2 Optimisation

2.1 Gradient of vector-valued functions

i. First method:

$$J(\mathbf{w}) = \mathbf{a}^\top \mathbf{w} = \sum_{k=1}^{n} a_k w_k \qquad \Longrightarrow \qquad \frac{\partial J(\mathbf{w})}{\partial w_i} = a_i$$

Hence

$$\nabla J(\mathbf{w}) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \mathbf{a}.$$

Second method:

$$J(\mathbf{w} + \epsilon \mathbf{h}) = \mathbf{a}^\top (\mathbf{w} + \epsilon \mathbf{h}) = \underbrace{\mathbf{a}^\top \mathbf{w}}_{J(\mathbf{w})} + \epsilon \underbrace{\mathbf{a}^\top \mathbf{h}}_{\nabla J^\top \mathbf{h}}$$

Hence we find again $\nabla J(\mathbf{w}) = \mathbf{a}$.

ii. First method: We start with

$$J(\mathbf{w}) = \mathbf{w}^\top \mathbf{A} \mathbf{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i A_{ij} w_j$$

Hence,

$$\begin{aligned} \frac{\partial J(\mathbf{w})}{\partial w_k} &= \sum_{j=1}^{n} A_{kj} w_j + \sum_{i=1}^{n} w_i A_{ik} \\ &= \sum_{j=1}^{n} A_{kj} w_j + \sum_{i=1}^{n} w_i (\mathbf{A}^\top)_{ki} \\ &= \sum_{j=1}^{m} \left(A_{kj} + (\mathbf{A}^\top)_{kj} \right) w_j \end{aligned}$$

where we have used that the entry in row i and column k of the matrix $\mathbf{A}$ equals the entry in row k and column i of its transpose $\mathbf{A}^\top$. It follows that

$$\begin{aligned} \nabla J(\mathbf{w}) &= \begin{pmatrix} \sum_{j=1}^{n} \left(A_{1j} + (\mathbf{A}^\top)_{1j} \right) w_j \\ \vdots \\ \sum_{j=1}^{n} \left(A_{nj} + (\mathbf{A}^\top)_{nj} \right) w_j \end{pmatrix} \\ &= (\mathbf{A} + \mathbf{A}^\top) \mathbf{w}, \end{aligned}$$

where we have used that sums like $\sum_j B_{ij} w_j$ are equal to the i-th element of the matrix-vector product $\mathbf{B}\mathbf{w}$.

Second method:

$$\begin{aligned} J(\mathbf{w}+\epsilon\mathbf{h}) &= (\mathbf{w}+\epsilon\mathbf{h})^\top\mathbf{A}(\mathbf{w}+\epsilon\mathbf{h}) \\ &= \mathbf{w}^\top\mathbf{A}\mathbf{w} + \mathbf{w}^\top\mathbf{A}(\epsilon\mathbf{h}) + \epsilon\mathbf{h}^\top\mathbf{A}\mathbf{w} + \underbrace{\epsilon\mathbf{h}^\top\mathbf{A}\epsilon\mathbf{h}}_{O(\epsilon^2)} \\ &= \mathbf{w}^\top\mathbf{A}\mathbf{w} + \epsilon(\mathbf{w}^\top\mathbf{A}\mathbf{h} + \mathbf{w}^\top\mathbf{A}^\top\mathbf{h}) + O(\epsilon^2) \\ &= \underbrace{\mathbf{w}^\top\mathbf{A}\mathbf{w}}_{J(\mathbf{w})} + \epsilon(\underbrace{\mathbf{w}^\top\mathbf{A} + \mathbf{w}^\top\mathbf{A}^\top}_{\nabla J(\mathbf{w})^\top})\mathbf{h} + O(\epsilon^2) \end{aligned}$$

where we have used that $\mathbf{h}^\top\mathbf{A}\mathbf{w}$ is a scalar so that $\mathbf{h}^\top\mathbf{A}\mathbf{w} = (\mathbf{h}^\top\mathbf{A}\mathbf{w})^\top = \mathbf{w}^\top\mathbf{A}^\top\mathbf{h}$. Hence

