Slides for the MFAI (Aug-Dec 2024) Lectures slides for lectures from Sep 25 - Nov 2, 2024

C R Subramanian

CSE Dept, SECS, Indian Institute of Technology, Bhubaneswar.

Lectures on Sep 25 (11am-12pm), Sep 28 (14:40-16:40)

- **Example Motivation :**
- Given : $P = \{(\vec{x}_i, y_i) : \vec{x}_i \in \mathcal{R}^d, y_i \in \mathcal{R})\}_{i=1,\ldots,n}$;
- Find : a function $y = f(\vec{x}) = \vec{a} \cdot \vec{x}$, $\vec{a} \in \mathcal{R}^d$, minimising
- total sum of squares of errors $E(\vec{a}) = \sum_{i=1}^{n} (y_i - \vec{a} \cdot \vec{x}_i)^2$.
-
- Want to find a $\vec{a} \in \mathcal{R}^d$ which minimises $E(\vec{a})$.
-
- When $d = 1$, $E(a)$ becomes a continuous function of one variable a.
- The minimiser a^* and the minimum value $E(a)$ can be computed in $O(n)$ time.

Limits and Continuity

- $f : O \to \mathcal{R}$ is a function. O is an open set. Let $a \in O$.
- limit of $f(x)$ as x approaches a is L if
- $\forall \epsilon > 0 \; \exists \delta > 0$ such that $0 < |x - a| < \delta \Rightarrow |f(x) - L| \leq \epsilon$.
- Denoted by : $Lt_{x \to a} f(x) = L$.
- Left limit : $Lt_{x \to a^-} f(x) = L$. $(-\delta < x - a < 0)$
- Right limit : $Lt_{x \to a^+} f(x) = L$. $(0 < x - a < \delta)$.
-
- L exists if and only if left- and right- limits exist and equal L.
- Example : $f(x) = [x]$ does not have limits when x is an integer ; both left- and right- limits of f exist at integers.
- $f(x) = 1/x$ has limits everywhere but not at $x = 0$. Both left and right limits do not exist at $x = 0$.

Limits and Continuity

- f is *continuous* at a if $f(a)$ is defined and $Lt_{x \to a}\, f(x) = f(a)$.
- $x, x^2, x^3, \sin(x), \cos(x), e^x, |x|$ - continuous everywhere.
- $f(x) = [x]$ continuous everywhere except at integers
- $f(x) = x^{-1}$ continuous everywhere except at $x = 0$.
- f and g are continuous at a. Then, $f + g$, $f - g$, $f \cdot g$ are continuous at a. $g(a) \neq 0 \Rightarrow f/g$ cont. at a.
-
- f is continuous at a, $\quad g$ is continuous at $f(a) \quad \Rightarrow$ $h(x) = g(f(x))$ is continuous at a.
- $sin\left(e^{x^2}\right)$, $e^{sin(x^2)}$ and $\left(e^{sin(x)}\right)^2$ are continuous everywhere.
-
- f is cont. over $[a, b]$ with $f(a) < f(b)$. Then, $\forall c \in (f(a), f(b))\ \ \exists x \in (a, b)$ such that $f(x) = c$.
- f is continuous over $[a, b]$ implies f is bounded over $[a, b]$.
- f is continuous over $[a, b]$ implies f achieves its min and max.

Differentiability

- f is *differentiable* at a if $Lt_{x \to a} \frac{f(x)-f(a)}{x-a}$ exists.
- limit is the *derivative* of f at a, denoted by
- $f'(a)$, $f^{(1)}(a)$, $\frac{df(a)}{dx}$.
- $x, x^2, x^3, e^x, \sin(x), \cos(x)$ -differentiable at every $x \in \mathcal{R}$.
- $|x|$ is differentiable everywhere except at $x = 0$.
-
- f is differentiable at $a \Rightarrow f$ is continuous at a.
- Converse need not be true : $|x|$ and $x = 0$, for example.
- Left-derivative : same except we focus on $x < a$.
- Right-derivative : same except we focus on $x > a$.
- For $|x|$, $f'_L(0) = -1$ and $f'_R(0) = +1$.

