Lecture 3 2023-09-02

Last time: propagating ODEs

Today: smooth, unconstrained optimization

$$\min_{x} f(x)$$

subject to: $g_i(x) \leq 0 \quad i = 1, \dots, n_{ineq}$

$h_i(x) = 0 \quad i = 1, \dots, n_{eq}$

---

gradient: if $f: \mathbb{R}^n \to \mathbb{R}$ and $f \in \mathcal{C}^1$

i.e., f has a 1^{st} derivative

$$\nabla f = \begin{pmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{pmatrix} \in \mathbb{R}^n$$

Hessian: if $f: \mathbb{R}^n \to \mathbb{R}$ and $\underline{f \in \mathcal{C}^2}$

f has 2nd derivative

$$\mathcal{H} = \nabla_{xx} f = \begin{pmatrix} \partial f / \partial x_1^2 & \partial f / \partial x_1 \partial x_2 & \cdots & \partial f / \partial x_1 \partial x_n \\ \vdots & & & \\ \partial f / \partial x_n \partial x_1 & \cdots & \cdots & \partial f / \partial x_n^2 \end{pmatrix}$$

symmetric matrix in $\mathbb{R}^{n \times n}$

usage: construct 2nd-order approximation about $\bar{x}$

$$f(x) \approx f(x) + \nabla_x f(\bar{x})^T \delta x + \frac{1}{2} \delta x^T \nabla_{xx} f(\bar{x}) \delta x$$

+ higher order terms

---

e.g. $f(x) = c^T x$ where $c \in \mathbb{R}^n$

$\rightarrow \nabla_x f(x) = c \in \mathbb{R}^n$

Jacobean: if $f: \mathbb{R}^n \to \mathbb{R}^m$ and $f \in \mathcal{C}^1$

$$J = \begin{matrix} \partial f_1/\partial x & \partial f_1/\partial x_1^2 & \partial f_1/\partial x_1 \partial x_n \\ \vdots & & \\ \partial f_m/\partial x & \partial f_m/\partial x_m \partial x_1 & \partial f_m/\partial x_m \partial x_n \end{matrix}$$

usage: $f(x) \approx f(\bar{x}) + J \delta x$ + H.O.T.

e.g. if $A = \begin{pmatrix} — a_1 — \\ \vdots \\ — a_m — \end{pmatrix} \in \mathbb{R}^{m \times n}$

$f(x) = Ax \rightarrow \dfrac{\partial f}{\partial x} = A$

$$\min_{x} \; f(x)$$

$$\mathcal{P} \qquad \text{subj. to: } \; g_i(x) \leq 0 \quad i = 1, \dots, n_{ineq}$$
$$h_i(x) = 0 \quad i = 1, \dots, n_{eq}$$

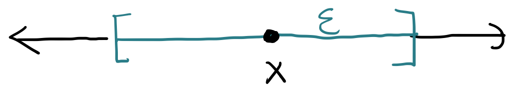
feasible set for $\mathcal{P}$:

$$\Omega = \{ x \in \mathbb{R}^n \mid g_i(x) \leq 0 \;\; \forall i \in [1, n_{ineq}] \text{ and } h_i(x) = 0 \;\; \forall i \in [1, n_{eq}] \}$$

ε- ball:

$$B_\varepsilon(x) = \{ y \in \mathbb{R}^n \mid d(x,y) \leq \varepsilon \}$$

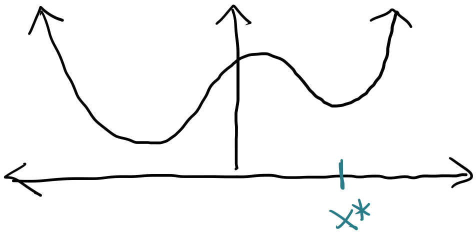
scalar case



Local minimizer: $x^* \in \Omega$ is a local minimizer of

$\mathcal{P}$ if $\exists\, B_\varepsilon(x^*)$ for $\varepsilon > 0$ such that:

$$f(y) \geq f(x^*) \quad \forall\, y \in B_\varepsilon(x^*)$$



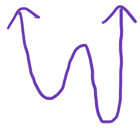
Global minimizer: $x^* \in \Omega$ is a global minimizer of $\mathcal{P}$ if $f(y) \geq f(x^*)$ $\forall\, y \in \Omega$

Salient questions:

1. How to determine if x^* is a local minimizer?
2. How many local minima are there?
3. Is a local minimizer the global one?