$$\nabla J(\mathbf{w})^\top = \mathbf{w}^\top\mathbf{A} + \mathbf{w}^\top\mathbf{A}^\top = \mathbf{w}^\top(\mathbf{A}+\mathbf{A}^\top)$$

and

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

iii. The easiest way to calculate the gradient of $J(\mathbf{w}) = \mathbf{w}^\top\mathbf{w}$ is to use the previous question with $\mathbf{A} = \mathbf{I}$ (the identity matrix). Therefore

$$\nabla J(\mathbf{w}) = \mathbf{I}\mathbf{w} + \mathbf{I}^\top\mathbf{w} = \mathbf{w} + \mathbf{w} = 2\mathbf{w}.$$

iv. Note that $||\mathbf{w}||_2 = \sqrt{\mathbf{w}^\top\mathbf{w}}$.

First method: We use the chain rule

$$\frac{\partial J(\mathbf{w})}{\partial w_k} = \frac{\partial\sqrt{\mathbf{w}^\top\mathbf{w}}}{\partial\mathbf{w}^\top\mathbf{w}}\frac{\partial\mathbf{w}^\top\mathbf{w}}{\partial w_k}$$

and that

$$\frac{\partial\sqrt{\mathbf{w}^\top\mathbf{w}}}{\partial\mathbf{w}^\top\mathbf{w}} = \frac{1}{2\sqrt{\mathbf{w}^\top\mathbf{w}}}$$

The derivatives $\partial\mathbf{w}^\top\mathbf{w}/\partial w_k$ were calculated in the question above so that

$$\nabla J(\mathbf{w}) = \frac{1}{2\sqrt{\mathbf{w}^\top\mathbf{w}}}2\mathbf{w} = \frac{\mathbf{w}}{||\mathbf{w}||_2}$$

Second method: Let $f(\mathbf{w}) = \mathbf{w}^\top\mathbf{w}$. From the previous question, we know that

$$f(\mathbf{w}+\epsilon\mathbf{h}) = f(\mathbf{w}) + \epsilon 2\mathbf{w}^\top\mathbf{h} + O(\epsilon^2).$$

Moreover,

$$\begin{aligned} \sqrt{z+\epsilon u + O(\epsilon^2)} &= \sqrt{z} + \frac{1}{2\sqrt{z}}(\epsilon u + O(\epsilon^2)) + O(\epsilon^2) \\ &= \sqrt{z} + \epsilon\frac{1}{2\sqrt{z}}u + O(\epsilon^2) \end{aligned}$$

With $z = f(\mathbf{w})$ and $u = 2\mathbf{w}^\top\mathbf{h}$, we thus obtain

$$\begin{aligned} J(\mathbf{w}+\epsilon\mathbf{h}) &= \sqrt{f(\mathbf{w}+\epsilon\mathbf{h})} \\ &= \sqrt{f(\mathbf{w})} + \epsilon\frac{1}{2\sqrt{f(\mathbf{w})}}2\mathbf{w}^\top\mathbf{h} + O(\epsilon^2) \\ &= \sqrt{f(\mathbf{w})} + \epsilon\frac{\mathbf{w}^\top}{\sqrt{f(\mathbf{w})}}\mathbf{h} + O(\epsilon^2) \\ &= J(\mathbf{w}) + \epsilon\frac{\mathbf{w}^\top}{\sqrt{||\mathbf{w}||_2}}\mathbf{h} + O(\epsilon^2) \end{aligned}$$

so that

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

v. Either the chain rule or the approach with the Taylor expansion can be used to deal with the outer function f. In any case:

$$\nabla J(\mathbf{w}) = f'(||\mathbf{w}||_2)\nabla||\mathbf{w}||_2 = f'(||\mathbf{w}||_2)\frac{\mathbf{w}}{||\mathbf{w}||_2},$$

where f' is the derivative of the function f.

