Stochastic Gradient Descent (SGD) on Quantile Estimation

Yiping Su
Yiping.Su@anu.edu.au

Supervisor: Cheng Soon Ong

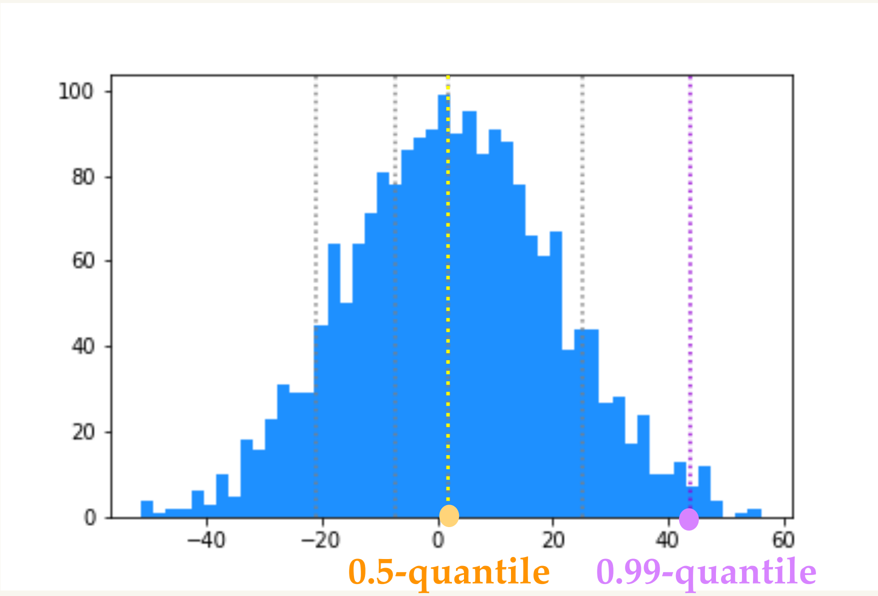
Overview

- **Motivation:**
    - Applications in quantile estimation of data stream
    - Restrictions on memory
- **SGD on quantile estimation**
    - The SGD algorithm
    - Equivalence between SGD and Frugal-1U
    - Step size adaptation
        - DH-SGD
    - Multi-quantile estimation
- **Summary and Future Work**

Motivation & Background

- **Quantiles** can help to characterize a data distribution.

- Definition: the **τ-quantile** is the cutting point that divides the distribution by τ (e.g. 0.5-quantile is the median)



- **Data stream**:

  Large amount of data is not available all at once, instead the **data points come in sequence** in a stream-like form.

- **Quantile computation?**

  -> restriction on **memory** space and **computation**

  -> sorting and computation is **not a feasible solution**

Quantile estimation on data streams

- **Applications**:
  - Network monitoring
  - Data Mining

Motivation

Memory restriction of big size data stream

- Problem: huge amount of data is not able to be stored.
- Current algorithms: space-efficient solutions

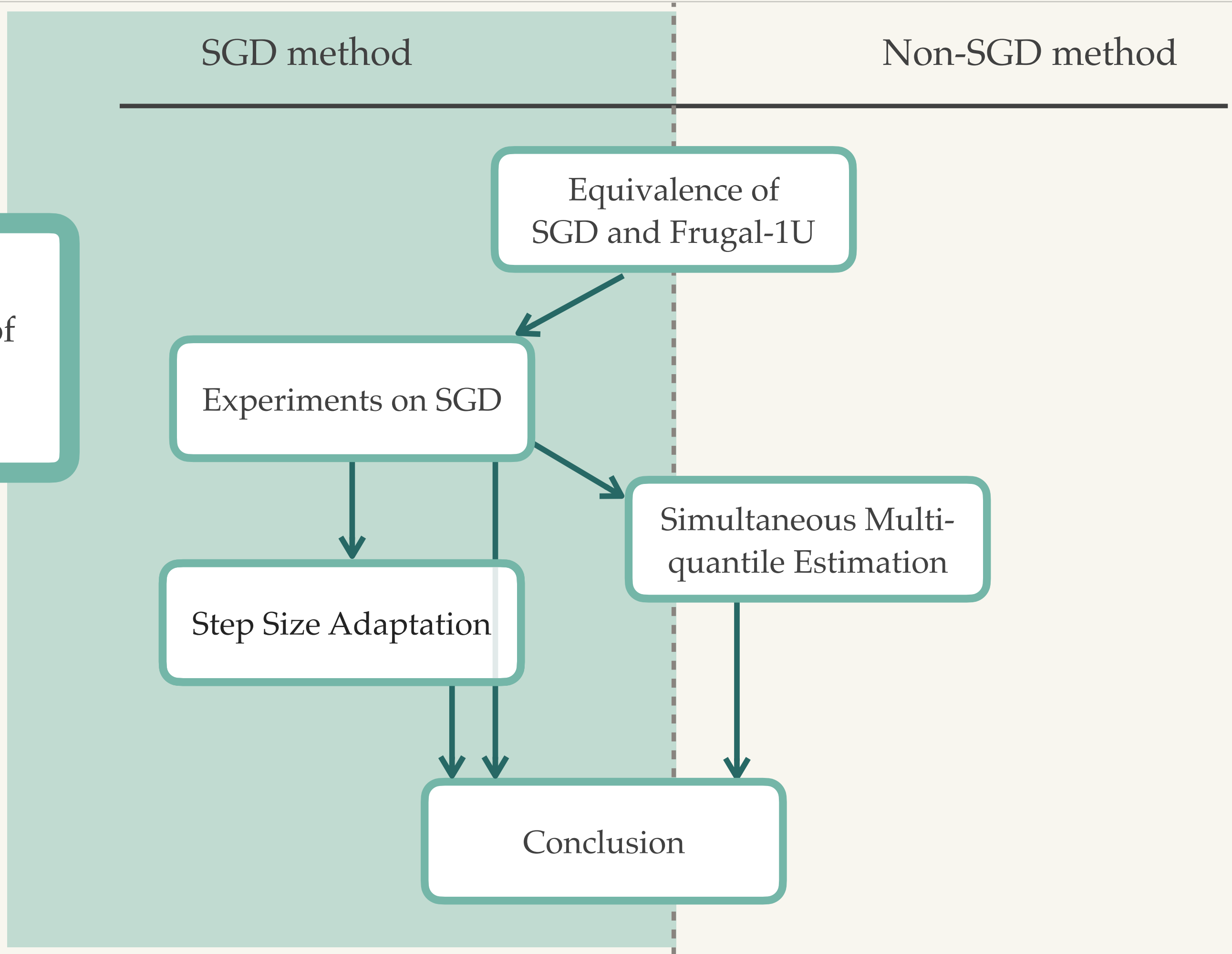
Algorithm	**Space complexity**
GK algorithm [1]	$O(\frac{1}{\epsilon}\log(\epsilon N))$
Q-Digest [2]	$O(\frac{1}{\epsilon}\log U)$
Count-Min sketch [3]	$O(\frac{1}{\epsilon}\log^2 N \log(\frac{\log N}{\phi\sigma}))$
Work of Felber and Ostrovsky [4]	$O(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$

