Optimization for Machine Learning

September 30, 2024

Today: Lecture on advanced gradient descent

Tomorrow: Tutorial / lab (bring laptops if you can)

Afterwards: A long break until December

$\Rightarrow$ Part 1 of homework assignment

$\Rightarrow$ Updated lecture notes with exercises

ADVANCED ASPECTS OF GRADIENT DESCENT

Setup: $\underset{w \in \mathbb{R}^d}{\text{minimize}}\ f(w) \qquad f : \mathbb{R}^d \to \mathbb{R}$

Assumptions:

- f is bounded below
  $(\exists \bar{f} \in \mathbb{R},\ f(w) \geq \bar{f}$ for any $w \in \mathbb{R}^d)$

  ⚠ $\bar{f}$ is not necessarily equal to $\min_{w \in \mathbb{R}^d} f(w)$

- f is $\mathcal{C}_L^{1,1}$ (the gradient of f exists $\forall w \in \mathbb{R}^d$ and it is L-Lipschitz continuous)

$$\forall (w,v) \in (\mathbb{R}^d)^2, \quad \|\nabla f(v) - \nabla f(w)\| \leq L \|v - w\|$$

("If v is close to w, then $\nabla f(v)$ is close to $\nabla f(w)$")

Ex) Linear regression, logistic regression

Remark: In general, functions are only $\mathcal{C}_L^{1,1}$ on subsets of $\mathbb{R}^d$
$\Rightarrow$ The analysis of gradient descent for $\mathcal{C}_L^{1,1}$ functions extends to more general classes of functions

Key inequality: If f is $C_L^{1,1}$, then for any $(w, v) \in (\mathbb{R}^d)^2$,

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

function we want to minimize (as a function of v)

For fixed w, this a function of v that is quadratic, convex, easy to minimize

↳ Given $w \in \mathbb{R}^d$, if we find v such that
$f(w) + \nabla f(w)^T (v - w) + \frac{L}{2} \|v - w\|^2 < f(w)$, then
the inequality guarantees $f(v) < f(w)$ without having to compute $f(v)$

↳ The function $v \mapsto f(w) + \nabla f(w)^T (v - w) + \frac{L}{2} \|v - w\|^2$
is minimized at $v^* = w - \frac{1}{L} \nabla f(w)$

$\Longrightarrow$ This is precisely gradient descent with stepsize $\frac{1}{L}$!

$u^T(\alpha v) = \alpha (u^T v)$
$\|\alpha v\|^2 = \alpha^2 \|v\|^2$

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

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

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

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

$$= f(w) - \frac{1}{L}\|\nabla f(w)\|^2 + \frac{1}{2L}\|\nabla f(w)\|^2 = f(w) + \underbrace{\left[-\frac{1}{L}+\frac{1}{2L}\right]}_{-\frac{1}{2L}}\|\nabla f(w)\|^2$$

$$\left(\nabla f(w)^T \nabla f(w) = \|\nabla f(w)\|^2\right)$$

$$= f(w) - \frac{1}{2L}\|\nabla f(w)\|^2 < f(w) \quad \text{if } \nabla f(w) \neq 0_{\mathbb{R}^d}$$

$$\underset{v\in\mathbb{R}^d}{\text{minimize}} \; \underbrace{f(w) + \nabla f(w)^T(v-w) + \frac{L}{2}\|v-w\|^2}_{g(v)} \quad \text{for fixed } w \in \mathbb{R}^d$$

g is convex (sum of convex functions: linear + norm^2)

$$v^* \in \underset{v\in\mathbb{R}^d}{\operatorname{argmin}}\, g(v) \iff \nabla g(v^*) = 0$$

$$\iff \nabla f(w) + L(v^* - w) = 0$$

$$\iff L v^* = Lw - \nabla f(w)$$

$$\iff v^* = w - \frac{1}{L}\nabla f(w)$$

Gradient formula (generic rules)

- The gradient of $w \mapsto x^T w + y$ is x
- the gradient of $w \mapsto a\|w-u\|^2$ is $2a(w-u)$

For $C_L^{1,1}$ functions, applying gradient descent (with $\frac{1}{L}$) corresponds to solving a sequence of convex, quadratic optimization problems

