LECTURE 1

NUMERICAL OPTIMIZATION

Jan 5th, 2026

- Textbook: Nocedal & Wright (Numerical Optimization 2nd Ed.)

* NONLINEAR PROGRAMING:

$$\begin{cases} \min f(x) \\ \text{s.t. } c(x)=0,\ h(x)\le 0 \end{cases} \quad \text{or} \quad \begin{cases} \min f(x) \\ \text{s.t. } c(x)=0,\ h(x)+y=0 \end{cases}$$

$y \ge 0$ ← SLACK variables

$$\text{or} \quad \begin{cases} \min f_1(z) \\ \text{s.t. } c_1(z)=0 \\ x \in \Omega = \mathbb{R}^n \times \mathbb{R}^m \end{cases}$$

REMARK: If the C^r fcts f, g, h, c are nonlinear ⇒ NONLINEAR PROBLEM

EXAMPLE: LAGRANGIAN APPROACH

$$\min_{\substack{x,y;\ c(x)=0 \\ h(x)+y=0}} f(x) \Rightarrow \min_{x,y\ge 0} f(x) + \lambda^T c(x) + \nu^T\left[h(x)+y\right] + c\,\|c(x)\|^2 + c\,\|h(x)+y\|^2$$

REMARK:

- Nonlinear program w/ ineq. constraints
- Nonlinear program w/ bound constraints (including "=")

↳ Bad idea if convex

- Minimization w/ bound constraints ⟶ Lagrange Theory
- Unconstrained minimization ⟶ Gradient Projection

Linear Algebra ⟶ Newton's Method

1-dim Optimization ⟶ Line Search

---

* NEWTON'S METHOD

- GOAL: Approximate the problem w/ the quadratic version at the current iterate (there are also constrained versions).

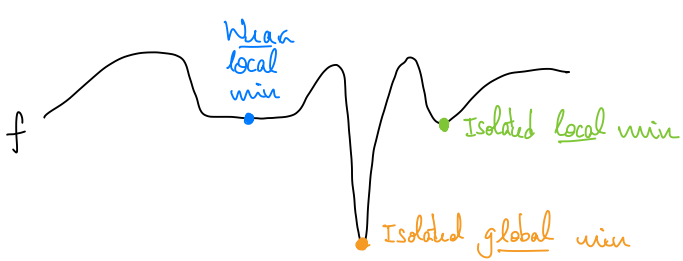
$$\min_x f(x) \Rightarrow \min_d f(x+d) \approx \min_d f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2_{xx} f(x)\, d$$

$$\Rightarrow \boxed{x^+ := x - \left[\nabla^2_{xx} f(x)\right]^{-1} \nabla f(x)}$$

THEN: Solve the current quadratic approximation and keep moving w/out blowing up.

SOLUTIONS TO MIN / MAX PROBLEMS

separated by a minus sign



OBS: The type of minima depends on the starting point.

DEF: A point x_* is a global minimizer iff $f(x_*) \leq f(x) \ \forall x$.

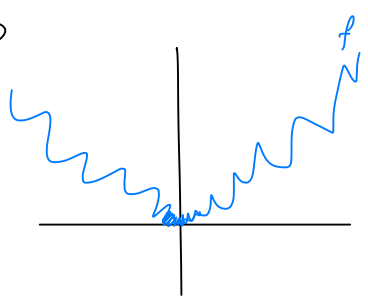
DEF: A point x_* is a local minimizer if $\exists N \ni x_*$ s.t.
$f(x_*) \leq f(x) \quad \forall x \in N \setminus \{x_*\}$.

DEF: A point x_* is a strict local minimizer (or strong local minimizer) if $\exists N \ni x_*$ s.t. $f(x_*) < f(x) \ \forall x \in N \setminus \{x_*\}$.

DEF: A point x_* is an isolated local minimizer if $\exists N \ni x_*$ s.t. x_* is the only local minimizer in N.

EXAMPLE: $f(x) = x^4 \cos\left(\frac{1}{x}\right) + 2x^4$, $f(0) = 0$

↳ $f \in C^3(\mathbb{R}, \mathbb{R}_+)$



All isolated local minimizers are strict.

Non-isolated local minimizer at zero.

LECTURE 2

OPTIMALITY CONDITIONS

Jan 7th 2026

THM: (1st order Necessary Condition) If x_* is a local minimizer of $f \in C^1$ on an open nhbd of x_*, then $\nabla f(x_*) = 0$.

IMPORTANT

Pf: $f: \mathbb{R} \to \mathbb{R}$, x_* is local min.

$$f'(x_*) = \lim_{x_n \nearrow x_*} \frac{f(x_n) - f(x_*)}{x_n - x_*} \leq 0$$

$f(x_n) - f(x_*) \geq 0$, $x_n - x_* < 0$

$$f'(x_*) = \lim_{x_n \searrow x_*} \frac{f(x_n) - f(x_*)}{x_n - x_*} \geq 0$$

$\Rightarrow f'(x_*) = 0$.

Note: Taylor series

$$f''(x_*) = \lim_{h \to 0} \frac{f(x_* + h) + f(x_* - h) - 2f(x_*)}{h^2}$$

If h is small enough, then $f''(x_*) \geq 0$ b/c x_* is a min. and $x_* + h$, $x_* - h$ are in a nhbd of x_*.

For $f: \mathbb{R}^n \to \mathbb{R}$. Look at one line at the time through x_*

$\Leftrightarrow$ Look at $g(t) = f(x_* + td)$ $\forall t, d \in \mathbb{R}^n$

$\Rightarrow 0$ is local min. for g $\forall d \in \mathbb{R}^n$.

Chain Rule: $\frac{d}{dt} g(t) = \nabla f(x_* + td)^T d$

$$\frac{d^2}{dt^2} g(t) = d^T \nabla^2_{xx} f(x_* + td)\, d$$

$\Rightarrow$ 1st optimality condition:

$$g'(0) = 0 \iff \nabla f(x_*)^T d = 0 \quad \forall d \in \mathbb{R}^n$$

$$g''(0) \geq 0 \iff d^T \nabla^2_{xx} f(x_*) d \geq 0 \quad \forall d \in \mathbb{R}^n$$

Therefore: If x_* is a local minimum, then

$$\nabla f(x_*) d = 0 \quad \forall d \in \mathbb{R}^n \iff \nabla f(x_*) = 0$$

$$d^T \nabla^2_{xx} f(x_*) d \geq 0 \quad \forall d \in \mathbb{R}^n \iff \nabla^2_{xx} f(x_*) \geq 0$$

Also proved ✓ ■

THM: (2nd order Necessary Condition) If x_* is a local minimizer of f and $\nabla^2 f$ exists and is continuous in an open nhbd of x_*, then $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) \geq 0$.

THM: (2nd order Sufficient Conditions) Suppose $\nabla^2 f$ is continuous in an open nhbd of x_* and $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) > 0$. Then x_* is a strict local minimizer of f.

THM: When f is convex, any local minimizer x_* is a global minimizer of f. If, in addition, f is differentiable, then any stationary point x_* is a global minimizer of f.

THM: (Taylor) $f \in C^1(\mathbb{R}^n, \mathbb{R})$ and $p \in \mathbb{R}^n$. Then,

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad \text{(Intermediate Value Thm)}$$

for some $t \in (0,1)$. If $f \in C^2(\mathbb{R}^n, \mathbb{R})$ we have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\, p \, dt, \quad \text{(Fund. Thm. Calc.)}$$

and

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp)\, p, \quad \text{(Intermediate value Thm)}$$

for some $t \in (0,1)$.

LOCAL OPTIMALITY CONDITIONS $\min_x f(x)$, $f \in C^2$.

- First order necessary condition: $\nabla f = 0$
- Second order necessary condition: $\nabla^2_{xx} f(x) \succeq 0$
- Second order sufficient condition: $\nabla^2_{xx} f(x) \succ 0$

(prove by looking at every line through origin)

REMARK: Global optimality is NP hard.

EXCEPTION: f convex $\rightsquigarrow \nabla^2_{xx} f(x) \succeq 0 \quad \forall x \in \mathbb{R}^n$.

UPSHOT: If $f \in C^2$, then solving for a local sol^n is getting close to solving a differentiable system of eqs:

$$x = \operatorname{argmin} f \iff \nabla f(x) = 0.$$

For such situations we have Newton's method and we replace an analytical problem with a linear algebra problem
— However, we cannot hope to get global min. this way in general.

RATES OF CONVERGENCE

(In general, cannot have only finitely many # of operations. So, only hope is to produce a seq. that $x_k \to x_*$)

- **Q-CONVERGENCE** (Quotient Convergence):

1) Q-LINEAR $\exists r \in [0,1)$ $\dfrac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \leq r$ $\forall k$ suff. large

e.g.: $r^k = x_k - x_*$, $r < 1$
$\Rightarrow \|r^{k+1}\| / \|r^k\| \leq r$

2) Q-SUPERLINEAR $\lim\limits_{k\to\infty} \dfrac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0$

e.g.: $r^{k^2} = x_k - x_*$, $r < 1$.

3) Q-QUADRATIC $\dfrac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^2} \leq M$ $\forall k$ suff. large and $M > 0$

e.g.: $r^{2^k} = x_k - x_*$, $r < 1$.

- **R-CONVERGENCE** (Root Convergence):

- The distance to sol^n is dominated by Q-convergent seq. $\forall k$

$\|x_k - x_*\| \leq v_k$, where $\{v_k\}$ converges in Q-(order)

where (order) is either linear, superlinear, quadratic.

- The origin of the name "root" convergence has to do with: for linear, it implies

$$\limsup_{k \to \infty} \sqrt[k]{\|x_k - x_*\|} \leq r < 1 .$$

GOALS: Unconstrained Optimization

- Derive efficient algorithms to solve the problem

$$\min_x f(x), \quad f \in C^2,$$

and fast.

- Solve #1 ≡ guarantee convergence to a point that satisfies the 1st order NECESSARY conditions.

  Solve #2 ≡ guarantee convergence to a point that satisfies the 2nd order NECESSARY conditions.

- Fast #1 ≡ Typically, if point also SUFFICIENT, then local convergence should be Newton-like (e.g., quadratic or superlinear).

  Fast #2 ≡ Make sure the linear algebra is efficient and fits with the optimization (e.g.: solving for Newton directions results in DESCENT).

Overview of Algorithms

- **1st order methods:** – Compute gradients
  – Those are descent directions for the function, go along them to make progress.
- **2nd order methods:** – Compute Hessians
  – Rely on fact that certain subproblems are tractable

Subproblems: linear systems of eqs.; e.g. Newton's direction

$$\nabla^2_{xx} f(x)\, d + \nabla f(x) = 0$$

– If the matrix is positive-definite, d is the descent direction.