- Design a method with **O(1) space complexity**?

  — Implement the machine learning approach of **Stochastic Gradient Descent (SGD)**

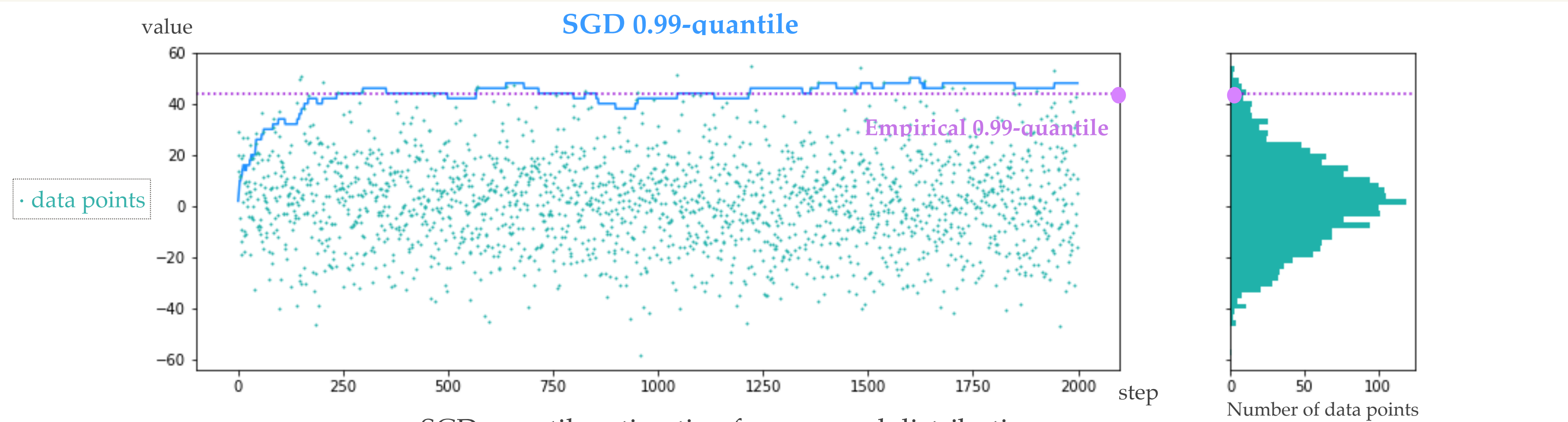
The Contribution of My Work

New part:
The introduction of SGD to quantile estimation



SGD on Quantile Estimation

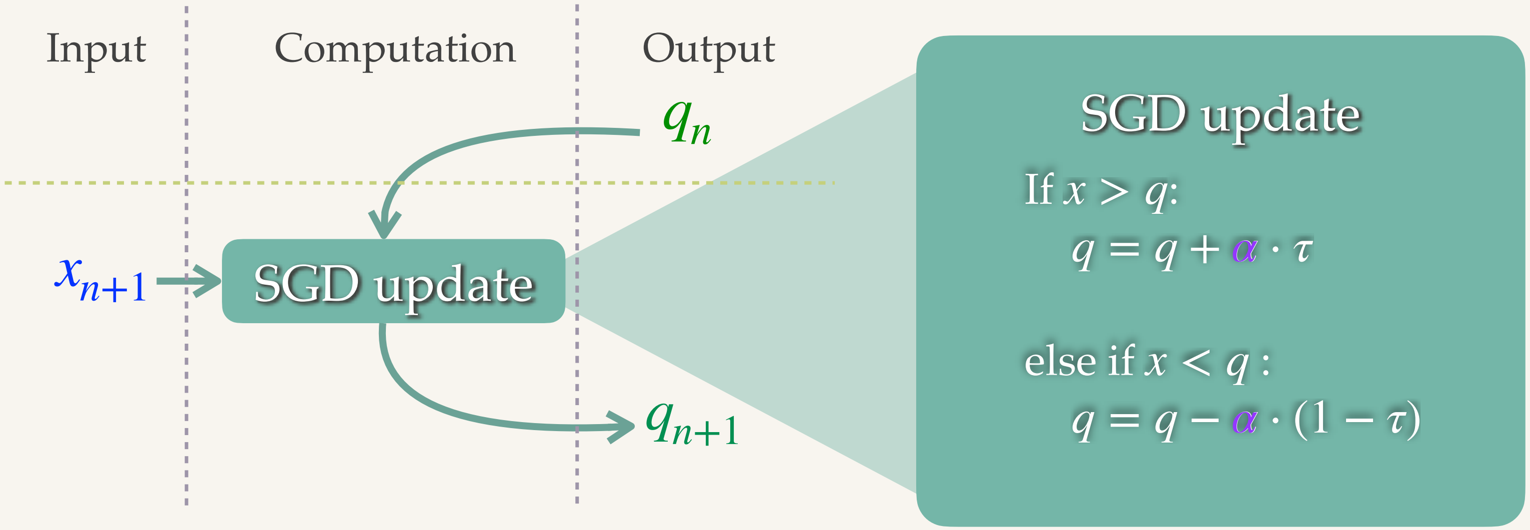
- **Stochastic Gradient Descent (SGD)**
  - **Gradient descent**: a convex optimization method, takes gradient from the **entire dataset**.
  - But for data streams, the entire data set is unavailable.
  - ***Stochastic* Gradient Descent**: stochastic approximation of gradient descent, using gradient from only the **latest coming data**.



