OPTIMIZATION FOR MACHINE LEARNING

October 4, 2023

Today: More on gradient descent + Nonconvexity

Tomorrow: Regularization and proximal gradient

After tomorrow:
- → Lecture notes
- → Course project
- → Lectures resume in November

GRADIENT DESCENT AND NONCONVEXITY

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

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

$\alpha_k > 0$ "stepsize"

(1) Choosing the stepsize ("Tuning the learning rate")

Recall: when $f \in C^{1,1}_L$ (C^1 + ∇f is L-Lipschitz continuous),

$\alpha_k = \frac{1}{L}$ is a good stepsize choice because it guarantees $\quad f(w_{k+1}) \leq f(w_k) - \frac{1}{2L} \|\nabla f(w_k)\|^2$

(Any $\alpha_k \in \left(0, \frac{2}{L}\right)$ provides similar guarantees)

In practice, the value of L may { be too expensive to compute. / be unknown / not exist

In this situation, there are 3 main families of stepsize choices:

(i) Constant stepsizes

$\alpha_k = \alpha > 0$ $\qquad$ (e.g. $\alpha = 0.1$, $\alpha = 0.01$, $\alpha = 0.001$, ...)

→ For α sufficiently small, this choice leads to convergence (guarantees decrease at every iteration)

→ For $C^{1,1}_L$ functions, $\alpha < \frac{2}{L}$ works!

→ Very popular choice but difficult to calibrate in advance

(2) Decreasing stepsizes

$\alpha_k \to 0$ $(k \to \infty)$

(e.g. $\alpha_k = \frac{1}{k+1}$, $\alpha_k = \frac{\alpha_0}{(k+1)^a}$ with $\alpha_0 > 0$, $a > 0$)

→ For k sufficiently large, α_k will be small enough to guarantee decrease at iteration k and thus convergence

→ For $C^{1,1}_L$ functions, $\alpha_k < \frac{2}{L}$ for k sufficiently large

(e.g. $\frac{1}{k+1} < \frac{2}{L}$ for $k > \frac{L}{2} - 1$)

→ Less popular than constant stepsizes yet useful in stochastic settings (see Lectures 5+6)

(3) Adaptive stepsizes (≈ learning rate scheduling)

Idea: . α_k is chosen according to $(w_k, f(w_k), \nabla f(w_k))$,

typically done via a line search

. the GD step has the form $\underbrace{w_k - \alpha \nabla f(w_k)}_{\text{function of } \alpha}$

We would like the best possible value of α in terms of function value $f(w_k - \alpha \nabla f(w_k))$ as small as possible

Exact line search:

$\alpha_k \in \operatorname{argmin}_{\alpha > 0} f(w_k - \alpha \nabla f(w_k))$

↳ often too expensive in practice, usually replaced by an approximation

Armijo backtracking line search

start with $\alpha_k = \bar{\alpha} > 0$ (e.g. $\bar{\alpha} = 1$)

while $\left(f(w_k - \alpha_k \nabla f(w_k)) > f(w_k) - c\,\alpha_k \|\nabla f(w_k)\|^2 \right)$

$\alpha_k \leftarrow \theta \alpha_k$ ($\theta \in (0,1)$ e.g. $1/2$)

($c \in (0, 1/2)$ e.g. 0.0001 or 0.5)

Backtracking: $\bar{\alpha}$, $\theta\bar{\alpha} < \bar{\alpha}$, $\theta^2\bar{\alpha}$, ... $\to 0$

$\theta \in (0,1)$

↳ Process based on a sufficient decrease condition

- The condition will be violated ($f(w_k - \alpha_k \nabla f(w_k)) \leq f(w_k) - c\alpha_k \|\nabla f(w_k)\|^2$) for sufficiently small α_k.
- If $f \in C_L^{1,1}$ and $c = \frac{1}{2}$, will occur for $\alpha_k < \frac{2}{L}$

$\Rightarrow$ For $\alpha_k = \frac{1}{L}$, we recover the guarantee

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

↳ This strategy is more expensive than using constant or decreasing stepsizes because it requires function evaluations ($f(w_k)$ + $f(w_k - \alpha \nabla f(w_k))$ for every value α used in the backtracking line search) $\Rightarrow$ In ML (mostly in deep learning), this is often considered as prohibitive

