Optimization Background (b)

Mingyi Hong

University Of Minnesota

Outline

- Basic Concepts in Nonlinear Optimization
- How to Analyze Algorithm Convergence
- Basic Concepts on Graph Theory

Constrained Problems

Constrained Problem

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- Here f is continuously differentiable, X is a convex set

Constrained Problem

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- Here f is continuously differentiable, X is a convex set
- Convex set X means we allow the following types of constraints
  - $g(x) \leq 0$ where $g(x)$ is a convex function
  - $h(x) = 0$ where $h(x)$ is an affine function: $Cx + d = 0$

Constrained Problem

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- Here f is continuously differentiable, X is a convex set
- Convex set X means we allow the following types of constraints
  - $g(x) \leq 0$ where $g(x)$ is a convex function
  - $h(x) = 0$ where $h(x)$ is an affine function: $Cx + d = 0$
- Why $g(x) = 0$ is not a convex set? Consider $x_1^2 + x_2^2 = 1$

Optimality Conditions

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- How to characterize the global/local optimal solution x^*?
- Suppose X is a convex set; f is continuously differentiable

- **Claim (a)** If x^* is a local minimum of the above constrained minimization problem, then (necessary condition)

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall \ x \in X$$

Optimality Conditions

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- How to characterize the global/local optimal solution x^*?
- Suppose X is a convex set; f is continuously differentiable

- **Claim (a)** If x^* is a local minimum of the above constrained minimization problem, then (necessary condition)

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall \ x \in X$$

- **Claim (b)** If further assume that, f is a convex function, this condition is also sufficient for x^* to minimize f over X

Optimality Conditions

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- How to characterize the global/local optimal solution x^*?
- Suppose X is a convex set; f is continuously differentiable

- **Claim (a)** If x^* is a local minimum of the above constrained minimization problem, then (necessary condition)

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \; \forall \; x \in X$$

- **Claim (b)** If further assume that, f is a convex function, this condition is also sufficient for x^* to minimize f over X

- **Note:** If f is nonconvex, solutions satisfying this condition is called stationary point (nowhere to move)

Proof of Claim a

- Suppose that $\nabla f(x^*)'(x - x^*) < 0$ for some $x \in X$.
- We are going to show that $f(x^*)$ is not a local min [graph]
- How to show? Let's go towards x!
- By the MVT, for every $\epsilon > 0$ there exists an $s \in [0, 1]$ such that $$f(x^* + \epsilon(x - x^*)) = f(x^*) + \epsilon \nabla f(x^* + s\epsilon(x - x^*))'(x - x^*).$$ [at x^*, go towards x with a small step ϵ]
- Since ∇f is continuous, for sufficiently small $s > 0$, $$g(s) := \nabla f(x^* + s\epsilon(x - x^*))'(x - x^*) < 0,$$
- The above two imply that there exists small $\epsilon > 0$ small such that $$f(x^* + \epsilon(x - x^*)) < f(x^*)$$
- The vector $x^* + \epsilon(x - x^*)$ is feasible for all $\epsilon \in [0, 1]$ because X is convex, contradicting the local optimality of x^*.

Proof of Claim b

Claim (b) If further assume that, f is a convex function, this condition is also sufficient for x^* to minimize f over X

- Using the convexity of f

$$f(x) \geq f(x^*) + \nabla f(x^*)'(x - x^*)$$

for every $x \in X$.

- If the condition $\nabla f(x^*)'(x - x^*) \geq 0$ holds for all $x \in X$,we obtain $f(x) \geq f(x^*)$.

- So x^* minimizes f over X.

Example: Optimization over a Simplex

Let us consider the following simplex constraint:
$X = \{x \mid x \geq 0, \sum_{i=1}^{n} x_i = r\}$, where $r > 0$ is given scalar.

- Necessary condition for $x^* = (x_1^*, \ldots, x_n^*)'$ to be a local min:

$$\sum_{i=1}^{n} \frac{\partial f(x^*)}{\partial x_i}(x_i - x_i^*) \geq 0, \forall x_i \geq 0 \text{ with } \sum_{i=1}^{n} x_i = r$$

- Fix i with $x_i^* > 0$ and let j be any other index. Use x with $x_i = 0$, $x_j = x_j^* + x_i^*$, and $x_m = x_m^*$ for all $m \neq i, j$:

$$\left(\frac{\partial f(x^*)}{\partial x_j} - \frac{\partial f(x^*)}{\partial x_i}\right)x_i^* \geq 0,$$

$$x_i^* > 0 \Longrightarrow \frac{\partial f(x^*)}{\partial x_i} \leq \frac{\partial f(x^*)}{\partial x_j}, \forall j.$$

Projection Over Convex Set

- Projection is a very important operation to deal with constraints [figure]
- **Question:** What if a point is out of the feasible set X?
- **Answer:** "Project" it back to X!

Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;

Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;
- **Problem:** Find a point $x \in X$, such that it is closest to z

Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;
- **Problem:** Find a point $x \in X$, such that it is closest to z

Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;
- **Problem:** Find a point $x \in X$, such that it is closest to z
- Consider the following:

$$f(x) = \|z - x\|^2, \text{subject to } x \in X. \tag{1.1}$$

Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;
- **Problem:** Find a point $x \in X$, such that it is closest to z
- Consider the following:

$$f(x) = \|z - x\|^2, \text{subject to } x \in X. \tag{1.1}$$

- This problem has a unique solution, and we denote it as $x^* = \text{proj}[z]$ (the projection of z); Optimality condition?

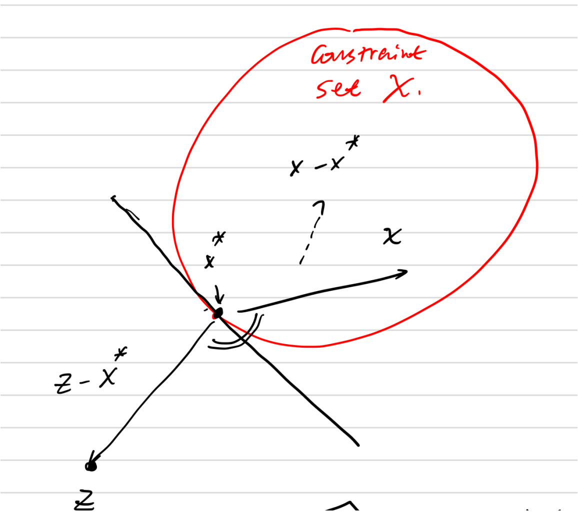
Projection Over Convex Set

- Let $z \in R^n$ and a closed convex set X be given;
- **Problem:** Find a point $x \in X$, such that it is closest to z
- Consider the following:

$$f(x) = \|z - x\|^2, \text{subject to } x \in X. \tag{1.1}$$

- This problem has a unique solution, and we denote it as $x^* = \text{proj}[z]$ (the projection of z); Optimality condition?
- $x^* = \text{proj}[z] \Longleftrightarrow$ The angle between $z - x^*$ and $x - x^*$ is greater or equal to 90 degrees for all $x \in X$, or $(z - x^*)'(x - x^*) \leq 0$

Projection Over Convex Set



Projection Over Convex Set

- **Key Property:** The mapping $f : \mathbb{R}^n \mapsto X$ defined by $f(z) = \text{proj}[z]$ is continuous and satisfying the following property, that is,

$$\|\text{proj}[z] - \text{proj}[y]\| \leq \|z - y\|, \; \forall z, y \in R^{n.}$$

- Intuitively is this true?