SGD quantile estimation for a normal distribution

SGD on Quantile Estimation

Update of quantile for a new coming data:



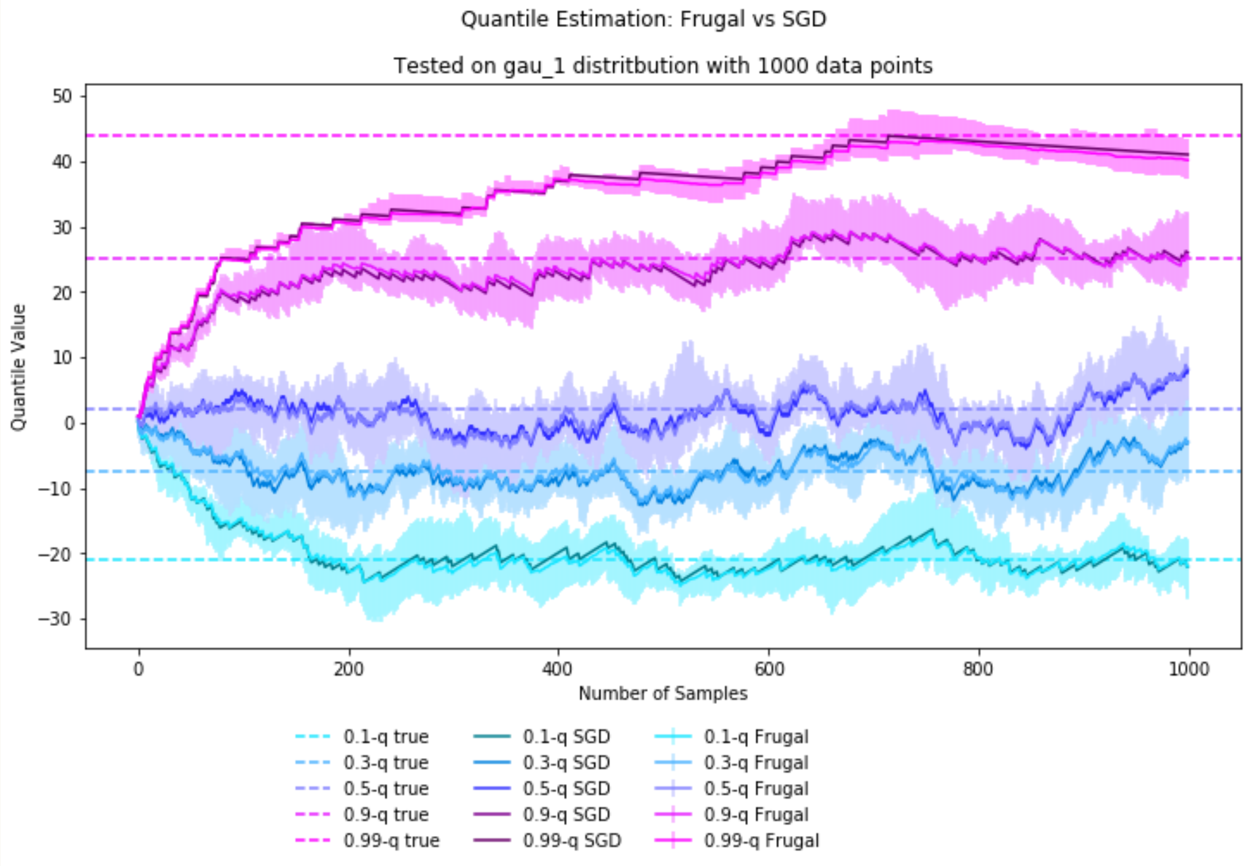
α is the step size for each step

Computationally simple and cheap

Space complexity **O(1)** since only one unit of memory is needed for the current quantile estimate q_n

SGD is Equivalent to Frugal-1U

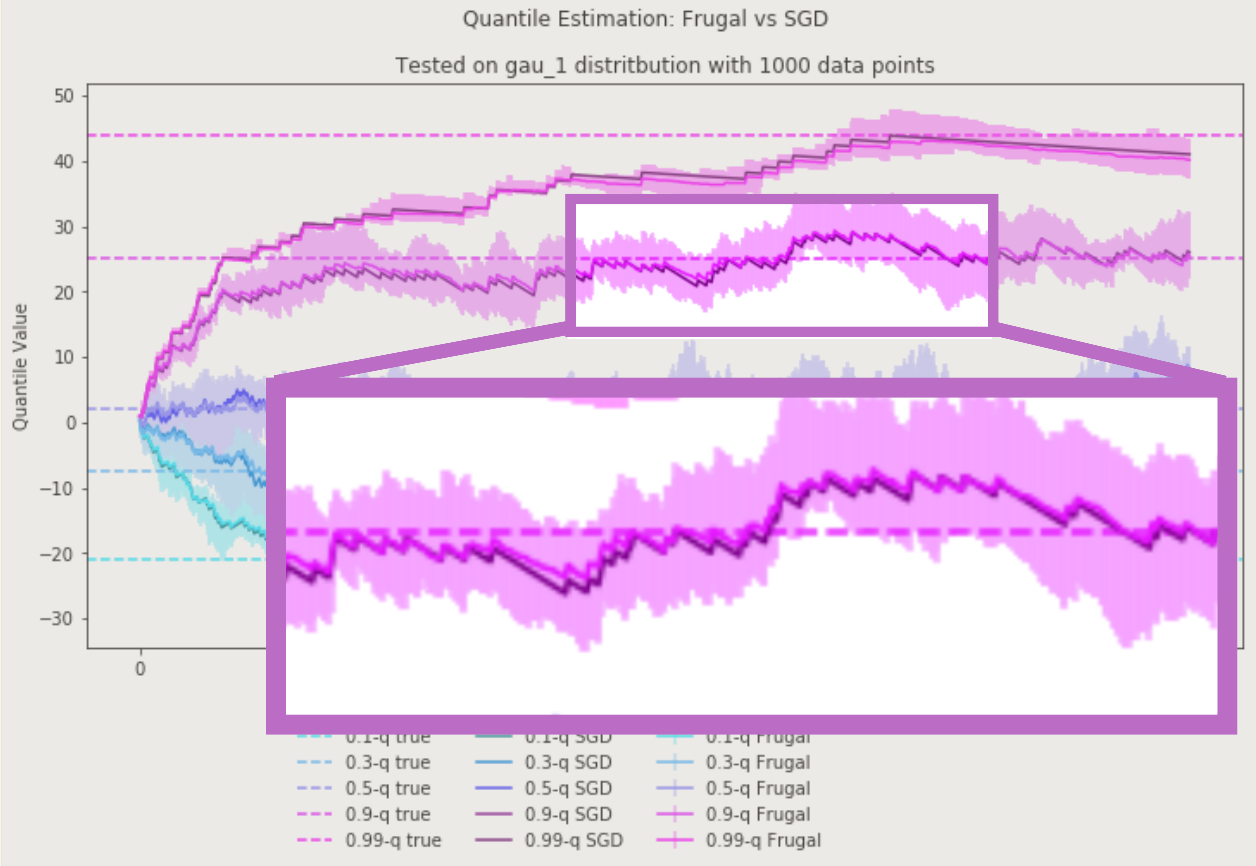
- **Frugal-1U**[7] is the current **state of the art** with **O(1)** space complexity that is very close to the SGD algorithm.
- **Empirically** very similar performances.
- **Theoretically**, the equivalence holds when SGD step size $\alpha = 1$.
- The result expectation of Frugal-1U is the same as SGD.



