2 Feb 2022

Gradient Descent

Announcement: Winter storm coming!

Check Cornell's operating status.
If classes canceled, that means online lectures will be canceled too!

Stay safe!

Example. (Differential vs. Gradient)

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = 4x_1^2 + x_2^2$$

$$\frac{\partial f}{\partial x_1} = 8x_1 \qquad \frac{\partial f}{\partial x_2} = 2x_2$$

Deriving a formula for df_x:

$$f(x+y) = 4(x_1+y_1)^2 + (x_2+y_2)^2$$

$$= 4x_1^2 + 8x_1y_1 + 4y_1^2 + x_2^2 + 2x_2y_2 + y_2^2$$

$$= \underbrace{\left(4x_1^2 + x_2^2\right)}_{f(x)} + \underbrace{\left(8x_1y_1 + 2x_2y_2\right)}_{df_x(y)} + \underbrace{\left(4y_1^2 + y_2^2\right)}_{g(y)}$$

$$df_x(\vec{y}) = 8x_1y_1 + 2x_2y_2 = \begin{bmatrix} 8x_1 & 2x_2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Now suppose $\mathbb{R}^2$ is given the non-standard inner product $\langle x, y\rangle \stackrel{\text{def}}{=} 2x_1y_1 + x_2y_2$.

What is ∇f_x?

We know it is def'd as the image of df_x under the isomorphism $(\mathbb{R}^2)^* \to \mathbb{R}^2$ induced by $\langle \cdot, \cdot \rangle$.

In other words, it means ∇f_x is the unique vector in $\mathbb{R}^2$ satisfying

$$\forall y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \langle \nabla f_x, y\rangle = df_x(y) = 8x_1y_1 + 2x_2y_2.$$

Denote ∇f_x by $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ we are seeking z_1, z_2 that satisfy

$$\forall y \qquad 2z_1y_1 + z_2y_2 = 8x_1y_1 + 2x_2y_2$$

$$\therefore \quad \begin{aligned} z_1 &= 4x_1 \\ z_2 &= 2x_2 \end{aligned} \quad \Longrightarrow \quad \nabla f_x = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}$$

Compare with the gradient w.r.t. standard inner product, $\begin{bmatrix} 8x_1 \\ 2x_2 \end{bmatrix}$.

Gradient Descent.

To search for a point in vector space V where $f: V \to \mathbb{R}$ is minimized, make a sequence of steps $x_0, x_1, x_2, \ldots$ each moving in the direction of $-\nabla f$.

We will be analyzing the following algorithm parameterized by $\eta > 0$. Greek letter ETA

(η = "step size")

