Advanced gradient descent and convex optimization

I) Convexity
II) Gradient descent (GD) on convex problems
III) Acceleration

I] Setup

objective function

Pb: minimize $f(w)$ $\quad f: \mathbb{R}^d \to \mathbb{R}$

problem variables $\to$ $w \in \mathbb{R}^d$

$\operatorname{argmin}_{w \in \mathbb{R}^d} f(w) \subseteq \mathbb{R}^d$: set of solutions

$\min_{w \in \mathbb{R}^d} f(w)$: minimum value of f / of the problem

$\in \mathbb{R} \cup \{-\infty, +\infty\}$

$-\infty$: f unbounded $\quad$ $+\infty$: f not defined on $\mathbb{R}^d$

1) We will assume that f is convex and continuously differentiable.

Continuously differentiable (C^1): ∇f exists at every point and $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$ is continuous

If $w^* \in \mathbb{R}^d$ is a local minimum, then

$\|\nabla f(w^*)\| = 0$

Convexity: $f \in C^1$ is convex iff

$\forall (u,v) \in (\mathbb{R}^d)^2, \quad f(u) \geq f(v) + \nabla f(v)^T (u - v)$

Property: If f is C^1 and convex,

$[w^* \in \mathbb{R}^d$ global minimum of $f] \iff \|\nabla f(w^*)\| = 0$

Def: $f \in C^1$ is μ-strongly convex (for $\mu > 0$) if $\forall (u,v) \in (\mathbb{R}^d)^2$,

$$f(u) \geq f(v) + \nabla f(v)^T (u - v) + \frac{\mu}{2} \|u - v\|^2$$

Property: If $f \in C^1$ and μ-strongly convex, it has a unique global minimum.

Question: Given f C^1 and (at least) convex, can we design a procedure to compute a global minimum of f?

Goals: ① Design an algorithm that provably converges towards a solution

For convex problems, this may mean:

a) Generate $\{w_k\}_k$ such that $\|\nabla f(w_k)\| \xrightarrow[k \to \infty]{} 0$

($\Leftrightarrow$ CV towards global min thanks to convexity)

b) $\{w_k\}_k$ such that $f(w_k) \xrightarrow[k \to +\infty]{} \min_{w \in \mathbb{R}^d} f(w)$

c) $\{w_k\}_k$ such that $w_k \xrightarrow[k \to +\infty]{} w^*$ where $w^* \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$

② Efficiently compute an approximate solution to the problem

$\Rightarrow$ Popular metrics: complexity / convergence rate

* (Worst-case) complexity: Given a tolerance / level of accuracy, how much does it cost to reach this level of accuracy

Ex) $\forall \varepsilon > 0$, $f(w_K) - \min_{w \in \mathbb{R}^d} f(w) \leq \varepsilon$ in at most $O(\varepsilon^{-1})$ iterations

$O(A) =$ constant $\times A$ (constant does not depend on A)

* Convergence rate: Given a budget (of iterations, e.g.), what is the level of accuracy that we can expect?

Ex) $\forall K \geq 1$, $f(w_K) - \min_{w \in \mathbb{R}^d} f(w) \leq O\left(\frac{1}{K}\right)$

II] Gradient descent for convex and strongly convex problems

Idea: w^* global minimum $\Leftrightarrow \nabla f(w^*) = 0$

If $\nabla f(w) \neq 0$, w not global (nor local) minimum, and there must exist a better point nearby

$\Rightarrow$ To find such a point, we look into the direction of steepest descent $(-\nabla f(w))$

GD iteration: $k = 0$ Pick $w_0 \in \mathbb{R}^d$

$k \geq 0$ $w_{k+1} = w_k - \alpha_k \nabla f(w_k)$

$\alpha_k > 0$ stepsize* (*learning rate)

How to choose α_k?

