Gradient Descent Algorithm for Convex function

November 16, 2023

L-Smoothness I

L-Lipschitz

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

L-smoothness a function L-smooth if it is continously differentiable and its gradient is Lipschitz continuous

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

Smoothness assures the gradient cannot change too quickly.

Examples

$f(x) = x^2$ is L-smooth with $L = 2$

$$|f'(x) - f'(y)| = |2x - 2y| \leq 2|x - y|$$

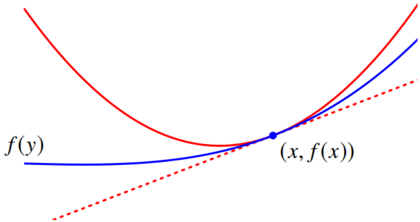
for any x and y.

$f(x) = |x|$ is non-smooth but convex.

L-Smoothness II

Quadratic upper bound.
If f is L-smooth, then for any $x \in \mathbf{intdom}(f)$ and $y \in \mathbf{dom}(f)$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}||y - x||^2$$



L-Smoothness III

Proof: Let $h = y - x$ and define a function $g(t)$ as

$$g(t) = f(x + th) - f(x) - t\nabla f(x)^T h$$

Then g is differentiable and

$$\begin{aligned} g'(t) &= (\nabla f(x + th) - \nabla f(x))^T h \\ &\leq ||\nabla f(x + th) - \nabla f(x)||\,||h|| \\ &\leq Lt||h||^2 \end{aligned}$$

Thus it follows that

$$g(1) = g(0) + \int_0^1 g'(t)dt \leq g(0) + \frac{L}{2}||h||^2(1 - 0) = \frac{L}{2}||h||^2$$

L-Smoothness IV

Note that $g(0) = 0$ and $g(1) = f(y) - f(x) - \nabla f(x)^T h$.

$$f(y) \leq f(x) + \nabla f(x)^T h + \frac{L}{2} \|y - x\|^2$$

Gradient update step to L-smooth function I

Gradient descent

$$x_{k+1} = x_k - t\nabla f(x_k)$$

If f is L-smooth then

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

Insert $y = x - \frac{1}{L}\nabla f(x)$ to the **Quadratic upper bound**

$$\begin{aligned} f(y = x - \frac{1}{L}\nabla f(x)) &\leq f(x) - \frac{1}{L}\langle \nabla f(x), \nabla f(x)\rangle + \frac{L}{2}\|\frac{1}{L}\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \end{aligned}$$

Gradient update step to L-smooth function II

When $x = x^\star$, since $f(x^\star) \le f(y)$ for all y

$$f(x^\star) - f(x) \le f(x - \frac{1}{L}\nabla f(x)) - f(x) \le -\frac{1}{2L}\|\nabla f(x)\|_2^2$$

it means

$$f(x^\star) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

Gradient Descent I

If f is convex and L-smooth, and $x^\star$ is a minimum of f, then for step size $t \in (0, \frac{1}{L}]$, the update sequence

$$x_{k+1} = x_k - t\nabla f(x_k)$$

satisfies

$$f(x_k) - f(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2tk}$$

-

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - t\nabla f(x_k) - x^\star\|^2 \\
&= \|x_k - x^\star\|^2 + t^2\|\nabla f(x_k)\|^2 + 2t\nabla f(x_k)^T(x^\star - x_k)
\end{aligned}$$

Gradient Descent II

- Apply the first-order condition for convexity

$$f(x^\star) \geq f(x_k) + \nabla f(x_k)^T(x^\star - x_k)$$

$$\nabla f(x_k)^T(x^\star - x_k) \leq f(x^\star) - f(x_k)$$

-

$$\frac{t}{2}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

-

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 + 2t(f(x^\star) - f(x_{k+1}))$$

or

$$f(x_{k+1}) - f(x^\star) \leq \frac{\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2}{2t}$$

