Streaming Algorithms for Support-Aware Histograms

Justin Chen, Piotr Indyk, Tal Wagner

Histograms in Data Streams

- Stream (insertion/deletion) of data points x_1, x_2, …, x_m in [n], defining an empirical distribution P over [n]

Histograms in Data Streams

- Stream (insertion/deletion) of data points x_1, x_2, …, x_m in [n], defining an empirical distribution P over [n]
- Histogram (piecewise constant) approximations of P are useful summaries of the distribution
- Succinctly capture locality in the distribution and are easily interpretable

Histograms in Data Streams

- Stream (insertion/deletion) of data points $x_1, x_2, \ldots, x_m$ in [n], defining an empirical distribution P over [n]
- Histogram (piecewise constant) approximations of P are useful summaries of the distribution
- Succinctly capture locality in the distribution and are easily interpretable
- Let H(k) be the set of all k-piece histograms over [n]
- **Goal:** Using small space, find a $f \in H(k')$ s.t.

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

Choosing an error function

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

Choosing an error function

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

○ L1 error over the *domain*: $\mathrm{err}(f, P) = \sum_{i \in [n]} |f(i) - P(i)|$

Choosing an error function

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- L1 error over the *domain*: $\mathrm{err}(f, P) = \sum_{i \in [n]} |f(i) - P(i)|$
  - Measures error across all domain elements, regardless of whether approximating those elements is important for downstream applications
  - Simple, sparse distributions cannot be approximated well by histograms under this notion of error

Choosing an error function

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

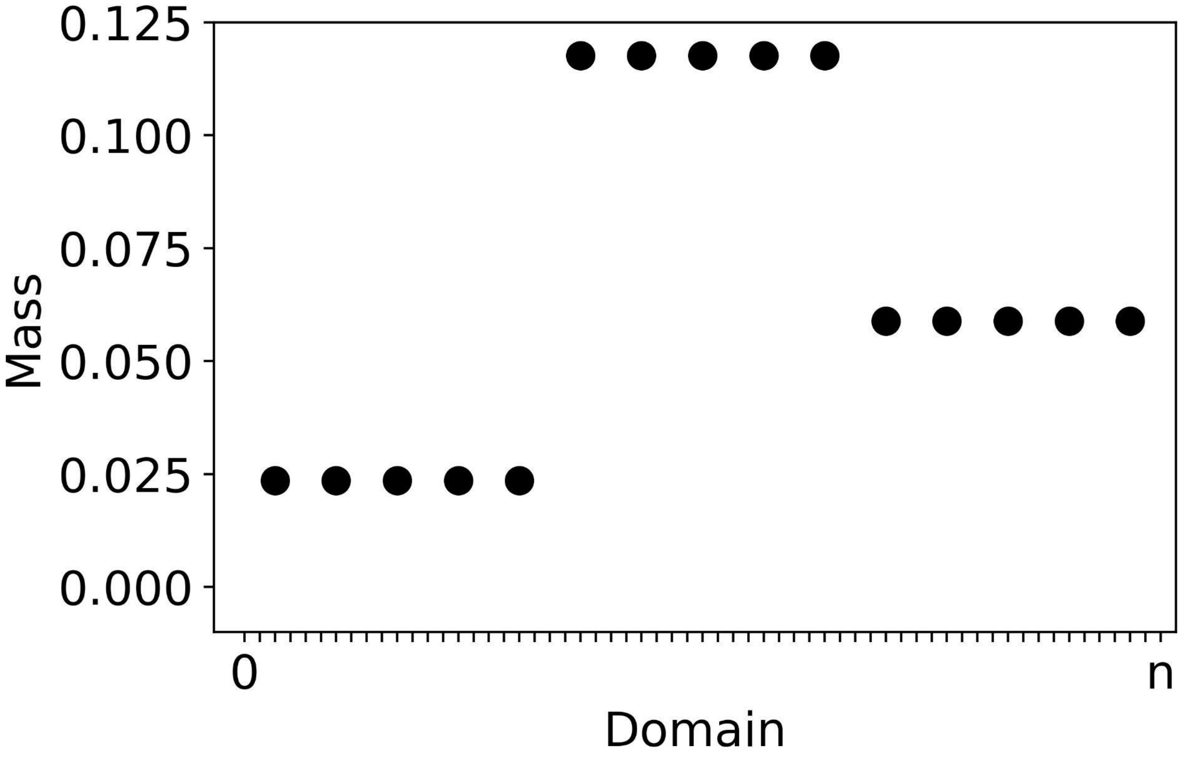
- L1 error over the *domain*: $\mathrm{err}(f, P) = \sum_{i \in [n]} |f(i) - P(i)|$
  - Measures error across all domain elements, regardless of whether approximating those elements is important for downstream applications
  - Simple, sparse distributions cannot be approximated well by histograms under this notion of error
- L1 error over the *support*: $\mathrm{err}(f, P) = \sum_{i \in \mathrm{supp}(P)} |f(i) - P(i)|$

