Unconstrained Minimization (Sec 9.1 – 9.4 + my SGD notes)

no dual

Aim: Solve the unconstrained min problem.

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

want: algorithm to find minimum of a point

$f: \mathbb{R}^n \to \mathbb{R}$ convex & twice continuously diff^ble.

(problem is solvable

Assume $\exists\, x^* \in \mathbb{R}^n$ s.t. $f(x^*) = \min_{x \in \mathbb{R}^n} f(x) = p^*$

↖ there exist a minimizer

Necessary and sufficient condition for x^* to be optimal

system of n equations $\nabla f(x^*) = 0$. stationarity

$\Rightarrow$ Solving the unconstrained problem $\Leftrightarrow$ solving the system of n equations in n unknowns.

↳ might be non linear etc.

Iterative algorithms: Produces a sequence $x^{(k)} \in \text{dom } f$ (iteration number) $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$

s.t. $f(x^{(k)}) \to p^*$ as $k \to \infty$.

Termination criterion: i) $f(x^{(k)}) - p^* \leq \varepsilon$. (in practice we do not know what this is)

ii) $\|\nabla f(x^{(k)})\| \leq \varepsilon$. (when gradient is sufficiently small)

Initial point $x^{(0)} \in \text{dom } f$.

Sublevel set $S = \{x \in \mathbb{R}^n : f(x) \leq f(x^{(0)})\}$ is closed

quadratic

Example 1: $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T P x + q^T x + r$ $\quad P \in S_+^n, q \in \mathbb{R}^n, r \in \mathbb{R}$.

(+ve definite, vector)

$\nabla f(x^*) = 0 \Rightarrow \quad Px^* + q = 0$

Since $P \in S_{++}^n$, P^{-1} exists, invertible $\quad x^* = -P^{-1}q$. (no need for optimisation algorithm ↳ closed form exist)

Example 2: Unconstrained Geometric Programming → cannot write down closed form expression for x^*

$\min_{x \in \mathbb{R}^n} f(x) \qquad f(x) = \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right)$

Stationarity condition: $\nabla f(x^*) = 0$.

$$\frac{\sum_{i=1}^m a_i \exp(a_i^T x + b_i)}{\sum_{i=1}^m \exp(a_i^T x + b_i)} = 0.$$

(always +ve)

$$\sum_{i=1}^m a_i \exp(a_i^T x + b_i) = 0.$$

$a_i \in \mathbb{R}^n$, $\exp(a_i^T x + b_i)$: scalar, $0 \in \mathbb{R}^n$ → system of n equations

No analytical solution, so we must use an iterative algorithm to get close to x^*.

Preview:

i) Condition. → strong convexity, → smoothness

ii) Gradient Descent.

works → prove convergence (how fast)

more conditions, the better the gradient descent (faster)

first order

iii) Stochastic Gradient Descent.

iv) Steepest Descent

v) Newton's method. 2nd order

---

normal convexity $\nabla^2 f(x) \geq 0 \quad \forall x \in S$

Strong Convexity

Assume $f: \mathbb{R}^n \to \mathbb{R}$ $m>0$ m-strongly convex on $S = \{x : f(x) \leq f(x^{(0)})\}$

i.e., Hessian $\nabla^2 f(x) \geq mI \quad \forall x \in S.$ (strictly larger than 0 (some lower bound); m convex $\Rightarrow$ convex)

Ex: If f is m-strongly convex, consider the new function $g(x) = f(x) - \underbrace{\frac{m}{2}\|x\|^2}_{\text{convex}}$. Then g is convex.

Implications: By Taylor's theorem $\forall x, y \in S$, $\exists z \in S$ on the line segment connecting x & y s.t.

control to Hessian

$$f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T \underbrace{\nabla^2 f(z)}_{\geq mI} (y-x).$$

If f is m-strongly convex,

$$f(y) \geq f(x) + (y-x)^T \nabla f(x) + \frac{m}{2}\|y-x\|^2 \quad \forall x, y \in S \quad -(*)$$

When $m=0$, $(*)$ becomes $f(y) \geq f(x) + (y-x)^T \nabla f(x)$, which is exactly convexity for diff^{ble} functions.

