Outline

1. Optimization Problems
2. One-Dimensional Optimization
3. Multi-Dimensional Optimization
Optimization

- Given function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), and set \( S \subseteq \mathbb{R}^n \), find \( x^* \in S \) such that \( f(x^*) \leq f(x) \) for all \( x \in S \).

- \( x^* \) is called **minimizer** or **minimum** of \( f \).

- It suffices to consider only minimization, since maximum of \( f \) is minimum of \( -f \).

- **Objective** function \( f \) is usually differentiable, and may be linear or nonlinear.

- **Constraint** set \( S \) is defined by system of equations and inequalities, which may be linear or nonlinear.

- Points \( x \in S \) are called **feasible** points.

- If \( S = \mathbb{R}^n \), problem is **unconstrained**.
General continuous optimization problem:

\[
\min f(x) \quad \text{subject to} \quad g(x) = 0 \quad \text{and} \quad h(x) \leq 0
\]

where \( f : \mathbb{R}^n \to \mathbb{R}, \ g : \mathbb{R}^n \to \mathbb{R}^m, \ h : \mathbb{R}^n \to \mathbb{R}^p \)

- **Linear programming**: \( f, g, \) and \( h \) are all linear
- **Nonlinear programming**: at least one of \( f, g, \) and \( h \) is nonlinear
Examples: Optimization Problems

- Minimize weight of structure subject to constraint on its strength, or maximize its strength subject to constraint on its weight
- Minimize cost of diet subject to nutritional constraints
- Minimize surface area of cylinder subject to constraint on its volume:

\[
\min_{x_1, x_2} f(x_1, x_2) = 2\pi x_1 (x_1 + x_2)
\]

subject to \( g(x_1, x_2) = \pi x_1^2 x_2 - V = 0 \)

where \( x_1 \) and \( x_2 \) are radius and height of cylinder, and \( V \) is required volume
Local vs Global Optimization

- \( x^* \in S \) is **global minimum** if \( f(x^*) \leq f(x) \) for all \( x \in S \)

- \( x^* \in S \) is **local minimum** if \( f(x^*) \leq f(x) \) for all feasible \( x \) in some neighborhood of \( x^* \)
Global Optimization

- Finding, or even verifying, global minimum is difficult, in general.

- Most optimization methods are designed to find local minimum, which may or may not be global minimum.

- If global minimum is desired, one can try several widely separated starting points and see if all produce same result.

- For some problems, such as linear programming, global optimization is more tractable.
Existence of Minimum

- If $f$ is continuous on closed and bounded set $S \subseteq \mathbb{R}^n$, then $f$ has global minimum on $S$.

- If $S$ is not closed or is unbounded, then $f$ may have no local or global minimum on $S$.

- Continuous function $f$ on unbounded set $S \subseteq \mathbb{R}^n$ is **coercive** if

  \[
  \lim_{\|x\| \to \infty} f(x) = +\infty
  \]

  i.e., $f(x)$ must be large whenever $\|x\|$ is large.

- If $f$ is coercive on closed, unbounded set $S \subseteq \mathbb{R}^n$, then $f$ has global minimum on $S$. 
An Example of a Coercive Function

\[
x = -1.6:1:1.5; \quad y = -1.6:1:1.7;
\]
\[
[X,Y]=\text{ndgrid}(x,y);
\]
\[
u = -3*\exp(-X.*X-Y.*Y).*\sin(3*X-Y) + 1.2*(X-Y).^2+(X+Y).^2;
\]

Goes to \(\infty\) as \(||x|| \to \infty\)

Level Sets (contours, isosurfaces)
An Example of a Non-Coercive Function

- What’s wrong with this case?

\[ f(x) = (x_1 - x_2)^2 \]
Level Sets

- **Level set** for function $f : S \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is set of all points in $S$ for which $f$ has some given constant value.

- For given $\gamma \in \mathbb{R}$, **sublevel set** is
  
  $$L_\gamma = \{ x \in S : f(x) \leq \gamma \}$$

- If continuous function $f$ on $S \subseteq \mathbb{R}^n$ has nonempty sublevel set that is closed and bounded, then $f$ has global minimum on $S$.

- If $S$ is unbounded, then $f$ is coercive on $S$ if, and only if, all of its sublevel sets are bounded. *(No open contours…)*
Uniqueness of Minimum

- Set $\mathcal{S} \subseteq \mathbb{R}^n$ is **convex** if it contains line segment between any two of its points.

- Function $f : \mathcal{S} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is **convex** on convex set $\mathcal{S}$ if its graph along any line segment in $\mathcal{S}$ lies on or below chord connecting function values at endpoints of segment.

- Any local minimum of convex function $f$ on convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is global minimum of $f$ on $\mathcal{S}$.

- Any local minimum of **strictly** convex function $f$ on convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is **unique** global minimum of $f$ on $\mathcal{S}$.
Convex Domains and Convex Functions

- Convex
- Nonconvex
- Convex
- Nonconvex

- Nonconvex
- Convex
- Strictly convex
First-Order Optimality Condition

- For function of one variable, one can find extremum by differentiating function and setting derivative to zero.

- Generalization to function of $n$ variables is to find critical point, i.e., solution of nonlinear system

$$\nabla f(x) = 0$$

where $\nabla f(x)$ is gradient vector of $f$, whose $i$th component is $\partial f(x)/\partial x_i$.

- For continuously differentiable $f : S \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$, any interior point $x^*$ of $S$ at which $f$ has local minimum must be critical point of $f$.

- But not all critical points are minima: they can also be maxima or saddle points.
First-Order Optimality Condition

- But not all critical points are minima: they can also be maxima or saddle points

- Saddle points in higher dimensions are more common than zero-slope inflection points in 1D

Note the open contours…. 
First-Order Optimality Condition

- For function of one variable, one can find extremum by differentiating function and setting derivative to zero.
- Generalization to function of \( n \) variables is to find critical point, i.e., solution of nonlinear system

\[
\nabla f(x) = 0
\]

- We use the solution techniques of Chapter 5 to solve this nonlinear system.
Second-Order Optimality Condition

- For twice continuously differentiable $f: S \subseteq \mathbb{R}^n \to \mathbb{R}$, we can distinguish among critical points by considering the Hessian matrix $H_f(x)$ defined by

$$\{H_f(x)\}_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

which is symmetric.

- At critical point $x^*$, if $H_f(x^*)$ is
  - positive definite, then $x^*$ is minimum of $f$
  - negative definite, then $x^*$ is maximum of $f$
  - indefinite, then $x^*$ is saddle point of $f$
  - singular, then various pathological situations are possible.
Classifying Critical Points

- Example 6.5 from the text:

\[ f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6. \]

- Gradient:

\[ \nabla f(x) = \begin{bmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{bmatrix}. \]

- Hessian is symmetric:

\[
H_f = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\
\frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_2}
\end{bmatrix}
= \begin{bmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{bmatrix}.
\]

- Critical points, solutions to \( \nabla f(x) = 0 \) are

\[
x^* = [1, -1]^T, \quad \text{and} \quad x^* = [2, -3]^T.
\]
• At first critical point, $x^* = [1, -1]^T$,

$$H_f(1, -1) = \begin{bmatrix} 18 & 12 \\ 12 & 6 \end{bmatrix},$$

does not have positive eigenvalues, so $[1, -1]^T$ is a saddle point.

• At second critical point, $x^* = [2, -3]^T$,

$$H_f(2, -3) = \begin{bmatrix} 30 & 12 \\ 12 & 6 \end{bmatrix},$$

does have positive eigenvalues, so $[2, -3]^T$ is a local minimum.
Example 6.5 from text

\[ f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6. \]
Example 6.5 from text

\[ f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6. \]

- First critical point, \( x^* = [1, -1]^T \), is a saddle point. Hessian not SPD.
Example 6.5 from text

\[ f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6. \]

- First critical point, \( x^* = [1, -1]^T \), is a saddle point. Hessian not SPD.
Example 6.5 from text

\[ f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6. \]

- Second critical point, \( x^* = [2, -3]^T \), is a local minimum. Hessian is SPD.
Second-Order Optimality Condition

- For twice continuously differentiable \( f : S \subseteq \mathbb{R}^n \to \mathbb{R} \), we can distinguish among critical points by considering the **Hessian matrix** \( H_f(x) \) defined by:

\[
\{H_f(x)\}_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}
\]

which is symmetric.

- At critical point \( x^* \), if \( H_f(x^*) \) is:
  - positive definite, then \( x^* \) is minimum of \( f \)
  - negative definite, then \( x^* \) is maximum of \( f \)
  - indefinite, then \( x^* \) is saddle point of \( f \)
  - singular, then various pathological situations are possible
Example of Singular Hessian in 1D

- Consider,

\[ f(x) = (x - 1)^4 \]

\[ \nabla f = \frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial x} = 4(x - 1)^3 = 0 \text{ when } x = x^* = 1. \]

\[ H_{ij} = \frac{\partial^2 f}{\partial x_i x_j} = \frac{\partial^2 f}{\partial x^2} = 12(x - 1)^2 = 0. \]

Here, the Hessian is singular at \( x = x^* \) and \( x^* \) is a minimizer of \( f \).

- However, if

\[ f(x) = (x - 1)^3 \]

\[ \nabla f = \frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial x} = 3(x - 1)^2 = 0 \text{ when } x = x^* = 1. \]

\[ H_{ij} = \frac{\partial^2 f}{\partial x_i x_j} = \frac{\partial^2 f}{\partial x^2} = 6(x - 1) = 0. \]

Here, the Hessian is singular at \( x = x^* \) and \( x^* \) is a not a minimizer of \( f \).
Constrained Optimality

- If problem is constrained, only \textit{feasible} directions are relevant.

- For equality-constrained problem

\[
\min f(x) \quad \text{subject to} \quad g(x) = 0
\]

where \( f: \mathbb{R}^n \to \mathbb{R} \) and \( g: \mathbb{R}^n \to \mathbb{R}^m \), with \( m \leq n \), necessary condition for feasible point \( x^* \) to be solution is that negative gradient of \( f \) lie in space spanned by constraint normals,

\[
-\nabla f(x^*) = J_g^T(x^*)\lambda
\]

where \( J_g \) is Jacobian matrix of \( g \), and \( \lambda \) is vector of \textit{Lagrange multipliers}.

