Notes On Advanced Optimization

Usman Anwar

These notes are a work in progress! I expect to complete them (and may be re-organize them) once I have some time on my hand (presumably in late February). The primary reference for these notes are the wonderful lectures by Constantine Caramanis. The content covered in these informal notes is also covered by Sébastien Bubeck in chapter 3 and 4 of his monograph on convex optimization. https://sites.google.com/site/burlachenkok/ee364b

Contents

1 Convex Function Definitions And Their Equivalence

A function f is said to be convex if and only if its domain is convex and

1. (**Definition 1**) $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for $0 \leq \theta \leq 1$ and $x, y \in \text{Dom} f$

2. **(Definition 2)** $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all y
3. **(Definition 3)** $\nabla^2 f(x) \succeq 0$

Note that definition 2 is only applicable when gradient or sub-gradient of $f(x)$ exists and definition 3 only applies to twice differentiable functions.

1.1 Definition 3 implies Definition 2

We will show this by showing that definition 3 implies that $f(x)$ is monotone and that if $f(x)$ is monotone, then definition 2 must hold.

$$\int_0^t (x-y)^\top \nabla^2 f(tx + (1-t)y) dt = \int_0^t \frac{d}{dt}(\nabla f(t(x-y)+y)) dt \tag{1}$$

$$= \nabla f(x) - \nabla f(y) \tag{2}$$

$$\int_0^t (x-y)^\top \nabla^2 f(tx + (1-t)y)(x-y) dt = (\nabla f(x) - \nabla f(y))^\top (x-y) \tag{3}$$

Note that RHS is of form $x^\top A x$, if $\nabla^2 f(x) \succeq 0$ then RHS must always be greater than 0. This gives us that $f(x)$ is monotone i.e.

$$(\nabla f(x) - \nabla f(y))^\top (x-y) \geq 0 \tag{4}$$

To prove that monotonicity implies definition 2

$$\int_0^1 \frac{d}{dt} f(t(y-x)+x) dt = \int_0^1 \nabla f(t(y-x)+x)(y-x) dt \tag{5}$$

$$f(y) - f(x) = \int_0^1 \nabla f(t(y-x)+x)(y-x) dt \tag{6}$$

$$f(y) = f(x) + \int_0^t h(t) dt \tag{7}$$

where $h(t) = \nabla f(t(y-x)+x)(y-x)$. We note that least value of $h(t)$ will occur for $t = 0$. This is because as we move from x to y on a straight line, because of the monotonicity property, $\nabla f(z)$ will increase if $y > x$ or $\nabla f(z)$ will decrease if $x < y$. Hence,

$$f(y) \geq f(x) + \nabla f(x)(y-x) \tag{8}$$

1.2 Definition 2 implies Definition 3

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) \tag{9}$$

$$f(x+td) \geq f(x) + \nabla f(x)^\top td \tag{10}$$

$$f(x) + t\nabla f(x)^\top d + \frac{t^2}{2} d^\top \nabla^2 f(x) d + o(t^2) \geq f(x) + t\nabla f(x)^\top d \tag{11}$$

$$d^\top \nabla^2 f(x) d + \frac{2}{t^2} o(t^2) \geq 0 \tag{12}$$

Then by letting $t \to 0$

$$d^\top \nabla^2 f(x) d \geq 0 \tag{13}$$
$$\nabla^2 f(x) \succeq 0 \tag{14}$$

1.3 Definition 2 implies Definition 1

$$f(x) \geq f(\theta x + (1-\theta)y) + \nabla f(\theta x + (1-\theta)y)(1-\theta)(y-x) \tag{15}$$
$$f(\theta x + (1-\theta)y) \leq f(x) - \nabla f(\theta x + (1-\theta)y)(1-\theta)(y-x) \tag{16}$$

Similarly,

$$f(\theta x + (1-\theta)y) \leq f(y) + \nabla f(\theta x + (1-\theta)y)(\theta)(y-x) \tag{17}$$

We multiply the first relation by θ and the second by $1-\theta$ and add the above two equations to get

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \tag{18}$$

1.4 Some Other Common Results

Proofs of these results can be found in examples of Boyd's book

- Max of convex functions is convex
- Min of convex functions is not guaranteed to be convex
- Largest element of a vector is convex
- Largest eigenvalue of symmetric matrix is convex

2 Smooth Convex Functions

A convex function $f(x)$ is called β-smooth if its gradient is Lipschitz continuous with parameter β i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x-y\|_2 \tag{19}$$

Intuitively, this means that $f(x)$ can be bounded above by a quadratic (in addition to being bounded by linear function due to standard convexity). We show this in the following two proofs.