2.2 Newton's method

i. We first write f as

$$\begin{aligned} f(\mathbf{w}) &= c + \mathbf{g}^\top(\mathbf{w}-\mathbf{w}_0) + \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T\mathbf{H}(\mathbf{w}-\mathbf{w}_0) \\ &= c - \mathbf{g}^\top\mathbf{w}_0 + \frac{1}{2}\mathbf{w}_0^\top\mathbf{H}\mathbf{w}_0 + \\ &\quad \mathbf{g}^\top\mathbf{w} + \frac{1}{2}\mathbf{w}^\top\mathbf{H}\mathbf{w} - \frac{1}{2}\mathbf{w}_0^\top\mathbf{H}\mathbf{w} - \frac{1}{2}\mathbf{w}^\top\mathbf{H}\mathbf{w}_0 \end{aligned}$$

Using now that $\mathbf{w}^\top\mathbf{H}\mathbf{w}_0$ is a scalar and that $\mathbf{H}$ is symmetric, we have

$$\mathbf{w}^\top\mathbf{H}\mathbf{w}_0 = (\mathbf{w}^\top\mathbf{H}\mathbf{w}_0)^\top = \mathbf{w}_0^\top\mathbf{H}^\top\mathbf{w} = \mathbf{w}_0^\top\mathbf{H}\mathbf{w}$$

and hence

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

With the results from 2.1 and the fact that $\mathbf{H}$ is symmetric, we thus obtain

$$\begin{aligned} \nabla f(\mathbf{w}) &= \mathbf{g} - \mathbf{H}^\top\mathbf{w}_0 + \frac{1}{2}(\mathbf{H}^\top\mathbf{w} + \mathbf{H}\mathbf{w}) \\ &= \mathbf{g} - \mathbf{H}\mathbf{w}_0 + \mathbf{H}\mathbf{w} \end{aligned}$$

The expansion of $f(\mathbf{w})$ due to the $\mathbf{w}-\mathbf{w}_0$ terms is a bit tedious. It is simpler to note that gradients define a linear approximation of the function. We can more efficiently deal with $\mathbf{w}-\mathbf{w}_0$ by changing the coordinates and determine the linear approximation of f as a function of $\mathbf{v} = \mathbf{w}-\mathbf{w}_0$, i.e. locally around the point $\mathbf{w}_0$. We then have

$$\begin{aligned} \tilde{f}(\mathbf{v}) &= f(\mathbf{v}+\mathbf{w}_0) \\ &= c + \mathbf{g}^\top\mathbf{v} + \frac{1}{2}\mathbf{v}^\top\mathbf{H}\mathbf{v} \end{aligned}$$

With 2.1, the derivative is

$$\nabla_{\mathbf{v}}\tilde{f}(\mathbf{v}) = \mathbf{g} + \mathbf{H}\mathbf{v}$$

and the linear approximation becomes

$$\tilde{f}(\mathbf{v}+\epsilon\mathbf{h}) = c + \epsilon(\mathbf{g}+\mathbf{H}\mathbf{v})^\top\mathbf{h} + O(\epsilon^2)$$

The linear approximation for $\tilde{f}$ determines a linear approximation of f around $\mathbf{w}_0$, i.e.

$$f(\mathbf{w}+\epsilon\mathbf{h}) = \tilde{f}(\mathbf{w}-\mathbf{w}_0+\epsilon\mathbf{h}) = c + \epsilon(\mathbf{g}+\mathbf{H}(\mathbf{w}-\mathbf{w}_0))^\top\mathbf{h} + O(\epsilon^2)$$

so that the derivative for f is

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \mathbf{g} + \mathbf{H}(\mathbf{w}-\mathbf{w}_0) = \mathbf{g} - \mathbf{H}\mathbf{w}_0 + \mathbf{H}\mathbf{w},$$

which is the same result as before.

ii. We set the gradient to zero and solve for $\mathbf{w}$:

$$\mathbf{g} + \mathbf{H}(\mathbf{w} - \mathbf{w}_0) = 0 \quad \leftrightarrow \quad \mathbf{w} - \mathbf{w}_0 = -\mathbf{H}^{-1}\mathbf{g}$$

so that

$$\mathbf{w}^* = \mathbf{w}_0 - \mathbf{H}^{-1}\mathbf{g}.$$

As we assumed that $\mathbf{H}$ is positive definite, the inverse $\mathbf{H}$ exists (and is positive definite too).

