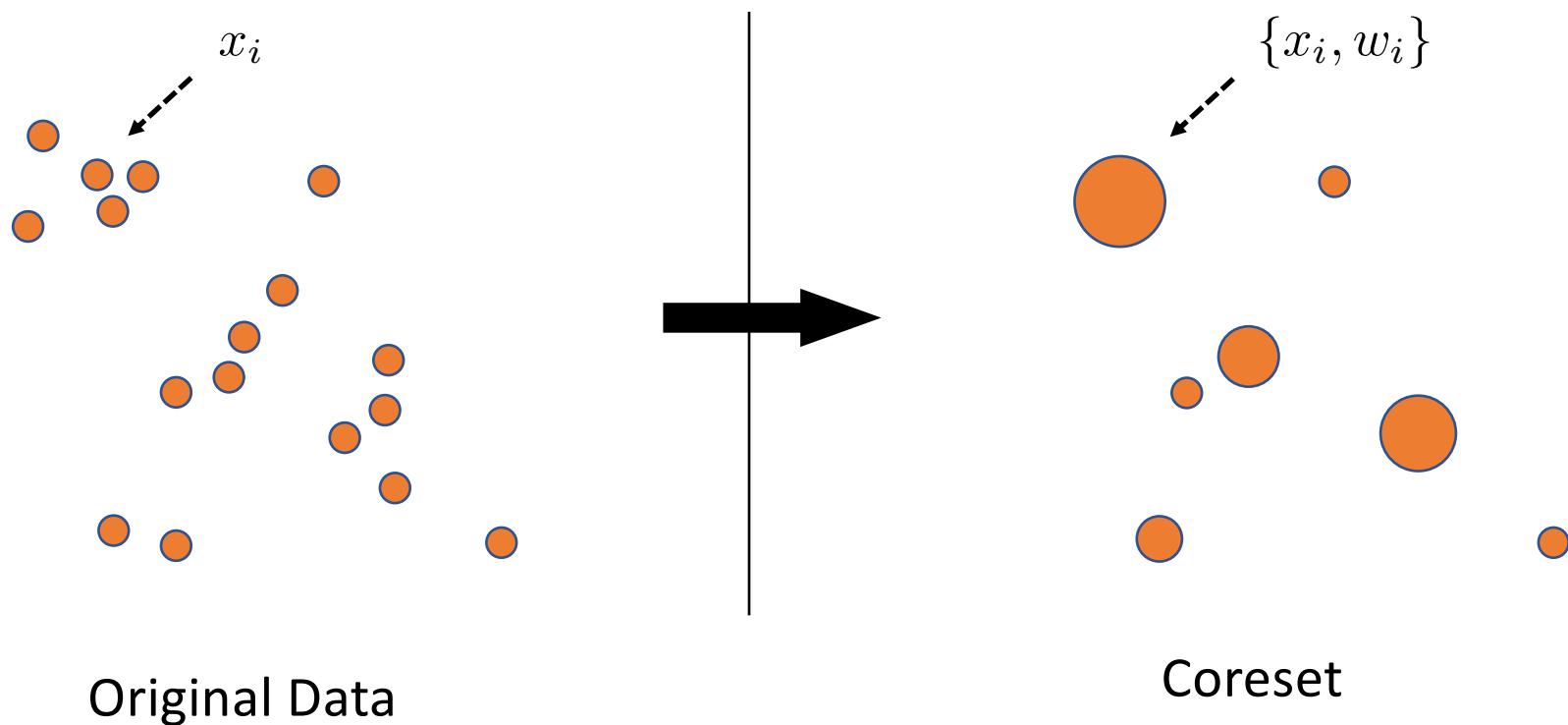


Coresets, Discrepancy, and Sketches in Machine Learning

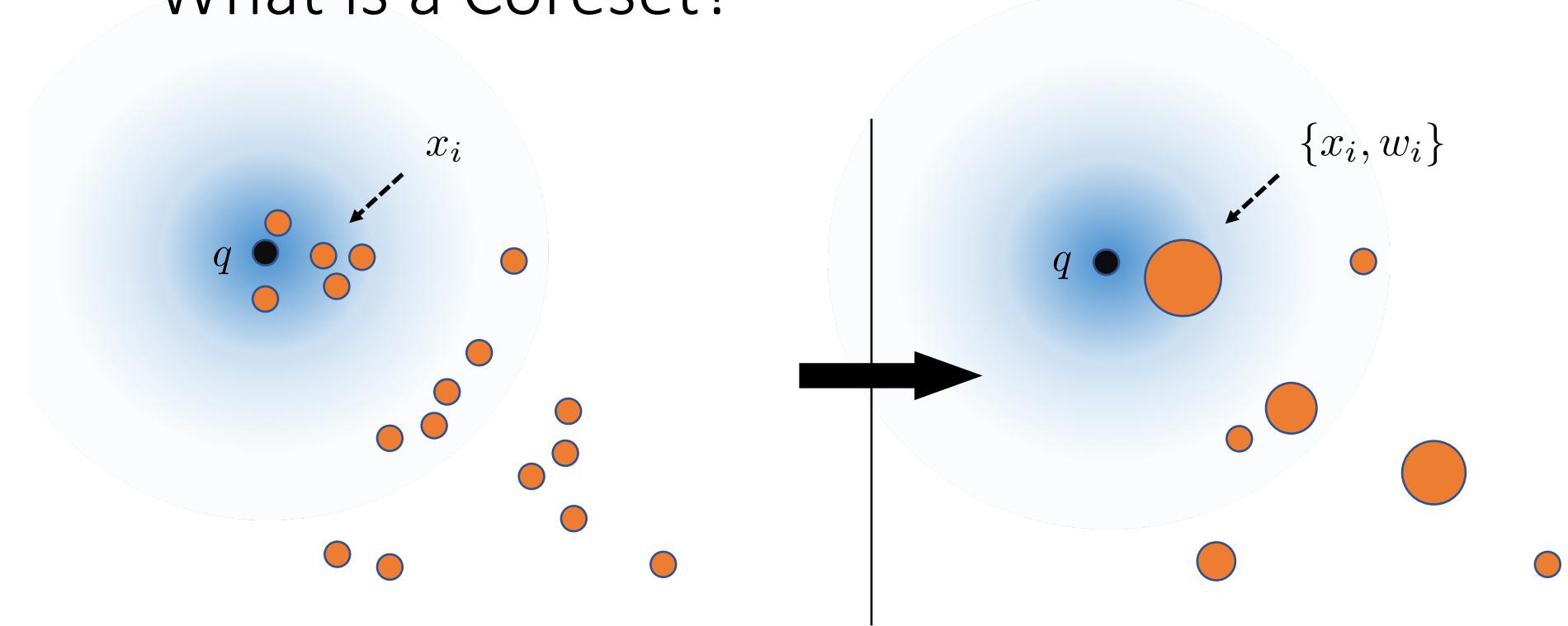
