Lecture 35

Final

Review (I)

Final exam on Fri 12/18

Practice exam will be posted soon
Algorithms!

0. Gaussian Elimination
   \[ Ax = b \]

0. Gram-Schmidt
   \( \{ v_1, \ldots, v_n \} \)

0. Least Squares / Projections
   \( x = \sum s \)

0. SVD / PCA

0. Power Method
   \[ v_k = A^k v_0 \]

0. Perceptron

0. Gradient Descent

0. SGD
   \[ \min_x \sum_i F_i(x) \]
You should have a good sense of what all these terms mean, and how they relate to each other.

Which algorithms to use when, and why.

Understand what does not make sense, e.g.

- divide two vectors $\frac{u}{v}$, $u, v \in \mathbb{R}^n$
- eigenvalues of a nonsquare matrix
- gradient of a matrix
- quadratic form $x^TQx$ with a rectangular
LEAST SQUARES: Two Flavors

OVERDETERMINED $Ax = b$ has no solution

Define approx. least sq sol by

$$\min_x \|Ax - b\|^2$$

UNDERDETERMINED $Ax = b$ has infinitely many solutions

Define minimum norm (LS) solution by

$$\min \|x\|^2 \quad \text{st.} \ Ax = b$$
OVERDETERMINED

\[ \mathbf{A} \mathbf{x} = \mathbf{b} \]

has no solution

Define approx. least sq sol by

\[ \min_{\mathbf{x}} \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2 \]

\[ \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2 \quad \forall \mathbf{x} \in \text{range of } \mathbf{A} \]

How to solve it?

**PROJECTION FORMULA**

(project b onto the range of A)

**SVD**

\[ \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \]

\[ \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2 = \| \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} - \mathbf{b} \|^2 \]

\[ = \| \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} - \mathbf{U}^T \mathbf{b} \|^2 \]

\[ = \| \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} - \mathbf{\Sigma} \mathbf{V}^T \mathbf{b} \|^2 \]

\[ = \| \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} - \mathbf{\Sigma} \mathbf{V}^T \mathbf{b} \|^2 \]

\[ = \| \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} - \mathbf{b} \|^2 \]

\[ \uparrow \text{ easy to solve} \]

**OPTIMIZATION**

\[ \min_{\mathbf{x}} \frac{1}{2} (\mathbf{A} \mathbf{x} - \mathbf{b})^T (\mathbf{A} \mathbf{x} - \mathbf{b}) \]

\[ \frac{1}{2} \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2 \]

(optimality conditions)

\[ \nabla F(\mathbf{x}) = 0 \quad \mathbf{A}^T (\mathbf{A} \mathbf{x} - \mathbf{b}) = 0 \quad \text{(normal eqs)} \]

\[ \mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b} \]

\[ \Rightarrow \mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \]
(Aside)

**Cool Linear Algebra Fact #731**

How many **orthogonal** unit-norm vectors can I find in \( \mathbb{R}^n \)?

E.g. in \( \mathbb{R}^2 \):

\[
\begin{pmatrix}
1
0
\end{pmatrix}, \begin{pmatrix}
0
1
\end{pmatrix}
\]

(2 vectors)

In \( \mathbb{R}^n \) \( \rightarrow \) at most \( n \) orthogonal vectors.

(Why? Orthogonal \( \Rightarrow \) LI)

But, what if I allow them to be "almost orthogonal"?
In high dimensions, a LOT more.

For instance, in $\mathbb{R}^{10^6}$, can find $1.17 \times 10^{33}$ vectors with $89^\circ \leq \angle (v_i, v_j) \leq 91^\circ$

(Compare against $n = 10^6 (!)$)
**Quadratic Programs**

A convex quadratic function of $x$:

$$\min_x q(x)$$

Subject to:

$$Ax = b$$

Linear equations (subspace)

- $\|Ax - b\|^2$
- $\|x\|^2$
- $x^TQx$
- $\frac{1}{2} x^TQx + b^Tx$
- ...

(matrix $Q$ is symmetric)

$$q(x) = (Ax - b)^T(Ax - b) = x^T A^TA x + \ldots$$

$q(x)$ must be convex, i.e.,

Matrix $Q$ is PSD ($\lambda_i \geq 0$)
Convex Programs

\[ \begin{align*}
\min_{x} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0
\end{align*} \]

\( f(x) \) are convex functions
\( g_i(x) \) are convex functions

Convex sets

Convex functions

\[ \forall \lambda \in [0,1], \quad f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) \]

\( f : \mathbb{R}^n \to \mathbb{R} \)

If convex:

Local minimum \( \Rightarrow \) Global minimum

\( x^* \) global optimum \( \iff \) \( \nabla f(x^*) = 0 \)
**Nonnegative Least Squares**

\[ \min \| Ax - b \|^2 \quad (\text{NNLS}) \]

**Constraints:**
\[ x \geq 0 \]

*Not* a least squares problem (or the type we have seen).

---

**Classification:**

\[ c^T x + b = 0 \]

\[ \min \| c \|^2 \]

\[ y (c^T x + b) \geq -1 \]

**Convex Problems \Rightarrow Good Algorithms**
Snow White distributed 399 liters of milk among the seven dwarfs. The first dwarf then distributed the contents of his pail evenly to the pails of other six dwarfs. Then the second did the same, and so on. After the seventh dwarf distributed the contents of his pail evenly to the other six dwarfs, it was found that each dwarf had exactly as much milk in his pail as at the start.

a) What was the initial distribution of the milk?
b) Generalize to N dwarfs.

\[ T_i = \begin{bmatrix} 2 & 1 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 1 & 1 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \Rightarrow T_t \ldots - T_2 T_1 \]
\[ x_{k+1} = x_k - \alpha \nabla f(x_k) \]
**Gradient Descent**

(\textit{convex functions})

- Converges for a range of stepsizes

\[ 0 < \gamma < \frac{2}{L} \]

\[ 0 < m \leq \lambda(x) \leq L \]

- Eigens of Hessian

- For strictly convex functions (i.e., \( m > 0 \))
  rate depends on condition number \( \frac{L}{m} \)

- Accelerated methods (e.g., heavy ball)
  can achieve improvements

- For large, data-oriented problems,
  SGD is a good option
Next time:

Overview of some research directions + other courses at MIT