Optimisation Basics Tutorial

Tirtharaj Dash

Dept. of CS & IS and APPCAIR
BITS Pilani, Goa Campus

August 23, 2021

Basic Optimisation I

Gradient of a scalar function Let $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$ and let $f(\mathbf{x})$ be a **scalar function** of $\mathbf{x}$. Then the derivative of $f(\mathbf{x})$ w.r.t. $\mathbf{x}$, called the **gradient vector** or **gradient** of $f(\mathbf{x})$ is a column vector denoted by

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \text{ or } \nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right]^\mathsf{T}$$

Gradient of a vector function Let $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$ and let $\mathbf{f}(\mathbf{x})$ be a **vector function** of $\mathbf{x}$, denoted by $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}, \ldots, f_m(\mathbf{x}]^\mathsf{T}$. Then, the derivative of $\mathbf{f}(\mathbf{x})$ w.r.t. $\mathbf{x}$, called the **Jacobian matrix** or **Jacobian** of $\mathbf{f}(\mathbf{x})$, is an $m \times n$ matrix denoted by

$$\mathbf{J_f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{x}}^\mathsf{T} f_1(\mathbf{x}) \\ \vdots \\ \nabla_{\mathbf{x}}^\mathsf{T} f_m(\mathbf{x}) \end{bmatrix}$$

Basic Optimisation III

Hessian of a scalar function Let $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$ and let $f(\mathbf{x})$ be a **scalar function** of $\mathbf{x}$. Then the second derivative of $f(\mathbf{x})$, called the **Hessian matrix** or **Hessian** of $f(\mathbf{x})$, is an $n \times n$ matrix denoted by

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

which is:

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1}\right) & \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_n}\right) \\ \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_n}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_1}\right) & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_n}\right) \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{x}}^\mathsf{T} \frac{\partial f}{\partial x_1} \\ \vdots \\ \nabla_{\mathbf{x}}^\mathsf{T} \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Basic Optimisation IV

Gradient of a function (1) Let $\mathbf{c} = [c_1, \ldots, c_n]^\mathsf{T}$ and $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$. Then the gradient of a linear scalar function $f(\mathbf{x}) = \mathbf{c}^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{c}$ w.r.t. $\mathbf{c}$

$$\nabla_{\mathbf{c}} f(\mathbf{x}) = \mathbf{x}$$

Gradient of a function (2) If $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{x}$, then

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{x}$$

Gradient of a function (3) If $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$, then

$$\nabla_{\mathbf{x}} f = 2\mathbf{A}\mathbf{x}$$

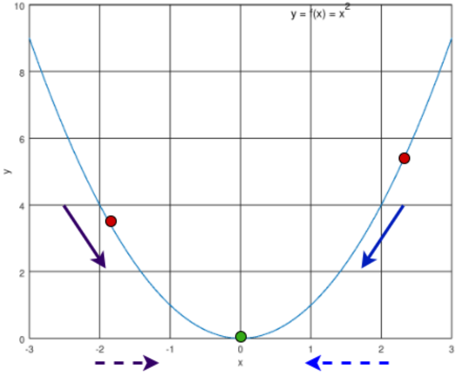
Basic Optimisation V

Let's look at minimisation problems for functions that are continuous and differentiable.

- If the derivative of the function is positive, the function is increasing.
  - Don't move in that direction, because you'll be moving away from a minimum.
- If the derivative of the function is negative, the function is decreasing.
  - Keep going, since you're getting closer to a minimum.

Basic Optimisation VI

Let $f(x) = x^2$. The function looks like this:



The arrows show movement of next functional value, and the dotted arrows show the corresponding direction of movement of x.

Basic Optimisation VII

Here is a very simple gradient descent procedure:

1. Initialize x to some value
2. **while** stopping criterion is not met
    1. Calculate the gradient of the function, $\nabla_x f$
    2. $x := x - \eta \nabla_x f$
3. **return** x

Notice step 2.2. above: x will move right, if $\nabla_x f$ is negative, and it will move left, if $\nabla_x f$ is positive.