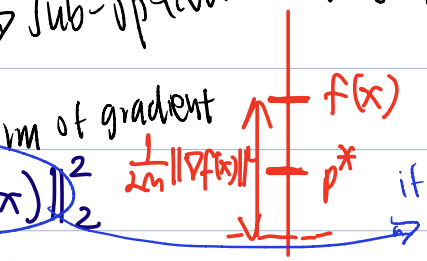
m is the strong convexity constant

PL inequality (Polyak-Łojasiewicz).

used to characterise sub-optimality gap

Lemma: If f is m-strongly convex, norm of gradient

smallest possible $p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$



if this small then we are close to optimal (eg. p^*)

Rmk: If the gradient $\nabla f(x)$ is small, then x is nearly optimal

Pf: Since f is m-strongly convex,

function of y

unconstrained so smallest is p^*

$$f(y) \geq f(x) + (y-x)^T \nabla f(x) + \frac{m}{2}\|y-x\|^2.$$

RHS is a (convex, quadratic) of y

Minimize both sides w.r.t. y.

∇ RHS $(y) = 0$

$\nabla f(x) + m(y-x) = 0$

$y = x - \frac{1}{m}\nabla f(x)$

y that minimises the RHS

$$p^* \geq f(x) + \left(-\frac{1}{m}\nabla f(x)\right)^T \nabla f(x) + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(x)\right\|^2$$

$$= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$$

(qed).

Rmk: If the gradient $\nabla f(x)$ is small, then x is nearly optimal

$$p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$$

$\|\nabla f(x)\| \leq \sqrt{2m\varepsilon} \longrightarrow f(x) - p^* \leq \varepsilon.$

If norm small, very close to optimality

x^* is global optimal solution
x is the current solution

The m-strongly convex condition can also be used to derive a bound on $\|x - x^*\|_2$.
↳ sub-optimality of x itself

$$f(y) \geq f(x) + (y-x)^T \nabla f(x) + \frac{m}{2}\|y-x\|^2 \qquad \forall x, y \in S.$$

set $y = x^*$

$$p^* = f(x^*) \geq f(x) + \underbrace{(x^*-x)}_{a}{}^T \underbrace{\nabla f(x)}_{b} + \frac{m}{2}\|x^*-x\|^2$$

$$\geq f(x) - \|x^*-x\|\,\|\nabla f(x)\| + \frac{m}{2}\|x^*-x\|^2.$$

cauchy ineq in opp direction

$\|a\|\|b\| \geq a^T b \geq -\|a\|\|b\|$

Since $p^* \leq f(x)$, $\forall x$,

$$-\|x^*-x\|\,\|\nabla f(x)\| + \frac{m}{2}\|x^*-x\|^2 \leq 0.$$

$$\|x^*-x\| \leq \frac{2}{m}\|\nabla f(x)\|_2$$

gradient = 0 characterises the optimality of x^*

Rmk: If $\nabla f(x)$ is small, then x is close to x^*.

Upper Bound on Hessian

If $\exists M > 0$ s.t. $\nabla^2 f(x) \preceq MI$ $\forall x \in S$, we say that f is M-smooth.

hold if level sets are closed and bounded

Lemma: (Descent lemma)

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

Pf: By Taylor's theorem, $\exists z \in [x,y]$ s.t.

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \underbrace{\nabla^2 f(z)}_{\preceq MI} (y-x)$$

$$\leq f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|^2 \quad ////.$$

(convex quadratic)

A consequence of descent lemma is

minimize over all possible y

$$p^* \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2$$

is a counterpart to the PL inequality: $p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$

---

f is both m-strongly convex & M-smooth, $0 \leq m \leq M < \infty$

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x \in S.$$

fast gradient descent

The ratio $\kappa = \frac{M}{m}$ is called the condition number of $\nabla^2 f(x)$.

Descent Methods

k is the iteration number

Produce a sequence $\{x^{(k)}\}_{k \in \mathbb{N}}$ s.t.

$$x^{(k+1)} = x^{(k)} + t^{(k)} \underbrace{\Delta x^{(k)}}_{\text{search direction.}} \quad \text{(descent direction)}$$

At iteration k, want to move in the direction of the search direction

learning rate / step size

(have many types - most common is negative gradient)

$$x^+ = x + t(\Delta x)$$

x^+: next iteration

Downhill → we want to minimise function

Descent methods: $f(x^+) < f(x)$ unless x is already optimal. (next iteration derives something better)

