CS 598 EVS: Tensor Computations
Matrix Computations Background

Edgar Solomonik

University of Illinois at Urbana-Champaign
Conditioning

- Absolute Condition Number:

- (Relative) Condition Number:
Posedness and Conditioning

- What is the condition number of an ill-posed problem?
Matrix Condition Number

- The matrix condition number \( \kappa(A) \) is the ratio between the max and min distance from the surface to the center of the unit ball (norm-1 vectors) transformed by \( A \):

- The matrix condition number bounds the worst-case amplification of error in a matrix-vector product:
Singular Value Decomposition

- The singular value decomposition (SVD)

- Condition number in terms of singular values
Visualization of Matrix Conditioning

\[ \{ x : x \in \mathbb{R}^2, \| x \|_2 = 1 \} \xrightarrow{A} \{ Ax : x \in \mathbb{R}^2, \| x \|_2 = 1 \} \]

\[ \kappa(A) = \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \]

\[ A = U \begin{bmatrix} \sigma_{\text{max}} \\ \sigma_{\text{min}} \end{bmatrix} V^T \]

\[ \| A \|_2 = \sigma_{\text{max}} \]

\[ 1/\| A^{-1} \|_2 = \sigma_{\text{min}} \]
Linear Least Squares

- Find $x^* = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$ where $A \in \mathbb{R}^{m \times n}$:

  - Given the SVD $A = U \Sigma V^T$ we have $x^* = \underbrace{V \Sigma^{\dagger} U^T}_A b$, where $\Sigma^{\dagger}$ contains the reciprocal of all nonzeros in $\Sigma$, and more generally $\dagger$ denotes pseudoinverse:
Normal Equations

- Normal equations are given by solving $A^T A x = A^T b$:

- However, solving the normal equations is a more ill-conditioned problem than the original least squares algorithm.
Solving the Normal Equations

- If $A$ is full-rank, then $A^T A$ is symmetric positive definite (SPD):

- Since $A^T A$ is SPD we can use Cholesky factorization, to factorize it and solve linear systems:
**QR Factorization**

- If $A$ is full-rank there exists an orthogonal matrix $Q$ and a unique upper-triangular matrix $R$ with a positive diagonal such that $A = QR$

- A reduced QR factorization (unique part of general QR) is defined so that $Q \in \mathbb{R}^{m \times n}$ has orthonormal columns and $R$ is square and upper-triangular

- We can solve the normal equations (and consequently the linear least squares problem) via reduced QR as follows
Computing the QR Factorization

- The Cholesky-QR algorithm uses the normal equations to obtain the QR factorization

- Orthogonalization-based methods are most efficient and stable for QR factorization of dense matrices
Eigenvalue Decomposition

- If a matrix $A$ is diagonalizable, it has an *eigenvalue decomposition*

- $A$ and $B$ are *similar*, if there exist $Z$ such that $A = ZBZ^{-1}$
## Similarity of Matrices

<table>
<thead>
<tr>
<th>matrix</th>
<th>similarity</th>
<th>reduced form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>real symmetric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hermitian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>real</td>
<td></td>
<td></td>
</tr>
<tr>
<td>diagonalizable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>arbitrary</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rayleigh Quotient

- For any vector \( x \) that is close to an eigenvector, the \textit{Rayleigh quotient} provides an estimate of the associated eigenvalue of \( A \):
Krylov subspace methods work with information contained in the $n \times k$ matrix

$$K_k = \begin{bmatrix} x_0 & Ax_0 & \cdots & A^{k-1}x_0 \end{bmatrix}$$

A is similar to companion matrix $C = K_n^{-1} A K_n$. 

Krylov Subspaces

- Given $Q_k R_k = K_k$, we obtain an orthonormal basis for the Krylov subspace,

$$\mathcal{K}_k(A, x_0) = \text{span}(Q_k) = \{p(A)x_0 : \deg(p) < k\},$$

where $p$ is any polynomial of degree less than $k$.

- The Krylov subspace includes the $k - 1$ approximate dominant eigenvectors generated by $k - 1$ steps of power iteration:
The $k \times k$ matrix $H_k = Q_k^T A Q_k$ minimizes $\|AQ_k - Q_k H_k\|_2$:

- $H_k$ is upper-Hessenberg, because the companion matrix $C_n$ is upper-Hessenberg:
Rayleigh-Ritz Procedure

- The eigenvalues/eigenvectors of $H_k$ are the *Ritz values/vectors*:

- The Ritz vectors and values are the *ideal approximations* of the actual eigenvalues and eigenvectors based on only $H_k$ and $Q_k$:
Low Rank Matrix Approximation

- Given a matrix $A \in \mathbb{R}^{m \times n}$ seek rank $r < m, n$ approximation

