2 Optimisation

2.1 Gradient of vector-valued functions

For a function J that maps a column vector $\mathbf{w} \in \mathbb{R}^n$ to $\mathbb{R}$, the gradient is defined as

$$\nabla J(\mathbf{w}) = \begin{pmatrix} \frac{\partial J(\mathbf{w})}{\partial w_1} \\ \vdots \\ \frac{\partial J(\mathbf{w})}{\partial w_n} \end{pmatrix},$$

where $\partial J(\mathbf{w})/\partial w_i$ are the partial derivatives of $J(\mathbf{w})$ with respect to the i-th element of the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$ (in the standard basis). Alternatively, it is defined to be the column vector $\nabla J(\mathbf{w})$ such that

$$J(\mathbf{w} + \epsilon \mathbf{h}) = J(\mathbf{w}) + \epsilon \left(\nabla J(\mathbf{w})\right)^\top \mathbf{h} + O(\epsilon^2) \tag{2.1}$$

for an arbitrary perturbation $\epsilon \mathbf{h}$. This phrases the derivative in terms of a first-order, or affine, approximation to the perturbed function $J(\mathbf{w} + \epsilon \mathbf{h})$. The derivative ∇J is a linear transformation that maps $\mathbf{h} \in \mathbb{R}^n$ to $\mathbb{R}$ [see Chapter 9, for a formal treatment of derivatives][1].

Use either definition to determine $\nabla J(\mathbf{w})$ for the following functions where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function.

i. $J(\mathbf{w}) = \mathbf{a}^\top \mathbf{w}$.

ii. $J(\mathbf{w}) = \mathbf{w}^\top \mathbf{A} \mathbf{w}$.

iii. $J(\mathbf{w}) = \mathbf{w}^\top \mathbf{w}$.

iv. $J(\mathbf{w}) = ||\mathbf{w}||_2$.

v. $J(\mathbf{w}) = f(||\mathbf{w}||_2)$.

[1] Walter Rudin. Principles of Mathematical Analysis. McGraw Hill, 3rd edition edition, 1976.

2.2 Newton's method

Assume that in the neighbourhood of $\mathbf{w}_0$, a function $J(\mathbf{w})$ can be described by the quadratic approximation

$$f(\mathbf{w}) = c + \mathbf{g}^\top(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}_0),$$

where $c = J(\mathbf{w}_0)$, $\mathbf{g}$ is the gradient of J with respect to $\mathbf{w}$, and $\mathbf{H}$ a symmetric positive definite matrix (e.g. the Hessian matrix for $J(\mathbf{w})$ at $\mathbf{w}_0$ if positive definite).

i. Use Task 2.1 to determine $\nabla f(\mathbf{w})$.

ii. A necessary condition for $\mathbf{w}$ being optimal (leading either to a maximum, minimum or a saddle point) is $\nabla f(\mathbf{w}) = 0$. Determine $\mathbf{w}^*$ such that $\nabla f(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}^*} = 0$. Provide arguments why $\mathbf{w}^*$ is a minimiser of $f(\mathbf{w})$.

iii. In terms of Newton's method to minimise $J(\mathbf{w})$, what do $\mathbf{w}_0$ and $\mathbf{w}^*$ stand for?

2.3 Gradient of matrix-valued functions

For functions J that map a matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ to $\mathbb{R}$, the gradient is defined as

$$\nabla J(\mathbf{W}) = \begin{pmatrix} \frac{\partial J(\mathbf{W})}{\partial W_{11}} & \cdots & \frac{\partial J(\mathbf{W})}{\partial W_{1m}} \\ \vdots & \vdots & \vdots \\ \frac{\partial J(\mathbf{W})}{\partial W_{n1}} & \cdots & \frac{\partial J(\mathbf{W})}{\partial W_{nm}} \end{pmatrix}.$$

Alternatively, it is defined to be the matrix ∇J such that

$$J(\mathbf{W} + \epsilon\mathbf{H}) = J(\mathbf{W}) + \epsilon \operatorname{tr}(\nabla J^\top \mathbf{H}) + O(\epsilon^2) \tag{2.2}$$
$$= J(\mathbf{W}) + \epsilon \operatorname{tr}(\nabla J \mathbf{H}^\top) + O(\epsilon^2) \tag{2.3}$$

This definition is analogue to the one for vector-valued functions in (2.1). It phrases the derivative in terms of a linear approximation to the perturbed objective $J(\mathbf{W} + \epsilon\mathbf{H})$ and, more formally, $\operatorname{tr} \nabla J^\top$ is a linear transformation that maps $\mathbf{H} \in \mathbb{R}^{n \times m}$ to $\mathbb{R}$.