– Quadratic optimization problems w/ ONE quadratic constraint

e.g.:

$$\min \; f(x) + \nabla f(x)^T d + \frac{1}{2} \nabla^2_{xx} f(x)\, d$$
$$\text{s.t.} \quad d^T d \le \rho^2$$

Line Search vs. Trust Region

(Both iterative)

- **Line search:** – Find a "good" direction p_k
  – Need to make sure f decreases along p_k (at least for a while)
  – Solve approximately the 1-dim. min. problem

$$\min_{\alpha > 0} f(x_k + \alpha p_k)$$

- TRUST REGION:

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p$$

  - Construct a quadratic model of the problem
  - Minimize the model subject to a trust-region constraint (most commonly of spherical or elliptical shape).
  - Manage trust region.

REMARK:
- For the trust-region, the "search" direction depends on the size of the trust-region.
- Particularly pronounced is the situation of far from spherical level sets and model contours.

LINE SEARCH: STEEPEST DESCENT

- Taylor series: for some $t \in (0, \alpha)$

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \frac{1}{2} \alpha^2 p^T \nabla^2_{xx} f(x_k + tp)\, p$$

- The fastest immediate descent direction is the sol^{n} of the problem

$$\min_p \; p^T \nabla f_k \quad \text{s.t.} \quad \|p\| = 1$$

- Of course, that is proportional to the negative gradient and is called STEEPEST DESCENT DIRECTION.

PLUSES: - Need only 1st derivative information

MINUSES: - Horribly slow (zig/zag effect)
- Q-linear convergence

LECTURE 3

Jan 12th 2026 (Ch 3)

LINE SEARCH METHODS

LINE SEARCH IDEA: At the current pt. x_k find direction d_k to do 1-dim. minimization

$$\Leftrightarrow \quad x_{k+1} = \operatorname{argmin}_{\alpha} f(x_k + \alpha d_k)$$

REMARK: Because line search always decreases f, we will have an accumulation point (cannot diverge if bdd below) — unlike Newton's method!

* NEWTON DIRECTION

- Approximate f at x_k w/ quadratic model with positive-definite Hessian

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p =: m_k(p).$$

- Its minimizer is the Newton direction and it's a descent direction:

$$p_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k.$$

- PROS: Extremely fast (Q-superlinear) when close enough to sol^{n}.

- MINUSES: Needs 2nd derivatives

- Obs: When not doing lin-search, called (NEWTON'S METHOD)

* SECANT CONDITION: QUASI-NEWTON METHOD

- Do Taylor series of the gradient

$$\nabla f_{k+1} = \nabla f_k + \nabla^2 f_k (x_{k+1} - x_k) + o(\|x_{k+1} - x_k\|)$$

- The Hessian has an approximate secant condition

$$\nabla^2 f_k (x_{k+1} - x_k) \approx \nabla f_{k+1} - \nabla f_k .$$

$$\Rightarrow B_{k+1} s_k = y_k ,$$

$$s_k = x_{k+1} - x_k$$

$$y_k = \nabla f_{k+1} - \nabla f_k$$

} Quasi-Newton Methods

- Descent directions:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} , \qquad \boxed{p_k = -B_k^{-1} \nabla f_k}$$

- PLUSES: No 2nd derivatives and Q-superlinear convergence
- MINUSES: Storage heavy

REMARK: All descent methods get stuck at saddle pts.
⇝ Trust region.
↑ In principle, the only way to converge to 2nd order stationary pts.

LINE SEARCH METHODS

- Each iteration of line search computes a search direction p_k and then decides how far to move along that direction.
- Each iteration is given by

$$x_{k+1} = x_k + \alpha_k p_k$$

$\alpha_k > 0$ STEP LENGTH

- Most line search algorithms require p_k to be a descent direction (i.e., $p_k^T \nabla f_k < 0$). Often of the form $p_k = -B_k^{-1} \nabla f_k$, where B_k symmetric nonsingular.
- E.g.: - Steepest descent $\longrightarrow B_k = \mathbb{1}$

- Newton's method $\longrightarrow$ $B_k = \nabla^2 f(x_k)$

- Quasi-Newton $\longrightarrow$ B_k is approx. to Hess by low-rank matrices.

> REMARK: When $p_k = -B_k^{-1} \nabla f_k$,
>
> $$p_k^T \nabla f_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0 \implies p_k \text{ is descent direction.}$$

* STEP LENGTH

- Ideal choice: minimizer of $\phi(\alpha) = f(x_k + \alpha p_k)$, $\alpha > 0$
  - ↳ Want to reduce f but not spend too much time going in the same direction

- Usually the most expensive part of the computation (usually requires a lot of evals. of f and ∇f).

- Simple necessary condition to ensure α_k gives reduction in f:

$$f(x_k + \alpha_k p_k) < f(x_k).$$

- Popular sufficient condition: Wolfe conditions

↳ Useful for proofs and analyses (hard to implement)

WOLFE CONDITIONS: Inexact search condition to ensure a sufficient condition for α_k to decrease the objective f:

1) For some $c_1 \in (0,1)$ (usually $\sim 10^{-4}$),

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k$$

(Armijo Condition)

i.e.: reduction in f should be proportional to both the step length α_k and the directional derivative $\nabla f_k^T p_k$.

2) To rule out unacceptably short steps, we introduce the second requirement: α_k must be such that

$\phi''(\alpha_k)$ → $\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k$, (Curvature Condition)

for some constant $c_2 \in (c_1, 1)$.

– Strong Wolfe Condition: α_k must satisfy

- $f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k$
- $|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq c_2 |\nabla f_k^T p_k|$

for $0 < c_1 < c_2 < 1$.

REMARK: • Curvature condition very hard to implement.

• Can go around this difficulty in the curvature condition by BACKTRACKING:

ALGORITHM: (BACKTRACKING)

Choose $\tilde{\alpha} > 0$, $\rho \in (0,1)$, $c \in (0,1)$

Set $\alpha \leftarrow \tilde{\alpha}$

Repeat until $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k$

$\alpha \leftarrow \rho \alpha$;

end (repeat)

Terminate with $\alpha_k = \alpha$.

- Obs: ρ can also vary at each iteration $\rho \in [\rho_{low}, \rho_{high}]$.
- This ensures the step size is optimaly long enough.

CONVERGENCE OF LINE SEARCH METHODS

THM: (Zoutendijk) Consider any line-search iteration $x_{k+1} = x_k + \alpha_k p_k$ where p_k is a descent direction and α_k satisfies the Wolfe conditions. Suppose f is bounded below in $\mathbb{R}^n$ and f is C^1 in an open set N containing the level set $L := \{x : f(x) \leq f(x_0)\}$, where x_0 is the starting pt. of the iteration. Assume ∇f is Lipschitz on N. Then,

(Zoutendijk Condition)

$$\sum_{k \geq 0} \cos^2 \theta_k \, \|\nabla f_k\|^2 = \sum_{k \geq 0} \left(\frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \, \|p_k\|} \right) \|\nabla f_k\|^2 < +\infty .$$

θ_k: Angle between p_k and steepest descent $-\nabla f_k$

- Zoutendijk shows, e.g., that steepest descent is globally convergent; and how p_k can deviate from steepest descent and still be globally convergent.

⚠ IMPORTANT

- Zoutendijk's condition implies that

$$\cos^2\theta_k \|\nabla f_k\|^2 \longrightarrow 0. \quad (*)$$

This can, in turn, be used to show global convergence for search algorithms.

- If our method for choosing p_k in $x_{k+1} = x_k + \alpha_k p_k$ ensures that $\theta_k := \arccos\left(-\nabla f_k^T p_k / \|\nabla f_k\| \|p_k\|\right)$ is bounded away from 90°, $\exists \delta > 0$ s.t.

$$\cos\theta_k \geq \delta > 0 \quad \forall k.$$

$$(*) \Longrightarrow \lim_{k\to\infty} \|\nabla f_k\| = 0.$$

Upshot: $\{\|\nabla f_k\|\}$ converge to zero provided that the search directions are never too close to orthogonality w/ the grad.. Moreover, steepest descent gives a seq. of grads that converges to zero, provided it uses a line search satisfying Wolfe or Goldstein (see 3.2 of Nocedal & Wright)

EXAMPLE: NEWTON-LIKE METHOD $\begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ p_k = -B_k^{-1} \nabla f_k \end{cases}$

Assume $\kappa(B_k) = \|B_k\| \|B_k^{-1}\| \leq M \quad \forall k$

$\Rightarrow \cos\theta_k \geq \frac{1}{M}$

$\Rightarrow \lim_{k \to \infty} \|\nabla f_k\| = 0$.

$\Rightarrow$ Newton and quasi-Newton are globally convergent provided $\{B_k\}$ have bounded condition #, are positive definite, and the step lengths satisfy the Wolfe conditions.

LECTURE 4

(Ch 3)

Jan 14 2026

CONVERGENCE OF LINE SEARCH & TRUST REGIONS

RATE OF CONVERGENCE

1) CONVERGENCE RATE OF STEEPEST DESCENT (EXACT LINE SEARCH)

- If objective fct is quadratic: $f(x) = \frac{1}{2} x^T Q x - b^T x$ ($Q^T = Q > 0$)

  $\Rightarrow \nabla f(x) = Qx - b$

- Line search

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}, \quad x_{k+1} = x_k - \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k$$

- In the Mahalanobis norm $\|x\|_Q^2 = x^T Q x$,

$$\|x_{k+1} - x_*\|_Q^2 = \left[1 - \frac{\left(\nabla f_k^T \nabla f_k\right)^2}{\left(\nabla f_k^T Q \nabla f_k\right)\left(\nabla f_k^T Q^{-1} \nabla f_k\right)} \right] \|x_k - x_*\|_Q^2$$

$Q = Q^{1/2} Q^{1/2}$
$Q = U \Lambda U^T$

Cauchy-Schwartz w/ $u = Q^{1/2} \nabla f_k$, $v = Q^{-1/2} \nabla f_k$

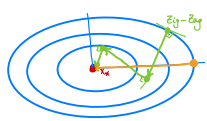
THM: When steepest descent with exact line search above is applied to the quadratic objective,

$$\|x_{k+1} - x_*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 \|x_k - x_*\|_Q^2$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are eigenvals. of Q.