- Predefined sequence (constant, decreasing)
- Adaptive way (e.g. via line search)
  - ⊖ Induces additional cost in general

$\Rightarrow$ Choice of α_k is generally guided by our assumptions on f

Here, we will assume that f is continuously differentiable with L-Lipschitz continuous gradient

(Other names: L-smooth, L-Lipschitz continuously differentiable, $C^{1,1}_L$)

∇f L-Lipschitz continuous $\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|$

Prop: If $f \in C^{1,1}_L$, then $\forall (u, v) \in (\mathbb{R}^d)^2$

$$(*) \quad f(u) \leq f(v) + \nabla f(v)^T (u - v) + \frac{L}{2}\|u - v\|^2$$

($\neq$ convexity inequality
$f(u) \geq f(v) + \nabla f(v)^T (u - v)$ $\left(+ \frac{\mu}{2}\|u - v\|^2 \right)$)

Proposition (Descent property).
Suppose that $f \in C^{1,1}_L$.
Run gradient descent with $\alpha_k = \frac{1}{L}$.
Then, $\forall k \in \mathbb{N}$,

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

Proof: Use the inequality $(*)$ with
$u = w_{k+1} = w_k - \alpha_k \nabla f(w_k) = w_k - \frac{1}{L}\nabla f(w_k)$
$v = w_k$

$$f(w_{k+1}) \leq f(w_k) + \nabla f(w_k)^T \left(-\frac{1}{L}\nabla f(w_k)\right) + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(w_k)\right\|^2$$

$$= f(w_k) - \frac{1}{L}\nabla f(w_k)^T \nabla f(w_k) + \frac{L}{2} \times \frac{1}{L^2}\|\nabla f(w_k)\|^2$$

$$= f(w_k) + \left(-\frac{1}{L} + \frac{1}{2L}\right)\|\nabla f(w_k)\|^2$$

$$= f(w_k) - \frac{1}{2L}\|\nabla f(w_k)\|^2$$

NB: $w_k - \frac{1}{L}\nabla f(w_k) \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(w_k) + \nabla f(w_k)^T (w - w_k) + \frac{L}{2}\|w - w_k\|^2 \right\}$

$(*)$ with $v = w_k$

Th (Convergence rate of GD on convex functions)

Suppose that f is $C^{1,1}_L$ and convex.

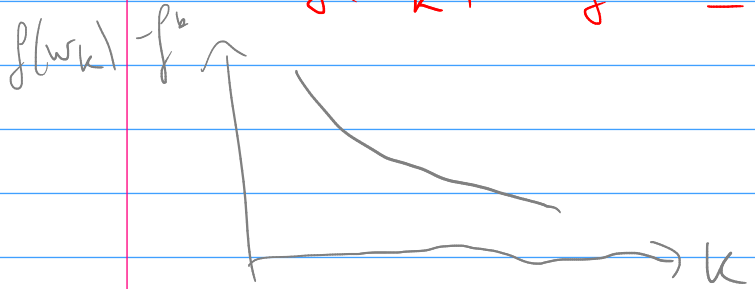
Suppose that $\operatorname{argmin}_{w \in \mathbb{R}^d} f(w) \neq \emptyset$ and

let $w^* \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$ and $f^* = f(w^*) = \min_{w \in \mathbb{R}^d} f(w)$

Run GD with $\alpha_k = \frac{1}{L}$ starting with $w_0 \in \mathbb{R}^d$.

Then, $\forall K \geq 1$,

$$f(w_K) - f^* \leq \frac{L}{2} \|w_0 - w^*\|^2 \times \frac{1}{K} = O\left(\frac{1}{K}\right)$$



If you increase K to $10K$, you are 10 times more accurate

Th (Convergence rate of GD for strongly convex functions)

Let $f \in C^{1,1}_L$ and μ-strongly convex.

Let $w^* \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$ and $f^* = f(w^*)$

Run GD with $\alpha_k = \frac{1}{L}$ starting from $w_0 \in \mathbb{R}^d$.

