3007 Notes: Algorithm

Xydd

July 2025

Following Notes will include both unconstrained and constrained convex optimization problem, i.e.,

$$\min_x f(x) \quad s.t. \quad x \in \Omega$$

where the Ω here is R^n for unconstrained problem.

1 Gradient Descent

1.1 Direction

First, I want to talk about the initial point for gradient descent, Mean Value Theorem or first order Taylor,

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x) d + o(d)$$

Here we just neglect, $o(d)$, then with $\alpha > 0$, to let $f(x^{k+1}) > f(x^k)$, if $d = -\nabla f(x^k)$ the RHS must smaller then LHS. So we have,

$$d^k = -\nabla f(x^k)$$

as our direction. Notice as $\nabla f(x)^T d \leq 0$, the d^k generated from this method must be a descent direction for Convex Problem

1.2 Step Size

So, intuitively we now have negative gradient as our direction, what about step size?

You may choose a fixed step size for simplicity, i.e., $\alpha^k = \alpha \quad \forall k$. But to make it shrink down fast and accurately, we have 2 method: exact line search and backtracking method.

For the first one, our intuition is to shrink thing down the fastest. so having x^k, d^k, we want,

$$\alpha^k = arg\min_{\alpha \geq 0} f(x^k + \alpha d^k)$$

Remark 1: this method depend on how we solve above equation, it may not have closed form solution!
Remark 2: in this method, highly likely, $d^k \perp d^{k+1}$, as $\nabla_{\alpha^k} f(x^k + \alpha^k d^k) = 0$, so that $\nabla f(x^k + \alpha^k d^k)^T d^k = -(d^{k+1})^T d^k = 0$.

For the second backtracking, before we start, I want to first introduce Some Condition, define $\phi(\alpha) = f(x^k + \alpha d^k)$. First

$$\phi(\alpha) < \phi(0) = f(x^k)$$

But this is only an necessary condition to let $f(x^{k+1}) < f(x^k)$, to let it down faster, we will have tangent line $l(\alpha) = f(x^k) + \nabla f_k^T d^k \alpha$, but most likely we will have, $l(\alpha) < \phi(\alpha)$, so the solution is an empty set. As a compromised way, people use,

$$\phi(\alpha) <= f(x^k) + C\nabla f_k^T d^k \alpha \quad \forall C \in (0, 1)$$

to find good α and this condition is called **Armijo condition**.

To find specific α, we just need to have an $\alpha^{(0)}$ and keep multiple $\sigma \in (0, 1)$, until it satisfy Armijo condition.

Remark: here we need to pick $\alpha^{(0)}, \sigma, C$, and in GD, people usually set $\alpha^{(0)} = 1$

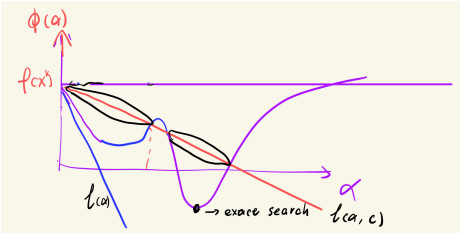

Figure 1: Geometric Meaning

1.3 Convergency

On the above, I talked main procedure of GD, here comes theoretical part about why and when this algorithm converges.

Okay, so before we start, **Lipschitz Contentious** and **Descent Lemma** need to be talked. It is said that, $\forall x, y \in R^n$, if $\exists L \geq 0$ such that,

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

then $\nabla f(x)$ is **Lipschitz Contentious** and f is called **Lipschitz smooth**.

Second about Descent Lemma, give $\phi(\alpha) = f(x + \alpha d)$, $y = x + d$, then f(y) $= \phi(1)$, and assume f is Lipschitz smooth. Notice,

$$\int_0^1 \phi'(\alpha) - \phi'(0) d\alpha = \phi(1) - \phi(0) - \phi'(0)$$

So, we have,

$$f(y) = f(x) + \nabla f(x)^T d + \int_0^1 (\nabla f(x + \alpha d) - \nabla f(x))^T d\mathbf{d}\alpha$$

$$\begin{aligned}
f(y) &= f(x) + \nabla f(x)^T d + \int_0^1 (\nabla f(x + \alpha d) - \nabla f(x))^T d\mathbf{d}\alpha && (1)\\
&\leq f(x) + \nabla f(x)^T d + \int_0^1 ||\nabla f(x + \alpha d) - \nabla f(x)|| * ||d|| \mathbf{d}\alpha && (2)\\
&\leq f(x) + \nabla f(x)^T d + ||d|| \int_0^1 L||\alpha d|| \mathbf{d}\alpha \quad \text{(Lipschitz smooth f)} && (3)\\
&= f(x) + \nabla f(x)^T d + \frac{L}{2}||d||^2 && (4)\\
&= f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}||y - x||^2 && (5)
\end{aligned}$$

Done proof, So if we have f is Lipschitz smooth with L, then we have,

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

This is so called **Descent Lemma**.

