CS 7301 ADVANCED OPTIMIZATION IN ML - ASSIGNMENT 1

Rohith Peddi, RXP190007

May 17, 2021

1

(7 points total) In this assignment, we will compute the Gradient and Hessian for a few supervised learning loss functions. Given training examples $(x_1, y_1), , (x_n, y_n)$ where $x_i \epsilon R^m$ is the feature vectors and y_i is the label

1.1

(1 Point) Compute the Gradient of the Hinge/SVM Loss:

$$L(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}$$

Here $y_i \epsilon \{1, +1\}$

<u>SOLUTION</u>

$$L(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\} \tag{1}$$

$$max\{0, 1 - y_i w^T x_i\} = \begin{cases} 0 & 1 - y_i w^T x_i < 0 \\ 1 - y_i w^T x_i & 1 - y_i w^T x_i \geq 0 \end{cases}$$

Can also be written using indicator function

$$I_{1 - y_i w^T x_i \geq 0} = \begin{cases} 0 & 1 - y_i w^T x_i < 0 \\ 1 & 1 - y_i w^T x_i \geq 0 \end{cases} \tag{2}$$

Gradient of smooth Hinge Loss

$$\nabla_w L(w) = \nabla_w (\sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}) = \sum_{i=1}^{n} -y_i x_i I_{1 - y_i w^T x_i \geq 0} \tag{3}$$

1.2

(2 points) Compute the gradient and hessian of smooth SVM Loss:

$$L(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}^2$$

Here $y_i \epsilon \{1, +1\}$

<u>SOLUTION</u>

$$L(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}^2 \tag{4}$$

Using the same indicator function as above to represent max function (2).

$$max\{0, 1 - y_i w^T x_i\}^2 = \begin{cases} 0 & 1 - y_i w^T x_i < 0 \\ (1 - y_i w^T x_i)^2 & 1 - y_i w^T x_i \geq 0 \end{cases}$$

$$max\{0, 1 - y_i w^T x_i\}^2 = (1 - y_i w^T x_i)^2 I_{1 - y_i w^T x_i \geq 0}$$

Gradient of smooth hinge loss

$$\nabla_w L(w) = \nabla_w (\sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}^2) = \sum_{i=1}^{n} -2 y_i x_i (1 - y_i w^T x_i) I_{1 - y_i w^T x_i \geq 0} \tag{5}$$

Hessian of smooth hinge loss

$$\begin{aligned} \nabla_w^2 L(w) &= \nabla_w^2 (\sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}) \\ &= \sum_{i=1}^{n} 2 y_i^2 x_i x_i^T I_{1 - y_i w^T x_i \geq 0} \end{aligned}$$

As $y_i \epsilon \{-1, +1\} \implies \mathrm{y}_i^2 = 1$

$$\nabla_w^2 L(w) = \sum_{i=1}^{n} 2 x_i x_i^T I_{1 - y_i w^T x_i \geq 0} \tag{6}$$

1.3

(2 points) Compute the gradient and hessian of Least sqaure loss:

$$L(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

Here $y_i \epsilon R$

<u>SOLUTION</u>

$$L(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 \tag{7}$$

Gradient of least square loss

$$\nabla_w L(w) = \nabla_w (\sum_{i=1}^{n} (y_i - w^T x_i)^2) = \sum_{i=1}^{n} -2x_i (y_i - w^T x_i) \tag{8}$$

Hessian of least square loss

$$\nabla_w^2 L(w) = \nabla_w^2 (\sum_{i=1}^{n} (y_i - w^T x_i)^2) = \sum_{i=1}^{n} 2x_i x_i^T \tag{9}$$

1.4

(2 points) Compute the gradient of simple 2 layer function:

$$L(w) = \sum_{i=1}^{n} (y_i - max(0, w^T x_i + b))^2$$

Here $y_i \epsilon R$

<u>SOLUTION</u>

$$L(w) = \sum_{i=1}^{n} (y_i - max(0, w^T x_i + b))^2 \tag{10}$$

max function can again be written using an Indicator function.

$$max\{0, w^T x_i + b\} = \begin{cases} 0 & w^T x_i + b < 0 \\ w^T x_i + b & w^T x_i + b \geq 0 \end{cases}$$

$$max\{0, w^T x_i + b\} = (w^T x_i + b) I_{w^T x_i + b \geq 0}$$

Gradient of simple 2 layer function

$$\nabla_w L(w) = \nabla_w (\sum_{i=1}^{n} (y_i - max(0, w^T x_i + b))^2) = \sum_{i=1}^{n} -2x_i(y_i - max(0, w^T x_i + b)) I_{w^T x_i + b \geq 0} \tag{11}$$

Hessian of simple 2 layer function

$$\nabla_w^2 L(w) = \nabla_w^2 (\sum_{i=1}^{n} (y_i - max(0, w^T x_i + b))^2) = \sum_{i=1}^{n} 2x_i x_i^T I_{w^T x_i + b \geq 0} \tag{12}$$

2

(12 points total) This question will focus on proving the convexity and in some cases, finding the (sub)gradients for gradient descent like optimization.