Then, $\forall K \in \mathbb{N}$,

$$f(w_K) - f^* \leq \underbrace{\left(1 - \frac{\mu}{L}\right)^K}_{\in (0,1)} \left(f(w_0) - f^*\right)$$

$\left(1 - \frac{\mu}{L}\right)^K \xrightarrow[K \to +\infty]{} 0$

↳ $\left(1 - \frac{\mu}{L}\right)^K$ goes to 0 faster than $\frac{1}{K}$

"Solving strongly convex problems is easier than convex problems."

Proof of the convergence rate of GD in the convex case

Setup: $f \in C^{1,1}_L$ and convex

$\forall (u,v) \in \mathbb{R}^d$,

(*) $f(u) \leq f(v) + \nabla f(v)^T (u-v) + \frac{L}{2}\|u-v\|^2$

(*) $f(u) \geq f(v) + \nabla f(v)^T (u-v)$

GD run with $\alpha_k = 1/L$ starting from w_0

↳ Because of (*) + $\alpha_k = 1/L$ $\forall k$, the descent property holds, i.e.

(1) $\forall k \geq 0$, $f(w_{k+1}) \leq f(w_k) - \frac{1}{2L}\|\nabla f(w_k)\|^2 \leq f(w_k)$

↳ Because of convexity, (*) applied with $v = w_k$ and $u = w^*$ ($\in \operatorname{argmin}_w f(w)$)

$f(w^*) \geq f(w_k) + \nabla f(w_k)^T (w^* - w_k)$

$\hookrightarrow$

(2) $f(w_k) \leq f(w^*) + \nabla f(w_k)^T (w_k - w^*)$

For every $k \in \mathbb{N}$, (1) + (2) gives

$$f(w_{k+1}) \leq f(w^*) + \nabla f(w_k)^T (w_k - w^*) - \frac{1}{2L}\|\nabla f(w_k)\|^2$$

"Trick": Notice that

$$\nabla f(w_k)^T (w_k - w^*) - \frac{1}{2L}\|\nabla f(w_k)\|^2 = \frac{L}{2}\left(\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right)$$

Hence, $\forall k \geq 0$

$$(*) \quad f(w_{k+1}) \leq f(w^*) + \frac{L}{2}\left(\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right)$$

Our goal is to bound $f(w_K) - f(w^*)$ for some $K \geq 1$.

We sum all inequalities of the form $(*)$ over $k \in \{0, \dots, K-1\}$

$$\sum_{k=0}^{K-1} f(w_{k+1}) \leq K f(w^*) + \frac{L}{2} \sum_{k=0}^{K-1} \left(\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right)$$

Telescopic sum

$$\sum_{k=0}^{K-1} (a_k - a_{k+1}) = a_0 - a_K$$

$$\sum_{k=0}^{K-1} f(w_{k+1}) \leq K f(w^*) + \frac{L}{2}\left(\|w_0 - w^*\|^2 - \underbrace{\|w_K - w^*\|^2}_{\geq 0}\right)$$

$$\sum_{k=0}^{K-1} f(w_{k+1}) \leq K f(w^*) + \frac{L}{2}\|w_0 - w^*\|^2$$

$$\sum_{k=0}^{K-1} \left(f(w_{k+1}) - f(w^*)\right) \leq \frac{L}{2}\|w_0 - w^*\|^2$$

Finally, per the descent property, we have

$$f(w_0) \geq f(w_1) \geq \dots \geq f(w_{K-1}) \geq f(w_K)$$

Therefore, $\sum_{k=0}^{K-1} f(w_{k+1}) - f(w^*) \geq K\left(f(w_K) - f(w^*)\right)$

and we obtain

$$K\left[f(w_K) - f(w^*)\right] \leq \frac{L}{2}\|w_0 - w^*\|^2$$

$$f(w_K) - \underbrace{f(w^*)}_{f^*} \leq \frac{L}{2}\|w_0 - w^*\|^2 \times \frac{1}{K}$$

Q.E.D.

