1 Regularizers and Examples

1.1 Sparsity Regularizer: $l_1$ norm

In order to avoid overfitting, $l_1$ norm can be used to induce sparsity in our model. (That is, we want to constrain some parameters to be 0 in some cases.) To begin with, let’s consider ridge regression.

$$
\min_{\omega \in \mathbb{R}} \frac{1}{2} \left( \omega - b \right)^2
$$

Recall that this is an unconstrained convex optimization problem, optimal solution occurs when the first derivative at that point is 0. It is obvious that the optimal solution $\omega^* = b$. Hence, $\omega^* = 0$ if and only if that $b = 0$. Our goal is to constrain the optimal solution, $\omega^*$, to be 0 for some range of $b$.

$$
\min_{\omega \in \mathbb{R}} \frac{1}{2} (\omega - b)^2 + \lambda \omega^2
$$

Solve $(\omega^* - b) + 2\lambda \omega = 0$, we can get $\omega^* = \frac{b}{1 + 2\lambda}$, where $\lambda \neq -\frac{1}{2}$. It is easy to see that $\omega^* = 0$ if and only if $b = 0$. So it shows that $l_2$ norm doesn’t promote sparsity.

Now let’s consider $l_1$ norm.

$$
\min_{\omega \in \mathbb{R}} \frac{1}{2} (\omega - b)^2 + \lambda |\omega|
$$
Let \( f(\omega) = \frac{1}{2}(\omega - b)^2 + \lambda|\omega| \). We can see that \( f(\cdot) \) is differentiable if \( \omega \neq 0 \). Suppose that \( \omega^* \) is a differentiable point. we define that \( \text{sgn}(x) = \frac{x}{|x|} \). We solve \( (\omega^* - b) + \lambda \text{sgn}(\omega^*) = 0 \) and get \( \omega^* = b - \lambda \text{sgn}(\omega^*) \). Hence we have

\[
\omega^* = \begin{cases} 
  b - \lambda & \text{if } b - \lambda > 0 \\
  b + \lambda & \text{if } b + \lambda < 0 \\
  0 & \text{if } -\lambda \leq b \leq \lambda 
\end{cases}
\]

Thus, \( \omega = 0 \) if and only if \( -\lambda \leq b \leq \lambda \). We can see that \( l_1 \) norm helps make the minimizer, \( \omega^* \), to 0 in a relatively larger range of \( b \).

On the other hand, in practice, \( \lambda \) is usually chosen by using cross validation.

### 1.2 Low Rank Structure in Hidden Partition Problem

Basically, all the vertices in graph are partitioned into groups with the same size. Let \( A \) be an adjacency matrix. Hidden partition problem can be formulated as

\[
\max_x \sum_{i,j} A_{ij}(1 + x_i x_j) \tag{4}
\]

subject to \( \sum_{i=1}^{n} x_i = 0 \) (The size of two groups must be the same)

\( x_i \in \{+1, -1\} \ \forall i \) (vertex \( i \) is either in group 1 or group 2)

By moving the first constraint into the objective function, we obtain

\[
\max_x x^T Ax - \lambda x^T 11^T x \tag{5}
\]

subject to \( x_i \in \{-1, +1\} \ \forall i \)

Let \( \bar{X} = x^T x \). In the next step, we want to use \( \bar{X} \) to replace \( x \) in (5). It follows from the objective function of (5) that

\[
x^T Ax - \lambda x^T 11^T x \tag{6}
\]

\[
= \text{Tr}(x^T Ax) - \lambda \text{Tr}(x^T 11^T x) \\
= \text{Tr}(Axx^T) - \lambda \text{Tr}(11^T xx^T)
\]
Let $-\tilde{A} = A - \lambda 11^T$. Then we want to build constraints of $\tilde{X}$ such that $\tilde{X}$ preserves the property of $x$ in (5). Note that $\tilde{X}$ is positive semidefinite, and the rank of it is 1. Also, $X_{ii} = x_i^2$ and $x_i \in \{-1, +1\}$ implies that $X_{ii} = 1$. Now the hidden partition problem can transformed into

\[
\min_{\tilde{X}} \langle \tilde{A}, \tilde{X} \rangle \quad \text{subject to} \quad \tilde{X} \succeq 0, \quad \tilde{X}_{ii} = 1, \quad \text{rank}(\tilde{X}) \leq 1
\]

(7) is not a convex optimization problem, because rank($\tilde{X}$) is non-convex. (rank($\tilde{X}$) is $l_0$ norm of singular value of $\tilde{X}$). In practice, $l_0$ norm can be approximated by $l_1$ norm.

