Nearly Linear Sparsification of ℓ_p Subspace Approximation

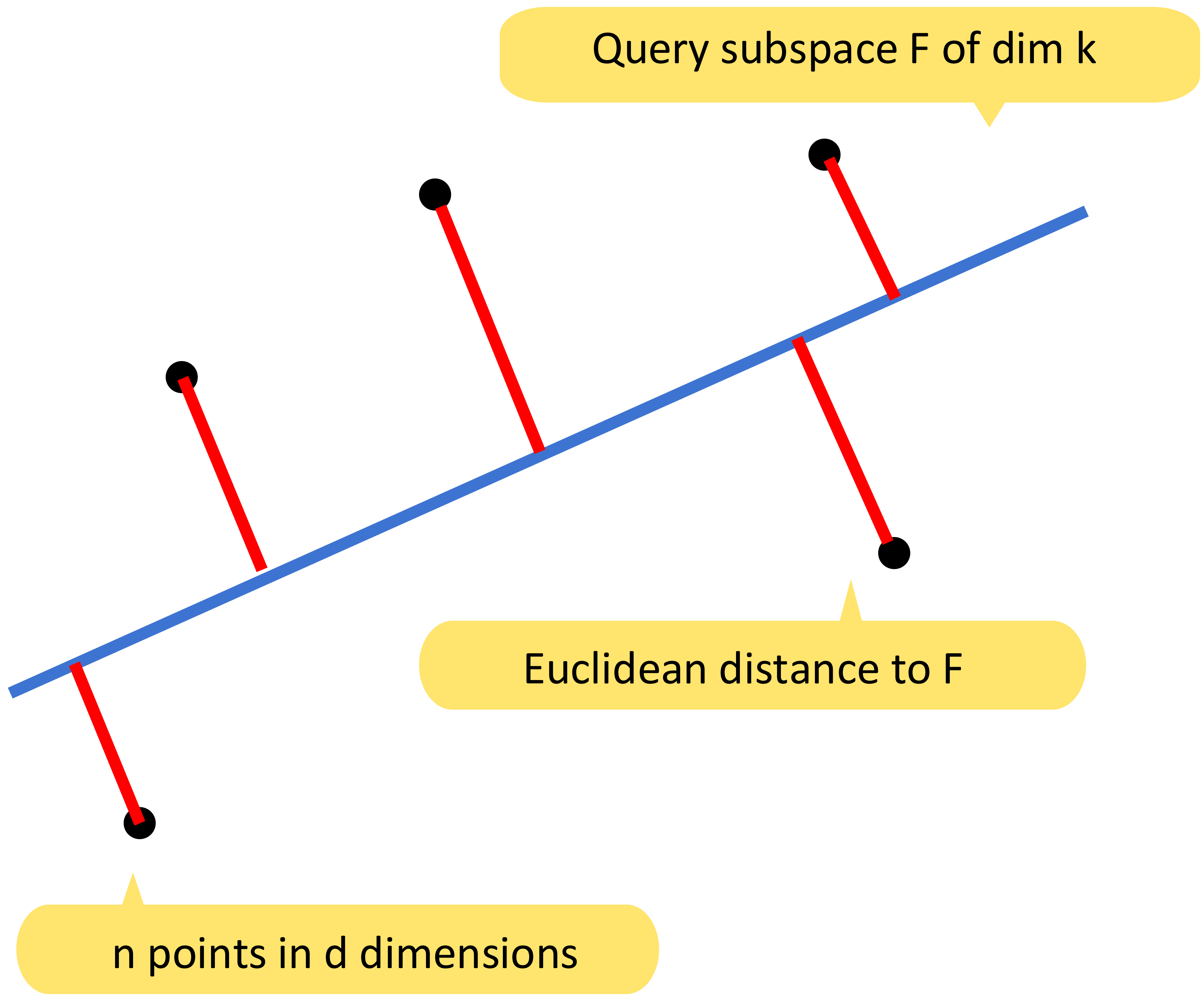
David Woodruff

Carnegie Mellon University
