EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Randomized Linear Algebra and Optimization
Lecture 18: Randomized Kernel Approximations, Effective Dimension and Nystrom Method

Recap: Low-rank matrix approximations

- Singular Value Decomposition (SVD)
- $A = U\Sigma V^T$
- takes $O(nd^2)$ time for $A \in R^{n\times d}$
- best rank-k approximation is $A_k := U_k\Sigma_k V_k^T = \sum_{i=1}^k \sigma_i u_i v_i^T$
- $\|A - A_k\|_2 \leq \sigma_{k+1}$

Recap: Randomized low-rank matrix approximations

idea: sample some rows/sketch $A \in \mathbb{R}^{n \times d}$ to get $C \in \mathbb{R}^{n \times m}$

- $C = AS$ where $S \in \mathbb{R}^{d \times m}$ is a sampling/sketching matrix
- we have $AA^T \approx CC^T$

then consider the best approximation of A

in the range of $C = AS$

$$\min_X \|CX - A\|_F$$

- also called CX decomposition
- $\tilde{A}_m := CX^* = CC^\dagger A$ is a randomized rank-m approximation

$$\underbrace{(AS)(AS)^\dagger A} \approx A$$

Recap: Randomized Singular Value Decomposition

- CX decomposition provides the approximation

$$(AS)(AS)^{\dagger}A \approx A$$

$$Q\cancel{R}\,\cancel{R}^{-1}Q^T$$

- calculate QR decomposition of $AS = QR$
- then $QQ^TA \approx A$, i.e., Q approximates the range space of A
- calculate the SVD $\boxed{Q^TA} = U\Sigma V^T$ $\Rightarrow QQ^TA = \underbrace{Q \cdot U}_{U'}\Sigma V^T = U'\Sigma V^T$
- approximate SVD of A is $A \approx (QU)\Sigma V^T$

Analysis of Randomized Low Rank Approximations

- CX decomposition: form sketch AS, and find the best approximation of A in the range of AS

$$X^* = \arg\min_X \|ASX - A\|_F^2 = (AS)^\dagger A$$

- approximation $ASX^* = (AS)(AS)^\dagger A \approx A$
- yields randomized SVD : $AS = QR$ and $Q^TA = U\Sigma V^T$
- Let $A = U\Sigma V^T$ and $A_k = \sum_{i=1}^k \sigma_k u_k v_k^T$, i.e., best rank-k approximation of A
- note that

optimality x^*

$$\|AS\underbrace{(AS)^\dagger A}_{X^*} - A\|_F^2 \leq \|AS(A_kS)^\dagger A_k - A\|_F^2$$

transpose

$$= \|A_k^T(S^TA_k^T)^\dagger S^TA^T - A^T\|_F^2$$

$\min \|S(Ax - B)\|_F^2$

$\|A(SA)^\dagger SB - B\|_F$

Analysis of Randomized Low Rank Approximations

- approximation error

$$
\begin{aligned}
\|AS\underbrace{(AS)^{\dagger}A}_{X^*}-A\|_F^2 &\leq \|AS(A_kS)^{\dagger}A_k - A\|_F^2 \\
&= \|A_k^T(S^TA_k^T)^{\dagger}S^TA^T - A^T\|_F^2 \\
&= \|A_k^T\tilde{X} - A^T\|_F^2
\end{aligned}
$$

where

$$
\tilde{X} := \arg\min_X \|S^TA_k^TX - S^TA^T\|_F^2
$$

Analysis of Randomized Low Rank Approximations

- approximation error

$$
\begin{aligned}
\|AS\underbrace{(AS)^{\dagger}A}_{X^*}-A\|_F^2 &\leq \|AS(A_kS)^{\dagger}A_k - A\|_F^2 \\
&= \|A_k^T(S^TA_k^T)^{\dagger}S^TA^T - A^T\|_F^2 \\
&= \|A_k^T\tilde{X} - A^T\|_F^2
\end{aligned}
$$

where

$$
\tilde{X} := \arg\min_X \|S^TA_k^TX - S^TA^T\|_F^2
$$

- identical to sketching the Generalized Least Squares problem

$$
\min_X \|A_k^TX - A^T\|_F^2
$$

Generalized Least Squares Problems

$$\min_{X} \|AX - B\|_F^2$$

- Least Squares problem with multiple right-hand-sides

  $B = [b_1, \ldots, b_r]$

  $X = [x_1, \ldots, x_r]$

$$\min_{x_1,\ldots,x_r} \sum_{i=1}^{r} \|Ax_i - b_i\|_2^2$$

- optimal solution

$$\begin{aligned} X^* &= [x_1^*, \ldots, x_r^*] \\ &= [A^\dagger b_1, \ldots, A^\dagger b_r] \\ &= A^\dagger B \end{aligned}$$

