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 $\kappa(\boldsymbol{A})$ is the ratio between the max and min distance from the surface to the center of the unit ball (norm-1 vectors) transformed by $\boldsymbol{A}$:

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

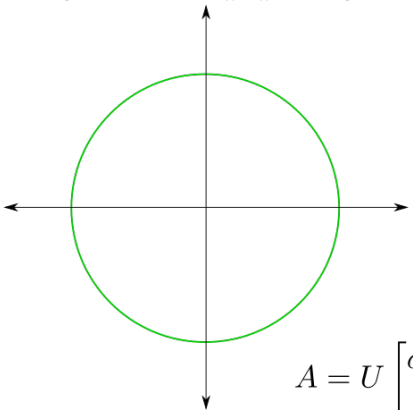
Singular Value Decomposition

- The singular value decomposition (SVD)

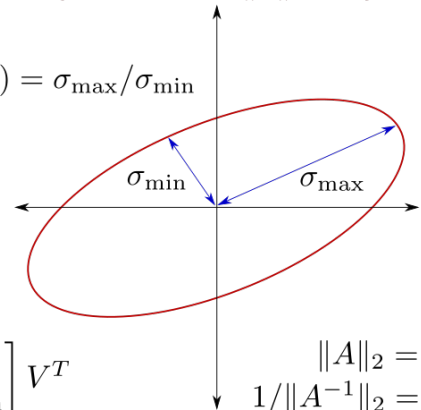
- Condition number in terms of singular values

Visualization of Matrix Conditioning

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



$$\kappa(A) = \sigma_{\max}/\sigma_{\min}$$



$$A = U \begin{bmatrix} \sigma_{\max} & \\ & \sigma_{\min} \end{bmatrix} V^T$$

$$\|A\|_2 = \sigma_{\max}$$
$$1/\|A^{-1}\|_2 = \sigma_{\min}$$

Linear Least Squares

- Find $\boldsymbol{x}^\star = \text{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2$ where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$:

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

Normal Equations

Demo: *Normal equations vs Pseudoinverse*
Demo: *Issues with the normal equations*

- *Normal equations* are given by solving $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$:

- However, solving the normal equations is a more ill-conditioned problem then the original least squares algorithm

Solving the Normal Equations

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

- Since $\boldsymbol{A}^T\boldsymbol{A}$ is SPD we can use Cholesky factorization, to factorize it and solve linear systems:

QR Factorization

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

- A reduced QR factorization (unique part of general QR) is defined so that $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ has orthonormal columns and $\boldsymbol{R}$ 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 $\boldsymbol{A}$ is diagonalizable, it has an *eigenvalue decomposition*

- $\boldsymbol{A}$ and $\boldsymbol{B}$ are *similar*, if there exist $\boldsymbol{Z}$ such that $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^{-1}$

Similarity of Matrices

matrix	*similarity*	*reduced form*
SPD		
real symmetric		
Hermitian		
normal		
real		
diagonalizable		
arbitrary		

Rayleigh Quotient

- For any vector x that is close to an eigenvector, the *Rayleigh quotient* provides an estimate of the associated eigenvalue of A:

Introduction to Krylov Subspace Methods

- *Krylov subspace methods* work with information contained in the $n \times k$ matrix

$$\boldsymbol{K}_k = \begin{bmatrix} \boldsymbol{x_0} & \boldsymbol{A x_0} & \cdots & \boldsymbol{A}^{k-1}\boldsymbol{x_0} \end{bmatrix}$$

- $\boldsymbol{A}$ is similar to *companion matrix* $\boldsymbol{C} = \boldsymbol{K}_n^{-1}\boldsymbol{A}\boldsymbol{K}_n$:

Krylov Subspaces

- Given $\boldsymbol{Q}_k\boldsymbol{R}_k = \boldsymbol{K}_k$, we obtain an orthonormal basis for the Krylov subspace,

$$\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{x}_0) = span(\boldsymbol{Q}_k) = \{p(\boldsymbol{A})\boldsymbol{x}_0 : deg(p) < k\},$$

  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:

Krylov Subspace Methods

- The $k \times k$ matrix $\boldsymbol{H}_k = \boldsymbol{Q}_k^T \boldsymbol{A} \boldsymbol{Q}_k$ minimizes $||\boldsymbol{A}\boldsymbol{Q}_k - \boldsymbol{Q}_k \boldsymbol{H}_k||_2$:

- $\boldsymbol{H}_k$ is upper-Hessenberg, because the companion matrix $\boldsymbol{C}_n$ is upper-Hessenberg:

Rayleigh-Ritz Procedure

- The eigenvalues/eigenvectors of $\boldsymbol{H}_k$ are the *Ritz values/vectors*:

