Summer School in Peking University
Modern Optimization

Professor: Yurri Nesterov
Scribe: Alex

Contents

1 General Nonlinear Optimization

1.1 General formulation of the problem

Let x be an n-dimension real vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$, S be a subset of $\mathbb{R}^n$, $f_i(x), 1 \leq i \leq m$.

$$\begin{aligned} &\min_x f_0(x) \\ &\text{subject to:} \quad f_j(x) \leq 0 \end{aligned} \tag{1}$$

We say $Q = \{x : f_j(x) \leq 0\}$ is the feasible region.

We say a problem is feasible if $Q \neq$, and strictly feasible if there exists x: $f_j(x) < 0$ for every inequality.

We can then come to the problem:

$$\min f_0(x) \quad \text{s.t.} \quad a_j \leq f_j(x) \leq b_j$$

Initialization...

We want to bond Numerical Method $\mathscr{M}$ and Problem $\mathscr{P}$. Best $\mathscr{M}$ for a single problem is easy. So we need:

- Description of a class of problems $\mathscr{P} \subset \mathscr{F}$;
- Description of an oracle $\mathscr{O}$, which provides $\mathscr{M}$ some information of $\mathscr{P}$.

We can define the performance of $\mathscr{M}$ by the worst problem $\mathscr{P}_w$ for $\mathscr{M}$. Performance of $\mathscr{M}$ on $\mathscr{P}$ is the total amount of Computational efforts to solve $\mathscr{P}$ using $\mathscr{M}$.

Exact solution or ϵ-approximate solution.

Generate iterative scheme:

- Starting point: x_0 and accuracy ϵ;
- Initialization: $k = 0$, $l_{-1} = 0$
- l_k is the k-step information of the problem, x_k can be obtained by l_{k-1}.
- Check the stopping condition.

Blackbox Concept:

- The only information available is the answer.
- The oracle is local.

We need to understand the oracle!

- Zero-order oracle: $f(x)$;
- First-order oracle: $f(x)$ and $f'(x)$;
- Second-order oracle: $f(x)$, $f'(x)$ and $f''(x)$

1.2 Uniform Grid Method

Problem Formultaion: $\min f(x)$ for $x \in \mathbb{B}^n$ where $\mathbb{B}^n = \{0 \le x_i \le 1\}$. The Lipchitz condition is:

$$|f(x) - f(y)| \le L\|x - y\|_\infty$$

Scheme of $\mathscr{UG}(p)$: Form p^n points $x_{i_1,\dots,i_n} = (\frac{1}{2p} + \frac{i_1}{p}, \dots, \frac{1}{2p} + \frac{i_n}{p})$

Theorem 1.1 (Upper bound)

If f^* is the minima, and $\bar{x} \in \mathbb{B}^n$, $f(\bar{x}) - f^* \le \frac{L}{2p}$.

Theorem 1.2 (Lower bound)

The analysis complexity is at least $\left[\frac{2L}{\epsilon}\right]^n$.

That means Uniform Grid Method is optimal on our problem class. However, the lower bound means the problem is unsolvable (e.g. if $L = 2$, $\epsilon = 0.01$, $n = 10$, still needs 10^{15} seconds to finish the calculation).

In another field, for example integeration. We can use monte-carlo to estimate.

1.3 Rules of the game

Simulated annealing, neural networks, genetic algorithms...

Our target becomes to find a local minimum and the function is differentiable.

Then we'll focus on the convex optimization, where global minima is available. However the rate depends on dimension.

And at last we will care about structural optimization, where we explore the fast rate converging to the global minimum. The rate depends on the structure of the problem.

2 Local Methods in Unconstrained Optimization

We need to point out that the view is totally different from above, which is more care about optimize with local information instead of the global landscape.

2.1 Preparations

Here our goal is to find the global minimum of the differentiable function $f(x)$ with fast rate. We try to find a sequence $\{x_k\}$ satisfying: $f(x_{k+1}) \le f(x_k)$.

The first order approximation is:

$$f(y) = f(x) + \langle f'(x), y - x\rangle + o(\|y - x\|).$$

We give the definition of the set of directions:

$$S(f, x) = \{s \in \mathbb{R}^n | s = \lim_{y_k \to x, f(y_k) = r} \frac{y_k - x}{\|y_k - x\|}\}.$$

We can observe that: for $s \in S(f, x)$

$$\langle f'(x), s\rangle = 0.$$

In other words, the tangent direction is orthonormal to the gradient direction in geography. This implies that $-f'(x)$ is the direction in which the function decays most rapidly. Further, we can observe that: $\langle f'(x^*), s\rangle = 0$ for every direction s, which implies $f'(x^*) = 0$.

And the second-order approximation is:

$$f(y) = f(x) + \langle f'(x), y - x\rangle + \frac{1}{2}\langle f''(x)(y-x), y-x\rangle + o(\|y-x\|^2).$$

The second-order optimization conditions are:

$$f'(x^*) = 0 \quad \text{and} \quad f''(x^*) \succeq 0.$$

Remark: if further $f''(x^*) \succ 0$, then x^* is strict local minimum.