Edo Liberty – Research Director, Amazon

Zohar Karnin – Principal Scientist, Amazon

What is a Coreset?



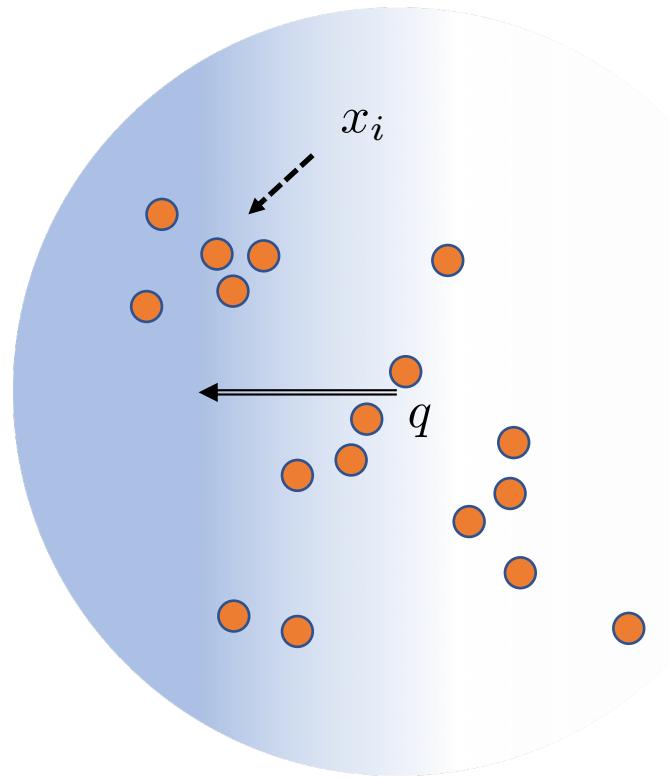
What is a Coreset?