Claim 1: For a β-smooth function, $g(x) = \frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex.
Proof: To show this result, we note that a function is convex iff it is monotone. Hence, we will show that $g(x)$ is monotone.

$$(\nabla g(x) - \nabla g(y))^\top (x-y) = (\beta(x-y) - \nabla f(x) + \nabla f(y))^\top (x-y) \tag{20}$$
$$= \beta \|x-y\|_2^2 - (\nabla f(x) - \nabla f(y))^\top (x-y) \tag{21}$$

By using Cauchy Schwartz $u^\top v \leq \|u\|_2 \|v\|_2$

$$
\begin{aligned}
(\nabla g(x) - \nabla g(y))^\top (x-y) &\geq \beta \|x-y\|_2^2 - \|\nabla f(x) - \nabla f(y)\|_2 \|x-y\|_2 && (22)\\
&\geq \beta \|x-y\|_2^2 - \beta \|x-y\|_2 \|x-y\|_2 && (23)\\
&\geq 0 && (24)
\end{aligned}
$$

Hence, $g(x)$ is monotone and threfore convex.

Claim 2: For a differentiable β-smooth convex function

$$
f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{\beta}{2} \|y-x\|_2^2 \tag{25}
$$

Proof: We use the fact that $g(x) = \frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex when f is β-smooth. Using first order condition of convexity on $g(x)$

$$
\begin{aligned}
g(y) &\geq g(x) + \nabla g(x)^\top (y-x) && (26)\\
\frac{\beta}{2}\|y\|_2^2 - f(y) &\geq \frac{\beta}{2}\|x\|_2^2 - f(x) + (\beta x - \nabla f(x))(y-x) && (27)\\
f(y) &\leq f(x) + \nabla f(x)^\top (y-x) + \frac{\beta}{2}\|y\|_2^2 - \frac{\beta}{2}\|x\|_2^2 + \beta\|x\|_2^2 - \beta x^\top y && (28)\\
f(y) &\leq f(x) + \nabla f(x)^\top (y-x) + \frac{\beta}{2}\|y-x\|_2^2 && (29)
\end{aligned}
$$

Claim 3: For a twice differentiable β-smooth convex function

$$
\nabla^2 f(x) \preceq \beta \mathbf{I} \tag{30}
$$

Proof: We again use the fact that $g(x) = \frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex when f is β-smooth. Using second order condition of convexity on $g(x)$

$$
\begin{aligned}
\nabla^2 g(x) &\succeq 0 && (31)\\
\beta - \nabla^2 f(x) &\succeq 0 && (32)\\
\nabla^2 f(x) &\preceq \beta \mathbf{I} && (33)
\end{aligned}
$$

Claim 4 (Strict decrease in $\mathbf{f(x)}$ on gradient step): Unless already converged, given sufficiently small step size η, gradient step results in strict decrease of value of $f(x)$ i.e. $f(x^t) < f(x^{t-1})$ whenver $f(x)$ is first order differentiable and β-smooth convex function.

Proof: We let $x^{t-1} = x$ to avoid clutter.

$$
f(x^t) \leq f(x) + \nabla f(x)^\top (x^t - x) + \frac{\beta}{2}\left\|x^t - x\right\|_2^2 \tag{34}
$$

Using the gradient descent update

$$f(x^t) \le f(x) + \nabla f(x)^\top(-\eta \nabla f(x)) + \frac{\beta}{2} \|\eta \nabla f(x)\|_2^2 \tag{35}$$

$$f(x^t) \le f(x) - \eta \|\nabla f(x)\|_2^2 + \frac{\eta^2 \beta}{2} \|\nabla f(x)\|_2^2 \tag{36}$$

$$f(x^t) \le f(x) - \eta \|\nabla f(x)\|_2^2 + \frac{\eta^2 \beta}{2} \|\nabla f(x)\|_2^2 \tag{37}$$

$$f(x^t) \le f(x) - \eta \left(1 - \frac{\eta \beta}{2}\right) \|\nabla f(x)\|_2^2 \tag{38}$$

For $\eta < \frac{2}{\beta}$, we have $f(x^t) < f(x)$.

