Stationary point

- A point x is a stationary point of a smooth function f if

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

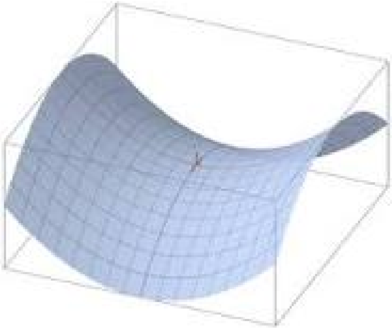
- Example: $f(x) = 40 + x_1^3(x_1 - 4) + 3(x_2 - 5)^2$:

$$\nabla f(x) = \begin{pmatrix} x_1^2(4x_1 - 12) \\ 6(x_2 - 5) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- Stationary points: $\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 5 \end{pmatrix}$

Stationary points

Every stationary point is either local maximum, local minimum, or a saddle point.



Stationary points

To distinguish between stationary points and local maximum, local minimum, we need second order Taylor expansion, let x be a stationary point:

$$
\begin{aligned}
f(x+\alpha d) &\approx f(x) + \alpha\langle \nabla f(x), d\rangle + \frac{\alpha^2}{2} d^T \nabla^2 f(x) d. \\
&= f(x) + \frac{\alpha^2}{2} d^T \nabla^2 f(x) d
\end{aligned}
$$

A direction d satisfying: $d^T \nabla^2 f(x) d < 0$ at stationary point x implies that $f(x+\alpha d) < f(x)$

- Hessian matrix of a smooth function f is negative semidefinite at every local maximum.
- Hessian matrix of a smooth function f is positive semidefinite at every local minimum.

Global minima of convex function

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. Suppose x^* is a local minimum, then x^* is a global minimum.

Proof: Consider an arbitrary point y and all the points between x and y: $\alpha x + (1-\alpha)y$, $0 < \alpha < 1$. Since x is a local minimum, α can be choosen so that $\alpha x + (1-\alpha)y$ is in small neighborhood of x and : $f(x) \leq f(\alpha x + (1-\alpha)y)$

By convexity of the function f,

$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$

We conclude: $f(x) \leq f(y) \forall y$.

Gradient descent method

- Iterative methods for minimizing a function $f(x)$
- For each iteration k generate a point x^{k+1} that is hopefully a *better* candidate for the final solution than x^k.
- Find direction d^k and step length α_k.
- Ideal situation for a direction d^k: $f(x^{k+1}) < f(x^k)$

Gradient descent method

- Consider the first order Taylor expansion of f at pointx^k:

$$\begin{aligned} f_{T1}(x^k + \alpha d^k) &= f(x^k) + \langle \nabla f(x^k), x^k + \alpha d^k - x^k \rangle. \\ &= f(x^k) + \alpha \langle \nabla f(x^k), d^k \rangle. \end{aligned}$$

- A good direction will satisfy: $\langle \nabla f(x^k), d^k \rangle < 0$.
- $d^k = -\nabla f(x^k)$

Gradient descent method

Method of gradient descent (steepest descent) with constant step length

- **Initialization**: Starting point $x^0 \in \mathbb{R}^n$, stepsize: α, tolerance ϵ, iteration number $k = 0$, maxIter
- Repeat:
  1. k=k+1;
  2. Calculate $d^k = -\nabla f(x^k)$.
  3. Set $x^{k+1} = x^k + \alpha d^k$.
  4. Test stopping criteria. If some tolerance $\leq \epsilon$ or maximum number of iterations is obtained.
- Output a point x^{k+1}, hopefully minimize function f(x).

Analysis of the method

Assume that function f has a derivative, $\nabla f(x)$, which also is a continuous function. Suppose there exists a constant M such that $\forall x, y \in \mathbb{R}^n$: (Lipschitz continuity)

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|.$$

Assume function f is bounded from below. If the stepsize α satistifies: $0 < \alpha < \frac{1}{M}$, then the sequence x^k generated by the method of gradient descent satisfied:

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

Gradient descent method

Mean value theorem: f continuous on [a,b] and differentiable on (a,b) then exists c: $a < c < b$ s.t:

$$f'(c) = \frac{f(a) - f(b)}{a - b}$$

Apply this, consider two points x^k,x^{k+1}, then exists: $\bar{x}$ in between those points such that:

$$f(x^{k+1}) - f(x^k) = f(x^k + \alpha d^k) - f(x^k) = \alpha \langle \nabla f(\bar{x}), d^k \rangle.$$

$$f(x^{k+1}) = f(x^k) + \alpha \langle \nabla f(\bar{x}) - \nabla f(x^k), d^k \rangle + \alpha \langle \nabla f(x^k), d^k \rangle$$

