Recitation 1

18-460/18-660 Optimization

Chaoyi Pan

Carnegie Mellon University

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.
Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f(\mathbf{x}) \\ \partial_{x_2} f(\mathbf{x}) \\ \vdots \\ \partial_{x_n} f(\mathbf{x}) \end{bmatrix}$$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.
Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f(\mathbf{x}) \\ \partial_{x_2} f(\mathbf{x}) \\ \vdots \\ \partial_{x_n} f(\mathbf{x}) \end{bmatrix}$$

Example. $f(\mathbf{x}) = 3x_1^2 + 5 \log x_2$. $\nabla f(\mathbf{x}) = \begin{bmatrix} \quad \\ \quad \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.
Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f(\mathbf{x}) \\ \partial_{x_2} f(\mathbf{x}) \\ \vdots \\ \partial_{x_n} f(\mathbf{x}) \end{bmatrix}$$

Example. $f(\mathbf{x}) = 3x_1^2 + 5\log x_2$. $\nabla f(\mathbf{x}) = \begin{bmatrix} 6x_1 \\ \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f(\mathbf{x}) \\ \partial_{x_2} f(\mathbf{x}) \\ \vdots \\ \partial_{x_n} f(\mathbf{x}) \end{bmatrix}$$

Example. $f(\mathbf{x}) = 3x_1^2 + 5\log x_2$. $\nabla f(\mathbf{x}) = \begin{bmatrix} 6x_1 \\ 5x_2^{-1} \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Because $f \in \mathbf{C}^2$, $\partial^2_{x_ix_j} f = \partial^2_{x_jx_i} f$, and $\nabla_f^2(\mathbf{x})$ is symmetric.

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Because $f \in \mathbf{C}^2$, $\partial^2_{x_ix_j} f = \partial^2_{x_jx_i} f$, and $\nabla_f^2(\mathbf{x})$ is symmetric.

Example. $f(\mathbf{x}) = x_1^3 + 5x_2 + 7x_1x_2$. $\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \quad \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Because $f \in \mathbf{C}^2$, $\partial^2_{x_ix_j} f = \partial^2_{x_jx_i} f$, and $\nabla_f^2(\mathbf{x})$ is symmetric.

Example. $f(\mathbf{x}) = x_1^3 + 5x_2 + 7x_1x_2$. $\nabla_f^2(\mathbf{x}) = \begin{bmatrix} 6x_1 & \\ & \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Because $f \in \mathbf{C}^2$, $\partial^2_{x_ix_j} f = \partial^2_{x_jx_i} f$, and $\nabla_f^2(\mathbf{x})$ is symmetric.

Example. $f(\mathbf{x}) = x_1^3 + 5x_2 + 7x_1x_2$. $\nabla_f^2(\mathbf{x}) = \begin{bmatrix} 6x_1 & 7 \\ 7 & \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable.
Hessian:

$$\nabla_f^2(\mathbf{x}) = \begin{bmatrix} \partial^2_{x_1x_1} f(\mathbf{x}) & \partial^2_{x_1x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \partial^2_{x_2x_1} f(\mathbf{x}) & \partial^2_{x_2x_2} f(\mathbf{x}) & \cdots & \partial^2_{x_1x_n} f(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{x_nx_1} f(\mathbf{x}) & \partial^2_{x_nx_2} f(\mathbf{x}) & \cdots & \partial^2_{x_nx_n} f(\mathbf{x}) \end{bmatrix}$$

Because $f \in \mathbf{C}^2$, $\partial^2_{x_ix_j} f = \partial^2_{x_jx_i} f$, and $\nabla_f^2(\mathbf{x})$ is symmetric.

Example. $f(\mathbf{x}) = x_1^3 + 5x_2 + 7x_1x_2$. $\nabla_f^2(\mathbf{x}) = \begin{bmatrix} 6x_1 & 7 \\ 7 & 0 \end{bmatrix}$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}}(\mathbf{y} - \mathbf{x})\mathrm{d}t$$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}} (\mathbf{y} - \mathbf{x}) \mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}}(\mathbf{y}-\mathbf{x})\mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$*.*

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}}(\mathbf{y} - \mathbf{x})\mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$*.*

$\mathbf{z}(0) = \mathbf{x},\ \mathbf{z}(1) = \mathbf{y}.$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}}(\mathbf{y} - \mathbf{x})\mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$.

