Data-Driven Linear Complexity Low-Rank Approximation of General Kernel Matrices

Yuanzhe Xi
Department of Mathematics
Emory University

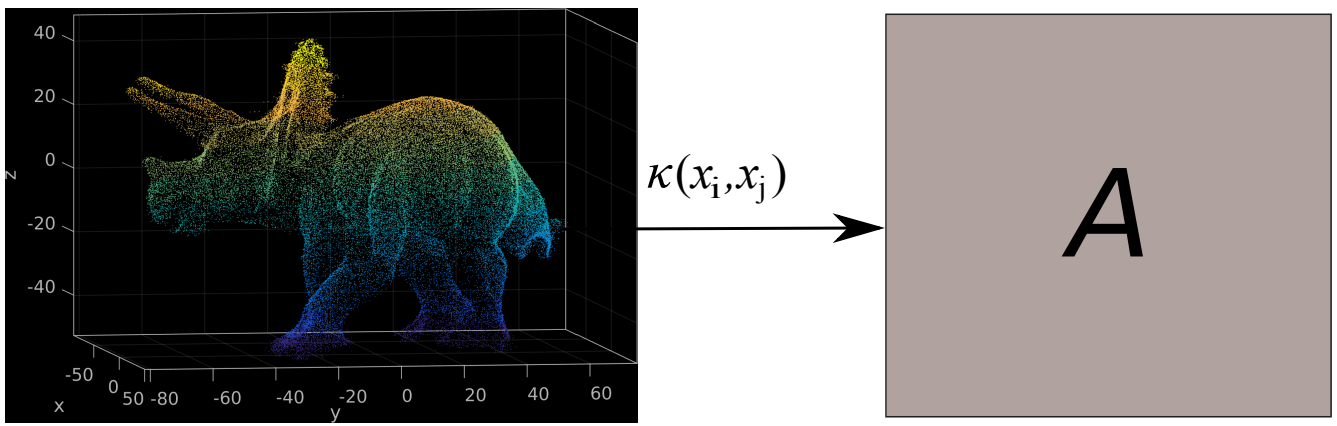
Joint work with Difeng Cai (Emory), Edmond Chow (Georgia Tech)

Fast Direct Solver 21, Purdue University
October 23, 2021
Supported by NSF OAC 2003720

Dense kernel matrices

- X, Y are data sets in $\mathbb{R}^d$

$$A = [\kappa(x,y)]_{x\in X, y\in Y} = \begin{bmatrix} \kappa(x_1,y_1) & \kappa(x_1,y_2) & \dots & \kappa(x_1,y_n) \\ \kappa(x_2,y_1) & \kappa(x_2,y_2) & \dots & \kappa(x_2,y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \kappa(x_m,y_1) & \kappa(x_m,y_2) & \dots & \kappa(x_m,y_n) \end{bmatrix}$$



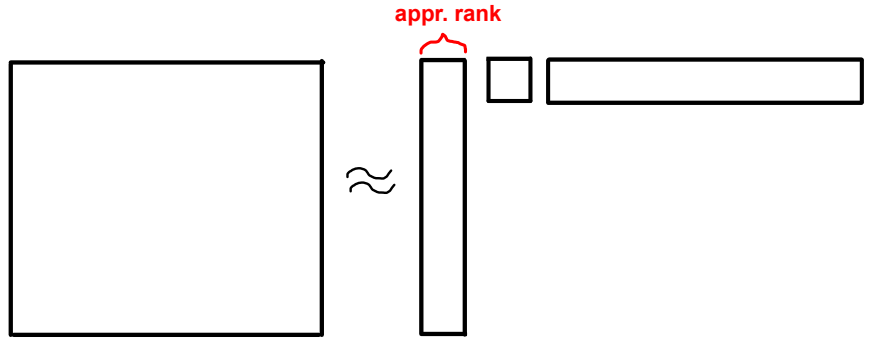
http://www.geo.tuwien.ac.at/downloads/pg/pctools

- Canonical kernels $\kappa(x,y)$:

$$\sqrt{1+\tfrac{1}{10}|x-y|^2},\ \frac{1}{|x-y|},\ \frac{e^{i|x-y|}}{|x-y|},\ e^{-|x-y|/\sigma}, e^{-|x-y|^2/\sigma^2}, \dots$$

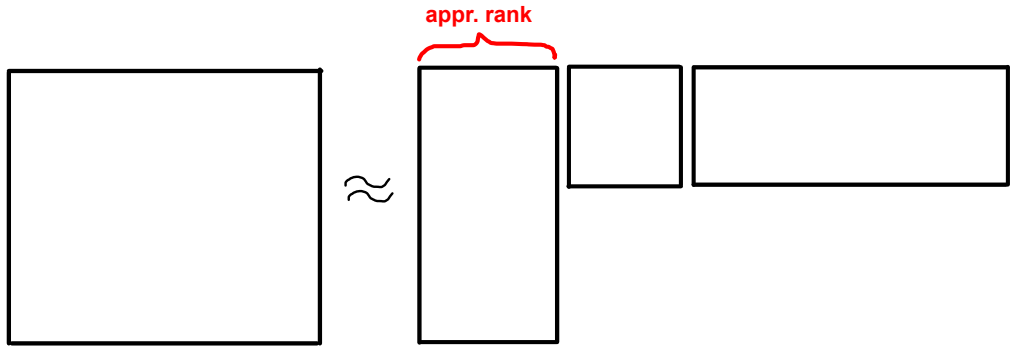
Low-rank approximation techniques

1 Algebraic methods (SVD, RRQR): small rank, high cost



2 Analytic methods: Taylor expansion, interpolation [Hackbush, Borm, Le Borne], spherical harmonic [Greengard, Rokhlin]

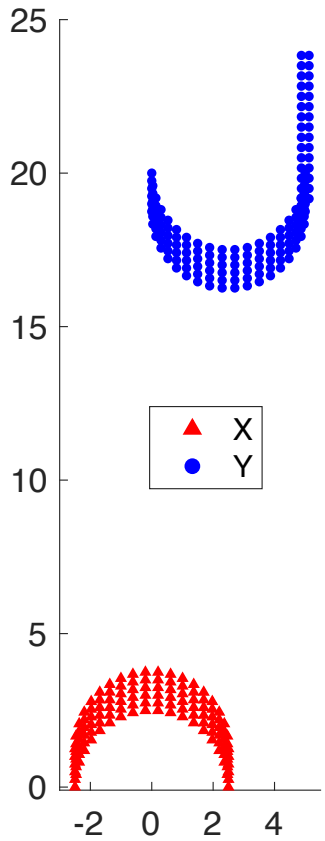
$$\kappa(x,y) \approx \sum_{k,l=0}^{r-1} c_{k,l}\phi_k(x)\psi_l(y), \quad \forall (x,y) \in X \times Y$$



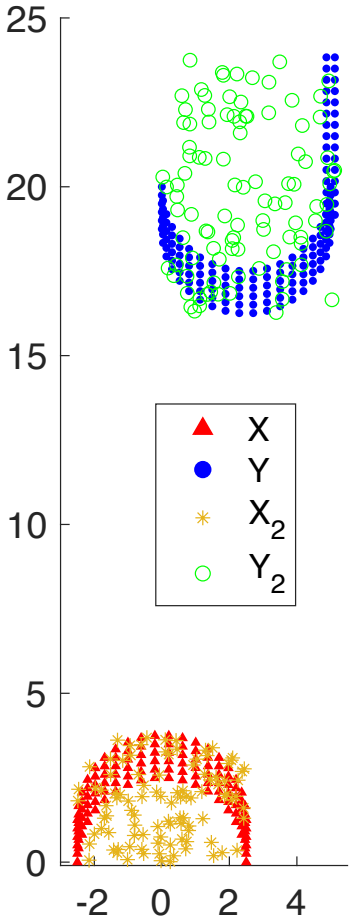
Hybrid compression techniques

3 Hybrid compression techniques: proxy point method, proxy surface method [Ying, Biros, Martinsson, Gillman, Darve, Xing, Chow, et.al.]

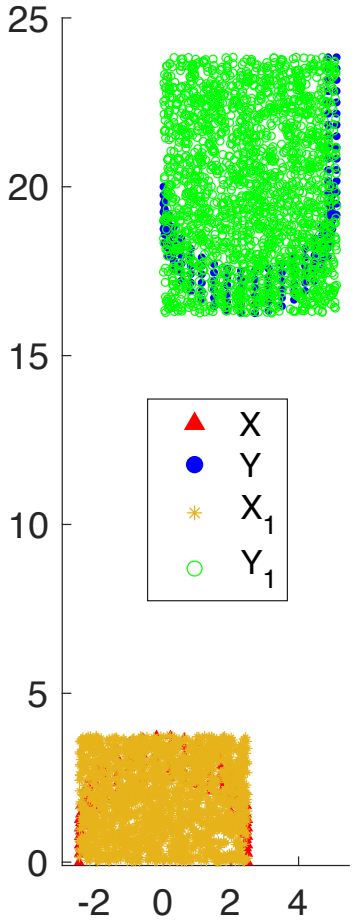
Compressing small kernel matrices from points sampled on an extended continuous domain $\Omega_X \times \Omega_Y$



