MATH466/MATH766
Math of machine learning
02/10-02/12 Lecture 9-10 Optimization Algorithm

References:

- Convex Optimization by Stephen Boyd and Lieven Vandenberghe
- Lecture notes by Prof. Ryan Tibshirani
  https://www.stat.cmu.edu/~ryantibs/convexopt-F15/

Todays contents:

- strong convexity and Lipschitz smoothness
- convergence and convergence rate
- gradient descent
- subgradient method

Important concepts:

- strong convexity and Lipschitz smoothness
- sublinear convergence, linear convergence

Recommend reading:

-

0. Preparation

0.1. strongly convexity

- f is μ-strongly convex ($\mu > 0$) iff $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex
- Strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex
  (* $f(x) = x^4$, $f(x) = e^x$ (* $f(x) = x$
- 0-th, 1-st and 2-nd equivalent conditions

  (0) $f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y) - \frac{\theta(1-\theta)\mu}{2}\|x-y\|^2$ $\forall x, y$ $\theta \in [0,1]$

  (1) $f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2$, $\forall x, y$

  (2) $\nabla^2 f(x) - \mu I \succeq 0$, $\forall x$

0.2. Lipschitz smoothness

- A differentiable function f is said to be L-Lipschitz differentiable or L-Lipschitz smooth if for some $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \quad \forall x, y. \qquad (*)$$

  (This definition does not assume convexity of f.)

- 1-st and 2-nd order necessary conditions

  (*) $\Rightarrow$ (1) $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|^2$

  (*) $\Rightarrow$ (2) $\nabla^2 f(x) \preceq LI$

Interpretation:

convexity	$\sim$	gradient is non-decreasing
strong convexity	$\sim$	the growth of gradient has a low-bound
Lipschitz smoothness	$\sim$	the growth of gradient has an upper-bound

1. Convergence and Convergence Rate

unconstrained problem. $\min_{x \in \mathbb{R}^n} f(x)$.

The study of an optimization algorithm mainly focus on

local / global — convergence and complexity. — # of iteration — convergence rate

— per iteration

- local convergence v.s. global convergence

Algorithm $\mathcal{A}$ has local convergence if there exists $r > 0$

such that when $d(x^0, x^*) < r$,

the $\{x^k\}$ generated by algorithm $\mathcal{A}$ satisfies

$f(x^k) \to f(x^*)$, and $\text{dist}(x^k, x^*) \to 0$.

Algorithm $\mathcal{A}$ has global convergence if

for any initialization x^0,

the $\{x^k\}$ generated by algorithm $\mathcal{A}$ satisfies

$f(x^k) \to f(x^*)$, and $\text{dist}(x^k, x^*) \to 0$.

- Convergence rate (assume $x^k \to \bar{x}$)

(1) sublinear rate if $\lim_{k \to +\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = 1$.

e.g. $x^k = \frac{1}{k+1}$, to have $\|x^k - \bar{x}\| < \varepsilon$,

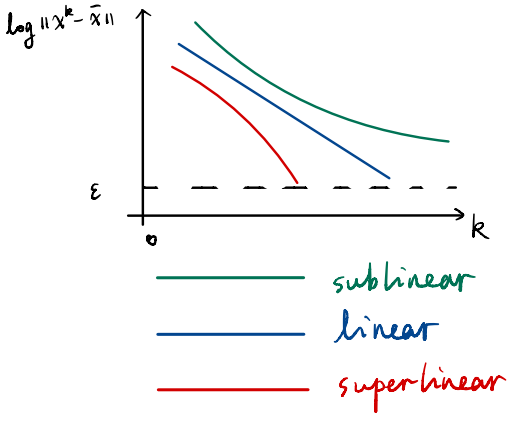
we need $k = O(\frac{1}{\varepsilon})$

(2) linear rate if $\lim_{k \to +\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} \le r \in (0,1)$.

e.g. $x^k = e^{-k}$, to have $\|x^k - \bar{x}\| < \varepsilon$, we need $k = O(\log \frac{1}{\varepsilon})$

(3) superlinear rate if $\lim_{k \to +\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = 0$

e.g. $x^k = e^{-2^k}$ to have $\|x^k - \bar{x}\| < \varepsilon$, we need $k = O(\log_2 \log \frac{1}{\varepsilon})$



2. Descent Method and Gradient Descent

Consider unconstrained smooth convex optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

A natural idea of algorithm: $x^{(k+1)} = x^{(k)} + t^{(k)} p^{(k)}$

($t^{(k)}$: step size; $p^{(k)}$: search direction)

Q1. What is a good search direction?

Q2. What is a proper step-size?