$\mathbf{z}(0) = \mathbf{x},\ \mathbf{z}(1) = \mathbf{y}$.

$$h'(t) \ = \ \nabla f(\mathbf{z}(t))^{\mathsf{T}} \begin{bmatrix} z_1'(t) \\ \vdots \\ z_n'(t) \end{bmatrix}$$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^\mathsf{T} (\mathbf{y} - \mathbf{x}) \mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$.

$\mathbf{z}(0) = \mathbf{x}, \ \mathbf{z}(1) = \mathbf{y}.$

$$h'(t) = \nabla f(\mathbf{z}(t))^\mathsf{T} \begin{bmatrix} z_1'(t) \\ \vdots \\ z_n'(t) \end{bmatrix} = \nabla f(\mathbf{z}(t))^\mathsf{T} \begin{bmatrix} y_1 - x_1 \\ \vdots \\ y_n - x_n \end{bmatrix}$$

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^{\mathsf{T}}(\mathbf{y} - \mathbf{x})\mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$.

$\mathbf{z}(0) = \mathbf{x},\ \mathbf{z}(1) = \mathbf{y}$.

$$h'(t) \ = \ \nabla f(\mathbf{z}(t))^{\mathsf{T}} \begin{bmatrix} z_1'(t) \\ \vdots \\ z_n'(t) \end{bmatrix} \ = \ \nabla f(\mathbf{z}(t))^{\mathsf{T}} \begin{bmatrix} y_1 - x_1 \\ \vdots \\ y_n - x_n \end{bmatrix} \ =$$

$\nabla f(\mathbf{z}(t))^{\mathsf{T}}(\mathbf{y} - \mathbf{x})$.

Recap from Calculus

$f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^\mathsf{T}(\mathbf{y}-\mathbf{x})\mathrm{d}t$$

Integration of *projection* of gradients along the line through $\mathbf{x}$ and $\mathbf{y}$.

Proof. *Let* $\mathbf{z}(t) = t\mathbf{y} + (1-t)\mathbf{x}$ *and* $h(t) = f(\mathbf{z}(t))$.

$\mathbf{z}(0) = \mathbf{x},\ \mathbf{z}(1) = \mathbf{y}$.

$h'(t) \ = \ \nabla f(\mathbf{z}(t))^\mathsf{T} \begin{bmatrix} z_1'(t) \\ \vdots \\ z_n'(t) \end{bmatrix} \ = \ \nabla f(\mathbf{z}(t))^\mathsf{T} \begin{bmatrix} y_1 - x_1 \\ \vdots \\ y_n - x_n \end{bmatrix} \ = \ \nabla f(\mathbf{z}(t))^\mathsf{T}(\mathbf{y}-\mathbf{x})$.

$f(\mathbf{y}) - f(\mathbf{x}) = h(1) - h(0) = \int_0^1 h'(t)\mathrm{d}t = \int_0^1 \nabla f\big(t\mathbf{y} + (1-t)\mathbf{x}\big)^\mathsf{T}(\mathbf{y}-\mathbf{x})\mathrm{d}t.$

Recap from Calculus

$f : \mathbb{R} \to \mathbb{R}$ is *twice* continuously differentiable.
For any $x, y \in \mathbb{R}$, there exists $t \in [0, 1]$ such that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(ty + (1 - t)x)(y - x)^2.$$

Recap from Calculus

$f : \mathbb{R} \to \mathbb{R}$ is *twice* continuously differentiable.
For any $x, y \in \mathbb{R}$, there exists $t \in [0, 1]$ such that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(ty + (1 - t)x)(y - x)^2.$$

What if f is multi-variate?

Recap from Calculus

$f : \mathbb{R} \to \mathbb{R}$ is *twice* continuously differentiable.
For any $x, y \in \mathbb{R}$, there exists $t \in [0, 1]$ such that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(ty + (1 - t)x)(y - x)^2.$$

What if f is multi-variate?
Hint: Define $h(s) = f\left(s\mathbf{y} + (1 - s)\mathbf{x}\right)$, and do Taylor's expansion of $h(1)$ on $s = 0$.