Claim 5 (Bound On Suboptimality Of Iterates): If f is β-smooth

$$\frac{1}{2\beta} \|\nabla f(x)\|_2^2 \overset{(a)}{\le} f(x) - f(x^*) \overset{(b)}{\le} \frac{\beta}{2} \|x - x^*\|_2^2 \tag{39}$$

Proof: In order to prove (b), we simply apply the quadratic upper bound property with $y = x$ and $x = x^*$.

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \tag{40}$$

$$f(x) \le f(x^*) + \nabla f(x^*)^\top (x - x^*) + \frac{\beta}{2} \|x - x^*\|_2^2 \tag{41}$$

$$f(x) - f(x^*) \le \frac{\beta}{2} \|x - x^*\|_2^2 \tag{42}$$

To prove (a), we proceed as follows

$$f(x^*) \le f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \tag{43}$$

$$f(x^*) \le \min_y f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \tag{44}$$

$$f(x^*) \le \min_y h(y) \tag{45}$$

In order to minimize LHS, we use optimality condition for unconstrained differentiable convex function

$$\nabla h(y) = 0 \tag{46}$$

$$\nabla f(x) + \beta(y - x) = 0 \tag{47}$$

$$y = x - \frac{\nabla f(x)}{\beta} \tag{48}$$

So

$$f(x^*) \le f(x) + \nabla f(x)^\top \left(-\frac{\nabla f(x)}{\beta}\right) + \frac{1}{2\beta} \|\nabla f(x)\|_2^2 \tag{49}$$

$$\frac{1}{2\beta} \|\nabla f(x)\|_2^2 \le f(x) - f(x^*) \tag{50}$$

Claim 6 (Co-coercivity): If f is β-smooth then $(\nabla f(x) - \nabla f(y))^\top (x-y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|_2^2$.
Proof:

$$f(y) - (f(x) + \nabla f(x)^\top (y-x)) = (f(y) - \nabla f(x)^\top y) - (f(x) - \nabla f(x)^\top x) \tag{51}$$
$$= f_x(y) - f_x(x) \tag{52}$$
$$\tag{53}$$

We note here that $f_x(z) = (f(z) - \nabla f(x)^\top z)$ will also be a β=smooth function if $f(x)$ is β-smooth. Further, we note that $f_x(z)$ is minimized by $z = x$. Therefore, $f_x(y) - f_x(x) \geq \frac{1}{2\beta} \|\nabla f_x(y)\|_2^2$ by previous claim. This gives

$$f(y) - f(x) + \nabla f(x)^\top (y-x)) \geq \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|_2^2 \tag{54}$$

Similarly, we can get

$$f(x) - f(y) + \nabla f(y)^\top (x-y)) \geq \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|_2^2 \tag{55}$$

Adding up these two, we have the following result

$$(\nabla f(x) - \nabla f(y))^\top (x-y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|_2^2 \tag{56}$$

3 Strongly Convex Function

A differentiable function f is called strongly called with paramter α if $g(x) = f(x) - \frac{\alpha}{2} \|x\|_2^2$ is convex.