REMARKS: • The ratio is $< 1 \Rightarrow$ **linear convergence**

- The ratio gets worse as Q gets more ill-conditioned; i.e., the ratio gets closer to 1.
  - ↳ $\kappa(Q) = \frac{\lambda_n}{\lambda_1}$ increases $\Rightarrow$ zig-zag



2) CONVERGENCE OF NEWTON'S METHOD

- line search with $p_k^N = -\left[\nabla^2 f_k\right]^{-1} \nabla f_k$.
- Since Hessian not always $\succ 0$, p_k^N may not always be a descent direction. $\Rightarrow$ Newton is locally convergent.
- Can modify Hess to make it $\succ 0$ in a trust region, if necessary, and make p_k^N a descent ↳ Modified Newton

THM: Suppose f is C^2 and $\nabla^2 f(x)$ is Lipschitz in a nhbd of a sol^n x_* at which the sufficient conditions are satisfied. Consider the iteration $x_{k+1} = x_k + p_k^N$, where $p_k^N = -[\nabla^2 f_k]^{-1} \nabla f_k$. Then

(i) If the starting pt. x_0 is suff. close to x_*, then $x_k \to x_*$.

(ii) The rate of convergence of $\{x_k\}$ is quadratic

(iii) The seq. of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.

- Lipschitz weaker than C^3: $\|\nabla^2 f(x_2) - \nabla^2 f(x_1)\| \leq L\|x_2 - x_1\|$

Pf: $\nabla^2 f(x_*) \succ 0 \Rightarrow \nabla^2 f(x) \succ 0 \quad \forall x \in B_r(x_*)$ by continuity

$$x_{k+1} - x_* = x_k - x_* - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

$$\Rightarrow x_{k+1} - x_* = \nabla^2 f(x_k)^{-1} \left[\nabla^2 f(x_k)(x_k - x_*) - \nabla f(x_k) \right]$$

But $\nabla f(x_*) = 0$ from optimality condition.

$$\Rightarrow x_{k+1} - x_* = \nabla^2 f(x_k)^{-1} \underbrace{\left[\nabla^2 f(x_k)(x_k - x_*) - \nabla f(x_k) + \nabla f(x_*) \right]}_{=: r(x_k)}$$

Fundamental Thm of Calculus: $\nabla f(x_* + t(x_k - x_*)) =: g(t)$ for some $t \in [0,1] \rightsquigarrow g: \mathbb{R} \to \mathbb{R}^n$.

$$g(1) - g(0) = \int_0^1 g'(z)\, dz$$

$$\overset{CR}{\iff} \nabla f(x_k) - \nabla f(x_*) = \int_0^1 \nabla^2 f\big(x_* + z(x_k - x_*)\big)(x_k - x_*)\, dz$$

$$\Rightarrow r(x_k) = \nabla^2 f(x_k)(x_k - x_*) - \int_0^1 \nabla^2 f\big(x_* + z(x_k - x_*)\big)(x_k - x_*)\, dz$$

$$= \int_0^1 \Big[\nabla^2 f(x_*) - \nabla^2 f\big(x_* + z(x_k - x_*)\big)\Big](x_k - x_*)\, dz$$

$$\Rightarrow \|r(x_k)\| \le \int_0^1 \big\|\nabla^2 f(x_*) - \nabla^2 f\big(x_* + z(x_k - x_*)\big)\big\| \, \|x_k - x_*\|\, dz$$

Lipschitz $\hookrightarrow$ $$\le \|x_k - x_*\|^2 \int_0^1 \gamma z(1 - z)\, dz$$

$$= \frac{1}{2} L \|x_k - x_*\|^2 .$$

Since $\nabla^2 f(x_*)$ is nonsingular $\exists r > 0$ s.t. $\|\nabla^2 f_k^{-1}\| \le 2\|\nabla^2 f(x_*)^{-1}\|$ $\forall x \in B_r(x_*)$. Then

$$\|\underbrace{x_k + p_k^N}_{= x_{k+1}} - x_*\| \le \underbrace{L\|\nabla^2 f(x_*)^{-1}\|}_{\tilde{L}} \|x_k - x_*\|^2$$

$$= \tilde{L}\, \|x_k - x_*\|^2 .$$

If $L\|\nabla^2 f(x_*)^{-1}\|\, \|x_k - x_*\| = \tilde{L}\|x_k - x_*\| \le \frac{1}{2}$

Proves (i) & (ii) at the same time

$$\Rightarrow L\|\nabla^2 f(x_*)^{-1}\|\, \|x_{k+1} - x_*\| \le \tilde{L}^2 \|x_k - x_*\|^2 \le \frac{1}{4}$$

$$\tilde{L}\, \|x_{k+1} - x_*\|$$

(i) $x_k \to x_*$ is at least q-linear.

(ii) $$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^2} \leq L \|\nabla^2 f(x_*)^{-1}\| \implies \text{q-quadratic}.$$

(iii) ⚠ SEE PROOF IN PG 95 OF NOCEDAL & WRIGHT ■

REMARK: • As Newton iterates approach the sol^n x_*, the Wolfe conditions (or Goldstein) will accept step length $\alpha_k = 1$ for all large k.

3) CONVERGENCE OF QUASI-NEWTON

- Quasi-Newton $\rightsquigarrow$ search direction $\equiv p_k = -B_k^{-1} \nabla f_k$ where $B_k^T = B_k > 0$ $\forall k$ is updated according to quasi-Newton.
- Assume α_k is inexact line search and satisfies Wolfe.

THM: Suppose $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and consider $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies the Wolfe conditions with $c_1 \leq 1/2$. If the seq. $\{x_k\}$ converges to a point x_* s.t. $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) \succ 0$ and if the search direction satisfies

$$\lim_{k \to \infty} \frac{\| \nabla f_k + \nabla^2 f_k p_k \|}{\| p_k \|} = 0 .$$

Then

(i) The step length $\alpha_k = 1$ is admissible for all k greater than some k_0.

(ii) If $\alpha_k = 1 \ \forall k > k_0$, then $x_k \longrightarrow x_*$ superlinearly.

Note: if $c_1 > \frac{1}{2}$ then line search can exclude the minimizer of a quadratic.

$\Rightarrow$ unit steps may not be admissible.

- p_k quasi-Newton $\Rightarrow \lim_{k \to \infty} \frac{\| \nabla f_k + \nabla^2 f_k p_k \|}{\| p_k \|} = 0$

$$\Updownarrow$$

$$\lim_{k \to \infty} \frac{\| (B_k - \nabla^2 f(x^*)) p_k \|}{\| p_k \|} = 0$$

LECTURE 5

Jan 21, 2026

1) NEWTON WITH HESSIAN MODIFICATION

- Away from sol^n, Hess (f) may not be positive definite
  $\Rightarrow \nabla^2 f(x_k) p_k^N = -\nabla f(x_k)$ may not be descent !
- Overcome this with Gaussian elimination.

ALGORITHM (Line Search Modified Hess Newton)

Given initial x_0;

for $k = 0, 1, 2, \ldots$

- Factorize $B_k = \nabla^2 f(x_k) + E_k$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is suff. positive definite; otherwise E_k is chosen to ensure B_k is suff. pos. def.
- Solve $B_k p_k = -\nabla f(x_k)$
- Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$ where α_k satisfies Wolfe, Goldstein, or Armijo backtracking

end.

□

NOTE: If $\kappa(B_k) = \|B_k\| \|B_k^{-1}\| \leq C$, $C > 0$ $\forall k \in \mathbb{N} \cup \{0\}$ we have global convergence:

THM: Let f be C^2 on $D \subset \mathbb{R}^n$, and let the starting pt. x_0 of the algorithm above be such that the level set $L = \{x \in D : f(x) \leq f(x_0)\}$ is compact. Then, if

$$\kappa(B_k) = \|B_k\| \|B_k^{-1}\| \leq C, \quad C > 0 \quad \forall k \in \mathbb{N} \cup \{0\},$$

we have

$$\lim_{k \to \infty} \nabla f(x_k) = 0 .$$

2) EIGENVALUE MODIFICATION

- **Ex**: $\nabla f_k = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix}$ and $\nabla^2 f_k = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}$

  $\Rightarrow$ Indefinite

- Spectral Decomposition: $Q = \mathbb{1}$, $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$.

  Then $$\nabla^2 f_k = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T .$$

- Pure Newton step $\rightsquigarrow$ $p_k^N = (-0.1, 1, 2)^T$ NOT DESCENT !

  $\nabla f_k^T p_k^N > 0$ ←

- **SOLUTION**: replace $\nabla^2 f_k$ by a pos. def. approx. B_k in which

all negative eigenvalues in $\nabla^2 f_k$ are replaced by a small positive $\delta > 0$ that is somewhat larger than $\varepsilon_{machine}$ (e.g., $\delta = \sqrt{\varepsilon_{machine}}$).

- In the example, we modify the Hessian to be

$$B_k = \sum_{i=1}^{2} \lambda_i q_i q_i^T + \delta q_3 q_3^T$$

$$= \text{diag}(10, 3, \underbrace{10^{-8}}_{\leftarrow \sqrt{\varepsilon_{machine}}})$$

$$\succ 0, \text{ numerically.}$$

- Moreover, the curvature along the ev's q_1 and q_2 has been preserved.

- Search direction now is:

$$p_k = -B_k^{-1} \nabla f_k$$

$$= -\sum_{i=1}^{2} \frac{1}{\lambda_i} q_i (q_i^T \nabla f_k) - \frac{1}{\delta} q_3 (q_3^T \nabla f_k)$$

$$\approx -(2 \cdot 10^8) q_3.$$

- For small δ, p_k is nearly parallel to q_3 and is very long.

↳ not really compatible w/ Newton's method

Many ways to modify this (e.g., increase δ, etc)

!

- If $A = A^T$ and $A = Q \Lambda Q^T$, then the correction matrix ΔA of minimum $\|\cdot\|_F$ that ensures $\lambda_{\min}(A + \Delta A) \geq \delta$ is given by

$$\Delta A = Q \operatorname{diag}(\tau_i) Q^T, \quad \tau_i = \begin{cases} 0, & \lambda_i \geq \delta \\ \delta - \lambda_i, & \lambda_i < \delta \end{cases}.$$

REMARK: ΔA is not diagonal in general and the modified matrix is then

$$A + \Delta A = Q(\Lambda + \operatorname{diag}(\tau_i)) Q^T.$$

- If $A = A^T$ and $A = Q \Lambda Q^T$, the correction matrix ΔA of minimum $\|\cdot\|_2$ such that $\lambda_{\min}(A + \Delta A) \geq \delta$ is given by

$$\Delta A = \tau \mathbb{1}, \quad \tau := \max\{0, \delta - \lambda_{\min}(A)\}.$$

Then $A + \Delta A = A + \tau \mathbb{1}$

3) Bunch-Parlett or Modified Symm. Indefinite

- Any $A = A^T$ can be written as

$$PAP^T = LBL^T$$

P: permutation matrix

L: unit lower triang.