Now we consider the function whose gradient is Lipchitz continuous (we denote it as $C_L^{2,1}$):

$$\|f'(x) - f'(y)\| \leq L\|x - y\|.$$

It is obvious that $f \in C_L^{2,1}$ iff $f''(x) \leq L$.

Here we give some examples:

- $f(x) = \alpha + \langle a, x\rangle + \langle x, Ax\rangle$, where $f'(x) = a$, and $f''(x) = A$.
- ...

However, we do not always need the $C_L^{2,1}$ condition in our analysis. Similarly, we have:

$$|f(y) - f(x) - \langle f'(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|^2.$$

We may also consider the better smoothness of the function f:

$$\|f''(x) - f''(y)\| \leq M\|x - y\|.$$

Similarly, if $f \in C_L^{2,2}$, we have:

$$\|f'(y) - f'(x) - f''(x)(y - x)\| \leq \frac{M}{2}\|y - x\|^2$$

2.2 Gradient Methods

The scheme is: $x_{k+1} = x_k - h_k f'(x_k)$, $h_k > 0$ is called the stepsize. The rules for stepsize choosing includes:

- Fixed stepsize: $h_k = h$ or $h_k = \frac{h}{\sqrt{k+1}}$
- Full-relaxation: $h_k = \text{argmin}_h f(x_k) - hf'(x_k)$
- Goldstein-Armijo rule: A complex condition to avoid local minimum.

We now care the global convergence. Our main inequality is:

$$f(y) \leq f(x) + \langle f'(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 \leq f(x) - \frac{1}{2L}\|f'(x)\|^2.$$

Now we consider $x_{k+1} = x_k - h_k f'(x_k)$. We can prove that:

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x_k)\|^2.$$

We sum them under $k = 0, ..., N$:

$$\frac{w}{L}\sum_{k=0}^{N}\|f'(x_k)\|^2 \leq f(x_0) - f(x_N) \leq f(x_0) - f(x^*).$$

where w is a constant. We denote that $g_k = \|f'(x_k)\|$ and $g_N^* = \min_k g_k$, then:

$$g_N^* \leq \frac{1}{\sqrt{N+1}}[\frac{L}{w}(f(x_0) - f(x^*))]^{1/2}$$

In practice, we often adopt ϵ-solution, where: $\|f'(x)\| \leq \epsilon$. So the upper complexity bound is $O(\frac{1}{\epsilon^2})$

Now assume that: x_0 approaches to x^*, and the hessian of x^* G can be bounded. Then we can obtain a exponential rate:

$$x_{k+1} = x_k - h_k G_k x_k,$$

where $G_k = \int_0^1 f''(x^* + \tau(x_k - x^*))\,\mathrm{d}\tau$ is the matrix which is near to G. And we can bound $I - h_k G_k$:

$$\alpha I \prec I - h_k G_k \prec \beta I,$$

where $0 < \alpha < \beta < 1$. and that gives:

$$\|x_n - x^*\| \leq \beta^n \|x_0 - x^*\|.$$

We need to point out that the exponential rate $O(\log(\frac{1}{\epsilon}))$ requires the proper initialization and the benign properties of the landscape. (That's not surprising for exponential rate can be obtained under the linear regression regime.)

2.3 Newton Methods

Our motivation is that:

$$\phi(t + \delta t) \approx \phi(t) + \phi'(t)\delta t.$$

We can write this as:

$$t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)}.$$

and turn it into our regime:

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$$

We can replace the $[f''(x_k)]^{-1}$ and introduce the quasi-newton methods. In order to ensure the convergence, we introduce the damped newton method: $x_{k+1} = x_k - \alpha_k [f''(x_k)]^{-1} f'(x_k)$.

Now we talk about the convergence of the newton methods. The idea is mostly in accordance with the exponential rate. We write:

$$x_{k+1} - x^* = G_k [f''(x_k)]^{-1}(x_k - x^*)$$

and prove the bound of the matrix.

3 Smooth Convex Functions

A general smooth functions is hopeless for optimization, and what is a good class? – Convex functions.

Of course there exists many definitions of convex functions:

$$f(y) \leq f(x) + \langle f'(x), y - x\rangle,$$

or if we assume $f \in C^2$:

$$f''(x) \succeq 0,$$

or we don't need the C^1 assumption, we assume $x < y < z$:

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y},$$

which can be written in the form:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

The key property of the convex functions is that if $f'(x^*) = 0$, then $f(x^*)$ is the global minimum, i.e. $f(x) \geq f(x^*)$.

Another important property is that if $f(x)$ is convex, then $\phi(x) = f(Ax + b)$ is convex.

Recall the Lipchitz condition, we can easily get the result:

$$0 \leq f(y) - f(x) - \langle f'(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2. \tag{2}$$

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2L}\|f'(y) - f'(x)\|^2. \tag{3}$$

$$\langle f'(y) - f'(x), y - x \rangle \geq \frac{1}{L}\|f'(y) - f'(x)\|^2. \tag{4}$$

Proof: Express these equalities in their integral forms.

