HW2 Handwritten Assignment Solution

Lecturor: Pei-Yuan Wu
TAs: Yuan-Chia Chang(Problem 5), Chun-Lin Huang(Problem 1, 2, 3, 4)

November 2023, First Edition

Problem 1 (Convolution)(0.5%)

As mentioned in class, image size may change after convolution layers. Consider a batch of image data with shape $(B, W, H, input\_channels)$, how will the shape change after the following convolution layer:

$$Conv2D\,(input\_channels, output\_channels, kernel\_size = (k_1, k_2), stride = (s_1, s_2), padding = (p_1, p_2))$$

For simplicity, the padding tuple means that p_1 pixels are padded on both left and right sides, and p_2 pixels are padded on both top and bottom sides.
Solution: The output is $(B, W', H', output\_channels)$, where

$$W' = \left\lfloor \frac{W + 2 * p_1 - k_1}{s_1} + 1 \right\rfloor$$
$$H' = \left\lfloor \frac{H + 2 * p_2 - k_2}{s_2} + 1 \right\rfloor$$

Note that you should give some explanation to get all points.

Problem 2 (Batch Normalization)(1%)

Batch normalization [1] is a popular trick for training deep networks nowadays, which aims to preserve the distribution within hidden layers and avoids vanishing gradient issue. The alogrithm can be written as below:

Algorithm 1 Batch Normalization

Input Feature from data points over a mini-batch $B = (x_i)_{i=1}^m$
Output $y_i = BN_{\gamma,\beta}(x_i)$

procedure BatchNormalize(B, γ, β)
  $\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^m x_i$ ▷ mini-batch mean
  $\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^m (x_i - \mu_B)^2$ ▷ mini-batch variance
  for $i \leftarrow 1$ to m **do**
    $\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$ ▷ normalize
    $y_i \leftarrow \gamma \hat{x}_i + \beta$ ▷ scale and shift
  end for
  return
end procedure

During training we need to backpropagate the gradient of loss ℓ through this transformation, as well as compute the gradients with respect to the parameters γ, β. Towards this end, please write down the close form expressions for $\frac{\partial \ell}{\partial x_i}$, $\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$ in terms of x_i, μ_B, σ_B^2, $\hat{x}_i$, y_i (given by the forward pass) and $\frac{\partial \ell}{\partial y_i}$ (given by the backward pass).

- Hint: You may first write down the close form expressions of $\frac{\partial \ell}{\partial \hat{x}_i}$, $\frac{\partial \ell}{\partial \sigma_B^2}$, $\frac{\partial \ell}{\partial \mu_B}$, and then use them to compute $\frac{\partial \ell}{\partial x_i}$, $\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$.

Solution: We use the gradient descent to update γ and β:

$$\gamma \leftarrow \gamma - \eta \frac{\partial l}{\partial \gamma}$$
$$\beta \leftarrow \beta - \eta \frac{\partial l}{\partial \beta}$$

where η is a learning rate and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^m \left(\frac{\partial l}{\partial y_i} \frac{\partial y_i}{\partial \gamma} \right) = \sum_{i=1}^m \frac{\partial l}{\partial y_i} \hat{x_i}$$
$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^m \left(\frac{\partial l}{\partial y_i} \frac{\partial y_i}{\partial \beta} \right) = \sum_{i=1}^m \frac{\partial l}{\partial y_i}$$

Note that we sum from 1 to m because we are working with mini-batches. Now,

we derive some important terms by chain rule:

$$\frac{\partial l}{\partial \hat{x_i}} = \frac{\partial l}{\partial y_i}\frac{\partial y_i}{\partial \hat{x_i}} = \frac{\partial l}{\partial y_i}\gamma$$
$$\frac{\partial l}{\partial \sigma_B^2} = \frac{\partial l}{\partial \hat{x_i}}\frac{\partial \hat{x_i}}{\partial \sigma_B^2} = -\frac{1}{2}\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}$$

$$\begin{aligned}\frac{\partial l}{\partial \mu_B} &= \frac{\partial l}{\partial \hat{x_i}}\frac{\partial \hat{x_i}}{\partial \mu_B} \\ &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\frac{\partial}{\partial \mu_B}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{1}{2}} \\ &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\left[\frac{\partial (x_i-\mu_B)}{\partial \mu_B}(\sigma_B^2+\epsilon)^{-\frac{1}{2}} + (x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B}\right] \\ &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\left[\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + (x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B}\right]\end{aligned}$$