SGD is Equivalent to Frugal-1U

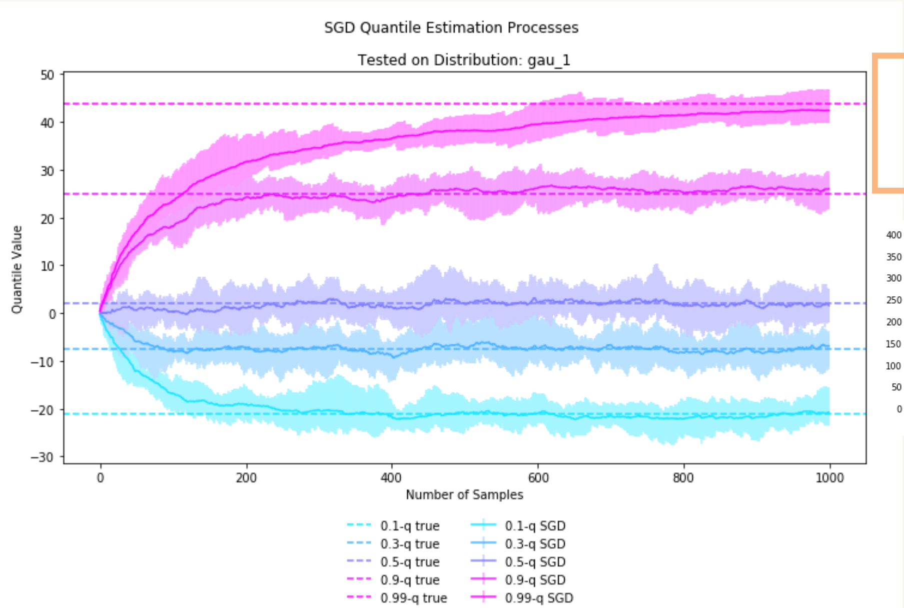
- **Frugal-1U**[7] is the current **state of the art** with **O(1)** space complexity that is very close to the SGD algorithm.

- **Empirically** very similar performances.
- **Theoretically**, the equivalence holds when SGD step size $\alpha = 1$.
- The result expectation of Frugal-1U is the same as SGD.

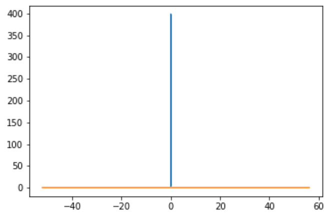


SGD Convergence Experiment - Trend

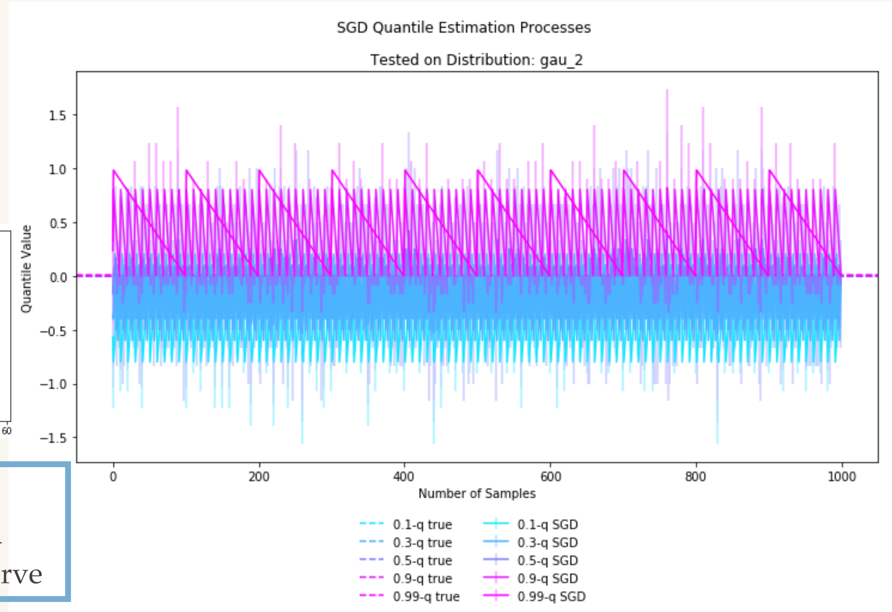
Same SGD method (step size = 1) on **different gaussian distributions**



