Tutorial 4 Problems

CSC311, Fall 2021

1 Gradient Descent Intuition

Suppose we are trying to optimize the loss function $f(x) = \frac{1}{2}x^T Ax$, where $x \in \mathbb{R}^2$

1. Let $A = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ and $x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
   What are the first two iterates of gradient descent, with a learning rate $\eta = 0.1$?
   (Solution: We have

$$\begin{aligned} x_{n+1} &= x_n - \eta \nabla f(x_n) \\ &= x_n - \eta A x_n \\ &= (I - \eta A)x_n \\ &= \begin{bmatrix} 1-4\eta & 0 \\ 0 & 1-\eta \end{bmatrix} x_n \end{aligned}$$

   So in general:

$$x_n = \begin{bmatrix} (1-4\eta)^n & 0 \\ 0 & (1-\eta)^n \end{bmatrix} x_0$$

   giving us $x_1 = \begin{bmatrix} 0.6 \\ 0.9 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 0.36 \\ 0.81 \end{bmatrix}$.
   Note that the later parts of this problem are good for building intuition but less important for purposes of the course. You can get more detailed explainations in the first part of the Distill article: `https://distill.pub/2017/momentum/`)

2. For which learning rates will gradient descent converge? The convergence speed is determined by how the error decreases in the "slowest" direction. What learning rate leads to the fastest convergence?
   (Solution: We need $|1-4\eta| < 1$ and $|1-\eta| < 1$. This holds when $\eta \in (0, \frac{1}{2})$.
   The convergence rate is given by $\max\{|1-4\eta|, |1-\eta|\}$. We want to choose η to minimize this rate, which happens when they are equal. From $|1-4\eta| = |1-\eta|$, we get a learning rate $\eta^* = 0.4$. This is where we "bounce around" in the direction of higher curvature.)

3. Suppose we choose the optimal learning rate. How many steps of gradient descent does it take for both components to be less than 1e-3 (0.001)?
   (Solution: This occurs when $(0.6)^n < 0.001$, or $\log_{0.6}(0.001) \approx 13.52$. So 14 iterations.)

4. Repeat the previous two parts with $A = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}$.
   (Solution: The optimal learning rate is now $\frac{2}{101}$. Notice it is much smaller when there is a larger difference between the two directions. For part (c), we now have $\log_{\frac{99}{101}}(0.001) \approx 345.37$. So 346 iterations.)

2 Sum of Convex Functions

Prove that the sum of two convex functions is convex.
(Solution: Let f and g be convex functions. Consider $h = f + g$. We have

$$\begin{aligned} h(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\leq \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\ &= \lambda h(x) + (1-\lambda)h(y) \end{aligned}$$

for all x, y and $\lambda \in (0,1)$. So h is convex.)