Appendix A: Mathematical Background

Norms: vector

- Inner product

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i \qquad \text{for } x, y \in \mathbf{R}^n$$

- Euclidean norm or l_2-norm

$$\|x\|_2 = (x^T x)^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}$$

- Angle

$$\angle(x, y) = \cos^{-1}\left(\frac{x^T y}{\|x\|_2 \|y\|_2}\right)$$

Norms: matrix

- Inner product

$$\langle X, Y \rangle = \mathbf{tr}(X^T Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij} \quad \text{for } X, Y \in \mathbf{R}^{m \times n}$$

- Frobenius norm

$$\|X\|_F = \left(\mathbf{tr}(X^T X)\right)^{1/2} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 \right)^{1/2}$$

Examples

- Sum-absolute-value, or ℓ_1-norm

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$

- Chebyshev or $\ell\infty$-norm

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

- ℓ_p-norm

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p} \quad \text{with } p \geq 1$$

- ℓ_1-norm

$$\lim_{p \to \infty} \|x\|_p = \max\{|x_1|, \ldots, |x_n|\}$$

Examples

- Sum-absolute-value norm

$$\|X\|_{\text{sav}} = \sum_{i=1}^{m}\sum_{j=1}^{n} |X_{ij}|$$

- Maximum-absolute-value norm

$$\|X\|_{\text{mav}} = \max\{|X_{ij}| \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}$$

Analysis: Interior point

- Interior point

An element $x \in C \subseteq \mathbf{R}^n$ is called an *interior* point of C if there exists an $\epsilon > 0$ for which

$$\{y \mid \|y - x\|_2 \leq \epsilon\} \subseteq C,$$

i.e., there exists a ball centered at x that lies entirely in C.

  - The set of all points interior to *C* is called the interior of *C* and is denoted **int***C*
  - *Open:* A set C is *open* if $\textbf{int}\, C = C$, *i.e.*, every point in C is an interior point.
  - *Closed:*

A set $C \subseteq \mathbf{R}^n$ is *closed* if its complement $\mathbf{R}^n \setminus C = \{x \in \mathbf{R}^n \mid x \notin C\}$ is open.

Analysis: Closure and Boundary

- Closure of a set C: $\mathbf{cl}\, C = \mathbf{R}^n \setminus \mathbf{int}(\mathbf{R}^n \setminus C)$

  i.e., the complement of the interior of the complement of C. A point x is in the closure of C if for every $\epsilon > 0$, there is a $y \in C$ with $\|x - y\|_2 \leq \epsilon$.

- Boundary of the set C: $\mathbf{bd}\, C = \mathbf{cl}\, C \setminus \mathbf{int}\, C$

A *boundary point* x (*i.e.*, a point $x \in \mathbf{bd}\, C$) satisfies the following property: For all $\epsilon > 0$, there exists $y \in C$ and $z \notin C$ with

$$\|y - x\|_2 \leq \epsilon, \qquad \|z - x\|_2 \leq \epsilon,$$

i.e., there exist arbitrarily close points in C, and also arbitrarily close points not in C.

Supremum and infimum

- Supremum $\sup C$

Suppose $C \subseteq \mathbf{R}$. A number a is an *upper bound* on C if for each $x \in C$, $x \leq a$. The set of upper bounds on a set C is either empty (in which case we say C is unbounded above), all of $\mathbf{R}$ (only when $C = \emptyset$), or a closed infinite interval $[b, \infty)$.

The number b is called the *least upper bound* or *supremum* of the set C

- Infimum $\inf C = -\sup(-C)$

A number a is a lower bound on $C \subseteq \mathbf{R}$ if for each $x \in C$, $a \leq x$.

Functions

- Notation $f : A \to B$

we mean that f is a function on the set $\mathbf{dom}\, f \subseteq A$ into the set B; in particular we can have $\mathbf{dom}\, f$ a proper subset of the set A.

  - Example

    $$f : \mathbf{R}^n \to \mathbf{R}^m$$

    means that f maps (some) n-vectors into m-vectors; it does not mean that $f(x)$ is defined for every $x \in \mathbf{R}^n$.

  - $f : \mathbf{S}^n \to \mathbf{R}$

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

Derivatives

- Definition

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ and $x \in \mathbf{int\,dom}\, f$. The function f is differentiable at x if there exists a matrix $Df(x) \in \mathbf{R}^{m\times n}$ that satisfies

$$\lim_{z\in\mathbf{dom}\, f,\; z\neq x,\; z\to x} \frac{\|f(z) - f(x) - Df(x)(z-x)\|_2}{\|z-x\|_2} = 0, \qquad \text{(A.4)}$$

in which case we refer to $Df(x)$ as the *derivative* (or *Jacobian*) of f at x. (There can be at most one matrix that satisfies (A.4).) The function f is *differentiable* if $\mathbf{dom}\, f$ is open, and it is differentiable at every point in its domain.