2.1

(4 points) Define the regularized Hinge/SVM loss as:

$$L_H(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\} + R(w)$$

Here $y_i \epsilon R$ and $R(w)$ is a Norm. Is $L_H(w)$ convex ? Why?

What about smooth regularized SVM Loss:

$$L_S(w) = \sum_{i=1}^{n} max\{0, 1 - y_i w^T x_i\}^2 + R(w)$$

Is $L_S(w)$ convex? Why ?

SOLUTION

A function f is said to be convex if it satisfies

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \forall x, y \epsilon dom(f) \tag{13}$$

Known properties :

- As R(w) is a norm it is convex and satisfies (13)
- Sum of two convex functions is convex

Lets check the convexity of $L_H(w)$:
$L_H(w)$ satisfies the (13) condition if it is convex i,e

$$L_H(\lambda w_1 + (1-\lambda)w_2) \leq \lambda L_H(w_1) + (1-\lambda) L_H(w_2), \forall w_1, w_2 \epsilon dom(L_H)$$

Considering $L_H(w1)$ and $L_H(w2)$

$$\lambda L_H(w_1) + (1-\lambda)L_H(w_2) = \lambda(\sum_{i=1}^{n} max\{0, 1 - y_i w_1^T x_i\} + R(w_1)) + (1-\lambda)(\sum_{i=1}^{n} max\{0, 1 - y_i w_2^T x_i\} + R(w_2))$$

$$= \lambda R(w_1) + (1-\lambda)R(w_2) + \lambda(\sum_{i=1}^{n} max\{0, 1 - y_i w_1^T x_i\}) + (1-\lambda)(\sum_{i=1}^{n} max\{0, 1 - y_i w_2^T x_i\})$$

$$= \lambda R(w_1) + (1-\lambda)R(w_2) + \sum_{i=1}^{n} max\{0, \lambda(1 - y_i w_1^T x_i)\} + \sum_{i=1}^{n} max\{0, (1-\lambda)(1 - y_i w_2^T x_i)\}$$

$$= \lambda R(w_1) + (1-\lambda)R(w_2) + \sum_{i=1}^{n} max\{0, \lambda(1 - y_i w_1^T x_i)\} + max\{0, (1-\lambda)(1 - y_i w_2^T x_i)\}$$

$$\geq R(\lambda w_1 + (1-\lambda)w_2) + \sum_{i=1}^{n} max\{0, \lambda(1 - y_i w_1^T x_i) + (1-\lambda)(1 - y_i w_2^T x_i)\}$$

$$\geq R(\lambda w_1 + (1-\lambda)w_2) + \sum_{i=1}^{n} max\{0, \lambda - \lambda y_i w_1^T x_i + 1 - \lambda - (1-\lambda) y_i w_2^T x_i\}$$

$$\geq R(\lambda w_1 + (1-\lambda)w_2) + \sum_{i=1}^{n} max\{0, 1 - y_i(\lambda w_1 + (1-\lambda)w_2)x_i\}$$

Properties used :

$$max\{0, \lambda f(w_1)\} + max\{0, (1-\lambda)f(w_2)\} \geq max\{0, \lambda f(w_1) + (1-\lambda)f(w_2)\}$$
$$\lambda R(w_1) + (1-\lambda)R(w_2) \geq R(\lambda w_1 + (1-\lambda)w_2)$$

Thus $L_H(w)$ satisfies the convexity property and can be called as convex function.

Now consider $L_S(w)$ smooth regularized SVM loss
Known properties :

1. As R(w) is norm, it is convex and satisfies (13)
2. Sum of convex functions is convex
3. Sum of psd matrices is a psd matrix
4. $\nabla_x^2 f(x) \geq 0 \implies$ f is convex

From Q1, we have Hessian for smooth SVM loss as (6)

$$\nabla_w^2 L_S(w) = \sum_{i=1}^{n} 2x_i x_i^T I_{1-y_i w^T x_i \geq 0} + \nabla_w^2 R(w)$$

Observations :

1. $x_i x_i^T$ is a psd matrix and $x_i x_i^T \geq 0$
2. As R(w) is norm and is convex, it satisfies $\nabla_w^2 R(w) \geq 0$

Thus,

$$\nabla_w^2 L_S(w) \geq 0 \tag{14}$$

So, smooth regularized SVM Loss is convex.

2.2

(2 points) Consider a 2 Layer function:

$$L(w_1, w_2, b) = \sum_{i=1}^{n} (y_i - w_1 max(0, w_2^T x_i + b))^2 + R(w)$$

Here $y_i \epsilon R$ and $R(w)$ is a Norm. Is $L(w_1, w_2, b)$ convex ? Why?

SOLUTION

$L(w_1, w_2, b)$ is not a convex function.

Observation :

1. We can prove that expression is non convex, if we can find a certificate which violates the convexity.
2. $L(w_1, w_2, b)$ involves product of parameters w_1, w_2, such expressions involving product of terms tend to be non convex.

