Optimization for Machine Learning

- In many situation, a problem in ML ends up in solving an optimization problem, without an explicit solution.
- Optimization methods are needed to solve these problems.
- Efficiency and accuracy of these methods are crucial to the performance of models in ML.
- Form: $min_x f(x)$:$g(x) \leq 0$, h(x) =0, $x \in X$, $x \in \mathbb{R}^n$
- Linear regression: $\min_\beta \frac{1}{2}\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_1^i + \cdots \beta_p x_p^i)]^2$

Optimization problems

- $Minimize_x : f(x) = ax^2 + bx + c$. Solution: $x^* = \frac{-b}{2a}$
- Find the maximum and minimum values of $f(x, y) = 81x^2 + y^2$ subject to the constraint: $4x^2 + y^2 = 9,\ -3 \leq y \leq 3, -3/2 \leq x \leq 3/2$.

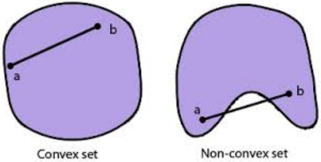
$$162x = 8x\lambda$$
$$2y = 2y\lambda$$
$$4x^2 + y^2 = 9$$

$y = 0 \rightarrow x = 3/2$ or$x = -3/2$
$\lambda = 1 \rightarrow x = 0 \rightarrow y = 3$ or $y = -3$.

Types of optimization problem

- Unconstrained optimization: $\min_x f(x)$
- Constrained optimization: $\min_x f(x)$: such that: $x \in C \subset \mathbb{R}^n$, $g(x) \leq 0, h(x) = 0$.
- Linear programming: $\min_x \langle f, x \rangle : Ax \leq b, Bx = c$, $A, B \in \mathbb{R}^{m \times n}$
  Food and budget problem.
- Quadratic programming: $\min_x \frac{1}{2} x^T Q x + \langle f, x \rangle : Ax \leq b, Bx = c,\ A, B \in \mathbb{R}^{m \times n}$

- Convex set: $X \subset \mathbb{R}^n$ is convex if $\forall x^1, x^2 \in X$: $\alpha x^1 + (1-\alpha)x^2 \in X$, $0 < \alpha < 1$.
  E.g: A disk in $\mathbb{R}^2$:$\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$ is a convex set.
  Interval $[0, 1]$ is a convex set.

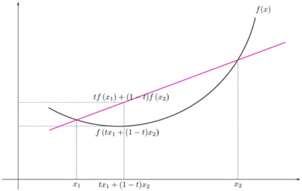


- Let $X_i \subset \mathbb{R}^n$, then $X = \cap_i X_i$ is a convex set.
- Let $X, Y \subset \mathbb{R}^n$ are convex sets, and a,b are real numbers, then set $Z = aX + bY$ is a convex set.

Convex function

- A function f is convex if and only if for all points x^1, x^2 and for all $0 \leq t \leq 1$:

$$f(tx^1 + (1-t)x^2) \leq tf(x^1) + (1-t)f(x^2).$$

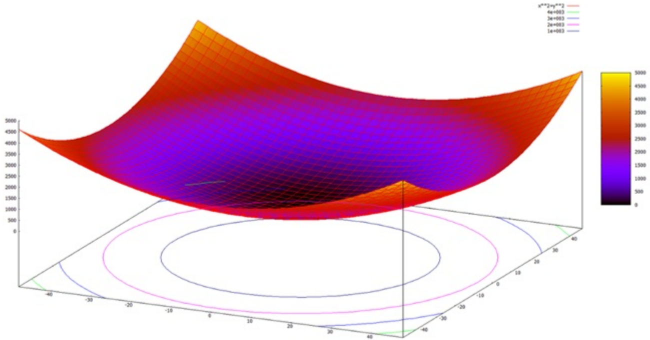


- Examples: $f(x) = x^2$, $f(x) = \|x\|_1$.
- A function f is strictly convex if and only if for all points x^1, x^2 and for all $0 \leq t \leq 1$:

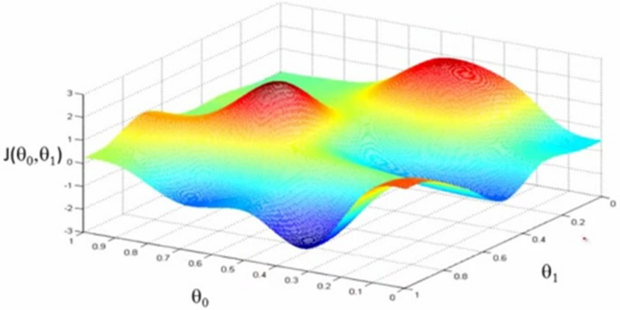
$$f(tx^1 + (1-t)x^2) < tf(x^1) + (1-t)f(x^2).$$

- Let $f_i(x)$ are convex functions, c_i are positive scalars then $g(x) = \sum_{i=1}^{n} c_i f_i(x)$ is a convex function.