- Partial derivatives

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

Gradient

- Definition

When f is real-valued (*i.e.*, $f : \mathbf{R}^n \to \mathbf{R}$) the derivative $Df(x)$ is a $1 \times n$ matrix, *i.e.*, it is a *row* vector. Its transpose is called the *gradient* of the function:

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

which is a (column) vector, *i.e.*, in $\mathbf{R}^n$. Its components are the partial derivatives of f:

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

The first-order approximation of f at a point $x \in \mathbf{int\,dom}\, f$ can be expressed as (the affine function of z)

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

Example

As a simple example consider the quadratic function $f : \mathbf{R}^n \to \mathbf{R}$,

$$f(x) = (1/2)x^T Px + q^T x + r,$$

where $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$. Its derivative at x is the row vector $Df(x) = x^T P + q^T$, and its gradient is

$$\nabla f(x) = Px + q.$$

Example

$$f(X) = \log\det X, \qquad \mathbf{dom}\, f = \mathbf{S}^n_{++}. \qquad\qquad \nabla f(X) = X^{-1}$$

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

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

Chain rule

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $x \in \textbf{int dom} f$ and $g : \mathbf{R}^m \to \mathbf{R}^p$ is differentiable at $f(x) \in \textbf{int dom} g$. Define the composition $h : \mathbf{R}^n \to \mathbf{R}^p$ by $h(z) = g(f(z))$. Then h is differentiable at x, with derivative

$$Dh(x) = Dg(f(x))Df(x). \tag{A.5}$$

As an example, suppose $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R} \to \mathbf{R}$, and $h(x) = g(f(x))$. Taking the transpose of $Dh(x) = Dg(f(x))Df(x)$ yields

$$\nabla h(x) = g'(f(x))\nabla f(x). \tag{A.6}$$

Composition with affine function

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable, $A \in \mathbf{R}^{n\times p}$, and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^p \to \mathbf{R}^m$ as $g(x) = f(Ax + b)$, with $\mathbf{dom}\, g = \{x \mid Ax + b \in \mathbf{dom}\, f\}$. The derivative of g is, by the chain rule (A.5), $Dg(x) = Df(Ax + b)A$.

When f is real-valued (*i.e.*, $m = 1$), we obtain the formula for the gradient of a composition of a function with an affine function,

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

For example, suppose that $f : \mathbf{R}^n \to \mathbf{R}$, $x, v \in \mathbf{R}^n$, and we define the function $\tilde{f} : \mathbf{R} \to \mathbf{R}$ by $\tilde{f}(t) = f(x + tv)$. (Roughly speaking, $\tilde{f}$ is f, restricted to the line $\{x + tv \mid t \in \mathbf{R}\}$.) Then we have

$$D\tilde{f}(t) = \tilde{f}'(t) = \nabla f(x + tv)^T v.$$

Example

Example A.2 Consider the function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^n$ and

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

where $a_1, \ldots, a_m \in \mathbf{R}^n$, and $b_1, \ldots, b_m \in \mathbf{R}$. We can find a simple expression for its gradient by noting that it is the composition of the affine function $Ax + b$, where $A \in \mathbf{R}^{m \times n}$ with rows $a_1^T, \ldots, a_m^T$, and the function $g : \mathbf{R}^m \to \mathbf{R}$ given by $g(y) = \log(\sum_{i=1}^m \exp y_i)$. Simple differentiation (or the formula (A.6)) shows that

$$\nabla g(y) = \frac{1}{\sum_{i=1}^m \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix}, \qquad \text{(A.7)}$$

so by the composition formula we have

$$\nabla f(x) = \frac{1}{\mathbf{1}^T z} A^T z$$

where $z_i = \exp(a_i^T x + b_i)$, $i = 1, \ldots, m$.

Second derivative

In this section we review the second derivative of a real-valued function $f : \mathbf{R}^n \to \mathbf{R}$. The second derivative or *Hessian matrix* of f at $x \in \mathbf{int\,dom}\, f$, denoted $\nabla^2 f(x)$, is given by

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

provided f is twice differentiable at x, where the partial derivatives are evaluated at x. The *second-order approximation* of f, at or near x, is the quadratic function of z defined by

$$\widehat{f}(z) = f(x) + \nabla f(x)^T (z - x) + (1/2)(z - x)^T \nabla^2 f(x)(z - x).$$

Example

$$f(x) = (1/2)x^T Px + q^T x + r,$$

where $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$. Its gradient is $\nabla f(x) = Px + q$, so its Hessian is given by $\nabla^2 f(x) = P$. The second-order approximation of a quadratic function is itself.

Example

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

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

For $Z \in \mathbf{S}^n_{++}$ near $X \in \mathbf{S}^n_{++}$, and $\Delta X = Z - X$

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

from the first-order approximation of the gradient above, the quadratic form can be expressed as $-\mathbf{tr}(X^{-1}UX^{-1}V)$ $(\log x)'' = -1/x^2$

$$
\begin{aligned}
f(Z) &= f(X + \Delta X) \\
&\approx f(X) + \mathbf{tr}(X^{-1}\Delta X) - (1/2)\,\mathbf{tr}(X^{-1}\Delta X X^{-1}\Delta X) \\
&\approx f(X) + \mathbf{tr}\left(X^{-1}(Z - X)\right) - (1/2)\,\mathbf{tr}\left(X^{-1}(Z - X)X^{-1}(Z - X)\right)
\end{aligned}
$$