Summary of MA 574 Fall 2024

Part I: Convergence of Gradient Descent

Nesterov's acceleration

Convergence of line search method

Part II: subgradient method

proximal point method

(fast) proximal gradient descent

splitting methods for $\min_x f(x)+g(x)$

PDHG, ADMM, Douglas-Rachford

Part III: randomized coordinate descent

Stochastic gradient descent

Part IV: Riemannian gradient descent

All these methods are first order methods.

using gradient

Usually we can prove at least $\lim \|\nabla f(x_k)\|=0$

But a critical point may not be a global minimizer.

First-order methods almost always avoid strict saddle points

Full Length Paper | Series B | Published: 18 February 2019
Volume 176, pages 311–337, (2019) Cite this article

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan & Benjamin Recht

This paper covers first order methods

such as
- gradient descent
- proximal point method
- coordinate descent
- Riemannian gradient descent

Today we go through the main ideas of this paper

Step I: Dynamical System

Step II: Show that gradient descent with only special initial guess can converge to saddle

Dynamical System for Optimization

Brief intro to stability analysis of dynamical system

Deterministic dynamical system

$$\varphi: T \times X \to X$$
$$(t, x) \mapsto \varphi(t, x)$$

Example: $T = \{0, 1, 2, 3, \dots\}$ $\quad \min f(x)$

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$
$$g = I - \eta \nabla f$$
$$\varphi(t, x) = \underbrace{g \circ g \circ \cdots \circ g}_{t\text{-fold}}(x)$$

Linearization near $x_* = 0$: $\nabla f(x) = \nabla f(x) - \nabla f(x_*) \approx \underbrace{H}_{\text{Hessian}}(x - x_*)$

Linearization gives $\Phi(t, x) = A^t x$

Assume $x_* = 0$ $\qquad = (I - \alpha H)^t x$ $\quad$ power t is an integer

Equilibrium

$$\varphi(t, \bar{x}) = \bar{x}, \quad \forall t \in T$$

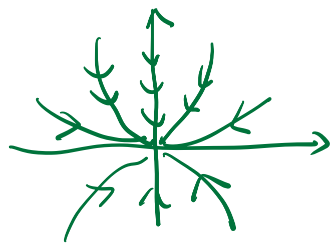
Example: $\bar{x} - \eta \nabla f(\bar{x}) = \bar{x} \iff \nabla f(\bar{x}) = 0$

Stability near Equilibrium

Example: $x_t = A^t x_0$, $A \in \mathbb{R}^{2\times 2}$, $x_0 \in \mathbb{R}^2$

① Let λ_1, λ_2 be eigenvalues of A

$\lambda_1, \lambda_2 \in \mathbb{R}$ $|\lambda_1| < 1$, $|\lambda_2| < 1$, then stable meaning $x_t \to \bar{x} = 0$ for any $x_0 \neq 0$

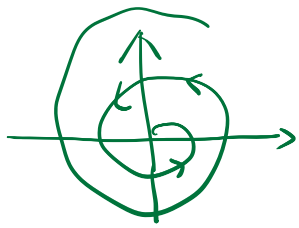


sink

② $\lambda_1, \lambda_2 \in \mathbb{R}$ $|\lambda_1| > 1$, $|\lambda_2| > 1$, source

unstable

③ $\lambda_1, \lambda_2 = re^{\pm i\theta}$, $r < 1$

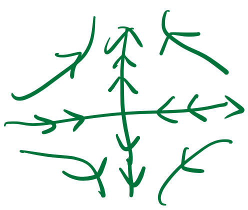


spiral sink

④ $\lambda_1, \lambda_2 = re^{\pm i\theta}$, $r > 1$

spiral source

⑤ $\lambda_1, \lambda_2 \in \mathbb{R}$ $|\lambda_1| < 1$, $|\lambda_2| > 1$



saddle

$\min_x f(x) = \frac{1}{2} x^T H x$

$\nabla^2 f(x) = H$

$(I - \eta \nabla f(x))^t = (I - \eta H)^t$

<u>Reference</u> Micheal Scharb 1987

Global stability of dynamical system

<u>Theorem</u> $g: X \to X$, $\varphi(t, x) = \underbrace{g \circ \cdots \circ g}_{t\text{-fold}}(x) = g^t(x)$

X ↳ finite dim linear space

If ① $g(0) = 0$, $g'(0) = A$

② $X = X_1 \oplus X_2$ (A-invariant decomposition)

$$A(X_1) \subseteq X_1 \qquad A(X_2) \subseteq X_2$$

③ $\|(A|_{X_1})^{-t}\| \le c\lambda^t$, $\lambda \in [0,1)$

$\|(A|_{X_2})^{t}\| \le c\lambda^t$, $\forall t \in T = \mathbb{N}$

A^{-1} on X_1 } is a contraction
A on X_2 }

A on X_1 is an expansion

(Example: $A = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$ $X_1 = \operatorname{Span}\left\{\begin{bmatrix}1\\0\end{bmatrix}\right\}$
$X_2 = \operatorname{Span}\left\{\begin{bmatrix}0\\1\end{bmatrix}\right\}$)

then Locally near 0, there exists Local manifolds M_1, M_2 which are tangent to X_1, X_2 s.t.