- Pic on board

Projection Over Convex Set

- **Proof**: From the optimality condition

$$\langle z - x^*, x - x^* \rangle \leq 0, \quad \forall\, x \in X \tag{1.2}$$

Projection Over Convex Set

- **Proof**: From the optimality condition

$$\langle z - x^*, x - x^* \rangle \leq 0, \quad \forall\, x \in X \tag{1.2}$$

- x^* is defined as proj$[z]$, so

$$\langle z - \text{proj}[z], x - \text{proj}[z] \rangle \leq 0, \quad \forall\, x \in X \tag{1.3}$$

Projection Over Convex Set

- **Proof**: From the optimality condition

$$\langle z - x^*, x - x^* \rangle \leq 0, \quad \forall\, x \in X \tag{1.2}$$

- x^* is defined as proj$[z]$, so

$$\langle z - \text{proj}[z], x - \text{proj}[z] \rangle \leq 0, \quad \forall\, x \in X \tag{1.3}$$

- similarly, for another point $y \in \mathbb{R}^n$

$$\langle y - \text{proj}[y], x - \text{proj}[y] \rangle \leq 0, \quad \forall\, x \in X \tag{1.4}$$

Projection Over Convex Set

- **Proof**: From the optimality condition

$$\langle z - x^*, x - x^* \rangle \leq 0, \quad \forall\, x \in X \tag{1.2}$$

- x^* is defined as proj$[z]$, so

$$\langle z - \text{proj}[z], x - \text{proj}[z] \rangle \leq 0, \quad \forall\, x \in X \tag{1.3}$$

- similarly, for another point $y \in \mathbb{R}^n$

$$\langle y - \text{proj}[y], x - \text{proj}[y] \rangle \leq 0, \quad \forall\, x \in X \tag{1.4}$$

- plugin $x = \text{proj}[y]$ in (1.3), and $x = \text{proj}[z]$ in (1.4), and sum:

$$\langle y - \text{proj}[y] - (z - \text{proj}[z]), \text{proj}[z] - \text{proj}[y] \rangle \leq 0, \quad \forall\, x \in X$$

Projection Over Convex Set

- **Proof**: From the optimality condition

$$\langle z - x^*, x - x^* \rangle \le 0, \quad \forall\, x \in X \tag{1.2}$$

- x^* is defined as proj$[z]$, so

$$\langle z - \mathsf{proj}[z], x - \mathsf{proj}[z] \rangle \le 0, \quad \forall\, x \in X \tag{1.3}$$

- similarly, for another point $y \in \mathbb{R}^n$

$$\langle y - \mathsf{proj}[y], x - \mathsf{proj}[y] \rangle \le 0, \quad \forall\, x \in X \tag{1.4}$$

- plugin $x = $ proj$[y]$ in (1.3), and $x = $ proj$[z]$ in (1.4), and sum:

$$\langle y - \mathsf{proj}[y] - (z - \mathsf{proj}[z]), \mathsf{proj}[z] - \mathsf{proj}[y] \rangle \le 0, \quad \forall\, x \in X$$

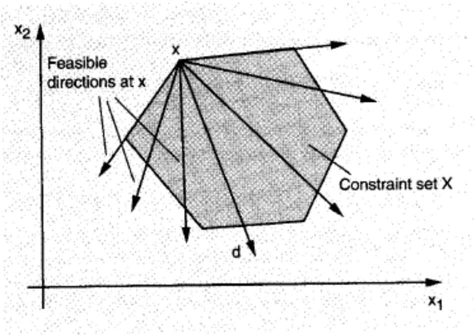
- move terms around, we obtain [then one more step to go using Cauchy-Swartz]

$$\|\mathsf{proj}[z] - \mathsf{proj}[y]\|^2 \le \langle y - z, \mathsf{proj}[z] - \mathsf{proj}[y] \rangle$$

Feasible Directions

- A feasible direction at an $x \in X$ is a vector $d \neq 0$ such that $x + \alpha d$ is feasible for all sufficiently small $\alpha > 0$

- The set of feasible directions at x is the set of all $(z - x)$ where $z \in X$, $z \neq x$ (for convex sets).



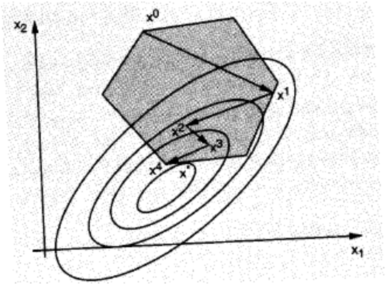
Feasible Directions Method

- A feasible direction method:

$$x^{r+1} = x^r + \alpha_r d^r,$$

where d^r is:

1. A feasible descent direction, i.e., $\nabla f(x^r)' d^r < 0$;
2. $\alpha_r > 0$ is such that $x^{r+1} \in X$ (similar as GD).



Gradient Projection (GP) Methods

- **Special case I:** Stepsize rules for α_r: let us fix $s_r \equiv s$, then we choose α_r by Limited minimization, or Armijo rule.

Gradient Projection (GP) Methods

- **Special case I:** Stepsize rules for α_r: let us fix $s_r \equiv s$, then we choose α_r by Limited minimization, or Armijo rule.

- **Special case II:** Stepsize rules for s_r: let us fix $\alpha_r \equiv 1$, we can use Armijo rule, or constant stepsize; the iteration becomes simply

$$\bar{x}^r = \mathsf{proj}_X[x^r - s_r \nabla f(x^r)]$$

Gradient Projection (GP) Methods

- **Special case I:** Stepsize rules for α_r: let us fix $s_r \equiv s$, then we choose α_r by Limited minimization, or Armijo rule.

- **Special case II:** Stepsize rules for s_r: let us fix $\alpha_r \equiv 1$, we can use Armijo rule, or constant stepsize; the iteration becomes simply

$$\bar{x}^r = \mathsf{proj}_X[x^r - s_r \nabla f(x^r)]$$

- Variant I is precisely the feasible direction method we mentioned; Variant II is actually simple and more popular

Gradient Projection (GP) Methods

- **Special case I:** Stepsize rules for α_r: let us fix $s_r \equiv s$, then we choose α_r by Limited minimization, or Armijo rule.