Remark: We can write the Lipchitz condition as: $f''(x) \preceq LI_n$.

Under the conditions now, we can get the $O(\frac{1}{n})$ rate, by iterate the inequality:

$$f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) + \langle f'(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \tag{5}$$

$$\leq f(x_t) - f(x^*) - \frac{1}{2L}\|f'(x_t)\|^2 + \frac{L}{2}(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2). \tag{6}$$

And the lower bound can also be obtained by constructing a "bad" function:

$$f_k(x) = \frac{L}{4}\left\{\frac{1}{2}[(x^{(1)})^2 + \sum_{i=1}^{k-1}(x^{(i+1)} - x^{(i)})^2 + (x^{(k)})^2)] - x^{(1)}\right\}$$

and prove that:

$$\|x_k - x^*\|^2 \geq \frac{1}{8}\|x_0 - x^*\|^2,$$

for $1 \leq k \leq \frac{n-1}{2}$.

Then we introduce the strong convex functions:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2,$$

which can be interpreted as the smallest eigenvalue of the Hessian being μ. We can then prove that $f(x) \leq f(x^*) + \frac{\mu}{2}\|x - x^*\|^2$.

We can obtain the exponential rate under strong convex condition, i.e.:

$$\|x_k - x^*\|^2 \leq \frac{L}{2}(\frac{\kappa - 1}{\kappa + 1})^{2k}\|x_0 - x^*\|^2, \tag{7}$$

where $\kappa = \frac{L}{\mu}$.

Here we introduce a more initial view to understand it – PL condition:

$$\|f'(x)\|^2 \geq 2\mu(f(x) - f(x^*)).$$

and we can observe that:

$$\frac{\mathrm{d}f(x_t)}{\mathrm{d}t} = -\|\nabla f(x_t)\|^2 \leq -\mu f(x_t).$$

Today we recall the traditional convex analysis. The gradient method is not optimal for smooth-convex function space or smooth-strong-convex function space.

4 First Order Optimization

Today we'll focus on finding an optimal method. The motivation is easy – let's combine the 2 methods above together!

Now we introduce the estimation sequences: $\{\phi_k(x)\}$ is called estimating sequence of $f(x)$ if $\lambda_k \to 0$, and for any $x \in \mathbb{R}^n$ and $k \leq 0$ we have:

$$\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x).$$

We can view the definition as $\phi_k(x)$ being bounded by a convex-combination net. And if $f(x_k) \leq \min \phi_k(x)$, then we have:

$$f(x_k) - f^* \leq \lambda_k(\phi_0(x^*) - f^*).$$

The scheme is in the k-th iteration:

- Compute α_k from the equation: $L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$, and $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu$.
- We calculate $y_k = \frac{\alpha_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + \alpha_k\mu}$ compute $f(y_k)$ and $f'(y_k)$.
- Set $x_{k+1} = y_k - h_k f'(y_k)$ such that

$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}f'(y_k)^2.$$

- $v_{k+1} = \frac{1}{\gamma_{k+1}}[(1 - \alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k f'(y_k)]$

Remark: the method combines the 2-step method and momentum. When $\alpha_k = 0$, the method is traditional gradient method.

A natural question is what about the constant step scheme? The scheme is:

- Set $x_{k+1} = y_k - \frac{1}{L}f'(y_k)$.
- Solve the equation:

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}.$$

  and set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$. Then set $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$.

This scheme is more like the momentum (Heavy-Ball) scheme. From physics view we give the mechanical equation

$$\alpha x''(t) = -f'(x) - \beta x'(t),$$

whose discrete form is:

$$\alpha[(x_{k+1} - x_k) - (x_k - x_{k-1})] = -f'(x) - \beta(x_k - x_{k-1}),$$

and simplify:

$$x_{k+1} = x_k + (1 - \frac{\beta}{\alpha})(x_k - x_{k-1}) - \frac{1}{\alpha} f'(x_k).$$

The difference is that the former utilizes the information of 2 steps.

The main result is the rate of convergence:

$$f(x_k) - f(x) \leq L \min \left\{ (1 - \sqrt{\frac{\mu}{L}})^k, \frac{4}{(k+2)^2} \right\} \|x_0 - x^*\|^2. \tag{8}$$

We notice 3 points:

- There are 2 rate of the convergence. Poly-rate works when t is small, while exp-rate works when t is large.
- The κ is optimized to the $\sqrt{\kappa}$. For the β balance the decay rate of different directions.
- The $O(\frac{1}{t^2})$ is in accordance with nesterov momentum rate.

5 General Convex Programming

Our problem can be generally described as:

$$\min_{x \in Q} \{f_0(x) : f_i(x) \leq 0, i = 1, ..., m\},$$

where Q is a convex set and $f_i(x)$ are convex functions. For example, we consider target function $f(x) = \max_{1 \leq j \leq p} \phi_j(x)$, where ϕ_j are convex and p is large.

