Notice

- Homework assignments Week 13:
  - 1 programming work (6%).
  - Due date: 21 p.m. on December 25, 2021.

Outline

- **Lecture 6: Primal and dual problems**
  - Lagrangian dual problems
- **Lecture 7: Alternating Direction Method of Multipliers**
  - The basic form
  - Application I: Sparse group lasso estimation
  - Application II: Support vector machines
  - Application III: Splitting across examples
- **Lecture 8: More on convex analysis**
  - Subdifferentials and subgradients
- **Solutions to quizzes in Week 12 and homework assignments in Week 11**

Computation in Data Science: Week 13
Lecture 6

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Data Science Degree Program

National Taiwan University

December 15, 2021

Dual Problems

- **Constrained optimization problems:**
  - Consider the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & l(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{C}, \end{aligned} \tag{1}$$

    where $l : \mathbb{R}^p \mapsto \mathbb{R}$ is the objective function of the problem (1), and $\mathcal{C}$ is defined as

$$\begin{aligned} \mathcal{C} \quad = \quad & \Big\{ \mathbf{x} : \mathbf{x} \in \mathbb{R}^p, f_i(\mathbf{x}) \leq 0 \text{ for } i = 1, 2, \cdots, m, \\ & \text{and } h_j(\mathbf{x}) = 0 \text{ for } j = 1, 2, \cdots, n \Big\}. \end{aligned} \tag{2}$$

  - $\mathcal{C}$ is assumed to be nonempty.

Dual Problems

- **The basic idea:**
  - Introduce a set of variables $\mathbf{u} = (u_1, u_2, \cdots, u_m)$ and $\mathbf{v} = (v_1, v_2, \cdots, v_n)$ to formulate the following function:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = l(\mathbf{x}) + \sum_{i=1}^{m} u_i f_i(\mathbf{x}) + \sum_{j=1}^{n} v_j h_j(\mathbf{x}). \tag{3}$$

Here (3) is called the **Lagrangian** corresponding to the primal problem (1), and the quantities u_i's and v_j's are called the **Lagrange multipliers**.

With (3) we formulate another optimization problem as

$$\begin{array}{ll} \max\limits_{\mathbf{u},\mathbf{v}} & \left\{ \min\limits_{\mathbf{x}} \left[l(\mathbf{x}) + \sum_{i=1}^{m} u_i f_i(\mathbf{x}) + \sum_{j=1}^{n} v_j h_j(\mathbf{x}) \right] \right\} \\ \text{subject to} & u_i \geq 0 \text{ for } i = 1, 2, \cdots, m. \end{array} \tag{4}$$

The optimization problem (4) is called the **Lagrangian dual of** (1).

Dual Problems

- **Duality gap:**
  - Define
$$L(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x}} \left[l(\mathbf{x}) + \sum_{i=1}^{m} u_i f_i(\mathbf{x}) + \sum_{j=1}^{n} v_j h_j(\mathbf{x}) \right].$$
  - Assume $\mathbf{x}^{**}$ is a solution to the **primal problem** (1) and $(\mathbf{u}^{**}, \mathbf{v}^{**})$ is a solution to the Lagrangian dual problem (4). We have
$$\text{opt}^{\text{dual}} = L(\mathbf{u}^{**}, \mathbf{v}^{**}) \leq L(\mathbf{x}^{**}, \mathbf{u}^{**}, \mathbf{v}^{**}) \leq l(\mathbf{x}^{**}) = \text{opt}^{\text{primal}},$$
which implies the **duality gap** $\text{opt}^{\text{primal}} - \text{opt}^{\text{dual}} \geq 0$.
  - *Remark:* If $\text{opt}^{\text{primal}} = \text{opt}^{\text{dual}}$, then we say **strong duality** holds for the primal problem (1) and dual problem (4).

Dual Problems

- **The Karush-Kuhn-Tucker conditions:**
  - Now assume $l(\mathbf{x})$, $f_i(\mathbf{x})$'s and $h_j(\mathbf{x})$'s are differentiable. Consider the following conditions for point $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*)$:

$$
\begin{aligned}
& f_i(\mathbf{x}^*) \leq 0 \text{ for } i = 1, 2, \cdots, m \textbf{ (Primal feasibility)}, \\
& h_j(\mathbf{x}^*) = 0 \text{ for } j = 1, 2, \cdots, n \textbf{ (Primal feasibility)}, \\
& u_i^* \geq 0 \text{ for } i = 1, 2, \cdots, m \textbf{ (Dual feasibility)}, \\
& u_i^* f_i(\mathbf{x}^*) = 0 \text{ for } i = 1, 2, \cdots, m \textbf{ (Complementary slackness)}, \\
& \nabla l(\mathbf{x}^*) + \sum_{i=1}^{m} u_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^{n} v_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0} \textbf{ (Stationarity)}.
\end{aligned}
$$

  These conditions are called the **Karush-Kuhn-Tucker conditions** or **KKT conditions**.

Dual Problems

- **Implications of the Karush-Kuhn-Tucker conditions:**
  - The KKT conditions are **necessary conditions** for checking whether a point is a solution to an optimization problem.
  - The KKT conditions further turn out to be **sufficient conditions** when

    i). l, f_i's and h_j's are **differentiable**;

    ii). l is **convex**;

    iii). $\mathcal{C}$ is a **convex set**;

    iv). h_j's are **affine maps on x**.

    If i), ii), iii) and iv) are satisfied and the point $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*)$ satisfies the KKT conditions, then $\mathbf{x}^*$ is a solution to the primal problem (1), and $(\mathbf{u}^*, \mathbf{v}^*)$ is a solution to the dual problem (4).

    In this situation, $\mathbf{x}^*$ is called the **primal optimal point** and $(\mathbf{u}^*, \mathbf{v}^*)$ is called the **dual optimal point**.