Left Sketching Generalized Least Squares Problems

- original problem

$$X^* := \arg\min_X \|AX - B\|_F^2$$

- form sketches of the data SA and SB, e.g.,

  uniform row sampling, weighted sampling, Gaussian, ± 1 i.i.d, CountSketch, FJLT...

$$\hat{X} := \arg\min_X \|SAX - SB\|_F^2$$

$$\begin{aligned}\hat{X}_i &= \arg\min_{x_i} \|SAx_i - Sb_i\|_2^2 \\ &= (SA)^\dagger (Sb_i)\end{aligned}$$

- left-sketch applied to simple Least Squares problem $\min_{x_i} \|Ax_i - b_i\|_2^2$

Recall Gaussian Sketch Analysis

- Let $A \in \mathbb{R}^{n\times d}$, $S \in \mathbb{R}^{m\times n}$ be i.i.d. Gaussian

$$x^* := \arg\min_{x\in\mathbb{R}^d} \underbrace{\|Ax - b\|_2^2}_{f(x)} \quad \text{and} \quad \tilde{x} = \arg\min_{x\in\mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Conditioned on the matrix SA

$$A(\tilde{x} - x^*) \sim N\Big(0, \frac{f(x^*)}{m} A(A^T S^T SA)^{-1} A\Big)$$

Recall Gaussian Sketch Analysis

- Let $A \in \mathbb{R}^{n\times d}$, $S \in \mathbb{R}^{m\times n}$ be i.i.d. Gaussian

$$x^* := \arg\min_{x\in\mathbb{R}^d} \underbrace{\|Ax-b\|_2^2}_{f(x)} \quad \text{and} \quad \tilde{x} = \arg\min_{x\in\mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Conditioned on the matrix SA

$$A(\tilde{x}-x^*) \sim N\Big(0, \frac{f(x^*)}{m} A(A^T S^T SA)^{-1} A\Big)$$

- taking expectation over SA, and using $\mathbb{E}\,[(A^T S^T SA)^{-1}] = (A^T A)^{-1}\frac{m}{m-d-1}$ we get

$$\begin{aligned}\mathbb{E}\|A(\tilde{x}-x^*)\|_2^2 &= \frac{f(x^*)}{m-d-1} trA(A^T A)^{-1} A \\ &= f(x^*)\frac{\mathbf{rank}(A)}{m-d-1} = f(x^*)\frac{d}{m-d-1}\end{aligned}$$

Left Sketching Generalized Least Squares Problems

- original problem and left-sketch

$$X^* := \arg\min_X \|AX - B\|_F^2 \quad \text{and} \quad \hat{X} := \arg\min_X \|SAX - SB\|_F^2$$

- x_i: i-th column of $\hat{X}$ satisfies

$$\hat{x}_i = \arg\min_{x_i} \|SAx_i - Sb_i\|_2^2$$

- For a Gaussian sketching matrix S we have

$$\mathbb{E}\|A(\hat{x}_i - x_i^*)\|_2^2 = \|Ax_i^* - b_i\|_2^2 \frac{d}{m-d-1}$$

implies

$$\begin{aligned}\mathbb{E}\|A(\hat{X} - X^*)\|_F^2 &= \sum_{i=1}^{r} \|Ax_i^* - b_i\|_2^2 \frac{d}{m-d-1} \\ &= \|AX^* - B\|_F^2 \frac{d}{m-d-1}\end{aligned}$$