Let us consider f as a function of $\mathbf{v}$ around $\mathbf{w}^*$, i.e. $\mathbf{w} = \mathbf{w}^* + \mathbf{v}$. With $\mathbf{w}^* + \mathbf{v} - \mathbf{w}_0 = -\mathbf{H}^{-1}\mathbf{g} + \mathbf{v}$, we have

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

Since $\mathbf{H}$ is positive definite, we have that $(-\mathbf{H}^{-1}\mathbf{g} + \mathbf{v})^\top \mathbf{H}(-\mathbf{H}^{-1}\mathbf{g} + \mathbf{v}) > 0$ for all $\mathbf{v}$. Hence, as we move away from $\mathbf{w}^*$, the function increases quadratically, so that $\mathbf{w}^*$ minimises $f(\mathbf{w})$.

iii. The equation

$$\mathbf{w}^* = \mathbf{w}_0 - \mathbf{H}^{-1}\mathbf{g}.$$

corresponds to one update step in Newton's method where $\mathbf{w}_0$ is the current value of $\mathbf{w}$ in the optimisation of $J(\mathbf{w})$ and $\mathbf{w}^*$ is the updated value. In practice rather than determining the inverse $\mathbf{H}^{-1}$, we solve

$$\mathbf{H}\mathbf{p} = \mathbf{g}$$

for $\mathbf{p}$ and then set $\mathbf{w}^* = \mathbf{w}_0 - \mathbf{p}$. The vector $\mathbf{p}$ is the search direction, and it is possible include a step-length α so that the update becomes $\mathbf{w}^* = \mathbf{w}_0 - \alpha\mathbf{p}$. The value of α may be set by hand or can be determined via line-search methods.

2.3 Gradient of matrix-valued functions

i. First method: With $J(\mathbf{W}) = \sum_{i=1}^{n}\sum_{j=1}^{m} u_i W_{ij} v_j$ we have

$$\frac{\partial J(\mathbf{W})}{W_{kl}} = u_k v_l = (\mathbf{u}\mathbf{v}^\top)_{kl}$$

and hence

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

Second method:

$$\begin{aligned} J(\mathbf{W}+\epsilon\mathbf{H}) &= \mathbf{u}^\top(\mathbf{W}+\epsilon\mathbf{H})\mathbf{v} \\ &= J(\mathbf{W}) + \epsilon\mathbf{u}^\top\mathbf{H}\mathbf{v} \\ &= J(\mathbf{W}) + \epsilon\,\mathrm{tr}(\mathbf{u}^\top\mathbf{H}\mathbf{v}) \\ &= J(\mathbf{W}) + \epsilon\,\mathrm{tr}(\mathbf{v}\mathbf{u}^\top\mathbf{H}) \end{aligned}$$

Hence:

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

ii. Expanding the objective function gives $J(\mathbf{W}) = \mathbf{u}^\top\mathbf{W}\mathbf{v} + \mathbf{u}^\top\mathbf{A}\mathbf{v}$. The second term does not depend on $\mathbf{W}$. With the previous question, the derivative thus is

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

iii. First method:

$$\begin{aligned} \frac{\partial J(\mathbf{W})}{\partial W_{ij}} &= \sum_{k=1}^{n} \frac{\partial}{\partial W_{ij}} f(\mathbf{w}_k^\top\mathbf{v}) \\ &= f'(\mathbf{w}_i^\top\mathbf{v}) \frac{\partial}{\partial W_{ij}} \underbrace{\mathbf{w}_i^\top\mathbf{v}}_{\sum_{j=1}^{m} W_{ij}v_j} \\ &= f'(\mathbf{w}_i^\top\mathbf{v}) v_j \end{aligned}$$

Hence

$$\nabla J(\mathbf{W}) = f'(\mathbf{W}\mathbf{v})\mathbf{v}^\top,$$

where f' operates element-wise on the vector $\mathbf{W}\mathbf{v}$.