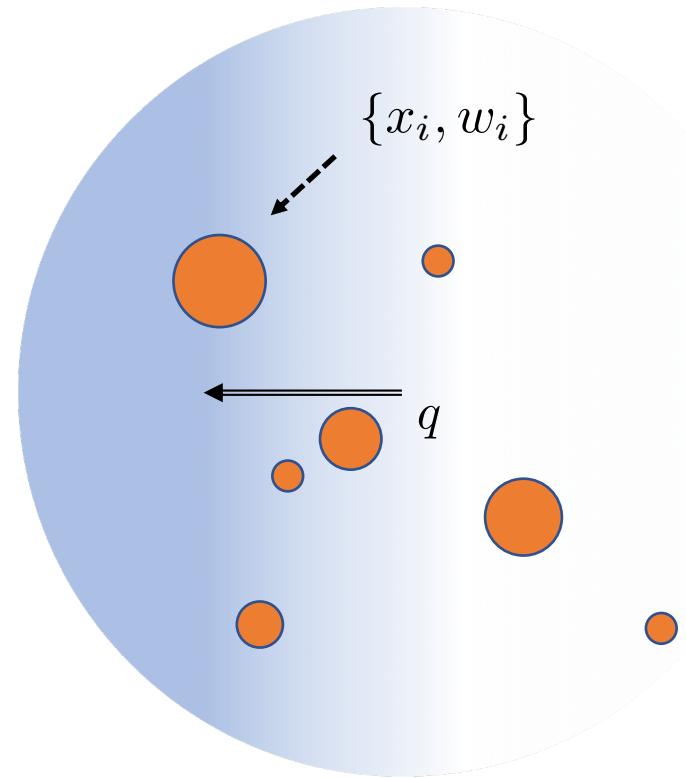
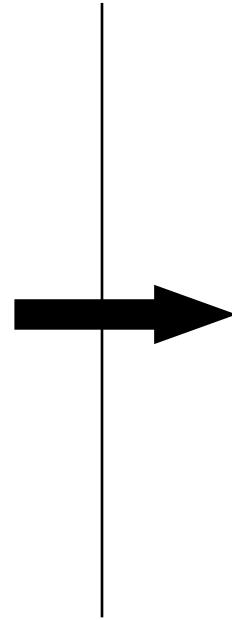
$$F(q) = \sum_i f(x_i, q)$$

$$\tilde{F}(q) = \sum_{i \in S} w_i f(x_i, q)$$

What is a Coreset?



