Optimization Notes

Sepand Kashani
Audiovisual Communications Laboratory (LCAV)
School of Computer and Communication Sciences (IC)
École Polytechnique Fédérale de Lausanne (EPFL)
sepand.kashani@epfl.ch

Abstract. While optimization is well studied for real-valued functions $f : \mathbb{R}^N \to \mathbb{R}$, many physical problems are (partially) specified in terms of complex-valued functions $f_c : \mathbb{C}^N \to \mathbb{C}^M$. Current optimization packages have limited support for such functions. In particular it is unclear how to define algorithmic differentiation w.r.t. complex-valued functions and arguments. This document is a collection of working notes on the topic.

Key words. First-order Methods, Algorithmic Differentiation

1. Preliminaries.

1.1. Conventions.

Throughout this document, we adopt the following conventions:

- Vectors are denoted with bold lowercase letters: $\mathbf{y}$.
- Matrices are denoted with bold uppercase letters: $\mathbf{A}$.
- If $\mathbf{A} \in \mathbb{C}^{M \times N}$, $\mathbf{a}_k \in \mathbb{C}^M$ denotes the k-th column of $\mathbf{A}$.
- The i-th entry of vector $\mathbf{y}$ is denoted $[\mathbf{y}]_i$.
- The (i, j)-th entry of matrix $\mathbf{A}$ is denoted $[\mathbf{A}]_{ij}$.
- The conjugation operator is denoted by overlining a vector or a matrix respectively: $\overline{\mathbf{a}}$, $\overline{\mathbf{A}}$.
- The modulus of a complex number $z \in \mathbb{C}$ is denoted by $|z|$.
- The real/imaginary parts of matrix $\mathbf{A}$ are denoted $\Re\{\mathbf{A}\}$, $\Im\{\mathbf{A}\}$, or $\mathbf{A}_R$, $\mathbf{A}_I$, respectively.

1.2. Hadamard, Kronecker and Khatri-Rao products.

The Hadamard product is the element-wise multiplication operator:

Definition 1.1 (Hadamard product). *Let* $\mathbf{A} \in \mathbb{C}^{M \times N}$ *and* $\mathbf{B} \in \mathbb{C}^{M \times N}$. *The* Hadamard product $\mathbf{A} \odot \mathbf{B} \in \mathbb{C}^{M \times N}$ *is defined as*

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} [\mathbf{B}]_{ij} .$$

Moreover, we denote by $\mathbf{A}^{\odot k}$ *the product sequence* $\underbrace{\mathbf{A} \odot \cdots \odot \mathbf{A}}_{k\times}$.

The Kronecker product generalises the vector outer product to matrices, and represents the tensor product between two finite-dimensional linear maps:

Definition 1.2 (Kronecker product). *Let* $\mathbf{A} \in \mathbb{C}^{M_1 \times N_1}$ *and* $\mathbf{B} \in \mathbb{C}^{M_2 \times N_2}$. *The* Kronecker product $\mathbf{A} \otimes \mathbf{B} \in \mathbb{C}^{M_1 M_2 \times N_1 N_2}$ *is defined as*

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} [\mathbf{A}]_{11}\,\mathbf{B} & \cdots & [\mathbf{A}]_{1N_1}\,\mathbf{B} \\ \vdots & \ddots & \vdots \\ [\mathbf{A}]_{M_1 1}\,\mathbf{B} & \cdots & [\mathbf{A}]_{M_1 N_1}\,\mathbf{B} \end{bmatrix}.$$

The main properties of the Kronecker product are [3]:

$$(\mathbf{A} \otimes \mathbf{B})^H = \mathbf{A}^H \otimes \mathbf{B}^H, \tag{1.1}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}), \tag{1.2}$$

$$(\mathbf{A} \otimes \mathbf{B}) \odot (\mathbf{C} \otimes \mathbf{D}) = (\mathbf{A} \odot \mathbf{C}) \otimes (\mathbf{B} \odot \mathbf{D}). \tag{1.3}$$

The Khatri-Rao product finally, is a column-wise Kronecker product:

Definition 1.3 (Khatri-Rao product). *Let* $\mathbf{A} \in \mathbb{C}^{M_1 \times N}$ *and* $\mathbf{B} \in \mathbb{C}^{M_2 \times N}$. *The* Khatri-Rao product $\mathbf{A} \circ \mathbf{B} \in \mathbb{C}^{M_1 M_2 \times N}$ *is defined as*

$$\mathbf{A} \circ \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \ldots, \mathbf{a}_N \otimes \mathbf{b}_N].$$