Mean: 2
SD: **18**
Flat bell curve



Mean: 0
SD: **0.001**
Steep bell curve

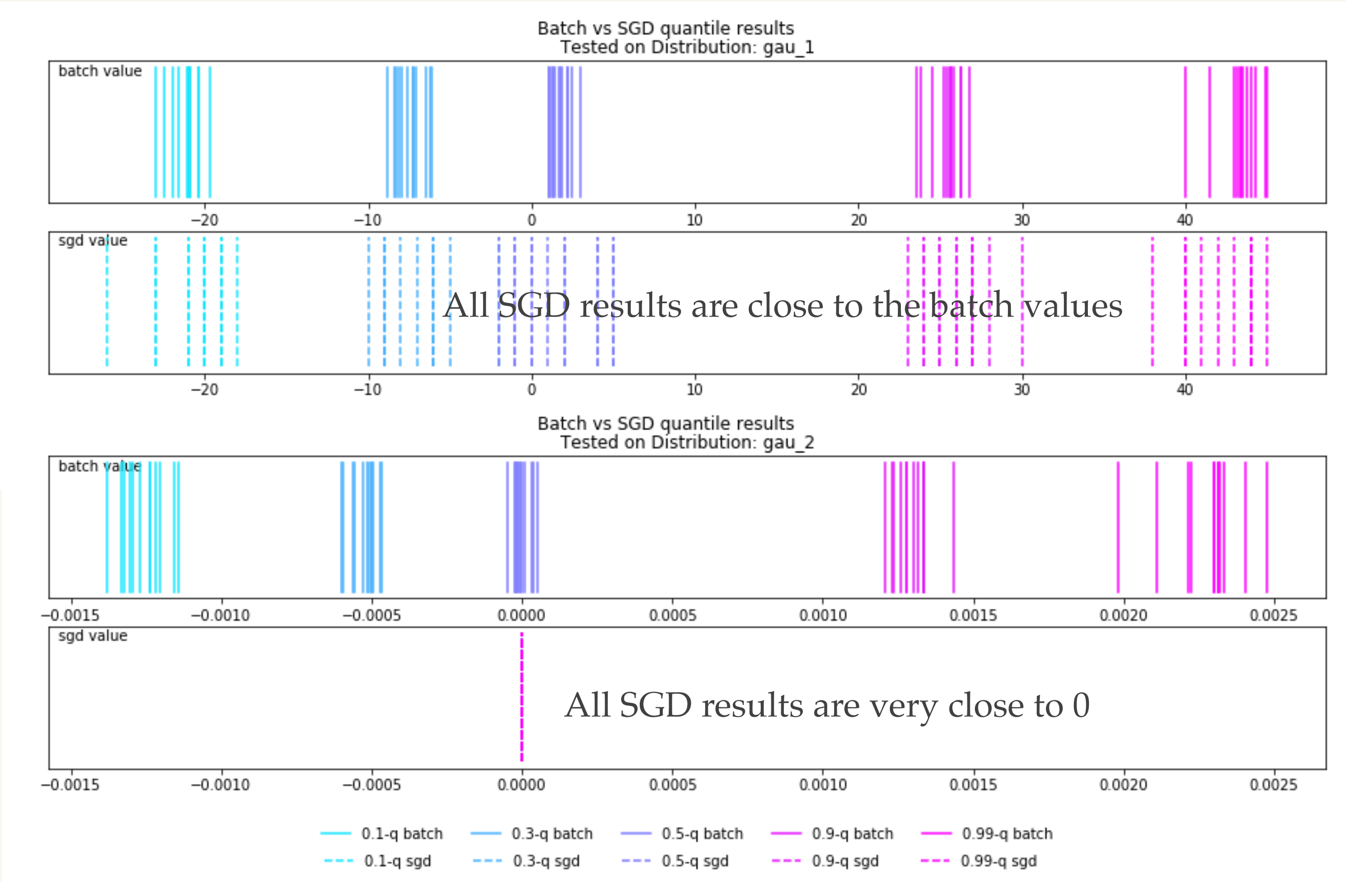


When the basic setting works: each quantile estimate steps to the true quantile value and remains stable around it.
When the basic setting fails: need to fix the problems of **step sizes, quantile crossing**, and convergence

SGD Convergence Experiment - Results

- **Technical details: Does it "work" or not?**

- For each quantile value, compare the SGD estimate results with the **calculated batch_values**