Recap from Calculus

$f : \mathbb{R} \to \mathbb{R}$ is *twice* continuously differentiable.
For any $x, y \in \mathbb{R}$, there exists $t \in [0, 1]$ such that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2} f''(ty + (1 - t)x)(y - x)^2.$$

What if f is multi-variate?
Hint: Define $h(s) = f\big(s\mathbf{y} + (1 - s)\mathbf{x}\big)$, and do Taylor's expansion of $h(1)$ on $s = 0$.

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathsf{T}}(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^{\mathsf{T}} \nabla_f^2 \big(ty + (1 - t)x\big)(\mathbf{y} - \mathbf{x})$$

Cauchy-Schwarz Inequality

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2.$$

Cauchy-Schwarz Inequality

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{x}^\mathsf{T}\mathbf{y} \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Proof. *Consider (univariate) quadratic function*

$$h(t) = \frac{1}{2}(t\mathbf{x} + \mathbf{y})^\mathsf{T}(t\mathbf{x} + \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 t^2 + \mathbf{x}^\mathsf{T}\mathbf{y}t + \frac{1}{2}\|\mathbf{y}\|_2^2.$$

Cauchy-Schwarz Inequality

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{x}^T \mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Proof. *Consider (univariate) quadratic function*

$$h(t) = \frac{1}{2}(t\mathbf{x} + \mathbf{y})^T (t\mathbf{x} + \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 t^2 + \mathbf{x}^T \mathbf{y} t + \frac{1}{2}\|\mathbf{y}\|_2^2.$$

$$h(t) \geq 0, \qquad \forall t \in \mathbb{R}.$$

Cauchy-Schwarz Inequality

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{x}^\mathsf{T}\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Proof. *Consider (univariate) quadratic function*

$$h(t) = \frac{1}{2}(t\mathbf{x} + \mathbf{y})^\mathsf{T}(t\mathbf{x} + \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 t^2 + \mathbf{x}^\mathsf{T}\mathbf{y}t + \frac{1}{2}\|\mathbf{y}\|_2^2.$$

$$h(t) \geq 0, \qquad \forall t \in \mathbb{R}.$$

Reminder: quadratic function $ax^2 + bx + c$ *has less than 2 distinct real roots* $\Leftrightarrow \Delta = b^2 - 4ac \leq 0$.

Cauchy-Schwarz Inequality

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{x}^{\mathsf{T}}\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Proof. *Consider (univariate) quadratic function*

$$h(t) = \frac{1}{2}(t\mathbf{x} + \mathbf{y})^{\mathsf{T}}(t\mathbf{x} + \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 t^2 + \mathbf{x}^{\mathsf{T}}\mathbf{y}t + \frac{1}{2}\|\mathbf{y}\|_2^2.$$

$h(t) \geq 0, \qquad \forall t \in \mathbb{R}.$

Reminder: quadratic function $ax^2 + bx + c$ has less than 2 distinct real roots $\Leftrightarrow \Delta = b^2 - 4ac \leq 0$.

$0 \geq \Delta = \left(\mathbf{x}^{\mathsf{T}}\mathbf{y}\right)^2 - \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2.$

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

A differentiable $f : \mathcal{D} \to \mathbb{R}$ has L-Lipschitz continuous gradients if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.$$

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

A differentiable $f : \mathcal{D} \to \mathbb{R}$ has L-Lipschitz continuous gradients if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.$$

Gradient doesn't have unbounded difference between two close points.

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

A differentiable $f : \mathcal{D} \to \mathbb{R}$ has L-Lipschitz continuous gradients if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.$$

Gradient doesn't have unbounded difference between two close points.

Note. *Differentiable $f : \mathcal{D} \to \mathbb{R}$ is m-strongly convex $\implies$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \geq m\|\mathbf{y} - \mathbf{x}\|_2.$$

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

A differentiable $f : \mathcal{D} \to \mathbb{R}$ has L-Lipschitz continuous gradients if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.$$

Gradient doesn't have unbounded difference between two close points.