Given: differentiable function $f: V \to \mathbb{R}$

inner product $\langle \cdot, \cdot \rangle$ on V

initial point $x_0 \in V$

step size $\eta > 0$,

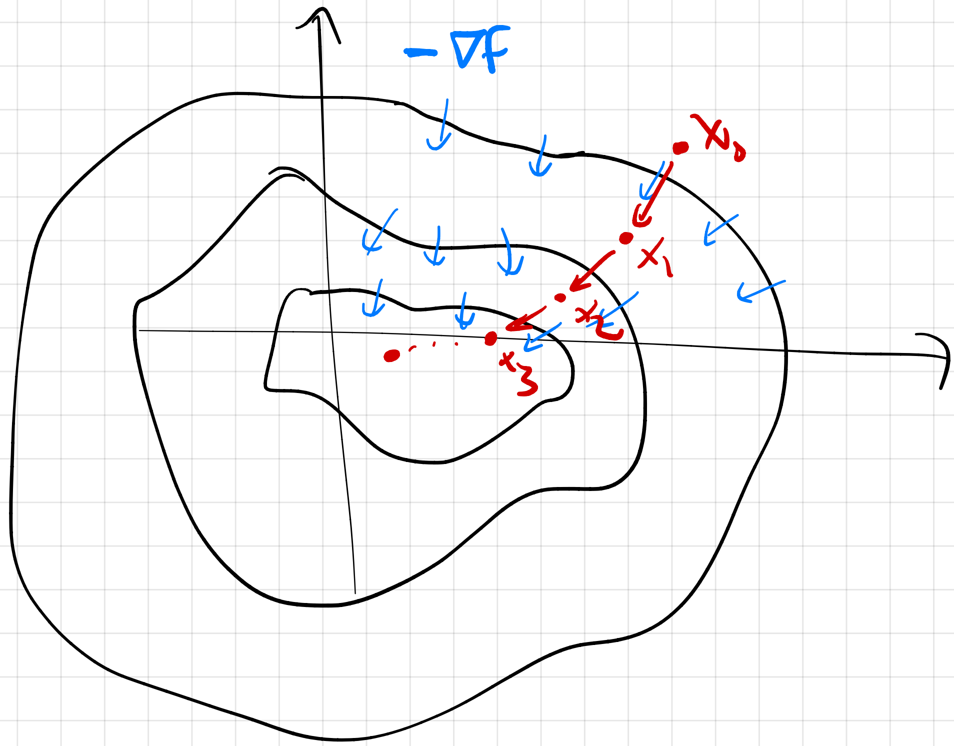
iteration count T

for $t = 1, 2, \ldots, T$:

$$x_t = x_{t-1} - \eta \nabla f_{x_{t-1}}$$

endfor

output $\hat{x} \in \{x_0, x_1, \ldots, x_T\}$ where f attains the smallest value observed.



Analysis when f satisfies: (for some $D, L > 0$)

1. f convex and differentiable
2. f is L-Lipschitz:

$$\forall x, y \quad |f(x) - f(y)| \leq L \cdot \|x - y\|_2$$

3. initial point x_0 satisfies $\|x_0 - x^*\|_2 \leq D$.

   where x^* is a point at which f is minimized.

Analysis will keep track of $\boxed{\Phi(t) \overset{\text{def}}{=} \|x_t - x^*\|_2^2.}$

How to bound $\Phi(t+1)$ given $\Phi(t)$?

Let $x = x_t$.

$$\langle a-b, a-b\rangle = \langle a,a\rangle + \langle b,b\rangle - 2\langle a,b\rangle$$

$$a = x - x^* \quad b = \eta \nabla f_x.$$

$$f(x+y) \geq f(x) + df_x(y)$$

$$\Phi(t+1) = \|x_{t+1} - x^*\|_2^2$$

$$= \langle x_{t+1} - x^*, x_{t+1} - x^* \rangle$$

$$= \langle x - \eta \nabla f_x - x^*, x - \eta \nabla f_x - x^* \rangle$$

$$= \langle x - x^*, x - x^* \rangle + \eta^2 \langle \nabla f_x, \nabla f_x \rangle - 2\eta \langle \nabla f_x, x - x^* \rangle$$

$$\leq \Phi(t) + \eta^2 L^2 + 2\eta \, df_x(x^* - x)$$

$$\leq \Phi(t) + \eta^2 L^2 + 2\eta \left(f(x^*) - f(x) \right)$$

<u>Conclusion:</u> if $f(x) > f(x^*) + \varepsilon$,

$$\Phi(t+1) \leq \Phi(t) + \eta^2 L^2 - 2\varepsilon\eta$$

Set $\eta = \frac{\varepsilon}{L^2}$ $\quad \eta^2 L^2 = \frac{\varepsilon^2}{L^2}$ $\quad 2\varepsilon\eta = \frac{2\varepsilon^2}{L^2}$

$$\Phi(t+1) \leq \Phi(t) - \varepsilon^2 / L^2.$$

$\Phi(0) \leq D^2$. $\quad \Phi(t) \geq 0$.

Within the first $T = D^2 L^2 / \varepsilon^2$ iterations, there must be a t s.t.

$$\Phi(t+1) > \Phi(t) - \varepsilon^2/L^2 .$$

$\Longrightarrow$ at that time t,

$f(x_t)$ was $\leq$ $f(x^*) + \varepsilon$.

$\therefore$ GD with $\eta = \varepsilon/L^2$

$T = D^2L^2/\varepsilon^2$

is guaranteed to reach an ε-optimal point $\hat{x}$.