Then we will apply these 2 properties to give out some fact. For some constant step size α, $x^{k+1} - x^k = -\alpha \nabla f(x^k)$, then,

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

So if we have fixed $\alpha \in (0, \frac{2}{L})$ with f is Lipschitz smooth, the fixed step size method will always decrease.

Question: Above proof only shown decreasing property, but does it converge to $f(x^*)$? The question still remains.

It's not hard to proof convergency, $f(x^k) - f(x^{k+1}) \geq c||\nabla f(x^k)||^2$, so

$$\sum_{k=0}^{\infty} f(x^k) - f(x^{k+1}) \geq \sum_{k=0}^{\infty} c||\nabla f(x^k)||^2$$

$$\frac{f(x^0 - x^*)}{c} \geq \sum_{k=0}^{\infty} ||\nabla f(x^k)||^2$$

By property of convergent sequence, we have $\lim_{t\to\infty} ||\nabla f(x^k)||^2 = 0$, so $\lim_{t\to\infty} \nabla f(x) = 0$. Work Done!

So, as a conclusion, if we have fixed $\alpha \in (0, \frac{2}{L})$ with f is Lipschitz smooth, the fixed step size method will always converge.
Remark 1: convergence with rate $\mathcal{O}(\frac{1}{k})$ under this case, which is called sub-linear convergence.

If you want it to convergent faster as linear, you need the f to be strongly convex and Lipschitz smooth, then for constant/ exact/ armijo linear search, the question converges to unique global minimum. Since the proof is too complicated for this course, I will not mention it here.

2 Newton's Method

2.1 General Cases

As learned in high school, $x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}$, here the idea is just the same with high dimensional Taylor expansion.

$$f(x + \Delta x) = f(x) + \nabla f(x)^T \Delta x + o(|\Delta x|)$$

$$\nabla f(x + \Delta x) = \nabla f(x) + \nabla^2 f(x)\Delta x + o(|\Delta x|)$$

Let $\Delta x = x^* - x$, neglect small-oh term,

$$\nabla f(x^*) = 0 = \nabla f(x) + \nabla^2 f(x)(x^* - x)$$

$$x^* = x - \nabla^2 f(x)^{-1} \nabla f(x)$$

So we will use $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$.

About step size α, here we usually apply backtracking.

Remark 1: Notice if f is convex, then d^k in newton's method must be a descent direction. ($\nabla^2 f(x) \succeq 0$)

Remark 2: Here we assume $\nabla^2 f(x)$ is invertible.

Remark 3: The convergency for this method depend badly on the initial point and the rate is quadratic, much better compare to linear GD method. But regard the time complexity in computing inverse Hessian, we prefer GD in high-dimensional cases.

Remark 4: For maximize problem, GD changes its sign but Newton does not, ($\nabla f(x)^T d = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \geq 0$)

2.2 Quasi-Newton and BFGS

The the last part, I mention the time complexity of inverse Hessian, to improve it, some approximate methods will be introduced. (no detail in MAT3007 Course)

Replace $\nabla^2 f(x)^{-1}$ with H_k,

$$x^{k+1} = x^k - \alpha^k H_k \nabla f(x^k)$$

For this H_k, we have,

$$H_{k+1} = H_k - \frac{d_k g_k^T H_k + H_k g_k d_k^T}{d_k^T g_k} + (1 + \frac{g_k^T H_k g_k}{d_k^T g_k}) \frac{d^k d_k^T}{d_k^T g_k}$$

where $g_k = \nabla f(x^{k+1}) - \nabla f(x^k), \quad d_k = x^{k+1} - x^k$. Above is so called BFGS Method. (Broyden–Fletcher–Goldfarb–Shanno).

3 Projected Gradient

Notice above we only talked no constraint optimization, now assume the question is,

$$\min_x f(x)$$
$$s.t \quad x \in \Omega$$

It is possible that our $x^{(k+1)}$ is not inside Ω, so we will use euclidean projection to project it back. So now we first find $y = P_\Omega(x)$,

$$\min_y \frac{1}{2}||y - x|| \quad s.t. \quad y \in \Omega$$

Some examples: Box/ Linear/ Ball.