Note. *Differentiable $f : \mathcal{D} \to \mathbb{R}$ is m-strongly convex $\implies$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \geq m\|\mathbf{y} - \mathbf{x}\|_2.$$

Proof.

Lipschitz Functions

$f : \mathcal{D} \to \mathbb{R}$ is L-Lipschitz continuous if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|y - x\|_2.$$

A differentiable $f : \mathcal{D} \to \mathbb{R}$ has L-Lipschitz continuous gradients if for *every* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2.$$

Gradient doesn't have unbounded difference between two close points.

Note. *Differentiable $f : \mathcal{D} \to \mathbb{R}$ is m-strongly convex $\implies$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \geq m\|\mathbf{y} - \mathbf{x}\|_2.$$

Proof.

Eigenvalues and Eigenvectors

Let $A \in \mathbb{R}^{n \times n}$. A nonzero vector v is an eigenvector with eigenvalue $\lambda \in \mathbb{C}$ if

$$Av = \lambda v.$$

Eigenvalues and Eigenvectors

Let $A \in \mathbb{R}^{n \times n}$. A nonzero vector v is an eigenvector with eigenvalue $\lambda \in \mathbb{C}$ if

$$Av = \lambda v.$$

If A is real symmetric ($A = A^\mathsf{T}$), then

- All eigenvalues are real.
- There exists an orthonormal basis of eigenvectors.
- A admits the eigendecomposition $A = Q \Lambda Q^\mathsf{T}$ with Q orthogonal, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Rayleigh Quotient and Inequalities (Symmetric A)

For $A = A^{\mathsf{T}}$ and any nonzero $x \in \mathbb{R}^n$, the Rayleigh quotient satisfies

$$\lambda_{\min}(A) \;\leq\; \frac{x^{\mathsf{T}} A x}{x^{\mathsf{T}} x} \;\leq\; \lambda_{\max}(A).$$

Rayleigh Quotient and Inequalities (Symmetric A)

For $A = A^{\mathsf{T}}$ and any nonzero $x \in \mathbb{R}^n$, the Rayleigh quotient satisfies

$$\lambda_{\min}(A) \;\le\; \frac{x^{\mathsf{T}}Ax}{x^{\mathsf{T}}x} \;\le\; \lambda_{\max}(A).$$

Equivalent statements:

- $A \succeq 0$ (positive semidefinite, PSD) $\Leftrightarrow x^{\mathsf{T}}Ax \ge 0$ for all x.
- $A \preceq B \Leftrightarrow B - A \succeq 0$ (matrix Loewner order).

Positive Semidefinite (PSD) and Notation $A \succeq B$

- $A \succeq 0$ (PSD) $\Leftrightarrow x^{\mathsf{T}}Ax \geq 0$ for all $x \in \mathbb{R}^n$.
- $A \succ 0$ (PD) $\Leftrightarrow x^{\mathsf{T}}Ax > 0$ for all nonzero x.
- $A \succeq B$ means $A - B \succeq 0$.

Positive Semidefinite (PSD) and Notation $A \succeq B$

- $A \succeq 0$ (PSD) $\Leftrightarrow x^{\mathsf{T}} A x \geq 0$ for all $x \in \mathbb{R}^n$.
- $A \succ 0$ (PD) $\Leftrightarrow x^{\mathsf{T}} A x > 0$ for all nonzero x.
- $A \succeq B$ means $A - B \succeq 0$.

For symmetric A:

$$A \succeq 0 \;\Leftrightarrow\; \lambda_i(A) \geq 0 \;\forall i \;\Leftrightarrow\; A = Q \Lambda Q^{\mathsf{T}},\; \Lambda \succeq 0.$$

Singular Value Decomposition (SVD)

For any $A \in \mathbb{R}^{m \times n}$,

$$A = U \Sigma V^{\mathsf{T}},$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal with nonnegative entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (singular values).