- **Special case II:** Stepsize rules for s_r: let us fix $\alpha_r \equiv 1$, we can use Armijo rule, or constant stepsize; the iteration becomes simply

$$\bar{x}^r = \mathsf{proj}_X[x^r - s_r \nabla f(x^r)]$$

- Variant I is precisely the feasible direction method we mentioned; Variant II is actually simple and more popular

- When we refer to "GP" method, usually variant II; **Question:** connection between Variant II and GD?

Perform the Projection

- Solve

$$\min \frac{1}{2}\|x - y\|^2, \quad \text{s.t.. } x \geq 0$$

- **Solution** [graphically]

$$x_i^* = y_i, \ \text{ if } y_i \geq 0, \quad x_i^* = 0 \text{ otherwise}, \quad \forall\ i$$

or compact we can write $y = [y]^+$ (means taking the positive part)

Perform the Projection

- Solve

$$\min \frac{1}{2}\|x - y\|^2, \quad \text{s.t.. } x \geq 0$$

- **Solution** [graphically]

$$x_i^* = y_i, \ \text{ if } y_i \geq 0, \quad x_i^* = 0 \text{ otherwise}, \quad \forall \ i$$

or compact we can write $y = [y]^+$ (means taking the positive part)

- **Why**? Check optimality condition

$$\sum_{i=1}^{K} \langle x_i^* - y_i, x_i - x_i^* \rangle \geq 0, \ \forall \ x \geq 0$$

- If $y_i \geq 0$, then $\langle x_i^* - y_i, x_i - x_i \rangle = 0$
- if $y_i \leq 0$, then $\langle 0 - y_i, x_i - 0 \rangle \geq 0$
- Verified that the [solution] satisfies the optimality condition

Convergence of GP (Version I)

- Fix s, if α_r is chosen by the limited minimization rule or Armijo rule, every limit point of $\{x^r\}$ is stationary; [Prop. 2.3.1 in book]
- **Proof:** Show that the direction sequence $\overline{x}^r - x^r$ is gradient related. Assume x^r is a nonstationary solution. Must prove

$$\nabla f(x^r)'(\overline{x}^r - x^r) < 0.$$

- Note that $\{\overline{x}^r - x^r\}_{r\in K}$ is given by $\mathrm{proj}_X[x^r - s\nabla f(x^r)] - x^r$. Using properties of projection

$$(x^r - s\nabla f(x^r) - \overline{x}^r)'(x - \overline{x}^r) \le 0, \text{ for all } x \in X.$$

- Applying this relation with $x = x^r$ (why this is true? see last example of part Lecture 6(a)),

$$\nabla f(x^r)'(\overline{x}^r - x^r) \le -\frac{1}{s} \| x^r - \mathrm{proj}_X[x^r - s\nabla f(x^r)] \|^2 < 0$$

Convergence of GP (Version II)

- Similar conclusion for constant stepsize $\alpha_r = 1, s_r = s$; under a Lipschitz condition on ∇f, and for small enough stepsize

$$0 \leq s \leq \frac{2}{L}$$

.

- See Prop. 2.3.2 in book (V2)

- Similar conclusion for Armijo rule for this case.

Convergence Rate

- Consider a strongly convex quadratic function
  $f(x) = \frac{1}{2}x'Ax + b'x$, with $A \succ 0$.

Convergence Rate

- Consider a strongly convex quadratic function $f(x) = \frac{1}{2}x'Ax + b'x$, with $A \succ 0$.
- $\exists$ a unique solution $x^* \in X$ satisfying $x^* = \text{proj}_X[x^* - s\nabla f(x^*)]$, for any $s > 0$ (why?)

Convergence Rate

- Consider a strongly convex quadratic function $f(x) = \frac{1}{2}x'Ax + b'x$, with $A \succ 0$.
- $\exists$ a unique solution $x^* \in X$ satisfying $x^* = \mathrm{proj}_X[x^* - s\nabla f(x^*)]$, for any $s > 0$ (why?)
- So we have

$$
\begin{aligned}
\| x^{r+1} - x^* \| &= \| \mathrm{proj}_X[x^r - s\nabla f(x^r)] - \mathrm{proj}_X[x^* - s\nabla f(x^*)] \| \\
&\le \| (x^r - x^*) - s(\nabla f(x^r) - \nabla f(x^*)) \| \\
&= \| (I - sA)(x^r - x^*) \| \\
&\le \|I - sA\| \|(x^r - x^*)\| \\
&\le \max\{|1 - s\lambda_{\min}|, |1 - s\lambda_{\max}|\} \| x^r - x^* \| \\
&\le \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right) \| x^r - x^* \| = \left(1 - \frac{2}{\kappa + 1}\right) \| x^r - x^* \|
\end{aligned}
$$

In the last inequality we choose

$$s = \frac{2}{\lambda_{\min} + \lambda_{\max}}$$

Lagrangian Multipliers

Lagrangian Multiplier

$$
\begin{aligned}
\text{minimize} \quad & f(x) \\
\text{s.t.} \quad & h_i(x) = 0, \quad i = 1, \cdots, m \qquad (2.1) \\
& g_j(x) \leq 0, \quad j = 1, \cdots, n \qquad (2.2)
\end{aligned}
$$

Reminder: The problem is called convex problem if

- $f(x)$ is a convex function
- $h_i(x)$ is an affine function, i.e., $h_i(x) = Ax + b$
- $g_j(x)$ is a convex function

The Lagrangian Function

- The Lagrangian can be formed using the Lagrangian multipliers $\lambda_i \geq 0$ and $\nu_i \in \mathbb{R}$

$$L(x, \lambda, \nu) = f(x) + \underbrace{\sum_{j=1}^{n} \lambda_j g_j(x)}_{\text{inequality constraints}} + \underbrace{\sum_{i=1}^{m} \nu_i h_i(x)}_{\text{equality constraints}}$$

The Lagrangian Function

- The Lagrangian can be formed using the Lagrangian multipliers $\lambda_i \geq 0$ and $\nu_i \in \mathbb{R}$

$$L(x, \lambda, \nu) = f(x) + \underbrace{\sum_{j=1}^{n} \lambda_j g_j(x)}_{\text{inequality constraints}} + \underbrace{\sum_{i=1}^{m} \nu_i h_i(x)}_{\text{equality constraints}}$$

- The Lagrangian dual function

$$L^*(\lambda, \nu) = \inf_{x \in X} L(x, \lambda, \nu) = \inf_{x \in X} f(x) + \sum_{j=1}^{n} \lambda_j g_j(x) + \sum_{i=1}^{m} \nu_i h_i(x)$$

