236330 - Introduction to Optimization: Homework #2

May 5, 2020

Amit Rotner
123456789
Or Steiner
123456789

Gradient Descent method and Newton method

Task 1 – Convex sets and functions:

Q1:

Show that if f_1 and f_2 are convex functions on a convex domain C, then $g(x) = max_{i=1,2}f_i(x)$

is also a convex function.

Solution:

We know that f_1 is a convex function, hence $\forall x_1, x_2 \in C$, $\forall \alpha \in [0,1]$:

$$f_1\left(\alpha x_1 + (1-\alpha)\, x_2\right) \leq \alpha f_1\left(x_1\right) + (1-\alpha)\, f_1\left(x_2\right)$$

In addition, f_1 is a convex function, hence $\forall x_1, x_2 \in C$, $\forall \alpha \in [0,1]$:

$$f_2\left(\alpha x_1 + (1-\alpha)\, x_2\right) \leq \alpha f_2\left(x_1\right) + (1-\alpha)\, f_2\left(x_2\right)$$

Now, let $x_1, x_2 \in C$, $\alpha \in [0,1]$:

$$\begin{aligned} g\left(\alpha x_1 + (1-\alpha)\, x_2\right) &= max_{i=1,2}f_i\left(\alpha x_1 + (1-\alpha)\, x_2\right) \\ &\leq max_{i=1,2}\left(\alpha f_i\left(x_1\right) + (1-\alpha)\, f_i\left(x_2\right)\right) \\ &\leq \alpha max_{i=1,2}f_i\left(x_1\right) + (1-\alpha)\, max_{i=1,2}f_i\left(x_2\right) \\ &= \alpha g\left(x_1\right) + (1-\alpha)\, g\left(x_2\right) \end{aligned}$$

Therefore, $g\left(x\right)$ is a convex function.

Q2:

Let $f\left(x\right)$ be a convex function defined over a convex domain C.

Show that the level set $L = \{x \in C : f\left(x\right) \leq \alpha\}$ is convex.

Solution:

We know that f is a convex function, hence $\forall x_1, x_2 \in C$, $\forall \beta \in [0,1]$:

$$f\left(\beta x_1 + (1-\beta)\, x_2\right) \le \beta f\left(x_1\right) + (1-\beta)\, f\left(x_2\right)$$

Let $x_1, x_2 \in L$:

From the definition of L, $f\left(x_1\right) \le \alpha$ and $f\left(x_2\right) \le \alpha$.

Now, let $\beta \in [0,1]$:

$$\begin{aligned} f\left(\beta x_1 + (1-\beta)\, x_2\right) &\le \beta f\left(x_1\right) + (1-\beta)\, f\left(x_2\right) \\ &\le \beta\alpha + (1-\beta)\,\alpha \\ &= \alpha \end{aligned}$$

Therefore, $\beta x_1+(1-\beta)\, x_2 \in L$. Hence, L in a convex set.

Q3:

Let $f\left(x\right)$ be a smooth and twice differentiable convex function continuously.

Show that $g\left(x\right) = f\left(Ax\right)$ is convex, where A is a matrix of appropriate size.

Check positive semi-definiteness of the Hessian.

Solution:

We know that f is a convex function, hence $\forall x_1, x_2 \in C$, $\forall \beta \in [0,1]$:

$$f\left(\beta x_1 + (1-\beta)\, x_2\right) \le \beta f\left(x_1\right) + (1-\beta)\, f\left(x_2\right)$$

Let $x_1, x_2 \in L, \beta \in [0,1]$:

$$\begin{aligned} g\left(\beta x_1 + (1-\beta)\, x_2\right) &= f\left(A\beta x_1 + A\,(1-\beta)\, x_2\right) \\ &= f\left(\beta A x_1 + (1-\beta)\, A x_2\right) \\ &\le \beta f\left(Ax_1\right) + (1-\beta)\, f\left(Ax_2\right) \\ &= \beta g\left(Ax_1\right) + (1-\beta)\, g\left(Ax_2\right) \end{aligned}$$

Therefore, $g\left(x\right)$ is a convex function.

As we have seen in HW1, the Hessian of g equals to:

$$H(x) = A^T \nabla^2 f(Ax) A$$

Given that f is smooth and twice differentiable convex function continuously, $\nabla^2 f$ is PSD.