Singular Value Decomposition (SVD)

For any $A \in \mathbb{R}^{m \times n}$,

$$A = U \Sigma V^{\mathsf{T}},$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal with nonnegative entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (singular values). Useful facts:

- $\sigma_i(A) = \sqrt{\lambda_i(A^{\mathsf{T}} A)}$.
- $\|A\|_2 = \sigma_{\max}(A)$, $\|A\|_F^2 = \sum_i \sigma_i^2$.
- Best rank-k approximation (Eckart–Young): $A_k = U_{:,1:k} \Sigma_{1:k,1:k} V_{:,1:k}^{\mathsf{T}}$ minimizes $\|A - X\|_F$ over $\mathrm{rank}(X) \leq k$.

Common Inequalities

- Cauchy–Schwarz: $|x^\mathsf{T} y| \leq \|x\|_2 \, \|y\|_2$.
- For PSD A: $|x^\mathsf{T} A y| \leq \sqrt{x^\mathsf{T} A x} \, \sqrt{y^\mathsf{T} A y}$.
- Submultiplicativity: $\|Ax\|_2 \leq \|A\|_2 \, \|x\|_2$.
- For symmetric A: $\lambda_{\min}(A) \|x\|_2^2 \leq x^\mathsf{T} A x \leq \lambda_{\max}(A) \|x\|_2^2$.

Use Hessian Matrix: Example 1

Prove that $f(x, y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

Use Hessian Matrix: Example 1

Prove that $f(x, y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

$$\nabla_f^2(x, y) = \begin{bmatrix} 2a & 1 \\ 1 & 2a \end{bmatrix}.$$

Use Hessian Matrix: Example 1

Prove that $f(x, y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

$$\nabla_f^2(x, y) = \begin{bmatrix} 2a & 1 \\ 1 & 2a \end{bmatrix}.$$

Let λ be its eigenvalue.

Use Hessian Matrix: Example 1

Prove that $f(x,y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

$$\nabla_f^2(x,y) = \begin{bmatrix} 2a & 1 \\ 1 & 2a \end{bmatrix}.$$

Let λ be its eigenvalue.

$$0 = \begin{vmatrix} 2a - \lambda & 1 \\ 1 & 2a - \lambda \end{vmatrix} = (2a - \lambda)^2 - 1.$$

Use Hessian Matrix: Example 1

Prove that $f(x, y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

$$\nabla_f^2(x, y) = \begin{bmatrix} 2a & 1 \\ 1 & 2a \end{bmatrix}.$$

Let λ be its eigenvalue.

$$0 = \begin{vmatrix} 2a - \lambda & 1 \\ 1 & 2a - \lambda \end{vmatrix} = (2a - \lambda)^2 - 1.$$

$$\lambda = 2a - 1, \; 2a + 1.$$

Use Hessian Matrix: Example 1

Prove that $f(x, y) = xy + a(x^2 + y^2)$ is convex if and only if $a \geq 1/2$.

$$\nabla_f^2(x, y) = \begin{bmatrix} 2a & 1 \\ 1 & 2a \end{bmatrix}.$$

Let λ be its eigenvalue.

$$0 = \begin{vmatrix} 2a - \lambda & 1 \\ 1 & 2a - \lambda \end{vmatrix} = (2a - \lambda)^2 - 1.$$

$$\lambda = 2a - 1,\ 2a + 1.$$

Both λ's are non-negative iff $a \geq 1/2$.

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \qquad\qquad\qquad \right]^{\mathsf{T}}.$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \quad - \frac{2e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \right]^{\mathsf{T}}.$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1-2\beta_2}}{1+e^{\beta_1-2\beta_2}} \quad -\frac{2e^{\beta_1-2\beta_2}}{1+e^{\beta_1-2\beta_2}}\right]^T.$$

$$\nabla_f^2(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\beta_1-2\beta_2}}{\left(1+e^{\beta_1-2\beta_2}\right)^2} & \\ & \end{bmatrix}$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \quad -\frac{2e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}}\right]^T.$$

$$\nabla_f^2(\boldsymbol{\beta}) \;=\; \begin{bmatrix} \frac{e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} & -\frac{2e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} \\ -\frac{2e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} & \end{bmatrix}$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \quad -\frac{2e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}}\right]^{\mathsf{T}}.$$

