Notes on Convex Optimization
Gradient Descent on Well-conditioned Functions

Chunpai Wang
cwang25@albany.edu

Feb. 10, 2017

1 Lipschitz Continuous, α-Strongly Convex, and β-Smooth

Denote $K \in \mathbb{R}^d$ a bounded domain, and compact set in Euclidean space, which is bounded and closed. We also denote by D an upper bound on the diameter of domain K,

$$\forall x, y \in K, \quad \|x - y\| \leq D \tag{1}$$

We denote by $G > 0$ an upper bound on the norm of the subgradients of f over K, i.e., $\|\nabla f(x)\| \leq G$ for all $x \in K$. Such an upper bound implies that the function is Lipschitz continuous with parameter G, that is, for all $x, y \in K$

$$|f(x) - f(y)| \leq G\|x - y\| \tag{2}$$

We say f is α-strongly convex, then $\forall x, y \in K$

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{\alpha}{2}\|x - y\|^2 \tag{3}$$

The second derivative of f has lower bound α. The hessian $\nabla^2 f(x)$ has lower bound means the largest eigenvalue of hessian is lower bounded. We say f is β -smooth, then $\forall x, y \in K$

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{\beta}{2}\|x - y\|^2 \tag{4}$$

The second derivative of f has upper bound β. Thus, we have

$$\alpha I \preceq \nabla^2 f(x) \preceq \beta I \tag{5}$$

From this, we can observe that f is β-smooth is equivalent to a Lipschitz condition over the gradients of f,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\| \tag{6}$$

When the function f is both α-strongly convex and β-smooth, we denotes γ as the well-condition number,

$$\gamma = \frac{\alpha}{\beta} \leq 1$$

.

- $f(x) = x$, f is convex and smooth. Check the second derivative, this is not strongly convex.
- $f(x) = x + x^2$, f is convex, strongly convex, and smooth.
- $f(x) = \exp(-x)$, f is convex, but not strongly convex or smooth. Because $f''(x) = \exp(-x)$ has no lower bound.
- Let f be α_1-strongly convex and g be α_2-strongly convex. Then $f + g$ is $(\alpha_1 + \alpha_2)$-strongly convex.
- Let f be β_1-smooth and g be β_2-smooth. Then $f + g$ is $(\beta_1 + \beta_2)$-smooth.

2 Convergence Analysis of Gradient Descent on γ-well Conditioned Function

We denote

$$h_t = f(x_t) - \min_{x \in K} f(x) \tag{7}$$

$$h_1 = f(x_1) - \min_{x \in K} f(x) \tag{8}$$

where x_1 is the initial value of x in gradient descent method. The target of convergence analysis is to compare the value h_{t+1} and h_t or to look for a upper bound of $h_{t+1} - h_t$, then we can conclude the convergence rate based on the upper bound of h_{t+1}, which is associated with t.

Theorem 1. *Consider $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Assume f is γ-well conditioned (i.e. α-strongly convex and β-smooth). Then, the gradient descent algorithm with learning rate $\eta_t = \frac{1}{\beta}$ converges as*

$$h_{t+1} = f(x_{t+1}) - \min_{x \in K} f(x) \leq h_1 \cdot e^{-\gamma t} \tag{9}$$

Proof. First, we have

$$\begin{aligned}
h_{t+1} - h_t &= f(x_{t+1}) - f(x_t) && (10)\\
&= f(x_t - \eta_t \nabla f(x_t)) - f(x_t) && (11)\\
&\leq -\eta_t \nabla f(x_t)^\top \nabla f(x_t) + \frac{\beta}{2}\eta_t^2 \|\nabla f(x_t)\|^2 && (12)\\
&= -\frac{1}{2\beta}\|\nabla f(x)\|^2 && (13)\\
&\leq -\frac{1}{\alpha\beta}(2\alpha h_t) && (14)\\
&= -\frac{\alpha}{\beta} h_t && (15)
\end{aligned}$$

where the inequality (8) is based on β-smoothness, since we are looking for an upper bound. The second inequality is based on the fact that

