Optimization Methods (CS1.404), Spring 2024
Lecture 25

Naresh Manwani

Machine Learning Lab, IIIT-H

April 22nd, 2024

The Gradient Projection Method

- The stationarity condition $\mathbf{x}^* = P_C(\mathbf{x}^* - s\nabla f(\mathbf{x}^*))$ motivates to solve the optimization problem

$$
\begin{aligned}
(P) \quad & \min \; f(\mathbf{x}) \\
& s.t. \; \mathbf{x} \in C
\end{aligned}
$$

Gradient Projection Method

Input: $\epsilon > 0$ (tolerance parameter)
Initialization: Pick $\mathbf{x}_0 \in C$ arbitrarily
General Steps: For $k = 0, 1, 2, \dots$ execute the following steps:

1. Pick a stepsize t_k by a line search procedure.
2. Set $\mathbf{x}_{k+1} = P_C(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))$.
3. If $\|\mathbf{x}_k - \mathbf{x}_{k+1}\| \leq \epsilon$, then STOP and $\mathbf{x}_{k+1}$ is the output.

Sufficient Decrease Lemma for Projected Gradient Descent

Lemma

Suppose that $f \in \mathbb{C}^{1,1}(C)$, where C is a closed convex set. Then for any $\mathbf{x} \in C$ and $t \in (0, 2/L)$, the following inequality will hold.

$$f(\mathbf{x}) - f(P_C(\mathbf{x} - t\nabla f(\mathbf{x}))) \geq t\left(1 - \frac{Lt}{2}\right)\|\frac{1}{t}(\mathbf{x} - P_C(\mathbf{x} - t\nabla f(\mathbf{x})))\|^2.$$

- When $C = \mathbb{R}^n$, the obtained inequality is exactly the same as the one obtained for unconstrained case.

Gradient Mapping

- We define gradient mapping as

$$G_M(\mathbf{x}) = M\left[\mathbf{x} - P_C\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)\right]$$

  where $M > 0$.
- When $C = \mathbb{R}^n$, we see that $G_M(\mathbf{x}) = \nabla f(\mathbf{x})$. So, the gradient mapping is an extension of usual gradient operation.
- We see that $G_M(\mathbf{x}) = \mathbf{0}$ if and only $\mathbf{x} = P_C\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)$. This happens if and only if $\mathbf{x}$ is a stationary point of optimization problem (P).
- Thus, we can use $\|G_M(\mathbf{x})\|$ as an optimality measure.
- Thus, the sufficient decrease property of projected gradient method essentially states that

$$f(\mathbf{x}) - f(P_C(\mathbf{x} - t\nabla f(\mathbf{x}))) \geq t\left(1 - \frac{Lt}{2}\right)\|G_{\frac{1}{t}}(\mathbf{x})\|^2.$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Sufficient decrease of consecutive function values

Lemma

Consider the optimization problem

$$(P): \quad \min f(\mathbf{x}) \quad s.t.\ \mathbf{x} \in C$$

where C is a closed convex set and $f \in \mathbb{C}_L^{1,1}$. Let $\mathbf{x}_k,\ k = 0, 1, 2, \ldots$ be the sequence generated by the project gradient algorithm for solving (P) using either constant stepsize $t_k = \tilde{t} \in (0, \frac{2}{L})$ or by a stepsize chosen using backtracking procedure with parameters (s, α, β) satisfying $s > 0, \alpha \in (0,1), \beta(0,1)$. Then for $k \geq 0$

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq M\|G_d(\mathbf{x}_k)\|^2$$

where

$$M = \begin{cases} \tilde{t}(1 - \frac{L\tilde{t}}{2}), & \text{for constant step size} \\ \min\left(s, \frac{2\beta(1-\alpha)}{L}\right), & \text{for Backtracking} \end{cases}$$

and

$$d = \begin{cases} \frac{1}{\tilde{t}}, & \text{for constant step size} \\ \frac{1}{s}, & \text{for Backtracking.} \end{cases}$$

Convergence of the Gradient Projection Method

Theorem

Consider the optimization problem

$$(P): \quad \min f(\mathbf{x}) \\ s.t.\ \mathbf{x} \in C$$

where C is a closed convex set and $f \in \mathbb{C}_L^{1,1}$ is bounded below. Let $\mathbf{x}_k,\ k = 0, 1, 2, \ldots$ be the sequence generated by the project gradient algorithm for solving (P) using either constant stepsize $t_k = \tilde{t} \in (0, \frac{2}{L})$ or by a stepsize chosen using backtracking procedure with parameters (s, α, β) satisfying $s > 0, \alpha \in (0, 1), \beta(0, 1)$. Then we have the following:

- The sequence $\{f(\mathbf{x}_k)\}_{k=0}^{\infty}$ is nonincreasing. In addition, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ unless $\mathbf{x}_k$ is a stationary point of (P).
- $G_d(\mathbf{x}_k) \to \mathbf{0}$ as $k \to \infty$. where

$$d = \begin{cases} \frac{1}{\tilde{t}}, & \text{for constant step size} \\ \frac{1}{s}, & \text{for Backtracking.} \end{cases}$$

Equality Constrained Minimization problem

- Consider following convex optimization problem with equality constraints,

$$\min \ f(\mathbf{x}) \quad s.t. \ \ A\mathbf{x} = \mathbf{b},$$

  where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and twice continuously differentiable, and $A \in \mathbb{R}^{p \times n}$ with $rank(A) = p < n$.
- The assumptions on A mean that there are fewer equality constraints than variables, and that the equality constraints are independent.
- We will assume that an optimal solution $\mathbf{x}^*$ exists and let $p^* = \inf\{f(\mathbf{x}) \mid A\mathbf{x} = \mathbf{b}\} = f(\mathbf{x}^*)$.
- A point $\mathbf{x}^* \in dom(f)$ is optimal for if and only if there is a $\boldsymbol{\mu}^* \in \mathbb{R}^p$ such that

$$A\mathbf{x}^* = \mathbf{b}$$
$$\nabla f(\mathbf{x}^*) + A^\top \boldsymbol{\mu}^* = \mathbf{0}$$

- Above are KKT optimality conditions. There are $n + p$ equations in the $n + p$ variables $\mathbf{x}^*, \boldsymbol{\mu}^*$.

Equality Constrained Convex Quadratic Minimization

- Consider the equality constrained convex quadratic minimization problem

$$\begin{aligned} &\min\ f(\mathbf{x}) = (1/2)\mathbf{x}^\top P\mathbf{x} + \mathbf{q}^\top \mathbf{x} + r \\ &s.t.\ A\mathbf{x} = \mathbf{b}, \end{aligned}$$

  where $P \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite matrix, $A \in \mathbb{R}^{p\times n}$, $\mathbf{q} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^p$.
- $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ is a KKT point of this problem if it satisfies the following conditions.

$$\begin{aligned} &A\mathbf{x}^* = \mathbf{b} \\ &P\mathbf{x}^* + \mathbf{q} + A^\top \boldsymbol{\mu}^* = \mathbf{0} \end{aligned}$$

- We can rewrite the above KKT conditions is the matrix form as follows.

$$\begin{bmatrix} P & A^\top \\ A & \mathbf{0}_{p\times p} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \boldsymbol{\mu}^* \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ \mathbf{b} \end{bmatrix}$$

- There are $n+p$ linear equations in the $n+p$ variables $\mathbf{x}^*, \boldsymbol{\mu}^*$. The coefficient matrix is called the KKT matrix.

Equality Constrained Convex Quadratic Minimization

- When the KKT matrix is nonsingular, there is a unique optimal primal-dual pair $(\mathbf{x}^*, \boldsymbol{\mu}^*)$.
- If the KKT matrix is singular, but the KKT system is solvable, any solution yields an optimal pair $(\mathbf{x}^*, \boldsymbol{\mu}^*)$.
- If the KKT system is not solvable, the quadratic optimization problem is unbounded below or infeasible.
  - In this case, there exist $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^p$ such that $P\mathbf{v} + A^\top \mathbf{w} = \mathbf{0}$, $A\mathbf{v} = \mathbf{0}$, $-\mathbf{q}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{w} > 0$.
  - Let $\hat{\mathbf{x}}$ be any feasible point. The point $\mathbf{x} = \hat{\mathbf{x}} + t\mathbf{v}$ is feasible for all t and

$$\begin{aligned} f(\hat{\mathbf{x}} + t\mathbf{v}) &= f(\hat{\mathbf{x}}) + t(\mathbf{v}^\top P\hat{\mathbf{x}} + \mathbf{q}^\top \mathbf{v}) + (1/2)t^2\mathbf{v}^\top P\mathbf{v} \\ &= f(\hat{\mathbf{x}}) + t(-\hat{\mathbf{x}}^\top A^\top \mathbf{w} + \mathbf{q}^\top \mathbf{v}) - (1/2)t^2\mathbf{w}^\top A\mathbf{v} \\ &= f(\hat{\mathbf{x}}) + t(-\mathbf{b}^\top \mathbf{w} + \mathbf{q}^\top \mathbf{v}) \end{aligned}$$

  which decreases without bound as $t \to \infty$.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Non-Singularity of KKT Matrix

Claim

KKT matrix is nonsingular if P is positive definite on the null space of A. Equivalently, $A\mathbf{x} = \mathbf{0}, \mathbf{x} \neq \mathbf{0} \Rightarrow \mathbf{x}^\top P\mathbf{x} > 0$.

