Nonconvex optimization

October 14, 2021

I) Nonconvex problems

II) Gradient descent on nonconvex problems

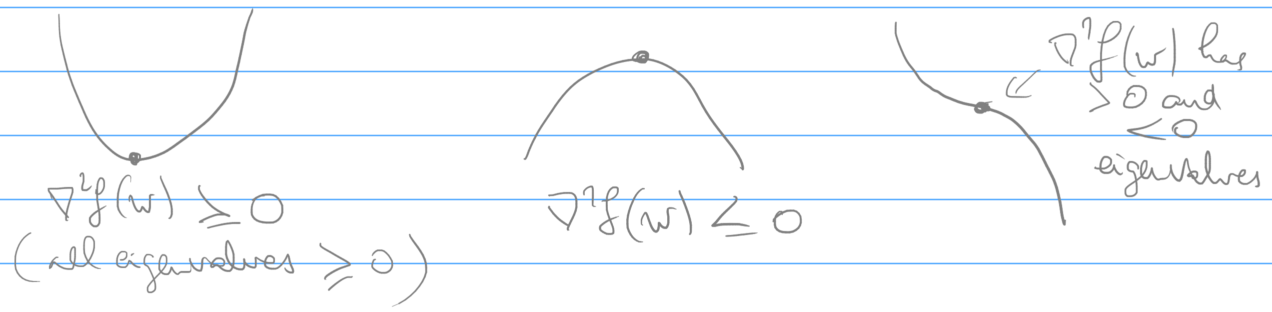
⚠ Correction : Clarified the statement on almost-sure convergence of gradient des

I] Nonconvex problems

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \ f(w) \qquad f : \mathbb{R}^d \to \mathbb{R}$$
f C^1, possibly C^2

f C^2 : $\forall w \in \mathbb{R}^d$, $\nabla f(w) \in \mathbb{R}^d$ and $\nabla^2 f(w) \in \mathbb{R}^{d \times d}$ are well-defined and the mappings $w \mapsto \nabla f(w)$ and $w \mapsto \nabla^2 f(w)$ are continuous

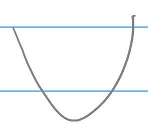
NB: If $f \in C^2$, $\nabla^2 f(w)$ is a symmetric matrix and its eigenvalues represent the curvature of f at w



↳ Optimality conditions

* First-order necessary condition : ($f \in C^1$)
$$[w^* \text{ local minimum of } f] \Rightarrow \|\nabla f(w^*)\| = 0$$

* Second-order necessary condition ($f \in C^2$)
$$[w^* \text{ local minimum of } f] \Rightarrow \begin{cases} \|\nabla f(w^*)\| = 0 \\ \nabla^2 f(w^*) \succeq 0 \end{cases}$$



* Second-order sufficient condition ($f \in C^2$)
$$\begin{bmatrix} \|\nabla f(w^*)\| = 0 \\ \nabla^2 f(w^*) \succ 0 \end{bmatrix} \Rightarrow [w^* \text{ local minimum of } f]$$

If $\|\nabla f(w)\| \neq 0$ or $\nabla^2 f(w)$ has negative eigenvalues, then it is possible to move away from w towards a better point

① Either in the direction $-\nabla f(w)$ (if $\neq 0$)
$\simeq$ GD step

② Or in a direction of negative curvature:
$v \in \mathbb{R}^d$, $\underbrace{v^T}_{1\times d} \underbrace{\nabla^2 f(w)}_{d\times d} \underbrace{v}_{d\times 1} < 0$

Th) If f is C^2 and convex, then we always have $\nabla^2 f(w) \succeq 0 \ \forall w$

If f is μ-strongly convex, then
$\nabla^2 f(w) \succeq \mu I_d$ identity matrix $\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} d$

all eigenvalues of $\nabla^2 f(w)$ are $\geq \mu$

A, B symmetric real matrices: $A \succeq B \iff A - B \succeq 0$ (eigenvalues of $A-B$ are ≥ 0)
Löwner order

↳ For convex functions, the second-order derivative does not help in defining additional directions of decrease (no negative curvature directions)
(global minima $\iff$ points with zero gradient)

↳ For nonconvex (i.e. not convex) problems, if $\bar{w}$ is such that $\|\nabla f(\bar{w})\| = 0$ ("a critical point"), then the Hessian can help in figuring out the nature of $\bar{w}$:

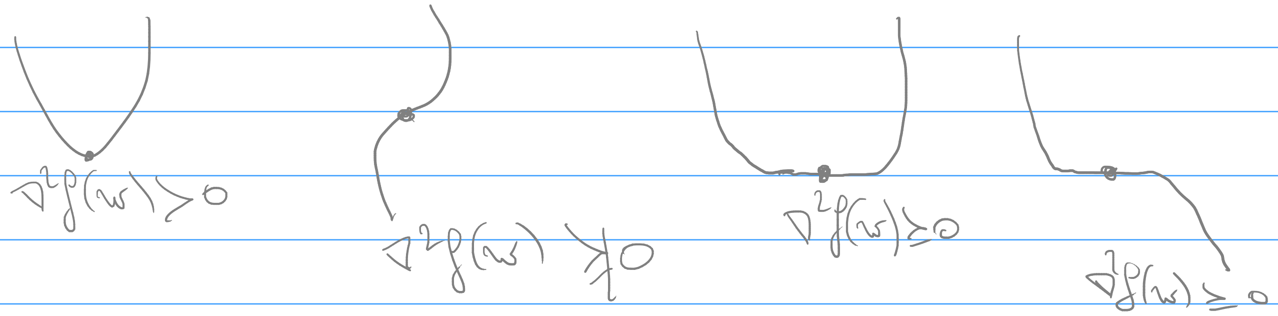
* $\nabla^2 f(\bar{w}) \succ 0 \implies \bar{w}$ local minimum
(from second-order sufficient condition)

$\nabla^2 f(w) \not\succeq 0$

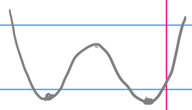
* $\nabla^2 f(\bar{w})$ has at least one negative eigenvalue $\implies \bar{w}$ either a local maximum or a saddle point

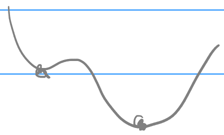
↳ We can "escape" that point using negative curvature

* $\nabla^2 f(\bar{w}) \succeq 0$: $\bar{w}$ can be a local minimum or a (high-order) saddle point



↳ In nonconvex optimization, we have to worry about saddle points (and local maxima)
$\Rightarrow$ Distinction between nonconvex problems with "good saddle points" and "bad saddle points", i.e. between problems where methods can be stuck at saddle points and problems where they converge to a local optimum



↳  In nonconvex optimization, local min $\neq$ global min

Still, can distinguish between problems for which any local minimum is also global, and those for which there exist spurious local minima (with a function value significantly higher than the global optimum)

Many nonconvex problems in data science possess a favorable structure for optimization:

oftentimes under assumptions on the data

* All local minima are global
* All saddle points are strict (the Hessian at these points has a negative eigenvalue)

Examples

(1) Eigenvalue optimization

Task: Compute the minimum eigenvalue of some symmetric matrix $H \in \mathbb{R}^{d \times d}$

Application: PCA (Principal Component Analysis)

$X = [x_i^T]_{i=1}^{n} \in \mathbb{R}^{n \times d}$ data matrix

$$\mathbb{R}^{d\times d} \ni C = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(x_i - \bar{x})}_{d\times 1} \underbrace{(x_i - \bar{x})^T}_{1 \times d}$$) empirical covariance matrix

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Principal component of X (direction of largest variability) is given by computing the maximum eigenvalue of C, which is equivalent to finding the minimum eigenvalue of $H = -C$

do not change the problem

Pb: $\underset{w \in \mathbb{R}^d}{\text{minimize}}$ $\frac{1}{2} w^T H w$ subject to $\|w\|^2 = 1$

Nonconvex problem (with nonconvex objective when $H \not\succeq 0$)
(in PCA, $H \preceq 0$: all eigenvalues are ≤ 0)

→ But all local minima are global
→ All saddle points are strict if H has no zero eigenvalues (e.g. $H \prec 0$)

$\Rightarrow$ This problem (and others of the same family like the trust-region subproblem) can be solved to global optimality despite being nonconvex

$\hookrightarrow$ The set of critical points of the problem (1st order condition satisfied) contains 0 and the eigenvectors of H ($Hw = \lambda w$)

$\Rightarrow$ Only the eigenvectors corresponding to the minimum eigenvalue of H are local minima, hence they are also global minima (best possible function value)

II) Gradient descent on nonconvex problems

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

Assumptions: * $f \in \mathcal{C}_L^{1,1}$

$(\|\nabla f(w) - \nabla f(u)\| \leq L\|w-u\|)$

* $\exists \bar{f} \in \mathbb{R}, \ \forall w \in \mathbb{R}^d, \ f(w) \geq \bar{f}$

($\bar{f}$ lower bound on f)

(GD) Gradient Descent iteration: $\begin{cases} k=0 & w_0 \in \mathbb{R}^d \\ k \geq 0 & w_{k+1} = w_k - \alpha_k \nabla f(w_k) \\ & (\alpha_k > 0) \end{cases}$

Th) Suppose we apply GD under the assumptions on f above with $\alpha_k = \frac{1}{L}$. Then, $\forall K \geq 1$

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

Convergence rate of GD on nonconvex problems

↳ Weaker guarantee than in the convex case
($\min_k \|\nabla f(w_k)\|$ vs $f(w_k) - f^*$)
⇒ Only guarantees that you converge towards a point with 0 gradient
⚠ could be a saddle point
↳ Slower rate $\left(\frac{1}{\sqrt{K}}\right)$ than in the convex case $\left(\frac{1}{K}\right)$ or in the μ-strongly convex case $\left((1-\frac{\mu}{L})^K\right)$