Dual Problems

- **An example of the Lagrangian dual problem:**
  - Now consider the following optimization problem:

$$\begin{array}{rl} \text{minimize} & l(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{C}\mathbf{x} = \mathbf{d}. \end{array}$$

  - By definition, the dual objective is

$$\begin{aligned} L(\mathbf{u},\mathbf{v}) &= \min_{\mathbf{x}} \{ l(\mathbf{x}) + \mathbf{u}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) + \mathbf{v}^T(\mathbf{C}\mathbf{x} - \mathbf{d}) \} \\ &= -\mathbf{b}^T\mathbf{u} - \mathbf{d}^T\mathbf{v} + \min_{\mathbf{x}} \{ l(\mathbf{x}) + \mathbf{x}^T(\mathbf{A}^T\mathbf{u} + \mathbf{C}^T\mathbf{v}) \} \\ &= -\mathbf{b}^T\mathbf{u} - \mathbf{d}^T\mathbf{v} - \max_{\mathbf{x}} \{ \mathbf{x}^T(-\mathbf{A}^T\mathbf{u} - \mathbf{C}^T\mathbf{v}) - l(\mathbf{x}) \} \\ &= -\mathbf{b}^T\mathbf{u} - \mathbf{d}^T\mathbf{v} - l^*(-\mathbf{A}^T\mathbf{u} - \mathbf{C}^T\mathbf{v}). \end{aligned}$$

  - Therefore the Lagrangian dual problem is

$$\begin{array}{rl} \text{maximize} & -\mathbf{b}^T\mathbf{u} - \mathbf{d}^T\mathbf{v} - l^*(-\mathbf{A}^T\mathbf{u} - \mathbf{C}^T\mathbf{v}) \\ \text{subject to} & \mathbf{u} \ge \mathbf{0}. \end{array}$$

Dual Problems

- **Why derives the Lagrangian dual problem?**
  - The Lagrangian dual problem provides a **lower bound** for the primal problem.
  - When strong duality holds, solving the Lagrangian dual problem is equivalent to solving the primal problem.
  - In some cases it may be easier to solve the dual problem than to solve the primal problem.
- **Why derives the KKT conditions?**
  - The KKT conditions provide a guide for how the possible **primal** and **dual** optimal points should behave, which is key to formulate **stopping criteria** for relevant optimization algorithms.

Computation in Data Science: Week 13 Lecture 7

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Data Science Degree Program

National Taiwan University

December 15, 2021

ADMM

- Now consider the following optimization problem:

$$\begin{aligned} \text{minimize}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad & l(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}) \\ \text{subject to} \quad & \mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} = \mathbf{b}, \end{aligned} \tag{5}$$

  where $\mathbf{U}$ is an $m \times p$ matrix, $\mathbf{V}$ is an $m \times q$ matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\gamma} \in \mathbb{R}^q$ and $\mathbf{b} \in \mathbb{R}^m$.

- **The augmented Lagrangian:** For (5) it is

$$\begin{aligned} \tilde{L}_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}) \quad = \quad & l(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}) + \mathbf{a}^T(\mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} - \mathbf{b}) \\ & + \frac{\rho}{2}||\mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} - \mathbf{b}||_2^2, \end{aligned} \tag{6}$$

  where $\mathbf{a}$ is an m-dimensional vector of Lagrange multipliers.

- *Remark:* The last term of (6) is a **coupling quadratic term** that connects the linear transform of $\mathbf{U}\boldsymbol{\beta}$ and $\mathbf{V}\boldsymbol{\gamma}$.

ADMM

- Now let $\mathbf{v} = \mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} - \mathbf{b}$. Putting $\mathbf{v}$ into the last two terms of (6) yields

$$\begin{aligned}
\mathbf{a}^T\mathbf{v} + \frac{\rho}{2}||\mathbf{v}||_2^2 &= \frac{\rho}{2}\left(\mathbf{v}^T\mathbf{v} + \frac{2}{\rho}\mathbf{a}^T\mathbf{v} + \frac{1}{\rho^2}\mathbf{a}^T\mathbf{a}\right) - \frac{1}{2\rho}\mathbf{a}^T\mathbf{a} \\
&= \frac{\rho}{2}||\mathbf{v} + (1/\rho)\mathbf{a}||_2^2 - \frac{1}{2\rho}||\mathbf{a}||_2^2. \qquad (7)
\end{aligned}$$

  Putting $\boldsymbol{\alpha} = (1/\rho)\mathbf{a}$ into (7) and replacing the last two terms of (6) with the scaled form (7), we obtain

$$\begin{aligned}
& \tilde{L}_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}) \\
= \; & L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \\
= \; & l(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}) + \frac{\rho}{2}||\mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} - \mathbf{b} + \boldsymbol{\alpha}||_2^2 - \frac{\rho}{2}||\boldsymbol{\alpha}||_2^2. \qquad (8)
\end{aligned}$$

- Here we call $\boldsymbol{\beta}$ the **primal variables**, $\boldsymbol{\gamma}$ the **auxiliary variables** and $\boldsymbol{\alpha}$ the **dual variables**.

ADMM

- **An iterative scheme:**
  - The scaled augmented Lagrangian (8) can be minimized by using the following iterative scheme to find a minimizer:

$$\begin{aligned}
\boldsymbol{\beta}^{r+1} &= \arg\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \frac{\rho}{2} ||\mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma}^r - \mathbf{b} + \boldsymbol{\alpha}^r||_2^2 \right\} \\
\boldsymbol{\gamma}^{r+1} &= \arg\min_{\boldsymbol{\gamma}} \left\{ g(\boldsymbol{\gamma}) + \frac{\rho}{2} ||\mathbf{U}\boldsymbol{\beta}^{r+1} + \mathbf{V}\boldsymbol{\gamma} - \mathbf{b} + \boldsymbol{\alpha}^r||_2^2 \right\} \\
\boldsymbol{\alpha}^{r+1} &= \boldsymbol{\alpha}^r + \mathbf{U}\boldsymbol{\beta}^{r+1} + \mathbf{V}\boldsymbol{\gamma}^{r+1} - \mathbf{b}. \qquad (9)
\end{aligned}$$

  - The iterative scheme (9) is an example of the **alternating direction method of multipliers**.
  - *Remark:* Each line in (9) updates a set of parameters given that all other parameters are *fixed*.

