COSC 594/650
Nonconvex optimization

Dongsheng Ding
dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Recap: what convexity buys

Problem: $\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \; f(\theta)$, f convex, L-smooth; minimizer exists

A ZERO GRADIENT IS A CERTIFICATE

$$\nabla f(\theta) \;=\; 0 \quad \Longleftrightarrow \quad f(\theta) \;=\; f^\star$$

Method	convex	σ-strongly convex
GD gap, $\alpha = 1/L$ $\kappa = L/\sigma$	$O(1/k)$ Lecture 8	$O\left((1-1/\kappa)^k\right)$ Lecture 7
SGD (expected)	$\mathbb{E}[f(\bar{\theta}_K) - f^\star] = O(1/\sqrt{K})$ Lecture 20	$\mathbb{E}\|\theta_k - \theta^\star\|^2 = O(1/k)$ Lecture 19

Convexity: local information certifies global optimality

QUESTION: What can we guarantee without convexity?

Nonconvex problems in machine learning

Learning: $\underset{\theta}{\text{minimize}} \quad f(\theta)$, θ: all trainable parameters

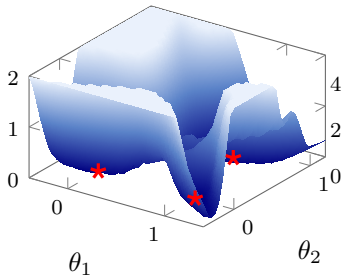
Problem	Objective $f(\theta)$	Parameters
Neural network training	$\frac{1}{m}\sum_{i=1}^{m} \ell\left(h_\theta(x_i), y_i\right)$	Network weights and biases
Low-rank matrix factorization	$\frac{1}{2}\|M - uv^\top\|_F^2$	$\theta = (u, v)$ Factors with r columns
Single neuron	$\frac{1}{m}\sum_{i=1}^{m} \left(\phi(\theta^\top x_i) - y_i\right)^2$	$\theta \in \mathbb{R}^d$ ϕ: ReLU or sigmoid

A **convex prediction loss** need not be convex in θ.

What nonconvex landscapes look like

Neural network training

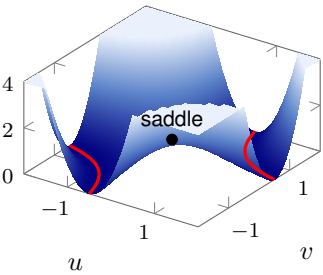
plane through three trained networks



three low-loss fits, barriers

Matrix factorization

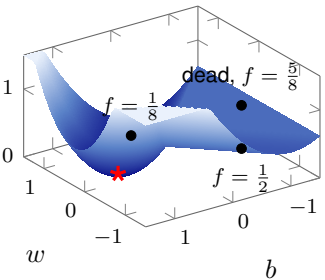
$f(\theta) = (1 - uv)^2$



curve of minima, saddle

Single ReLU neuron

$\theta = (w, b)$, data $(1, 1)$, $(-1, \frac{1}{2})$



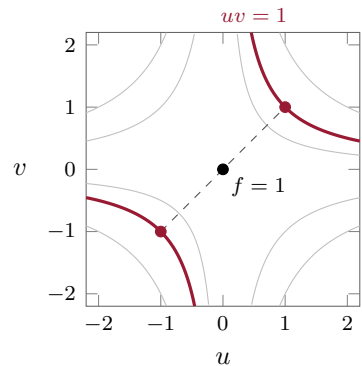
spurious minima, dead region

Nonconvex: many minima, barriers, saddles, flat regions

Symmetric minimizers

Example: $f(\theta) = (1 - uv)^2$, $\theta = (u, v)$ – linear network $x \mapsto v\,(ux)$, data $(1, 1)$

- SYMMETRIC MINIMIZERS
  - $(1, 1)$ and $(-1, -1)$ – both give $f = 0$
  - their midpoint $(0, 0)$ – $f = 1$
- SYMMETRIES IN LEARNING
  - factorization $uv^\top$ – flip to $(-u, -v)$
  - networks – permute hidden units