Intuitively, if a function is strongly convex we can only lower bound it with a quadratic approximation. This can be seen by rewriting the above condition appropriately

$$g(y) \geq g(x) + \nabla g(x)^\top (y-x) \tag{57}$$
$$f(y) - \frac{\alpha}{2} \|y\|_2^2 \geq f(x) - \frac{\alpha}{2} \|x\|_2^2 + \nabla f(x)^\top (y-x) - \alpha x^\top (y-x) \tag{58}$$
$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\alpha}{2} \|y\|_2^2 - \frac{\alpha}{2} \|x\|_2^2 + \alpha \|x\|_2^2 - \alpha x^\top y \tag{59}$$
$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\alpha}{2} \|y-x\|_2^2 \tag{60}$$

If $f(x)$ is twice differentiable strongly convex function, we have the following lower bound on Hessian

$$\nabla^2 f(x) \succeq \alpha \mathbf{I} \tag{61}$$

While often functions are both strongly convex and smooth, this is not necessary.

- Supremum of two convex functions is strongly convex but not smooth.

- Is is also possible to be smooth but not strongly convex. An easy example of this is a piecewise linear function which has been smoothened at the junctions.

Also note that adding or subtracting a linear term does not affect strong convexity and smoothness properties.

Claim 1 (Bound On sub-optimality of any finite x): If f is strongly convex with parameter α

$$\frac{\alpha}{2}\|x-x^*\| \overset{(a)}{\leq} f(x)-f(x^*) \overset{b}{\leq} \frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \tag{62}$$

Proof: First we prove (a) by using the quadratic lower bound property of strongly convex functions.

$$f(x) \geq f(x^*)+\nabla f(x^*)(x-x^*)+\frac{\alpha}{2}\|x-x^*\|_2^2 \tag{63}$$

$$f(x)-f(x^*) \geq \frac{\alpha}{2}\|x-x^*\|_2^2 \tag{64}$$

For proving part (b), we note that the minimum value $f(x^*)$ must be greater than the lowest value of the quadratic lower bound.

$$f(x^*) \geq \min_y f(x)+\nabla f(x)^\top(y-x)+\frac{\alpha}{2}\|y-x\|_2^2 \tag{65}$$

$$f(x^*) \geq f(x)-\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \tag{66}$$

$$f(x)-f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \tag{67}$$

Claim 2 (Coercivity): if f is α-strongly convex then

$$(\nabla f(x)-\nabla f(y))^\top(x-y) \geq \alpha\|x-y\|_2^2 \tag{68}$$

Proof: We recall that if $f(x)$ is strongly convex, then $g(x)=f(x)\frac{\alpha}{2}\|x\|_2^2$ is also convex. Applying the monotonicity of gradient of $g(x)$ easily gives the desired result

$$(\nabla g(x)-\nabla g(y))^\top(x-y) \geq 0 \tag{69}$$

$$(\nabla f(x)-\nabla f(y)-(\alpha x \alpha y))^\top(x-y) \geq 0 \tag{70}$$

$$(\nabla f(x)-\nabla f(y))^\top(x-y) \geq \alpha\|x-y\|_2^2 \tag{71}$$

4 Optimization Algorithms

4.1 Sub-Gradient Method

5 Convergence Rates

5.1 Sub-Gradient Method

The sub-gradient descent algorithm iteratively applies the rule $x_t = x_{t-1}-\eta g_t$ where g_t is the subgradient of $f(x)$ evaluated at x_t. We assume that everywhere sub-gradient is bounded by some value G.

$$\|x_{t+1} - x^*\|_2^2 = \|x_t - \eta g_t - x^*\|_2^2 \tag{72}$$

$$= \|x - x^*\|_2^2 + \eta^2 \|g_t\|_2^2 - 2\eta g_t^\top (x_t - x^*) \tag{73}$$

$$\leq \|x - x^*\|_2^2 + \eta^2 G^2 - 2\eta (f(x_t) - f(x^*)) \tag{74}$$

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta} \left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\eta}{2} G^2 \tag{75}$$

$$f(x_{t-1}) - f(x^*) \leq \frac{1}{2\eta} \left(\|x_{t-1} - x^*\|_2^2 - \|x_t - x^*\|_2^2 \right) + \frac{\eta}{2} G^2 \tag{76}$$

$$\vdots \tag{77}$$

$$f(x_1) - f(x^*) \leq \frac{1}{2\eta} \left(\|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2 \right) + \frac{\eta}{2} G^2 \tag{78}$$