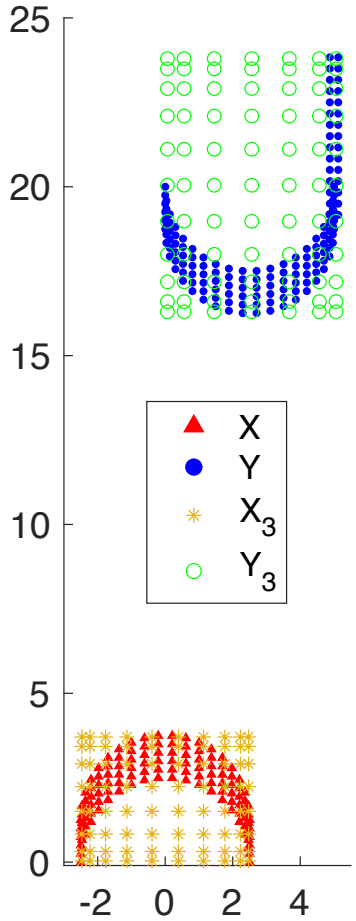
Three types of samples



(b) 100 rand pts



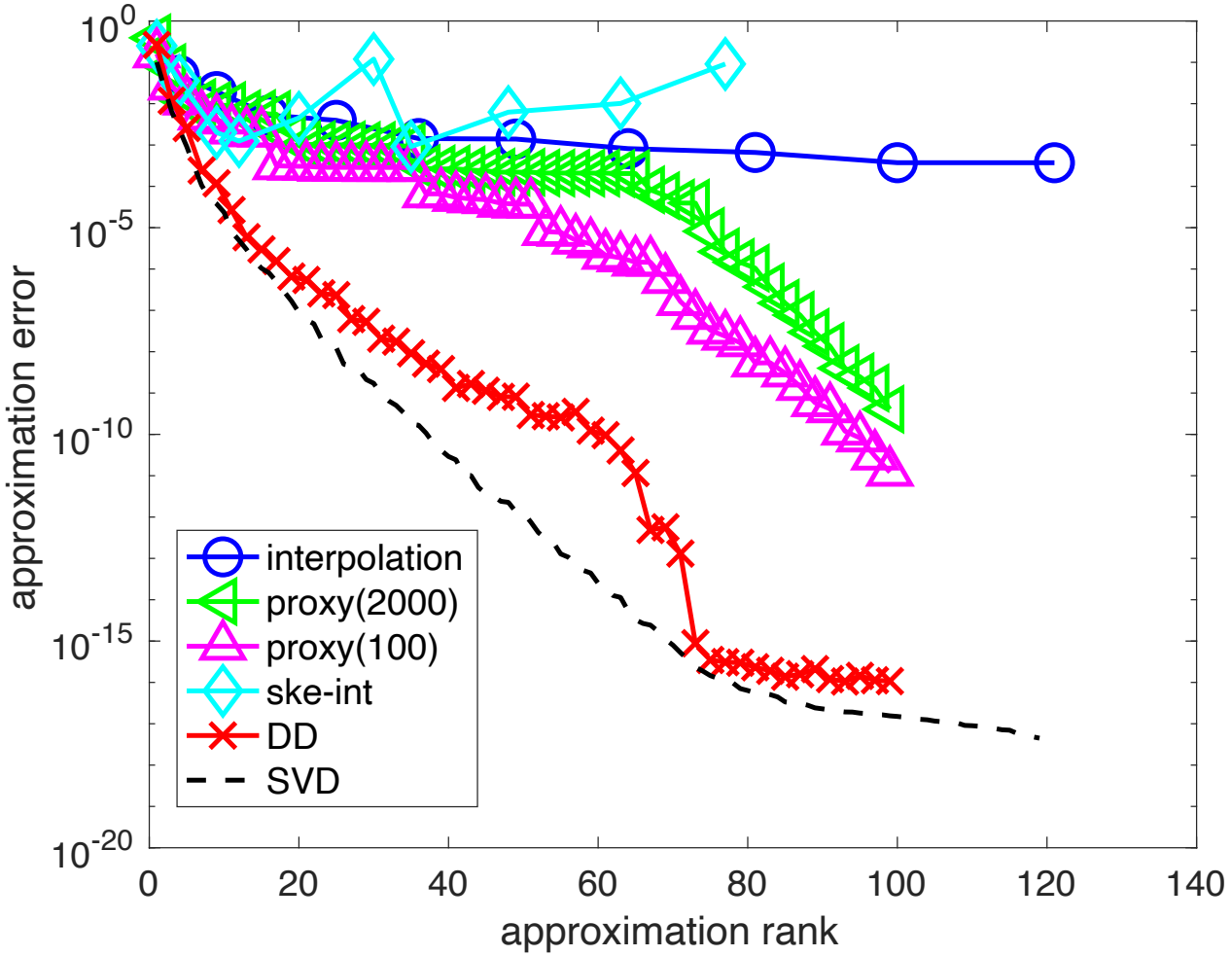
(c) 2000 rand pts (proxy point, Xing, Chow, 2020)



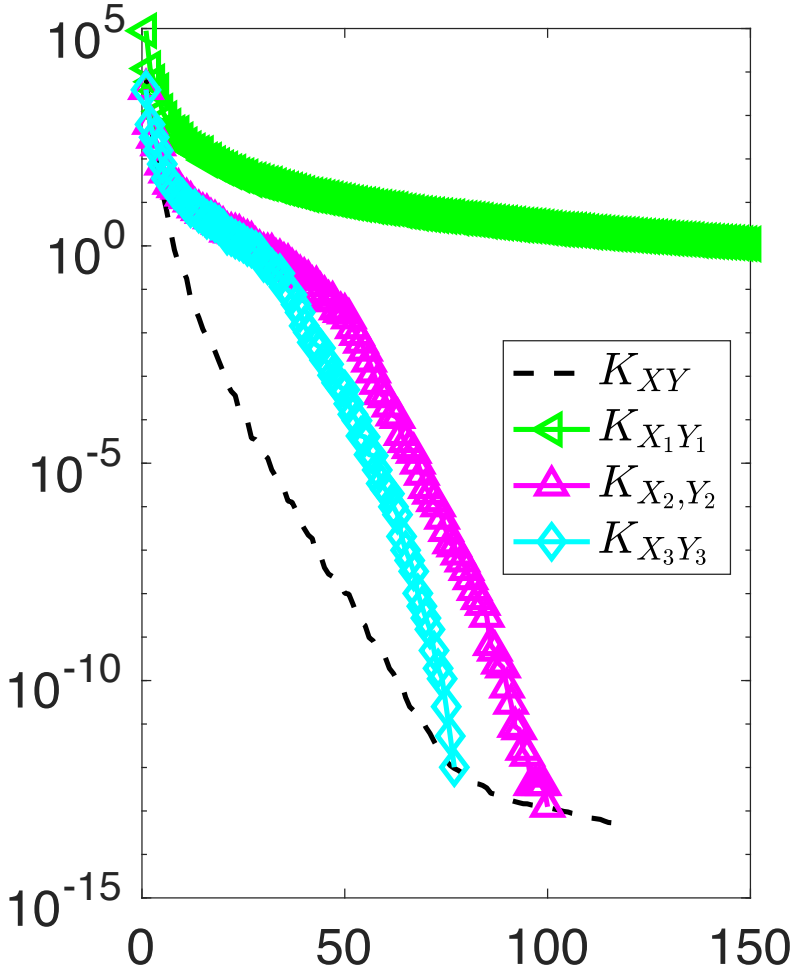
(d) 10×10 Cheby pts (Skeletonized interpolation, Darve, 2020)

Approximation error

$$\kappa(x,y) = \sqrt{1 + 100|x - y + a|^2} \quad with\ a=[0,20]^T$$



Negative effect of ignoring data geometry



Singular values of K_{XY} (120-by-150), $K_{X_1Y_1}$ (2000-by-2000), $K_{X_2Y_2}$ (100-by-100), and $K_{X_3Y_3}$ (100-by-100)

Subset selection-based methods

Given $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^d$ and κ, find a subset $S_X \in X$ and a subset $S_Y \in Y$ such that

$$K_{XY} \approx K_{XS_Y} K^{+}_{S_X S_Y} K_{S_Y Y}$$

where $K_{XS} = [\kappa(x, y)]_{x \in K, y \in S}$.

Similar formats to some existing factorization:

- Randomized CUR/Nyström factorization via uniform, leverage score and k-means based sampling [M. W. Mahoney; P. Drineas, Musco]
- Nearest neighbor-based importance sampling [G. Biros]
- Pseudoskeleton [E. E.Tyrtyshnikov, D. Kressner]
- Interpolative decomposition [P.G. Martinsson]
- DEIM [D. Sorensen, M. Embree]

Geometrical understanding of S_X and S_Y on approx. accuracy

Error estimate for two-sided scheme

Theorem (Cai, Chow, XY, 2021)

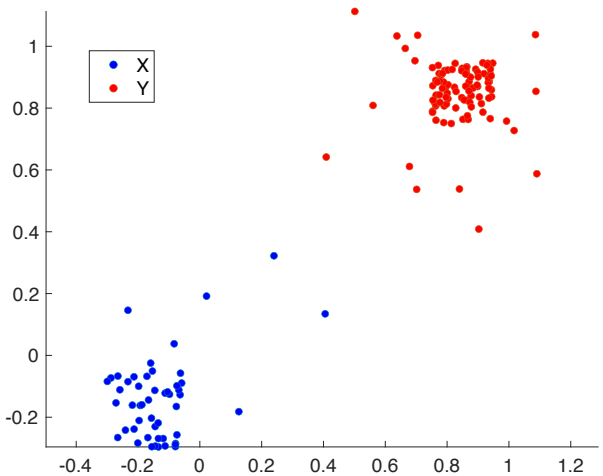
Given X and Y, $\kappa(x,y) \in C(\mathbb{R}^d \times \mathbb{R}^d)$ is Lipschitz continuous with Lipschitz constant L, then

$$\|K_{XY} - K_{XS_Y} K^{+}_{S_X S_Y} K_{S_X Y}\|_{\max} \leq \sqrt{2} L \max\{\delta_{X,S_X}, \delta_{Y,S_Y}\} + \sqrt{r_2} L \delta_{X,S_X} + \sqrt{r_1} L \delta_{Y,S_Y} + \|K^{+}_{S_1 S_2}\| \sqrt{r_1 r_2} L^2 \delta_{X,S_X} \delta_{Y,S_Y}$$

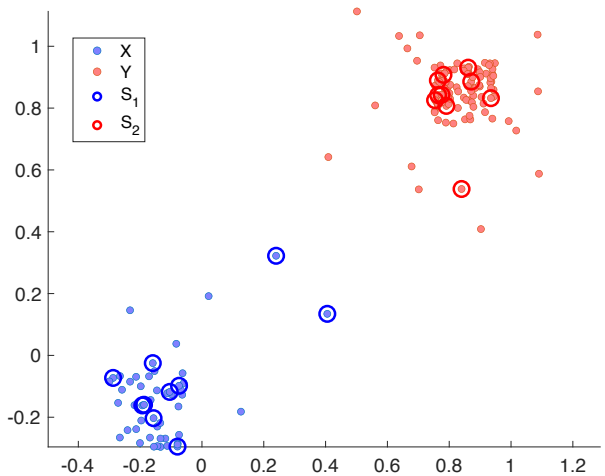
where $\delta_{X,S_X} = \max_{x \in X} dist(x, S_X)$ *and* $\delta_{Y,S_Y} = \max_{y \in Y} dist(y, S_Y)$.

Subset quality indicators

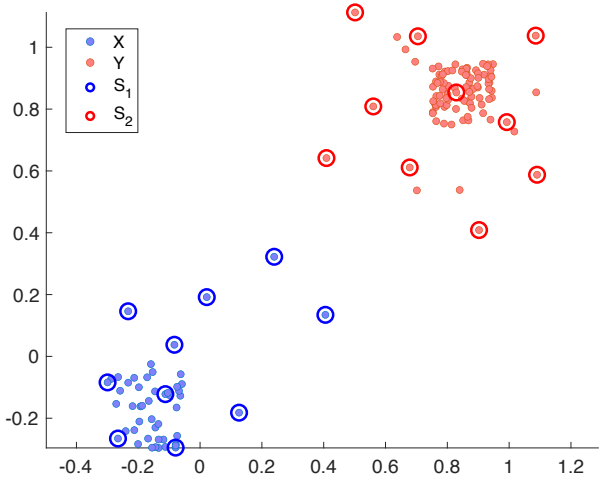
indicator 3 = δ_{X,S_1}, indicator 4 = δ_{Y,S_2}, indicator 5 = $\|K^{+}_{S_1 S_2}\|$



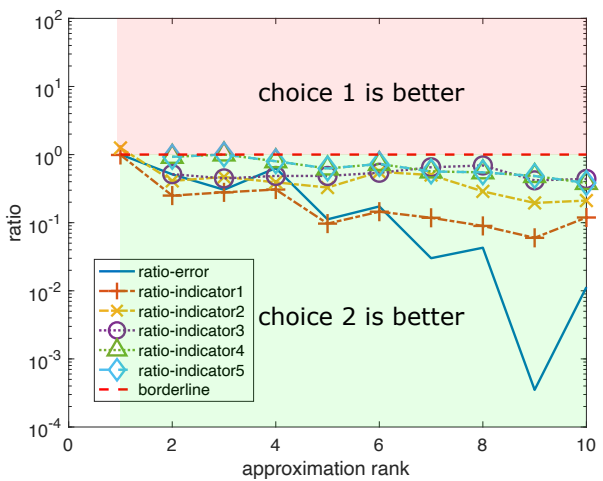
(a) Datasets X, Y



(b) Choice 1



(c) Choice 2



One-sided scheme

Given $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^d$ and κ,

1. Select a subset $S_Y \in Y$
2. Perform interpolative decomposition [P.G. Martinsson] to obtain
$$K_{XS_Y} = UK_{\mathcal{I}S_Y}$$
3. Form $K_{\mathcal{I}Y}$ such that
$$K_{XY} \approx UK_{\mathcal{I}Y}$$