Today, we'll derive the sub-gradient, which is available in any functional analysis textbook.

As a example we give the derivation of KKT-condition:

Theorem 5.1 (KKT-condition)

The conditions of the convex program

$$\min_{\mathbf{x} \in Q} \{f_0(\mathbf{x}) : f_i(\mathbf{x}) \leq 0, i = 1, ..., m\},$$

in $\mathbf{x} \in \mathbb{R}^N$ and $\boldsymbol{\lambda} \in \mathbb{R}^M$ are:

$$\begin{cases} f_i(\mathbf{x}) \leq 0, \\ \boldsymbol{\lambda} \geq \mathbf{0}, \\ \lambda_m f_m(\mathbf{x}) = 0, \\ \nabla f_0(\mathbf{x}) + \sum_{i=1}^M \lambda_i \nabla f_i(\mathbf{x}) = \mathbf{0}. \end{cases}$$

Proof: First we denote the feasible set $\mathscr{C} = \{x \in \mathbb{R}^N : f_m(\mathbf{x}) \leq 0,\ m = 1, ..., M\}$.

Suppose the KKT conditions hold for $\mathbf{x}^*$ and $\boldsymbol{\lambda}$:

$$\lambda_1 > 0, ..., \lambda_R > 0, \quad \lambda_{R+1} = 0, ..., \lambda_M = 0.$$

Consider $x \in \mathscr{C}$, then

$$f_m(\mathbf{x}^* + \theta(\mathbf{x} - \mathbf{x}^*)) \leq 0 = f_m(\mathbf{x}^*).$$

Thus for $1 \leq m \leq R$:

$$\langle \mathbf{x} - \mathbf{x}^*, \nabla f_m(\mathbf{x}^*) \rangle \leq 0,$$

we have:

$$\langle \mathbf{x} - \mathbf{x}^*, \nabla f_0(\mathbf{x}^*) \rangle \geq 0.$$

Thus:

$$f_0(\mathbf{x}) \geq f_0(\mathbf{x}^*) + \langle \mathbf{x} - \mathbf{x}^*, \nabla f(x^*) \rangle \geq f^*(\mathbf{x})$$

Remark: we can understand it as if index i is useful for restriction, the λ_i works, otherwise, it doesn't work.

6 Methods of Nonsmooth Optimization

First we consider convex and Local-Lipchitz-continuous functions. We consider first order black box, which means $g(x) = \partial f(x)$ is available. Our method is to generate the sequence $\{x_k\}$, where:

$$x_k \in x_0 + \text{Lin}\{g(x_0), ..., g(x_{k-1})\}.$$

Here we consider the worst functions in the world: fix μ as a constant, define $f_k(x) = \gamma \max_{1 \leq i \leq k} x^{(i)} + \frac{\mu}{2}\|x\|^2$, $k = 1, ..., n$. Then:

$$\partial f_k(x) = \mu x + \gamma \text{Conv}\{e_i | i \in I(x)\},$$

where $I(x) = \{j | 1 \leq j \leq k, x^{(j)} = \max_{1 \leq i \leq k} x^{(i)}\}$.

We can easily prove that the f_k is Lipchitz on $B_2(0, \rho)$. And the minimal of the function f_k is:

$$\begin{cases} x_k^{(i)} = -\frac{\gamma}{\mu k}, & 1 \leq i \leq k; \\ x_k^{(i)} = 0, & k+1 \leq i \leq n. \end{cases}$$

Let's look at the optimization process. Assume $x_0 = 0$, and $g_0 = e_1$. Then $x_1 = (*, ..., 0)$, and if "$*$" > 0, $g_1 = e_1$; else, $g_1 = e_2$. Therefore, for $1 \leq i \leq k$, $f_k(x_i) \geq \gamma \max\limits_{1 \leq j \leq k} x_i^{(j)} = 0$. That implies the each component is activated in turns.

We define $\mathscr{P}(x_0, R, M)$ is function space which:

- $\|x - x_0\| \leq R$.
- The Lipchitz constant of the functions is M.

Theorem 6.1 (Lower Bounds)

For any class in $\mathscr{P}(x_0, R, M)$, there exists a function f, such that:

$$f(x_k) - f^* \geq \frac{MR}{2(1 + \sqrt{k+1})},$$

for any methods that:

$$x_k \in x_0 + \text{Lin}\{g(x_0), ..., g(x_{k-1})\}.$$

Proof: Let's choose $f(x) = f_{k+1}(x)$ with $\gamma = \frac{\sqrt{k+1}M}{1+\sqrt{k+1}}$ and $\mu = \frac{M}{(1+\sqrt{k+1})R}$, which satisfies that:

$$\mu R + \gamma = M.$$

Notice that $f^* = -\frac{\gamma^2}{2\mu(k+1)}$ and $f(x_k) \geq 0$, which leads to the proof of the theorem.

Remark: the core of the lower bound is to find that for nonsmooth functions, the gradient method activates the component slowly.