Computer Science Department

Taisuke (Tai) Yasuda

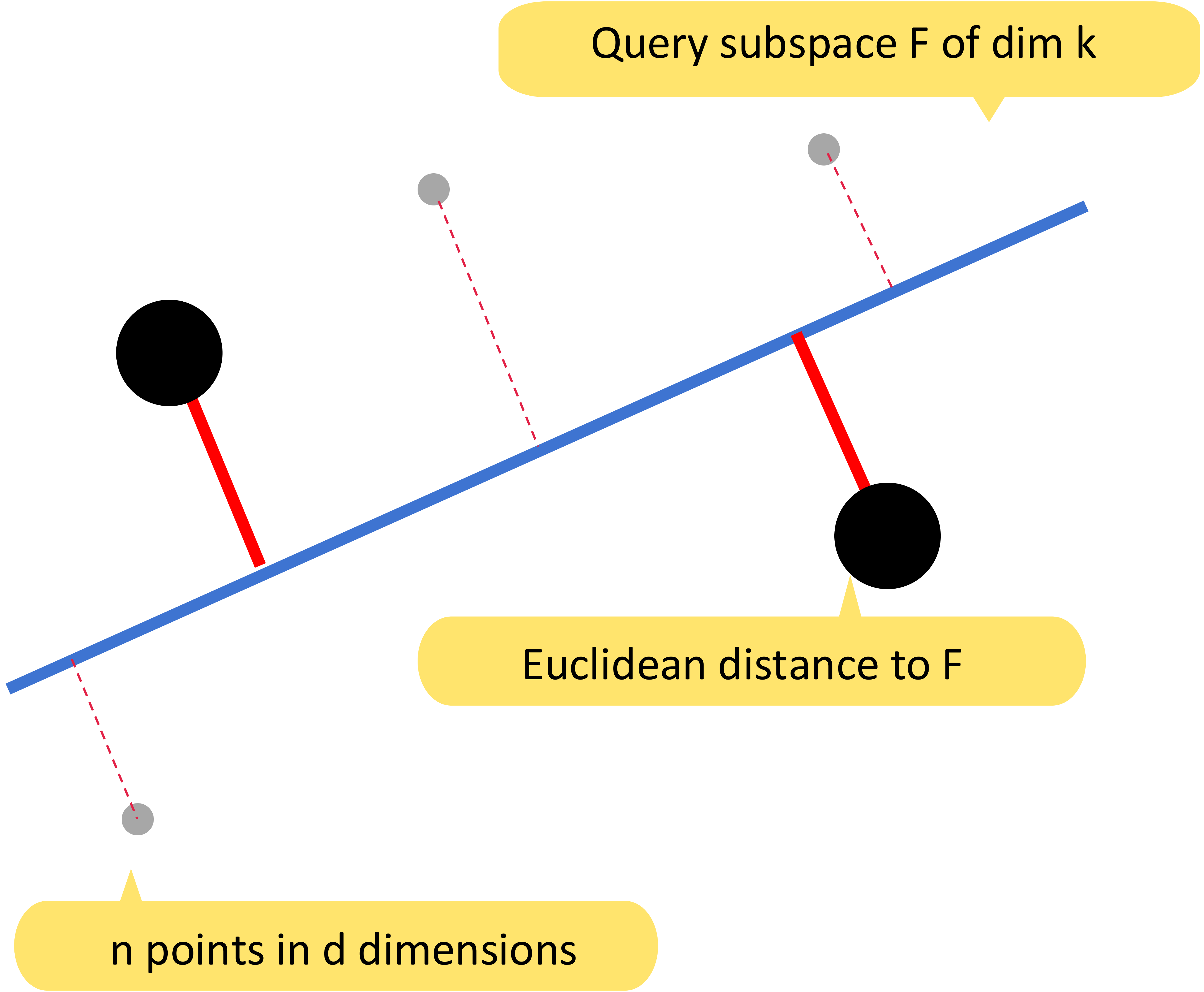
Carnegie Mellon University
Computer Science Department → The Voleon Group