Left Sketching Optimality Gap

- suppose that $\mathbf{rank}(A) = r$
- original problem and left-sketch

$$X^* := \arg\min_X \|AX - B\|_F^2 \quad \text{and} \quad \hat{X} := \arg\min_X \|SAX - SB\|_F^2$$

$$\mathbb{E}\|A(\hat{X} - X^*)\|_F^2 = \|AX^* - B\|_F^2 \, \frac{r}{m - r - 1}$$

$$\begin{aligned}
\mathbb{E}\|A\hat{X} - B\|_F^2 &= \mathbb{E}\|AX^* - B + A(\hat{X} - X^*)\|_F^2 \\
&= \|AX^* - B\|_F^2 + \mathbb{E}\|A(\hat{X} - X^*)\|_F^2 \\
&= \|AX^* - B\|_F^2 \, (1 + \frac{r}{m - r - 1}) \\
&= \|AX^* - B\|_F^2 \, \frac{m - 1}{m - r - 1}
\end{aligned}$$

Back to Randomized Low Rank Approximations

- approximation error

$$
\begin{aligned}
\mathbb{E}\|\underbrace{AS}_{\text{rank} \le m}\underbrace{(AS)^{\dagger}A}_{X^*}-A\|_F^2 &\le \mathbb{E}\|AS(A_kS)^{\dagger}A_k - A\|_F^2 \\
&= \|A_k^T(S^TA_k^T)^{\dagger}S^TA^T - A^T\|_F^2 \\
&= \mathbb{E}\|A_k^T\tilde{X} - A^T\|_F^2 \\
&\le \frac{m-1}{m-k-1}\|\underbrace{A_k^T(A_k^T)^{\dagger}A^T}\, - A^T\|_F^2 \\
&\le \frac{m-1}{m-k-1}\|A(A_k^{\dagger}A_k - I)\|_F^2 \\
&\le \frac{m-1}{m-k-1}\|\underbrace{A_k - A}\|_F^2
\end{aligned}
$$

$$\min_{X} \|A_k^T X - A^T\|_F$$

projector onto the top k s.v subspace

Approx error of exact SVD,

$1+\epsilon$

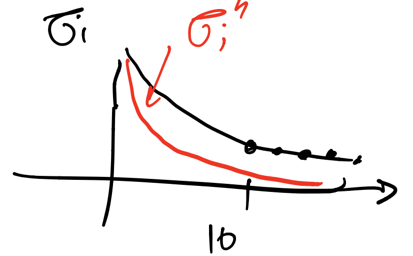
Randomized Low Rank Approximation and Randomized SVD Error Bound

- CX decomposition and randomized SVD
- $AS(AS)^{\dagger}A \approx A$
- final Frobenius norm error bound

$$\mathbb{E}\|AS(AS)^{\dagger}A - A\|_F^2 \leq \frac{m-1}{m-k-1}\|A_k - A\|_F^2$$

- valid for any $k \in \{1, \ldots, \mathbf{rank}(A)\}$

Randomized Low Rank Approximation and Randomized SVD Error Bound



- CX decomposition and randomized SVD
- $AS(AS)^{\dagger}A \approx A$
- final Frobenious norm error bound

$\sigma_i(A^TA) = \sigma_i^2(A)$

$A^TA \cdot A^TA = \sigma_i^4(A)$

$$\mathbb{E}\|AS(AS)^{\dagger}A - A\|_F^2 \leq \frac{m-1}{m-k-1}\|A_k - A\|_F^2$$

- valid for any $k \in \{1, ..., \mathbf{rank}(A)\}$
- define the oversampling factor $\ell := m - k - 1$

$$\|AS(AS)^{\dagger}A - A\|_F^2 \leq (1 + \frac{k}{\ell})\|\underbrace{A_k - A}\|_F^2$$

$$\left\|\sum_{i=k+1}^{r} \sigma_i u_i v_i^T\right\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

Reducing the Error: Power Iteration

- error bounds depend on tail singular values

$$\|A_k - A\|_F^2 = \sum_{j=k+1}^{\mathbf{rank}(A)} \sigma_j^2$$

$AA^T AS$

- idea: compute the sketch of $(AA^T)^q A$

$$C = \underbrace{(AA^T)^q A}_{} S$$

$\text{s.v} = (\sigma_i(A))^{2q+1}$

where q is an integer parameter

- $CC^\dagger A \approx A$

  $CC^\dagger$ approximates the range of A better for $q \geq 1$

- singular values of $(AA^T)^q A$ are $\sigma_i(A)^{2q+1}$

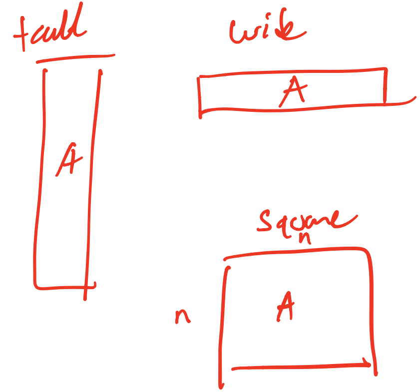
  where $\sigma_i(A)$ are the singular values of A

resembles the power method

$$x_{t+1} = \frac{A^T A x_t}{\|A^T A x_t\|_2} \rightarrow \text{largest e.v.}$$

(if it [illegible])