Therefore, $\forall x \in \mathbb{R}^n : x^T \nabla^2 f x \geq 0$.

$\forall x \in \mathbb{R}^n$:

$$x^T H x = x^T A^T \nabla^2 f A x \overset{z=Ax}{=} z^T \nabla^2 f z \geq 0$$

Hence, H is PSD.

Q4:

Phrase and prove Jensen's inequality for the discrete case.

Solution:

Jensen's inequality:

Let $f : C \to \mathbb{R}$ be a convex function and let $x_1, x_2, \ldots, x_n \in C$.

If $\alpha_1, \alpha_2, \ldots, \alpha_n$ are positive numbers such that $\sum_{i=1}^{n} \alpha_i = 1$ then:

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$$

Proof, by induction:

- For $n = 2$: $a_2 = 1 - \alpha_1$.
  The statement:
  $$f(\alpha_1 x_1 + (1 - \alpha_1) x_2) \leq \alpha_1 f(x_1) + (1 - \alpha_1) f(x_2)$$
  is true by the convexity of f.

- Suppose that the statement is true for some n, we need to prove that it's true for $n+1$:
  Let $\alpha_1, \alpha_2, \ldots, \alpha_{n+1}$ be positive numbers such that $\sum_{i=1}^{n+1} \alpha_i = 1$.
  By the convexity of f:

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \alpha_i x_i\right) &= f\left(\alpha_1 x_1 + \sum_{i=2}^{n+1} \alpha_i x_i\right) \\ &= f\left(\alpha_1 x_1 + (1-\alpha_1) \sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i\right) \\ &\leq \alpha_1 f(x_1) + (1-\alpha_1) f\left(\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i\right) \end{aligned}$$

  Since $\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} = 1$, by the induction hypotheses we get:

$$\begin{aligned} \alpha_1 f(x_1) + (1-\alpha_1) f\left(\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i\right) &\leq \alpha_1 f(x_1) + (1-\alpha_1) \sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} f(x_i) \\ &= \alpha_1 f(x_1) + \sum_{i=2}^{n+1} \alpha_i f(x_i) \\ &= \sum_{i=1}^{n+1} \alpha_i f(x_i) \end{aligned}$$

$\blacksquare$

Q5:

Using Jensen inequality, prove arithmetic geometric mean inequality:

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Where $\forall i,\ x_i > 0$.

Solution:

We know that $log(\cdot)$ is a concave function. Hence, $-log(\cdot)$ is a convex function.

Therefore, using Jensen's inequality, where $\forall i:\ \alpha_i = \frac{1}{n}$ yields:

$$\begin{aligned}
log\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) &= log\left(\sum_{i=1}^{n}\alpha_i x_i\right)\\
&\geq \sum_{i=1}^{n}\alpha_i log\left(x_i\right)\\
&= \frac{1}{n}log\left(x_1\right)+\cdots+\frac{1}{n}log\left(x_n\right)\\
&= \frac{1}{n}log\left(x_1\cdot x_2\cdots x_n\right)\\
&= log\left(\sqrt[n]{x_1\cdot x_2\cdots x_n}\right)
\end{aligned}$$

Since $log\left(\cdot\right)$ is strictly increasing,

$$\frac{x_1+x_2+\cdots+x_n}{n} \geq \sqrt[n]{x_1\cdot x_2\cdots x_n}$$

■

Task 2 – Gradient Descent Analytical Convergence:

Q6:

Let $f\left(x\right)=\frac{1}{2}x^TQx-b^Tx+c$ be the function to minimize, where $Q\succ 0$ is a symmetric matrix.

1. We will define the condition number of a positive definite matrix A as $\theta\triangleq\frac{\lambda_{max}}{\lambda_{min}}$.
   Write an upper bound on the convergence ratio β that we found in the tutorial,
   using $\theta\left(Q\right)$ - the condition number of Q.
2. Assume that the step size can be modified at any iteration.
   Find the optimal step size α_k^* .