- This condition says we cannot reduce objective function without violating constraints.
Constrained Optimality, continued

- **Lagrangian function** \( \mathcal{L}: \mathbb{R}^{n+m} \to \mathbb{R} \), is defined by

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

- Its gradient is given by

\[
\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix}
\nabla f(x) + J_g^T(x) \lambda \\
g(x)
\end{bmatrix}
\]

- Its Hessian is given by

\[
H_{\mathcal{L}}(x, \lambda) = \begin{bmatrix}
B(x, \lambda) & J_g^T(x) \\
J_g(x) & 0
\end{bmatrix}
\]

where

\[
B(x, \lambda) = H_f(x) + \sum_{i=1}^{m} \lambda_i H_{g_i}(x)
\]
Equality Constraint: $g(x) = 0$

- Equality constraint:
  \[
  g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix} = 0.
  \]

- Lagrangian: $\mathcal{L}$:
  \[
  \mathcal{L} = f(x) + g(x)^T \lambda = f(x) + \sum_{i=1}^{m} g_i(x)^T \lambda_i.
  \]

- Gradient:
  \[
  \nabla \mathcal{L} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial x_1} \\ \frac{\partial \mathcal{L}}{\partial x_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial x_n} \\ \frac{\partial \mathcal{L}}{\partial \lambda_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \lambda_m} \end{pmatrix} = \begin{pmatrix} \nabla f + \sum_{i=1}^{m} \frac{\partial g_i}{\partial x_i} \lambda_i \\ g_1 \\ g_2 \\ \vdots \\ g_m \end{pmatrix} = \begin{pmatrix} \nabla f + J_g^T \lambda \\ g(x) \end{pmatrix}.
  \]
• Hessian:

\[ \nabla H_\mathcal{L}(x, \lambda) = \begin{pmatrix} H_f + \sum_{i=1}^{m} H_{g_i} \lambda_i & J_g^T \\ J_g & 0 \end{pmatrix} = \begin{pmatrix} B(x, \lambda) & J_g^T \\ J_g & 0 \end{pmatrix} \]
Together, necessary condition and feasibility imply critical point of Lagrangian function,

\[ \nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + J_g^T(x)\lambda \\ g(x) \end{bmatrix} = 0 \]
Equality Constraints: Case $g(x)=$scalar.

$$\min f(x) \text{ subject to } g(x) = 0.$$ 

- Lagrangian

$$\mathcal{L}(x, \lambda) := f(x) + \lambda g(x)$$
$$\lambda := \text{Lagrange multiplier}$$

$$\nabla \mathcal{L} = \left( \begin{array}{c} \frac{\partial \mathcal{L}}{\partial x_i} \\ \frac{\partial \mathcal{L}}{\partial \lambda} \end{array} \right) = \left( \begin{array}{c} \nabla f + \lambda \nabla g \\ g \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \end{array} \right)$$

at critical point $(x^*, \lambda^*)$.

- Note, $\nabla f(x^*) \neq 0$.

- Instead, $\nabla f(x^*) = -\lambda^* \nabla g(x^*)$.

- The gradient of $f$ at $x^*$ is a multiple of the gradient of $g$. 
Constrained Optimality Example, \( g(\mathbf{x}) \) a Scalar.

- **Comment on:** “This condition says we cannot reduce objective function without violating constraints.”
- **A key point is that:** \( \mathbf{x}^*, \nabla f \) is parallel to \( \nabla g \)
- **If there are multiple constraints, then** \( \nabla f \) in \( \text{span}\{\nabla g_k\} = J_g^T \lambda \)

\[
-\nabla f(\mathbf{x}^*) = J_g^T(\mathbf{x}^*) \lambda \\
-\frac{\partial f}{\partial x_i} = \sum_{k=1}^{m} \frac{\partial g_k}{\partial x_i} \lambda_k
\]
Constrained Optimality Example, $g(x)$ a Scalar.

- Here, we see the gradients of $f$ and $g$ at a point that is not $x^*$.
- Clearly, we can make more progress in reducing $f(x)$ by moving along the $g(x)=0$ contour until $\nabla f$ is parallel to $\nabla g$. 
Matlab examples

- const_examp.m
- co_const.m
Example: Two Equality Constraints

\[ \min f(x) \text{ subject to } g_1(x) = 0 \text{ and } g_2(x) = 0. \]

- Lagrangian
  \[ \mathcal{L}(x, \lambda) := f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x). \]

- First-order conditions
  \[ \nabla \mathcal{L} = \left( \begin{array}{c} \frac{\partial \mathcal{L}}{\partial x_i} \\ \frac{\partial \mathcal{L}}{\partial \lambda_k} \end{array} \right) = \left( \begin{array}{c} \frac{\partial f}{\partial x_i} + \sum_k \lambda_k \frac{\partial g_k}{\partial x_i} \\ 0 + \left( \begin{array}{c} g_1 \\ g_2 \end{array} \right) \end{array} \right) \]
  \[ = \left( \begin{array}{c} \nabla f + J_g^T \lambda \\ g \end{array} \right) = 0. \]
Constrained Optimality, continued

- Together, necessary condition and feasibility imply critical point of Lagrangian function,

\[
\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + J_g^T(x)\lambda \\ g(x) \end{bmatrix} = 0
\]

- Hessian of Lagrangian is symmetric, but not positive definite, so critical point of \( \mathcal{L} \) is saddle point rather than minimum or maximum

- Critical point \((x^*, \lambda^*)\) of \( \mathcal{L} \) is constrained minimum of \( f \) if \( B(x^*, \lambda^*) \) is positive definite on null space of \( J_g(x^*) \)

- If columns of \( Z \) form basis for null space, then test projected Hessian \( Z^T B Z \) for positive definiteness
Together, necessary condition and feasibility imply critical point of Lagrangian function,

$$\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + J_g^T(x)\lambda \\ g(x) \end{bmatrix} = 0$$

Hessian of Lagrangian is symmetric, but not positive definite, so critical point of $\mathcal{L}$ is saddle point rather than minimum or maximum.

Critical point $(x^*, \lambda^*)$ of $\mathcal{L}$ is constrained minimum of $f$ if $B(x^*, \lambda^*)$ is positive definite on null space of $J_g(x^*)$

- **Note:** null space of $J_g(x^*)$ is the set of vectors tangent to the constraint surface (or surfaces, if more than one $g$ present).
Constrained Optimality Example, $g(x)$ a Scalar.

- Comment on: “This condition says we cannot reduce objective function without violating constraints.”

- A key point is that, $x^*$, $\nabla f$ is parallel to $\nabla g$

- If there are multiple constraints, then $\nabla f$ in $\text{span}\{\nabla g_k\} = J_g^T \lambda$

\[
-\nabla f(x^*) = J_g^T(x^*) \lambda
\]

\[
-\frac{\partial f}{\partial x_i} = \sum_{k=1}^{m} \frac{\partial g_k}{\partial x_i} \lambda_k
\]
Constrained Optimality, continued

- If inequalities are present, then KKT optimality conditions also require nonnegativity of Lagrange multipliers corresponding to inequalities, and complementarity condition

Karush-Kuhn-Tucker

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

\[ g(x^*) = 0, \]

\[ h(x^*) \leq 0, \]

\[ \lambda_i^* \geq 0, \quad i = m + 1, \ldots, m + p \quad \text{inequality Lagrange multipliers} \]

\[ \lambda_i^* h_i(x^*) = 0, \quad i = m + 1, \ldots, m + p \quad \text{complementarity condition}. \]
Inequality Constraint and Sign of $\lambda$

- A key point is that, at $x^*$, $\nabla f$ is parallel to $\nabla h$
- If there are multiple constraints, then $\nabla f$ in $\text{span}\{\nabla h_k\} = J_h^T \lambda$

$$-\nabla f(x^*) = J_h^T(x^*)\lambda$$

$$-\frac{\partial f}{\partial x_i} = \sum_{k=1}^{m} \frac{\partial h_k}{\partial x_i} \lambda_k$$

$h(x) = 0$

$h(x) < 0$
Inequality Constraint and Sign of $\lambda$

- If constraint is active ($h(x^*) = 0$) then $\nabla f$ and $\nabla h$ point in opposite directions.
- $\nabla f + \lambda^* \nabla h = 0$.
- $\lambda^* > 0$.

\[
\begin{align*}
-\nabla f(x^*) &= J_h^T(x^*)\lambda \\
-\frac{\partial f}{\partial x_i} &= \sum_{k=1}^{m} \frac{\partial h_k}{\partial x_i} \lambda_k
\end{align*}
\]
Sensitivity and Conditioning

- Function minimization and equation solving are closely related problems, but their sensitivities differ.

- In one dimension, absolute condition number of root $x^*$ of equation $f(x) = 0$ is $1/|f'(x^*)|$, so if $|f(\hat{x})| \leq \epsilon$, then $|\hat{x} - x^*|$ may be as large as $\epsilon/|f'(x^*)|$

- For minimizing $f$, Taylor series expansion

\[
f(\hat{x}) = f(x^* + h) = f(x^*) + f'(x^*)h + \frac{1}{2} f''(x^*)h^2 + O(h^3)
\]

shows that, since $f'(x^*) = 0$, if $|f(\hat{x}) - f(x^*)| \leq \epsilon$, then $|\hat{x} - x^*|$ may be as large as $\sqrt{2\epsilon}/|f''(x^*)|$

- Thus, based on function values alone, minima can be computed to only about half precision.
Consider $f=1-\cos(x)$
Consider $f=1-\cos(x)$
Consider \( f = 1 - \cos(x) \)

- As you zoom in, \( f(x) \to 0 \) quadratically, but \( x \to x^* \) only linearly.

- So, if \( |f(x) - f(x^*)| \approx \epsilon \), then \( \|x - x^*\| = O(\epsilon^{1/2}) \)
Consider $f = 1 - \cos(x)$

- As you zoom in, $f(x) \to 0$ quadratically, but $x \to x^*$ only linearly

- So, if $|f(x) - f(x^*)| \approx \epsilon$, then $\|x - x^*\| = O(\epsilon^{1/2})$
Sensitivity of Minimization

Terminate search when

\[ |f(x^* + \Delta x) - f(x^*)| \leq \varepsilon. \]

Taylor series:

\[ f(x^* + \Delta x) = f(x^*) + \Delta x f'(x^*) + \frac{\Delta x^2}{2} f''(x^*) + O(\Delta x^3) \]

\[ \frac{\Delta x^2}{2} \approx \frac{f(x^* + \Delta x) - f(x^*)}{f''(x^*)} \]

\[ |\Delta x| \approx \sqrt{\frac{2\varepsilon}{|f''(x^*)|}} \]

• So, if \( \varepsilon \approx \varepsilon_M \), can expect accuracy to approximately \( \sqrt{\varepsilon_M} \).