1.3. Matrix identities.

$\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \circ \mathbf{B}$ are often too large to be stored in memory. However it is not the matrix itself that is of interest in many circumstances, but rather the effect of a linear map such as $f(\mathbf{x}) = (\mathbf{A} \otimes \mathbf{B})\mathbf{x}$. The matrix identities below allow us to evaluate $f(\mathbf{x})$ without ever having to compute large intermediate arrays. They make use of the vectorisation operator:

Definition 1.4 (Vectorisation). *Let* $\mathbf{A} \in \mathbb{C}^{M \times N}$. *The* vectorisation operator $\operatorname{vec}(\cdot)$ *reshapes a matrix into a vector by stacking its columns:*

$$[\operatorname{vec}(\mathbf{A})]_{M(j-1)+i} = [\mathbf{A}]_{ij}.$$

Conversely, the matricisation operator $\operatorname{mat}_{M,N}(\cdot)$ *reshapes a vector into a matrix:*

$$[\operatorname{mat}_{M,N}(\mathbf{a})]_{ij} = [\mathbf{a}]_{M(j-1)+i}.$$

Commonly used matrix identities are the following [2, 5]:

$$\operatorname{vec}(\mathbf{ABC}) = \left(\mathbf{C}^T \otimes \mathbf{A}\right) \operatorname{vec}(\mathbf{B}) \tag{1.4}$$

$$\operatorname{vec}(\mathbf{A}\operatorname{diag}(\mathbf{b})\mathbf{C}) = \left(\mathbf{C}^T \circ \mathbf{A}\right) \mathbf{b} \tag{1.5}$$

$$\langle \mathbf{A}, \mathbf{B} \rangle_F = \operatorname{tr}\left(\mathbf{A}^H \mathbf{B}\right) = \operatorname{vec}(\mathbf{A})^H \operatorname{vec}(\mathbf{B}) \tag{1.6}$$

$$\operatorname{vec}(\mathbf{b}\mathbf{a}^T) = \mathbf{a} \otimes \mathbf{b} \tag{1.7}$$

The following nonstandard matrix identities are proved in Appendix A:

$$(\mathbf{A} \circ \mathbf{B})^H \operatorname{vec}(\mathbf{C}) = \operatorname{diag}\left(\mathbf{B}^H \mathbf{C} \overline{\mathbf{A}}\right) \tag{1.8}$$

$$(\mathbf{A} \otimes \mathbf{B})^H (\mathbf{C} \otimes \mathbf{D}) \operatorname{vec}(\mathbf{E}) = \operatorname{vec}(\mathbf{B}^H \mathbf{D} \mathbf{E} \mathbf{C}^T \overline{\mathbf{A}}) \tag{1.9}$$

$$(\mathbf{A} \circ \mathbf{B})^H (\mathbf{C} \circ \mathbf{D}) \, \mathbf{e} = \operatorname{diag}\left(\mathbf{B}^H \mathbf{D} \operatorname{diag}(\mathbf{e}) \mathbf{C}^T \overline{\mathbf{A}}\right) \tag{1.10}$$

$$(\mathbf{A} \circ \mathbf{B})^H (\mathbf{C} \circ \mathbf{D}) = \mathbf{A}^H \mathbf{C} \odot \mathbf{B}^H \mathbf{D} \tag{1.11}$$

2. Algorithmic Differentiation.

Algorithmic differentiation (AD) [1] is an efficient procedure to evaluate *numerical* derivatives of mathematical expressions using a few symbolic building blocks in conjuction with the chain rule.

Definition 2.1 (Jacobian matrix). *Let $f : \mathbb{R}^N \to \mathbb{R}^M$. The* Jacobian matrix $\mathbf{D}_f \in \mathbb{R}^{M \times N}$ *is*

$$\mathbf{D}_f = \begin{bmatrix} \frac{\partial [f]_1}{\partial [\mathbf{x}]_1} & \cdots & \frac{\partial [f]_1}{\partial [\mathbf{x}]_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial [f]_M}{\partial [\mathbf{x}]_1} & \cdots & \frac{\partial [f]_M}{\partial [\mathbf{x}]_N} \end{bmatrix}.$$

Definition 2.2 (Chain rule (real case)). *Let $f : \mathbb{R}^N \to \mathbb{R}^M$, $g : \mathbb{R}^M \to \mathbb{R}^P$ and $h = g \circ f$. Then*

$$\mathbf{D}_h(\mathbf{x}) = \mathbf{D}_g(\mathbf{f}) \, \mathbf{D}_f(\mathbf{x}) \in \mathbb{R}^{P \times N},$$

with $\mathbf{f} = f(\mathbf{x}) \in \mathbb{R}^M$.