Differentiability

- **Algebra :**
- f and g are defined over $\mathcal{R}$.
- $f'(a)$ and $g'(a)$ exist for $a \in \mathcal{R}$.
- $(f \pm g)'(a) = f'(a) \pm g'(a)$.
- $(f \cdot g)'(a) = f(a) \cdot g'(a) + f'(a) \cdot g(a)$.
- $\left(\frac{f}{g}\right)'(a) = \frac{g(a) \cdot f'(a) - f(a) \cdot g'(a)}{g(a)^2}$ provided $g(a) \neq 0$.
-
- **Chain Rule :**
- Suppose $Range(f) \subseteq Domain(g)$; $f'(a)$, $g'(f(a))$ exist.
- $(g(f))'(a)$ exists and equals $g'(f(a)) \cdot f'(a)$.
- Familiar version :
- $y = f(x)$, $z = g(y)$, $z = g(f(x)) \Rightarrow \frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$.

Differentiability

- For $x \in \mathcal{R}$, $B(x, \delta) := \{y \in \mathcal{R} : 0 \leq |y - x| < \delta\}$.
- f is *twice-differentiable* at a if
- (i) for some $\delta > 0$, $f'(x)$ exists for every $x \in B(a, \delta)$
- (ii) derivative of $f'(x)$ $(= Lt_{x \to a} \frac{f'(x) - f'(a)}{x - a})$ exists at a.
- Second derivative is denoted by $f''(a)$, $f^{(2)}(a)$ and $\frac{df^2(a)}{dx^2}$.
-
- Generally, for $k \geq 1$, f is *k-times differentiable* at a if
- (i) for some $\delta > 0$, $f^{(k-1)}(x)$ exists for every $x \in B(a, \delta)$
- (ii) $f^{(k-1)}(x)$ is differentiable at a.
- k-th derivative denoted by $f^{(k)}(a)$ or $\frac{df^k(a)}{dx^k}$.

Differentiability

- $x, x^2, x^3, e^x, \sin(x), \cos(x)$ - k-times differentiable for every $k \geq 1$ and everywhere.
- $f(x) = \log_e x$ - $f^{(k)}(x)$ exists for every $k \geq 1$ for every $x > 0$.
-
- a is a local minimum / local maximum of f if
- $f(a) \leq f(x)$ / $f(a) \geq f(x)$
- for every $x \in B(a, \delta)$ for some $\delta > 0$.
-
- $f : O \to \mathcal{R}$, O is open.
- $a \in O$ is a global minimum / global maximum of f over O if
- $f(a) \leq f(x)$ / $f(a) \geq f(x)$ for every $x \in O$.
- Every global optimum is also a local optimum.

Differentiability and optima

- If a is a local optimum for f, then $f'(a) = 0$.
-
- Necessary but not sufficient.
- Example : $f(x) = x^3$ for $x < 0$ and $f(x) = x^2$ for $x \geq 0$.
- $f'(0) = 0$ but 0 is neither a local minimum nor a local maximum for f.
-
- a is a *saddle point* if $f'(a) = 0$ but a is not a local optimum.
- $f'(a) = 0$ - a is a critical point.

Differentiability and optima

- $f'(a) = 0$ and $f''(a) > 0 \Rightarrow a$ is a local minimum for f.
- sufficient but not necessary.
- Eg : $f(x) = -x^3$ for $x \leq 0$ and $f(x) = x^3$ for $x > 0$.
- 0 is global minimum for f. But, $f'(0) = f''(0) = 0$.
-
- $g'(a) = 0$ and $g''(a) < 0 \Rightarrow a$ is a local maximum for g.
- sufficient but not necessary.
- Eg : $g(x) = -f(x)$
- 0 is global maximum for g. But, $g'(0) = g''(0) = 0$.
-

Taylor's Approximation Formula

- f'' exists and is continuous over $(a-\delta, a+\delta)$ for some $\delta > 0$.
- **Taylor's first-order approximation formula :**
- $f(x) = f(a) + f'(a)(x-a) + E_1(x)$, $\forall x \in B(a, \delta)$
- where $E_1(x) = \int_a^x (x-t)f''(t)dt \to 0$ as $x \to a$.
-
- $E_1(x) = \frac{f''(c)(x-a)^2}{2}$ for some $c \in (a, x)$.
- $f(a+h) = f(a) + hf'(a) + o(h)$ as $h \to 0$.
- $f(a+h) \approx f(a) + hf'(a)$ as $h \to 0$.
-
- differentiability $\iff$ local linearizability.

