Gradient Descent Convergence Rates

$\min_{x \in \mathbb{R}^n} f(x)$; f convex, ∇f Lipschitz from $\|x-x^*\|$ to $\|\nabla f(x)\|$

Gradient Descent

Find that from no minimum (any) at $x_0 \in \mathbb{R}^n$

Find $x_k \in \mathbb{R}^n$ s.t.

$$f(x) - f(x^*) \le \epsilon$$

Algorithm

Choose $x_0 \in \mathbb{R}^n$

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

for $k = 0, 1, \ldots$ and $\gamma > 0$

Convergence Rates:

Plain GD Iterative Optimization methods usually have poor global convergence rates. Best from GD

Let ϵ_k denote the error in iteration k

In minimization, typically

$$\epsilon_k = f(x_k) - f(x^*)$$

Linear Convergence

$$\epsilon_{k+1} \le c\,\epsilon_k \quad \text{for all}$$

sufficient large k

$\Rightarrow$ to hit large accuracy ϵ, we need $O(\log(1/\epsilon))$ iterations

where $O(\log(1/\epsilon))$ iterations

Sublinear Convergence

$$\epsilon_{k+1} \le c\,\epsilon_k^{\alpha} \quad \text{for}$$

sufficiently large k

$\Rightarrow$ to hit accuracy ϵ, we need $O(\log\log(1/\epsilon))$ iterations

Variable Analysis of Gradient Descent

Study order characterization of convexity

f is convex iff

$$f(y) \ge f(x) + \nabla f(x)^T (y-x)$$

for all $x, y \in \mathrm{dom}(f)$

For a convex f:

$$\nabla f(x_1) - \nabla f(x_2) = (x_1 - x_2)^T \nabla^2 f(x_2)$$

Let $g = \nabla f(x)$

Also by (4) $\quad y = x - \frac{1}{L} g$

$$\Rightarrow \frac{1}{L}(x - x_1)^T (x_1 - x_2)$$

$$= \frac{1}{2L}(x - x_1)^T (x_1 - x_2)$$

$$= \frac{1}{2L}(x - x_1)^T + (x_1 - x_2)^T$$

$$= \frac{1}{2L}(x - x_1)^T (x - x_1) + (x_1 - x_2)^T (x_1 - x_2)$$

$$= \frac{1}{2L}\left(\|x - x_1\|^2 + \|x_1 - x_2\|^2 - \|x - x_2\|^2\right)$$

Summing over k

$$\sum_{k=0}^{T} g_k^T (x_k - x^*) \ge \sum_{k=0}^{T} \frac{1}{2L}\|x_k - x^*\|^2 + \frac{1}{2L}\left(\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2\right)$$

$$\ge \sum_{k=0}^{T} \frac{1}{2L}\|x_k - x^*\|^2 - \frac{1}{2L}\|x_0 - x^*\|^2$$

Using (2)

$$\sum_{k=0}^{T} f(x_k) - f(x^*) \le \sum_{k=0}^{T} \frac{1}{2}\|g_k\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 \quad (5)$$

Effectively Convex Functions

A function f is μ-strongly convex

$$f(x) - f(x^*) \le \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall x, y \in X$$

Note: $f \in \mathcal{F}_{\mu,L}$ convex $\forall \mu, L$

$\nabla_x \nabla^2 f(x) \succeq \mu I$

$\nabla_x \nabla^2 f(x) \preceq L I$; $\mu \le L$ $\forall x$

Then combining $\nabla^2 f(x) \succeq \mu I$

Gradient Descent Rate

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

Then (5)

$$\sum_{k=0}^{T} (f(x_k) - f(x^*)) \le \frac{1}{2\mu}\|\nabla f(x)\|^2 + \frac{L}{2}\|x_0 - x^*\|^2$$

We want γ such that

$$\gamma(x) = \sum_{k=0}^{T} \frac{1}{2\mu}\|\nabla f(x)\|^2 + \frac{L}{2}\|x_0 - x^*\|^2$$

is minimized

$$\gamma'(x) = 0$$

$$\Rightarrow \frac{\partial \gamma}{\partial x} = \frac{\partial \gamma}{\partial x} = 0$$

$$\Rightarrow \frac{\partial \gamma}{\partial x} = \frac{\partial \gamma}{\partial x} = 0$$

$$\Rightarrow x^* = \frac{\mu}{L} \Rightarrow \boxed{\gamma = \frac{\mu}{L}}$$

$\gamma(\mathbb{R}^n) = \mathbb{R}\mathbb{R}^n$

Dividing by T, gives the result $\Rightarrow$

$\Rightarrow$ To achieve $\min_{k=0}^{T} (f(x_k) - f(x^*)) \le \epsilon$

$$T \ge \frac{\mu \|x_0 - x^*\|^2}{2\epsilon}$$

For Smooth convex functions:

Let f be differentiable, ∇f is L-smooth

be convex and $L \in \mathbb{R}_+$ f is called L-smooth provided

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x - y\|^2 \quad ; \forall x, y \in \mathbb{R}^n$$

Lemma: Let $f: \mathbb{R}^n \to \mathbb{R}$ be differentiable by L-smooth with $\nabla f = \frac{1}{L}$, gradient descent

gives

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

(sufficient decrease lemma)

Proof:

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

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|\nabla f(x_k)\|^2$$

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

Then: Let $f: \mathbb{R}^n \to \mathbb{R}$ convex, differentiable, has L-smooth

gives $\nabla f = \frac{1}{L}$

Then gradient descent gives

$$f(x_T) - f(x^*) \le \frac{1}{LT}\|x_0 - x^*\|^2, \forall T \ge 0$$

Proof:

$$\frac{1}{2L}\sum_{k=0}^{T} \|\nabla f(x_k)\|^2 \le \sum_{k=0}^{T} (f(x_k) - f(x_{k+1})) = f(x_0) - f(x_T)$$

with $\nabla f = \frac{1}{L}$; (B) gives

$$\sum_{k=0}^{T} (f(x_k) - f(x^*)) \le \frac{1}{2L}\sum_{k=0}^{T} \|\nabla f(x_k)\|^2 + \frac{L}{2}\|x_0 - x^*\|^2$$

$$\le f(x_0) - f(x_T) + \frac{L}{2}\|x_0 - x^*\|^2$$

(from (1))

$$\Rightarrow \sum_{k=0}^{T} (f(x_k) - f(x^*)) \le \frac{L}{2}\|x_0 - x^*\|^2$$

Note $f(x_{k+1}) \le f(x_k)$ for each $0 \le k \le T$ (by sufficient decrease lemma)

taking average

$$f(x_T) - f(x^*) \le \frac{1}{T}\sum_{k=0}^{T} (f(x_k) - f(x^*)) \le \frac{L}{2T}\|x_0 - x^*\|^2 \quad \square$$

Implication: for $\epsilon^2 > \|x_0 - x^*\|^2$

reach $T \ge \frac{\epsilon^2 L}{2\epsilon}$

Smooth & Strongly convex functions

f is convex, differentiable, and let $\mu \in \mathbb{R}_+$ f is called μ-strongly convex

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|x - y\|^2 \quad ; \forall x, y \in X \quad (\text{Same (1)})$$

μ (without that)

Let $f \in \mathcal{F}_{\mu,L}$ be μ-strongly convex, then f is of form

$$f(x) = \frac{\mu}{2}\|x - x_0\|^2 + c \quad \text{where } x_0 \in \mathbb{R}^n, c \in \mathbb{R}$$

Idea: If we have unique x^* $\min_{x \in \mathbb{R}^n} x_0 = x^*$

$$g_k^T (x_k - x^*) = \nabla f(x_k)^T (x_k - x^*) \ge f(x_k) - f(x^*) + \frac{\mu}{2}\|x_k - x^*\|^2$$

(By strong convexity)

$$f(x_k) - f(x^*) \le \frac{1}{2\gamma}\left(\gamma^2 \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x_k - x^*\|^2$$

$$\|x_{k+1} - x^*\|^2 \le 2\gamma^2 (f(x^*) - f(x_k)) + \gamma^2 \|\nabla f(x_k)\|^2 + (1 - \mu\gamma)\|x_k - x^*\|^2 \quad \square$$

Then: Let $f \in \mathcal{F}_{\mu,L}$ convex, differentiable be L-smooth, μ-strong

Choosing $\gamma = \frac{1}{L}$ gradient descent with satisfying x_k satisfies

i) $\|x_{k+1} - x^*\|^2 \le \left(1 - \frac{\mu}{L}\right)\|x_k - x^*\|^2 \to 0$

ii) $f(x_k) - f(x^*) \le \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2 \to 0$

Proof:

By sufficient decrease lemma

$$f(x^*) - f(x_k)$$

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

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

Using $\frac{1}{L} = \frac{1}{\gamma L}$, multiplying by $2\gamma^2$ and rearranging we get

$$2\gamma(x_k + (x^* - f(x_k))) + \gamma^2 \|\nabla f(x_k)\|^2 \le 0$$

By (C)

$$\|x_{k+1} - x^*\|^2 \le \left(1 - \frac{\mu}{L}\right)\|x_k - x^*\|^2$$

$$= \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2$$

and

$$\|x_T - x^*\|^2 \le \left(1 - \frac{\mu}{L}\right)^T \|x_0 - x^*\|^2$$

(Part (i))

For (ii) Using Smooth def

$$\nabla f(x^*) = 0$$

$$f(x_T) - f(x^*)$$

$$\le \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}\|x_T - x^*\|^2$$

$$= \frac{L}{2}\|x_T - x^*\|^2 \quad \square$$

Require $T \ge \frac{L}{\mu} \ln\left(\frac{\epsilon^2}{2\epsilon}\right)$