Example 2.3. Let $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 = (\gamma \circ \beta \circ \alpha)(\mathbf{x})$, with

$$\begin{aligned} \alpha : \mathbb{R}^N &\to \mathbb{R}^M & \beta : \mathbb{R}^M &\to \mathbb{R}^M & \gamma : \mathbb{R}^M &\to \mathbb{R} \\ \mathbf{x} &\to \mathbf{A}\mathbf{x} & \mathbf{a} &\to \mathbf{y} - \mathbf{a} & \mathbf{b} &\to \|\mathbf{b}\|_2^2 \end{aligned}$$

Then $\nabla_{\mathbf{x}} f \in \mathbb{R}^{1 \times N}$ is given by

$$\begin{aligned} \nabla_{\mathbf{x}} f = \mathbf{D}_f(\mathbf{x}) &= \mathbf{D}_\gamma(\mathbf{b}) \, \mathbf{D}_\beta(\mathbf{a}) \, \mathbf{D}_\alpha(\mathbf{x}) \\ &= (2\mathbf{b}^T) \, (-I_M) \, \mathbf{A} \\ &= -2\mathbf{b}^T \mathbf{A}, \end{aligned}$$

where $\mathbf{a} = \alpha(\mathbf{x}) \in \mathbb{R}^M$ and $\mathbf{b} = \beta(\mathbf{a}) \in \mathbb{R}^M$.

While well developed for real-valued functions $f : \mathbb{R}^N \to \mathbb{R}^M$, generalization of Definition 2.2 to complex-valued functions $f : \mathbb{C}^N \to \mathbb{C}^M$ is not straightforward. The generalization makes use of the hat operator:

Definition 2.4 (Hat operator). *Let $f : \mathbb{C}^N \to \mathbb{C}^M$. The* hat operator $\widehat{\ }$ *maps f onto its counterpart $\hat{f}$ expressed solely in terms of real-valued expressions:*

$$f : \mathbb{C}^N \to \mathbb{C}^M \qquad \hat{f} : \mathbb{R}^{2N} \to \mathbb{R}^{2M}$$

$$\mathbf{x}_R + j\mathbf{x}_I \to f_R(\mathbf{x}_R + j\mathbf{x}_I) + j f_I(\mathbf{x}_R + j\mathbf{x}_I) \qquad \begin{bmatrix} \mathbf{x}_R \\ \mathbf{x}_I \end{bmatrix} \to \begin{bmatrix} f_R(\mathbf{x}_R, \mathbf{x}_I) \\ f_I(\mathbf{x}_R, \mathbf{x}_I) \end{bmatrix}$$

Example 2.5 (Linear map).

$$f : \mathbb{C}^N \to \mathbb{C}^M \qquad \hat{f} : \mathbb{R}^{2N} \to \mathbb{R}^{2M}$$

$$\mathbf{x}_R + j\mathbf{x}_I \to \mathbf{A}\mathbf{x} \qquad \begin{bmatrix} \mathbf{x}_R \\ \mathbf{x}_I \end{bmatrix} \to \begin{bmatrix} \mathbf{A}_R\mathbf{x}_R - \mathbf{A}_I\mathbf{x}_I \\ \mathbf{A}_R\mathbf{x}_I + \mathbf{A}_I\mathbf{x}_R \end{bmatrix}$$

See [4] for some useful properties of the hat operator.

Definition 2.6 (Chain rule (complex case)). *Let $f : \mathbb{C}^N \to \mathbb{C}^M$, $g : \mathbb{C}^M \to \mathbb{C}^P$ and $h = g \circ f$. Then*

$$\mathbf{D}_{\hat{h}}(\hat{\mathbf{x}}) = \mathbf{D}_{\hat{g}}(\hat{\mathbf{f}})\, \mathbf{D}_{\hat{f}}(\hat{\mathbf{x}}) \in \mathbb{R}^{2P \times 2N}, \quad \textit{with}$$

$$\mathbf{D}_{\hat{f}}(\hat{\mathbf{x}}) = \begin{bmatrix} \frac{\partial f_R}{\partial \mathbf{x}_R}(\mathbf{x}_R, \mathbf{x}_I) & \frac{\partial f_R}{\partial \mathbf{x}_I}(\mathbf{x}_R, \mathbf{x}_I) \\ \frac{\partial f_I}{\partial \mathbf{x}_R}(\mathbf{x}_R, \mathbf{x}_I) & \frac{\partial f_I}{\partial \mathbf{x}_I}(\mathbf{x}_R, \mathbf{x}_I) \end{bmatrix},$$

where $\mathbf{f} = f(\mathbf{x}) \in \mathbb{C}^M$.