Taylor's Approximation Formula

- f''' exists and is continuous over $(a-\delta, a+\delta)$ for some $\delta > 0$.
- **Taylor's second-order approximation formula :**
- $f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2} + E_2(x), \forall x \in B(a, \delta)$
- where $E_2(x) = \frac{1}{2} \cdot \int_a^x (x-t)^2 f'''(t)dt \to 0$ as $x \to a$.
-
- $E_2(x) = \frac{f'''(c)(x-a)^3}{6}$ for some $c \in (a, x)$.
- $f(a+h) = f(a) + hf'(a) + \frac{h^2 f''(a)}{2} + o(h^2)$ as $h \to 0$.
- $f(a+h) \approx f(a) + hf'(a) + \frac{h^2 f''(a)}{2}$ as $h \to 0$.

Taylor's Approximation Formula

- $f^{(n+1)}()$ exists, continuous over $(a-\delta, a+\delta)$ for some $\delta > 0$.
- **Taylor's nth-order approximation formula :**
- $f(x) = \sum_{j=0}^{n} \frac{f^{(j)}(a)(x-a)^j}{j!} + E_n(x)$, $\forall x \in B(a, \delta)$
- where $E_n(x) = \frac{1}{n!} \cdot \int_a^x (x-t)^n f^{(n+1)}(t) dt \to 0$ as $x \to a$.
- $f^{(0)}(a) = f(a)$.
-
- $E_n(x) = \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$ for some $c \in (a, x)$.
- $f(a+h) = \sum_{j=0}^{n} \frac{f^{(j)}(a)h^j}{j!} + o(h^n)$ as $h \to 0$.
- $f(a+h) \approx \sum_{j=0}^{n} \frac{f^{(j)}(a)h^j}{j!}$ as $h \to 0$.

Lectures on 16/10/2024

- **Taylor's Formula - illustrations**
- e^x is infinitely differentiable over $\mathcal{R}$.
- $e^x = 1 + x + \frac{x^2}{2} + \ldots + \frac{x^n}{n!} + o(x^n)$, $x \to 0$.
-
- $\log(1+x)$ is infinitely differentiable for every $x > -1$.
- $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \ldots + (-1)^{n-1}\frac{x^n}{n} + o(x^n)$, $x \to 0$.
-
- $\cos x$ is infinitely differentiable for every $x \in \mathcal{R}$.
- $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^n\frac{x^{2n}}{(2n)!} + o(x^n)$, $x \to 0$.
-

Taylor's series

- f is infinitely differentiable over $(a-\delta, a+\delta)$ for some $\delta > 0$.
- **Taylor series expansion for $f(x)$:**
- $f(x) = f(a) + f'(a)(x-a) + \ldots + \frac{f^{(n)}(a)(x-a)^n}{n!} + \ldots.$
-
- $f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)(x-a)^j}{j!}, \forall x \in B(a, \delta)$
-
- $f(a+h) = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)h^j}{j!}, \forall h \in (-\delta, \delta).$
-
- infinite differentiability is necessary but not sufficient.

Optimisation :

- **Problem** : Minimise (or Maximise) $f(x)$ subject to $x \in \Omega$.
- Given : oracle access to computing $f(x)$, $f'(x)$ and $f''(x)$
- and oracle access to testing "$x \in \Omega$?" :
- Goal : Find a $x \in \Omega$ optimising $f(x)$.
-
- **A General Optimisation Algorithm :**

1. Start with an initial guess x.
2. **while** x is not an optimal solution **do**
3. Determine a search direction p ;
4. $x \leftarrow x + p$. **endwhile**
5. Return x.

Optimisation :

- Repeatedly check for local optimality ;
- Check if $f'(x) = 0$ and if $f''(x) \neq 0$.
- Calls for finding zeroes of $f'(x)$.
- Search direction p is guided by the optimality check.
- In special cases like Linear Programs or semi-definite
- programs, other direct and efficient approaches available.
-
- Checking global optimality is a much harder problem.

Newton's Method for finding zeroes :

- Given oracle access to computing f and f',
  Goal : To compute a x^* satisfying $f(x^*) = 0$.
- **Newton's Method for finding roots :**

1. Start with an initial guess x.
2. **while** $f(x) \neq 0$ **and** $f'(x) \neq 0$ **do**
3. $\quad p \leftarrow -\frac{f(x)}{f'(x)}$; $x \leftarrow x + p$. **endwhile**
4. Return x.

-
- One can replace $f(x) \neq 0$ by $|f(x)| > \epsilon$, small ϵ.