6.1 The Motivation of the Optimization

We now consider f is a convex function, we now use the sub-gradient to optimize the problem. The negative information is: 1) $-g(x)$ may not decrease the $f(x)$; 2)$g \nrightarrow 0$ when $x \to x^*$. The positive information is:

$$\langle g(x), x - x^* \rangle \leq 0.$$

We introduce: $S_k = \{x | \langle x_i - x, g(x_i) \rangle \geq 0, \quad \text{for } i = 0, ..., k\}$ as the localization set of the problem generated by the sequence $\{x_i\}_{i=0}^k$. For all k, we have $x^* \in S_k$.

To have the similar Lipchitz property, we define $\mu_f(x,t) = \max_{\|y-x\| \leq t}(f(y) - f(x))$. And if we define $v(x,y) = \frac{1}{\|g(y)\|}\langle g(y), y - x \rangle$, and we have a substitution of Lipchitz Lemma:

$$f(y) - f(x) \leq \mu_f(x, v_f(x,y)).$$

Proof of the Lemma: 1) $\langle g(y), y - x \rangle \leq 0$, then $f(x) \geq f(y)$;
2) $\langle g(y), y - x \rangle > 0$, then for $z = x + v_f(x,y)\frac{g(y)}{\|g(y)\|}$. We have: $f(z) \geq f(y) + \langle g(y), z - y \rangle = f(y)$,

$$f(y) - f(x) \leq f(z) - f(x) \leq \mu_f(x, \|z - x\|) = \mu_f(x, v_f(x,y)).$$

Remark: The intuition of the lemma is that the directions is non-decreasing in a convex landscape. For:

$$f(y) \leq f(x) + \langle g(y), y - x \rangle = f(x).$$

The corollary of the lemma is for Lipchitz continuous function f (M), we have:

$$f(y) - f(x) \leq M v_f(x,y)_+.$$

6.2 Scheme

Our sub-gradient scheme is:

- Choose $x_0 \in Q$, and $\{h_k\}$ which decays to 0.
- $x_{k+1} = \pi_Q(x_k - h_k \frac{g_k}{\|g_k\|})$ (projection on Q).

We output $f_k^* = \min\limits_{0 \leq i \leq k} f(x_i)$, then with the corollary above, we have:

$$f_k^* - f^* \leq M v_k^*$$

The scheme is almost the "SAM"!

Theorem 6.2 (Upper Bounds)

Let f be Lipchitz continuous on ball $B_2(x^*, R)$ with the constant M. Then:

$$f_k^* - f^* \leq M \frac{R^2 + \sum_{i=0}^k h_i^2}{2\sum_{i=0}^k h_i}.$$

Proof: We define $r_i = \|x_i - x^*\|$, and:

$$r_{i+1}^2 = \|\pi_Q(x_i - h_k \frac{g_i}{\|g_i\|}) - x^*\|^2 \leq \|x_i - h_k \frac{g_i}{\|g_i\|} - x^*\|^2 = r_i^2 - 2h_i v_i + h_i^2.$$

Then we define $v_i = v_f(x_i, x^*)$ and $v_k^* = \min\limits_{0\le i\le k} v_i$:

$$r_0^2 + \sum_{i=0}^{k} h_i^2 \ge 2\sum_{i=0}^{k} h_i v_i + r_{k+1}^2 \ge 2v_k^* \sum_{i=0}^{k} h_i,$$

so we induce that:

$$v_k^* \le M\frac{R^2 + \sum_{i=0}^{k} h_i^2}{2\sum_{i=0}^{k} h_i}.$$

And we use the lemma:

$$f_k^* - f \le Mv_k^*,$$

and derive the results.

Remark: Of course we want $\sum_{k=1}^{\infty} h_k = \infty$ and $\sum_{k=1}^{\infty} h_k^2 \le C$. For example, if we choose $h_i = \frac{r}{\sqrt{i+1}}$, then the upper bound is $\frac{R^2+r^2\log(k+1)}{2r\sqrt{k+1}}$.

6.3 Other Methods

Center of Gravity Method: we define that $cg(S) = \frac{1}{\mathrm{vol}(S)}\int_S x\,\mathrm{d}x$, and $S_{k+1} = \{x|\langle g(x_k), x_k - x\rangle \ge 0\}$.

Our scheme is:

- Choose $x_k = cg(S_k)$;
- Compute $f(x_k), g(x_k)$;
- Compute $S_{k+1} = \{x|\langle g(x_k), x_k - x\rangle \ge 0\}$.

Our main results is

Theorem 6.3 (Upper Bounds)

if f is M-Lipchitz and convex on the ball $B_2(x^*, D)$, we have:

$$f_k^* - f^* \le MD\left(1 - \frac{1}{e}\right)^{\frac{k}{n}}.$$

Proof sketch: We prove the 2 results below:

- $v_k^* \le D\left(\frac{\mathrm{vol}(S_k)}{\mathrm{vol}(Q)}\right)^{\frac{1}{n}}$;
- $\frac{\mathrm{vol}(S_{k+1})}{\mathrm{vol}(S_k)} \le 1 - \frac{1}{e}$.