GD: $w_0 \in \mathbb{R}^d$, $w_{k+1} = w_k - \frac{1}{L}\nabla f(w_k)$ $k = 0, 1, \ldots$

$$w_0 \to \text{Pick } w_1 \in \underset{v \in \mathbb{R}^d}{\operatorname{argmin}}\ f(w_0) + \nabla f(w_0)^T (v - w_0) + \frac{L}{2}\|v - w_0\|^2$$

$$\iff w_1 = w_0 - \frac{1}{L}\nabla f(w_0)$$

$$w_1 \to \text{Pick } v_2 \in \underset{v \in \mathbb{R}^d}{\operatorname{argmin}}\ f(w_1) + \nabla f(w_1)^T (v - w_1) + \frac{L}{2}\|v - w_1\|^2$$

$$\iff w_2 = w_1 - \frac{1}{L}\nabla f(w_1)$$

$w_2 \to \quad \ldots$

Q) What can we guarantee after K iterations of gradient descent?

For any $k = 0, \ldots, K-1$, we have shown that

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

Therefore

$$f(w_K) \leq f(w_{K-1}) - \frac{1}{2L}\|\nabla f(w_{K-1})\|^2$$

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

$$\vdots$$

$$f(w_K) \leq f(w_0) - \frac{1}{2L}\sum_{k=0}^{K-1}\|\nabla f(w_k)\|^2$$

Since f is bounded below, $f(w_K) \geq \bar{f}$

Hence $$\bar{f} \leq f(w_0) - \frac{1}{2L} \sum_{k=0}^{K-1} \|\nabla f(w_k)\|^2$$

$$\Leftrightarrow \quad \frac{1}{2L} \sum_{k=0}^{K-1} \|\nabla f(w_k)\|^2 \leq f(w_0) - \bar{f}$$

$$\Leftrightarrow \quad \sum_{k=0}^{K-1} \|\nabla f(w_k)\|^2 \leq 2L\left(f(w_0) - \bar{f}\right)$$

Using $$\sum_{k=0}^{K-1} \|\nabla f(w_k)\|^2 \geq K \left(\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \right)^2$$

$$\left(\sum_{k=0}^{K-1} a_k^2 \geq \sum_{k=0}^{K-1} (\min_k a_k)^2 = K \left(\min_k a_k\right)^2 \right)$$

Overall, we obtain that

$$K \left(\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \right)^2 \leq 2L\left(f(w_0) - \bar{f}\right)$$

$$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \leq \frac{\sqrt{2L(f(w_0) - \bar{f})}}{\sqrt{K}}$$

Theorem Apply gradient descent to minimize $\underset{w \in \mathbb{R}^d}{f(w)}$ with f $C_L^{1,1}$ and the stepsize for gradient descent chosen as $\frac{1}{L}$.

Then, for any $K \geq 1$,

$$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \leq \frac{\sqrt{2L(f(w_0) - \bar{f})}}{\sqrt{K}} = O\left(\frac{1}{\sqrt{K}}\right)$$

"A constant times $1/\sqrt{K}$"

$\Rightarrow$ We say that the convergence rate for gradient descent on nonconvex functions is $\frac{1}{\sqrt{K}}$

$\Rightarrow$ As $K \to \infty$, $\min_{0 \le k \le K-1} \|\nabla f(w_k)\| \to 0$: the method converges to a point with zero gradient

(Recall that $w^* \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w) \Rightarrow \nabla f(w^*) = 0_{\mathbb{R}^d}$)

Remark: Sometimes K is predetermined (budget criterion)

· But it can also be selected to achieve some accuracy/ some convergence guarantee

For example, if we want to compute w_k with $\|\nabla f(w_k)\| \le \varepsilon$, then choosing $K \geq 2L(f(w_0) - \bar{f})\varepsilon^{-2}$ guarantees that GD will compute such a point

$\rightarrow$ If L and $\bar{f}$ are known, can use this value in practice

$\rightarrow$ otherwise, this value is used as a guidance on the performance of the method