• Question: Is a small \( f'' \) good? Or bad?
Example: Minimize Cost over Some Design Parameter, $x$
Example: Minimize Cost over Some Design Parameter, $x$

While having $f''(x^*)$ small makes it difficult to find the \textit{optimal} $x^*$, it in fact is a happy circumstance because it gives you liberty to add additional constraints at no cost.

So, often, having a broad minimum in practice is \textit{good}. 

\begin{itemize}
  \item While having $f''(x^*)$ small makes it difficult to find the \textit{optimal} $x^*$, it in fact is a happy circumstance because it gives you liberty to add additional constraints at no cost.
  \item So, often, having a broad minimum in practice is \textit{good}.
\end{itemize}
Methods for One-Dimensional Problems

- Demonstrate:
  - basic techniques
  - bracketing
  - convergence rates

- Useful for *line search* in multi-dimensional problems
For minimizing function of one variable, we need “bracket” for solution analogous to sign change for nonlinear equation.

Real-valued function $f$ is *unimodal* on interval $[a, b]$ if there is unique $x^* \in [a, b]$ such that $f(x^*)$ is minimum of $f$ on $[a, b]$, and $f$ is strictly decreasing for $x < x^*$, strictly increasing for $x^* < x$.

Unimodality enables discarding portions of interval based on sample function values, analogous to interval bisection.
Golden Section Search

Optimization Problems

One-Dimensional Optimization

Multi-Dimensional Optimization

Golden Section Search, continued

\[ \tau = \left( \frac{\varphi - 1}{2} \right) \]

\[ x_1 = a + (1 - \tau)(b - a) \]

\[ f_1 = f(x_1) \]

\[ x_2 = a + \tau(b - a) \]

\[ f_2 = f(x_2) \]

\[ f_1 > f_2 \]

\[ \text{while } (b - a) > \text{tol} \]

\[ \text{if } f_1 > f_2 \]

\[ a = x_1 \]

\[ x_1 = x_2 \]

\[ f_1 = f_2 \]

\[ x_2 = a + \tau(b - a) \]

\[ f_2 = f(x_2) \]

\[ \text{else} \]

\[ b = x_2 \]

\[ x_2 = x_1 \]

\[ f_2 = f_1 \]

\[ x_1 = a + (1 - \tau)(b - a) \]

\[ f_1 = f(x_1) \]

\[ \text{end} \]

\[ \text{end} \]
Golden Section Search

\[ \tau = \frac{\phi}{2} = \frac{1}{2} (1 + \sqrt{5}) \]

\[ x_1 = a + (1 - \tau)(b - a) \]

\[ f_1 = f(x_1) \]

\[ x_2 = a + \tau(b - a) \]

\[ f_2 = f(x_2) \]

while \((b - a) > \text{tol}) do

if \(f_1 > f_2\) then

\[ a = x_1 \]

\[ x_1 = x_2 \]

\[ f_1 = f(x_1) \]

else

\[ b = x_2 \]

\[ x_2 = x_1 \]

\[ f_2 = f(x_2) \]

end

end
Golden Section Search

\[ \alpha = \frac{\phi - 1}{2} \]
\[ x_1 = a + \alpha (b - a) \]
\[ f_1 = f(x_1) \]

while \((b - a) > \text{tol} \)

if \( f_1 > f_2 \)
\[ a = x_1 \]
\[ x_1 = x_2 \]
\[ f_1 = f_2 \]
\[ x_2 = a + \alpha (b - a) \]
\[ f_2 = f(x_2) \]

else
\[ b = x_2 \]
\[ x_2 = x_1 \]
\[ f_2 = f(x_2) \]
\[ x_1 = a + \alpha (b - a) \]
\[ f_1 = f(x_1) \]

end

Michael T. Heath
Scientific Computing 21 / 74

\( f_1 > f_2 \)
Golden Section Search

New sample pattern, \([a,x_1,x_2,b]\) is similar to original.
Golden Section Search

- Suppose $f$ is unimodal on $[a, b]$, and let $x_1$ and $x_2$ be two points within $[a, b]$, with $x_1 < x_2$.

- Evaluating and comparing $f(x_1)$ and $f(x_2)$, we can discard either $(x_2, b]$ or $[a, x_1)$, with minimum known to lie in remaining subinterval.

- To repeat process, we need compute only one new function evaluation.

- To reduce length of interval by fixed fraction at each iteration, each new pair of points must have same relationship with respect to new interval that previous pair had with respect to previous interval.
To accomplish this, we choose relative positions of two points as $\tau$ and $1 - \tau$, where $\tau^2 = 1 - \tau$, so

$$\tau = \left(\sqrt{5} - 1\right)/2 \approx 0.618 \text{ and } 1 - \tau \approx 0.382$$

Whichever subinterval is retained, its length will be $\tau$ relative to previous interval, and interior point retained will be at position either $\tau$ or $1 - \tau$ relative to new interval.

To continue iteration, we need to compute only one new function value, at complementary point.

This choice of sample points is called golden section search.

Golden section search is safe but convergence rate is only linear, with constant $C \approx 0.618$.
Golden Section Search

- \( f(x) \) unimodal on \([a,b]\).
- Subdivide \([a,b]\) into 3 parts.
- If \( f_1 > f_2 \) discard \((a,x_1)\)
- Choose new point in larger of remaining two intervals: \((x_1,x_2)\) or \((x_2,b)\).
  - \( \rightarrow x_1 \) and \( x_2 \) should be closer to center and not at 1/3, 2/3.

![Diagram of Golden Section Search]

\( a \) \hspace{1cm} \( x_1 \) \hspace{1cm} \( x_2 \) \hspace{1cm} \( b \)

\( f_1 > f_2 \)
Golden Section Geometry

- Want new section \([1-\tau, \tau]\) to have same relation as \([0, 1-\tau]\) to \([0, 1]\).

\[
\frac{\tau - (1 - \tau)}{\tau} = \frac{1 - \tau}{1}
\]

\[
2\tau - 1 = \tau - \tau^2
\]

\[
\tau^2 + \tau - 1 = 0
\]

\[
\tau = \frac{-1 + \sqrt{1+4}}{2} = \frac{\sqrt{5} - 1}{2} = 0.618
\]
Golden Section Search, continued

\[ \tau = \left(\sqrt{5} - 1\right) / 2 \]

\[ x_1 = a + (1 - \tau)(b - a); \quad f_1 = f(x_1) \]

\[ x_2 = a + \tau(b - a); \quad f_2 = f(x_2) \]

while \((b - a) > \text{tol}\) do

if \((f_1 > f_2)\) then

\[ a = x_1 \]
\[ x_1 = x_2 \]
\[ f_1 = f_2 \]
\[ x_2 = a + \tau(b - a) \]
\[ f_2 = f(x_2) \]

else

\[ b = x_2 \]
\[ x_2 = x_1 \]
\[ f_2 = f_1 \]
\[ x_1 = a + (1 - \tau)(b - a) \]
\[ f_1 = f(x_1) \]

end

end

Michael T. Heath

Scientific Computing
Example: Golden Section Search

Use golden section search to minimize

\[ f(x) = 0.5 - x \exp(-x^2) \]
<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$f_1$</th>
<th>$x_2$</th>
<th>$f_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.764</td>
<td>0.074</td>
<td>1.236</td>
<td>0.232</td>
</tr>
<tr>
<td>0.472</td>
<td>0.122</td>
<td>0.764</td>
<td>0.074</td>
</tr>
<tr>
<td>0.764</td>
<td>0.074</td>
<td>0.944</td>
<td>0.113</td>
</tr>
<tr>
<td>0.652</td>
<td>0.074</td>
<td>0.764</td>
<td>0.074</td>
</tr>
<tr>
<td>0.584</td>
<td>0.085</td>
<td>0.652</td>
<td>0.074</td>
</tr>
<tr>
<td>0.652</td>
<td>0.074</td>
<td>0.695</td>
<td>0.071</td>
</tr>
<tr>
<td>0.695</td>
<td>0.071</td>
<td>0.721</td>
<td>0.071</td>
</tr>
<tr>
<td>0.679</td>
<td>0.072</td>
<td>0.695</td>
<td>0.071</td>
</tr>
<tr>
<td>0.695</td>
<td>0.071</td>
<td>0.705</td>
<td>0.071</td>
</tr>
<tr>
<td>0.705</td>
<td>0.071</td>
<td>0.711</td>
<td>0.071</td>
</tr>
</tbody>
</table>
**Successive Parabolic Interpolation**

- Fit quadratic polynomial to three function values
- Take minimum of quadratic to be new approximation to minimum of function

![Graph showing successive parabolic interpolation]

- New point replaces oldest of three previous points and process is repeated until convergence
- Convergence rate of successive parabolic interpolation is superlinear, with $r \approx 1.324$
Use successive parabolic interpolation to minimize

\[ f(x) = 0.5 - x \exp(-x^2) \]
### Example, continued

<table>
<thead>
<tr>
<th>$x_k$</th>
<th>$f(x_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.500</td>
</tr>
<tr>
<td>0.600</td>
<td>0.081</td>
</tr>
<tr>
<td>1.200</td>
<td>0.216</td>
</tr>
<tr>
<td>0.754</td>
<td>0.073</td>
</tr>
<tr>
<td>0.721</td>
<td>0.071</td>
</tr>
<tr>
<td>0.692</td>
<td>0.071</td>
</tr>
<tr>
<td>0.707</td>
<td>0.071</td>
</tr>
</tbody>
</table>

Successive Parabolic Interpolation is Newton’s Method applied to a quadratic model of the data. $r = 1.324$

We turn to Newton’s method next, $r = 2$. 