Error estimate for one-sided scheme

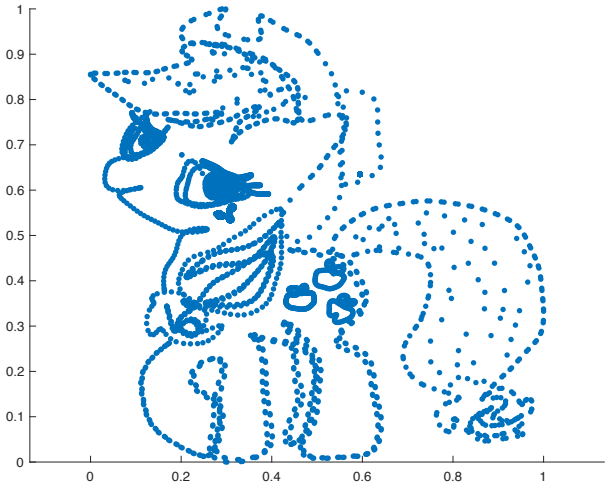
Theorem

Given X and Y, $\kappa(x, y) \in C(\mathbb{R}^d \times \mathbb{R}^d)$ is Lipschitz continuous with Lipschitz constant L, then

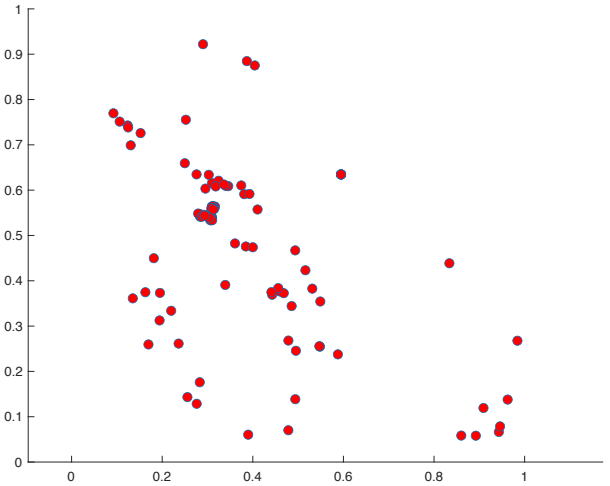
$$\begin{aligned}\|K_{XY} - UK_{\mathcal{I}Y}\|_{\max} \leq \sqrt{2}L\delta_{Y,\mathrm{S}} + (1+2r)\sqrt{m}L\delta_{Y,\mathrm{S}} \\ + 2r\sqrt{2}L\max\{\delta_{X,\mathcal{I}}, \delta_{Y,\mathrm{S}}\},\end{aligned}$$

where $m = card(X), r = card(\mathcal{I})$.

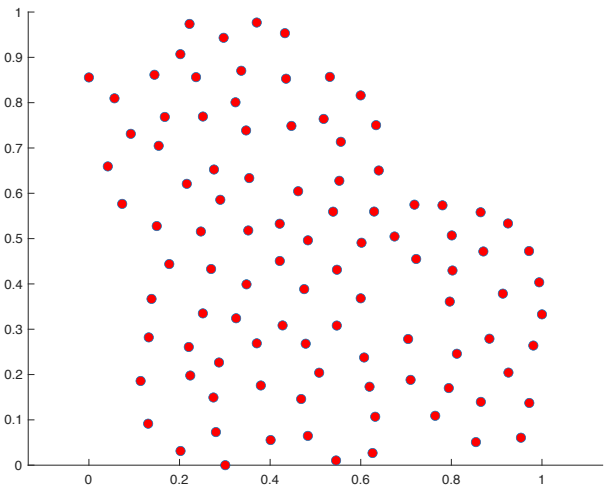
Linear complexity subset selection schemes



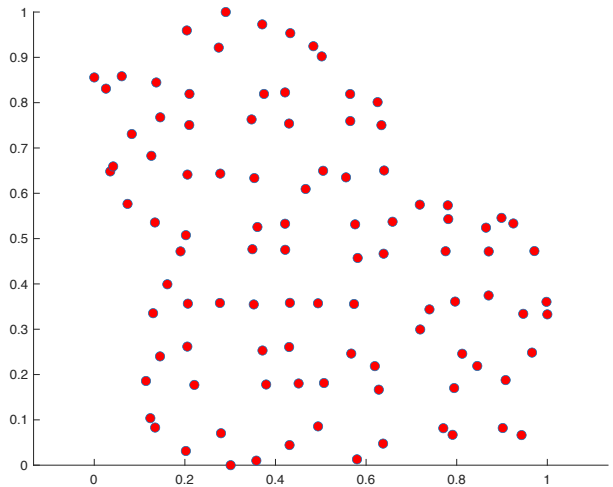
(e) Pony dataset



(f) Uniform sampling



(g) Farthest point sampling



(h) Anchor net method (Cai, Nagy, Xi, 2021)

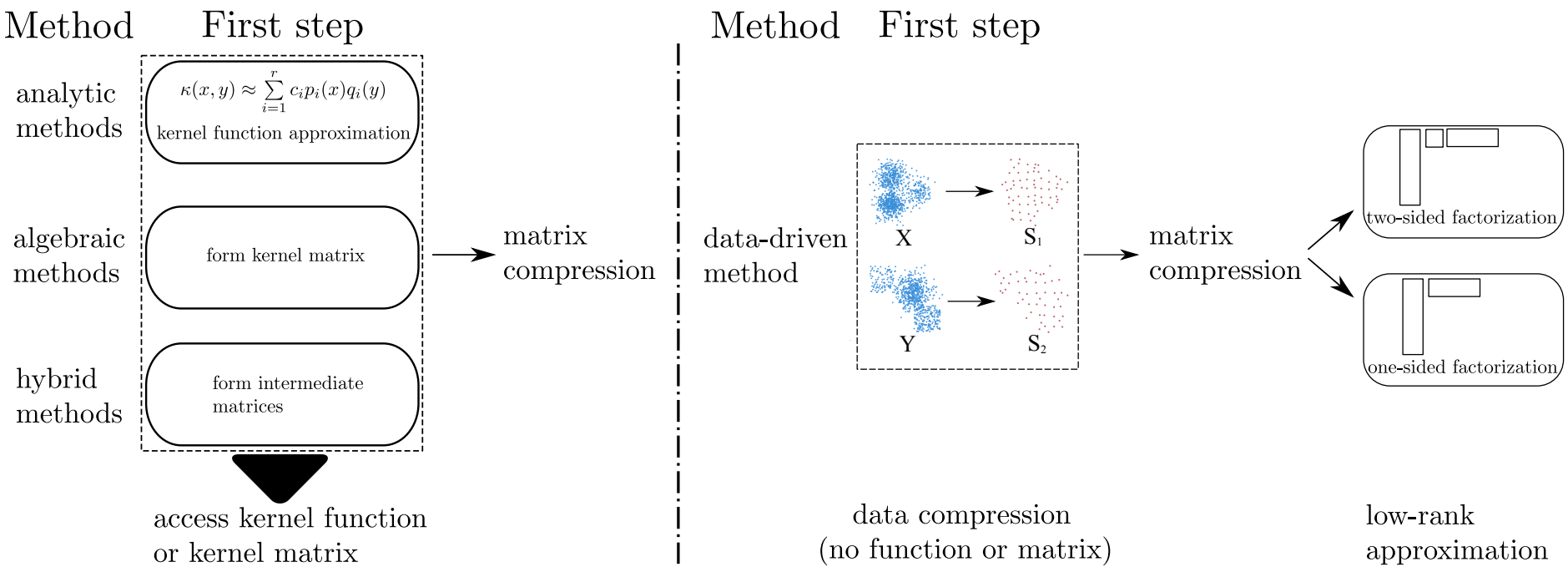
Complexity analysis

Select r points from n points in $\mathbb{R}^d$

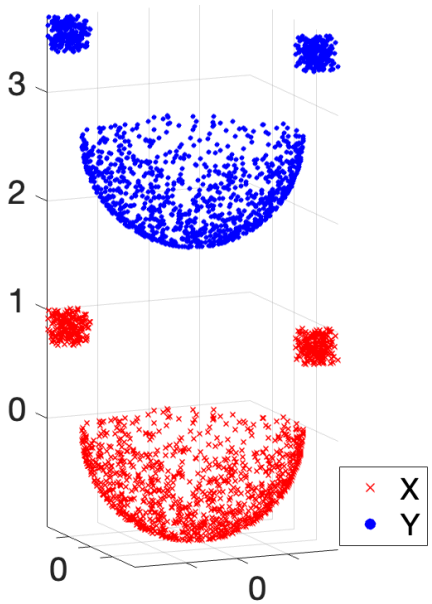
1. Farthest point sampling (FPS): $O(dr^2n)$
2. Anchor net sampling (ANC): $O(drn)$

Computational complexity for data-driven methods

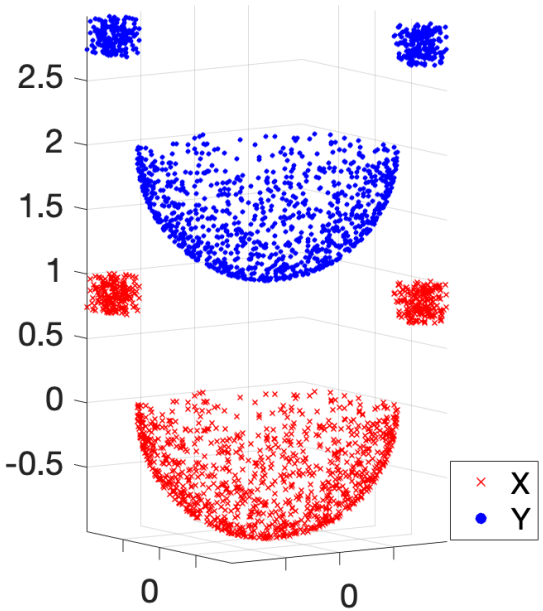
1. Two-sided scheme: $O(dr(m+n))$ (ANC), or $O(dr^2(m+n))$(FPS)
2. One-sided scheme: $O(dr^2(m+n))$


Method
First step
analytic methods
$\kappa(x,y) \approx \sum_{i=1}^{r} c_i p_i(x) q_i(y)$
kernel function approximation
algebraic methods
form kernel matrix
hybrid methods
form intermediate matrices
matrix compression
access kernel function or kernel matrix
Method
First step
data-driven method
X
S_1
Y
S_2
matrix compression
two-sided factorization
one-sided factorization
data compression (no function or matrix)
low-rank approximation