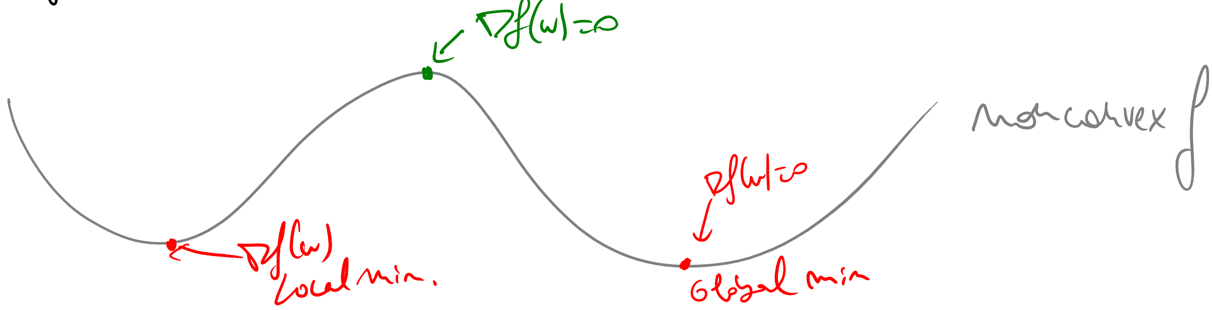
$$\varepsilon = 10^{-6} \longrightarrow 2L(f(w_0) - \bar{f})\,\varepsilon^{-2} \text{ iterations}$$

$$\frac{\varepsilon}{10} = 10^{-7} \longrightarrow 2L(f(w_0) - \bar{f})\varepsilon^{-2} \times 100 \text{ iterations}$$

Remark: ($\|\nabla f(w_k)\| = 0$ and nonconvexity)

$\hookrightarrow$ The theorem shows that GD converges to a point with zero gradient.

In theory, since f is not assumed to be convex, a point with zero gradient is not necessarily a minimum!



so GD can converge to points that are not minima in theory. In practice, however, this almost never happens, i.e. GD tends to converge to local or even global minima.

$\Longrightarrow$ Explained very recently (through sophisticated math. arguments)

Theorem (Lee et al. 2015) ⚠ INFORMAL VERSION

Suppose that we run GD with w_0 chosen randomly in $\mathbb{R}^d$. Then, with probability 1, GD converges to a local minimum.

$\longrightarrow$ Result actually holds for a subclass of nonconvex functions

$\longrightarrow$ This subclass includes classical problems in ML

- Low-rank matrix completion
- Shallow neural networks
- Phase retrieval

$\Rightarrow$ See the homework!

Convergence rates for gradient descent

↳ GD can be applied to any C^1 function, regardless of the function being convex or nonconvex

⇒ All you need is a gradient and a stepsize

↳ when the function is convex, GD has stronger theoretical guarantees.

↳ when the function is strongly convex, GD has even better guarantees.

Def: $f:\mathbb{R}^d \to \mathbb{R}$ C^1 is μ-strongly convex for some $\mu > 0$ if $\forall (w,v) \in (\mathbb{R}^d)^2$, $f(v) \geq f(w) + \nabla f(w)^T (v-w) + \frac{\mu}{2} \|v-w\|^2$

→ Any μ-strongly convex function is convex

→ The converse is not true: $w \mapsto x^T w + y$ is convex but not strongly convex

→ If f is $C_L^{1,1}$ and μ-strongly convex, then $\forall (v,w) \in (\mathbb{R}^d)^2$,

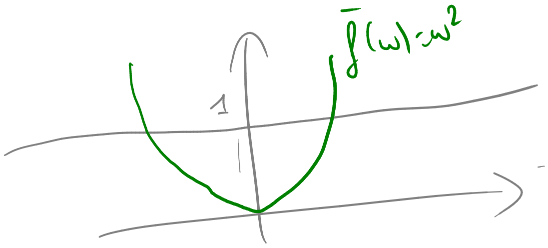
$$f(w)+\nabla f(w)^T(v-w)+\frac{\mu}{2}\|v-w\|^2 \le f(v) \le f(w)+\nabla f(w)^T(v-w)+\frac{L}{2}\|v-w\|^2$$

↳ At every w, you can lower and upper bound f by quadratic functions

↳ Implies $L \geq \mu$

Proposition: If $f:\mathbb{R}^d \to \mathbb{R}$ is C^1 and μ-strongly convex, it has a unique minimum, which is the only point $w^* \in \mathbb{R}^d$ such that $\nabla f(w^*) = 0_{\mathbb{R}^d}$

→ Convex functions can have infinitely many global minima (ex: constant functions)