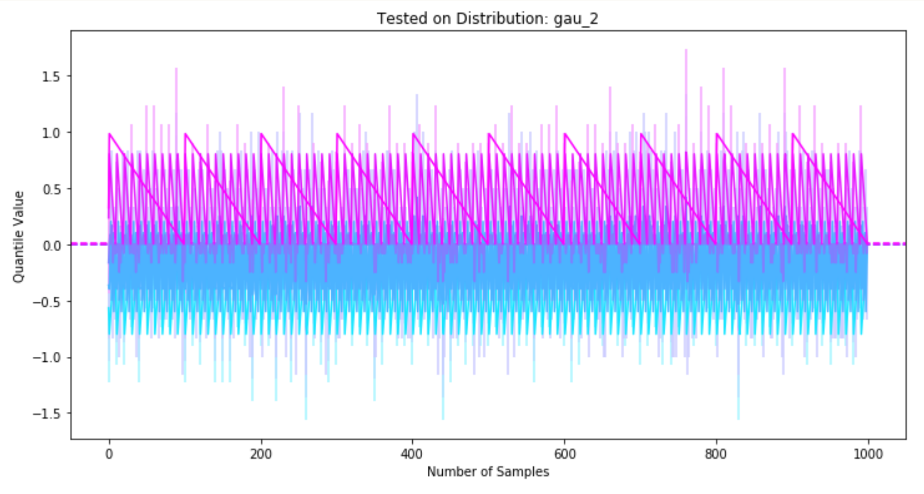


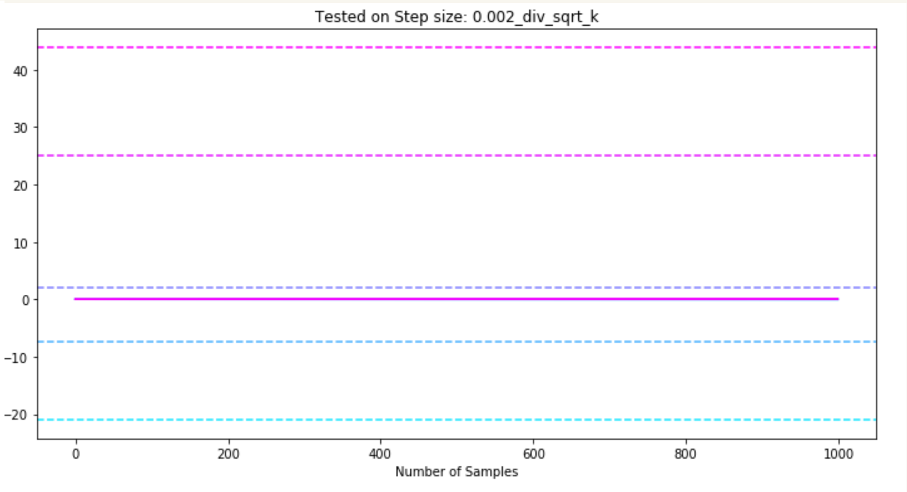
Mean: 2
SD: **18**
Flat bell curve

Mean: 0
SD: **0.001**
Steep bell curve

Problem 1: Selection of Step Size

- There is no single setting of SGD step size that fits all input data streams





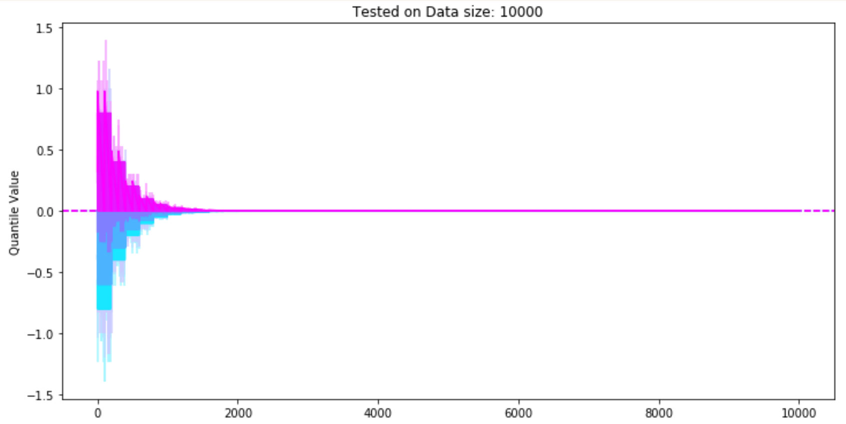
Step size	Fast convergence	Small fluctuation after convergence	Other problems
Big	✓	✗	Quantile crossing, super inaccurate estimation
Small	✗	✓	Might takes much more (e.g, 10000+ times) data to finally converge

Method 1: DH-SGD

- **More flexible** adjustive step size that changes with regards to the distributions?

- *Doubling and Halving SGD* **(DH-SGD):**
  - Intuition: **Change** step size according to the **latest estimation record**.
  - Implementation Details:
    - For each **intervals of** C **updates**, record the number of increasing and decreasing updates
    - **Double** the step size if it is too small, and **halve** the step size if it is too big.

Constant step size $\alpha_n = 1$

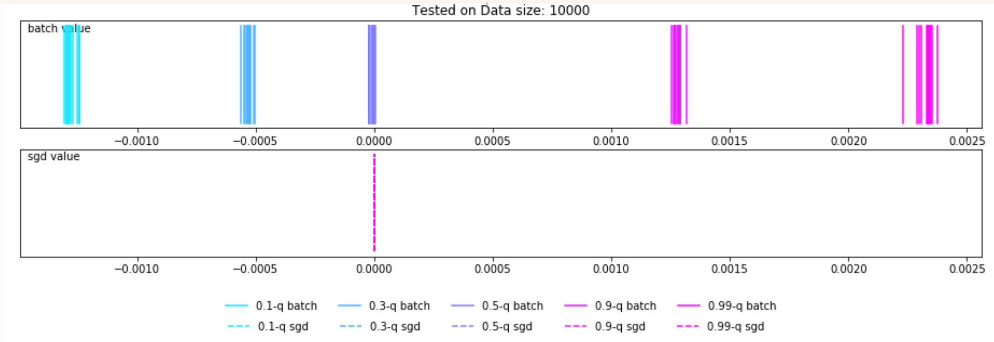


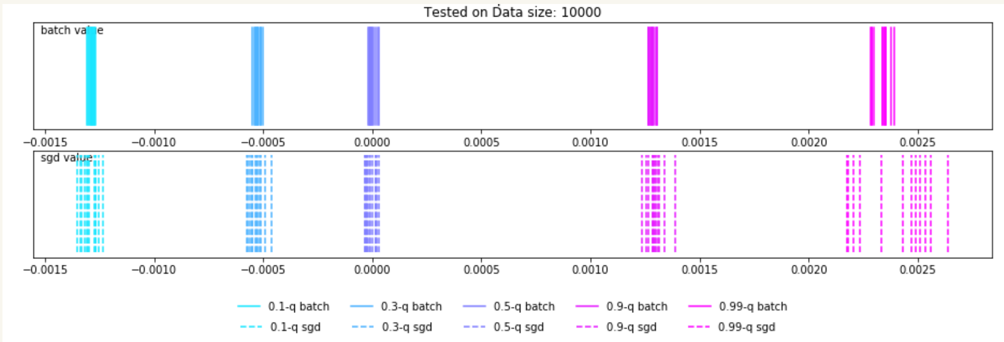
DH-SGD Adjustive step size based on previous update trends

Method 1: DH-SGD

- **More flexible** adjustive step size that changes with regards to the distributions?

- *Doubling and Halving SGD* **(DH-SGD):**
  - Intuition: **Change** step size according to the **latest estimation record.**
  - Implementation Details:
    - For each **intervals of** C **updates**, record the number of increasing and decreasing updates
    - **Double** the step size if it is too small, and **halve** the step size if it is too big.