① $x \in M_1 \cap B \iff \exists\, x_t \to 0$, s.t. $g^t(x_t) = x$

B: Local ball in X

$x_t = (g^{-1})^t(x) \to 0$

② $x \in M_2 \cap B \iff x_t = g^t(x) \to 0$ ← exponential rate

③ If $x \notin M_2 \cap B$, $\exists t$ s.t. $g^t(x) \notin B$

Remark: ① Only a Local result

② If $X = X_s \oplus X_c \oplus X_u$

$$\begin{array}{ccc} |\lambda|<1 & |\lambda|=1 & |\lambda|>1 \\ M_s & M_c & M_u \end{array}$$

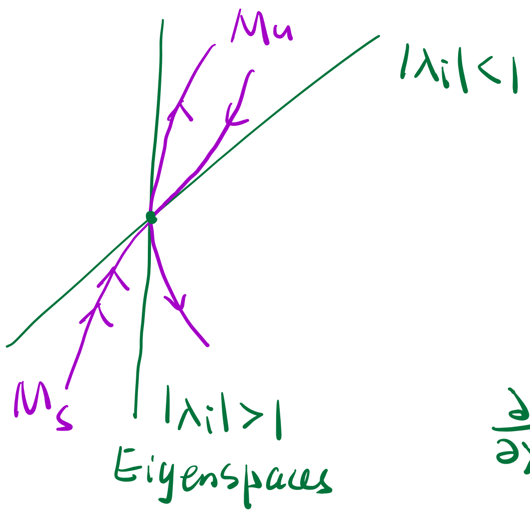
For center manifold M_c, then

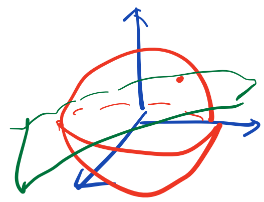
1) $g^t(X)$ escapes B (subexponential rate)

2) $g^t(X) \to 0$ (subexponential rate)

3) $g^t(X) \in B$, but not converging to 0.

Now just focus on a saddle point X_*





strict saddle points

$\nabla f(x_*) = 0 \quad \lambda_{min}(\nabla^2 f(x_*)) < 0$

$$\frac{\partial}{\partial x}(I - \eta \nabla f(x_*)) = I - \eta \nabla^2 f(x_*)$$

Gradient Descent $\quad x_{t+1} = x_t - \alpha \nabla f(x_t) = g(x_t)$

$$g = I - \alpha \nabla f$$

Linearization : $A = I - \alpha H$

$$H = \nabla^2 f(x_*)$$

$$\lambda_i(A) = 1 - \alpha \lambda_i(H)$$

If $\lambda_i(H) < 0$, then $\lambda_i(A) > 1$

$\Rightarrow$ the unstable manifold M_u is non-trivial

$\dim(M_u) \geq 1$

Assume $\lambda_i(H) \neq 0, \forall i \Rightarrow |\lambda_i(A)| \neq 1, \forall i$

$\Rightarrow$ no center manifold

$\Rightarrow \dim(M_s) = d - \dim(M_u) \leq d-1 < d$

$\Rightarrow$ Measure $(M_s \cap B) = 0$

saddle point

If $x_t \to x_*$, then $\exists\, t_0$ s.t.

Theorem from last time

$$x_{t_0} \in M_s \cap B$$

$$x_{t_0 - 1} = g^{-1}(x_{t_0}) \in g^{-1}(M_s \cap B)$$

$$\Rightarrow \quad x_0 \in g^{-t_0}(M_s \cap B)$$

$$\Rightarrow \quad x_0 \in \bigcup_{t=0}^{\infty} g^{-t}(M_s \cap B)$$

$$\Rightarrow \quad \text{measure}\left(\bigcup_{t=0}^{\infty} g^{-t}(M_s \cap B)\right)$$

$$\leq \sum_{t=0}^{\infty} \text{measure}\left(g^{-t}(M_s \cap B)\right)$$

$\Bigg($ **Lemma**: If $\lambda_i(Dg(x)) \neq 0$, $\forall x \in \mathbb{R}^d$

then $\text{meas}(B) = 0 \Rightarrow \text{meas}(g^{-1}(B)) = 0$ $\Bigg)$

$$= 0$$

Countable Union of Measure Zero Sets has measure 0

Theorem If x_* is a strict saddle,

then $\{x_0 \mid x_t \to x_*\}$ has measure 0.

So if we take a random initial guess, then the probability of this initial guess lying in the set $\{x_0 \mid x_t \to x_*\}$ is zero.