Want: $f(x^{(k+1)}) \leq f(x^{(k)})$

$$f(x^{(k+1)}) = f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)} + s p^{(k)})^T p^{(k)} \qquad s \in (0, t^{(k)})$$

- Descent direction: if $\nabla f(x^{(k)})^T p^{(k)} < 0$, then $p^{(k)}$ is a descent direction
- Line-search for stepsize: if $p^{(k)}$ is a descent direction
  then when $t^{(k)}$ is small enough, $f(x^{(k+1)}) \leq f(x^{(k)})$

But we don't want $t^{(k)}$ to be too small (why?)

for given $x^{(k)}$ and direction $p^{(k)}$

① exact line search: (in general not practical)

solve $\min_{t>0} f(x^{(k)} + t p^{(k)})$ for t 1D opt

② backtracking line search: pick $\alpha \in (0, 0.5)$, $\beta \in (0,1)$, $\gamma > 0$

$t \leftarrow \gamma$

(Armijo's cond.) while $f(x^{(k)} + t p^{(k)}) > f(x^{(k)}) + \alpha t \nabla f(x^{(k)})^T p^{(k)}$

there are other versions of line search | $t \leftarrow \beta t$

α: the fraction of decrease in linearized f we accept

typical choice: $\alpha \in (0.01, 0.3)$ less ⟷ more greedy

β: the amount of shrinkage when condition doesn't hold

typical choice: $\beta \in (0.1, 0.8)$ less ⟷ more crude

Gradient descent

$p^{(k)} = -\nabla f(x^{(k)})$

Alg. GD

Initialization $x^{(0)}$

While $\|\nabla f(x^{(k)})\|_2 > \varepsilon$

$\quad x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)})$

Prop. If f is L-smooth, $\nabla f(x^{(k)}) \neq 0$, then for any $t \in (0, \frac{2}{L})$

$$f(x^{(k+1)}) < f(x^{(k)}).$$

Pf.

$$f(x^{(k+1)}) \leq f(x^{(k)}) + t \nabla f(x^{(k)})^T p^{(k)} + \frac{L}{2} \|t p^{(k)}\|^2$$

$$= f(x^{(k)}) - t \|\nabla f(x^{(k)})\|^2 + \frac{L}{2} t^2 \|\nabla f(x^{(k)})\|^2$$

$$= f(x^{(k)}) - (-\frac{L}{2} t^2 + t) \|\nabla f(x^{(k)})\|^2$$

Since $\|\nabla f(x^{(k)})\|^2 < 0$

And when $t \in (0, \frac{2}{L})$, $-\frac{L}{2} t^2 + t < 0$, $f(x^{(k+1)}) < f(x^{(k)})$ □

If f is L-smooth, we usually pick $t = \frac{1}{L}$.

With this choice, we have the following convergence results.

Thm 1. Assume f is L-smooth, and $f(x) \geq f^*$ for any x

then $$\min \{\|\nabla f(x^{(k)})\|^2\}_{k=0}^{k-1} \leq \frac{2L (f(x^{(0)}) - f^*)}{K}$$

Thm 2. In addition to Thm 1 assumptions,

if f is convex and $f(x) \geq f(x^*)$ for any x.

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

Thm 3. In addition to Thm 1, 2 assumptions,

if f is μ-strongly convex.

then $$f(x^{(k)}) - f(x^*) \leq (1 - \frac{\mu}{L})^k (f(x^{(0)}) - f(x^*))$$

and $$\|x^{(k)} - x^*\|^2 \leq (1 - \frac{\mu}{L})^k \|x^{(0)} - x^*\|^2$$

3. Subgradient Method

direction $p^{(k)} = g^{(k)} \in \partial f(x^{(k)})$

Alg. SubGD

Initialization $x^{(0)}$

While $\|g^{(k)}\|_2 > \varepsilon$

$\quad x^{(k+1)} = x^{(k)} - t^{(k)} g^{(k)}$

Subgradient is NOT Necessarily a descent method

i.e. $f(x^{(k+1)}) \le f(x^{(k)})$ may not hold

e.g. $f(x) = |x|$, $x^{(k)} = 0$, $g^{(k)} = 1$

as long as $t > 0$, $f(x^{(k+1)}) > f(x^{(k)})$

so we keep track of best iterate $x_{best}^{(k)}$ among $x^{(0)}, \cdots, x^{(k)}$ so far.

i.e. $f(x_{best}^{(k)}) = \min_{k=0,\cdots,k} f(x^{(k)})$

or consider an average output $\bar{x}^{(k)} := \frac{1}{k} \sum_{k=1}^{k} x^{(k)}$

Thm. Assume that $f(x) \ge f(x^*)$ for any x

and f is G-Lipschitz continuous i.e. $|f(x) - f(y)| \le G\|x-y\|_2$

Then $f(x_{best}^{(k)}) - f(x^*) \le \dfrac{R^2 + G^2 \sum_{k=1}^{k} t_k^2}{2 \sum_{k=1}^{k} t_k}$, $(R = \|x^{(1)} - x^*\|)$

If we pick a fixed step-size t, then $\lim_{k \to +\infty} f(x_{best}^{(k)}) \le f(x^*) + \dfrac{G^2 t}{2}$

The optimal value may NOT be achieved in the limit.

Smaller step-size may reduce the gap but requires more iterations.

Solution: diminishing step-size e.g. $t_k = \frac{t}{k}$

Convergence rate: To have $\dfrac{R^2 + G^2 \sum_{k=1}^{k} t_k^2}{2 \sum_{k=1}^{k} t_k} \le \varepsilon$, need at least $\left(\frac{RG}{\varepsilon}\right)^2$ iterations

$\dfrac{R^2 + G^2 \sum_{k=1}^{k} t_k^2}{2 \sum_{k=1}^{k} t_k} \ge \dfrac{RG}{\sqrt{K}}$, "=" when $t_k = \dfrac{R/G}{\sqrt{k}}$