1.3 More Examples

1) Netflix Prize
In order to recommend movies to customers, Netflix wants to estimate the ratings of all the movies from them. Suppose we have $n$ movies and $m$ customers. Let $A$ be an $n$ by $m$ Matrix, which stores the ratings of various movies from all the $m$ customers. In particular, $A_{ij}$ represents the rating of movie $i$ from customer $j$. $A_{ij}$ is blank if the rating of movie $i$ from customer $j$ is unknown. Note that the rating matrix, $A$, is very sparse, because nobody can watch every movie. The rank of $A$ is usually around 20 to 50. Let $\Omega$ denote the set of the indices of all the known ratings, $A'_{ij}$.s. Let $X = \{x_{ij}\}$ denote the estimated rating matrix. We want to minimize the error of estimates and potential of overfitting. Mathematically speaking, we want to use a relatively low-rank matrix $X$ to estimate the rating matrix $A$. Now the corresponding optimization can be formulated as

\[
\min_X \frac{1}{2} \sum_{(i,j) \in \Omega} (A_{ij} - X_{ij})^2 + \lambda \text{rank}(X)
\]
which is equivalent to

$$\min_{X} \frac{1}{2} \| P_{\Omega}(A - X) \|^2_F + \lambda X$$  \hspace{1cm} (9)$$

where $P_{\Omega}$ is projection

2) Haplotype Phasing

A haplotype is the sequence of nucleotides along a single chromosome. The human genome is made up of 23 pairs of chromosomes. For a chromosome with $k$ variants, we can represent its haplotype as a string from the set $\{A, C, G, T\}^k$. With current technology, it is difficult to separate a pair of chromosomes, and we often get the two haplotypes mixed together. In fact, we assume that variants are bi-allelic, that is each variant takes one of two possible allelic values. Therefore, without loss of generality, we can represent haplotypes as a string from the set $\{-, +\}$, where $-$ and $+$ represent two possible allelic values at each variant location.

The objective of the haplotype phasing problem is to recover the two haplotypes from mother and father out of the $2^k$ possible haplotypes of an individual. We only know whether two signs on two different locations is the same or not. Moreover, we consider there are some errors in the observation. Let $A$ be the matrix of observations. We define that

$$A_{ij} = \begin{cases} 
0 & \text{if signs on location } i \text{ and location } j \text{ are not compared} \\
+1 & \text{signs on location } i \text{ and location } j \text{ are the same} \\
-1 & \text{signs on location } i \text{ and location } j \text{ are different} 
\end{cases}$$  \hspace{1cm} (10)$$

For instance, suppose we have a haplotype $\{+++\}$.

The $A$ matrix can be

$$A = \begin{bmatrix} 
+1 & -1 & 0 & 0 \\
-1 & +1 & 0 & -1 \\
0 & 0 & +1 & 0 \\
0 & -1 & 0 & +1 
\end{bmatrix}$$  \hspace{1cm} (11)$$

$A_{14}$ is $-1$, because location 1 has sign $+$ and location 4 has sign $+1$. Since we incorrectly observe the signs on location 1 and location 2, $A_{12}$ becomes $-1$.

The decision variables are as follows

$$x_i : \text{sign on location } i \hspace{0.5cm} x_i \in \{-1, +1\}$$
Given A, the matrix of observations, we want to maximize consistence.

\[
\max_{x \in \mathbb{R}^k} \sum_{i,j} x_i x_j A_{ij} \quad (12)
\]

subject to \( x_i \in \{-1, +1\} \ \forall i \)

which is equivalent to

\[
\max_{X \in \mathbb{R}^{k \times k}} < A, X > \quad (13)
\]

subject to \( X_{ii} = 1 \ \forall i \)

\( X \succeq 0 \ \text{rank}(X) \leq 1 \)