The Lagrangian Function

- The Lagrangian can be formed using the Lagrangian multipliers $\lambda_i \geq 0$ and $\nu_i \in \mathbb{R}$

$$L(x,\lambda,\nu) = f(x) + \underbrace{\sum_{j=1}^{n} \lambda_j g_j(x)}_{\text{inequality constraints}} + \underbrace{\sum_{i=1}^{m} \nu_i h_i(x)}_{\text{equality constraints}}$$

- The Lagrangian dual function

$$L^*(\lambda,\nu) = \inf_{x\in X} L(x,\lambda,\nu) = \inf_{x\in X} f(x) + \sum_{j=1}^{n} \lambda_j g_j(x) + \sum_{i=1}^{m} \nu_i h_i(x)$$

- The Dual Problem (where $\lambda := \{\lambda_i\}$, $\nu := \{\nu_i\}$)

$$\max_{\lambda,\nu} \quad L^*(\lambda,\nu), \quad \text{s.t. } \lambda \geq 0 \tag{2.3}$$

The Lagrangian Function

- The Lagrangian can be formed using the Lagrangian multipliers $\lambda_i \geq 0$ and $\nu_i \in \mathbb{R}$

$$L(x, \lambda, \nu) = f(x) + \underbrace{\sum_{j=1}^{n} \lambda_j g_j(x)}_{\text{inequality constraints}} + \underbrace{\sum_{i=1}^{m} \nu_i h_i(x)}_{\text{equality constraints}}$$

- The Lagrangian dual function

$$L^*(\lambda, \nu) = \inf_{x \in X} L(x, \lambda, \nu) = \inf_{x \in X} f(x) + \sum_{j=1}^{n} \lambda_j g_j(x) + \sum_{i=1}^{m} \nu_i h_i(x)$$

- The Dual Problem (where $\lambda := \{\lambda_i\}$, $\nu := \{\nu_i\}$)

$$\max_{\lambda, \nu} \quad L^*(\lambda, \nu), \quad \text{s.t. } \lambda \geq 0 \tag{2.3}$$

- λ_i and ν_i's can be viewed as "prices" for violating the constraints

The Lagrangian Function

- Let f^* be the optimal value of $f(x)$
- The Lagrangian dual L^* is
  - A concave function: even when the original problem is not convex (why?)
  - A lower bound: for $\lambda \geq 0$, $L^*(\lambda, \nu) \leq f^*$

An Example

$$\begin{aligned} \text{minimize} \quad & \|x\|^2 \\ \text{s.t.} \quad & Ax = b \end{aligned} \tag{2.4}$$

- Lagrangian: $L(x, \nu) = \|x\|^2 + \langle \nu, (Ax - b) \rangle$

An Example

$$\begin{aligned} \text{minimize} \quad & \|x\|^2 \\ \text{s.t.} \quad & Ax = b \end{aligned} \tag{2.4}$$

- Lagrangian: $L(x, \nu) = \|x\|^2 + \langle \nu, (Ax - b) \rangle$
- Recall $L^*(\nu) = \inf_x L(x, \nu)$

An Example

$$
\begin{aligned}
\text{minimize} \quad & \|x\|^2 \\
\text{s.t.} \quad & Ax = b
\end{aligned}
\tag{2.4}
$$

- Lagrangian: $L(x, \nu) = \|x\|^2 + \langle \nu, (Ax - b) \rangle$
- Recall $L^*(\nu) = \inf_x L(x, \nu)$
- Setting the gradient to 0, $x = -\frac{1}{2} A^T \nu$

An Example

$$
\begin{aligned}
\text{minimize} \quad & \|x\|^2 \\
\text{s.t.} \quad & Ax = b
\end{aligned}
\tag{2.4}
$$

- Lagrangian: $L(x, \nu) = \|x\|^2 + \langle \nu, (Ax - b) \rangle$
- Recall $L^*(\nu) = \inf_x L(x, \nu)$
- Setting the gradient to 0, $x = -\frac{1}{2} A^T \nu$
- Hence the dual

$$L^*(\nu) = L\left(-\frac{1}{2} A^T \nu, \nu\right) = -\frac{1}{4} \nu^T A A^T \nu - \nu^T b$$

An Example

$$\begin{aligned} \text{minimize} \quad & \|x\|^2 \\ \text{s.t.} \quad & Ax = b \end{aligned} \tag{2.4}$$

- Lagrangian: $L(x, \nu) = \|x\|^2 + \langle \nu, (Ax - b) \rangle$
- Recall $L^*(\nu) = \inf_x L(x, \nu)$
- Setting the gradient to 0, $x = -\frac{1}{2} A^T \nu$
- Hence the dual

$$L^*(\nu) = L\left(-\frac{1}{2}A^T\nu, \nu\right) = -\frac{1}{4}\nu^T A A^T \nu - \nu^T b$$

- $L^*(\nu)$ is a concave function

Duality – A General Theory

- Let d^* be the optimal objective of the dual problem

Duality – A General Theory

- Let d^* be the optimal objective of the dual problem
- Weak duality: $d^* \leq f^*$
    - Always true
    - Non-trivial lower bound for hard problems
    - Useful in approximation algorithms

Duality – A General Theory

- Let d^* be the optimal objective of the dual problem
- Weak duality: $d^* \leq f^*$
  - Always true
  - Non-trivial lower bound for hard problems
  - Useful in approximation algorithms
- Strong duality: $d^* = f^*$
  - Does not hold in general
  - If holds, sufficient to solve the dual
  - How to check if it holds?

Duality – A General Theory

- Let d^* be the optimal objective of the dual problem
- Weak duality: $d^* \leq f^*$
  - Always true
  - Non-trivial lower bound for hard problems
  - Useful in approximation algorithms
- Strong duality: $d^* = f^*$
  - Does not hold in general
  - If holds, sufficient to solve the dual
  - How to check if it holds?
- Constraint qualification
  - Normally true for convex problems
  - True if the problem is convex; And it is strictly feasible, i.e. there exists a $x \in X$ such that

$$h_i(x) = 0, \quad g_j(x) < 0$$

  - The above condition is known as the Slater's condition

Equality Constrained Problem

Let us first consider the following equality constrained problem

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 0, \quad i = 1, ..., m. \end{array}}$$

where $f : R^n \mapsto R,\ h_i : R^n \mapsto R, i = 1, ..., m$, are continuously differentiable function.