ADMM

- **Stopping criteria:**
  - **KKT conditions:** If $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ is a solution to the primal problem of (5) and $\boldsymbol{\alpha}^*$ is a solution to the Lagrangian dual problem associated with (5), then the following conditions must be satisfied:
    - **The primal feasibility condition:**

$$\mathbf{U}\boldsymbol{\beta}^* + \mathbf{V}\boldsymbol{\gamma}^* = \mathbf{b}.$$

    - **The stationarity condition:**

$$\mathbf{0} \in \begin{bmatrix} \partial l(\boldsymbol{\beta}^*) + \rho \mathbf{U}^T \boldsymbol{\alpha}^* \\ \partial g(\boldsymbol{\gamma}^*) + \rho \mathbf{V}^T \boldsymbol{\alpha}^* \end{bmatrix}.$$

ADMM

- **Stopping criteria (contd):**
  - One can show that $\boldsymbol{\beta}^{r+1}$ should satisfy the following relationship:

$$\begin{aligned}\mathbf{0} &\in \partial l(\boldsymbol{\beta}^{r+1}) + \rho \mathbf{U}^T(\mathbf{U}\boldsymbol{\beta}^{r+1} + \mathbf{V}\boldsymbol{\gamma}^r - \mathbf{b} + \boldsymbol{\alpha}^r) \\ &= \partial l(\boldsymbol{\beta}^{r+1}) + \rho \mathbf{U}^T \boldsymbol{\alpha}^r + \rho \mathbf{U}^T \mathbf{U}\boldsymbol{\beta}^{r+1} - \rho \mathbf{U}^T \mathbf{b} + \rho \mathbf{U}^T \mathbf{V}\boldsymbol{\gamma}^r. \quad (10)\end{aligned}$$

    This is the **stationarity condition** for the first line of the iterative scheme (9).

  - On the other hand, from the third line of (9) we have

$$\mathbf{U}^T\boldsymbol{\alpha}^{r+1} = \mathbf{U}^T\boldsymbol{\alpha}^r + \mathbf{U}^T\mathbf{U}\boldsymbol{\beta}^{r+1} + \mathbf{U}^T\mathbf{V}\boldsymbol{\gamma}^{r+1} - \mathbf{U}^T\mathbf{b}. \quad (11)$$

    Plugging in (11) into (10) and rearranging it yields

$$\rho \mathbf{U}^T\mathbf{V}(\boldsymbol{\gamma}^{r+1} - \boldsymbol{\gamma}^r) \in \partial l(\boldsymbol{\beta}^{r+1}) + \rho \mathbf{U}^T\boldsymbol{\alpha}^{r+1}.$$

ADMM

- **Stopping criteria (contd):**
  - Now define

$$\begin{aligned}\mathbf{t}^{r+1} &= \mathbf{U}\boldsymbol{\beta}^{r+1} + \mathbf{V}\boldsymbol{\gamma}^{r+1} - \mathbf{b},\\ \mathbf{s}^{r+1} &= \rho\mathbf{U}^T\mathbf{V}(\boldsymbol{\gamma}^{r+1} - \boldsymbol{\gamma}^r).\end{aligned}$$

  - As suggested by Boyd et al. (2010), one may use the following criteria:

$$\begin{aligned}\frac{||\mathbf{t}^r||_2}{\sqrt{m}} &\leq \epsilon^{\text{abs}} + \frac{10^{-4}}{\sqrt{m}}\max\{||\mathbf{U}\boldsymbol{\beta}^r||_2, ||\mathbf{V}\boldsymbol{\beta}^r||_2, ||\mathbf{b}||_2\},\\ \frac{||\mathbf{s}^r||_2}{\sqrt{p}} &\leq \epsilon^{\text{abs}} + \frac{10^{-4}}{\sqrt{p}}||\rho\mathbf{U}^T\boldsymbol{\alpha}^r||_2\end{aligned}$$

    to stop the iterative scheme (9).
  - Here $\mathbf{t}^r$ and $\mathbf{s}^r$ are called the **primal residual** and the **dual residual** at the rth iteration, respectively.
  - As shown by Boyd et al. (2010), when both l and g are convex,

$$l(\boldsymbol{\beta}^r) + g(\boldsymbol{\gamma}^r) - \text{ opt}^{\text{primal}} \leq ||\rho\boldsymbol{\alpha}^r||_2||\mathbf{t}^r||_2 + ||\boldsymbol{\beta}^r - \boldsymbol{\beta}^*||_2||\mathbf{s}^r||_2.$$

ADMM

- **Choices of ρ:**
  - The scale parameter ρ measures the penalty to which the constraints or the **primal feasibility** (The constraints $\mathbf{U}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\gamma} = \mathbf{b}$) are violated.
  - We can either **fix** ρ during the iteration or vary it to adjust the ratio between the primal and dual residuals.
  - Boyd et al. (2010) provided an adaptive method for computing ρ.