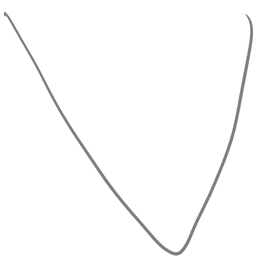


$f(w)=1 \;\forall w$ $\quad \arg\min f(w) = \mathbb{R}$

f is convex but not strongly convex

$\arg\min \bar{f}(w) = \{0\}$

μ-strongly convex with small μ



μ-strongly convex with large μ

NB: If f C^1 μ-strongly convex had two global minima w_1 and w_2, we would have

$$f(w_2) \geq f(w_1) + \underbrace{\nabla f(w_1)^T(w_2-w_1)}_{\overset{\shortparallel}{0} \ (w_1 \text{ minimum})} + \frac{\mu}{2}\|w_2-w_1\|^2$$

$$f(w_2) \geq f(w_1) + \underbrace{\frac{\mu}{2}\|w_2-w_1\|^2}_{>0 \text{ if } w_2 \neq w_1} > f(w_1) \quad \text{if } w_1 \neq w_2$$

This contradicts the fact that $w_2 \in \operatorname{argmin}_w f(w)$

Q) What is the convergence rate of GD on convex functions and strongly convex functions?

Setup: $\min_{w \in \mathbb{R}^d} f(w)$, f $C^{1,1}_L$, run GD with stepsize $\frac{1}{L}$ for $K \geq 1$ iterations.

(1) For nonconvex functions, we showed above that

$$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \leq O\left(\frac{1}{\sqrt{K}}\right)$$

$O(A)$: a constant times A

(2) For convex functions, we can show that

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

This rate is better than the rate for nonconvex functions!

- $\frac{1}{K}$ decreases faster than $\frac{1}{\sqrt{K}}$

• $f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \longrightarrow 0$ is a stronger guarantee than $\|\nabla f(w_k)\| \to 0$

③ If f is μ-strongly convex, we can also show that

$$f(w_K) - \min_{w \in \mathbb{R}^d} f(w) \leq O\left(\left(1-\frac{\mu}{L}\right)^K\right)$$

• Same guarantee than convex functions: $f(w_K) - \min_{w \in \mathbb{R}^d} f(w) \to 0$

• But much better rate: $\left(1-\frac{\mu}{L}\right)^K$ vs $\frac{1}{K}$

Ex) $\mu = 1$, $L = 2$, $K = 100$

$$\left(1-\frac{\mu}{L}\right)^K = \left(\frac{1}{2}\right)^K = \frac{1}{2^{100}} \quad \text{vs} \quad \frac{1}{K} = \frac{1}{100}$$

• Even stronger guarantee:

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

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

$w_K - w^* \to 0$: The iterates of GD converge to the minimum of f!

Takeaways

→ Strongly convex functions are easier to minimize than convex functions, and convex functions are easier to minimize than nonconvex ones!

→ For the same algorithm (GD in our case), the difference between these 3 classes of functions can be seen in the convergence rates

→ Convergence rates are a modern tool to compare optimization algorithms and choose the fastest method for a given class of functions

The fastest methods for minimizing $C_L^{1,1}$ functions

- If f is nonconvex, the fastest rate is $\frac{1}{\sqrt{K}}$, which is the rate of gradient descent.
- If f is convex, the fastest rate is $\frac{1}{K^2}$, which is faster than the $1/K$ rate of gradient descent
- If f is μ-strongly convex, the fastest rate is $\left(1-\sqrt{\frac{\mu}{L}}\right)^K$, which is faster than the $\left(1-\frac{\mu}{L}\right)^K$ rate of gradient descent.

$$\left(L \geq \mu \Rightarrow \frac{\mu}{L} \in (0,1] \Leftrightarrow \sqrt{\frac{\mu}{L}} \geq \frac{\mu}{L} \Leftrightarrow 1-\sqrt{\frac{\mu}{L}} \leq 1-\frac{\mu}{L} \right)$$

→ GD is not the fastest method for convex/strongly convex $C_L^{1,1}$ functions

⇒ This result was shown before the fastest methods were discovered

⇒ For a while, there did not exist an algorithm with the best convergence rates

⇒ Today, we know of such an algorithm: accelerated gradient, aka Nesterov's method

↳ History of accelerated gradient

* 1964: Polyak's method (Heavy ball / Gradient descent with momentum)