Second method:

$$\begin{aligned} J(\mathbf{W}) &= \sum_{k=1}^{n} f(\mathbf{w}_k^\top\mathbf{v}) \\ &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v}), \end{aligned}$$

where $\mathbf{e}^{[k]}$ is the unit row vector that is zero everywhere but for element k which equals one. We now perform a perturbation of $\mathbf{W}$ by $\epsilon\mathbf{H}$.

$$\begin{aligned} J(\mathbf{W}+\epsilon\mathbf{H}) &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}(\mathbf{W}+\epsilon\mathbf{H})\mathbf{v}) \\ &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v}+\epsilon\mathbf{e}^{[k]}\mathbf{H}\mathbf{v}) \\ &= \sum_{k=1}^{n} (f(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v}) + \epsilon f'(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v})\mathbf{e}^{[k]}\mathbf{H}\mathbf{v} + O(\epsilon^2) \\ &= J(\mathbf{W}) + \epsilon \left(\sum_{k=1}^{n} f'(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v})\mathbf{e}^{[k]}\right)\mathbf{H}\mathbf{v} + O(\epsilon^2) \end{aligned}$$

The term $f'(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v})\mathbf{e}^{[k]}$ is a row vector that equals $(0,\ldots,0,f'(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v}),0,\ldots,0)$. Hence, we have

$$\sum_{k=1}^{n} f'(\mathbf{e}^{[k]}\mathbf{W}\mathbf{v})\mathbf{e}^{[k]}) = f'(\mathbf{W}\mathbf{v})^\top$$

where f' operates element-wise on the column vector $\mathbf{W}\mathbf{v}$. The perturbed objective function thus is

$$\begin{aligned} J(\mathbf{W}+\epsilon\mathbf{H}) &= J(\mathbf{W}) + \epsilon f'(\mathbf{W}\mathbf{v})^\top\mathbf{H}\mathbf{v} + O(\epsilon^2) \\ &= J(\mathbf{W}) + \epsilon \operatorname{tr}\left(f'(\mathbf{W}\mathbf{v})^\top\mathbf{H}\mathbf{v}\right) + O(\epsilon^2) \\ &= J(\mathbf{W}) + \epsilon \operatorname{tr}\left(\mathbf{v}f'(\mathbf{W}\mathbf{v})^\top\mathbf{H}\right) + O(\epsilon^2) \end{aligned}$$

Hence, the gradient is the transpose of $\mathbf{v}f'(\mathbf{W}\mathbf{v})^\top$, i.e.

$$\nabla J(\mathbf{W}) = f'(\mathbf{W}\mathbf{v})\mathbf{v}^\top$$

iv. We first verify the hint:

$$\begin{aligned} \left(\mathbf{W}^{-1} - \epsilon\mathbf{W}^{-1}\mathbf{H}\mathbf{W}^{-1} + O(\epsilon^2)\right)(\mathbf{W}+\epsilon\mathbf{H}) &= \mathbf{I} + \epsilon\mathbf{W}^{-1}\mathbf{H} - \epsilon\mathbf{W}^{-1}\mathbf{H} + O(\epsilon^2) \\ &= \mathbf{I} + O(\epsilon^2) \end{aligned}$$

Hence the identity holds up to terms smaller than ϵ^2, which is sufficient we do not care about terms of order ϵ^2 and smaller in the definition of the gradient given by

$$J(\mathbf{W}+\epsilon\mathbf{H}) = J(\mathbf{W}) + \epsilon\operatorname{tr}(\nabla J^\top\mathbf{H}) + O(\epsilon^2) \tag{2.1}$$

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

. Let us thus make a first-order approximation of the perturbed objective $J(\mathbf{W}+\epsilon\mathbf{H})$:

$$\begin{aligned} J(\mathbf{W}+\epsilon\mathbf{H}) &= \mathbf{u}^\top(\mathbf{W}+\epsilon\mathbf{H})^{-1}\mathbf{v} \\ &\overset{\text{hint}}{=} \mathbf{u}^\top(\mathbf{W}^{-1} - \epsilon\mathbf{W}^{-1}\mathbf{H}\mathbf{W}^{-1} + O(\epsilon^2))\mathbf{v} \\ &= \mathbf{u}^\top\mathbf{W}^{-1}\mathbf{v} - \epsilon\mathbf{u}^\top\mathbf{W}^{-1}\mathbf{H}\mathbf{W}^{-1}\mathbf{v} + O(\epsilon^2) \\ &= J(\mathbf{W}) - \epsilon\operatorname{tr}\left(\mathbf{u}^\top\mathbf{W}^{-1}\mathbf{H}\mathbf{W}^{-1}\mathbf{v}\right) + O(\epsilon^2) \\ &= J(\mathbf{W}) - \epsilon\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{v}\mathbf{u}^\top\mathbf{W}^{-1}\mathbf{H}\right) + O(\epsilon^2) \end{aligned}$$

Comparison with (2.1) gives

$$\nabla J^\top = -\mathbf{W}^{-1}\mathbf{v}\mathbf{u}^\top\mathbf{W}^{-1}$$

and hence

$$\nabla J = -\mathbf{W}^{-\top}\mathbf{u}\mathbf{v}^\top\mathbf{W}^{-\top},$$

where $\mathbf{W}^{-\top}$ is the transpose of the inverse of $\mathbf{W}$.

2.4 Gradient of the log-determinant

i. As in Task 1.3, let $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ be the eigenvalue decomposition of $\mathbf{W}$ (with $\mathbf{V}^\top = \mathbf{U}^{-1}$). Then $\mathbf{\Lambda} = \mathbf{V}^\top\mathbf{W}\mathbf{U}$ and

$$\begin{aligned}\lambda_n &= \mathbf{e}^{[n]}\mathbf{\Lambda}\mathbf{e}^{(n)} \\ &= \mathbf{e}^{[n]}\mathbf{V}^\top\mathbf{W}\mathbf{U}\mathbf{e}^{(n)} \\ &= (\mathbf{V}\mathbf{e}^{(n)})^\top\mathbf{W}\mathbf{U}\mathbf{e}^{(n)} \\ &= \mathbf{v}_n^\top\mathbf{W}\mathbf{u}_n,\end{aligned}$$

where $\mathbf{e}^{(n)}$ is the standard basis (unit) vector with a 1 in the n-th slot and zeros elsewhere, and $\mathbf{e}^{[n]}$ is the corresponding row vector.

ii. With 2.3, we have

$$\nabla_{\mathbf{W}}\lambda_n(\mathbf{W}) = \nabla_{\mathbf{W}}\mathbf{v}_n^\top\mathbf{W}\mathbf{u}_n = \mathbf{v}_n\mathbf{u}_n^\top.$$

iii. In Task 1.4, we have shown that $\det(\mathbf{W}) = \prod_i \lambda_i$ and hence $|\det(W)| = \prod_i |\lambda_i|$.

(i) If $\mathbf{W}$ is positive definite, its eigenvalues are positive and we can drop the absolute values so that $|\det(W)| = \prod_i \lambda_i$.

(ii) If $\mathbf{W}$ is a matrix with real entries, then $\mathbf{W}\mathbf{u} = \lambda\mathbf{u}$ implies $\mathbf{W}\bar{\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}$, i.e. if λ is a complex eigenvalue, then $\bar{\lambda}$ (the complex conjugate of λ) is also an eigenvalue. Since $|\lambda|^2 = \lambda\bar{\lambda}$,

$$|\det(\mathbf{W})| = \left(\prod_{\lambda_i\in\mathbb{C}} \lambda_i\right)\left(\prod_{\lambda_j\in\mathbb{R}} |\lambda_j|\right).$$

Now we can write $J(\mathbf{W})$ in terms of the eigenvalues:

$$\begin{aligned}J(\mathbf{W}) &= \log|\det(\mathbf{W})| \\ &= \log\left(\prod_{\lambda_i\in\mathbb{C}} \lambda_i\right)\left(\prod_{\lambda_j\in\mathbb{R}} |\lambda_j|\right) \\ &= \log\left(\prod_{\lambda_i\in\mathbb{C}} \lambda_i\right) + \log\left(\prod_{\lambda_j\in\mathbb{R}} |\lambda_j|\right) \\ &= \sum_{\lambda_i\in\mathbb{C}} \log\lambda_i + \sum_{\lambda_j\in\mathbb{R}} \log|\lambda_j|.\end{aligned}$$