Applications of ADMM

- **Sparse group lasso estimation:**
  - Suppose $\mathbf{X} = (\mathbf{X}_{[1]}, \mathbf{X}_{[2]}, \cdots, \mathbf{X}_{[m]})$, where $\mathbf{X}_{[j]}$ is an $n \times p_j$ matrix, and $\mathbf{X}$ is an $n \times p$ matrix, where $p = \sum_{j=1}^{m} p_j$. We model response $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{j=1}^{m} \mathbf{X}_{[j]}\boldsymbol{\beta}_j + \boldsymbol{\epsilon},$$

    where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m)$, and $\boldsymbol{\epsilon}$ is an n-dimensional vector of residuals.
  - We want to do "variable selection" that can simultaneously select $\mathbf{X}_{[j]}$ and columns of $\mathbf{X}_{[j]}$, i.e. both between-group and within-group sparsity are considered.
  - The sparse group lasso estimation aims to simultaneously do between-group and within-group variable selection:

$$\widehat{\boldsymbol{\beta}}^{\mathsf{SGL}} = \arg\min_{\boldsymbol{\beta}_j' s} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^{m} \mathbf{X}_{[j]}\boldsymbol{\beta}_j \right\|_2^2 + \sum_{j=1}^{m} \Big(\lambda_j ||\boldsymbol{\beta}_j||_2 + \lambda_0 ||\boldsymbol{\beta}_j||_1 \Big) \right\}. \quad (12)$$

Applications of ADMM

- **Sparse group lasso estimation (contd):**
  - We reformulate the optimization problem (12) by introducing a set of auxiliary vectors $\boldsymbol{\gamma}_j$'s and a set of equality constraints $\boldsymbol{\gamma}_j = \boldsymbol{\beta}_j$:

$$\begin{aligned} \text{minimize}_{\boldsymbol{\beta}_j' s} \quad & \left\{ \frac{1}{2} \left|\left| \mathbf{y} - \sum_{j=1}^{m} \mathbf{X}_{[j]} \boldsymbol{\beta}_j \right|\right|_2^2 + \sum_{j=1}^{m} \Big(\lambda_j ||\boldsymbol{\gamma}_j||_2 + \lambda_0 ||\boldsymbol{\gamma}_j||_1 \Big) \right\} \\ \text{subject to} \quad & \boldsymbol{\gamma}_j = \boldsymbol{\beta}_j \text{ for } j = 1, 2, \cdots, m. \end{aligned} \tag{13}$$

  - The augmented Lagrangian corresponding to (13) is

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \quad = \quad & \frac{1}{2} \left|\left| \mathbf{y} - \sum_{j=1}^{m} \mathbf{X}_{[j]} \boldsymbol{\beta}_j \right|\right|_2^2 + \sum_{j=1}^{m} \Big(\lambda_j ||\boldsymbol{\gamma}_j||_2 + \lambda_0 ||\boldsymbol{\gamma}_j||_1 \Big) \\ & + \frac{\rho}{2} \sum_{j=1}^{m} \Big(||\boldsymbol{\beta}_j - \boldsymbol{\gamma}_j + \boldsymbol{\alpha}_j||_2^2 - ||\boldsymbol{\alpha}_j||_2^2 \Big), \end{aligned} \tag{14}$$

    where $\boldsymbol{\alpha}_j$'s are newly introduced scaled dual variables.

Applications of ADMM

- **Sparse group lasso estimation (contd):**
  - We use the following iterative scheme to find a solution to the problem of minimizing (14):

$$
\begin{aligned}
\boldsymbol{\beta}^{r+1} &= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^{m} \mathbf{X}_{[j]} \boldsymbol{\beta}_j \right\|_2^2 + \frac{\rho}{2} \sum_{j=1}^{m} \left(||\boldsymbol{\beta}_j - \boldsymbol{\gamma}_j^r + \boldsymbol{\alpha}_j^r||_2^2 \right) \right\} \\
\boldsymbol{\gamma}_j^{r+1} &= \arg\min_{\boldsymbol{\gamma}_j} \left(\frac{\lambda_j}{\rho} ||\boldsymbol{\gamma}_j||_2 + \frac{\lambda_0}{\rho} ||\boldsymbol{\gamma}_j||_1 + \frac{1}{2} ||\boldsymbol{\gamma}_j - \boldsymbol{\beta}_j^{r+1} - \boldsymbol{\alpha}_j^r||_2^2 \right) \\
& \quad \text{for } j = 1, 2, \cdots, m \\
\boldsymbol{\alpha}_j^{r+1} &= \boldsymbol{\alpha}_j^r + \boldsymbol{\beta}_j^{r+1} - \boldsymbol{\gamma}_j^{r+1} \text{ for } j, 1, 2, \cdots, m. \qquad (15)
\end{aligned}
$$

  - The first line of (15) has a closed form representation:

$$
\boldsymbol{\beta}^{r+1} = (\mathbf{X}^T \mathbf{X} + \rho \mathbf{I}_{p \times p})^{-1} [\mathbf{X}^T \mathbf{y} + \rho(\boldsymbol{\gamma}^r - \boldsymbol{\alpha}^r)], \qquad (16)
$$

    where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_m)$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m)$.

Applications of ADMM

- **Sparse group lasso estimation (contd):**
  - The second line of (15) also has a heuristic representation:

$$\boldsymbol{\gamma}_j^{r+1} = \left(||S_{\lambda_0/\rho} \circ (\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r)||_2 - \frac{\lambda_j}{\rho} \right)_+ \frac{S_{\lambda_0/\rho} \circ (\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r)}{||S_{\lambda_0/\rho} \circ (\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r)||_2}, \quad (17)$$

    where $S_{\lambda_0/\rho}(a)$ is the **soft thresholding operator**, i.e.

$$S_{\lambda_0/\rho}(a) = \ \mathsf{sign}(a)\left(|a| - \frac{\lambda_0}{\rho} \right)_+$$

    and $\circ$ is the elementarywise operator.

- *Remark 1:* Computation of (16) can be done by first computing the Cholesky decomposition of $\mathbf{X}^T\mathbf{X} + \rho\mathbf{I}_{p\times p}$, and then storing the lower triangular matrix of the Cholesky decomposition for reuse during the iteration.

- *Remark 2:* Computation of (17) is straightforward and usually can be done in $O(p)$ flops.

Applications of ADMM

- **Sparse group lasso estimation (contd):**
  - *Remark 3:* When $\lambda_j = 0$ for all j's, the problem (12) becomes the **lasso estimation problem**, and (17) becomes

$$\boldsymbol{\gamma}_j^{r+1} = S_{\lambda_0/\rho} \circ (\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r), \tag{18}$$

which is the **soft thresholding operator** or the **proximal operator of** l_1**-norm** on the vector $\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r$.