Let $\mathbf{e}^{(i)}$ be *column* vector which is everywhere zero but in slot i where it is 1. Moreover let $\mathbf{e}^{[j]}$ be a *row* vector which is everywhere zero but in slot j where it is 1. The outer product $\mathbf{e}^{(i)}\mathbf{e}^{[j]}$ is then a matrix that is everywhere zero but in row i and column j where it is one. For $\mathbf{H} = \mathbf{e}^{(i)}\mathbf{e}^{[j]}$, we obtain

$$\begin{aligned} J(\mathbf{W} + \epsilon\mathbf{e}^{(i)}\mathbf{e}^{[j]}) &= J(\mathbf{W}) + \epsilon \operatorname{tr}((\nabla J)^\top \mathbf{e}^{(i)}\mathbf{e}^{[j]}) + O(\epsilon^2) \\ &= J(\mathbf{W}) + \epsilon\mathbf{e}^{[j]}(\nabla J)^\top \mathbf{e}^{(i)} + O(\epsilon^2) \\ &= J(\mathbf{W}) + \epsilon\mathbf{e}^{[i]}\nabla J\mathbf{e}^{(j)} + O(\epsilon^2) \end{aligned}$$

Note that $\mathbf{e}^{[i]}\nabla J\mathbf{e}^{(j)}$ picks the element of the matrix ∇J that is in row i and column j, i.e. $\mathbf{e}^{[i]}\nabla J\mathbf{e}^{(j)} = \partial J/\partial W_{ij}$.

Use either of the two definitions to find $\nabla J(\mathbf{W})$ for the functions below, where $\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m, \mathbf{A} \in \mathbb{R}^{n \times m}$, and $f : \mathbb{R} \to \mathbb{R}$ is differentiable.

i. $J(\mathbf{W}) = \mathbf{u}^\top \mathbf{W}\mathbf{v}$.

ii. $J(\mathbf{W}) = \mathbf{u}^\top (\mathbf{W} + \mathbf{A})\mathbf{v}$.

iii. $J(\mathbf{W}) = \sum_n f(\mathbf{w}_n^\top \mathbf{v})$, where $\mathbf{w}_n^\top$ are the rows of the matrix $\mathbf{W}$.

iv. $J(\mathbf{W}) = \mathbf{u}^\top \mathbf{W}^{-1}\mathbf{v}$.
[*Hint:* $(\mathbf{W} + \epsilon\mathbf{H})^{-1} = \mathbf{W}^{-1} - \epsilon\mathbf{W}^{-1}\mathbf{H}\mathbf{W}^{-1} + O(\epsilon^2)$]

2.4 Gradient of the log-determinant

The goal of this exercise is to determine the gradient of

$$J(\mathbf{W}) = \log|\det(\mathbf{W})|.$$

i. Show that the n-th eigenvalue λ_n can be written as

$$\lambda_n = \mathbf{v}_n^\top \mathbf{W}\mathbf{u}_n,$$

where $\mathbf{u}_n$ is the nth eigenvector and $\mathbf{v}_n$ the nth column vector of $\mathbf{U}^{-1}$, with $\mathbf{U}$ being the matrix with the eigenvectors $\mathbf{u}_n$ as columns.

ii. Calculate the gradient of λ_n with respect to $\mathbf{W}$, i.e. $\nabla\lambda_n(\mathbf{W})$.

iii. Write $J(\mathbf{W})$ in terms of the eigenvalues λ_n and calculate $\nabla J(\mathbf{W})$.

iv. Show that

$$\nabla J(\mathbf{W}) = (\mathbf{W}^{-1})^\top.$$

2.5 Descent directions for matrix-valued functions

Assume we would like to minimise a matrix-valued function $J(\mathbf{W})$ by gradient descent, i.e. the update equation is

$$\mathbf{W} \leftarrow \mathbf{W} - \epsilon \nabla J(\mathbf{W}),$$

where ϵ is the step-length. The gradient $\nabla J(\mathbf{W})$ was defined in Task 2.3. It was there pointed out that the gradient defines a first order approximation to the perturbed objective function $J(\mathbf{W} + \epsilon \mathbf{H})$. With (2.2),

$$J(\mathbf{W} - \epsilon \nabla J(\mathbf{W})) = J(\mathbf{W}) - \epsilon \operatorname{tr}(\nabla J(\mathbf{W})^\top \nabla J(\mathbf{W})) + O(\epsilon^2)$$

For any (nonzero) matrix $\mathbf{M}$, it holds that

$$\begin{aligned}
\operatorname{tr}(\mathbf{M}^\top \mathbf{M}) &= \sum_i (\mathbf{M}^\top \mathbf{M})_{ii} \\
&= \sum_i \sum_j (\mathbf{M}^\top)_{ij} (\mathbf{M})_{ji} \\
&= \sum_i \sum_j M_{ji} M_{ji} \\
&= \sum_{ij} (M_{ji})^2 \\
&> 0,
\end{aligned}$$

which means that $\operatorname{tr}(\nabla J(\mathbf{W})^\top \nabla J(\mathbf{W})) > 0$ if the gradient is nonzero.

Hence,

$$J(\mathbf{W} - \epsilon \nabla J(\mathbf{W})) < J(\mathbf{W})$$

for small enough ϵ. Consequently, $\nabla J(\mathbf{W})$ is a descent direction.

Show that $\mathbf{A}^\top \mathbf{A} \nabla J(\mathbf{W}) \mathbf{B} \mathbf{B}^\top$ for non-zero matrices $\mathbf{A}$ and $\mathbf{B}$ is also a descent direction or leaves the leaves the objective invariant.