We know $\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2$, so the second term in bracket is

$$\begin{aligned}(x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B} &= (x_i-\mu_B)\frac{-1}{2}(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\frac{\partial(\sigma_B^2+\epsilon)}{\partial \mu_B} \\ &= \frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\frac{\partial\left(\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2+\epsilon\right)}{\partial \mu_B} \\ &= \frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right)\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial l}{\partial \mu_B} &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \underbrace{\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}}_{\frac{\partial l}{\partial \sigma_B^2}}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right) \\ &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x_i}}\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \frac{\partial l}{\partial \sigma_B^2}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right)\end{aligned}$$

To derive $\frac{\partial l}{\partial x_i}$, we use the chain rule $\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x_i}}\frac{\partial \hat{x_i}}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_i} + \frac{\partial l}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i}$. Now, calculate the remaining term:

$$\frac{\partial \hat{x_i}}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2+\epsilon}}$$
$$\frac{\partial \mu_B}{\partial x_i} = \frac{1}{m}$$
$$\frac{\partial \sigma_B^2}{\partial x_i} = \frac{2(x_i-\mu)}{m}$$

That is,

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i}\frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial l}{\partial \sigma_B^2}\frac{2(x_i - \mu)}{m} + \frac{1}{m}\frac{\partial l}{\partial \mu_B}$$

Problem 3 (Constrained Mahalanobis Distance Minimization Problem)(1.5%)

1. Let $\Sigma \in R^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in R^m$. Please construct a set of points $x_1, ..., x_n \in R^m$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \Sigma, \quad \frac{1}{n}\sum_{i=1}^{n}x_i = \mu$$

- Find the relation between set of points and (μ, Σ) and (μ, Σ) is known

Solution: $WLOG$(Without Loss of Generality), let $\mu = 0$. Since Σ is a symmetric positive semi-definite matrix, we can perform eigen decomposition as follows:

$$\Sigma = UDU^T = \sum_{i=1}^{m}(d_i u_i u_i^T).$$

where U and U^T are orthogonal matrix. Let $n = 2m$ and construct a set of points $x_1, ..., x_m, ..., x_{2m}$ where $x_i = \sqrt{d_i}u_i \quad$ and $\quad x_{m+i} = -\sqrt{d_i}u_i \; \forall\, 1 \le i \le m$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}x_i = \mu = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T = \sum_{i=1}^{m}(d_i u_i u_i^T) = UDU^T = \Sigma$$

Note that lots of students just use the covariance of eigen decomposition to construct $\{x_i\}_{i=1}^n$. However, It should satisfy the condition that $\frac{1}{n}\sum_{i=1}^{n}x_i = \mu$

2. Let $1 \le k \le m$, solve the following optimization problem (and justify with proof):
minimize $\quad Trace(\Phi^T\Sigma\Phi)$
subject to $\quad \Phi^T\Phi = I_k$
variables $\quad \Phi \in R^{m\times k}$