Choosing an error function

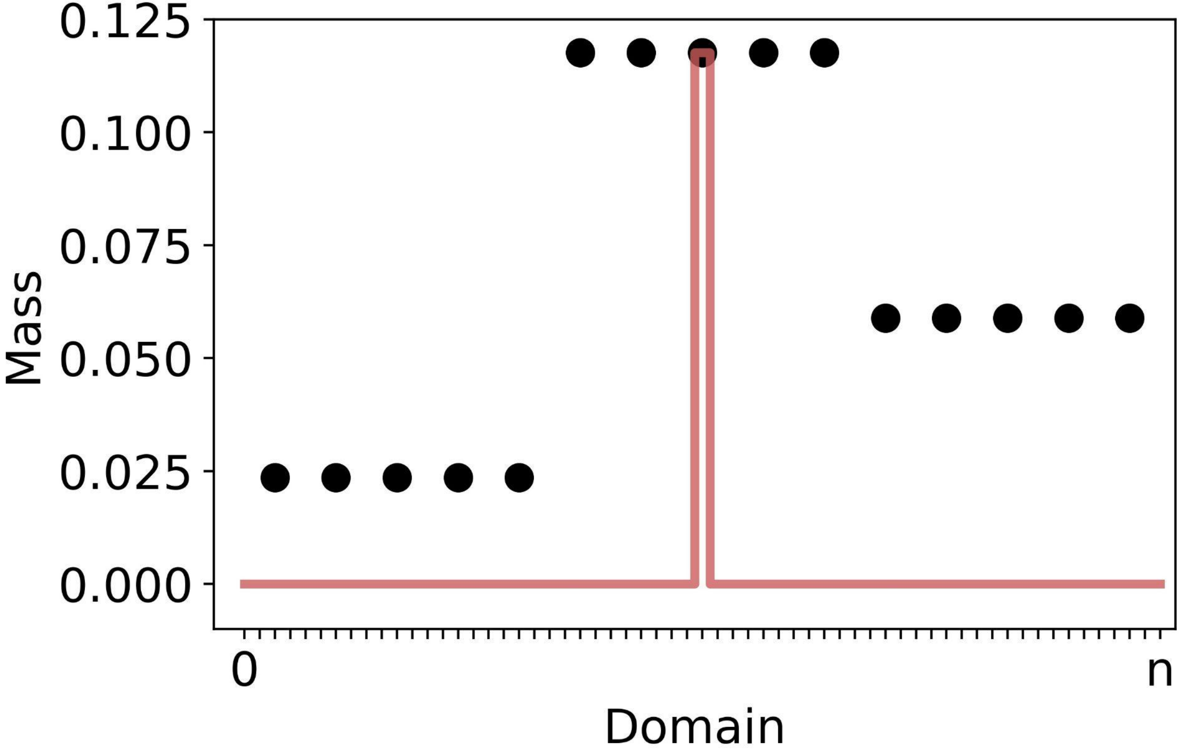
$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- L1 error over the *domain*: $\mathrm{err}(f, P) = \sum_{i \in [n]} |f(i) - P(i)|$
    - Measures error across all domain elements, regardless of whether approximating those elements is important for downstream applications
    - Simple, sparse distributions cannot be approximated well by histograms under this notion of error
- L1 error over the *support*: $\mathrm{err}(f, P) = \sum_{i \in \mathrm{supp}(P)} |f(i) - P(i)|$
    - “Support-aware” error is a natural definition that captures simple structure in sparse data

A simple example (3-piece histogram)