- The Ritz vectors and values are the *ideal approximations* of the actual eigenvalues and eigenvectors based on only $\boldsymbol{H}_k$ and $\boldsymbol{Q}_k$:

Low Rank Matrix Approximation

- Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 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 $f(\boldsymbol{x})$

- Quadratic optimization $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$

Basic Multidimensional Optimization Methods

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

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

  $$||\boldsymbol{e}_{k+1}||_{\boldsymbol{A}} =$$

  - 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 $\boldsymbol{A}$, 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 $\boldsymbol{x}_k - \boldsymbol{x}_{k-1}$):

- The *heavy ball method*, which uses constant $\alpha_k = \alpha$ and $\beta_k = \beta$, achieves better convergence than steepest descent:

Conjugate Gradient Method

- The *conjugate gradient method* is capable of making the optimal (for a quadratic objective) choice of α_k and β_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 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$ (which satisfies optimality condition $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$) within the Krylov subspace of $\boldsymbol{A}$:

CG and Krylov Optimization

The solution at the kth step, $\boldsymbol{y}_k = ||\boldsymbol{b}||_2 \boldsymbol{T}_k^{-1} \boldsymbol{e}_1$ is obtained by CG from $\boldsymbol{y}_{k+1}$ with a single matrix-vector product with $\boldsymbol{A}$ and vector operations with $O(n)$ cost

CG for QPs / lin-sys

$f(x) = \frac{1}{2} x^T A x - x^T b$

$\nabla f(x) = 0 \Rightarrow Ax = b$

x_k $r = Ax - b$

$\text{span}\{b, Ab, A^2 b, \dots A^{k-1} b\}$

Preconditioning

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

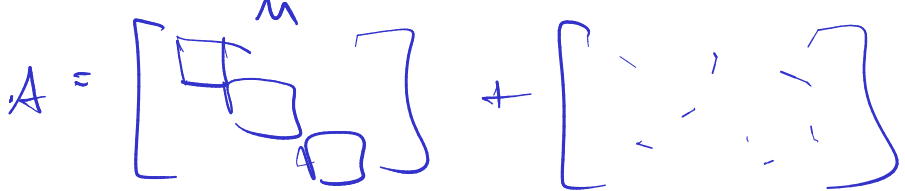
$$\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{b}$$

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

→ goal: reduce # of CG iterations to achieve accuracy

$$\|e_{k+1}\|_A = \|e_0\| \frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)+1}}$$

- Common preconditioners select parts of $\boldsymbol{A}$ or perform inexact factorization



$$A = X D X^{-1}$$

$$A A = X D X^{-1} X D X^{-1} = X D^2 X^{-1}$$

$$A^k = X D^k X^{-1}$$

$$p(A) = I + 2A - A^4 = X p(D) X^{-1}$$

$$p(x) = 1 + 2x - x^4$$

$$\mathcal{K}_{k-1}(A, b) = \operatorname{span}\{b, Ab, \dots A^{k-1} b\} = \{p_{k-1}(A) b : p_{k-1} \in \Pi_{k-1}$$

Conjugate Gradient Convergence Analysis

- In previous discussion, we assumed $\boldsymbol{K}_n$ is invertible, which may not be the case if $\boldsymbol{A}$ 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

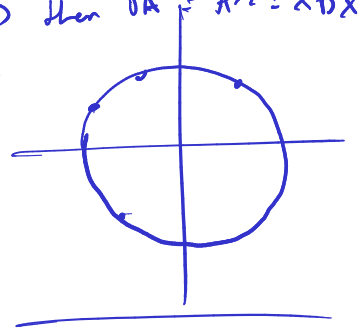
$\|x\|_A^2 = x^T A x$

A spd

$\|x\|_A^2 = x^T \sqrt{A}\sqrt{A} x$

A is SPD then $\sqrt{A} = A^{1/2} = X D^{1/2} X^{-1}$

2-norm

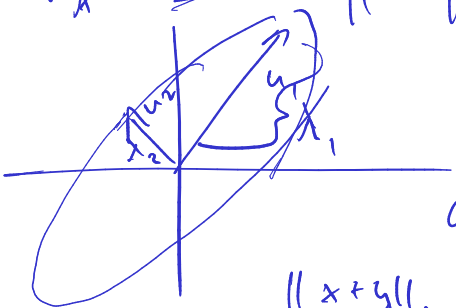


$v Q Q^T v = v^T v$

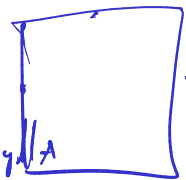
$\|v\| = \|\ \|_2$

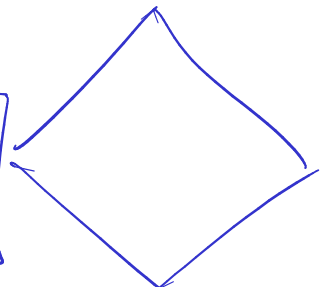
QQ^T

1-norm



∞-norm





$\|x+y\|_A \leq \|x\|_A + \|y\|_A$

Conjugate Gradient Convergence Analysis (II)