Adding all these relations for $t = 1$ till $t = T$ and then dividing by T, we have

$$\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2\eta T} \left(\|x_1 - x^*\|_2^2 - \|x_T - x^*\|_2^2 \right) + \frac{\eta}{2} G^2 \tag{79}$$

$$f(\frac{1}{T} \sum_{t=1}^{T} x_t) - f(x^*) \leq \frac{1}{2\eta T} \left(\|x_1 - x^*\|_2^2 \right) + \frac{\eta}{2} G^2 \tag{80}$$

$$f(\frac{1}{T} \sum_{t=1}^{T} x_t) - f(x^*) \leq \frac{1}{2\eta T} R^2 + \frac{\eta}{2} G^2 \tag{81}$$

Summary

- If we plan to run sub-gradient method for T iterations, best step size is $\eta \approx \frac{1}{\sqrt{T}}$.
- Error after T iterations $\approx \frac{1}{\sqrt{T}}$. This means that to have error less than or equal to η, we need $\frac{1}{\eta^2}$ iterations.
- While the sub-gradient method is not the descent method, we can still be η close to the the optimal solution if we run the sub-gradient method for $\frac{1}{\eta^2}$ iterations and then **average over all the iterates**.
- Sub-gradient method is dimension free i.e. convergence analysis is only dependent on number of iterations and not on the dimensionality of $f(x)$.

5.2 Projected Sub-gradient Method

Projected sub-gradient method enjoys similar guarantees as the sub-gradient method due to the fact that projection onto a convex set is always a contraction. Specifically, we note that each iteration of projected sub-gradient performs following two computations

- $y_{t+1} = x_t - \eta g_t$

- $x_{t+x} = \mathcal{P}_{\mathcal{X}}(y_{t+1})$, $\mathcal{P}_{\mathcal{X}}$ is the projection operator onto convex set $\mathcal{X}$.

By doing similar calculations as the previous section, we can obtain the following relation with pertinent change highlighted in red

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2\right) + \frac{\eta}{2}G^2 \tag{82}$$

We note that $\|y_{t+1} - x^*\|_2^2 \geq \|x_{t+1} - x^*\|_2^2$, so, we have

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\eta}{2}G^2 \tag{83}$$

This in turn allows using the same analaysis as last section to show that projected sub-gradient method has the convergance rate of $\approx \frac{1}{\sqrt{T}}$.

5.3 Proximal Gradient Algorithm

Proximal operator is an operator associated with a proper, semi-continuous function $f(x)$ and is defined by

$$\text{prox}_{\eta f} f(v) = \underset{x \in \mathcal{X}}{\text{argmin}}\left(f(x) + \frac{1}{2}\|x - v\|_2^2\right)$$

. Consider the optimization problem to have the form $\min_x f(x) + h(x)$ where $f(x)$ is differentiable and $h(x)$ is non-smooth convex function. Then, using prox to denote the proximal operator, we can write the proximal gradient algorithm update as

$$x_{t+1} = \text{prox}_{\eta h}(x_t - \eta \nabla f(x_t)).$$

In the specific case where $h(x) = \mathbb{I}(x)$ i.e. indicator function over convex set $\mathcal{X}$, proximal gradient algorithm and projected sub-gradient method are equivalent.

Note that the proximal gradient algorithm does not uses oracle model of computation but instead utilizes intimate information about the strcture of the problem.

Examples Of Prox Operator

1. ℓ_1-norm: For $f(x) = \|x\|_1 = \sum x_i$

$$\text{prox}_{\eta f}(x) = \underset{u}{\text{argmin}}\, \|u\|_1 + \frac{1}{2\eta}\|u - x\|_2^2 \tag{84}$$

$$\left(\text{prox}_{\eta f}(x)\right)_i = \begin{cases} x_i - \eta & \text{if} \quad x_i \geq \eta \\ 0 & \text{if} \quad |x_i| \leq \eta \\ x_i + \eta & \text{if} \quad x_i \leq -\eta \end{cases} \tag{85}$$

   This proximal operator is ofen called soft thresholding operator or shrinkage operator due to its tendency to pull the value into $[-\eta, \eta]$ range.