(2) Gradient descent on nonconvex problems

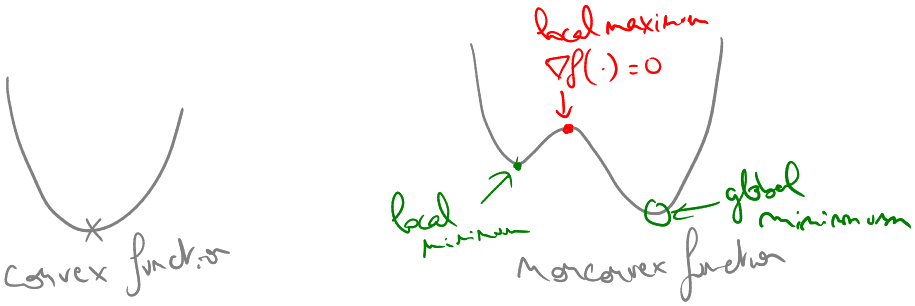
↳ When the objective is convex, GD converges to a minimum value and we can derive convergence rates of the form

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

↳ That guarantee is obtained thanks to convexity of f, and in particular

- $\forall w \in \mathbb{R}^d$, $[\|\nabla f(w)\| = 0] \iff [w$ is a global minimum of $f]$
- Any local minimum of a convex function is a global minimum.

↳ These properties no longer hold for general, nonconvex functions



- For nonconvex functions, there may exist local minima that are not global ("spurious local minima")
- For nonconvex functions, there exist first-order stationary points (i.e. $w \in \mathbb{R}^d$ such that $\|\nabla f(w)\| = 0$) that are not local minima:
  - → Local maxima
  - → Saddle points (maxima in some directions, minima in others)

For these reasons, when applying GD to a nonconvex problem, the goal is to converge to a first-order stationary point

↳ Consistent with the GD iteration

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

If $\nabla f(w_k) = 0$, then $w_{k+1} = w_k$

Th. Suppose that f is $C_L^{1,1}$ and run GD with $\alpha_k = \frac{1}{L}$ $\forall k \geq 0$ for $K \geq 1$ iterations. If f is bounded below ($\exists \bar{f} \in \mathbb{R}$ such that $f(w) \geq \bar{f}$ $\forall w$)

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

"GD converges on nonconvex functions at a rate $\frac{1}{\sqrt{K}}$"

↳ As $K \to \infty$, the rate shows that $\|\nabla f(w_k)\| \to 0$

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

↳ Weaker result than for convex functions
- Applies to $\|\nabla f(w_k)\|$ instead of $f(w_k) - \min_{w \in \mathbb{R}^d} f(w)$
- Rate is worse $\frac{1}{\sqrt{K}}$ vs $\frac{1}{K}$

(3) Benign nonconvexity

↳ Easy to construct examples of nonconvex functions for which GD fails to converge to a minimum

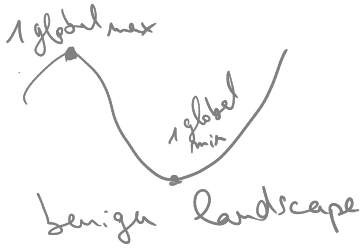
↳ But in practice, GD usually converges to a minimum !

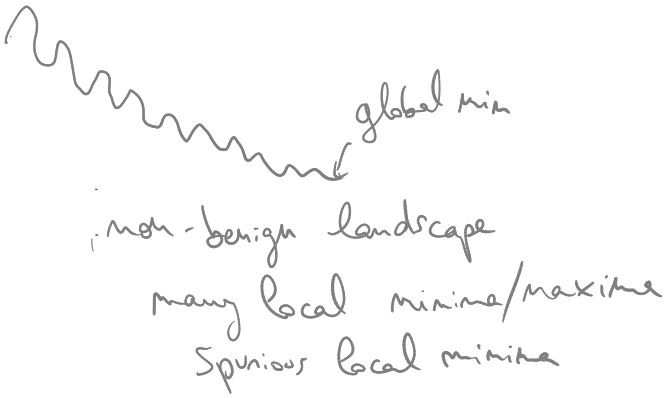
↳ For specific ML problems, GD converges to a minimum (generally local, but often global) and this can be explained by looking at the landscape of those problems
↳ Set of all points with zero gradient (1st-order stationary points)