B: block diag. with blocks of dim 1 or 2

- B has the same signature as A (# of positive and negative eigenvals.) → Sylvester's Law of Inertia

  # positive ew's = # positive 1×1 blocks + # 2×2 blocks

- Amount of computation is $\approx$ Cholesky.

- Ensure the modified factors are $\succ 0$:

  (i) Compute the factorization $PAP^T = LBL^T$

  (ii) Compute $B = Q\Lambda Q^T$ (not expensive b/c B is block diag.)

  (iii) Construct F such that $L(B+F)L^T$ is suff. positive definite.

  (iv) Define, for $i = 1, \dots, n$,

$$F := Q\,\mathrm{diag}(z_i)\,Q^T, \quad z_i := \begin{cases} 0, & \lambda_i \geq \delta \\ \delta - \lambda_i, & \lambda_i < \delta \end{cases}$$

where λ_i are the ev's of B.

- The matrix F is thus the modification of minimum $\|\cdot\|_F$ that guarantees the ev's of $B+F$ are $\geq \delta$. This factorization produces:

$$P(A+E)P^T = L(B+F)L^T,$$

where $E = P^T L F L^T P$.

↑ not diag. in general

TRUST REGION METHODS

- Outperforms line search in # of function evaluations ⇒ Better if computing the fct. is expensive.
- Quadratic Model: $m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p$ (Taylor expansion at x_k)
- Natural choices: $B_k = \nabla^2 f_k$ and $g_k = \nabla f_k$
- TRUST REGION PROBLEM:

$$\begin{cases} \min\limits_{p \in \mathbb{R}^n} & m_k(p) \\ \text{subject to} & \|p\| \leq \Delta_k \end{cases}$$

- Difference between $m_k(p)$ and $f(x_k+p)$ is $O(\|p\|^2)$
  $\Rightarrow$ small if p is small

- Approximation error in the model fct m_k when $B_k = \nabla^2 f(x_k)$ is $O(\|p\|^3)$ $\Rightarrow$ very accurate when p is small

---

LECTURE 6

Jan 26, 2026

THM: The vector p^* is a global solution of the trust-region problem

$$\min_{p \in \mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \quad \text{s.t. } p^T p \leq \Delta^2,$$

if and only if p^* is feasible and there is a scalar $\lambda \geq 0$ s.t. the following conditions hold:

1) $(B + \lambda \mathbb{1})p^* = -g$;

2) $\lambda(\Delta - \|p^*\|) = 0$;

3) $(B + \lambda \mathbb{1}) \succeq 0$.

Pf: (Homotopy Argument) Suppose $\exists\, \lambda$ s.t. $(B+\lambda \mathbb{1})p^* = -g$
$\lambda(\Delta - \|p^*\|) = 0$

Continuously adjust λ. Consider p^* to be a fct. of λ: $p^*(\lambda)$.

$$p^*(\lambda) = -(B+\lambda I)^{-1} g \xrightarrow{\lambda \to \infty} 0 \quad (*)$$

Optimality cond^{n} $\min_p \frac{1}{2} p^T (B+\lambda I) p + g^T p$

ASIDE

Obs: Inactive trust region (inactive trust region)

$\min_p \frac{1}{2} p^T B p + g^T p$ s.t. $p^T p \le \Delta^2$. If $p^T p < \Delta$, then unconstrained. $\Rightarrow B p^* + g = 0$

$B \succeq 0$.

$\Rightarrow \lambda = 0 \Rightarrow p^* = -B^{-1} g$

- If trust-region is active: $B + \lambda I \succeq 0 \Rightarrow$ convex problem

- Now, eigendecomposition of B: $B q_i = \lambda_i q_i$,

$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$.

$(B+\lambda I) q_i = (\lambda_i + \lambda) q_i$

$(B+\lambda I) = Q \hat{\Lambda} Q^T$, $Q \in O(n)$

$(B+\lambda I)^{-1} = Q \hat{\Lambda}^{-1} Q^T$.

$$(B+\lambda I)^{-1} a = Q \hat{\Lambda}^{-1} Q^T a = \sum_{i=1}^{n} \frac{q_i^T a \, q_i}{\lambda_i} = \sum_{i=1}^{n} \frac{q_i^T a \, q_i}{\lambda_i + \lambda}.$$

$(*) \Rightarrow (B + \lambda I)\, p^*(\lambda) = -g$

$$\Rightarrow \boxed{p^*(\lambda) = -\sum_{i=1}^{n} \frac{q_i^T g\, q_i}{\lambda_i + \lambda}}$$

So, if $\lambda \to +\infty$, then $\|p^*(\lambda)\| \to 0$.

- Assume λ_1 is not degenerate.

$$p^*(\lambda) = -\sum_{i=2}^{n} \frac{q_i^T g\, q_i}{\lambda + \lambda_i} - \underbrace{\frac{q_1^T g\, q_1}{\lambda_1 + \lambda}}_{\to +\infty}$$

On $(-\lambda_1, \infty)$ $p^*(\lambda)$ is continuous.

$\lambda \searrow \lambda_1$ then $\|p^*(\lambda)\| = +\infty$

continuity $\Rightarrow \exists \lambda^*$ s.t. $\|p^*(\lambda)\| = \Delta \Rightarrow$ sol^n of trust region

$\square$

$\|p(\lambda)\| < \Delta \quad \forall \lambda \in (-\lambda_1, \infty)$

HARD CASE: $\lambda \geq -\lambda_1 \Rightarrow \lambda = -\lambda_1$. (EDGE CASE)

May have double roots

$$p(z) = -\sum_{j : \lambda_j \neq \lambda_1} \frac{q_j^T g}{\lambda_j - \lambda_1} + z\, q_1$$

$\exists z : \|p(z)\| = \Delta$?

"Inexact" Trust Regions

- Use Cauchy pt. as yardstick
  $\Leftrightarrow$ solve trust-region problem for linear instead of quadratic:

$$p_k^s = \underset{p \in \mathbb{R}^n}{\operatorname{argmin}} \; f_k + g_k^T p$$

$$\text{s.t. } \|p\| \leq \Delta_k .$$

- Calculate $\tau_k = \underset{\tau \geq 0}{\operatorname{argmin}} \; m_k(\tau p_k^s)$, s.t. $\|\tau p_k^s\| \leq \Delta_k$.
- Set Cauchy point: $p_k^c = \tau_k p_k^s$.

- Useful when B_k is invertible but not positive definite.

↑ "DOGLEG"

LECTURE 7 (Ch 5)

Jan 28 2026

CONJUGATE GRADIENT: (KRYLOV METHOD)

Avoid factorizing matrices ⇝ use methods that only use matrix-vector multiplication.

⇓

Reduces cost

- GRAM-SCHMIDT:

  - Classical Algorithm: Given $a_1, \dots, a_n$,

```
for k = 1, ..., n
    q_k = a_k
    for j = 1, ..., k-1
        r_jk = q_j^T a_k
        q_k = q_k - r_jk q_j
    end
    r_kk = ||q_k||^2
    q_k = q_k / r_kk
end
```

  - q_k, r_{jk} → Reduced QR.

* CONJUGATED = ORTHOGONAL

• $A^T = A \succ 0 \Rightarrow Ax = b \Leftrightarrow \min_{x \in \mathbb{R}^n} \varphi(x) := \frac{1}{2} x^T A x - b^T x$

$\nabla \varphi(x) = Ax - b =: r(x)$

• So, at $x = x_k$, $r_k = A x_k - b$.

IDEA: Generate cheaply a set of vectors with a property known as conjugacy.

DEF: A set of nonzero vectors $\{p_0, p_1, \dots, p_l\}$ is said to be A-conjugate w.r.t. $A^T = A \succ 0$ if

$$p_i^T A p_j = 0 \quad \forall\, i \neq j.$$

NOTE: Easily, any set of conjugate vectors is also lin. independent.

• Importance of conjugacy is that we can minimize $\varphi(\cdot)$ in n steps by successively minimizing along the individual directions in a conjugate set

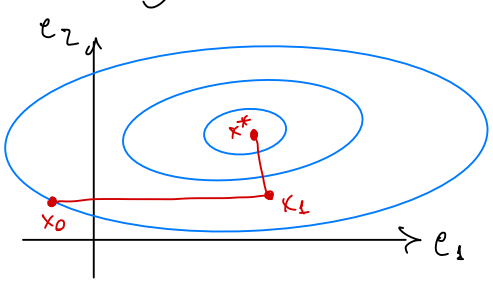
• Given starting $x_0 \in \mathbb{R}^n$ and a conjugate set $\{p_0, \dots, p_{n-1}\}$, set $x_{k+1} = x_k + \alpha_k p_k$, where α_k is the 1-dimensional

minimizer of the quadratic $\varphi(\cdot)$ along $x_k + \alpha p_k$.

Explicitly, $$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}.$$

THM: $\forall x_0 \in \mathbb{R}^n$, the sequence $\{x_k\}$ generated by the conjugate direction algorithm $x_{k+1} = x_k + \alpha_k p_k$, with $\alpha_k = -r_k^T p_k / p_k^T A p_k$, converges to sol^n x^* of $Ax = b$ in at most n steps.

NOTE: If A is diagonal in $\varphi(x) = \frac{1}{2} x^T A x - b^T x$, the contours of the fct $\varphi(\cdot)$ are ellipses whose axes are aligned with coordinate directions



← Can find minimizer by doing 1-dim. minimizations along the coord. directions $e_1, \ldots, e_n$.

- If A is not diagonal, the contours are still elliptical but they are usually no longer aligned with coordinate directions. ⇒ Minimization along these directions no longer gives sol^n in a finite # of iterations

GOAL: Transform A into diagonal and minimize!

- Change of variables: $\hat{x} := S^{-1}x$, where

$$S := \begin{bmatrix} | & | & & | \\ p_0 & p_1 & \cdots & p_{n-1} \\ | & | & & | \end{bmatrix} \quad \leftarrow A\text{-conjugate directions}$$

- Then

$$\hat{q}(\hat{x}) \overset{\text{def}}{=} q(S\hat{x}) = \frac{1}{2}\hat{x}^T S^T A S \hat{x} - b^T S \hat{x}.$$

$p_i^T A p_j = 0 \implies S^T A S$ is diagonal

$\implies$ Can find the min. of $\hat{q}$ by performing n 1-dimensional minimizations along the coordinate directions of $\hat{x}$.

- NOTE: When A is diagonal, k iterations give the first k coordinates of the solution (i.e., x^* in $e_1, \dots, e_k$). This actually also works when A is not necessarily diagonal.