ℓ_p Subspace Approximation



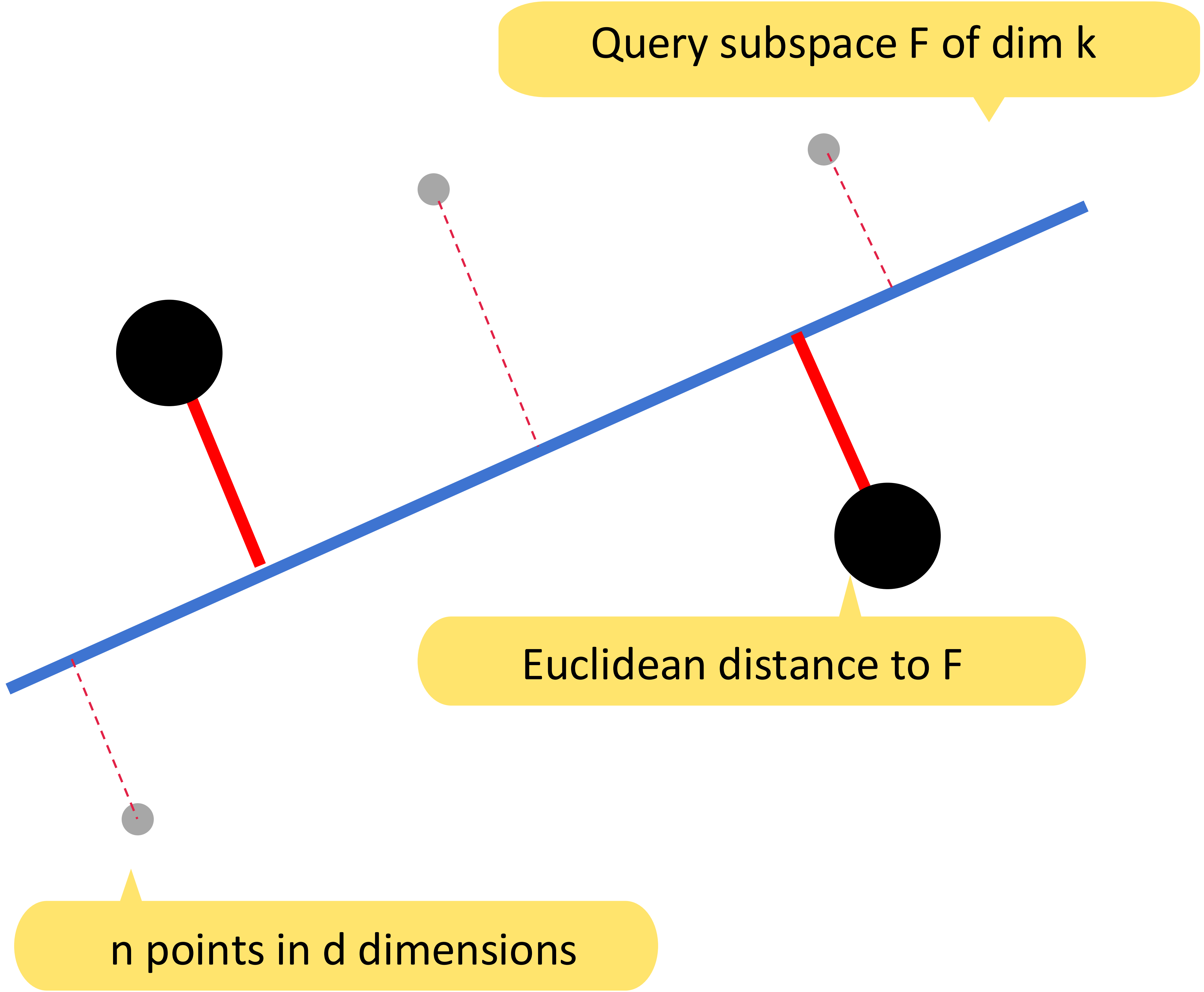
- Cost of a k-dimensional subspace F: ℓ_p norm of the distances to F
- Goal: minimize the cost over subspaces F
- p = 2: PCA
- p = 1
  - "Median hyperplane problem"
  - Rotationally invariant L1 PCA
- p = ∞
  - "Center hyperplane problem"
  - Generalizes extent/containment problems: enclosing sphere, cylinder

Coresets for ℓ_p Subspace Approximation



- Coreset: weighted subset of the points whose cost approximates the cost of the entire set for every subspace F up to $(1+\varepsilon)$ factors
- Prior results:
  - [Feldman-Langberg 2011]
    - Coreset of size $\mathrm{poly}(k, d, \varepsilon^{-1})$
  - [Sohler-Woodruff 2018]
    - Coreset of size $k^{\max\{1,\frac{p}{2}\}}\ \mathrm{poly}(\varepsilon^{-1})$
    - Needs an additional coordinate, exp. time
  - [Huang-Vishnoi 2020]
    - Coreset of size $\mathrm{poly}(k, \varepsilon^{-1})$
    - No additional coordinate, input sparsity time
  - Main question: best of both worlds?