Solution:

We start by finding x^*:

$$\begin{aligned}
df = Qx-b=0\\
Qx^*=b\\
x^*=Q^{-1}b
\end{aligned}$$

$\nabla f\left(x^*\right)=0$ Hence, $\nabla f\left(x^*\right)=Qx^*-b$

From the definition of the gradient decent method,

$$x_{k+1}=x_k-\alpha\nabla f\left(x_k\right)=x_k-\alpha\left(Qx_k-b\right)=\left(I-\alpha Q\right)x_k+\alpha b$$

1. From the above, we get:

$$x_{k+1} = (I - \alpha Q)\, x_k + \alpha b = (I - \alpha Q)\, x_k + \alpha Q x^*$$

And therefore,

$$x_{k+1} - x^* = (I - \alpha Q)\, x_k + \alpha Q x^* - x^* = (I - \alpha Q)\, x_k + (\alpha Q - I)\, x^* = (I - \alpha Q)\,(x_k - x^*)$$

As we have learned in Numerical Algorithms, $\|Ax\| \leq \|A\| \cdot \|x\|$ for every norm.
Hence:

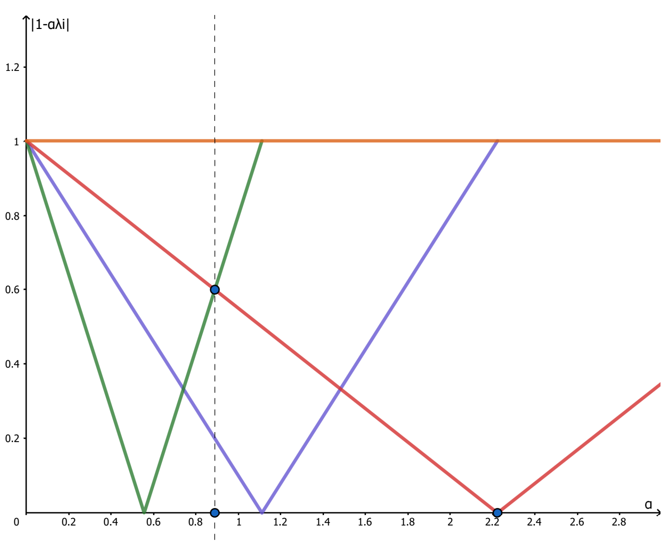
$$\|(I - \alpha Q)\,(x_k - x^*)\|_2 \leq \|(I - \alpha Q)\|_2 \cdot \|(x_k - x^*)\|_2$$

$$\begin{aligned}
\beta &\triangleq \frac{\|x_{k+1} - x^*\|_2}{\|x_k - x^*\|_2} \\
&= \frac{\|(I - \alpha Q)\,(x_k - x^*)\|_2}{\|x_k - x^*\|_2} \\
&\leq \frac{\|(I - \alpha Q)\|_2 \cdot \|(x_k - x^*)\|_2}{\|x_k - x^*\|_2} \\
&= \|(I - \alpha Q)\|_2 \\
&= \sigma_{max}\,(I - \alpha Q) \\
&= max_i\,\{|1 - \alpha\lambda_i|\}
\end{aligned}$$

Hence, we need to find:

$$\alpha_{opt} = argmin_\alpha max_i\,\{|1 - \alpha\lambda_i|\}$$

We can solve this problem graphically:



Looking on the graph we see that the optimal solution satisfies:

$$1-\alpha_{opt}\lambda_{min}=\alpha_{opt}\lambda_{max}-1$$
$$\alpha_{opt}=\frac{2}{\lambda_{min}+\lambda_{max}}$$

Therefore,

$$\begin{aligned}\beta &\leq max_i\left\{\left|1-\alpha_{opt}\lambda_i\right|\right\}\\ &=max\left\{1-\alpha_{opt}\lambda_{min},1-\alpha_{opt}\lambda_{max}\right\}\\ &=1-\alpha_{opt}\lambda_{min}\\ &=1-\frac{2}{\lambda_{min}+\lambda_{max}}\lambda_{min}\\ &=\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}\\ &=\frac{\theta\left(Q\right)-1}{\theta\left(Q\right)+1}\end{aligned}$$

2. We know from above that:

$$x_{k+1}=x_k-\alpha_k\nabla f\left(x_k\right)=x_k+\alpha_k d_k$$

The optimal step is given by:

$$\alpha_k^*=argmin_{\alpha_k}f\left(x_{k+1}\right)=argmin_{\alpha_k}f\left(x_k-\alpha_k\nabla f\left(x_k\right)\right)=argmin_{\alpha_k}f\left(x_k+\alpha_k d_k\right)$$