THM: (Expanding Subspace Minimization) $x_0 \in \mathbb{R}^n$ starting pt. and $\{x_k\}$ generated by the conjugate gradient algorithm $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_k = -r_k^T p_k / p_k^T A p_k$. Then, if $r_{k+1} = r_k + \alpha_k A p_k$. We have $r_k^T p_i = 0 \quad \forall i = 0, \dots, k-1$

and x_k is the minimizer of $\varphi(x) = \frac{1}{2} x^T A x - b^T x$ over the set

$$\left\{ x \in \mathbb{R}^n : x = x_0 + \text{span}\{p_0, \dots, p_{k-1}\} \right\}.$$

⚠ NOTE: Current residual r_k is orthogonal to all previous search directions

- How to choose conjugate direction set ?
  - Eigenvectors $v_1, \dots, v_n$ of A are mutually orthogonal as well as conjugate w.r.t. A.
    ↳ Possible choice, but VERY expensive for large-scale applications...
  - Use a Gram-Schmidt-esque algorithm to produce a conjugate set of directions (instead of orthonormal directions).
    ↳ Also expensive though...

* CONJUGATE GRADIENT : Generates a set of conjugate vectors and it can compute a new conjugate vector p_k by only using the previous one p_{k-1}.

i.e., it doesn't need to know $p_0, \dots, p_{k-2}$

$\Rightarrow$ Very cost-effective!

- In conjugate gradient, each direction p_k is chosen to be a linear combination of the negative residual $-r_k$, which, by definition $\nabla \varphi(x) = Ax - b =: r(x)$, is the steepest descent direction for φ; and of p_{k-1}.

— PRELIMINARY CG (not the most cost-effective; cf. pg 107 of N&W)

- Write $p_k = -r_k + \beta_k p_{k-1}$, where

$$\beta_k := \frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}$$

} Enforces p_{k-1} and p_k are conjugate w.r.t. A

Choose $p_0 = -\nabla \varphi(x_0)$ (steepest descent at x_0)

* See Algorithms 5.1 (preliminary CG) and 5.2 (actual cost-effective CG) for actual implementations.

$r_k \in \mathcal{K}_{k+1}(A, r_0)$

$p_k \in \mathcal{K}_{k+1}(A, r_0)$

Convergence Rate of CG

- Exact arithmetic $\Rightarrow$ CG terminates in n steps.
- Note: $x_{k+1} = x_0 + \alpha_0 p_0 + \cdots + \alpha_k p_k$

  $= x_0 + \gamma_0 r_0 + \gamma_1 A r_0 + \cdots + \gamma_k A^k r_0$.
- Let $p_k^*(z) = \gamma_k z^k + \cdots + \gamma_1 z + \gamma_0$.

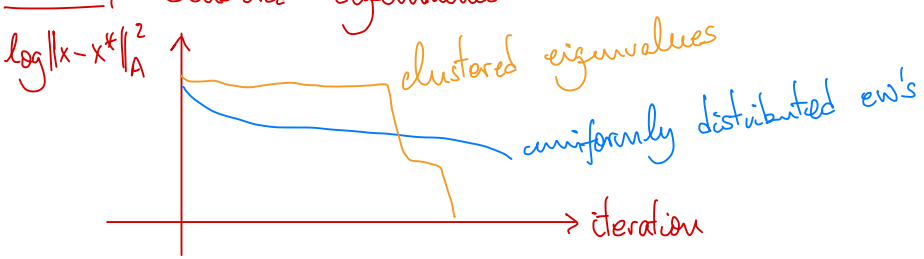
$\Rightarrow$ $\boxed{x_{k+1} = x_0 + p_k^*(A) r_0}$.

Thm: If A (SPD) has d distinct eigenvalues, then CG terminates at the solution in *at most* d iterations.

Thm: If $A = A^T > 0$ has eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, then

$$\|x_{k+1} - x^*\|_A^2 \le \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1} \right)^2 \|x_0 - x^*\|_A^2 .$$

Warning: Clustered eigenvalues



UPSHOT: $\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \lambda_n / \lambda_1$

$$\|x_k - x^*\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x_0 - x^*\|_A .$$

PRECONDITIONING: $x \mapsto Cx =: \hat{x}$.

$$\hat{\phi}(\hat{x}) = \frac{1}{2} \hat{x}^T \left(C^{-T} A C^{-1} \right) \hat{x} - \left(C^{-T} b \right)^T \hat{x} .$$

$\min_{\hat{x}} \hat{\phi}(\hat{x}) \iff$ solve $\left(C^{-T} A C^{-1} \right) x = C^{-T} b$.

- Convergence rate depends on ew's of $C^{-T}AC^{-1}$ (not of A...)
- Want to choose C so that $\kappa(C^{-T}AC^{-1}) \ll \kappa(A)$. 
- Can also choose C so that ew's of $C^{-T}AC^{-1}$ are clustered.
- See Algorithm 5.3 for Preconditioned CG.

↑ uses preconditioner $M := C^T C$

- Choosing the preconditioner: inexpensive M, easy to solve $My = r$, etc... Common choice: INCOMPLETE CHOLESKY

$M = \tilde{L}\tilde{L}^T$ $\quad \in \quad A \approx \tilde{L}\tilde{L}^T$, $\tilde{L} \approx L$ where $LL^T = A$.

$C^{-T}AC^{-1} = \tilde{L}^{-1} A \tilde{L}^{-T} \approx \mathbb{1}$.

LECTURE 8 (Ch 6) QUASI-NEWTON METHODS

Feb 2, 2026

FINAL

Davidon's Quasi-Newton ⇒ one of the best methods for optimization

BFGS METHOD (Broyden, Fletcher, Goldfarb, Shanno)

- Quadratic model for iterate x_k:

$$m_k(p) := f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p .$$

$B_k = B_k^T \succ 0$ $n \times n$
updated at each k

$$\underset{p}{\operatorname{argmin}}\ m_k(p) =: p_k = -B_k^{-1} \nabla f_k$$

Approximate Hessian

$$x_{k+1} = x_k + \alpha_k p_k .$$

Wolfe

- Next iterate:

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p .$$

CONDITIONS ON B_{k+1}:

1) $\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k .$

$$\iff \boxed{B_{k+1} \alpha_k p_k = \nabla f_{k+1} - \nabla f_k} \quad !$$

(SECANT EQUATION)

2) B_{k+1} must map $x_{k+1} - x_k = \alpha_k p_k$ into $\nabla f_{k+1} - \nabla f_k$ for secant equation to make sense. So, we must have

$$(x_{k+1} - x_k)^T (\nabla f_{k+1} - \nabla f_k) > 0 \qquad \text{(CURVATURE CONDITION)}$$

Obs: $\nabla f_{k+1} - \nabla f_k \approx \nabla^2 f_k (x_{k+1} - x_k) \overset{!}{>} 0$

REMARK: If we use Wolfe line search, curvature condition holds automatically!

$$(\nabla f_{k+1} - \nabla f_k)^T (x_{k+1} - x_k) \underset{\text{Wolfe}}{\geq} (c_2 - 1) \nabla f_k^T (x_{k+1} - x_k) > 0$$

! DAVIDON - FLETCHER POWELL UPDATE: The sol^n to minimum update problem that is SPD is the DFP update

$$B_{k+1} = \underbrace{(I - \rho_k y_k s_k^T) B_k (I - \rho_k s_k y_k^T)}_{\geq 0} + \underbrace{\rho_k y_k y_k^T}_{\geq 0}, \quad \geq 0$$

where $\rho_k := \dfrac{1}{y_k^T s_k}$, $\begin{cases} y_k := \nabla f_{k+1} - \nabla f_k \\ s_k := x_{k+1} - x_k \end{cases}$

- Algorithm 6.1 $\rightarrow$ BFGS method. SEE THM 6.5 & 6.4

LECTURE 8 (Ch 7)

Feb 4, 2026

- Prove Thms in book
- Redo HWs
- Do book problems (Ch 1, 2, 3, 4, 5, 6)
- CG $\leftrightarrow$ minimization over polynomials

LARGE SCALE UNCONSTRAINED OPTIMIZATION

- **INEXACT NEWTON METHODS:**
  - Newton seeks to solve $\nabla^2 f_k p_k^N = -\nabla f_k$. (Symmetric $n \times n$ system)
  - INEXACT Newton:

$$r_k := \nabla^2 f_k p_k + \nabla f_k$$

Inexact step

  - Usually we terminate $\|r_k\| \leq \eta_k \|\nabla f_k\|$

$0 < \eta_k < 1 \quad \forall k$

FORCING SEQUENCE

  - CHOICE OF FORCING SEQUENCE η_k AFFECTS CONVERGENCE OF INEXACT NEWTON.

THM (7.1): Say $\nabla^2 f(x)$ is continuous around minimizer x^* and $\nabla^2 f(x_*) \succ 0$. Consider $x_{k+1} = x_k + p_k$, where the step p_k satisfies $\|r_k\| \overset{\text{def}}{=} \|\nabla^2 f_k p_k + \nabla f_k\| \leq \eta_k \|\nabla f_k\|$, and $\eta_k \leq \eta \ \forall k$ for some $\eta \in [0, 1)$. If x_0 is sufficiently close to x^*, then $x_k \to x^*$ and

$$\|\nabla^2 f(x^*)(x_{k+1} - x^*)\| \leq \hat{\eta} \|\nabla^2 f(x^*)(x_k - x^*)\|$$

for some $\hat{\eta}$ s.t. $\eta < \hat{\eta} < 1$.

- If the forcing sequence $\eta_k \xrightarrow{k\to\infty} 0$, then we solve the subproblem exactly and approach the Newton step.

THM 7.2: Suppose conditions of 7.1 hold, and assume that the iterates x_k generated by inexact Newton are s.t. $x_k \to x^*$. Then, the rate of convergence is SUPERLINEAR if $\eta_k \to 0$. Moreover, if $\nabla^2 f(x)$ is Lipschitz around x^* and $\eta_k = O(\|\nabla f_k\|)$ then convergence is QUADRATIC.

- NOTE: To get superlinear convergence, set, e.g.,

$$\eta_k = \min\left(\frac{1}{2}, \sqrt{\|\nabla f_k\|}\right).$$

To get quadratic convergence, use

$$\eta_k = \min\left(\frac{1}{2}, \|\nabla f_k\|\right).$$

} Conservative choice ⇓ Many iterations in the beginning

KRYLOV METHODS (7.2)

- Deal w/ indefiniteness of matrices (CG works only for positive definite matrices...)

IDEA: Proceed with CG until we find a negative inner product.

