Optimization for the information and data sciences

Mahdi Soltanolkotabi

High Dimensional derivatives and Matrix Calculus

Ming Hsieh Department of Electrical Engineering

USC University of Southern California

Derivatives

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $x \in$**int dom** f. The derivative of f at x denoted by $Df(x)$ is an $m \times n$ matrix defined as

$$Df(x)ij = \frac{\partial f_i(x)}{\partial x_j}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n.$$

Derivatives

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $x \in$**int dom** f. The derivative of f at x denoted by $Df(x)$ is an $m \times n$ matrix defined as

$$Df(x)ij = \frac{\partial f_i(x)}{\partial x_j}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n.$$

The affine function of z given by

$$f(x) + Df(x)(z - x)$$

is called the *first-order approximation* of f at (or near) x.

The gradient

When f is real-valued (i.e., $f : \mathbb{R}^n \to \mathbb{R}$) the derivative is a $1 \times n$ row vector. Gradient is column vector

$$\nabla f(x) = Df(x)^T$$

with components the partial derivatives

$$\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1, 2, \ldots, n.$$

The gradient

When f is real-valued (i.e., $f : \mathbb{R}^n \to \mathbb{R}$) the derivative is a $1 \times n$ row vector. Gradient is column vector

$$\nabla f(x) = Df(x)^T$$

with components the partial derivatives

$$\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1, 2, \ldots, n.$$

first order approximation of f at a point x is the affine function of z given by

$$f(x) + \nabla f(x)^T (z - x).$$

Examples of gradients

Quadratic function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) = \frac{1}{2}x^T P x + q^T x + r.$$

Examples of gradients

Quadratic function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) = \frac{1}{2}x^T P x + q^T x + r.$$

Gradient is equal to

$$\nabla f(x) = (P + P^T)x + q.$$

log det function

logdet function $f : \mathcal{S}^n \to \mathbb{R}$, given by

$$f(X) = \log \det X, \quad \textbf{dom} f = \mathcal{S}^n_{++}.$$

log det function

logdet function $f : \mathcal{S}^n \to \mathbb{R}$, given by

$$f(X) = \log \det X, \quad \mathbf{dom} f = \mathcal{S}^n_{++}.$$

Note that

$$\begin{aligned} f(Z) = \log \det Z &= \log \det (X + Z - X) \\ &= \log \det \left(X^{1/2} \left(I + X^{-1/2} (Z - X) X^{-1/2} \right) X^{1/2} \right) \\ &= \log \det X + \log \det \left(\left(I + X^{-1/2} (Z - X) X^{-1/2} \right) \right) \\ &= \log \det X + \sum_{i=1}^{n} \log(1 + \lambda_i) \\ &\approx \log \det X + \sum_{i=1}^{n} \lambda_i \\ &= \log \det X + \mathbf{tr} \left(X^{-1/2} (Z - X) X^{-1/2} \right) \\ &= \log \det X + \mathbf{tr} \left(X^{-1} (Z - X) \right) \\ &= f(X) + \mathbf{tr} \left(X^{-1} (Z - X) \right). \end{aligned}$$

Thus $\nabla f(X) = X^{-1}$.

Chain rule

Suppose

- $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \textbf{int dom } f$ and
- $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(x) \in \textbf{int dom } g$
- composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(z) = g(f(z))$

$$Dh(x) = Dg(f(x))Df(x)$$

Chain rule

Suppose

- $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in$ **int dom** f and
- $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(x) \in$ **int dom** g
- composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(z) = g(f(z))$

$$Dh(x) = Dg(f(x))Df(x)$$

- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $h(x) = g(f(x))$

$$\nabla h(x) = g'(f(x))\nabla f(x).$$

Composition with affine function

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $A \in \mathbb{R}^{n \times p}$, and $b \in \mathbb{R}^n$.
- Define $g : \mathbb{R}^p \to \mathbb{R}^m$ as $g(x) = f(Ax + b)$, with **dom** g=$\{x | Ax + b \in$ **dom** $f\}$

Composition with affine function

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $A \in \mathbb{R}^{n \times p}$, and $b \in \mathbb{R}^n$.
- Define $g : \mathbb{R}^p \to \mathbb{R}^m$ as $g(x) = f(Ax + b)$, with **dom** g=$\{x | Ax + b \in$ **dom** $f\}$
- **Q:** What is $\nabla g(x)$?