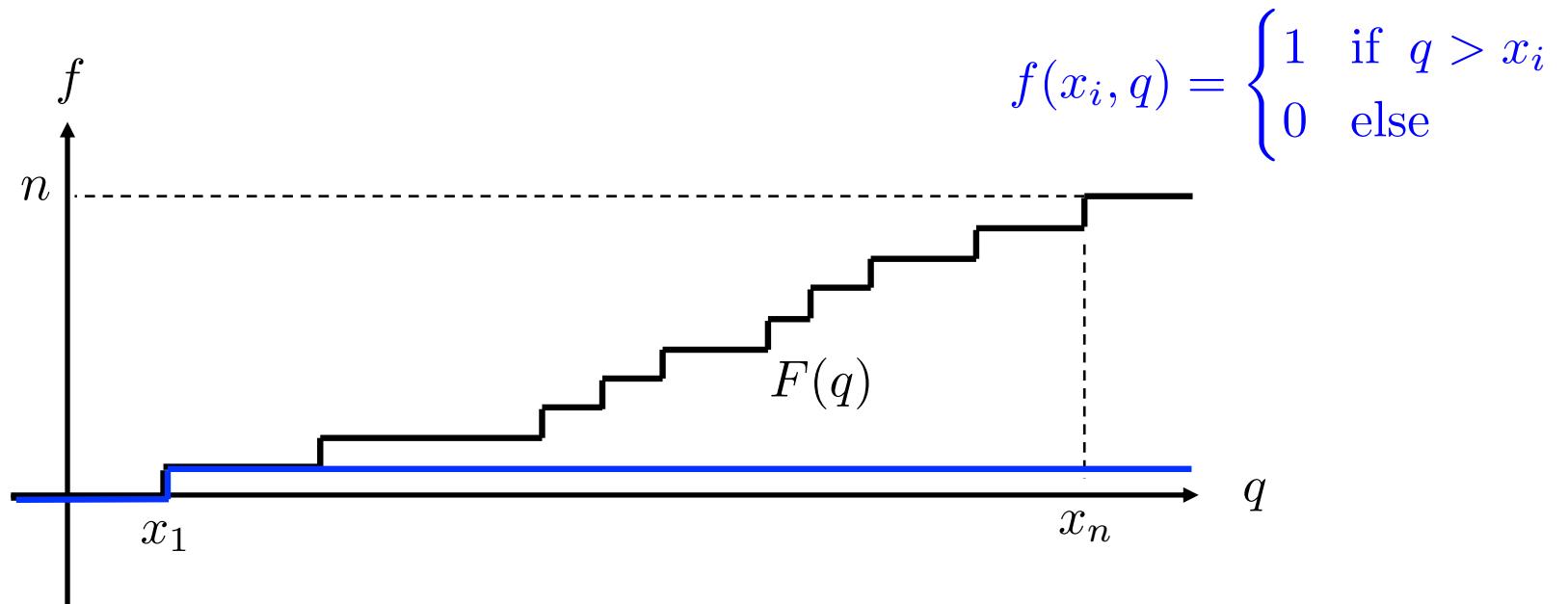
$$F(q) = \sum_i f(x_i, q)$$



$$\tilde{F}(q) = \sum_{i \in S} w_i f(x_i, q)$$

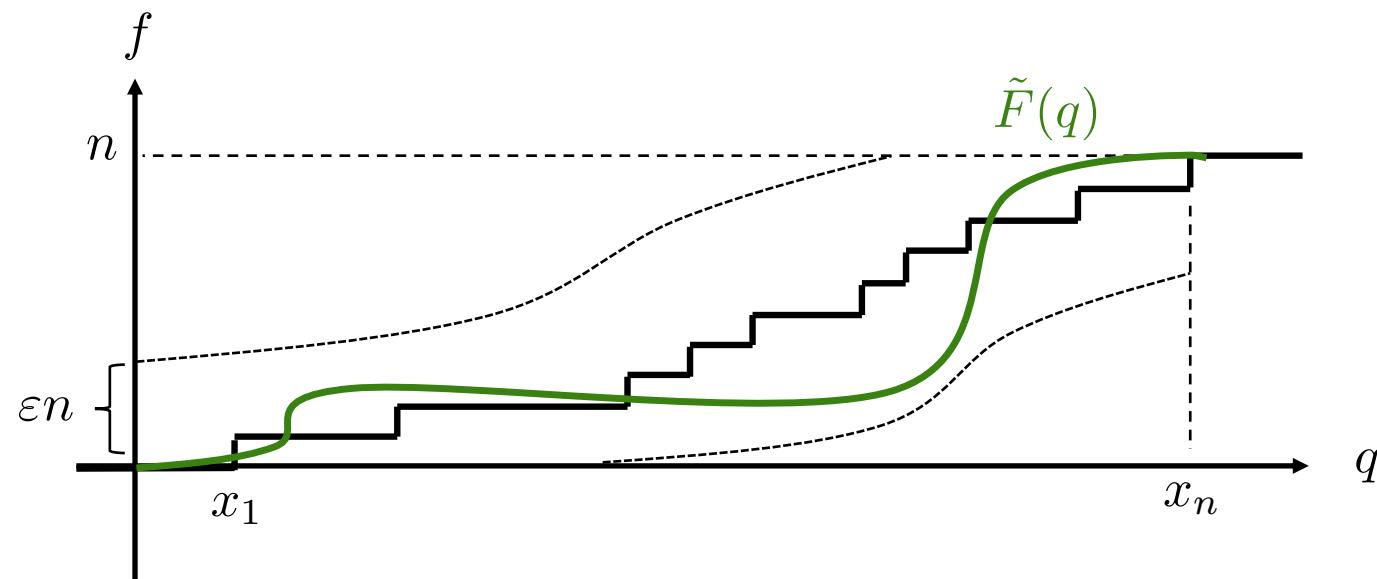
Approximate CDF

The (empirical) CDF is given by $F(q) = \sum_i f(x_i, q)$



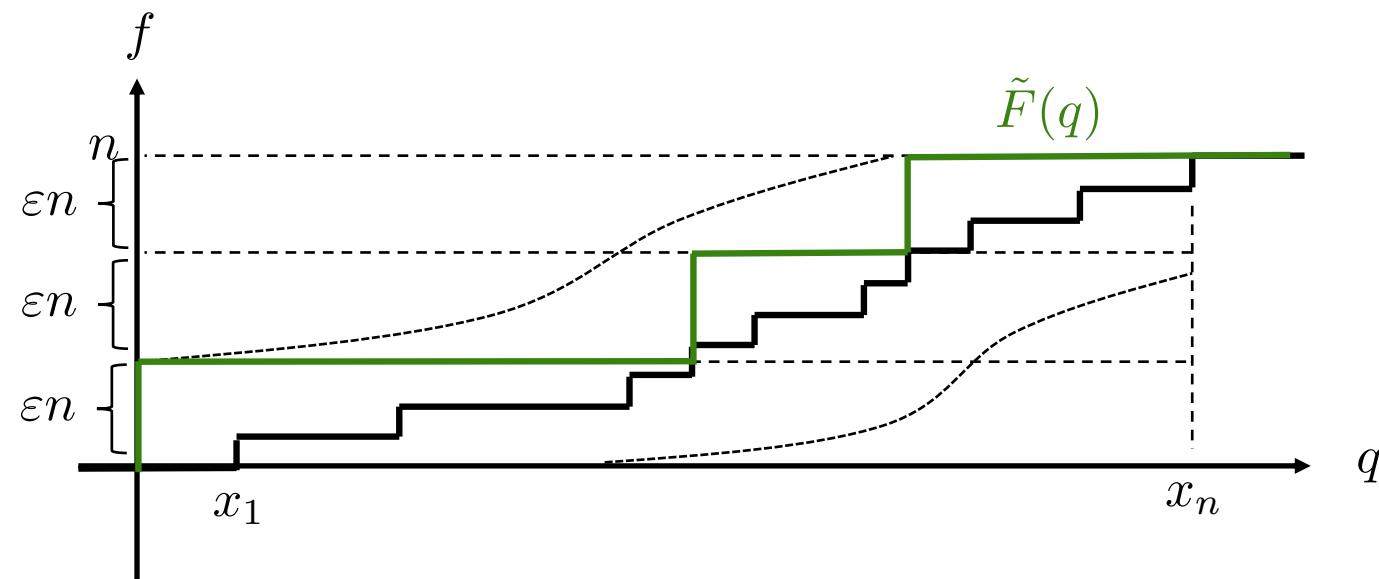
Approximate CDF

An approximate CDF is given by $|\tilde{F}(q) - F(q)| \leq \varepsilon n$



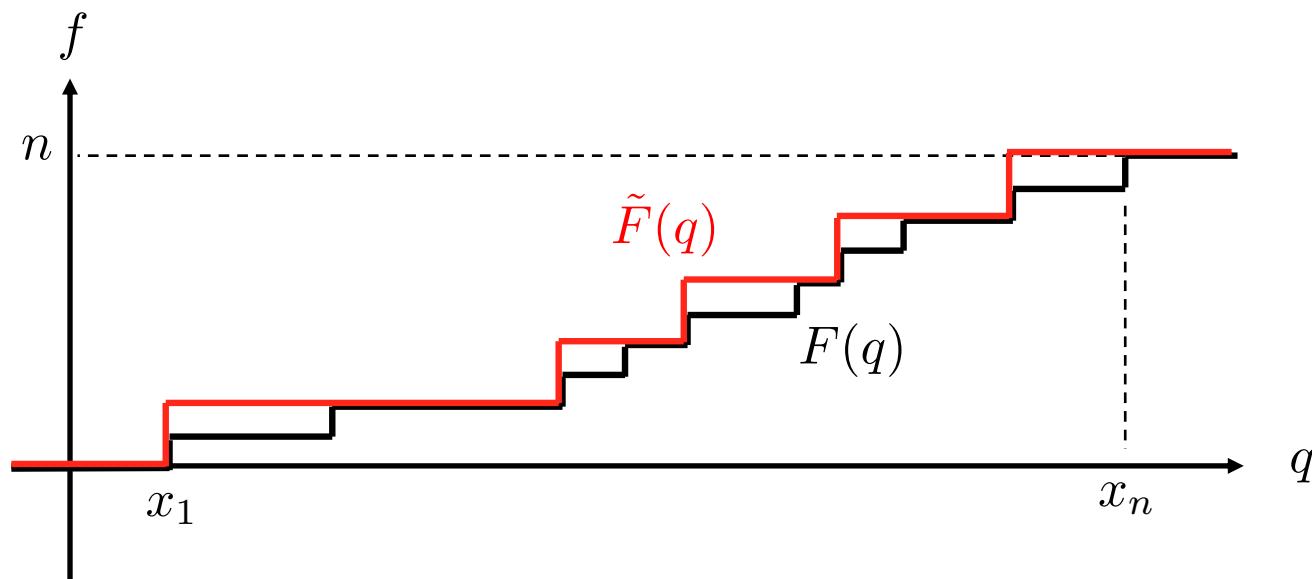
Approximate CDF

There is a trivial coresset of size $1/\varepsilon$



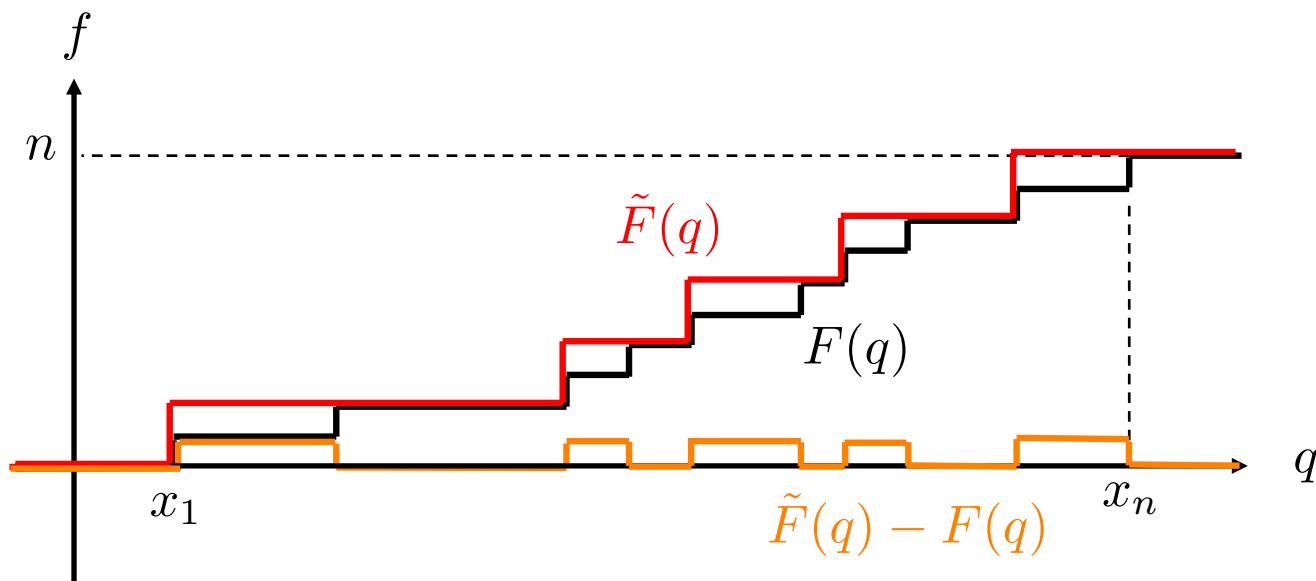
Approximate CDF

One way to get there is as like this: $\tilde{F}(q) = \sum_{i \in S} 2f(x_i, q)$



Approximate CDF

The Discrepancy is $\tilde{F}(q) - F(q) = \sum_i \sigma_i f(x_i, q) \leq 1$