Equality Constrained Problem

Lagrange Multiplier Theorem

- Let x^* be a local min and a regular point [$\nabla h_i(x^*)$: linearly independent]. Then there exist unique scalars $\lambda_1^*, \cdots, \lambda_m^*$ such that

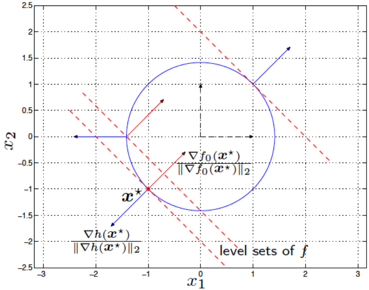
$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

- If in addition f and h are twice continuously differentiable,

$$y' \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y \geq 0, \quad \forall y \text{ s.t. } \nabla h_i(x^*)' y = 0$$

Characterizes a set of necessary conditions for local min.

Example 1



Consider the problem

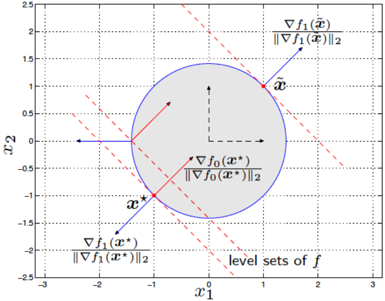
$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} & f_0(\boldsymbol{x}) = x_1 + x_2 \\ \text{subject to} & h(\boldsymbol{x}) = x_1^2 + x_2^2 - 2 = 0. \end{array}$$

This is a problem with a linear objective function $f(\boldsymbol{x})$ and one nonlinear equality constraint $h(\boldsymbol{x}) = 0$. At the solution $\boldsymbol{x}^\star$, the gradient of the constraint $\nabla h(\boldsymbol{x}^\star)$ is orthogonal to the level set of the function at $\boldsymbol{x}^*$, and hence $\nabla h(\boldsymbol{x}^\star)$ and $\nabla f_0(\boldsymbol{x}^\star)$ are parallel $i.e.$, there is a scalar $\nu^\star$ such that

$$\nabla f_0(\boldsymbol{x}^\star) + \nu^\star \nabla h(\boldsymbol{x}^\star) = \mathbf{0}.$$

Clearly, in this example $\boldsymbol{x}^\star$ is regular (because $\nabla h(\boldsymbol{x}^\star) \neq \mathbf{0}$).

Example 2



Consider the problem

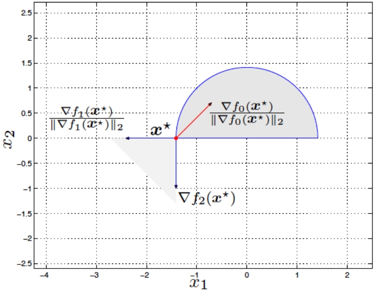
$$\begin{aligned} \underset{x \in \mathbb{R}^2}{\text{minimize}} \quad & f_0(x) = x_1 + x_2 \\ \text{subject to} \quad & f_1(x) = x_1^2 + x_2^2 - 2 \leq 0. \end{aligned}$$

This is a problem with a linear objective function $f(x)$ and one nonlinear inequality constraint $f_1(x) \leq 0$. At the solution $x^\star$, the gradient of the constraint $\nabla f_1(x^\star)$ is orthogonal to the level set of the function at x^*, and the following equality holds

$$\nabla f_0(x^\star) + \lambda^\star \nabla f_1(x^\star) = \mathbf{0},$$

for $\lambda^\star = \frac{1}{2} \geq 0$. Note that at the point $\tilde{x} = (1, 1)$, $\nabla f_0(\tilde{x}) + \lambda \nabla f_1(\tilde{x}) = \mathbf{0}$ holds as well, however, $\lambda = -\frac{1}{2} \leq 0$.

Example 3



Consider the problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \quad & f_0(\boldsymbol{x}) = x_1 + x_2 \\ \text{subject to} \quad & f_1(\boldsymbol{x}) = x_1^2 + x_2^2 - 2 \leq 0, \\ & f_2(\boldsymbol{x}) = -x_2 \leq 0. \end{aligned}$$

At the solution $\boldsymbol{x}^\star = (-\sqrt{2}, 0)$, $-\nabla f_0(\boldsymbol{x}^\star)$ belongs to the normal cone to the feasible set at point $\boldsymbol{x}^\star$, hence, there is $\boldsymbol{\lambda}^\star \geq \mathbf{0}$ that satisfies

$$\nabla f_0(\boldsymbol{x}^\star) + \lambda_1^\star \nabla f_1(\boldsymbol{x}^\star) + \lambda_2^\star \nabla f_2(\boldsymbol{x}^\star) = \mathbf{0}.$$

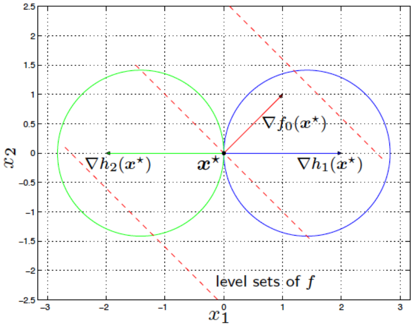
Example 4

Consider the problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \quad & f_0(\boldsymbol{x}) = x_1 + x_2 \\ \text{subject to} \quad & h_1(\boldsymbol{x}) = (x_1+1)^2 + x_2^2 - 2 = 0. \\ & h_2(\boldsymbol{x}) = (x_1-1)^2 + x_2^2 - 2 = 0. \end{aligned}$$

This is a problem with a linear objective function $f_0(\boldsymbol{x})$ and two nonlinear equality constraints $h_1(\boldsymbol{x}) = 0$, $h_2(\boldsymbol{x}) = 0$. At the solution $\boldsymbol{x}^\star$ (which is in fact the only feasible point), no linear combination of the gradients of the two constraints is equal to $\nabla f_0(\boldsymbol{x}^\star)$, *i.e*, there is no $\boldsymbol{\nu}^\star$ such that

$$\nabla f_0(\boldsymbol{x}^\star) + \nu_1^\star \nabla h_1(\boldsymbol{x}^\star) + \nu_2^\star \nabla h_2(\boldsymbol{x}^\star) = \mathbf{0}.$$



Summary

- From these examples, we know that the first-order necessary condition we just mentioned (by using Lagrangian multipliers to characterize) may or may not hold true

- That is the reason why, in the previous statement, we have to make the following critical assumption [regularity condition]:

  The vectors of $\nabla h_i(x^*)$'s are all linearly independent.

Equality Constrained Problem

Lagrange Multiplier Theorem (necessary condition)

- Let x^* be a local min and a regular point[the vectors of $\nabla h_i(x^*)$'s are all linearly independent]. Then there exist unique scalars $\lambda_1^*, ..., \lambda_m^*$ such that

$$\nabla f(x^*) = -\sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*).$$

- **Interpretation I.** A a local optimal solution, the gradient of the function can be expressed to a linear combination of the gradients of the constraints

Question: Is this always possible to finds these $\{\lambda_i^*\}$?