2. For $f(x) = \frac{1}{2}x^\top Qx + q^\top x + q_0$ with $Q \succeq 0$,

$$\text{prox}_{\eta f}(x) = (I + \eta Q)^{-1}(x - \eta q)$$

3. For $f(x) = \sum f_i(x_i)$

$$\left(\text{prox}_{\eta f}(x)\right)_i = \text{prox}_{\eta f_i}(x_i)$$

5.3.1 Properties Of Prox Operator

1. **Prox is a contraction**

$$\|\text{prox}_h(x) - \text{prox}_h(y)\| \leq \|x - y\|_2 .$$

2. **Gradient Mapping**
Given $f(x) = g(x) + h(x)$ to minimize where $g(x)$ is smooth; we define a gradient mapping $G_\eta(x) = \frac{1}{\eta}\left(x - \text{prox}_{\eta h}(x - \eta \nabla g(x))\right)$. This lets us write the update for proximal gradient algorithm as $x_{t+1} = x_t - \eta G_\eta(x_t)$.
Note that in general $G_\eta(x) \notin \delta f(x_t)$.

3. **Optimal solutions are the only fixed points of the prox grad update**
Alternatively, we can say that $G_\eta(x) = 0$ iff x minimizes $f(x) = g(x) + h(x)$.

4. g is β-smooth, α-strongly convex function (α can be zero) and $\eta \leq \frac{1}{\beta}$, we have the following lemma

$$f(x - \eta G_\eta(x)) \leq f(z) + G_\eta(x)^\top (x - z) - \frac{\eta}{2}\left\|G_\eta(x)\right\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2 .$$

Proof:

$$f(x - \eta G_\eta(x)) = g(x - \eta G_\eta(x)) + h(x - \eta G_\eta(x)) \tag{86}$$

We note that

$$\begin{aligned} g(x - \eta G_\eta(x)) &\leq g(x) - \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2}\left\|G_\eta(x)\right\|_2^2 && (87) \\ &\leq g(z) - \nabla g(z)^\top (z - x) - \frac{\alpha}{2}\|z - x\|_2 - && (88) \\ &\qquad \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2}\left\|G_\eta(x)\right\|_2^2 && (89) \end{aligned}$$

Also we note that $G_\eta(x) - \nabla g(x) \in \delta h(x - \eta G_\eta(x))$. This result can be obtained by making the observation that that $x - \eta G_\eta(x) = \text{prox}_{\eta h}(x - \eta \nabla g(x))$ and using the identity $u = \text{prox}_{\eta h}(x) \equiv x - u \in \eta \delta h(x)$. Then by using the definition of convexity on h, we have

$$h(x - \eta G_\eta(x)) \leq h(z) - (G_\eta(x) - \nabla g(x))^\top (z - (x - \eta G_\eta(x))) \tag{90}$$

This cumulatively gives us the lower bound on $f(x - \eta G_\eta(x))$:

$$f(x - \eta G_\eta(x)) \leq g(z) - \nabla g(z)^\top (z - x) - \frac{\alpha}{2} \|z - x\|_2 - \tag{91}$$

$$\eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2} \left\|G_\eta(x)\right\|_2^2 + h(z) \tag{92}$$

$$- (G_\eta(x) - \nabla g(x))^\top (z - (x - \eta G_\eta(x))) \tag{93}$$

$$\leq f(z) + G_\eta(x)^\top (x - z) - \frac{\alpha}{2} \|z - x\|_2 - \frac{\eta}{2} \left\|G_\eta(x)\right\|_2^2 \tag{94}$$

5.3.2 Rate Of Convergence

We consider the case where g is $\beta - smooth$ but not strongly convex i.e. $\alpha = 0$.