Question:

For the function $f(x_i, q) = \begin{cases} 1 & \text{if } q > x_i \\ 0 & \text{else} \end{cases}$

1) We have that $\min_{\sigma} \max_q \left| \sum_i \sigma_i f(x_i, q) \right| \leq 1$

2) We have a coresnet of size $1/\varepsilon$

Does this generalize?

Answer: Yes

Definition: Class Discrepancy $D_m = \min_{\sigma} \max_q \frac{1}{m} \left| \sum_{i=1}^m \sigma_i f(x_i, q) \right|$

Lemma: For any function whose Class Discrepancy is $D_m = O(c/m)$

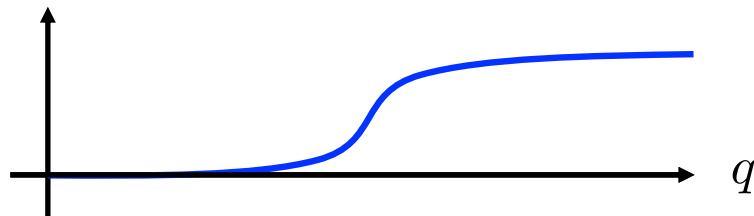
Its coresset complexity is $O(c/\varepsilon)$

Its *streaming* coresset complexity is $O\left(c/\varepsilon \cdot \log^2(\varepsilon n/c)\right)$