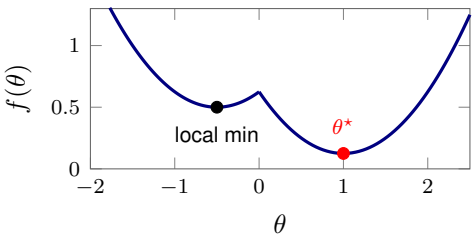


A nonoptimal midpoint rules out convexity

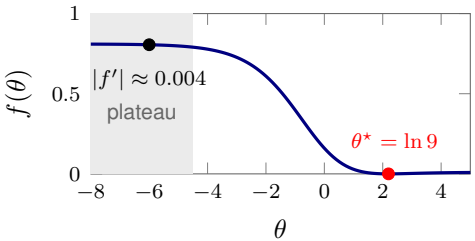
Spurious local minima and plateaus

Single neuron: $f(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\phi(\theta x_i) - y_i)^2, \quad \theta \in \mathbb{R}$

ReLU, data $(1,1)$, $(-1,\frac{1}{2})$



sigmoid, data $(1, 0.9)$



WHY GRADIENT DESCENT CAN STRUGGLE

- ReLU – $\alpha = 0.5$, $\theta_0 < 0$: $\theta_k \to -\frac{1}{2}$, four times the optimal loss
- sigmoid – at $\theta = -6$, slope ≈ -0.004, yet $f \approx 0.81$ and $f^\star = 0$

Local information can mislead: wrong basins, flat regions

Global minimization: how many queries?

Class: $f : [0,1]^d \to \mathbb{R}, \quad |f(\theta) - f(\theta')| \le M\|\theta - \theta'\|_\infty$

Goal: $f(\hat{\theta}) - f^\star \le \epsilon, \quad 0 < \epsilon < M/2$

Oracle model: deterministic value queries; return a queried point

$N^\star(\epsilon)$: fewest queries needed to guarantee the goal for *every* f

$$\left\lfloor \frac{M}{2\epsilon} \right\rfloor^d \;\le\; N^\star(\epsilon) \;\le\; \left\lceil \frac{M}{2\epsilon} \right\rceil^d$$

- UPPER AND LOWER BOUNDS
  - ⋆ upper – grid search succeeds within this budget for *every* f
  - ⋆ lower – *every* method needs at least this many queries for *some* f

Nesterov (2018), *Lectures on Convex Optimization*, Theorems 1.1.1–1.1.2

Upper bound: grid search

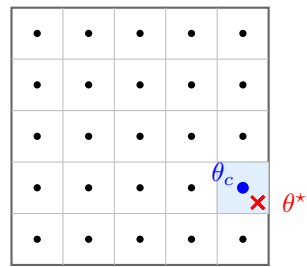
Grid: $p = \lceil M/(2\epsilon) \rceil$, divide $[0,1]^d$ into p^d equal cells

- QUERY EVERY CELL CENTER

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in G_p} f(\theta), \quad G_p = \{\text{all cell centers}\}$$

- COMPARE WITH A MINIMIZER
  - $\star$ $\theta^\star$'s cell center: $\|\theta_c - \theta^\star\|_\infty \le \frac{1}{2p}$

$$\begin{aligned} f(\hat{\theta}) - f^\star &\le f(\theta_c) - f^\star \\ &\le M\|\theta_c - \theta^\star\|_\infty \le \frac{M}{2p} \le \epsilon \end{aligned}$$



grid G_p

Sufficient budget: $N^\star(\epsilon) \le p^d = \lceil M/(2\epsilon) \rceil^d$ value queries

A weaker goal: stationary points

ϵ-stationary point: $\|\nabla f(\theta)\| \le \epsilon$ (not a guarantee of a minimum)

- SMALL GRADIENTS IN OUR EXAMPLES

Loss	Point	∇f	eigenvalues of $\nabla^2 f$	Type
$(1-uv)^2$	$uv = 1$	0	$0,\ 2(u^2+v^2)$	global minima
$(1-uv)^2$	$(0,0)$	0	± 2	saddle
ReLU neuron	$\theta = -\frac{1}{2}$	0	1	spurious local min
sigmoid neuron	$\theta = -6$	≈ 0	≈ 0	plateau, $f \approx 0.81$

Settle for $\|\nabla f(\theta)\| \le \epsilon$, not for $f(\theta) \le f^\star + \epsilon$

Gradient descent without convexity

Assumptions: $\|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|, \quad L > 0, \quad f^\star > -\infty$

Gradient descent: $\theta_{k+1} = \theta_k - \frac{1}{L}\nabla f(\theta_k), \quad k = 0, 1, 2, \ldots$

- DESCENT LEMMA (LECTURE 8)

$$f(\theta') \;\leq\; f(\theta) + \nabla f(\theta)^\top(\theta' - \theta) + \frac{L}{2}\|\theta' - \theta\|^2$$