$$\nabla_f^2(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} & -\frac{2e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} \\ -\frac{2e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} & \frac{4e^{\beta_1 - 2\beta_2}}{\left(1 + e^{\beta_1 - 2\beta_2}\right)^2} \end{bmatrix}$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\beta_1-2\beta_2}}{1+e^{\beta_1-2\beta_2}} & -\frac{2e^{\beta_1-2\beta_2}}{1+e^{\beta_1-2\beta_2}} \end{bmatrix}^{\mathsf{T}}.$$

$$\nabla_f^2(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\beta_1-2\beta_2}}{\left(1+e^{\beta_1-2\beta_2}\right)^2} & -\frac{2e^{\beta_1-2\beta_2}}{\left(1+e^{\beta_1-2\beta_2}\right)^2} \\ -\frac{2e^{\beta_1-2\beta_2}}{\left(1+e^{\beta_1-2\beta_2}\right)^2} & \frac{4e^{\beta_1-2\beta_2}}{\left(1+e^{\beta_1-2\beta_2}\right)^2} \end{bmatrix} \propto_+ \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix}.$$

Use Hessian Matrix: Example 2

Prove that $f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

$$\nabla f(\boldsymbol{\beta}) = \left[\frac{e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}} \quad -\frac{2e^{\beta_1 - 2\beta_2}}{1 + e^{\beta_1 - 2\beta_2}}\right]^{\mathsf{T}}.$$

$$\nabla_f^2(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\beta_1 - 2\beta_2}}{(1 + e^{\beta_1 - 2\beta_2})^2} & -\frac{2e^{\beta_1 - 2\beta_2}}{(1 + e^{\beta_1 - 2\beta_2})^2} \\ -\frac{2e^{\beta_1 - 2\beta_2}}{(1 + e^{\beta_1 - 2\beta_2})^2} & \frac{4e^{\beta_1 - 2\beta_2}}{(1 + e^{\beta_1 - 2\beta_2})^2} \end{bmatrix} \propto_+ \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix}.$$

The eigenvalues are 0 *and* $\dfrac{5e^{\beta_1 - 2\beta_2}}{(1 + e^{\beta_1 - 2\beta_2})^2} \geq 0.$

$\nabla_f^2(\boldsymbol{\beta}) \succeq 0$. f *is convex.*

Use Hessian Matrix: Example 2

$f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

Extension. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\top \boldsymbol{\beta}}\right)$ always convex for any constant vector $\mathbf{c}$?

Use Hessian Matrix: Example 2

$f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

Extension. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ always convex for any constant vector $\mathbf{c}$?

$$\nabla_f^2(\boldsymbol{\beta}) = \frac{e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)^2} \begin{bmatrix} c_1^2 & c_1 c_2 & \cdots & c_1 c_n \\ c_1 c_2 & c_2^2 & \cdots & c_2 c_n \\ \vdots & \vdots & \ddots & \vdots \\ c_1 c_n & c_2 c_n & \cdots & c_n^2 \end{bmatrix} \propto_+ \mathbf{c}\mathbf{c}^\mathsf{T}.$$

Use Hessian Matrix: Example 2

$f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

Extension. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ always convex for any constant vector $\mathbf{c}$?

$$\nabla_f^2(\boldsymbol{\beta}) = \frac{e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)^2} \begin{bmatrix} c_1^2 & c_1 c_2 & \cdots & c_1 c_n \\ c_1 c_2 & c_2^2 & \cdots & c_2 c_n \\ \vdots & \vdots & \ddots & \vdots \\ c_1 c_n & c_2 c_n & \cdots & c_n^2 \end{bmatrix} \propto_+ \mathbf{c}\mathbf{c}^\mathsf{T}.$$

Why is $\nabla_f^2(\boldsymbol{\beta})$ always positive semi-definite?

Use Hessian Matrix: Example 2

$f(\beta_1, \beta_2) = \log\left(1 + e^{\beta_1 - 2\beta_2}\right)$ is convex over $\mathbb{R}^2$.