Def (informal)
A function f has "benign landscape" or "benign nonconvexity" if
- f has no spurious local minima (might imply that all local minima are global)

and/or
- first-order points (aka critical points) are either local minima or strict saddle points

and/or
- GD converges to local minima on these problems





Example of benign nonconvex problem: Low-rank matrix approximation

(1) $$\underset{W \in \mathbb{R}^{d\times d}}{\text{minimize}} \quad \frac{1}{2} \sum_{(i,j)\in S} (W_{ij} - M_{ij})^2$$

$M \in \mathbb{R}^{d\times d}$ data matrix

$S \subseteq \{1,\dots,d\}\times\{1,\dots,d\}$ index of observed entries in M

- Convex problem
- many global optima
  ⇒ one solution: $W_{ij} = M_{ij}$ if $(i,j)\in S$
  $W_{ij} = 0$ otherwise

↳ (1) does not provide any way to approximate the missing entries of M.

↳ But in data analysis, we hypothesize that M can be fully approximated from a subset of its coefficients

Typically $d \gg 1$ and M is assumed to be of rank $r \ll d$ ⇒ Low-rank matrix

(M is rank r if $\exists \bar{U} \in \mathbb{R}^{d\times r}, \exists \bar{V} \in \mathbb{R}^{d\times r}$ such that $M = \bar{U}\bar{V}^T$)

⇒ Can encode all the information in M using $2dr$ entries instead of $d^2 \gg 2dr$

↳ Assuming M is rank r, one tries to build a rank-r approximation of a matrix using $2dr$ variables $\ll d^2$

$$\underset{\substack{U\in\mathbb{R}^{d\times r}\\ V\in\mathbb{R}^{d\times r}}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in S} \left([UV^T]_{ij} - M_{ij}\right) =: f(U,V)$$

UV^T: approximate of M, low rank by construction

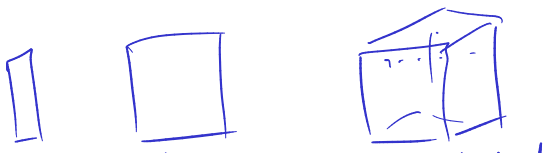
⟶ Nonconvex problem!

⟶ For this particular problem, f has only global minima or saddle points.

If $|S|$ is large enough ($\mathcal{O}(\log(d))$), can guarantee that GD will converge to a global minimum

Other examples of benign nonconvex problems

- Tensor factorization/completion



(Ex) Rank-1 tensor approximation

$$\underset{\substack{u\in\mathbb{R}^d \\ \|u\|=1}}{\text{minimize}} \ \frac{1}{2}\sum_{(i,j,k,\ell)\in S}\left(u_i u_j u_k u_\ell - T_{ijkl}\right)^2$$

T_{ijkl}: tensor data

Tensor approximation

$$\left[u_i u_j u_k u_\ell\right]_{(i,j,k,\ell)\in\{1,\dots,d\}^4} \in \mathbb{R}^{d\times d\times d\times d}$$

- Phase retrieval
- Eigenvalue calculation (connected to PCA)

$$\underset{\substack{\|w\|=1 \\ w\in\mathbb{R}^d}}{\text{minimize}} \ w^T\Sigma w \qquad \Sigma \in \mathbb{R}^{d\times d}$$

$\Sigma = \Sigma^T$ symmetric matrix

Solution set: set of all eigenvectors corresponding to the minimum eigenvalue of Σ

$$\cdot\ \bar{w} \in \underset{\|w\|=1}{\operatorname{argmin}} \ w^T\Sigma w \iff \Sigma\bar{w} = \underbrace{\lambda_{\min}(\Sigma)}_{\text{minimum eigenvalue}}\bar{w}$$

$$\Sigma = P\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_d\end{pmatrix}P^T \qquad P\in\mathbb{R}^{d\times d},\ P=P^T$$

$\lambda_1 \geq \dots \geq \lambda_d = \lambda_{\min}(\Sigma)$ are called the eigenvalues of Σ

All local minima are global

GD converges to such minima for a well-chosen initialization

Theorem (informal) (2016-2019)

Let f be benign nonconvex. Suppose that we run GD from a randomly chosen initial point $w_0 \in \mathbb{R}^d$. Then, with probability 1 (aka almost surely), GD converges to a minimum.