Introduction to Machine Learning
Lecture 7

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Convex Optimization

Differentiation

- Definition: let $f\colon X \subseteq \mathbb{R}^N \to \mathbb{R}$ be a differentiable function, then the gradient of f at $\mathbf{x} \in X$ is defined by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_N}(x) \end{bmatrix}.$$

- Definition: let $f\colon X \subseteq \mathbb{R}^N \to \mathbb{R}$ be a twice differentiable function, then the Hessian of f at $\mathbf{x} \in X$ is defined by

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right]_{1 \leq i,j \leq N}.$$

Unconstrained Optimization

(Fermat, 1629)

- **Theorem**: let $f\colon X \subseteq \mathbb{R}^N \to \mathbb{R}$ be a differentiable function. If f admits a local extremum at $x^* \in X$, then

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

  - x^* is a stationary point.
  - a local minimum is a global minimum if the function is convex.

Convexity

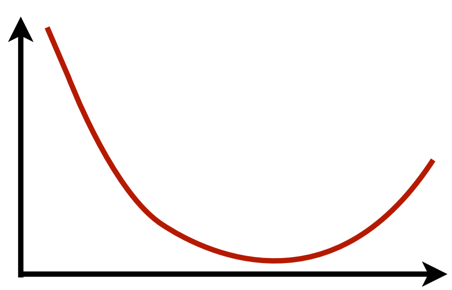
- Definition: $X \subseteq \mathbb{R}^N$ is said to be convex if for any two points $x, y \in X$ the segment $[x, y]$ lies in X:

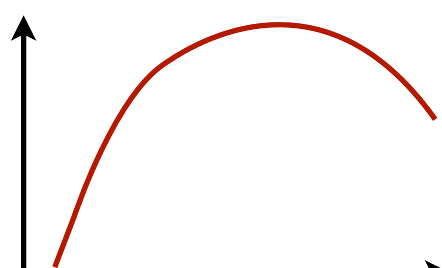
$$\{\alpha x + (1-\alpha)y, 0 \leq \alpha \leq 1\} \subseteq X.$$

- Definition: let X be a convex set. A function $f : X \to \mathbb{R}$ is said to be convex if for all $x, y \in X$ and $\alpha \in [0, 1]$,

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

With a strict inequality, f is said to be strictly convex. f is said to be concave when $-f$ is convex.

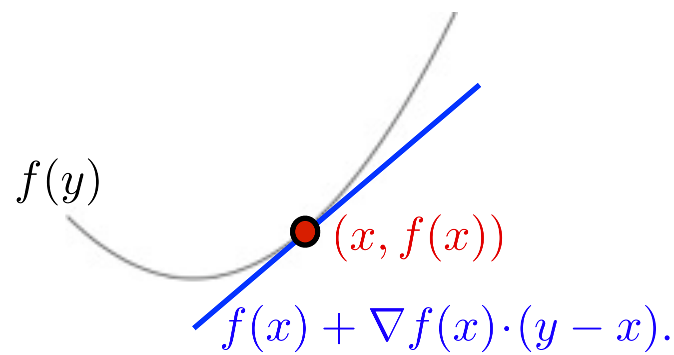




Properties of Convex Functions

- **Theorem**: let f be a differentiable function. Then, f is convex iff $\mathrm{dom}(f)$ is convex and

$$\forall x, y \in \mathrm{dom}(f),\ \ f(y) - f(x) \geq \nabla f(x) \cdot (y - x).$$



- **Theorem**: let f be a twice differentiable function. Then, f is convex iff its Hessian is positive semi-definite:

$$\forall x \in \mathrm{dom}(f),\ \ \nabla^2 f(x) \succeq 0.$$

Constrained Optimization Problem

- Problem: Let $X \subseteq \mathbb{R}^N$ and $f, g_i \colon X \to \mathbb{R}$, $i \in [1, m]$. A constrained optimization problem has the form:

$$\min_{\mathbf{x} \in X} \; f(\mathbf{x})$$
$$\text{subject to: } \; g_i(\mathbf{x}) \leq 0, i \in [1, m].$$

  - no convexity assumption.
  - can be augmented with equality constraints.
  - primal problem.
  - optimal value p^*.