Extension. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right)$ always convex for any constant vector $\mathbf{c}$?

$$\nabla_f^2(\boldsymbol{\beta}) = \frac{e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right)^2} \begin{bmatrix} c_1^2 & c_1 c_2 & \cdots & c_1 c_n \\ c_1 c_2 & c_2^2 & \cdots & c_2 c_n \\ \vdots & \vdots & \ddots & \vdots \\ c_1 c_n & c_2 c_n & \cdots & c_n^2 \end{bmatrix} \propto_+ \mathbf{c}\mathbf{c}^{\mathsf{T}}.$$

Why is $\nabla_f^2(\boldsymbol{\beta})$ always positive semi-definite?

Because for every $\mathbf{x}, \boldsymbol{\beta} \in \mathbb{R}^n$,

$$\mathbf{x}^{\mathsf{T}} \nabla_f^2(\boldsymbol{\beta}) \mathbf{x} = \frac{e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right)^2} \mathbf{x}^{\mathsf{T}} \mathbf{c}\mathbf{c}^{\mathsf{T}} \mathbf{x} = \frac{e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right)^2} \left(\mathbf{x}^{\mathsf{T}} \mathbf{c}\right)^2 \geq 0.$$

Use Hessian Matrix: Example 2

$f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right)$ is always convex over $\mathbb{R}^n$.

Extension II. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^{\mathsf{T}}\boldsymbol{\beta}}\right) - \mathbf{m}^{\mathsf{T}}\boldsymbol{\beta} + \text{Const}$ always convex for any constant vector $\mathbf{c}, \mathbf{m}$?

Use Hessian Matrix: Example 2

$f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ is always convex over $\mathbb{R}^n$.

Extension II. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right) - \mathbf{m}^\mathsf{T}\boldsymbol{\beta} + \text{Const}$ always convex for any constant vector $\mathbf{c}, \mathbf{m}$?

Application: Logistic Regression. $X_i \in \mathbb{R}^n, y_i \in \{0, 1\}$ *are given data.*

Use Hessian Matrix: Example 2

$f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ is always convex over $\mathbb{R}^n$.

Extension II. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right) - \mathbf{m}^\mathsf{T}\boldsymbol{\beta} + \text{Const}$ always convex for any constant vector $\mathbf{c}, \mathbf{m}$?

Application: Logistic Regression. $X_i \in \mathbb{R}^n, y_i \in \{0, 1\}$ *are given data.*

We want a good $\boldsymbol{\beta}$ *for predicting* y *with* $\hat{y} \in [0, 1]$ *given another* X.

Use Hessian Matrix: Example 2

$f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ is always convex over $\mathbb{R}^n$.

Extension II. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right) - \mathbf{m}^\mathsf{T}\boldsymbol{\beta} + \text{Const}$ always convex for any constant vector $\mathbf{c}, \mathbf{m}$?

Application: Logistic Regression. $X_i \in \mathbb{R}^n, y_i \in \{0, 1\}$ *are given data.*

We want a good $\boldsymbol{\beta}$ *for predicting* y *with* $\hat{y} \in [0, 1]$ *given another* X.

With a good $\boldsymbol{\beta}^*$, *we predict that*

$$\hat{y} = \frac{1}{1 + e^{-X^\mathsf{T}\boldsymbol{\beta}^*}}.$$

Use Hessian Matrix: Example 2

$f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right)$ is always convex over $\mathbb{R}^n$.

Extension II. Is $f(\boldsymbol{\beta}) = \log\left(1 + e^{\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}\right) - \mathbf{m}^\mathsf{T}\boldsymbol{\beta} + \text{Const}$ always convex for any constant vector $\mathbf{c}, \mathbf{m}$?

Application: Logistic Regression. $X_i \in \mathbb{R}^n, y_i \in \{0, 1\}$ *are given data.*

We want a good $\boldsymbol{\beta}$ *for predicting* y *with* $\hat{y} \in [0, 1]$ *given another* X.

With a good $\boldsymbol{\beta}^*$, *we predict that*

$$\hat{y} = \frac{1}{1 + e^{-X^\mathsf{T}\boldsymbol{\beta}^*}}.$$

Maximum Likelihood Estimation (MLE) gives that