Composition with affine function

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $A \in \mathbb{R}^{n \times p}$, and $b \in \mathbb{R}^n$.
- Define $g : \mathbb{R}^p \to \mathbb{R}^m$ as $g(x) = f(Ax + b)$, with **dom** g=$\{x | Ax + b \in$ **dom** $f\}$
- **Q:** What is $\nabla g(x)$?
- **Answer:**
  chain rule $\Rightarrow \quad Dg(x) = Df(Ax + b)A \quad \Rightarrow \quad \nabla g(x) = A^T \nabla f(Ax + b)$

Example

Consider $f : \mathbb{R}^n \to \mathbb{R}$, with **dom** $f = \mathbb{R}^n$ and

$$f(x) = \log \sum_{i=1}^{m} \mathsf{exp}\left(a_i^T x + b_i\right)$$

Example

Consider $f : \mathbb{R}^n \to \mathbb{R}$, with **dom** $f = \mathbb{R}^n$ and

$$f(x) = \log \sum_{i=1}^{m} \exp\left(a_i^T x + b_i\right)$$

Note that $f(x) = g(Ax + b)$

- with $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(y) = \log\left(\sum_{i=1}^{m} e^{y_i}\right)$
- $A \in \mathbb{R}^{m \times n}$ with rows $a_1^T, \ldots, a_m^T$

Example

Consider $f : \mathbb{R}^n \to \mathbb{R}$, with **dom** $f = \mathbb{R}^n$ and

$$f(x) = \log \sum_{i=1}^{m} \mathsf{exp}\left(a_i^T x + b_i\right)$$

Note that $f(x) = g(Ax + b)$

- with $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(y) = \log\left(\sum_{i=1}^m e^{y_i}\right)$
- $A \in \mathbb{R}^{m \times n}$ with rows $a_1^T, \ldots, a_m^T$

$$\nabla g(y) = \frac{1}{\sum_{i=1}^m e^{y_i}} \begin{bmatrix} e^{y_1} \\ \vdots \\ e^{y_m} \end{bmatrix}$$

Example

Consider $f : \mathbb{R}^n \to \mathbb{R}$, with **dom** $f = \mathbb{R}^n$ and

$$f(x) = \log \sum_{i=1}^{m} \exp\left(a_i^T x + b_i\right)$$

Note that $f(x) = g(Ax + b)$

- with $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(y) = \log\left(\sum_{i=1}^m e^{y_i}\right)$
- $A \in \mathbb{R}^{m\times n}$ with rows $a_1^T, \ldots, a_m^T$

$$\nabla g(y) = \frac{1}{\sum_{i=1}^m e^{y_i}} \begin{bmatrix} e^{y_1} \\ \vdots \\ e^{y_m} \end{bmatrix}$$

so by composition formula we have

$$\nabla f(x) = \frac{1}{\mathbf{1}^T \boldsymbol{z}} A^T z \quad \text{where} \quad z_i = e^{a_i^T x + b_i}, i = 1, 2, \ldots, m.$$

Hessians

Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ denoted by $\nabla^2 f(x)$ is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i = 1, 2, \dots, n, \quad j = 1, 2, \dots n,$$

Hessians

Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ denoted by $\nabla^2 f(x)$ is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots n,$$

The quadratic function of z given by

$$f(x) + \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$$

is called the *second-order approximation* of f at (or near) x.

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- $\nabla^2 f(x) = D\nabla f(x)$

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- $\nabla^2 f(x) = D\nabla f(x)$

Examples

- $f : \mathbb{R}^n \to \mathbb{R}$, with $f(x) = \frac{1}{2}x^T Px + q^T x + r$

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- $\nabla^2 f(x) = D\nabla f(x)$

Examples

- $f : \mathbb{R}^n \to \mathbb{R}$, with $f(x) = \frac{1}{2}x^T P x + q^T x + r$
  - $\nabla f(x) = \frac{1}{2}(P + P^T)x + q$

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- $\nabla^2 f(x) = D\nabla f(x)$

Examples

- $f : \mathbb{R}^n \to \mathbb{R}$, with $f(x) = \frac{1}{2}x^T Px + q^T x + r$
  - $\nabla f(x) = \frac{1}{2}(P + P^T)x + q$
  - $\nabla^2 f(x) = \frac{1}{2}(P + P^T)$

Hessian as derivative of derivative

$f : \mathbb{R}^n \to \mathbb{R}$ then

- gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- $\nabla^2 f(x) = D\nabla f(x)$

Examples

- $f : \mathbb{R}^n \to \mathbb{R}$, with $f(x) = \frac{1}{2}x^T P x + q^T x + r$
  - $\nabla f(x) = \frac{1}{2}(P + P^T)x + q$
  - $\nabla^2 f(x) = \frac{1}{2}(P + P^T)$
- $f(X) = \log \det(X)$ then $\nabla f(X) = X^{-1}$

$$\begin{aligned}
Z^{-1} &= (X + Z - X)^{-1} \\
&= \left(X^{1/2}\left(I + X^{-1/2}(Z - X)X^{-1/2}\right)X^{1/2}\right)^{-1} \\
&= X^{-1/2}\left(I + X^{-1/2}(Z - X)X^{-1/2}\right)^{-1}X^{-1/2} \\
&\approx X^{-1/2}\left(I - X^{-1/2}(Z - X)X^{-1/2}\right)^{-1}X^{-1/2} \\
&= X^{-1} - X^{-1}(Z - X)X^{-1}
\end{aligned}$$

Thus

$$\nabla^2 f(X)[U, V] = -\mathbf{tr}(X^{-1}UX^{-1}V)$$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

- **composition with affine function**
  Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A$$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

