EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Thursday, March 11 2021

Randomized Linear Algebra and Optimization
Lecture 19: Kernel Matrices, Effective Dimension, Nystrom Method and Random Fourier Features

Approximating Large Square Matrices

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

  ℓ_2 (Tikhonov) regularization

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

- alternative form

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

Sketching Regularized Problems

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

- 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*

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

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

Sketching Regularized Problems

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

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

- approximation ratio $f(\hat{x}) \leq f(x^*)(1+\epsilon)$
  when $m \geq \text{constant} \times d_e(\lambda)$
  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
- $d_e(0) = \mathrm{rank}(A)$

Hessian Sketching for Regularized Problems

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

- sketched Newton iterations

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

- $\big(\nabla^2 f(x_t)\big)^{1/2} S^T S \big(\nabla^2 f(x_t)\big)^{1/2} + \lambda I$ 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

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

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

- 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 $K\alpha \approx y$
- examples:

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

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

- Kernel Ridge Regression

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

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

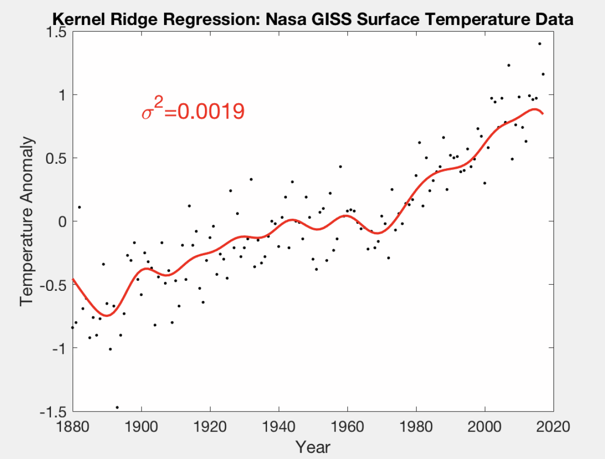
- maps data to higher dimension

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

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

Application: Kernel Regression

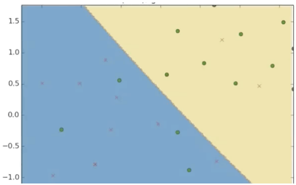
$$\text{Gaussian Kernel} \qquad K_{ij} = e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$$

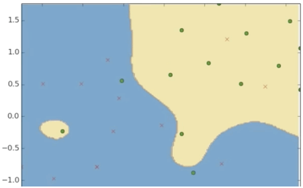


Application: Kernel Classification

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

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





Nystrom Method

- 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
- Recall the CX decomposition $\tilde{A} = (AS)(AS)^\dagger A \approx A$ we used in randomized SVD
- Consider approximating $A^T A$ via $\tilde{A}^T \tilde{A}$

$$\begin{aligned}\left((AS)(AS)^\dagger A\right)^T (AS)(AS)^\dagger A &= A^T (AS)(AS)^\dagger (AS)(AS)^\dagger A \\ &= A^T (AS)(AS)^\dagger A \\ &= A^T AS(S^T A^T AS)^{-1} S^T A^T A\end{aligned}$$

- randomized low rank approximation of K is given by

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

- 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

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

- 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

- 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

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

- 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 $w_1, ..., w_m \sim N(0, I)$ i.i.d.
- Define feature vectors

$$h(x) = \frac{1}{\sqrt{m}} \begin{bmatrix} e^{-jw_1^T x} \\ e^{-jw_2^T x} \\ \dots \\ e^{-jw_m^T x} \end{bmatrix}$$

- then we have

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

- Kernel matrix can be approximated via a rank m matrix, i.e., $K_{ij} \approx \frac{1}{m} \sum_{i=1}^{m} e^{jw_i^T (x_i - y_i)} = \langle h(x_i), h(x_j) \rangle$
  Rahimi and Recht, Random Features for Large-Scale Kernel Machines, 2007

Random Fourier Features

- The embedding is a nonlinear sketch:

  Let $A = [x_1, x_2, \ldots, x_n]^T$, define $\tilde{A} := \frac{1}{\sqrt{m}} \exp(-iAS)$

  where $\exp(\cdot)$ is the entrywise scalar exponential function.

  We have $\tilde{A}^T \tilde{A} \approx K$ since $\mathbb{E}\tilde{A}^T \tilde{A} = K$
- 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 $\|\tilde{A}^T \tilde{A} - \mathbb{E}\tilde{A}^T \tilde{A}\|_2$ can be controlled via matrix concentration bounds since $\tilde{A}^T \tilde{A}$ is a sum of m i.i.d. matrices.
- equivalently, we may use random nonlinear features $h(x)$ in linear models, e.g., least squares, logistic regression, SVM etc.
- usually faster than Nystrom but less accurate