Approximating Large Square Matrices



- Large and square matrices $A \in \mathbb{R}^{n \times n}$
- Regularized Least Squares

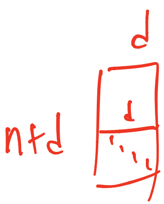
  ℓ_2 (Tikhonov) regularization

$\lambda > 0$

$$x = \underbrace{(A^T A + \lambda I)^{-1}}_{\text{always invertible}} A^T b$$

$$\min_x \|Ax - b\|_2^2 + \lambda \|x\|_2^2$$

- alternative form



$$= \min_x \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2$$

$$\|\sqrt{\lambda} I x\|_2^2 = \lambda \|x\|_2^2$$

Sketching Regularized Problems

$\nabla^2 f(x) = A^T A + \lambda \cdot I$

$\mathbb{E}\, A^T S^T S A + \lambda I = A^T A + \lambda I$ Unbiased!

$$\min_x \Big\| \underbrace{\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}}_{\tilde{A}} x - \underbrace{\begin{bmatrix} b \\ 0 \end{bmatrix}}_{\tilde{b}} \Big\|_2^2$$

$(\tilde{A}^T S^T S \tilde{A})^{-1} \tilde{A}^T S^T S \tilde{b}$ might not be invertible

$\boxed{\frac{d}{m-d-1} \cdot OPT}$

- Left sketch $\min_x \|S\tilde{A}x - S\tilde{b}\|_2^2$ approximates the solution when sketch dimension $m > d + 1$, e.g., for Gaussian S
- Sketch dimension can be smaller if we use a *partial sketch*

$SA = U\Sigma V^T$ $\boxed{O(m^2 d)}$

$$\min_x \|SAx - Sb\|_2^2 + \lambda \|x\|_2^2$$

$\hat{x} = (A^T S^T S A + \lambda I)^{-1} A^T S^T S b = V(\Sigma^2 + \lambda I)^{-1} V^T A^T S^T S b$

- the term $\sqrt{\lambda} I$ is not sketched/subsampled

$O(d^3)$ $(A^T S^T S A + \lambda I)^{-1}$ $\gg$ $O(m^2 d)$ SVD of SA

always invertible!

$m = 100$

$A : n \times d$

$n = 1M$

$d = 1M$

Sketching Regularized Problems

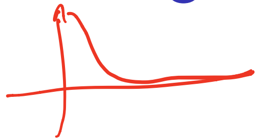
$d_e(\lambda)$: effective dim, at level λ.

$$x^* = \arg\min_x \underbrace{\|Ax - b\|_2^2 + \lambda\|x\|_2^2}_{f(x)}$$

$$\hat{x} = \arg\min_x \|SAx - Sb\|_2^2 + \lambda\|x\|_2^2$$

- approximation ratio $f(\hat{x}) \leq f(x^*)(1+\epsilon)$ when $m \geq \text{constant} \times d_e(\lambda)/\epsilon^2$ for i.i.d. Gaussian, sub-Gaussian and FJLT sketch (ignoring log factors)
- $d_e(\lambda) = \sum_{i=1}^{d} \frac{\sigma_i(A)^2}{\sigma_i(A)^2+\lambda}$ is the *effective dimension* of A
- $d_e(0) = \text{rank}(A)$ o.w $d_e(\lambda) < \text{rank}(A)$

Hessian Sketching for Regularized Problems



f : convex function

$\nabla^2 f$ is P.S.D.