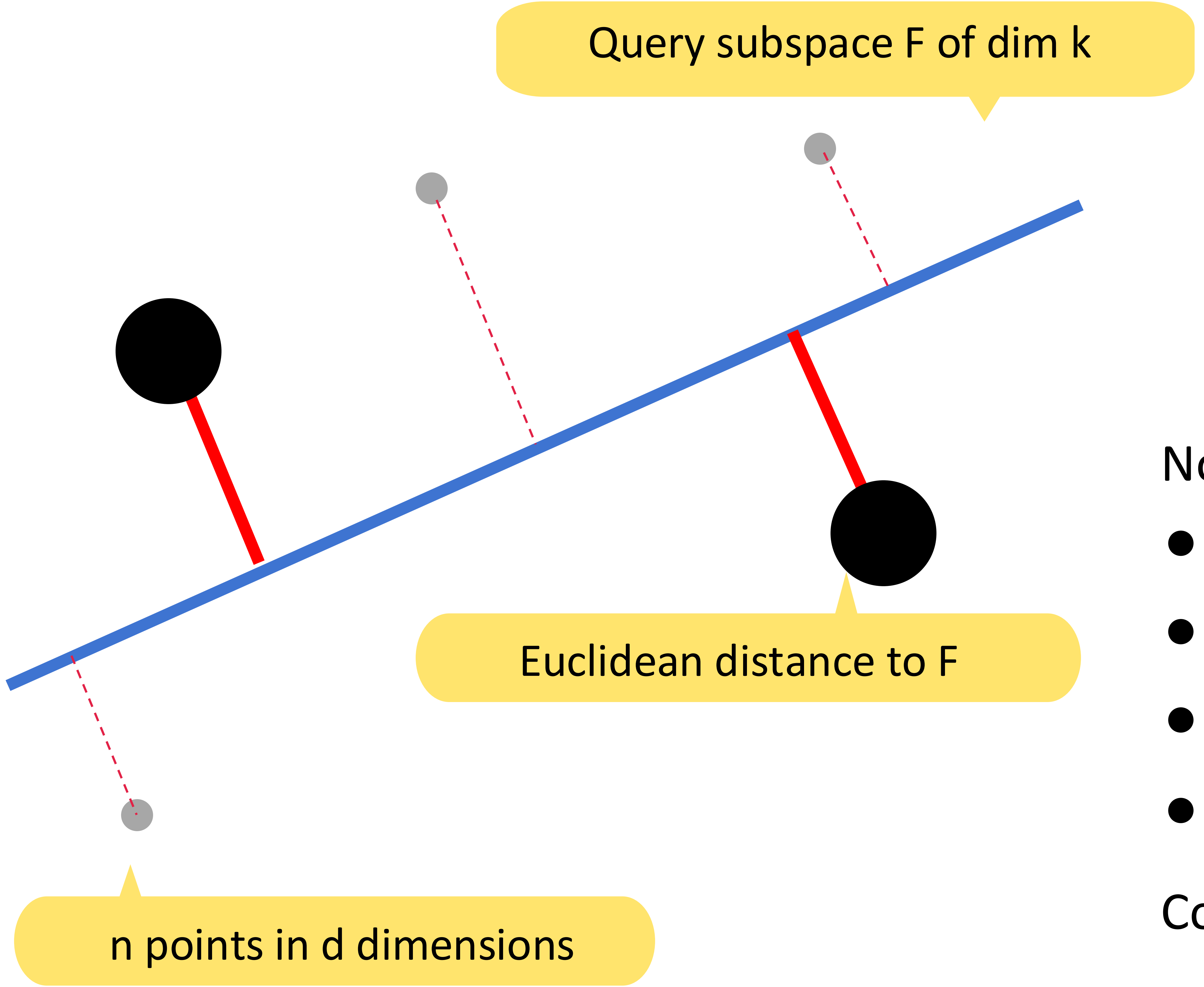
Coresets for ℓ_p Subspace Approximation



Theorem. There is an algorithm which constructs a true coreset of size $k^{\max\{1,\frac{p}{2}\}}\,\mathrm{poly}(\varepsilon^{-1})$ in input sparsity time.

