GRADIENT METHODS
GRADIENT = STEEPEST DESCENT

Convex Function

Iso-contours

gradient

negative gradient
Gradient descent

\[ x^{k+1} = x^k - \tau \nabla f(x^k) \]
STEPSIZE RESTRICTION

\[ f(x) = \frac{\alpha}{2} x^2 \]

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

\[ x^{k+1} = x^k - \tau(\alpha x^k) \]

Restriction:

\[ \tau < \frac{2}{\alpha} \]
Lipschitz Constant for gradient

\[ \| \nabla f(x) - \nabla f(y) \| \leq M \| x - y \| \]

\[ f(y) \leq f(x) + (y - x)^T \nabla f(x) + \frac{M}{2} \| y - x \|^2 \]

When Hessian exists:

\[ M \geq \| \nabla^2 f(x) \| \]
STABILITY RESTRICTION

If you know the Lipschitz Constant for the gradient

\[ f(y) \leq f(x) + (y - x)^T \nabla f(x) + \frac{M}{2} \|y - x\|^2 \]

\[ f(x - \tau \nabla f(x)) \leq f(x) - \tau \nabla f(x)^T \nabla f(x) + \frac{M \tau^2}{2} \|\nabla f(x)\|^2 \]

\[ = f(x) - \tau \|\nabla f(x)\|^2 + \frac{M \tau^2}{2} \|\nabla f(x)\|^2 \]

\[ = f(x) + \frac{M \tau^2 - 2\tau}{2} \|\nabla f(x)\|^2 \]

\[ M \tau^2 - 2\tau < 0 \Rightarrow \tau < \frac{2}{M} \]

proves convergence in terms of gradient
LINE SEARCH METHODS

• Choose search direction: $d$
• **SEARCH** for stepsize that satisfies **SOME** inequality:

• Update iterate: $x^{k+1} = x^k + \tau d$

**Armijo condition**

$$f(x^k + \tau d) \leq f(x^k) + \alpha(\tau d)^T \nabla f(x^k), \quad \alpha < 1$$
Choose search direction: \( d = -\nabla f(x^k) \)

Find stepsizes \( \tau \) satisfying Wolf conditions

\[
f(x^k + \tau d) \leq f(x^k) + \alpha (\tau d)^T \nabla f(x^k), \quad \alpha < 1
\]

\[
d^T \nabla f(x^k + \tau d) > \beta d^T \nabla f(x^k), \quad \alpha < \beta < 1
\]

Update iterate: \( x^{k+1} = x^k + \tau d \)

**Theorem**

Suppose we have the uniform bound for a convex function

\[
\frac{\nabla f(x_k)^T d}{\|\nabla f(x_k)\| \|d\|} > c
\]

then

\[
\lim_{k \to \infty} \nabla f(x^k) \to 0
\]
LINE SEARCH IN REAL LIFE

**Backtracking/Armijo line search**

Choose search direction: $d = -\nabla f(x^k)$

While $f(x^k + \tau d) \geq f(x^k) + \alpha(\tau d)^T \nabla f(x^k)$

$\tau \leftarrow \tau / 2$

Update iterate $x^{k+1} = x^k + \tau d$

**Exact line search**

Choose search direction: $d = -\nabla f(x^k)$

$\tau = \min_{\tau} f(x^k + \tau d)$

Update iterate $x^{k+1} = x^k + \tau d$
Consider this simple model problem...

\[ f(x) = \frac{\alpha}{2} x^T x \]

\[ x^{k+1} = x^k - \tau \nabla f(x^k) = x^k - \tau (\alpha x^k) \]

**Best choice:**  \( \tau = \frac{1}{\alpha} \)

**Barzilai-Borwein Model:**

\[ f(x) \approx \frac{\alpha}{2} x^T x \]

Figure out best approximation
BB METHOD

Barzilai-Borwein Model:

\[ f(x) \approx \frac{\alpha}{2} x^T x \]

\[ \nabla f(y) - \nabla f(x) \approx \alpha (y - x) \]