Solution: Let $\phi_1, \cdots, \phi_k$ be the columns of Φ. Then

$$Trace(\Phi^T\Sigma\Phi) = \sum_{i=1}^{k}\phi_i^T\Sigma\phi_i = \sum_{i=1}^{k}\phi_i^T(\sum_{j=1}^{m}(d_j u_j u_j^T))\phi_i = \sum_{j=1}^{m}d_j\sum_{i=1}^{k}\langle u_j, \phi_i\rangle^2 = \sum_{j=1}^{m}c_j d_j$$

where $\langle\cdot,\cdot\rangle$ is standard inner product in Euclidean space, $c_j := \sum_{i=1}^{k}\langle u_j, \phi_i\rangle^2$ for each $j = 1, \cdots, m$ and $d_1 \ge d_2 \ge \cdots \ge d_m \ge 0$ Claim: $0 \le c_j \le 1$ and $\sum_{j=1}^{m}c_j = k$ Clearly, $c_j \ge 0$. Extending $\phi_1, \cdots, \phi_k$ to $\phi_1, \cdots, \phi_k, \phi_{k+1}, \cdots, \phi_m$ for $\mathbb{R}^m$. Then, for each $j = 1, \cdots, m$

$$c_j = \sum_{i=1}^{k}\langle u_j, \phi_i\rangle^2 \le \sum_{i=1}^{m}\langle u_j, \phi_i\rangle^2 = 1$$

Finally, since $u_1, \cdots, u_m$ is an orthonormal basis for $\mathbb{R}^m$,

$$\sum_{j=1}^{m} c_j = \sum_{j=1}^{m} \sum_{i=1}^{k} \langle u_j, \phi_i \rangle^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \langle u_j, \phi_i \rangle^2 = \sum_{i=1}^{k} \|\phi_i\|_2^2 = k$$

Hence, the minimum value of $\sum_{j=1}^{m} c_j d_j$ over all choice of $c_1, c_2, \cdots, c_m \in [0, 1]$ with $\sum_{j=1}^{k} c_j = k$ is $d_{m-k+1}, \cdots, d_m$. This is achieved when $c_1, \cdots, c_{m-k} = 0$ and $c_{m-k+1} = \cdots = c_m = 1$

References

[1] Sergey Ioffe and Christian Szegedy (2015), "Batch Normalization: Accelerating Deep Network Training b y Reducing Internal Covariate Shift", Arxiv:1502.03167

Problem 4 (Convergence of K-means Clustering) (1.5%)

In the K-means clustering algorithm, we are given a set of n points $x_i \in \mathbb{R}^d, i \in \{1, \ldots, n\}$ and we want to find the centers of k clusters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ by minimizing the average distance from the points to the closest cluster center. In general, $n \geq k$. Define function $\mathcal{C} : \{1, \cdots, n\} \to \{1, 2, \cdots, k\}$ assigns one of k clusters to each point in the data set such that $\mathcal{C}(i) = q$ if the i-th data point x_i is assigned to the q-th cluster where $i \in \{1, 2, \cdots, n\}$ and $q \in \{1, 2, \cdots, k\}$

Formally, we want to minimize the following loss function

$$L(\mathcal{C}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \left\| x_i - \mu_{\mathcal{C}(i)} \right\|_2^2 = \sum_{q=1}^{k} \sum_{i:\mathcal{C}(i)=q} \|x_i - \mu_q\|_2^2$$

The K-means algorithm:

Algorithm 2 K-means algorithm

Initialize cluster center $\mu_j, j = 1, 2, \cdots, k$ (k random x_n from data set)
Repeat:

1. Fix $\boldsymbol{\mu}$, update $\mathcal{C}(i)$ for each i that minimizes L. Formally, consider a data point x_i, and let $\mathcal{C}(i)$ be the assignment from the previous iteration and $C^*(i)$ be the new assignment obtained as: $C^*(i) = \arg\min_{j=1,\cdots,k} \|x_i - \mu_j\|_2^2$

2. Fix $\mathcal{C}$, update the centers μ_j which satisfies

$$|\{i : \mathcal{C}(i) = j\}| \, \mu_j = \sum_{i:\mathcal{C}(i)=j} x_i,$$

   for each j, where $|\{i : \mathcal{C}(i) = j\}|$ is the number of elements of set $\{i : \mathcal{C}(i) = j\}$.(i.e. Set the cluster centres to be the means of the points in each cluster.)