Proof:

- Assume that $A\mathbf{x} = \mathbf{0}, \mathbf{x} \neq \mathbf{0} \Rightarrow \mathbf{x}^\top P\mathbf{x} > 0$.
- Let $(\mathbf{x}, \boldsymbol{\mu})$ be a vector in the null space of the KKT matrix.
- Thus, $\begin{bmatrix} P & A^\top \\ A & \mathbf{0}_{p\times p} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix} = \mathbf{0}$.
- This means, $P\mathbf{x} + A^\top \boldsymbol{\mu} = \mathbf{0}$ and $A\mathbf{x} = \mathbf{0}$.
- Thus, $\mathbf{x}^\top (P\mathbf{x} + A^\top \boldsymbol{\mu}) = 0$, which implies $\mathbf{x}^T P\mathbf{x} + (A\mathbf{x})^\top \boldsymbol{\mu} = 0$. But using $A\mathbf{x} = \mathbf{0}$, we get $\mathbf{x}^T P\mathbf{x} = 0$. But, as per assumption P is positive definite on the null space of A, we get $\mathbf{x} = 0$.
- Now using the equation $P\mathbf{x} + A^\top \boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{x} = \mathbf{0}$, we get $A^\top \boldsymbol{\mu} = \mathbf{0}$. But, this gives $\boldsymbol{\mu} = \mathbf{0}$ as A is full rank matrix.
- Thus, the only null space of the KKT matrix is vector $(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{0}$. Thus, KKT matrix is nonsingular.

Example: Equality Constrained Convex Quadratic Minimization

- Consider the following optimization problem

$$\min\ (x_1 - 2)^2 + 2(x_2 - 1)^2 - 5$$
$$s.t.\ x_1 + 4x_2 = 3$$

- KKT Conditions give

$$\begin{aligned} 2x_1 + \mu &= 4 \\ 4x_2 + 4\mu &= 4 \\ x_1 + 4x_2 &= 3 \end{aligned}$$

which is equivalent to $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 4 & 4 \\ 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \mu \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix}$.

- The solution of these KKT condition is $x_1^* = \frac{5}{3},\ x_2^* = \frac{1}{3},\ \mu^* = \frac{2}{3}$.

Eliminating Equality Constraints

- One general approach to solving the equality constrained problem is to eliminate the equality constraints, and then solve the resulting unconstrained problem using methods for unconstrained minimization.
- We first find a matrix $F \in \mathbb{R}^{n\times(n-p)}$ and vector $\hat{\mathbf{x}} \in \mathbb{R}^n$ that parametrize the (affine) feasible set:

$$\{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}\} = \hat{\mathbf{x}} + \{\mathbf{x} \mid A\mathbf{x} = \mathbf{0}\} = \{F\mathbf{z} + \hat{\mathbf{x}} \mid \mathbf{z} \in R^{n-p}\}.$$

- Here,
  - $\hat{\mathbf{x}}$ is any feasible point. Thus, $A\hat{\mathbf{x}} = \mathbf{b}$.
  - $F \in \mathbb{R}^{n\times(n-p)}$ is composed of $(n-p)$ basis vectors of null space of A. Thus, $AF = \mathbf{0}_{p\times(n-p)}$.
- Thus, the reduced unconstrained problem is

$$\min_{\mathbf{z}} \; f_{new}(\mathbf{z}) = \min_{\mathbf{z}} \; f(\hat{\mathbf{x}} + F\mathbf{z})$$

Eliminating Equality Constraints

$\mathbf{x}^* = \hat{\mathbf{x}} + F\mathbf{z}^*$ fulfils the KKT conditions with $\boldsymbol{\mu}^* = -(AA^\top)^{-1}A\nabla f(\mathbf{x}^*)$.

