COSC 594/650
Polyak-Łojasiewicz condition

Dongsheng Ding
dongshed@utk.edu

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

Recap: gradient descent without convexity

Smoothness: ∇f is L-Lipschitz, $f^\star = \inf f > -\infty$

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

- ONE STEP

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

- AFTER K STEPS

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

  - ⋆ but a stationary point can be a saddle, with $f > f^\star$

Small gradients, but no control of $f(\theta_k) - f^\star$

QUESTION: **When does a small gradient force a small gap?**

Look back: the strongly convex proof

Strong convexity: $f(\theta') \geq f(\theta) + \nabla f(\theta)^\top (\theta' - \theta) + \frac{\sigma}{2}\|\theta' - \theta\|^2$

- MINIMIZE BOTH SIDES OVER θ'

$$f^\star \;\geq\; f(\theta) - \frac{1}{2\sigma}\|\nabla f(\theta)\|^2$$

  - ⋆ RHS is minimized at $\theta' = \theta - \nabla f(\theta)/\sigma$

- COMBINE WITH ONE STEP (LECTURE 7)

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

The linear rate used strong convexity through one inequality only

The Polyak-Łojasiewicz inequality

$$\mu\textbf{-PL:} \quad \frac{1}{2}\|\nabla f(\theta)\|^2 \;\geq\; \mu\,(f(\theta) - f^\star) \quad \text{for all } \theta, \quad \mu > 0$$

- WHAT IT SAYS
  - $\nabla f(\theta) = 0 \Rightarrow f(\theta) = f^\star$ – every stationary point is a global minimizer
  - $\|\nabla f(\theta)\| \leq \epsilon \Rightarrow f(\theta) - f^\star \leq \epsilon^2/(2\mu)$
- WHAT IT DOES NOT SAY
  - convexity, or a unique minimizer
  - σ-strongly convex $\Rightarrow$ μ-PL with $\mu = \sigma$ – the previous slide

PL: the gradient norm certifies the gap

Linear convergence without convexity

Assumptions: ∇f is L-Lipschitz, f is μ-PL

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

- ONE STEP: DESCENT LEMMA, THEN PL

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

- CONVERGENCE RATE

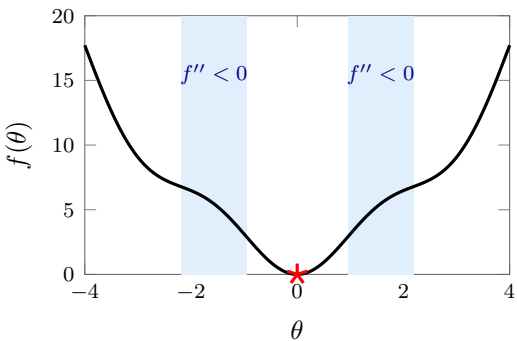
$$f(\theta_k) - f^\star \;\leq\; \left(1 - \frac{\mu}{L}\right)^k (f(\theta_0) - f^\star)$$

  - ϵ-accurate after $\frac{L}{\mu}\log\frac{f(\theta_0)-f^\star}{\epsilon}$ steps – as for strongly convex f

Smoothness and PL give the strongly convex rate

A nonconvex PL function

Example: $f(\theta) = \theta^2 + 3\sin^2\theta, \quad \theta \in \mathbb{R}, \quad f^\star = 0$ at $\theta^\star = 0$



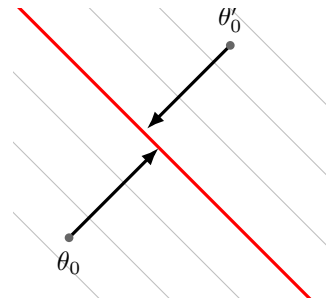
- NONCONVEX
  - $f'' = 2 + 6\cos 2\theta \in [-4, 8]$
- BUT PL
  - $f'(\theta) = 0$ only at $\theta = 0$
  - $\mu = 1/32, \quad L = 8$

PL allows nonconvexity, not bad stationary points

PL with many minimizers: least squares

Problem: $f(\theta) = \frac{1}{2}\|A\theta - b\|^2, \quad A \in \mathbb{R}^{m \times d}$ of rank $0 < r < d$

- NOT STRONGLY CONVEX
  - the minimizers (red) form an affine set
- BUT PL
  - $\mu =$ smallest positive eigenvalue of $A^\top A$
  - gradient descent converges to the minimizer nearest θ_0



PL holds on a whole valley of minimizers

Nonlinear least squares

Problem: $f(\theta) = \frac{1}{2}\|r(\theta)\|^2$, residuals $r : \mathbb{R}^d \to \mathbb{R}^m$, $f^\star = 0$

Jacobian: $J(\theta) \in \mathbb{R}^{m \times d}$, $\nabla f(\theta) = J(\theta)^\top r(\theta)$

- THE GRADIENT NORM

$$\|\nabla f(\theta)\|^2 \;=\; r(\theta)^\top J(\theta) J(\theta)^\top r(\theta) \;\geq\; \lambda_{\min}\left(J(\theta)J(\theta)^\top\right) \cdot 2f(\theta)$$

- PL WHENEVER $J(\theta)J(\theta)^\top \succeq \mu I$ FOR ALL θ
  - ⋆ possible only if $m \leq d$ – at least as many parameters as equations
  - ⋆ over-parameterized models that can fit the data exactly