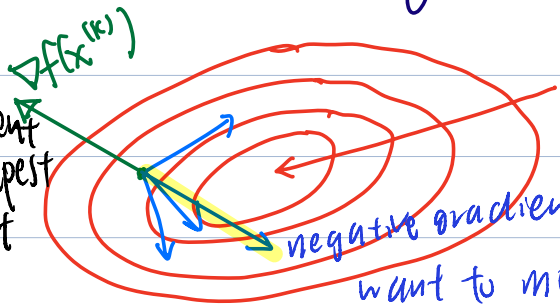
From convexity, $\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0 \Rightarrow f(y) \geq f(x^{(k)})$

$\Downarrow$

$f(y) \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (y - x^{(k)})$ first order condition

Search direction $\Delta x^{(k)}$ must satisfy $\nabla f(x^{(k)})^T (\Delta x^{(k)}) < 0$. (go downhill; direction at this iteration)

need search direction to be different direction than gradient



Such a direction is known as a descent direction.

General Descent Method

(guess starting point)

Given a starting point $x^{(0)} \in \text{dom } f$

repeat

1. Determine a descent direction Δx $\Delta x^{(k)} = -\nabla f(x^{(k)})$
2. Line search: Choose a step size t t_k (how much to move)
3. Update: $x^+ = x + t\Delta x$

(can tune t as a function of M and m)

until some stopping criterion is satisfied. (stop 100 iterations or gradient sufficiently small?)

Exact line search: $t = \underset{s \geq 0}{\text{argmin}}\ f(x + s\Delta x)$

optimisation problem in an optimisation problem

(already chosen direction)

tune these hyperparameters

Backtracking line search: (Armijo rule)

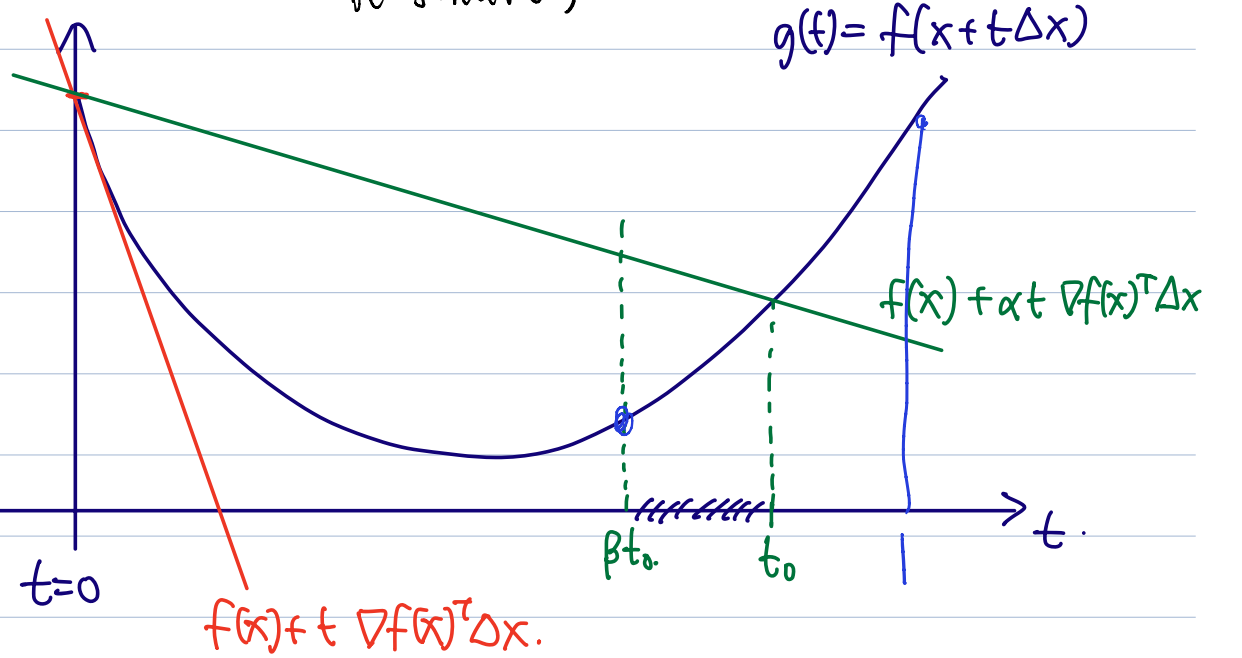
Given a descent direction Δx. $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.