Cauchy Schwarz inequality:
$\alpha \langle \nabla f(\bar{x}) - \nabla f(x^k), d^k \rangle \leq \alpha \|\nabla f(\bar{x}) - \nabla f(x^k)\| \|d^k\|$.
Lipschitz: $\|\nabla f(\bar{x}) - \nabla f(x^k)\| \leq M \|\bar{x} - x^k\|$

Analysis of gradient descent method

$$f(x^{k+1}) \leq f(x^k) + \alpha\langle \nabla f(x^k), d^k\rangle + \alpha M\|\bar{x} - x^k\|\|\nabla f(x^k)\|$$

With $d^k = -\nabla f(x^k)$:

$$f(x^{k+1}) \leq f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \alpha M\|\bar{x} - x^k\|\|\nabla f(x^k)\|$$

Since $\bar{x}$ is between x^k, x^{k+1},
$\|\bar{x} - x^k\| \leq \|x^{k+1} - x^k\| = \alpha\|\nabla f(x^k)\|$:

$$f(x^{k+1}) \leq f(x^k) - \alpha(1 - \alpha M)\|\nabla f(x^k)\|^2.$$

Since $0 < \alpha < \frac{1}{M}$, $f(x^{k+1}) \leq f(x^k)$. Sequence $f(x^k)$ is decreasing and bounded from below, so it has a limit point.

$$0 \leq \alpha(1 - \alpha M)\|\nabla(f(x^k)\|^2 \leq f(x^k) - f(x^{k+1}).$$

Gradient descent - Example

Example 1

f(x) = 7x - log(x), $x^* = \frac{1}{7}$

$\nabla f(x) = 7 - \frac{1}{x}, d^k = \frac{1}{x} - 7.$

Example 2

$f(x_1, x_2) = -log(1 - x_1 - x_2) - log(x_1) - log(x_2)$

$$\nabla f(x) = \begin{pmatrix} \frac{1}{1-x_1-x_2} - \frac{1}{x_1} \\ \frac{1}{1-x_1-x_2} - \frac{1}{x_2} \end{pmatrix}$$

Gradient descent method - Pros and cons

- Easy to implement.
- Fast (if function evaluation and calculate gradient are easy enough).
- With small steplength, it can be painfully slow when it is close to the optimal solution.

Newton's method

- Consider a quadratic function: $f(x) = \frac{1}{2}x^T Q x + b^T x$
- If Q has an inverse (when is that?), then we can find the minimum of this function by: $x^* = -Q^{-1}b$
- Consider second order Taylor approximation:

$$f_{T2}(x+d) = f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2} d^T \nabla^2 f(x) d.$$

  To minize this function, we can get the solution $d^* = -(\nabla^2 f(x))^{-1} \nabla f(x)$.
- For Newton's method, we use search direction: $d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$.

Newton's method

Assume that $\nabla^2 f(x)$ is invertible at each iteration.
There is no guarantee that $f(x^{k+1}) < f(x^k)$.

- Initialization: Starting point x^0, k=0, small ϵ.
- Repeat
    1. $d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$. If $\|d^k\| < \epsilon$, stop,
    2. Set $x^{k+1} = x^k + \alpha^k d^k, k = k + 1$;

If Newton's method starting sufficiently close to the optimal solution, it will converge very fast. It can be implemented together with gradient descent method.

Example

Example 1 f(x) = 7x - log(x), $x^* = \frac{1}{7}$

$\nabla f(x) = 7 - \frac{1}{x}$, $\nabla^2 f(x) = \frac{1}{x^2}$.

$d = -(\nabla^2 f(x))^{-1} \nabla f(x) = (\frac{1}{x^2})^{-1}(7 - \frac{1}{x}) = x - 7x^2$.

Example 2 $f(x_1, x_2) = -log(1 - x_1 - x_2) - log(x_1) - log(x_2)$

$$\nabla f(x) = \begin{pmatrix} \frac{1}{1-x_1-x_2} - \frac{1}{x_1} \\ \frac{1}{1-x_1-x_2} - \frac{1}{x_2} \end{pmatrix}$$

$$H = \begin{pmatrix} \frac{1}{(1-x_1-x_2)^2} + \frac{1}{x_1^2} & \frac{1}{(1-x_1-x_2)^2} \\ \frac{1}{(1-x_1-x_2)^2} & \frac{1}{(1-x_1-x_2)^2} + \frac{1}{x_2^2} \end{pmatrix}$$