Convex functions



Not convex



$J(\theta_0,\theta_1)$
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ_0
0
0.2
0.4
0.6
0.8
1
θ_1

Gradient and Hessian

- $f : \mathbb{R}^{n\times 1} \rightarrow R$ is a differentiable function that takes an input as a vector x of size $n \times 1$ and output a real number. Gradient of f is the vector of partial derivatives that has size $n \times 1$: $\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$
- $\nabla(f+g)(x) = \nabla f(x) + \nabla g(x)$, $\nabla(af(x)) = a\nabla f(x)$

Gradient and Hessian

- Example: $f(x) = 5x_1^2x_2^3$, we compute the gradient as:

$$\frac{\partial f(x)}{\partial x_1} = 10x_1x_2^3; \frac{\partial f(x)}{\partial x_2} = 15x_1^2x_2^2$$

Let $x = (1, -2)$ then $\nabla f(x) = \begin{pmatrix} 10(-2)^3 \\ 15(-2) \end{pmatrix} = \begin{pmatrix} -80 \\ -30 \end{pmatrix}$

Hessian

- If the function f is twice differentiable, the Hessian matrix $\nabla^2 f(x)$ is the $n \times n$ matrix of partial derivatives:

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial f(A)}{\partial x_1 x_1} & \frac{\partial f(A)}{\partial x_1 x_2} & \dots & \frac{\partial f(A)}{\partial x_1 x_n} \\ \frac{\partial f(A)}{\partial x_2 x_1} & \frac{\partial f(A)}{\partial x_2 x_2} & \dots & \frac{\partial f(A)}{\partial x_2 x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f(A)}{\partial x_n x_1} & \frac{\partial f(A)}{\partial x_n x_2} & \dots & \frac{\partial f(A)}{\partial x_n x_n} \end{pmatrix}.$$

Hessian

- $f(x) = 5x_1^2 x_2^3$, we compute the Hessian as:

$$\frac{\partial f(x)}{\partial x_1 \partial x_1} = 10x_2^3; \frac{\partial f(x)}{\partial x_1 \partial x_2} = 30x_1 x_2^2$$
$$\frac{\partial f(x)}{\partial x_2 \partial x_1} = 30x_1 x_2^2; \frac{\partial f(x)}{\partial x_2 \partial x_2} = 30x_1^2 x_2$$

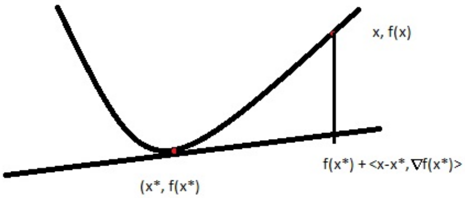
At x=(1,-2), $\nabla^2 f(x) = \begin{pmatrix} -80 & 120 \\ 120 & -60 \end{pmatrix}$

- Notice that the Hessian a symmetric matrix.

Taylor series expansion

- Let $f : \mathbb{R}^n \to R$ be a differentiable function, a point x^*,then the first order or linear Taylor series approximation of f at point x^* is:

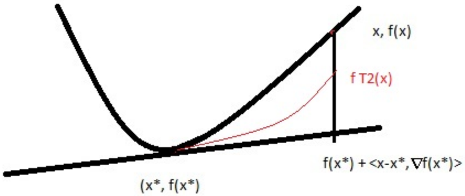
$$f_{T1}(x) = f(x^*) + \langle x - x^*, \nabla f(x^*) \rangle$$



Second order or quadratic Taylor series

- The second order or quadratic Taylor series of f at x^*:

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



Taylor series approximation

- Example: $f(x_1, x_2) = x_1 ln(x_2) + 2, x^* = (-3, 1)$
- $\nabla f(x^*) = \begin{pmatrix} ln(x_2) \\ \frac{x_1}{x_2} \end{pmatrix} = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$
- First order approximation:

$$f_{T1}(x^*) = f(x^*) + \langle x - x^*, \nabla f(x^*) \rangle$$

  Consider a point $y = (4, 2)$ then
  $f_{T1}(y) = 2 + \langle y - x^*, \nabla f(x^*) \rangle$= 2+7 * 0 + 1*(-3) = -1.
- Second order Taylor approximation:
  $\nabla^2 f(x^*) = \begin{pmatrix} 0 & \frac{1}{x_2} \\ \frac{1}{x_2} & -\frac{x_1}{x_2^2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 3 \end{pmatrix}$

Taylor approx

- Second order approx:

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

$$f_{T2}(y) = -1 + \frac{1}{2}[7, 1]^T \begin{pmatrix} 0 & 1 \\ 1 & 3 \end{pmatrix} [7, 1] = -1 + 8.5 = 7.5$$

Convexity

- Assume function f is differentiable, then f is onvex if and only if for all x,y: $f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle$
- Assume function f has gradient and Hessian then, f is a convex function if and only if its Hessian $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^n$. If the Hessian positive definite for all $x \in \mathbb{R}^n$, then f is strictly convex.

Solutions of optimization problem

- $\min_x f(x)$
- Local optimal solution: A point is a local optimal if and only if there exists some $\delta > 0$ such that $\forall z : \|x - z\|_2 \leq \delta$, we have $f(x) \leq f(z)$.
- Global optimal solution: A point is a global optimal if and only if $f(x) \leq f(z) \forall z$

Optimality for unconstrained differentiable functions

- Assume f is differentiable at a point x^*, if f attains its local minimum at x^* then $\nabla f(x^*) = 0$.
  If f is a convex function and $\nabla f(x^*) = 0$, then x^* is a global minimum of f.
- $f(x) = \frac{1}{2}x^T Qx + f^T x$, and Q is positive definite then the global minimum of f satisfy:$\nabla f(x) = Qx + b = 0$. or $x = Q^{-1}b$.

Iterative methods in optimization

- Iterative methods is a computational procedure that generates a sequence of points that are impoving approximate solutions for a problem.
- Initialization: function f, x^0 is a starting point, k=0 is a number indicating the current number of iteration
- At a iteration k:
  1. Find a direction d_k with some procedure.
  2. Find a proper step length α_k and update $x_{k+1} = x_k + \alpha_k d_k$
  3. Check for stopping condition.
  4. k=k+1