Ex: Line Search Newton-CG (Alg 7.1); CG-Trust Region (Alg 7.2 Steihaug); L-BFGS two loop recursion (Alg 7.4)

Summary of Optimization

Q: How do you get global convergence to stationary points?

A: Line search through Zoutendijk or trust region

MIDTERM: PROVE ZOUTENDIJK FOR ANY LINE SEARCH SCHEME (lecture → Wolfe; HW → backtracing; need to show for Goldstein?)

Q: How to get quadratic convergence?

A: Newton

Q: How to get convergence to second-order stationary pts?

A: Trust-region (only exact line-search & exact Hessian) (Dogleg doesn't work)

Q: Don't want to factorize matrix in Newton.

A: CG

Q: What if matrix is not positive def.?

A: Newton-CG

Q: Superlinear convergence without computing Hessian?

A: Quasi-Newton (e.g. BFGS)

Q: Large problems with many variables? A: Preconditioning

LECTURE 9

(Ch 7, 10)

Feb 8, 2026

- Newton-CG, BFGS, trust region

LECTURE 10

(Ch 12, 15)

Feb 16, 2026

* CONSTRAINED OPTIMIZATION

PROBLEM: $\min_{x \in \mathbb{R}^n} f(x)$ subject to $x \in \Omega$

FEASIBLE SET:

$$\Omega = \{x : c_i(x) = 0, \ i \in \mathcal{E} \ ; \ c_i(x) \geq 0, \ i \in \mathcal{I}\}$$

$$\min_{x \in \Omega} f(x)$$

f very differentiable

LOCAL VS GLOBAL SOLUTIONS:

- Constraints make the problem simpler since the search space is smaller

- But may be more complicated
  $\min (x_2 + 100)^2 + 0.01 x_1^2$ subject to $x_2 \geq \cos x_1$
- Unconstrained $\rightsquigarrow$ one min
  Constrained $\rightsquigarrow$ many minima

DEF: (1) x^* is local solution of $\min_{x \in \Omega} f(x)$ if $x^* \in \Omega$ and $\exists N \ni x^*$ s.t. $f(x) \geq f(x^*)$ $\forall x \in N \cap \Omega$.

(2) x^* is an isolated local solution if $x^* \in \Omega$ and $\exists N \ni x^*$ s.t. x^* is the only local sol^n in $N \cap \Omega$.

(3) x^* is a strict local solution (also called strong local solution) if $x^* \in \Omega$ and $\exists N \ni x^*$ s.t. $f(x) > f(x^*)$ $\forall x \in N \cap \Omega$ with $x \neq x^*$.

NOTE: $\min f(x)$, $f(x) = \max \{x^2, x\}$

$$\Leftrightarrow \begin{cases} \min t \\ \text{s.t. } \max\{x^2, x\} \leq t \end{cases} \Leftrightarrow \begin{cases} \min t \\ \text{s.t. } x^2 \leq t, \; x \leq t \end{cases}$$

* No 1st order Descent:

- To make sure we're at a minimum, at least the following should happen:
  - If we have a feasible arc the gradient fct should not decrease from a min. point.
  - i.e., we should have that the system of inequalities:

$$\nabla f(x^*)^T d < 0 \quad \text{has no solution.}$$

* Tangent & Linearized Cone

- Tangent Cone

$$T_\Omega(x) := \left\{ d : \exists \{z_k\} \subset \Omega,\ z_k \to x,\ \exists \{t_k\} \subset \mathbb{R}_{>0},\ t_k \searrow 0,\ \lim_{k\to\infty} \frac{z_k - x}{t_k} = d \right\}$$

- Linearized Feasible Direction Set

$$F(x) := \left\{ d : d^T \nabla c_i(x) = 0,\ i \in \mathcal{E},\ d^T \nabla c_i(x) \geq 0,\ i \in \mathcal{I} \cap \mathcal{A} \right\}$$

($\mathcal{A}$: active)

$\Rightarrow T_\Omega(x) \subset F(x)$ ALWAYS.

WHEN DO WE HAVE $T_\Omega(x) = F(x)$?

Ex: $\begin{cases} \min x_1 + x_2 \\ \text{s.t. } x_1^2 + x_2^2 - 2 = 0 \end{cases}$

If "=" holds, we say "constraints are qualified".

Take $(1,1)$.

$T_\Omega(x) \parallel \{ x_1 + x_2 = 2 \} \rightsquigarrow d_1 + d_2 = 0.$

$\Rightarrow$ Eq. is the same as the eq. of linearized feasible directions: $2x_1 d_1 + 2x_2 d_2 = 0$.

$\Rightarrow$ In this case, the sets are equal.

Ex: $\begin{cases} \min x_1 + x_2 \\ \text{s.t. } (x_1^2 + x_2^2 - 2)^2 = 0 \end{cases}$

- $T_\Omega(x)$ is dependent on the feasible set and not of the representation, so it doesn't change, it's still $d_1 + d_2 = 0$.
- Feasible directions:

$$4(x_1^2 + x_2^2 - 2) x_1 d_1 + 4(x_1^2 + x_2^2 - 2) x_2 d_2 = 0$$

$$= 0 \cdot d_1 + 0 \cdot d_2 = 0$$

$\Rightarrow F(x) = \mathbb{R}^2 \neq T_\Omega(x)$

* INEQUALITY CONSTRAINTS: ACTIVE SET

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad \begin{cases} c_i(x) = 0 & i \in \mathcal{E} \\ c_i(x) \geq 0 & i \in \mathcal{I} \end{cases}.$$

DEF: (ACTIVE SET) Say $x \in \Omega$, then

$$\mathcal{A}(x) := \mathcal{E} \cup \{ i \in \mathcal{I} : c_i(x) = 0 \}.$$

DEF: (LICQ) Given $x \in \Omega$ and $\mathcal{A}(x)$, we say LICQ (linear independence constraint qualification) holds if the set of active constraint gradients

$$\{ \nabla c_i(x) : i \in \mathcal{A}(x) \}$$

is linearly independent.

THM: $x^* \in \Omega$ feasible, then

1) $T_\Omega(x^*) \subset F(x^*)$

2) If LICQ holds at x^*, then $T_\Omega(x^*) = F(x^*)$

$$\Rightarrow T_\Omega(x) = F(x).$$

LAGRANGIAN: $\mathcal{L}(x, \lambda) := f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$.

THM: (KKT – 1st ORDER OPTIMALITY CONDITIONS) If x^* is a local sol^n to $\min_{x \in \Omega} f(x)$, f and c_i are C^1, and LICQ holds at x^*, then $\exists \lambda^* = (\lambda_i^*)_{i \in \mathcal{E} \cup \mathcal{I}}$ Lagrange multiplier vector s.t., at (x^*, λ^*),

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$$

$$c_i(x^*) = 0, \quad i \in \mathcal{E}$$

$$c_i(x^*) \geq 0, \quad i \in \mathcal{I} \qquad \text{(KKT)}$$

$$\lambda_i^* \geq 0, \quad i \in \mathcal{I}$$

$$\lambda_i^* c_i(x^*) = 0, \quad i \in \mathcal{E} \cup \mathcal{I}$$

assume only C^1

THM: If x_* is a local sol^n of $\min_{x \in \Omega} f(x)$, then

$$\nabla f(x^*)^T d \geq 0 \quad \forall d \in T_\Omega(x^*).$$

Pf: (Taylor expand over feasible directions.)

□

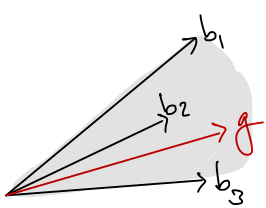
FARKAS LEMMA: Let $K := \{ By + Cw : y \geq 0 \}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{n \times p}$. This cone K is closed (not trivial).

Given any vector $g \in \mathbb{R}^n$, we have either $g \in K$

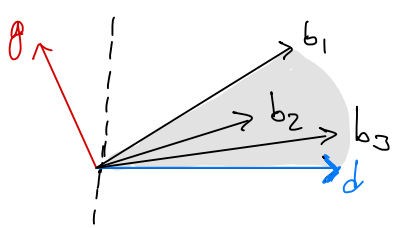
or $\exists d \in \mathbb{R}^n$ s.t. $g^T d < 0$, $B^T d \geq 0$, $C^T d = 0$.

Never both.

↳ Either g is in a convex cone C, or, else, it is on the other side of a hyperplane from the cone C; the hyperplane can be chosen to go through the origin.



OR



DEF: (Critical Cone)

$$C(x^*, \lambda^*) = \{ w \in T_\Omega(x^*) : \nabla f(x^*)^T w = 0 \}.$$

DEF: (Strict Complimentarity) Given a local sol^n x^* of $\min_{x \in \Omega} f(x)$ and λ^* satisfying (KKT), the strict complimentarity condition holds if exactly one of λ_i^* and $c_i(x^*)$ is zero for each index $i \in \mathcal{I}$; i.e., $\lambda_i^* > 0 \quad \forall i \in \mathcal{I} \cap \mathcal{A}(x)$.

THM: (2^{nd} order Necessary Conditions) Suppose x^* is a local sol^n of $\min_{x \in \Omega} f(x)$ and that the LICQ condition is satisfied. Let λ^* be the Lagrange multiplier vector for which KKT holds. Then

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \geq 0 \quad \forall w \in C(x^*, \lambda^*).$$

THM: (2^{nd} order sufficient conditions) Suppose that for some feasible $x^* \in \mathbb{R}$ $\exists \lambda^*$ s.t. KKT holds. Suppose

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w > 0 \quad \forall \, 0 \neq w \in C(x^*, \lambda^*)$$

Then, x^* is a strict local sol^n for $\min_{x \in \Omega} f(x)$.

LECTURE 11

(Ch 15)

Feb 18, 2026

* LINEAR PROGRAMING

- Problem with linear objective and linear equality and inequality constraints.

- Standard Form: $\min c^T x$
  $\text{s.t. } Ax = b$
  $x \geq 0$

NOTE: Any linear programming problem can be turned into this form.