Basic Optimisation VIII

1. Using gradient descent, obtain the value of x that minimizes $f(x) = (x-2)^2 - 5$. Starting value of $x = 3$ and $\eta = 1$.

   Answer. Derivative of f w.r.t. x: $\nabla f = 2(x-2)$

   - $x = 3$: $\nabla f|_{x=3} = 2$; $x = 3 - 2 = 1$; $f(1) = -4$
   - $x = 1$: $\nabla f|_{x=1} = -2$; $x = 1 - (-2) = 3$; $f(3) = -4$.
   - ... gets repeated.

Basic Optimisation IX

2. Solve the same question with same starting point, but with $\eta = 0.5$.

Answer. Derivative of f w.r.t. x: $\nabla f = 2(x-2)$

- $x = 3$: $\nabla f|_{x=3} = 2$; $x = 3 - 0.5 \times 2 = 2$; $f(2) = -5$
- $x = 2$: $\nabla f|_{x=2} = 0$; $x = 2 - 0.5 \times 0 = 2$; $f(2) = -5$.
- $x = 2$: $\nabla f|_{x=2} = 0$; $x = 2 - 0.5 \times 0 = 2$; $f(2) = -5$.
- Value of f doesn't change further. So, stopping criterion met. Return $x = 2$. This is same as the exact solution i.e. Find root of $\nabla f = 0$.

Basic Optimisation X

Gradient descent is guaranteed to eventually find a local minimum if:

- the learning rate is set appropriately (sometimes, using adaptive learning rate); $\eta \in [0.0001, 1]$.
- a finite local minimum exists (i.e. the function doesn't keep decreasing forever).

Basic Optimisation XI

Various stopping criteria for gradient descent:

- Stop when the norm of the gradient is below some threshold, θ

$$||\nabla f|| < \theta$$

  This is checking the distant the gradient is from the origin, $\mathbf{0}$.

- Maximum number of iterations is reached.

Basic Optimisation XII

It is straightforward to extend the gradient descent procedure to scalar functions with multiple variables.

3. Let $f(x_1, x_2) = 3x_1^2 - 2x_1x_2 + x_2^2 - 5$. Initial values $x_1 = 1$, $x_2 = 1$. Fix $\eta = 1$.

   Answer. Present value of f: $f(1,1) = 3 - 2 + 1 - 5 = -3$. The partial derivatives are:

$$\nabla_{x_1} f = 6x_1 - 2x_2$$
$$\nabla_{x_2} f = 2x_2 - 2x_1$$

Basic Optimisation XIII

Update the present $x_{1,2}$:

$$\begin{aligned} x_1 &= x_1 - \eta \nabla_{x_1} f \\ &= 1 - (6 - 2) = -3 \\ x_2 &= x_2 - \eta \nabla_{x_2} f \\ &= 1 - (2 - 2) = 1 \end{aligned}$$

New value of f: $f(-3, 1) = 29$. Update the present $x_{1,2}$ using gradients:

$$\begin{aligned} x_1 &= x_1 - \eta \nabla_{x_1} f \\ &= -3 - (-18 - 2) = 17 \\ x_2 &= x_2 - \eta \nabla_{x_2} f \\ &= 1 - (-6 - 2) = 9 \end{aligned}$$

New value of f: $f(17, 9) = 637$.

Basic Optimisation XIV

4. Solve the above question with $\eta = 0.1$.

Answer. Update the present $x_{1,2}$:

$$x_1 = 1 - 0.1(6 - 2) = 0.6$$
$$x_2 = 1 - 0.1(2 - 2) = 1$$

New value of f: $f(0.6, 1) = -4.12$. Update the present $x_{1,2}$ using gradients:

$$x_1 = 0.6 - 0.1(3.6 - 2) = 0.44$$
$$x_2 = 1 - 0.1(2 - 1.2) = 0.92$$

New value of f: $f(0.44, 0.92) = -4.38$.