Gradient Descent III

- Sum over N steps

$$\sum_{k=0}^{N-1}(f(x_{k+1}) - f(x^\star)) \leq \frac{\|x_0 - x^\star\|^2 - \|x_N - x^\star\|^2}{2t} \leq \frac{\|x_0 - x^\star\|^2}{2t}$$

- Since $f(x_{k+1}) \leq f(x_k)$

$$f(x_N) - f(x^\star) \leq \frac{1}{N}\sum_{k=0}^{N-1}(f(x_{k+1}) - f(x^\star)) \leq \frac{\|x_0 - x^\star\|^2}{2tN}$$

Gradient descent for Strongly Convex function I

If f is m-strongly convex and L-smooth, and $x^\star$ is a minimum of f , $\{x_{k+1} = x_k - t\nabla f(x_k)\}$ satisfies

$$f(x_k) - f(x^\star) \leq \frac{L(1-mt)^k}{2}\|x_0 - x^\star\|^2$$

- Note that since $f(x)$ is m-strong convex and L-smooth

$$\frac{m}{2}\|x-y\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{L}{2}\|x-y\|^2$$

- Similar to the proof without strong convexity

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|x_k - t\nabla f(x_k) - x^\star\|^2 \\ &= \|x_k - x^\star\|^2 + t^2\|\nabla f(x_k)\|^2 + 2t\nabla f(x_k)^T(x^\star - x_k)\end{aligned}$$

Gradient descent for Strongly Convex function II

- Apply the first-order condition for m-strong convexity

$$f(x_k)^T(x^\star - x_k) \leq f(x^\star) - f(x_k) - \frac{m}{2}\|x_k - x^\star\|^2$$

- together with

$$\frac{t}{2}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

-

$$\|x_{k+1} - x^\star\|^2 \leq (1 - mt)\|x_k - x^\star\|^2 + 2t(f(x^\star) - f(x_{k+1}))$$

- since $f(x^\star) \leq f(x_{k+1})$

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &\leq (1 - mt)\|x_k - x^\star\|^2 \\ &\leq (1 - mt)^{k+1}\|x_0 - x^\star\|^2\end{aligned}$$

Gradient descent for Strongly Convex function III

- $$\begin{aligned} f(x_{k+1}) - f(x^\star) &\leq \frac{(1-mt)\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2}{2t} \\ &\leq \frac{L(1-mt)^{k+1}}{2}\|x_0 - x^\star\|^2 \end{aligned}$$

Example for Quadratic function I

$$\min_x f(x) = \frac{1}{2} x^T Q x$$

where $Q = \mathrm{diag}(m, L)$, $L > m > 0$, f is L-smooth and m-strongly convex.
The gradient descent step

$$\begin{aligned} x_{k+1} &= x_k - t\nabla f(x_k) \\ &= (I - tQ)x_k \\ &= (I - tQ)^{k+1} x_0 \\ &= \begin{bmatrix} (1 - mt)^{k+1} x_{01} \\ (1 - Lt)^{k+1} x_{02} \end{bmatrix} \end{aligned}$$

Example for Quadratic function II

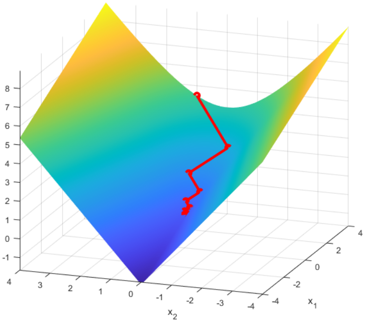
We can normalize Q so that $Q = \mathrm{diag}(1, \gamma)$ and initialize $x_0 = [\gamma, 1]^T$

$$x_k = \begin{bmatrix} (1-t)^k \gamma \\ (1-\gamma t)^k \end{bmatrix}$$

Convergence speed depends on γ

Nondifferentiable Function

$$f(x) = \begin{cases} \sqrt{x_1^2 + \gamma x_2^2} & |x_2| \leq x_1, \\ \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} & |x_2| > x_1 \end{cases}$$



Initialize $x_0 = [\gamma, 1]$, the GD algorithm with exact line search converges to $f(0) = 0$, non-optimal point