Lagrangian/Lagrange Function

- Definition: the Lagrange function or Lagrangian associated to a constraint problem is the function defined by:

$$\forall \mathbf{x} \in X, \forall \boldsymbol{\alpha} \geq 0, L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}).$$

  - α_is are called Lagrange or dual variables.

Lagrange Dual Function

- **Definition**: the (Lagrange) dual function associated to the constraint optimization problem is defined by

$$\forall \boldsymbol{\alpha} \geq 0, F(\boldsymbol{\alpha}) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\alpha})$$
$$= \inf_{\mathbf{x} \in X} f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}).$$

  - F is always concave: Lagrangian is linear with respect to α and $\inf$ preserves concavity.
  - $\forall \boldsymbol{\alpha} \geq 0, F(\boldsymbol{\alpha}) \leq p^*$: for a feasible $\mathbf{x}$,

$$f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

Dual Optimization Problem

- Definition: the dual (optimization) problem associated to the constraint optimization is

$$\max_{\boldsymbol{\alpha}} \quad F(\boldsymbol{\alpha})$$
$$\text{subject to:} \quad \boldsymbol{\alpha} \geq 0.$$

  - always a convex optimization problem.
  - optimal value d^*.

Weak and Strong Duality

- Weak duality: $d^* \leq p^*$.
  - always holds (clear from previous observations).
- Strong duality: $d^* = p^*$.
  - does not hold in general.
  - holds for convex problems with constraint qualifications.

Constraint Qualification

- Definition: Assume that $\mathbf{int}\, X \neq \emptyset$. Then, the following is the strong constraint qualification or Slater's condition:

$$\exists\, \overline{\mathbf{x}} \in \mathbf{int}\, X \colon\ g(\overline{\mathbf{x}}) < 0.$$

- Definition: Assume that $\mathbf{int}\, X \neq \emptyset$. Then, the following is the weak constraint qualification or Slater's condition:

$$\exists\, \overline{\mathbf{x}} \in \mathbf{int}\, X \colon\ \forall i \in [1, m], \big(g_i(\overline{\mathbf{x}}) < 0\big) \vee \big(g_i(\overline{\mathbf{x}}) = 0 \wedge g_i \text{ affine}\big).$$

Saddle Point - Sufficient Condition

(Lagrange, 1797)

- Theorem: Let P be a constrained optimization problem over $X = \mathbb{R}^N$. If $(\mathbf{x}^*, \boldsymbol{\alpha}^*)$ is a saddle point, that is $\forall \mathbf{x} \in \mathbb{R}^N, \forall \boldsymbol{\alpha} \geq 0,\ L(\mathbf{x}^*, \boldsymbol{\alpha}) \leq L(\mathbf{x}^*, \boldsymbol{\alpha}^*) \leq L(\mathbf{x}, \boldsymbol{\alpha}^*),$ then it is a solution of P.

- Proof: By the first inequality,

$$\forall \boldsymbol{\alpha} \geq 0, L(\mathbf{x}^*, \boldsymbol{\alpha}) \leq L(\mathbf{x}^*, \boldsymbol{\alpha}^*) \Rightarrow \forall \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha} \cdot g(\mathbf{x}^*) \leq \boldsymbol{\alpha}^* \cdot g(\mathbf{x}^*)$$

$$(\text{use } \alpha \rightarrow +\infty \text{ then } \alpha \rightarrow 0) \Rightarrow \boxed{g(\mathbf{x}^*) \leq 0 \wedge \boldsymbol{\alpha}^* \cdot g(\mathbf{x}^*) = 0.}$$

  - In view of that, the second inequality gives

$$\forall \mathbf{x}, L(\mathbf{x}^*, \boldsymbol{\alpha}^*) \leq L(\mathbf{x}, \boldsymbol{\alpha}^*) \Rightarrow \forall \mathbf{x}, f(\mathbf{x}^*) \leq f(\mathbf{x}) + \boldsymbol{\alpha}^* \cdot g(\mathbf{x}).$$

Thus, for all x such that $g(x) \leq 0$, $\boxed{f(\mathbf{x}^*) \leq f(\mathbf{x}).}$

Saddle Point - Necessary Conditions

- Theorem: Assume that f and g_i , $i \in [1, m]$, are convex functions and that Slater's condition holds. If $\mathbf{x}$ is a solution of the constrained optimization problem, then there exists $\alpha \geq 0$ such that $(\mathbf{x}, \boldsymbol{\alpha})$ is a saddle point of the Lagrangian.