Categories:

- Convex vs. non-convex 

- Smooth vs. non-smooth  

- Continuous vs. discrete $x \in \mathbb{R}^n$ $x \in \mathbb{Z}^n$

- Constrained vs. unconstrained

- Deterministic vs. stochastic

Today's focus: smooth and unconstrained

$\overline{P}$

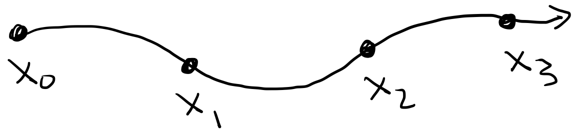
$$\min_x f(x)$$

$x \in \mathbb{R}^n$ (unconstrained)

$f \in \mathcal{C}^2$ (smooth)

Comment on notation: we've used

subscripts to denote "time"



we'll now use superscript to denote iteration:

$$x^{(0)} \xrightarrow{\text{gradient step}} x^{(1)} \xrightarrow{\text{gradient step}} x^{(2)}$$

How does our definition of a local minimizer pertain to solving $\mathcal{P}$?

we want $f(x^*)$ such that $f(x^* + \delta x) \geq f(x^*)$ "close" to x^* $\rightarrow$

$f(x^* + \delta x) \approx f(x^*) + \nabla f(x^*)^T \delta x$

$\rightarrow \quad f(x^* + \delta x) - f(x^*)$

$= f(x^*) + \nabla f(x^*)^T \delta x - f(x^*)$

$= \nabla f(x^*)^T \delta x \geq 0$

we want this to be true $\forall\, \delta x \rightarrow \nabla f(x^*) = 0$

How to solve for this? Use iterative approach

pseudo-code: given $x^{(k)}$

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$$

if $\|\nabla f(x^{(k)})\|_2 \leq \varepsilon_{tol}$

terminate

α is the step size or learning rate

→ two ways we'll discuss today:

1. Newton's method
2. Backtracking line search

1. Newton's method: use 2^{nd}-order Taylor expansion

$$f(\bar{x} + \delta x) \approx f(\bar{x}) + \nabla f(\bar{x})^T \delta x + \frac{1}{2} \delta x^T \underbrace{\nabla_{xx} f(\bar{x})}_{H} \delta x$$

$$= f(\bar{x}) + \nabla f(\bar{x})^T \delta x + \frac{1}{2} \delta x^T H \delta x \quad := \bar{f}$$

$\rightarrow \quad \nabla_x \bar{f} = \nabla f(\bar{x}) + H\,\delta x = 0$

$\rightarrow \quad \delta x = -H^{-1}\nabla f(\bar{x}) = -\left(\nabla_{xx} f(\bar{x})\right)^{-1}\delta x$

$$\alpha = \left(\nabla_{xx} f(\bar{x})\right)^{-1}$$

$\rightarrow$ if $f(x)$ is quadratic, Newton method will converge in one iteration.

Since Newton method computes inverse of $\nabla_{xx} f(\bar{x})$ requires $\Theta(n^3)$ flops $\rightarrow$ intractable for large n

2. backtracking line search: pick largest $\alpha > 0$ such that

$$f(\bar{x} - \alpha \nabla f(\bar{x})) < f(\bar{x})$$

two stronger alternatives include:

Wolfe conditions: if $d^{(k)}$ is descent direction:

1. $$f(x^{(k)} + \alpha d^{(k)}) \leq f(x^{(k)}) + c_1 \alpha d^{(k)T} \nabla f(x^{(k)})$$ Armijo condition

2. $$-d^{(k)T} \nabla f(x^{(k)} + \alpha d^{(k)}) \leq -c_2 d^{(k)T} \nabla f(x^{(k)})$$ curvature condition

Strong Wolfe conditions:

includes preceding two conditions (Armijo + curvature) and adds:

3. $$|d^{(k)T} \nabla f(x^{(k)} + \alpha d^{(k)})| \leq c_2 |d^{(k)T} \nabla f(x^{(k)})|$$

Going back to $\mathcal{P}$, pseudo-code:

given $x^{(0)}$

for $k = 1, \dots, N_{max}$ $\quad d^{(k)} = -\nabla_x f(x^{(k-1)})$

choose α

Line Search$(x^{(k-1)}, \nabla f(x^{(k-1)}))$

Newton$(\nabla f(x^{(k-1)}), \nabla_{xx} f(x^{(k-1)}))$

$x^{(k)} = x^{(k-1)} + \alpha d^{(k)}$

if $\|d^{(k)}\|_2 \le \varepsilon_{tol}$

terminate