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

$$e(z) = \left(x - \rho_{k-1}(A)Ax\right)^T A(\ldots)$$

$$= \left(\left(I - \rho_{k-1}(A)A\right)x\right)^T A(\ldots)$$

$q_k(A)$

$q_k(\alpha) = 1 - \rho_{k-1}(\alpha)\alpha$, $q_k(0) = 1$

- We can bound the objective based on the eigenvalues of $A = Q\Lambda Q^T$ using the identity $p(A) = Qp(\Lambda)Q^T$,

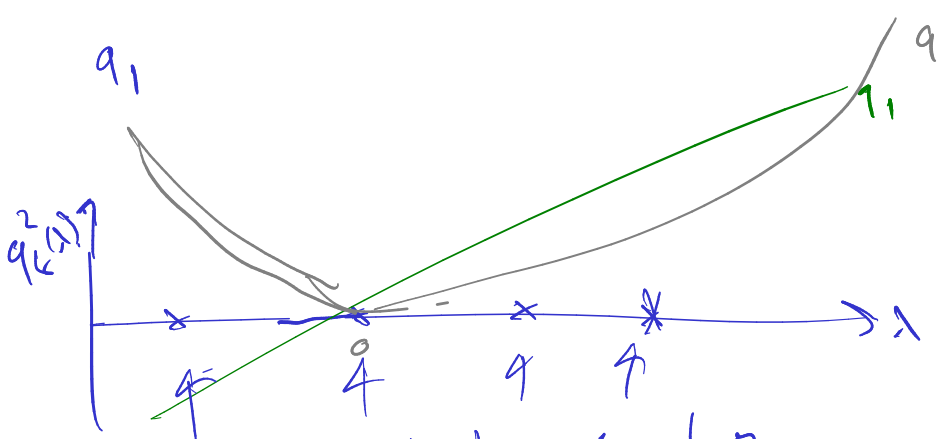
$$e(z) = x^T \left(I - \rho_{k-1}(A)A\right) A$$

$$e(z) = x^T Q\, q_k(\Lambda)\, \Lambda\, q_k(\Lambda) Q^T x$$

$$\leq \max_{\lambda_i \in \lambda(A)} \left(q_k(\lambda_i)\right)^2 x^T Q \Lambda Q^T x$$

A

$e(x_0)$



then $\exists\ q_m \in \Pi_m$ with 0s at
$q_m(\lambda) = \prod_{j=1}^{m}(\lambda_j - \lambda) \Big/ \prod_{j=1}^{m} \lambda_j \quad | \quad \lambda_1 \dots \lambda_m$

Conjugate Gradient Convergence Analysis (III)

- Using our bound on the square of the residual norm $\phi(\boldsymbol{z})$, 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 $\mathcal{Q}_m$ (the set of polynomials q_m of degree m with $q_m(0) = 1$)

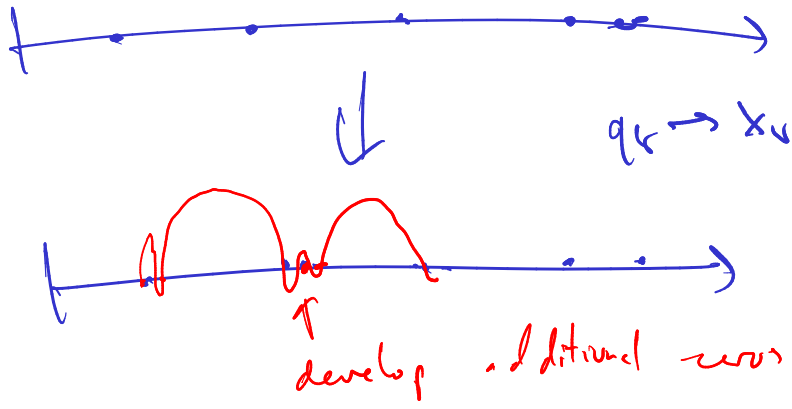
Round-off Error in Conjugate Gradient

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

  - classic view of CG (as direct) direct
  - due to round-off, not so exact
  - robust as an iterative methods

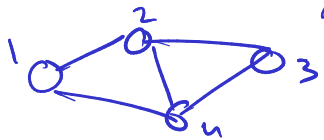
- Due to round-off CG may stagnate / have plateaus in convergence