### 1.4 Low rank Structure and Singular Norm

\[
\min_X f(X) \quad (14)
\]

subject to \( \text{rank}(X) \leq r \)

(14) is not a convex optimization problem, because \( \text{rank}(X) := ||\sigma||_0 \) (namely, number of nonzero singular values) is not convex. But we can use nuclear norm, \( ||X||_* := \sum_i \sigma_i(X) = ||\sigma||_1 \), to estimate \( ||\sigma||_0 \). Thus (14) can be approximated by a convex optimization problem below.

\[
\min_X f(X) + \lambda ||X||_* \quad (15)
\]

Note that \( ||X||_* \) is non-differentiable. In the one dimensional case, \( \sigma(x) = |x| \). And \( |x| \) is non-differentiable at 0.

### 2 Convex but Non-smooth Optimization

#### 2.1 Subgradient

Basically, subgradient gives an affine global underestimate of the function, \( f \). Subgradient is very in solving non-smooth convex optimization problem.
Definition 1 Vector $g$ is a subgradient of $f : \mathbb{R}^d \to \mathbb{R}$ at $x$ if $f(y) \geq f(x) + \langle g, y - x \rangle$, for all $y$.

Definition 2 Subdifferential, $\partial f(x)$, is a set of all the subgradients of $f$ at $x$.

Example: $f(x) = |x|$

$$\partial f(x) = \begin{cases} 
{+1} & \text{if } x > 0 \\
{-1} & \text{if } x < 0 \\
[-1, +1] & \text{if } x = 0 
\end{cases} \quad (16)$$

Theorem 1 Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f : S \to \mathbb{R}$ be convex. Then for $x \in \text{int} S$, there exists a subgradient of $f$ at $x$. Let’s give the formal definition of subgradients and subdifferential. Then we will prove some important properties of subdifferential.

Proof

Since $f$ is a convex function, $\text{epi} f := \{(x, y) : x \in S, y \geq f(x)\}$ is convex. Noting that $(\bar{x}, f(\bar{x}))$ belongs to the boundary of $\text{epi} f$, there exists a supporting hyperplane which supports $\text{epi} f$ at $\bar{x}$. That is

$$\langle \xi_0, x - \bar{x} \rangle + \mu[y - f(\bar{x})] \leq 0 \quad \forall (x, y) \in \text{epi} f \quad (17)$$

$\mu$ is not positive, because $\lim_{y \to \infty} \mu[y - f(\bar{x})] = \infty$, it implies that we can choose sufficiently large $y$ such that the left hand side of (17) is greater than 0.

Then we need to show that $\mu$ can’t be 0. By contradiction, suppose that $\mu = 0$. Then $\langle \xi_0, x - \bar{x} \rangle$. Since $\bar{x} \in \text{int} S$, there exists $\lambda > 0$ such that $\bar{x} + \lambda \xi_0 \in S$. $\langle \xi_0, \bar{x} + \lambda \xi_0 - \bar{x} \rangle = -\lambda ||\xi_0||_2^2 \leq 0$ implies that $\xi_0 = 0$, which contradicts the fact that $(\xi_0, \mu)$ is nonzero vector.

Thus, $\mu < 0$. Let $\xi := \frac{\xi_0}{|\mu|}$ We divide (13) by $|\mu|$ and get

$$\langle \xi, x - \bar{x} \rangle - y + f(\bar{x}) \leq 0 \quad \forall (x, y) \in \text{epi} f \quad (18)$$

By letting $f(x) = y$, we get $f(x) \geq f(\bar{x}) + \langle \xi, x - \bar{x} \rangle$. Therefore, this shows that the subdifferential of convex function on convex domain is nonempty.
Proposition 1 For convex function, the subdifferential is convex and closed.

Step 1: To show the subdifferential is convex.
Suppose \( g_1, g_2 \in \partial f(x) \).
By definition, we have
\[
\begin{align*}
  f(y) &\geq f(x) + < g_1, y - x > \\
  f(y) &\geq f(x) + < g_2, y - x > 
\end{align*}
\]
We pick \( \alpha \in (0, 1) \). By multiplying (15) and (16) by \( \alpha \) and \( 1 - \alpha \), respectively, we get
\[
\begin{align*}
  f(y) &= \alpha f(y) + (1 - \alpha) f(y) \geq \alpha f(x) + (1 - \alpha) f(x) + < \alpha g_1, y - x > + \\
        &< (1 - \alpha) g_2, y - x > = f(x) + < \alpha g_1 + (1 - \alpha) g_2, y - x > 
\end{align*}
\]
Thus, it shows that \( \alpha g_1 + (1 - \alpha) g_2 \in \partial f(x) \). Since \( g_1, g_2 \) and \( \alpha \) are arbitrary, \( \partial f(x) \) is convex.