Let us consider the certificate

$$n = 1, \lambda = 0.5, w_{11} = 1, w_{12} = 0.5, w_{21} = 2, w_{22} = 4, y_1 = 0, b_1 = 0, b_2 = 0, x_1 = 1$$

We need to show that it violates convexity property i,e

$$\lambda L(w_{11}, w_{21}, b_1) + (1-\lambda)L(w_{12}, w_{22}, b_2) < L(\lambda w_{11} + (1-\lambda)w_{12}, \lambda w_{21} + (1-\lambda)w_{22}, \lambda b_1 + (1-\lambda)b$$

$$0.5 * (0 - max\{0, 2\})^2 + 0.5 * (0 - 0.5 * max\{0, 4\})^2 < (0 - 0.75 * max\{0, 3\})^2$$

$$2 + 2 < 5.0625$$

As we found a certificate which violates the convexity property.
The above function is non convex in nature.

2.3

(4 points) Recall the logistic loss:

$$L_{log}(w) = \sum_{i=1}^{n} log(1 + exp(-y_i w^T x_i))$$

Denote P as a polyhedron and define a Polyhedral regularization as $R(w) = f_p(|w|)$ where $f_p(w) = max_{y \epsilon P} y^T w$.

Is the function $L(w) = \sum_{i=1}^{n} log(1 + exp(-y_i w^T x_i)) + R(w)$ convex ? Why ? Compute the gradient of L(w).

SOLUTION

Known properties:

1. Sum of convex functions is convex

First let us check for convexity of Logistic loss.
Inorder to check the convexity we compute Hessian.

$$g(w) = \sum_{i=1}^{n} log(1 + exp(-y_i w^T x_i))$$

$$\nabla_w g(w) = \sum_{i=1}^{n} \frac{-y_i x_i exp(-y_i w^T x_i)}{(1 + exp(-y_i w^T x_i)}$$

$$\nabla_w g(w) = \sum_{i=1}^{n} \frac{-y_i x_i}{(1 + exp(y_i w^T x_i)}$$

$$\nabla_w^2 g(w) = \sum_{i=1}^{n} \frac{exp(y_i w^T x_i)}{(1 + exp(y_i w^T x_i))^2} y_i^2 x_i x_i^T$$

As the hessian involves $x_i x_i^T$ which is a positive semi-definite matrix.
And since sum of psd matrices is a psd matrix. $\nabla_w^2 g(w) \geq 0$
Logistic loss term is convex in nature.
Now, consider the regularizer term.

$$R(w) = f_P(|w|) = max_{y \epsilon P} y^T |w|$$

$f_P(w) = max_{y \epsilon P} y^T w$

Observation

1. $f_P(w)$ it is considering point-wise max over linear functions. So it is convex.

2. Following the slides of conditional gradient descent dual norm is defined as $||x||_* = \max_{||z||<1} z^T x$

3. $L_\infty-$norm is the dual norm of L_1- norm

In order to find the gradient, we need a specific polyhedron first.

$$\nabla_w L_{log}(w) = \sum_{i=1}^{n} \frac{-y_i x_i}{(1 + exp(y_i w^T x_i)} + \nabla_w R(w)$$

Considering polyhedron as $L_1 - norm$, and its dual norm shall be $L_\infty - norm$

$$\implies f_P(w) = ||w||_\infty = \max_{||y||<1} y^T w$$

As partial derivative of $L_\infty norm$ is

$$\frac{\partial}{\partial w_i} ||w||_\infty = sign(w_j)\delta_{kj}$$ where δ_{kj} is the Kronecker delta function

2.4

(2 points) Softmax estimator for contextual bandits:

$$SM(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{exp(\theta^T x_i^{a_i})}{\sum_{j=1}^{k} exp(\theta^T x_i^j)}$$

convex ? Why ?

SOLUTION

It is not convex function. Can be shown by considering the hessian

$$\nabla_\theta SM(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{x_i^{a_i} \alpha exp(\theta^T x_i^{a_i}) - exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)}{\alpha^2}$$

$$\nabla_\theta SM(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{x_i^{a_i} exp(\theta^T x_i^{a_i})}{\alpha} - \frac{exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)}{\alpha^2}$$

$$\nabla_\theta^2 SM(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{\alpha exp(\theta^T x_i^{a_i})(x_i^{a_i})^T (x_i^{a_i}) - (x_i^{a_i})^T \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)}{\alpha^2}$$
$$- \frac{\alpha^2 (exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} (x_i^j)(x_i^j)^T exp(\theta^T x_i^j) + x_i^{a_i} exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)) - exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k}}{\alpha^4}$$

$$\nabla_\theta^2 SM(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{exp(\theta^T x_i^{a_i})(x_i^{a_i})^T (x_i^{a_i})}{\alpha} - \frac{(x_i^{a_i})^T \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)}{\alpha^2}$$
$$- \frac{(exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} (x_i^j)(x_i^j)^T exp(\theta^T x_i^j) + x_i^{a_i} exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j))}{\alpha^2}$$
$$- \frac{exp(\theta^T x_i^{a_i}) \sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j)(2 \sum_{j=1}^{k} exp(\theta^T x_i^j))(\sum_{j=1}^{k} x_i^j exp(\theta^T x_i^j))}{\alpha^4}$$

Here the hessian consists of summation of non psd matrices. As it involves terms like $(x_i^{a_i})(x_i^j)^T$ which need not be psd.