3.1 How and Why this works

So our general idea is to let $x^{k+1} = P_\Omega(x^k - \lambda^k \nabla f(x^k))$.
Notice that,

$$x^{k+1} = x^k + [P_\Omega(x^k - \lambda^k \nabla f(x^k)) - x^k]$$

Set $d^k = P_\Omega(x^k - \lambda^k \nabla f(x^k)) - x^k$. Given $\alpha \in (0, 1)$,

$$x^{k+1} = (1 - \alpha)x^k + \alpha P_\Omega(x^k - \lambda^k \nabla f(x^k)) = x^k - \alpha d^k$$

So if we can guarantee convexity of Ω and $x^k \in \Omega$, we can say that this x^{k+1} is also inside Ω. And Above method is so called **Projected Gradient**

Theorem (Feasibility for Projected Gradient): If Ω is convex and $x^0 \in \Omega$, then for all iterations k, x^k will be inside this feasible region Ω given step size

$\in (0,1)$.

Next about the descent property, similar with GD part, we need first proof some lemmas.

Theorem (FONC for Problems with Convex Constraints): if f is differentiable on a open set contains Ω which is convex and closed set, given x^* as a local minimizer, then we have,

$$\nabla f(x^*)^T(x - x^*) \geq 0 \quad \forall x \in \Omega$$

Proof:

By definition, $\exists \epsilon > 0, f(x^*) \leq f(x) \quad \forall x \in B_\epsilon(x^*)$.
Let $x \in \Omega$, $\lambda \in (0,1)$,

$$x(\lambda) = \lambda x + (1-\lambda)x^* = x^* + \lambda(x - x^*) \in \Omega$$

Since x can be any point inside Ω, this $x(\lambda)$ also have this property. Now to let $x(\lambda) \in B_\epsilon(x^*)$,

$$||x(\lambda) - x^*|| = \lambda||x - x^*|| < \epsilon$$

So, if $\lambda \in [0, \frac{\epsilon}{||x-x^*||}]$, we have,

$$\frac{f(x(\lambda)) - f(x^*)}{\lambda} \geq 0$$

By definition of derivative, when λ approach to 0, we have,

$$\nabla f(x^*)^T(x - x^*) \geq 0$$

Proof Done!

Next We apply this theorem to show some other property.

Set $f(y) = \frac{1}{2}||y - x||$, we have

$$(y^* - x)^T(y - y^*) \geq 0$$

where $y^* = P_\Omega(x)$.
So we can conclude that, (x,y here $\in R^n$)

$$(P_\Omega(x) - x)^T(z - P_\Omega(x)) \geq 0 \quad \forall z \in \Omega$$

$$(P_\Omega(y) - y)^T(z - P_\Omega(y)) \geq 0 \quad \forall z \in \Omega$$

Pick $z = P_\Omega(y)$ and $P_\Omega(x)$. Add together, we have,

$$(P_\Omega(x) - P_\Omega(y))(P_\Omega(y) - P_\Omega(x) + x - y) \geq 0$$

$$||P_\Omega(x) - P_\Omega(y)||^2 \leq (P_\Omega(x) - P_\Omega(y))^T(x - y)$$

$$||P_\Omega(x) - P_\Omega(y)||^2 \leq ||P_\Omega(x) - P_\Omega(y)|| * ||x - y||$$

$$||P_\Omega(x) - P_\Omega(y)|| \leq ||x - y||$$

So, $P_\Omega : R^n \to R^n$ is Lipschitz continuous with L=1.

Now we can proof why this $d = P_\Omega(x - \lambda\nabla f(x)) - x$ is a descent direction if x is not a stationary point.

$$\begin{aligned}
\nabla f(x)^T d &= -\frac{1}{\lambda}(x - \lambda\nabla f(x) - x)^T(P_\Omega(x - \lambda\nabla f(x)) - P_\Omega(x)) \\
&\leq -\frac{1}{\lambda}||x - \lambda\nabla f(x) - x|| * ||P_\Omega(x - \lambda\nabla f(x)) - P_\Omega(x)|| \\
&\leq -\frac{1}{\lambda}||P_\Omega(x - \lambda\nabla f(x)) - P_\Omega(x)||^2 \quad \text{(Lipschitz Continuous)} \\
&= -\frac{1}{\lambda}||d||^2 \leq 0
\end{aligned}$$

That's all for the algorithm part. Although some proof might be tricky, but the main idea is nothing but Taylor expansion, Cauchy inequality and triangle inequality.