- Eckart-Young (optimal low-rank approximation by SVD) theorem
Rank Revealing Matrix Factorizations

- Computing the SVD
- QR with column pivoting
Orthogonal Iteration

- For sparse matrices, QR factorization creates fill, so must revert to iterative methods

- Orthogonal iteration interleaves deflation and power iteration
Randomized SVD

- Orthogonal iteration for SVD can also be viewed as a randomized algorithm
Generalized Nyström Algorithm

- The generalized Nyström algorithm provides an efficient way of computing a sketched low-rank factorization
Multidimensional Optimization

- Minimize $f(x)$

- Quadratic optimization $f(x) = \frac{1}{2} x^T A x - b^T x$
Basic Multidimensional Optimization Methods

- Steepest descent: minimize $f$ in the direction of the negative gradient:

- Given quadratic optimization problem $f(x) = \frac{1}{2}x^T Ax + b^T x$ where $A$ is symmetric positive definite, the error $e_k = x_k - x^*$ satisfies

$$\|e_{k+1}\|_A =$$

- When sufficiently close to a local minima, general nonlinear optimization problems are described by such an SPD quadratic problem.
- Convergence rate depends on the conditioning of $A$, since
Gradient Methods with Extrapolation

- We can improve the constant in the linear rate of convergence of steepest descent by leveraging *extrapolation methods*, which consider two previous iterates (maintain *momentum* in the direction $x_k - x_{k-1}$):

- The *heavy ball method*, which uses constant $\alpha_k = \alpha$ and $\beta_k = \beta$, achieves better convergence than steepest descent:
The *conjugate gradient method* is capable of making the optimal (for a quadratic objective) choice of $\alpha_k$ and $\beta_k$ at each iteration of an extrapolation method:

- *Parallel tangents* implementation of the method proceeds as follows
Krylov Optimization

- Conjugate gradient (CG) finds the minimizer of \( f(x) = \frac{1}{2} x^T A x - b^T x \) (which satisfies optimality condition \( A x = b \)) within the Krylov subspace of \( A \):
The solution at the $k$th step, $y_k = \|b\|_2 T_k^{-1} e_1$ is obtained by CG from $y_{k+1}$ with a single matrix-vector product with $A$ and vector operations with $O(n)$ cost.
Preconditioning

- Convergence of iterative methods for $Ax = b$ depends on $\kappa(A)$, the goal of a preconditioner $M$ is to obtain $x$ by solving

$$M^{-1}Ax = M^{-1}b$$

with $\kappa(M^{-1}A) < \kappa(A)$

- Common preconditioners select parts of $A$ or perform inexact factorization
Conjugate Gradient Convergence Analysis

- In previous discussion, we assumed $K_n$ is invertible, which may not be the case if $A$ has $m < n$ distinct eigenvalues, however, in exact arithmetic CG converges in $m - 1$ iterations$^1$

$^1$This derivation follows *Applied Numerical Linear Algebra* by James Demmel, Section 6.6.4
Conjugate Gradient Convergence Analysis (II)

- Using $z = \rho_{k-1}(A)Ax$, we can simplify $\phi(z) = (x - z)^T A(x - z)$ as

- We can bound the objective based on the eigenvalues of $A = Q\Lambda Q^T$ using the identity $p(A) = Qp(\Lambda)Q^T$. 
Conjugate Gradient Convergence Analysis (III)

- Using our bound on the square of the residual norm $\phi(z)$, we can see why CG converges after $m - 1$ iterations if there are only $m < n$ distinct eigenvalues.

- To see that the residual goes to 0, we find a suitable polynomial in $Q_m$ (the set of polynomials $q_m$ of degree $m$ with $q_m(0) = 1$).
Round-off Error in Conjugate Gradient

- CG provides strong convergence guarantees for SPD matrices in exact arithmetic

- Due to round-off CG may stagnate / have plateaus in convergence
Graph and Matrix Duality

- graphs have a natural correspondence with sparse matrices

- matrix-based representations of graphs can be used to devise algorithms
Graph Partitioning from Eigenvectors

- The Laplacian matrix provides a model of interactions on a graph that is useful in many contexts

- The second-smallest-eigenvalue eigenvector of the Laplacian (the Fiedler vector), gives a good partitioning of the graph
Newton’s Method

- Newton’s method in $n$ dimensions is given by finding minima of $n$-dimensional quadratic approximation using the gradient and Hessian of $f$:
Nonlinear Least Squares

- An important special case of multidimensional optimization is *nonlinear least squares*, the problem of fitting a nonlinear function $f_x(t)$ so that $f_x(t_i) \approx y_i$:

- We can cast nonlinear least squares as an optimization problem to minimize residual error and solve it by Newton’s method:
Gauss-Newton Method