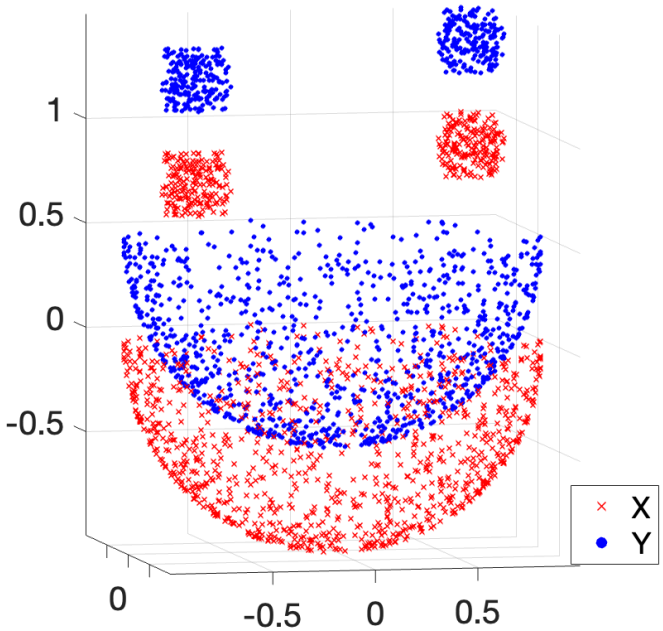
Datasets in three dimensions



(j) Dataset 1



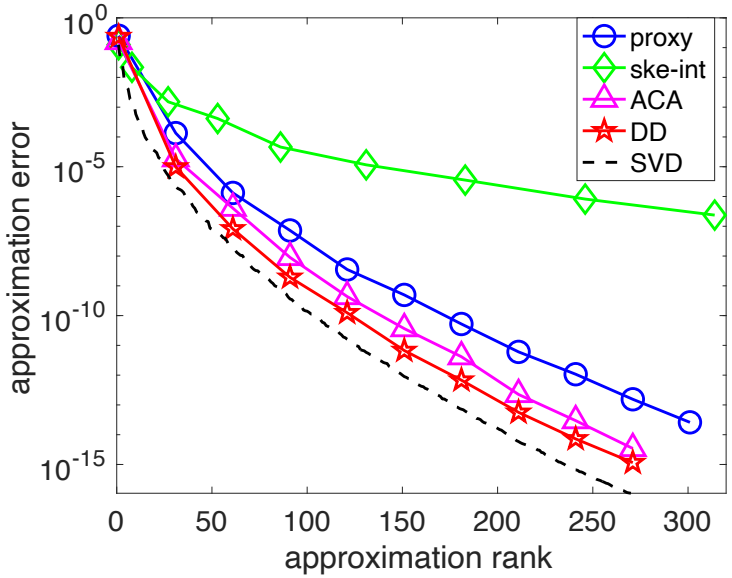
(k) Dataset 2



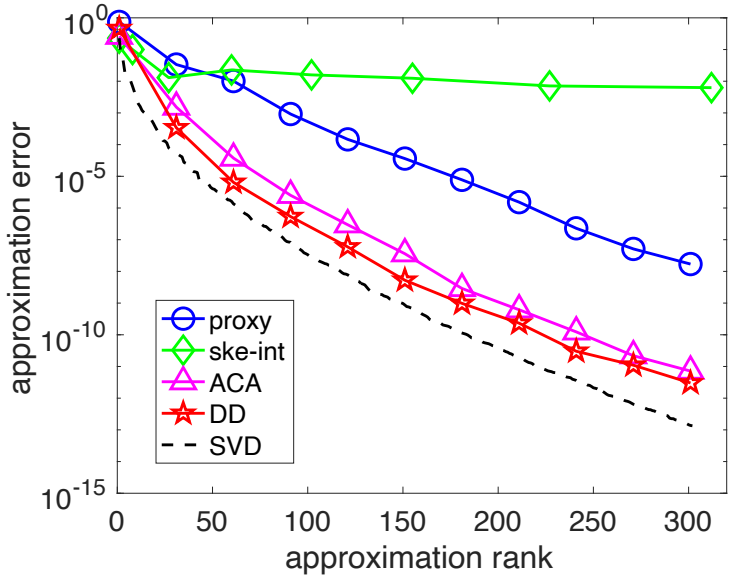
(l) Dataset 3

Y: vertical shift of X by 2.7, 2.0, 0.5 with dist(X, Y) = 1.00, 0.43, 0.12

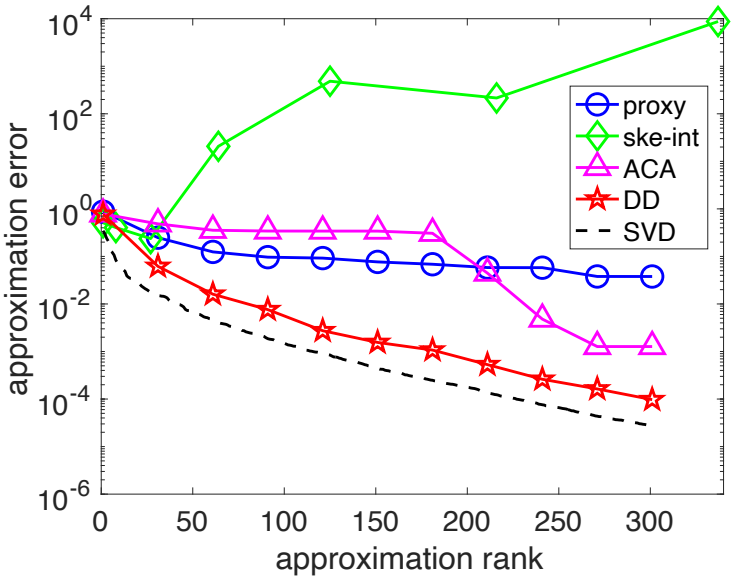
KERNEL FUNCTION $\kappa(x, y) = 1/|x - y|$



(a) Dataset 1

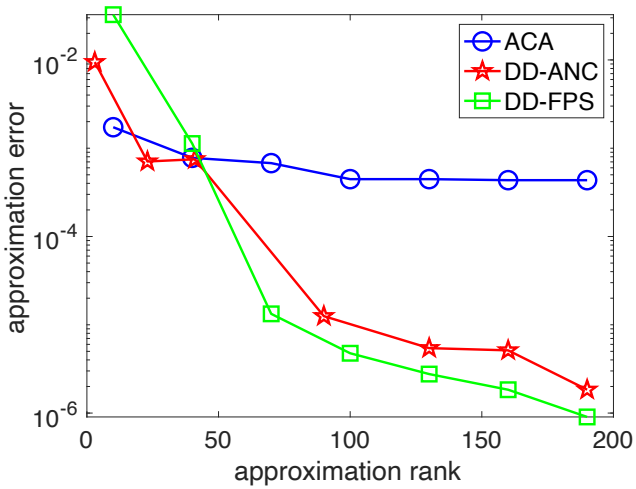


(b) Dataset 2

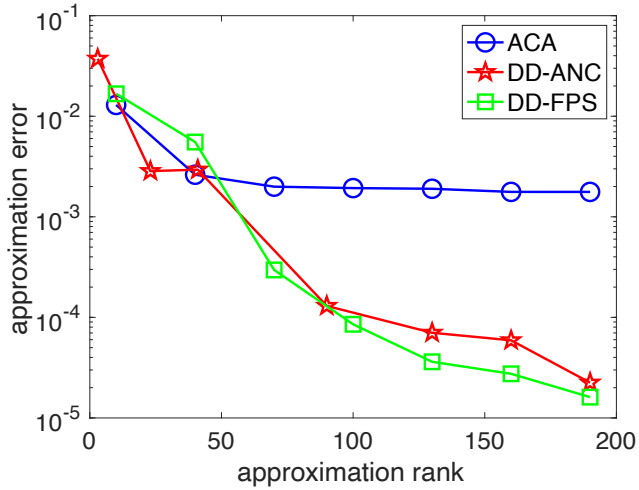


COVERTYPE ($d = 54$)[1]

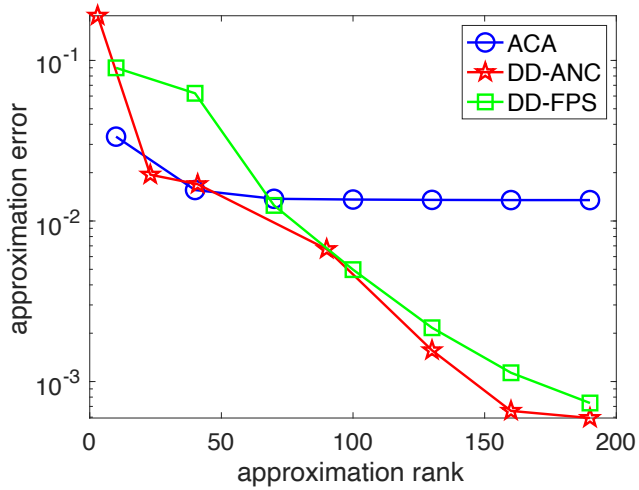
Consider Gaussian kernel $\kappa(x, y) = \exp(-|x - y|^2/\sigma^2)$



(a) $\sigma = \sigma_1 \approx 74.46$

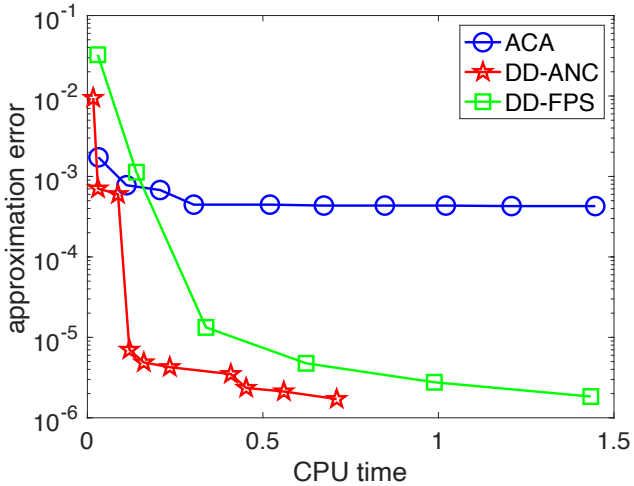


(b) $\sigma = \sigma_2 \approx 37.23$

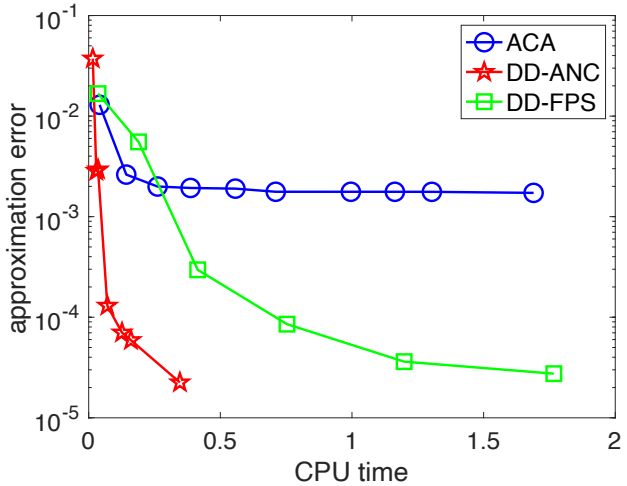


(c) $\sigma = \sigma_3 \approx 14.89$

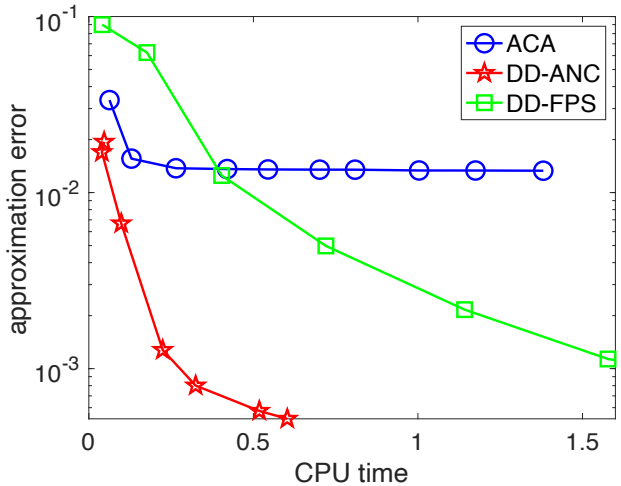
COVERTYPE ($d = 54$)[2]