$$\min_x f(Ax) + \lambda \|x\|_2^2$$

- sketched Newton iterations

$\underline{\nabla^2 f} \quad + \lambda I$

$$x_{t+1} = \arg\min_x \frac{1}{2} \|S(\nabla^2 f(x_t))^{1/2} \overbrace{\cancel{x}}^{(x-x_t)}\|_2^2 + (x - x_t)^T \nabla f(x_t) + \frac{\lambda}{2}\|x\|_2^2$$

$(x-x_t)^T (\nabla^2 f)^{1/2} S^T S (\nabla^2 f)^{1/2} (x-x_t)$

- $(\nabla^2 f(x_t))^{1/2} S^T S (\nabla^2 f(x_t))^{1/2} + \lambda I$ is invertible for all m when $\lambda > 0$
- similar guarantees involving the effective dimension of the Hessian matrix
- $\lambda = 0$ requires $m > d$ for invertibility

Kernel Matrices

$\|A\alpha - b\|_2$

$A = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$

- Large square matrices $K \in \mathbb{R}^{n \times n}$
- Kernel Ridge Regression

$$\min_{\alpha} ||K\alpha - y||_2^2 + \underbrace{\lambda \alpha^T K \alpha}_{\text{smoothness}}$$

- K is called the **kernel matrix**
- $K = \kappa(x_i, x_j)$ where $x_1, ..., x_n \in \mathbb{R}^d$ are data vectors
  κ is the **kernel function**
- prediction at a point x is $\sum_{i=1}^{n} \kappa(x_i, x)\alpha_i$, i.e, predictions on the training set are $K\alpha \approx y$

  $K = A^T A$

- examples:

  sobolev kernel

  Gaussian kernel $K_{ij} = \kappa(x_i, x_j) = e^{-\frac{1}{\sigma^2}\|x_i - x_j\|_2^2}$

  $\gamma < 1$

  Polynomial kernel $K_{ij} = \kappa(x_i, x_j) = (x_i^T x_j)^r$

  $r=1$ linear kernel $\Rightarrow$ LS boils down to

  n^γ

- Kernel matrices typically have low effective dimension, e.g.,
- Gaussian kernel has $d_e(\lambda) = O(\sqrt{\log n})$ for $\lambda = \sqrt{\frac{\log n}{n}}$. This choice of λ provides optimal statistical guarantees

Kernel Trick

$x_i^T x_j$

liner
least sq

$K_{ij} = (x_i^T y_j)^2$

- Kernel Ridge Regression

$$\min_{\alpha} ||K\alpha - y||_2^2 + \lambda \alpha^T K \alpha$$

example: polynomial kernel (degree 2)
$K_{ij} = \kappa(x_i, x_j) = (x_i^T x_j)^2 = \left(\sum_j x_{ij} x_{ij} \right)^2$

- maps data to higher dimension

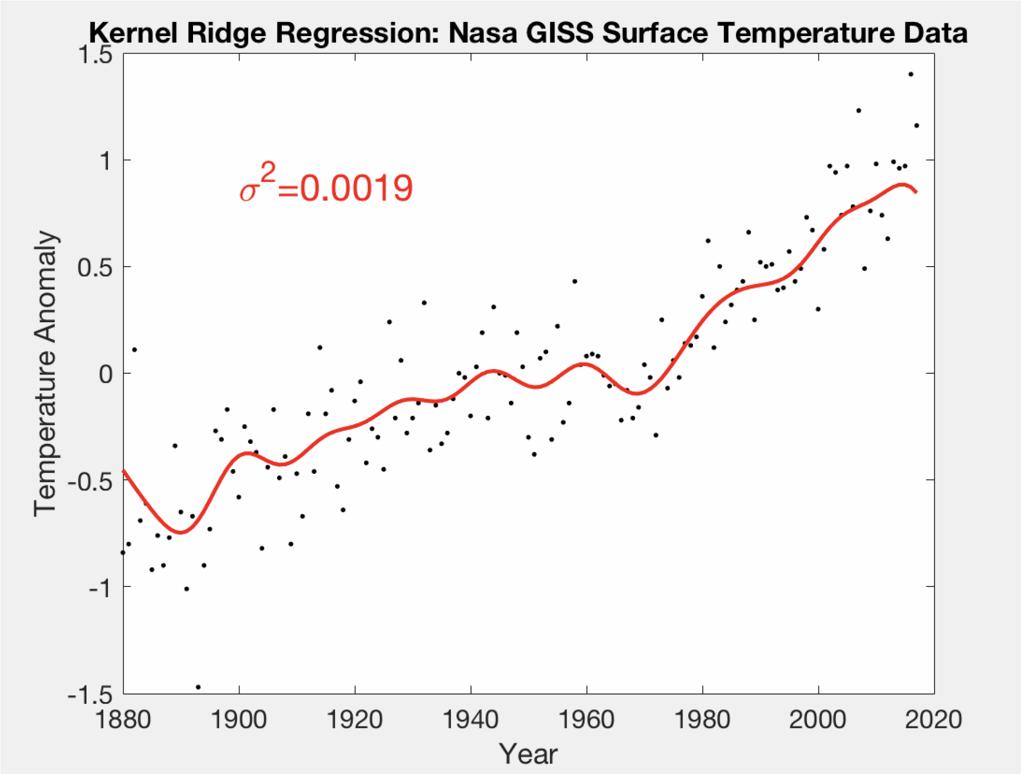
$$A = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \\ x_{n1} & \dots & x_{nd} \end{bmatrix} \rightarrow$$