---

*Not monotone*  
*Superlinear*
function x=parab(a,b,c);

fa=f(a); fb=f(b); fc=f(c); x=1; fx=f(x);

for k=1:15;
    num = ((fb-fc)*(b-a)^2-(fb-fa)*(b-c)^2); % m'
    den = 2*((fb-fc)*(b-a)-(fb-fa)*(b-c)); % m''
    x=b-num/den; fx = f(x);

    c=b; fc=fb; b=a; fb=fa; a=x; fa=fx; % Push c off
end;
function x=parab(a,b,c);
    xx=a:.001:c; yy=f(xx);
    hold off; plot(xx,yy,'r-'); hold on;

    format compact; format longe;
    fa=f(a); fb=f(b); fc=f(c); x=1; fx=f(x);

    for k=1:15;
        xo=x; fo=fx;

        num = ((fb-fc)*(b-a)^2-(fb-fa)*(b-c)^2); % m'
        den = 2*((fb-fc)*(b-a)-(fb-fa)*(b-c)); % m''
        x=b-num/den; fx = f(x);

        c=b; fc=fb; b=a; fb=fa; a=x; fa=fx; % Push c off
        dx=xo-x; df=fo-fx; [ x fx dx df ] % Plot

        kk(k)=k; dk(k)=abs(dx);fk(k)=abs(df);
        plot(x,fx,'kx'); hold on
        if abs(df) < 20*eps; break; end;

    end;
    pause
    hold off; semilogy(kk,dk,'r.-',kk,fk,'b.-')

    function fx=f(x);
    exx = exp(-x.*x); fx = .5 - x.*exx;
    fx = 1./sin(x);
    fx = exp(x) + 1./x;
Newton’s Method

- Another local quadratic approximation is truncated Taylor series
  
  \[ f(x + h) \approx f(x) + f'(x)h + \frac{f''(x)}{2} h^2 \]

- By differentiation, minimum of this quadratic function of \( h \) is given by \( h = -f'(x) / f''(x) \)

- Suggests iteration scheme
  
  \[ x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \]

  which is Newton’s method for solving nonlinear equation \( f'(x) = 0 \)

- Newton’s method for finding minimum normally has quadratic convergence rate, but must be started close enough to solution to converge

  newton1d.m
Example: Newton’s Method

- Use Newton’s method to minimize $f(x) = 0.5 - x \exp(-x^2)$

- First and second derivatives of $f$ are given by
  
  $f'(x) = (2x^2 - 1) \exp(-x^2)$

  and

  $f''(x) = 2x(3 - 2x^2) \exp(-x^2)$

- Newton iteration for zero of $f'$ is given by
  
  $x_{k+1} = x_k - \frac{(2x_k^2 - 1)}{(2x_k(3 - 2x_k^2))}$

- Using starting guess $x_0 = 1$, we obtain

<table>
<thead>
<tr>
<th>$x_k$</th>
<th>$f(x_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.132</td>
</tr>
<tr>
<td>0.500</td>
<td>0.111</td>
</tr>
<tr>
<td>0.700</td>
<td>0.071</td>
</tr>
<tr>
<td>0.707</td>
<td>0.071</td>
</tr>
</tbody>
</table>
format compact; format longe;

x=1; f=0;

for k=1:10;
    fo = f; exx = exp(-x.*x); f = .5 - x.*exx;

    fp = (2.*x.*x-1).*exx;
% fpp = 2.*x.*(3-2.*x.*x).*exx;
    s = -(2.*x.*x-1)/(2.*x.*(3-2.*x.*x));
    x = x+s;

    y = f-fo; [k x f s y] % print convergence

end;
As with nonlinear equations in one dimension, slow-but-sure and fast-but-risky optimization methods can be combined to provide both safety and efficiency.

Most library routines for one-dimensional optimization are based on this hybrid approach.

Popular combination is golden section search and successive parabolic interpolation, for which no derivatives are required.
function x=parab(a,b,c);
xx=a:.001:c; yy=f(xx); hold off; plot(xx,yy,'r-');
format compact; format longe;
fa=f(a); fb=f(b); fc=f(c);
x=1; fx=f(x);
for k=1:90;
    num = ((fb-fc)*(b-a)^2-(fb-fa)*(b-c)^2);
    den = ((fb-fc)*(b-a)-(fb-fa)*(b-c));
    xo=x; fo=fx;
    x=b-.5*num/den; fx = f(x);
    if x>b;
        if fx>fb; c=x; fc=fx; else;
            a=b; fa=fb;b=x; fb=fx; end;
        else
            if fx>fb; a=x; fa=fx; else;
                c=b; fc=fb;b=x; fb=fx; end;
        end;
    dx=xo-x; df=fo-fx;
    [ x fx dx df ]
    kk(k)=k; dk(k)=abs(dx);fk(k)=abs(df);
    plot(x,fx,'kx'); hold on
    if abs(df)<20*eps; break; end;
end;
pause
hold off; semilogy(kk,dk,'r-',kk,fk,'b-')

function fx=f(x);
exx = exp(-x.*x); fx = .5 - x.*exx;
fx = 1./sin(x);
fx = exp(x) + 1./x;
Convergence Behavior

Question 1: What type of convergence is this?

Question 2: What does this plot say about conditioning of

$$\min_x f(x)$$
Methods for Multi-Dimensional Problems
Steepest Descent Method

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function of $n$ real variables.
- At any point $x$ where the gradient vector is nonzero, the negative gradient, $-\nabla f(x)$, points downhill toward lower values of $f$.
- In fact, $-\nabla f(x)$ is the locally direction of steepest descent: $f$ decreases more rapidly along the direction of negative gradient than along any other.

**Steepest descent** method: starting from initial guess $x_0$, successive approximate solutions given by

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

where $\alpha_k$ is the line search parameter that determines how far to go in the given direction.
Steepest Descent, continued

- Given descent direction, such as negative gradient, determining appropriate value for $\alpha_k$ at each iteration is one-dimensional minimization problem

$$\min_{\alpha_k} f(x_k - \alpha_k \nabla f(x_k))$$

that can be solved by methods already discussed

- Steepest descent method is very reliable: it can always make progress provided gradient is nonzero

- But method is myopic in its view of function's behavior, and resulting iterates can zigzag back and forth, making very slow progress toward solution

- In general, convergence rate of steepest descent is only linear, with constant factor that can be arbitrarily close to 1
Example, continued
Example: Steepest Descent

- Use steepest descent method to minimize
  \[ f(x) = 0.5x_1^2 + 2.5x_2^2 \]

- Gradient is given by \( \nabla f(x) = \begin{bmatrix} x_1 \\ 5x_2 \end{bmatrix} \)

- Taking \( x_0 = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \), we have \( \nabla f(x_0) = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \)

- Performing line search along negative gradient direction,
  \[ \min_{\alpha_0} f(x_0 - \alpha_0 \nabla f(x_0)) \]
  
  exact minimum along line is given by \( \alpha_0 = 1/3 \), so next approximation is \( x_1 = \begin{bmatrix} 3.333 \\ -0.667 \end{bmatrix} \)
Example, continued

<table>
<thead>
<tr>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$\nabla f(x_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.000</td>
<td>15.000</td>
<td>5.000 5.000</td>
</tr>
<tr>
<td>3.333</td>
<td>6.667</td>
<td>3.333 -3.333</td>
</tr>
<tr>
<td>2.222</td>
<td>2.963</td>
<td>2.222 2.222</td>
</tr>
<tr>
<td>1.481</td>
<td>1.317</td>
<td>1.481 -1.481</td>
</tr>
<tr>
<td>0.988</td>
<td>0.585</td>
<td>0.988 0.988</td>
</tr>
<tr>
<td>0.658</td>
<td>0.260</td>
<td>0.658 -0.658</td>
</tr>
<tr>
<td>0.439</td>
<td>0.116</td>
<td>0.439 0.439</td>
</tr>
<tr>
<td>0.293</td>
<td>0.051</td>
<td>0.293 -0.293</td>
</tr>
<tr>
<td>0.195</td>
<td>0.023</td>
<td>0.195 0.195</td>
</tr>
<tr>
<td>0.130</td>
<td>0.010</td>
<td>0.130 -0.130</td>
</tr>
</tbody>
</table>
Example, continued
Newton’s Method

- Broader view can be obtained by local quadratic approximation, which is equivalent to Newton’s method.

- In multidimensional optimization, we seek zero of gradient, so *Newton iteration* has form

\[ \mathbf{x}_{k+1} = \mathbf{x}_k - H_f^{-1}(\mathbf{x}_k) \nabla f(\mathbf{x}_k) \]

where \( H_f(\mathbf{x}) \) is *Hessian* matrix of second partial derivatives of \( f \),

\[ \{H_f(\mathbf{x})\}_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \]
Newton’s Method, continued

- Do not explicitly invert Hessian matrix, but instead solve linear system

\[
H_f(x_k)s_k = -\nabla f(x_k)
\]

for Newton step \(s_k\), then take as next iterate

\[
x_{k+1} = x_k + s_k
\]

- Convergence rate of Newton’s method for minimization is normally quadratic

- As usual, Newton’s method is unreliable unless started close enough to solution to converge
Example: Newton’s Method

- Use Newton’s method to minimize
  \[ f(x) = 0.5x_1^2 + 2.5x_2^2 \]

- Gradient and Hessian are given by
  \[ \nabla f(x) = \begin{bmatrix} 2x_1 \\ 5x_2 \end{bmatrix} \quad \text{and} \quad H_f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \]

- Taking \( x_0 = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \), we have \( \nabla f(x_0) = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \)

- Linear system for Newton step is
  \[ \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} s_0 = \begin{bmatrix} -5 \\ -5 \end{bmatrix} , \] so
  \[ x_1 = x_0 + s_0 = \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \begin{bmatrix} -5 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \] which is exact solution for this problem, as expected for quadratic function
In principle, line search parameter is unnecessary with Newton’s method, since quadratic model determines length, as well as direction, of step to next approximate solution.

When started far from solution, however, it may still be advisable to perform line search along direction of Newton step $s_k$ to make method more robust (damped Newton).

Once iterates are near solution, then $\alpha_k = 1$ should suffice for subsequent iterations.
Newton’s Method, continued

- If objective function $f$ has continuous second partial derivatives, then Hessian matrix $H_f$ is symmetric, and near minimum it is positive definite.

- Thus, linear system for step to next iterate can be solved in only about half of work required for LU factorization.