So, the function is non-convex in nature as positive semi definiteness is not guaranteed during summation.

3

(8 points) Recall that Prox operator of a function h is

$$prox_h(z) = argmin_x \frac{1}{2t}||x-z||^2 + h(x)$$

Compute the Prox operator for the following functions. Assume $\lambda > 0$ wherever applicable.

<u>SOLUTION</u>

3.1

$$g_1(x) = 0, \text{ if } x \neq 0 \text{ and } -\lambda, \text{ if } x = 0$$

Above equation can be written as follows

$$g_1(x) = \begin{cases} 0 & x \neq 0 \\ -\lambda & x = 0 \end{cases}$$

$$prox_{g_1}(z) = argmin_x(\frac{1}{2t}||x-z||^2 + g_1(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x-z||^2 & x \neq 0 \\ -\lambda + \frac{||z||^2}{2t} & x = 0 \end{cases}$$

$$\text{clearly } \frac{1}{2t}||x-z||^2 \geq 0, \forall t > 0$$

Minimum is obtained when $x = z$

$$min_x f(x) = min(0, -\lambda + \frac{||z||^2}{2t}), \text{and is obtained for } argmin_x f(x) = (z, 0)$$

$$min_x f(x) = \begin{cases} 0 & -\lambda + \frac{||z||^2}{2t} > 0, argmin_x f(x) = z \\ -\lambda + \frac{||z||^2}{2t} & -\lambda + \frac{||z||^2}{2t} < 0, argmin_x f(x) = 0 \\ 0 & -\lambda + \frac{||z||^2}{2t} = 0, argmin_x f(x) = 0, z \end{cases}$$

$$prox_{g_1}(z) = \begin{cases} z & ||z||^2 > 2\lambda t \\ 0 & ||z||^2 < 2\lambda t \\ 0, z & ||z||^2 = 2\lambda t \end{cases}$$

3.2

$$g_2(x) = 0, \text{if } x \neq 0 \text{ and } \lambda, \text{ if } x = 0$$

Above equation can be written as follows

$$g_2(x) = \begin{cases} 0 & x \neq 0 \\ \lambda & x = 0 \end{cases}$$

$$prox_{g_2}(z) = argmin_x(\frac{1}{2t}||x-z||^2 + g_2(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x-z||^2 & x \neq 0 \\ \lambda + \frac{||z||^2}{2t} & x = 0 \end{cases}$$

$$\text{clearly } \frac{1}{2t}||x-z||^2 \geq 0, \forall t > 0$$

Minimum is obtained when $x = z$

$$min_x f(x) = min(0, \lambda + \frac{||z||^2}{2t}), \text{and is obtained for } argmin_x f(x) = (z, 0)$$

As we know $\lambda > 0 \implies \lambda + \frac{||z||^2}{2t} > 0$

$$min_x f(x) = 0, \text{and is obtained for } argmin_x f(x) = z$$

$$prox_{g_2}(z) = z \tag{15}$$

3.3

$$g_3(x) = \lambda x^3, \text{if } x \geq 0 \text{ and } \infty \text{ otherwise}$$

Above equation can be written as follows

$$g_3(x) = \begin{cases} \lambda x^3 & x \geq 0 \\ \infty & x < 0 \end{cases}$$

$$prox_{g_3}(z) = argmin_x(\frac{1}{2t}||x-z||^2 + g_3(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x-z||^2 + \lambda x^3 & x \geq 0 \\ \infty & x < 0 \end{cases}$$

Clearly minimum for $f(x)$ can only be obtained in the region $x \geq 0$
In order to find the minimum we differentiate $f(x)$ w.r.t x

$$\begin{aligned} \nabla_x f(x) &= \nabla_x(\frac{1}{2t}||x-z||^2 + \lambda x^3) \\ &= \frac{x-z}{t} + 3\lambda x^2, \forall x \geq 0 \end{aligned}$$

$$\text{Solving for x, } \nabla_x f(x) = 0, x \geq 0, \text{we get}$$

$$\begin{aligned} \frac{x-z}{t} + 3\lambda x^2 = 0 \\ 3\lambda t x^2 + x - z = 0 \\ x = \frac{-1+\sqrt{1+12\lambda zt}}{6\lambda t}, \frac{-1-\sqrt{1+12\lambda zt}}{6\lambda t} \end{aligned}$$

$$\frac{-1-\sqrt{1+12\lambda zt}}{6\lambda t} < 0 \implies x \neq \frac{-1-\sqrt{1+12\lambda zt}}{6\lambda t}$$

$$\begin{aligned} x = \frac{-1+\sqrt{1+12\lambda zt}}{6\lambda t} &\geq 0 \\ \implies -1+\sqrt{1+12\lambda zt} &\geq 0 \\ \implies 1+12\lambda zt &\geq 1 \\ \implies z &\geq 0 \end{aligned}$$