On each iteration, define:

\[ \Delta g = \nabla f(x^{k+1}) - \nabla f(x^k) \]

\[ \Delta x = x^{k+1} - x^k \]

The model tells us

\[ \Delta g \approx \alpha \Delta x \]

\[ \min_{\alpha} \frac{1}{2} \| \alpha \Delta x - \Delta g \|^2 \]
BB METHOD

\[
\min_\alpha \frac{1}{2} \| \alpha \Delta x - \Delta g \|^2
\]

\[
\Delta x^T \Delta x \alpha - \Delta x^T \Delta g = 0
\]

\[
\alpha = \frac{\Delta x^T \Delta g}{\Delta x^T \Delta x} \quad \tau = \alpha^{-1} = \frac{\Delta x^T \Delta x}{\Delta x^T \Delta g}
\]
BB METHOD

- Choose search direction:  \( d = -\nabla f(x^k) \)
- Compute BB step:  \( \tau^k = \frac{\Delta x^T \Delta x}{\Delta x^T \Delta g} \)
- While  \( f(x^k + \tau^k d) \geq f(x^k) + \tau^k \alpha d^T \nabla f(x^k) \)
  \[ \tau^k \leftarrow \frac{\tau^k}{2} \]
- Update iterate:  \( x^{k+1} = x^k + \tau d \)

For some smooth problems, BB method is superlinear; i.e. for large enough \( k \)
\[
f(x^{k+1}) - f^* \leq C(f(x^{k+1}) - f^*)^p
\]
for some \( C > 0 \), and \( p > 1 \)

This rate is asymptotic, and not global
CONVERGENCE RATE: EXACT LINE SEARCH

Strong convexity  \[ f(y) \geq f(x) + (y - x)^T \nabla f(x) + \frac{m}{2} \|y - x\|^2 \]

Lipschitz bound  \[ f(y) \leq f(x) + (y - x)^T \nabla f(x) + \frac{M}{2} \|y - x\|^2 \]

\[ y = x^{k+1}, \quad x = x^k \]

\[ y - x = x^{k+1} - x^k = -\tau \nabla f(x^k) \]

\[ f(x^{k+1}) \leq \min_{\tau} f(x^k) - \tau \nabla f(x^k)^T \nabla f(x^k) + \frac{M\tau^2}{2} \|\nabla f(x^k)\|^2 \]
CONVERGENCE RATE: EXACT LINE SEARCH

\[ f(x^{k+1}) \leq \min_{\tau} f(x^k) - \tau \nabla f(x^k)^T \nabla f(x^k) + \frac{M\tau^2}{2} \| \nabla f(x^k) \|^2 \]

\[ \tau = \frac{1}{M} \]

\[ f(x^{k+1}) \leq f(x^k) - \frac{1}{2M} \| \nabla f(x^k) \|^2 \]

\[ f(x^{k+1}) - f^* \leq f(x^k) - f^* - \frac{1}{2M} \| \nabla f(x^k) \|^2 \]

Optimality gap

Write in terms of optimality gap so we get relative change
STRONG CONVEXITY BOUND

\[ f(x^{k+1}) - f^* \leq f(x^k) - f^* - \frac{1}{2M} \| \nabla f(x^k) \|^2 \]

Strong convexity constant

\[ f(y) \geq f(x) + (y - x)^T \nabla f(x) + \frac{m}{2} \| y - x \|^2 \]

\[ f(x^*) \geq \min_y f(x^k) + (y - x^k)^T \nabla f(x^k) + \frac{m}{2} \| y - x^k \|^2 \]