The algorithm stops when no change in loss function occurs during the assignment step.

Suppose that the algorithm proceeds from iteration t to $t + 1$.

1. Consider the points $z_1, z_2, \cdots, z_m$, where $m \geq 1$. and for $i \in \{1, 2, \cdots, m\}, z_i \in \mathbb{R}^d$. Let $\bar{z} = \frac{1}{m} \sum_{i=1}^{m} z_i$ be the mean of these points, and let $z \in \mathbb{R}^d$ be an arbitrary point in the same (d-dimensional) space. Then

$$\sum_{i=1}^{m} \|z_i - z\|_2^2 \geq \sum_{i=1}^{m} \|z_i - \bar{z}\|_2^2$$

Solution:

$$\begin{aligned}
\sum_{i=1}^{m} \left\|\boldsymbol{z}^i - \boldsymbol{z}\right\|^2 &= \sum_{i=1}^{m} \left\|\left(\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right) + \left(\overline{\boldsymbol{z}} - \boldsymbol{z}\right)\right\|^2 \\
&= \sum_{i=1}^{m} \left(\left\|\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right\|^2 + \|\overline{\boldsymbol{z}} - \boldsymbol{z}\|^2 + 2\left(\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right) \cdot \left(\overline{\boldsymbol{z}} - \boldsymbol{z}\right)\right) \\
&= \sum_{i=1}^{m} \left\|\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right\|^2 + \sum_{i=1}^{m} \|\overline{\boldsymbol{z}} - \boldsymbol{z}\|^2 + 2\sum_{i=1}^{m} \left(\boldsymbol{z}^i \cdot \overline{\boldsymbol{z}} - \boldsymbol{z}^i \cdot \boldsymbol{z} - \overline{\boldsymbol{z}} \cdot \overline{\boldsymbol{z}} + \overline{\boldsymbol{z}} \cdot \boldsymbol{z}\right) \\
&= \sum_{i=1}^{m} \left\|\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right\|^2 + m\|\overline{\boldsymbol{z}} - \boldsymbol{z}\|^2 + 2(m\overline{\boldsymbol{z}} \cdot \overline{\boldsymbol{z}} - m\overline{\boldsymbol{z}} \cdot \boldsymbol{z} - m\overline{\boldsymbol{z}} \cdot \overline{\boldsymbol{z}} + m\overline{\boldsymbol{z}} \cdot \boldsymbol{z}) \\
&= \sum_{i=1}^{m} \left\|\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right\|^2 + m\|\overline{\boldsymbol{z}} - \boldsymbol{z}\|^2 \\
&\geq \sum_{i=1}^{m} \left\|\boldsymbol{z}^i - \overline{\boldsymbol{z}}\right\|^2 .
\end{aligned}$$

2. Show that $L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t) \leq L(\mathcal{C}^t, \boldsymbol{\mu}^t)$ i.e. The first step in K-means clustering
Solution: It follows directly from the logic of the algorithm: $\mathcal{C}^t$ and $\mathcal{C}^{t+1}$ are different only if there is a point that finds a closer cluster center in $\boldsymbol{\mu}^t$ than the one assigned to it by $\mathcal{C}^t$:

$$L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t\right) = \sum_{i=1}^{n} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^t_{C^{t+1}(i)}\right\|^2 < \sum_{i=1}^{n} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^t_{C^t(i)}\right\|^2 = L\left(\mathcal{C}^t, \boldsymbol{\mu}^t\right)$$

3. Show that $L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}) \leq L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t)$ i.e. The second step in K-means clustering. (Hint: Use the result of (a)) *Solution:* Use the result in (1):

$$\begin{aligned}
L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}\right) &= \sum_{i=1}^{n} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^{t+1}_{C^{t+1}(i)}\right\|^2 \\
&= \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, \mathcal{C}^{t+1}(i)=k'} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^{t+1}_{C^{t+1}(i)}\right\|^2 \\
&\leq \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, \mathcal{C}^{t+1}(i)=k'} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^{t}_{C^{t+1}(i)}\right\|^2 \text{ (by 1)} \\
&= \sum_{i=1}^{n} \left\|\boldsymbol{x}^i - \boldsymbol{\mu}^{t}_{C^{t+1}(i)}\right\|^2 \\
&= L\left(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t\right).
\end{aligned}$$