* QUADRATIC PROGRAMMING PROBLEMS

$$\begin{cases} \min_x q(x) := \frac{1}{2} x^T G x + x^T c \\ \text{subject to } a_i^T x = b_i, \quad i \in \mathcal{E} \\ \qquad\qquad a_i^T x \geq b_i, \quad i \in \mathcal{I} \end{cases}$$

- Optimality Conditions (linear $G = 0$):
  - Original form: $Gx - A^T\lambda + c = 0$
    $Ax - b \geq 0$
    $(Ax-b)_i \lambda_i = 0, \quad i = 1, \dots, m$
    $\lambda \geq 0$
  - Slack form: $Gx - A^T\lambda + c = 0$
    $Ax - y - b = 0$

$$y_i \lambda_i = 0, \quad i = 1, \dots, m$$

$$(y, \lambda) \geq 0$$

- PERTURBED KKT conditions as a nonlinear system:

$$F(x, y, \lambda; \sigma, \mu) = \begin{bmatrix} Gx - A^T\lambda + c \\ Ax - y - b \\ Y\Lambda e - \sigma\mu e \end{bmatrix} = 0$$

- Solve successively while taking $\mu \to 0$:

$$F(x(\mu), y(\mu), \lambda(\mu); \mu, \sigma) = 0, \quad y_i(\mu)\lambda_i(\mu) = \mu\sigma > 0$$

$$\mu \to 0 \Rightarrow (x(\mu), y(\mu), \lambda(\mu)) \to (x^*, y^*, \lambda^*)$$

satisfies usual KKT

- BEST METHODS FOR LP AND CONVEX QP:
  - Very large sparse LP and sparse convex QP can be very efficiently solved by interior point methods on GPU.
  - Need factorization

- Holy Grail: preconditioned CG for interior pt.

* PENALTY METHODS:

Idea: replace the constraints by a penalty term

- Inexact penalties: parameter driven to infinity to recover solution.

Ex: $x^* = \operatorname{argmin} f(x)$ subject to $c(x) = 0$

$\Leftrightarrow x^\mu = f(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x)$, $x^* = \lim_{\mu \to \infty} x^\mu$.

- Exact but nonsmooth penalty $\rightsquigarrow$ the penalty parameter can stay finite.

Ex:

$x^* = \operatorname{argmin} f(x)$ subject to $c(x) = 0$

$\Leftrightarrow f(x) + \mu \sum_{i \in \mathcal{E}} |c_i(x)|$, $\mu \geq \mu_0$.

- AUGMENTED LAGRANGIAN METHODS: Mix Lagrangian pt. of view

$x^* = \operatorname{argmin} f(x)$ subject to $c(x) = 0$

$$x^{\mu,\lambda} = f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i(x)^2$$

$$x^* = \lim_{\lambda \to \lambda^*} x^{\mu,\lambda}, \quad \mu \geq \mu_0 > 0.$$

* SEQUENTIAL QUADRATIC PROGRAMMING

- Successively solve quadratic programs:

$$\begin{cases} \min\limits_{p} \ \frac{1}{2} p^T B_k p + \nabla f_k \\ \text{subject to} \ \nabla c_i(x_k)^T d + c_i(x_k) = 0, \ i \in \mathcal{E} \\ \qquad\qquad \nabla c_i(x_k)^T d + c_i(x_k) \geq 0, \ i \in \mathcal{I} \end{cases}$$

- ANALOGOUS OF NEWTON'S FOR THE CASE OF CONSTRAINTS IF $B_k = \nabla^2_{xx} \mathcal{L}(x_k, \lambda_k)$

↳ Solve this subproblems with extensions of simplex e.g.: BFGS which makes it convex.

* Interior Point Methods

- Reduce the ineq. constraints w/ a barrier:

$$\begin{cases} \min_{x,s} \; f(x) - \mu \sum_{i=1}^{m} \log s_i \\ \text{subject to } c_i(x) = 0, \; i \in \mathcal{E} \\ \qquad\qquad c_i(x) - s_i = 0, \; i \in \mathcal{I} \end{cases}$$

- Can also use a penalty:

$$\min_{x,s} \; f(x) - \mu \sum_{i=1}^{m} \log s_i + \frac{1}{2\mu} \sum_{i \in \mathcal{I}} \left[c_i(x) - s_i \right]^2 + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i(x)^2 .$$

- NOTE: can solve this as a seq. of unconstrained problems.

- NOTE: Don't need to start w/ a feasible point.

LECTURE 12

Feb 23 2026

* MERIT FUNCTIONS & FILTERS

$$\phi(x) = w_1 f(x) + w_2 \left[\sum_{i \in \mathcal{E}} |c_i(x)| + \sum_{i \in \mathcal{I}} \max\{-c_i(x), 0\} \right]$$

for weights $w_1, w_2 > 0$

- Can scale the merit fct so that the weight of objective is 1.
- Infeasibility measure is called "penalty parameter"

- Can monitor progress by ensuring that $\phi(x)$ decreases, as in unconstrained optimization.

$$\Rightarrow \phi_1(x, \mu) = f(x) + \mu \sum_{i \in \mathcal{E}} |c_i(x)| + \mu \sum_{i \in \mathcal{I}} \max\{-c_i(x), 0\}$$

PENALTY PARAMETER

(15.1)

DEF: A merit function $\phi(x, \mu)$ is EXACT iff $\exists \mu^* > 0$ s.t., for any $\mu > \mu^*$, any local sol^n of the nonlinear program is a local minimizer of $\phi(x, \mu)$.

THM: (17.3) ℓ^1 merit fct $\phi_1(x,\mu)$ is exact and the threshold value μ^* is given by

$$\mu^* = \max\{|\lambda_i^*|,\ i \in \mathcal{E} \cup \mathcal{I}\}.$$

* SMOOTH EXACT PENALTY: Fletcher's augmented Lagrangian:

$$\phi_F(x,\mu) = f(x) - \lambda(x)^T c(x) + \frac{1}{2}\mu \sum_i c_i(x)^2,$$

$$\lambda(x) = \left[A(x)A(x)^T\right]^{-1} A(x)\nabla f(x).$$

- Both smooth and exact, but impractical b/c of the linear solve.

* SMOOTH INEXACT PENALTY: Augmented Lagrangian

$$\phi(x) = f(x) - \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i(x)^2$$

- Update of the Lagrange multiplier is needed.

* Armijo Line-Search for Nonsmooth Merit Functions

$$\phi_1(x,\mu) = f(x) + \mu \sum_{i\in\mathcal{E}} |c_i(x)| + \mu \sum_{i\in\mathcal{I}} \max\{-c_i(x), 0\}.$$

Line Search:

$$\phi(x_k,\mu) - \phi(x_k + \beta^m p_k, \mu) \geq -\rho\beta^m D(\phi(x_k,\mu), p_k)$$

$$\phi(x_k,\mu) - \phi(x_k+\beta^m p_k, \mu) \geq -\eta_1 (m(0) - m(p_k)).$$

Trust-Region: $0 < \eta_1 < 0.5$.

* Choose penalty parameter: Tricky & depends on penalty fct.

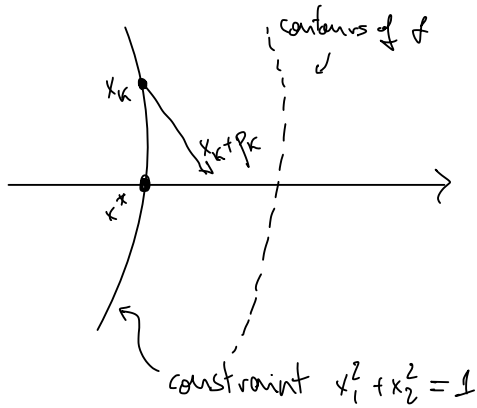
- Adaptive criterion: if optimality gets ahead of feasibility make penalty param. more stringent.
- e.g.: l_1 fct, with $\mu^* = \max\{|\lambda_i^*| : i\in\mathcal{E}\cup\mathcal{I}\}$, take max. of current value of multipliers plus safety factor

* MARATOS EFFECT: Newton step may not be compatible with penalty.

consider $\begin{cases} \min f(x_1, x_2) = 2(x_1^2 + x_2^2 - 1) - x_1 \\ \text{s.t. } x_1^2 + x_2^2 - 1 = 0 \end{cases}$.

- From any feasible pt., Newton-type step increases both infeasibility and objective function.

- So, fast convergence does not occur when coupled with globalization mechanisms.



Solutions?

(a) Use Fletcher's fct that does not suffer from this problem, but both filter and line-search suffer from this.

(b) Do curvilinear search:

- $A_k p_k + c(x_k) = 0$

Correction $\longrightarrow$ $A_k \hat{p}_k + c(x_k + p_k) = 0$

$$\hat{p}_k = -A_k^T (A_k A_k^T)^{-1} c(x_k + p_k)$$

Update $\longrightarrow$ curvilinear line search:

$x_k + p_k + \hat{p}_k \qquad x_k + z p_k + z^2 p_k$

$$\left.\begin{aligned} c(x_k + p_k + \hat{p}_k) &= O(\|x_k - x^*\|) \\ c(x_k + p_k) &= O(\|x_k - x^*\|^2) \end{aligned}\right\} \Rightarrow$$

Corrected Newton step is more likely to be accepted.

Ch 16: Solving Quadratic Programs (QP) w/ Bound Constraints

Problem:

$$\begin{cases} \min\limits_x \; q(x) = \frac{1}{2} x^T G x + x^T c \\ \text{s.t.} \quad l \le x \le u \end{cases}$$

- Like in trust-region, look for Cauchy pt. based on a projection on the feasible set.
- G does not need to be symm. pos. def..
- **Projection Operator:**

$$P(x, l, u)_i := \begin{cases} l_i, & x_i < l_i \\ x_i, & x_i \in [l_i, u_i] \\ u_i, & x_i > u_i \end{cases}$$

- SEARCH PATH: Create a piecewise linear path which is feasible (as opposed to the linear one in unconstrained case) by projecting the gradient:

$$x(t) = P(x - tg, l, u)$$
$$g = Gx + c.$$

- COMPONENT-WISE

$$\tilde{t}_i = \begin{cases} (x_i - u_i)/g_i, & \text{if } g_i < 0 \text{ and } u_i < \infty \\ (x_i - l_i)/g_i, & \text{if } g_i > 0 \text{ and } l_i > -\infty \\ \infty, & \text{otherwise} \end{cases}.$$

Thm, on each component:

$$x_i(t) = \begin{cases} x_i - tg_i, & t \le \tilde{t}_i \\ x_i - \tilde{t}_i g_i, & \text{otherwise} \end{cases}.$$

PATH: $0 < t_1 < t_2 < t_3 < \dots$

DIRECTION: $x(t) = x(t_{j-1}) + (\Delta t)\, p^{j-1}$,

$$\Delta t := t - t_{j-1} \in [0, t_j - t_{j-1}],$$

$$p_i^{j-1} = \begin{cases} -g_i, & t_{j-1} < \tilde{t}_i \\ 0, & \text{else} \end{cases}.$$

- LINE SEARCH REVISITED: Along each piece $[t_{j-1}, t_j]$ find the minimum of the quadratic $\frac{1}{2} x^T G x + c^T x$.