(d) $\sigma = \sigma_1 \approx 74.46$



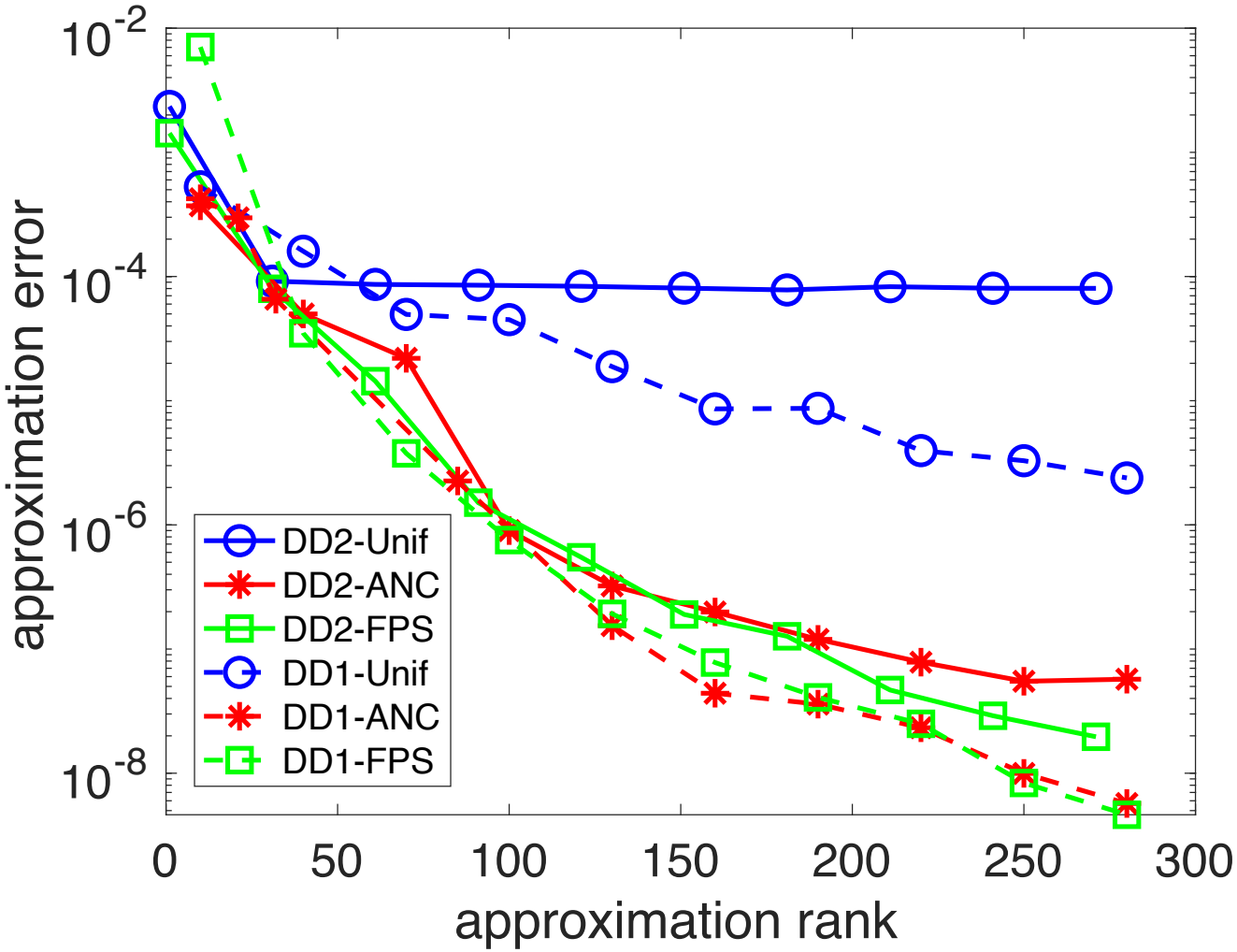
(e) $\sigma = \sigma_2 \approx 37.23$



(f) $\sigma = \sigma_3 \approx 14.89$

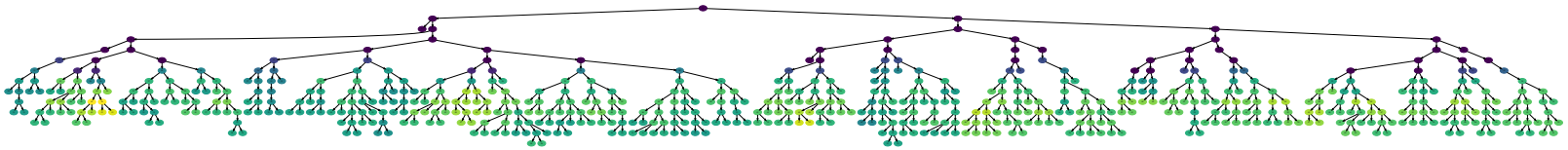
[2]UC Irvine Machine Learning Repository

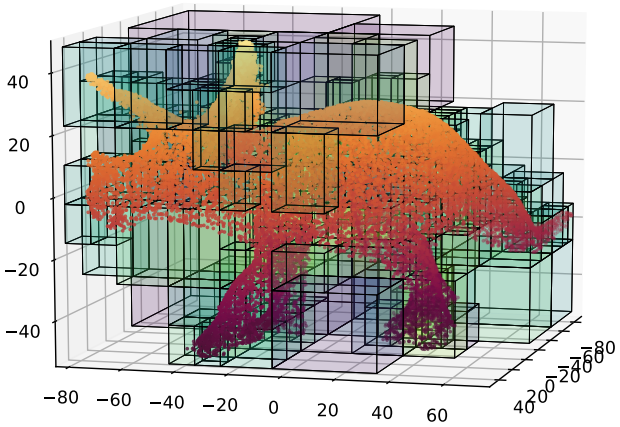
Two-sided V.S. one-sided

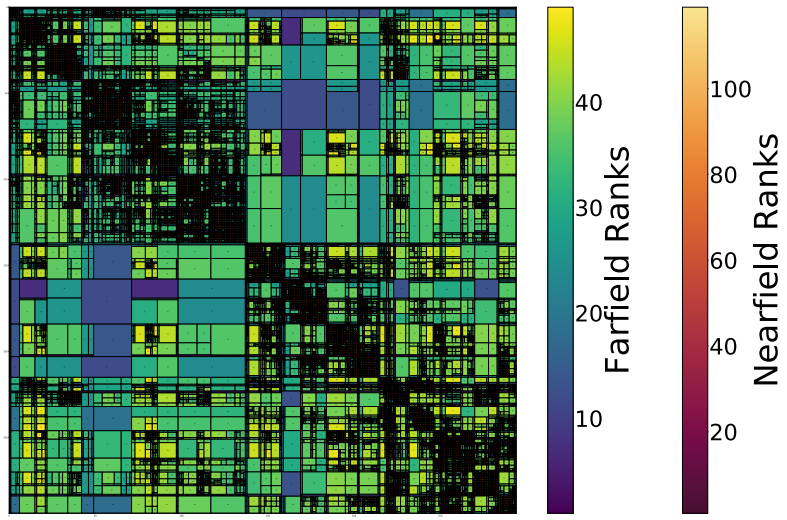


Hierarchical matrices

When kernel has singularity, low-rank properties are associated with certain off-diagonal blocks

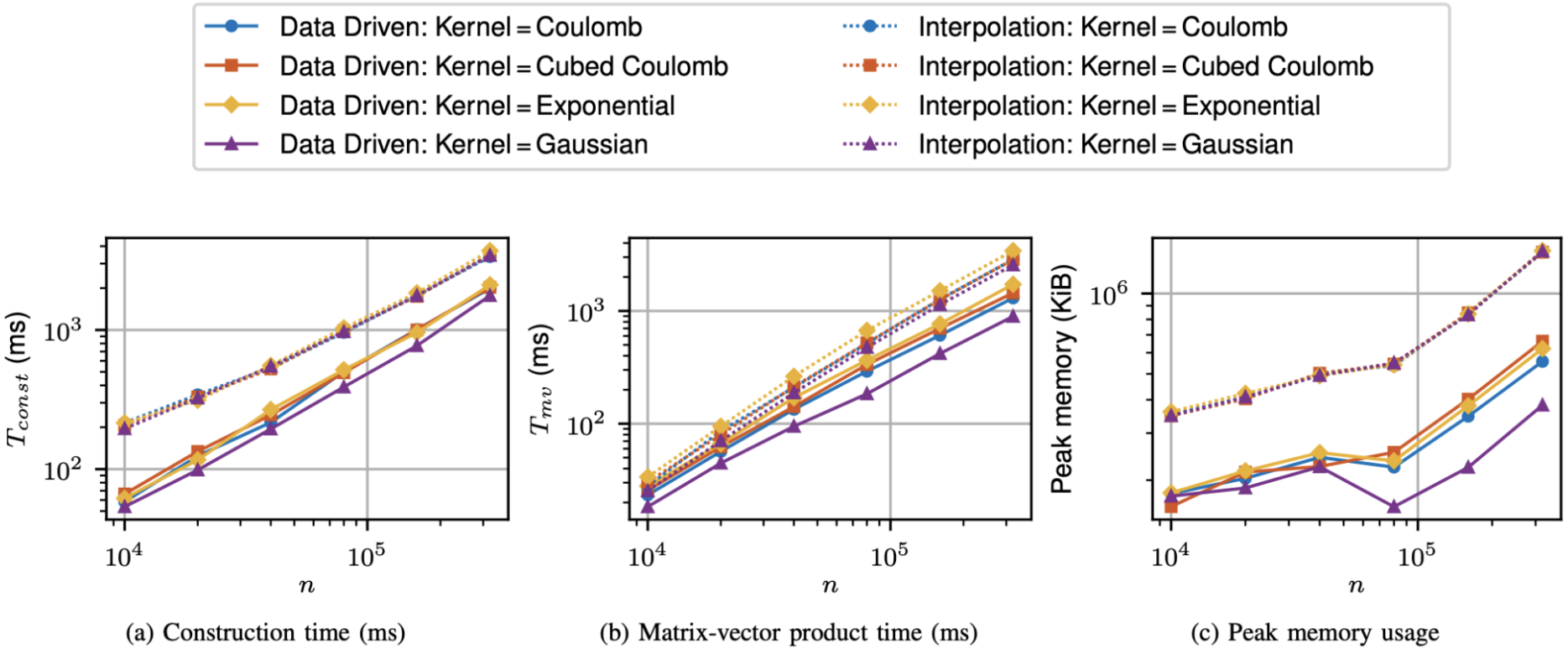






Examples: Fast Multipole Method (Rokhlin 1985, Greengard, Rokhlin 1987), $\mathcal{H}^2$ matrix (Hackbusch et al. 2001-)

Data-driven vs Interpolation



(a) Construction time (ms)