Newton's Method - Analysis :

- Analysis of Newton's Method :
-
- x_0 = initial guess ; x_k = guess after k iterations ;
- $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$; $e_k = x_k - x^*$;
- $0 = f(x_k) - e_k f'(x_k) + \frac{f''(\eta_k)e_k^2}{2} \Rightarrow e_k = \frac{f(x_k)}{f'(x_k)} + \frac{f''(\eta_k)e_k^2}{2f'(x_k)}$.
- $e_{k+1} = e_k - \frac{f(x_k)}{f'(x_k)} = e_k^2 \cdot \frac{f''(\eta_k)}{2f'(x_k)} \rightarrow e_k^2 \cdot \frac{f''(x^*)}{2f'(x^*)}$, as $x_k \rightarrow x^*$;
-
- $\forall$ large k, $e_k \approx \left(e_0 \cdot \left(\frac{f''(x^*)}{2f'(x^*)}\right)\right)^{2^k} \cdot \frac{2f'(x^*)}{f''(x^*)}$.
- If $\{x_k\} \rightarrow x^*$, the convergence rate is quadratic with rate constant $\frac{f''(x^*)}{2f'(x^*)}$, that is, $Lt_{k\rightarrow\infty} \frac{|e_{k+1}|}{|e_k|^2} = \frac{f''(x^*)}{2f'(x^*)}$.
- Works fine if x_0 is reasonably close to x^* and rate constant is not too big.

Unconstrained Optimisation in 1D :

- Given : oracle access to computing f' and f'' :
- Optimising $f \iff$ repeatedly finding roots of $f'(x) = 0$.
- Optimising strictly convex $f \iff$ finding a root of $f'(x) = 0$.
-
- By applying Newton's Method for finding roots,
- can find approximations to a root of $f'(x) = 0$ with
- quadratic convergence rate and rate constant $\frac{f'''(x^*)}{2f''(x^*)}$
- where x^* is a root of $f'(x) = 0$.

Gradient-Descent Method :

- Assumption : $|f''(x)| \leq L$ for $x \in [a, b]$.
- Given oracle access to computing f and f',
  Goal : To compute a x^* satisfying $f'(x^*) = 0$.
-

1. Start with an initial guess x. Define $\gamma \leftarrow L^{-1}$.
2. **while** $f'(x) \neq 0$ **do** $x \leftarrow x - \gamma f'(x)$ **endwhile**
3. Return x.

-

Gradient-Descent - Analysis :

- x_k = value of x after k iterations ; $x_{k+1} = x_k - \gamma f'(x_k)$.
- $f(x_{k+1}) \leq f(x_k) - \gamma f'(x_k)^2 + \frac{L\gamma^2 f'(x_k)^2}{2} = f(x_k) - \frac{f'(x_k)^2}{2L}$.
- $f(x_k) < f(x_k)$ for each k. $\{f(x_k)\}_k$ is a decreasing sequence converging to a limit a.
- $\{x_k\}_k$ converges to a limit x^* satisfying $f(x^*) = Lt_k f(x_k)$.
- $Lt_k\ f'(x_k)^2 \leq 2L \cdot Lt_k\,(f(x_k) - f(x_{k+1})) = 0 \Rightarrow f'(x^*) = 0$.
- A local optimum or a saddle point can be approached arbitrarily closely.

Scalar and Vector functions

- $f : \mathcal{R}^n \to \mathcal{R}^m, \; n, m \geq 1$.
- $m = 1$ - real-valued or scalar functions/fields.
- $m > 1$ - vector-valued or vector functions/fields.
- $n = 1$ and $m > 1$ - trajectories (say, of a projectile in 3-space).
-
- $\vec{x} \in \mathcal{R}^d$. $||\vec{x}||_2 = \sqrt{x_1^2 + \ldots + x_d^2}$ - L_2-norm of x.
- $\vec{x}, \vec{y} \in \mathcal{R}^d$. $d_2(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2$ - L_2-distance.
-
- $f : \mathcal{R}^n \to \mathcal{R}^m$. $\vec{a} \in \mathcal{R}^n$, $\vec{l} \in \mathcal{R}^m$.
- $Lt_{\vec{x} \to \vec{a}} f(x) = \vec{l}$ if, $\forall \epsilon > 0$, $\exists \delta > 0$
- satisfying $d_2(f(\vec{x}), \vec{l}) \leq \epsilon$ whenever $0 < d_2(\vec{x}, \vec{a}) \leq \delta$.
- f is continuous at $\vec{a}$ if $Lt_{\vec{x} \to \vec{a}} \vec{f(x)} = f(\vec{a})$.