When $\lambda_0 = 0$, the problem (12) becomes the **group lasso estimation problem**, and (17) becomes

$$\boldsymbol{\gamma}_j^{r+1} = \left(||\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r||_2 - \frac{\lambda_j}{\rho} \right)_+ \frac{\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r}{||\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r||_2}, \tag{19}$$

which is the **proximal operator of the** l_2**-norm** on the vector $\boldsymbol{\beta}_j^{r+1} + \boldsymbol{\alpha}_j^r$.

Applications of ADMM

- **Quiz:**

1. Consider soft thresholding operator:

$$S_\lambda(a) = \text{sign}(a)\Big(|a| - \lambda\Big)_+.$$

Which of the following statements are *true*?

a. $S_\lambda(a)$ is the solution to the following problem:

$$\min_x \frac{1}{2}(x-a)^2 + \lambda x.$$

b. $S_\lambda(a)$ is the solution to the following problem:

$$\min_x \frac{\lambda}{2}(x-a)^2 + |x|.$$

c. $S_\lambda(a)$ is the solution to the following problem:

$$\min_x \frac{1}{2}(x-a)^2 + \lambda|x|.$$

Applications of ADMM

- **Support vector machines:**
  - Remember the support vector machine problem:

$$\min_{\boldsymbol{\theta},\alpha} \quad \frac{1}{2}||\boldsymbol{\theta}||_2^2$$
$$\text{subject to} \quad -y_i(\mathbf{x}_i^T\boldsymbol{\theta}+\alpha) \le -1 \text{ for } i=1,2,\cdots,n. \tag{20}$$

  - **Regularized estimation form for (20):** To do that, we first introduce a set of **slack variables** $\xi_i \ge 0$ for $i=1,2,\cdots,n$ to the inequality constraints to *relax* the problem (20) as

$$\min_{\boldsymbol{\theta},\alpha} \quad \frac{1}{2}||\boldsymbol{\theta}||_2^2 + \frac{1}{\lambda}\sum_{i=1}^{n}\xi_i$$
$$\text{subject to} \quad -y_i(\mathbf{x}_i^T\boldsymbol{\theta}+\alpha) - \xi_i \le -1, \xi_i \ge 0 \text{ for } i=1,2,\cdots,n. \tag{21}$$

    where $\lambda > 0$ is a tuning parameter adjusting the impact of the penalty function $\sum_{i=1}^{n}\xi_i$.

  - The problem (21) *approximates* to (20) when $\xi_i \to 0$ for all i's. This makes sense since we add a penalty term $(1/\lambda)\sum_{i=1}^{n}\xi_i$ to indicate that we want ξ_i's as small as possible.
  - ξ_i can be approximated as

$$\xi_i \approx \max\{0, 1-y_i(\mathbf{x}_i^T\boldsymbol{\theta}+\alpha)\}. \tag{22}$$

Applications of ADMM

- **Support vector machines (contd):**
  - With (22) we can trun the problem (21) as a regularized estimation problem:

$$(\widehat{\boldsymbol{\theta}}, \widehat{\alpha}) = \arg\min_{\boldsymbol{\theta}, \alpha} \left\{ \sum_{i=1}^{n} \max\{0, 1 - y_i(\mathbf{x}_i^T \boldsymbol{\theta} + \alpha)\} + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2 \right\}. \qquad (23)$$

  - (23) is a regularized estimation form for the SVM problem (20).
  - The function

$$l(f) = \max\{0, 1 - f\}$$

    is called the **hinge loss** of f.
  - We can write (23) in ADMM form as

$$\begin{aligned} \min_{\boldsymbol{\theta}, \alpha} \quad & (l \circ [\mathbf{y} \circ (\mathbf{X}\boldsymbol{\theta} + \alpha\mathbf{1})])^T \mathbf{1} + \frac{\lambda}{2} ||\boldsymbol{\gamma}||_2^2 \\ \text{subject to} \quad & \boldsymbol{\theta} = \boldsymbol{\gamma}. \end{aligned} \qquad (24)$$

Applications of ADMM

- **Support vector machines (contd):**
  - The augmented Lagrangian corresponding to (24) is

$$\begin{aligned} L(\boldsymbol{\theta}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\tau}) &= (l \circ [\mathbf{y} \circ (\mathbf{X}\boldsymbol{\theta} + \alpha \mathbf{1})])^T \mathbf{1} + \frac{\lambda}{2} ||\boldsymbol{\gamma}||_2^2 \\ &+ \frac{\rho}{2} ||\boldsymbol{\theta} - \boldsymbol{\gamma} + \boldsymbol{\tau}||_2^2 - \frac{\rho}{2} ||\boldsymbol{\tau}||_2^2. \end{aligned} \tag{25}$$

    The iterative scheme for solving the problem of minimizing (25) is:

$$\begin{aligned} (\boldsymbol{\theta}^{r+1}, \alpha^{r+1}) &= \arg\min_{\boldsymbol{\theta}} \left\{ (l \circ [\mathbf{y} \circ (\mathbf{X}\boldsymbol{\theta} + \alpha \mathbf{1})])^T \mathbf{1} + \frac{\rho}{2} ||\boldsymbol{\theta} - \boldsymbol{\gamma}^r + \boldsymbol{\tau}^r||_2^2 \right\} \\ \boldsymbol{\gamma}^{r+1} &= \arg\min_{\boldsymbol{\gamma}} \left\{ \frac{\lambda}{2\rho} ||\boldsymbol{\gamma}||_2^2 + \frac{1}{2} ||\boldsymbol{\gamma} - \boldsymbol{\theta}^{r+1} - \boldsymbol{\tau}^r||_2^2 \right\} \\ &= \frac{\boldsymbol{\theta}^{r+1} + \boldsymbol{\tau}^r}{\lambda/\rho + 1} \\ \boldsymbol{\tau}^{r+1} &= \boldsymbol{\tau}^r + \boldsymbol{\theta}^{r+1} - \boldsymbol{\gamma}^{r+1}. \end{aligned}$$

