta P'(x)}{\Delta P(x) + \Delta^2 P'(x)}\right]$$

In surrounding of x_m, i.e. $P(x_m)=0.5$, for $\Delta \to 0$

$$g(x \quad \hat{g}(x) \approx \left[\frac{2P'(x_m)}{\Delta}(-2x + 2x_m)\right] \quad P'(x_m)$$

$$g'(x_m) \quad \approx \quad \frac{}{\Delta}$$

Proof: Lemma I

The solution of the differential equation

$$\pi'_x(x) - \left[\frac{1 - 2P(x) - \Delta P'(x)}{\Delta P(x) + \Delta^2 P'(x)}\right] \pi_x(x) = 0$$

In the surrounding of x_m and for $\Delta \rightarrow 0$ is:

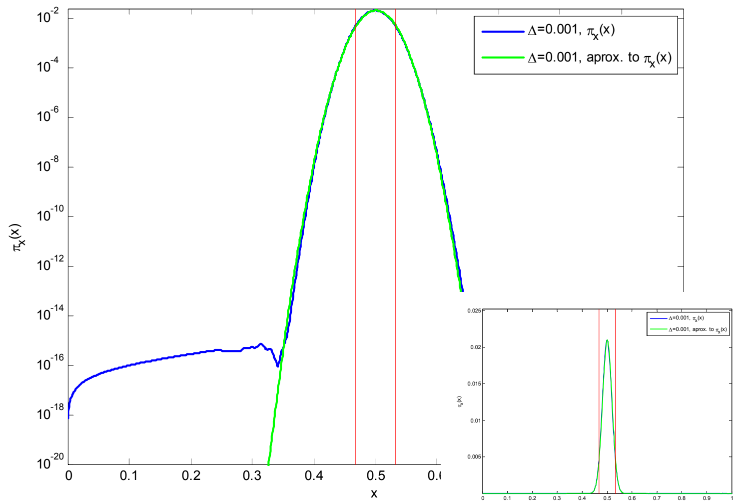
$$\pi_x(x) \sim \exp\left(-\frac{2P'(x_m)}{\Delta}(x - x_m)^2\right)$$

Proof: Corollary

According to Lemma I, all the probability mass of $\pi_x(x)$ is concentrated in the peak that behaves as $\sqrt{\frac{\Delta}{P'(x)}}$

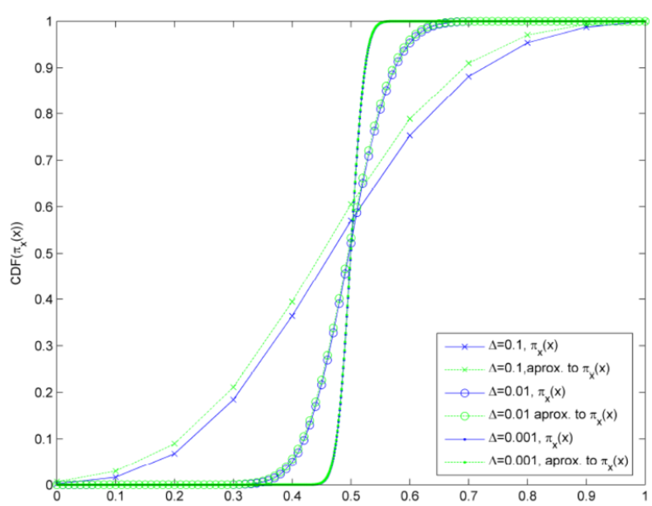
i.e., approaches zero as $\Delta^{1/2}$

Proof: Approximation quality

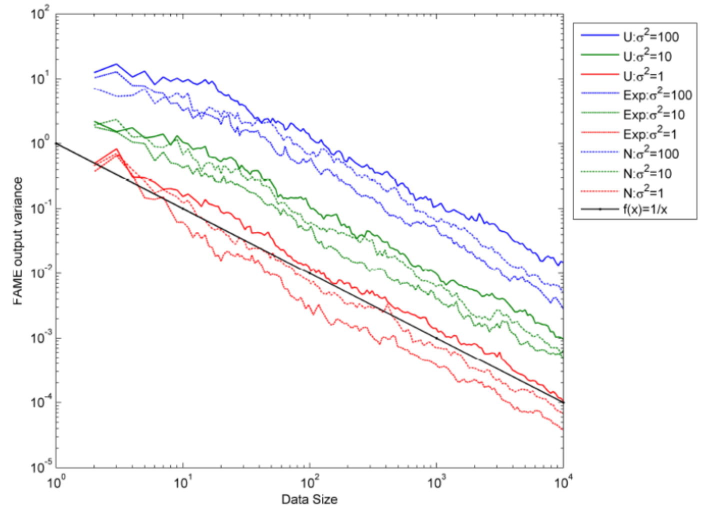


~exp(ln(2)/2)

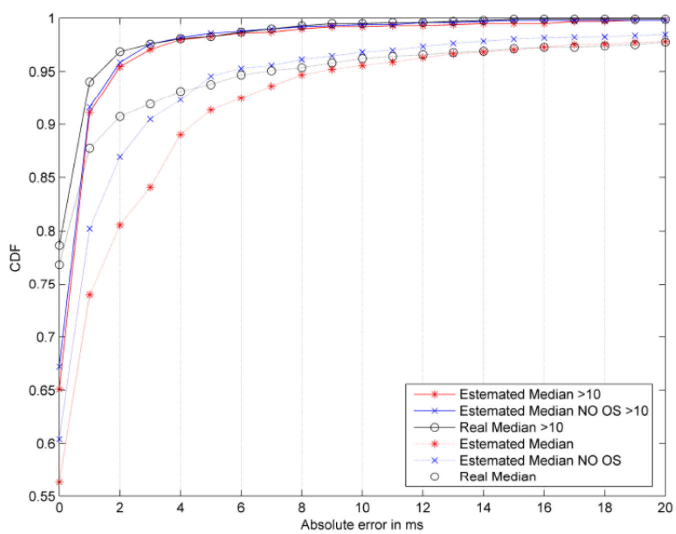
Proof: Approximation quality



Fame Convergence Rate



FAME: Test On Real Data



Summary

- Works well for a stream of 10,000,000s samples of ~1,600,000 r.v.
  - Some r.v. have a few samples
  - Some r.v. have 1000s of samples
  - All samples are mixed