Clearly, for $z < 0$, $f(x)$ has minimum at x = 0.

$$prox_{g_3}(z) = \begin{cases} \frac{-1+\sqrt{1+12\lambda zt}}{6\lambda t} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{16}$$

3.4

$$g_4(x) = 0, \text{if } 0 \leq x \leq \lambda \text{ and } \infty \text{ otherwise}$$

Above equation can be written as follows

$$g_4(x) = \begin{cases} 0 & 0 \leq x \leq \lambda \\ \infty & otherwise \end{cases}$$

$$prox_{g_4}(z) = argmin_x(\frac{1}{2t}||x - z||^2 + g_4(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x - z||^2 & 0 \leq x \leq \lambda \\ \infty & otherwise \end{cases}$$

We only have to consider the region, $0 \leq x \leq \lambda$ for finding the min.
Inorder to find $argmin_x f(x)$ we differentiate and check for critical points

$$\nabla_x f(x) = \nabla_x(\frac{1}{2t}||x - z||^2) = \frac{x - z}{t} = 0 \implies x = z \text{ and } 0 \leq x \leq \lambda$$

So,

$$prox_{g_4}(z) = argmin_x f(x) = \begin{cases} 0 & z < 0 \text{ as, f is increasing in } x\epsilon[0, \lambda] \\ z & 0 \leq z \leq \lambda \\ \lambda & z > \lambda \text{ as, f is decreasing in } x\epsilon[0, \lambda] \end{cases} \tag{17}$$

3.5

$$g_5(x) = -logx, \text{if } x > 0 \text{ and } \infty \text{ otherwise}$$

Above equation can be written as follows

$$g_5(x) = \begin{cases} -logx & x > 0 \\ \infty & otherwise \end{cases}$$

$$prox_{g_5}(z) = argmin_x(\frac{1}{2t}||x - z||^2 + g_5(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x - z||^2 - logx & x > 0 \\ \infty & otherwise \end{cases}$$

We only have to consider the region, $x > 0$ for finding the min.
Inorder to find $argmin_x f(x)$ we differentiate and check for critical points

$$\nabla_x f(x) = \nabla_x(\frac{1}{2t}||x - z||^2 - logx) = \frac{x - z}{t} - \frac{1}{x} = 0 \implies x^2 - zx - t = 0$$

$$x = \frac{z + \sqrt{z^2 + 4t}}{2}, \frac{z - \sqrt{z^2 + 4t}}{2}$$

as $z - \sqrt{z^2 + 4t} < 0 \implies x = \dfrac{z + \sqrt{z^2 + 4t}}{2}$ is the only root $\forall z \epsilon R$ s.t $x > 0$.

So,

$$prox_{g_5}(z) = \frac{z + \sqrt{z^2 + 4t}}{2} \tag{18}$$

3.6

$$g_6(x) = \lambda|x|$$

Above equation can be written as follows

$$g_6(x) = \begin{cases} -\lambda x & x < 0 \\ 0 & x = 0 \\ \lambda x & x > 0 \end{cases}$$

$$prox_{g_6}(z) = argmin_x(\frac{1}{2t}||x - z||^2 + g_6(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x - z||^2 - \lambda x & x < 0 \\ \frac{1}{2t}||z||^2 & x = 0 \\ \frac{1}{2t}||x - z||^2 + \lambda x & x > 0 \end{cases}$$

Inorder to find $argmin_x f(x)$ we differentiate and check for critical points

$$\nabla_x f(x) = \nabla_x \begin{cases} \frac{1}{2t}||x - z||^2 - \lambda x & x < 0 \\ \frac{1}{2t}||z||^2 & x = 0 \\ \frac{1}{2t}||x - z||^2 + \lambda x & x > 0 \end{cases}$$

$$= \begin{cases} \frac{x-z}{t} - \lambda & x < 0 \\ \frac{-z}{t} + [-\lambda, \lambda] & x = 0 \\ \frac{x-z}{t} + \lambda & x > 0 \end{cases}$$

$$\implies \begin{cases} x = \lambda t + z & x < 0 \\ z = \lambda t, -\lambda t & x = 0 \text{ Considering sub gradients} \\ x = \lambda t - z & x > 0 \end{cases}$$

$$prox_{g_6}(z) = argmin_x f(x) \qquad = \begin{cases} \lambda t + z & z < -\lambda t \\ 0 & z \\ \lambda t - z & x > 0 \end{cases}$$