III) Acceleration

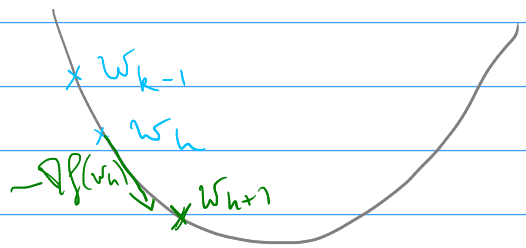
Question: Given a class of optimization problems ($C^{1,1}_L$ and convex/strongly convex) what is the fastest algorithm in terms of convergence rate that relies only on 1 gradient evaluation per iteration ?

comparable to GD

Answer: Accelerated gradient techniques / Gradient methods with momentum

Idea: At every iteration, we continue moving in the direction of the previous step, and we combine this move with one in the (negative) gradient direction.

* locally, $-\nabla f(\cdot)$ is a good direction
* The step you just took (in a negative gradient direction) may still be good for your current iterate



III - 1) Heavy-ball method (Polyak, 1964)

Iteration

$w_0 \in \mathbb{R}^d$, $w_{-1} = w_0$

(HB) $\forall k \geq 0$, $w_{k+1} = \underbrace{w_k - \alpha_k \nabla f(w_k)}_{\text{GD step}} + \underbrace{\beta_k (w_k - w_{k-1})}_{\text{Momentum step}}$

Guarantees: If f is a μ-strongly convex quadratic and $C^{1,1}_L$, run (HB) with

$$\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2} \quad \text{and} \quad \beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$$

Then $\forall k \geq 0$,

$$f(w_k) - f^* \leq O\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^k\right)$$

→ Optimal CV rate for strongly convex quadratics

(Recall: for GD, $f(w_k) - f^* \leq O\left(\left(1-\frac{\mu}{L}\right)^k\right)$)

$\mu \leq L \quad 1-\frac{\mu}{L} \geq 1-\sqrt{\frac{\mu}{L}}$

so $\left(1-\sqrt{\frac{\mu}{L}}\right)^k \to 0$ faster than $\left(1-\frac{\mu}{L}\right)^k$

⚠ The analysis does not extend to general strongly convex functions

⚠ In heavy ball, you can have $f(w_{k+1}) > f(w_k)$

III 2) Nesterov's method (strongly convex)

1983

Yurii Nesterov

Iteration $w_0 \in \mathbb{R}^d$, $w_{-1} = w_0$

$$\forall k \geq 0, \quad w_{k+1} = w_k - \alpha_k \nabla f(w_k + \beta_k(w_k - w_{k-1})) + \beta_k(w_k - w_{k-1})$$

Idea: Do the momentum step before the gradient step

(Heavy ball: gradient step before momentum)

Convergence rate: $f \in C^{1,1}_L$, μ-strongly convex (not necessarily quadratic)

Run Nesterov's method (aka Nesterov's accelerated gradient) with $\alpha_k = \frac{1}{L}$ (like GD) and $\beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$ (like HB)

Then, $\forall K \geq 0$,

$$f(w_K) - f^* \leq \frac{L+\mu}{2} \|w_0 - w^*\|^2 \left(1 - \sqrt{\frac{\mu}{L}}\right)^K$$

$$= O\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^K\right)$$

optimal for strongly convex functions

Proof: Technical but the idea is to work with a sequence of quadratic functions that approximate f more closely as $k \to +\infty$

II - 3) Nesterov's method for convex functions

Same as before but $(\beta_k)_k$ is predefined independently of f!

$$t_0 = 0, \quad t_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right), \quad \beta_k = \frac{t_k - 1}{t_{k+1}}$$

CV rate: If $f \in C^{1,1}_L$ convex, run Nesterov's method with the above choice for β_k and $\alpha_k = \frac{1}{L}$, then $\forall K \geq 0$,

$$f(w_K) - f^* \leq \frac{2L\|w_0 - w^*\|^2}{(K+1)^2} = O\left(\frac{1}{K^2}\right)$$

(GD: $O(1/K)$)

optimal for convex functions