Proof: Since f is $C_L^{1,1}$ and $\alpha_k = \frac{1}{L}$, the descent property holds: (see 12/10 lecture)

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

Re-arranging the terms gives

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

Summing $(*)$ for $k = 0, 1, \dots, K-1$, we obtain:

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

On one hand, $\|\nabla f(w_k)\|^2 \geq \left(\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\|\right)^2 \quad \forall k \leq K-1$

OTOH, $\sum_{k=0}^{K-1}\left(f(w_k) - f(w_{k+1})\right) = f(w_0) - f(w_K)$
$$\leq f(w_0) - \bar{f}$$
$$\left(f(w) \geq \bar{f} \quad \forall w \in \mathbb{R}^d\right)$$

Hence, $K \times \left[\min_{0\le k\le K-1} \|\nabla f(w_k)\|\right]^2 \le 2L\,(f(w_0)-\bar{f})$

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

Q.E.D.

complexity **Corollary**: GD computes a point w_k such that $\|\nabla f(w_k)\| \le \varepsilon$ $(\varepsilon > 0)$ after at most $O(\varepsilon^{-2})$ iterations

$\Rightarrow$ Unlike in the convex case, this bound is sharp for $C_L^{1,1}$ functions

2010 It exists $f \in C_L^{1,1}$ such that GD takes at least $O(\varepsilon^{-2})$ iterations

↳ Common wisdom in nonconvex optimization

- In terms of CV rates, GD is slow (but cannot do faster)
- It can converge to saddle points or even local maxima

Ex) $f(w) = -w^2$, $d = 1$

$w_0 = 0$

GD $w_k = w_0 = 0$ $\forall k$

- In practice, GD actually performs much better and generally reaches local minima (and therefore avoids saddle points)

(2015) Thm) Run GD on f C^2 with $\alpha_k < 1/L$. Then, GD will converge almost surely to $\bar{w}$ such that $\begin{cases} \|\nabla f(\bar{w})\| = 0 \\ \nabla^2 f(\bar{w}) \succeq 0 \end{cases}$

for almost every w_0

Idea: Consider the Lebesgue measure on $\mathbb{R}^d$.

The set $\{ w_0 \mid \text{GD cv to } \bar{w} \text{ with } \nabla^2 f(\bar{w}) \not\succeq 0 \}$ has zero Lebesgue measure

$\Rightarrow$ Drawing w_0 according to the Lebesgue measure leads almost certainly to a vector w_0 such that GD converges towards a point satisfying the second-order necessary conditions

⚠ Asymptotic result (needs $k \to \infty$)

In fact, a subsequent result showed that GD can take an arbitrarily large number of iterations to escape the vicinity of a strict saddle point

strict $\nabla^2 f(w) \not\succeq 0$

$$\underbrace{w_0 \to w_1 \cdots \to w_k}_{\|\nabla f(w_k)\| \leq \varepsilon \text{ after } O(\varepsilon^{-2}) \text{ iterations}} \qquad \underbrace{\qquad w_{\bar{k}}}_{\bar{k} - k \text{ can be arbitrarily large}} \dashrightarrow \bar{w}_{\text{local}} \quad k \to \infty$$

$w_{\bar{k}}$: better value than the saddle point

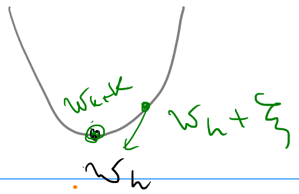
↳ One common way of avoiding saddle points / improving the outcome of GD

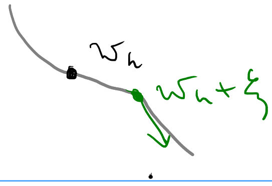
→ Run GD twice, once with a fixed w_0 and once with a random w_0 (then take the best of the two outcomes)

↳ In practice, when $\|\nabla f(w_k)\|$ gets small, adding some noise to the input (ex: Gaussian) can help in escaping saddle points faster!

$$w_0 \to w_1 \to \cdots \to w_k \to w_k + \xi \to w_{k+1} \to \cdots$$

$$\|\nabla f(w_k)\| \leq \varepsilon \qquad \xi \sim \mathcal{B}(0, r)$$





$\sim B(0, r)$ Uniform distribution over the ball centered at 0 and of radius r

Th (2017): Let $f \in C^2$ nonconvex, and $C^{1,1}_L$. Run GD with fixed stepsize $\alpha > 0$ and noise injection (adding $\xi \sim B(0, r)$). Then, under some assumptions on α, r, for any $\varepsilon > 0$, GD will find w_k such that

stronger guarantee than GD

$$\begin{cases} \|\nabla f(w_k)\| \leq \varepsilon \\ \nabla^2 f(w_k) \succeq -\sqrt{L\varepsilon}\, I_d \end{cases}$$

all eigenvalues of the Hessian are $\geq -\sqrt{L\varepsilon}$

in at most

$$O\left(\varepsilon^{-2} \ln^4(d\,\varepsilon^{-2})\right)$$

with high probability

$\Rightarrow$ Randomness can improve theoretical guarantees (in probability)