Its randomized *streaming* coresset complexity is $O\left(c/\varepsilon \cdot \log^2 \log(|Q_\varepsilon|/\delta)\right)$

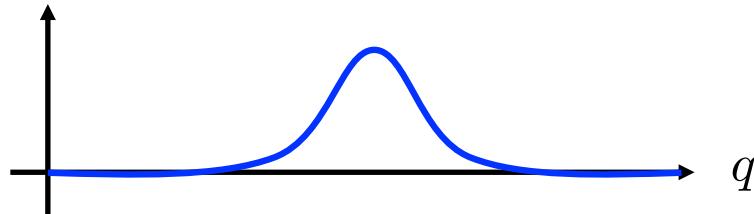
Bounding the Class Discrepancy

$$f(x, q) = 1/(1 + e^{x-q})$$



$$D_m = c/m$$

$$f(x, q) = \exp(-(x - q)^2)$$

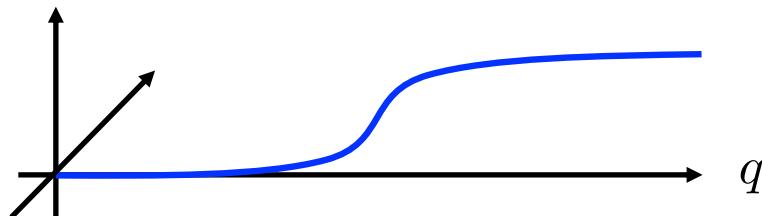


$$D_m = c/m$$

Bounding the Class Discrepancy

Sigmoid Activation Regression

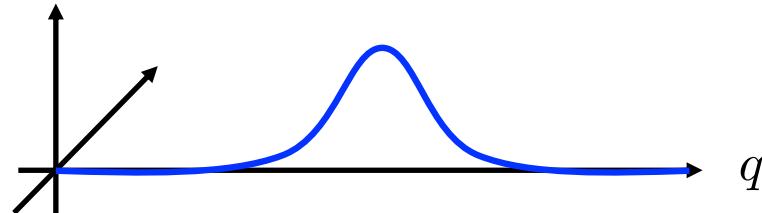
$$f(x, q) = 1/(1 + e^{-\langle x, q \rangle})$$



$$D_m = ?$$

Gaussian Kernel Density

$$f(x, q) = \exp(-\|x - q\|^2)$$



$$D_m = ?$$

Interesting Connection

Class Discrepancy

$$D_m = \min_{\sigma} \max_q \frac{1}{m} \left| \sum_{i=1}^m \sigma_i f(x_i, q) \right|$$