Equality Constrained Problem

Lagrange Multiplier Theorem (necessary condition)

- Let x^* be a local min and a regular point[the vectors of $\nabla h_i(x^*)$'s are all linearly independent]. Then there exist unique scalars $\lambda_1^*, ..., \lambda_m^*$ such that

$$\nabla f(x^*) = -\sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*).$$

- **Interpretation I.** A a local optimal solution, the gradient of the function can be expressed to a linear combination of the gradients of the constraints

- **Interpretation II.** The cost gradient $\nabla f(x^*)$ is orthogonal to the subspace of first order feasible variations

$$V(x^*) = \{\Delta x \mid \nabla h_i(x^*)' \Delta x = 0\} \tag{2.5}$$

Question: Is this always possible to finds these $\{\lambda_i^*\}$?

Sufficiency Condition

- Second Order Suffciency Conditions: Let $x^* \in R^n$ and $\lambda \in R^m$ satisfy

$$\nabla_x L(x^*, \lambda^*) = 0, \ \nabla_\lambda L(x^*, \lambda^*) = 0,$$
$$y' \nabla^2_{xx} L(x^*, \lambda^*) y > 0, \ \forall y \neq 0 \text{ with } \nabla h(x^*)' y = 0.$$

Then x^* is a strict local minimum.

Sufficiency Condition

- Example:

$$\begin{array}{ll} \text{minimize} & -(x_1x_2 + x_2x_3 + x_1x_3) \\ \text{subject to} & x_1 + x_2 + x_3 = 3. \end{array}$$

We have that $x_1^* = x_2^* = x_3^* = 1$ and $\lambda^* = 2$ satisfy the 1st order conditions. Also

$$\nabla^2_{xx} L(x^*, \lambda^*) = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}$$

We have for all $y \neq 0$ with $\nabla h(x^*)'y = 0$ or $y_1 + y_2 + y_3 = 0$,

$$y'\nabla^2_{xx} L(x^*, \lambda^*)y = -y_1(y_2+y_3)-y_2(y_1+y_3)-y_3(y_1+y_2) = y_1^2+y_2^2+y_3^2 > 0$$

Subgradient Algorithms

General Story

- Many optimization problems can be formulated by

$$\min_{\mathbf{x}} \quad \underbrace{\sum_{i=1}^{N} \ell_i(\mathbf{x}; \mathbf{a}_i, b_i)}_{\text{empirial loss on training}} \quad + \quad \underbrace{r(\mathbf{x})}_{\text{regularization}} \tag{3.1}$$

- "The finite-sum problem"
- Optimize the data one by one? Handle nonsmoothness?

- To deal with the nonsmooth function, need a new notion "subgradient"; To handle multiple terms, "stochastic" gradient method (SGD)

More on Convexity

- Recall that we learned that if a function $f(x)$ is *twice differentiable*, then it is convex over a set X iff

$$\nabla^2 f(x) \succeq 0, \; \forall \; x \in X$$

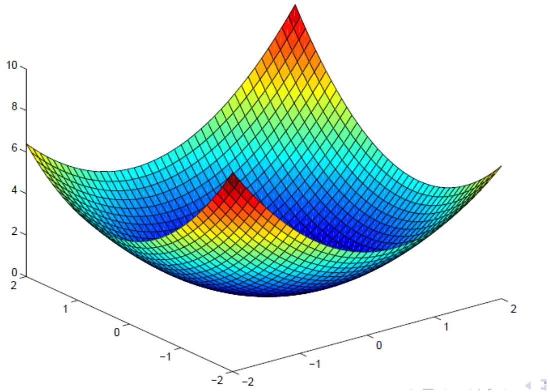


Figure 3.1: Illustration of a convex quadratic function.

More on Convexity

- What if a function is non-differentiable? Still convex?
- How does the function $\max\{0, 1-x\}$ look like? Convex? Why?
- **The subgradient**, $\partial f(x)$ of f at x is a set that

$$\partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x\rangle, \forall\ y\}$$

- **Claim**: IF $\partial f(x)$ has a single element at x, then $f(x)$ is differentiable at x

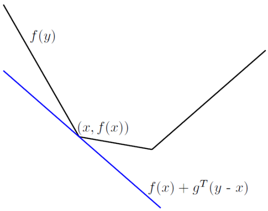


Figure 3.2: A function that is convex but non-differentiable.

Exercise

- What's the subgradient of the following functions (draw a figure)

$$|x|, \quad \max\{0, 1-x\}$$

Exercise

- What's the subgradient of the following functions (draw a figure)

$$|x|, \quad \max\{0, 1-x\}$$

- We have (sign(x) represents the "sign function")

$$\partial|x| = \text{sign(x)}, \text{ if } x \neq 0, \quad \partial|x| = [-1,\ 1], \text{ if } x = 0$$

Exercise

- What's the subgradient of the following functions (draw a figure)

$$|x|, \quad \max\{0, 1 - x\}$$

- We have (sign(x) represents the "sign function")

$$\partial|x| = \text{sign(x)},\ \text{if } x \neq 0, \quad \partial|x| = [-1,\ 1],\ \text{if } x = 0$$

- We have

$$\partial \max\{0, 1 - x\} = 0,\ \text{if } x > 1, \quad \partial \max\{0, 1 - x\} = -1,\ \text{if } x < 1$$

$$\partial \max\{0, 1 - x\} = [0,\ -1],\ \text{if } x = 1$$

Exercise

- What's the subgradient of the following functions (draw a figure)

$$|x|, \quad \max\{0, 1-x\}$$

- We have (sign(x) represents the "sign function")

$$\partial|x| = \text{sign(x)}, \text{ if } x \neq 0, \quad \partial|x| = [-1,\ 1], \text{ if } x = 0$$

- We have

$$\partial \max\{0, 1-x\} = 0, \text{ if } x > 1, \quad \partial \max\{0, 1-x\} = -1, \text{ if } x < 1$$
$$\partial \max\{0, 1-x\} = [0,\ -1], \text{ if } x = 1$$

- In particular, for the hinge loss $h(\mathbf{x}) = \max\{0, 1 - b_i \mathbf{a}_i^T \mathbf{x}\}$, the following is one subgradient

$$\partial h(\mathbf{x}) = -b_i \mathbf{a}, \text{ if } b_i(\mathbf{a}_i^T \mathbf{x}) < 1, \quad \partial h(\mathbf{x}) = 0, \text{ Otherwise}$$

Exercise

- What's the subgradient of the following functions (draw a figure)

$$|x|, \quad \max\{0, 1-x\}$$

- We have (sign(x) represents the "sign function")

$$\partial |x| = \text{sign(x)}, \text{ if } x \neq 0, \quad \partial |x| = [-1,\ 1], \text{ if } x = 0$$