- **composition with affine function**
  Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A$$

Example: $f(x) = \log \sum_{i=1}^m \mathrm{exp}(a_i^T x + b_i)$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

- **composition with affine function**
  Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A$$

Example: $f(x) = \log \sum_{i=1}^m \exp(a_i^T x + b_i)$
can think of it as $f(x) = g(Ax + b)$ with $g(y) = \log\left(\sum_{i=1}^m e^{y_i}\right)$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

- **composition with affine function**
  Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A$$

Example: $f(x) = \log \sum_{i=1}^m \text{exp}(a_i^T x + b_i)$
can think of it as $f(x) = g(Ax + b)$ with $g(y) = \log \left(\sum_{i=1}^m e^{y_i}\right)$

- Note that $\nabla^2 g(y) = \textbf{diag}\left(\nabla g(y)\right) - \nabla g(y)\nabla g(y)^T$

Chain rule for second derivative

Suppose $f : \mathbb{R}^n \to R$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

- **composition with scalar function**
  $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

- **composition with affine function**
  Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A$$

Example: $f(x) = \log \sum_{i=1}^m \exp(a_i^T x + b_i)$
can think of it as $f(x) = g(Ax + b)$ with $g(y) = \log \left(\sum_{i=1}^m e^{y_i}\right)$

- Note that $\nabla^2 g(y) = \textbf{diag}\left(\nabla g(y)\right) - \nabla g(y)\nabla g(y)^T$
- by composition $\nabla^2 f(x) = A^T \left(\frac{1}{(\mathbf{1}^T \boldsymbol{x})}\textbf{diag}(z) - \frac{1}{(\mathbf{1}^T \boldsymbol{x})^2} zz^T\right) A$
  where $z_i = \exp(a_i^T x + b_i)$, $i = 1, 2, \ldots, m$.

Matrix calculus cheat sheet-simple stuff

$$
\begin{aligned}
\partial \mathbf{A} &= 0 && (\mathbf{A} \text{ is a constant}) \\
\partial(\alpha \mathbf{X}) &= \alpha \partial \mathbf{X} \\
\partial(\mathbf{X}+\mathbf{Y}) &= \partial \mathbf{X} + \partial \mathbf{Y} \\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial \mathbf{X}) \\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial \mathbf{X})\mathbf{Y} + \mathbf{X}(\partial \mathbf{Y}) \\
\partial(\mathbf{X} \circ \mathbf{Y}) &= (\partial \mathbf{X}) \circ \mathbf{Y} + \mathbf{X} \circ (\partial \mathbf{Y}) \\
\partial(\mathbf{X} \otimes \mathbf{Y}) &= (\partial \mathbf{X}) \otimes \mathbf{Y} + \mathbf{X} \otimes (\partial \mathbf{Y}) \\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1} \\
\partial(\det(\mathbf{X})) &= \mathrm{Tr}(\mathrm{adj}(\mathbf{X})\partial \mathbf{X}) \\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial \mathbf{X}^T &= (\partial \mathbf{X})^T \\
\partial \mathbf{X}^H &= (\partial \mathbf{X})^H
\end{aligned}
$$

Matrix calculus cheat sheet-Derivatives of determinant

$$
\begin{aligned}
\frac{\partial \det(\mathbf{Y})}{\partial x} &= \det(\mathbf{Y})\operatorname{Tr}\left[\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right] \\
\sum_k \frac{\partial \det(\mathbf{X})}{\partial X_{ik}} X_{jk} &= \delta_{ij}\det(\mathbf{X}) \\
\frac{\partial^2 \det(\mathbf{Y})}{\partial x^2} &= \det(\mathbf{Y})\left[\operatorname{Tr}\left[\mathbf{Y}^{-1}\frac{\partial \frac{\partial \mathbf{Y}}{\partial x}}{\partial x}\right]\right. \\
&\quad + \operatorname{Tr}\left[\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right]\operatorname{Tr}\left[\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right] \\
&\quad \left. - \operatorname{Tr}\left[\left(\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right)\left(\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right)\right]\right]
\end{aligned}
$$

Matrix calculus cheat sheet-grad and hessians

$$
\begin{aligned}
f &= \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} \\
\nabla_{\mathbf{x}} f = \frac{\partial f}{\partial \mathbf{x}} &= (\mathbf{A} + \mathbf{A}^T)\mathbf{x} + \mathbf{b} \\
\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} &= \mathbf{A} + \mathbf{A}^T
\end{aligned}
$$

Matrix calculus cheat sheet-first order

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}) &= \mathbf{I} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) &= \mathbf{A}^T \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A}\mathbf{X}\mathbf{B}) &= \mathbf{A}^T\mathbf{B}^T \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A}\mathbf{X}^T\mathbf{B}) &= \mathbf{B}\mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}^T\mathbf{A}) &= \mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A}\mathbf{X}^T) &= \mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A} \otimes \mathbf{X}) &= \mathrm{Tr}(\mathbf{A})\mathbf{I}
\end{aligned}
$$