Median Calculation

Yuval Shavitt
School of Electrical Engineering

Median Estimation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Calculation Complexity</th>
<th>Storage Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Blum, Floyd, Pratt, Rivest, and Tarjan [1974]</td>
<td>O(N)</td>
<td>O(N)</td>
</tr>
<tr>
<td>Improvements [1975-2001]</td>
<td>Improves by a factor</td>
<td>O(N)</td>
</tr>
<tr>
<td>Munro and Patterson [1980]</td>
<td>p passes</td>
<td>O(N^{1/p})</td>
</tr>
<tr>
<td>Battiato et al. [2000]</td>
<td>O(N)</td>
<td>in-place O(N)</td>
</tr>
<tr>
<td>Rousseeuw et al. [1990]</td>
<td>O(N log(N))</td>
<td>O(log(N))</td>
</tr>
<tr>
<td>Remedian¹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional works in article [1997-2002]¹</td>
<td>O(N)</td>
<td>O(N^{1/2}) ;O(log^2(N))</td>
</tr>
<tr>
<td>Greenwald &amp; Khanna [2001]</td>
<td>O(N)</td>
<td>O(log(eN)/ε)</td>
</tr>
<tr>
<td>FAME</td>
<td>O(N)</td>
<td>2</td>
</tr>
</tbody>
</table>

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

\[ M_0 = D_0 / 2, b \]
\[ D_1 = D_0 \]
\[ M_1 = M_0 + \text{Step}_0 \]
\[ \text{Step}_1 = \text{Step}_0 \]
\[ D_2 = D_1 - \text{Step}_1 \]
\[ M_2 = M_1 - \text{Step}_2 \]
\[ \text{Step}_3 = \text{Step}_2 / 2 \]

\[ M = M_3 \]
Fame Formal Description

**Algorithm 1**: Fast Algorithm for Median Estimation

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

4. **For each new item** \(i\):  
5. if $M > data(i)$ then  
6. $M = M - step$  
7. else if $M < data(i)$ then  
8. $M = M + step$  
9. end if

10. if $|data(i) - M| < step$ then  
11. $step = step/2$  
12. end if

Proof: 1-D Markov to FAME
Proof of correctness

- We define a random process
  \[ X_{n+1} = X_n + \text{step} \cdot \text{sign}(x - X_n) \]
- \( x \sim P(x < X) \)

Proof of correctness

Lemma I
Let \( C \) be a Markov chain as defined above, then for \( \Delta \to 0 \), the steady state probability distribution in the median surrounding behaves as:
\[
\pi_x(x) \sim \exp(-2P'(x_m)/\Delta \cdot (x-x_m)^2)
\]

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

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 $\Delta \rightarrow 0$ 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:

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

Proof: Lemma I

We define:

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

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

$$g(x) \approx \frac{2P'(x_m)}{\Delta}(-2x + 2x_m) P'(x_m)$$

$$g'(x_m) \approx \frac{-2}{\Delta}$$
Proof: Lemma I

The solution of the differential equation

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

In the surrounding of \(x_m\) and for \(\Delta \to 0\) 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

\[ \Delta = 0.001, \pi(x) \approx \exp \left( \frac{\ln(2)}{2} \right) \]
Fame Convergence Rate

![Graph showing convergence rate]

FAME: Test On Real Data

![Graph showing test results on real data]
Summary

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