$t \leftarrow 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$ (function still too big)

$t \leftarrow \beta t$ (set t to be smaller)

restricting to a certain direction: still convex



Backtracking line search stops with a step size of t that is $t = 1$ or $t \in (\beta t_0, t_0)$

α: fraction of the ↓ in f predicted by linear extrapolation that we will accept.

If α large, then loop have to go many times

↑ make obtuse angle with gradient

Gradient Descent $\Delta x = -\nabla f(x)$.

Stopping criterion: $\|\nabla f(x)\| \le \eta$ $\eta > 0$.
(sufficiently small eg. 10^{-3})

Thm: If GD is implemented on a function f that is m-strongly convex & M-smooth (i.e., $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in S$), with exact line search,

linear convergence

after k iterations $f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$

(less than 1) $c = 1 - \frac{m}{M}$ (contraction factor) ← want this close to 0!

every iteration contract by c (want c to be small)

Pf: From the M-smoothness

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|^2$$

Let $y = x - t\nabla f(x)$ ($= x^+$, next iteration) $\quad \tilde{f}(t) = f(x - t\nabla f(x))$. ($x - t\nabla f(x) = x^+$)

$$f(x - t\nabla f(x)) \le f(x) + \nabla f(x)^T(-t\nabla f(x)) + \frac{M}{2}\|-t\nabla f(x)\|^2$$

$\|$

$$\tilde{f}(t) \le f(x) + \nabla f(x)^T(-t\nabla f(x)) + \frac{M}{2}\|-t\nabla f(x)\|^2$$

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

↑ min over all t.

Since we use exact line search, min LHS over all t

$t_{exact} = \underset{s \ge 0}{\operatorname{argmin}}\, f(x - s\nabla f(x))$

choose step length to minimize this

$$f(x^+) = f(t_{exact}) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|^2$$

diff w.r.t t set to 0.
$t^* = \frac{1}{M}$

Subtract p^* on both sides.

$$f(x^+) - p^* \le f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|^2.$$

Since f is m-strongly convex, PL ineq. holds

$$\|\nabla f(x)\|^2 \ge 2m(f(x) - p^*)$$

replace gradient with PL ineq

$$f(x^+) - p^* \le f(x) - p^* - \frac{1}{2M}(2m(f(x) - p^*))$$
$$= \boxed{\left(1 - \frac{m}{M}\right)}(f(x) - p^*)$$

amount of improvement in 1 iteration

Applying this k times

$$f(x^{(k)}) - p^* \le \left(1 - \frac{m}{M}\right)^k (f(x^{(0)}) - p^*)$$ ////.

Cor: We have $f(x^{(k)}) - p^* \le \varepsilon$ if we use

$$k = \left\lceil \frac{\log((f(x^{(0)}) - p^*)/\varepsilon)}{\log\left(\frac{1}{c}\right)} \right\rceil$$ ceiling

iterations.

condition # ↑ denominator small

ratio of initial to final suboptimality.

Rmk: If $\frac{M}{m}$ is large

want condition number to be small

if condition number is bad ⇒ we need many iterations

$$\log\left(\frac{1}{c}\right) = \log\left(\frac{1}{1-m/M}\right) = -\log\left(1-\frac{m}{M}\right) \approx \frac{m}{M}$$

small.

$$\log(1-x) \approx -x$$

Analysis of backtracking line search.

Thm: f is m-strongly convex & M-smooth. Use Armijo or backtracking line search with (α, β). Then

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$$

$$c = 1 - \min\left\{2m\alpha, 2\beta\alpha\frac{m}{M}\right\} < 1.$$

Steepest Descent

$$f(x+v) \cong f(x) + \nabla f(x)^T v$$

$\nabla f(x)^T v$ is the directional derivative of f at x in the direction v.

↗ can decrease $f(x+v)$ function by a lot

Choose v to make $\nabla f(x)^T v$ as negative as possible.

$$\Delta x_{nsd} = \operatorname{argmin}\left\{\nabla f(x)^T v : \|v\| = 1\right\}$$

↑ normalized steepest descent

$\|\cdot\|$ suitably chosen norm. So cannot choose v to be very long vector

- choose different norm ⇒ get diff type of direction

Unnormalized version $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$

$\|x\|_* = \sup\{y^T x : \|y\| \le 1\}$. dual norm

Rmk: $\nabla f(x)^T \Delta x_{sd} = \nabla f(x)^T \left(\|\nabla f(x)\|_* \Delta x_{nsd}\right)$
$= -\|\nabla f(x)\|_*^2$ by definition of dual norm

Ex: Euclidean norm: $\|\cdot\| = \|\cdot\|_2$

$\Delta x_{sd} = \operatorname{argmin}\{\nabla f(x)^T v : \|v\|_2 = 1\}$
$= -\nabla f(x)$ (a, b; b constrained to be norm 1)
↳ achieved when b is a?