Step 2: To show the subdifferential is closed.
Suppose we have a convergent sequence \( \{ g_k \} \subset \partial f(x) \). Hence, \( f(y) \geq f(x) + < g_k, y - x > \) for all \( k \). Let \( h(g) := f(x) + < g_k, y - x > \) and \( \lim_{k \to \infty} g_k = \bar{g} \). Since \( h(g) \) is continuous function, \( f(y) \geq f(x) + < \bar{g}, y - x > \).

Proposition 2 Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f : S \to \mathbb{R} \) be convex. Suppose that \( f \) is differentiable at \( \bar{x} \in \text{int}S \). Then the subdifferential of \( f \) at \( \bar{x} \) is the singleton set \( \{ \nabla f(\bar{x}) \} \).

Proof Let \( g \) be a subgradient of \( f \) at \( \bar{x} \). By definition, we have
\[
\begin{align*}
  f(\bar{x} + \lambda d) &\geq f(\bar{x}) + \lambda < g, d > 
\end{align*}
\]
By the first order Taylor expansion, we have
\[
\begin{align*}
  f(\bar{x} + \lambda d) &= f(\bar{x}) + \lambda < \nabla f(\bar{x}), d > + o(\lambda ||d||_2) 
\end{align*}
\]
Subtracting the equation from the inequality, we have
\[
0 \geq \lambda < g - \nabla f(x), d > - o(\lambda ||d||_2)
\]
We divide the inequality above by $\lambda$ and let $\lambda \to 0^+$, if follows that $g - \nabla f(\bar{x}), d \geq 0$. Choosing $d = g - \nabla f(\bar{x})$, we will have $||g - \nabla f(\bar{x})||_2^2 \leq 0$, which implies $g = \nabla f(\bar{x})$.

Therefore, it shows that the subdifferential of $f$ at $\bar{x}$ is the singleton set $\{\nabla f(\bar{x})\}$.

The next proposition was not discussed in the class.

**Proposition 3** Let $f : R^n \to R$ be convex function. For every $x \in R^n$, we have $f'(x; y) = \max_{d \in \partial f(x)} d^T y$, $\forall y \in R^n$.

where $f'(x; y) := \lim_{\alpha \to 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$

For the proof, please refer to Bertsekas’s “Convex Analysis and Optimization”. We are going to use Proposition 3 to show that that if $f$ is convex and $\partial f(x) = \{g\}$, then $f$ is differentiable at $x$ and $g = \nabla f(x)$.

**Proposition 4** Let $f : R^n \to R$ be convex function. If $\partial f(x) = \{g\}$, then $f$ is differentiable at $x$ and $g = \nabla f(x)$.

**Proof**

By the definition of directional derivative, we see that $f$ is differentiable at $x$ if and if the directional derivative $f'(x; y) = \nabla f(x)^T y$. According to Proposition 3, we have

$$f'(x; y) = \max_{d \in g} d^T y = g^T y$$  \hspace{1cm} (24)

Since $f'(x; y) = g^T y$ for all $y \in R^n$, we conclude that $f$ is differentiable at $x$.

**Proposition 5** $\partial f(x) = \gamma \partial f(x)$ for $\gamma > 0$.

**Proof**

Pick $g \in \partial f(x)$. By definition,

$$f(y) \geq f(x) + g, y - x$$  \hspace{1cm} (25)

By multiplying both sides by $\gamma$, we have

$$\gamma f(y) \geq \gamma f(x) + \langle \gamma g, y - x \rangle$$  \hspace{1cm} (26)
Hence, it implies that $\gamma g$ is the subgradient of $\gamma f$ at $x$. So it shows that $\gamma \partial f(x) \subset \partial \gamma f(x)$. And the proof of opposite direction is similar.

References