Optimality:

\[ y - x^k = -\frac{1}{m} \nabla f(x^k) \]

\[ f(x^*) \geq f(x^k) - \frac{1}{m} \| \nabla f(x^k) \|^2 + \frac{1}{2m} \| \nabla f(x^k) \|^2 \]

\[ 2m(f(x^k) - f(x^*)) \leq \| \nabla f(x^k) \|^2 \]
STRONG CONVEXITY BOUND

\[ f(x^{k+1}) - f^* \leq f(x^k) - f^* - \frac{1}{2M} \| \nabla f(x^k) \|^2 \]

\[ 2m(f(x^k) - f(x^*)) \leq \| \nabla f(x^k) \|^2 \]

\[ f(x^{k+1}) - f^* \leq f(x^k) - f^* - \frac{m}{M} (f(x^k) - f^*) \]

\[ f(x^{k+1}) - f^* \leq \left( 1 - \frac{m}{M} \right) (f(x^k) - f^*) \]

What’s this?
STRONG CONVEXITY BOUND

\[ f(x^{k+1}) - f^* \leq \left( 1 - \frac{m}{M} \right) (f(x^k) - f^*) \]

By induction

\[ f(x^k) - f^* \leq \left( 1 - \frac{m}{M} \right)^k (f(x^0) - f^*) \]

**Theorem**

Gradient with exact line search satisfies **linear convergence** when the objective is strongly convex

\[ f(x^k) - f^* \leq \left( 1 - \frac{1}{\kappa} \right)^k (f(x^0) - f^*) \]

Condition number
GRADIENT DESCENT USES CONTOURS

Gradients are perpendicular to contours

\[ \kappa = 2 \]
POOR CONDITIONING: CONTOUR INTERPRETATION

\[ \kappa = 100 \]  

Contours are (almost) parallel!
GRADIENT DESCENT: WEAKLY CONVEX PROBLEMS

All methods so far assume strong convexity…

For strongly convex problems:

\[ f(x^k) - f^* \leq \left(1 - \frac{m}{M}\right)^k (f(x^0) - f^*) \]

For weakly convex problems:

\[ f(x^{k+1}) - f^* \leq \frac{M\|x^0 - x^*\|^2}{k + 1} \]
WHAT’S THE BEST WE CAN DO?

**Goal:** put worst-case bound on convergence of *first-order* methods

**Definition:** a method is first-order if

\[ x_k \in \text{span}\{x^0, \nabla f(x^0), \nabla f(x^1), \ldots, \nabla f(x^{k-1})\} \]

**Theorem:** For any first-order method, there is a smooth function with Lipschitz constant \( M \) such that

\[
f(x^k) - f(x^*) \geq \frac{3M \|x^0 - x^*\|^2}{32(k + 1)^2} \]

Nemirovsky and Yudin, 82’
PROOF

The worst function in the world

\[ f_n(x) = \frac{L}{4} \left( \frac{1}{2} x_1^2 + \frac{1}{2} \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + \frac{1}{2} x_n^2 - x_1 \right) \]

minimizer

\[ x_i^* = 1 - i/(n + 1) \]

\[ f_n(x^*) = \frac{L}{8} \left( \frac{1}{n+1} - 1 \right) \]
PROOF

The worst function in the world

\[ f_n(x) = \frac{L}{4} \left( \frac{1}{2} x_1^2 + \frac{1}{2} \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + \frac{1}{2} x_n^2 - x_1 \right) \]

\[ f_n(x^*) = \frac{L}{8} \left( \frac{1}{n+1} - 1 \right) \]

After iteration \( k \)
\[ x_i = 0, \text{ for } i > k \]