$$f(x_{t+1}) = f(x_t - \eta G_\eta(x_t)) \leq f(x^*) + G_\eta(x)^\top (x_t - x^*) - \frac{\eta}{2} \left\|G_\eta(x_t)\right\|_2^2 \tag{95}$$

$$f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta} \left[2(\eta G_\eta(x))^\top (x_t - x^*) - \left\|\eta G_\eta(x_t)\right\|_2^2\right] \tag{96}$$

$$\leq \frac{1}{2\eta} \left[\|x_t - x^*\|_2^2 - \left[\|x_t - x^*\|_2^2 - 2(\eta G_\eta(x))^\top (x_t - x^*) + \left\|\eta G_\eta(x_t)\right\|_2^2\right]\right] \tag{97}$$

$$= \frac{1}{2\eta} \left[\|x_t - x^*\|_2^2 - \left\|x_t - x^* - \eta G_\eta(x_t)\right\|_2^2\right] \tag{98}$$

$$= \frac{1}{2\eta} \left[\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right] \tag{99}$$

By telescoping this gives

$$f(x_T) - f(x^*) \leq \frac{1}{T} (\sum_{t=1}^{T} f(x_t) - f(x^*)) \tag{100}$$

$$\leq \frac{1}{2\eta T} \left(\|x_1 - x^*\|_2^2 - x_T - x^{*2}_2\right) \tag{101}$$

$$\leq \frac{1}{2\eta T} \|x_1 - x^*\|_2^2 \tag{102}$$

When the function is also strongly convex i.e. $\alpha \geq 0$, we also have the $\frac{\alpha}{2} \|x_t - x^*\|_2^2$ term, this gives

$$f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta} \left[(1 - \eta\alpha) \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right] \tag{103}$$

If we choose $\eta = \frac{1}{\beta}$, we have

$$\|x_{t+1} - x^*\|_2^2 \leq (1 - \eta\alpha) \|x_t - x^*\|_2^2 \tag{104}$$

$$\|x_{t+1} - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|x_t - x^*\|_2^2 \tag{105}$$

$$\|x_{t+1} - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^t \|x_1 - x^*\|_2^2 . \tag{106}$$

These results show that proximal gradient descent has $\mathcal{O}(\frac{1}{t})$ rate of convergence (as opposed to $\mathcal{O}(\frac{1}{\sqrt{t^2}})$ convergance rate for sub-gradient method).

5.3.3 Iterative Shrinkage Thresholding Algorithm (ISTA)

$$y_i = a_i x + \zeta_i$$

. In general, we need n measurements $n > p$, where p is the dimensionality. But what if we know that x has only s non-zeros i.e. x is sparse. To enforce sparsity, we add ℓ_1 regularization term. Under some technical assumptions, if x is sparse, we only need $n \approx s \log p$ measurements.

$$\hat{x} = \text{argmin} \, \|Ax - y\|_2^2 + \lambda x_1$$

It is easy to show that sub-gradient method would require ϵ^2 steps to get an error of ϵ.

But we can use proximal gradient algorithm as ℓ_1-regularization term has an easy to compute prox operator. We choose $eta = \frac{1}{\beta}$. This gives us the following formulation:

$$x_{t+1} = \underset{x}{\text{argmin}} \, \lambda \|x\|_1 + \frac{\beta}{2} \left\| x - (x_t - \frac{1}{\beta} \nabla g(x_t)) \right\|_2^2 = \text{prox}_{\frac{\lambda}{\beta} \cdot \|\cdot\|_1}(x_t - \frac{1}{\beta} \nabla g(x_t)) \tag{107}$$

$$= \text{prox}_{\frac{\lambda}{\beta} \cdot \|\cdot\|_1}(x_t - 2A^\top(Ax - y)) \tag{108}$$

Thus we have the linear convergance rate, which is a significant improved over plain sub-gradient method.

6 Definitions

6.1 Dual Norm

6.2 Frenchel Conjugate

6.3 Lipschitz Function

A function f is Lipschitz with parameter L wrt norm $\|\cdot\|$ if

$$\|f(x) - f(y)\| \leq L \|x - y\| \tag{109}$$

If f is convex, then the above definition is equivalent to

$$\|\nabla f(x)\|_* \leq L \tag{110}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

6.4 β-Smooth Function

A function f is Lipschitz with parameter L wrt norm $\|\cdot\|$ if

$$\|f(x) - f(y)\| \leq L \|x - y\| \tag{111}$$

If f is convex, then the above definition is equivalent to

$$\|\nabla f(x)\|_* \leq L \tag{112}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.