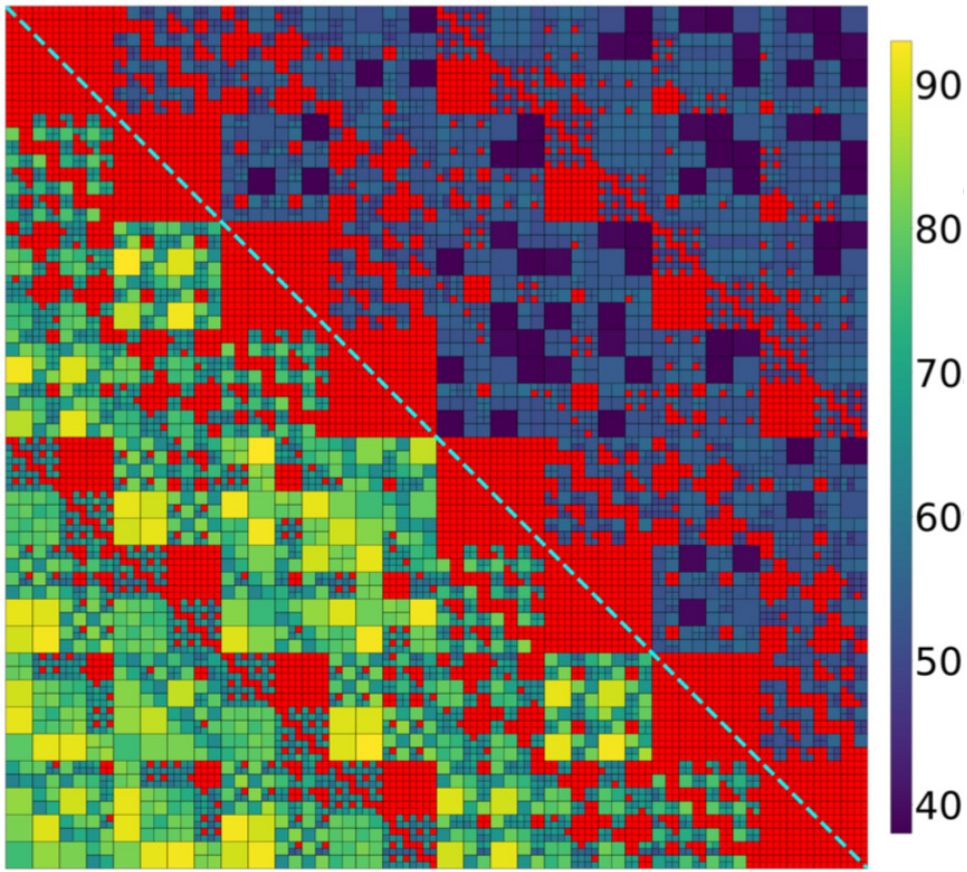
(b) Matrix-vector product time (ms)

(c) Peak memory usage

DD (solid line) vs Interpolation (dotted line)

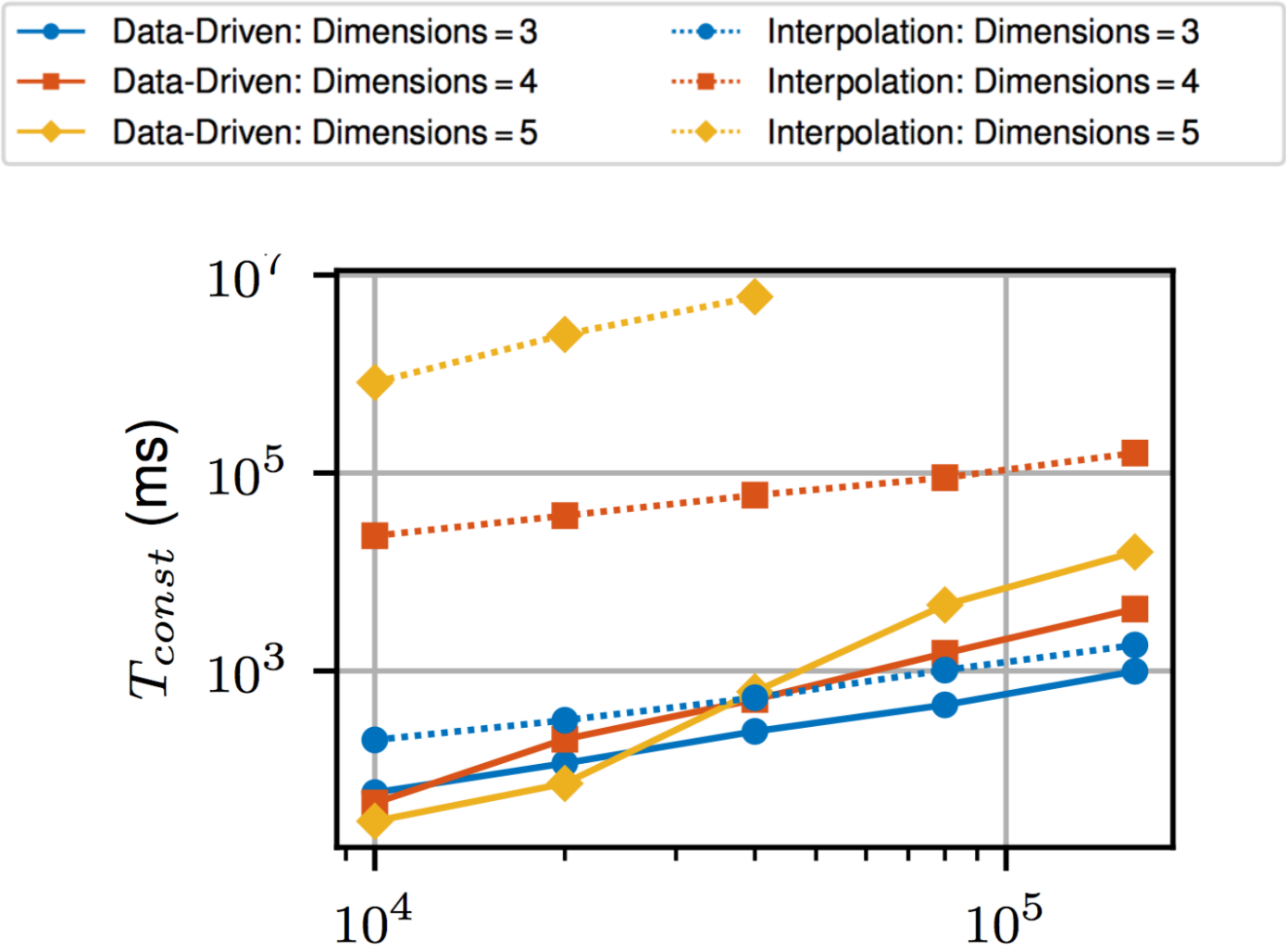
Data-driven vs Interpolation

Coulomb potential, 3D cube, fixed tolerance: $1E-7$



Rank comparison. Lower triangular: interpolation; upper triangular: DD

Data-driven sampling vs Interpolation



Matvec in high dimensions

Conclusions and future work

Data-driven hierarchical matrix methods

1. Data-driven sampling
2. General kernel functions
3. Efficient hierarchical matrix construction

Future work

1. Efficient preconditioning techniques
2. Tensor kernel functions
3. High dimensional problems

References

1. D. Cai, L. Erlandson, E. Chow and YX, Accelerating Parallel Hierarchical Matrix-Vector Products via Data-Driven Sampling, 34th IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2020
2. D. Cai, J. Nagy, and YX, Fast determinstic approximation of general symmetric kernel matrices in high dimensions, 2021, arXiv:2102.05215
3. D. Cai, E. Chow and YX, Data-driven linear complexity low-rank approximation of general kernel matrices: A Geometric Approach, submitted, 2021

Low discrepancy sets

Definition

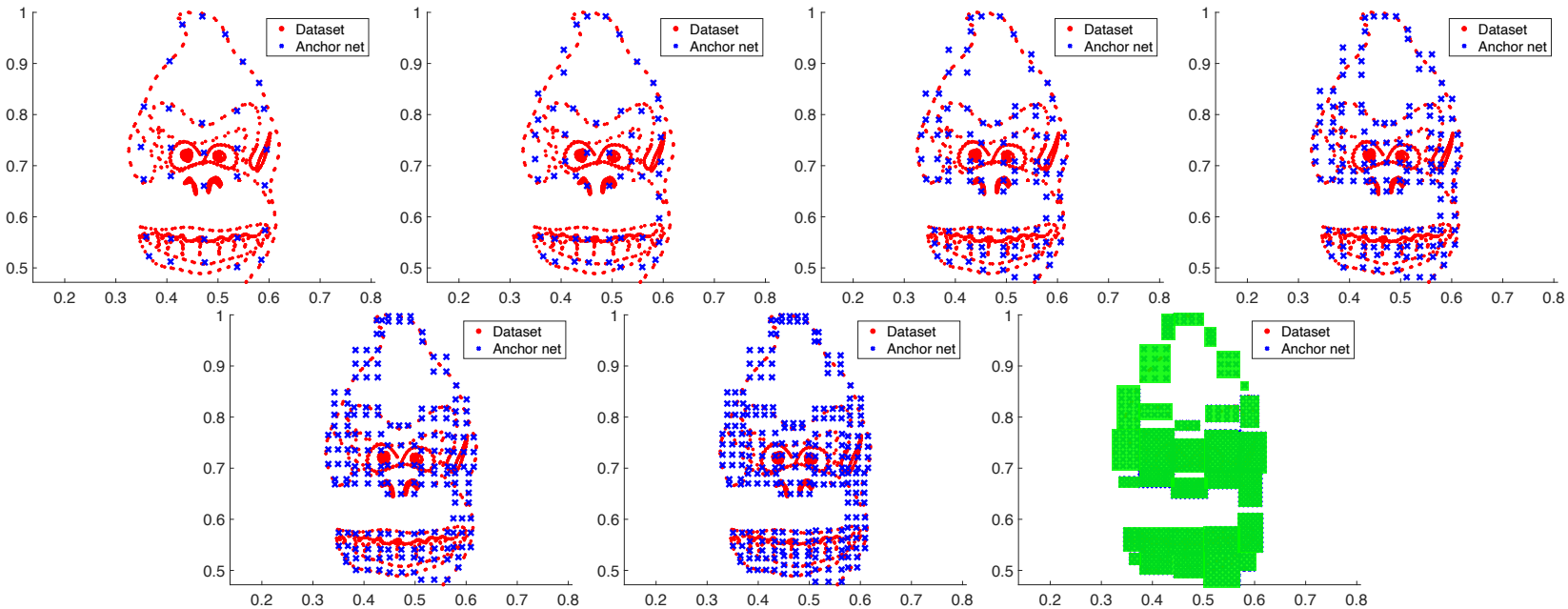
The star discrepancy $D_N^*(\mathcal{A})$ of $\mathcal{A} = \{x_1, \ldots, x_N\} \subset [0,1]^d$ is defined by

$$D_N^*(\mathcal{A}) := \sup_{J \in \mathcal{J}_1} |\#(\mathcal{A} \cap J)/N - \lambda(J)|,$$

where $\mathcal{J}_1$ is the family of all boxes in $[0,1)^d$ of the form $\prod_{i=1}^d [0, a_i)$.

Examples: Halton sequences, digital nets with $D_N^*(\mathcal{A}_N) = O(N^{-1}(\log N)^d)$ with $\mathcal{A}_N$ being the first N terms.