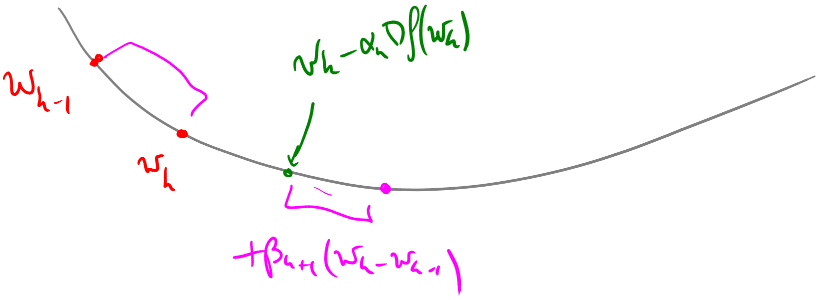
GD: $w_{k+1} = w_k - \alpha_k \nabla f(w_k)$ (typically $\alpha_k = \frac{1}{L}$)

Heavy ball: $w_{k+1} = w_k - \alpha_k \nabla f(w_k) + \beta_{k+1}(w_k - w_{k-1})$

with $\beta_{k+1} \geq 0$, $w_{-1} = w_0$

Idea: Incorporate information from the previous iteration ($w_{k-1} \longrightarrow w_k$) into the calculation of w_{k+1}

$\Rightarrow$ Tool: Momentum term $\beta_{k+1}(w_k - w_{k-1})$



↳ Polyak showed that this method is the fastest for $C_L^{1,1}$, strongly convex quadratic functions (special case)

$\Rightarrow$ Unfortunately, Polyak's method can fail on other convex functions.

* 1983: Yurii Nesterov introduces the accelerated gradient method

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

- Still combine gradient steps and momentum like in heavy ball
- But you add the momentum before computing the gradient

$\Rightarrow$ Nesterov's method can be written as

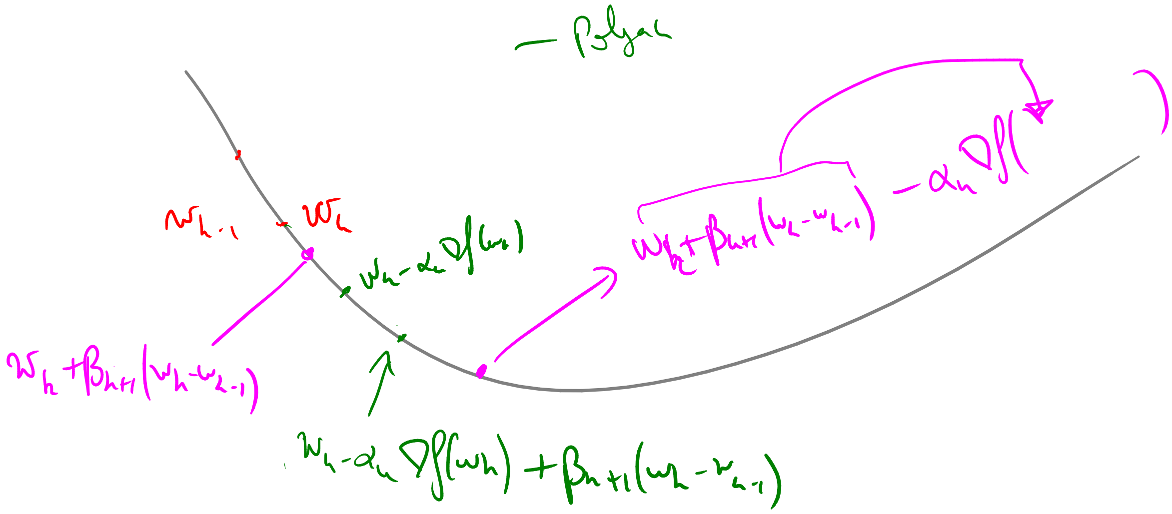
$$z_{k+1} = w_k + \beta_{k+1}(w_k - w_{k-1})$$

$$w_{k+1} = z_{k+1} - \alpha_k \nabla f(z_{k+1})$$

↳ Nesterov showed that this method has the fastest convergence rate possible for convex and strongly convex functions.

⚠ α_k and β_{k+1} need to be chosen carefully

μ-strongly convex: $\alpha_k = \frac{1}{L}$ $\quad \beta_{k+1} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$



convex case: $\alpha_k = \frac{1}{L}$, β_{k+1} chosen independently of f and L!