Proof:

- Clearly, $\mathbf{x}^* = \hat{\mathbf{x}} + F\mathbf{z}^*$ is primal feasible as $A\mathbf{x}^* = A\hat{\mathbf{x}} + AF\mathbf{z}^* = \mathbf{b} + \mathbf{0} = \mathbf{b}$.
- Consider the matrix $\begin{bmatrix} F^\top \\ A \end{bmatrix}$. For such a matrix,

$$\begin{bmatrix} F^\top \\ A \end{bmatrix}(\nabla f(\mathbf{x}^*) + A^\top \boldsymbol{\mu}^*) = \begin{bmatrix} F^\top \nabla f(\mathbf{x}^*) + F^\top A^\top \boldsymbol{\mu}^* \\ A\nabla f(\mathbf{x}^*) + AA^\top \boldsymbol{\mu}^* \end{bmatrix}$$
$$= \begin{bmatrix} F^\top \nabla f(\mathbf{x}^*) - F^\top A^\top (AA^\top)^{-1} A\nabla f(\mathbf{x}^*) \\ A\nabla f(\mathbf{x}^*) - AA^\top (AA^\top)^{-1} A\nabla f(\mathbf{x}^*) \end{bmatrix}$$
$$= \begin{bmatrix} \nabla f_{new}(\mathbf{z}^*) - (AF)^\top (AA^\top)^{-1} A\nabla f(\mathbf{x}^*) \\ A\nabla f(\mathbf{x}^*) - A\nabla f(\mathbf{x}^*) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

- Since, matrix $\begin{bmatrix} F^\top \\ A \end{bmatrix}$ has full rank, $\begin{bmatrix} F^\top \\ A \end{bmatrix}(\nabla f(\mathbf{x}^*) + A^\top \boldsymbol{\mu}^*) = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ implies $\nabla f(\mathbf{x}^*) + A^\top \boldsymbol{\mu}^* = \mathbf{0}$. Thus, KKT conditions are satisfied by $\mathbf{x}^* = \hat{\mathbf{x}} + F\mathbf{z}^*$ and $\boldsymbol{\mu}^* = -(AA^\top)^{-1}A\nabla f(\mathbf{x}^*)$.

Newton Method for Equality Constraint Problem

- Here we describe an extension of Newton's method to include linear equality constraint.
- The methods are almost the same except for two differences:
  - the initial point must be feasible $A\mathbf{x} = \mathbf{b}$,
  - the Newton step must be a feasible direction $A\Delta\mathbf{x}_{nt} = \mathbf{0}$
- $\Delta\mathbf{x}_{nt}$ solves the second order approximation of f at $\mathbf{x}$ with variable $\mathbf{v}$.
- Let $\hat{f}(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{v} + \frac{1}{2}\mathbf{v}^\top \nabla^2 f(\mathbf{x})\mathbf{v}$. Then, $\Delta\mathbf{x}_{nt}$ is the minimizer of the following problem.

$$\begin{aligned} \min \ & f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{v} + \frac{1}{2}\mathbf{v}^\top \nabla^2 f(\mathbf{x})\mathbf{v} \\ s.t. \ & A(\mathbf{x}+\mathbf{v}) = \mathbf{b} \end{aligned}$$

which is same as

$$\begin{aligned} \min \ & f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{v} + \frac{1}{2}\mathbf{v}^\top \nabla^2 f(\mathbf{x})\mathbf{v} \\ s.t. \ & A\mathbf{v} = \mathbf{0} \end{aligned}$$

Newton Method for Equality Constraint Problem

- $\Delta \mathbf{x}_{nt}$ satisfies KKT Optimality conditions for this optimization problem, which are:

$$\begin{aligned} &\nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x}) \Delta \mathbf{x}_{nt} + A^\top \boldsymbol{\mu} = \mathbf{0} \\ &A \Delta \mathbf{x}_{nt} = \mathbf{0} \end{aligned}$$

  where $\boldsymbol{\mu}$ is the associated dual variable for the quadratic problem.
- Thus, Newton step $\Delta \mathbf{x}_{nt}$ at $\mathbf{x}$ is the solution of following linear system

$$\begin{bmatrix} \nabla^2 f(\mathbf{x}) & A^\top \\ A & \mathbf{0}_{p \times p} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x}_{nt} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} -\nabla f(\mathbf{x}) \\ \mathbf{0} \end{bmatrix}$$

Newton Method for Equality Constraint Problem

- **Newton Decrement:** Newton decrement is defined as

$$\lambda(\mathbf{x}) = (\Delta\mathbf{x}_{nt}^\top \nabla f(\mathbf{x})^2 \Delta\mathbf{x}_{nt})^{\frac{1}{2}}$$