- Far from minimum, $H_f(x_k)$ may not be positive definite, so Newton step $s_k$ may not be *descent direction* for function, i.e., we may not have

  $$\nabla f(x_k)^T s_k < 0$$

- In this case, alternative descent direction can be computed, such as negative gradient or direction of negative curvature, and then perform line search.
cg4.m demo
Quasi-Newton Methods

- Newton’s method costs $O(n^3)$ arithmetic and $O(n^2)$ scalar function evaluations per iteration for dense problem.
- Many variants of Newton’s method improve reliability and reduce overhead.
- **Quasi-Newton** methods have form
  \[ x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k) \]
  where $\alpha_k$ is line search parameter and $B_k$ is approximation to Hessian matrix.
- Many quasi-Newton methods are more robust than Newton’s method, are superlinearly convergent, and have lower overhead per iteration, which often more than offsets their slower convergence rate.
Could use Broyden’s method to seek zero of gradient, but this would not preserve symmetry of Hessian matrix.

Several secant updating formulas have been developed for minimization that not only preserve symmetry in approximate Hessian matrix, but also preserve positive definiteness.

Symmetry reduces amount of work required by about half, while positive definiteness guarantees that quasi-Newton step will be descent direction.
One of the most effective secant updating methods for minimization is **BFGS**

\[ x_0 = \text{initial guess} \]
\[ B_0 = \text{initial Hessian approximation} \]

for \( k = 0, 1, 2, \ldots \)

\[ \text{Solve } B_k s_k = -\nabla f(x_k) \text{ for } s_k \]
\[ x_{k+1} = x_k + s_k \]
\[ y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \]
\[ B_{k+1} = B_k + \frac{(y_k y_k^T)/(y_k^T s_k)}{(s_k^T B_k s_k)} - \frac{(B_k s_k s_k^T B_k)/(s_k^T B_k s_k)}{(s_k^T B_k s_k)} \]

end
In practice, factorization of $B_k$ is updated rather than $B_k$ itself, so linear system for $s_k$ can be solved at cost of $O(n^2)$ rather than $O(n^3)$ work.

Unlike Newton’s method for minimization, no second derivatives are required.

Can start with $B_0 = I$, so initial step is along negative gradient, and then second derivative information is gradually built up in approximate Hessian matrix over successive iterations.

BFGS normally has superlinear convergence rate, even though approximate Hessian does not necessarily converge to true Hessian.

Line search can be used to enhance effectiveness.
Example: BFGS Method

- Use BFGS to minimize $f(x) = 0.5x_1^2 + 2.5x_2^2$
- Gradient is given by $\nabla f(x) = \begin{bmatrix} x_1 \\ 5x_2 \end{bmatrix}$
- Taking $x_0 = [5 \ 1]^T$ and $B_0 = I$, initial step is negative gradient, so

$$x_1 = x_0 + s_0 = \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \begin{bmatrix} -5 \\ -5 \end{bmatrix} = \begin{bmatrix} 0 \\ -4 \end{bmatrix}$$

- Updating approximate Hessian using BFGS formula, we obtain

$$B_1 = \begin{bmatrix} 0.667 & 0.333 \\ 0.333 & 0.667 \end{bmatrix}$$

- Then new step is computed and process is repeated
Example: BFGS Method

<table>
<thead>
<tr>
<th>$\mathbf{x}_k$</th>
<th>$f(\mathbf{x}_k)$</th>
<th>$\nabla f(\mathbf{x}_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.000</td>
<td>15.000</td>
<td>5.000 5.000</td>
</tr>
<tr>
<td>0.000</td>
<td>40.000</td>
<td>0.000 −20.000</td>
</tr>
<tr>
<td>−2.222</td>
<td>2.963</td>
<td>−2.222 2.222</td>
</tr>
<tr>
<td>0.816</td>
<td>0.350</td>
<td>0.816 0.408</td>
</tr>
<tr>
<td>−0.009</td>
<td>0.001</td>
<td>−0.009 −0.077</td>
</tr>
<tr>
<td>−0.001</td>
<td>0.000</td>
<td>−0.001 0.005</td>
</tr>
</tbody>
</table>

- Increase in function value can be avoided by using line search, which generally enhances convergence.

- For quadratic objective function, BFGS with exact line search finds exact solution in at most $n$ iterations, where $n$ is dimension of problem. <interactive example>
Another method that does not require explicit second derivatives, and does not even store approximation to Hessian matrix, is *conjugate gradient* (CG) method.

CG generates sequence of conjugate search directions, implicitly accumulating information about Hessian matrix.

For **quadratic objective function**, CG is theoretically exact after at most $n$ iterations, where $n$ is dimension of problem.

CG is effective for general unconstrained minimization as well.
\[ x_0 = \text{initial guess} \]
\[ g_0 = \nabla f(x_0) \]
\[ s_0 = -g_0 \]

for \( k = 0, 1, 2, \ldots \)

Choose \( \alpha_k \) to minimize \( f(x_k + \alpha_k s_k) \)

\[ x_{k+1} = x_k + \alpha_k s_k \]
\[ g_{k+1} = \nabla f(x_{k+1}) \]
\[ \beta_{k+1} = (g_{k+1}^T g_{k+1}) / (g_k^T g_k) \]
\[ s_{k+1} = -g_{k+1} + \beta_{k+1} s_k \]

end

- Alternative formula for \( \beta_{k+1} \) is

\[ \beta_{k+1} = ((g_{k+1} - g_k)^T g_{k+1}) / (g_k^T g_k) \]
Conjugate Gradient Method for Optimization

\[ f(x) = \frac{1}{2} x^T H x - x^T b + c \]

\[ \nabla f = H x - b \]

Find \( x^* \) s.t. \( \nabla f = 0 \) \( \Rightarrow \) \[ H x^* = b \]

Let

\[ x_k = x_{k-1} + \alpha_k s_{k-1} \]

\[ = x_0 + \alpha_0 s_0 + \alpha_1 s_1 + \ldots + \alpha_{k-1} s_{k-1} \]

\[ = \begin{bmatrix} s_0 & s_1 & \ldots & s_{k-1} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{k-1} \end{bmatrix} \text{ if } x_0 = 0 \]

Idea:

Find \( \alpha_k \) s.t.

\[ S_k x_k = x^* \]

Least Squares Problem — with a hitch.

We don't know \( x^* \), we only know:

\[ H x^* = b \]

Recall:

Linear least squares:

\[ \min_{\alpha} \left\| S \alpha - x^* \right\|^2 \rightarrow s_i^T (S \alpha - x^*) = 0 \text{ for } i = 0, \ldots, k-1 \]

\[ s_i^T S \alpha = s_i^T x^* = (s_i, x^*) \text{ for } i = 0, \ldots, k-1 \]

We need a computable rhs: \( (s_i, x^*) \).

Use the \( H \) inner-product:

\[ \min_{\alpha} \left\| S \alpha - x^* \right\|^2_H \rightarrow (s_i, S \alpha)_H = (s_i, x^*_H) \text{ for } i = 0, \ldots, k-1 \]
Conjugate Gradient Method for Optimization

\[(s_i, x^*)_H = s_i^T H x^* = s_i^T b\]

\[(s_i, s_j)_H = s_i^T H \left( \sum_{j=0}^{k-1} \alpha_j s_j \right) = s_i^T H x_{k-1}\]

As with linear least squares, it helps if

\[(s_i, s_j)_H = \delta_{ij} \quad \text{orthonormal basis}\]

- in H-inner product,

Then we have:

\[(s_i, s \alpha)_H = \alpha_i = s_i^T b \iff \alpha = S^T b\]

So:

\[x_k = \sum_{j=0}^{k-1} s_j \alpha_j = S \alpha = SS^T b\]

\[= \sum_{j=0}^{k-2} s_j \alpha_j + s_{k-1} \alpha_{k-1}\]

\[= x_{k-1} + s_{k-1} \alpha_{k-1}\]

Thus, as we augment the search space,

\[f(s_0, s_1, \ldots, s_{k-2})\]

we have

\[x_k = x_{k-1} + s_{k-1} \alpha_{k-1} = \text{arg min } f(x_{k-1} + s_{k-1} \alpha_{k-1})\]

is the best fit in \((s_0, s_1, \ldots, s_{k-2})\) in H-norm.

If \(H\) is \(n \times n\), then \(x_n = x^*\)
Conjugate Gradient Method for Optimization

To enforce \( s_k \perp \{s_0, s_1, \ldots, s_{k-1}\} \)

Let \( s_k = -g_k - P_{k-1} g_k \)

\[ s_k = -g_k + \sum_{j=1}^{k-1} \beta_j s_j \]

\[ s_k^T H s_i = 0 \quad i = 1, \ldots, k-1 \]

\[ s_k^T H (s_i + \beta_j s_j) = -s_k^T H g_k \]

\[ \beta_i = \frac{s_k^T H g_k}{s_i^T H s_i} = \frac{g_k^T H s_i}{s_i^T H s_i} \]

Recall \( q_k = H x_k - H x^* = H s_k \)

\[ e_k^T H v = 0 \quad v \in \{s_0, s_1, \ldots, s_{k-1}\} \]

\[ q_0 = H x_0 - \frac{1}{2} \nabla F(x_0) \]

\[ s_0 = -q_0 \]

\[ x_1 = x_0 + \alpha_0 s_0 \]

\[ q_1 = \nabla F(x_1) = H x_1 - \frac{1}{2} \]

\[ = H x_0 + g_0 \]

\[ = -\alpha_0 H q_0 + g_0 \quad \in \mathbb{P}_1(H) q_0 \]

\[ s_1 = -q_1 + \beta_0 s_0 \quad \in \mathbb{P}_1(H) q_0 \]

\[ q_2 = H (x_1 + \alpha_1 s_1) - \frac{1}{2} \]

\[ = q_1 + \alpha_1 H s_1 \quad \in \mathbb{P}_2(H) q_0 \]

\[ s_2 = -q_2 + \sum_{j=0}^{1} \beta_j s_j \quad \in \mathbb{P}_2(H) q_0 \]
Conjugate Gradient Method for Optimization

\[ \beta_{k-1} = \frac{g_k^T H s_{k-1}}{s_{k-1}^T H s_{k-1}} \]

\[ s_k = -g_k + \sum_{j=0}^{k-1} \beta_j s_j \]

\[ \alpha_k = \frac{\beta_k}{\beta_{k-1}} \]

\[ g_k = g_k - \alpha_k H s_k \]

\[ g_k^T g_k = g_k^T g_k + \alpha_k H s_k \]

\[ (s_k^T H s_k) \alpha_k = s_k^T g_{k-1} \]

\[ s_k^T H s_{k-1} \alpha_{k-1} = s_k^T g_{k-1} \]

\[ \alpha_{k-1} = \frac{s_{k-1}^T g_{k-1}}{s_k^T H s_{k-1}} \]

\[ g_k^T H s_{k-1} = \frac{g_k^T g_k}{\alpha_k} = \frac{s_{k-1}^T H s_{k-1} g_{k-1}}{s_k^T H s_{k-1}} \]

\[ \frac{g_k^T H s_{k-1}}{s_k^T H s_{k-1}} = \frac{g_k^T g_k}{s_{k-1}^T s_{k-1}} \]

\[ g_k^T g_k = g_k^T g_k - g_k^T H s_{k-1} \]

\[ g_k^T H s_{k-1} = g_k^T g_k - g_k^T H s_{k-1} \]

\[ g_k^T H s_{k-1} = g_k^T g_k - g_k^T g_k \]

\[ g_k^T H s_{k-1} = g_k^T g_k - g_k^T g_k \]
Conjugate Gradient Iteration for Optimization