Note that the chain rule is only defined in terms of $\hat{f}$. In particular, it is generally *not* possible to “unhat” $\mathbf{D}_{\hat{f}} : \mathbb{R}^{2M} \to \mathbb{R}^{2N}$. However, in the special case of functions $f : \mathbb{C}^N \to \mathbb{R}^M$, the short-hand complex-valued quantity $\mathbf{D}_f(\mathbf{x}_R + j\mathbf{x}_I) = \frac{\partial f}{\partial \mathbf{x}_R}(\mathbf{x}) + j\frac{\partial f}{\partial \mathbf{x}_I}(\mathbf{x})$ is sometimes useful.

Example 2.7. Let $f(\mathbf{x}) = \mathbf{1}^T(\mathbf{y} - \mathbf{A}\mathbf{x}) = (\gamma \circ \beta \circ \alpha)(\mathbf{x})$, with

$$\alpha : \mathbb{C}^N \to \mathbb{C}^M \qquad \beta : \mathbb{C}^M \to \mathbb{C}^M \qquad \gamma : \mathbb{C}^M \to \mathbb{C}$$

$$\mathbf{x} \to \mathbf{A}\mathbf{x} \qquad \mathbf{a} \to \mathbf{y} - \mathbf{a} \qquad \mathbf{b} \to \mathbf{1}^T\mathbf{b}$$

Then $\nabla_{\hat{\mathbf{x}}} \hat{f} \in \mathbb{R}^{2 \times 2N}$ is given by

$$\begin{aligned} \nabla_{\hat{\mathbf{x}}} \hat{f} = \mathbf{D}_{\hat{f}}(\hat{\mathbf{x}}) &= \mathbf{D}_{\hat{\gamma}}(\hat{\mathbf{b}})\, \mathbf{D}_{\hat{\beta}}(\hat{\mathbf{a}})\, \mathbf{D}_{\hat{\alpha}}(\hat{\mathbf{x}}) \\ &= \begin{bmatrix} \mathbf{1}_M^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_M^T \end{bmatrix} \begin{bmatrix} -I_M & \mathbf{0} \\ \mathbf{0} & -I_M \end{bmatrix} \begin{bmatrix} \mathbf{A}_R & -\mathbf{A}_I \\ \mathbf{A}_I & \mathbf{A}_R \end{bmatrix} \\ &= -\begin{bmatrix} \mathbf{1}_M^T\mathbf{A}_R & -\mathbf{1}_M^T\mathbf{A}_I \\ \mathbf{1}_M^T\mathbf{A}_I & \mathbf{1}_M^T\mathbf{A}_R \end{bmatrix}, \end{aligned}$$

where $\mathbf{a} = \alpha(\mathbf{x}) \in \mathbb{C}^M$ and $\mathbf{b} = \beta(\mathbf{a}) \in \mathbb{C}^M$. This expression *cannot* be further reduced to obtain a valid expression for $\nabla_{\mathbf{x}} f$.

Example 2.8. Let $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 = (\delta \circ \beta \circ \alpha)(\mathbf{x})$, with

$$\alpha : \mathbb{C}^N \to \mathbb{C}^M \qquad \beta : \mathbb{C}^M \to \mathbb{C}^M \qquad \delta : \mathbb{C}^M \to \mathbb{R}$$

$$\mathbf{x} \to \mathbf{A}\mathbf{x} \qquad \mathbf{a} \to \mathbf{y} - \mathbf{a} \qquad \mathbf{b} \to \|\mathbf{b}\|_2^2$$