Over-parameterized least squares: PL without convexity

Where PL fails: back to the saddle

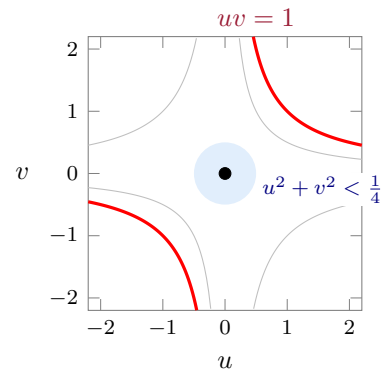
Example: $f(\theta) = \frac{1}{2}(uv-1)^2, \quad r(\theta) = uv - 1, \quad J(\theta) = (v,\ u)$

- AN EXACT IDENTITY

$$\frac{1}{2}\|\nabla f(\theta)\|^2 \;=\; \left(u^2 + v^2\right) f(\theta)$$

  - ⋆ pointwise coefficient $u^2 + v^2$

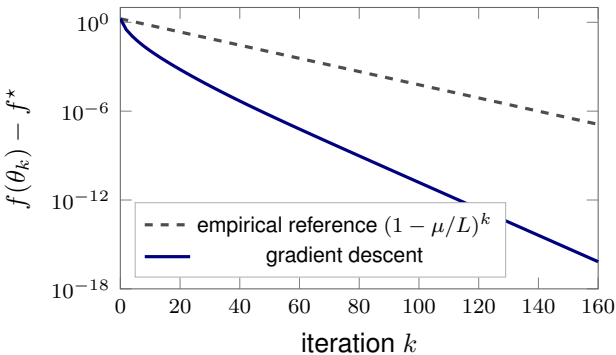
- AT THE ORIGIN
  - ⋆ $J = 0$ – $\nabla f = 0$, yet $f = \frac{1}{2}$
  - ⋆ outside the disk – PL with $\mu = 1/4$



The saddle violates PL

Experiment: an over-parameterized model

Setup: $r(\theta) = \tanh(X\theta) - y$, $m = 50$ equations, $d = 200$ parameters, $\theta_0 = 0$

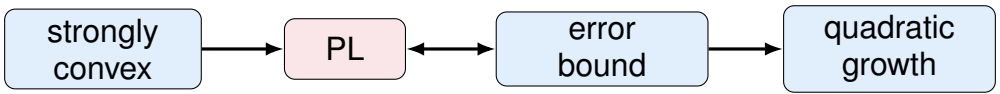


- NONCONVEX
  - ⋆ indefinite Hessians
- ALONG THE PATH
  - ⋆ $\mu = 0.20$, $L = 2.06$
  - ⋆ reference $1 - \mu/L = 0.90$
  - ⋆ observed 0.81 per step

Linear convergence on a nonconvex problem

The PL family

Solution set: $\Theta^\star = \operatorname{argmin} f$, distance $\operatorname{dist}(\theta, \Theta^\star)$



- DEFINITIONS
  - ⋆ error bound – $\|\nabla f(\theta)\| \geq c \operatorname{dist}(\theta, \Theta^\star)$
  - ⋆ quadratic growth – $f(\theta) - f^\star \geq \frac{c}{2} \operatorname{dist}(\theta, \Theta^\star)^2$
- FOR L-SMOOTH f
  - ⋆ PL $\Leftrightarrow$ error bound $\Rightarrow$ quadratic growth – all equivalent if f is convex

PL is the classical error bound, in another form

Stochastic gradient under PL

Problem: $f = \frac{1}{m}\sum_{i=1}^{m} f_i$, each f_i L-smooth, f μ-PL

Iteration: $\theta_{k+1} = \theta_k - \alpha_k \nabla f_{i_k}(\theta_k)$, $\mathbb{E}\|\nabla f_{i_k}(\theta_k)\|^2 \le M^2$ for all k

- CONSTANT STEP $0 < \alpha < 1/(2\mu)$

$$\mathbb{E}\left[f(\theta_k)\right] - f^\star \;\le\; (1 - 2\mu\alpha)^k \left(f(\theta_0) - f^\star\right) + \frac{LM^2\alpha}{4\mu}$$

- SHRINKING STEP $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$

$$\mathbb{E}\left[f(\theta_k)\right] - f^\star \;\le\; \frac{LM^2}{2\mu^2 k}, \qquad k \ge 1$$

  - ⋆ both: i_k uniform independently of the past

Same regimes as Lecture 19, now for the gap

Convex, strongly convex, PL

L-smooth	Stationary $\Rightarrow$ global?	GD gap[†]	SGD expected gap
smooth only	no	best: $O(1/\sqrt{k})$ Lecture 23	–
convex	yes	$O(1/k)$ Lecture 8	avg. $O(1/\sqrt{k})$ Lecture 20
μ-PL	yes, possibly many	$(1-\mu/L)^k$ Today	$O(1/k)$ Today
σ-strongly convex	yes, one minimizer	$(1-\sigma/L)^k$ Lecture 7	$O(1/k)$ Lecture 19[*]

[†] Smooth only: best gradient norm. [*] Lecture 19: squared distance; smoothness gives the gap.

PL: the strongly convex rates, without convexity

Questions?

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