- Theorem: Assume that f and g_i , $i \in [1, m]$, are convex differentiable functions and that the weak Slater's condition holds. If $\mathbf{x}$ is a solution of the constrained optimization problem, then there exists $\alpha \geq 0$ such that $(\mathbf{x}, \boldsymbol{\alpha})$ is a saddle point of the Lagrangian.

Kuhn-Tucker's Theorem

(Karush 1939; Kuhn-Tucker, 1951)

- **Theorem**: Assume that $f, g_i \colon X \to \mathbb{R}, i \in [1, m]$ are convex and differentiable and that the constraints are qualified. Then $\overline{\mathbf{x}}$ is a solution of the constrained program iff there exists $\overline{\boldsymbol{\alpha}} \geq 0$ such that:

$$\left.\begin{aligned}
&\nabla_{\mathbf{x}} L(\overline{\mathbf{x}}, \overline{\boldsymbol{\alpha}}) = \nabla_{\mathbf{x}} f(\overline{\mathbf{x}}) + \overline{\boldsymbol{\alpha}} \cdot \nabla_{\mathbf{x}} g(\overline{\mathbf{x}}) = 0 \\
&\nabla_{\boldsymbol{\alpha}} L(\overline{\mathbf{x}}, \overline{\boldsymbol{\alpha}}) = g(\overline{\mathbf{x}}) \leq 0 \\
&\overline{\boldsymbol{\alpha}} \cdot g(\overline{\mathbf{x}}) = \sum_{i=1}^{m} \overline{\boldsymbol{\alpha}}_i g(\overline{\mathbf{x}}_i) = 0.
\end{aligned}\right\} \text{KKT conditions}$$

- **Note**: Last two conditions equivalent to

$$\big(g(\overline{\mathbf{x}}) \leq 0\big) \wedge \big(\underbrace{\forall i \in [1, m], \bar{\alpha}_i g_i(\overline{\mathbf{x}}) = 0}_{\text{complementary conditions}}\big).$$

- Since the constraints are qualified, if $\overline{\mathbf{x}}$ is solution, then there exists $\overline{\boldsymbol{\alpha}}$ such that $(\overline{\mathbf{x}}, \overline{\boldsymbol{\alpha}})$ is a saddle point. In that case, the three conditions are verified (for the 3rd condition see proof of sufficient condition slide).
- Conversely, assume that the conditions are verified. Then, for any $\mathbf{x}$ such that $g(\mathbf{x}) < 0$,

$$
\begin{aligned}
f(\mathbf{x}) - f(\overline{\mathbf{x}}) &\geq \nabla_{\mathbf{x}} f(\overline{\mathbf{x}}) \cdot (\mathbf{x} - \overline{\mathbf{x}}) && \text{(convexity of } f\text{)} \\
&\geq -\sum_{i=1}^{m} \overline{\boldsymbol{\alpha}}_i \nabla_{\mathbf{x}} g_i(\overline{\mathbf{x}}) \cdot (\mathbf{x} - \overline{\mathbf{x}}) && \text{(first condition)} \\
&\geq -\sum_{i=1}^{m} \overline{\boldsymbol{\alpha}}_i [g_i(\mathbf{x}) - g_i(\overline{\mathbf{x}})] && \text{(convexity of } g_i\text{s)} \\
&\geq -\sum_{i=1}^{m} \overline{\boldsymbol{\alpha}}_i g_i(\mathbf{x}) \geq 0. && \text{(third and second condition)}
\end{aligned}
$$

References

- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*, Cambridge University Press, 2004.