Consider function with \( n=2k \) variables

\[ f_{2k}(x^k) = f_k(x^k) \geq \frac{L}{8} \left( \frac{1}{k+1} - 1 \right) \]

\[ f_{2k}(x^*) = \frac{L}{8} \left( \frac{1}{2k+1} - 1 \right) \]
Consider function with $n=2k$ variables

$$f_{2k}(x^k) = f_k(x^k) \geq \frac{L}{8} \left( \frac{1}{k+1} - 1 \right)$$

$$f_{2k}(x^*) = \frac{L}{8} \left( \frac{1}{2k+2} - 1 \right)$$

lowest possible energy on step $k$

$$\frac{f_{2k}(x^k) - f_{2k}(x^*)}{\|x^0 - x^*\|^2} \geq \frac{L}{8} \left( \frac{1}{k+1} - \frac{1}{2k+2} \right) \geq \frac{3L}{32(k+1)^2}$$

true solution

start at 0
NESTEROV’S METHOD ACHIEVES THE BOUND

Gradient

\[ f(x^{k+1}) - f^* \leq \frac{\|x_0 - x^*\|^2}{\tau (k + 1)} \]

Nesterov

\[ f(x^{k+1}) - f^* \leq \frac{2\|x_0 - x^*\|^2}{\tau (k + 1)^2} \]

Nemirovski and Yundin ’83
NESTEROV’S METHOD

\[
\tau < \frac{1}{L\nabla f}, \quad \alpha^0 = 1
\]

\[
x^{k+1} = y^k - \tau \nabla f(y^k)
\]

\[
\alpha^{k+1} = \frac{1 + \sqrt{4(\alpha^k)^2 + 1}}{2}
\]

\[
y^{k+1} = x^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}}(x^{k+1} - x^k)
\]
OPTIMAL CONVERGENCE

**Theorem:** The objective error of the kth iterate of Nester’s method is bounded by

$$f(x^k) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{(k + 1)^2}$$

This **worst** case doesn’t mean much… in practice adaptive convergence is faster.
POOR CONDITIONING: FUNCTION INTERPRETATION

\[ f(x) = \sum \frac{d_i}{2} x_i^2 = x^T D x \]

\[ d_i = \frac{1}{3} \]

\[ d_i = 50 \]
POOR CONDITIONING: FUNCTION INTERPRETATION

\[ f(x) = \sum \frac{d_i}{2} x^2_i = x^T D x \]

\[ d_i = \frac{1}{3} \]

\[ x_i^{k+1} = x_i - \tau d_i x_k \]

Small step \( \tau = \frac{1}{50} \)

Big step \( \tau = \frac{1}{50} \)
minimize $\frac{1}{2} x^T H x + g^T x$

minimize $\frac{1}{2} x^T U D U^T x + g^T x$

$y \leftarrow U^T x$

minimize $\frac{1}{2} y^T D y + (U^T g)^T x = \sum \frac{d_i}{2} y_i^2 + (U^T g)_i y_i$
PRECONDITIONERS

\[ x = Py \]

\[ \min f(x) \quad \Rightarrow \quad \min f(Py) \]

\[ H \quad \Rightarrow \quad P^T HP \]
THE "BEST" PRECONDITIONER

minimize \( f(x) \) \quad x = Py \quad \text{minimize} \quad f(Py)

\[ H \quad \text{Hessians} \quad P^T HP \]

\[ P = H^{-1/2} \]

\[ P^T HP = H^{-1/2} HH^{-1/2} = I \]
SVD INTERPRETATION

\[ f(x) = \frac{1}{2} x^T H x = \frac{1}{2} x^T U D U^T x = \frac{1}{2} x^T U D^{1/2} D^{1/2} U^T x = \frac{1}{2} y^T y \]

\[ y = D^{1/2} U^T x \]

\[ x = U D^{-1/2} y = P y \]

Rotation  Stretch

\[ P \]
NEWTON’S METHOD

minimize \( f(x) \)  \( \Rightarrow \)  minimize \( f(H^{-1/2}y) \)

\[
x = H^{-1/2}y
\]

\[
y^{k+1} = y^k - H^{-1/2} \nabla f(H^{-1/2}y^k)
\]

change variables back to \( x \)
\[
H^{-1/2}y^{k+1} = H^{-1/2}y^k - H^{-1/2}H^{-1/2} \nabla f(H^{-1/2}y^k)
\]

\[
x^{k+1} = x^k - H^{-1} \nabla f(x^k)
\]