Scalar and Vector functions

- Suppose $f(\vec{x}) \to \vec{f}$ and $g(\vec{x}) \to \vec{g}$ when $\vec{x} \to \vec{a}$.
- Then, $f(\vec{x}) \pm g(\vec{x}) \to \vec{f} \pm \vec{g}$ as $\vec{x} \to \vec{a}$.
- $\alpha f(\vec{x}) \to \alpha \vec{f}$ as $\vec{x} \to \vec{a}$ for every $\alpha \in \mathcal{R}$.
- $||f(\vec{x})||_2 \to ||\vec{f}||_2$ as $\vec{x} \to \vec{a}$.
- $f(\vec{x}) \cdot g(\vec{x}) \to \vec{f} \cdot \vec{g}$ as $\vec{x} \to \vec{a}$.
-
- $f : \mathcal{R}^n \to \mathcal{R}^m$ defined by $f(\vec{x}) = (f_1(\vec{x}), \ldots, f_m(\vec{x}))$ for each x.
- f is continuous at $\vec{a}$ if and only if each f_i is continuous at $\vec{a}$.
-

Scalar and Vector functions

- Suppose $f : \mathcal{R}^n \to \mathcal{R}^m$ and $g : \mathcal{R}^m \to \mathcal{R}^p$. Define $h = g \cdot f : \mathcal{R}^n \to \mathcal{R}^p$ by $h(\vec{x}) = g(f(\vec{x}))$ for each $x \in \mathcal{R}^n$.
- Suppose also that f is continuous at $\vec{a} \in \mathcal{R}^n$ and g is continuous at $f(\vec{a}) \in \mathcal{R}^m$. Then, h is continuous at $\vec{a}$.
-
- $f_1, f_2, f_3, f_4 : \mathcal{R}^2 \to \mathcal{R}$ be defined by
- $f_1(x, y) = \sin(x^2 y)$; $f_2(x, y) = \log_e(x^2 + y^2)$;
- $f_3(x, y) = \frac{e^{x+y}}{x+y}$;
- f_1 is continuous everywhere ;
- f_2 is continuous everywhere except at $(0, 0)$.
- f_3 is continuous everywhere except on the line $x + y = 0$.
-
- $f(x, y) = \frac{xy}{x^2+y^2}$ for $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$.
- f is continuous as a function of x alone and as a function of y alone but not as function of x and y both.

Differentiability of Scalar functions

- $f : \mathcal{R}^n \rightarrow \mathcal{R}$ may have different derivatives along different directions at a point $\vec{a}$.
- Focus on specific directions $\vec{y}$.
-
- $\vec{a}, \vec{y} \in \mathcal{R}^n$. The derivative of f at $\vec{a}$ along $\vec{y}$ is defined as
- $Lt_{h \rightarrow 0} \frac{f(\vec{a}+h\vec{y})-f(\vec{a})}{h}$. Denoted by $f'(\vec{a}, \vec{y})$ or $\frac{df(\vec{a})}{d\vec{y}}$.
- $f'(\vec{a}, \vec{0}) = 0$ always for any $\vec{a}$.
- When $||\vec{y}||_2 = 1$, $f'(\vec{a}, \vec{y})$ is the directional derivative of f at $\vec{a}$.
- When $\vec{y} = e_i$ along x_i axis, $f'(\vec{a}, e_i) = \frac{\partial f(\vec{a})}{\partial x_i}$.
- the gradient of f at $\vec{a}$ is the vector
- $\nabla f(\vec{a}) = \left(\frac{\partial f(\vec{a})}{\partial x_1}, \ldots, \frac{\partial f(\vec{a})}{\partial x_n} \right)$.

Differentiability of Scalar functions

- Existence of directional derivatives $f'(\vec{a}, \vec{y})$ for each $\vec{y}$ does not guarantee f is continuous at $\vec{a}$.
- Example : $f(y) = \frac{xy^2}{x^2+y^4}$ for $x \neq 0$ and $f(0, y) = 0$ for all y.
- $f'((0,0), \vec{y})$ exists for each $y \in \mathcal{R}^n$.
- Along the parabola $x = y^2$, $f(x, y) = 1/2$ and f is not continuous at $(0, 0)$.
-
- f is **differentiable** at $\vec{a}$ if, for some $r > 0$, there exist a LT
- $T_{\vec{a}} : \mathcal{R}^n \rightarrow \mathcal{R}$ and a scalar function $E_{\vec{a}}(\vec{y})$ such that
- $f(\vec{a} + \vec{y}) = f(\vec{a}) + T_{\vec{a}}(\vec{y}) + ||\vec{y}||_2 \cdot E_{\vec{a}}(\vec{y})$ holds true for all $||\vec{y}|| < r$ and $E_{\vec{a}}(\vec{y}) \rightarrow 0$ as $||\vec{y}|| \rightarrow 0$.
- $T_{\vec{a}}$ is the *Total Derivative* of f at $\vec{a}$, denoted also by $f'(\vec{a})$.