Illustration of Anchor nets



Anchor net

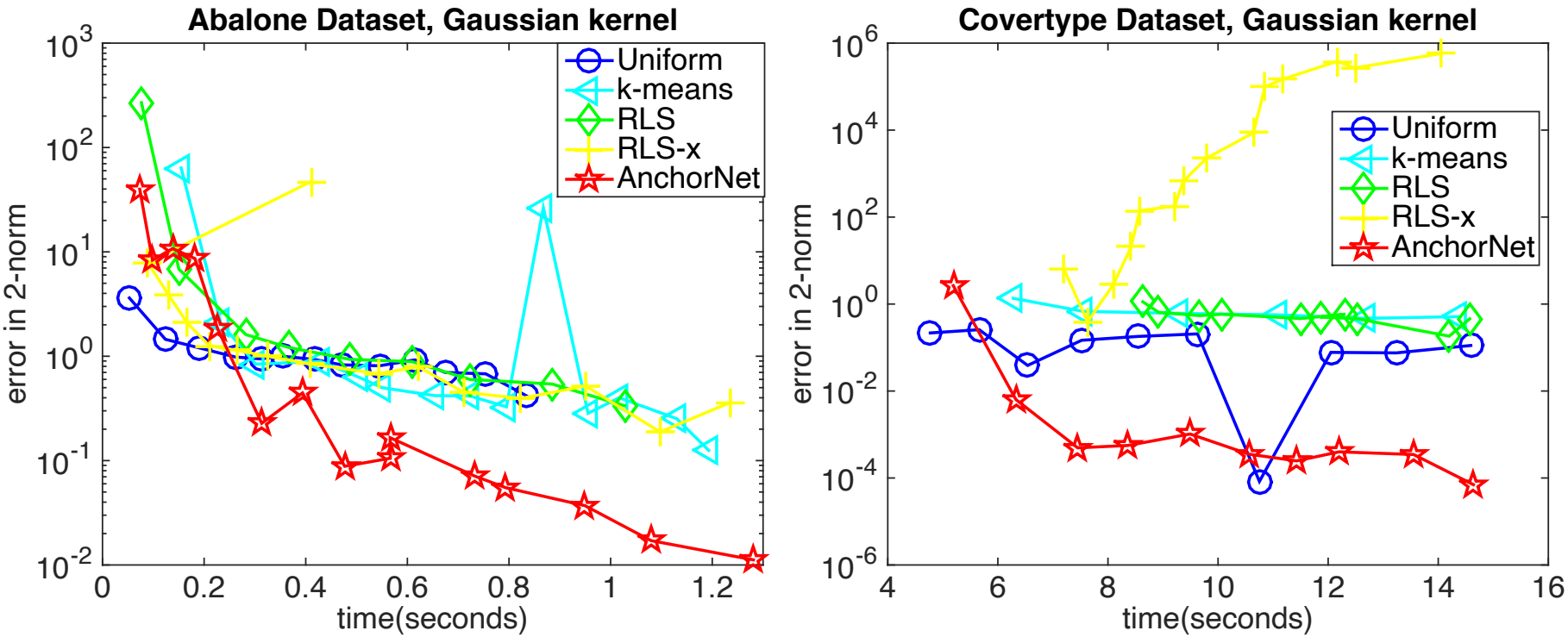
Property (Anchor net $\mathcal{A}_X$)

The Anchor net $\mathcal{A}_X \subset [0,\infty)^d$ *parametrized by* $p = 0, 1, 2, \ldots$ *satisfies*

1. *For* $\Omega = \bigcap_{\epsilon>0} \limsup_{N\to\infty} \left\{x \in \mathbb{R}^d : dist_\infty(x, \mathcal{A}_X) \leq \epsilon\right\}$. *Then* $\lambda(\Omega) > 0$ *and* $X \subset \Omega$.
2. $D_N^*(\mathcal{A}) := \sup_{J\in\mathcal{J}_1} |\#(\mathcal{A}\cap J)/N - \lambda(J)|$.

1. $\mathcal{A}_X$ densely surrounds X as p increases
2. Uniformity of points in $\mathcal{A}_X$ is guaranteed

DATA: ABALONE ($d = 8$), COVERTYPE ($d = 54$)[3]



Error-Time plot for approximating Gaussian kernel matrix

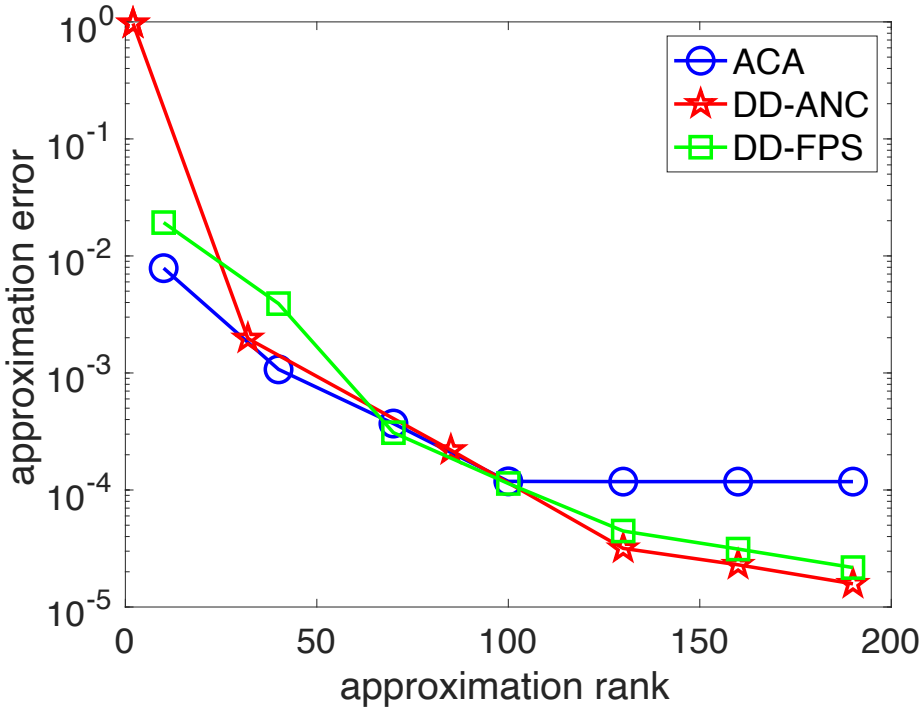
RLS: Recursive Nystrom method (C. Musco, C. Musco, 2017)

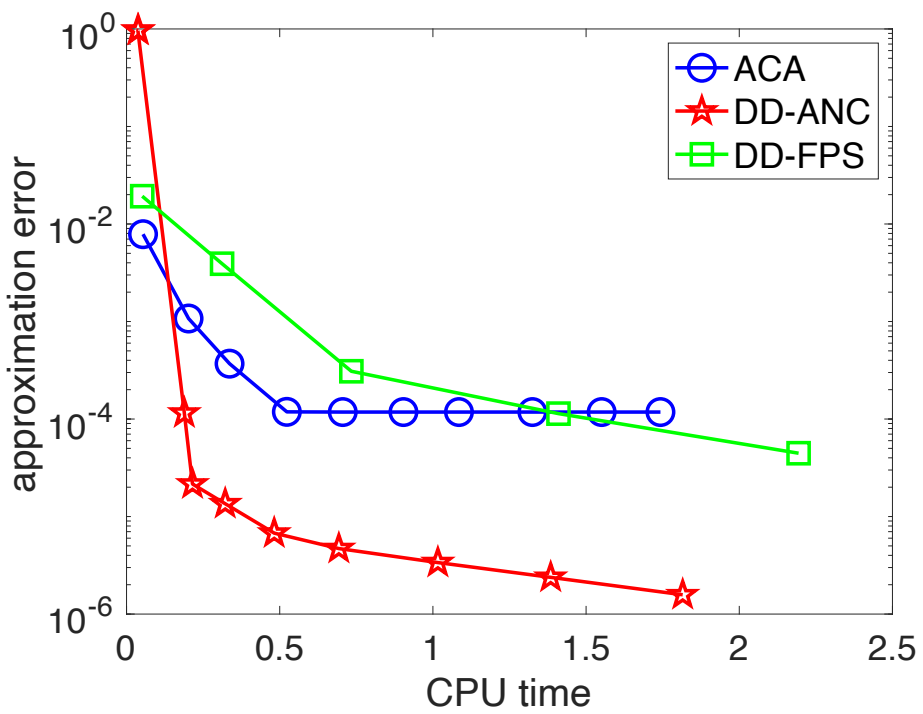
RLS-x: Accelerated RLS (Neurips 2017 Spotlights)

[3]UC Irvine Machine Learning Repository

Data: Gas Sensor($d = 128$)[4]