3.7

$$g_7(x) = a^T x + b$$

$$\begin{aligned}
prox_{g_7}(z) &= argmin_x(\frac{1}{2t}||x - z||^2 + g_7(x)) \\
\text{Consider } f(x) &= \frac{1}{2t}||x - z||^2 + a^T x + b \\
\nabla_x f(x) &= \nabla_x(\frac{1}{2t}||x - z||^2 + a^T x + b) \\
&\implies \frac{x - z}{t} + a = 0 \\
&\implies x = z - ta
\end{aligned}$$

$$prox_{g_7}(z) = z - ta \tag{19}$$

3.8

$$g_8(x) = \lambda|x|^3$$

Above equation can be written as follows

$$g_6(x) = \begin{cases} -\lambda x^3 & x < 0 \\ 0 & x = 0 \\ \lambda x^3 & x > 0 \end{cases}$$

$$prox_{g_8}(z) = argmin_x(\frac{1}{2t}||x-z||^2 + g_8(x))$$

$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x-z||^2 - \lambda x^3 & x < 0 \\ \frac{1}{2t}||z||^2 & x = 0 \\ \frac{1}{2t}||x-z||^2 + \lambda x^3 & x > 0 \end{cases}$$

Inorder to find $argmin_x f(x)$ we differentiate and check for critical points

$$\nabla_x f(x) = \nabla_x(\frac{1}{2t}||x-z||^2 + g_8(x))$$
$$= \begin{cases} \frac{x-z}{t} - 3\lambda x^2 & x < 0 \\ \frac{-z}{t} & x = 0 \\ \frac{x-z}{t} + 3\lambda x^2 & x > 0 \end{cases}$$

Equations obtained are as follows

$$\begin{cases} x - z - 3t\lambda x^2 = 0 & x < 0 \\ z = 0 & x = 0 \\ x - z + 3t\lambda x^2 = 0 & x > 0 \end{cases}$$
$$\implies \begin{cases} x = \frac{1 \pm \sqrt{1-12\lambda tz}}{6t\lambda} & x < 0 \\ z = 0 & x = 0 \\ x = \frac{-1 \pm \sqrt{1+12\lambda tz}}{6t\lambda} & x > 0 \end{cases}$$
$$\implies \begin{cases} \text{for } x < 0 & x = \frac{1-\sqrt{1-12\lambda tz}}{6t\lambda} \text{ possible root with } z < 0 \\ \text{for } x = 0 & z = 0 \\ \text{for } x > 0 & x = \frac{-1+\sqrt{1+12\lambda tz}}{6t\lambda} \text{ possible root with } z > 0 \end{cases}$$

$$prox_{g_8}(z) = \begin{cases} \frac{1-\sqrt{1-12\lambda tz}}{6t\lambda} & z < 0 \\ 0 & z = 0 \\ \frac{-1+\sqrt{1+12\lambda tz}}{6t\lambda} & z > 0 \end{cases} \tag{20}$$

4

(4 points) Compute the Projection Operator

$$P_C(z) = prox_{I_C}(z) = argmin_x \frac{1}{2t}||x-z||^2 + I_C(x)$$

for the following constraints:

SOLUTION

4.1

$$C = \{x\epsilon R^n : x \geq 0\}$$

$$I_C(x) = \begin{cases} 0 & x \geq 0 \\ \infty & x < 0 \end{cases}$$

$$P_C(z) = prox_{I_C}(z) = argmin_x(\frac{1}{2t}||x-z||^2 + I_C(x))$$
$$\text{Consider } f(x) = \begin{cases} \frac{1}{2t}||x-z||^2 & x \geq 0 \\ \infty & x < 0 \end{cases}$$

Clearly we have to consider only the region $x \geq 0$ in order to find $min_x f(x)$

$$\begin{aligned} \nabla_x f(x) &= \nabla_x(\frac{1}{2t}||x-z||^2) \text{ and } x \geq 0 \\ &= \frac{x-z}{t} = 0 \\ &\implies x = z \\ &\text{as } x \geq 0 \text{ and } x = z \implies z \geq 0 \end{aligned}$$

$$P_C(z) = prox_{I_C}(z) = \begin{cases} z & z \geq 0 \\ 0 & z < 0 \text{ ,As } z < 0, \ \frac{1}{2t}||x-z||^2 \text{ increases continuously } \implies argmin_x f(x) = 0 \end{cases}$$

$$P_C(z) = prox_{I_C}(z) = \begin{cases} z & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{21}$$

4.2

$$C = \{x \epsilon R^n : ||x - c|| \leq R\}$$

$$P_C(z) = prox_{I_C}(z) = argmin_x \frac{1}{2t}||x - z||^2 + I_C(x)$$

Constraint can also be written as

$$C = \{x \epsilon R^n : ||x - c||_2^2 \leq R^2\}$$

as $||x - c|| > 0$
We construct the Lagrangian as follows:

$$g(x, \lambda) = \frac{1}{2t}||x - z||^2 + \lambda(||x - c||^2 - R^2)$$

Two cases arise based on z

1. $z \epsilon C \implies$ Constraints are not active
2. z is outside $C \implies$ Constraints are active