Usually: $D_m = O(c/m)$

Governs: Coreset Complexity

Rademacher Complexity

$$R_m = \mathbb{E}_{\sigma} \max_q \frac{1}{m} \left| \sum_{i=1}^m \sigma_i f(x_i, q) \right|$$

Usually: $K_m \approx O(c/\sqrt{m})$

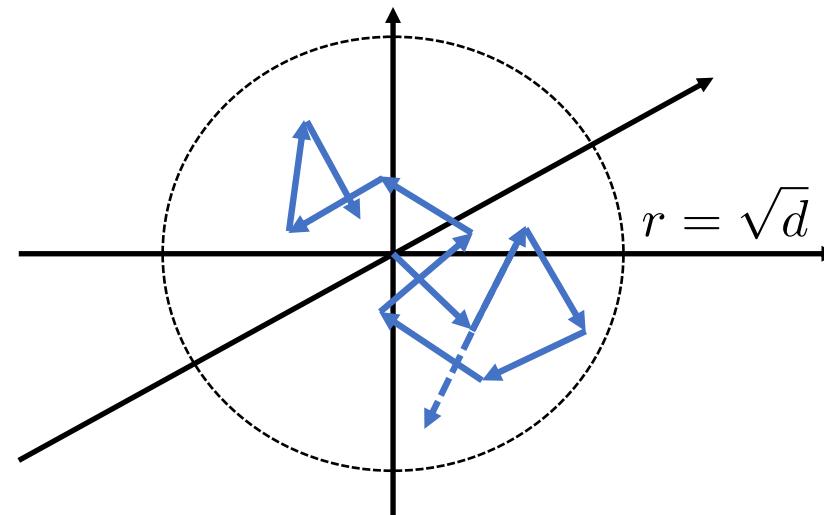
Governs: Sample Complexity

Look at techniques for bounding the Rademacher Complexity for inspiration...

Bounding sums of vectors

$$\min_{\sigma} \left\| \sum_{i=1}^n \sigma_i x_i \right\| \leq \sqrt{d}$$

Does not depend on n



$$\mathbb{E}_{\sigma} \left\| \sum_{i=1}^n \sigma_i x_i \right\| \approx \sqrt{n}$$

That's encouraging.....

Bounding sums of matrices

$$\min_{\sigma} \left\| \sum_{i=1}^n \sigma_i x_i x_i^T \right\| \leq ?$$

I did not know the answer to this one for a while...

Bounding sums of matrices

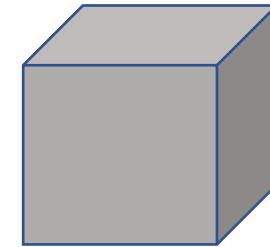
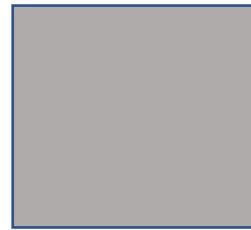
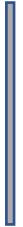
$$\min_{\sigma} \left\| \sum_{i=1}^n \sigma_i x_i x_i^T \right\| = O(\sqrt{d})$$

Proof [*Due to Nikhil Bansal in private communication*]

This is a clever application of Banaszczyk's theorem together with standard bounds on the spectral norm of random matrices.

It's also tight.

Bounding sums of all tensors powers



$$x = x^{\otimes 1}$$

$$xx^T = x^{\otimes 2}$$

$$\text{outer}(x, x, x) = x^{\otimes 3}$$

Lemma [Karnin, L]: For any set of vectors $x_i \in \mathbb{R}^d$ there exist signs σ such that for all k simultaneously

$$\left\| \sum_{i=1}^n \sigma_i x_i^{\otimes k} \right\| \leq \underbrace{\sqrt{d} \cdot \text{poly}(k)}_{\text{Still does not depend on } n!}$$

Bounding the Class Discrepancy

Lemma: [Karnin, L]: The Class Discrepancy of any analytic function of the dot product is $O(\sqrt{d}/m)$

Proof:

$$\begin{aligned} \sum_i \sigma_i f(\langle x_i, q \rangle) &= \sum_i \sigma_i \underbrace{\sum_k \alpha_k \langle x_i, q \rangle^k}_{\text{Taylor expansion}} \\ &= \sum_k \alpha_k \left\langle \sum_i \sigma_i x_i^{\otimes k}, q^{\otimes k} \right\rangle \\ &\leq \sum_k |\alpha_k| \cdot \left\| \sum_i \sigma_i x_i^{\otimes k} \right\| \stackrel{\text{Constant if } f \text{ is analytic}}{\leq} \sqrt{d} \sum_k |\alpha_k| \cdot \text{poly}(k) \end{aligned}$$