Newton direction

There are many interpretations…
GEOMETRIC INTERPRETATION

minimize \( f(x) \approx f(x^k) + (x - x^k)^T g + (x - x^k)^T H(x - x^k) \)

Derivative: \( g + H(x - x^k) = 0 \)

\( x^{k+1} = x^k - H^{-1}g \)

\[ f(x^k) + (x - x^k)^T g + (x - x^k)^T H(x - x^k) \]
CLASSICAL NEWTON METHOD

Newton’s method = Algorithm for root finding

\[ g(x) = 0 \]

\[ x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)} \]

For nonlinear system of equations

\[ g(x) \approx g(x^k) + (x - x^k)^T \nabla g(x^k) \]

\[ x^{k+1} = x^k - \nabla g(x^k)^{-1} g(x) \]
Newton Method for Optimization

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

Taylor’s theorem

\[ f(x) \approx (x - x^k)^T \nabla f(x^k) + \frac{1}{2} (x - x^k)^T \nabla^2 f(x^k)(x - x^k) \]

Derivative

**Linear** approximation of gradient

\[ \nabla f(x) \approx \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = g + H(x - x^k) = 0 \]

\[ x^{k+1} = x^k - H^{-1}g \]
SIMPLE RATE ANALYSIS

Bound the error: \[ e^k = x^k - x^* \]

Assume \( f \in C^2 \), \( e^k \) is sufficiently small, and strong convexity.

\[
0 = \nabla f(x^*) = \nabla f(x^k - e^k) = \nabla f(x^k) - H e^k + O(\|e^k\|^2)
\]

Multiply by inverse hessian

\[
0 = H^{-1} \nabla f(x^k) - e^k + O(\|e^k\|^2)
\]

note: \( e^{k+1} = e^k - H^{-1} \nabla f(x^k) \)

\( e^{k+1} = O(\|e^k\|^2) \)

Fletcher, Practical Methods of Optimization
DAMPED NEWTON METHOD

- Choose search direction: \( d = -\nabla^2 f(x^k)^{-1} \nabla f(x^k) \)

- Find stepsize \( \tau \) satisfying Wolf conditions
  \[
  f(x^k + \tau d) \leq f(x^k) + \alpha(\tau d)^T \nabla f(x^k), \quad \alpha < 1
  \]

- Update iterate: \( x^{k+1} = x^k + \tau d \)

\( \tau < 1 \) = Damped step
\( \tau = 1 \) = Full Newton

Predicted objective change in Newton direction
DAMPED NEWTON METHOD

**Theorem**

Suppose we have a Lipschitz constant for the Hessian

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H \|x - y\|$$

When the gradient gets small enough to satisfy

$$\|\nabla f(x^k)\| \leq 3(1 - 2\alpha) \frac{m^2}{L_H}$$

….the unit stepsize is an Armijo step, and

$$\tau = 1$$

$$\frac{L_H}{2m^2} \|\nabla f(x^{k+1})\| \leq \left( \frac{L_H}{2m^2} \|\nabla f(x^k)\| \right)^2$$

see Boyd and Vandenberghe
NEWTON IN THE WILD

In practice Hessian might be indefinite

If you start at a point of negative curvature, the Newton goes up!

(negative) search direction: must form an acute angle with the gradient

\[ g^T d > 0 \]

We can use any SPD matrix to approximate Hessian

\[ d = \hat{H}^{-1} g \]

\[ g^T d = g^T \hat{H}^{-1} g > 0 \]
MODIFIED HESSIAN

We can use any SPD matrix to approximate Hessian

Levenberg–Marquardt

$$\hat{H} = H + \gamma I$$

Add a large enough multiple of the identity to the Hessian that it becomes SPD

Modified Cholesky

Begin with partial Cholesky

$$H \rightarrow L \left( \begin{array}{cc} I & 0 \\ 0 & B \end{array} \right) L^T$$

Replace B with a SPD matrix

See Fang & O’leary, ‘06