- Consider the objective function

\[ f(x) = \frac{1}{2} x^T H x = x^T b + c \]

\[ = \frac{1}{2} (x - x^*)^T H (x - x^*) + \tilde{c}. \]

- Let’s take the case \( H = \text{constant} \), which is essentially the case for \( x \approx x^* \) as it represents the first two terms in the Taylor series about \( x^* \).

- The gradient is

\[ \nabla f = Hx - b, \]

and the minimizer is the point \( x^* \) such that \( \nabla f(x^*) = 0 \),

\[ \implies Hx^* = b. \]

- Near the minimizer, need to solve a linear system.
Update Algorithm

• For steepest descents, Newton’s method, CG, etc., the update algorithm is of the form:

\[
x_{k+1} = x_k + \alpha_k s_k
\]

\[
= x_0 + \alpha_0 s_0 + \alpha_1 s_1 + \cdots + \alpha_k s_k.
\]

• If \( x_0 = 0 \),

\[
x_{k+1} = \begin{bmatrix} s_0 & s_1 & \cdots & s_k \end{bmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}
\]

\[
= S_k a_k.
\]

• That is, \( x_{k+1} \) is a linear combination of the preceding \( k + 1 \) search directions \( s_j \).

• Main Idea: Find \( a_k = [\alpha_0 \ \alpha_1 \ \cdots \ \alpha_k]^T \) such that

\[
S_k a_k \approx x^*.
\]
• **Main Idea:** Find $a_k = [\alpha_0 \alpha_1 \cdots \alpha_k]^T$ such that

$$S_k a_k \approx x^*.$$

• Use linear least squares, with a computable inner product.

• Since we only know $\nabla f = Hx - b = Hx - Hx^* = H(x - x^*)$, rather than $x^*$ itself, it makes sense to use the $H$ inner product:

• Actually, we won’t even need to know $H$ or $b$!

• For the moment, we proceed as if we do know them.
Linear Least Squares in $H$ Inner-Product.

- In the $H$-norm, the closest element in our search space ($= \text{span}\{s_0 \cdots s_k\}$) is given by the vector $S_k a_k$ satisfying,

$$
\min_{a_k} [S_k a_k - x^*]_H \implies s_i^T H (S_k a_k - x^*) = 0, \quad i = 0, \ldots, k.
$$

$$
S_i^T H S_k a_k = S_i^T H x^* = s_i^T b
$$

$$(S_k^T H S_k) a_k = S_k^T b.
$$

- If $s_i^T H s_j = 0$ for $i \neq j$, then $S_k^T H S_k$ is diagonal:

$$
(S_k^T H S_k) a_k = \begin{bmatrix}
    s_0^T H s_0 \\
    \cdots \\
    s_k^T H s_k
\end{bmatrix}
\begin{pmatrix}
    \alpha_0 \\
    \alpha_1 \\
    \vdots \\
    \alpha_k
\end{pmatrix}
= \begin{pmatrix}
    s_0^T b \\
    s_1^T b \\
    \vdots \\
    s_k^T b
\end{pmatrix}.
• Solving for $\alpha_k$ becomes trivial:

$$\alpha_k = \frac{s_k^T b}{s_k^T H s_k}.$$ 

- Note that, for optimization problems, we don’t know $H$ nor $b$.
- We can find $\alpha_k$, however, through a line search in the direction $s_k$.

• The main point is that each update step yields the best fit in $S_k$, independent of the previous steps,

$$x_{k+1} = \alpha_0 s_0 + \alpha_1 s_1 + \cdots + \alpha_{k-1} s_{k-1} + \alpha_k s_k$$

$$= x_k + \alpha_k s_k.$$ 

• Thus, we have a short-term recurrence to go from $x_k$ to $x_{k+1}$ that yields the optimal solution.

• The only requirement for optimality is: $s_k \perp_H S_{k-1}$. 
• Thus, set:

\[ s_{k+1} = - \left( g_{k+1} - \sum_{j=1}^{k} \gamma_j s_j \right) \]

with

\[ \gamma_j = \frac{s_j^T H g_{k+1}}{s_j^T H s_j} = 0. \]

• As we saw in Chapter 3 notes, \( \gamma_j = 0 \) for \( j < k \) because \( g_{k+1} = H e_{k+1} \) and \( e_{k+1} \perp_H H S_j, j < k \).

• So, we also have a short-term recurrence for \( s_{k+1} \):

\[ s_{k+1} = -g_{k+1} + \beta_{k+1} s_k, \]

with

\[ \beta_{k+1} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \text{ or } \beta_{k+1} = \frac{(g_{k+1} - g_k)^T g_{k+1}}{g_k^T g_k}. \]
Summary of CG for Optimization

- Conjugate gradient iteration produces the projection of $x^*$ onto span $(s_0 \ s_1 \ldots s_k)$

- For constant SPD Hessian, $H$, CG iteration will be exact after:
  - $n$ iterations, or $m < n$ iterations if $H$ has only $m$ distinct eigenvalues.

- For nonconstant $H$, CG is not exact. However $H \rightarrow$ constant as $x \rightarrow x^*$

- **Important** to restart orthogonalization process after $n$ iterations.
  - Otherwise, $S^THS$ will be $k \times k$, with $k > n$, and will be singular.

- Method does not need $H$, nor $b$, only $\nabla f$, plus line search in the direction $s_k$ (not $g_k := \nabla f$, which would be steepest descent).
Conjugate Gradient Method, continued

\[ x_0 = \text{initial guess} \]
\[ g_0 = \nabla f(x_0) \]
\[ s_0 = -g_0 \]

\textbf{for} \ k = 0, 1, 2, \ldots \\
\hspace{1cm} \text{Choose} \ \alpha_k \ \text{to minimize} \ f(x_k + \alpha_k s_k)\\
\hspace{1cm} x_{k+1} = x_k + \alpha_k s_k \\
\hspace{1cm} g_{k+1} = \nabla f(x_{k+1}) \\
\hspace{1cm} \beta_{k+1} = (g_{k+1}^T g_{k+1})/(g_k^T g_k) \\
\hspace{1cm} s_{k+1} = -g_{k+1} + \beta_{k+1} s_k \\
\textbf{end} \\

\textbullet \ \text{Alternative formula for} \ \beta_{k+1} \ \text{is} \\
\hspace{1cm} \beta_{k+1} = ((g_{k+1} - g_k)^T g_{k+1})/(g_k^T g_k)
Example: Conjugate Gradient Method

- Use CG method to minimize \( f(x) = 0.5x_1^2 + 2.5x_2^2 \)
- Gradient is given by \( \nabla f(x) = \begin{bmatrix} x_1 \\ 5x_2 \end{bmatrix} \)
- Taking \( x_0 = [5, 1]^T \), initial search direction is negative gradient,
  \[
  s_0 = -g_0 = -\nabla f(x_0) = \begin{bmatrix} -5 \\ -5 \end{bmatrix}
  \]
- Exact minimum along line is given by \( \alpha_0 = 1/3 \), so next approximation is \( x_1 = [3.333, -0.667]^T \), and we compute new gradient,
  \[
  g_1 = \nabla f(x_1) = \begin{bmatrix} 3.333 \\ -3.333 \end{bmatrix}
  \]
Example, continued

- So far there is no difference from steepest descent method.

- At this point, however, rather than search along new negative gradient, we compute instead

\[ \beta_1 = \frac{g_1^T g_1}{g_0^T g_0} = 0.444 \]

which gives as next search direction

\[ s_1 = -g_1 + \beta_1 s_0 = \begin{bmatrix} -3.333 \\ 3.333 \end{bmatrix} + 0.444 \begin{bmatrix} -5 \\ -5 \end{bmatrix} = \begin{bmatrix} -5.556 \\ 1.111 \end{bmatrix} \]

- Minimum along this direction is given by \( \alpha_1 = 0.6 \), which gives exact solution at origin, as expected for quadratic function.

\textit{cg4.m demo}
Another way to reduce work in Newton-like methods is to solve linear system for Newton step by iterative method.

Small number of iterations may suffice to produce step as useful as true Newton step, especially far from overall solution, where true Newton step may be unreliable anyway.

Good choice for linear iterative solver is CG method, which gives step intermediate between steepest descent and Newton-like step.

Since only matrix-vector products are required, explicit formation of Hessian matrix can be avoided by using finite difference of gradient along given vector.
Nonlinear Least Squares
Nonlinear Least Squares

A bit of notation...

- Recall *linear* least squares problem: *Find* $\mathbf{x}$ *such that*

$$\mathbf{y} = \mathbf{A}\mathbf{x} \approx \mathbf{b}$$

for $m \times n$ matrix with $m > n$.

- The *nonlinear* least squares problem is almost the same: *Find* $\mathbf{x}$ *such that*

$$\mathbf{y} = \mathbf{A}(\mathbf{x})\mathbf{x} \approx \mathbf{b}.$$

- Here, $\mathbf{A} = \mathbf{A}(\mathbf{x})$ is a function of the unknown parameters $\mathbf{x} = [x_1 \ x_2 \ \ldots \ x_n]^T$.

- In many cases, cannot readily write residual evaluation as $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$.