Assume that the real-valued λ_j are non-zero so that

$$\begin{aligned}\nabla_{\mathbf{W}}\log|\lambda_j| &= \frac{1}{|\lambda_j|}\nabla_{\mathbf{W}}|\lambda_j| \\ &= \frac{1}{|\lambda_j|}\text{sign}(\lambda_j)\nabla_{\mathbf{W}}\lambda_j\end{aligned}$$

Hence

$$\begin{aligned}\nabla J(\mathbf{W}) &= \sum_{\lambda_i\in\mathbb{C}} \nabla_{\mathbf{W}} \log \lambda_i + \sum_{\lambda_j\in\mathbb{R}} \nabla_{\mathbf{W}} \log |\lambda_j| \\ &= \sum_{\lambda_i\in\mathbb{C}} \frac{1}{\lambda_i}\nabla_{\mathbf{W}}\lambda_i + \sum_{\lambda_i\in\mathbb{R}} \frac{1}{|\lambda_i|}\mathrm{sign}(\lambda_i)\nabla_{\mathbf{W}}\lambda_i \\ &= \sum_{\lambda_i\in\mathbb{C}} \frac{\mathbf{v}_i\mathbf{u}_i^\top}{\lambda_i} + \sum_{\lambda_i\in\mathbb{R}} \frac{\mathrm{sign}(\lambda_i)\mathbf{v}_i\mathbf{u}_i^\top}{|\lambda_i|} \\ &= \sum_{\lambda_i\in\mathbb{C}} \frac{\mathbf{v}_i\mathbf{u}_i^\top}{\lambda_i} + \sum_{\lambda_i\in\mathbb{R}} \frac{\mathbf{v}_i\mathbf{u}_i^\top}{\lambda_i} \\ &= \sum_i \frac{\mathbf{v}_i\mathbf{u}_i^\top}{\lambda_i}.\end{aligned}$$

iv. This follows from Task 1.3 where we have found that

$$\mathbf{W}^{-1} = \sum_i \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top.$$

Indeed:

$$\nabla J(\mathbf{W}) = \sum_i \frac{\mathbf{v}_i\mathbf{u}_i^\top}{\lambda_i} = \sum_i \frac{1}{\lambda_i}(\mathbf{u}_i\mathbf{v}_i^\top)^\top = \left(\mathbf{W}^{-1}\right)^\top.$$

2.5 Descent directions for matrix-valued functions

As in the introduction to the question, we appeal to (2.1) to obtain

$$\begin{aligned}J(\mathbf{W} - \epsilon\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}\mathbf{B}^\top) &= J(\mathbf{W}) - \epsilon\,\mathrm{tr}(\nabla J(\mathbf{W})^\top\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}\mathbf{B}^\top) + O(\epsilon^2) \\ &= J(\mathbf{W}) - \epsilon\,\mathrm{tr}(\mathbf{B}^\top\nabla J(\mathbf{W})^\top\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}) + O(\epsilon^2),\end{aligned}$$

where $\mathrm{tr}(\mathbf{B}^\top\nabla J(\mathbf{W})^\top\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B})$ takes the form $\mathrm{tr}(\mathbf{M}^\top\mathbf{M})$ with $\mathbf{M} = \mathbf{A}\nabla J(\mathbf{W})\mathbf{B}$. With

$$\mathrm{tr}(\mathbf{M}^\top\mathbf{M}) > 0,$$

we thus have $\mathrm{tr}(\mathbf{B}^\top\nabla J(\mathbf{W})^\top\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}) > 0$ if $\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}$ is non-zero, and hence

$$J(\mathbf{W} - \epsilon\mathbf{A}^\top\mathbf{A}\nabla J(\mathbf{W})\mathbf{B}\mathbf{B}^\top) < J(\mathbf{W})$$

for small enough ϵ. We have equality if $\mathbf{A}\nabla J(\mathbf{W})\mathbf{B} = 0$, e.g. if the columns of $\mathbf{B}$ are all in the null space of ∇J.