Differentiability of Scalar functions

- f is differentiable at $\vec{a} \Longrightarrow T_{\vec{a}}(\vec{y}) = f'(\vec{a}, \vec{y})$ for each $\vec{y}$.
- Also, $T_{\vec{a}}(\vec{y}) = \nabla\, f(\vec{a}) \cdot \vec{y} = \sum_{i=1}^{n} \frac{\partial f(\vec{a})}{\partial x_i} \cdot y_i$ for each $\vec{y}$.
- f is differentiable at $\vec{a} \Longrightarrow f$ is continuous at $\vec{a}$.
-
- f is differentiable at $\vec{a} \Longrightarrow$ Taylor's first order formula :
- $f(\vec{a} + \vec{y}) = f(\vec{a}) + \nabla f(\vec{a}) \cdot \vec{y} + ||y|| E_{\vec{a}}(\vec{y}),\ ||y|| < r$.
-
- When $||y|| = 1$, $f'(\vec{a}, \vec{y}) = ||\nabla f(\vec{a})|| \cdot \cos(\theta)$ where
- $\theta =$ angle between $\nabla f(\vec{a})$ and $\vec{y}$.
- $f'(\vec{a}, \vec{y}) =$ component of $\nabla f(\vec{a})$ in the direction of $\vec{y}$.

Sufficiency for Differentiability and Chain Rule

- f is a scalar function over $\mathcal{R}^n$ and $\vec{a} \in \mathcal{R}^n$.
- If all first-order partial derivatives exist at all points in an open neighborhood around $\vec{a}$ and they are continuous at $\vec{a}$, then f is differentiable at $\vec{a}$.
-
- **Chain Rule :** $r : O \to S$, $f : S \to \mathcal{R}$, $O \subseteq \mathcal{R}$, $S \subseteq \mathcal{R}^n$.
- Suppose $r'(t)$ exists and $f'(r(t))$ exists. Then, for $g = f \cdot r$,
- $g'(t)$ exists and $g'(t) = \nabla f(r(t)) \cdot r'(t)$.
-
- Write $r(t) = (r_1(t), \ldots, r_n(t))$.
- $r'(t) = (r_1'(t), \ldots, r_n'(t))$.
- $\nabla f(r(t)) = \left(\frac{\partial f(r(t))}{\partial r_1}, \ldots, \frac{\partial f(r(t))}{\partial r_n}\right)$.
- $g'(t) = \sum_{i=1}^{n} \frac{\partial f(r(t))}{\partial r_i} \cdot \frac{dr_i(t)}{dt}$.

Higher-order derivatives for Scalar functions

- $f : O \to \mathcal{R}$, $O \subseteq \mathcal{R}^n$, O is open.
- Suppose $f'(\vec{x})$ exists for every $\vec{x} \in B(\vec{a}, r)$.
- Derivative of f' at $\vec{a}$, if it exists, is the second-derivative $f''(\vec{a})$.
- Our Focus : Second-order partial derivatives - $\frac{\partial^2 f(\vec{a})}{\partial x_i \partial x_j}$.
- Hessian (denoted by $\nabla^2 f(\vec{a})$) is the matrix $\left(\frac{\partial^2 f(\vec{a})}{\partial x_i \partial x_j}\right)_{i,j}$.
- Hessian is symmetric if the second-order pds are continuous.