Applications of ADMM

- **Splitting across examples:**
  - The **Splitting across Examples** is a distributed optimization technique useful in situations when there are large amounts of examples (samples) but relatively small number of features in model fitting tasks.
  - Consider the following optimization problem:

$$\text{minimize} \qquad \sum_{i=1}^{m} l_i(\mathbf{X}_i\boldsymbol{\beta}) + g(\boldsymbol{\beta}), \tag{26}$$

    where $\mathbf{X}_i$ is an $n_i \times p$ matrix, and $\boldsymbol{\beta}$ is a p-dimensional vector.
  - **Example 1:** We have

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} \Rightarrow \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 = \frac{1}{2}\sum_{i=1}^{m}||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}||_2^2.$$

Applications of ADMM

- **Splitting across examples (contd):**
  - To use the idea of ADMM, we introduce the following equality constraints:

$$\boldsymbol{\theta}_i - \boldsymbol{\beta} = 0 \text{ for } i = 1, 2, \cdots, m.$$

  - With the above constraints, we can reformulate the optimization problem (26) as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} l_i(\mathbf{X}_i \boldsymbol{\theta}_i) + g(\boldsymbol{\beta}) \\ \text{subject to} \quad & \boldsymbol{\theta}_i = \boldsymbol{\beta} \text{ for } i = 1, 2, \cdots, m. \end{aligned} \tag{27}$$

  The augmented Lagrangian for (27) is

$$\begin{aligned} L_\rho(\boldsymbol{\theta}_i's, \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^{m} l_i(\mathbf{X}_i \boldsymbol{\theta}_i) + g(\boldsymbol{\beta}) \\ &\quad + \frac{\rho}{2} \sum_{i=1}^{m} ||\boldsymbol{\theta}_i - \boldsymbol{\beta} + \boldsymbol{\alpha}_i||_2^2 - \frac{\rho}{2} \sum_{i=1}^{m} ||\boldsymbol{\alpha}_i||_2^2. \end{aligned} \tag{28}$$

Applications of ADMM

- **Splitting across examples (contd):**
  - We can minimize (28) using the following iterative scheme:

$$
\begin{aligned}
\boldsymbol{\theta}_i^{r+1} &= \arg\min_{\boldsymbol{\theta}_i} \left\{ l_i(\mathbf{X}_i\boldsymbol{\theta}_i) + \frac{\rho}{2}||\boldsymbol{\theta}_i - \boldsymbol{\beta}^r + \boldsymbol{\alpha}_i^r||_2^2 \right\} \text{ for } i = 1, 2, \cdots, m \\
\boldsymbol{\beta}^{r+1} &= \arg\min_{\boldsymbol{\beta}} \left\{ g(\boldsymbol{\beta}) + \frac{m\rho}{2} \left|\left| \boldsymbol{\beta} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\theta}_i^{r+1} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\alpha}_i^r \right|\right|_2^2 \right\} \\
\boldsymbol{\alpha}_i^{r+1} &= \boldsymbol{\alpha}_i^r + \boldsymbol{\theta}_i^{r+1} - \boldsymbol{\beta}^{r+1} \text{ for } i = 1, 2, \cdots, m.
\end{aligned} \tag{29}
$$

  - **Derivation of the second line of (29):** Note that to solve the second line of (29), the following first-order condition must be satisfied:

$$
\begin{aligned}
& \mathbf{0} \in \partial g(\boldsymbol{\beta}^{r+1}) - \rho\left(\sum_{i=1}^m \boldsymbol{\theta}_i^{r+1} - m\boldsymbol{\beta}^{r+1} + \sum_{i=1}^m \boldsymbol{\alpha}_i^r \right) \\
\Rightarrow \quad & \mathbf{0} \in \partial g(\boldsymbol{\beta}^{r+1}) + m\rho\left(\boldsymbol{\beta}^{r+1} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\theta}_i^{r+1} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\alpha}_i^r \right),
\end{aligned}
$$

which is equivalent to solving the following optimization problem:

$$
\boldsymbol{\beta}^{r+1} = \arg\min_{\boldsymbol{\beta}} \left\{ g(\boldsymbol{\beta}) + \frac{m\rho}{2} \left|\left| \boldsymbol{\beta} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\theta}_i^{r+1} - \frac{1}{m}\sum_{i=1}^m \boldsymbol{\alpha}_i^r \right|\right|_2^2 \right\}. \tag{30}
$$

Computation in Data Science: Week 13 Lecture 8

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Data Science Degree Program

National Taiwan University

December 15, 2021

Subdifferentials

- **Motivation:** For a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ that is **convex** and **differentiable**, one has

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$$

  for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, i.e. the first-order characterization of convex functions.

- What if f is convex but is *not* differentiable at some points in $\mathbb{R}^p$?

Subdifferentials

- **Definition:** Assume $\mathcal{C} \in \mathbb{R}^p$ is a non-empty set. For a function $f : \mathcal{C} \mapsto \mathbb{R}$, the subdifferential of f at $\mathbf{x} \in \mathcal{C}$ is defined by

$$\begin{aligned} \partial f(\mathbf{x}) &= \{\mathbf{u} : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{u}^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{y} \in \mathcal{C}\} \\ &= \bigcap_{\mathbf{y} \in \mathcal{C}} \Big\{\mathbf{u} : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{u}^T(\mathbf{y} - \mathbf{x})\Big\}, \end{aligned} \tag{31}$$

  that is, the subdifferential $\partial f(\mathbf{x})$ evaluated at $\mathbf{x}$ is a set of all p-dimensional vectors $\mathbf{u}$ such that the inequality

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{u}^T(\mathbf{y} - \mathbf{x})$$

  holds for *any* $\mathbf{y} \in \mathcal{C}$.

- **Theorem (existence of subdifferentials):** Assume $\mathcal{C}$ is a convex set. A function $f : \mathcal{C} \mapsto \mathbb{R}$ is a convex function if and only if the corresponding subdifferentials are **non-empty at any** $\mathbf{x} \in \mathcal{C}$, i.e. $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \mathcal{C}$.