The method is optimal in finite dimension, and the rate depend on n (CoD). It's impractical for $cg(S)$ is hard to compute.

6.4 Smoothing Technique

We recall that our main difficulties are:

- $g \in \partial f(x)$ is not a descent direction at x.
- $g \in \partial f(x^*)$ does not imply $g = 0$.

The subgradient method is simple and has low memory requirements, however the convergence rate is $O(\frac{1}{\sqrt{k}})$, which is slow!

Observe that the smooth minimization complexity is $O(\frac{L(f)}{k^2})$. We need to develop a method to smooth the nonsmooth functions.

For function f, we define its Fenchel conjugate:

$$f_*(s) = \max_{x\in\mathbb{R}^n}[\langle s, x\rangle - f(x)].$$

The conjugate is well-defined for:

$$f(x) = \max_{s}[\langle s, x\rangle - f_*(s)],$$

which means $(f_*(s))_* = f(x)$. (An intuitive understanding is: for s, x, we have $s = f'(x)$ and $x = f'_*(s)$.) The $f_*(s)$ is a closed function with $\mathrm{dom} f_* = \{f'(x) : x \in \mathbb{R}^n\}$.

We then define:

$$f_\mu(x) = \max_{s\in\mathrm{dom} f_*}[\langle s, x\rangle - f_*(s) - \frac{\mu}{2}\|s\|^2].$$

if $\mu = 0$, then $f_\mu = f$. The μ is to describe the degree of the smoothing.

We find that $f'_\mu(x) = s_\mu(x)$ and $x = f'_*(s_\mu(x)) + \mu s_\mu(x)$ (the "$\mu s_\mu(x)$" is important here to balance the bad derivation!) .Denote x^1, x^2, $s^i = f'_\mu(x^i)$, and:

$$\|x^1 - x^2\|^2 = \|f'_*(s^1) - f_*(s^2)\|^2 + 2\mu\langle f'_*(s_1) - f'_*(s^2), s^1 - s^2\rangle + \mu^2\|s_1 - s^2\|^2$$

which means $f_\mu \in C^{1,1}_{1/\mu}$. And the difference between f and f_μ is not big:

$$f(x) \le f_\mu(x) \le f(x) - \mu D^2,$$

where $D = \mathrm{diam}(\mathrm{dom} f_*)$.

If we can find a smooth ϵ-approximation $f_\epsilon(x)$ with $L(f_\epsilon) = O(\frac{1}{\epsilon})$. We need a convenient algorithm.

6.5 Adjoint Problem

For example: Consider $f(x) = \max_{1\le j\le m}|\langle a_j, x\rangle - b_j|$. We can write the optimization into a continuous version:

$$f(x) = \max_{u\in\mathbb{R}^m}\left\{\sum_{i=1}^m u_i(\langle a_j, x\rangle - b_j) \quad : \quad \sum_{i=1}^m |u_i| \le 1\right\}.$$

From this intuition we can formulate the general form:

$$f(x) = \hat{f}(x) + \max_u\{\langle Ax, u\rangle_2 - \hat{\phi}(u) : u \in Q_2\}$$

where:

- $\hat{f}(x)$ is differentiable and convex on Q_1.
- Q_2 is closed convex and bounded.
- $\hat{\phi}(u)$ is continuous function on Q_2.
- A is a linear operator.

We introduce the prox function $d_2(u)$ where:

$$d_2(v) \geq d_2(u)\langle \nabla d_2(u), v-u\rangle + \frac{1}{2}\sigma_2\|v-u\|_2^2$$

Fix μ, we define:

$$f_\mu(x) = \max_u[\langle Ax, u\rangle_2 - \hat{\phi}(u) - \mu d_2(u)].$$

Theorem 6.4

$f_\mu(x)$ is convex and differentiable, where:

$$L(f_\mu) = \frac{1}{\mu\sigma_2}\|A\|_{1,2}^2,$$

where $\|A\|_{1,2} = \max_{x,u}\{\langle Ax, u\rangle : \|x\|_1 = 1, \|u\|_2 = 1\}$.

7 Self-concordant Functions

Motivation: the core is to

- Choose a basic method (e.g. Newton Method)
- Describe a set of problems for which the basic method is very efficient.
- Describe a class of problems for which their models can be created in a computable form.

Here we focused on the Newton Method. Recall that:

$$x_{k+1} = x_k - [f''(x_k)]^{-1}f'(x_k).$$

Our standard theorem is: If we assume that:

- $f''(x^*) \succeq lI_n$.
- $\|f''(x) - f''(y)\| \leq M\|x-y\|$.
- Close initialization: $\|x_0 - x^*\| \leq \bar{r} = \frac{2l}{3M}$.