- The Hessian for nonlinear least squares problems has the form:

- The *Gauss-Newton* method is Newton iteration with an approximate Hessian:
Constrained Optimization Problems

We now return to the general case of constrained optimization problems:

$$\min_x f(x) \quad \text{subject to} \quad g(x) = 0 \quad \text{and} \quad h(x) \leq 0$$

Generally, we will seek to reduce constrained optimization problems to a series of simpler optimization problems:
Lagrangian Duality

- The Lagrangian function with constraints \( g(x) = 0 \) and \( h(x) \leq 0 \) is

\[
\mathcal{L}(x, \lambda) = f(x) + \lambda^T [g(x) - h(x)]
\]

- The Lagrangian dual problem is an unconstrained optimization problem:

\[
\max_{\lambda} q(\lambda), \quad q(\lambda) = \begin{cases} 
\min_x \mathcal{L}(x, \lambda) & \text{if } \lambda \geq 0 \\
-\infty & \text{otherwise}
\end{cases}
\]

The unconstrained optimality condition \( \nabla q(\lambda^*) = 0 \), implies

\[
\max \left( \lambda^*, \begin{bmatrix} h(x^*) \\ g(x^*) \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

\[
\mathcal{L}(x^*, \lambda^*) = f^* = q^*
\]

\[
f^* = q^*
\]

Strong duality
Optimality and Complementarity Slackness Condition

Consider the inequality-constrained optimization problem, \( h(x) \leq 0 \),

\[
\mathcal{L}(x, \lambda) = f(x) + \lambda^T h(x)
\]

▷ The pair \( x^* \) and \( \lambda^* \) are a primal-dual optimal solution if \( x^* \) is feasible, \( \lambda^* \geq 0 \), and strong duality holds, \( f(x^*) = \max \{q(\lambda^*)\} = \mathcal{L}(x^*, \lambda^*) \)

\[
\mathcal{L}(x^*, \lambda^*) = f(x^*) - \lambda^* h(x^*)
\]

\[\lambda_i = 0 \quad \text{or} \quad h_i(x^*) = 0\]
Sequential Quadratic Programming

- *Sequential quadratic programming (SQP)* reduces a nonlinear equality constrained problem to a sequence of constrained quadratic programs via a Taylor expansion of the Lagrangian function $\mathcal{L}_f(x, \lambda) = f(x) + \lambda^T g(x)$:

$$
\mathcal{L}_f(x, \lambda) = f(x) + \lambda^T g(x)
$$

- SQP ignores the constant term $\mathcal{L}_f(x_k, \lambda_k)$ and minimizes $s$ while treating $\delta$ as a Lagrange multiplier:
Barrier functions provide an effective way of working with inequality constraints $h(x) \leq 0$:

- Interior point methods additionally incorporate Lagrangian optimization
Karush-Kuhn-Tucker (KKT) conditions

Consider the linear-constrained Quadratic program (QP): Its Lagrangian

\[
f(x) = \frac{1}{2} x^T H x + x^T b \quad A x = b \quad C x \geq d
\]

function may be used to derive an interior point method. The first-order optimality (KKT) conditions are

\[
\nabla f(x) = \lambda \quad \lambda, u \geq 0
\]

\[
A x - b = 0 \quad C x - d \geq 0 \Rightarrow C x - d - s = 0
\]

\[
s \geq 0 \quad u^T (C x - d) = 0
\]
Primal-dual Interior Point Method (IPM)

Solve perturbed KKT conditions after introducing slack variables $s \in \mathbb{R}^{m_2}$

$\nabla f = 0 \mid s, \lambda, \nu \geq 0 \mid Ax - b = 0 \mid Cx - d - s = 0$

$m$ - barrier parameter

$s \cdot v_i = \mu <s, v_i> - \frac{1}{\dim(s)}$

$\mu \mu_i \to 0 \quad s^* \text{ for KKT}$

$\text{guess } s^0 \text{ for original problem}$

control path
Interior Point Method (IPM): KKT system

Newton’s method applied to KKT equations results in linear systems

\[
\frac{\partial f(x)}{\partial x} s_k = -\nabla f(x_k)
\]

Newton for minimizing \( f(x) \) = Newton for solving NLS with \( Df(x) = 0 \)

These linear systems become ill-conditioned as the interior point method approaches converges.
\[ f(x) = \frac{1}{2} x^T U x \]

\[ A x = b \]

\[ x \geq 0 \quad \Rightarrow \quad C x \geq 0 \]

\[ A \in \mathbb{R}^{m \times n} \quad u \in \mathbb{R}^{n \times n} \quad C \in \mathbb{R}^{n \times n} \]

with \( C = I \)