Optimality conditions $\nabla_x g = 0$ and $\nabla_\lambda g = 0$

$$\nabla_\lambda g = 0 \implies ||x - c||^2 - R^2 = 0 \implies ||x - c||^2 = R^2 \tag{22}$$

$$\nabla_x g = 0 \implies \frac{x - z}{t} + 2\lambda(x - c) = 0 \tag{23}$$

As we have $(x - c)^T(x - c) = R^2$, we multiply $(x - c)^T$ on both sides

$$\begin{aligned} (x - c)^T(\frac{x - z}{t} + 2\lambda(x - c)) = 0 \\ \frac{(x - c)^T(x - z)}{t} + 2\lambda(x - c)^T(x - c) = 0 \\ \frac{(x - c)^T(x - z)}{t} + 2\lambda R^2 = 0 \end{aligned}$$

$$\lambda = \frac{(x - c)^T(z - x)}{2R^2 t} \tag{24}$$

Substituting this in (23) we get

$$
\begin{aligned}
\frac{x-z}{t} + 2\frac{(x-c)^T(z-x)}{2R^2t}(x-c) &= 0 \\
\frac{(x-c)-(z-c)}{t} + 2\frac{(x-c)^T((z-c)-(x-c))}{2R^2t}(x-c) &= 0 \\
\frac{(x-c)-(z-c)}{t} + 2\frac{(x-c)^T(z-c)-(x-c)^T(x-c)}{2R^2t}(x-c) &= 0 \\
\frac{(x-c)-(z-c)}{t} + \frac{(x-c)^T(z-c)-R^2}{R^2t}(x-c) &= 0 \\
\frac{(x-c)-(z-c)}{t} + \frac{(x-c)^T(z-c)(x-c)}{R^2t} - \frac{(x-c)}{t} &= 0 \\
\frac{(x-c)^T(z-c)(x-c)}{R^2t} &= \frac{(z-c)}{t} \\
(x-c)^T(z-c)(x-c) &= R^2(z-c) \\
\underline{(x-c)^T(z-c)}(x-c) &= \underline{R^2}(z-c) \text{ Underlined parts are not vectors}
\end{aligned}
$$

$\implies$ (x-c) and (z-c) should be along same direction

$$x-c = \theta(z-c)$$

Substituting back we get

$$
\begin{aligned}
\theta^2(z-c)^T(z-c)(z-c) &= R^2(z-c) \\
\theta^2||z-c||^2(z-c) &= R^2(z-c) \\
\theta &= \frac{R}{||z-c||}
\end{aligned}
$$

$$x = c + \theta(z-c) \implies x = c + \frac{R}{||z-c||}(z-c) \tag{25}$$

Above holds for the second case when $||z-c|| > R$. For the case when $||z-c|| \leq R$ Minimum value occurs at x = z. Therefore

$$P_C(z) = \begin{cases} z & ||z-c|| \leq R \\ c + \frac{R}{||z-c||}(z-c) & ||z-c|| > R \end{cases} \tag{26}$$

4.3

$$C = \{x \epsilon R^n : a^T x \geq b\}$$

We construct the Lagrangian as follows:

$$g(x, \lambda) = \frac{1}{2t}||x - z||^2 + \lambda(b - a^T x)$$

Two cases arise based on z

1. $z \epsilon C \implies$ Constraints are not active
2. z is outside $C \implies$ Constraints are active

Optimality conditions $\nabla_x g = 0$ and $\nabla_\lambda g = 0$

$$\nabla_\lambda g = 0 \implies b - a^T x = 0 \implies b = a^T x \tag{27}$$

$$\nabla_x g = 0 \implies \frac{x - z}{t} - \lambda a = 0 \tag{28}$$

Multiplying by a^T on both sides of (28) gives us

$$\begin{aligned} a^T(\frac{x - z}{t} - \lambda a) &= 0 \\ \frac{a^T x - a^T z}{t} - \lambda a^T a &= 0 \\ \frac{b - a^T z}{t} &= \lambda a^T a \\ \implies \lambda &= \frac{b - a^T z}{t a^T a} \end{aligned}$$

Substituting this back in (28) gives us

$$\begin{aligned} \frac{x - z}{t} - \lambda a &= 0 \\ \frac{x - z}{t} - \frac{b - a^T z}{t a^T a} a &= 0 \\ x &= z + \frac{b - a^T z}{t a^T a} a \end{aligned}$$

Above holds for the second case when $a^T z < b$. For the case when $a^T z > b$ Minimum value occurs at x = z. Therefore

$$P_C(z) = \begin{cases} z & a^T z \geq b \\ z + \frac{b - a^T z}{t a^T a} a & a^T z < b \end{cases} \tag{29}$$