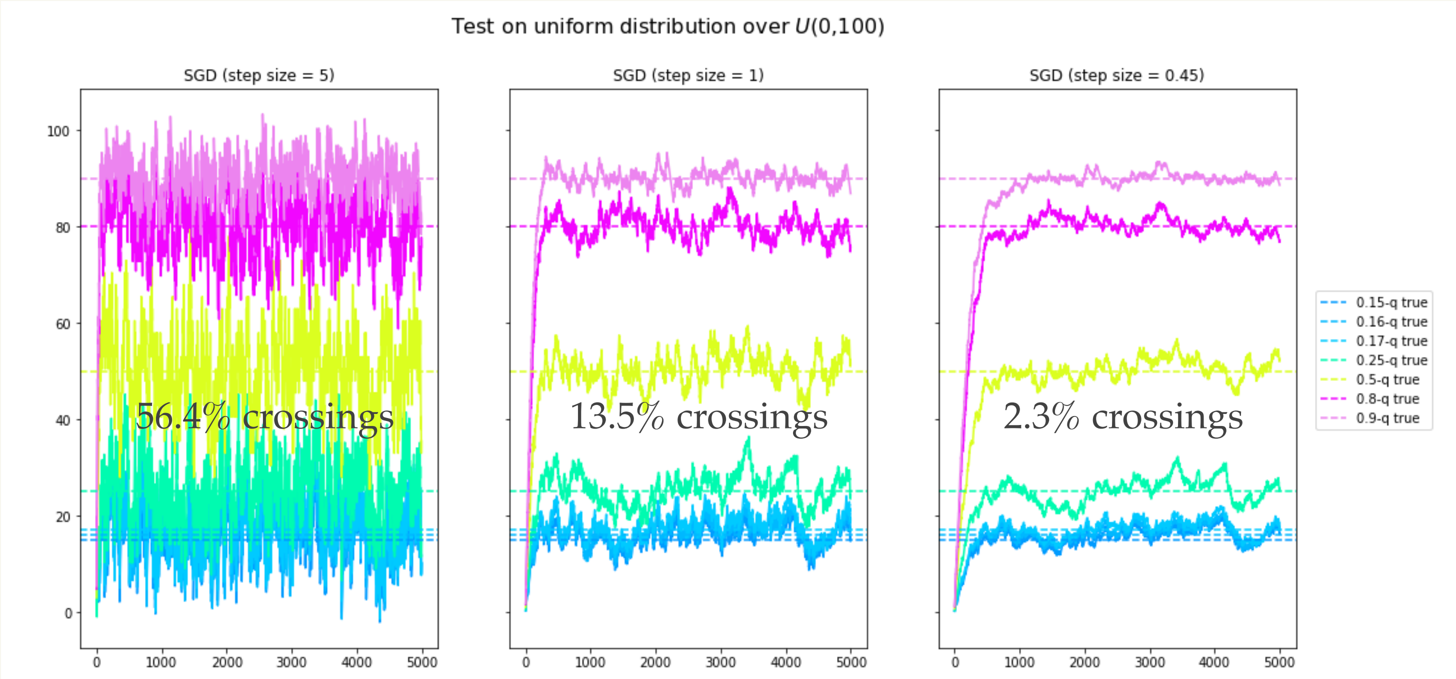


Constant step size $\alpha_n = 1$

DH-SGD Adjustive step size based on previous update trends

Problem 2: Multi-quantile crossing

Monotone property: for a smooth distribution, we have $q_n^{(1)} < q_n^{(2)} < \ldots < q_n^{(K)}$

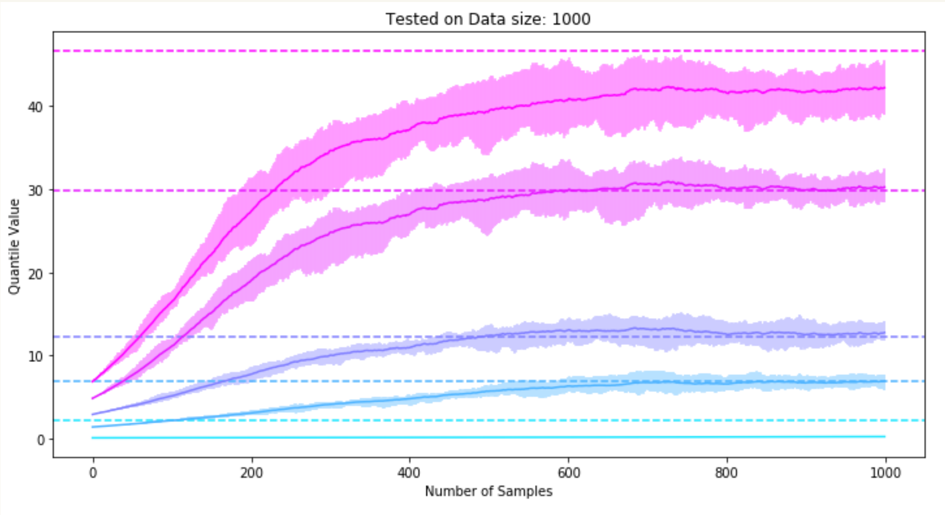


Missing Opportunity: We can use the extra information that the quantiles are in increasing order

Method 2: shiftQ or Extended P^2

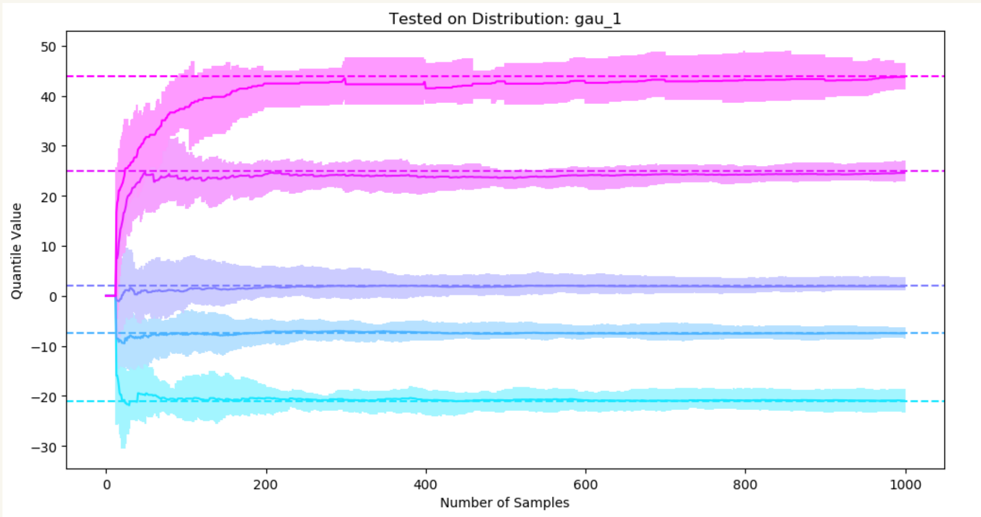
We have **not yet** come up with a SGD solution to it

