OM-S20-22: Nonlinear Optimization (Cont.)

C. V. Jawahar

IIIT Hyderabad

http://preon.iiit.ac.in/om_quiz

07 Apr 2020

Convex Sets: A set C is *convex* if the line segment between any two points in C lies in C, *i.e.*, if for any x_1, $x_2 \in C$ and any θ with $0 \leq \theta \leq 1$, we have

$$y = \theta x_1 + (1 - \theta)x_2 \ \in \ C \tag{1}$$

.

Convex Function: A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if $\mathbf{dom(f)}$ is a convex set and if for all $x, y \in \mathbf{dom}(f)$, and θ with $0 \ \leq \ \theta \leq \ 1$, we have

$$f(\theta x + (1 - \theta)y) \ \leq \ \theta f(x) + (1 - \theta)f(y). \tag{2}$$

Strictly convex function when $\leq$ is replaced by $<$.
Convex optimization is the optimization/minimization of a convex function over a convex set.

Example Properties

- Interesection of two convex sets is convex
- Convex Hull: Convex combination of points in the set.
- If f is a convex function, then αf is also a convex function for $\alpha \geq 0$
- If f and f are convex then:
  - $f + g$ is convex
  - $\max(f, g)$ is convex
- *Eg. LP is convex optimization* Note that half spaces are convex and LP is an optimization over a set of intersctions of the half spaces. Also linear objective functions are convex functions.
- What about IP? What about LP Relaxation?

Problem

- Problem:

$$\min_x f(x)$$

- Unconstrained optimization.
- Popular to "Regularize" the solution with additional requirements/desirabilities:

$$\min_x f(x) + \lambda h(x)$$

- The objective function f could be convex or non-convex.

Gradient Descent

- Iterative solution starting with x^0

$$x^k \leftarrow x^{k-1} - t_k \nabla f(x^{k-1})$$

- Simple intuitive method.
- The parameter t_k the step size or learning rate. Convergence may depend on this learning rate. If t_k is too large, solution may diverge or oscillate.
- GD based methods need an initialization. Non-nonvex could be highly sensitive about starting point. Doesn't it affect convex?

Gradient Descent

Assume we are at x and we "approximate" a neighbouring point y as:

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} ||y - x||_2^2$$

We want to find the best such y. We differentiate wrt y and equate to zero:

$$\nabla f(x) + \frac{1}{2t}(2y - 2x) = 0$$

$$y = x - t\nabla f(x)$$

or

$$x^k \leftarrow x^{k-1} - t_k \nabla f(x^{k-1})$$

Possibly easier to appreciate geometrically!!

From Taylor's Series (Recap): Solve $f(x) = 0$

We are at x and we want to estimate the function value at a nearby point y.

$$f(y) = f(x) + \frac{f'(x)}{1!}(y - x) + \frac{f''(x)}{2!}(y - x)^2 + \frac{f^{(3)}(x)}{3!}(y - x)^3 + \cdots .$$

The first order approximation takes the first two terms in the series and approximates the function

$$f(y) = f(x) + \frac{f'(x)}{1!}(y - x)$$

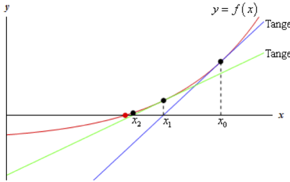
We are interested in y s.t. $f(y) = 0$: i.e.,

$$y - x = \frac{-f(x)}{f'(x)} \Rightarrow y = x - \frac{f(x)}{f'(x)}$$

$$x^{k+1} = x^k - f(x)/f'(x)$$

Newton's Method: Geometric Interpretation (Recap)

$$x^1 = x^0 - \frac{f(x^0)}{f'(x^0)} \text{ or } x^{k+1} = x_k - \frac{f(x^k)}{f'(x^k)}$$



If we are close to the root, then $|x - x^*|$ is small, which means that $|x - x^*|^2 \ll |x - x^*|$, hence we make the approximation:

$$0 \approx f(x) + (x^* - x)\, f'(x), \leftrightarrow x^* \approx x - \frac{f(x)}{f'(x)}$$

$f : \mathbf{R^n} \rightarrow \mathbf{R^m}$

$$\overline{f}(y) = \overline{f}(x) + J_x(y - x)$$

Where Jacobian J is an $m \times n$ matrix and can be given as

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Putting $f(y) = 0$ in equation we get:

$$\overline{0} = \overline{f}(\overline{x}) + J_x(\overline{y} - \overline{x})$$

Let $\overline{y} - \overline{x} = \overline{s}$

$$0 = f(x) + J_x s$$

$$J_x s = -f(x)$$

Procedure

$\overline{x}^0 \leftarrow$ Initial guess
$k = 0, 1, 2, \cdots$ do
Solve $J_{x_k} s = -f(x_k)$ for s
$x_{k+1} = x_k + s$

Newton's Method $f(x) = 0$ to $\min g(x)$

- Consider the problem of solving $f(x) = 0$. This arise when we optimize $g(x)$ leading to:

$$f(x) = \nabla g(x) = 0$$

- $\nabla g(x)$ is gradient and a vector (when x is a vector)
- f' in the Newton's then become Jacobian of the gradient. i.e., Hessian.

Therefore,

$$x^{k+1} = x^k - f(x)/f'(x)$$

becomes:

$$x^{k+1} = x^k - H^{-1}(g(x))\nabla g(x)$$

Where Hessain is the Jacobian of Gradient. Hessian in the denominator is the inverse.