- Instead, seek parameters $x_j$, $j = 1, \ldots, n$ that minimize the 2-norm of the residual vector with components

$$r_i(\mathbf{x}) = b_i - f(t_i, \mathbf{x}),$$

with $f(t, \mathbf{x})$ a nonlinear function of $\mathbf{x}$. 
Examples: Linear Least Squares

**Model:** \[ f(t, \mathbf{x}) = x_1 \sin 2\pi t + x_2 \sin 4\pi t. \]

- This problem is *linear* in the unknown model parameters, \( x_1 \) and \( x_2 \).
- (It is nonlinear in independent parameter, \( t \), however.)
- In graph, the \( b_i \) are given data as a function of \( t_i \).
- The linear least squares problem is

\[
\begin{bmatrix}
\sin 2\pi t_1 & \sin 4\pi t_1 \\
\sin 2\pi t_2 & \sin 4\pi t_2 \\
\vdots \\
\sin 2\pi t_n & \sin 4\pi t_n
\end{bmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
\approx
\begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix}.
\]
Examples: Nonlinear Least Squares

Model: \( f(t, x) = x_1 \sin x_2 \pi t + x_3 \sin x_4 \pi t. \)

• This problem is nonlinear in the unknown model parameters, \( x = [x_1 \ x_2 \ x_3 \ x_4]^T. \)

• The nonlinear least squares problem is:

\[
\text{Find} \ [x_1 \ x_2 \ x_3 \ x_4] \text{ that minimizes } ||r||, \text{ with } \\
\quad r := b - f(t_i, x).
\]

• Note: The text uses \( y := b \) (which is different from the role of \( y \) in Chapter 3.)
Nonlinear Least Squares

- Given data \((t_i, y_i)\), find vector \(x\) of parameters that gives “best fit” in least squares sense to model function \(f(t, x)\), where \(f\) is nonlinear function of \(x\)

- Define components of *residual function*

\[
r_i(x) = y_i - f(t_i, x), \quad i = 1, \ldots, m
\]

so we want to minimize \(\phi(x) = \frac{1}{2} r^T(x) r(x)\)

- Gradient vector is \(\nabla \phi(x) = J^T(x) r(x)\) and Hessian matrix is

\[
H_\phi(x) = J^T(x) J(x) + \sum_{i=1}^{m} r_i(x) H_i(x)
\]

where \(J(x)\) is Jacobian of \(r(x)\), and \(H_i(x)\) is Hessian of \(r_i(x)\)
Nonlinear Least Squares

- Given data \((t_i, y_i)\), find vector \(x\) of parameters that gives “best fit” in least squares sense to model function \(f(t, x)\), where \(f\) is nonlinear function of \(x\)

- Define components of *residual function*

\[
r_i(x) = y_i - f(t_i, x), \quad i = 1, \ldots, m
\]

so we want to minimize \(\phi(x) = \frac{1}{2} r^T(x) r(x)\)

- Gradient vector is \(\nabla \phi(x) = J^T(x) r(x)\) and Hessian matrix is

\[
H_\phi(x) = J^T(x) J(x) + \sum_{i=1}^{m} r_i(x) H_i(x)
\]

where \(J(x)\) is Jacobian of \(r(x)\), and \(H_i(x)\) is Hessian of \(r_i(x)\)
Nonlinear Least Squares, continued

- Linear system for Newton step is

\[
\begin{pmatrix}
J^T(x_k)J(x_k) + \sum_{i=1}^{m} r_i(x_k)H_i(x_k)
\end{pmatrix} s_k = -J^T(x_k)r(x_k)
\]

- \(m\) Hessian matrices \(H_i\) are usually inconvenient and expensive to compute

- Moreover, in \(H_\phi\) each \(H_i\) is multiplied by residual component \(r_i\), which is small at solution if fit of model function to data is good
This motivates Gauss-Newton method for nonlinear least squares, in which second-order term is dropped and linear system

$$J^T(x_k)J(x_k)s_k = -J^T(x_k)r(x_k)$$

is solved for approximate Newton step $s_k$ at each iteration.

This is system of normal equations for linear least squares problem

$$J(x_k)s_k \approx -r(x_k)$$

which can be solved better by QR factorization.

Next approximate solution is then given by

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

and process is repeated until convergence.
Nonlinear Least Squares

• \( r_i = y_i - f(t_i, \mathbf{x}) \) (\( \mathbf{x} \) are the unknowns.)

Example: \( f(t, \mathbf{x}) = x_1 e^{x_2 t} \).

• Minimization problem:

\[
\phi(\mathbf{x}) = \frac{1}{2} \mathbf{r}^T \mathbf{r} = \frac{1}{2} \sum_{i=1}^{m} r_i^2.
\]

• Stationary point:

\[
\nabla \phi(\mathbf{x}) = \frac{\partial}{\partial x_j} \phi
\]

\[
= \frac{\partial}{\partial x_j} \left[ \frac{1}{2} \sum_{i=1}^{m} r_i^2 \right]
\]

\[
= \sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial r_i}{\partial x_j} r_i = J^T(\mathbf{x}) \mathbf{r}(\mathbf{x}),
\]

with

\[
J_{ij} := \frac{\partial r_i}{\partial x_j}.
\]
• Hessian:

\[
H_{\phi}(x) = \frac{\partial}{\partial x_k} \nabla \phi
\]

\[
= \frac{\partial}{\partial x_k} \left( \sum_{i=1}^{m} \frac{\partial r_i}{\partial x_j} r_i \right)
\]

\[
H_{\phi,jk} = \sum_{i=1}^{m} \frac{\partial r_i}{\partial x_j} \frac{\partial r_i}{\partial x_k} + \sum_{i=1}^{m} r_i \frac{\partial^2 r_i}{\partial x_j x_k}
\]

\[
H_{\phi}(x) = J^T(x)J(x) + \sum_{i=1}^{m} r_i(x)H_i(x).
\]

• \(H_{\phi}\) is the sum of SPD matrices (near \(x^*\), at least), where the \(H_i\)s are the Hessians associated with each residual component, \(r_i(x)\).

• The contributions of \(H_i\) are weighted by the residual component, \(r_i(x)\) \(\rightarrow 0\) as \(x \rightarrow x^*\).

• Since we typically spend most of the work near \(x^*\), it is convenient to neglect the \(H_i\)s when considering the Hessian of \(\phi\).
• Newton Step:

\[ x_{k+1} = x_k + s_k \]
\[ H_\phi s_k = -J_k^T r_k \]

• Search direction via approximate Hessian:

\[
\begin{align*}
\left( J_k^T J_k + \sum_{i=1}^m r_i H_{r_i} \right) s_k &= -J(x)^T r_k \\
( J_k^T J_k ) s_k &= -J(x)^T r_k 
\end{align*}
\]

• Equivalent to normal equations, linear least squares:

\[ J_k s_k \approx -r_k. \]

• Solve via (reduced) QR instead of Gaussian elimination:

\[
\begin{align*}
Q_1 R_1 &= J_k \\
Q_1^T Q_1 R_1 s_k &= -Q_1^T r_k \\
s_k &= -R_1^{-1} Q_1^T r_k \\
x_{k+1} &= x_k + s_k.
\end{align*}
\]

`nllsq_test.m`
Nonlinear Least Squares

\[ r_i = y_i - f(t_i, \hat{x}) \]

\[ \phi(\hat{x}) = \frac{1}{2} \mathbf{r}^T \mathbf{r} = \frac{1}{2} \sum_{i=1}^{m} r_i^2 \]

\[ \nabla \phi = \mathbf{J}^T \mathbf{r} \]

\[ \nabla \phi = \frac{\partial}{\partial x_j} \phi(\mathbf{r}(\hat{x})) \]

\[ = \frac{\partial}{\partial x_j} \left( \frac{1}{2} \sum_{i=1}^{m} r_i^2 \right) \]

\[ = \sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial f(t_i, \hat{x})}{\partial x_j} r_i = \mathbf{J}^T \mathbf{r} \]

\[ \mathbf{J}_{ij} := \frac{\partial r_i}{\partial x_j} \]

\[ H_{jk} = \sum_{i=1}^{m} \frac{\partial^2 r_i}{\partial x_j \partial x_k} + \sum_{i=1}^{m} r_i \frac{\partial^2 r_i}{\partial x_j \partial x_k} \]

\[ \mathbf{H} = \mathbf{J}^T \mathbf{J} + \sum_{i=1}^{m} r_i \mathbf{H}_i \]
Nonlinear Least Squares

**Newton Step**

\[ x_{k+1} = x_k + s_k \]

\[ H_\phi s_k = -J_k^T r_k \]

\[ (J_k^T J_k + \sum_{i=1}^{\kappa} \beta_i r_i H_i r_i) s_k = -J_k^T r_k \]

\[ (J_k^T J_k) s_k = -J_k^T r_k \]

Solve via \( QR = J \)

\[ Q_1^T Q_1 R s_k = -Q_1^T r_k \]

\[ s_k = -R_1^{-1} Q_1^T r_k \]

\[ x_{k+1} = x_k + s_k \]
Example: Gauss-Newton Method

- Use Gauss-Newton method to fit nonlinear model function

\[ f(t, x) = x_1 \exp(x_2 t) \]

to data

\[
\begin{array}{ccccc}
   t & 0.0 & 1.0 & 2.0 & 3.0 \\
   y & 2.0 & 0.7 & 0.3 & 0.1 \\
\end{array}
\]

- For this model function, entries of Jacobian matrix of residual function \( r \) are given by

\[
\{ J(x) \}_{i,1} = \frac{\partial r_i(x)}{\partial x_1} = -\exp(x_2 t_i)
\]

\[
\{ J(x) \}_{i,2} = \frac{\partial r_i(x)}{\partial x_2} = -x_1 t_i \exp(x_2 t_i)
\]
Example, continued

- If we take $x_0 = [1 \ 0]^T$, then Gauss-Newton step $s_0$ is given by linear least squares problem

\[
\begin{bmatrix}
-1 & 0 \\
-1 & -1 \\
-1 & -2 \\
-1 & -3
\end{bmatrix}
\begin{bmatrix}
s_0
\end{bmatrix}
= 
\begin{bmatrix}
-1 \\
0.3 \\
0.7 \\
0.9
\end{bmatrix}
\]

whose solution is $s_0 = \begin{bmatrix} 0.69 \\ -0.61 \end{bmatrix}$

- Then next approximate solution is given by $x_1 = x_0 + s_0$, and process is repeated until convergence
**Example, continued**

<table>
<thead>
<tr>
<th>$x_k$</th>
<th>$|r(x_k)|^2_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1.690</td>
<td>-0.610</td>
</tr>
<tr>
<td>1.975</td>
<td>-0.930</td>
</tr>
<tr>
<td>1.994</td>
<td>-1.004</td>
</tr>
<tr>
<td>1.995</td>
<td>-1.009</td>
</tr>
<tr>
<td>1.995</td>
<td>-1.010</td>
</tr>
</tbody>
</table>

< interactive example >
Gauss-Newton method replaces nonlinear least squares problem by sequence of linear least squares problems whose solutions converge to solution of original nonlinear problem.