Other people's quantile estimation methods, both are **non-SGD methods**



shiftQ algorithm

Note: It works only when all data are strictly positive

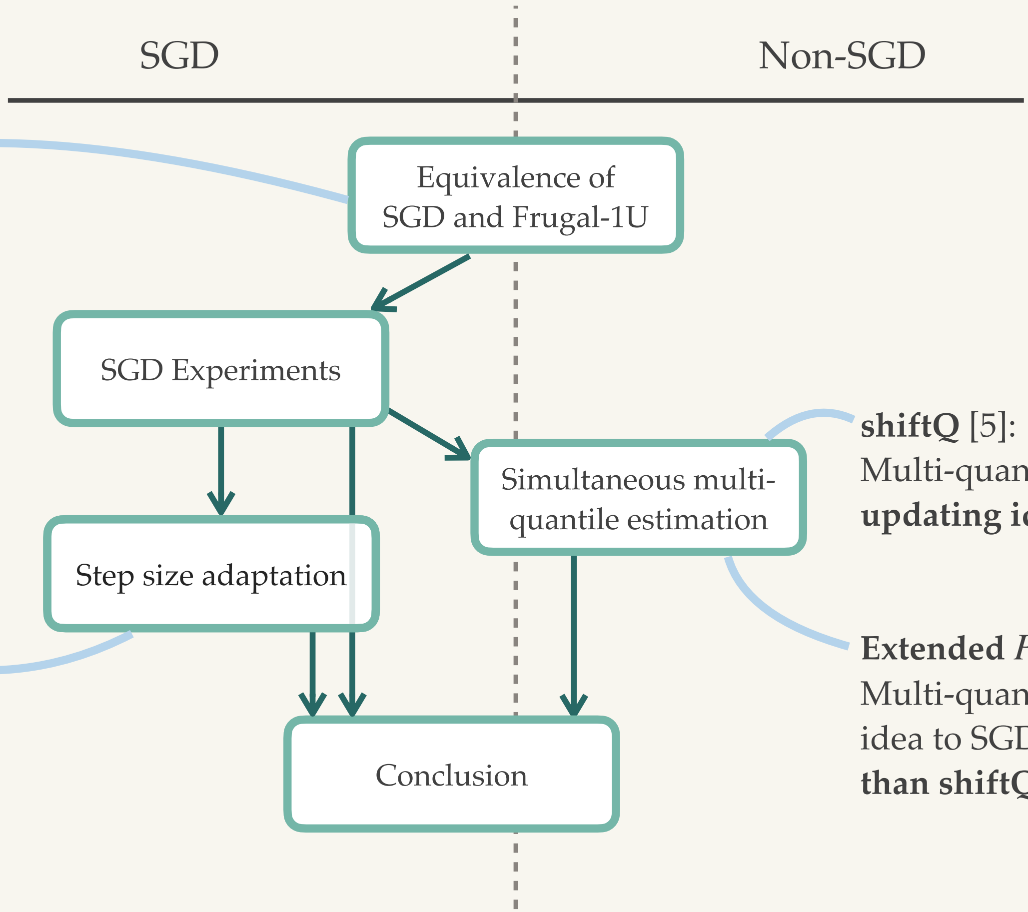


Extended P^2 algorithm

Missing methods in this talk

Frugal-1U:
Very similar to SGD, but approaches the quantile estimation **from a different aspect**.

SAG:
Step size adaptation method that implements the **machine learning** approach SAG for **better convergence** rate.



shiftQ [5]:
Multi-quantile method, uses similar **updating idea to SGD**

Extended P^2 [6]:
Multi-quantile method, uses very different idea to SGD, but has **better performance than shiftQ**.

Summary & Future Work

- SGD works, and is equivalent to Frugal-1U
- Different settings of experiments affect the performance of SGD
- Proposed step size adaptation algorithms (DH-SGD, SAG) are effective

- Improve the DH-SGD algorithm
- Multi-quantile SGD estimation development
- Experiments on real data

References

[1] Michael B. Greenwald and Sanjeev Khanna. “Quantiles and Equi-Depth Histograms over Streams.” en. In: *Data Stream Management*. Ed. by Mi- nos Garofalakis, Johannes Gehrke, and Rajeev Rastogi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 45–86. isbn: 978-3-540-28607-3 978-3- 540-28608-0.

[2] Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, and Subhash Suri. “Medians and beyond: New Aggregation Techniques for Sensor Networks.” In: Proceedings of the 2nd International Conference on Em- bedded Networked Sensor Systems. SenSys ’04. Baltimore, MD, USA: Associa- tion for Computing Machinery, Nov. 2004, pp. 239–249. isbn: 978-1-58113-879-5.

[3]Graham Cormode and S. Muthukrishnan. “An Improved Data Stream Summary: The Count-Min Sketch and Its Applications.” en. In: Journal of Algorithms 55.1 (Apr. 2005), pp. 58–75. issn: 01966774.

[4]David Felber and Rafail Ostrovsky. “A Randomized Online Quantile Sum- mary in O((1/ε)Log (1/ε)) Words.” en. In: *Theory of Computing* 13.1 (2017), pp. 1–17. issn: 1557-2862.

[5]Raj Jain and Imrich Chlamtac. 1985. The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations. Commun. ACM 28, 10 (October 1985), 1076–1085.

[6]Hugo Lewi Hammer, Anis Yazidi, and Håvard Rue. “Joint Tracking of Multiple Quantiles Through Conditional Quantiles.” In: *arXiv:1902.05428 [stat]* (Feb. 2019).

[7]Qiang Ma, S. Muthukrishnan, and Mark Sandler. “Frugal Streaming for Es- timating Quantiles:One (or Two) Memory Suffices.” en. In: *arXiv:1407.1121 [cs]* (July 2014). arXiv: 1407.1121 [cs].