$$\tilde{A} := \begin{bmatrix} x_{11} & \dots & x_{1d} & x_{11}^2 & \dots & x_{1d}^2 \\ \vdots & & & & & \\ x_{n1} & \dots & x_{nd}^2 & x_{11}^2 & \dots & x_{nd} \end{bmatrix}$$

Application: Kernel Regression

$f(x) = \sum e^{-\frac{\|x - x_i\|_2^2}{2\sigma^2}} \cdot \alpha_i$

Gaussian Kernel $\qquad K_{ij} = e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$



Application: Kernel Classification

$\ell(\cdot, y) = (\cdot - y)^2$

$\min \|K\alpha - y\|_2^2 + \lambda \alpha^T K \alpha$

partial sketch: $\|SK\alpha - Sy\|_2^2 + \lambda \alpha^T K \alpha$

$(KS^TSK + \lambda K)^{-1} K S^T S y$

right sketching

$\|K \cdot S z - y\|_2^2 + \lambda z^T S^T K S z$

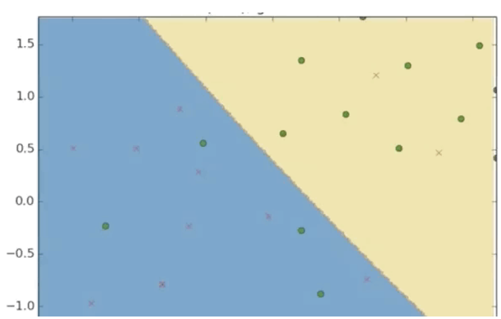
$z^* = (S^T K^2 S + S^T K S)^{-1} S K y$

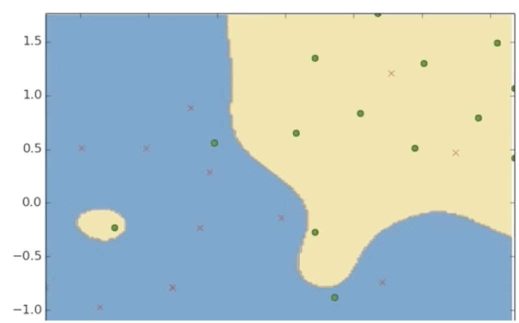
$\hat{\alpha} = S z^*$

$$\min_{\alpha} \sum_{i=1}^{n} \ell(K\alpha, y) + \lambda \alpha^T K \alpha$$

$f($ RKHS

linear kernel $K_{ij} = x_i^T x_j$ gaussian kernel $K_{ij} = e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$



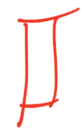


$KS(KS)^{\dagger}K \approx K$

Nystrom Method

sym. matrix

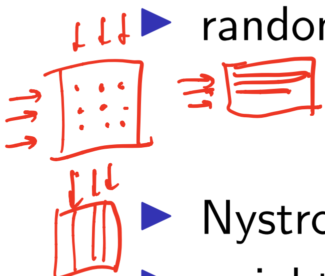
- We need a symmetric approximation. CX decomposition is not symmetric.
- Most kernel matrices are positive semidefinite, i.e., $K = A^T A$ for some matrix A — A might be hard to compute
- Recall the CX decomposition $\tilde{A} = (AS)(AS)^{\dagger}A \approx A$ we used in randomized SVD