- Newton decrement is related to $f(\mathbf{x}) - \min_{\mathbf{y}:A\mathbf{y}=\mathbf{b}} \hat{f}(\mathbf{y})$ as:

$$f(\mathbf{x}) - \min_{\mathbf{y}:A\mathbf{y}=\mathbf{b}} \hat{f}(\mathbf{y}) = \frac{1}{2}\lambda(\mathbf{x})^2$$

  where $\hat{f}$ is second order approximation of f at $\mathbf{x}$.
- This gives an estimate of $f(\mathbf{x}) - f^*$ using quadratic approximation.
- The Newton decrement comes up in the line search as well, since the directional derivative of f in the direction $\Delta\mathbf{x}_{nt}$ is

$$\frac{\partial}{\partial t} f(\mathbf{x} + t\Delta\mathbf{x}_{nt})|_{t=0} = \nabla f(\mathbf{x})^\top \Delta\mathbf{x}_{nt} = -\Delta\mathbf{x}_{nt}^\top \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{nt} = -\lambda(\mathbf{x})^2.$$

  Here, we used the Newton step property $\nabla^2 f(\mathbf{x})\Delta\mathbf{x}_{nt} = -\nabla f(\mathbf{x})$.

Newton Method for Equality Constraint Problem

- We say that $\mathbf{v} \in \mathbb{R}^n$ is a feasible direction for $A(\mathbf{x}+\mathbf{v}) = \mathbf{b}$ if $A\mathbf{v} = \mathbf{0}$.
- We say that $\mathbf{v}$ is a descent direction for f at $\mathbf{x}$, if for small $t > 0$, $f(\mathbf{x} + t\mathbf{v}) < f(\mathbf{x})$.
- The Newton step $\Delta\mathbf{x}_{nt}$ is always a feasible descent direction (except when $\mathbf{x}$ is optimal, in which case)$\Delta\mathbf{x}_{nt} = \mathbf{0}$).

Newton Method for Equality Constraint Problem

Newton Method for Equality Constraint Problem

Input: starting point $\mathbf{x} \in domain(f)$ with $A\mathbf{x} = \mathbf{b}$, tolerance $\epsilon > 0$.
while $\frac{1}{2}\lambda(\mathbf{x})^2 > \epsilon$ **do**
  Compute the Newton step $\Delta\mathbf{x}_{nt}$ and decrement $\lambda(\mathbf{x})$.
  Choose step size t by backtracking line search.
  Update $\mathbf{x} = \mathbf{x} + t\Delta\mathbf{x}_{nt}$.
end while

- The method is called a feasible descent method, since all the iterates are feasible, with $f(\mathbf{x}(k+1)) < f(\mathbf{x}(k))$ (unless $\mathbf{x}(k)$ is optimal).

Precise Algorithm: Newton Method for Equality Constraint Problem

Newton Method for Equality Constraint Problem

Input: starting point $\mathbf{x}_0 \in domain(f)$ with $A\mathbf{x}_0 = \mathbf{b}$, tolerance $\epsilon > 0$, Maximum number of iterations K.

for $k = 0, 1, \ldots, K$ **do**

  Find $\Delta\mathbf{x}_k$ and $\boldsymbol{\mu}_k$ as

$$\begin{bmatrix} \Delta\mathbf{x}_k \\ \boldsymbol{\mu}_k \end{bmatrix} = \begin{bmatrix} \nabla^2 f(\mathbf{x}_k) & A^\top \\ A & \mathbf{0}_{p\times p} \end{bmatrix}^{-1} \begin{bmatrix} -\nabla f(\mathbf{x}_k) \\ \mathbf{0} \end{bmatrix}$$

  if $\lambda(\mathbf{x}_k) \leq \epsilon$ **then**

    return $\mathbf{x}_k$

  end if

  Choose step size t by backtracking line search.

  Update $\mathbf{x}_{k+1} = \mathbf{x}_k + t\Delta\mathbf{x}_k$.

end for

Convergence of Feasible Newton Method for Equality Constraint Problem

- The iterates $\mathbf{x}(k)$ are the same as those of the Newton algorithm for the eliminated unconstrained problem

$$\hat{f}(\mathbf{z}) := f(\mathbf{x}_0 + F\mathbf{z}), \quad \mathbf{x}(k) = \mathbf{x}_0 + F\mathbf{z}(k)$$

  as they fulfil the KKT conditions of the quadratic approximation.
- Thus convergence is the same as in the unconstrained case.