- *Remark:* We will call any vector $\mathbf{u} \in \partial f(\mathbf{x})$ the **subgradient** of f evaluated at $\mathbf{x}$. A subgradient is an element in the subdifferential.

Subdifferentials

- **First-order characterization:** If $f : \mathcal{C} \mapsto \mathbb{R}$ is convex and differentiable, then we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$. This implies $\partial f(\mathbf{x}) = \nabla f(\mathbf{x})$.

- **Solutions to a convex optimization problem:** If $l : \mathcal{C} \mapsto \mathbb{R}$ is convex and $\mathbf{x}^* \in \mathcal{C}$ minimizes l, then we must have

$$l(\mathbf{y}) \geq l(\mathbf{x}^*) = l(\mathbf{x}^*) + \mathbf{0}^T(\mathbf{y} - \mathbf{x}^*) \text{ for all } \mathbf{y} \in \mathcal{C},$$

which implies

$$\mathbf{0} \in \partial l(\mathbf{x}^*).$$

- **Composition with an affine mapping:** Assume $l : \mathcal{C} \mapsto \mathbb{R}$ is convex. Define $\mathbf{f} = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}$ with $\boldsymbol{\beta} \in \mathbb{R}^p$, and $g(\boldsymbol{\beta}) = l(\mathbf{X}\boldsymbol{\beta})$. Then we have $g(\boldsymbol{\beta})$ is a convex function of $\boldsymbol{\beta}$ and

$$\partial g(\boldsymbol{\beta}) = \left\{ \mathbf{u} : \mathbf{u} = \mathbf{X}^T\mathbf{v}, \mathbf{v} \in \partial l(\mathbf{f}), \mathbf{f} = \mathbf{X}\boldsymbol{\beta} \right\}.$$

Subdifferentials

- **Example 1 (The l_1-norm):**
  - Consider the l_1-norm $||\mathbf{x}||_1$ with $\mathbf{x} \in \mathbb{R}^p$. Since $||.||_1$ is convex on $\mathbb{R}^p$, if $\mathbf{u}$ is a subgradient of $||\mathbf{x}||_1$, we must have

    $$(\mathbf{y} - \mathbf{x})^T \mathbf{u} + ||\mathbf{x}||_1 \leq ||\mathbf{y}||_1$$

    for any $\mathbf{y} \in \mathbb{R}^p$.

    Case 1: Obviously for $x_j \neq 0$ we have

    $$\frac{\partial ||\mathbf{x}||_1}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\sum_{j=1}^{p} |x_j|\right) = \text{ sign}(x_j).$$

    Case 2: For $x_j = 0$, if u_j is a subgradient of $|x_j|$, we must have

    $$(y - 0)u_j + 0 \leq |y| \Rightarrow yu_j \leq |y| \Rightarrow u_j \in [-1, 1],$$

    i.e. u_j is any point in $[-1, 1]$.

    Case 3: For $\mathbf{x} = \mathbf{0}$, if $\mathbf{u}$ is a subgradient of $||\mathbf{x}||_1$, then for any $\mathbf{y} \in \mathbb{R}^p$ we must have

    $$\mathbf{y}^T \mathbf{u} \leq ||\mathbf{y}||_1 \Rightarrow \max_{\mathbf{u}} \mathbf{y}^T \mathbf{u} \leq ||\mathbf{y}||_1 \Rightarrow ||\mathbf{y}||_1 ||\mathbf{u}||_\infty \leq ||\mathbf{y}||_1 \Rightarrow ||\mathbf{u}||_\infty \leq 1.$$

  - From the above results we conclude for $\mathbf{x} \in \mathbb{R}^p$,

    $$\partial ||\mathbf{x}||_1 = \{\mathbf{u} : ||\mathbf{u}||_\infty \leq 1 \text{ and } \mathbf{u}^T \mathbf{x} = ||\mathbf{x}||_1\}.$$

Subdifferentials

- **Example 2 (The l_2-norm):**
  - Consider the l_2-norm $||\mathbf{x}||_2$ with $\mathbf{x} \in \mathbb{R}^p$.

    Case 1: Obviously for $\mathbf{x} \neq \mathbf{0}$ we have

    $$\frac{\partial ||\mathbf{x}||_2}{\partial \mathbf{x}} = \frac{\mathbf{x}}{||\mathbf{x}||_2}.$$

    Case 2: For $\mathbf{x} = \mathbf{0}$, if $\mathbf{u}$ is a subgradient of $||\mathbf{x}||_2$, then for any $\mathbf{y} \in \mathbb{R}^p$ we must have

    $$\mathbf{y}^T\mathbf{u} \leq ||\mathbf{y}||_2 \Rightarrow \max_{\mathbf{u}} \mathbf{y}^T\mathbf{u} \leq ||\mathbf{y}||_2 \Rightarrow ||\mathbf{y}||_2||\mathbf{u}||_2 \leq ||\mathbf{y}||_2 \Rightarrow ||\mathbf{u}||_2 \leq 1.$$

    From the above results we conclude for $\mathbf{x} \in \mathbb{R}^p$,

    $$\partial||\mathbf{x}||_2 = \{\mathbf{u} : ||\mathbf{u}||_2 \leq 1 \text{ and } \mathbf{u}^T\mathbf{x} = ||\mathbf{x}||_2\}.$$

Subdifferentials

- **Example 3 (The rectified linear unit (ReLU)):**
  - Consider the rectified linear unit $f(x) = \max\{0, x\}$ with $x \in \mathbb{R}$. The function is a convex function of $x \in \mathbb{R}$ but is not differentiable at $x = 0$.