$f : \mathbb{C}^N \to \mathbb{C}^M$	$\mathbf{D}_{\hat{f}} : \mathbb{R}^{2M} \to \mathbb{R}^{2N}$
$\alpha\mathbf{x},\ \alpha \in \mathbb{C}$	$\begin{bmatrix} \alpha_R & -\alpha_I \\ \alpha_I & \alpha_R \end{bmatrix} \otimes I_N$
$\mathbf{x} + \mathbf{y},\ \mathbf{y} \in \mathbb{C}^N$	$I_2 \otimes I_N$
$\overline{\mathbf{x}}$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes I_N$
$\mathbf{A}\mathbf{x},\ \mathbf{A} \in \mathbb{C}^{M \times N}$	$\begin{bmatrix} \mathbf{A}_R & -\mathbf{A}_I \\ \mathbf{A}_I & \mathbf{A}_R \end{bmatrix}$
$\mathbf{a} \odot \mathbf{x},\ \mathbf{a} \in \mathbb{C}^N$	$\begin{bmatrix} \operatorname{diag}(\mathbf{a}_R) & -\operatorname{diag}(\mathbf{a}_I) \\ \operatorname{diag}(\mathbf{a}_I) & \operatorname{diag}(\mathbf{a}_R) \end{bmatrix}$
$\mathbf{a} \otimes \mathbf{x},\ \mathbf{a} \in \mathbb{C}^K$	$\begin{bmatrix} \mathbf{a}_R \otimes I_N & -\mathbf{a}_I \otimes I_N \\ \mathbf{a}_I \otimes I_N & \mathbf{a}_R \otimes I_N \end{bmatrix}$
$\mathbf{x} \otimes \mathbf{a},\ \mathbf{a} \in \mathbb{C}^K$	$\begin{bmatrix} I_N \otimes \mathbf{a}_R & I_N \otimes -\mathbf{a}_I \\ I_N \otimes \mathbf{a}_I & I_N \otimes \mathbf{a}_R \end{bmatrix}$

Table 1

Jacobian matrices of commonly-used operators in optimization. These can be chained using Definition 2.6 to evaluate numerical gradients of arbitarily-complex functions.

Then $\nabla_{\hat{\mathbf{x}}} \hat{f} \in \mathbb{R}^{2 \times 2N}$ is given by

$$\begin{aligned} \nabla_{\hat{\mathbf{x}}} \hat{f} = \mathbf{D}_{\hat{f}}(\hat{\mathbf{x}}) &= \mathbf{D}_{\hat{\delta}}(\hat{\mathbf{b}})\ \mathbf{D}_{\hat{\beta}}(\hat{\mathbf{a}})\ \mathbf{D}_{\hat{\alpha}}(\hat{\mathbf{x}}) \\ &= \begin{bmatrix} 2\mathbf{b}_R^T & 2\mathbf{b}_I^T \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} -I_M & \mathbf{0} \\ \mathbf{0} & -I_M \end{bmatrix} \begin{bmatrix} \mathbf{A}_R & -\mathbf{A}_I \\ \mathbf{A}_I & \mathbf{A}_R \end{bmatrix} \\ &= -2 \begin{bmatrix} \Re\left\{\mathbf{b}^T \overline{\mathbf{A}}\right\} & \Im\left\{\mathbf{b}^T \overline{\mathbf{A}}\right\} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \end{aligned}$$

where $\mathbf{a} = \alpha(\mathbf{x}) \in \mathbb{C}^M$ and $\mathbf{b} = \beta(\mathbf{a}) \in \mathbb{C}^M$. This expression *can* be further reduced to obtain a valid expression for $\nabla_{\mathbf{x}} f = \mathbf{D}_f(\mathbf{x}) = -2\mathbf{b}^T \overline{\mathbf{A}} \in \mathbb{C}^{1 \times N}$.

Remark 2.9 (Implementation note). Since optimization algorithms require (sums of) loss functions of the form $f : \mathbb{C}^N \to \mathbb{R}$, in practice we will always be able to express gradients using the shorthand form $\nabla_{\mathbf{x}} f \in \mathbb{C}^{1 \times N}$ after applying Definition 2.6.

Table 1 provides symbolic closed-form expressions for most common operators encountered in optimization.

Appendix A. Proofs.