$-\|a\|\|b\| \le a^T b \le \|a\|\|b\|$

Steepest descent method with Euclidean norm is nothing but gradient descent!

Quadratic norm: $P \in S^n_{++}$ $\|z\|_P = \sqrt{z^T P z}$

$\Delta x_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-1/2} P^{-1} \nabla f(x)$.

$\Delta x_{sd} = -P^{-1} \nabla f(x)$

Can be regarded as a gradient method applied to the problem after a change of coordinates: $\bar{x} = P^{1/2} x$.

l_1-norm

$$\Delta x_{nsd} = \operatorname{argmin}\{\nabla f(x)^T v : \|v\|_1 = 1\}.$$
$$= -\operatorname{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i$$

e_i: i^{th} standard basis vector.

$$\Delta x_{sd} = \Delta x_{nsd} \|\nabla f(x)\|_\infty = -\frac{\partial f}{\partial x_i} e_i$$

vector

At each iteration, choose a component of $\nabla f(x)$ that has the largest abs value, decrease or increase the corresponding comp. of x by $\operatorname{sign}((\nabla f(x))_i)$.

only move in that coordinate rather than full gradient

$\Rightarrow$ coordinate descent algorithm.

Stochastic Gradient Descent

expensive to compute gradient

iid (x_i, y_i) $\quad \text{Loss}(\theta; D) = \frac{1}{m}\sum_{i=1}^{m} \text{Loss}(\theta; x_i, y_i)$

$i = 1, \dots, m$ $\quad$ finite sum.

expensive to calculate full gradient (alot of training samples)

not assuming m-s.c.

not using full gradient.

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

pick 1 training sample from m

SGD: $\rightarrow$ $i_k \in [m]$ randomly. $\quad P(i_k = i) = \frac{1}{m} \quad \forall i \in [m]$

$\rightarrow$ $x^{(k+1)} = x^{(k)} - t_k \nabla f_{i_k}(x^{(k)})$

$\mathbb{E}[\nabla f_i(x)] = \nabla f(x)$

unbiased

one sample gradient

one sample representative of whole dataset

Assumption: 1) $p^* = \inf_{x \in \mathbb{R}^n} f(x) > -\infty$ (finite)

2) f: M-smoothness $\nabla^2 f(x) \preccurlyeq MI \ \forall x$ ($MI - \nabla^2 f(x)$ is positive semi-definite)

3) Variance is bounded

$\sup_x \mathbb{E}[\|\nabla f_i(x)\|^2] \leq \sigma^2$

only the i (index) is random

t_k step size

Thm: Assume 1), 2), 3), run SGD for T iterations.

$$\min_{k=0,1,\dots,T} \mathbb{E}[\|\nabla f(x^{(k)})\|^2] \leq \frac{f(x^{(0)}) - p^*}{\sum_{k=0}^{T-1} t_k} + \frac{M\sigma^2}{2} \frac{\sum_{k=0}^{T-1} t_k^2}{\sum_{k=0}^{T-1} t_k}.$$

find 1 iteration such that we are close to having 0 gradient

want this to be small

(or: 1) $t_k = t > 0$, $\sigma^2 = 0$ (no noise, essentially doing gradient descent)

constant

$\sum_{k=0}^{T-1} t_k = tT$. First term gives $\frac{f(x^{(0)}) - p^*}{tT}$.

Convergence rate of $O(\frac{1}{T})$.