$$\|\nabla f(x)\|^2 \geq 2\alpha h_t \tag{16}$$

since according to α-strongly convex, we have

$$\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2 && (17)\\
&\geq \min_z \left\{ f(x) + \nabla f(x)(z - x) + \frac{\alpha}{2}\|z - x\|^2 \right\} && (18)\\
&= f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2 && (19)
\end{aligned}$$

where the last equation is because $z = x - \frac{1}{\alpha}\nabla f(x)$.
Since we have $h_{t+1} - h_t = -\frac{\alpha}{\beta} h_t$, we can get

$$\begin{aligned}
h_{t+1} &\leq (1 - \frac{\alpha}{\beta}) h_t && (20)\\
&= (1 - \gamma) h_t && (21)\\
&\leq (1 - \gamma)[(1 - \gamma) h_{t-1}] && (22)\\
&\leq (1 - \gamma)^t \cdot h_1 && (23)\\
&\leq (e^{-\gamma})^t \cdot h_1 && (24)
\end{aligned}$$

where the last inequality is due to the fact

$$1 + x \leq (1 + \frac{x}{n})^n \xrightarrow{n=\infty} e^x \tag{25}$$

We can see that h_{t+1} is upper bounded by $h_1 \cdot e^{-\gamma}$, since γ is positive, we can say it converges exponentially fast, a.k.a. linear convergence rate. □

Now, if we slightly change the learning rate, we can see a significant improvement of convergence rate.

Theorem 2. *For constrained minimization of γ-well conditioned functions and $\eta_t = \frac{\alpha}{2(\beta - \alpha)}$, gradient descent algorithm converges as*

$$h_{t+1} \leq h_1 \cdot e^{-\frac{\gamma t}{4}} \tag{26}$$

Proof. □

3 Examples

Now, we would like to introduce some examples such that we can apply our theorems above to more general functions. However, our approach may not show tighter bound than analyzing GD from scratch.

3.1 Reduction to Smooth, Non-Strongly Convex Functions

Assume f is non-strongly convex, but β-smooth, we cannot apply our theorems directly, and we need to make a new function that is well-conditioned, i.e.

$$g(x) = f(x) + \frac{\tilde{\alpha}}{2}\|x_1\| \tag{27}$$

Now, we can apply gradient descent method with learning rate $\eta_t = \frac{1}{\beta}$ on function g to get a meaningful convergence rate, since g is $\tilde{\alpha}$-strongly convex, and $(\tilde{\alpha} + \beta)$-smooth.

Lemma 3. *For β-smooth convex functions f, apply gradient descent on function g with parameter $\tilde{\alpha} = \frac{\beta \log t}{D^2 t}$ converges as*

$$h_{t+1} = \mathcal{O}(\frac{\beta \log t}{t}) \tag{28}$$

Proof. □

3.2 Reduction to Stronlgy Convex, Non-Smooth Functions

Smoothing cannot be obtained by simple addition of a smooth (or any other) function. Instead, we need a smoothing operation, which amounts to taking a local integral of the function, as follows.
Let f be G-Lipschitz continuous and α-strongly convex. Define for any $\theta > 0$

$$\hat{f}_\theta = E_{v\sim B}[f(x + \theta v)] \tag{29}$$

where $B = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ is the Euclidean ball and $v \sim B$ denotes a random variable drawn from the uniform distribution over B. We will prove that the function $\hat{f}_\theta$ is a smooth approximation to $f : \mathbb{R}^d \to \mathbb{R}$, it is both smooth and close in value to f, as given in the following lemma.

Lemma 4. *$\hat{f}_\theta$ has the following properties:*

1. *If f is α-strongly convex, then so is $\hat{f}_\theta$*
2. *$\hat{f}_\theta$ is $\frac{d \cdot G}{\theta}$-smooth, where d denotes by the dimension of the variable x.*
3. *$|\hat{f}_\theta(x) - f(x)| \leq \theta G$ for all $x \in K$.*

Proof. First, since $\hat{f}_\theta$ is an average of α-strongly convex functions, it is also α-strongly convex. In order to prove smoothness, we will use *Stokes' theorem* from calculus: For all $x \in \mathbb{R}^d$ and for a vector random variable v which is uniformly distributed over the Euclidean sphere $S = \{y \in \mathbb{R}^d : \|y\| = 1\}$

$$E_{v\sim S}[f(x + \theta v)v] = \frac{\theta}{d}\nabla \hat{f}_\theta(x) \tag{30}$$