- Previously unknown whether coresets of this size even exist
- Notes on techniques:
  - Sampling uses ridge leverage scores (RLS)
    - Surprising, since RLS seems highly specific to the ℓ_2/Frobenius norm
  - Key ideas to make RLS work for ℓ_p norms: flattening
- Seamlessly handles online/streaming settings

Coresets for ℓ_p Subspace Approximation



Theorem. There is an algorithm which constructs a true coreset of size $k^{\max\{1,\frac{p}{2}\}}\,\mathrm{poly}(\varepsilon^{-1})$ in input sparsity time.

Notation

- A : n x d matrix with the n points in the rows
- S : n x n diagonal matrix of coreset weights
- P_F : projection matrix onto subspace F
- $\|\cdot\|_{p,2}$: (p,2) norm (ℓ_p norm of ℓ_2 norm of rows)

Coreset guarantee: $\|SA(I - P_F)\|_{p,2} = (1 \pm \varepsilon)\|A(I - P_F)\|_{p,2}$

Technical Ingredients

Proof Sketch

- Representative subspace theorem [Sohler-Woodruff 2018]
  - Informally: ℓ_p subspace approximation in d dimensions can be approximated by an instance in $k \cdot \mathrm{poly}(1/\varepsilon)$ dimensions
    - Thus, our task is to preserve an (unknown) subspace of dimension $k \cdot \mathrm{poly}(1/\varepsilon)$ via sampling
- Sampling algorithm: ridge leverage scores [Cohen-Musco-Musco 2017]
  - For any d-dimensional x, we can preserve $\|Ax\|_p$ up to small *additive* error
  - Problem with additive error: loses $\mathrm{poly}(n)$ factors
- Fix for the additive error: two different types of flattening tricks

Representative Subspace Theorem

- [Sohler-Woodruff 2018] Informally: ℓ_p subspace approximation in d dimensions can be approximated by an instance in $s = k \cdot \mathrm{poly}(1/\varepsilon)$ dimensions
  - There exists a subspace S of dimension $s = k \cdot \mathrm{poly}(1/\varepsilon)$ (the representative subspace) s.t. for any query subspace F, the cost can be approximately decomposed into...
    - The cost to project onto S
    - The cost within S
- More formally, for any k-dimensional subspace F, $\|A(I - P_F)\|_{p,2} = (1 \pm \varepsilon)\|[A'(I - P_F), b]\|_{p,2}$
  - Here, $A' = AP_S$ and b is the vector of costs to project onto S
- Thus, our task is to preserve an *unknown* subspace of dimension $s = k \cdot \mathrm{poly}(1/\varepsilon)$ via sampling
- For this talk, pretend like s is just k

Ridge Leverage Scores

- [Cohen-Musco-Musco 2017] Constructing S via weighted sampling by ridge leverage scores gives nearly optimal coresets for $p = 2$
- Score for i-th row: $a_i^\top (A^\top A + \lambda I)^{-1} a_i$, for $\lambda = \frac{1}{k}\|A - A_k\|_F^2$
- Key property: *additive-multiplicative subspace embedding*

$p = 2$

Lemma. If S is constructed by RLS sampling with $\tilde{O}(\frac{k}{\varepsilon^2})$ rows, then for all $x \in \mathbb{R}^d$, we have
$\|SAx\|_2^2 = \|Ax\|_2^2 \pm \varepsilon(\|Ax\|_2^2 + \lambda\|x\|_2^2)$.

$p = 1$

Lemma. If S is constructed by root RLS sampling with $\tilde{O}(\frac{n^{1/2}k^{1/2}}{\varepsilon^2})$ rows, then for all $x \in \mathbb{R}^d$, we have
$\|SAx\|_1 = \|Ax\|_1 \pm \varepsilon(\|Ax\|_1 + \lambda^{1/2}\|x\|_1)$.