4. Use the result in (b) and (c) to show that the loss of k-means clustering algorithm is monotonic decreasing.(Hint: Show that the sequence $\{l_t\}$, where $l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t)$, which is monotone decreasing $(l_{t+1} \leq l_t, \forall t)$ and bounded below $(l_t \geq 0)$. Then, we use monotone convergence theorem for sequences, $\{l_t\}$ converges.) *Solution:* Define the sequence $\{l_t\}$, where $l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t)$. By previous results, we have

$$l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t) \leq L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}) = l_{t+1}$$

for all t. Hence, $\{l_t\}$ is a monotonic decreasing sequence. Note that we apply **monotonic convergence theorem of sequence** to prove the sequence is convergence, which does not guarantee this algorithm could find the **global** minimum, just a **local** minimum.

5. Show that the k-means clustering algorithm converges in finitely many steps. *Solution:* There are at most k^N ways to partition N data points into k clusters. Then, this algorithm converges in finitely many steps. Note that the upper bound (k^N) may not tight.

Problem 5 (Gradient Descent Convergence) (1.5%)

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Also, f is β-smoothness and α-strongly convex.

$$\beta - smoothness : \beta > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

$$\alpha - strongly\ convex : \alpha > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Then we propose a gradient descent algorithm

1. Find a initial $\boldsymbol{\theta}^0$.
2. Let $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\ \forall n \geq 0$, where $\eta = \frac{1}{\beta}$.

The following problems lead you to prove the gradient descent convergence.

1. Prove the property of β-smoothness function

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \leq \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

   (a) Define $g : \mathbb{R} \to \mathbb{R}, g(t) = f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))$. Show that $f(\boldsymbol{x}) - f(\boldsymbol{y}) = \int_0^1 g'(t)\, dt$.
   Solution: $f(\boldsymbol{x} - \boldsymbol{y}) = g(1) - g(0) = \int_0^1 g'(t) dt$.

   (b) Show that $g'(t) = \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})$.
   Solution: Let $\boldsymbol{z} = \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})$.
   $g'(t) = \frac{dg(t)}{dt} = \frac{df(\boldsymbol{z})}{dt} = \sum_{i=1}^n \frac{df(\boldsymbol{z})}{\boldsymbol{z}_i}\frac{d\boldsymbol{z}_i}{dt} = \sum_{i=1}^n \frac{df(\boldsymbol{z})}{\boldsymbol{z}_i}\frac{d(\boldsymbol{y}_i + t(\boldsymbol{x}_i - \boldsymbol{y}_i))}{dt} = \sum_{i=1}^n \frac{df(\boldsymbol{z})}{\boldsymbol{z}_i}(\boldsymbol{x}_i - \boldsymbol{y}_i) = \nabla f(\boldsymbol{z})^T(\boldsymbol{x} - \boldsymbol{y}) = \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})$

   (c) Show that $|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})| \leq \int_0^1 |(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})|\, dt$.
   Solution:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})| = |\int_0^1 g'(t)dt - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})|$$

$$= |\int_0^1 \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})dt - \int_0^1 \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})dt|$$

$$= |\int_0^1 (\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^T - \nabla f(\boldsymbol{y})^T)(\boldsymbol{x} - \boldsymbol{y})dt|$$

(d) By Cauchy-Schwarz inequality and the definition of β-smoothness, show that $|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})| \leq \frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, hence we get

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) \leq \frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