Then:

$$\|x_{k+1} - x^*\| \leq \frac{\|x_k - x^*\|}{2(l - M\|x_k - x^*\|)}$$

Observe if $f \in C^3$, we can write the third assumption in a new perspective:

$$\|f'''(x)[u]\| \leq M\|u\|,$$

which means:

$$|\langle f'''(x)[u]v, v\rangle| \leq M\|u\|\|v\|^2.$$

Note: The $f'''(x)$ is a tensor $\in R^{n\times n\times n}$.

We denote:

$$Df(x)[u] = \langle f'(x), u\rangle, \tag{9}$$
$$D^2f(x)[u] = \langle f''(x)u, u\rangle, \tag{10}$$
$$D^3f(x)[u] = \langle f'''(x)[u]u, u\rangle, \tag{11}$$

and

$$\|u\|^2_{f''(x)} = \langle u, f''(x)u\rangle,$$

and

$$\|u\|_x = \langle u, f''(x)u\rangle$$

and

$$\lambda_f(x) = \langle [f''(x)]^{-1} f'(x), f'(x)\rangle.$$

We call a function is self-concordant function if:

$$D^3 f(x)[u,u,u] \leq M_f \|u\|^3_{f''(x)}.$$

that means the landscape is continuous enough for Newton-methods. In specific, we can intuitively understand the denotation by observing $f(x) = \langle x, A(x)x\rangle$, where $A(x)$ is Lipchitz-continuous.

We here present some properties of the self-concordant functions:

- If $f_i(x)$ are self-concordant, then $\alpha f_1(x) + \beta f_2(x)$ is also self-concordant.
- If f is self-concordant, then $\phi(x) = f(Ax)$ is self-concordant.
- For any point $\mathbf{x} \in \partial(\mathrm{dom} f)$, $x_k \to \mathbf{x}$, then $f(x_k) \to \infty$. (The intuition is that, if there is an upper bound, then $f(x_k)$ has a limit, but the $(\mathbf{x}, \bar{f}) \notin \mathrm{epi}(f)$.)
- Let f is self-concordant. If $\mathrm{dom}(f)$ contains no straight line. Then $f''(x)$ is non-degenerate for any $x \in \mathrm{dom}(f)$. (Otherwise the straight line is in accordance with the degenerate component.)

Conclusion: we focus on a local setting, where the initialization is close to the target and the hessian didn't change acutely. The advantage is we don't require a global continuity of the function. The continuity is a property that inherits the properties from some points, and the convexity points the way of the global minima.

Now we give the derivation: First, the hessian is continuous:

$$(1 - \|y - x\|_x)^2 f''(x) \preceq f''(y) \preceq \frac{1}{(1 - \|y - x\|_x)^2} f''(x).$$

Then we calculate the integral of the hessian:

$$\langle f'(y) - f'(x), y - x\rangle \geq \frac{\|y - x\|^2_x}{1 + \|y - x\|_x},$$

and further we have:

$$f(y) - f(x) \geq \langle f'(x), y - x\rangle + w(\|y - x\|_x).$$

On the other hand:

$$\langle f'(y) - f'(x), y - x\rangle \leq \frac{\|y - x\|^2_x}{1 - \|y - x\|_x},$$

and further we have:

$$f(y) - f(x) \leq \langle f'(x), y - x\rangle + w_*(\|y - x\|_x).$$

Here $w(x) = x - \log(1+x)$ and $w_*(x) = -x - \log(1-x)$. Putting them into the optimization formula and we have:

$$\begin{aligned} f(y) &\geq f(x) + \langle f'(x), y - x\rangle + w(\|y - x\|_x) && (12)\\ &\geq f(x) - \|f'(x)\|_x \|y - x\|_x + w(\|y - x\|_x) && (13)\\ &= f(x) - \lambda_f(x)\|y - x\|_x + w(\|y - x\|_x) && (14) \end{aligned}$$

So the level set $L_f(f(x)) = \{y|f(y) \leq f(x)\}$ is bounded, which implies the minimum of $f(x)$ x_f^* exists. And we have:

$$f(y) \geq f(x_f^*) + w(\|y - x_f^*\|_x)$$

Now we introduce the most important scheme:

$$x_{k+1} = x_k - \frac{1}{1 + \lambda_f(x_k)}[f''(x_k)]^{-1} f'(x_k)$$

which has the intuition that if the landscape is more steep we adopt more cautious step. We have:

$$f(x_{k+1}) \leq f(x_k) - w(\lambda_f(x_k)).$$

Our main result is:

Theorem 7.1 (Convergence)

Let $\lambda_f(x) < 1$, and we have:

$$w(\lambda_f) \leq f(x) - f(x_f^*) \leq w_*(\lambda_f) \tag{15}$$

$$w'(\lambda_f) \leq \|x - x_f^*\| \leq w'_*(\lambda_f) \tag{16}$$

The λ_f is converging and x is converging.
In conclusion, the method can be realized in a 2-stage dynamics:

- First stage: $\lambda_f(x_k) \geq \beta$, we adopt the damped Newton Method, and $f(x_{k+1} \leq f(x_k) - w(\beta)$. The first stage will end in bounded steps.
- Second stage: $\lambda_f(x_k) \leq \beta$, and $\lambda_f(x_{k+1}) < \lambda_f(x_k)$ for some β.