□

We can see that there is $\nabla \hat{f}_\theta(x)$ in the formula above. Recall that a function f is β-smooth if and only if gradient of f is Lipschitz continuous for all $x, y \in K$, which is

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\| \tag{31}$$

Now, in order to show $\hat{f}_\theta$ is smooth, our target becomes to find the upper bound of second derivative of $\hat{f}_\theta$:

$$\begin{aligned}
\|\nabla \hat{f}_\theta(x) - \nabla \hat{f}_\theta(y)\| &= \frac{d}{\theta}\|E_{v\sim S}[f(x + \theta v)v] - E_{v\sim S}[f(y + \theta v)v]\| && (32)\\
&= \frac{d}{\theta}\|E_{v\sim S}[f(x + \theta v)v - f(y + \theta v)v]\| && (33)\\
&\leq \frac{d}{\theta}E_{v\sim S}\|[f(x + \theta v)v - f(y + \theta v)v]\| && (34)\\
&\leq \frac{d}{\theta}G\|x - y\|E_{v\sim S}[\|v\|] && (35)\\
&= \frac{d \cdot G}{\theta}\|x - y\| && (36)
\end{aligned}$$

where (33) based on linearity of expectation, (34) based on the Jensen's inequality on norm, (35) based on Lipschitz continuity, and the (36) is based on stoke theorem.
The third property, namely that $\hat{f}_\theta$ is a good approximation to f

$$\begin{aligned}|\hat{f}_\theta(x) - f(x)| &= |E_{v\sim B}[f(x+\theta v) - f(x)]| && (37)\\ &\leq E_{v\sim B}[|f(x+\theta v) - f(x)|] && (38)\\ &\leq E_{v\sim B}[G\|\theta v\|] && (39)\\ &\leq G\theta && (40)\end{aligned}$$

where the (37) is based on the definition of $\hat{f}$, (38) is based on the Jensen's inequality, (39) is based on the f is G-Lipschitz, and (40) is based on $v \in B$.

Lemma 5. *Apply gradient descent on $\hat{f}_\theta$, $T, \{\eta_t = \theta = \frac{dG}{\alpha}\frac{\log t}{t}\}$, it converges as*

$$h_t = \mathcal{O}\left(\frac{G^2 d \log t}{\alpha t}\right) \quad (41)$$

Proof. Lemma 4 shows that $\hat{f}_\theta$ is a good approximation to the original function f, and the function $\hat{f}_\theta$ is γ-well-conditioned for $\gamma = \frac{\alpha\theta}{dG}$. Hence,

$$\begin{aligned}h_{t+1} &= f(x_{t+1}) - f(x^*) && (42)\\ &= [f(x_{t+1}) - \hat{f}_\theta(x_{t+1})] + [\hat{f}_\theta(x_{t+1}) - \hat{f}_\theta(x^*)] + [\hat{f}_\theta(x) - f(x^*)] && (43)\\ &\leq \theta G + [\hat{f}_\theta(x_{t+1}) - \hat{f}_\theta(x^*)] + \theta G && (44)\\ &\leq \hat{h}_1 \cdot e^{-\frac{\gamma t}{4}} + 2\theta G && (45)\\ &= \hat{h}_1 \cdot e^{-\frac{\alpha\theta t}{4dG}} + 2\theta G && (46)\\ &= \hat{h}_1 \cdot e^{-\frac{\alpha t}{4dG}\frac{dG}{\alpha}\frac{\log t}{t}} + 2\theta G && (47)\\ &= \hat{h}_1 \cdot e^{-\frac{\log t}{4}} + 2\frac{dG^2}{\alpha}\frac{\log t}{t} && (48)\\ &= \mathcal{O}\left(\frac{G^2 d \log t}{\alpha t}\right) && (49)\end{aligned}$$

□

Lemma 6. *If f is neither strongly convex, not β-smooth, but Lipschitz continuous, applying gradient descent on f will converge as*

$$h_{t+1} = \mathcal{O}(\frac{\log t}{\sqrt{t}}) \quad (50)$$