- PLUG IN $\theta = \theta_k$, $\theta' = \theta_{k+1}$

$$f(\theta_{k+1}) \;\leq\; f(\theta_k) - \frac{1}{2L}\|\nabla f(\theta_k)\|^2$$

  - ⋆ uses smoothness only – no convexity

Every step lowers f by at least $\|\nabla f(\theta_k)\|^2/(2L)$

Rate to an ϵ-stationary point

- SUM OVER $k = 0, \ldots, K-1$, $K \geq 1$

$$\frac{1}{2L} \sum_{k=0}^{K-1} \|\nabla f(\theta_k)\|^2 \;\leq\; f(\theta_0) - f(\theta_K) \;\leq\; f(\theta_0) - f^\star$$

- CHOOSE THE ITERATE WITH THE SMALLEST GRADIENT NORM

$$\min_{0 \leq k < K} \|\nabla f(\theta_k)\| \;\leq\; \sqrt{\frac{2L\,(f(\theta_0) - f^\star)}{K}}$$

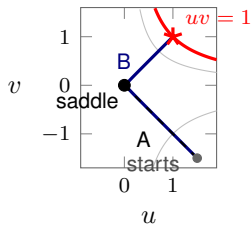
  - $K \geq \lceil 2L\,(f(\theta_0) - f^\star)\,/\epsilon^2 \rceil$ steps – no explicit d

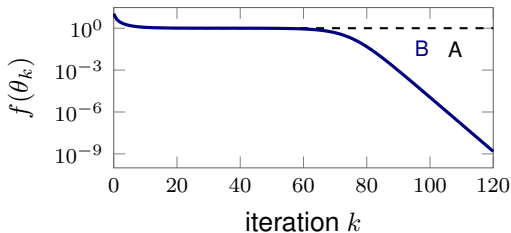
GD finds an ϵ-stationary iterate in $O(1/\epsilon^2)$ steps

Gradient descent near a saddle

Example: $f(\theta) = (1 - uv)^2$, $\theta = (u, v)$, gradient descent with $\alpha = 0.05$

Starts: $\theta_0^A = (1.5, -1.5)$, $\theta_0^B = (1.5, -1.499)$ (10^{-3} apart)





- TWO NEARBY INITIALIZATIONS, TWO OUTCOMES
  - ⋆ A – symmetry preserves $u = -v$: converges to the saddle $(0, 0)$, $f \to 1$
  - ⋆ B – its component along $(1, 1)$ grows: escapes toward $uv = 1$, $f \to 0$

Second-order stationary points

Lipschitz Hessian: $\|\nabla^2 f(\theta) - \nabla^2 f(\theta')\| \leq \rho\|\theta - \theta'\|, \quad \rho > 0$

- SMALL GRADIENTS CAN OCCUR AT SADDLES
  - ⋆ at the previous saddle: $\nabla f(0,0) = 0$, but $\lambda_{\min}(\nabla^2 f(0,0)) = -2$
- ϵ-SECOND-ORDER STATIONARITY, $\epsilon > 0$

$$\|\nabla f(\theta)\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\theta)) \geq -\sqrt{\rho\epsilon}$$

- PERTURBED GRADIENT DESCENT
  - ⋆ small gradient: add a small random perturbation, then resume GD

Jin et al. (2017), *How to Escape Saddle Points Efficiently*.

Three goals and their costs

LANDSCAPES

- ⋆ symmetric minimizers, spurious local minima, plateaus, saddles

Goal	Assumptions	Method	Queries / steps
$f(\theta) - f^\star \le \epsilon$	f M-Lipschitz on $[0,1]^d$	grid search	$\lceil M/(2\epsilon) \rceil^d$
$\|\nabla f(\theta)\| \le \epsilon$	∇f L-Lipschitz	GD	$\lceil 2L\Delta/\epsilon^2 \rceil$
$+\ \lambda_{\min}(\nabla^2 f(\theta)) \ge -\sqrt{\rho\epsilon}$	$+\ \nabla^2 f$ ρ-Lipschitz, $\epsilon \le L^2/\rho$	perturbed GD	$\tilde{O}\left(L\Delta/\epsilon^2\right)$

$0 < \Delta = f(\theta_0) - f^\star < \infty$, $\epsilon > 0$. Last row: probability $\ge 1 - \delta$, $0 < \delta < 1$; $\tilde{O}$ hides logarithms.

Weaker goals, stronger assumptions: lower costs

Questions?

(**Email: dongshed@utk.edu**)