  - Reduces to analyzing a 1-dimension quadratic form of t on an interval.

  - If the min. is on the right end of interval, we continue.

  - If not, we found the local min. and the Cauchy point = the first local minimizer (different from Trust-region and not guaranteed to find global min)

- SUBSPACE MINIMIZATION:

  - Active Set of Cauchy Point

$$A(x^c) := \{ i : x_i^c = l_i \text{ or } x_i^c = u_i \}$$

  - Solve subspace minimization problem:

$$\begin{cases} \min_x q(x) = \frac{1}{2} x^T G x + x^T c \\ \text{s.t. } x_i = x_i^c, \quad i \in A(x^c) \\ l_i \le x_i \le u_i, \quad i \notin A(x^c) \end{cases}$$

← No need to solve exactly, use truncated CG to terminate if an inactive var. reaches bound

ALGORITHM 16.5 (Gradient Projection Method for QP).

- When projection does not progress, x is a fixed pt. of the gradient projection.
- On each component, either the grad. is 0 or the x-component is at the right bound.

AUGMENTED LAGRANGIAN (Ch 17)

Augmented lagrangian:

$$\mathcal{L}_A(x, \lambda, \mu) := f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i(x)^2.$$

Obs: if $\lambda = \lambda^*$, $\mu = \mu_0 \Rightarrow \nabla \mathcal{L}_A(x^*, \lambda^*, \mu) = 0$

$$\nabla^2 \mathcal{L}_A(x^*, \lambda^*, \mu) = \nabla^2 \mathcal{L}(x^*, \lambda^*, \mu) + \mu (\nabla c(x^*))^T (\nabla c(x^*))$$

$\Rightarrow$ x^* is stationary pt. for Aug Lag for exact multipliers and is a min for μ suff. large!

because:

$$\nabla^2 \mathcal{L}_A(x^*, \lambda^*, \mu) \simeq \begin{bmatrix} Y & Z \end{bmatrix}^T \nabla^2 \mathcal{L}(x^*, \lambda^*) \begin{bmatrix} Y & Z \end{bmatrix} + \mu \left(\nabla c(x^*)\, Y\right)^T \left(\nabla c(x^*)\, Y\right)$$

$$= \begin{bmatrix} Z^T \nabla^2 \mathcal{L}(x^*, \lambda^*) Z & * \\ * & * + \mu \left(\nabla c(x^*) Y\right)^T \left(\nabla c(x^*) Y\right) \end{bmatrix}$$

$\succeq 0$ for μ suff. large

- At current estimate, solve

$$0 \simeq \nabla \mathcal{L}_A(x_k, \lambda^k, \mu_k) = \nabla f_k - \sum_{i \in \mathcal{E}} \left[\lambda_i^k - \mu_k c_i(x_k)\right] \nabla c_i(x_k)$$

Obvious choice: $\lambda_i^{k+1} = \lambda_i^k - \mu_k c_i(x_k)$ for all $i \in \mathcal{E}$.

- Inequality? Use slacks

$c_i(x) \geq 0$, $i \in \mathcal{I}$ $\longrightsquigarrow$ $c_i(x) - s_i = 0$, $s_i \geq 0$

$\forall i \in \mathcal{I}$.

- Problem: $\min_{x \in \mathbb{R}^n} f(x)$

s.t. $c_i(x) = 0$, $i = 1, \dots, m$, $\ell_i \leq x \leq u$.

- New Aug Lag:

$$\mathcal{L}_A(x, \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{m} c_i(x)^2$$

ALG. 17.4 for bound-constrained Lag. method.

NOTE: - Projected Gradient for QP
  - Projected grad. for LP
  - CG

- Trust-region logic for bound constrained NLP
- KKT inspired Aug lag def. and multiplier.
- Only linearly convergent

LECTURE 13

Feb 25 2026

Probability basic facts & Stochastic Gradient Descent.

LECTURE 14 — STOCHASTIC GRADIENT

Mar 2 2026

PROBLEM: $f: \mathbb{R}^n \to \mathbb{R}$ smooth & convex.

$$\min_{x \in \mathbb{R}^n} f(x).$$

Use $g(x, \xi)$ unbiased estimator of $\nabla f(x)$:

$$\nabla f(x) = \mathbb{E}_{\xi}\left[g(x, \xi) \right].$$

STOCHASTIC GRADIENT DESCENT:

$$x_{k+1} = x_k - \alpha_k \, g(x_k, \xi_k), \quad \xi_k \text{ iid}, \quad \alpha_k > 0 \text{ step size}$$

$\Rightarrow$ Move in a direction that equals $-\nabla f(x_k)$ in expectation.

$\rightsquigarrow f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \rightsquigarrow$ SGD selects i_k to descent:

- Cyclic Incremental Gradient: $i_k = (k \bmod N) + 1$
- Randomized: $i_k = \text{Uniform}\{1, \dots, N\}$.

↑ usually better convergence.

- **Empirical Risk Minimization (ERM)**

$$R[f] = \mathbb{E}_{(x,y)\sim P}\left[\underbrace{l}_{\text{loss}}(f(x), y)\right]$$

$$R_{emp}[f] = \frac{1}{N}\sum_{i=1}^{N} l(f(x_i), y_i).$$

$$\nabla f(x) = \mathbb{E}[g(x, \xi)]$$

- COMPUTATIONAL COST: full gradient: $O(N)$.
  SGD: $O(1)$ ← good

Upshot:
- SGD trades accuracy per step for very cheap iterations:
  convergence is slower $\rightarrow O\left(\frac{1}{\sqrt{k}}\right)$ instead of linear.

Ex: $f(x) = \frac{1}{2N}\sum_{i=1}^{N} (x - w_i)^2$

$$\nabla f_i(x) = x - w_i, \quad \alpha_k = \frac{1}{k+1}.$$

$$x_{k+1} = \frac{k}{k+1} x_k + \frac{1}{k+1} \omega_{k+1}$$

$$\Rightarrow \boxed{x_k = \frac{1}{k} \sum_{j=1}^{k} \omega_j}$$

$$f(x) = \frac{1}{2} \mathbb{E}\left[(x-\omega)^2\right]$$

$$\Rightarrow \quad x^* = \mathbb{E}[\omega] = \operatorname*{argmin}_x f(x) .$$

Note: $f(x_k) - f(x^*) = \frac{\mathrm{Var}(\omega)}{2k} = O\left(\frac{1}{k}\right)$.

ROBBINS – MONRO CONDITIONS:

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \rightsquigarrow \text{ reachability}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty \quad \rightsquigarrow \text{ noise suppression}$$

- RANDOMIZED KACZMARZ : least squares with

$$f(x) = \frac{1}{2N} \sum_{i=1}^{N} (a_i^T x - b_i)^2 .$$

- Assume: sol^n x^* exists

- Update: $x_{k+1} = x_k - a_{i_k} (a_{i_k}^T x_k - b_{i_k})$.

Note: $x_{k+1} - x^* = \underbrace{(\mathbb{1} - a_{i_k} a_{i_k}^T)}_{M_{i_k}} (x_k - x^*)$.

$$\Rightarrow \quad x_{k+1} - x^* = M_{i_k} (x_k - x^*) .$$

- Assume: $\mathbb{E}[M_i] = \rho \, \mathbb{1}$, $0 < \rho < 1$.

$$\Rightarrow \quad \mathbb{E}[x_{k+1} - x^*] = \rho (x_k - x^*)$$

$$\Rightarrow \quad \mathbb{E}[x_k - x^*] = \rho^k (x_0 - x^*) .$$

Now, on the other hand: by independence

$$\mathbb{E}[x_{k+1} - x^*] = \mathbb{E}[M_{i_k}] \, \mathbb{E}[x_k - x^*] .$$

$$\Rightarrow \quad \mathbb{E}\left[x_k - x^*\right] = \left(\mathbb{E}[M_i]\right)^k (x_0 - x^*)$$

$$= \rho^k (x_0 - x^*).$$

□

KEY ASSUMPTIONS :

$$\mathbb{E}\left[\|g(x,\xi)\|^2\right] \leq L_g^2 \|x - x^*\|^2 + B^2$$

- $B > 0 \Rightarrow$ persistent noise $\Rightarrow$ sublinear
- $B = 0 \Rightarrow$ vanishing noise $\Rightarrow$ linear possible

Note : . Persistent noise $\Rightarrow O\left(\frac{1}{k}\right)$ limit

. Vanishing noise $\Rightarrow$ possible linear convergence

SGD: $x_{k+1} = x_k - \alpha_k\, g(x_k, \xi_k)$,

$\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x)$

for a convex objective f.

CONVERGENCE: Must control the size of the stochastic directions $g(x,\xi)$:

- if g is too noisy, descent info. is swamped.
- need to find a bound for: growth w/ distance versus noise floor.
- Assume $\exists\, L_g \geq 0$ and $B \geq 0$ constants s.t.

$$\mathbb{E}_\xi\left[\|g(x,\xi)\|^2\right] \leq L_g^2 \|x - x^*\|^2 + B^2 \quad \forall x.$$

2nd MOMENT ASSUMPTION

L_g: controls how $\|g\|$ can grow as we move away from x^*

B: noise floor (nonzero when randomness persists at x^*)

NOTE: Unbiasedness: $\nabla f(x) = \mathbb{E}_\xi[g(x,\xi)]$

Jensen: $\|\nabla f(x)\|^2 = \|\mathbb{E}_\xi[g(x,\xi)]\|^2$

$\leq \mathbb{E}_\xi\left[\|g(x,\xi)\|^2\right]$.

$$\Rightarrow \boxed{\|\nabla f(x)\|^2 \le L_g^2 \|x - x^*\|^2 + B^2}$$

* Logistic Regression: For $y_i \in \{0,1\}$

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \left(-y_i a_i^T x + \log\left(\underbrace{1 + \exp(a_i^T x)}_{>0 \ \forall x}\right)\right)$$

- $\xi = i$ uniformly $\{1, \dots, N\}$, the stochast gradient is

$$g(x, i) = -y_i a_i + \frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)} a_i$$

$\sigma(t) := \frac{e^t}{1+e^t} \in (0,1)$. Thus

$$g(x,i) = (-y_i + \sigma(a_i^T x)) a_i$$

$$\Rightarrow \|g(x,i)\| \le |-y_i + \sigma(a_i^T x)| \, \|a_i\|$$

$$\le \|a_i\|$$

i.e., $\|g(x,i)\|^2 \le \|a_i\|^2$

i.e., $\mathbb{E}_\xi[\|g(x,\xi)\|^2] \le \max_{1 \le i \le N} \|a_i\|^2$.