Newton's Method for Optimization

Algorithm 5.1. NEWTON'S METHOD FOR UNCONSTRAINED MINIMIZATION.

given initial x, tolerance $\epsilon > 0$
repeat

1. Evaluate $\nabla g(x)$ and $\nabla^2 g(x)$.
2. **if** $\|\nabla g(x)\| \leq \epsilon$, **return** x.
3. Solve $\nabla^2 g(x) v = -\nabla g(x)$.
4. $x := x + v$.

until a limit on the number of iterations is exceeded

Backtracking: The Problem

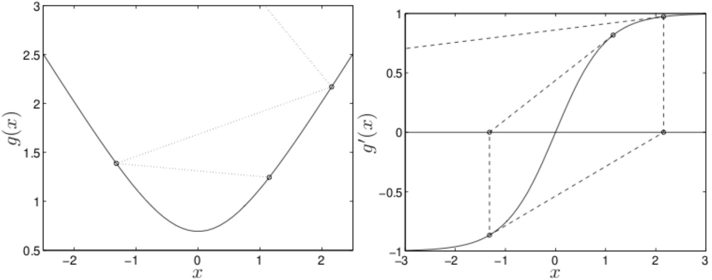


Figure 5.6 The solid line in the left figure is $g(x) = \log(\exp(x) + \exp(-x))$. The circles indicate the function values at the successive iterates in Newton's method, starting at $x^{(0)} = 1.15$. The solid line in the right figure is the derivative $g'(x)$. The dashed lines in the right-hand figure illustrate the first interpretation of Newton's method.

Solution

- There is nothing seriously wrong with the (newton's) update direction (say v^k).
- attempt full Newton $x^k + v^k$. If $g(x^k + v^k)$ is higher than $g(x^k)$, we reject the update,
- try $x^k + \frac{1}{2}v^k$ instead. if that also fails, try $\frac{1}{4}$
- until an acceptable value of t is found.
- We are searching over a line $x + tv$ for various values of t such as $t = 1, \frac{1}{2}, \frac{1}{4}, \ldots$. Thus the name line search.
- The purpose of the line search is to find a step size t such that $g(x + tv)$ is sufficiently less than $g(x)$.
- How does it lead to the condition?:

$$g(x + tv) \leq g(x) + \alpha t \nabla g(x)^T v$$

See Lec 1 video around 40.00

Backtracking

Algorithm 5.2. NEWTON'S METHOD WITH LINE SEARCH.

given initial x, tolerance $\epsilon > 0$, parameter $\alpha \in (0, 1/2)$.
repeat

1. Evaluate $\nabla g(x)$ and $\nabla^2 g(x)$.
2. **if** $\|\nabla g(x)\| \le \epsilon$, **return** x.
3. Solve $\nabla^2 g(x) v = -\nabla g(x)$.
4. $t := 1$.
   while $g(x + tv) > g(x) + \alpha t \nabla g(x)^T v$, $t := t/2$.
5. $x := x + tv$.

until a limit on the number of iterations is exceeded

Min $h(t)$. See Video at time 40.00

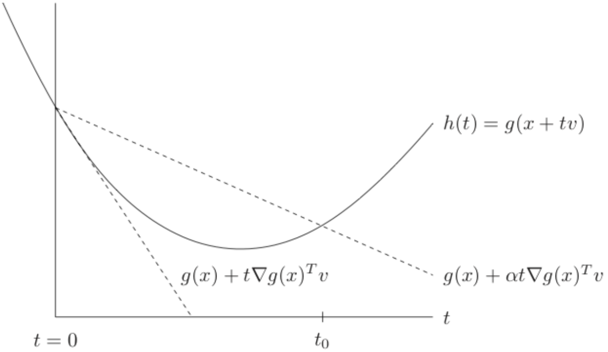


Figure 5.7 Backtracking line search. The curve shows g, restricted to the line over which we search. The lower dashed line shows the linear extrapolation of g, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that $g(x + tv)$ lies below the upper dashed line, *i.e.*, $0 \le t \le t_0$. The line search starts with $t = 1$, and divides t by 2 until $t \le t_0$.

Newton for non-Convex

Algorithm 5.3. Newton's method for nonconvex local minimization.

given initial x, tolerance $\epsilon > 0$, parameter $\alpha \in (0, 1/2)$.
repeat

1. Evaluate $\nabla g(x)$ and $\nabla^2 g(x)$.
2. **if** $\|\nabla g(x)\| \leq \epsilon$, **return** x.
3. **if** $\nabla^2 g(x)$ is positive definite, solve $\nabla^2 g(x) v = -\nabla g(x)$ for v
   else, $v := -\nabla g(x)$.
4. $t := 1$.
   while $g(x + tv) > g(x) + \alpha t \nabla g(x)^T v$, $t := t/2$.
5. $x := x + tv$.

until a limit on the number of iterations is exceeded

Convergence Analysis

- GD for convex: $O(\frac{1}{K})$

$$f(x^k) - f(x^*) \leq \frac{1}{k} \cdot \frac{1}{2t} ||x^0 - x^*||_2^2$$

- GD for strictly convex: $O(c^k)$

$$f(x^k) - f(x^*) \leq c^k \cdot \frac{L}{2} ||x^0 - x^*||_2^2$$

- GD for non-convex: $O(\frac{1}{\sqrt{K}})$ Gradients reduce at this rate.

Convergence Analysis (cont.) GD vs Newton

- GD convereges in $O(c^k)$
- Newton's method (convex; twice differentiable): $O(\frac{1}{2}^{2^k})$

Comparisons (GD vs Newton's)

- Computational complexity
- Storage complexity
- Theoretical elegance
- GD rules the practical world!!

What next?

- View videos again. Hope you understand better, and almost complete.
- Let us come back to the channel for discussions. Post your questions and discuss.
- We shall also post some questions for the discussions by Friday.