If residual at solution is large, then second-order term omitted from Hessian is not negligible, and Gauss-Newton method may converge slowly or fail to converge.

In such “large-residual” cases, it may be best to use general nonlinear minimization method that takes into account true full Hessian matrix.
Levenberg-Marquardt Method

- Levenberg-Marquardt method is another useful alternative when Gauss-Newton approximation is inadequate or yields rank deficient linear least squares subproblem

- In this method, linear system at each iteration is of form

\[
(J^T(x_k)J(x_k) + \mu_k I)s_k = -J^T(x_k)r(x_k)
\]

where \( \mu_k \) is scalar parameter chosen by some strategy

- Corresponding linear least squares problem is

\[
\begin{bmatrix}
J(x_k) \\
\sqrt{\mu_k} I
\end{bmatrix}
\begin{bmatrix}
s_k \\
0
\end{bmatrix}
\approx
\begin{bmatrix}
-r(x_k) \\
0
\end{bmatrix}
\]

- With suitable strategy for choosing \( \mu_k \), this method can be very robust in practice, and it forms basis for several effective software packages

< interactive example >
Levenberg-Marquardt

- Standard Gauss-Newton step is like Newton’s method applied to truncated Hessian:
  \[
  (J_k^T J_k) s_k = -J(x)^T r_k
  \]

- Suffers from not being robust.

- Steepest descent would be:
  \[
  s_k = -(J(x))^T r_k = -\nabla \phi(x).
  \]

- Levenberg-Marquardt is a combination of these two:
  \[
  (J_k^T J_k + \mu_k I) s_k = -J(x)^T r_k,
  \]
  which, for large \( \mu_k \), corresponds to steepest descent.

- If \( \mu_k \to 0 \) as the iteration proceeds, recover Gauss-Newton and quadratic convergence.

\[\text{levenberg.m}\]
Equality-Constrained Optimization

- For equality-constrained minimization problem
  \[ \min f(x) \quad \text{subject to} \quad g(x) = 0 \]
  where \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^n \to \mathbb{R}^m \), with \( m \leq n \), we seek critical point of Lagrangian \( \mathcal{L}(x, \lambda) = f(x) + \lambda^T g(x) \)

- Applying Newton’s method to nonlinear system
  \[ \nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + J_g^T(x) \lambda \\ g(x) \end{bmatrix} = 0 \]
  we obtain linear system
  \[ \begin{bmatrix} B(x, \lambda) & J_g^T(x) \\ J_g(x) & 0 \end{bmatrix} \begin{bmatrix} s \\ \delta \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + J_g^T(x) \lambda \\ g(x) \end{bmatrix} \]
  for Newton step \((s, \delta)\) in \((x, \lambda)\) at each iteration
Lagrangian function \( \mathcal{L}: \mathbb{R}^{n+m} \rightarrow \mathbb{R} \), is defined by

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

Its gradient is given by

\[
\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix}
\nabla f(x) + J_{g}^T(x) \lambda \\
J_{g}(x) \lambda
\end{bmatrix}
\]

Its Hessian is given by

\[
H_{\mathcal{L}}(x, \lambda) = \begin{bmatrix}
B(x, \lambda) & J_{g}^T(x) \\
J_{g}(x) & O
\end{bmatrix}
\]

where

\[
B(x, \lambda) = H_f(x) + \sum_{i=1}^{m} \lambda_i H_{g_i}(x)
\]
Sequential Quadratic Programming

- Foregoing block $2 \times 2$ linear system is equivalent to quadratic programming problem, so this approach is known as *sequential quadratic programming*.

- Types of solution methods include:
  - *Direct solution* methods, in which entire block $2 \times 2$ system is solved directly.
  - *Range space* methods, based on block elimination in block $2 \times 2$ linear system.
  - *Null space* methods, based on orthogonal factorization of matrix of constraint normals, $J_g^T(x)$.
**Direct Method:**

Solve 2 x 2 Block System via Gaussian Elimination

\[
\begin{bmatrix}
B(x, \lambda) & J_g^T(x) \\
J_g(x) & 0
\end{bmatrix}
\begin{bmatrix}
s \\
\delta
\end{bmatrix}
= -
\begin{bmatrix}
\nabla f(x) + J_g^T(x) \lambda \\
g(x)
\end{bmatrix}
\]
Once Newton step \((s, \delta)\) determined, we need *merit function* to measure progress toward overall solution for use in line search or trust region.

Popular choices include *penalty function*

\[
\phi_\rho(x) = f(x) + \frac{1}{2} \rho \ g(x)^T g(x)
\]

and *augmented Lagrangian function*

\[
\mathcal{L}_\rho(x, \lambda) = f(x) + \lambda^T g(x) + \frac{1}{2} \rho \ g(x)^T g(x)
\]

where parameter \(\rho > 0\) determines relative weighting of optimality vs feasibility.

Given starting guess \(x_0\), good starting guess for \(\lambda_0\) can be obtained from least squares problem

\[
J_g^T(x_0) \lambda_0 \cong -\nabla f(x_0)
\]
Methods just outlined for equality constraints can be extended to handle inequality constraints by using *active set* strategy.

Inequality constraints are provisionally divided into those that are satisfied already (and can therefore be temporarily disregarded) and those that are violated (and are therefore temporarily treated as equality constraints).

This division of constraints is revised as iterations proceed until eventually correct constraints are identified that are binding at solution.
Active Constraint

- If constraint is active \( h(x^*) = 0 \) then \( \nabla f \) and \( \nabla h \) point in opposite directions.
- \( \nabla f + \lambda^* \nabla h = 0 \).
- \( \lambda^* > 0 \).

\[
-\nabla f(x^*) = J_h^T(x^*)\lambda
\]

\[
-\frac{\partial f}{\partial x_i} = \sum_{k=1}^{m} \frac{\partial h_k}{\partial x_i} \lambda_k
\]

- \( h(x) = 0 \)
- \( h(x) < 0 \)
Penalty Methods

- Merit function can also be used to convert equality-constrained problem into sequence of unconstrained problems

- If \( x^{*\rho} \) is solution to

\[
\min_{x} \phi_{\rho}(x) = f(x) + \frac{1}{2} \rho g(x)^T g(x)
\]

then under appropriate conditions

\[
\lim_{\rho \to \infty} x^{*\rho} = x^{*}
\]

- This enables use of unconstrained optimization methods, but problem becomes ill-conditioned for large \( \rho \), so we solve sequence of problems with gradually increasing values of \( \rho \), with minimum for each problem used as starting point for next problem

< interactive example >
For inequality-constrained problems, another alternative is *barrier function*, such as

\[
\phi_\mu(x) = f(x) - \mu \sum_{i=1}^{p} \frac{1}{h_i(x)}
\]

or

\[
\phi_\mu(x) = f(x) - \mu \sum_{i=1}^{p} \log(-h_i(x))
\]

which increasingly penalize feasible points as they approach boundary of feasible region.

Again, solutions of unconstrained problem approach \( x^* \) as \( \mu \to 0 \), but problems are increasingly ill-conditioned, so solve sequence of problems with decreasing values of \( \mu \).

Barrier functions are basis for *interior point* methods for linear programming.
Barrier Functions for Inequality Constraints

- Solve modified unconstrained minimization problem,

\[
\phi_\mu(x) = f(x) - \mu \sum_{i=1}^{p} \frac{1}{h_i(x)}
\]

- **Starting in the feasible region!**

- As \(\mu \to 0\) it appears that the constraint is weaker.

- However, the constraint goes to \(\infty\) for any \(\mu \neq 0\), so a small \(\mu\) simply means the fence is *steeper*, which makes the minimization problem more challenging.

\[
\begin{align*}
\mu &= 1.0 \\
\mu &= 0.1
\end{align*}
\]

**feasible region, \(h < 0\)**

\[
\begin{align*}
f(x) &= \frac{1}{2}x_1^2 + \frac{5}{2}x_2^2 \\
h(x) &= 1 - x_1 + x_2 < 0
\end{align*}
\]
With barrier methods, simply augment the original (unconstrained) objective function, \( f(x) \), with the penalty function.

Then, start in the feasible region.

`cg_barrier.m`
Example: Constrained Optimization

Consider quadratic programming problem

$$\min_x f(x) = 0.5x_1^2 + 2.5x_2^2$$

subject to

$$g(x) = x_1 - x_2 - 1 = 0$$

Lagrangian function is given by

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x) = 0.5x_1^2 + 2.5x_2^2 + \lambda(x_1 - x_2 - 1)$$

Since

$$\nabla f(x) = \begin{bmatrix} x_1 \\ 5x_2 \end{bmatrix} \quad \text{and} \quad J_g(x) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

we have

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla f(x) + J^T_g(x)\lambda = \begin{bmatrix} x_1 \\ 5x_2 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
So system to be solved for critical point of Lagrangian is

\[
\begin{align*}
    x_1 + \lambda &= 0 \\
    5x_2 - \lambda &= 0 \\
    x_1 - x_2 &= 1
\end{align*}
\]

which in this case is linear system

\[
\begin{bmatrix}
1 & 0 & 1 \\
0 & 5 & -1 \\
1 & -1 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\lambda
\end{bmatrix} =
\begin{bmatrix}
0 \\
0 \\
1
\end{bmatrix}
\]

Solving this system, we obtain solution

\[
x_1 = 0.833, \quad x_2 = -0.167, \quad \lambda = -0.833
\]
Example, continued

contours of $0.5x_1^2 + 2.5x_2^2$

line $x_1 - x_2 = 1$