Proof. (1.8)

$$
\begin{aligned}
\left[(\mathbf{A}\circ\mathbf{B})^H \operatorname{vec}(\mathbf{C})\right]_i &= \langle(\mathbf{A}\circ\mathbf{B})_i, \operatorname{vec}(\mathbf{C})\rangle = (\mathbf{a}_i\otimes\mathbf{b}_i)^H \operatorname{vec}(\mathbf{C}) \\
&\overset{(1.7)}{=} \operatorname{vec}(\mathbf{b}_i\mathbf{a}_i^T)^H \operatorname{vec}(\mathbf{C}) \overset{(1.6)}{=} \operatorname{tr}\left(\overline{\mathbf{a}}_i\mathbf{b}_i^H\mathbf{C}\right) \\
&= \operatorname{tr}\left(\mathbf{b}_i^H\mathbf{C}\overline{\mathbf{a}}_i\right) = \left[\mathbf{B}^H\mathbf{C}\overline{\mathbf{A}}\right]_{ii} = \left[\operatorname{diag}\left(\mathbf{B}^H\mathbf{C}\overline{\mathbf{A}}\right)\right]_i
\end{aligned}
$$

∎

Proof. (1.9)

$$
\begin{aligned}
(\mathbf{A}\otimes\mathbf{B})^H(\mathbf{C}\otimes\mathbf{D})\operatorname{vec}(\mathbf{E}) &\overset{(1.1)}{=} \left(\mathbf{A}^H\otimes\mathbf{B}^H\right)(\mathbf{C}\otimes\mathbf{D})\operatorname{vec}(\mathbf{E}) \\
&\overset{(1.2)}{=} \left[\left(\mathbf{A}^H\mathbf{C}\right)\otimes\left(\mathbf{B}^H\mathbf{D}\right)\right]\operatorname{vec}(\mathbf{E}) \\
&\overset{(1.4)}{=} \operatorname{vec}(\mathbf{B}^H\mathbf{D}\mathbf{E}\mathbf{C}^T\overline{\mathbf{A}})
\end{aligned}
$$

∎

Proof. (1.10)

$$
\begin{aligned}
(\mathbf{A}\circ\mathbf{B})^H(\mathbf{C}\circ\mathbf{D})\,\mathbf{e} &\overset{(1.5)}{=} (\mathbf{A}\circ\mathbf{B})^H\operatorname{vec}\left(\mathbf{D}\operatorname{diag}(\mathbf{e})\mathbf{C}^T\right) \\
&\overset{(1.8)}{=} \operatorname{diag}\left(\mathbf{B}^H\mathbf{D}\operatorname{diag}(\mathbf{e})\mathbf{C}^T\overline{\mathbf{A}}\right)
\end{aligned}
$$

∎

Proof. (1.11)

$$
\begin{aligned}
\left[(\mathbf{A}\circ\mathbf{B})^H(\mathbf{C}\circ\mathbf{D})\right]_{ij} &= \langle\mathbf{a}_i\otimes\mathbf{b}_i, \mathbf{c}_j\otimes\mathbf{d}_j\rangle \overset{(1.7)}{=} \left\langle\operatorname{vec}(\mathbf{b}_i\mathbf{a}_i^T), \operatorname{vec}(\mathbf{d}_j\mathbf{c}_j^T)\right\rangle \\
&\overset{(1.6)}{=} \operatorname{tr}\left(\overline{\mathbf{a}_i}\mathbf{b}_i^H\mathbf{d}_j\mathbf{c}_j^T\right) = \operatorname{tr}\left(\mathbf{b}_i^H\mathbf{d}_j\mathbf{c}_j^T\overline{\mathbf{a}_i}\right) \\
&= \langle\mathbf{b}_i, \mathbf{d}_j\rangle\langle\mathbf{a}_i, \mathbf{c}_j\rangle\,.
\end{aligned}
$$

When put in matrix form, the above yields

$$
(\mathbf{A}\circ\mathbf{B})^H(\mathbf{C}\circ\mathbf{D}) = \mathbf{A}^H\mathbf{C}\odot\mathbf{B}^H\mathbf{D}.
$$

∎

REFERENCES

[1] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, *Automatic differentiation in machine learning: a survey*, Journal of machine learning research, 18 (2018).

[2] K. Jinadasa, *Applications of the matrix operators vech and vec*, Linear Algebra and its Applications, 101 (1988), pp. 73–79.

[3] S. Liu and G. Trenkler, *Hadamard, khatri-rao, kronecker and other matrix products*, International Journal of Information and Systems Sciences, 4 (2008), pp. 160–177.

[4] B. Rimoldi, *Principles of Digital Communication: A Top-Down Approach*, Cambridge University Press, 2016.

[5] A.-J. van der Veen and S. J. Wijnholds, *Signal processing tools for radio astronomy*, in Handbook of Signal Processing Systems, Springer, 2013, pp. 421–463.