Taylor's approximation

- $f : O \to \mathcal{R},\ O \subseteq \mathcal{R}^n,\ \vec{a} \in O.$
- second-order pds are continuous.
- $f(\vec{a}+\vec{p}) = f(\vec{a}) + \vec{p}^T \cdot \nabla f(\vec{a}) + \frac{\vec{p}^T \cdot \nabla^2 f(\vec{a}) \cdot \vec{p}}{2} + \ldots.$
-
- $f(\vec{a}+\vec{p}) = f(\vec{a}) + \vec{p}^T \cdot \nabla f(\vec{a}) + \frac{\vec{p}^T \cdot \nabla^2 f(\vec{\eta}) \cdot \vec{p}}{2}$
- for some $\vec{\eta} \in L(\vec{a}, \vec{a}+\vec{p})$.
-
- $f(\vec{a}+\vec{p}) = f(\vec{a}) + \sum_{i=1}^{n} \ p_i \frac{\partial f(\vec{a})}{\partial x_i} + \sum_{i,j=1}^{n} \ p_i p_j \frac{\partial^2 f(\vec{\eta})}{\partial x_i \partial x_j}$
-
- $f(\vec{a}+\vec{p}) = f(\vec{a}) + \sum_{i=1}^{n} \ p_i \frac{\partial f(\vec{a})}{\partial x_i} + o(||\vec{p}||)$ as $\vec{p} \to \vec{0}$.
-
- Linear approximation : $f(\vec{a}+\vec{p}) \approx f(\vec{a}) + \vec{p}^T \cdot \nabla f(\vec{a})$.

Example (from Griva, Nash and Sofer)

- Consider $f(x,y) = x^3 + 5x^2y + 7xy^2 + 2y^3$. Let $\vec{a} = (-2,3)$.
- $\nabla f(\vec{a}) = (3x^2 + 10xy + 7y^2, 5x^2 + 14xy + 6y^2)_{(-2,3)} = (15,-10)$.
- $\nabla^2 f(\vec{a}) = \begin{pmatrix} 6x+10y & 10x+14y \\ 10x+14y & 14x+12y \end{pmatrix}_{(-2,3)} = \begin{pmatrix} 18 & 22 \\ 22 & 8 \end{pmatrix}$
- Let $\vec{p} = (0.1, 0.2)$.
- $f(\vec{a}+\vec{p}) = f(-1.9, 3.2) \approx f(\vec{a}) + \vec{p}^T \cdot \nabla f(\vec{a}) + \frac{\vec{p}^T \cdot \nabla^2 f(\vec{a}) \cdot \vec{p}}{2}$.
- $f(-1.9, 3.2) \approx -20 - 0.5 + 0.69 = -19.81$
- Actual $f(-1.9, 3.2) = -19.755$.

Unconstrained minimisation of scalar functions

- f is a scalar function.
- $\vec{a}$ is a local minimum for $f \Rightarrow \nabla f(\vec{a})^T \cdot \vec{p} \geq 0$ for all $\vec{p}$.
- $\vec{a}$ is a local minimum for $f \Rightarrow \nabla f(\vec{a}) = \vec{0}$.
- Necessary but not sufficient.
- $\vec{a}$ is a local minimum for $f \Rightarrow \nabla^2 f(\vec{a})$ is positive semi-definite.
-
- Sufficiency :
- $\nabla f(\vec{a}) = \vec{0}$ and $\nabla^2 f(\vec{a})$ is positive definite $\Rightarrow \vec{a}$ is a local minimum.
- A matrix B is positive semi-definite ($B \succeq 0$) if $x^T Bx \geq 0$ for all $x \in \mathcal{R}^n$.
- A matrix B is positive definite ($B \succ 0$) if $x^T Bx > 0$ for all $x \neq \vec{0}$.

Unconstrained Minimization : Newton's Method

- $f : O \to \mathcal{R}$, $O \subseteq \mathcal{R}^n$, O is open set.
- Given oracle access to computing ∇f and $\nabla^2 f$,
  Goal : To compute a local minimizer $\vec{x}^*$ of f.
- **Newton's Method for Minimizing :**

1. Start with an initial guess $\vec{x}$.
2. **while** $\nabla f(\vec{x}) \neq \vec{0}$ **and** $\nabla^2 f(\vec{x}) \succ 0$ **do**
3. $p \leftarrow -\left(\nabla^2 f(\vec{x})\right)^{-1} \cdot \nabla f(\vec{x})$; $x \leftarrow x + p$. **endwhile**
4. Return x.

-
- In practice, one replaces $\nabla f(\vec{x}) \neq \vec{0}$ by $||\nabla f(\vec{x})|| > \epsilon$, small ϵ.