- For $p = 2$, the sampled subspace approximation cost pays the additive error s times: once for each dimension in the representative subspace
  - In this case, this is already a proof
- For $p \neq 2$, the additive error is multiplied by a factor of $s^{p/2}$

Problems with the Additive Error: p < 2

$p = 2$

Lemma. If S is constructed by RLS sampling with $\tilde{O}(\frac{k}{\varepsilon^2})$ rows, then for all $x \in \mathbb{R}^d$, we have
$\|SAx\|_2^2 = \|Ax\|_2^2 \pm \varepsilon(\|Ax\|_2^2 + \lambda\|x\|_2^2)$.

$p = 1$

Lemma. If S is constructed by root RLS sampling with $\tilde{O}(\frac{n^{1/2}k^{1/2}}{\varepsilon^2})$ rows, then for all $x \in \mathbb{R}^d$, we have
$\|SAx\|_1 = \|Ax\|_1 \pm \varepsilon(\|Ax\|_1 + \lambda^{1/2}\|x\|_1)$.

- Recall $\lambda = \frac{1}{k}\|A - A_k\|_F^2$ (even for $p \neq 2$!)
- WLOG assume that $d \leq n$
- Additive error for $p = 1$: $\lambda^{1/2}\|x\|_1 \leq \frac{n^{1/2}}{k^{1/2}}\|A - A_k\|_F\,\|x\|_2 \leq \frac{n^{1/2}}{k^{1/2}}\,\mathrm{OPT}\,\|x\|_2$
  - This is off by $n^{1/2}$!
  - The problem: bounding $\|\cdot\|_F \leq \|\cdot\|_{1,2}$ is loose

Problems with the Additive Error: p < 2

- The problem: bounding $\|\cdot\|_F \leq \|\cdot\|_{1,2}$ is loose
- The solution for $p < 2$: *flattening*
  - For a vector y, if we replace an entry y_i with t copies of $y_i/t^{1/p}$, the ℓ_p norm is preserved
  - We can flatten any y to a new vector y' so that
    - y' has length at most $2n$
    - $\|y'\|_p = \|y\|_p$
    - $\|y'\|_2 \leq O(n^{\frac{1}{2}-\frac{1}{p}})\|y\|_p$: this recovers the lost factor we need!
- For subspace approximation: efficiently compute a bicriteria approximation solution P' of rank $k' = O(k)$
  - Flatten A using the cost of the rows of $A(I - P')$, say B
  - Then, $\|B - B_{k'}\|_F \leq \|B(I - P')\|_F \leq O(n^{\frac{1}{2}-\frac{1}{p}})\|B(I - P')\|_{p,2} \leq O(n^{\frac{1}{2}-\frac{1}{p}})\mathrm{OPT}$
  - Additive error is small enough and completes the proof sketch

Problems with the Additive Error: p > 2

- For $p > 2$, we have a similar problem, and the previous idea does not work
- We will find a different source of flattening in the ridge leverage scores (RLS) lemma
  - RLS is just leverage score sampling on a concatenated matrix $[A; \lambda^{1/2} I]$
    - Additive error is the ℓ_p norm of $\left[A; \lambda^{1/2} I\right]x \Rightarrow$ additive error $\lambda^{p/2}\|x\|_p^p \leq n^{\frac{p}{2}-1}\lambda^{p/2}\|x\|_2^p$
  - In fact, the same proof applies if we replace I by any orthonormal matrix U
  - Idea: choose U to be random orthonormal matrix
    - Additive error is the ℓ_p norm of $\left[A; \lambda^{1/2} U\right]x \Rightarrow$ additive error $\lambda^{p/2}\|Ux\|_p^p \leq O(\lambda^{p/2})\|x\|_2^p$
      - Note: only for x in a small-dim space, which is all we need
- Additive error is small enough and completes the proof sketch

Conclusion

- We resolve the dependence of k in the coreset size for ℓ_p subspace approximation
- Techniques use a combination of ridge leverage scores and novel use of flattening
- Our techniques seamlessly handle online/streaming settings
- Open directions
  - Main question: resolving the dependence on ε
    - Currently, the exponent on ε is p^2
    - Conjecture: ε^2