Solution: By Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\int_0^1 (\nabla f(\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))^T - \nabla f(\boldsymbol{y})^T)(\boldsymbol{x}-\boldsymbol{y})dt| \\
&\leq \int_0^1 |(\nabla f(\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))^T - \nabla f(\boldsymbol{y})^T)(\boldsymbol{x}-\boldsymbol{y})|dt \\
&\leq \int_0^1 |(\nabla f(\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y})) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x}-\boldsymbol{y})|dt \\
&\leq \int_0^1 ||(\nabla f(\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))-\nabla f(\boldsymbol{y}))||_2||(\boldsymbol{x}-\boldsymbol{y})||_2 dt \quad \text{By Cauchy-Schwarz inequality} \\
&\leq \int_0^1 \beta||t(\boldsymbol{x}-\boldsymbol{y})||_2(\boldsymbol{x}-\boldsymbol{y})||_2 dt \leq \int_0^1 t\beta||(\boldsymbol{x}-\boldsymbol{y})||_2^2 dt = \beta||(\boldsymbol{x}-\boldsymbol{y})||_2^2 \int_0^1 tdt = \frac{\beta}{2}||(\boldsymbol{x}-\boldsymbol{y})||_2^2
\end{aligned}$$

2. Let $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})$ and use 1., prove that

$$f(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2$$

and

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2,$$

where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$.

Solution: By the fact that $f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) \leq \frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$,

$$f(\boldsymbol{x}) - f(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})) - \nabla f(\boldsymbol{x})^T\frac{1}{\beta}\nabla f(\boldsymbol{x}) \leq \frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2.$$

Also, $\nabla f(\boldsymbol{x})^T\frac{1}{\beta}\nabla f(\boldsymbol{x}) = \frac{1}{\beta}\|\nabla f(\boldsymbol{x})\|_2^2$. Hence,

$$f(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2$$

Owing to the fact that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) - \frac{1}{\beta}$, we get

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2.$$

3. Show that $\forall n \geq 0$,

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \eta^2 \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 - 2\eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^T (\boldsymbol{\theta}^n - \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

Solution:

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n) - \boldsymbol{\theta}^*\|_2^2$$

$$(\boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n) - \boldsymbol{\theta}^*)^T (\boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n) - \boldsymbol{\theta}^*) = \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \eta^2 \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 - 2\eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^T (\boldsymbol{\theta}^n - \boldsymbol{\theta}^*)$$

4. Use 2. and the definition of α-strongly convex to prove $\forall n \geq 0$

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 \leq (1 - \frac{\alpha}{\beta}) \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2,$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

Solution: By rearranging the inequality we get in 2., $\eta^2 \|\nabla f(\boldsymbol{\theta}^n)\|_2^2 \leq 2\beta\eta^2 (f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n))$.
Also, by the definition of α-strongly convex, $-2\eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^T (\boldsymbol{\theta}^n - \boldsymbol{\theta}^*) \leq -\alpha\eta ||\boldsymbol{\theta}^n - \boldsymbol{\theta}^*||_2^2 - 2\eta (f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n))$ Combined the above results and $\eta = \frac{1}{\beta}$, we get

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 \leq \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \frac{2}{\beta}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)) - \frac{\alpha}{\beta} ||\boldsymbol{\theta}^n - \boldsymbol{\theta}^*||_2^2 - \frac{2}{\beta}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)) = (1 - \frac{\alpha}{\beta}) \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2.$$

5. Use the above inequality to show that $\boldsymbol{\theta}^n$ will converge to $\boldsymbol{\theta}^*$ when n goes to infinity.

Solution: Because $\frac{\alpha}{\beta} > 0$, also, by the property of β-smoothness function and the definition of α strongly convex function, we get $\alpha \leq \beta$, which derives $\frac{\alpha}{\beta} < 1$ and $0 < 1 - \frac{\alpha}{\beta} < 1$. When n goes to infinity, $||\boldsymbol{\theta^n} - \boldsymbol{\theta^*}|| = 0$. Hence, $\boldsymbol{\theta^n}$ will converge to $\boldsymbol{\theta^*}$

Version Description

1. First Edition: Finish Problem Solution 1 to 5