- We have

$$\partial \max\{0, 1-x\} = 0, \text{ if } x > 1, \quad \partial \max\{0, 1-x\} = -1, \text{ if } x < 1$$
$$\partial \max\{0, 1-x\} = [0,\ -1], \text{ if } x = 1$$

- In particular, for the hinge loss $h(\mathbf{x}) = \max\{0, 1 - b_i \mathbf{a}_i^T \mathbf{x}\}$, the following is one subgradient

$$\partial h(\mathbf{x}) = -b_i \mathbf{a}, \text{ if } b_i(\mathbf{a}_i^T \mathbf{x}) < 1, \quad \partial h(\mathbf{x}) = 0, \text{ Otherwise}$$

- The subgradient of two functions
$\partial(h(\mathbf{x}) + g(\mathbf{x})) = \partial h(\mathbf{x}) + \partial g(\mathbf{x})$

Subgradient Method

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

- $f(x)$ is convex, but nonsmooth
- Subgradient Method

$$x^{r+1} = x^r - \alpha^r g^r$$

  where α^r is a stepsize, $g^r \in \partial f(x)$ is one of the subgradient
- **Note**: This is NOT a descent method!

Optimality

- Recall for f convex, differentiable,

$$f(x^\star) = \inf_x f(x) \Leftrightarrow 0 = \nabla f(x^\star)$$

- Generalization to nondifferentiable convex f:

$$f(x^\star) = \inf_x f(x) \Leftrightarrow 0 \in \partial f(x^\star)$$

proof. by definition (!)

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) \text{ for all } y \Leftrightarrow 0 \in \partial f(x^\star)$$

Subgradient Method: Convergence

- Main proof steps a bit different than gradient descent
- Assume that the subgradients are all bounded, $\|g^r\| \le G$
- We have

$$\begin{aligned}\|x^{r+1} - x^*\|^2 &= \|x^r - \alpha g^r - x^*\|^2 \\ &= \|x^r - x^*\|^2 - 2\alpha\langle g^r, x^r - x^*\rangle + \alpha^2\|g^r\|^2 \quad (3.2)\end{aligned}$$

Subgradient Method: Convergence

- Main proof steps a bit different than gradient descent
- Assume that the subgradients are all bounded, $\|g^r\| \le G$
- We have

$$\begin{aligned}\|x^{r+1} - x^*\|^2 &= \|x^r - \alpha g^r - x^*\|^2 \\ &= \|x^r - x^*\|^2 - 2\alpha\langle g^r, x^r - x^*\rangle + \alpha^2\|g^r\|^2 \quad (3.2)\end{aligned}$$

- Utilizing the definition of subgradient

$$f(x^*) \ge f(x^r) + \langle g^r, x^* - x^r\rangle = f(x^r) - \langle g^r, x^r - x^*\rangle \quad (3.3)$$

Subgradient Method: Convergence

- Main proof steps a bit different than gradient descent
- Assume that the subgradients are all bounded, $\|g^r\| \le G$
- We have

$$\begin{aligned}\|x^{r+1} - x^*\|^2 &= \|x^r - \alpha g^r - x^*\|^2 \\ &= \|x^r - x^*\|^2 - 2\alpha\langle g^r, x^r - x^*\rangle + \alpha^2\|g^r\|^2 \quad (3.2)\end{aligned}$$

- Utilizing the definition of subgradient

$$f(x^*) \ge f(x^r) + \langle g^r, x^* - x^r\rangle = f(x^r) - \langle g^r, x^r - x^*\rangle \quad (3.3)$$

- We obtain

$$\begin{aligned}&\|x^{r+1} - x^*\|^2 \le \|x^r - x^*\|^2 + 2\alpha[f(x^*) - f(x^r)] + \alpha^2\|g^r\|^2 \\ \Rightarrow &f(x^r) - f(x^*) \le \frac{\|x^r - x^*\|^2 - \|x^{r+1} - x^*\|^2}{2\alpha} + \frac{\alpha}{2}G^2\end{aligned}$$

Subgradient Method: Convergence

- Summing over $t = 1 \to T$, we have

$$\sum_{t=1}^{T} f(x^r) - f(x^*) \le \frac{1}{2\alpha} \sum_{r=1}^{T} \left(\|x^r - x^*\|^2 - \|x^{r+1} - x^*\|^2 \right) + \frac{T\alpha G^2}{2}$$

$$= \frac{1}{2\alpha} \|x^1 - x^*\|^2 - \frac{1}{2\alpha} \|x^{T+1} - x^*\|^2 + \frac{T\alpha}{2} G^2$$

- Now let $D = \|x^1 - x^*\|$, which is a constant, we have

$$f(x^{\text{best}}) - f(x^*) \le \frac{1}{T} \sum_{t=1}^{T} (f(x^r) - f(x^*)) \le \frac{1}{2\alpha T} D^2 + \frac{\alpha}{2} G^2$$

$f(x^{\text{best}})$ is the best objective function we have so far (up until T iterations)

Subgradient Method: Convergence

- Summing over $t = 1 \to T$, we have

$$\sum_{t=1}^{T} f(x^r) - f(x^*) \leq \frac{1}{2\alpha} \sum_{r=1}^{T} \left(\|x^r - x^*\|^2 - \|x^{r+1} - x^*\|^2 \right) + \frac{T\alpha G^2}{2}$$

$$= \frac{1}{2\alpha}\|x^1 - x^*\|^2 - \frac{1}{2\alpha}\|x^{T+1} - x^*\|^2 + \frac{T\alpha}{2} G^2$$

- Now let $D = \|x^1 - x^*\|$, which is a constant, we have

$$f(x^{\text{best}}) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} (f(x^r) - f(x^*)) \leq \frac{1}{2\alpha T} D^2 + \frac{\alpha}{2} G^2$$

  $f(x^{\text{best}})$ is the best objective function we have so far (up until T iterations)

- Set the stepsize $\alpha = \frac{1}{\sqrt{T}}$, we have

$$f(x^{\text{best}}) - f(x^*) \leq \frac{1}{2\sqrt{T}} \left(D^2 + G^2 \right)$$

Subgradient Method: Convergence

- Differences with the gradient descent
  1. Stepsize $\alpha = \frac{1}{\sqrt{T}}$, have to know how many iteration you are planning to run
  2. The rate is worse, $\mathcal{O}(1/\sqrt{T})$ (compared with gradient descent, $\mathcal{O}(1/T)$), for convex problems
  3. But can deal with nonsmooth function

Subgradient Method: Convergence

- Comparison of rates (running T iterations of all algorithms)

$$\begin{aligned} &\text{second order differentiable, strongly convex} \Rrightarrow && \approx \beta^T \\ &\text{differentiable, convex} \Rrightarrow && \approx (1/T) \\ &\text{nonsmooth, convex} \Rrightarrow && \approx (1/\sqrt{T}) \end{aligned}$$