Since $f\left(x\right)=\frac{1}{2}x^TQx-b^Tx+c$:

$$\begin{aligned}\alpha_k^* &=argmin_{\alpha_k}\frac{1}{2}\left(x_k+\alpha_k d_k\right)^TQ\left(x_k+\alpha_k d_k\right)-b^T\left(x_k+\alpha_k d_k\right)+c\\ &=argmin_{\alpha_k}\frac{1}{2}\left\|A\left(x_k+\alpha_k d_k\right)\right\|^2-b^T\left(x_k+\alpha_k d_k\right)+c\end{aligned}$$

$$\frac{df\left(x_k-\alpha_k d_k\right)}{\alpha_k}=d_k^TQ\left(x_k+\alpha_k d_k\right)-d_k^Tb=0$$

$$d_k^T\left(Qx_k-b\right)+\alpha_k d_k^TQd_k=0$$

Because $\nabla f\left(x_k\right)=Qx_k-b$, we get that $d_k=-\nabla f\left(x_k\right)=b-Q_k$.
Hence,

$$0=d_k^T\left(Qx_k-b\right)+\alpha_k d_k^TQd_k=-d_k^Td_k+\alpha_k d_k^TQd_k$$

$$\alpha_k d_k^TQd_k=d_k^Td_k$$

Assuming that $d_k \neq 0$ (otherwise, $x_k = x^*$), we have that $d_k^T Q d_k > 0$ because $Q \succ 0$.
Therefore:

$$\alpha_k^* = \frac{d_k^T Q d_k}{d_k^T d_k}$$

Q7:

Let there be a strongly convex function $f(x)$.

Prove that if $\forall x \in Dom(f): \ mI \succcurlyeq \nabla^2 f(x) \succcurlyeq MI$ then:

$$\frac{1}{2m} \|\nabla f(x))\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2M} \|\nabla f(x))\|_2^2$$

Solution:

From Taylor's multivariate theorem we know that:

$$\forall x, y \in \mathbb{R}^n, \ \exists z \in [x, y]: \ f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(z) (y - x)$$

Strong convexity implies that:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T M (y - x) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} M \|y - x\|_2^2$$

The above function minimized at $y^* = x - \frac{1}{M} \nabla f(x)$.

Therfore:

$$\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^T (y^* - x) + \frac{1}{2} M \|y^* - x\|_2^2 \\
&= f(x) + \nabla f(x)^T \left(x - \frac{1}{M} \nabla f(x) - x \right) + \frac{1}{2} M \left\| x - \frac{1}{M} \nabla f(x) - x \right\|_2^2 \\
&= f(x) - \frac{1}{M} \nabla f(x)^T \nabla f(x) + \frac{M}{2} \frac{1}{M^2} \|\nabla f(x)\|_2^2 \\
&= f(x) - \frac{1}{M} \|\nabla f(x)\|_2^2 + \frac{1}{2M} \|\nabla f(x)\|_2^2 \\
&= f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2
\end{aligned}$$

On the other side of the inequality, we get:

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T m (y-x) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} m \|y-x\|_2^2$$

The above function minimized at $y^* = x - \frac{1}{m}\nabla f(x)$.

Therfore:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (y^* - x) + \frac{1}{2} < \|y^* - x\|_2^2 \\ &= f(x) + \nabla f(x)^T \left(x - \frac{1}{m}\nabla f(x) - x\right) + \frac{1}{2} m \left\|x - \frac{1}{m}\nabla f(x) - x\right\|_2^2 \\ &= f(x) - \frac{1}{m}\nabla f(x)^T \nabla f(x) + \frac{m}{2}\frac{1}{m^2}\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{m}\|\nabla f(x)\|_2^2 + \frac{1}{2m}\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \end{aligned}$$

By choosing $y = x^*$ we get:

$$f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \geq f(x^*) \geq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

$$-\frac{1}{2m}\|\nabla f(x)\|_2^2 \geq f(x^*) - f(x) \geq -\frac{1}{2M}\|\nabla f(x)\|_2^2$$

$$\frac{1}{2m}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2M}\|\nabla f(x)\|_2^2$$

■