Results

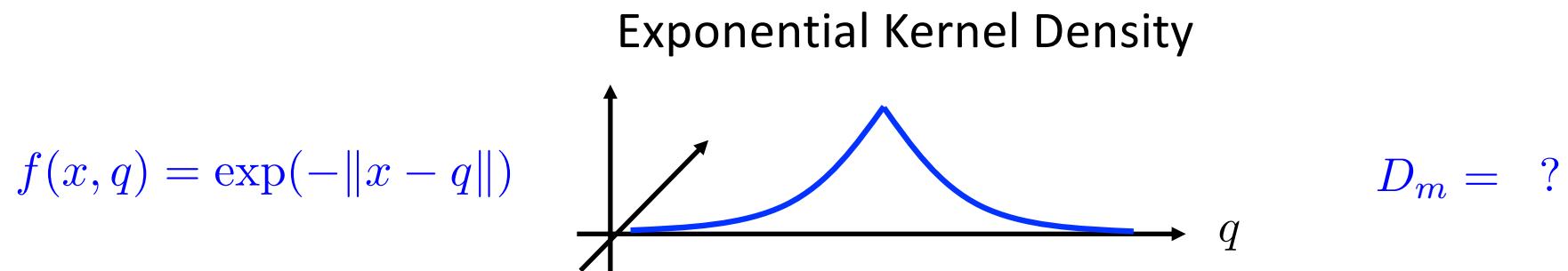
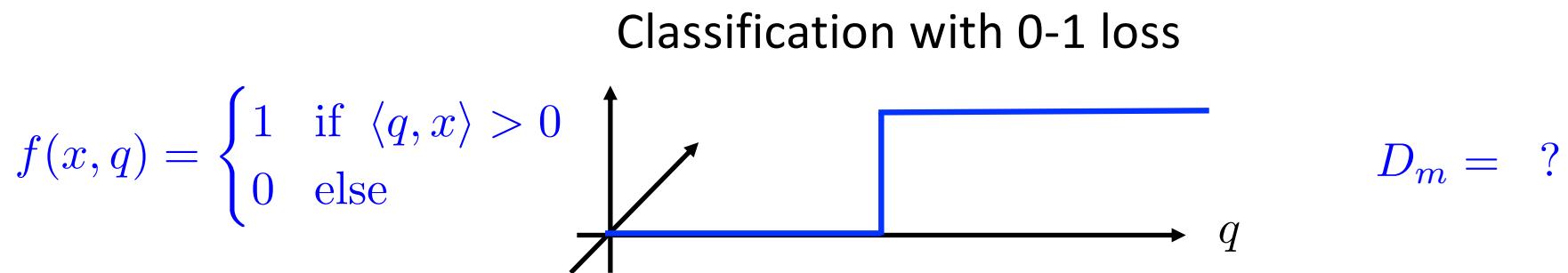
Resolves the open problem
See Philips an Tai 2018

- Sigmoid Activation Regression, Logistic Regression,
- Covariance approximation and Graph Laplacians Quadratic forms
- Gaussian Kernel Density estimation

All have the above have Class Discrepancy of $D_m = O(\sqrt{d}/m)$

- 1) coresets of size $O(\sqrt{d}/\varepsilon)$
- 2) Streaming Coresets of size $O\left(\sqrt{d}/\varepsilon \cdot \log^2\left(\varepsilon n/\sqrt{d}\right)\right)$
- 3) Randomized Streaming Coresets of size $O\left(\sqrt{d}/\varepsilon \cdot \log^2 \log(|Q_\varepsilon|/\delta)\right)$

Back to square one (same only different...)



</slides>

- Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.
- Jeff M Phillips. *Small and stable descriptors of distributions for geometric statistical problems*. PhD thesis, 2009.
- Jeff M. Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*
- Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. 2011
- Jeff M. Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates
- Elad Tolochinsky and Dan Feldman. Coresets for monotonic functions with applications to deep learning.
- Sariel Har-Peled, Dan Roth, and Dav Zimak. Maximum margin coresets for active and noise tolerant learning. *IJCAI 2007*
- Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *The 21st ACM Symposium on Computational Geometry, Pisa, Italy, June 6-8, 2005*
- Gurmeet Singh Manku, Sridhar Rajagopalan, and Bruce G. Lindsay. Random sampling techniques for space efficient online computation of order statistics of large datasets.
- Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression.
- Zohar S. Karnin, Kevin J. Lang, and Edo Liberty. Optimal quantile approximation in streams. *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016*
- Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003
- Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning
- Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Struct. Algorithms*, 12(4):351–360, July 1998