If assume M-smoothness (no m-s.c.), GD converges at a rate of $O(\frac{1}{T})$, much slower than c^T. exponential

2) SGD: $t_k = \frac{1}{\sqrt{k}}$ $\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \sim \sqrt{T}$. (approximation)

$\sum_{k=1}^{T} t_k^2 = \sum_{k=1}^{T} \frac{1}{k} = O(\log T)$.

$$\min_{k=0,1,\dots,T} \mathbb{E}[\|\nabla f(x^{(k)})\|^2] \leq O\left(\frac{\log T}{\sqrt{T}}\right)$$

slower than $(\frac{1}{T})$

worse than gradient descent without s-c assumption

minibatch set

→ $B_k \subseteq [n]$ randomly. $|B_k|$ = # machines

size

→ $x^{(k+1)} = x^{(k)} - t_k \frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x^{(k)})$ batch gradient

---

Pf: Only assume f is M-smooth. ⇒ Descent lemma.

$$f(x^{(k+1)}) \le f(x^{(k)}) + \nabla f(x^{(k)})^T (x^{(k+1)} - x^{(k)}) + \frac{M}{2} \|x^{(k+1)} - x^{(k)}\|^2$$

SGD does:

$$x^{(k+1)} = x^{(k)} - t_k \nabla f_{i_k}(x^{(k)})$$

only thing random is sample chosed

at iteration k, pick an index i_k

$$f(x^{(k+1)}) \le f(x^{(k)}) - t_k \nabla f(x^{(k)})^T \nabla f_{i_k}(x^{(k)}) + \frac{M t_k^2}{2} \|\nabla f_{i_k}(x^{(k)})\|^2$$

take expectation on both sides

every time see i_k, it is random and we take E

$$\mathbb{E}[f(x^{(k+1)})] \le f(x^{(k)}) - t_k \nabla f(x^{(k)})^T \mathbb{E}[\nabla f_{i_k}(x^{(k)})] + \frac{M t_k^2}{2} \mathbb{E}[\|\nabla f_{i_k}(x^{(k)})\|^2]$$

($\nabla f(x^{(k)})$ pulled out of $\mathbb{E}$)

unbiased ($\mathbb{E}[\nabla f_{i_k}(x^{(k)})] = \nabla f(x^{(k)})$)

$$\le f(x^{(k)}) - t_k \|\nabla f(x^{(k)})\|^2 + \frac{M t_k^2}{2} \sigma^2$$

$$\Rightarrow t_k \|\nabla f(x^{(k)})\|^2 \le f(x^{(k)}) - \mathbb{E}[f(x^{(k+1)})] + \frac{M t_k^2 \sigma^2}{2}$$

Run SGD for T iterations and sum up both sides.

Take expectation on both sides, then sum up

$$\sum_{k=1}^{T} t_{k-1}\, \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right] \le \sum_{k=1}^{T} \mathbb{E}\left[f(x^{(k)}) - f(x^{(k+1)})\right] + \frac{M\sigma^2}{2}\sum_{k=1}^{T} t_k^2$$

(c_k over $f(x^{(k)})$, c_{k+1} over $f(x^{(k+1)})$)

$$(c_1 - c_2) + (c_2 - c_3) + (c_3 - c_4) + \cdots + (c_T - c_{T+1})$$

telescoping

$$\sum_{k=1}^{T} t_{k-1}\, \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right] \le f(x^{(0)}) - \mathbb{E}\left[f(x^{(T)})\right] + \frac{M\sigma^2}{2}\sum_{k=1}^{T} t_k^2$$

$\mathbb{E}[f(x^{(T)})] \ge p^*$ (finite)

Lower Bound LHS by $\min_{k=0,1,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right]$.

$$\min_{k=0,1,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x^{(k)})\|^2\right] \le \frac{f(x^{(0)}) - p^*}{\sum_{k=0}^{T-1} t_k} + \frac{M\sigma^2}{2}\,\frac{\sum_{k=0}^{T-1} t_k^2}{\sum_{k=0}^{T-1} t_k}$$

$\sum_{k=0}^{T-1} t_k \uparrow \sqrt{T}$; $\sum_{k=0}^{T-1} t_k^2 \uparrow \log T$; $\sum_{k=0}^{T-1} t_k \uparrow \sqrt{T}$

$t_k = \frac{1}{\sqrt{k}}$ (step size)

$$\sum_{k=1}^{T} \frac{1}{k} \simeq \int_1^T \frac{1}{u}\,du = \log T \qquad = O\left(\frac{\log T}{\sqrt{T}}\right).$$

$$\sum_{k=1}^{T} \frac{1}{\sqrt{k}} = \int_1^T \frac{1}{\sqrt{u}}\,du \simeq \sqrt{T}$$

Summary: 1) Conditions on function (m-strongly convex, M-smooth)

2) Gradient descent has linear convergence

3) Steepest descent

4) Stochastic GD (M-smoothness)

assume less about function

↓

get worse results