4.4

$$C = \{x\epsilon R^n : ||x||_1 \leq R\}$$

We construct the Lagrangian as follows:

$$g(x, \lambda) = \frac{1}{2t}||x - z||^2 + \lambda(||x||_1 - R)$$

Two cases arise based on z

1. $z\epsilon C \implies$ Constraints are not active
2. z is outside $C \implies$ Constraints are active

From case - 1, when $z\epsilon C$ then it satisfies $||z||_1 \leq R$
In this case minimum occurs when $x = z$
Optimality conditions $\nabla_x g = 0$ and $\nabla_\lambda g = 0$

$$\nabla_\lambda g = 0 \implies ||x||_1 - R \implies ||x||_1 = R \tag{30}$$

$$\nabla_{x_i} g = 0 \implies \frac{x_i - z_i}{t} + \lambda sign(x_i) = 0 \tag{31}$$

From (31), as we know $\lambda > 0$

$$\lambda = \frac{x_i - z_i}{tsign(x_1)}$$

$$\implies \begin{cases} x_i < z_i & x_i \geq 0 \\ x_i > z_i & x_i < 0 \end{cases}$$

From both cases, it is evident that$|z_i| > |x_i|$
$\implies ||z||_1 > ||x||_1 = R$

$x_i = z_i - \lambda t sign(x_i)$

$$x_i = \begin{cases} z_i - \lambda t & x_i > 0 \\ 0 & x_i = 0 \text{ We need to consider sub gradient here} \\ z_i + \lambda t & x_i < 0 \end{cases}$$

$$x_i = \begin{cases} z_i - \lambda t & x_i > 0 \implies z > \lambda t \\ 0 & x_i = 0 \implies z \epsilon [\lambda t, \lambda t] \text{ Considering sub gradient} \\ z_i + \lambda t & x_i < 0 \implies z < -\lambda t \end{cases}$$

$x_i = sign(z_i)[|z_i| - \lambda t]_+$ where $(|z_i| - \lambda t)_+$ is the max operator

As we have $||x||_1 = R$

$$\implies \sum_{i=1}^{n} |x_i| = R$$

$$\implies \sum_{i=1}^{n} |sign(z_i)[|z_i| - \lambda t]_+| = R$$

$$\implies \sum_{i=1}^{n} [|z_i| - \lambda t]_+ = R$$

$$P_C(z) = \begin{cases} z & ||z||_1 \leq R \\ sign(z_i)[|z_i| - \lambda t]_+ & ||z||_1 > R \text{ and } \lambda \text{ satisfies } \sum_{i=1}^{n} [|z_i| - \lambda t]_+ = R \end{cases} \tag{32}$$

5

(3 Points) Implement numerically correct versions of the following functions:

SOLUTION

5.1

$L(x) = log(1 + exp(-x))$

There is a potential risk in evaluating large negative, positive values.
As we know $\lim_{y \to 0} log(1 + y) \approx y$.
So for $x >> 0$ we can write it as $exp(-x)$
So, this can be written as follows

$$f(x) = \begin{cases} x & x < -35 \\ exp(x) & x > 10 \\ log(1 + exp(-x)) & otherwise \end{cases}$$

$$f(x) = \begin{cases} x & x < -35 \\ exp(x) & x > 10 \\ logaddexp(0, -x) & otherwise \end{cases}$$

5.2

$L(x) = log(exp(x_1) + exp(x_2))$

$$L(x_1, x_2) = log(exp(x_1)(1 + \frac{exp(x_1)}{exp(x_2)}))$$

$$\implies log(exp(x_1)) + log((1 + \frac{exp(x_1)}{exp(x_2)}))$$

$$\implies x_1 + log(1 + exp(x_2 - x_1))$$

Now, consider $(x_2 - x_1)$ as y and apply the same from 5.1

$$\implies \begin{cases} x_1 + y & y > 35 \\ x_1 + exp(y) & y < -10 \\ x_1 + log(1 + exp(y)) & otherwise \end{cases}$$

$$\implies \begin{cases} x_2 & x_2 - x_1 > 35 \\ x_1 + exp(x_2 - x_1) & x_2 - x_1 < -10 \\ x_1 + log(1 + exp(x_2 - x_1)) & otherwise \end{cases}$$

5.3

$L(x) = \frac{exp(x_1)}{exp(x_1)+exp(x_2)}$

$$
\begin{aligned}
L(x) &= \frac{exp(x_1)}{exp(x_1) + exp(x_2)} \\
&\implies \frac{1}{1 + exp(x_2 - x_1)} \\
&\implies (1 + exp(x_2 - x_1))^{-1} \\
&\implies \begin{cases} 1 - exp(x_2 - x_1) & (x_2 - x_1) < 0 \\ (1 + exp(x_2 - x_1))^{-1} & 0 < (x_2 - x_1) < 35 \\ exp(-(x_2 - x_1)) & otherwise \end{cases}
\end{aligned}
$$

6 PROGRAMMING ASSIGNMENTS

Please find the assignment in the following colab link

https://colab.research.google.com/drive/11wpP$_N 3DP1T22SyxdAZnKi0J - 5BBdopK?usp = sharing$