    Case 1: Obviously for $x > 0$ we have

    $$\frac{d \max\{0, x\}}{dx} = \frac{dx}{dx} = 1.$$

    Case 2: For $x < 0$, we have

    $$\frac{d \max\{0, x\}}{dx} = \frac{d0}{dx} = 0.$$

    Case 3: For $x = 0$, if u is a subgradient of $\max\{0, x\}$, then for all $y \in \mathbb{R}$, we must have

    $$(y - 0)u + \max\{0, 0\} \leq \max\{0, y\} \Rightarrow yu \leq \max\{0, y\} \Rightarrow u \in [0, 1].$$

Subdifferentials

- **Computation of subdifferentials (contd):**
  - **Example 4 (The hinge loss):** The hinge loss of f is defined as $l(f) = \max\{0, 1 - f\}$ with $f \in \mathbb{R}$.
  - $l(f)$ is convex since for any $f, h \in \mathbb{R}$ and $\alpha \in [0,1]$,

$$\begin{aligned} \max\{0, 1 - [\alpha f + (1-\alpha)h]\} &= \max\{0, \alpha(1-f) + (1-\alpha)(1-h)\} \\ &\leq \max\{0, \alpha(1-f)\} + \max\{0, (1-\alpha)(1-h)\} \\ &= \alpha \max\{0, 1-f\} + (1-\alpha)\max\{0, 1-h\}. \end{aligned}$$

  - If u is a subgradient of $l(f)$, then for any $h \in \mathbb{R}$, the following inequality must be satisfied:

$$(h - f)u + \max\{0, 1 - f\} \leq \max\{0, 1 - h\}.$$

Case 1: Obviously for $f > 1$ we have

$$\frac{d \max\{0, 1-f\}}{df} = \frac{d0}{df} = 0.$$

Case 2: For $f < 1$, we have

$$\frac{d \max\{0, 1-f\}}{dx} = \frac{d(1-f)}{df} = -1.$$

Subdifferentials

- **Computation of subdifferentials (contd):**
  - **Example 4 (contd):**

    Case 3: For $f = 1$, if u is a subgradient of $l(f)$, then for all $h \in \mathbb{R}$, we must have

    $(h-0)u + \max\{0,0\} \leq \max\{0, 1-h\} \Rightarrow hu \leq \max\{0, 1-h\} \Rightarrow u \in [-1, 0]$.

  - Now let $\mathbf{f} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $n \times p$ matrix, and $\boldsymbol{\beta}$ is a p-dimensional vector. Consider the loss function with $l(f) = \max\{0, 1-f\}$:

    $$\sum_{i=1}^{n} l(f_i) = (l \circ \mathbf{X}\boldsymbol{\beta})^T \mathbf{1} = g(\boldsymbol{\beta}),$$

    where $\mathbf{1} = (1, 1, \cdots, 1)$.

    We have

    $$\partial g(\boldsymbol{\beta}) = \left\{ \mathbf{u} : \mathbf{u} = \mathbf{X}^T \mathbf{v}, \mathbf{v} \in \partial\{(l \circ \mathbf{f})^T \mathbf{1}\}, \mathbf{f} = \mathbf{X}\boldsymbol{\beta} \right\}.$$

    Here $\mathbf{u}$ is a p-dimensional vector, and $\mathbf{v}$ is an n-dimensional vector and can be expressed as

    $$(\mathbf{v})_i = \begin{cases} -1 \text{ if } \mathbf{x}_i^T \boldsymbol{\beta} < 1 \\ 0 \text{ if } \mathbf{x}_i^T \boldsymbol{\beta} > 1 \\ \text{any point} \in [-1, 0] \text{ otherwise} \end{cases}$$

    for $i = 1, 2, \cdots, n$.

Subdifferentials

- **About non-convex functions:**
  - For a non-convex function, subdifferentials may not exist for some points. For example, for $x \in \mathbb{R}$ consider the following function:

$$f(x) = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases} .$$

  - $f(x)$ is non-convex in x.
  - Assume the subdifferential of $f(x)$ evaluated at $x = 0$ is non-empty. Then for any $y > 0$ there exists a u in the subdifferential such that

$$\begin{aligned} (y-0)u + f(0) \leq f(y) \quad &\Rightarrow \quad u \leq \frac{f(y)}{y} = \frac{1}{y} \\ &\Rightarrow \quad u \in (-\infty, 1/y], \end{aligned}$$

i.e. u is dependent on y and therefore is contradict to the definition of the subdifferential of $f(x)$ evaluated at $x = 0$.

Subdifferentials

- **Quiz:**

1. Consider the l_∞-norm $||\mathbf{x}||_\infty$ with $\mathbf{x} \in \mathbb{R}^p$. Which of the following statements are *true*?

a. For $x_j = (\mathbf{x})_j$ with $x_j \neq 0$, we have

$$\frac{\partial ||\mathbf{x}||_\infty}{\partial x_j} = \begin{cases} 1 \text{ if } |x_j| = \max_{j'} |x_{j'}| \\ 0 \text{ otherwise} \end{cases} .$$

b. For $\mathbf{x} = \mathbf{0}$, if $\mathbf{u}$ is a subgradient of $||\mathbf{x}||_\infty$, we must have

$$(\mathbf{y} - \mathbf{0})^T \mathbf{u} + ||\mathbf{0}||_\infty \leq ||\mathbf{y}||_\infty \Rightarrow ||\mathbf{u}||_1 \leq 1$$

for any $\mathbf{y} \in \mathbb{R}^p$.

c. In general we have

$$\partial ||\mathbf{x}||_\infty = \{\mathbf{u} \in \mathbb{R}^p : 1 \leq ||\mathbf{u}||_1 < \infty \text{ and } \mathbf{u}^T \mathbf{x} = ||\mathbf{x}||_\infty\}.$$

Week 13

- **References:**

  - H. H. Bauschke and P. L. Combettes (2010). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer. The materials on sudifferentials and Fenchel dual problems are from this book.
  - S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, Vol. 3. No. 1, 1-122. The materials on ADMM in the slides are from this book. This book can be downloaded from the Internet.
  - S. Boyd and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press. The materials on Lagrangian dual problems in the slides are from this book. This book can be downloaded from the Internet.