AS full column rank

- Consider approximating $A^T A$ via $\tilde{A}^T\tilde{A}$

$$
\begin{aligned}
((AS)(AS)^{\dagger}A)^T(AS)(AS)^{\dagger}A &= A^T(AS)(\cancel{AS})^{\dagger}(\cancel{AS})(AS)^{\dagger}A \\
&= A^T(AS)(AS)^{\dagger}A \\
&= \underbrace{A^TA}_{K}S(S^T\underbrace{A^TA}_{K}S)^{-1}S^T\underbrace{A^TA}_{K}
\end{aligned}
$$



- randomized low rank approximation of K is given by

$$
\tilde{K} = KS(S^TKS)^{-1}S^TK \approx K
$$

- Nystrom Method: S is uniform column sampling
- weighted sampling or sketching can also be used

Generalized Nystrom Method

- Nystrom method can be generalized to non symmetric matrices
- Consider CX decomposition where $C = AS$ and S is a sketching matrix

$$\min_X \|ASX - A\|_F$$

- Apply another sketching matrix R on the left

any $$\min_X \|RASX - RA\|_F$$

- solution $X^* = (RAS)^\dagger RA$
- approximation of A is $AS(RAS)^\dagger RA \approx A$
- reduces to the Nystrom method when $R = S$ and $A = A^T$
- faster than CX and randomized SVD, less accurate

Random Fourier Features

$\nexists\ A^T S^T S A = A^T A$

$\nexists\ \exp(A^T S^T j)\exp(SA \cdot j) =$ Nonlinear

- Random approximations of kernel matrices
- Generate $w \sim N(0, I)$
- Define features $h(x) := e^{-jw^T x}$ where $j = \sqrt{-1}$

  it holds that

$$
\begin{aligned}
\mathbb{E}_w h(x) h(y)^* &= \mathbb{E}_w e^{-jw^T x} e^{+jw^T y} \\
&= \mathbb{E}_w e^{-jw^T (x-y)} \\
&= \int p(w) e^{-jw^T (x-y)} dw \\
&= e^{-\frac{1}{2}(x-y)^T (x-y)}
\end{aligned}
$$

✓ Gaussian kernel!

- where $p(w)$ is the multivariate Gaussian distribution
- **Bochner's Theorem:** Fourier transforms of probability distributions correspond to positive semidefinite kernels
- Gaussian distribution corresponds to the Gaussian kernel

Random Fourier Features

- Random approximations of kernel matrices
- Generate $w_1, \ldots, w_m \sim N(0, I)$ i.i.d.
- Define feature vectors

$$S = \begin{bmatrix} w_1^T \\ | \\ w_m^T \end{bmatrix}_{m \times d}$$

$$h(x) = \frac{1}{\sqrt{m}} \begin{bmatrix} e^{jw_1^T x} \\ e^{jw_2^T x} \\ \ldots \\ e^{jw_m^T x} \end{bmatrix} = \frac{1}{\sqrt{m}} \exp(Sx \cdot j)$$

- then we have

$$\langle h(x), h(y) \rangle = \frac{1}{m} \sum_{i=1}^{m} e^{jw_i^T (x-y)} \approx \mathbb{E}_w e^{jw^T (x-y)} = e^{-\frac{1}{2}(x-y)^T (x-y)}$$

$$= \kappa(x, y)$$

Rahimi and Recht, Random Features for Large-Scale Kernel Machines, 2007 $\tilde{A} = \exp(AS \cdot j) \Rightarrow \mathbb{E}\, \tilde{A}^T \tilde{A} = K$

Random Fourier Features

- The embedding is nonlinear $\frac{1}{\sqrt{m}} \exp(iXS)$
- can also be obtained using real valued embeddings
- Generate $w \sim N(0, I)$ i.i.d.
- $h(x) = \sqrt{2}\cos(w^T x + b)$ where $b \sim \text{Uniform}(0, 2\pi)$ also works
- the approximation error $\left\| \left\{ \frac{1}{m}\sum_{i=1}^m e^{jw_i^T(x_k - y_l)} - \mathbb{E}\frac{1}{m}\sum_{i=1}^m e^{jw_i^T(x_k - y_l)} \right\}_{kl} \right\|_2$ can be controlled via matrix concentration bounds