However if we take $f(x) = \langle x, Ax\rangle$, we can find that $\lambda_f(x) = f(x)$, the scheme doesn't align with our intuition.

7.1 Central Path

Define the penalty function:

$$f(t; x) = t\langle c, x\rangle + f(x),$$

with $t \geq 0$. Note that $f(t; x)$ is self-concordant in x. Define:

$$x^*(t) = \arg \min_{x \in \mathrm{dom} f} f(t; x).$$

The trajectory is called Central Path of the problem: $\min_x\{\langle c, x\rangle | x \in Q\}$. Maybe it relates to the string method of minimum free-energy path.

We can immediately have: $x^*(t) \to x^*$ as $t \to \infty$. And we want to follow the central path. Assume we have $x = x^*(t)$ and we want to increase t to $t_+ = t + \Delta t$. We want the x is still in the region of quadratic convergence:

$$\lambda_f(x_{t+\Delta t};)(x) \leq \beta \leq \bar{\lambda}$$

Notice that:

$$\lambda_f(x_{t+\Delta t};)(x) = \|t_+ c + f'(x)\| = |\Delta|\|c\|_x = \frac{|\Delta|}{t}\|f'(x)\|_x \leq \beta.$$

So we need to assume the value $\|f'(x)\|_x^2 = \langle [f''(x)]^{-1} f'(x), f'(x)\rangle$ is uniformly bounded on $\mathrm{dom} f$.

Here we give the definition of Self-concordant Barriers: If the set $\mathrm{dom}F$ satisfies:

$$\max_{u\in\mathbb{R}^n} \quad [2\langle F'(x), u\rangle - \langle F''(x)u, u\rangle] \leq \nu.$$

We notice that if $F''(x)$ is nondegenerate, then $\|f'(x)\|_x^2 = \langle [f''(x)]^{-1}f'(x), f'(x)\rangle \leq \nu$ as we let $u = F''(x)^{-1}F'(x)$. The definition can also induce the result:

$$\langle F'(x), u\rangle^2 \leq \nu\langle F''(x)u, u\rangle.$$

The path-following scheme gives the sample complexity as: $O(\sqrt{\nu}\log\frac{1}{\epsilon})$.

Conclusion: the method is used to make the attraction region affine invariant, and be related to the function instead of a fixed region. The main ingredient of the central path is to increase t and ensure every step the point is in the attraction region.

It is related to the function of the punishment.

8 Applications to Problems with Explicit Structure

A widely admitted fact is: in the optimization, proving general bounds makes no sense. We should focus on describing the problems precisely and providing effective algorithms for a specific class of problems. The modern optimization is highly related to the tasks and data.

8.1 Linear Programming

The problem formulation is:

$$\min_{x\in\mathbb{R}^n}\{\langle c, x\rangle | Ax = b\}.$$

We choose the barrier: $F(x) = -\sum_{i=1}^n \log x^{(i)}$, $\nu = n$. The path-following scheme gives the complexity $O(\sqrt{n}\log\frac{1}{\epsilon})$.

8.2 Quadratic Programming

The problem formulation is:

$$\min_{q_i(x)=\alpha_i+\langle a_i,x\rangle+\langle A_ix,x\rangle\leq\beta_i}\{\alpha_0 + \langle a_0, x\rangle + \langle A_0x, x\rangle\}.$$

We choose the barrier: $F(x;t) = -\sum_{i=1}^m \log(t_i - q_i(x)) - \sum_{i=1}^m \log(\beta_i - t_i)$, $\nu = 2m+1$. The path-following scheme gives the complexity $O(\sqrt{2m+1}\log\frac{1}{\epsilon})$.

8.3 Semi-definite Programming

The problem formulation is:

$$\min_{X\in\mathbb{P}^n}\{\langle C, X\rangle_F : \langle A_i, X\rangle_F = b_i, \quad i = 1, ..., m\}.$$

We choose the barrier: $F(X) = -\log\det X$, $\nu = n$. The path-following scheme gives the complexity $O(\sqrt{n}\log\frac{1}{\epsilon})$.

9 Conclusion

Our tour begins with the local methods such as gradient descent and newton methods, which utilizes the continuity of our target functions. However, we can't ensure the global convergence and the rate is usually $O(\frac{1}{\sqrt{T}})$.

Then we dive into convex optimization, which is more widely applied. We can immediately get $O(\frac{1}{T})$, and we have the exponential rate if strong-convexity holds. In the momentum with flexible step size, we can obtain $O(\frac{1}{T^2})$ rate.

Then we remove the smooth condition, where we provide 2 methods: sub-gradient method and smoothing technique.

At last we recall the Newton Method, and provide a function class. When the initialization is close to the target, the hessian is close to the target hessian. The analysis adopts the landscape theory, which has deeper insight in the optimization. We also introduce the central path theory, which can be applied to many specific problems.