- **Comment**: If the diminishing rule is chosen, then convergence is guaranteed, α^r is about $1/r$, or $1/\sqrt{r}$

SGD Outline

- Many machine learning problems minimizing empirical loss

$$\min f(\mathbf{x}) := \sum_{i=1}^{N} L_i(\mathbf{x}; \mathbf{a}_i, b_i)$$

- Computing the gradient

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{N} \nabla L_i(\mathbf{x}; \mathbf{a}_i, b_i) := \sum_{i=1}^{N} \nabla g_i(\mathbf{x})$$

- Per iteration effort needed for computing the gradient is $\mathcal{O}(N)$
- Overall complexity for reaching ϵ-OPT: $\mathcal{O}(\frac{N}{\epsilon})$
- When the number of data points is huge, too bad

SGD Algorithm

- **Motivation 1**. Reduce the dependence on the # of data points
- **Motivation 2**. Start building models quickly, without seeing all data point once
- **The SGD Method:** $r = 1, \cdots, T$
  1. Randomly pick a data point $i \in \{1, \cdots, N\}$
  2. Compute (sub)gradient $g^r \in \partial L_i(\mathbf{x}; \mathbf{a}_i, b_i)$
  3. $\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha^r g^r$
- Many important variants, will go through the analysis later

Some basics of graph theory

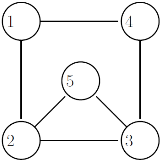
Basics of Graph Theory

- In this subsection, we present some basic notions in graph theory; these results will be used later for modeling interactions among the distributed users and agents

- Detailed results can be found in the following monograph: "F. R. K. Chung, Spectral Graph Theory"

Basic Notations

- Define an undirected and unweighted graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with $|\mathcal{V}| = M$ vertices and $|\mathcal{E}| = E$ edges
- Each agent can only communicate with its immediate neighbors
- Let $A \in \mathbb{R}^{E \times M}$ be the edge-node incidence matrix; $x := [x_1; \cdots; x_M] \in \mathbb{R}^M$
- If $e \in \mathcal{E}$ and it connects vertex i and j with $i > j$, then $A_{ev} = 1$ if $v = i$, $A_{ev} = -1$ if $v = j$ and $A_{ev} = 0$ otherwise.



$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}.$$

Basic Notations

- Let d_i be the degree of node i, i.e., the number of nodes it is connected with; $d_i = 0$ means node i is not connected with anyone
- Graph Laplacian

$$\mathcal{L} := P^{-1/2} A^T A P^{-1/2}, \text{ with } P = \text{diag}[d_1, \cdots, d_M]$$

$$[\mathcal{L}]_{ij} = \begin{cases} 1 & \text{if } i = j \\ -\frac{1}{\sqrt{d_i d_j}} & \text{if } (ij) \in \mathcal{E}, i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

- For notation simplicity, sometimes we use $i \sim j$ to denote node i and j are connected

Basic Notations

- Sometimes, you will see unnormalized graph Laplacian

$$L := A^T A,$$

$$[L]_{ij} = \begin{cases} d_i & \text{if } i = j \\ -1 & \text{if } (ij) \in \mathcal{E}, i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

- We have $L = P^{1/2} \mathcal{L} P^{1/2}$

Properties of Graph Laplacian

- Let y be a vector of size $|\mathcal{V}|$
- Then we have, the ith entry of $\mathcal{L}v$ is:

$$[\mathcal{L}v]_i = \frac{1}{\sqrt{d_i}} \sum_{j:j\sim i} \left(\frac{v(i)}{\sqrt{d_i}} - \frac{v(j)}{\sqrt{d_j}} \right) \tag{4.1}$$

$$[Lv]_i = \sum_{j:j\sim i} (v(i) - v(j)) \tag{4.2}$$

$$v^T L v = \sum_{i\sim j} (v(i) - v(j))^2 \geq 0. \tag{4.3}$$

Properties of Graph Laplacian

- Let $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots, \leq \lambda_M$ be the eigenvalues of $\mathcal{L}$
- Let $\underline{1}$ denote the constant vector with all entries 1
- We have the following properties
  - $P^{1/2}\underline{1}$ is an eigenfunction of $\mathcal{L}$ with eigenvalue 0
  - We have

$$\lambda_1 = \inf_{f \perp P^{1/2}\underline{1}} \frac{\sum_{u \sim v}(f(u) - f(v))^2}{\sum_v f(v)^2 d_v} \tag{4.4}$$

- Denote $\underline{\lambda}_{\min}(\mathcal{L}) := \lambda_1$, the smallest nonzero eigenvalue; $\lambda_{\max}(\mathcal{L}) = \lambda_M$
- The eigengap ($\underline{\lambda}_{\min}$: smallest non-zero eigenvalue)

$$\xi(\mathcal{L}) = \underline{\lambda}_{\min}(\mathcal{L}) / \lambda_{\max}(\mathcal{L}) \leq 1.$$

Examples

- For the complete graph on M vertices, the eigenvalues are 0 and $M/(M-1)$ (with multiplicity $M-1$)
- For the star graph on M vertices, the eigenvalues are $0, 1$ (with multiplicity $M-2$) and 2.
- For the path graph on M vertices, the eigenvalues are $1-\cos(\pi k/(M-1))$, for $k = 0, 1, \cdots, M-1$
- For the cycle graph on M vertices, the eigenvalues are $1-\cos(2\pi k/M)$, for $k = 0, 1, \cdots, M-1$
- For the M - cube graph on 2^n vertices, the eigenvalues are $2k/M$, for $k = 0, 1, \cdots, M$
- For the

Examples

In terms of the eigengaps ξ

1) **Complete Graph:** $\xi(\mathcal{L}) = 1$;

2) **Star Graph:** $\xi(\mathcal{L}) = 1/2$;

3) **Path Graph:** $\xi(\mathcal{L}) = \mathcal{O}(1/M^2)$.

4) **Grid Graph:** $\xi(\mathcal{L}) = \mathcal{O}(1/M)$.

5) **Random Geometric Graph:** Place the nodes uniformly in $[0,1]^2$ and connect any two nodes separated by a distance less than a radius $R \in (0,1)$. With high probability $\xi(\mathcal{L}) = \mathcal{O}\left(\frac{\log(M)}{M}\right)$.

For more discussion, see [Lemma 5, Duchi-12]

J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," 2012

Some other generic facts

- $\sum_i \lambda_i \leq M$
- $\lambda_i \leq 2$, for all $i \leq M$
- $\lambda_1 \geq \frac{1}{D \times \mathsf{vol}(G)}$, where D is the diameter of the graph, and

$$\mathsf{vol}(G) = \sum_{i=1}^{M} d_i.$$

Using the above result, it is easy to derive some of the eigengap results presented in the previous slide.