Unconstrained Minimization : Newton's Method

- Obtained by minimizing the RHS of the quadratic approximation :
- $f(\vec{x}) \approx f(\vec{x}_k) + \nabla f(x_k)(\vec{x} - \vec{x}_k) + \frac{(\vec{x}-\vec{x}_k)\nabla^2 f(\vec{x}_k)(\vec{x}-\vec{x}_k)}{2}$.
- $\nabla^2 f$ is Lipschitz continuous on O, that is, $||\nabla^2 f(\vec{x}) - \nabla^2 f(\vec{y})|| \leq L||\vec{x} - \vec{y}||, \forall \vec{x}, \vec{y} \in O$.
- $\vec{x}^*$ - minimizer of f and $\nabla^2 f(\vec{x}^*) \succ 0$.
- If $||\vec{x} - \vec{x}^*||$ is "sufficiently small", then $\{\vec{x}_k\}_k$ converges quadratically to $\vec{x}^*$.
-

Unconstrained Minimization : Gradient-Descent Method :

- Descent along direction of Steepest Descent, namely, $-\nabla f(\vec{x})$.
- Assumption : $||\nabla^2 f(\vec{x})|| \leq L$ for $x \in O$, for some $L > 0$.
- Given oracle access to computing $\nabla f()$ and $f()$,
  Goal : To compute a $\vec{x}^*$ satisfying $\nabla f(\vec{x}^*) = \vec{0}$.
-

1. Start with an initial guess x. Define $\gamma \leftarrow L^{-1}$.
2. **while** $\nabla f(\vec{x}) \neq \vec{0}$ **do** $\vec{x} \leftarrow \vec{x} - \gamma \nabla f(\vec{x})$ **endwhile**
3. Return $\vec{x}$.

-
- In practice, one replaces $\nabla f(\vec{x}) \neq \vec{0}$ by $||\nabla f(\vec{x})|| > \epsilon$, small ϵ.

Minimization of Scalar functions : Grad-Des. - Analysis :

- $\vec{x}_k$ = value of $\vec{x}$ after k iterations ; $\vec{x}_{k+1} = \vec{x}_k - \gamma \nabla(\vec{x}_k)$.
- $f(\vec{x}_{k+1}) \leq f(\vec{x}_k) - \gamma ||\nabla f(\vec{x}_k)||^2 + \frac{\gamma^2 ||\nabla^2 f(\vec{x}_k)|| \cdot ||\nabla f(\vec{x}_k)||^2}{2}$
- $\qquad\qquad = f(\vec{x}_k) - \frac{||\nabla f(\vec{x}_k)||^2}{2L}$.
- $f(\vec{x}_{k+1}) < f(\vec{x}_k)$ for each k. $\{f(\vec{x}_k)\}_k$ is a decreasing sequence converging to a limit a.
- As in the $1D$-case, $\{\vec{x}_k\}_k$ converges to a limit $\vec{x^*}$ satisfying $f(\vec{x^*}) = Lt_k f(\vec{x}_k)$.
- $Lt_k \, ||\nabla f(\vec{x}_k)||^2 \leq 2L \cdot Lt_k \, (f(\vec{x}_k) - f(\vec{x}_{k+1})) = 0$ $\Rightarrow \nabla f(\vec{x^*}) = \vec{0}$.
- A local optimum or a saddle point can be approached arbitrarily closely.

Gradient Descent with Backtracking Line Search :

- Presumes apriori knowledge of L. Possibly not available.
-
- $x_0 \leftarrow$ initial guess of $\vec{x}^*$; $n \leftarrow 0$;
- **while** $\nabla f(\vec{x}_n) \neq \vec{0}$ **do**
- $\gamma_n \leftarrow$ initial estimate of Step size γ ;
- **while** $f(\vec{x}_n - \gamma_n \nabla f(\vec{x}_n)) > f(\vec{x}_n) - \frac{\gamma_n ||\nabla f(\vec{x}_n)||^2}{2}$ **do**
- $\gamma_n \leftarrow \gamma_n / 2$ **endwhile**
- $\vec{x}_{n+1} \leftarrow \vec{x}_n - \gamma_n \nabla f(\vec{x}_n)$; $n \leftarrow n + 1$. **endwhile**
- Return $\vec{x}_n$.
-
- In practice, one replaces $\nabla f(\vec{x}_n) \neq \vec{0}$ by $||\nabla f(\vec{x}_n)|| > \epsilon$.
- Takes care of narrow, deep valleys and chooses γ adaptively.

Newton's method (NM) vs Gradient descent (GD)

- GD guarantees convergence while NM can fail if Hessian is not positive definite.
- NM provides quadratic rate of convergence if $\vec{x}_0$ is "reasonably close" local minimum.
- NM is computationally expensive (computing Hessian and its inverse) and also suffers from numerical instabilities.
- Where applicable, NM converges much faster than GD if we start within a suitable neighborhood.
- For GD, choose step size small in regions of greater variability of the gradient and large in regions of small variability.