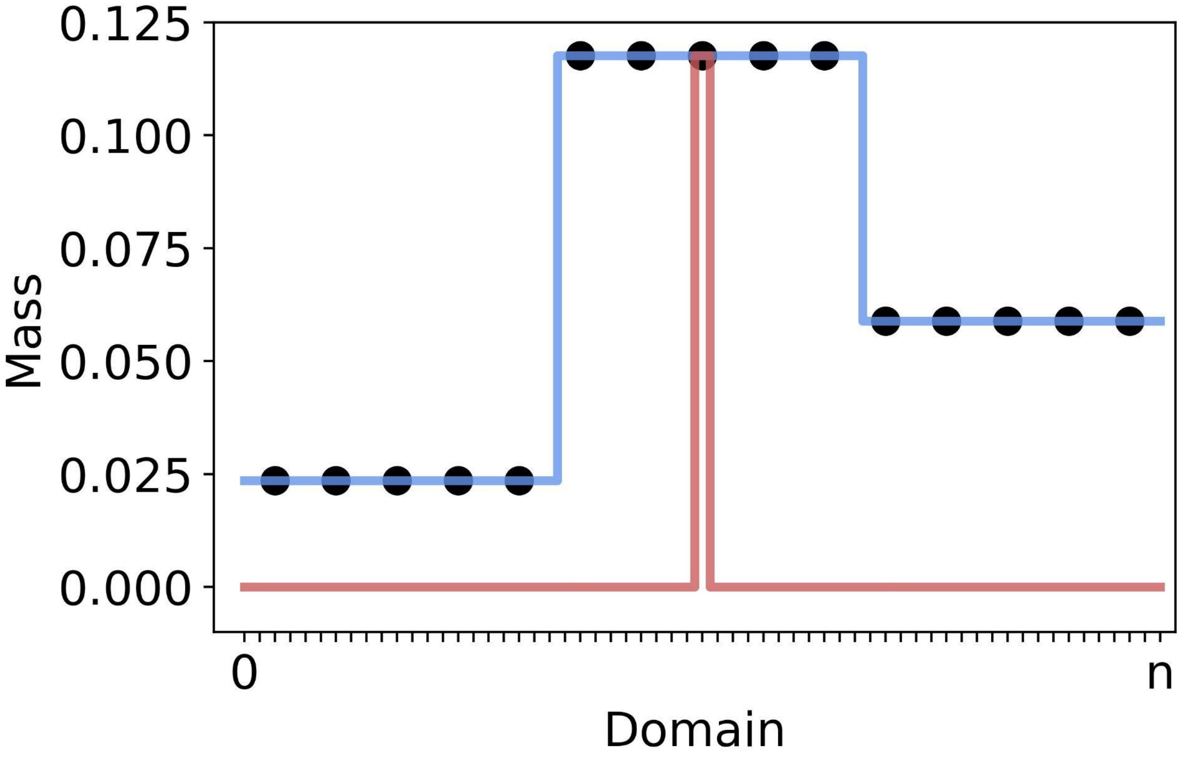


A simple example (3-piece histogram)



error over domain

A simple example (3-piece histogram)



error over domain

error over support

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)
- One pass algorithm using $O(\mathrm{sqrt}(n) \cdot k \cdot \varepsilon^{-3})$ space

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)
- One pass algorithm using $O(\mathrm{sqrt}(n) \cdot k \cdot \varepsilon^{-3})$ space
  - Complementary lower bound that $\Omega(\mathrm{sqrt}(n))$ space is required in one pass even for $k=2$

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)
- One pass algorithm using $O(\mathrm{sqrt}(n) \cdot k \cdot \varepsilon^{-3})$ space
  - Complementary lower bound that $\Omega(\mathrm{sqrt}(n))$ space is required in one pass even for k=2
- Two pass algorithm using $O(\log^2(n) \cdot k \cdot \varepsilon^{-3})$ space (exponential gap!)

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)
- One pass algorithm using $O(\mathrm{sqrt}(n) \cdot k \cdot \varepsilon^{-3})$ space
  - Complementary lower bound that $\Omega(\mathrm{sqrt}(n))$ space is required in one pass even for $k=2$
- Two pass algorithm using $O(\log^2(n) \cdot k \cdot \varepsilon^{-3})$ space (exponential gap!)
- Experiments on four datasets using our algorithms to find structure in real data

Streaming Support-Aware Histograms

$$\mathrm{err}(f, P) < \min_{f^* \in H(k)} \mathrm{err}(f^*, P) + \varepsilon$$

- Streaming algorithms for *support-aware* error cannot achieve multiplicative error guarantees (reduction from Set Disjointness)
- One pass algorithm using O(sqrt(n) · k
  - Complem pt(n)) space is required in one pass
- Two pass algorithm using $O(\log^2(n) \cdot k \cdot \varepsilon^{-3})$ space (exponential gap!)
- Experiments on four datasets using our algorithms to find structure in real data

Check out our paper and poster!