  - as CG on a larger matrix, whose eigenvalues are clustered around those of A

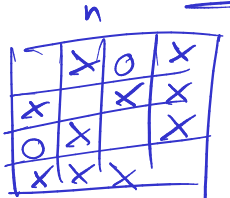

$q_k \to x_v$
develop additional errors

Graph and Matrix Duality

- graphs have have a natural correspondence with sparse matrices



adjacency



$G = (V, E)$ $|V| = n$

$w(u, v)$

- matrix-based representations of graphs can be used to devise algorithms

- combinatorial algorithms (BFS, SSSP BF)
- semirings $(\min, +)$ $(+, \cdot)$

$w = A \otimes v \Rightarrow w_i = \min_{j \in N(i)} a_{ij} + v_j$

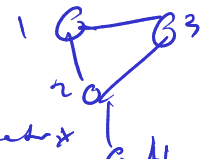
- $D = I \oplus A \oplus (A \otimes A) \oplus \dots \oplus (A^{\otimes n}) \Rightarrow d_{ij} = \min_{k \in \{1,\dots,n\}} \min_{u_1,\dots,u_{k-1}} w(i, u_1) + w(u_1, u_2) + \dots$

Graph Partitioning from Eigenvectors

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

L = Laplacian = $D - A$

(D: diagonal degree matrix; A: adjacency matrix)



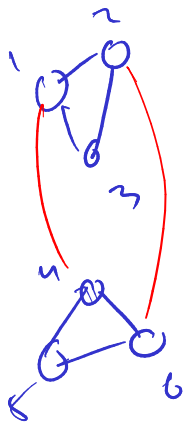
$$\begin{bmatrix} 2 & -1 & -1 & \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & \\ & -1 & & 1 \end{bmatrix}$$

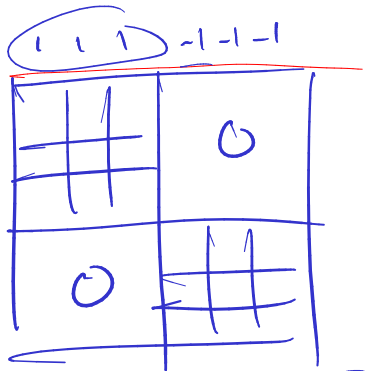
- The second-smallest-eigenvalue eigenvector of the Laplacian (the Fiedler vector), gives a good partitioning of the graph

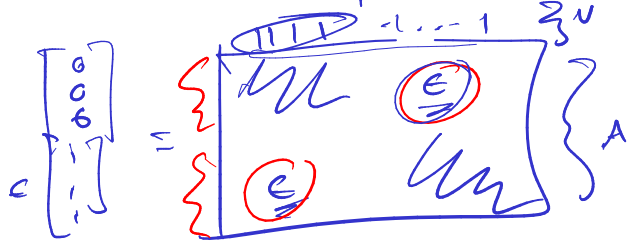
$$v = L\begin{bmatrix}1\\1\\1\\1\end{bmatrix} = D\begin{bmatrix}1\\1\\1\\1\end{bmatrix} - A\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

$v_i = d_i - d_i = 0$

L has at least 1 0 eigenvalue







$$\rho_A(v) = \frac{v^T A v}{v^T v} \approx \epsilon$$

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 $f_{\boldsymbol{x}}(t)$ so that $f_{\boldsymbol{x}}(t_i) \approx y_i$:

- 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:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{g}(\boldsymbol{x}) = \mathbf{0} \quad \text{and} \quad \boldsymbol{h}(\boldsymbol{x}) \leqslant \mathbf{0}$$

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

Lagrangian Duality

- The Lagrangian function with constraints $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{h}(\boldsymbol{x}) \leqslant \mathbf{0}$ is

- The Lagrangian dual problem is an unconstrained optimization problem:

$$\max_{\boldsymbol{\lambda}} q(\boldsymbol{\lambda}), \quad q(\boldsymbol{\lambda}) = \begin{cases} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) & \text{if } \boldsymbol{\lambda} \geqslant \mathbf{0} \\ -\infty & \text{otherwise} \end{cases}$$

  The unconstrained optimality condition $\nabla q(\boldsymbol{\lambda}^*) = \mathbf{0}$, implies

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

- SQP ignores the constant term $\mathcal{L}_f(\boldsymbol{x}_k, \boldsymbol{\lambda}_k)$ and minimizes $\boldsymbol{s}$ while treating $\boldsymbol{\delta}$ as a Lagrange multiplier:

Interior Point Methods

- Barrier functions provide an effective way of working with inequality constraints $\boldsymbol{h}(\boldsymbol{x}) \leqslant \mathbf{0}$:

- Interior point methods additionally incorporate Lagrangian optimization