Randomly sample 8k points for X and 10k points for Y





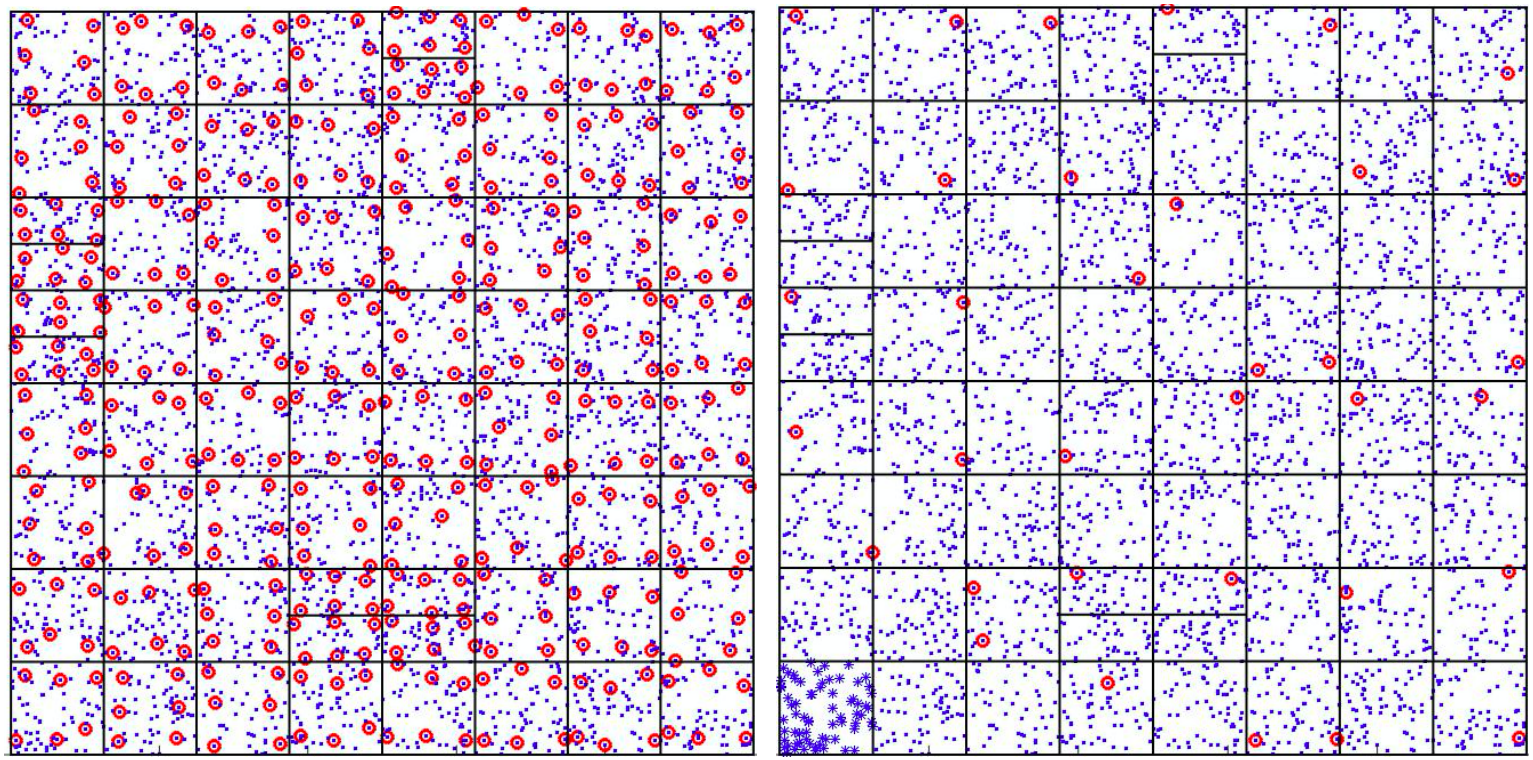
Gaussian kernel with $\sigma = 14.59$

ACA: S. Borm, L. Grasedyck, 2005

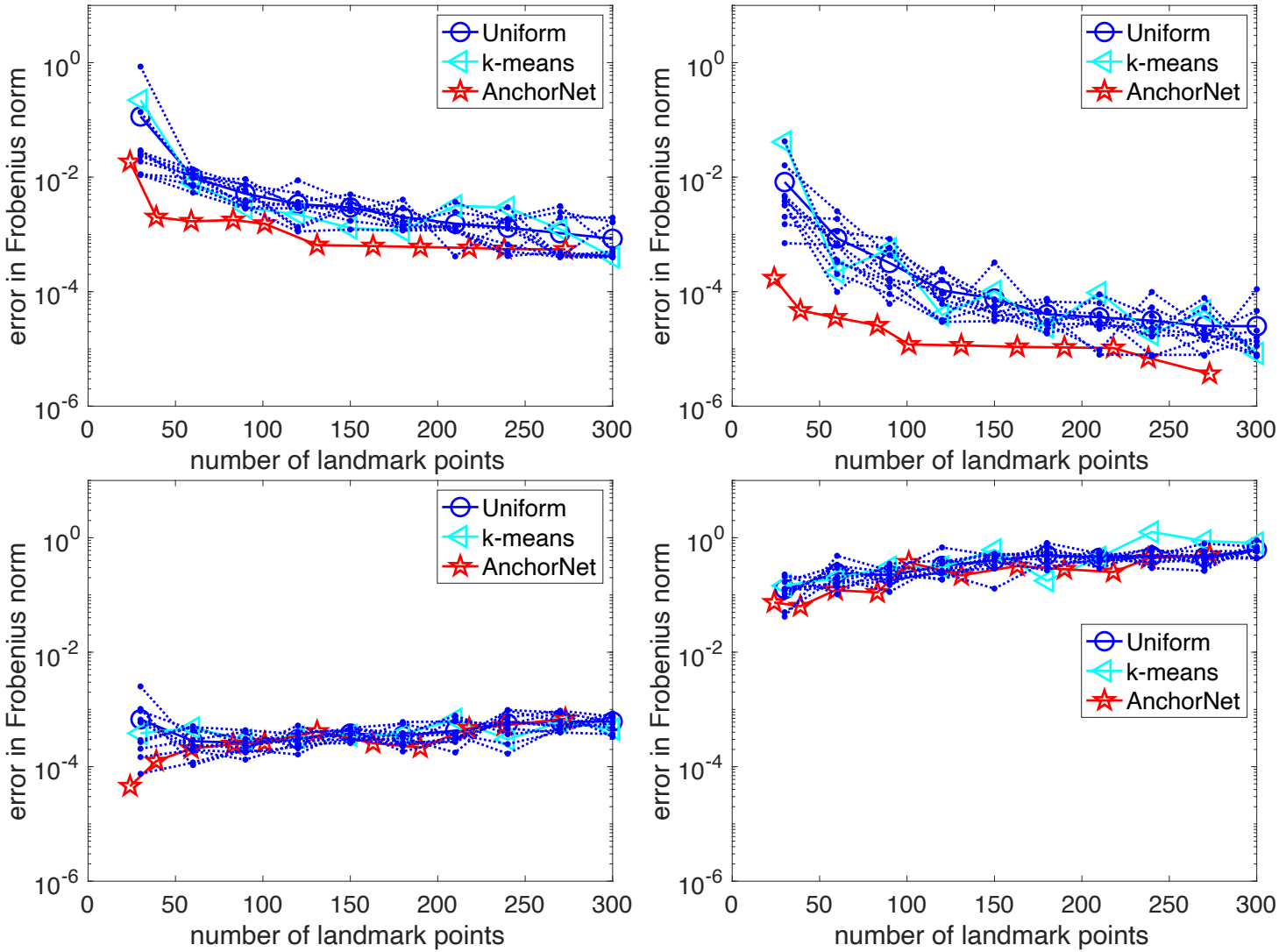
[4] UC Irvine Machine Learning Repository

Hierarchical data driven sampling

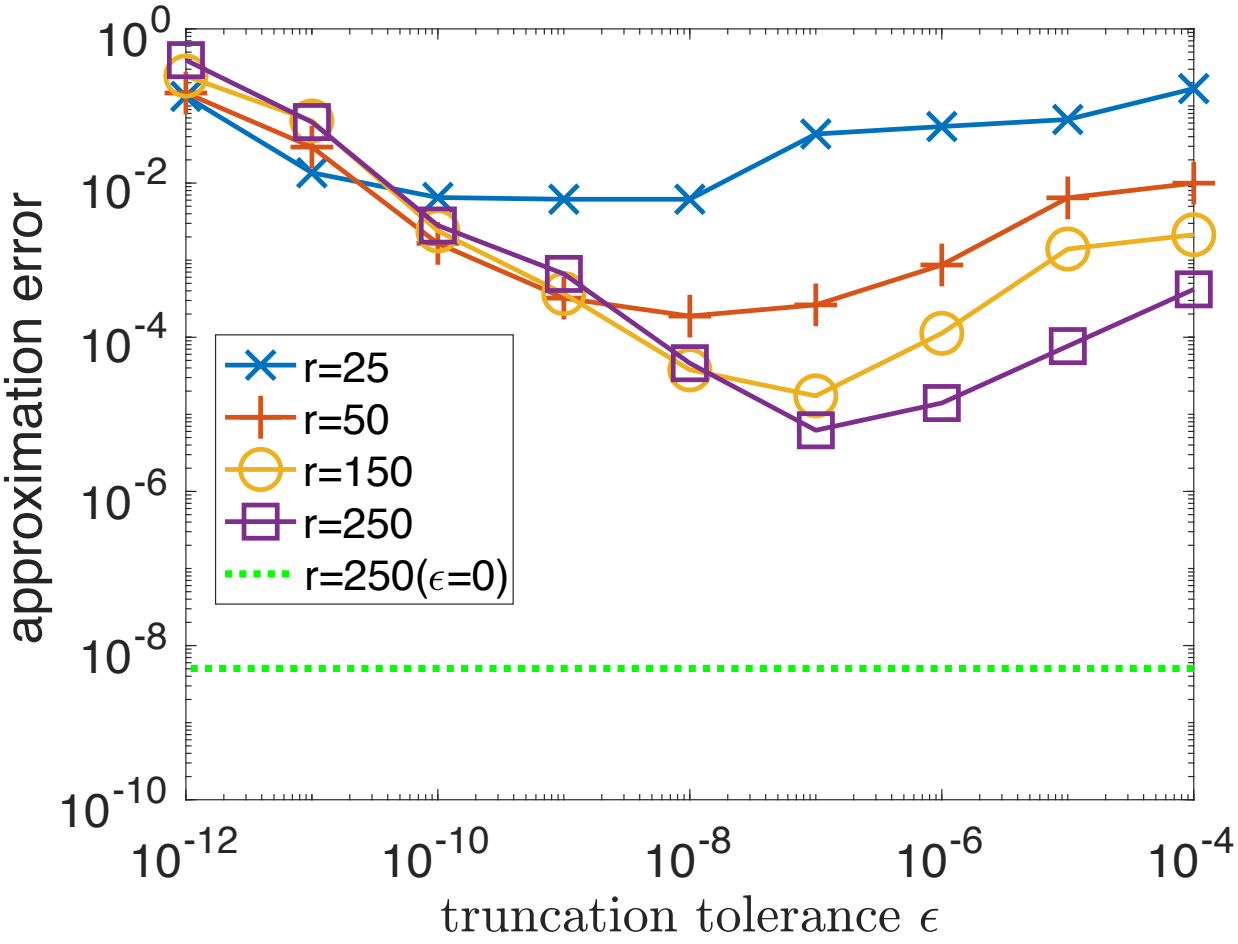
Bottom-up (left) and Top-down sampling (right)



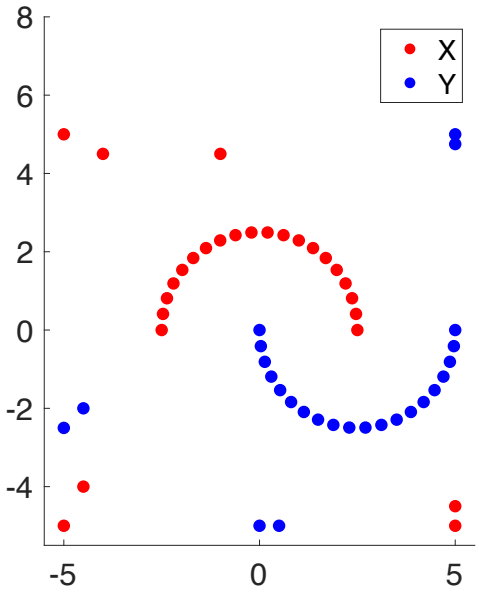
$\bar{U}_{\mathbf{i}} = [\kappa(x, y)]_{x \in X_i, y \in Y_i}$ (Kernel Independent)



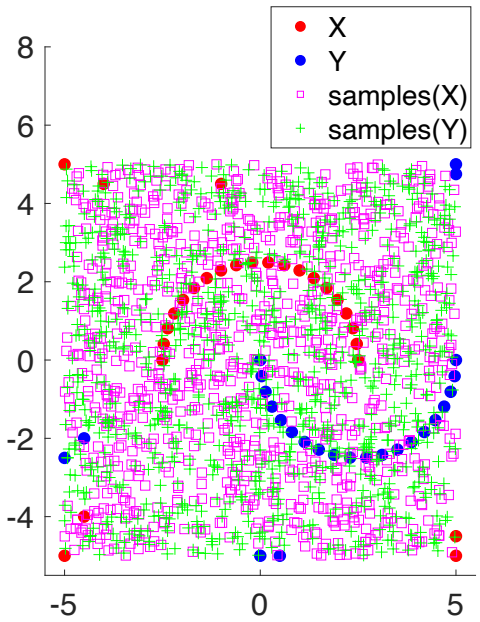
Approximation errors using ϵ-pseudoinverse. First four figures are error-rank plots of three methods with $\epsilon = 10^{-4}, 10^{-6}, 10^{-9}, 10^{-12}$, respectively.


approximation error
truncation tolerance ϵ
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-10}
10^{-8}
10^{-6}
10^{-4}
r=25
r=50
r=150
r=250
r=250(ϵ=0)

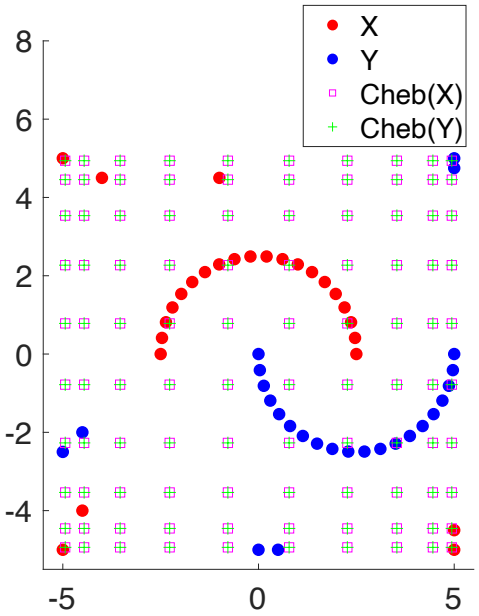
Proxy point approximation



(a) Datasets X, Y



(b) Proxy point method (Xing, Chow, 2020)



(c) Skeletonized interpolation (E. Darve 2019)

Singular value decay of Gaussian matrices